Politecnico di Torino

Dipartimento di Scienza Applicata e Tecnologia

Politecnico di Torino

Master Degree in Physics of Complex Systems

Thesis

Model-data fusion of chlorophyll fluorescence for reducing uncertainties in local-scale simulations of plant photosynthesis and transpiration

Advisor:
Prof. Alessandro Pelizzola

Co-Advisors:
Dr. Fabienne Maignan
Dr. Camille Abadie

Candidate:
Lorenzo Francesco Davoli

Session March-April 2023

Academic Year 2021–2022

Abstract

Over the last few years, more and more climate and environmental datasets have been produced and collected into structured databases, allowing Earth-System Models (ESMs) to exploit new local and global observations to increase their predictive power. A key aspect of terrestrial biosphere models is the description of water and carbon cycles, where vegetation is one of the main agents in carbon exchange and water transport between the surface and the atmosphere. Therefore, an accurate description of vegetation dynamics is required in every model aiming to properly represent the land-surface fluxes of these quantities. Photosynthesis plays a crucial role in this scope, since it is responsible for carbon assimilation in the plant (from atmospheric CO_2 absorption) and the release of water vapour through stomatal pores in the ambient air, the so-called plant transpiration, a by-product of the photosynthetic reaction. These phenomena occur at a microscopical scale inside or at the surface of the leaves, but they have important impacts also at larger scales and therefore they are usually collectively observed at stand or canopy level by using eddy covariance methods for the measurement of water vapour and CO_2 fluxes emerging from the canopy. The estimates of the water flux computed with this technique account for the total evapotranspiration process, including both evaporation and transpiration. The separation of this flux into evaporation and transpiration is affected by significant uncertainties. As a consequence, the research is focused on finding new methods or proxies which could provide measurements of plant transpiration fully unravelled from evaporative processes.

Two candidate databases for this role (SAPFLUXNET sap flow local measurements and TROPOMI Sun-Induced chlorophyll Fluorescence satellite observations) have been investigated in the present research to assess the potential impact of their data assimilation within the Organising Carbon and Hydrology In Dynamic Ecosystems (ORCHIDEE) land-surface model. 6 sites have been chosen as study cases, because they are the only ones covered for more than 3 years by both SAPFLUXNET and FLUXNET2015, the latter being the database providing the meteorological forcing data needed by ORCHIDEE for its local simulations.

SAPFLUXNET is a dataset containing tree sap flow observations collected over hundreds of measurement sites around the world. Sap flow in the plant stem is a very good proxy of the plant transpiration rate, in fact the great majority of the water absorbed by the plant from the soil is released through transpiration. An integration method has been devised to derive transpiration observations at canopy level starting from sets of single-plant sap flow measurements.

The TROPOspheric Monitoring Instrument (TROPOMI) is the unique sensor onboard the Copernicus Sentinel-5 Precursor satellite. It collects hyperspectral radiances on a global scale at a daily frequency, from which we can derive products

such as SIF. Its observations can be used to indirectly estimate plant photosynthetic activity, since the SIF radiation emitted by plants can be related to the photosynthetic reaction.

The main goal of the present research consists of the optimization of ORCHIDEE by using the ORCHIDEE Data Assimilation System (ORCHIDAS). The most important parameters in the computation of Gross Primary Production (GPP)[1], Latent Heat Flux (LE)[2], transpiration and SIF for each Plant Functional Type (PFT)[3] have been identified by Sensitivity Analysis (SA) and then optimized with respect to FLUXNET2015 GPP data, SAPFLUXNET transpiration measurements and TROPOMI SIF estimates. The assimilation of transpiration data in ORCHIDEE has improved the description of both transpiration and SIF with respect to the non-optimized model, while worsening its estimates for GPP mainly in boreal evergreen needleleaf forests. SIF assimilation has not produced a consistent increase in the model predictive power of transpiration and GPP, possibly due to remaining errors in the modeling of the SIF radiative transfer within the canopy. The GPP-optimized model has produced results very close to the original model ones. This is due to the fact that GPP is one of the quantities which are being used as a reference during the model development. Therefore its accuracy has been maximized in this process and it is already very high in its non-optimized version.

Keywords: Earth-System, Land-Surface Model, Data-Assimilation, Sun-Induced Fluorescence, Gross Primary Production, Plant Transpiration, Sap Flow

Title:	Model-data fusion of chlorophyll fluorescence for reducing uncertainties in local-scale simulations of plant photosynthesis and transpiration
Author:	Lorenzo Francesco Davoli
Advisors:	prof. Alessandro Pelizzola, Dr. Fabienne Maignan, Dr. Camille Abadie
Study programme:	Physics of Complex Systems
Institution:	DISAT - Department of Applied Science and Technology Politecnico di Torino
Year:	2023

[1] GPP is the total amount of carbon compounds produced by photosynthesis of plants in an ecosystem in a given period of time.[1]

[2] Latent heat is the energy absorbed by or released from a substance during a phase change from a gas to a liquid or a solid or vice versa, in this case it is associated to water evaporation in the stand.

[3] Plant Functional Types (PFTs) have been adopted by modellers to represent broad groupings of plant species that share similar characteristics (e.g. growth form) and roles (e.g. photosynthetic pathway) in ecosystem function.[2]

Acknowledgements

I'm extremely grateful to my supervisors prof. Alessandro Pelizzola and Dr. Fabienne Maignan, as well as my co-supervisor Dr. Camille Abadie, for their invaluable patience, their precious teaching and their thorough feedback.

This journey would not have been possible without the continuous support of prof. Alfredo Braunstein, whose presence has been fundamental from the beginning of my master in Turin until its very end in Paris.

I am deeply indebted to all the members of the MOSAIC team for supporting and sharing their precious knowledge during my stay in LSCE.

I would like to extend my sincere thanks to the whole MushRoom team (Mandresy, Maureen, Zac, Julien, Amelie, Joss, Karine) and to many other people (Nina, Lucas, Germain, Maxime, Fanny *et al.*) who made the time spent at LSCE brighter than my wildest expectations.

I am also heartily grateful to all the people who have played such a fundamental role in my life during this master degree, especially in Paris: Beppe, Samu, Manon, Mati, Gaia, Nicollo, Sonia, Marco, Sylvain, Sofi, Tanguy, Luis, Checco, Ani, Tsò, Silvia, Pacia, Stoppy. You all have made this experience unforgettable, and I am so lucky to have had the opportunity to cross your paths, even briefly.

Back to Italy, I would commit a gross mistake not to mention all those people who, for a long time, I considered the reason why I called Reggio Emilia "my home": B, Pol, Carra, Andre (aka Ferro), Carlo, Anna, Lusu, P, Chia, Gue, Co, Laura, Margherite (tutte e settordici), Ali, Magno, Andre (aka Persi), Ste, Nilde. As one by one most of you went your own ways, I realised that nothing went missing for me. For me, you were my actual home, all the time, from the beginning. You were, you are, and I hope you will always be.

Talking about homes, I think a couple of people are missing, two important ones for sure: Cristina and Pietro. My greatest supporters, my most rigorous critics, and we could find pros and cons in both of these aspects. Anyways, both of you are undoubtedly part of me, and I am grateful and proud you are.

Finally, I want to thank Elena, for being such an incredible and still deeply human person, a passionate friend and a fun lover.

dedicato a coloro che, scoprendosi ultimi, divennero primi:
prof.ssa e doppio Premio Nobel Maria Skłodowska Curie,
don Lorenzo Milani, don Pino Puglisi,
le Aquile Randagie Kelly e Baden,
Mariasilvia Spolato

Preface

Before you lies the master thesis "Model-data fusion of chlorophyll fluorescence for reducing uncertainties in local-scale simulations of plant photosynthesis and transpiration". It has been written to present the research carried out by the graduand as part of the final project work within the master degree in Physics of Complex Systems by Politecnico of Turin.

The aforementioned research is the result of a 5-month long internship hosted by the Laboratoire des Sciences du Climat et de l'Environnement (LSCE), located by the CEA of Paris-Saclay (Orme Des Merisiers), and founded by the CNES TOSCA FORGE project. The graduand has taken part in the research activities of the MOSAIC team, under the supervision of Dr. Fabienne Maignan and cotutored by Dr. Camille Abadie.

The choice of this specific research project has been driven by the personal interest of the graduand in the scopes of environmental sciences and modelling techniques. Moreover, this experience has provided a precious opportunity of developing a complete research activity, with a wide set of experiences and insights in the scope of Earth-System Models (ESMs), from the collection of the data to the evaluation of model performance, passing through the development of simulations, experiments, analysis methods and optimisation algorithms. In fact, the research has been structured as a preliminary study of two recent databases, namely SAPFLUXNET and TROPOMI, in order to verify the potential role they could play within the model ORCHIDEE. The research direction and its objectives have evolved and developed along the way, driven by the new information learnt in the discovery process itself. This flexibility has allowed the graduand to explore several aspects of the research in Environmental Sciences depending on his interests and knowledge.

The thesis and the associated dissertation have been designed for a public with a solid scientific background in STEM higher education and modelling, but with no specific knowledge regarding environmental physics and biology. The fundamental concepts which are involved in the research have been described at the level needed to understand their role in the dissertation, while appropriate references are given for those who want to dive into the more technical details.

Reggio Emilia,
January 2023

Lorenzo Francesco Davoli
lorenzofrancesco.davoli@studenti.polito.it

Contents

List of Figures

List of Tables

Acronyms

BAI Basal Area Integration. ix, 24–28

BFA Brute Force Approach. 24, 25

ChlF Chlorophyll Fluorescence. 4

CMIP6 Coupled Model Intercomparison Project Phase 6. 1, 13

CNES Centre National d'Etudes Spatiales. v

DA Data Assimilation. 2, 4

ECMWF European Centre for Medium-Range Weather Forecasts. 6

ESMs Earth-System Models. ii, v, 1, 12–14

ET Evapo-Transpiration. xi, 3, 18, 39, 41, 44

FORGE Fusion modèle-données Orchidee-tRoposif pour une Gpp amélioréE. v

GPP Gross Primary Production. iii, 5, 8, 35, 38, 49, 56, 60

IPCC Intergovernmental Panel on Climate Change. 1, 12, 14

IPSL Institut Pierre-Simon Laplace. 13, 14

LAI Leaf Area Index. 19

LE Latent Heat Flux. iii, 5, 18, 38, 49, 56, 60

LSCE Laboratoire des Sciences du Climat et de l'Environnement. iv, v, 15, 16

NPQ Non-Photochemical Quenching. 4

ORCHIDAS ORCHIDEE Data Assimilation System. iii, 34, 60

ORCHIDEE Organising Carbon and Hydrology In Dynamic Ecosystems. ii, v, xiii, 2, 4, 5, 9, 10, 13, 14, 18, 60, 61

PAR Photosynthetic Active Radiation. 20

PCS Physics of Complex Systems. v

PFT Plant Functional Type. iii, xiii, 9, 10, 16, 17, 19, 21, 27, 28, 61, 71

PQ Photochemical Quenching. 4

PSI Photosystem I. 4, 19, 20

PSII Photosystem II. 4, 19, 20

RMSD Root Mean Square Deviation. 61

SA Sensitivity Analysis. iii, 16, 35, 36, 60

SIF Sun-Induced chlorophyll Fluorescence. ii, iii, 3, 4, 8, 12, 14, 16, 18–20, 31, 35, 38, 49, 56, 60, 71

STEM Science, Technology, Engineering and Math. v

TOC Top Of the Canopy. 19, 20

TOSCA Terre solide, Océan, Surfaces Continentales et Atmosphère. v

TROPOMI TROPOspheric Monitoring Instrument. ii, 8, 29

VPD Vapour Pressure Deficit. 18, 49, 54

Chapter 1

Introduction

1.1 Earth-system modeling: an overview

”The five IPCC assessment cycles since 1990 have comprehensively and consistently laid out the rapidly accumulating evidence of a changing climate system, with the Fourth Assessment Report in 2007 being the first to conclude that warming of the climate system is unequivocal.”[3] More than 70 years after the first studies on Climate Change[4], the last report from Intergovernmental Panel on Climate Change (IPCC) leaves no doubts on the existence and extent of an increasing environmental crisis, mostly due to anthropogenic greenhouse gases emissions. Nowadays climate change has finally become a key topic in public debate and policy-making. Therefore, the requests for reliable projections and a deeper understanding of the Earth-system at local and global scales are becoming more and more frequent and imperative.

The IPCC was established in 1988 to provide policymakers with regular scientific assessments on the current state of knowledge about climate change. Since 1988, the IPCC has had six assessment cycles and delivered six Assessment Reports, the most comprehensive scientific reports about climate change produced worldwide. Its results are obtained through the comparison of several Earth-System Models (ESMs), providing an extensive description of the most important physical and biogeochemical phenomena characterizing our environment on diverse spatial and temporal scales.

The Coupled Model Intercomparison Project Phase 6 (CMIP6) [5] has a fundamental role in the section ”IPCC Sixth Assessment Report: Physical Science Basis[6]” as a reference for climate projections. CMIP6 coordinates somewhat independent model inter-comparison activities and their experiments which have adopted a common infrastructure for collecting, organizing, and distributing outputs from models performing common sets of experiments.

Some of these models, referred to as "coupled models", do not need any external data (apart from initial conditions) to be run, since all the quantities of interest are dynamically computed during the simulations. The single components of the coupled models only simulate single parts of the Earth-system, such as the land surfaces or the atmosphere. In this case processes that are not simulated can play a role in the dynamics the models aim to reproduce (for instance wind speed in land-surface phenomena), therefore they need external datasets, called "forcing data", to provide the missing information. For land-surface models, forcing data mostly consist of meteorological variables such as temperature, precipitation and wind speed. They provide the necessary information about atmosphere phenomena (which are not simulated within the land surface model) needed by mechanistic models to fully develop continental water and carbon cycle dynamics.

In addition to forcing data, most of the predictive power of mechanistic models lies in the parametrization of many physical quantities involved in the model algorithm. Depending on the specificity of the phenomenon, some standard values can be found in the literature, while others have to be retrieved through approximations and guesses. This is especially true when dealing with phenomena which are still not fully understood, too complex to describe in a purely mechanistic way or very specific and dependent on the analysed context. All these features are often present in biological systems and ecosystems.

A possible way out of this issue consists of data-driven approaches for parameters optimisation, which allow obtaining educated guesses for small sets of parameters. If a real measure of some key aspect of the simulated system is available, classical or machine learning techniques allow to optimize the model through maximum likelihood estimation procedures denominated Data Assimilation (DA). The application of such a technique requires a careful design of each step involved in the process and a deep knowledge of the model specificities and features, as it is going to be shown in section 4.2.

1.2 Research goals

The main goal of this research is to improve the description of vegetation dynamics in the above-mentioned land-surface model ORCHIDEE, especially regarding plant transpiration. Transpiration[7] is a phenomenon which occurs at microscopic level, as a consequence of the photosynthetic activity of plant cells. It consists in the release of water vapour from the leaf surface in the atmosphere through stomatal pores, small openings which allow the exchange of water vapour and carbon dioxide with the ambient environment. In higher plants stands, transpiration accounts by itself for about three-quarters of the water that is va-

porized at the global land surface and one-eighth of that vaporized over the entire globe[7], and therefore it is one of the main component of latent heat. Transpiration is also a key factor in vegetation dynamics, representing one of the more accurate descriptors of plant activity as a whole[8]. In fact, transpiration is strictly related to photosynthesis. During photosynthesis, plants use sunlight to convert CO_2 and water into carbohydrates and O_2. They take up the carbon dioxide from the ambient air through stomata on the surface of their leaves and water from the soil through roots. Approximately 1% of the water extracted by plants from the soil is actually used for plant growth; the rest is released as water vapour to the atmosphere through plant transpiration, an unavoidable by-product of carbon exchange via stomata openings[8][9].

Being so deeply connected with all the aspects of the plant life-cycle, transpiration is strongly influenced by several environmental factors, also called "drivers" of the phenomenon, such as plant available water, shortwave incoming radiation, soil moisture, CO_2 concentration in the atmosphere and air temperature[10]. This aspect induces a huge variability and complexity in transpiration dynamics, which are highly specific for every ecosystem considered and therefore quite complex to study and model. Measuring directly at the leaf level a microscopical process and then up-scaling the results can produce large uncertainties, even at a single-plant level. Not to mention the difficulty of an up-scaling to canopy level, which is the scale of interest for most of the applications. The lack of such measures makes it quite difficult to train or tune models describing correctly and in a general way this phenomenon in its full complexity and variability.

Diverse strategies are currently being adopted to obtain transpiration estimates from other sources, called "proxies". The most established procedure uses Evapo-Transpiration (ET)[8] (the total water vapour flux from the land surface to the atmosphere) as a proxy for transpiration, assuming a fixed ratio between bare-soil water evaporation and plant transpiration contribution. This ratio is far from being precisely estimated though, and it is still affected by large uncertainties (depending on the studies, the transpiration contribution lays between 70% and 90% of ET). Therefore, other proxies are being tested to obtain estimates which are independent of bare-soil evaporation phenomena, and only related to plant activity. Amongst those, two in particular will be considered in the present research: single-plant sap flow and Sun-Induced chlorophyll Fluorescence (SIF).

A gradient in the water potential allows bringing the water collected by roots from the soil to the leaves, where photosynthesis and transpiration occur. As mentioned above, almost all the water transported in this way is released through transpiration, and only a small percentage is actually stored in plant tissues in the growth process. Therefore, sap flow can be considered a valuable proxy of the water vapour flux due to transpiration.

Sun-Induced chlorophyll Fluorescence (SIF) instead is directly related to pho-

tosynthesis. In fact, photosynthesis is controlled by two types of photosystems located in the leaves: Photosystem I (PSI) and Photosystem II (PSII). These photosystems are pigment-containing protein complexes where light is absorbed and electrons are transported. Most of the incoming solar radiation is absorbed and converted into energy for photosynthesis (Photochemical Quenching, PQ). Some of the energy is dissipated as heat for photoprotection (Non-Photochemical Quenching, NPQ), and a small fraction is re-emitted back as Chlorophyll Fluorescence (ChlF) at wavelengths between 650 and 850 nm[11]. Subsequently, the fluorescence signal which it is possible to measure from these photosystems is directly relatable to photosynthetic activity, and from there to transpiration[12].

The sap flow dataset integration method is a key-point investigated in the scope of this research, needed for the data to be statistically relevant, homogeneous and compatible with models estimates. For the sake of completeness, different integration methods are tested and compared amongst each others to find the most effective procedure. The core of the research project is precisely the DA of two recently published databases (namely SAPFLUXNET[13] single plant sap flow measurements and TROPOMI[14] SIF satellite observations) within the above-mentioned land-surface model ORCHIDEE, in order to assess their influence on the model predictive power. SAPFLUXNET data being local measurements, the whole research is focused on a set of 6 sites, namely those which are in common with both FLUXNET2015 (i.e. ORCHIDEE forcing files for local simulations, providing also important information on water- and carbon-related fluxes) and SAPFLUXNET.

Chapter 2

Meteorological data and fluxes

2.1 Reference databases

2.1.1 FLUXNET2015: meteorological data and fluxes

FLUXNET is an international "network of networks", tying together regional networks of Earth-system scientists and collecting their data into structured data products. FLUXNET research teams use the eddy covariance technique [15] to measure carbon, water, and energy fluxes between the biosphere and atmosphere.

The most recent global FLUXNET data product, FLUXNET2015[16], is hosted by the Lawrence Berkeley National Laboratory (USA) and is publicly available for download. Currently there are over 1000 active and historic flux measurement sites, dispersed across most climate spaces and representative biomes (Figure 2.1). The higher concentration of observation sites is found in western countries and Japan, while developing countries are less represented and present shorter time-series, since their sites have been activated more recently.

FLUXNET2015 data contain hourly or half-hourly time-series of several meteorological and fluxes observations, in the form of instantaneous, mean (over the time step interval) or cumulative values. In particular, the observations that have been used are GPP, Latent Heat Flux, air temperature T_{air}, near surface specific humidity of air Q_{air} and downward short wave radiation SW_{down}. Most of the time-series used in the scope of this research cover the timespan between 1998 and 2014. All data undergo quality tests to check their consistency and completeness. In case some entries are missing or fail the quality checks, a gap-filling procedure guarantees the completeness of the dataset. Also for this reason FLUXNET2015 has been chosen as the reference database regarding forcing data for ORCHIDEE local offline simulations and fluxes used in the model optimiza-

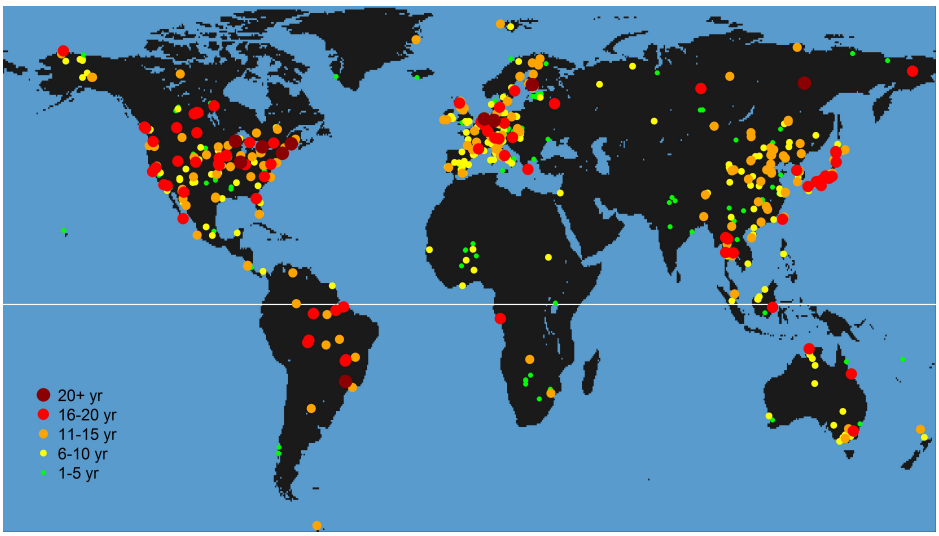


Figure 2.1: FLUXNET2015 observation towers map. Sites are represented by dots whose dimension and color indicate the duration of the period of activity of that measurement site. Source: https://fluxnet.org/about, last visit: 08/11/2022

tion.

The database provides also some correction for quantities which are known to fulfill specific conditions. For instance a correction to energy fluxes estimates is computed by imposing the energy balance closure[17], which is otherwise not fulfilled by the great majority of the data[18].

2.1.2 ERA5: atmospheric re-analysis

ERA5[19] is the fifth generation of European Centre for Medium-Range Weather Forecasts (ECMWF) atmospheric re-analysis of the global climate, produced by the Copernicus Climate Change Service (C3S) at ECMWF, covering the period from January 1950 to present.

ERA5 provides hourly estimates of a large number of atmospheric, land and oceanic climate variables. The data cover the Earth on a 30km grid and resolve the atmosphere using 137 levels from the surface up to a height of 80km. ERA5 combines vast amounts of historical observations into global estimates using advanced modelling and data assimilation systems. ORCHIDEE simulations from 2015 to present (not covered by FLUXNET2015) use these data as forcing files.

2.1.3 SAPFLUXNET: a proxy for transpiration estimates

SAPFLUXNET[13] is the first global tree sap flow database. It collects quality-controlled sub-daily transpiration data, derived from sap flow measurements, from more than 200 sites all over the world. Sap flow sensors track the diffusion of heat applied to the plant's conducting tissue using temperature sensors deployed in the plant's main stem, and derive the sap flux from it.

The transpiration is available in 3 formats, namely transpiration per sapwood area, transpiration per leaf area and single-plant transpiration. Only the latter is present for all the sites and it is usually computed through an integration over the plant sapwood area, where the sap flows. This measurement is the one exploited in the research, being the easier to up-scale to canopy level once the stand composition is known (see subsection 4.1.2). Datasets also include sub-daily time series of hydrometeorological drivers and metadata on the stand characteristics. SAPFLUXNET has a broad bioclimatic coverage, with woodland/shrubland and temperate forest biomes especially well represented (80% of the datasets), and covers the period between 1995 and 2018. The database covers only the growing season for most of the sites. Accompanying radiation and vapour pressure deficit data are available for most of the datasets, while on-site soil water content is available for about half of them.

For each site, the plants whose sap flow measurements have been collected are chosen in order to reproduce at best the actual stand composition, as verified through the preliminary analysis presented in subsection 4.1.2, together with the integration procedure needed to up-scale from single-plant to canopy level. This procedure also takes care of the most critical characteristic of this database, a widespread lack of entries, even for long periods. If not correctly handled this aspect could heavily compromise the reliability of the data.

Sap flow can be considered a suitable (and practical) proxy for transpiration, as shown by Koppa et al.[10]. In fact, a direct measure of single-plant transpiration is very complicated and expensive, since the process occurs at the leaf level and it can be highly inhomogeneous even across the canopy. On the other side, the transport of water (in the form of sap) from the roots through the trunk is almost entirely directed to leaves, where the water is exchanged with the environment as a consequence of photosynthetic reaction, which is coupled to transpiration due to stomatal opening.

2.1.4 TROPOMI: SIF satellite observations

The TROPOspheric Monitoring Instrument (TROPOMI)[14] is the satellite instrument on board the Copernicus Sentinel-5 Precursor satellite, launched on 13 October 2017 for a mission of at least seven years (the data are available from May 2018). Although the Sentinel-5P mission was mostly designed to monitor atmospheric phenomena, the TROPOMI apparatus enables to also estimate terrestrial Sun-Induced chlorophyll Fluorescence (SIF) at spatial and temporal resolutions (up to 7km × 3.5km pixels with a daily revisit) more suitable for land surface models than the characteristics of its predecessors, such as GOSAT (with a spatial resolution of 10.5km × 10.5km and 3-day revisit[1]) and GOME-2 (with a spatial resolution of 80km × 40km and daily revisit[2]).

SIF is an electromagnetic signal emitted by chlorophyll and related to photosynthetic activity when illuminated. This emission is characterized by a two-peak spectrum roughly covering the 650÷850 nm spectral range. The SIF estimates which have been used in the following come from the filling-in of solar Fraunhofer lines for the 743 nm fluorescence emission peak. Environmental disturbances due to light and water stress influence instantaneously SIF, which makes it a better proxy than classical reflectance-based vegetation indices for photosynthetic activity in certain situations[20].

SIF estimation from space-borne spectrometers requires both high spectral resolution and advanced retrieval schemes, since it constitutes only 0.5%÷2% of the radiance at the top of the canopy, which is mostly composed of reflected sunlight. The broad range of viewing-illumination geometries covered by TROPOMI's 2,600-km-wide swath introduces large directional effects that need to be considered[14]. SIF is already being proven to be able to improve model estimates of Gross Primary Production (GPP), the quantity of atmospheric carbon dioxide assimilated through photosynthesis [21]. According to Damm et al.[22]: "The novel Earth observation signal sun-induced chlorophyll fluorescence (SIF) is the most direct measure of plant photosynthesis and offers new pathways to advance estimates of transpiration". As explained in the paper, SIF by itself is not able to fully and correctly constrain transpiration in current land surface models yet, and therefore it needs to be coupled with other Earth observation data.

[1] source: https://www.eoportal.org/satellite-missions/gosat, last visit 24/02/2023

[2] source: https://www.esa.int/Applications/Observing_the_Earth/Meteorological_missions/MetOp/About_GOME-2, last visit 24/02/2023

Figure 2.2: Study sites chosen amongst FLUXNET2015 and SAPFLUXNET common sites, with a minimum of 3-year long timeseries and providing the stand information needed by the integration procedure. The tags refer to the FLUXNET2015 notation.

2.2 Study sites

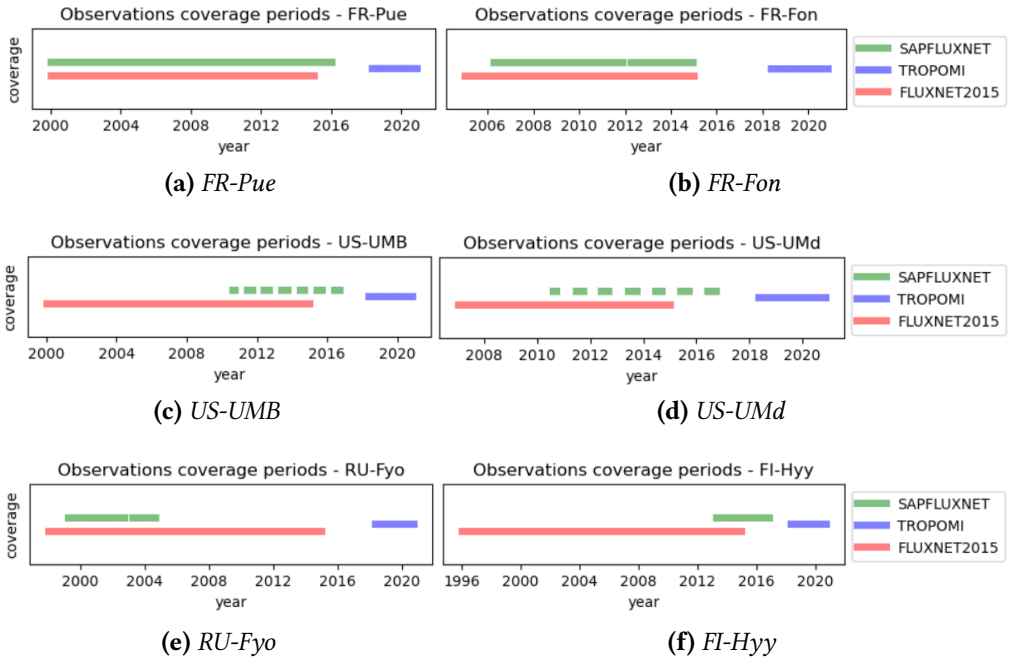
The need for comparing ORCHIDEE transpiration estimates with sap flow derived ones has led to the choice of 6 study sites (4 independent and 2 being part of a coordinated experiment), which are in common with SAPFLUXNET and FLUXNET2015 databases and provide at least a 3-year long overlapping period (see Figure 2.3 for coverage periods). The locations of these sites are presented in Figure 2.2, while some further details are reported in Table 2.1.

All these sites, except for the US sites which are part of a coupled experiment[3], present individual characteristics which make them unique with respect to the others. For instance, FI-Hyy presents the same biome and Plant Functional Type (see section 3.2 and Table 3.1 for more details) as RU-Fyo, but a different soil texture (see section 3.2 and Figure 3.2 for more details). Also FR-Fon and FR-Pue, located respectively at Fontainebleau and Puéchabon in France, differ quite a lot in their behaviours, because FR-Pue is an evergreen broad-leaved forest with

[3]Site US-UMB is artificially defoliated to study the response of the stand to external disturbances. US-UMB is the control sample, while US-UMd is the perturbed one[24].

Site tag	Country	Coordinates	Biome	PFT	Soil texture
FI-Hyy	Finland	61°84'74"N 24°29'47"E	Boreal forest	7	sandy loam
FR-Fon	France	48°47'63"N 2°78'01"E	Woodland/Shrubland	6	loam
FR-Pue	France	43°74'13"N 3°59'58"E	Woodland/Shrubland	5	clay loam
RU-Fyo	Russia	56°46'15"N 32°92'20"E	Boreal forest	7	loam
US-UMB	USA	45°55'97"N -84°71'33"E	Temperate forest	6	sand
US-UMd	USA	45°56'25"N -84°69'75"E	Temperate forest	6	sand

Table 2.1: Study sites characterization. The first column of ID tags reports the notation used in FLUXNET2015, while the second one refers to the SAPFLUXNET denomination. The PFT notation follows the ORCHIDEE system (Table 3.1)[23]. Soil texture is expressed according to the USDA soil texture classification (see Figure 3.2).

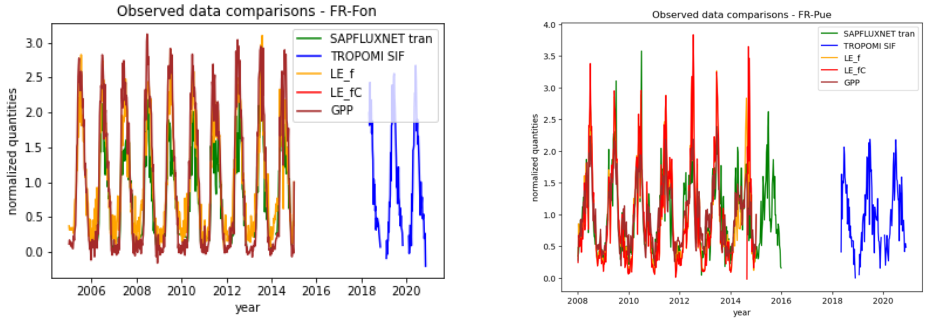


(a) *FR-Pue* **(b)** *FR-Fon*

(c) *US-UMB* **(d)** *US-UMd*

(e) *RU-Fyo* **(f)** *FI-Hyy*

Figure 2.3: Coverage of meteorological observations over the whole available periods for all the sites.



(a) *FR-Fon* **(b)** *FR-Pue*

Figure 2.4: Comparison amongst several flux time-series of sites FR-Fon (a) and FR-Pue (b). In 2011 a severe drought hit Europe, and its effects are visible from the data, where 2011 presents a local minimum for all the variables sampled in that period.

clay loam soil texture, while FR-Fon is a summergreen broad-leaved forest with loam soil texture. As it is shown in chapter 5, these difference cause a completely different behaviour between the two sites, but their geographical proximity allows to qualitatively observe for both of them the influence on GPP of a drought taking place in France in 2011[4], as shown in Figure 2.4.

Besides these natural and unavoidable specificities, the databases coverage changes from site to site, both in terms of time series duration and period, and the same holds for their reliability. For instance SAPFLUXNET data from US-UMB collect measurements for more than 50 different trees, while in some periods FI-Hyy presents information regarding only one tree at a time.

The small number of available sites, their diversity and the lack of long time series pose a challenging issue for the statistical analysis carried out during this research and they might cause a loss of comparability between different sites observations. Biological systems being so diverse in nature and climate studies being so expensive to carry out on large scales, these issues are not unusual in environmental sciences. Nonetheless the study of these phenomena, even in non-optimal conditions, has been useful to identify points of strength and weakness in the datasets, in the methodology and in the model itself, highlighting criticalities and new perspectives to better direct the future development of the research in this scope.

[4]about the drought of 2011 in Europe: https://www.eea.europa.eu/data-and-maps/figures/onset-of-the-2011-european, last visit: 01/12/2022

Chapter 3

Land-surface modelling

3.1 A short review on Earth-System Models and CMIP6

The development of ESMs is not only focused on the production of climate projections as those used in IPCC Assessments. It is also plenty of opportunities for indirectly investigating the presence of hidden or unknown processes and testing the correct description of the known ones. As it is going to be shown in section 5.3, the comparison of ORCHIDEE[1] estimates with transpiration and SIF observations has provided interesting information and insights on the nature of the described processes, and useful hints on the absence of important ones.

Moreover, ESMs are used to estimate quantities which are otherwise difficult to measure directly on the field. An example that will be treated in this research is the direct measurement of canopy transpiration, which would usually need a large amount of sensors for every single plant in order to be properly quantified. Instead, models solve this problem by simulating the whole-stand biotic and abiotic dynamics in their entirety, providing the opportunity for estimating a wide set of physical quantities, such as the aforementioned transpiration or SIF. Whenever it is possible, model estimates should be validated against a set of data to ensure the model is correctly describing the process. For instance, in the present research transpiration estimates have been compared with observations derived from SAPFLUXNET sap flow data in order to assess the performances of the model and improve them through parameters optimization.

The validation procedure is strongly dependent on the quality and quantity of the observations which are used to validate the model. Local measurements being usually expensive and mostly localized in high- or middle-income countries

[1] see section 3.2 for more details

(North America, Europe, Japan and China, see Figure 2.1), they can not provide a complete and sufficient coverage over the entire land-surface. For this reason satellite missions represent a fundamental source of global observations, which would partially solve this problem and produce high resolution and wide coverage data for most of the countries, a precious contribution for the development of ESMs. These types of dataset are also more suitable to be used for producing large (or even global) scale simulations, which are the most useful for the study of climate change on Earth.

Several models have been developed in the last few decades to simulate climate on a global scale. The Coupled Model Intercomparison Project Phase 6 (CMIP6) is a leading project in the scope of Earth-System modeling. It consists in a ensemble of "coupled models", simulating all the principal domains in Earth dynamics, their reciprocal interactions and transport phenomena. For example, Institut Pierre-Simon Laplace maintains and develops the coupled model *IPSL-CM6A-LR*[25], which gathers and couples three main models, LMDz, ORCHIDEE and NEMO, respectively describing atmosphere, land-surface and ocean dynamics.

All together, in the so-called "coupled" or "online mode", these models can describe most of Earth-System phenomena on a global scale, and they are still undergoing a continued development to include more and more processes in their simulations. ORCHIDEE by itself presents more than 50 active development branches, implementing new functions, phenomena and improving performance starting from the trunk version of the model[2].

CMIP6 models can also be run independently ("uncoupled") one from the others (so-called "offline mode"), when provided a meteorological and environmental forcing dataset, consisting of local or global observations and re-analysis data. For instance ORCHIDEE land-surface local simulations can be run for single sites covered by the FLUXNET2015 database[16] (providing atmospheric meteorological forcing data), while global simulations use ERA5 re-analysis data[19] (see section 2.1 for further details about databases). The resolution of these global simulations is usually on the order of $0.5^\circ \times 0.5^\circ$ in a latitude-longitude grid, while for the local simulation the grid cell depends on the site footprint. For the 6 sites analysed in this research the resolution of local simulations is $0.05^\circ \times 0.05^\circ$.

[2]for a complete list of all the development activities: https://forge.ipsl.jussieu.fr/orchidee/wiki/DevelopmentActivities, last visit: 29/11/2022

3.2 ORCHIDEE model: an overview

The Organising Carbon and Hydrology In Dynamic Ecosystems (ORCHIDEE) model is a model developed by Institut Pierre-Simon Laplace (IPSL) as part of the IPSL Earth-System Models (ESMs). Its simulations have contributed to the Nobel prize winning efforts of the IPCC, as well as to numerous landmark projects such as the Global Carbon Project, TRENDY[26], CMIP and related publications used by IPCC. The model is currently at its 4^{th} release, but the one used in this research is an implementation of version 2 including SIF computation. The soil description of this version is structured as an 11-layer profile which accurately describes the varying water stocks in the soil.

ORCHIDEE describes the functioning of the terrestrial biosphere and it can be deployed as a stand alone model or in a coupled set-up, where other models from IPSL contribute to the description of atmospheric and ocean dynamics, namely LMDz and NEMO (see section 3.1). It needs atmospheric meteorological data (precipitation, air temperature, wind, solar radiation, humidity) and atmospheric $CO2$ as forcing data, which can be provided by the coupling with the LMDz model or by an external dataset, for instance FLUXNET2015 or ERA5. The model can run global and local simulations up to 2020, using ERA5 re-analysis data. In the latter case it is suggested to use FLUXNET2015 data whenever is possible. If the site is not covered, or the period exceed 2014, ERA5 data are used to extend the simulations up to 2020.

ORCHIDEE solves the water and energy budgets and fast processes at a half-hourly time step, within the module *Sechiba*, while most of the carbon cycle and other slow processes (as carbon allocation and phenology) are computed on a daily basis by the module *Stomate* (Figure 3.1). Outputs and the corresponding output frequencies are specified by the user, depending on quantities of interest, time scales of the investigated phenomena and computational resources availability.

The process which is followed to produce local ORCHIDEE simulations consists in 3 phases: spin-up phase, transient phase and effective simulation.

The "spin-up" phase enables all carbon pools to stabilize towards a stationary state such that the net biome production oscillates around zero. A pseudo-analytical iterative estimation of the carbon pools allows the simulation to reach the equilibrium more quickly by inducing a 6-to-20 times faster convergence of the algorithm[28]. The spin-up phase simulates 360 years and it is performed by cycling the years available in the FLUXNET2015 forcing files over and over. Each spin-up run has a computation time of 5 hours approximately, corresponding to less than one minute of computational time per simulated year. The computation cost of the spin-up is also reduced by the fact that outputs and data are saved with

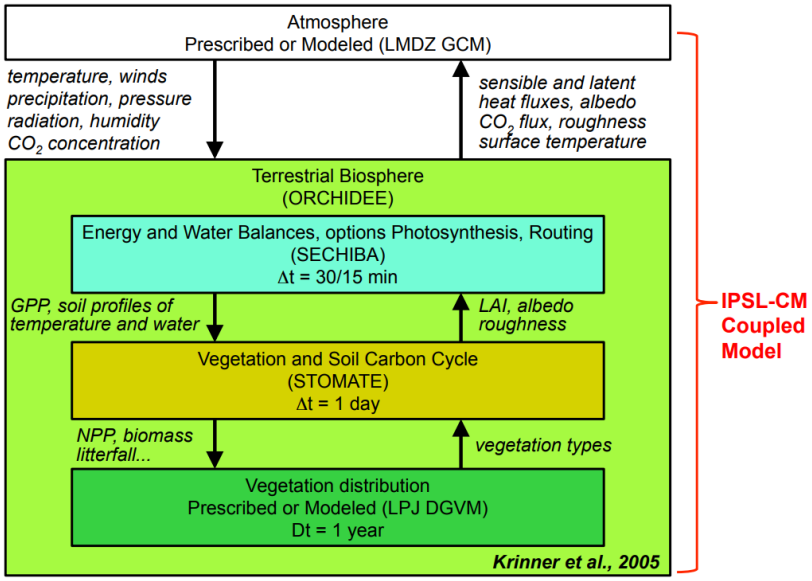


Figure 3.1: Structure of ORCHIDEE model and coupling with LMDz atmospheric model. Source: Krinner et al. (2005)[27]

a much smaller frequency than the actual simulation phase, since the product of this phase has a yearly output frequency.

The latter considerations hold also for the transient phase, which uses as initial state of its own run the outcome of the spin-up phase. Transient phase runs simulate 60 years between 1950 and 2010 cycling through the same forcing files as before, and they introduce disturbances such as climate change, land use change and increasing CO2 atmospheric concentrations in order to reproduce the current environmental conditions. The computation time is about 1 or 2 hours per run, with the same computational advantages of spin-up phase.

The actual simulations are run starting from the restart files generated at the end of the transient phase. They can simulate whatever period within the timespan covered by FLUXNET2015 or ERA5 data, used as a forcing files. A 10-15 years simulation with a half-hourly output frequency over a single site usually takes 20-30 hours to be completed.

All in all, the overall computational time required in this research to run all the simulations and optimization amounts to 3900 hours approximately, corresponding to 162 days of uninterrupted computation. It is important to mentions that these calculations are not considering all the failed, wrong, repeated simulations and the computational time used in the development and execution of the analysis software. Using the LSCE supercomputer cluster, called Obelix, made possible to run many simulations simultaneously, reducing the computational time approximately by a factor 6. As a consequence, the computational cost of

the whole process on the cluster amounts to about one month of computation in real life.

Many other features are currently being developed in ORCHIDEE, mainly concerning soil hydrology, carbon and nitrogen cycles, wetland and permafrost dynamics, crop phenology and further parallelizations of the simulations. The model is continuously developing and getting more and more accurate in the description of Earth-System dynamics, thanks to the contribution of many teams and developers working simultaneously on it. This cooperative development also helps to discover new perspectives and spot undetected errors: having several developers, researchers, students performing independent analyses and experiments allows a mutual validation and correction of the modules from many different points of view and backgrounds. For instance, the optimization of ORCHIDEE through TROPOMI SIF observations carried out in this research project has quantitatively brought out some inconsistencies in the newly added module of SIF computation, as reported in section 5.3. The downside of this aspect is the difficulty in maintenance and coordination between and within the different teams. For instance, in the middle of the present research it has been decided to change the model toward a new version, including SIF estimation, which has been developed recently by the LSCE MOSAIC team. Similarly, the software for Sensitivity Analysis (SA) used at the beginning of the optimization phase was not the most recent release, but an older version steadily used by the team. This version has been found not to be working properly for the set of experiments of this research, due to a bug which was already being fixed in the latest version. This kind of issues certainly increase the time spent on debugging and fixing the errors, but also allows members of different teams to learn much more about the model as a whole and not to focus only on their own research module.

The vegetation distribution can be modeled (with a yearly update frequency) or prescribed, and the ecosystem is described in terms of a range of Plant Functional Type (PFT) (Table 3.1). PFTs approximate the composition of the stand and define all the biological features used in the model to simulate the vegetation. The computation of carbon fluxes uses the so-called ”big-leaf” approach, assuming that canopy carbon fluxes have the same relative responses to the environment as any single sunlit leaf in the upper canopy, recently implemented with the addition of diffuse light contribution (see section 3.3 for further details). The model also computes its own phenology, i.e., the onset of leaves at the start of the growing season for deciduous species and their turnover and senescence. The soil texture is prescribed by a global soil map[29], but it can also be specified site by site whenever more precise information about soil texture is available, which is the case for the 6 study cases of this research. USDA soil texture classification is adopted (see Figure 3.2 for details).

Plant Functional Types	Long name
PFT1	Bare soil
PFT2	Tropical broadleaved evergreen forest
PFT3	Tropical broadleaved raingreen forest
PFT4	Temperate needleleaf evergreen forest
PFT5	Temperate broadleaved evergreen forest
PFT6	Temperate broadleaved summergreen forest
PFT7	Boreal needleleaf evergreen forest
PFT8	Boreal broadleaved summergreen forest
PFT9	Boreal needleleaf summergreen forest
PFT10	Temperate C3 grass
PFT11	C4 grass
PFT12	C3 agriculture
PFT13	C4 agriculture
PFT14	Tropical C3 grass
PFT15	Boreal C3 grass

Table 3.1: ORCHIDEE Plant Functional Types (PFTs).

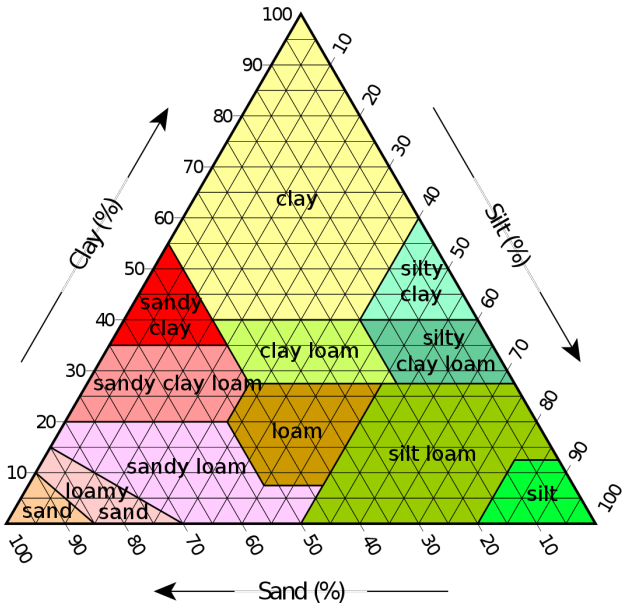


Figure 3.2: Diagram describing the soil textures in the USDA classification. Each soil is classified through its percentages of clay, sand and silt. Source: Wikimedia Commons, author: Mike Norton.

3.3 Transpiration and sun-induced fluorescence in ORCHIDEE

Photosynthesis and all components of the surface energy and water budgets (including transpiration and SIF) are calculated at a half-hourly resolution in the *Sechiba* section of ORCHIDEE. Photosynthesis main drivers are light availability, CO2 concentration, soil moisture, and temperature. These drivers are related to photosynthesis by following the approach of Yin and Struik (2009)[30]. This formulation describes the main photosynthesis processes (i.e. electron transport and carboxilation), and associated variables like the stomatal conductance, and the intercellular CO2 partial pressure. In the version of the model that has been used for this research, the calculations take in account the photosynthetic activity due to both direct and diffuse light absorption. In fact, several studies of in situ observations have found that the presence of a high diffuse light fraction (with respect to the overall irradiation level over the canopy) can enhance light use efficiency and photosynthesis in the plants[31].

Some mechanisms have been found to be involved. First, diffuse light isotropic nature allows the radiation to penetrate deeper in the canopy, reaching leaves that would be otherwise mostly shaded for their position or orientation, and limiting the waste over light-saturated sunlit leaves. Second, diffuse light is usually related to cloudy weather, and as a consequence it is often accompanied by less stressing temperatures and Vapour Pressure Deficit (VPD) for photosynthesis.

These correlated environmental factors may cause the photosynthesis under cloudier conditions to be more intense than the one estimated in a direct-light-only absorption case, making a proper description of diffuse light effects on vegetation fundamental for a correct evaluation of the canopy photosynthetic activity in ORCHIDEE as well[32]. ORCHIDEE calculates the fraction of diffuse light[33] as well as a process-based multilayer canopy light transmission model to effectively represent the contribution of diffuse light fraction on photosynthesis[34].

Transpiration is involved both in the calculation of energy balance and water cycle. It is strictly related to Latent Heat Flux (LE) and gives an important contribution to the description of ET processes[3]. Several environmental factors influence the transpiration process, such as water stress linked to soil water con-

[3]Evapotranspiration is the transfer of energy from the Earth's surface to the atmosphere in the form of latent heat, due to the evaporation of water from the ground and bodies of water, and the transpiration of water from plants. Source: https://energyeducation.ca/encyclopedia/Evapotranspiration, last visit: 06/01/2023

tent, air moisture, net radiation, wind and temperature. The characterisation of biological features of the stand is defined by the associated PFT.

The output variable of transpiration in ORCHIDEE consists in the flux of water transpired by the full canopy per stand area ($kg\,m^{-2}\,s^{-1}$). Transpiration T is computed through:

$$T = V \cdot \sum_{i \in PFTs} \beta_3^i \tag{3.1}$$

Here β_3^i represents the canopy transpiration resistance of PFT i component of the stand and V is defined as:

$$V = \Delta t \cdot (1 - \beta_1)(1 - \beta_5)\left(q_{s,sat} - q_a\right) \cdot \rho \cdot S_{TOC} \cdot q_c^{drag} \tag{3.2}$$

where Δt is the length of one timestep in the module *Sechiba* (s), β_1 and β_5 are respectively the snow sublimation resistance and floodplains resistance (dimensionless), $q_{s,sat}$ is the saturation humidity corresponding to a particular surface temperature (dimensionless), q_a is the specific humidity in the atmosphere immediately above the surface (dimensionless), ρ is the air density ($kg\,m^{-3}$), S_{TOC} the Top Of the Canopy (TOC) wind speed ($m\,s^{-1}$), q_c^{drag} the surface drag coefficient (dimensionless).

SIF computation is carried out independently for each grid cell and PFT. Its output variable consists in the upward flux of SIF ($W\,m^{-2}\,sr^{-1}\,\mu m^{-1}$) at the 740nm wavelength emerging from the TOC. This value is comprehensive of the contribution of the PSI and PSII (see section 1.2 for further details) activity of all the leaves layers in the canopy, and it is given by:

$$SIF = (SIF_{PSI} + SIF_{PSII}) \cdot veget_{max} \tag{3.3}$$

where SIF_{PSI} and SIF_{PSII} are respectively the contributions from PSI and PSII, and $veget_{max}$ the maximum vegetation fraction for the PFT in the cell (dimensionless).

These two contributions exhibit the same structure, as they differ only for the photosystem considered when computing the total leaf fluorescence flux $f_{lai,irr}^{\alpha}$ ($W\,m^{-2}\,sr^{-1}\,\mu m^{-1}$), where $\alpha \in \{PSI, PSII\}$, lai indicates the canopy layer (the use of the term lai refers to the Leaf Area Index, which is defined as the total one-sided green leaf area per unit of ground surface[35], in this case corresponding to a specific canopy layer) and irr indicates if the direct or diffuse light irradiation is considered. Its explicit formula reads:

$$f_{lai,irr}^{\alpha} = PAR_{lai} \cdot abs_{chl} \cdot a_{\alpha} \cdot \Phi_{\alpha} \cdot e_{eff} \tag{3.4}$$

The factors in Equation 3.4 represent the Photosynthetic Active Radiation PAR_{lai} ($W\,m^{-2}$) of the layer lai , the relative specific absorption coefficient of chlorophyll over the 400nm-750nm spectrum abs_{chl} (dimensionless), the absorption cross-section area of α a_α (dimensionless), the quantum yield of fluorescence Φ_α ($\mu mol_{photon}\,m^{-2}\,s^{-1}$) and finally the emission efficiency of α at 740nm e_{eff} (μm^{-1}). The contribution of PSI and PSII are not independent, a_{PSI} and a_{PSII} being related through the equation

$$a_{PSI} = 1 - a_{PSII}$$

while Φ_{psI} and Φ_{psII} through

$$\Phi_{psI} = c\,\Phi_{psII}$$

(c being fixed ratio).

As mentioned before, each system contribution is the result of the integration over all the canopy layers and irradiation types (direct or diffuse light). In fact, the integration is based on the evaluation of $\mathcal{F}_{lai,irr}$ ($W\,m^{-2}\,sr^{-1}\,\mu m^{-1}$), the upward fluorescence flux emerging from the lai-th layer (starting from the bottom one) both related to direct and diffuse light (represented by irr). Defining the bottom layer fluorescence flux

$$\mathcal{F}^{\alpha}_{0,irr} = \frac{1}{2\pi} f^{alpha}_{0,irr} \tag{3.5}$$

where the factor $\frac{1}{2\pi}$ emerges by considering only the upward fluorescence flux, SIF_α is given by

$$SIF_\alpha = \mathcal{F}^{\alpha}_{0,dir} + \sum_{i=1}^{n_{lai}-1} \mathcal{F}^{\alpha}_{i,dir} + \frac{1}{2\pi} f^{\alpha}_{n_{lai},dir} + \mathcal{F}^{\alpha}_{0,dif} + \sum_{i=1}^{n_{lai}-1} \mathcal{F}^{\alpha}_{i,dif} + \frac{1}{2\pi} f^{\alpha}_{n_{lai},dif} \tag{3.6}$$

n_{lai} being the total number of layers, representing also the TOC layer index. $\mathcal{F}^{\alpha}_{i,irr}$, with $i \in \{1, n_{lai} - 1\}$ are computed recursively, starting from the bottom layer and computing the contribution of the upper ones by considering their own absorption, emission and reflectance, based on the leaf absorption of 740nm wavelength radiation.

Unlike transpiration, SIF is only a diagnostic variable, which is not involved in any further computation. For this reason the aforementioned error in the modeling of the SIF radiative transfer within the canopy does not compromise the whole simulation, despite producing wrong values for SIF itself.

Chapter 4

Methodology of research

4.1 Preliminary analysis

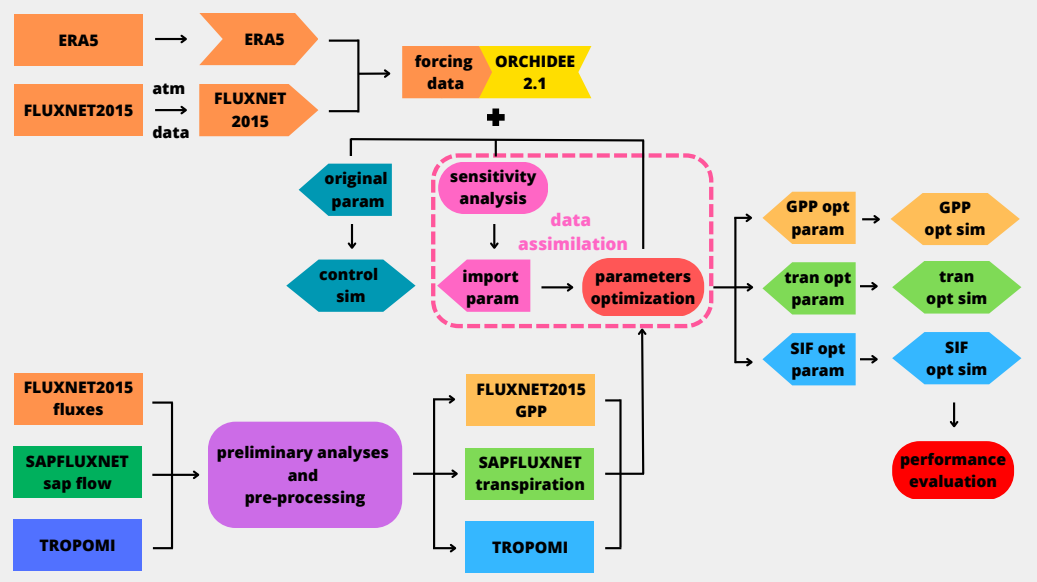
4.1.1 Importance of data preliminary studies

The research has been grounded on the study and comparison of 4 databases, introduced in section 2.1. A graphical representation of the research structure is displayed in Figure 4.1.

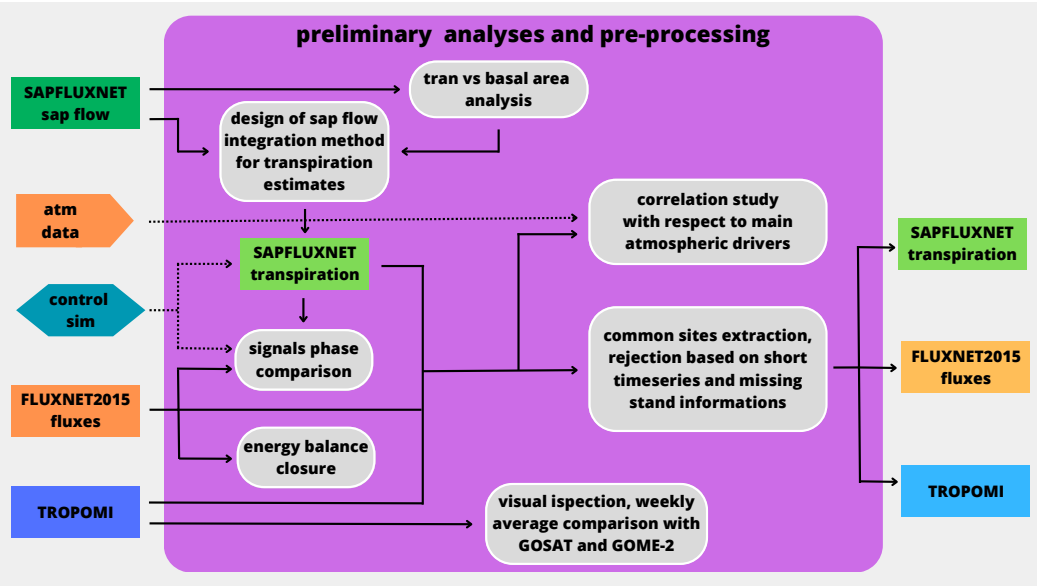
Each database required a proper analysis and study in order to determine its full potential and assess the comparability with the model and the other data. The deep diversity of the 4 sources in terms of data sampling frequency, format, coverage, and the specificity of the sites themselves emerged as a major challenge in the design of experiments which could effectively produce statistically relevant and comparable results, as will be explained in the following sections.

A graphical visualization of the data has been used to qualitatively evaluate the noise associated to the measurements. With this information the data has been averaged towards daily and weekly values, depending on the observations, in order to obtain a more distinguishable, less noisy and physically meaningful signal. As it is possible to observe in Figure 4.2, it is also important to mention that taking daily averages is also necessary whenever the diel cycle needs to be ruled out due to statistical reasons.

The preliminary analyses (Figure 4.1b) have also pointed out the difficulty of comparing results from different sites due to the lack of overlapping periods in SAPFLUXNET data. For example, the optimisation against transpiration data, introduced in section 4.2, has been performed over different periods from site to site, and therefore the differences emerging from the comparison between different PFT could be due to that. Also, it has not been possible to compute directly correlation between SIF and GPP or transpiration data because there

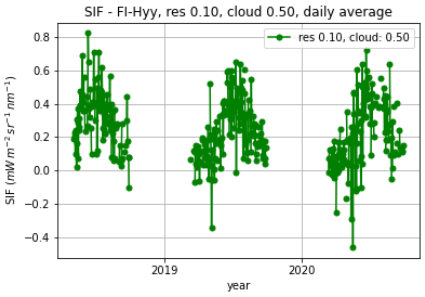


(a) *research flowchart*

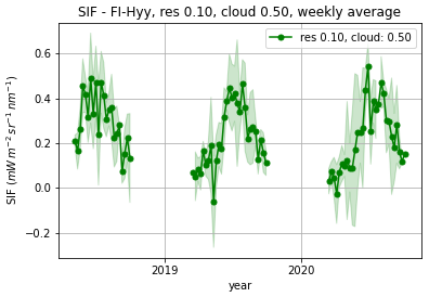


(b) *preliminary analysis and pre-processing detail*

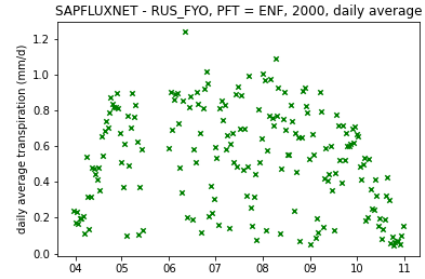
Figure 4.1: The diagram in (a) represents the main steps followed during the research, starting from the original databases and ORCHIDEE configuration and ending with the optimized models. The preliminary analysis and pre-processing procedures undergone by the databases presented in section 2.1 are displayed in (b).



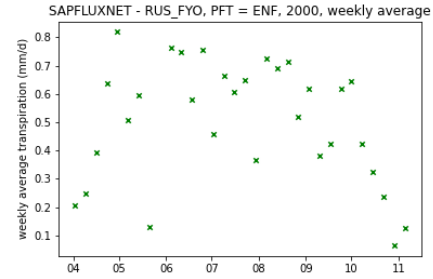
(a) *daily values*



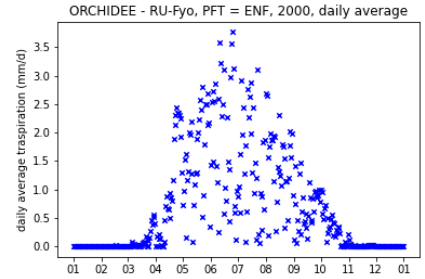
(b) *weekly means*



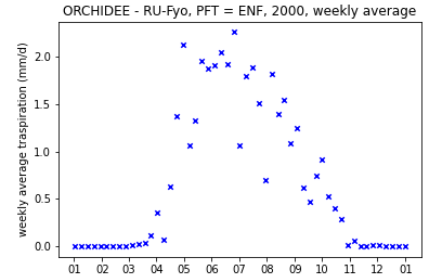
(c) *daily means*



(d) *weekly means*



(e) *daily means*



(f) *weekly means*

Figure 4.2: Comparison between TROPOMI SIF daily data (a) and their weekly average with shaded uncertainties (b) over 3 growing seasons (from 2018 to 2020). The signal is almost indistinguishable in the first case, while it is much more evident in the second one. The same considerations hold also for SAPFLUXNET observations and ORCHIDEE estimates, which are originally sampled with half-hourly frequency. Here they are presented in (c) and (e) as daily averages and in (d) and (f) as weekly ones over a year. The reported errors in (b) are the standard deviations of the daily samples for each week.

was no overlapping between their time-series, and this is one of the reasons why an indirect correlation through meteorological observations has been analyzed.

4.1.2 Transpiration integration methods

For each site SAPFLUXNET provides a set of single plant half-hourly transpiration measurements, in terms of vapour flux (usually around $1 \div 10\ kg\,m^{-2}s^{-1}$). The amount of plants simultaneously under observation is very variable, depending on the site and period considered. For example, FR-Pue measurements extend over the whole year, being the stand mostly composed by evergreen broad-leaved trees, which present a non-negligible transpiration also during the winter. On the contrary, RU-Fyo, FI-Hyy, US-UMB and US-UMd are under observation only during the growing season, or even for shorter periods in the case of the US sites. Some sites, for instance the US sites and FR-Pue, have an excellent coverage, with more than 20 trees belonging to diverse species simultaneously measured, while other sites observe less than 10 trees. In FI-Hyy only 4 trees are sampled, and in most of the time series only one or two are present due to missing data.

It is immediately clear how, in order to obtain a reliable up-scaling towards the canopy transpiration representation, a careful integration has to be performed. Fortunately, SAPFLUXNET metadata provide information about the stand (in particular the total basal area[1] and the stand plants density), its composition (basal area percentage occupied by each species) and the characteristics of all the plants under observation for the chosen sites (amongst these also the single plant diameter at breast high is reported, which can be used to estimate the plant individual basal area). By combining this information it is possible to properly weight the contribution of each plant and produce a full canopy transpiration estimate while discarding time steps where the coverage is not complete enough to guarantee a reliable result. This integration method has been referred to as Basal Area Integration (BAI). However, these metadata are not always available, therefore it is interesting to evaluate how the BAI method performs with respect to a simple arithmetic mean, which is the only up-scaling method one can carry out for sites where stand composition information is missing. This approach has been denominated Brute Force Approach (BFA).

The rejecting mechanism, which has been devised by the author, prevents the algorithm from producing estimates whenever the available data do not allow to obtain a statistically relevant description of the stand. This happens when more than half of the stand composition is not represented, i.e., when it is not possible to describe the behaviour of more than 50% of the trees in the stand because

[1] Basal area is the cross-sectional area of trees at breast height.

there are no measurements of any plants belonging to those species. The species percentage in the stand is computed from Equation 4.1:

$$p^s = \frac{p^s_{\mathcal{B}}}{\bar{\mathcal{B}}^s} \cdot \frac{\rho^{tot}_{\mathcal{B}}}{\rho^{tot}} \tag{4.1}$$

where p^s represents the percentage of plants belonging to species s in the stand, $p^s_{\mathcal{B}}$ the percentage of basal area $\mathcal{B}$ occupied by the species s, $\rho^{tot}_{\mathcal{B}}$ the total basal area per surface unit (dimensionless), $\bar{\mathcal{B}}^s$ the average basal area $\mathcal{B}$ of plants observed in the dataset belonging to species s (m^2), and finally ρ^{tot} the stand plant density (plants per m^2).

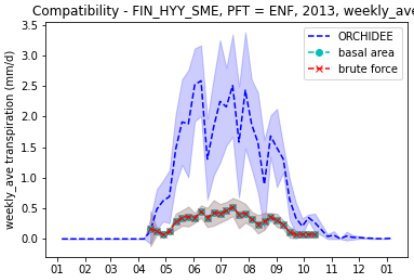
Once p^s is defined, the rejecting algorithm proceeds to evaluate the datasets time step by time step, checking that in each one the set of available measurements includes trees belonging to enough species to reach the acceptance threshold. The actual number of plants for each species is not taken in consideration here: as far as one plant is present its measurement is used to obtain a re-scaled estimate of the whole species contribution, given the species percentage p^s in the stand. However, this approximation can lead to unrealistic results whenever a small number of measures are available and their behaviour is atypical. This is the case shown in Figure 4.3a, where only one plant is under observation and its behaviour seems to be much less intense than ORCHIDEE predicted one. However, it is not possible to know in principle which between the plant measurements or ORCHIDEE estimates are more correct, and an epistemic point of view suggests to put more trust in the observations rather than the model. By observing other results from PFT7 stands like Figure 5.8, where a solid amount of plants is under observation, it appears that ORCHIDEE tends to overestimate transpiration also for this site, even if not as much as in the previous case. This difference may suggest that also an atypical behaviour of the single plant considered in Figure 4.3a is present, or some other phenomenon is not correctly being described causing a wider error in the estimates.

The canopy transpiration T_{can} ($mm\,d^{-1}$) at time step t is finally given by:

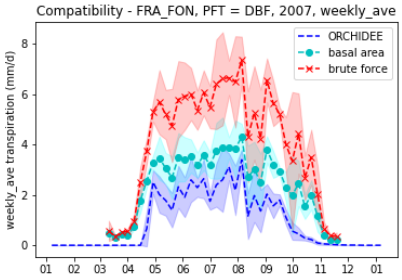
$$T_{can}(t) = \sum_s \left\{ \frac{\sum_{i \in s} \frac{T_i(t)}{\mathcal{B}_i}}{N_s(t)} \cdot \bar{\mathcal{B}}^s p^s \right\} \rho^{tot} \tag{4.2}$$

Here the new quantities are: the transpiration measure of the i^{th} tree of species s at time t $T_i(t)$ ($mm\,d^{-1}$), the total number of trees under observation belonging to species s at time t $N_s(t)$ and $\mathcal{B}_i$, the basal area of the i^{th} tree (m^2).

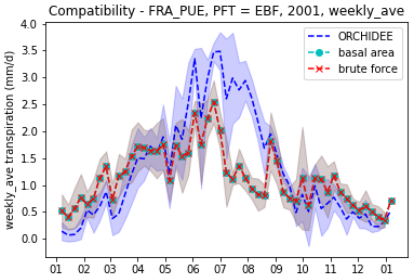
The results of both BAI and BFA have been compared, using ORCHIDEE estimates of transpiration as a reference (Figure 4.3).



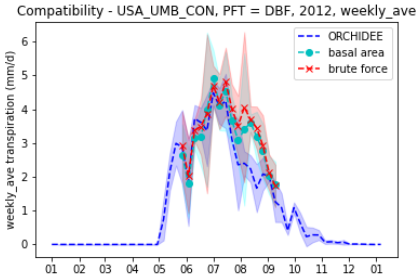
(a) *FI-Hyy data comparison*



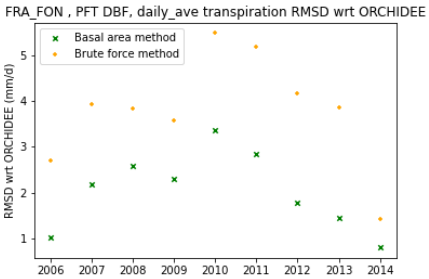
(b) *FR-Fon data comparison*



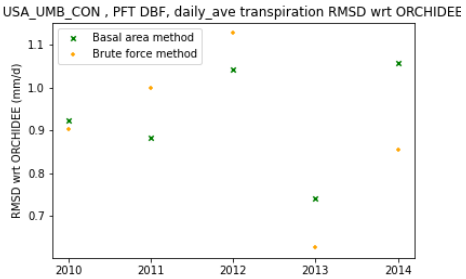
(c) *FR-Pue data comparison*



(d) *US-UMB data comparison*



(e) *FR-Fon error comparison*



(f) *US-UMB error comparison*

Figure 4.3: Subfigures (a) to (d): comparison of integration methods and ORCHIDEE simulations. In (a) it is possible to observe how the presence of a reduced set of measurement (in this case only 1) can lead to huge errors, no matter which method is used, due to the influence of atypical behaviour of the observed plant. In (b) the number of plants under observation is still small (only 8 plants), nevertheless the BAI considerably improves the final result. When increasing the data coverage, the two methods become more and more accurate and give similar results, as observed in (c) and (d). Subfigures (e) and (f): Root Mean Square Deviation (RMSD) of integrated transpiration data with respect to ORCHIDEE simulations, computed over single-year data. The aforementioned improvement of BAI on poorly-sampled sites is independent of the year considered (e), while in case of better coverage (f) it is not possible to clearly define which method works better. The reported errors in (a), (b), (c) and (d) consist in the standard deviation of the daily samples for each week.

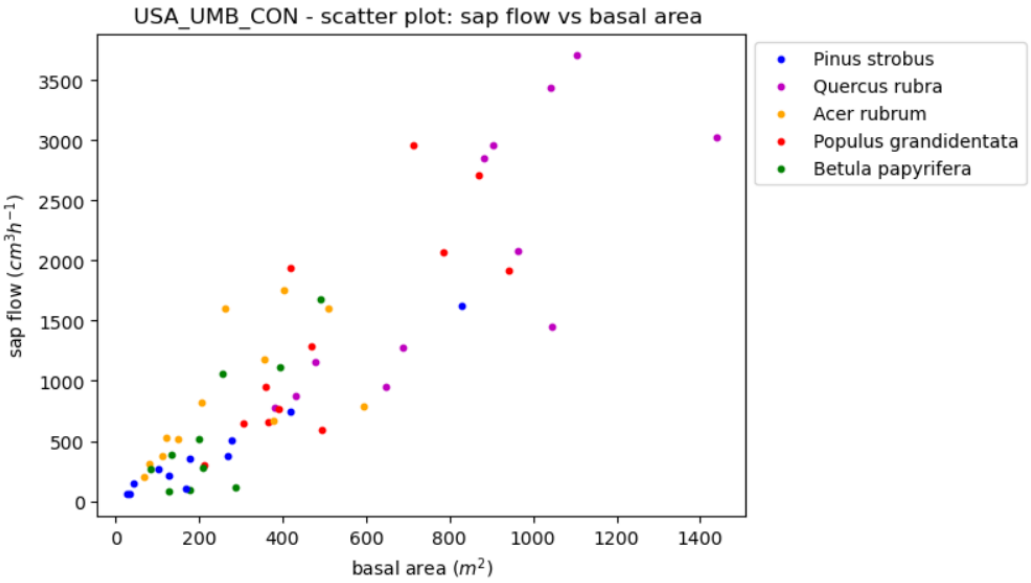


Figure 4.4: Scatter plot of average sap flow and basal area for all the plants in US-UMB. The color represents the species the plant belongs to. The hypothesis of a linear relation between the two quantities is compatible with the observed trend, with different slopes depending on the species considered.

The comparison does not allow to clearly assess which method performs better in full generality, due to the specificity of each site and the natural variability of each set of measurements. However, it seems possible to observe how for poorly-sampled data (as in the FR-Fon site) the BAI re-scaling consistently improves the agreement with respect to ORCHIDEE estimates. This is not the case when the quality of the data is extremely poor (as for FI-Hyy in 2013, where only one plant is sampled and its behaviour is quite atypical), or when the coverage is sufficiently good to properly describe the whole stand behaviour without the need of a re-scaling procedure (that is the case of US-UMB and FR-Pue).

Since BAI is physically motivated and guarantees a further protection from statistical errors due to a poor data coverage, it has been chosen as the reference integration method for SAPFLUXNET data in the following steps of the research. The two biggest criticalities of this method are the assumption of a linear dependency between sap flow and basal area and the approximation of neglecting the plant growth over the years. These assumptions are reasonable, but they are strictly dependent on the local PFT and site. The relation between basal area and sap flow has been investigated, and the linear hypothesis seems to be consistent in most of the cases, as shown in Figure 4.4 for US-UMB.

Another debatable point comes up when considering the comparison with

ORCHIDEE simulations. It consists in defining whether or not plants belonging to PFTs other than the one prescribed in ORCHIDEE, but nonetheless present in the real forest composition, should be considered in the computation. From one point of view, the model is built to simulate a certain stand composition (which can consist in a single PFT or a mix of different ones), and comparing data with a different one would inevitably lower its predictive power. On the other hand, the use of PFTs aims to approximate and simplify the description of the actual stand composition, even if it would include a certain amount of trees not belonging to the predominant species identified by the PFT itself.

In order to investigate these aspects two different experiments have been performed. First, two sets of ORCHIDEE simulations have been run, one presenting the FLUXNET2015 prescribed PFT, and the other one using a mix of different PFTs based on the real stand composition given by SAPFLUXNET stand data. It turned out the model exhibits little or no difference between the two configurations. The same comparison has been done between SAPFLUXNET data, where BAI has been performed first including only the species belonging to the prescribed PFT (normalizing the stand composition in order to replace the percentage of excluded species) and then considering all the species with no distinction in terms of PFT belonging. The major differences observed in this comparison are the ones showed in Figure 4.5, where the relative discrepancy corresponds to an almost uniform upward shift of about 7% of single-PFT-integrated data with respect to the more inclusive one. The negligible effects observed in this analysis are a validation of the prescribed PFT choice. Since the computational cost of a mixed-PFT simulation is higher than a single-PFT one, only this latter configuration has been used in the following.

4.1.3 Sun-induced fluorescence observations

SIF data available from TROPOMI satellite observations have been aggregated at a daily sample frequency and different options in terms of spatial resolution. In the present research the dataset with the $0.1^\circ \times 0.1^\circ$ resolution has been used, the closest to the chosen resolution of ORCHIDEE single site simulations ($0.05^\circ \times 0.05^\circ$). The cloud coverage used in the retrieval algorithm has been set to 50%, since SIF estimates are degraded above by the masking effect of clouds.

As shown in Figure 4.2 and Figure 4.6 the data acquisition algorithm produces some unrealistic negative values. According to the literature, this can be related to calibration issues or uncertainties in the retrieval algorithm[14]. Although these values could easily be identified and discarded, it has been decided to keep them in order to maintain a comparability with other analogous databases and not to create biased means. Moreover, by observing Figure 4.2b and Figure 4.6b it

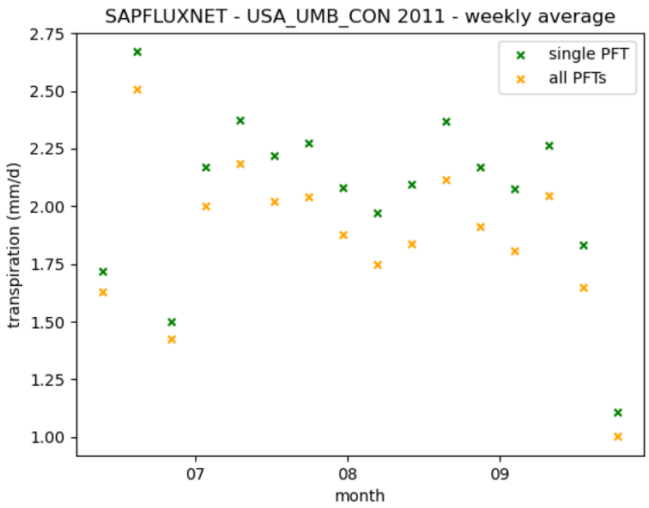
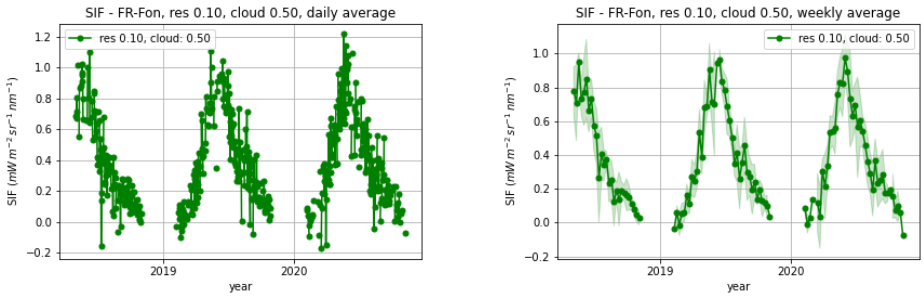


Figure 4.5: Comparison of SAPFLUXNET integrations including all PFTs and only the main one in data from US-UMB in summer 2011. In green the single-PFT-integrated data, in orange the ones including all the species. It is the major discrepancy observed amongst all sites and seasons, amounting to $\sim 7\%$.

is interesting to see how using weekly averages instead of daily observations not only reduces the noise of the data and produces a clearer signal, but also removes almost all the negative data points, retrieving a physically meaningful signal. For these reasons only weekly mean values have been used in the following.

GOSAT and GOME-2 satellites observations from Jeong et al.[36] (Figure 4.7), Joiner et al.[37] and Walther et al.[38] have been visually compared with TROPOMI data (in Figure 4.10). By this comparison it has been possible to assess TROPOMI correct observation of the effect of phenology on plant seasonal cycle. Indeed, deciduous forests (PFT6, FR-Fon and US-UMB) present a steeper increase in SIF during the start of the season, when leaf onset occurs, and a slower and steady decrease after the peak. Also, their emission intensity is consistently higher than evergreen trees one, as expected. On the other side, evergreen needle-leaf forests (PFT7, FI-Hyy and RU-Fyo) and evergreen broad-leaved ones (PFT5, FR-Pue) show a more symmetrical and smooth seasonal curve.

Looking at the US-UMB site, a 4-6 week phase delay of leaf onset, season peak and dormancy emerges in TROPOMI observations with respect to the average behaviour between 1999 and 2006 computed by Joiner et al.[37] from GOME-2 and GOSAT data (Figure 4.8). However, the leaf onset mostly depends on the local meteorology over a period of a few weeks, and therefore this discrepancy causes no concern, as TROPOMI observations of growing seasons only cover 2019 and 2020.



(a) *daily values* **(b)** *weekly means*

Figure 4.6: Comparison between TROPOMI SIF daily data (a) and their weekly average with shaded uncertainties (b) between May 2018 and December 2020 from the FR-Fon site. Re-sampling into weekly frequency not only improves the readability and comparability of the signal, but also corrects the presence of outliers with negative intensity measure due to calibration and retrieval algorithm structural uncertainties[14]. The reported errors in (b) consist in the standard deviations of the daily samples for each week.

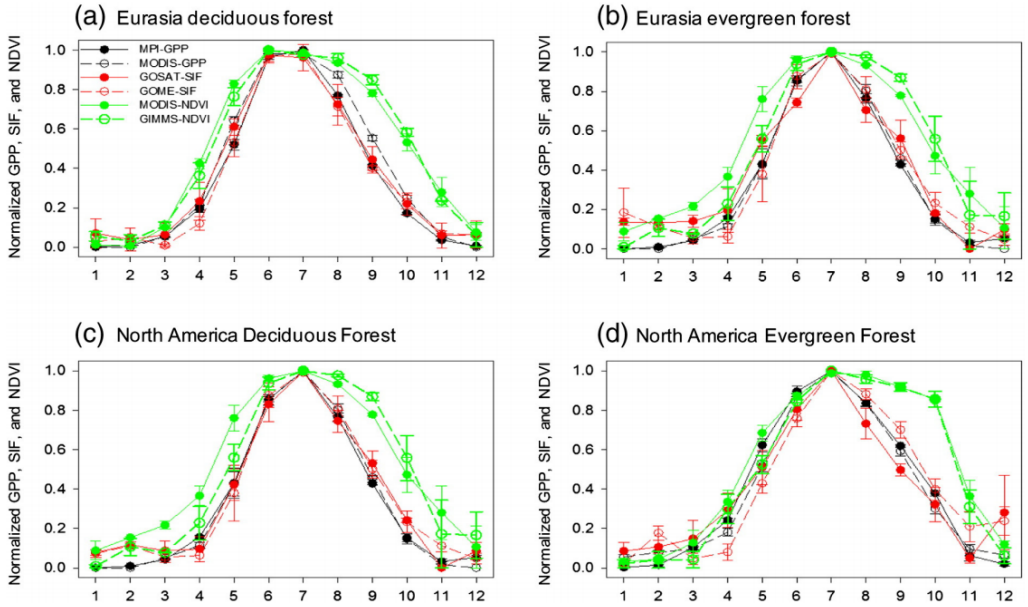


Figure 4.7: Normalized mean seasonal cycle of area averaged GOSAT SIF, GOME SIF and other greenness indices over northern temperate and boreal forests for the period 2010–2012, in deciduous and evergreen forests over Eurasia (a, b) and over North America (c, d). Source: Jeong et al.[36]

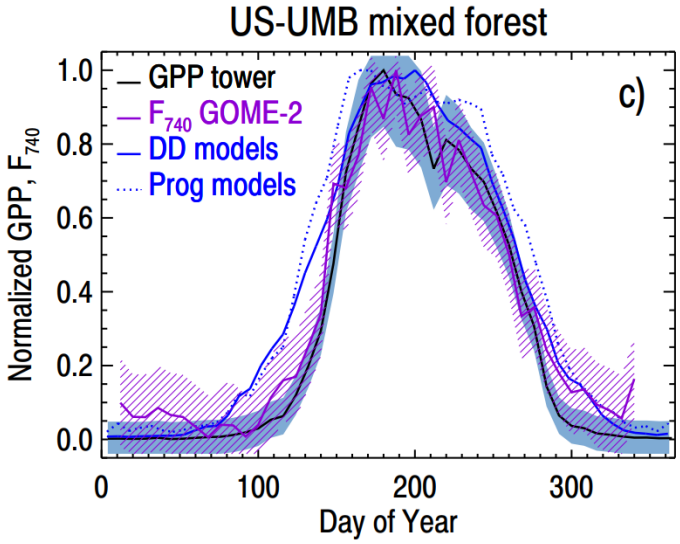


Figure 4.8: GPP and SIF measurements and multi-model ensemble estimates as day-of-the-year averages over US-UMB in the period 1999-2006. Source: Joiner et al.[37].

In terms of signal amplitude, TROPOMI estimates for the site RU-Fyo have been compared with those from GOME-2 by Walther et al.[38] (who report non-normalized average values between 2007 and 2012), showing an uniform underestimation of about 30% by TROPOMI data with respect to GOME-2 ones (Figure 4.9). This discrepancy is not caused by a difference in the treatment of negative values, because they are not uniformly distributed amongst the whole period, but it could be related to different calibrations and retrieval algorithms, or to different meteorological conditions.

4.1.4 Study of correlation

The link between transpiration and SIF through photosynthetic activity is well-known and physically reasonable. A certain degree of correlation is expected to be observed when comparing the behaviour of the two on a common stand and period. In fact SIF and transpiration are expected to covary, since they share some common abiotic drivers[22]. TROPOMI and SAPFLUXNET data do not cover any common period (TROPOMI starts from 2018, while SAPFLUXNET ends in 2016), so the correlations with their common drivers are the only tools which can be used to qualitatively assess if a covariance between transpiration and SIF exists in the observations.

The most commonly used correlation measure is Pearson correlation coeffi-

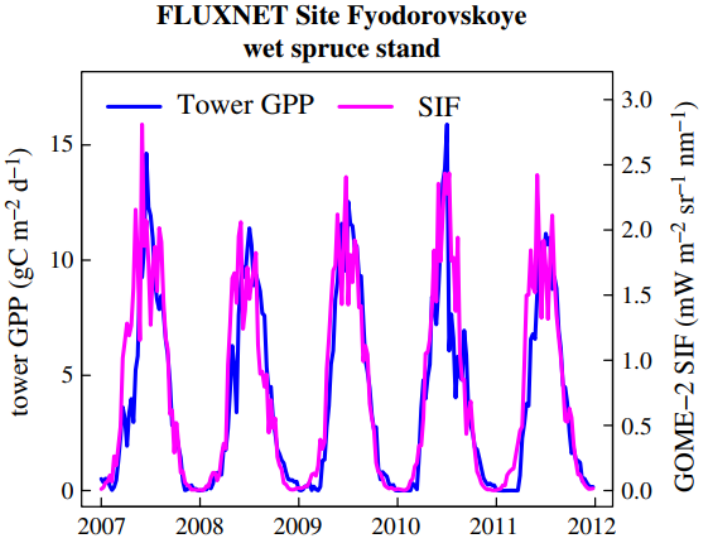


Figure 4.9: Comparison of flux tower GPP observations and the satellite SIF measurements from GOME-2 over Ru-Fyo. Source: Walther et al.[38].

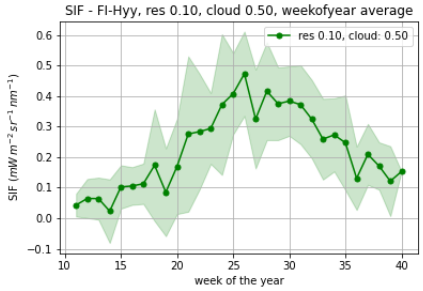
cient, defined for two variables x and y as:

$$\rho(x, y) = \frac{cov(x, y)}{\sigma_x \cdot \sigma_y} \tag{4.3}$$

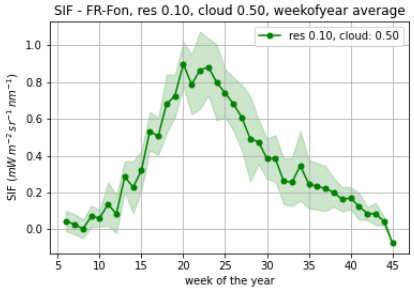
where $cov(x, y)$ is the covariance between x and y, while σ_α represent the standard deviation associated with the variable $\alpha \in \{x, y\}$.

If a numerical relation between two variables of interest is needed, using their correlation coefficient will give misleading results whenever another confounding variable (often called "controlling variable") numerically related to both the variables of interest is present. This could easily be the case when considering biological systems depending on environmental variables. Therefore an alternative measure is defined: the partial correlation coefficient. It aims to evaluate the degree of association between two random variables, while taking into account the effect of a set of controlling random variables. The formal definition is: "The correlation between the deviations of the values of a variate from their least square estimates by a regression function linear in terms of an external set of variates, with the corresponding deviations of another variate from its own regression function linear in the same external set."[B1].

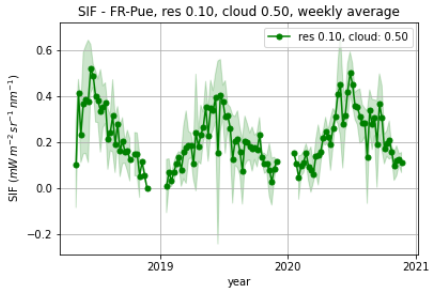
Using a more explicit mathematical formalism, the partial correlation between two random variables can be computed using linear regression. Define two random variables x and y, and a set of k control random variables z, belonging to $\mathbb{R}$. x_i, y_i and $\mathbf{z}_i$ represent the values of the i^{th} observation amongst the N sampled



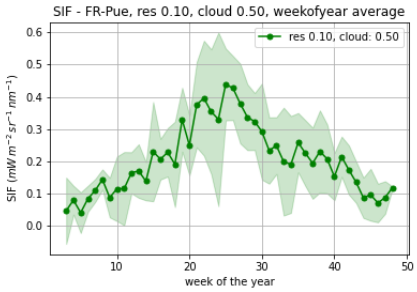
(a) *FI-Hyy week of the year*



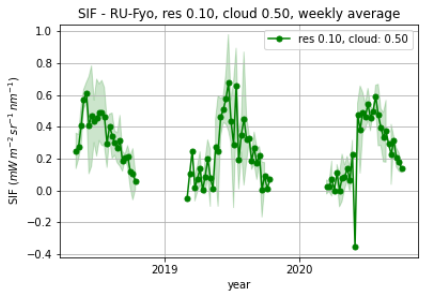
(b) *FR-Fon week of the year*



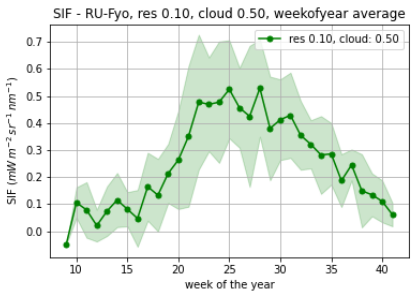
(c) *FR-Pue weekly means*



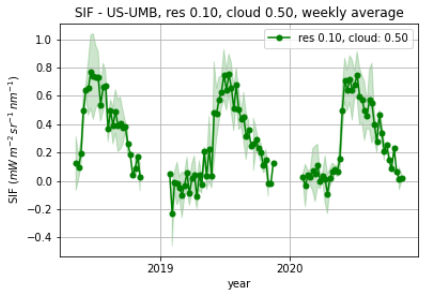
(d) *FR-Pue week of the year*



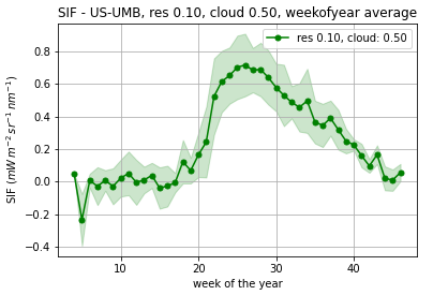
(e) *RU-Fyo weekly means*



(f) *RU-Fyo week of the year*



(g) *US-UMB weekly means*



(h) *US-UMB week of the year*

Figure 4.10: TROPOMI SIF observations and averages. The reported errors in the graphs consist in the standard deviations of the daily samples for each week or group of weeks (depending whether the weekly mean or the week-of-the-year mean is considered).

from their joint probability distribution.

The k dimensional linear regression coefficients $\mathbf{w}_x$ and $\mathbf{w}_y$ of the variables x and y with respect to $\mathbf{z}$ are given by:

$$\mathbf{w}_x = arg\,min_{\mathbf{w}} \left\{ \sum_{i=1}^{N} (x_i - \langle \mathbf{w}, \mathbf{z}_i \rangle)^2 \right\} \tag{4.4}$$

$$\mathbf{w}_y = arg\,min_{\mathbf{w}} \left\{ \sum_{i=1}^{N} (y_i - \langle \mathbf{w}, \mathbf{z}_i \rangle)^2 \right\} \tag{4.5}$$

where $\langle \cdot, \cdot \rangle$ represents a scalar product. Defining the residuals $R_{x,i}$ and $R_{y,i}$ as

$$R_{x,i} = x_i - \langle \mathbf{w}_x, \mathbf{z}_i \rangle \tag{4.6}$$

$$R_{y,i} = y_i - \langle \mathbf{w}_y, \mathbf{z}_i \rangle \tag{4.7}$$

the partial correlation of variables X and Y with control variables $\mathbf{Z}$ is given by:

$$\rho_{\mathbf{z}}(x, y) = \frac{\sum_{i=1}^{N} R_{x,i} R_{y,i}}{\sqrt{\sum_{i=1}^{N} R_{x,i}^2} \cdot \sqrt{\sum_{i=1}^{N} R_{y,i}^2}} \tag{4.8}$$

Both coefficients have been used to evaluate the correlation between transpiration and SIF with respect to air temperature, soil water content, vapour pressure deficit and air moisture, and their comparison is reported in section 5.2.

4.2 ORCHIDAS: a tool for sensitivity analysis and optimisation

Parametrization is a key aspect in models design. It is especially complex when dealing with a wide class of interacting phenomena, in many different external conditions, and producing several outputs which have to give the most accurate description of each process.

In ORCHIDEE the parametrization of the model can be obtained from values known in the literature or by optimizing the simulation results against a set of observations. The optimization procedures used to obtain the best parameter estimates are referred to as ”experiments” in the following. These experiments are performed through the ORCHIDEE Data Assimilation System (ORCHIDAS)[39]. This software allows to optimize a set of parameters by using a machine learning technique (specifically the Genetic Algorithm[B2]) to find the combination

of parameter values producing the most accurate estimate of a given observed quantity. This algorithm computational cost depends on both the length of observation time series and on the number of optimized parameters, and can be run simultaneously for multiple sites and using multiple variables as references to obtain a more general result.

Since not all the parameters influence every phenomena, and not with the same intensity, it is important to determine which parameters have to be included in the optimization procedure. This information is obtained through Sensitivity Analysis (SA), which is a first step to identify the key parameters before aiming at calibrating them. Indeed, focusing on the key model parameters for calibration both limits the computational cost of optimization and the risk of over-fitting.

SA is carried out by a built-in tool, which receives as inputs the set of candidate parameters to analyse and the variable of interest that the optimisation procedure aims at reproducing. Both in SA and optimization each parameter is characterized by a default value and a range of variation around that. There is no standard procedure in this definition, which is one of the most delicate aspects of the model, especially when introducing a new parameter. Unfortunately, the choice of a default value from literature is not always possible, since often the study of biological phenomena are strictly related to the specific environment and conditions the experiments are performed into, and they can not be properly employed in more general situations. When a parameter default value is roughly approximated and needs to be refined, the choice of a proper variation range is crucial. Taking an excessively narrow interval does not allow the optimization to really explore the influence of the parameter in the model, while giving too much freedom can drive the system towards a numerically optimal configuration which realizes completely un-physical dynamics. Usually these pathological behaviours are identified by a saturation of some parameters value at the extremes of the range, as it is shown in section 5.3.

The first selection of parameters included in the SA of GPP and SIF is based on previous SA performed with ORCHIDEE[40][41]. Experiments regarding transpiration measurements assimilation and SA had never been performed on ORCHIDEE, so its choice of the parameter set has followed a direct analysis of the parameters involved in transpiration calculation (see Figure 3.1). In general, the goal has been to include in the SA all the parameters that can possibly play a role in the phenomena related to transpiration, GPP, and SIF, in order to be sure not to exclude any important one. These parameters can be involved in the representation of photosynthesis, phenology, carbon allocation, conductance, respiration, biomass, or soil hydrology for example[41]. SA is quite fast to run, especially if compared with the optimization algorithm, and evaluating large sets of parameters is not prohibitive in terms of computational cost. Whenever little or no information was available to constrain the values of a parameter, an arbitrary

variation range of 40% of the parameter default value itself has been used.

Depending on the configuration, for each parameter a variable number of values are sampled within the variation range and are used to run independent simulations, in order to assess the influence of each trajectory (i.e., set of parameters, since the path in the space of parameters defines a specific value for each one of them) on the target quantity. The Morris method[42] has been used for the sensitivity analysis as it is relatively low time-consuming and enables to rank the parameters by importance. This qualitative method requires only a low number of simulations $(p+1)n$, with p the number of parameters and n the number of random trajectories generated, which has been set to 10 in this context. The computational time required from each SA to be completed is around 3 hours per site.

The optimization algorithm has a similar structure as well, but in addition it requires a set of data to optimize the model against. This procedure is often referred to as "data assimilation". In this algorithm the values assumed by the sets of parameters are not randomly sampled. Instead, the machine learning Genetic Algorithm is used to find the combination of parameters which better reproduces the observed behaviour of the variable used as a reference for the optimization (GPP, transpiration or SIF in this case). Its computational time is about 6 hours per optimization, depending on the length of the timeseries of the assimilated quantity and the number of parameters involved.

The goal of the research is to investigate how the optimization against these 3 different quantities improves the predictive power of the model. The availability of data is quite different amongst the 3 databases. GPP in FLUXNET2015 and transpiration derived from SAPFLUXNET cover usually a similar number of years. However, some SAPFLUXNET sites only reports data during the growing season, therefore the overall number of entries can still be quite different. These discrepancies are negligible when considering the TROPOMI SIF dataset, which covers the period between May 2018 and December 2020, with no gapfilling and showing frequently missing data during the cold season. Such an inhomogeneity between the databases is an issue, since the optimization performance, as with every machine learning technique, strongly depends on the amount of training data. In order to preserve comparability between the diverse optimizations, only 3 years of observations have been used as training data for all the experiments. This choice drastically reduces the performance, but it allows to compare the effective potential of each new database in terms of model optimization.

A neat dependence on PFTs of performances of ORCHIDEE estimates emerged from the comparison of ORCHIDEE transpiration estimates and SAPFLUXNET data, as shown in subsection 5.1.2. For this reason, the optimization procedure has been carried out on groups defined by PFT and operating a multi-site optimization. This approach is also meant to simultaneously optimize those parameters

which are PFT-dependent together with all the sites with that PFT and not independently one from the other.

The computational time of the optimization algorithm depends on the number of sites considered in the procedure. The FR-Pue single site optimization takes around 6 hours to be completed, while the 2-3 sites ones (respectively PFT7 and PFT6) take around 14 hours.

Chapter 5

Results and Discussion

5.1 Transpiration-related observations and discrepancy from ORCHIDEE estimates

5.1.1 Comparison between observations

Transpiration, canopy Latent Heat Flux (LE), GPP and SIF are all strictly related to plant photosynthetic activity and the resulting fluxes, and they are sampled with different frequencies. However, on a weekly scale it is possible to compare the average behaviour of these quantities and a good phase agreement is expected to be observed.

Unfortunately the aforementioned lack of overlapping time series from TROPOMI SIF data with the other two databases does not allow a direct comparison between all the quantities simultaneously, as shown in Figure 5.1. In fact, while it is possible to run simulations up to 2020 thanks to the re-analysis data from ERA5, this database does not include any flux, but only meteorological variables.

FLUXNET2015 not only contains a LE estimate from eddy covariance measures, but for some sites it also provides a LE estimate corrected by imposing energy balance closure[17], referred to as LE_fC in the following. LE_fC has been included in the comparison to have a qualitative idea of the consistency of LE data along the period.

In Figure 5.2 it is possible to observe the effects of the energy balance correction on LE and its correlation with transpiration. Indeed the early months of 1998 appearing in Figure 5.2a exhibit a behaviour which can not be related to vegetation activity, since out of the growing season the transpiration is almost quiescent. Evaporation being equivalent to LE (they only differ in terms of unit of measures), this trend could be related to snow sublimation after the end of

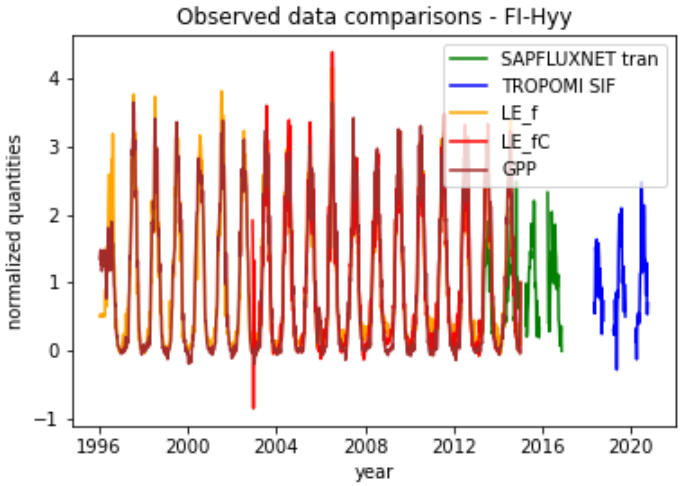
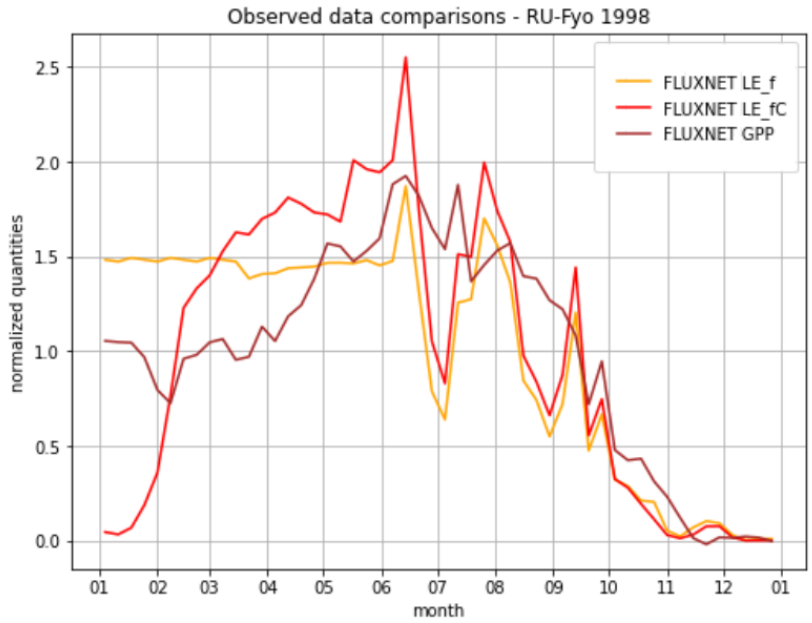


Figure 5.1: Plot of normalized values of transpiration-related quantities for the whole period of coverage of each database on site FI-Hyy. Amongst FLUXNET2015 estimates (LE_f, LE_fC and GPP), LE_f represents the Latent Heat flux and LE_fC its value when the energy balance closure correction is applied. It is possible to observe the different periods of coverage of each database. The series are normalized with respect to their own mean values.

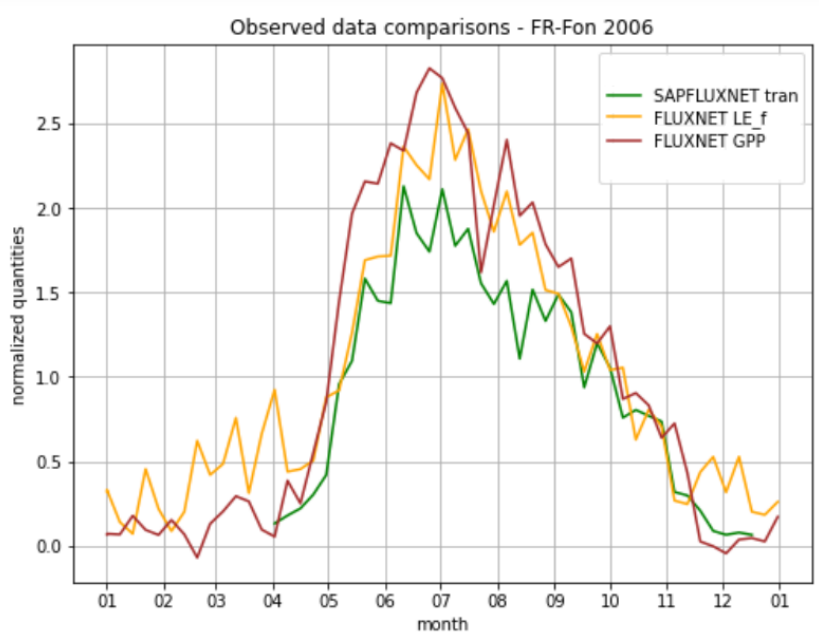
the winter, but this phenomenon can not account for the anomalous values of GPP, and the local meteorological observations do not highlight any substantial difference from the same period in other years within the timeseries. Also, LE and GPP being closely related to ET, it is unlikely to observe values close to the peak during the winter. This behaviour is effectively corrected in LE_fC, which shows a more realistic curve. The study of this phenomenon is outside the scope of the present article and warrants further research. The close phase agreement between transpiration and LE is particularly evident in Figure 5.2b, where their matching is almost perfect, especially around the season peak.

It is important to mention that, the measures being normalized through the seasonal mean, whenever the coverage is not complete on the whole year (for instance transpiration in Figure 5.2b) or a pathological behaviour alters the curve (as LE or GPP suggest in Figure 5.2a) the sample is re-scaled differently from other measurements. Therefore, the only information it is possible to extract from these graphs is about phase agreement, since the positions of peaks and minima are not influenced by the scale.

The results of this analysis are not sufficient to draw a big picture of a common behaviour amongst these quantities. It is clear that they share some drivers, mostly related to their seasonal cycle and the good agreement which emerges



(a) *RU-Fyo 1998*



(b) *FR-Fon 2006*

Figure 5.2: Comparison of normalized values of observations for transpiration, SIF, LE and GPP. It is interesting to observe in (a) how the energy balance correction on LE corrects the un-physical behaviour of the early stage of the year. In (b) the phase agreement between LE and transpiration is particularly evident, showing how these two quantities are strictly correlated.

frequently (for example between GPP and transpiration in Figure 5.3a, and Figure 5.4a), but several discrepancies suggest that other drivers differentiate their behaviour on shorter scales (Figure 5.3b).

Even if it is not possible to directly compare SIF observations to other quantities due to the lack of overlap between the databases, it is interesting to notice the similarity of the patterns observed in Figure 5.4a and Figure 5.4b. Also, FR-Pue being a PFT5 site (which implies a predominance of evergreen broad-leaved trees) at intermediate latitude (with a non negligible incoming radiation even during the cold season), all its observed quantities exhibit non-vanishing values over the whole year.

The comparison has been carried out also for ORCHIDEE estimates of the aforementioned quantities, with the goal of highlighting the most evident differences and common points between observations and simulations.

In Figure 5.5 it is possible to notice two features of the observations which are correctly reproduced in the simulations. First, the presence of a non-zero LE during the cold season, when photosynthetic activity (and therefore most of the transpiration process) is strongly reduced, but snow sublimation (very important for sites with presence of snow) and other ET processes are still active. Second, the trigger condition that has to be reached in order for photosynthesis to be enabled (both RU-Fyo and FI-Hyy are evergreen needle-leaf forests, so no leaf onset is occurring here), which is visible in the sharp transition between a quiescence to a non negligible activity observed for example in the simulation of April 2011 for FI-Hyy.

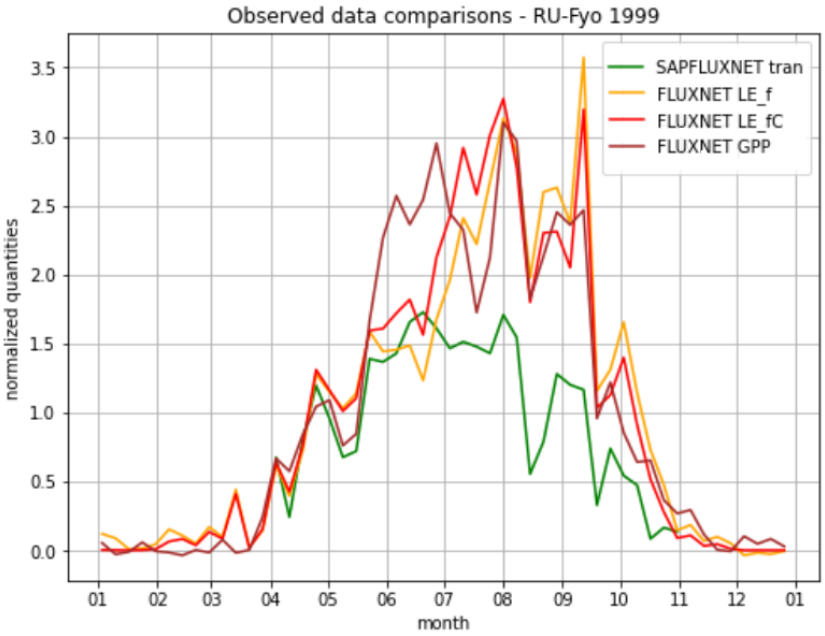
Simulated quantities exhibit similar behaviours, as shown in Figure 5.5, Figure 5.6 and Figure 5.7, proving the consistency of the model in describing these physically related quantities.

Also leaf onset for summergreen PFT can be observed in the simulations, as shown in Figure 5.6. Of course, depending on the meteorology and biome the onset can be triggered in different moments in the season, in this case in mid May 2011 for US-UMB and early April 2011 for FR-Fon.

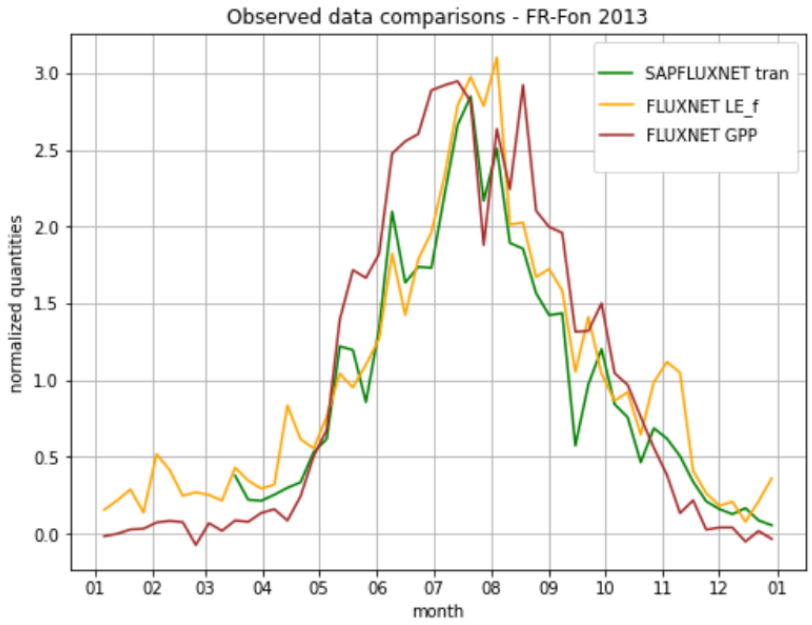
5.1.2 Transpiration matching between ORCHIDEE estimates and SAPFLUXNET observations

Some characteristic behaviours depending on the PFT belonging of the stand have emerged in the comparison between ORCHIDEE estimates of transpiration and SAPFLUXNET-derived observations. For instance, simulations of PFT7 present a distinctive tendency to overestimate transpiration, as shown in Figure 4.3a and Figure 5.8, even if their overall phase agreement is quite good.

Two of the most complex and important features of deciduous species for

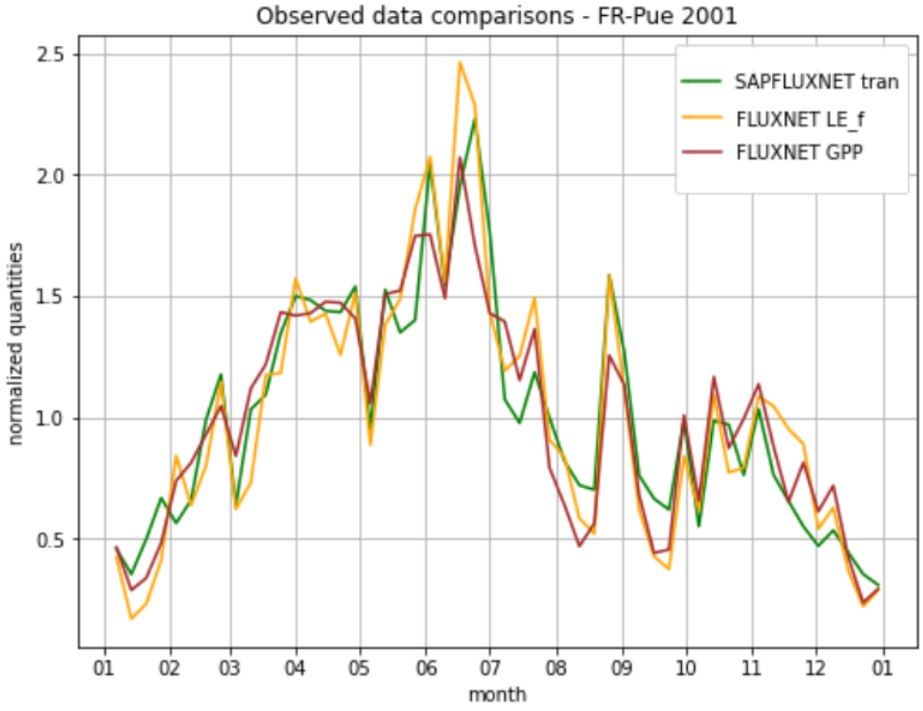


(a) *RU-Fyo 1999*

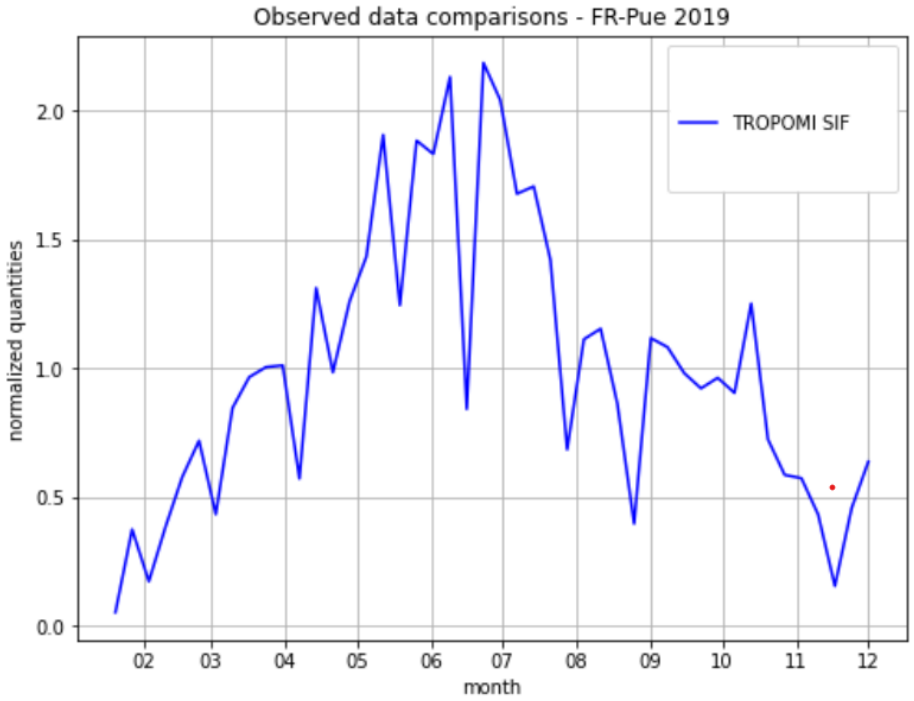


(b) *FR-Fon 2013*

Figure 5.3: Comparison of normalized values of observations for transpiration, LE and GPP. It is interesting to observe in (a) how the phase agreement for GPP and transpiration is very accurate (considering the different vertical stretch). On the opposite in (b) even if the seasonal cycle is clear and the amplitudes seem comparable, the curves behave quite differently.

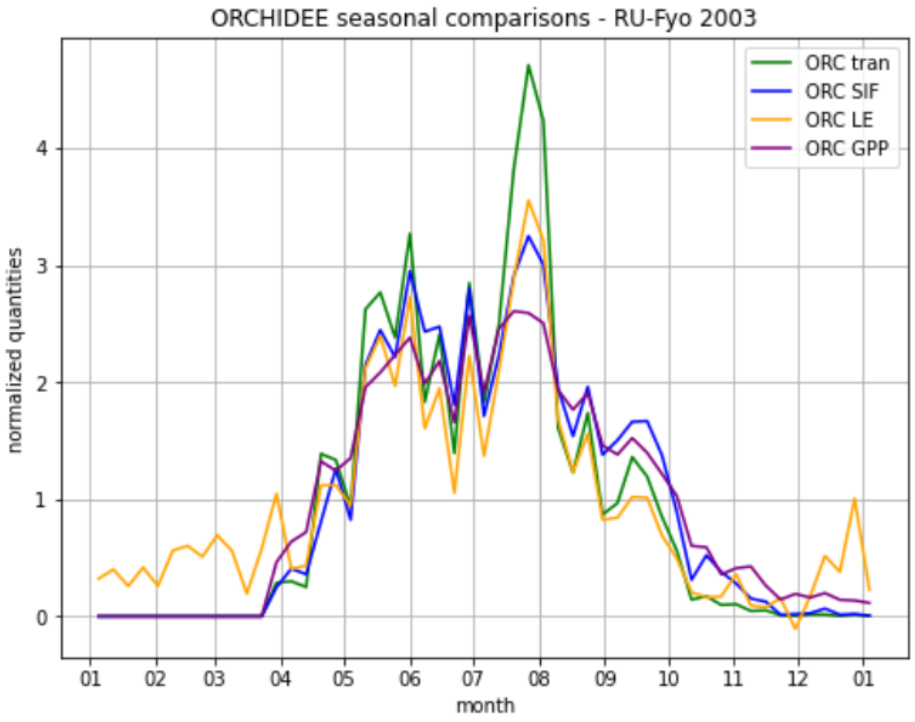


(a) *FR-Pue 2001*

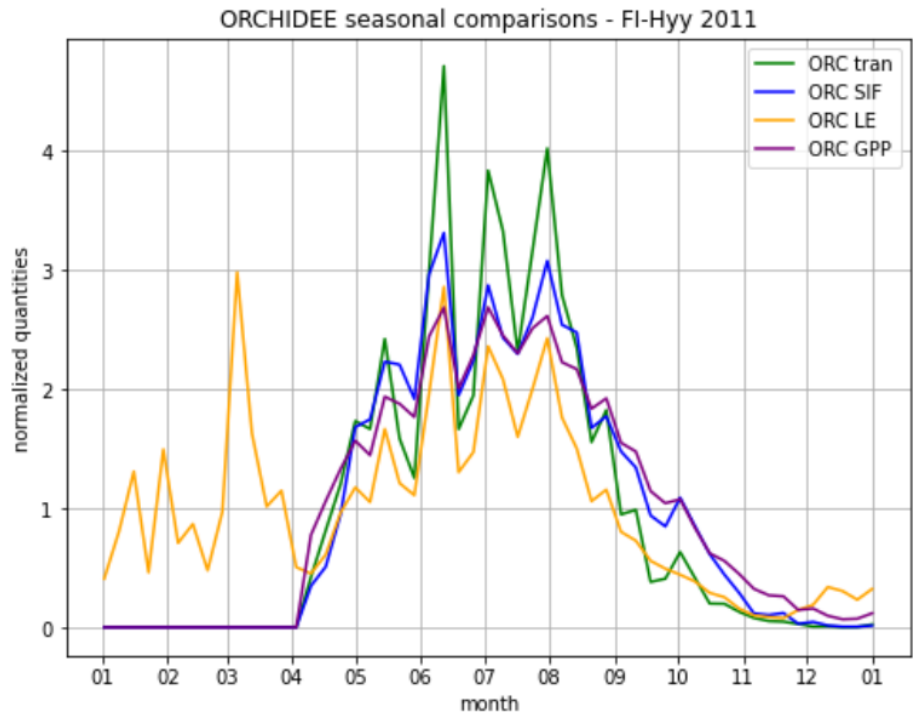


(b) *FR-Pue 2019*

Figure 5.4: In (a) comparison of normalized values of observations for transpiration, SIF, LE and GPP. In (b) the seasonal behaviour of SIF observed in FR-Pue 2019. It is interesting to notice the similarity of the patterns observed in (a) and (b), due to meteorological oscillations and the nature of the local biome, which lead photosynthetic activity to be present along the whole period.

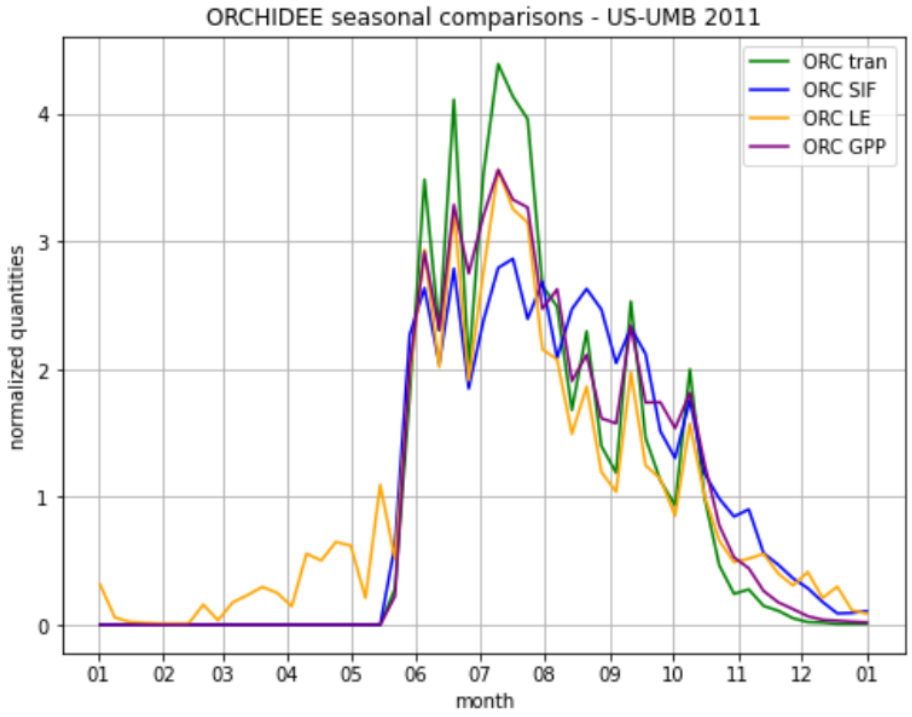


(a) *RU-Fyo 2003*

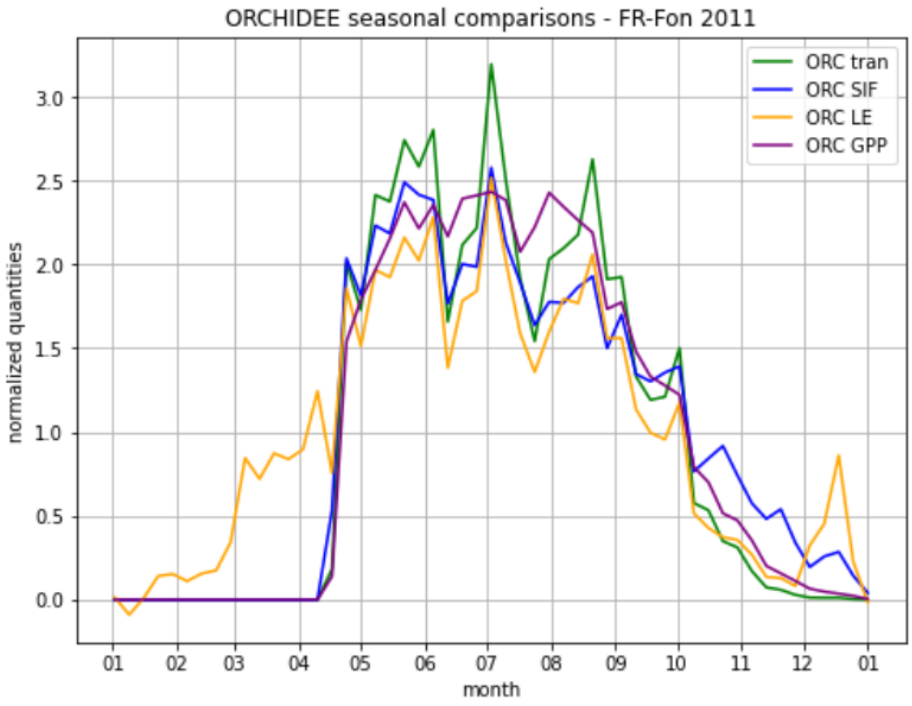


(b) *FI-Hyy 2011*

Figure 5.5: Comparison of normalized values of ORCHIDEE estimates for transpiration, SIF, LE and GPP. The existence of a snow contribution to ET during the cold season and of a photosynthetic activity activation mechanism is shown by the simulated behaviour in the first months of the year.



(a) *US-UMB 2011*



(b) *FR-Fon 2011*

Figure 5.6: Comparison of normalized values of ORCHIDEE estimates for transpiration, SIF, LE and GPP.

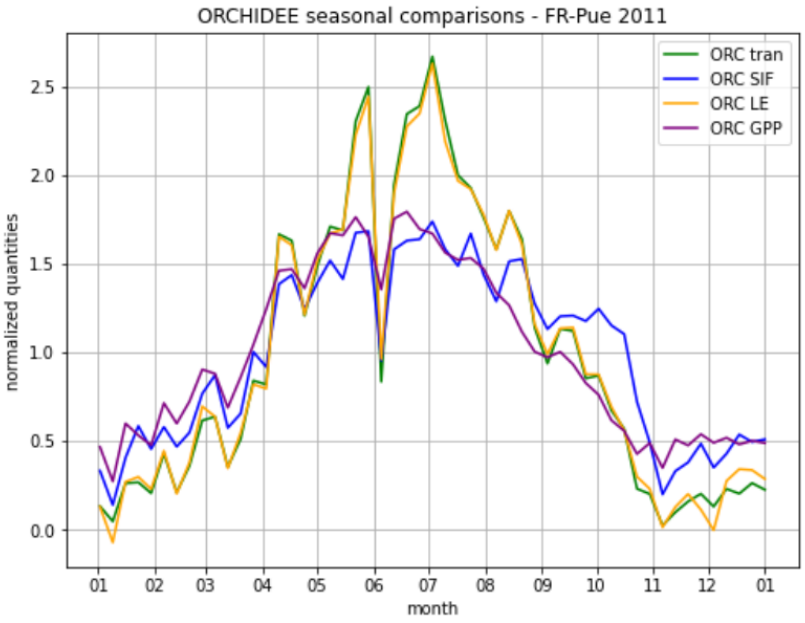


Figure 5.7: Comparison of normalized values of ORCHIDEE estimates for transpiration, SIF, LE and GPP. The aforementioned "everlasting" photosynthetic activity of PFT5 is correctly reproduced by the model.

a model to simulate are leaf onset and senescence. Unfortunately only a very narrow period within the growing season is sampled each year for the US-UMB and US-UMd sites, so it is not possible to draw any consideration regarding the accuracy in the realization of these two phenomena performed by the model from these sites. Also the accuracy of the model with respect to the observations is not completely satisfying for these sites (see Figure 5.9a as a reference).

By comparing the disturbed and control samples of US-UMB it is interesting to observe in Figure 5.9 how the 2 sites, despite them being approximately 10 km apart, present quite different behaviours both in the observations and simulations. In fact, their meteorological forcing files show some differences both in terms of time coverage (US-UMB measurement period is twice longer than the US-UMd one) and values, with data of precipitation and wind speed being substantially different all along the timeseries. This aspect is outside the scope of the present thesis and warrants further research.

The model performance on correctly reproducing leaf onset and senescence can be observed in the other PFT6 site, FR-Fon. In fact, a common behaviour amongst FR-Fon simulations is the presence of a late leaf onset and an early senescence, as the comparison with SAPFLUXNET observations shows in Figure 5.10, coupled with a persistent underestimation of transpiration levels, which seems

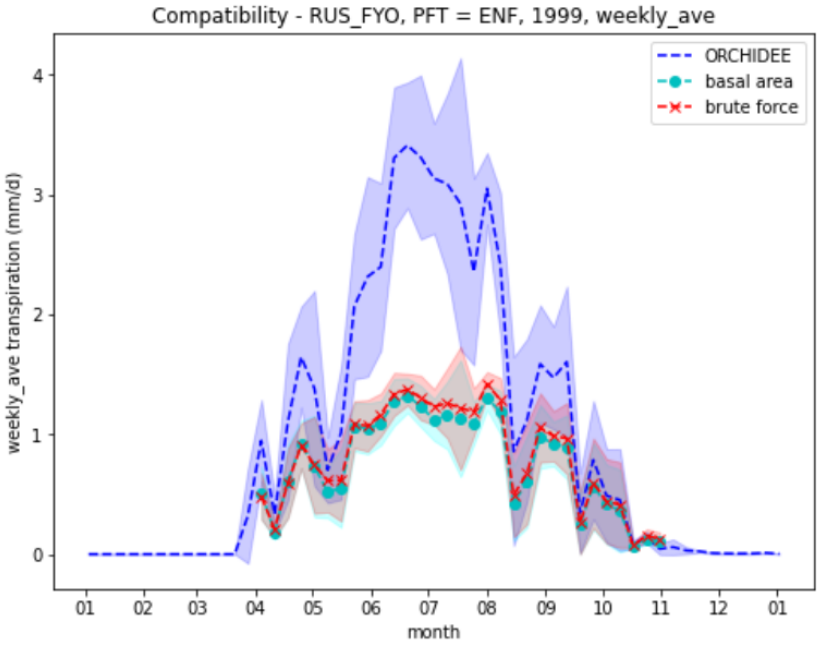
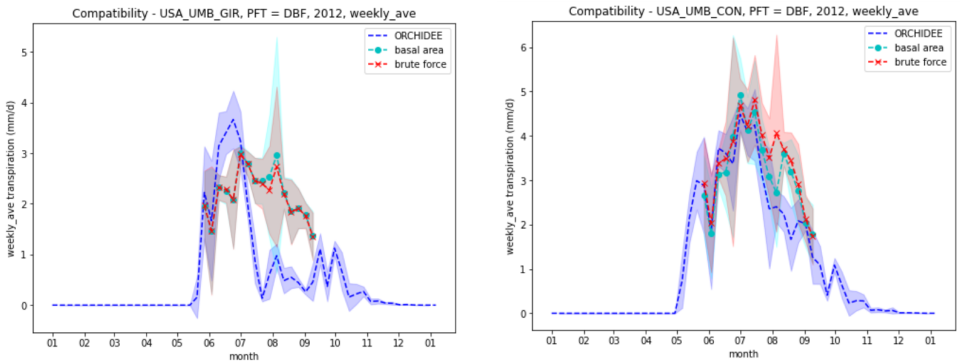


Figure 5.8: Transpiration SAPFLUXNET measures and ORCHIDEE estimates for RU-Fyo in 1999. The reported errors consist in the standard deviations of the half-hourly samples for each week.



(a) *US-UMd 2012* **(b)** *US-UMB 2012*

Figure 5.9: Transpiration SAPFLUXNET measures and ORCHIDEE estimates for US-UMd and UMB in 2012. The reported errors consist in the standard deviations of the half-hourly samples for each week.

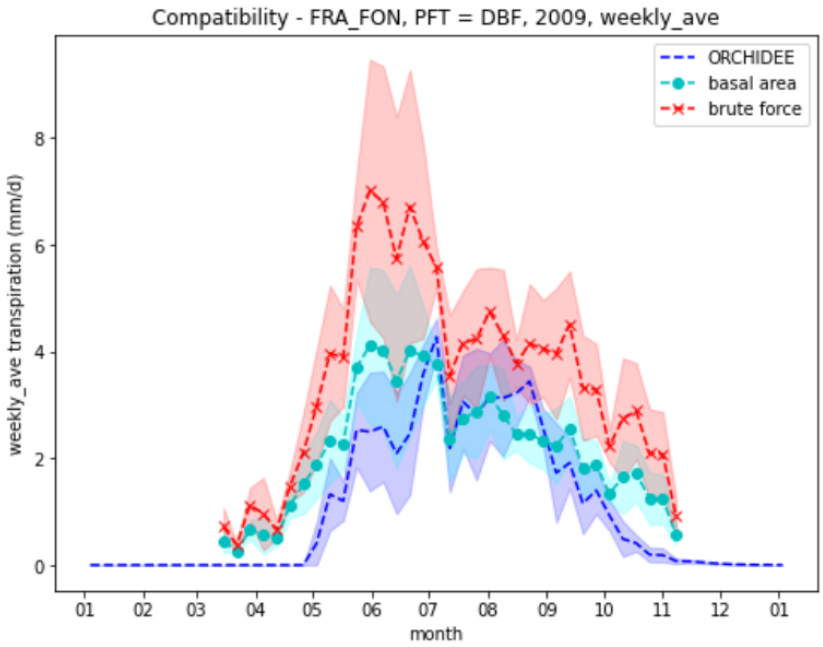
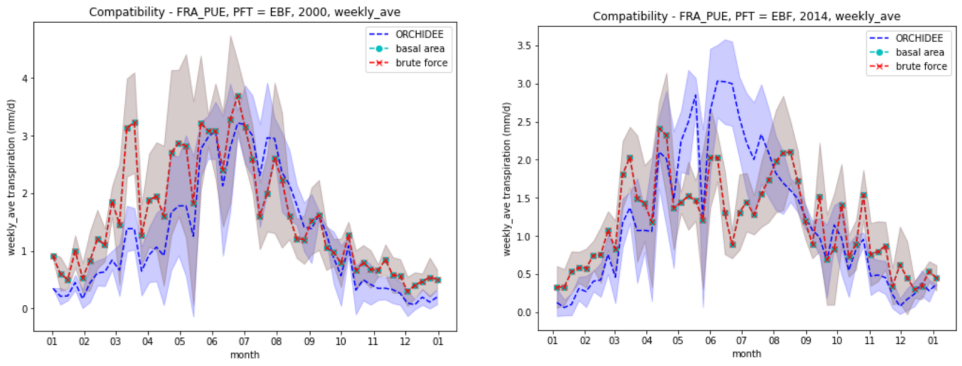


Figure 5.10: Transpiration SAPFLUXNET measures and ORCHIDEE estimates for FR-Fon in 2009. The reported errors consist in the standard deviations of the half-hourly samples for each week.

to suggest ORCHIDEE is not correctly describing the transpiration process with respect to the actual one.

In PFT5 simulations ORCHIDEE almost always reproduces the correct seasonal cycle, as observed in Figure 5.11. The estimates are mostly inside the error range, which on the other side is not a very restrictive condition to meet, since the uncertainties on both measurements and estimates are considerable. However, in some cases (for instance Figure 5.11b) very important features like seasonal peak and strong local minima are not correctly reproduced.



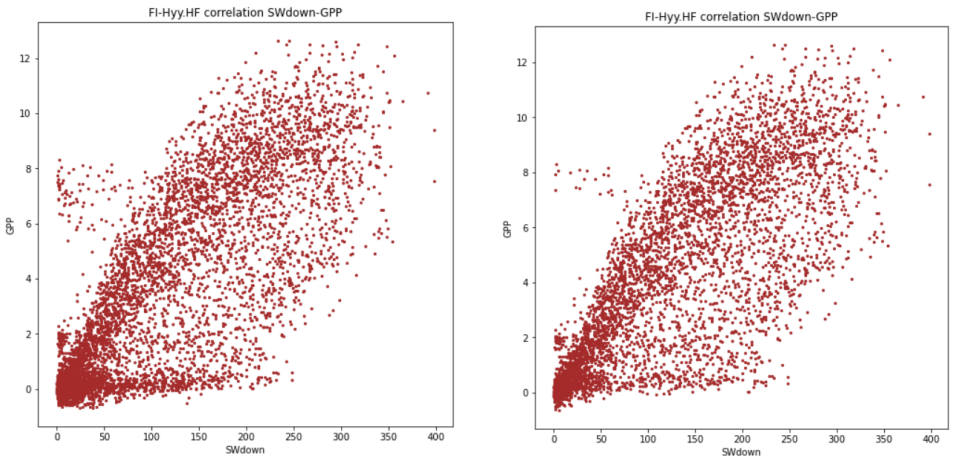
(a) *FR-Pue 2000* **(b)** *FR-Pue 2014*

Figure 5.11: Transpiration SAPFLUXNET measures and ORCHIDEE estimates for FR-Pue in 2000 and 2014. The reported errors consist in the standard deviations of the half-hourly samples for each week.

5.2 Pearson and partial correlations of studied quantities with respect to main meteorological drivers

5.2.1 Qualitative analysis

The most simple and direct way to infer correlations between variables is by the observation of scatter plots. In this case 4 variables from the observation databases (transpiration, LE, GPP, SIF) have been correlated to 4 drivers, namely air temperature T_{air}, near-surface specific humidity of air Q_{air}, downward short wave radiation SW_{down} and Vapour Pressure Deficit vpd. The variables being dependent on several drivers, it is important to realize that each scatter plot represents a projection along one axis of the multi-dimensional ensemble of data-points collected from the meteorological time series. Therefore, some apparent relations could emerge from the projection operation, which do not carry any physical information. For instance, in Figure 5.12a two trends are distinguishable: one is a logarithm-like growth, the other a plateau for almost vanishing values of GPP. While the log-like growth has a physically reasonable interpretation, photosynthetic activity being asymptotically related to the incoming solar radiation, the plateau is not interpretable without considering the high dimensionality of the problem. What is being observed is the presence of many data-points which, even for values of SW_{down} that would normally allow photosynthesis to happen, do not exhibit a significant GPP value because an other variable is blocking the



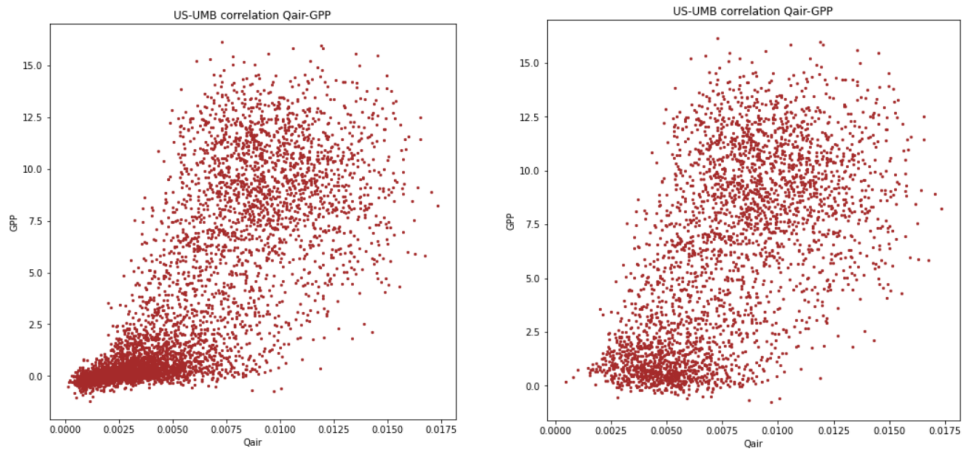
(a) *complete data* **(b)** *cropped data*

Figure 5.12: Scatter plot of SW_{down} (on the horizontal axis) and GPP (on the vertical axis) for FI-Hyy site. On the left the entire dataset is plotted. On the right the data-points below the temperature threshold are removed.

process, for instance water stress or low temperatures.

This is one of the reasons why the data show such a high dispersion and they do not seem to provide precise information about the underlying relation between SW_{down} and GPP, as it is possible to see also in Figure 5.13a for Q_{air} and GPP. Other sources of dispersion are the systematic and random uncertainties in the measurements, and those due to the extraction and pre-processing of fluxes and derived quantities from the observations.

The presence of a temperature threshold for photosynthetic activity is well-known in the literature, and it seems quite evident also from the data (for instance in Figure 5.15a, and also in Figure 5.12 and Figure 5.13, as it will be explained later). These data pose a structural problem in the evaluation of Pearson correlation, because they enforce the un-realistic hypothesis of GPP independence of T_{air} and Q_{air}. For this reason, a cropping has been applied to the dataset, removing all the data-points with a temperature entry under a certain threshold, which account for around 5%÷10% of the whole dataset. The phenomenon being dependent on the biological features of the plants, a different threshold has been chosen for each PFT, based on the observations, as reported in Table 5.1. For FR-Pue (PFT5) no threshold has been consistently observed, as shown for example in Figure 5.14. This is consistent with the aforementioned hypothesis of photosynthesis temperature threshold, FR-Pue being an evergreen forest in a temperate biome, therefore allowing continuous photosynthetic activity along the whole year.



(a) *complete data* **(b)** *cropped data*

Figure 5.13: Scatter plot of Q_{air} and GPP for US-UMB site. On the left the entire dataset is plotted. On the right the data-points below the temperature threshold are removed.

PFT	Phenology	$T_0(K)$
5	evergreen broad-leaved	NA
6	summergreen broad-leaved	278
7	evergreen needle-leaf	270

Table 5.1: Temperature threshold T_0 for photosynthetic activation depending on the PFT. The value for PFT5 (FR-Pue) is not clearly distinguishable in the scatter plots and therefore it has not been considered.

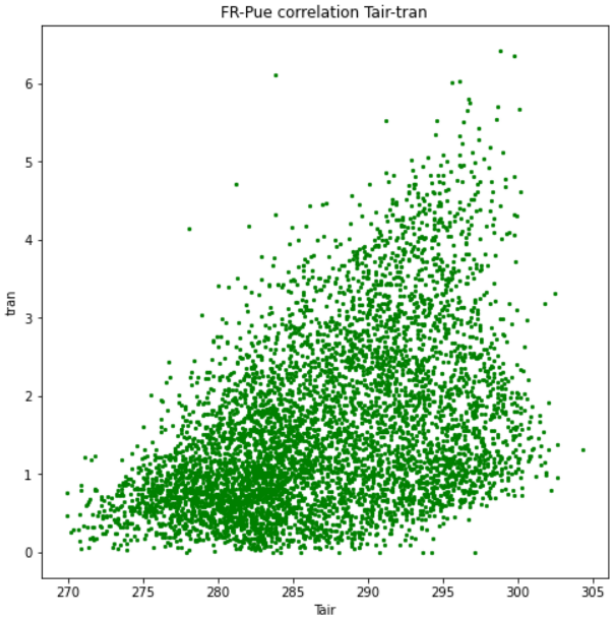
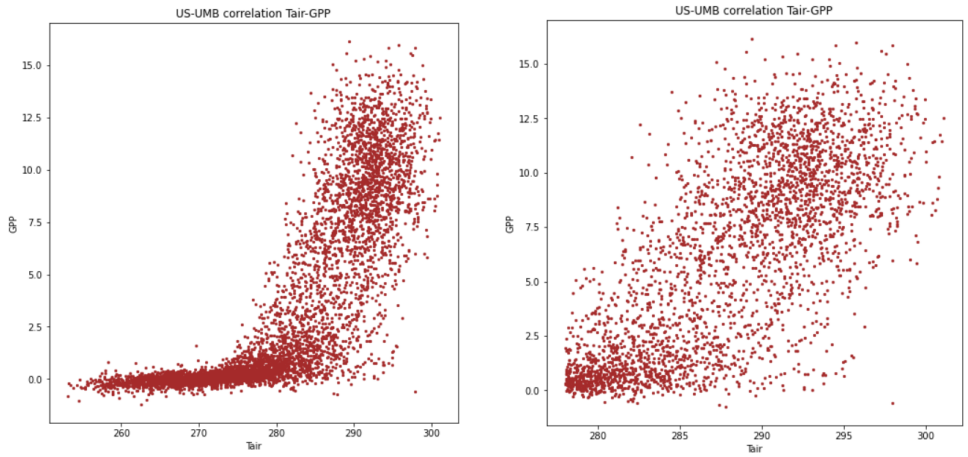


Figure 5.14: Scatter plot of T_{air} and transpiration for FR-Fon site. No temperature threshold is observed, as expected from PFT5 in a temperate biome.

The effect of the cropping is presented in Figure 5.12b, Figure 5.13b and Figure 5.15b. It is possible to see a much smaller density of low-GPP points in the plateau, which is therefore much less influential in the computation of correlation. In order to give a clearer grasp of the magnitude of the correction applied, the scatter plots of US-UMB T_{air} and GPP full and cropped data are showed in Figure 5.15. From now on corrected datasets are used if not specified otherwise.

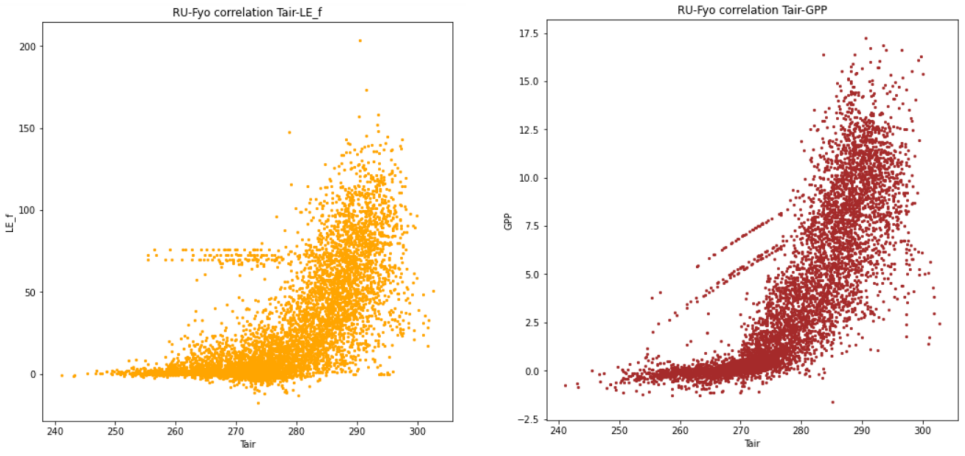
Another pathological behaviour has been observed in FLUXNET2015-derived quantities. It consists of the appearance of lines of points exhibiting a strong linear correlation, as shown in Figure 5.16. Those lines are immediately visible since they are located outside the main bulk of data-points, but they could also be present inside it without being noticeable. Since their correlation seems quite artificial and appears only in FLUXNET2015 data, which undertake a gap-filling procedure during the database formation, these points are most likely gap-filled data. Their behaviour strongly separates from real measurements, so it would be beneficial to remove these points before performing the quantitative analysis. Unfortunately, most of the entries from FLUXNET2015 are flagged as gap-filled or corrected in the metadata associated with the measurements, therefore it is impossible to neatly identify them. A rigorous correction being not applicable, it has been chosen not to remove these points from the dataset.

TROPOMI observations only comprise data during the growing season and



(a) *complete data* **(b)** *cropped data*

Figure 5.15: Scatter plot of T_{air} and GPP for US-UMB site. On the left the entire dataset is plotted. On the right the data-points below the temperature threshold are removed.



(a) *LE* **(b)** *GPP*

Figure 5.16: Scatter plot of T_{air} and LE (a) and GPP (b) for RU-Fyo site. The full dataset with no correction is shown. In both plots some series of quite likely gap-filled points are visible: in (a) for values of LE in $50 \div 100\, W\, m^2$ and T_{air} in $255 \div 280K$, in (b) for values of GPP in $2.5 \div 8.5\, gC\, m^{-2}$ per timestep and T_{air} in $255 \div 280K$.

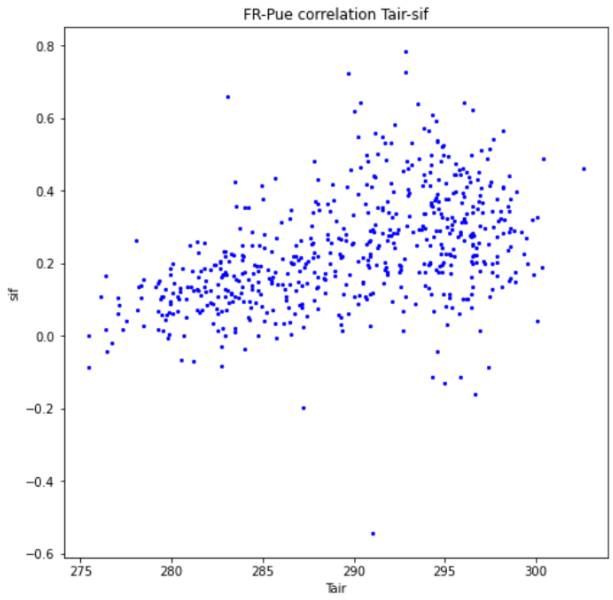


Figure 5.17: Scatter plot of T_{air} and SIF for FR-Pue.

around it, at the exclusion of the coldest months, between May 2018 and December 2020. As a consequence, the number of data-points is quite small compared to the other quantities and the correlation measure is less accurate, because it is more sensible to stochastic oscillations and aforementioned dispersion effects (see Figure 5.17). Vapour Pressure Deficit (here vpd) being extracted from SAPFLUNET database (which has no overlapping periods with the TROPOMI one), its correlation to SIF cannot be evaluated without deriving it from ERA5 re-analysis data, a procedure which would introduce new sources of uncertainties. As a consequence, this correlation is not available in the analysis.

5.2.2 Quantitative analysis

Once the discrete threshold behaviours have been taken into account and solved, the major issue in quantifying the correlation between each quantity and its drivers stems from the dispersion of the data-points due to the presence of several controlling variables, namely the other drivers.

Using Pearson correlation in these conditions could lead to misleading results, therefore also partial correlation coefficients have been computed. Unfortunately, some of the assumptions needed for partial correlation to be properly applied are not fulfilled by the data. For example, the presence of linear relations amongst all the variables is not guaranteed (see Figure 5.12 and Figure 5.16) and the requirement of a gaussian distribution of the measurements is not met for

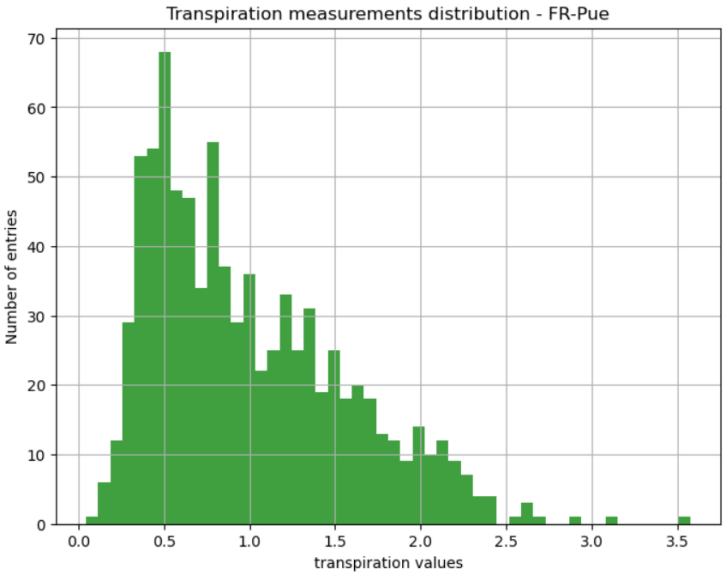


Figure 5.18: Histogram of the distribution of transpiration data for FR-Pue.

the majority of the quantities considered, as shown for transpiration data in Figure 5.18. Moreover, partial correlation is very sensible to outliers[43], which are quite common in these datasets as it is possible to observe from the previous scatter plots. As a consequence, an optimal measure of correlation is missing, and both have been considered in the following.

A wide discrepancy emerges comparing partial correlation coefficients and Pearson ones (see Table 5.2, Table 5.3, Table 5.4), in some cases even suggesting unrealistic physical relations or a substantial independence of some quantities with respect to important drivers. For example, in Table 5.3 for the FR-Fon site the partial correlation between temperature and SIF is negative and the almost vanishing value of the one between T_{air} and LE suggests an independence of the two, in stark contrast with the literature.

It is important to mention, though, that the partial correlation is correctly grasping the strong dependency of photosynthetic activity on the presence of solar radiation. In fact, comparing partial correlation coefficients of each quantity with respect to SW_{down} and all the other meteorological variables, it is possible to notice how SW_{down} results to be much more correlated than all the others. A possible explanation for this behaviour can be found by considering the nature and the importance of each meteorological driver in the biological process of photosynthesis. Q_{air} and vpd are both linked to the presence of water in the environment. Indeed, this a fundamental factor for the growth of the plants, but the presence of reservoirs and the length scale of water stress effects make these two drivers less strongly correlated to the photosynthetic activity on a diel scale.

Sites	Drivers	Correlation	tran	SIF	LE	GPP
FR-Pue	T_{air}	Pearson	0.44	0.46	0.45	0.36
		Partial	0.27	-0.08	0.30	0.25
	Q_{air}	Pearson	0.13	0.35	0.12	0.13
		Partial	-0.24	0.13	-0.24	-0.20
	SW_{down}	Pearson	0.70	0.58	0.69	0.61
		Partial	0.59	0.41	0.61	0.63
	vpd	Pearson	0.47	NA	0.53	0.31
		Partial	-0.36	NA	-0.31	-0.44

Table 5.2: Pearson and partial correlations for PFT5 site FR-Pue. NA labels for vpd and SIF correlation indicate the absence of overlapping periods and therefore it is impossible to evaluate a correlation between these two quantities.

In the same way, also the temperature is frequently involved in long time scale processes. For instance it is strongly related to the start and end of the growing season, where temperature constraints matter more. On the other side, SW_{down} represents the incoming solar radiation, which affects the photosynthetic activity instantaneously and it is an essential element for it to happen. As a consequence, it is not unreasonable for SW_{down} to be more correlated to the observable data in a partial correlation analysis, at least on a diel scale.

As previously mentioned, transpiration and SIF observations do not overlap on any period. As a consequence, an indirect measure of their mutual correlation has been found in the comparison of the correlations of both quantities with respect to the 4 drivers used above. For completeness, GPP and LE have been included in the comparison as well. For simplicity, the correlations with respect to vpd have not been considered, since its observations do not overlap with SIF ones. In this analysis the correlations of each quantity with the drivers are represented as a vector in the form:

$$\rho_x = (\rho(x, T_{air}),\ \rho(x, SW_{down}),\ \rho(x, Q_{air})) \tag{5.1}$$

where x represents one amongst the 4 observations (GPP, LE, tran, SIF) and $\rho(x, \alpha)$ its correlation coefficient with respect to the driver α. The agreement between the correlations of two quantities x and y with respect to the drivers is quantified by $d(x, y)$:

$$d(x, y) = \frac{1}{3} \sum_{\alpha} |\rho(x, \alpha) - \rho(y, \alpha)| \tag{5.2}$$

with $\alpha \in \{T_{air}, SW_{down}, Q_{air}\}$. The values of $d(x, y)$ for all couples (x, y) and sites are reported in Table 5.5.

Sites	Drivers	Correlation	tran	SIF	LE	GPP
FR-Fon	T_{air}	Pearson	0.70	0.27	0.75	0.74
		Partial	0.32	-0.41	0.02	0.16
	Q_{air}	Pearson	0.44	0.17	0.57	0.58
		Partial	-0.20	0.38	0.10	0.02
	SW_{down}	Pearson	0.72	0.70	0.75	0.75
		Partial	0.59	0.73	0.63	0.63
	vpd	Pearson	0.66	NA	0.61	0.52
		Partial	-0.22	NA	-0.01	-0.21
US-UMB	T_{air}	Pearson	0.25	0.57	0.73	0.73
		Partial	0.05	0.23	0.12	0.17
	Q_{air}	Pearson	-0.16	0.43	0.50	0.59
		Partial	-0.06	-0.03	0.07	0.07
	SW_{down}	Pearson	0.65	0.52	0.69	0.58
		Partial	0.37	0.38	0.56	0.53
	vpd	Pearson	0.58	NA	0.78	0.68
		Partial	0.12	NA	0.24	-0.02
US-UMd	T_{air}	Pearson	0.03	0.57	0.62	0.69
		Partial	-0.10	0.23	0.26	0.46
	Q_{air}	Pearson	-0.16	0.43	0.42	0.54
		Partial	0.08	-0.03	-0.12	0.29
	SW_{down}	Pearson	0.59	0.52	0.71	0.63
		Partial	0.43	0.38	0.48	0.52
	vpd	Pearson	0.48	NA	0.67	0.43
		Partial	0.25	NA	0.15	-0.37

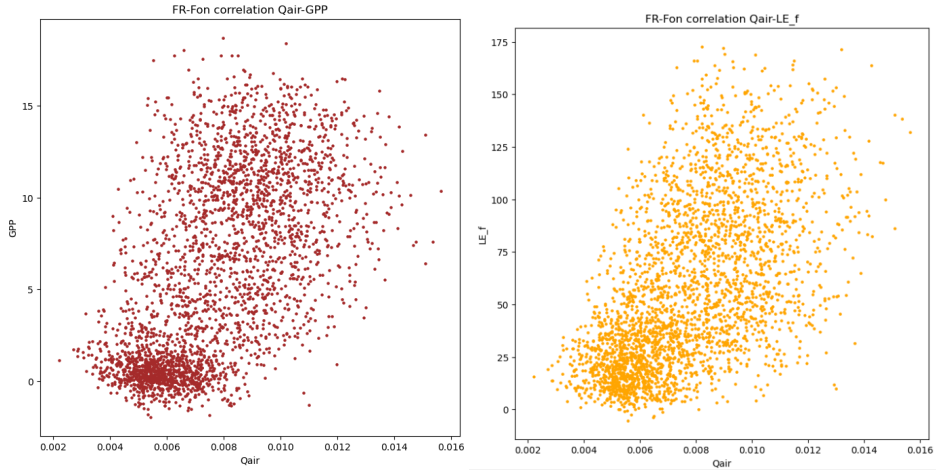
Table 5.3: Pearson and partial correlations for PFT6 sites. NA labels for vpd and SIF correlation indicate the absence of overlapping periods and therefore it is impossible to evaluate a correlation between these two quantities.

Sites	Drivers	Correlation	tran	SIF	LE	GPP
FI-Hyy	T_{air}	Pearson	0.43	0.56	0.79	0.84
		Partial	-0.16	0.14	0.19	0.44
	Q_{air}	Pearson	0.26	0.43	0.55	0.61
		Partial	0.21	0.04	0.02	-0.11
	SW_{down}	Pearson	0.56	0.42	0.81	0.80
		Partial	0.38	0.19	0.50	0.54
	vpd	Pearson	0.50	NA	0.76	0.71
		Partial	0.13	NA	-0.04	-0.33
RU-Fyo	T_{air}	Pearson	0.58	0.57	0.71	0.80
		Partial	0.11	0.05	0.04	0.12
	Q_{air}	Pearson	0.20	0.45	0.52	0.67
		Partial	-0.05	0.13	0.11	0.18
	SW_{down}	Pearson	0.72	0.56	0.79	0.76
		Partial	0.34	0.36	0.51	0.58
	vpd	Pearson	0.71	NA	0.65	0.63
		Partial	0.08	NA	-0.10	-0.22

Table 5.4: Pearson and partial correlations for PFT7 sites. NA labels for vpd and SIF correlation indicate the absence of overlapping periods and therefore it is impossible to evaluate a correlation between these two quantities.

FR-Pue	tran	SIF	LE	GPP
tran	-	0.12	0.01	0.06
SIF	0.30	-	0.11	0.12
LE	0.02	0.31	-	0.06
GPP	0.03	0.29	0.04	-
FR-Fon	**tran**	**SIF**	**LE**	**GPP**
tran	-	0.24	0.07	0.07
SIF	0.24	-	0.31	0.31
LE	0.07	0.27	-	0.01
GPP	0.07	0.35	0.08	-
US-UMB	**tran**	**SIF**	**LE**	**GPP**
tran	-	0.35	0.43	0.44
SIF	0.13	-	0.13	0.13
LE	0.13	0.13	-	0.07
GPP	0.13	0.10	0.03	-
US-UMd	**tran**	**SIF**	**LE**	**GPP**
tran	-	0.36	0.37	0.41
SIF	0.07	-	0.08	0.11
LE	0.15	0.09	-	0.09
GPP	0.27	0.21	0.12	-
FI-Hyy	**tran**	**SIF**	**LE**	**GPP**
tran	-	0.15	0.30	0.33
SIF	0.22	-	0.25	0.28
LE	0.22	0.25	-	0.04
GPP	0.36	0.28	0.14	-
RU-Fyo	**tran**	**SIF**	**LE**	**GPP**
tran	-	0.14	0.17	0.24
SIF	0.09	-	0.15	0.22
LE	0.14	0.06	-	0.09
GPP	0.16	0.11	0.07	-

Table 5.5: Distances between each couple of Pearson (upper triangular matrix in red) or partial (lower triangular matrix in green) correlations vectors ρ_x and ρ_y for all the sites. The distances are computed as the average discrepancy between the correlation coefficients of the 2 quantities with respect to each driver.



(a) *GPP* **(b)** *LE*

Figure 5.19: Scatter plots of GPP (a) and LE (b) with respect to Q_{air} for FR-Fon site.

GPP and LE are by far the two most closely correlated quantities both for Pearson and Partial correlations amongst all the sites, which is coherent with their similar behaviour. An example of that is observed in Figure 5.16 and in Figure 5.19, showing a very similar dependence of GPP and LE with respect to T_{air} and Q_{air}. A possible reason behind their similarity can be found in the common origin of these quantities, both derived by FLUXNET2015, which also implies the same spatial coverage. Indeed, this is not the case for TROPOMI and SAP-FLUXNET data. We also observe a relevant link between LE and transpiration, when looking at partial correlations. On the other hand, the SIF-GPP relation is not so strong, as depicted by their correlations.

5.3 Sensitivity analysis and model optimisation

Sensitivity Analysis results are summarized in the heat-maps in Figure 5.21 and Figure 5.22. For each observable, these maps report the 25 most relevant parameters amongst those involved in ORCHIDEE computation of the observable itself, on a scale from 1 (black, high sensitivity of the variable with respect to the parameter) to 0 (white, independent). Around 12 parameters, amongst the highest-ranked ones for each PFT, have been used later as parameters set for the optimization procedure with ORCHIDAS.

It is possible to notice a good level of agreement regarding the most influ-

ential parameters within each PFT. Indeed some differences still appear, which are probably due to features of the sites such as soil texture and climate. As a consequence, it has been decided to optimize the parameters grouping the simulations by PFT, extracting the union of the most important parameters for all the sites belonging to each PFT and then using their combined datasets to obtain a wider training set for the optimization algorithm.

The two US sites present similar correlation coefficients if compared with those of other sites, despite the artificial disturbance on US-UMd which causes the forest population to have more young trees than the US-UMB one. The standard deviations of their Pearson and partial correlation coefficients are respectively equal to 0.07 and 0.15, with the wider discrepancies regarding transpiration correlation coefficients (see Table 5.3). Amongst those coefficients, the largest errors affect the correlation with T_{air}, while the correlations with SW_{down} and Q_{air} are the most compatible. As it is possible to see in the scatter plots in Figure 5.20 for the US-UMB site, almost no correlation seems to be present for T_{air} (Figure 5.20a), while SW_{down} (Figure 5.20b) exhibits a much more evident one, and the same holds for the site US-UMd, showing from this point of view a similar behaviour. It has been chosen to use also the disturbed-site US-UMd data in the optimization procedure, in order to enrich the dataset and obtain more reliable results while not only representing mature forests, but also younger ones. Moreover, as mentioned in subsection 5.1.2, US-UMB and US-UMd simulations and observations are substantially different, so their contribution is comparable to that of 2 independent sites and doesn't pose a threat to the validity of the machine learning technique.

In Figure 5.22 the most influential parameter for all the sites is *alpha_NPQ_reversible*, a coefficient involved in the description of Non Photochemical Quenching (NPQ) phenomena. The relative rate constant for fluorescence k_F, which presents a direct proportionality to SIF in the computation, is the second most important parameter.

Three optimization procedures have been run, each one optimizing the model estimate of a different quantity among transpiration, GPP and SIF. The actual performance of the model after the optimization can be deduced by measuring the discrepancy of each model (the original and the three optimized ones) estimates from the three sets of observations. These comparisons are showed in Figure 5.23, where a normalized Root Mean Square Deviation (RMSD) has been used as measure of the distance of ORCHIDEE estimates from the observations.

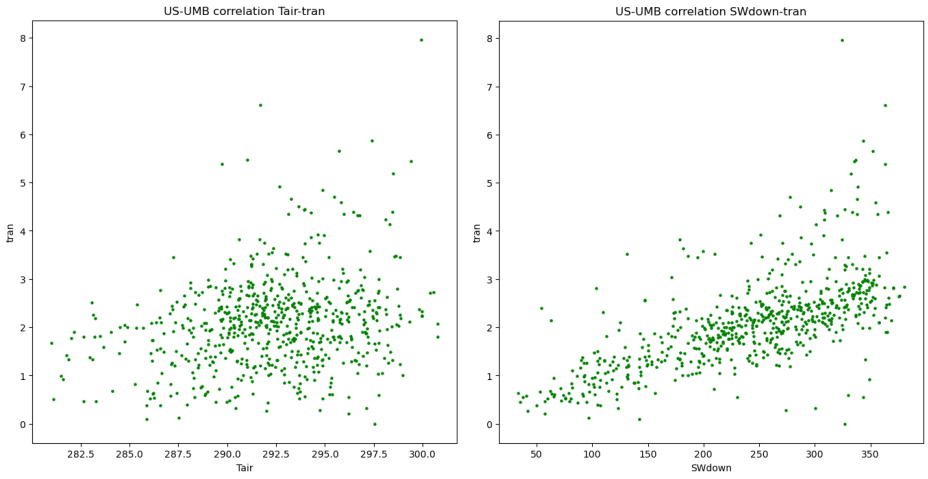
A few interesting features emerge from the diagrams. The original model and the one optimized by GPP data show the same performance in almost every site and quantity considered. The reason behind this similarity is found in the fact that GPP has been used as a reference quantity while building the model, and it is one of the quantities which the model simulates reasonably well. This

Parameter	Description (unit)
A1	Empirical factor involved in the calculation of fvpd
ARJV	a coefficient of the linear regression (a+bT) defining the Jmax25/Vcmax25 ratio ($\mu mol\,\mu mol_{CO2}^{-1}$)
ASJ	A coefficient of the linear regression (a+bT) defining the Entropy term for Jmax ($JK^{-1}mol^{-1}$)
ASV	A coefficient of the linear regression (a+bT) defining the Entropy term for Vcmax ($J\,K^{-1}mol^{-1}$)
B1	Empirical factor involved in the calculation of fvpd
BRJV	b coefficient of the linear regression (a+bT) defining the Jmax25/Vcmax25 ratio ($\mu mol\,\mu mol_{CO2}^{-1}(^{\circ}C)^{-1}$)
CLUMPING	Clumping index of leaves
CT	Heat transfer coefficient of the leaf
CWRR_N_VANGENUCHTEN	Van Genuchten coefficient n
G0	Residual stomatal conductance when irradiance approaches zero ($mol\,m^{-2}\,s^{-1}\,bar^{-1}$)
GB_REF	Leaf bulk boundary layer resistance ($s\,m^{-1}$)
GDDNCD_CURVE	Constant in the computation of critical GDD
GDDNCD_OFFSET	Constant in the computation of critical GDD
GDDNCD_REF	Reference value used in the computation of critical GDD ($days$)
HYDROL_HUMCSTE	Root profile (m)
KMC25	Michaelis-Menten constant of Rubisco for CO2 at 25°C (μbar)
LAI_MAX	Maximum LAI (Leaf Area Index) ($m^2\,m^{-2}$)
LAI_MAX_TO_HAPPY	SAI/LAI ratio, Larcher 1991
LEAFAGECRIT	Critical leaf age, tabulated ($days$)
SECHIBA_QSINT	Interception reservoir coefficient (m)
SLA	Specif leaf area ($m^2\,gC^{-1}$)
TAU_LEAFINIT	Time to attain the initial foliage using the carbohydrate reserve ($days$)
TAU_T2M_MONTH	Time constant for the monthly 2-meter temperature ($days$)
VCMAX25	Maximum rate of Rubisco activity-limited carboxylation at 25 °C ($\mu mol\,m^{-2}\,s^{-1}$)
VWC_FC	Volumetric water content field capacity
VWC_SAT	Saturated soil water content
ZSNOWTHRMCOND1	Snow thermal conductivity parameter 1 ($W\,m^5\,kg^{-2}K^{-1}$)
ZSNOWTHRMCOND_CVAP	Snow thermal conductivity (vapor) parameter c (K)

Table 5.6: Transpiration and LE SA parameters description. Dimensionless quantities have no unit of measure reported. For further details: https://orchidas.lsce.ipsl.fr/overview/orchidee.php

Parameter	Description (unit)
a_psII	Absorption cross-section area of PSII
alpha_NPQ_reversible	Reversible NPQ coefficient
ARJV	a coefficient of the linear regression (a+bT) defining the Jmax25/Vcmax25 ratio ($\mu mol\,\mu mol_{CO2}^{-1}$)
ASJ	A coefficient of the linear regression (a+bT) defining the Entropy term for Jmax ($JK^{-1}mol^{-1}$)
ASV	A coefficient of the linear regression (a+bT) defining the Entropy term for Vcmax ($J\,K^{-1}mol^{-1}$)
BRJV	b coefficient of the linear regression (a+bT) defining the Jmax25/Vcmax25 ratio ($\mu mol\,\mu mol_{CO2}^{-1}(^{\circ}C)^{-1}$)
BSJ	b coefficient of the linear regression (a+bT) defining the Entropy term for Jmax ($JK^{-1}mol^{-1}{}^{\circ}C^{-1}$)
CLUMPING	Clumping index of leaves
CWRR_N_VANGENUCHTEN	Van Genuchten coefficient n
D_JMAX	Energy of deactivation for Jmax ($Jmol^{-1}$)
E_JMAX	Energy of activation for Jmax ($Jmol^{-1}$)
GB_REF	Leaf bulk boundary layer resistance ($s\,m^{-1}$)
GDDNCD_CURVE	Constant in the computation of critical GDD
GDDNCD_REF	Reference value used in the computation of critical GDD ($days$)
HYDROL_HUMCSTE	Root profile (m)
k_F	Relative rate constant for fluorescence
k_P	Relative rate constant for photochemistry
KMC25	Michaelis-Menten constant of Rubisco for CO2 at 25°C (μbar)
KMO25	Michaelis-Menten constant of Rubisco for O2 at $25^{\circ}C$ (μbar)
LAI_MAX	Maximum LAI (Leaf Area Index) ($m^2\,m^{-2}$)
LAI_MAX_TO_HAPPY	SAI/LAI ratio, Larcher 1991
LEAFAGE_LASTMAX	Leaf age at which vmax falls below vcmax_opt in fraction of critical leaf age
LEAFAGECRIT	Critical leaf age, tabulated ($days$)
SCO25	Relative CO2/O2 specificity factor for Rubisco at $25^{\circ}C$
SLA	Specif leaf area ($m^2\,gC^{-1}$)
TAU_T2M_MONTH	Time constant for the monthly 2-meter temperature ($days$)
TAU_T2M_RACZKA	Speed of plant temperature adaptation defined in Raczka et al. (2019) ($days$)
TAU_T2M_WEEK	Time constant for the weekly 2-meter temperature ($days$)
TPHOTO_MIN	Minimum temperature for photosynthesis ($^{\circ}C$)
THETA	Convexity factor for response of J to irradiance
VCMAX25	Maximum rate of Rubisco activity-limited carboxylation at 25 °C ($\mu mol\,m^{-2}\,s^{-1}$)
VMAX_OFFSET	Minimum relative vcmax, offset

Table 5.7: SIF and GPP SA parameters description. Dimensionless quantities have no unit of measure reported. For further details: https://orchidas.lsce.ipsl.fr/overview/orchidee.php



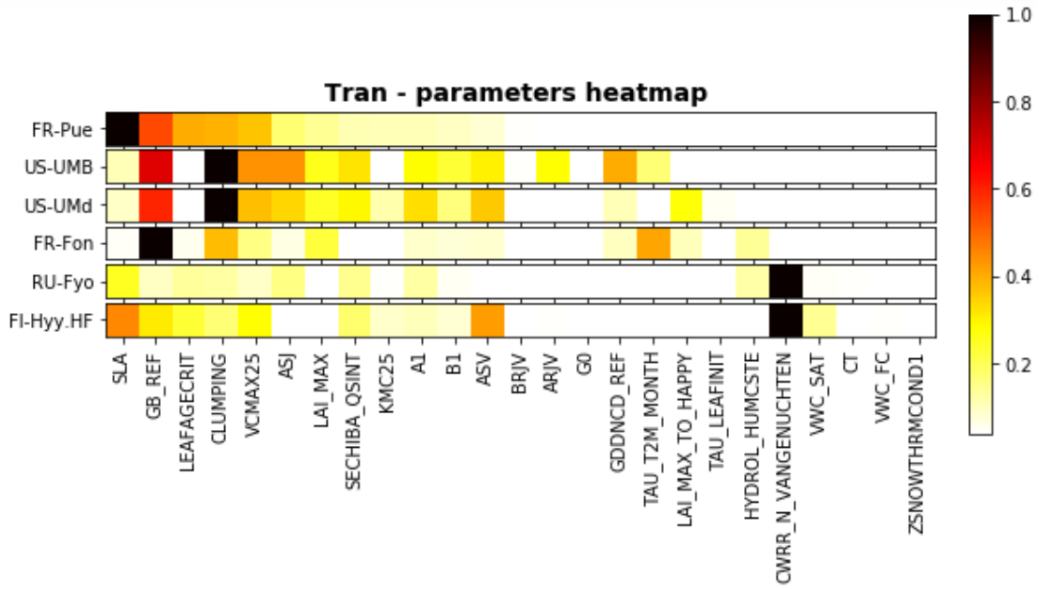
(a) T_{air} **(b)** SW_{down}

Figure 5.20: Scatter plots of transpiration with T_{air} (a) and SW_{down} (b) for the US-UMB site.

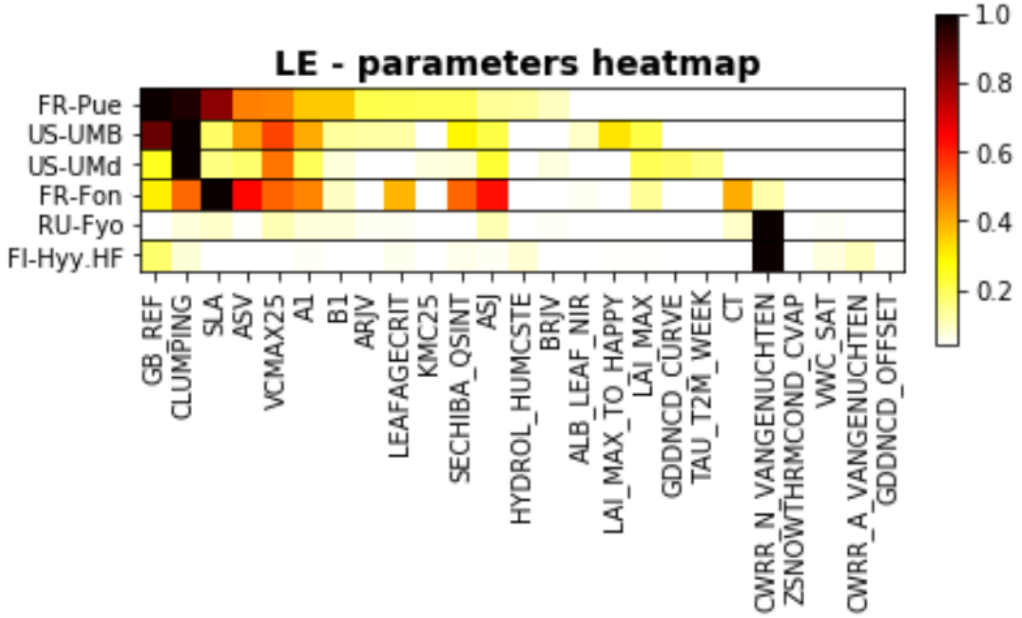
is indeed confirmed by observing how GPP is on average the quantity which is the most correctly reproduced by all the models, even those which are optimized with respect to transpiration and SIF.

The model optimized through the use of SIF data is almost always showing the worst performances in predicting GPP and transpiration, with the only exception being US-UMd. At the same time, the original model and the ones optimized by transpiration and GPP show a very low predictive power when trying to produce SIF estimates. This fact suggests a substantial difference of the SIF-informed model from all the others. During the last week of this research project the model has been found out to present some inconsistencies in the module computing SIF, which explains the source of the atypical behaviour of the SIF-optimized model. The computational time needed to re-run all the simulations and analyses is around 2-3 weeks, so it has not been possible to produce the new data with the corrected model by the end of the internship and SIF estimates presented in this research have to be revised. However, on a personal and professional dimension, it has been deeply interesting and educational to be able to effectively perceive and measure this problem from the comparison and study of model estimates against observations.

Another interesting point comes from the observation of the performances of the transpiration-optimized model. It has been found to produce better results than the original and GPP-based model when it comes to considering transpiration and SIF estimates (in particular in Figure 5.23e and f, PFT7 sites), while having almost the same performance in all the other cases. Therefore, transpira-

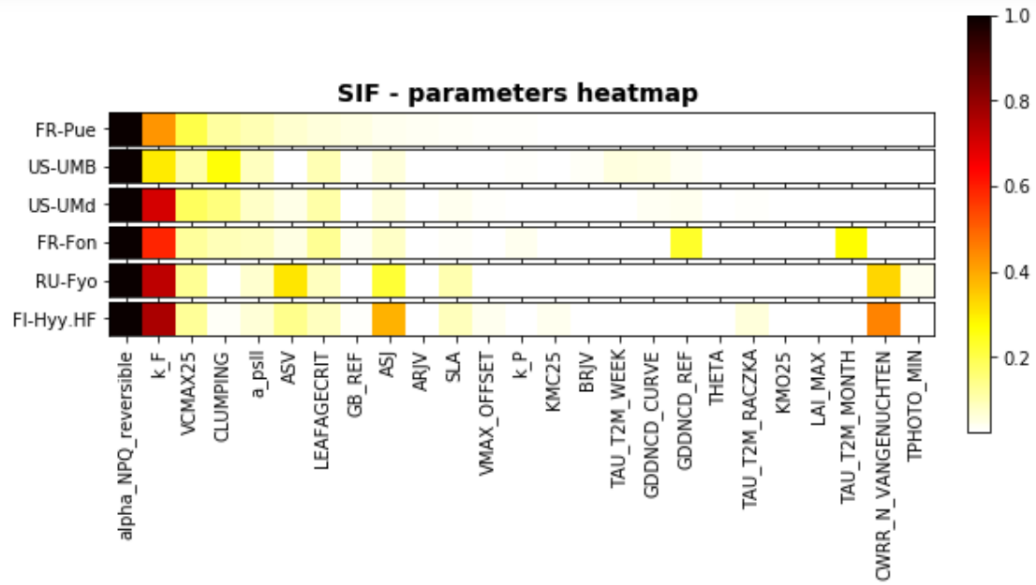


(a) *tran*

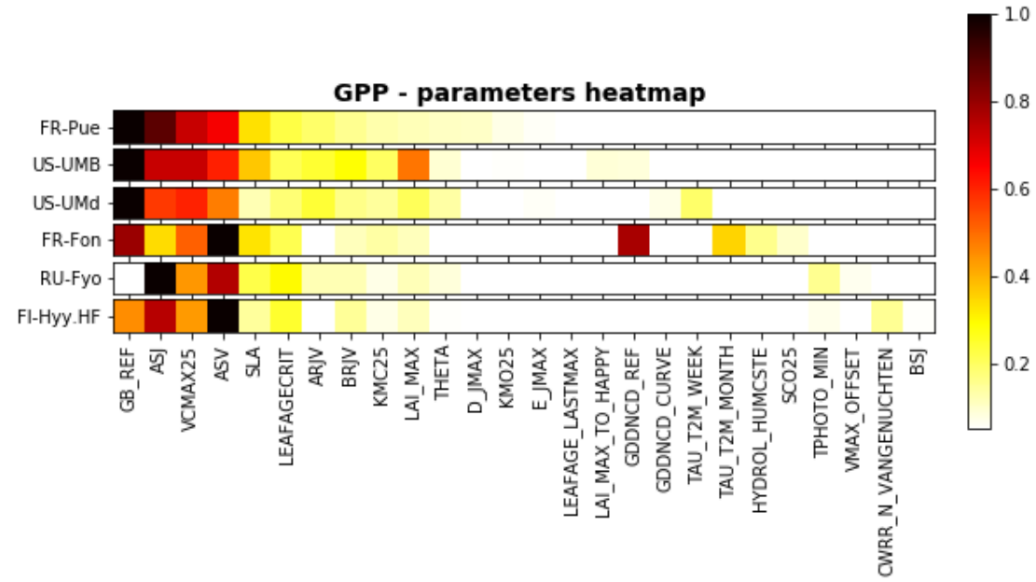


(b) *LE*

Figure 5.21: Sensitivity analysis results over transpiration and LE, on a scale from 1 (black, high influence) to 0 (white, no influence).

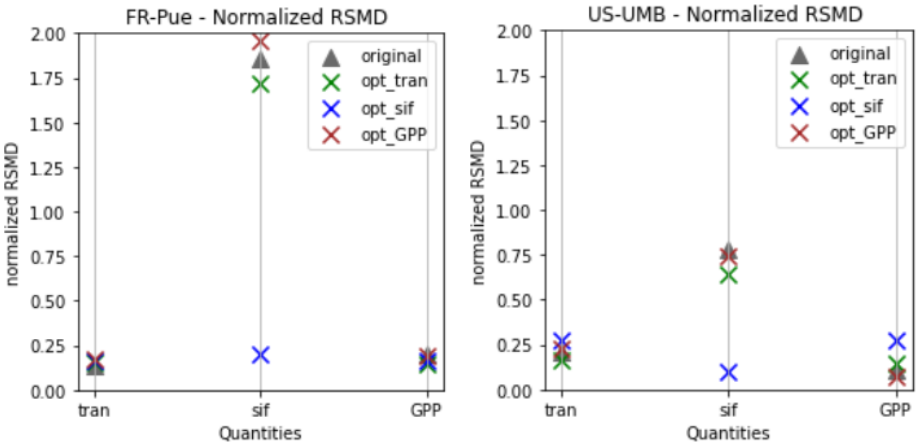


(a) *SIF*



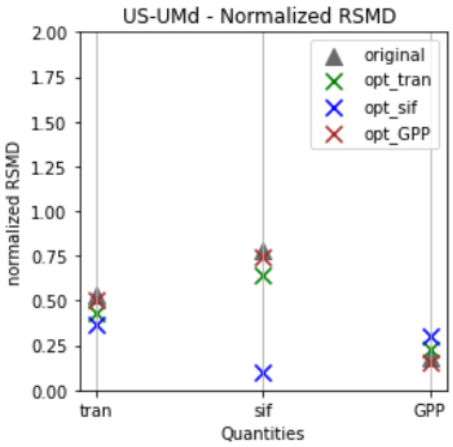
(b) *GPP*

Figure 5.22: Sensitivity analysis results over SIF and GPP, on a scale from 1 (black, high influence) to 0 (white, no influence).

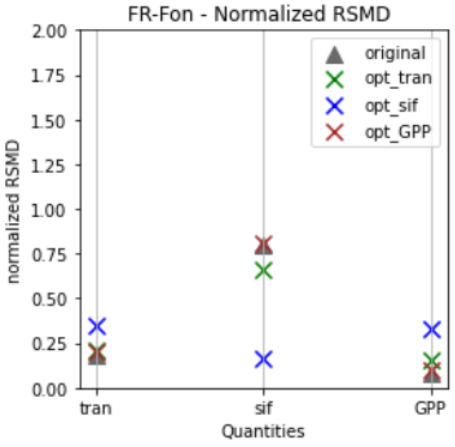


(a) *FR-Pue*

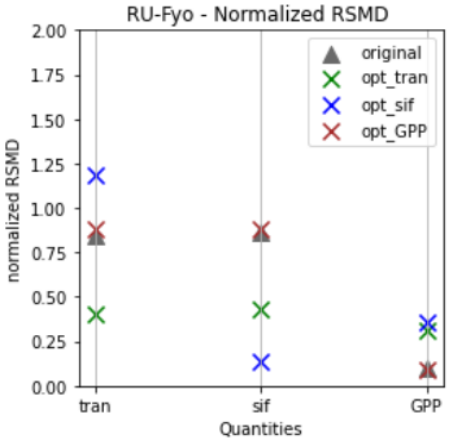
(b) *US-UMB*



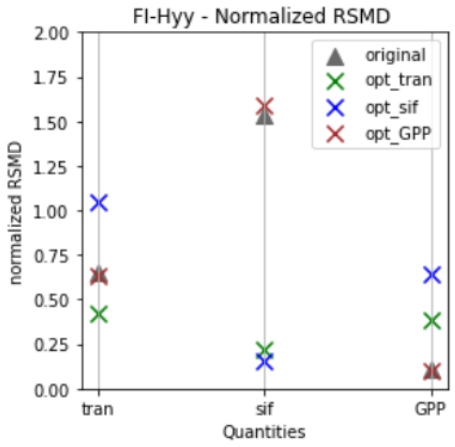
(c) *US-UMd*



(d) *FR-Fon*



(e) *RU-Fyo*



(f) *FI-Hyy*

Figure 5.23: Performance evaluation for original and optimized models obtained through the comparison between model estimates and observed values for transpiration, SIF and GPP.

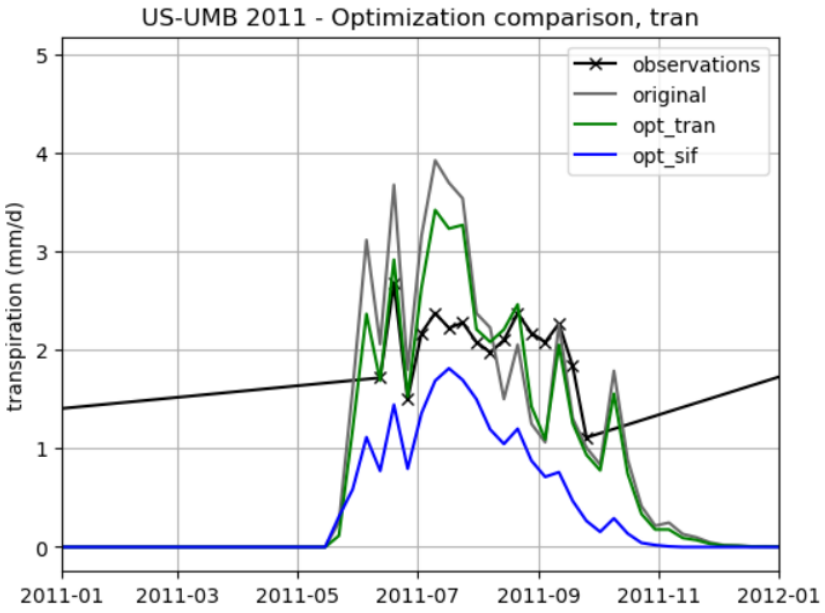
tion data seem to be able to correctly constrain the model, improving its overall predictive power. An example of this behaviour is observed in Figure 5.24.

The updated parameters are displayed in Figure 5.25. The observation of their new values can provide a validation of the success (or failure) of the optimization and an idea of the size of the changes that have been applied. From a quick inspection of the diagram, it is clear that the optimization procedure has not been completely successful. Indeed, several parameters have been optimized towards their upper or lower limits, as for *GB_REF* (leaf bulk boundary layer resistance, $s\,m^{-1}$) in Figure 5.25a for the GPP-based optimization, *B1_07* (which is an adimensional empirical factor involved in the calculation of the effect of leaf-to-air vapour difference on stomatal conductance g_s) in Figure 5.25b for the transpiration-based one or *k_F* (relative rate constant for fluorescence, s^{-1}) in Figure 5.25c using SIF.

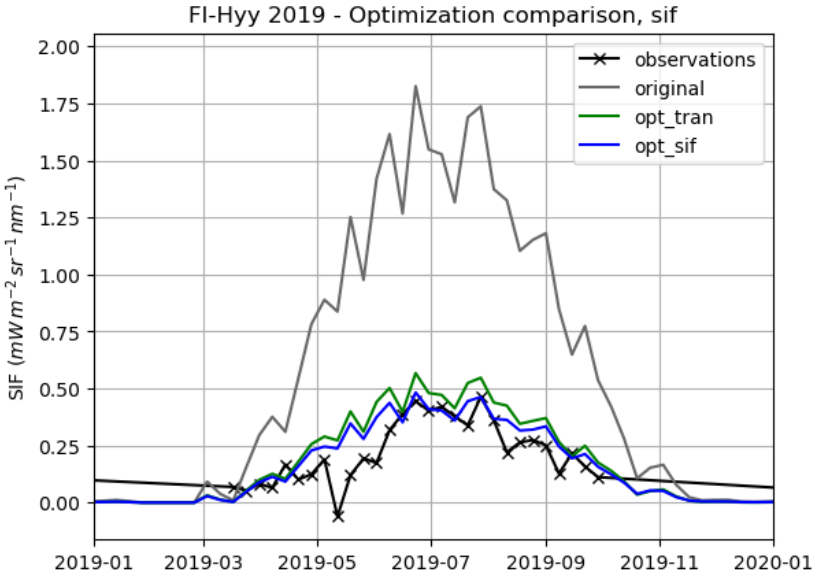
This behaviour indicates that either the range of variation which is used in the optimisation of the parameters is not adequate, too wide or too small, or the model does not include some important processes, instead it is trying to account for them by modifying the ones it possesses to reproduce the observations.

In the first case, the prescribed range for the parameter could have been set too wide. As a consequence, very unrealistic values of the parameters have been taken into consideration, and they might have produced a better numerical result within the algorithm execution than other more realistic ones by chance. In fact, there are several local minima in the parameters space that can be solutions for the optimization even if they are unrealistic, depending on the variation range of the parameters themselves. However, the stochastic nature of the genetic algorithm used for the optimization procedure should prevent the system from falling into local minima. If this is happening, decreasing the range could solve the problem by constraining the variable to assume values closer to the more realistic ones.

In the opposite situation the parameter variation interval could be too small. In fact, especially when dealing with new parameters not well known in the literature, the range is chosen almost arbitrarily. Therefore it is not sure whether the optimal value is included in the interval or not. In these cases, the range has to be widened, in order to allow the algorithm to explore more values for the parameter and find the optimal one.

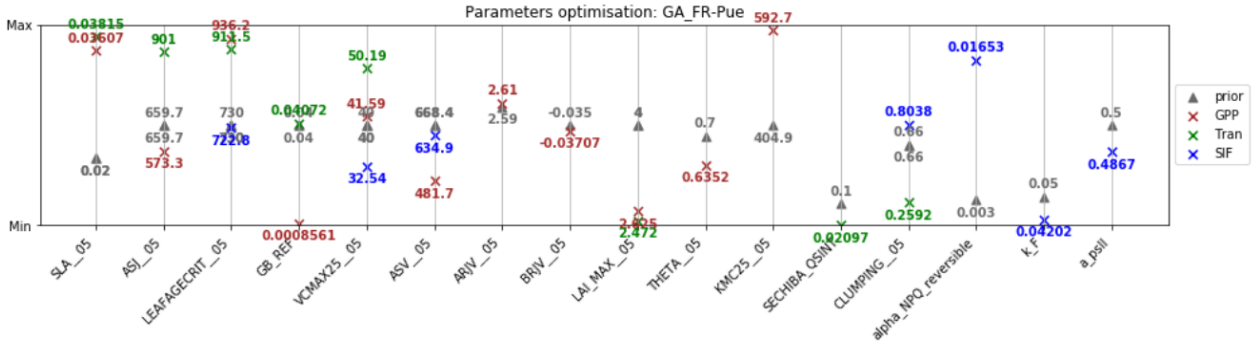


(a) *transpiration, US-UMB 2011*

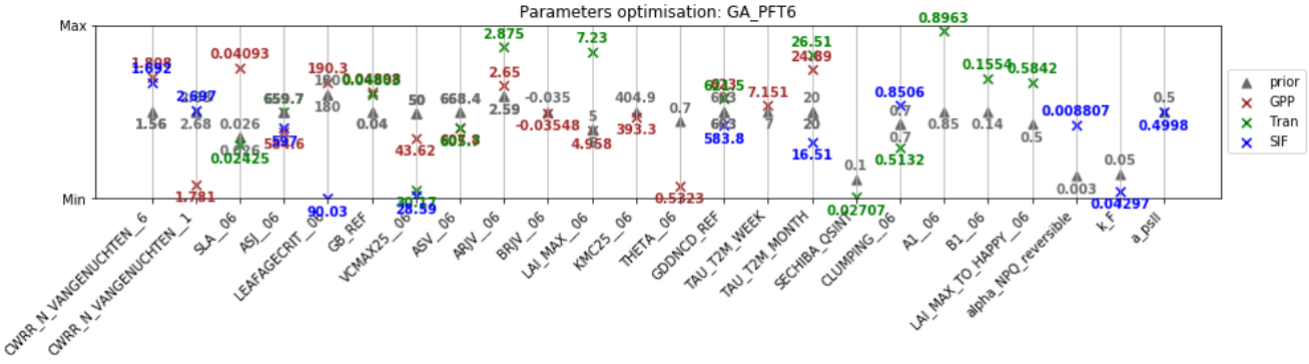


(b) *SIF, FI-Hyy 2019*

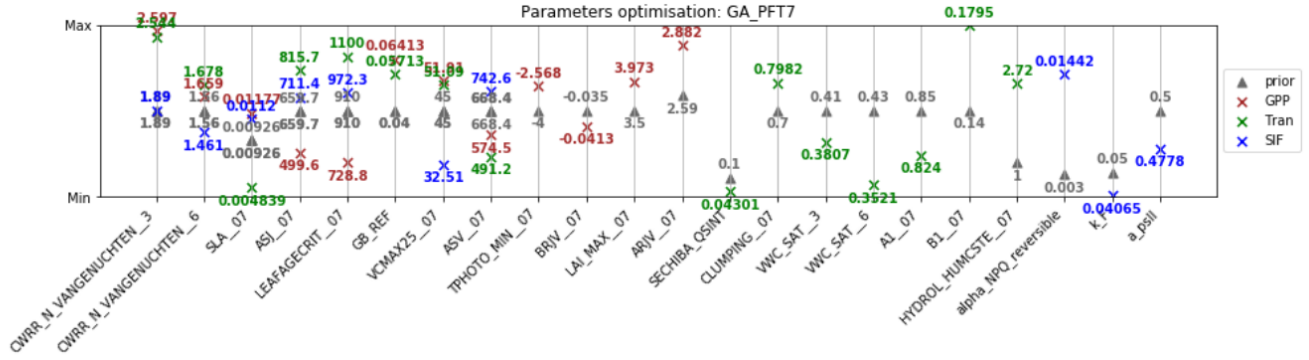
Figure 5.24: Comparison of transpiration (a) and SIF (b) estimates from the original and optimized models with respect to the observed data over the sites US-UMB and FI-Hyy.



(a) *PFT5*



(b) *PFT6*



(c) *PFT7*

Figure 5.25: Parameter updates after the optimization process. On the top (a) the results for PFT5, in the middle (b) those for PFT6 and at the bottom (c) PFT7. The original values are labelled in grey, those optimized with respect to GPP in red, transpiration in green and in blue the ones obtained from SIF. Each column corresponds to a parameter and it has an extension equal to the variation range of the parameter itself. It is possible to observe several saturations of the parameters during the optimization, for instance *SECHIBA_QSINT* shrinks to its minimum in (a) when considering the transpiration-informed optimization.

Chapter 6

Conclusions

Indeed, for the two databases that have been investigated and used in this research, SAPFLUXNET and TROPOMI, positive and negative aspects have been found, which have to be taken into account for a proper assimilation into ORCHIDEE.

Transpiration estimates from SAPFLUXNET sap flow data have produced good results in terms of optimization performance, and they have been proven to be potentially effective in constraining the model and increasing its predictive power, not only with respect to transpiration estimates but also with SIF ones. Given enought data, it would be possible to calibrate the model over all the PFTs and run global simulations optimized through local data. However, the need for choosing sites which overlap with FLUXNET2015 to run local simulations strongly reduces their number and the amount of available data which can be used to optimize the model. As a consequence, it is currently impossible to optimize the model over the other PFTs. A possible way out could be found using ERA5 re-analysis meteorological data to run simulations over SAPFLUNXET sites not covered by FLUXNET2015, as it has been done for the period between 2015 and 2020. However, the use of ERA5 re-analysis data implies a lower accuracy of the simulations, once again limiting the effectiveness of the optimization procedure.

The integration of sap flow single-plant data to transpiration at canopy level could be improved too. In particular the current method assumes a linear relation between basal area and sap flow. This is a realistic approximation for some plants, as shown by Güney, A. (2018)[44], but it is heavily dependent on the species and the environmental conditions. Therefore, a further step could improve the integration by adopting in the computation a single-tree basal area distribution of the stand and a more realistic relation between basal area and sap flow. This latter relation could be directly extrapolated from the data if not available in the literature.

The considerations it is possible to draw from SIF observations are limited by the inconsistencies in ORCHIDEE regarding the computation of this quantity, which is scheduled to be updated in early 2023. The investigation of the dataset has brought out some criticalities, like the presence of a very strong noise in the data and strong limitations due to the short time coverage, but also highlighted some interesting and promising aspects.

Regarding the amount of available data, including the last two years of observations of TROPOMI, currently not available for this research, would almost double the dataset and the mission will keep producing data for several years, further increasing the dimension of the dataset. Moreover, previous data from GOME-2 and GOSAT satellites observations could be integrated, even if they use different resolutions. The FLuorescence EXplorer (FLEX) satellite, planned to be launched in 2025, is going to provide spatially high-resolution measurements of SIF, which will be a complement to products from other existing satellite missions and high-temporal resolution products from upcoming geostationary missions. Its new observations could provide enough data to better constrain plant transpiration, assess the impacts of water stress on plants, and infer processes occurring in the root zone through the soil-plant water column[12].

Finally, the results of the partial correlation of SW_{down} and the measurements of correlation discrepancies between GPP and LE with respect to the other drivers have pointed out the potential effects of diverse time-scales and spatial resolutions in data analysis. These behaviours suggest putting specific care into the role of observation time and spatial resolution. Photosynthesis at the leaf level is instantaneously influenced by some drivers (such as solar radiation), differently from other observables which have longer response scales. Therefore, considering a shorter time scale in the optimization could be useful in order to properly describe this feature in the model. Since sub-daily data are not available for SIF, but they are for transpiration (through sap flow measurements), the use of these two contributions together (for example within a simultaneous optimization of the model with respect to both observations) could provide a stronger and more suitable constraint for the model vegetation description.

References

Articles

1 C. Cid-Liccardi, T. Kramer, M. Ashton and B. Griscom:
'Managing forest carbon in a changing climate',
in (2012), pp. 183–204.

2 S. D. Wullschleger, H. E. Epstein, E. O. Box, E. S. Euskirchen, S. Goswami, C. M. Iversen, J. Kattge, R. J. Norby, P. M. van Bodegom and X. Xu:
'Plant functional types in Earth system models: past experiences and future directions for application of dynamic vegetation models in high-latitude ecosystems',
Annals of Botany **114**, 1–16 (2014).

3 R. S. e. a. M. Pathak:
'Technical summary. in: climate change 2022: mitigation of climate change. contribution of working group iii to the sixth assessment report of the intergovernmental panel on climate change',
Cambridge University Press, Cambridge, UK and New York, NY, USA, 10.1017/9781009157926.002 (2022).

4 G. S. Callendar:
'Can carbon dioxide influence climate?',
Weather **4**, 310–314 (1949).

5 V. Eyring, S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer and K. E. Taylor:
'Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization',
Geoscientific Model Development **9**, 1937–1958 (2016).

6 P. Z. e. a. Masson-Delmotte V.:
'Annex ii: models. in climate change 2021: the physical science basis. contribution of working group i to the sixth assessment report of the intergovernmental panel on climate change',
Cambridge University Press, Cambridge, UK and New York, NY, USA, pp. 2087–2138, 10.1017/9781009157896.016 (2021).

7 S. von Caemmerer and N. Baker:
'The Biology of Transpiration. From Guard Cells to Globe',
Plant Physiology **143**, 3–3 (2007).

8 A. Damm, E. Paul-Limoges, E. Haghighi, C. Simmer, F. Morsdorf, F. Schneider, C. van der Tol, M. Migliavacca and U. Rascher:
'Remote sensing of plant-water relations: an overview and future perspectives',
Journal of Plant Physiology **227**, From aquaporin to ecosystem: Plants in the water cycle, 3–19 (2018).

9 J. Green, A. Konings and S. e. a. Alemohammad:
'Regionally strong feedbacks between the atmosphere and terrestrial biosphere.',
Nature Geosci **0**, 410–414 (2017).

10 A. Koppa, D. Rains, P. Hulsman, R. Poyatos and D. G. Miralles:
'A deep learning-based hybrid model of global terrestrial evaporation',
Nature communications **13**, 1–11 (2022).

11 A. Porcar-Castell, Z. Malenovský, T. Magney, S. Van Wittenberghe, B. Fernandez-Marin, F. Maignan, Y. Zhang, K. Maseyk, J. Atherton, L. Albert, T. Robson, F. Zhao, J. I. Garcia Plazaola, I. Ensminger, P. Rajewicz, S. Grebe, M. Tikkanen, J. Kellner, J. Ihalainen and B. Logan:
'Chlorophyll a fluorescence illuminates a path connecting plant molecular biology to earth-system science',
Nature Plants **7**, 10.1038/s41477-021-00980-4 (2021).

12 F. Jonard, S. De Cannière, N. Brüggemann, P. Gentine, D. Short Gianotti, G. Lobet, D. Miralles, C. Montzka, B. Pagán, U. Rascher and H. Vereecken:
'Value of sun-induced chlorophyll fluorescence for quantifying hydrological states and fluxes: current status and challenges',
Agricultural and Forest Meteorology **291**, 108088 (2020).

13 R. Poyatos, V. Granda, V. Flo, M. A. Adams, B. Adorján, D. Aguadé, M. P. M. Aidar, S. Allen and J. e. a. Martínez-Vilalta:
'Global transpiration data from sap flow measurements: the sapfluxnet database',
Earth System Science Data **13**, 2607–2649 (2021).

14 P. Köhler, M. J. Behrenfeld, J. Landgraf, J. Joiner, T. S. Magney and C. Frankenberg:
'Global Retrievals of Solar-Induced Chlorophyll Fluorescence at Red Wavelengths With TROPOMI', en,
Geophysical Research Letters **47**, 10.1029/2020GL087541 (2020).

15 M. Aubinet, T. Vesala and D. Papale:
'Eddy covariance: a practical guide to measurement and data analysis',
Eddy Covariance (2012).

16 G. Pastorello, C. Trotta, E. Canfora, H. Chu, D. Christianson, Y.-W. Cheah, C. Poindexter, M. Torn and e. a. Papale Dario:
'The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data',
Scientific Data **7**, 225 (2020).

17 O. Dare-Idowu, A. Brut, J. Cuxart, T. Tallec, V. Rivalland, B. Zawilski, E. Ceschia and L. Jarlan:
'Surface energy balance and flux partitioning of annual crops in southwestern france',
Agricultural and Forest Meteorology **308-309**, 108529 (2021).

18 K. Wilson, A. Goldstein, E. Falge, M. Aubinet, D. Baldocchi, P. Berbigier, C. Bernhofer, R. Ceulemans, H. Dolman, C. Field, A. Grelle, A. Ibrom, B. Law, A. Kowalski, T. Meyers, J. Moncrieff, R. Monson, W. Oechel, J. Tenhunen, R. Valentini and S. Verma:
'Energy balance closure at fluxnet sites',
Agricultural and Forest Meteorology **113**, FLUXNET 2000 Synthesis, 223–243 (2002).

19 H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, 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. Hólm, M. Janisková, S. Keeley, P. Laloyaux, P. Lopez, C. Lupu, G. Radnoti, P. de Rosnay, I. Rozum, F. Vamborg, S. Villaume and J.-N. Thépaut:
'The era5 global reanalysis',
Quarterly Journal of the Royal Meteorological Society **146**, 1999–2049 (2020).

20 J. Cao, Q. An, X. Zhang, S. Xu, T. Si and D. Niyogi:
'Is satellite sun-induced chlorophyll fluorescence more indicative than vegetation indices under drought condition?',
Science of The Total Environment **792**, 148396 (2021).

21 C. Bacour, F. Maignan, N. MacBean, A. Porcar-Castell, J. Flexas, C. Frankenberg, P. Peylin, F. Chevallier, N. Vuichard and V. Bastrikov:
'Improving estimates of gross primary productivity by assimilating solar-induced fluorescence satellite retrievals in a terrestrial biosphere model using a process-based sif model',
Journal of Geophysical Research: Biogeosciences **124**, 3281–3306 (2019).

22 A. Damm, E. Haghighi, E. Paul-Limoges and C. van der Tol:
'On the seasonal relation of sun-induced chlorophyll fluorescence and transpiration in a temperate mixed forest',
Agricultural and Forest Meteorology **304-305**, 108386 (2021).

23 B. Poulter, N. MacBean, A. Hartley, I. Khlystova, O. Arino, R. Betts, S. Bontemps, M. Boettcher, C. Brockmann, P. Defourny, S. Hagemann, M. Herold, G. Kirches, C. Lamarche, D. Lederer, C. Ottlé, M. Peters and P. Peylin:
'Plant functional type classification for earth system models: results from the european space agency's land cover climate change initiative',
Geoscientific Model Development **8**, 2315–2328 (2015).

24 C. M. Gough, B. S. Hardiman, L. E. Nave, G. Bohrer, K. D. Maurer, C. S. Vogel, K. J. Nadelhoffer and P. S. Curtis:
'Sustained carbon uptake and storage following moderate disturbance in a great lakes forest',
Ecological Applications **23**, 1202–1215 (2013).

25 O. Boucher, J. Servonnat, A. L. Albright, O. Aumont, Y. Balkanski, V. Bastrikov, J. Vialard, N. Viovy and N. e. a. Vuichard:
'Presentation and evaluation of the ipsl-cm6a-lr climate model',
Journal of Advances in Modeling Earth Systems **12**, e2019MS002010 10.1029/2019MS002010, e2019MS002010 (2020).

26 S. Sitch, P. Friedlingstein, N. Gruber, S. D. Jones, G. Murray-Tortarolo, A. Ahlström, S. C. Doney, H. Graven, C. Heinze, C. Huntingford, S. Levis, P. E. Levy, M. Lomas, B. Poulter, N. Viovy, S. Zaehle, N. Zeng, A. Arneth, G. Bonan, L. Bopp, J. G. Canadell, F. Chevallier, P. Ciais, R. Ellis, M. Gloor, P. Peylin, S. L. Piao, C. Le Quéré, B. Smith, Z. Zhu and R. Myneni:
'Recent trends and drivers of regional sources and sinks of carbon dioxide',
Biogeosciences **12**, 653–679 (2015).

27 G. Krinner, N. Viovy, N. de Noblet-Ducoudré, J. Ogée, J. Polcher, P. Friedlingstein, P. Ciais, S. Sitch and I. C. Prentice:
'A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system',
Global Biogeochemical Cycles **19**, https://doi.org/10.1029/2003GB002199 (2005).

28 R. Lardy, G. Bellocchi and J.-F. Soussana:
'A new method to determine soil organic carbon equilibrium',
Environmental Modelling & Software **26**, 1759–1763 (2011).

29 C. A. Reynolds, T. J. Jackson and W. J. Rawls:
'Estimating soil water-holding capacities by linking the food and agriculture organization soil map of the world with global pedon databases and continuous pedotransfer functions',
Water Resources Research **36**, 3653–3662 (2000).

30 X. Yin and P. Struik:
'C3 and c4 photosynthesis models: an overview from the perspective of crop modelling',
NJAS - Wageningen Journal of Life Sciences **57**, Recent Advances in Crop Growth Modelling, 27–38 (2009).

31 S. Hemming, T. Dueck, J. Janse and F. van Noort:
'The effect of diffuse light on crops',
Acta Horticulturae 2008 (2008) 801 **801**, 10.17660/ActaHortic.2008.801.158 (2007).

32 Y. Zhang, A. Bastos, F. Maignan, D. Goll, O. Boucher, L. Li, A. Cescatti, N. Vuichard, X. Chen, C. Ammann, M. A. Arain, T. A. Black, B. Chojnicki, T. Kato, I. Mammarella, L. Montagnani, O. Roupsard, M. J. Sanz, L. Siebicke, M. Urbaniak, F. P. Vaccari, G. Wohlfahrt, W. Woodgate and P. Ciais:
'Modeling the impacts of diffuse light fraction on photosynthesis in orchidee (v5453) land surface model',
Geoscientific Model Development **13**, 5401–5423 (2020).

33 A. Weiss and J. M. Norman:
'Partitioning solar radiation into direct and diffuse, visible and near-infrared components',
Agricultural and Forest Meteorology **34**, 205–213 (1985).

34 C. Spitters, H. Toussaint and J. Goudriaan:
'Separating the diffuse and direct component of global radiation and its implications for modeling canopy photosynthesis part i. components of incoming radiation',
Agricultural and Forest Meteorology **38**, 217–229 (1986).

35 H. Fang and S. Liang:
'Leaf area index models',
edited by S. E. Jørgensen and B. D. Fath, 2139–2148 (2008).

36 S.-J. Jeong, D. Schimel, C. Frankenberg, D. T. Drewry, J. B. Fisher, M. Verma, J. A. Berry, J.-E. Lee and J. Joiner:
'Application of satellite solar-induced chlorophyll fluorescence to understanding large-scale variations in vegetation phenology and function over northern high latitude forests',
Remote Sensing of Environment **190**, 178–187 (2017).

37 J. Joiner, Y. Yoshida, A. Vasilkov, K. Schaefer, M. Jung, L. Guanter, Y. Zhang, S. Garrity, E. Middleton, K. Huemmrich, L. Gu and L. Belelli Marchesini:
'The seasonal cycle of satellite chlorophyll fluorescence observations and its relationship to vegetation phenology and ecosystem atmosphere carbon exchange',
Remote Sensing of Environment **152**, 375–391 (2014).

38 S. Walther, M. Voigt, T. Thum, A. Gonsamo, Y. Zhang, P. Köhler, M. Jung, A. Varlagin and L. Guanter:
'Satellite chlorophyll fluorescence measurements reveal large-scale decoupling of photosynthesis and greenness dynamics in boreal evergreen forests',
Global Change Biology **22**, 2979–2996 (2016).

39 V. Bastrikov, N. MacBean, C. Bacour, D. Santaren, S. Kuppel and P. Peylin:
'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)',
Geoscientific Model Development **11**, 4739–4754 (2018).

40 P. Peylin, C. Bacour, N. MacBean, S. Leonard, P. Rayner, S. Kuppel, E. Koffi, A. Kane, F. Maignan, F. Chevallier, P. Ciais and P. Prunet:
'A new stepwise carbon cycle data assimilation system using multiple data streams to constrain the simulated land surface carbon cycle',
Geoscientific Model Development **9**, 3321–3346 (2016).

41 K. Mahmud, R. L. Scott, J. A. Biederman, M. E. Litvak, T. Kolb, T. P. Meyers, P. Krishnan, V. Bastrikov and N. MacBean:
'Optimizing carbon cycle parameters drastically improves terrestrial biosphere model underestimates of dryland mean net co2 flux and its inter-annual variability',
Journal of Geophysical Research: Biogeosciences **126**, e2021JG006400 2021JG006400, e2021JG006400 (2021).

42 F. Campolongo, J. Cariboni and A. Saltelli:
'An effective screening design for sensitivity analysis of large models',
Environmental Modelling & Software **22**, Modelling, computer-assisted simulations, and mapping of dangerous phenomena for hazard assessment, 1509–1518 (2007).

43 S. J. Devlin, R. Gnanadesikan and J. R. Kettenring:
'Robust estimation and outlier detection with correlation coefficients',
Biometrika **62**, 531–545 (1975).

44 A. Güney:
'Sapwood area related to tree size, tree age, and leaf area index in cedrus libani',
Bilge International Journal of Science and Technology Research **2**, 83–91 (2018).

Books

B1 F. H. C. Marriott and I. S. Institute.:
A dictionary of statistical terms / by f.h.c. marriott, English,
5th ed. (Longman Scientific & Technical Harlow, Essex, 1989), viii, 223 p.

B2 D. J. Murray-Smith:
6 - experimental modelling: system identification, parameter estimation and model optimisation techniques,
edited by D. J. Murray-Smith (Woodhead Publishing, 2012), pp. 165–214.