Learning from machine learning

Using machine learning to resolve the global carbon cycle

Auke van der Woude

March 2019
MSc thesis
Meteorology and Air Quality group, Wageningen University

This MSc thesis is written at the Meteorology and Air quality department of the Wageningen University and studies the use of data-based models, specifically gradient boosted trees, in resolving the carbon cycle. This is done in two distinct ways, of which the methods and results are described in two seperate parts. First, data-based models are used as a predictive model for predicting gross primary production (GPP). Secondly, data-based models are used to do model output statistics on a process-based model. In principle, the two parts can be read individually, although readers unfamiliar with data-based models and gradient boosted trees are encouraged to read Section 2.1.1 before reading Part II.

Gross primary production (GPP) is the largest land-carbon flux and resolving this flux provides insight in current climate change. Still, this flux is very uncertain, with estimates of its global value ranging from 107 to 175 Pg/year. In order to reduce this uncertainty, data-based models are becoming more and more popular. In these data-based models, variables correlating to GPP can be used to estimate GPP at local and global scale, without limitations by theories or simplifications. A variable that is strongly related to GPP is Sun Induced Fluorescence (SIF). In this research, the added quality of SIF to data-based models known as model trees to predict GPP is assessed. The yearly average total GPP is simulated to be 125-131 PgC/year. It is found that due to SIF both the model quality and model complexity increase slightly. The results of this research show that SIF is very important in predicting GPP, but the quality of the SIF dataset is limiting.

The large, but uncertain change in the biosphere carbon fluxes due to anthropogenic activity show the sensitivity of the terrestrial biosphere to human influence. Besides, the biosphere is very susceptible to short-term climate variability. These two processes affect the net ecosystem exchange (NEE), indicating the need of reliable NEE estimates. A prior, process-based model and observations of CO_2 can be used to constrain the NEE by data assimilation. Less biased prior fluxes result in smaller errors in the posterior. Model output statistics (MOS), where predictands are statistically related to predictors, can be used to reduce the bias in the prior model. In this research, data-based models are used to conduct MOS on the prior model of the CarbonTracker Europe data assimilation system. The data-based models are found to increase the predictive quality of the prior model at most continents, showing the potential of data-based models to improve carbon flux estimates.

Supervision by dr.ir. Liesbeth Florentie and prof. dr. Wouter Peters

Contents

Chapter 1

Introduction

1.1 Importance of the carbon cycle

Due to an imbalance in the sources and sinks of carbon, the CO_2 mixing ratio in the atmosphere has increased from about 280 to over 400 ppm in the past 150 years (Etheridge et al., 1996; Dlugokencky and Tans, 2015). This increase in CO_2, along with other gases, resulted in a changed global radiation budget (Friedlingstein et al., 2014; Jackson et al., 2015). In turn, this change in radiation budget increases global surface and ocean temperatures, which is also known as the global greenhouse effect (Schneider, 1989). The imbalance in sources and sinks is mainly due to CO_2 release from anthropogenic fossil fuel burning. Since the industrial revolution, about half of the anthropogenic emitted CO_2 has stayed in the atmosphere (Knorr, 2009). The rest of the emitted CO_2 is taken up by the biosphere or the ocean.

The biosphere is a major CO_2 sink (Schlesinger and Bernhardt, 2013). In the process of photosynthesis, plants take up CO_2 in order to produce the sugars they need to grow, survive and reproduce. This uptake is called Gross Primary Production (GPP). GPP is the main land CO_2 sink and together with respiration, the main land CO_2 source, it forms Net Ecosystem Exchange (NEE). NEE is a key driver of land-atmosphere CO_2 exchange (Schlesinger and Bernhardt, 2013) and therefore also a key driver of the fraction of anthropogenically emitted CO_2 that remains in the atmosphere. GPP and NEE thus plays a major role in past, present and future climate change.

Although the importance of these previously mentioned carbon fluxes on climate change is widely recognised (e.g. Friedlingstein et al. (2006)), the carbon cycle is still not well resolved and carbon flux estimates are very uncertain. GPP is one of the largest fluxes of carbon in the earth system, but the exact magnitude is uncertain, due to the simultaneity with respiration. NEE on the other hand is much smaller, but directly of influence on the CO_2 concentration and thus of key importance in understanding climate change. Small measurement errors in NEE may lead to faulty estimations of the carbon cycle and wrong predictions for where the climate is headed. Therefore, reliable estimates of global GPP and NEE are of the utmost importance.

Besides, the effect of climate change on carbon fluxes is subject to great uncertainties (Cox et al., 2000; Sitch et al., 2008). For example, the effect of temperature on GPP is still unknown. Tait and Schiel (2013), for example, found a Q10 of about 2 for GPP, whilst Raulier et al. (2000) found no real temperature effect on GPP. Also, in a warmer climate, nutrient availability might increase due to higher rates of decomposition. Predominantly in nutrient-limited ecosystems, this would result in increased GPP (Melillo et al., 1993). However, due to climate change, soil moisture content might decrease in some regions, thereby also decreasing GPP (Melillo et al., 1993). On the other hand, increased carbon concentrations in the atmosphere may facilitate carbon uptake by plants, and as such effectively increase GPP. This is known as the carbon fertilisation effect (Heimann and Reichstein, 2008). The combination of these (and other, not included here) effects is known as the carbon-climate feedback. In order to gain more insight in this carbon-climate feedback and in how the biosphere responds to climate change, the availability of reliable estimates for global GPP and NEE are desired.

1.2 Modelling the carbon cycle

Because global carbon cycle measurements are impossible, model studies need to be conducted in order to get reliable estimates of the magnitude and distribution of the carbon fluxes on earth. These studies are typically conducted using three different types of models (See also Table 1.1):

Process-based models A commonly used traditional method for global GPP and NEE estimation is the assessment of process-based models (see Cramer et al. (1999) for an overview). Process based carbon cycle models try to mimic how the global ecosystems works, based on theory and equations. However, these theories and equations always depend on simplification and computational power. Therefore, process-based models are prone to biases that lead

to faulty estimations of carbon fluxes (Jung et al., 2007). However, process based models can provide knowledge on interactions in the climate system and thereby increase our knowledge on past and future climate change.

Conceptual models Another method of predicting GPP is using so called *conceptual models*, where the model is generalised greatly and the true relation between parameters is not known, but fitted to reproduce a known value, such as observations. Conceptual models are based on thorough, but simplified system knowledge. Conceptual models are often used in combination with satellite products, such as the research done by Xiao et al. (2004b,a), or as simplification to increase usability, such as Landsberg and Waring (1997).

Data-based models As process-based models are derived from theory and depend on simplifications, they are prone to biases (Jung et al., 2007). Besides, conceptual models often tend to over-simplify or ignore the complex processes related to the carbon cycle, resulting in faulty and biased estimations as well. Therefore, in more recent researches studying the carbon cycle, data-based models have become popular (Jung et al., 2009; Beer et al., 2010; Jung et al., 2011; Bodesheim et al., 2018).

Creating, or rather training data based models is also known as machine learning. Data-based models use measured or observed data, rather than underlying processes, to derive a set of rules for the data. Based on these rules, derived from the available data, predictions can be made. A widely used and very simple example of a data-based model is the linear regression model: $y = ax + b$, where a and b are the derived rules, based on known x and y values. Using the calculated a and b, new y values can be estimated, based on observed x. Note that for this estimation of y, no knowledge about the interaction and underlying processes between y and x is needed, as only the quality of the estimation is of importance and not the underlying mechanics.

More complex well-known machine learning algorithms include neural networks, principle component analysis, Bayesian algorithms and model trees. Although it is outside the scope of this research to explain the machine learning algorithms in detail, model trees and neural networks both use regression in an extended form: model trees first split up the data into smaller datasets, so that a linear regression fits better through the smaller dataset, resulting in better predictions than linear regressions. Neural networks are a multi-layer implementation of regression models, inspired on how the human brain works. In a neural network, input data is processed by a layer of linear regression models. The processed data is forwarded to the next layer of linear regression models, until the final layer of regression models is reached and a output is generated. Because of the multiple layers, complex, non-linear processes can be resolved, which could not have been resolved using usual regression.

The benefit of machine-learned models is that they are not dependent on theories, simplifications or calibration. Therefore, they can make very accurate predictions without a bias towards existing theories or assumed simplifications. A drawback, however, is that most data-based models have a high black-box calibre and are very sensitive to variations in input data.

1.3 Improving GPP estimates

1.3.1 Local uncertainty

As stated before, GPP is the largest land carbon sink. However, as it goes hand in hand with respiration, GPP can not be measured directly. NEE can be partitioned in GPP and respiration and is measured locally by the eddy covariance (EC) method (Baldocchi, 2003). The EC method has become more and more popular over the past few years, also because it is the only continuous, non-destructive way to measure ecosystem carbon fluxes (Beer et al., 2010). World-wide EC measurements synthesised by the FLUXNET network (Baldocchi, 2003). The most recently produced FXUXNET dataset contains 15000 site months from over 200 sites (Miyata et al., 2018).

As these EC towers only measure NEE, GPP is inferred from the available data. However, as EC towers only meausure local carbon fluxes. Therefore, in order to get a reliable estimate of global GPP, models that calculate GPP at a global scale are needed. In these models, the local GPP estimations by FLUXNET can be used as validation data for a process based model or as training data for a data based model.

1.3.2 Global GPP estimations and uncertainties

Due to the various different models, simplifications, assumptions and data extraction methods, global GPP estimates are quite diverse. Table 1.1 shows an overview of some estimations of global GPP. As can be seen from the table, the estimates range from 107 to 175 PgC/year.

Table 1.1: Global yearly GPP predictions by various authors

Author	Global GPP (PgC/year)	Type of model
Zhao et al. (2005)	109.3	Satellite data
Piao et al. (2009)	133 ± 15	Process-based model (ORCHIDEE)
Piao et al. (2009)	151 ± 4	Process-based model (CLM4)
Piao et al. (2009)	140	Process-based model (TRIFFID)
Beer et al. (2010)	123 ± 8	Data-based models
Yuan et al. (2010)	110.5	Conceptual model
Welp et al. (2011)	150-175	Oxygen isotope analysis
Jung et al. (2011)	119 ± 6	Data-based model
Mao et al. (2012)	114	Satellite data
Koffi et al. (2012)	146 ± 19	Inverse modelling
Yebra et al. (2015)	107	Conceptual model
Zhang et al. (2017)	121.6-129.4	Conceptual model
Joiner et al. (2018)	140	Conceptual model
Bodesheim et al. (2018)	128.5	Data-based models

1.3.3 Opportunities to improve data-based models predicting GPP

Data-based models are inherently limited by the available data and can therefore be improved by higher quality data. Beer et al. (2010); Jung et al. (2011); Joiner et al. (2018) and Bodesheim et al. (2018) all used FLUXNET data combined with additional data sources with a global coverage, such as the fraction of absorbed photosynthetically active radiation (fAPAR), temperature and vegetation structure to make a prediction of global GPP. These estimates obtained using machine learning vary greatly however (Table 1.1). In order to decrease the uncertainty in global GPP estimates, new observations, such as satellite-observed sun-induced fluorescence (SIF), can be used as input for data-based models. As SIF is a by-product of photosynthesis, it correlates to GPP (Figure 1.1). Therefore, SIF provides new opportunities to improve and ascertain estimations of global GPP (Joiner et al., 2014; Duveiller and Cescatti, 2016). Therefore, Joiner et al. (2014) and Yoshida et al. (2015) have argued that estimates of GPP could be improved using SIF.

As stated, SIF is a by-product of photosynthesis. As plants photosynthesise, they take up light to combine CO_2 and water into sugars and oxygen. Most of the absorbed light energy is used to drive photosynthesis. However, the energy can also be dissipated, or it can be re-emitted as light with a longer wavelength. The latter process is called sun-induced fluorescence (SIF) (Maxwell and Johnson, 2000). As the three previously mentioned processes compete with each other (Baker, 2008), SIF is directly related to photosyntetic activity. Although the fraction of absorbed light that is re-emitted as SIF is only about 1%, these re-emitted photons can be detected by satellites such as GEOSAT or GOME-2 (Tol et al., 2014; Sun et al., 2015; Schaik, 2016). The benefit of SIF over other variables related to GPP, such as fAPAR, is that SIF is physiologically related to carbon uptake and shows a strong similarity to GPP anomalies and seasonal cycles in GPP (Joiner et al., 2014; Koren et al., 2018). The relation between SIF and GPP for two locations, one in Europe and one in Africa, is shown in Figure 1.1. The figure clearly shows that SIF correlates well with GPP. However, as shown in the bottom part of the figure, GPP and SIF do not always have the same correlation, as Parazoo et al. (2014); Li et al. (2018) have shown.

The high spatial variability of SIF is shown in Figure 1.2. The figure clearly shows higher SIF in the tropics and lower values near the poles, as is expected as GPP in the tropics is higher than the GPP in the mid-latitudes. The figure shows a lower SIF value, corresponding to a low GPP in the Sahara as well, both represented as low SIF values or no SIF values at all. High SIF values are observed in the northern hemisphere mid-latitudes, such as eastern United

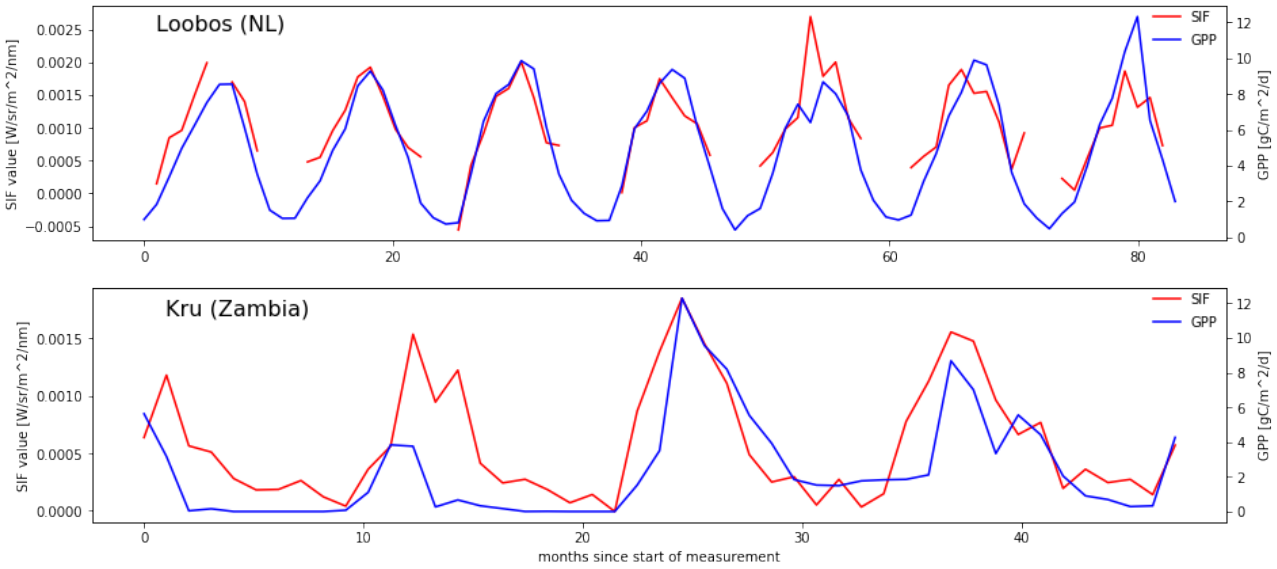

Figure 1.1: SIF observations and GPP measurements at the Loobos site in the Netherlands and Skukuza (Kru) site in Zambia. The relation between SIF and GPP is very clear in the Loobos site, however, a lot of SIF data is missing. In the Zambian site, the relation is less clear defined, although SIF and GPP follow the same trend .

States, corresponding to the expected high GPP values in the northern hemisphere summer (Huston and Wolverton, 2009; Jung et al., 2011).

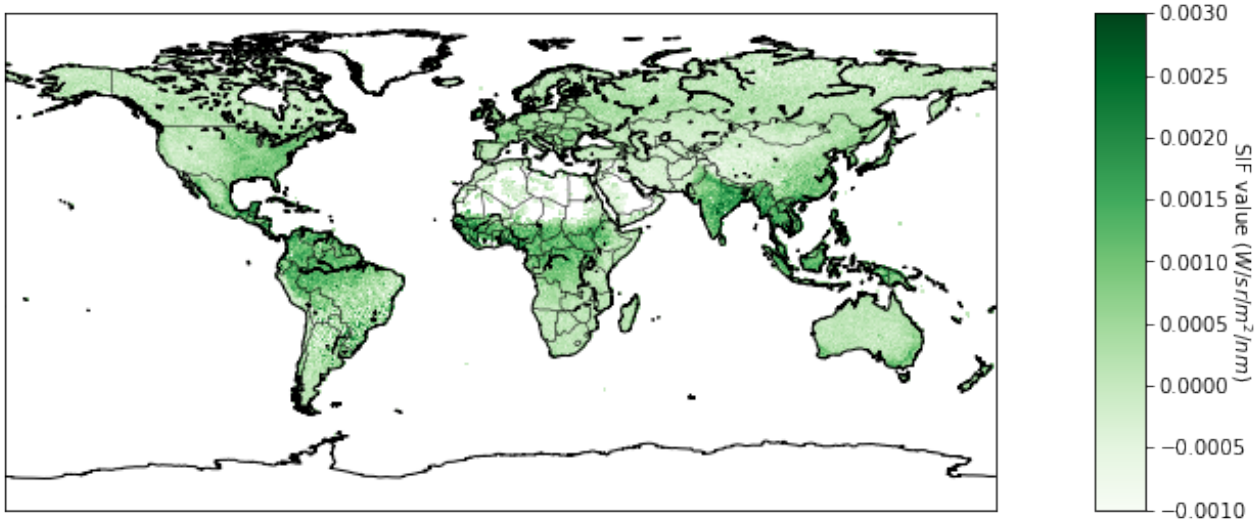

Figure 1.2: SIFTER satellite observations of August 2007. The spatial distribution of SIF is clear from the figure. In the tropical latitudes, more SIF is emitted. In contrast, in the polar latitudes, the observed SIF value is less.

In previous research, SIF has been used to assess drought stress in the Amazon (Lee et al., 2013; Koren et al., 2018) and estimate GPP (Duveiller and Cescatti, 2016). Duveiller and Cescatti (2016) found that using spatially down-scaled SIF to predict GPP can yield equally good results as dedicated GPP products, even without calibration. Koren et al. (2018) have shown that SIF can indicate reduced GPP due to El-Niño related droughts. Joiner et al. (2014) showed that data-driven models including fAPAR tend to overestimate GPP, as the photosynthetically-active period is over-estimated. Therefore, they argue that including SIF in global models can improve global GPP predictions.

1.3.4 Aim and objectives

In previous research, data-based models have been used to provide global GPP products that are as least as possible dependent on theories or simplifications. Also, SIF has previously been used as proxy for GPP. Therefore, it is expected that the inclusion of SIF observations into data-based models will improve estimates of total GPP fluxes, as well as estimates of the seasonal cycle, inter-annual variability and anomalies of GPP. The aim of this research is to predict GPP fluxes using a data-based model and assess the added value of SIF for these GPP predictions. To do so, we distinguish the following objectives:

- *To create a data-based model that estimates GPP on FLUXNET sites*

  FLUXNET data, complemented with meteorological data and data on the vegetation state and SIF will be used to train data-based models.

  The results will be used to assess which SIF features can provide additional information on the GPP, and GPP predictions made with and without SIF will be compared.

- *To assess what SIF features should be used in training the data-based models and how this affects the resulting GPP predictions*

  Finally, data-based models for global GPP predictions are made and applied. The predictive quality of these models, both with and without SIF, will be assessed and models with and without SIF will be compared to one another, to measurements and to state-of-the-art data-based models.

- *To successfully apply the data-based models for a global GPP prediction and assess the differences between models with and without SIF for global GPP fluxes*

The results are used to get a deeper understanding of both GPP and the potential of SIF for estimating GPP.

1.4 Improving estimates of NEE

Besides the unresolved uncertainties in GPP, there are also uncertainties in the sum of the plant uptake of carbon and the respiration: Net Ecosystem Exchange (NEE). As stated before, NEE is very small, but directly on the influence of atmospheric CO_2 levels and therefore of great importance in past, present and future climate. Contrary to GPP, NEE can be measured directly by eddy-covariance towers. However, these measurements do not have global coverage and are very susceptible to micro-meteorological variations. In order to assess global carbon fluxes and gain understanding in the mechanics of the carbon cycle, models with global coverage are needed.

1.4.1 NEE models and data assimilation

As NEE is of direct influence on the CO_2 concentrations in the atmosphere, CO_2 measurements can be used to validate models directly, but also to constrain CO_2 fluxes. This latter technique, using both observations and output of a process based model to find the state of the system, is known as data assimilation.

One example of a global carbon model using data assimilation is the CarbonTracker Europe (CTE) model (Peters et al., 2007; ESRL, 2005; van der Laan-Luijkx et al., 2017). In the CTE model, the CO_2 exchange between the surface and the atmosphere is represented as the sum of fossil fuel emissions, fires, terrestrial biosphere exchange and ocean exchange. Because CO_2 is a long-lived trace gas, atmospheric transport is important in resolving the CO_2 fluxes. In CTE, atmospheric transport of CO_2 is resolved by a global transport model. By comparing observed and simulated CO_2 concentrations, the inverse model calculates scaling factors for the biosphere and ocean carbon fluxes. By applying these scaling factors, a more robust estimate of the surface CO_2 fluxes is calculated. These more robust estimates are called the *posterior* flux; the unscaled fluxes the *prior*. Because scaling factors are applied to the prior

fluxes, the posterior depends on the prior. Therefore, a bias in the prior potentially results in a biased posterior, and inversely, a better prior results in smaller errors in the posterior fluxes.

One possible method for improving the prior model is to first derive a statistical relationship between the prior model and the posterior fluxes. This is known as model output statistics (MOS). To illustrate; if in a weather forecast model the model generally under-estimates precipitation with westerly winds by 2mm, the forecasts could be improved by increasing the expected precipitation with 2mm. This very principle can be applied to biosphere and carbon cycle models. In this thesis, the MOS is done by data-based models that predict the residuals between the prior and posterior flux. These models is then used to improve the prior NEE flux of CarbonTracker Europe. By doing so, the bias in the prior model can be reduced, resulting in a tighter fit of the posterior fluxes. Moreover, information on the driving forces of the mismatch between the prior and posterior can be obtained.

1.4.2 Aim and objectives

This results in the two main aims of the second part of this research:

1. Predict the mismatch between the prior and posterior NEE from CTE by using machine learning.

   It is expected that the data-based models improve simulated prior carbon fluxes. However, due to the complexity of the carbon cycle, the fluxes are not expected to be resolved perfectly.

2. To use data-based models to deepen our knowledge on the predicted carbon fluxes according to CTE.

   It is known that the residuals of the prior model show a strong seasonal cycle. Therefore, it is expected that variables that show a strong seasonal cycle can explain some of the variance in the model. However, as the residuals are not only seasonal, also other variables, such as land-use, are expected to explain some of the variance in the residuals.

. To reach the aims, the following objectives are distinguished:

- *To successfully set-up a data-based model that predicts the residuals of the process-based model*

  From these data-based models, information on the residuals can be obtained. It is investigated what variables are important to predict the residuals, and why these variables are so important.

- *To investigate whether there are variables that explain the residuals of the prior model, and if so, which variables are most effective in this.*

  With the most important variables, simpler models will be set-up in order to reduce the computational costs of CTE even further. The quality of these models is assessed.

- *To improve the quality of the prior NEE flux in the CTE model.*

  By using an improved prior flux, the resulting posterior carbon flux is less biased.

1.5 Outline

In the first part of this thesis, the research objectives posed in section 1.3.4 will be assessed. In order to do so, in chapter 2, the data and methods for this research will be elaborated. In chapter 3, the added value of SIF for local and global predictions of GPP will be discussed respectively. The methods, data and results are discussed in chapter 4. Finally, in chapter 5, a conclusion of Part I will be presented.

In the second part of this thesis, the research objectives as stated in section 1.4 will be addressed. The data and methods to do so are described in Chapter 6. The results are described and discussed in Chapter 7. A summarising discussion is presented in Chapter 8, followed by a conclusion in Chapter 9. Finally, a synthesis on using machine learning in simulating the carbon cycle is given.

Part I

Using SIF as input for data-based models for predicting GPP

Chapter 2

Methods and data

2.1 Methods

In this section, the methods used in this research are explained. First, the machine learning algorithm used is explained. Secondly, set-up and tuning of the model are elaborated. Finally, the model evaluation is discussed.

2.1.1 Set-up of the model

Following Jung et al. (2009); Beer et al. (2010); Jung et al. (2011) and Bodesheim et al. (2018), model tree ensembles were used to predict GPP from environmental drivers, such as temperature, leaf-area-index and the fraction of absorbed photosynthetically active radiation (fAPAR). Model trees are a machine learning algorithm that uses if/else statements to divide large datasets into smaller, more uniform subsets where the variance of the target value is low.

Model trees are used because of three key characteristics of model trees: 1) GPP is known to have a non-linear response to various variables, which can be resolved by the use of model trees. 2) Model trees have been used in the past and performed well (Frank et al., 1998; De'Ath, 2007; Jung et al., 2009; Beer et al., 2010; Jung et al., 2011; Bodesheim et al., 2018) 3) Model trees can be visualised, and therefore have a low black-box calibre. Besides, the importance of the variables used in the model can be assessed in model trees, even further decreasing the black-box calibre of the model.

An example of a dataset where a model tree is of use is shown in Figure 2.1. If the goal is to predict the sepal width of an iris, a normal linear regression through all data would probably result in an incorrect estimate. This is shown by the blue line in Figuere 2.1. However, first splitting the data on species of iris before doing a regression, would improve the estimate greatly (yellow, red and green line in Figure 2.1). A model tree follows this same line of reasoning of splitting up larger datasets into smaller datasets with a lower variance.

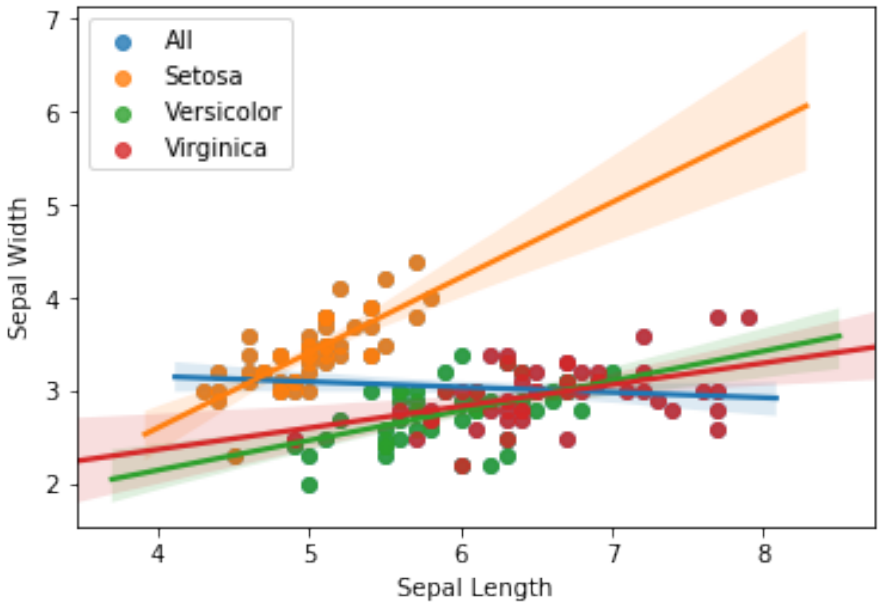

Figure 2.1: Example of a dataset where a model tree could be of use. The colour indicates the variety of iris. It is clear that the Setosa variety has a different sepal length/sepal width distribution than Virginica and Versicolor. The blue line shows the regression line through all data, indistinct of variety. The orange, green and red line show the regression through only data of Setosa, Versicolor and Virginica respectively

An example of a visualised model tree is shown in Figure 2.2. In this figure, the circles indicate splits (if/else statements) and the values in the squares indicate the target values the model is predicting (leaves). Note that all trees trained in this research will be deeper (i.e. have more splitting points) than the tree shown in Figure 2.2.

Despite the previously mentioned benefits, using model trees also comes with two potential problems. The first one occurs if the target value is imbalanced. For example, when 99% of the target values is 0, a model would score a 99% accuracy score when only predicting zeros. In this research, it is assumed that this is not the case and that the data continuous and balanced. The second possible problem is over-fitting. This is discussed below.

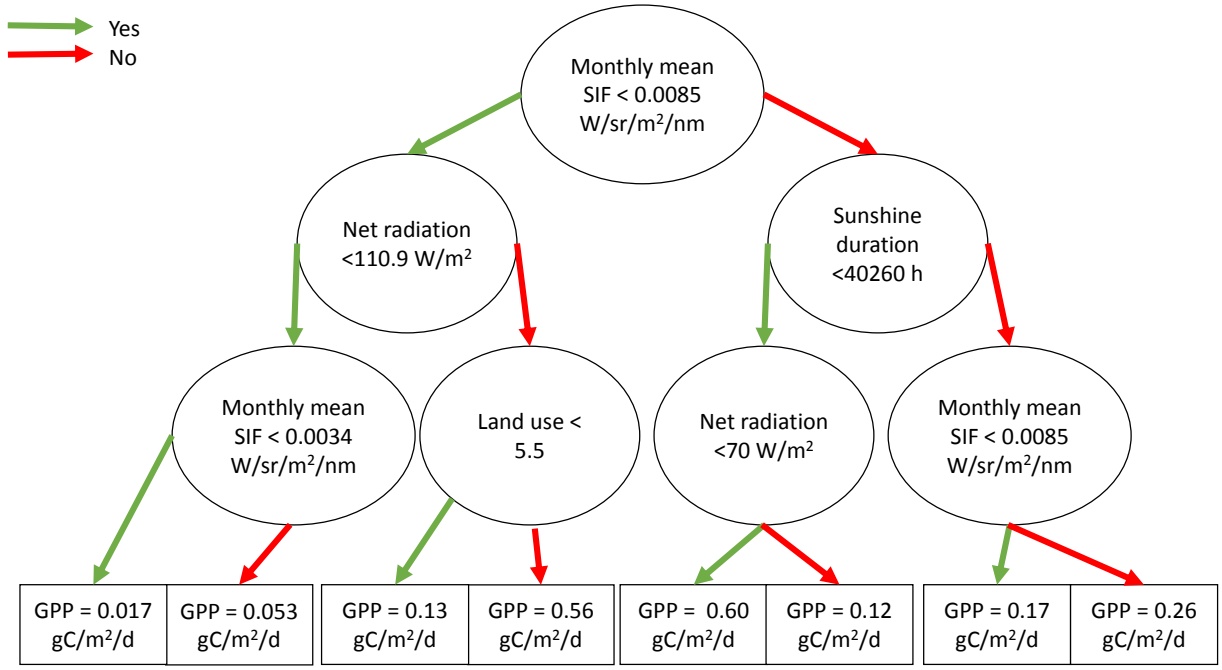

Figure 2.2: Example of an undeep model tree. In this tree, first, the data is split based on the monthly mean SIF value. If the monthly mean SIF of a datapoint is smaller than 0.00085 ($W/sr/m^2/nm$), the net radiation is checked for that data point and so forth. Ultimately, in the bottom of the figure, the square boxes indicate the predicted monthly mean GPP ($gC/m^2/d$) for that data point. This process is repeated for each data point in the dataset. Note that land-use is converted to an ordinal feature.

Prevention of over-fitting

One of the most predominant issues in machine learning is over-fitting. Over-fitting can be seen as the model not only having learned the trends and general structures in the data, but also the noise. This results in the model scoring excellent on data that has been used to train the model, but poorly on new data. In this section, the three ways that are used in this research to prevent the model from over-fitting are discussed.

The first method used to prevent over-fitting is to stop the building of the tree before all data is stratified in a separate class, preventing that a regression is done over only one data point. An example of a stop criterion is a certain minimum number of data points in a leave (with one data point per leave, the model has been over-fitted maximally) or maximum tree depth (i.e. the amount of consecutive if/else statements). Finding the optimal stop criteria is done using a so-called grid-search function, which uses a selection of stop criteria to train a model tree. The function then returns the stop criterion that performed best (i.e. results in the lowest testing error) (Buitinck et al., 2013).

The second way used to to prevent over-fitting is to train multiple trees on randomly selected training data and average the predictions by the trees. This is done by selecting multiple sets of train and test data at random. Then, a single tree is trained on a selection of training data. Because the sets are chosen at random, the training data varies per tree and every new tree is different. The combination of trees is called a *forest* or a *model tree ensemble*. The prediction of the forest is the averaged prediction of all trees. Due to their increased robustness, model tree

ensembles are used widely to predict a variety of features, for example classification (Frank et al., 1998) and numerical prediction (Jung et al., 2009).

Additionally, instead of building a forest consisting of fully grown trees, which all have a chance of being over-fitted, *gradient boosted* trees are used to decrease the chance of over-fitting in this research. Gradient boosted trees are a form of a model tree ensemble with shallower trees. The trees are built sequentially and predict the error of the previous tree, so it can be used to improve the previous estimates. This results in robust models that are less susceptible to over-fitting. For a more in-depth discussion, see (Chen and Guestrin, 2016).

The third way that is used in this research to prevent over-fitting is cross-validation. In the k-fold cross-validation used in this research, the training data is split up in k groups and for each group, a model tree ensemble is trained on the data of all but the k^{th} group. The tree is then tested using the data from k^{th} group. Subsequently, the model with the lowest root mean square error is selected (Krstajic et al., 2014).

Feature importance

For the assessment of the importance of different variables in predicting GPP, the *gain* index is used. Gain represents the decrease in mean square error (MSE) of the dataset, due to that respective variable (Lundberg and Lee, 2017). Although other measures of feature importance are implemented in the algorithm, only the gain is used as it is deemed the most informative.

Feature selection

As the model only uses the variables that increase the predictive quality of the model the most, variables that do not add to the quality of the model are left out. Also, variables that correlate with one another are left out, because they do not add additional information on the system as the improvement gained by splitting on one variable could also be achieved by splitting on the other. Selecting the important variables is called *feature selection*. Feature selection results in a more clear picture of GPP drivers and therefore a model that is easier to interpret, compared to a model without feature selection. Besides, decreasing the amount of explanatory variables decreases the computational costs (Li et al., 2016).

Because feature importances may interact, e.g. because features correlate, using all variables could result in faulty estimates of the feature importance. Therefore, in order to assess the most important variables in a model, recursive feature elimination is used. In recursive feature elimination, the variables with the lowest gain are dropped from a model, until a given number of variables is left. The number of allowed features is selected as such that the model is as simple as possible, but the predictive quality of the model does not suffer. This is done using the Akaike information criterion (AIC). The AIC is used to select the best model according to the AIC (Akaike, 1998; Posada and Buckley, 2004; Burnham and Anderson, 2004). The AIC is based on the predictive quality of the model and penalises model complexity and is calculated as

$$AIC = (2 \cdot numvars) + n \cdot log\left(\frac{\sum_{i=1}^{n}(y^i_{measured} - y^i_{predicted})^2}{n}\right), \tag{2.1}$$

where $numvars$ is the number of variables used in the model, n is the is the sample size, $y_{measured}$ are observed target values, and $y_{predicted}$ is the predicted target value, according to the data-based model.

The first term on the right hand side represents the penalty term, increasing with the number of variables. The second term on the right hand side is the logarithm of the mean squared error (logMSE), corrected by the sample size, which represents the predictive quality of the model. It is expected that the predictive quality of the model is higher when more variables are used, as the model is better fitted. The logMSE decreases with increasing model quality, also decreasing the AIC. Therefore, models with lower AIC are deemed to be better models, as they have high predictive quality and low complexity (Posada and Buckley, 2004).

As shown in Equation 2.1, the AIC of a model depends on the arbitrary sample size. Therefore, individual AIC values can not be interpreted, and the AIC is rescaled by

$$\Delta_i = AIC_i - AIC_{min}, \tag{2.2}$$

where AIC_i is the AIC of the i^{th} model and AIC_{min} is the minimum AIC of the models tested. Δ_i is the expected information loss by using the i^{th} model, rather than the model with the minimum AIC (Burnham and Anderson, 2004).

Because the Δ_i values do not depend on the scaling factors found in Equation 2.1, the AIC values can not be interpreted as a strength of evidence, but $Delta_i$ values can. Models with a $\Delta_i < 2$ have a high probability of being a better description of the system; whereas models with $\Delta_i > 2$ have a low probability of describing the system better than the model with AIC_{min} (Burnham and Anderson, 2004). Therefore, models with $\Delta_i < 2$ are used in the model and variable assessment.

Feature creation

Feature creation is the process by which variables, or features, are combined to create new variables that could explain more of the variance in a dataset. However, feature engineering could potentially bury the true potential of SIF under un-physical combinations of features. As the objective of this thesis is to assess the added quality of SIF for GPP, rather than a perfect GPP prediction, feature engineering is not applied in this thesis. Nevertheless, yearly maxima and minima of features are calculated and used as input as these have physical meaning.

2.1.2 Validation

To assess the quality of the machine learned predicitons (MLP), the predictions will be tested against local FLUXNET data (Baldocchi et al., 2001) and a state-of-the-art data-based model resolving GPP fluxes at a half-hourly scale (Bodesheim et al., 2018). This section will elaborate on how the model will be validated on both local and global scale.

Local (FLUXNET)

To train the model trees, the available FLUXNET data will be split in training (about 90%) and testing (about 10%) data. The model will be trained on the training data, with as target the GPP and validated on the test data. The metrics used to assess the model quality are Nash-Sutcliffe model efficiency (Moriasi et al., 2007) (from here on: model efficiency, ME), bias (where a positive bias indicates over-estimation of the GPP by the model), normalised root mean square error (NRMSE) and time-series of the prediction and the measured GPP value. The (unitless) NRMSE can be seen as the normalised variance of the errors. Note that for regression procedures, the ME is equal to the coefficient of determination (R^2) and is calculated following equation 2.3.

$$ME = 1 - \frac{\sum_{n=1}^{N}(GPP^n_{simulated} - GPP^n_{observed})^2}{\sum_{n=1}^{N}(GPP^n_{observed} - \overline{GPP}^n_{observed})^2}, \tag{2.3}$$

where $GPP^n_{simulated}$ is the simulated GPP on FLUXNET datapoint n and $GPP^n_{observed}$ is the observed GPP at the FLUXNET datapoint. A ME > 0 indicates that the model predictions are better than predicting the mean, and a ME of 1 indicates a perfect model.

NRMSE is the residual variance and is calculated using

$$NRMSE = \frac{\sqrt{\frac{\sum_{n=1}^{N}(GPP^n_{simulated} - GPP^n_{observed})^2}{N}}}{\overline{GPP_{observed}}}, \tag{2.4}$$

and the model bias is calculated according to

$$bias = 100 * \frac{\sum_{n=1}^{N}(GPP^n_{simulated} - GPP^n_{observed})}{\sum_{n=1}^{N}(GPP^n_{observed})}. \tag{2.5}$$

To assess the model quality and the model robustness, different tests will be done on the models trained on the FLUXNET measurements. All tests will be done both including and excluding SIF.

Firstly, as a reference model, the dataset is split randomly in training and testing data, where 10% of the site-months is left out of the training data and used as test data. As this selection is random, the train and test data and therefore the predictive quality of the model will vary. In order to obtain valid model statistics, this is done 100 times.

Secondly, to ensure that the model has used GPP values from all ranges of GPP, the training data is selected as such that it contains data from all quantiles of GPP. For this, the data is split up in 100 quantiles. From every quantile, a random 10 percent of the data is used as test data, whilst the rest is used as training data. This simulation is called *Equal ratio*. In addition, to assess the potential problem of lower GPP values being over-represented, models are trained with equal amounts of low and high GPP values. This simulation is called *Equal representation*. To do so, the GPP values are binned and from every bin 33 values are used as training data. The number 33 is chosen to use GPP values from as many bins as possible. Figure 2.3 shows the density plot of the original, training and testing data for the Equal representation simulation.

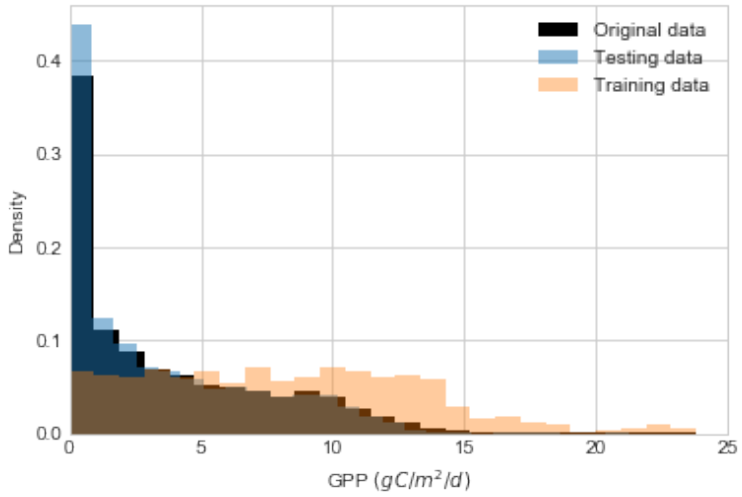

Figure 2.3: Density plot of the Equal representation data. The black bars show the original data, the orange bars the training data and the blue bars the testing data. Note that there is less less training data than testing data, and that the density does not indicate the amount of data-points, but the ratio of the distribution of the data points.

Thirdly, as the model is used to predict areas for which no local measurements are available, the model is used as well to predict the GPP of FLUXNET sites that have not been used for the training of the model. This will be done by randomly leaving 10% of the FLUXNET sites out of the training dataset. These sites are then used as test set. To get insights in the statistical variation, this is done 100 times as well. To address the possible problem of Europe being over-represented in the FLUXNET data (see Section 2.2), this will also be done for only sites from Europe. As there are only 66 sites in Europe, this analysis is done 50 times.

Fourthly, the model is used to predict global GPP from years of which no FLUXNET data is available yet. To test the quality of these predictions, one of the measured years will be left out of the training dataset. This year will then be used as test data. To get insight in a possible trend in errors, this is done for every year in the FLUXNET dataset.

Additionally, the model is used to predict 5 years of data, based on the previous three years. In doing so, the FLUXNET sites where no measurements are done before the 5 predicted years are excluded. Therefore, this simulation can be seen as a combination of the third and fourth simulation, as mentioned above.

Global validation

As a first validation of the global GPP product, FLUXNET data that is not taken into account in the training of the model due to missing values is used. If a GPP measurement is available in the dropped data, this measurement will be used to assess the quality of the global model. In total, just over 2600 site months can be used for validation this way. Figure 2.4 shows an overview of the amount of available months of FLUXNET observations for validation of the global GPP product. The figure also shows the predicted yearly GPP by Beer et al. (2010). The figure shows

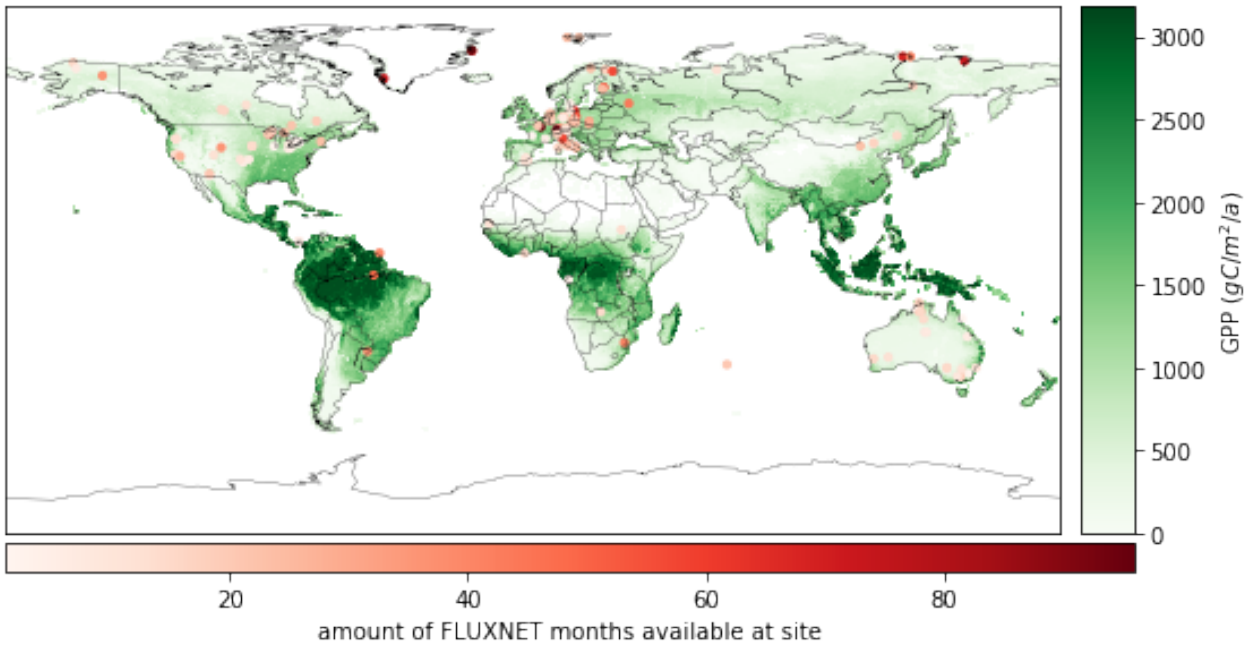

Figure 2.4: Global location of the FLUXNET sites that can be used for validation. The colour and of the dots indicates the amount of months available. The green shading shows the predicted GPP by Beer et al. (2010).

that there is a surplus of validation points available in western Europe, but the tropical rainforests, where GPP is predicted to be the highest, lacks validation points.

In addition to the validation on FLUXNET data, the global GPP product is compared to the products by Jung et al. (2011); Bodesheim et al. (2018). Jung et al. (2011) used a more complex algorithm, described in Jung et al. (2009). In this algorithm, a multiple linear regression is conducted in the leaves of the tree. Besides, the variables used can be indicated to be either used only for splits or both for splits and regression in the nodes. Additionally, they have used more (29) explanatory variables. The product by Bodesheim et al. (2018) is based on half-hourly observations. As these products are made using more advanced methods and more explanatory variables, they are used as benchmarks. In order to compare the predictive quality of these models, these GPP products will be compared to the FLUXNET data as well.

Unfortunately, after personal communication with both Paul Bodesheim and Martin Jung, it was concluded that they both did not asses feature importance in their research. Therefore, feature importance cannot be compared to current state-of-the-art models.

Workflow

An overview of the steps taken in this part of this thesis, including training and validation, using different methods, is shown in Figure 2.5. In this figure, the blue boxes indicate input data, the grey boxes indicate created data, the orange boxes represent models that are created using the created data and the green boxes represent output data. The red boxes indicate the results.

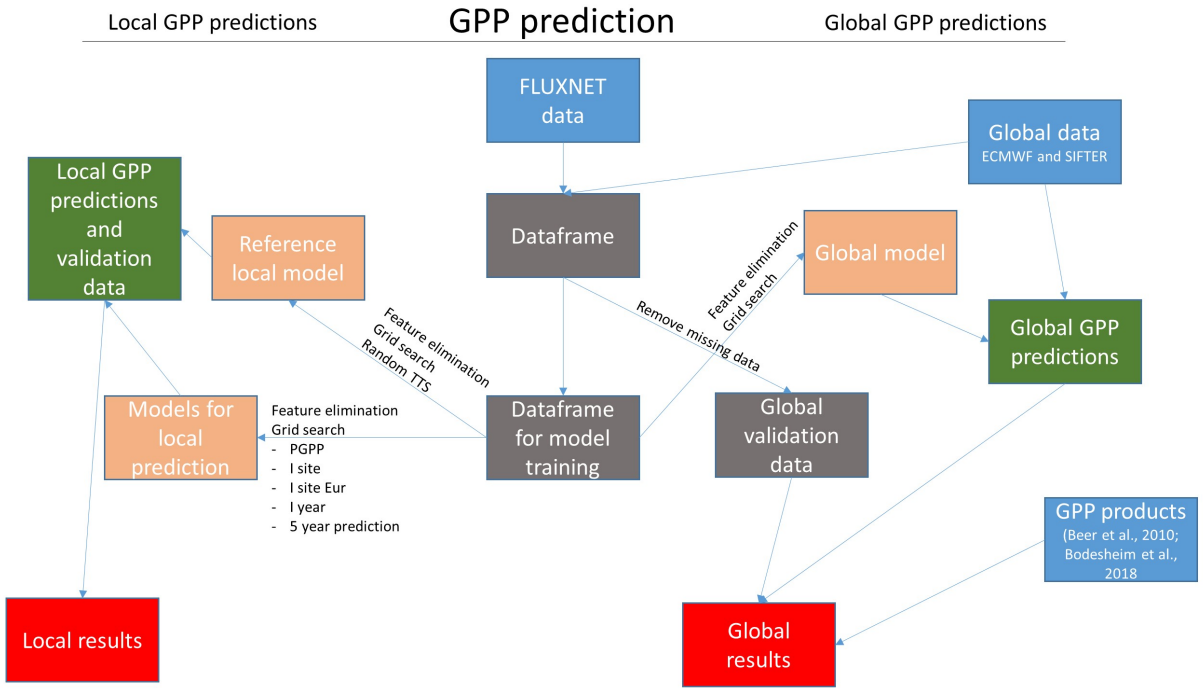

Figure 2.5: Basic flow diagram of the steps taken in this research. The left part of the figure shows the local validation, as described in Section 2.1.2, whilst the right side of the figure shows the global validation, as described in Section 2.1.2

2.2 Data used as input for the machine learning algorithms

In this section, the main data sources are shortly discussed. Also data pre-processing and the availability of the data are discussed. In order to capture seasonal variability, but reduce computational costs and chance of outliers, monthly averaged data is used in this research. As the SIFTER data is available on a 0.5 by 0.5 degree grid, this grid size is used.

2.2.1 Local data: FLUXNET

In order to get values for the drivers of GPP and GPP itself, the FLUXNET 2015 dataset (Miyata et al., 2018) is used. The dataset consists of 200 eddy-covariance towers, spread out over the entire world. Figure 2.6 shows the distribution of the measurement sites globally. A list of all sites, including the land-use, is shown in Appendix 12.1.

Partitioning of NEE into GPP

The FLUXNET data contains measurements of NEE, which are partitioned in GPP and respiration, according to $NEE = GPP - Respiration$. Estimates of GPP are derived from the observed NEE using the algorithms proposed by Lasslop et al. (2010) and Reichstein et al. (2005). In this research, GPP calculated according to the method described by Lasslop et al. (2010) is used, as this method calculates GPP using daytime data, removing the potential bias due to suppression of turbulence. Also, this method includes GPP limitation by water stress. In practice however, the two calculation methods result in very similar results, with a correlation factor of 0.97 and an average absolute difference of 5%. Following Lasslop et al. (2010), GPP is calculated according to

$$GPP = \alpha\beta\frac{R_g}{\alpha R_g + \beta}, \tag{2.6}$$

where α represents the initial slope of the GPP response to light, β is the maximum CO_2 uptake rate at light saturation and R_g is the global radiation. For the other used parameters and constants, see Lasslop et al. (2010). α and β are derived from the measured NEE and the calculated respiration: $GPP = NEE - respiration$, where

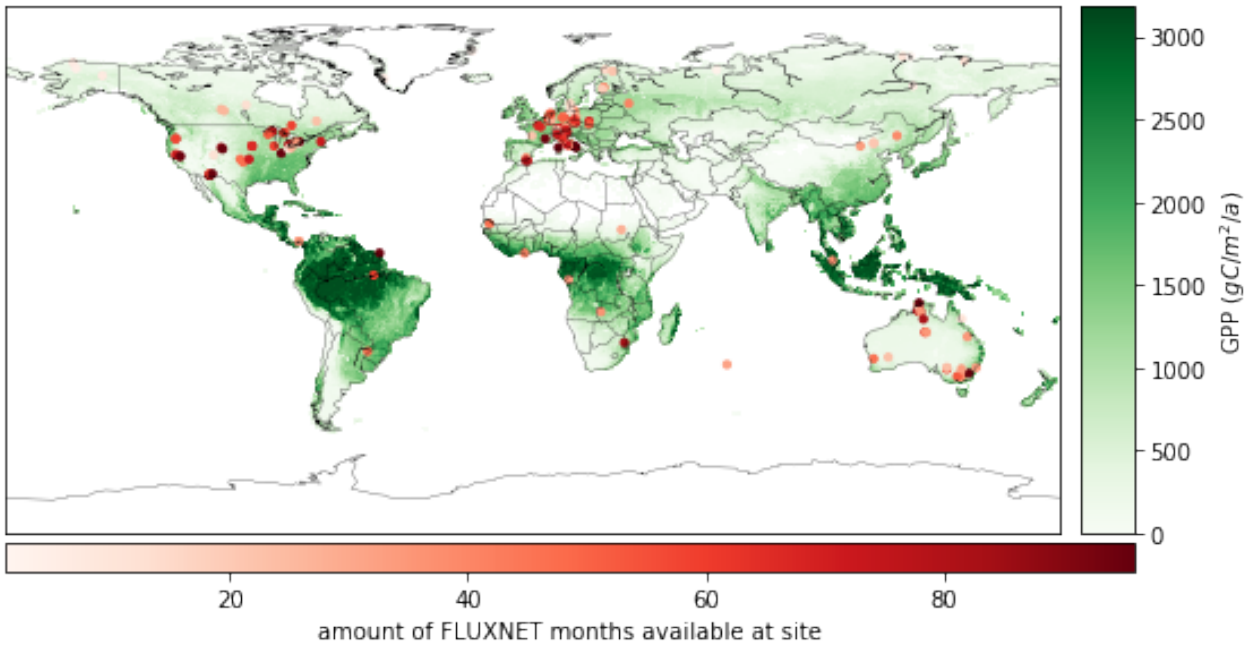

Figure 2.6: Global location of FLUXNET sites with measurements between 2007 and 2015. The colour of the dots indicates the amount of months available. The green shading indicates the GPP as predicted by Beer et al. (2010).

respiration is calculated according to

$$respiration = R_b \cdot e^{\left(\frac{1}{T_{ref} - T_0} - \frac{1}{T_{air} - T_0}\right)}, \tag{2.7}$$

where R_b is the base respiration at T_{ref}, which is equal to 15 degrees °C. T_0 is the temperature at wich no respiration takes place, which is assumed to be -46,02 degrees °C. T_{air} is the air temperature.

To account for decreased GPP in droughts, β in Equation 2.6 was made dependent on water vapour pressure deficit (VPD) by using an exponential decreasing function if the VPD is greater than 10 hPa Lasslop et al. (2010). Therefore, the GPP decrease during droughts can be captured. According to Lasslop et al. (2010), including the VPD resulted in the model being able to reproduce peak fluxes and removed the systematic pattern in the GPP residuals of the model.

Slight uncertainties in the variables in equation 2.6 and 2.7 result in a range of plausible GPP values. Therefore, in the FLUXNET 2015 dataset, a reference GPP has been selected based on the model efficiency of GPP calculations using a range of variables. The GPP is selected by starting with 40 different estimations for GPP, based on equations 2.6 and 2.7. The model efficiency has been calculated between the respective estimation and the other 39. The GPP estimation with the highest model efficiency is selected as reference (Flu, 2018). In this research, these reference GPP estimates are used as target values to train the model tree ensembles.

Data pre-processing

To ensure data quality, the FLUXNET data is already pre-processed by applying checks and filters to the raw data. NEE observations are both corrected for storage and are de-spiked (i.e. biased observations due to quasi-systematic measurement errors, such as water droplets on sonic anemometers are removed), according to the method described in Papale et al. (2006). Besides, as Loescher et al. (2006) state, eddy-covariance measurements during periods with low turbulence (indicated by low friction velocity u*) tend to underestimate NEE. Therefore, data obtained when u* was below a certain threshold was discarded. This was done using methods described by Barr et al. (2013) and Papale et al. (2006). Also, in periods with low turbulence, CO_2 can accumulate under the canopy. Therefore, in order to avoid false emission pulses after these periods, data obtained half an hour after periods with low turbulence was also discarded.

Other variables included in the FLUXNET dataset

Besides GPP and NEE, the FLUXNET data also contains data on, amongst other, meteorological variables such as the temperature, radiation and land use. These drivers of GPP are used to train the data-based models. However, FLUXNET does not measure for example SIF, precipitation and evaporation. Drivers that are not measured on the FLUXNET locations are added based on a global data. See Section 2.2.2 for an in-depth discussion on the global data and Table 2.2 for an overview of the data used.

FLUXNET sites are labelled with an IGBP land-use (Turner et al., 1994). Table 2.1 shows an overview of the land uses and the amount of towers in the respective land use. The table shows that Cropland, Mixed Forest, Evergreen Needleleaf Forest and Grasslands are over-represented, whereas Deciduous Needleleaf Forests and Snow and Ice are underrepresented. This, combined with temperate regions being over-represented, as seen in Figure 2.6, may result in the model performing better in temperate regions with land uses such as cropland and evergreen needle-leaved forests. This also increases the need for cross-validation, as described in section 2.1.1.

Table 2.1: Land uses and the amount of FLUXNET eddy-covariance towers in the respective land use

Land use	Number of sites
Evergreen Needleleaf Forest	38
Mixed Forests	31
Cropland	29
Grasslands	22
Woody Savannas	17
Open Shrublands	14
Cropland/Natural Vegetation Mosaic	14
Deciduous Broadleaf Forest	12
Evergreen Broadleaf Forest	10
Savannas	9
Closed Shrublands	5
Urban and Built-Up	3
Permanent Wetlands	3
Water	1
Snow and Ice	1
Deciduous Needleleaf Forest	1

If a site month contains a missing value in one of the variables used as inputs in the model trees, the site month is dropped. For the FLUXNET data alone, this results in about 4% of the site months to be left out. When also the SIFTER data is dropped, this increases to about 17.5%.

2.2.2 Global data

In order to assess global GPP, using the models trained and validated on the local FLUXNET data, global measurements of GPP drivers are used as input data for the model trees. This section will shortly describe the data sources. All global data is scaled to 0.5 by 0.5 and monthly averaged, except stated otherwise. Table 2.2 shows an overview of all global data. In order to prevent the model from predicting GPP of for example the arctic or the Sahara regions, regions where more than 6 months of SIFTER data are missing are neglected in this research and are deemed to have a GPP of 0.

SIF: SIFTER

The main focus of this research is to assess the added quality of data-based models when SIF is used as input variable. Currently, the most known remote-sensing SIF product is the product as retrieved by the algorithm proposed by Joiner et al. (2014) (NASA SIF). Joiner et al. (2018) have used this SIF retrieval from the GOME-2 satellite as input for machine learning algorithms to estimate GPP. However, the validity of this SIF product has been questioned by Zhang

et al. (2018), as they argued that the decreased NASA SIF signal found by Yang et al. (2018) during drought in the Amazon does not reflect the expected decrease of GPP.

In order to try to improve on the NASA SIF, Sanders et al. (2016) developed a different algorithm for retrieving SIF from GOME-2 observations. This algorithm is called the Sun-Induced Fluorescence of Terrestrial Ecosystems Retrieval (SIFTER). SIFTER initially used a wide spectral fitting window. Because of this, also more principle components were needed for the SIFTER algorithm than for NASA SIF. Also, the reference area differs between the two algorithms. Where NASA SIF uses cloudy ocean, SIFTER uses a selection of non-vegetated pixels in the Sahara desert (Sanders et al., 2016). For a more in-depth discussion, see Kooreman et al. (2011); Schaik (2016).

Based on the algorithm by Sanders et al. (2016), Schaik (2016) made some key changes to the original SIFTER algorithm. A small fitting window was used, as well as less principle components (8, contrasting the 35 of the previous version). Also the reference area was changed to a clouded ocean. Due to these changes, higher measurements of SIF over the tropics were retrieved. Schaik (2016) found that the SIFTER shows a reduction in SIF during a massive drought in India in 2009, whereas the NASA SIF product shows no decrease at all. The new SIFTER product also showed a higher correlation with GPP than both the original SIFTER product and NASA SIF.

Albeit the good results, SIFTER produces negative SIF values over both the desert and high-altitudes. Intuitively, negative SIF values are impossible, as even with no photosynthesis the SIF would be 0. Currently, no explanation has been found for this. Another drawback of the SIFTER data-set is that it contains quite some gaps. This is because pixels with a cloud cover of > 0.4 are not retrieved, but also pixels where no SIF signal is retrieved are filtered out. As this missing data occurs mostly over the polar latitudes and Sahara, where no GPP takes place, the missing data is used to prevent the model from over-estimating global GPP, as described in Section 2.1.2.

One of the key benefits of SIF is that the anomalies of SIF are corresponding very well to anomalies in GPP (Koren et al., 2018). Therefore, anomalies of SIF are generated. This is done by fitting a mean seasonal cycle through the seasonal cycle of the retrieved SIF. The anomalies are then calculated as the deviation from the seasonal cycle.

The currently available SIFTER data-set has data from 2007 to August of 2017, summing up to a total of 128 months. The data-set covers the entire globe with a grid-size of 0.5 by 0.5 degree. The SIFTER data is quite noisy on regional spatial scales, such as the Amazon region. As the GPP and SIF correlate well on the FLUXNET sites however, this does not appear to be a problem (see Figure 1.1). If this does appear to be a problem however, possibilities of smoothing or averaging the SIFTER data are assessed.

Land cover: MODIS

Land cover schematically characterises biomes and other properties closely linked to biosphere-atmosphere interactions. Therefore, it is important to have an accurate representation of the global land cover as input for predicting GPP. As the vegetation at FLUXNET sites is classified according to the IGBP land cover classification, this classification is used as input data for the model trees as well (Friedl et al., 2010). As the land cover is not expected to change significantly over a year, yearly data land-cover data is used in this research. Yearly data, assembled by the MODIS satellite, is freely available at a 0.05 by 0.05 degree resolution (Maccherone, 2005). The land cover is regridded to 0.5 by 0.5 degree by taking median of the land covers present in the 0.05 by 0.05 degree grid. The IGBP land cover classification system differentiates between 18 different land cover classes, as can be seen in Table 2.1. Note that, since there are no eddy-covariance measurements done over barren soil, this class is not included in the table, but there are towers where the land cover is unclassified. This is shown as the *null* class in Table 2.1.

LAI dataset: MODIS

The leaf area index (LAI) indicates how densely the vegetation is covered in leaves. It is defined as the one-sided green leaf area per unit of ground area (Myneni et al., 2002). As areas with more leaves can potentially sequester more carbon, LAI is a proxy for potential GPP. Also, LAI is a proxy for seasonal variability of ecosystems, as deciduous trees have a lower LAI in the winter, corresponding to a lower GPP.

As LAI dataset, MODIS data is used, as this dataset is consistent with all biome types and particularly with woody vegetation, which is expected to have the largest influence on GPP (Maccherone, 2005) . Daily global LAI at 0.05 by 0.05 degree is freely available (Maccherone, 2005) and averaged to 0.5 by 0.5 degree. The LAI data is retrieved from reflected radiation, accounting for the architecture of the foliage, optical properties of the vegetation and the ground and atmospheric conditions following Knyazikhin et al. (1998).

fAPAR dataset: MODIS

fAPAR is the absorbed fraction of the photosyntetic active radiation and can be seen as the amount of sunlight a plant has taken up for photosynthesis. Although fAPAR, contrary to SIF, is not directly related to photosynthesis, fAPAR has proven effective in predicting GPP in previous research (Beer et al., 2010; Jung et al., 2011). Therefore, fAPAR is included in this research as well. As MODIS fAPAR has been found to perform well by Olofsson and Eklundh (2007), global daily fAPAR is taken from MODIS observations at a 0.05 by 0.05 degree resolution and averaged to monthly means of 0.5 by 0.5 degree (Maccherone, 2005).

Meteorological data: ECMWF

Climate fields, such as mean temperature, precipitation and the (derivative of the) incoming radiation are taken from the ECMWF-ERA interim (ERAI) reanalysis (Berrisford et al., 2011). The ERAI archive is a reanalysis dataset of the global atmosphere, including hydrological cycle. The ERAI has been found to agree well with observations (Dee et al., 2011; Simmons et al., 2014). Although the finest temporal resolution available is 6 hours, the data also includes monthly means. These monthly means is used as input for the model tree. If needed, also mean annual climate is extracted from this dataset. The ECMWF data is also downloaded at 0.5 * 0.5 degree resolution.

The meteorological data is used to calculate climate data, such as the mean annual precipitation and the mean annual temperature.

For a full overview of the data and data sources used in this research, see Table 2.2.

Table 2.2: table: Variables used in this research, the databases the variables are taken from and the type of variability applied to the variables.

Realm	Variable	Database	variability
Climate	mean annual temperature	ECMWF	yearly
	mean annual precipitation sum	EMCWF	yearly
	mean annual evaporation	ECMWF	yearly
	mean annual sunshine hours	ECMWF	yearly
	mean monthly temperature	ECMWF	monthly
	mean monthly precipitation sum	ECMWF	monthly
	mean monthly evaporation	ECMWF	monthly
	mean monthly sunshine hours	ECMWF	monthly
Vegetation structure	maximum SIF of year	SIFTER	yearly
	minimum SIF of year	SIFTER	yearly
	mean annual SIF	SIFTER	yearly
	mean monthly LAI	MODIS	monthly
	maximum LAI of year	MODIS	yearly
	minimum LAI of year	MODIS	yearly
	mean monthly fAPAR	MODIS	monthly
	maximum fAPAR of year	MODIS	yearly
	minimum fAPAR of year	MODIS	yearly
	land cover	MODIS	yearly
Meteorology	mean monthly temperature	ECMWF	monthly
	mean monthly precipitation	ECMWF	monthly
	potential incoming solar radiation	ECMWF	monthly
	net radiation	ECMWF	monthly
	seasonality	ECMWF	monthly

Chapter 3

Results

3.1 predicting Local GPP

This section describes how well the models, trained with FLUXNET data, can predict GPP of data-points not included in the training of the models. This is done in 5 different ways. First, a summary of the results is shown. Then, the results and of the five different methods are explained in more detail. First, the predictive quality of the model is assessed for data that is split randomly into training and testing data. Secondly, the quality of the model is assessed for a model that is trained with data from every percentile of GPP. Thirdly, the quality of the model is assessed when predicting GPP at a station that has not been used as training data, as well as years that have not been used in training the model. Finally, a 5 year GPP prediction is made, based on 3 years of FLUXNET data. For all simulations, the quality of the model is also assessed when SIF is excluded from the data. For these assessments, the site-months used are the same for the model with and without SIF.

3.1.1 Summary

The model with SIF performs better than the model without SIF. This is most notably seen in the lower standard deviation for the percentage bias for the models with SIF in Table 3.1. Based on model efficiency however, the models without SIF perform better (Figure 3.1). Most notably, the models predicting independent sites with SIF included perform worse than the models without SIF. For the other models, leaving out SIF does barely change the model efficiency. Note however, that ME alone is not a good indicator of a good model.

For the simulations with a RTTS, which are used as reference, the models with and without SIF perform similar (Table 3.1). The predictive quality of the models was not increased by using GPP values of every percentile of GPP, as the standard deviation of the bias for these simulations is very high. This indicates that individual models tend to have a large bias. Table 3.1 also shows that the models with SIF that estimate GPP of years not included in the training (I year in Table 3.1) predict the GPP in these years with about the same quality as the RTTS. However, leaving out SIF reduces the model efficiency and increases the bias. It is clear as well that when trained on independent sites, the models perform worse than the independent year simulations, with a model efficiency of about 0.7 and a higher NRMSE. Here again, the models without SIF tend to have a large bias (both positive and negative, averaging out to about 0), which is not found in the model with SIF. The quality of the independent site predictions is not increased for a region with a higher FLUXNET measurement tower density; Europe. Moreover, the predictive quality even decreased with a higher NRMSE and a lower ME and a very high spread in bias, the latter predominantly for the simulations without SIF. A five year prediction, incorporating both the independent site and independent year simulations, did not result in a much worse predictive quality than the respective simulations. However, as this simulation as only been done once, the statistics on this simulation are uncertain.

The models without SIF use on average less variables than models with SIF, with the exception of the PGGP simulation. Although the reason for this is unclear, it might be caused by confounding variables. Another possibility is that the inclusion of SIF makes it possible for the model to resolve more complex relations, that increase model quality slightly. Because of these complex relations, more variables are used in the model. In the PGGP simulation, SIF is very important as explanatory variable. This is assessed in Section 3.3.2

Based on the above, the global GPP prediction is expected to have a predictive quality similar to the 5 year prediction. This is because there does not seem to be a decrease in model quality for a unknown site and year, compared to a independent site.

3.1.2 Random train/test-split

This section describes the results of models when the FLUXNET data is split randomly into training (90%) and testing (10%) data (RTTS). As the split is random, every model build is different, resulting in varying model efficiencies.

Table 3.1: Statistics of 100 models trained with a random train-test-split (RTTS), using percentiles of GPP to ensure the model to know every range of GPP values (PGPP), with leaving one year out and predicting that year (I year) and with leaving 10% of the sites out and predicting these sites (I site). The latter simulation is also done for only sites in Europe (I site Eur). A 5 year prediction has been done as well. All simulations are done 100 times and done both with and without SIF. Note that I year is not done 100 times, as there are only 8 years available in the combined FLUXNET and SIFTER data, the I site Eur has only been done 50 times and the 5 year prediction only once.

		NRMSE		ME		Norm. # var		percentage bias	
		mean	stdev	mean	stdev	mean	stdev	mean	stdev
RTTS	SIF	0.424	0.019	0.80	0.018	0.68	0.16	0.29	1.72
	No SIF	0.423	0.019	0.802	0.017	0.57	0.15	0.28	1.7
PGPP	SIF	0.47	0.002	0.76	0.0018	0.67	0.017	7.17	0.18
	No SIF	0.43	0.003	0.79	0.003	0.81	0.019	12.8	0.43
I year	SIF	0.47	0.027	0.80	0.026	0.76	0.15	0.15	1.15
	No SIF	0.53	0.028	0.73	0.037	0.70	0.165	-3.73	3.1
I site	SIF	0.52	0.068	0.66	0.10	0.69	0.18	0.32	0.76
	No SIF	0.563	0.073	0.69	0.076	0.72	0.18	1.32	8.87
I site Eur	SIF	0.52	0.09	0.527	0.387	0.66	0.22	3.11	13.65
	No SIF	0.536	0.13	0.66	0.16	0.64	0.23	2.14	14.65
5 year prediction	SIF	0.59		0.63		0.33		3.7	
	No SIF	0.59		0.63		0.38		3.3	

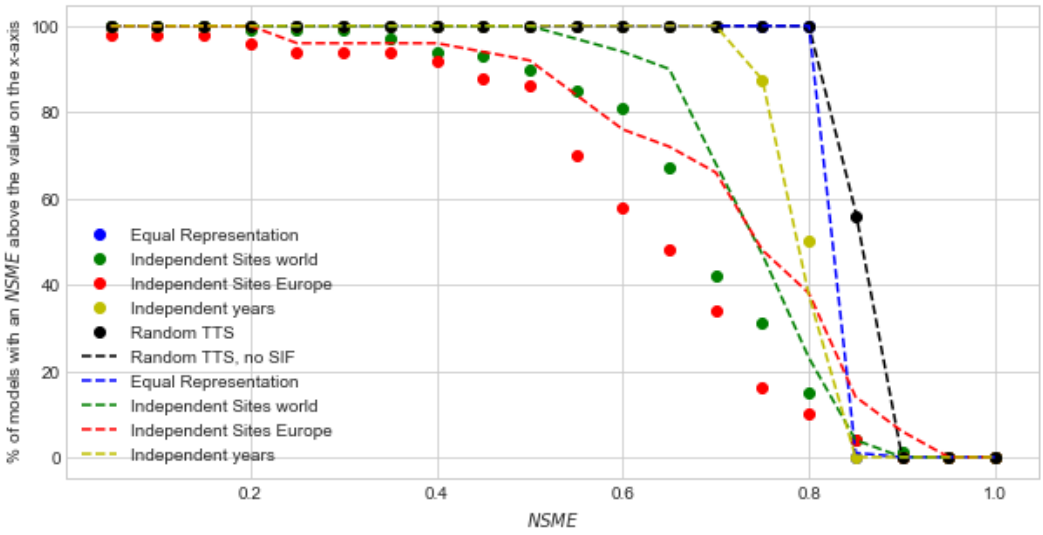

Figure 3.1: The model efficiency of the models used in this section. The dots indicate models with SIF, the dashed lines indicate models without SIF.

Therefore, 100 models are trained with random training and testing data and the model statistics of the 100 models are averaged. See Table 3.1 for these statistics.

Using a random train-test split, the mean model efficiency for both a model with and without SIF is about 0.8 (Table 3.1). Figure 3.2 shows a scatterplot of the measured GPP values at the FLUXNET sites versus the predicted values of one of the models. The red dashed line shows the 1 to 1 line, representing a perfect prediction. In the bottom pane, the residuals of this model are shown. The slope of the regression of the residuals is -0.2 and the intercept is 0.75 (Fig 3.2). This can be interpreted as the model overestimating GPP values of 0 by 0.75 $gC/m^2/d$ on average, but underestimating large GPP fluxes by 20% on average. This could ultimately result in an over-estimation of boreal and temperate GPP fluxes, but an underestimation of tropical GPP fluxes. This bias can be due to the high GPP values being under-represented in the training of the model. This will be assessed in the next section.

Table 3.1 shows that the models with and without SIF perform about the same, with an average NRMSE of 0.42. Besides, for both the models with and without SIF, the average percentage bias over the 100 models is about 0.3,

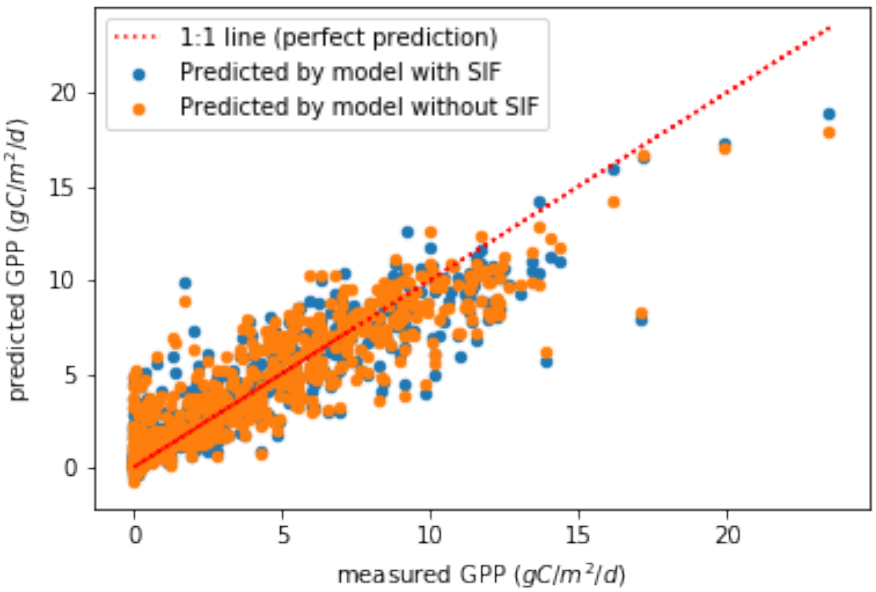

Figure 3.2: Scatterplot of the measured GPP at FLUXNET sites vs. the GPP predicted by models with and without SIF. The red dashed line indicates the 1:1 line.

indicating no real bias in the averaged models. However, the relatively high standard deviation shows that individual models tend to have a quite strong bias.

3.1.3 Train/test split based on GPP values

As stated in the previous section, randomly selecting training and testing data could result in biased estimations. In this section, the predictive quality of the MLP is assessed when the training data is selected based on the GPP value. This is done in two ways: 1) The ratio of high and low GPP values is the same for the training and testing data (Equal ratio); 2) the training data consists of the same amount of high and low GPP values (Equal representation). Due to the low amount of training and the high amount of testing data, the model statistics of the Equal representation are biased. Therefore, only the model statistics of the Equal ratio simulation are shown in Table 3.1.

Equal ratio The Equal ratio models perform about the same as the models with a random train-test-split. The model efficiency (0.76) for the model with SIF is slightly lower than the RTTS models with SIF (0.80) (Table 3.1).

Both the models with and without SIF tend to over-estimate the GPP. The model with SIF simulates GPP values to be about 7% too high, the model without SIF over-estimates GPP values by even more than 12% (Table 3.1). The low standard deviation shows that these over-estimations are systematic.

The aforementioned over-estimation of the GPP occurs only at low GPP values, smaller than 1 $gC/m^2/day$ (left panel of Figure 3.3). For the higher GPP values, the models under-estimate GPP. Both the models with and without SIF show a predicted GPP value of larger than 0, although the measured GPP is equal to 0. This is a side-effect of using percentiles of GPP, as not only the high values are taken into account in the training of the model, but the low values as well. This resulted in about 19% of the GPP values of the training and testing data being 0.

For the Equal ratio simulations, models trained without SIF use more variables (81%) than the models without SIF in the RTTS simulations (67%). Besides, the models without SIF use more variables than the models with SIF for this simulation (Table 3.1). This is contrary to the other simulations, where the simulations with SIF generally use more variables. Although the reason for this is not certain, it might be because for high GPP values, which are certainly represented in both the train and the test dataset, SIF is a very important explanatory variable. This is assessed in Section 3.3.2.

Equal representation For the simulations with an equal representation, the NRMSE is 0.78 and the ME is only 0.36 for a model with SIF. For a model without SIF, the NRMSE is 0.83 and the ME 0.38. Note that the model

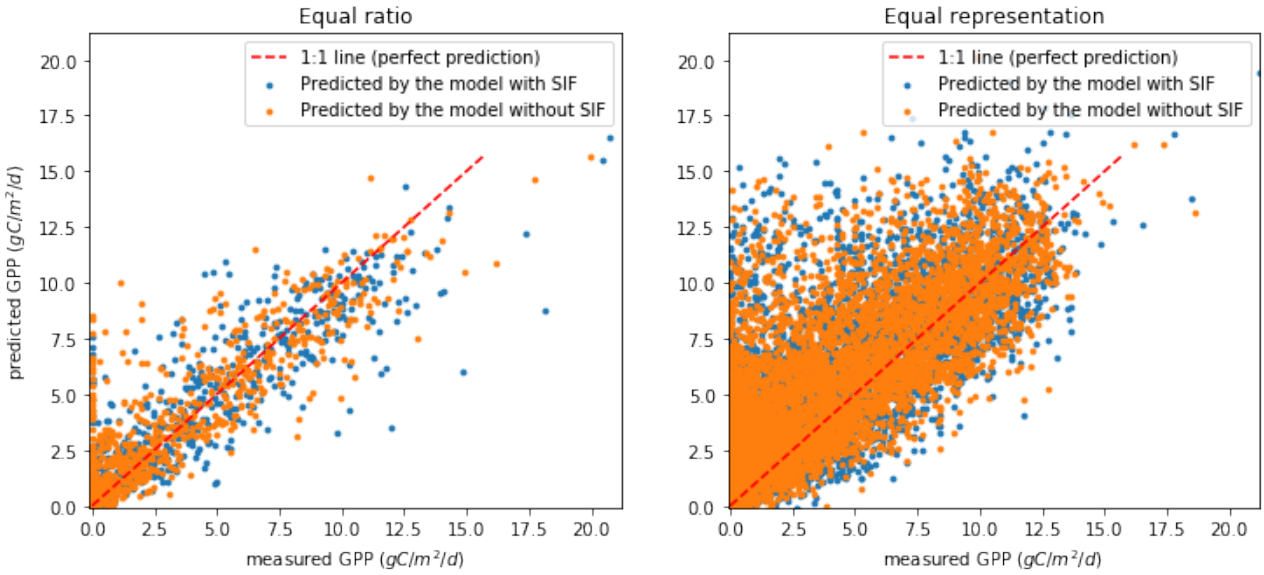

Figure 3.3: Left: Scatterplot of the measured and predicted GPP for the equal ratio simulation. The red dashed line indicates the 1:1 line. Right: Scatterplot of the measured and predicted GPP for the equal representation simulation. The red dashed line indicates the 1:1 line. For both panels, the blue dots indicate a model with SIF, the orange dots indicate a model without SIF.

is trained with only 8 percent of the data, which decreases the model quality. For this reason, the model statistics of the equal representation are not included in Table 3.1. Both the models with and without SIF have a percentage bias of about 40%, which is also because of the large amount of testing data with low GPP values.

The high NRMSE of the Equal representation is shown in the right panel in Figure 3.3 in the wide spread around the 1:1 line. For the Equal ratio (left panel), this spread is smaller.

3.1.4 Predicting an independent year

The used FLUXNET 2015 dataset contains data up to 2015. However, also the global GPP of 2016 will be predicted in this research. Therefore, the predictive quality of the models is assessed when independent years (I year) are predicted by the model. This is done by leaving one year out of the training dataset and using data from that year as testing data. This way, also the capacity of the models to fit the seasonal cycle of GPP is assessed.

The models predicting independent years with SIF perform on average almost the same as the RTTS (Table 3.1), albeit with a slightly higher NRMSE than the RTTS. The models predicting an independent year do require more variables however. The percentage bias is even lower for the I year simulations than for the RTTS. This might be due to the limited amount of models trained, resulting in skewed model statistics.

For the model without SIF however, predicting an independent year results in a model effiency that is about 10% lower than the RTTS. The NRMSE is about 20% higher for this simulation. Besides, the model without SIF estimates GPP values to be lower than the measured values, which is indicated by the negative percentage bias in Table 3.1.

Note that this simulation is done 8 times, as there are only 8 years available in the combined FLUXNET and SIFTER data. No clear trend in model performance was discovered over the years (not shown). However, analysis has shown that the predictions made by the model, both with and without SIF, are best when the seasonal cycles of the sites are well defined and unperturbed, and the inter-annual variability is low.

From the scatterplot in Figure 3.4, the same pattern emerges as from the RTTS simulations, where low GPP values were over-estimated and high GPP values were under-estimated by the model, for both a model with and without SIF. The higher NRMSE, as shown in Table 3.1 can be seen in the larger spread of the points around the 1:1 line. The model has trouble capturing the seasonality on the site shown in Figure 3.4. This is shown in the offset in predicted GPP between month 20 and 25. This offset results in a high RMSE and a low ME. The site-specific model efficiency is 0.57, which is lower than the average model efficiency. This indicates that for most sites, the GPP is resolved better than for this site.

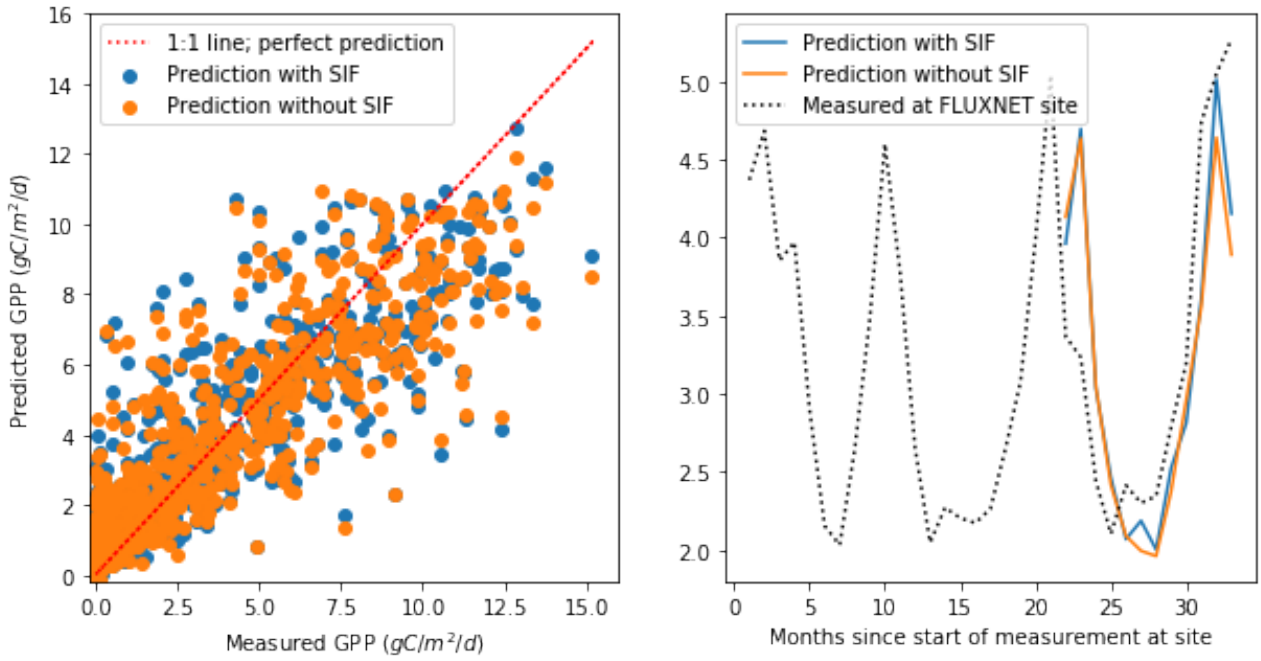

Figure 3.4: Left: Scatterplot of the predicted and measured GPP at FLUXNET sites for the year 2014. The blue dots indicate the prediction by the model with SIF, the orange dots the predictions by the model without SIF. The red dotted line indicates the 1:1 line, representing a perfect prediction. Right: The measured GPP at the AU-Gin site in Australia with land-use type 'Woody savannah'. The black dotted line indicates the measured timeseries of GPP. The blue line indicates the GPP prediction by a model with SIF, the orange line the prediction by a model without SIF.

3.1.5 Predicting an independent site

Because the penultimate goal of this research is to predict GPP at sites where no FLUXNET measurements are available, the model will be tested on randomly chosen sites that are not taken into account in the training of the model (I site).

Predicting independent sites results in a lower model efficiency (0.66 and 0.69) compared to the RTTS for models with and without SIF, respectively. The NRMSE for the models with SIF is about 10% lower than the NRMSE in the models without SIF. The most striking, however, is that the models without SIF have a standard deviation of the percentage bias of about 9%, indicating that individual models tend to have a very strong bias. The average percentage bias is only 1.3%, indicating that the positive and negative biases average out over the 100 models. Due to this high percentage bias, it can be concluded that the models with SIF perform better than the models without SIF, albeit the 3% lower model efficiency.

Although the time series shown in Figure 3.5 clearly shows that the predictions follow the observed trend, the highest GPP values are not well resolved, both by the model with and without SIF.

Figure 3.5 also shows that the model has trouble capturing inter-annual variability, as the measured GPP has more variability than the observed GPP, most notably in summer, when GPP values are high. This also results in a low ME.

To assess the predictive quality of the models for an area with a higher density of FLUXNET towers, also models are trained with as input FLUXNET data in Europe (I site Eur). These models were then used to predict GPP at European FLUXNET sites not taken into account in the training of the model. The model efficiency for the models with SIF is 0.53 for the prediction of European sites, but for models without SIF this model efficiency is 0.66. The low model efficiency can be contributed to the high standard deviation of the percentage biases, which is 13.65 and 14.65 % for models with and without SIF respectively. The average percentage bias is 3.1 and 2.1 respectively, indicating that the extreme values are averaged out (Table 3.1). An explanation for the low model efficiency and high standard deviation of bias is that in Europe many different towers are placed over a large variety of landscapes and land-uses, to be able to gain insight in unique ecosystems which are not always representative for their respective 0.5 by 0.5 degree grid-cell. Because of this, generalisation of the driving factors is harder, resulting in a more erroneous model.

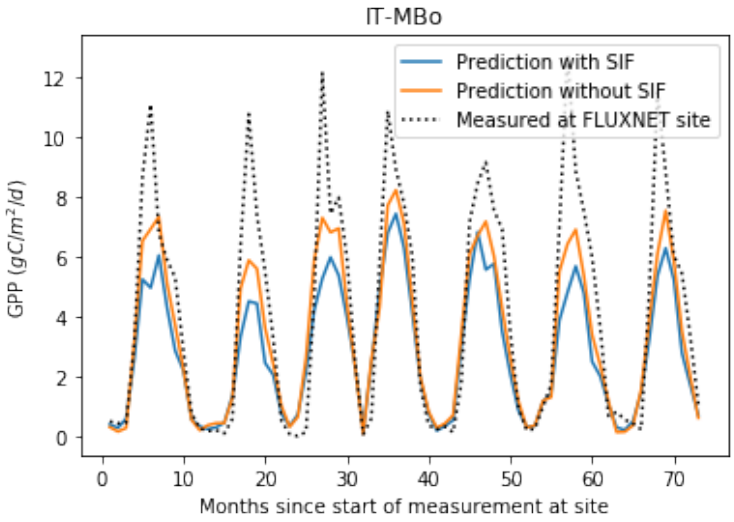

Figure 3.5: Time series of measured GPP (black dotted line), predicted GPP with SIF(blue line) and predicted GPP without SIF (orange line) for the FLUXNET site MBo in Italy with land-use type 'Grassland'

3.1.6 5 year prediction

Extending on the previous simulations, the models have been trained to predict the years 2010, 2011, 2012, 2013 and 2014, based on the data of 2007, 2008 and 2009. By doing so, the FLUXNET sites where the measurements started later than 2009 are not taken into account in the training data. Because of this, about 25% of the sites is also excluded from the training data. Therefore, this simulation can be seen as a combination of the previous two sections. The five year prediction using SIF resulted in a NRMSE of about 0.6 and a ME of 0.63 for both a model with and without SIF, which is about equal to the 'independent site' simulation. Note that, as this simulation is done only once, no standard deviation is calculated (Table 3.1). For this simulation, only 33% and 38% of the available variables are used for a model with and without SIF respectively. Compared to the other simulations done in this research, this is very little. One reason for this might be to the limited amount of data that is available for the training of the model. Because of the small amount of data, the variance in the data can be explained using less splits and therefore with a lower amount of variables.

Striking is that both the NRMSE and the ME are the same for both a model with and without SIF in this simulation. The main difference in performance is that the model without SIF tends to over-estimate the GPP slightly more than the model with SIF, albeit only 0.4% more. The positive percentage bias in both models may be contributed to by the fact that that the average GPP over the years in the FLUXNET data decreases and that the data is therefore skewed towards higher GPP values. However, also other models tend to over-estimate the GPP (Table 3.1).

Note that the model statistics of the 5 year prediction are only slightly worse than the independent site (I site) simulation. This leads to believe that the main model error is due to unknown sites and is only marginally increased by predicting new months or years.

A scatterplot of the predicted (both for a model with and without SIF) and measured GPP is shown in Figure 3.6. The figure shows the same behaviour as the plots for using a random train/test-split, percentiles of GPP and predicting an independent site (Figures 3.2, 3.3 and 3.5 respectively), where the GPP at low values is over-estimated, but at higher values the GPP is under-estimated. As the model is created with less training data, more validation data is left over. Visually, this results in a larger spread around the perfect prediction than for the other figures in this section.

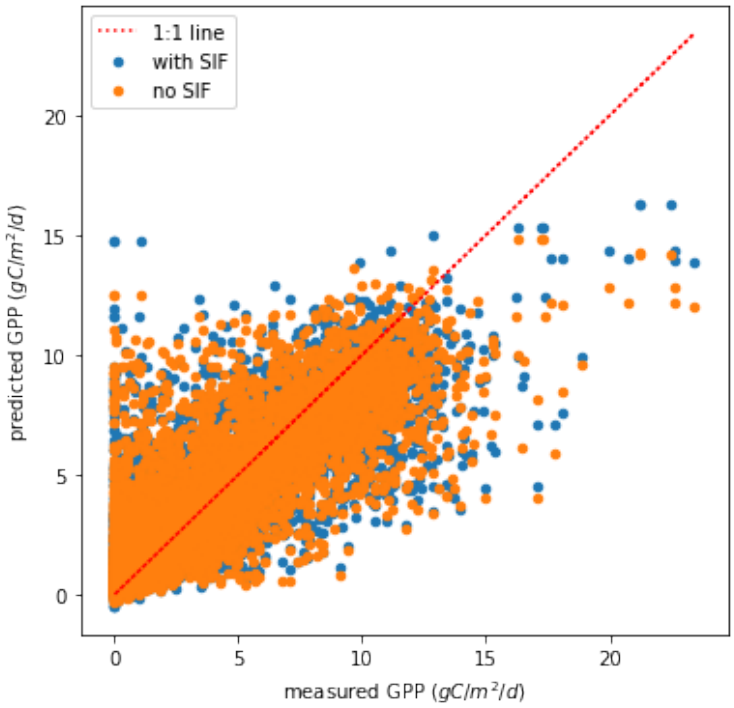

Figure 3.6: Scatterplot of the measured an predicted GPP, for both a model with and without SIF. The red dashed line indicates the 1 to 1 line, which is a perfect prediction.

3.2 Predicting global GPP

In this section, the results of the global GPP predictions are described and interpreted. First, the product created in this thesis (gGPP) is compared to the FLUXNET data that is not taken into account in the training of the model due to missing data. Secondly, the yearly total GPP is presented, followed by the seasonal cycle as simulated by the models. Finally, the inter-annual variability is presented. In this section, the GPP product created in this thesis is referred to as $gGPP_{SIF}$ and $gGPP_{no\ SIF}$ for models that respectively included and excluded SIF. If both products are indicated, this is referred to as the gGPP products.

3.2.1 Comparing to FLUXNET data

The $gGPP_{SIF}$ and $gGPP_{no\ SIF}$ both correlate well with the measured GPP at FLUXNET sites (Figure 3.7). The figure also shows the prediction by Bodesheim et al. (2018), which is the current state-of-the-art global GPP product (bodesheimGPP). bodesheimGPP has a higher NRMSE and ME than both the $gGPP_{SIF}$ and $gGPP_{no\ SIF}$ (Table 3.2). The higher NRMSE of bodesheimGPP is also shown in the figure in the larger spread around the 1:1 line.

Table 3.2: Model statistics for the global models with and without SIF, compared to the GPP product by Bodesheim et al. (2018)

	SIF	No SIF	Bodesheim
NRMSE	0.76	0.76	0.9
ME	0.67	0.67	0.53

It is noteworthy that the model statistics for the $gGPP_{SIF}$ and $gGPP_{no\ SIF}$ are the same. This indicates that the models perform the same, and including SIF does not increase the model quality.

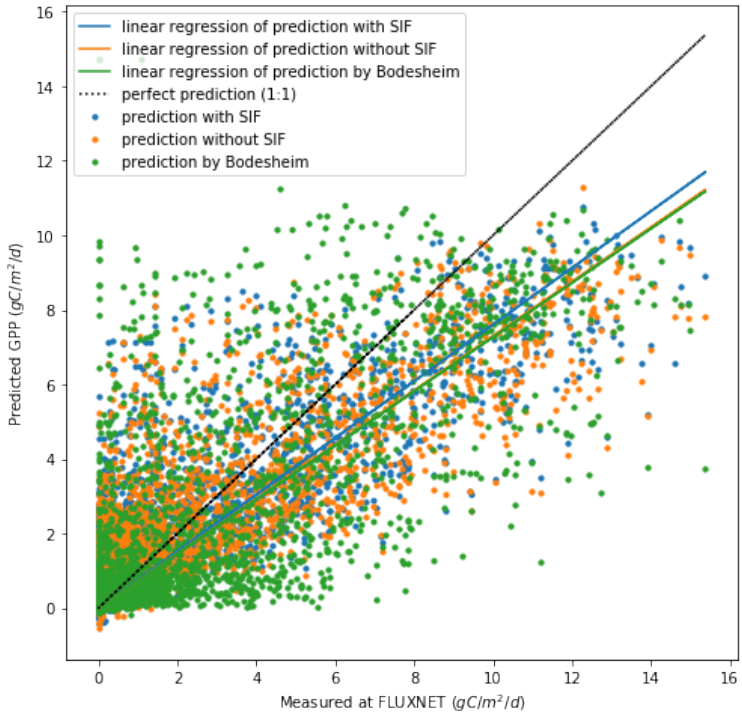

Figure 3.7: Scatterplot of the GPP measured at FLUXNET sites, predicted by Bodesheim et al. (2018) and by the gGPP product, both with and without SIF. The black line indicates the 1:1 line, representing a perfect prediction. The lines indicate the linear regression through the respective points.

3.2.2 Yearly total GPP

The yearly total GPP, averaged over the years 2007 to 2016, as predicted by the gGPP with SIF included is about 131 $\pm$ 1 PgC/year, and without SIF it is about 125 $\pm$ 1 PgC/year. Note that the both predicted GPP values correspond to current estimates of average yearly global GPP, as shown in Table 1.1. Also note that the standard deviation is calculated based on inter-annual variability, and not on the basis of multiple models as this would increase computational complexity too much.

It appears that the key difference between the gGPP products and the bodesheimGPP is that both gGPP products show a higher GPP in the northern mid-latitudes compared to Bodesheim et al. (2018) and lower values in the tropics (Figure 3.8). Note that the FLUXNET data coverage in the tropics is sparse and that these differences balance each other out, resulting in an average yearly total global GPP of about 130 PgC/year. The over-estimation of low GPP values was also found in the local GPP predictions in Section 3.1.

However, for the global prediction, the over-estimation of high-latitude GPP is also caused by the methods used in this research. The SIF dataset has large data-gaps for high-latitude winters. This is also shown in Figure 1.1. In the training of the models, the site-months where SIF is missing are removed, because the algorithm can not cope with missing data. Once trained however, the model can cope with missing data. It does so by always imposing that the missing data is smaller than the split value. Although this is very robust, the model is predicting values that are not taken into account in the training of the data. Because the site-months with missing data are also removed from the validation dataset for the local GPP predictions, this effect is only shown in the global GPP predictions.

Besides, the $gGPP_{SIF}$ predicts the GPP to be higher in the tropics than the $gGPP_{no\ SIF}$. Because the FLUXNET coverage in the tropics is sparse, this difference does not result in different model statistics (Table 3.2).

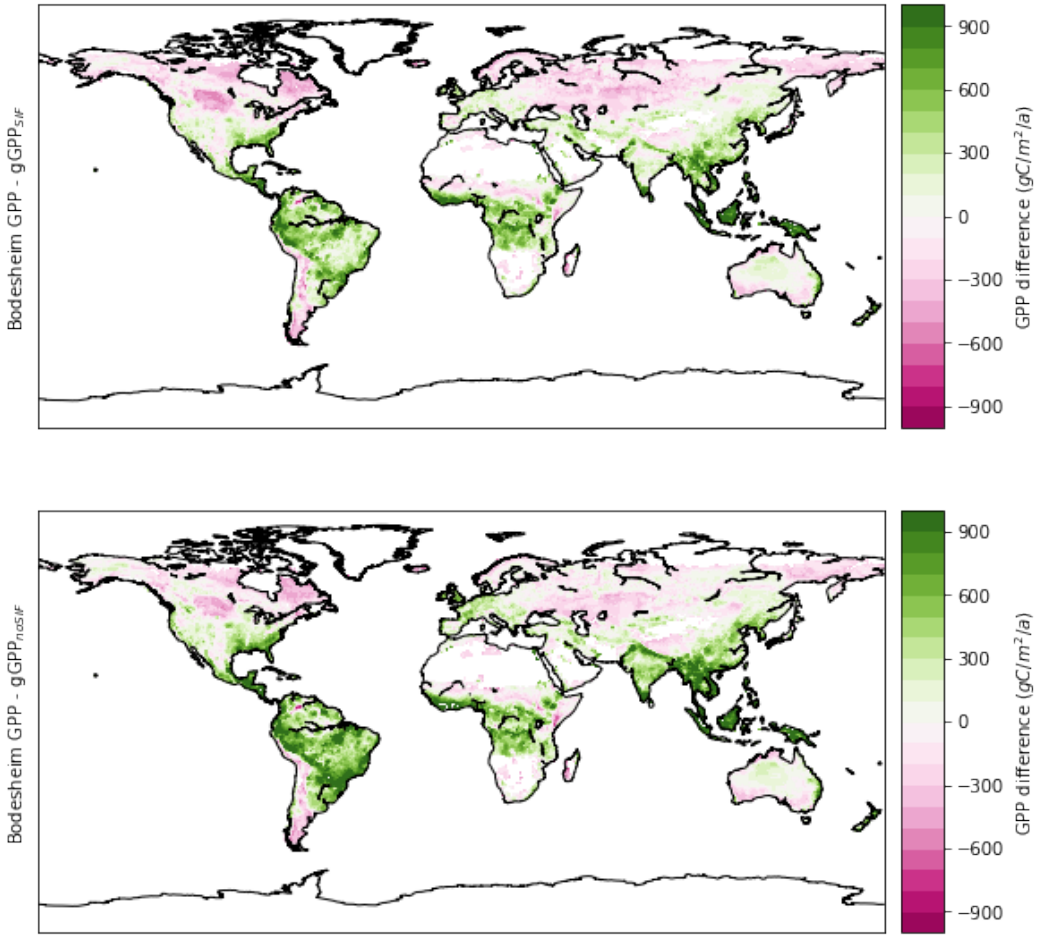

Figure 3.8: The difference between the product by Bodesheim et al. (2018) and the GPP as predicted in this thesis, calculated as the average yearly total GPP according to Bodesheim et al. (2018) minus the GPP calculated in this thesis, with and without SIF respectively

3.2.3 Seasonal cycle

Previous research has suggested that temperate ecosystems in summer have a NEE and GPP similar to tropical ecosystems (Huston and Wolverton, 2009; Jung et al., 2011). The predicted GPP shows a very strong seasonal cycle, also showing an estimated GPP of temperate ecosystems that is similar to that of tropical ecosystems. Both the simulation with and without SIF simulate such a strong seasonal cycle.

The simulated bodesheimGPP values in the tropics are about 10% larger than the GPP values according to the gGPP products (Figure 3.9). Besides, the product by Bodesheim et al. (2018) shows less GPP in the mid to high latitudes in the winter, whereas the gGPP both with and without SIF has a minimum zonal GPP of about 50gC/m^2/month for the mid to high latitudes. No GPP takes place in higher latitude winters, as plants do not photosynthesise. Therefore, the prediction by Bodesheim et al. (2018) seems more reasonable. This difference also explains Figure 3.8 more clearly, as the relative high GPP in the winter results in an over-estimation of total GPP in the mid-latitudes.

For the southern mid-latitudes, the GPP predicted by Bodesheim et al. (2018) shows very high GPP values larger than 250gC/m^2/month around 50°S in the southern hemisphere summer. This value is higher than the predicted GPP in the tropics (250gC/m^2/month) and higher than the predicted GPP values in northern hemisphere summer (225gC/m^2/month). Note however that the amount of land mass is very small at this latitude and that the zonal mean therefore depends stronly on a few datapoints.

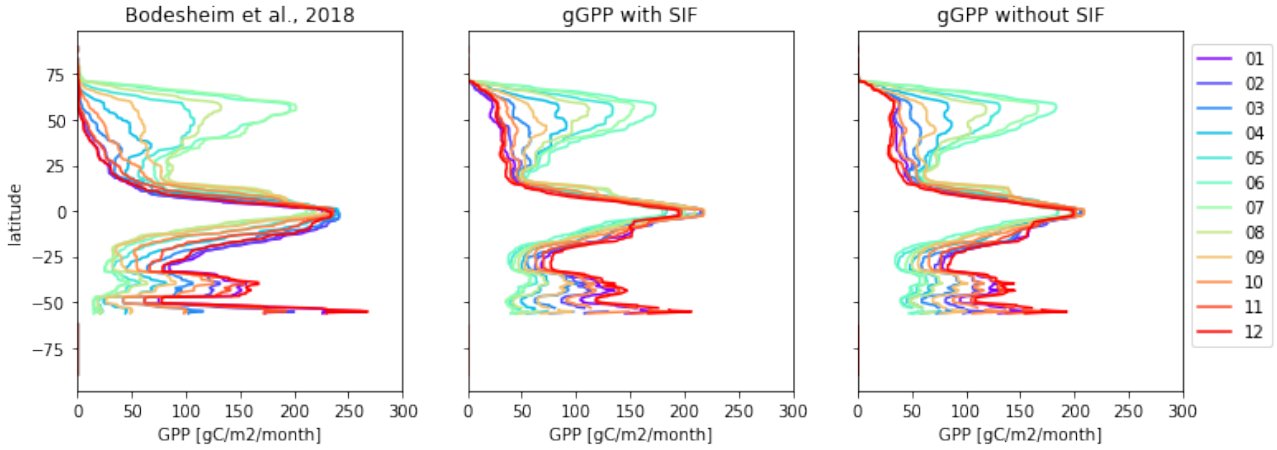

Figure 3.9: GPP over the latitudes. The colour indicates the month, with January being month 01 and December month 12. Left: according to Bodesheim et al. (2018), middle: gGPP with SIF, right: gGPP without SIF

3.2.4 Inter-annual variability

For changes in the global atmospheric CO_2 concentrations, the most important regions are the regions where the GPP has a very large inter-annual variability (IAV), as these regions can take up a lot of carbon the one year, but might be a source of CO_2 the next year (Poulter et al., 2014). Following Jung et al. (2011), in order to find the hotspots for IAV, the standard deviation of the yearly GPP was calculated and regions where the standard deviation exceeds the 90th percentile are dubbed hot-spots for IAV.

The main hotspots for IAV are found just outside of the the tropics (e.g. Eastern Australia, Southern Africa and central America) (Figure 3.10). It is noteworthy that the model with SIF predicts regions with higher IAV in the Amazon and in the southern Sahara. This is in line with the findings of Ahlström et al. (2015), who found that the hotspots for IAV are mainly found in semi-arid ecosystems. In these ecosystems, the carbon uptake is strongly influenced by circulation driven variations in precipitation and temperature (e.g. El Niño). The model without SIF supports this statement less than the model with SIF.

The main difference between these hotspots and the hotspots found by Jung et al. (2011) are that the IAV in South-America was found to be smaller than the IAV found in this research. The same holds for the IAV that was found for Australia. Note that these regions have a sparse FLUXNET coverage and that the model may therefore not be well suited to predict IAV in these regions. Besides, the SIF signal in these regions is subject to high uncertainty,

due to the presence of the Southern Atlantic Anomaly (SAA) (Arida, 2002). The SAA is an anomaly in Earth's magnetic field, resulting in distorted satellite retrievals.

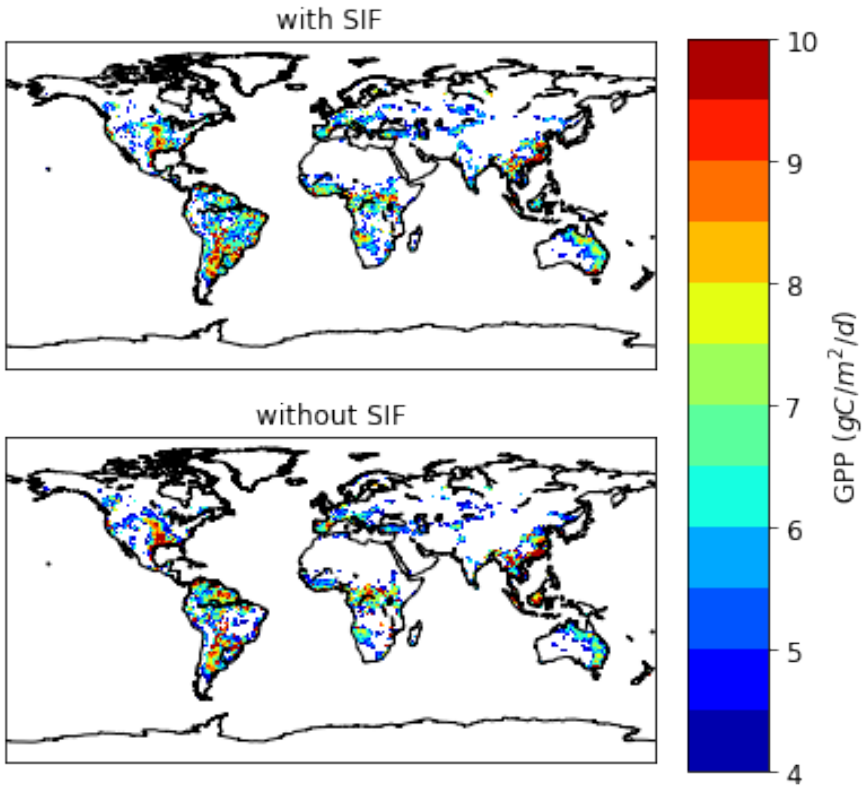

Figure 3.10: Hotspots for interannual variability in GPP, calculated as regions where the standard deviation of the GPP is above the 90th percentile.

3.3 Important variables in estimating GPP

In order to assess the most important variables in predicting GPP, the feature importance of the variables in the models is assessed. First, to find the best set of variables to predict GPP, the AIC is used. Secondly, a model is created using the three most important variables. In order to assess differences between GPP values, a distinction is made between high, middle and low GPP values.

3.3.1 Variables based on AIC

Minimising the Akaike Information Criterion (AIC) is a method of selecting the model that uses the least amount of variables, whilst having a high predictive quality. Because the train and test data vary throughout different simulations, the AIC is calculated for 100 simulations. The variables that resulted in the lowest AIC are deemed to be the most important.

The three most important variables in a model with SIF are Land Use, evaporation and SIF (Figure 3.11) These variables are used in all the models. For a model without SIF, the temperature replaces SIF as important variable, although temperature is not used in all the models. Seasonality and incoming potential shortwave radiation are more important when SIF is not included. However, most other variables are used more often when SIF is included. This indicates that the variance in GPP that is not explained by SIF can be explained by other variables, but also that there is no single variable that explains the incapability of SIF to capture variance in GPP. Besides, this implies that including SIF provides the opportunity for the model to resolve more complex relations between variables and GPP. From Section 3.1, it can be drawn that these relations only increase model quality slightly.

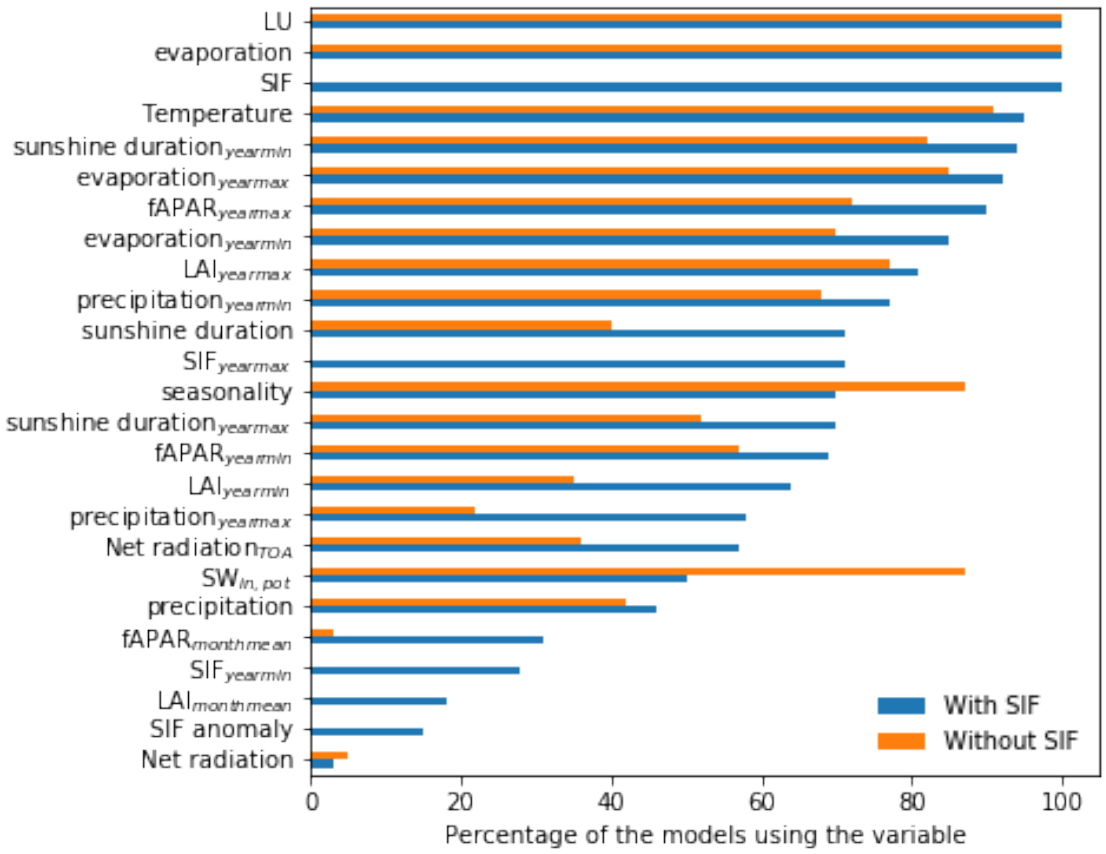

Figure 3.11: The percentage of the models using the respective features. The blue bars indicate models with SIF, the orange bars indicate models without SIF.

3.3.2 Three most important variables

For both a model with and without SIF, the most important variable (based on the gain) for estimating GPP is evaporation. Table 3.3 shows the model efficiency and NRMSE for both a model with and without SIF using only the three most important variables. These variables are listed as well.

Table 3.3: Model statistics and most important variables for a model with and without SIF. The variables are in order of descending gain

	SIF	No SIF
ME	0.61	0.68
NRMSE	0.64	0.61
Variables	Evaporation, Land Use, SIF	Land Use, Evaporation, Temperature

The table shows that the model with SIF has a lower model efficiency and a higher NRMSE than the model without SIF. This indicates that the model without SIF performs better than the model with SIF. Nevertheless, SIF has a higher gain than the potential incoming shortwave radiation, and is therefore not dropped. The evaporation is more important than land use when SIF is included.

3.3.3 Most important variables for different ranges of GPP

To distinguish the most important variables for different values of GPP, the GPP values are categorised as 'high' (the upper 33%), 'middle' (the middle 33%) or 'low' (the bottom 33%) and recursive feature elimination is done for these three categories. The most important variables are shown in Table 3.4

Table 3.4: Most important variables for different categories of GPP, including the average percentage of variables used in training the models and the model efficiency (ME), when only the three most important variables are used.

	SIF			No SIF		
	% of variables used	Top 3 vars (RFE)	ME	% of variables used	Top 3 vars	ME
High	73.3	Yearly maximum sunshine duration, SIF, Yearly maximum SIF	0.45	68.9	Yearly maximum sunshine duration, Air temperature, Yearly minimum LAI	0.45
Mid	70.0	Net radiation,Air temperature, Seasonality $SW_{pot,\ in}$	0.32	72.0	Net radiation,Air temperature, Seasonality $SW_{pot,\ in}$	0.34
Low	68.7	Net radiation,Air temperature, $SW_{pot,\ in}$	-0.07	62.9	Net radiation,Air temperature, $SW_{pot,\ in}$	-0.09

The table shows that for high GPP values, SIF is very important. Both monthly mean SIF (SIF in the table) and the yearly maximum SIF are amongst the 3 most important variables. For lower GPP values however, SIF is not of importance. The ME is low for all three categories of GPP, due to the limited amount of training data. The model efficiency for the lowest 33% of the GPP values is negative, indicating that the predictions are worse than predicting the average. This is mostly due to an over-estimation of the GPP at low GPP values, as also shown in section 3.1.

Chapter 4

Discussion and recommendations

In this research, data-based models were used to predict both local and global GPP. The results of this were presented in the previous chapters. First, the local GPP was estimated at FLUXNET sites, where GPP is derived from observed NEE. These results where used to make a model to predict global GPP, based on ECMWF meteorological data and satellite observed sun induced fluorescence (SIF). In the analysis of the results, a focus was put on the role of SIF in improving GPP estimates. Although the results were generally similar to observations and more complex data-based models, there is room for improvement. This section discusses first the methods and data used in the previous chapters followed by the results found in these chapters, as well as recommend improvements for potential follow-up research.

4.1 Discussion of the data

4.1.1 Uncertainty in FLUXNET GPP

The FLUXNET dataset is widely used to estimate, research and validate carbon fluxes (Friend et al., 2007; Jung et al., 2009; Beer et al., 2010; Reichstein et al., 2013; Joiner et al., 2018). The dataset is subject to state-of-the-art gap filling, noise filtering and flux partitioning (Loescher et al., 2006; Papale et al., 2006; Lasslop et al., 2010; Vuichard and Papale, 2015) (see also section 2.2.1). However, this dataset also has some minor drawbacks for this research, of which 3 intrinsic and one extrinsic will be discussed in this section.

Firstly, the FLUXNET GPP is calculated by the algorithm proposed by Lasslop et al. (2010), which is described in section 2.2.1. In this algorithm, no distinction is made between direct and diffuse radiation, although diffuse radiation is found to result in higher photosynthesis rates and therefore higher GPP than direct radiation (Lasslop et al., 2010). As the radiation seems to be an important variable, this is expected to result in a sligth error in GPP predictions. As with the current data no assessment of the difference between direct and diffuse radiation can be made, the sign of this error is uncertain.

Besides, as both GPP and respiration are derived from NEE, compensating errors can occur. This occurs when both GPP and respiration are over or under-estimated. As the FLUXNET data is monthly averaged, these compensating errors are expected to be averaged out. Therefore, it is assumed that no persistent bias is included in the flux partitioning algorithms. Therefore, compensating errors are expected to result in a small error, of which the sign is uncertain as well.

The third intrinsic drawback of the FLUXNET dataset is that all flux partitioning methods are dependent on the filtering of the NEE measurements (Lasslop et al., 2010). This filtering of the NEE measurements has been thoroughly done by the FLUXNET network, albeit based on arbitrary thresholds. These thresholds might influence the resulting GPP and therefore the outcome of the global GPP predictions greatly. The discarding of measurements after periods with a low u*, as proposed by Papale et al. (2006) is expected not to result in a large bias, as these periods usually occur during the night, in which no photosynthesis takes place.

An extrinsic limitation of the FLUXNET dataset for this research is the spatial distribution of the measurement sites. Over 100 measurement sites are located in the temperate regions of Europe and America, but only few sites are located in desert, savanna and tropical biomes. As the main GPP flux is expected and simulated to be in the tropical biomes, mostly the fact that only two FLUXNET towers are located in the Amazon is a large drawback for this study. More measurement sites in the tropics could provide additional insight in the drivers of GPP in the tropics and could therefore provide a more robust GPP prediction in tropical regions. Besides, more extensive input data results in a more robust GPP prediction in the tropics, as the models are better trained to predict higher GPP values. This also could affect local GPP predictions, as models are better trained to predict high GPP values if more measurements of high GPP are included in the training data. This potentially reduces the under-estimation of high GPP values by the models, resulting in better model statistics.

4.1.2 Quality of the SIF data

For this research, the SIF retrieval by Sanders et al. (2016); Schaik (2016) is used. Although this SIF retrieval (SIFTER) has shown to be able to resolve droughts (Schaik, 2016; Koren et al., 2018) better than the more widely used NASA SIF as described in Section 2.2, the SIF product is very noisy and has some large data gaps. Although this could be resolved by e.g. a median filter (Jones et al., 2018) and a poisson grid fill (Meier-Fleischer and Böttinger, 2018) respectively, a lot of the spatial variability would be lost in the process. Therefore, this has not been done in this research.

The satellite that measures SIF, GOME-2, has a footprint of 80 by 40km. In order to validate the SIF observations by the satellite, ground measurements on this scale should be conducted. On this scale however, this is near impossible. Therefore, the SIF retrieval cannot be validated. This results in large uncertainties, mainly over the tropics as tropical air contains more moisture and is more clouded, which both influence the SIF signal. The sign and magnitude of this influence are uncertain due to the lack of validation measurements.

Another source of uncertainty in the SIFTER data is the Southern Atlantic Anomaly (SAA) (Arida, 2002). The SAA is a decrease in the Earth's magnetic field, interacting with satellites. Due to the SAA, the uncertainty of the SIF signal is increased, mostly over South-Eastern Brazil. Again, the magnitude and the sign of the effect on the GPP predictions are uncertain.

The (combined) impacts of these uncertainties in SIF on the results found in this thesis might be quite high. If similar SIF signals correspond to different GPP values, the model cannot resolve GPP correctly, resulting in faulty GPP estimations. The sign of this potential error is uncertain, because the error in SIF signal is uncertain. As the SIF signal has the highest uncertainties over the tropics, this affects mostly the higher values of GPP, which could affect the estimated average global GPP values largely.

For a follow-up research, models could be trained and tested with SIF values within the bounds of uncertainty. This would help to assess the potential error in GPP. Besides, a distinction could be made between different regions, where SIFTER is expected to be of a better quality. In this research, only Europe has been distinguished. Although the quality of the SIFTER data is expected to be good over this region, the results did not improve.

A final remark in the SIFTER data is that the satellite is degrading, resulting in a decreased SIF signal (Zhang et al., 2018). Again, due to the lack of validation data, the magnitude of this is uncertain. However, analysis has shown that the yearly total global SIF signal is constant up unto the end of 2015, leading to believe that the error is negligible up unto 2015. Therefore, this is mainly expected to be an issue in the global GPP prediction for the years 2016 and 2017, resulting in GPP estimations that are too low. Koren et al. (2018) have countered this effect by increasing the SIF signal by 1% per year over the Amazon. However, the satellite degradation is not expected to be the same over the entire earth. Therefore, in a follow-up research, region-specific correction factors for the SIF signal could improve the results of this research.

In stead of using SIF, another option is the promising Near Infrared Refleciton of vegetation (NIRv). NIRv is the satellite-observed reflection of vegetation of near-infrared light, which is found to correlate well with GPP (Badgley et al., 2017, 2018). NIRv can be observed using the MODIS satellite, which has a very high spatial resolution, also tackling some of the issues that are addressed below.

4.1.3 Micro-meteorological and spatial variability

FLUXNET towers are measuring on smaller scales than the 0.5 by 0.5 degree grid. The average state of the grid cell may vary greatly from the measured or observed state at the FLUXNET tower. Because of this, biases may occur in the predicted GPP. This is mainly a limiting factor for variables that are very site-dependent, such as the land-use or the LAI, as the average condition of the grid-box may vary greatly from the site-specific conditions.

This is mainly expected to be a problem in Europe, as research groups set up eddy-covariance towers at a-typical sites, of which the characteristics are not captured by other EC towers. This results in GPP predictions that are worse than GPP predictions for global FLUXNET towers, as is shown by the results from the Independent Site Europe simulation (Table 3.1).

This drawback could be resolved is by calculating the GPP for each land-use class present in the respective grid-box, and calculate the weighed average of the predicted GPP of these land-uses. As the land-use map has a higher resolution (0.05 by 0.05 degree) than the other global data used (0.5 by 0.5), this is a viable option to reduce the uncertainty in GPP on a global scale. Due to temporal constrains, this has not been implemented in this thesis.

4.2 Discussion of the methods

4.2.1 Extrapolating capacity

As gradient boosted trees stratify the known data into small subsets in which one target value is predicted, gradient boosted trees can not extrapolate data outside their training data range. Therefore, the highest possible GPP value that can be predicted is the maximum of the data the model has seen in the training. This may be a problem in the tropics, as higher GPP may occur than the maximum that is measured in the FLUXNET dataset. In future research, this problem can be resolved by either including GPP observations from towers that are located in absolute hotspots for GPP. Another option for resolving this drawback are more elaborate model trees, in which a multiple linear regression is conducted in the leaves of the trees (Jung et al., 2009) or a different machine-learning algorithm, for example neural networks, which are capable of extrapolating. Additionally, a different implementation of gradient boosted trees could be used to reduce the bias. CatBoost (Dorogush et al., 2017) is an example of such an implementation, outperforming XGBoost (which is used in this thesis) if the model is properly tuned for categorical variables, such as land-use (Swalin, 2018).

4.2.2 Sub-optimal hyper-parameters

Great concern has been put in the prevention of over-fitting, whilst also preventing under-fitting. Due to the high computational costs that comes with finding the best hyper parameters, such as maximum tree depth, only a limited number of combinations of hyper-parameters is tested. Therefore, it is expected that the hyper-parameters are still sub-optimal. Although the predictive quality of the models could be slightly increased by optimising the hyper-parameters more, this would cost significantly more computational power and is expected to result in only a slight increase in performance and no significantly different results. Therefore, this is not done in this thesis.

A Bayesian approach to grid-search is used in the second part of this thesis (Section 6.2). Although this approach is expected to increase the model statistics slightly, preliminary analysis has shown that the hyper-parameters according to the Bayesian approach do not differ significantly from the hyper-parameters used in this part of this thesis. Therefore, also the results are not expected to be different.

4.2.3 Handling of missing data

As gradient boosted trees cannot be trained with data that includes missing values, these missing values are removed from the FLUXNET data. In predicting however, missing data is coped with by always assigning the missing data to be smaller than the split value. As the largest data-gaps in the SIF data are in the high-latitude winters, the trained models have not been trained on high-latitude winters, because this data is removed. This could result in faulty GPP predictions in these regions. For a follow-up research that uses SIFTER SIF, it is recommended to create a data-based model that can handle missing data in the input data. Due to temporal constraints, this has not been done in this research.

4.2.4 Feature engineering

A final remark that could improve the model quality in a follow-up research is feature engineering. Feature engineering is not applied in this research because the main goal of this research was to assess the improved quality due to SIF. However, by combining important features for resolving GPP, GPP predictions might be improved. Moreover, a more thorough overview of the potential of the SIFTER dataset could result from feature engineering.

4.3 Discussion of the results

Despite the potential drawbacks in the data and methods as discussed above, the model efficiency of the models for local GPP predictions is high (between 0.6 and 0.8) and the NRMSE is low (between 0.4 and 0.6). Therefore, it can be concluded that both the models with and without SIF used in this research can make a robust prediction of GPP at FLUXNET sites (Table 3.1).

4.3.1 Local GPP predictions

Comparing their results to local GPP observations, Jung et al. (2011); Bodesheim et al. (2018); Joiner et al. (2018) found model efficiencies between 0.6 and 0.8 for their models, using more complex models and more input data (see Section 2.1.2). The models trained in this thesis have about the same model efficiency, indicating that the models used in this thesis can be used as a solid benchmark to predict global GPP fluxes.

Despite the good model statistics, some patterns in the predictions of GPP were observed. First of all, the models over-estimate GPP for GPP observations around 0 $gC/m^2/d$, and under-estimate the values at GPP values higher than 15 $gC/m^2/d$. This pattern holds for all simulations; for both local and global GPP predictions and models with and without SIF. An explanation for this is the over-representation of low GPP values in FLUXNET data, and the under-representation of high GPP values. As there are many measurements with low GPP, the model is biased towards low GPP values. On the other hand, as there are many measurements with low GPP, the chance of some of these values being categorised wrongly increases. This results in both an over-estimation of low GPP values and an under-estimation of high GPP values.

This could be potentially improved by normalising the measured data, for example by log-normalising the GPP. In doing so, it is implicitly implied that the relation between GPP and drivers is not linear at lower GPP values. Although test have been done in this research, these were not extensive. Preliminary results have shown that log-normalising the GPP resulted in the models correctly predicting low GPP values, but under-estimating high GPP values even more. Future research could more extensively test this, also implementing other means of normalising the data, such as normalising using the reciprocal of the GPP. Additionally, the lower values of GPP could be normalised, and the higher values not. Another possibility is to tailor e.g. SIF, for example by exponentiation.

Nevertheless, normalising the GPP or tailoring SIF is not expected to affect the main finding of this research, namely that SIF does not appear to add significant value to the predictive quality of the models. The potential gain by including SIF is compensated for by other variables when SIF is excluded. Nevertheless, the standard deviation of the percentage bias is higher for models that are trained without SIF, indicating that including SIF does reduce the bias of the models.

4.3.2 Global GPP predictions

The predicted yearly total GPP by the models created in this thesis is between 125 - 131 PgC/year. This is very similar to GPP estimates by previous research (Table 1.1). When comparing monthly GPP fluxes, the model appears to have better model statistics than the state-of-the-art data-based model by Bodesheim et al. (2018). However, the models used in this thesis appear to over-estimate GPP in the mid to high-latitude winters, which is conform the previous statements that the model seems to over-predict GPP values at lower measured values. An explanation for this behaviour is that the SIF signal in winter, when GPP is low, shows some very large gaps. Therefore, this data is dropped in the training of the models and the models do not learn the decreased GPP in winter, resulting in an over-estimation of GPP. In a future research, this can be overcome by using a SIF dataset that shows less data gaps.

Similar to the local results, when comparing a global model with SIF to a model without SIF, the differences in model statistics are minor. This contradicts the conclusions by Joiner et al. (2014); Yoshida et al. (2015) that state that data-based models can benefit from adding SIF. However, finding that SIF is a very important variable in the data-based models, this research supports the finding by Joiner et al. (2018), stating that SIF could be used to resolve GPP fluxes, but the quality of the retrieval should be improved.

Two key differences were found between a global GPP product obtained by a model with and without SIF:

1. Adding SIF to the model increases the estimated total global GPP to about 131 ± 1 PgC/year, whilst without SIF this estimate is about 125 ± 1 PgC/year. This difference is mainly due to the simulated increase in GPP in the tropics for a model with SIF (Figure 3.8). Because these predictions are made using only one model with SIF and one without SIF, the standard deviations are calculated based on the predicted GPP over different years. This results in a very small standard deviation. The standard deviation might be increased by training more models. This could also result in better insight in the uncertainty in global total GPP.

2. More regions are found to have a high IAV when SIF is taken into account in training the model. This might be due to the direct relation of SIF to GPP, increasing the IAV. Most notably, the IAV is higher in the south-eastern part of Brazil if SIF is taken into account. However, in this region, the Southern Atlantic Anomaly

interacts with the satellite signal, resulting in a higher uncertainty of the SIF signal in this region. Therefore, the inter-annual variability in this region has a high uncertainty as well.

4.3.3 Feature importance

Although SIF was found to be compensated for when not included in the model trees, it has also been found to be the third most important variable when recursive feature elimination is conducted, after evaporation and land use. This shows that SIF explains a very large part of the variance in GPP. Besides, SIF was found to be very important for high GPP values, as both monthly mean SIF and yearly maximum SIF are amongst the three most important variables. For lower GPP values however, SIF was not found to be amongst the three most important variables. This indicates that SIF explains most of the variance in the higher GPP regions. This could be due to an enhanced SIF signal with higher GPP. It is important to note however that, for low GPP values, a lot of the data is dropped because SIF is too uncertain in winter.

One of the main motivations for using SIF is that its anomalies are found to correlate to anomalies in GPP (Yoshida et al., 2015; Wang et al., 2016; Koren et al., 2018). However, this research found that this correlation does not add sufficient information on the GPP to be of importance in the data-based models used in this research. Note that in this research the GPP is predicted, and not anomalies in GPP. As anomalies in SIF are found to correlate to anomalies in GPP, anomalies in SIF might aid in predicting anomalies in GPP.

Unfortunately, both Bodesheim et al. (2018) and Jung et al. (2011), whose research was used as a benchmark for comparing the results found in this thesis, did not research the feature importance in their research. Therefore, the feature importance found in this research cannot be compared to state-of-the-art research.

Chapter 5

Conclusion

In this research, data-based models predicting gross primary production (GPP) on both local and global scale were created to assess the added value of sun-induced fluorescence (SIF) for the quality of data-based models. The results of this thesis show that although SIF is amongst the most important features for predicting GPP, the added value of SIF is compensated for by other variables if SIF is not included in the training of the models.

This thesis consisted of three objectives, of which the main findings are.

- Data-based models can estimate GPP at FLUXNET sites using SIF, meteorological and vegetation data.

  GPP at FLUXNET sites was estimated using gradient boosted trees (Chen and Guestrin, 2016) for different training and testing data. The models were found to have a model efficiency between 0.8 and 0.6, but consistently over-estimating low GPP values and under-estimating high GPP values. Models trained and tested on data that was randomly split into training and testing data were found to perform best, model that were trained on sites in Europe predicting independent sites in Europe were found to perform the worse

- The monthly mean SIF is found to be an important variable in predicting GPP. However, the anomalies and the yearly minimum SIF are not found to be important. Yearly maximum GPP is found to be important to predict high GPP values.

  Although SIF anomalies have been theorised to correlate very well to anomalies in GPP (Joiner et al., 2014; Koren et al., 2018), SIF anomalies were not found to be important in the models. Also, both yearly maximum and minimum SIF were found to be unimportant in predicting GPP. Although SIF has been theorised to correlate to GPP, evaporation and land use were found to be more important.

  Models that were trained on data including SIF performed slightly better than models without SIF, with lower root mean square errors and lower biases. The model efficiency was generally higher for a model without SIF however.

- The estimated global GPP by a model with SIF does not differ significantly from the predicted GPP by a model without SIF.

  The addition of SIF has mostly been found to barely affect the estimated yearly total GPP, but the simulateed inter-annual variability is affected due to SIF. Yearly total GPP calculated by models that included SIF was found to be about 4% higher than yearly total GPP calculated by models not including SIF. The model with SIF estimates a higher inter-annual variability near the Amazon and Sahara than the model without SIF, which is conform the research by Ahlström et al. (2015).

All in all, data-based models that use SIF as an explaining variable can be used as a solid benchmark for predicting carbon fluxes. Although SIF has been found to improve GPP estimates at local scale, the data quality of the SIF needs to be improved to improve global GPP estimates.

Part II

Improving NEE estimations

Chapter 6

Methods and Data

In this section, the methods used for the second part of this research are elaborated. First, the inverse model CarbonTrackerEurope (CTE) is explained. For this, first, the general principle of CTE is elaborated. Then, the different modules of the CTE model are explained. As the focus of this thesis is on the residuals of the land carbon fluxes, the biosphere module is elaborated upon in detail. Secondly, the set-up of the data-based model is explained. Thirdly, the data used for training the models is elaborated.

6.1 The CTE inverse model

As stated before, data assimilation is a method of combining model output with observations, providing the opportunity to resolve the behaviour of the system better than either the observations or the model on its own (Stuart and Taeb, 2018). For the sake of brevity, the mathematics of the data inversion will not be discussed here (see Peters et al. (2005) and van der Laan-Luijkx et al. (2017) for a description of the set-up of the model used in this part of this thesis).

The total CO_2 flux as calculated by CTE is

$$F(x,y,t) = \lambda * F_{bio}(x,y,t) + \lambda * F_{ocean}(x,y,t) + F_{ff} + F_{fire}(x,y,t) \tag{6.1}$$

Where the λ is a set of scaling factors, calculated by the data-inversion. The λ is used to adjust the prior biosphere and ocean fluxes (F_{bio} and F_{ocean} respectively) to result in simulated CO_2 concentrations that are similar to the measured concentration. For this thesis, posterior flux F (Equation 6.1) is deemed to be the truth.

Equation 6.1 consists of 4 flux components on the right-hand side: The biosphere flux (F_{bio}), the ocean flux (F_{ocean}), the fossil fuel burning flux (F_{ff}) and the fire emissions (F_{fire}). Due to the high quality of the fossil fuel emission and fire emission inventories, only the land biosphere flux and ocean carbon flux are scaled by the calculated λ. As the focus of this thesis is on the NEE and therefore on the land biosphere flux, only the biosphere module is elaborated upon.

A schematic overview of the data pipeline of CTE is shown in Figure 6.1.

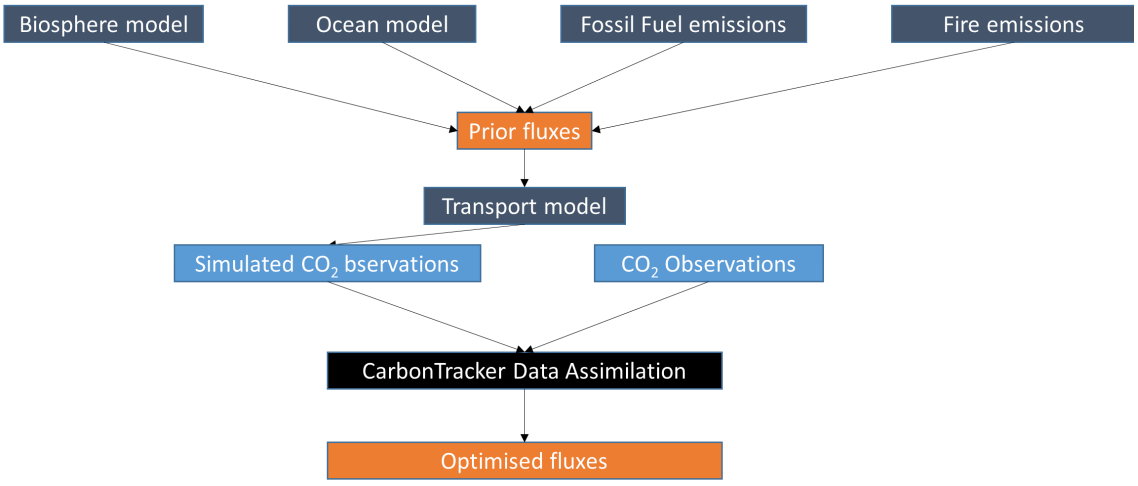

Figure 6.1: Schematic overview of the CTE processing pipeline. The dark blue boxes indicate models and inventories. The orange boxes indicate calculated fluxes. The light blue box indicates observations and the black box indicates the data-inversion.

6.1.1 Biosphere model

The process-based model that provides the prior flux estimates for CTE is SiBCASA, which is a combination of the Simple Biosphere model (SiB) and Carnegie-Ames-Stanford Approach (CASA) (Schaefer et al., 2008). SiB is a biosphere model that simulates the transfer of energy, carbon and momentum over timescales in the order of 10 minutes (Sellers et al., 1986, 1996a,b). On the other hand, CASA simulates the carbon fluxes through various pools, such as above and belowground biomass on daily to monthly timescales (Potter et al., 1993).

In order to calculate the total photosynthesis of a grid cell, the calculated leaf photosynthesis is scaled up to ecosystem level using the fraction of absorbed photosynthetic active radiation, which is derived from the normalised vegetation index (NDVI) (Sellers et al., 1996a,b). The upscaled photosynthesis is constrained by satellite-observed LAI. Besides, the model differentiates between C_3 and C_4 plants. The distribution of these plant types is based on the research by Still et al. (2003) and is assumed to be time-independent. An overview of both the in and output variables of SiBCASA is provided in Appendix 12.2.

In the biosphere model, also a fire module is included. This module is based on satellite-observed wildfires, releasing CO_2 into the atmosphere (Van der Velde et al., 2014). The carbon emitted in biomass burning is calculated based on remote sensing observations (mostly by the MODIS satellite), which are combined in the Global Fire Emissions Database (Giglio et al., 2013).

6.1.2 Transport model

Carbon is a long-lived gas. Due to motion in the atmosphere, carbon is constantly being distributed over the earth. Because of this, CO_2 that is emitted on one location can be measured at another location. As this can have a huge impact on the simulated carbon fluxes, it is important to resolve the atmospheric motions that transport CO_2. This is done by the TM5 Transport Model (Krol et al., 2005). In the TM5 model, regions for which a higher spatial resolution is required can be studied in more detail by applying a finer grid to those regions. In CTE, the finer grid of 1 by 1 degree is applied to both Europe and North-America. Over the remainder of the globe, a 3 by 2 degree grid is used. In both this research and CTE, the transport of carbon is assumed to be perfectly resolved by TM5 (Peters, 2018).

6.1.3 Observations

In order to scale the prior biosphere fluxes to match measured CO_2 concentrations, observations are needed. In CTE, these observations are provided by 46 institutions worldwide, summing up to 354 observations. These observations are included in the observation package GLOBALVIEWplusv3.2 (Peters et al., 2009; Peters, 2018). In order to minimise errors due to poor forecasting, these observations are filtered, based on random errors and a model-data mismatch. For the full filtering procedure, see the CarbonTracker documentation (Peters, 2018).

In GLOBALVIEWplus, the quality of observations is flagged by the data providers. Only data that are indicated to be suitable for assimilation are used. For most of the quasi-continuous measurements, the CO_2 mole fractions observed during the local afternoon are assimilated, as TM5 has more trouble resolving the stable (nighttime) boundary layer over land. For measurement sites at mountain tops, the observations from night-time are used, because this avoids up-slope winds that advect CO_2 mole fractions influenced by local vegetation or anthropogenic activities.

6.1.4 Calculation of the scaling factor

As shown in Equation 6.1, the scaling factor λ is used to scale prior (biosphere and ocean) fluxes to result in the observed CO_2 mole fractions. For the calculation of the scaling factor, CTE differentiates continental regions and ecoregions. The continental regions are defined by the TransCom definitions (Gurney et al., 2002). The ecoregions are defined as all grid-cells in that share the same Olson region (Olson et al., 2001) within the TransCom region (e.g. European broadleaved forests). The TransCom regions are shown in Figure 6.2.

In previous versions of CTE, one scaling factor was calculated for all grid boxes that share the same ecoregion (van der Laan-Luijkx et al., 2017). However, especially when ecoregions cover large areas and are therefore geographically far apart, ecoregion response to forcings might be quite different, for example due to varying management types throughout the region. Therefore, one scaling factor for an entire ecoregion is sub-optimal. A more realistic approach is to assume that the biosphere fluxes do not correlate in space, and a scaling factor is only applicable to one 1

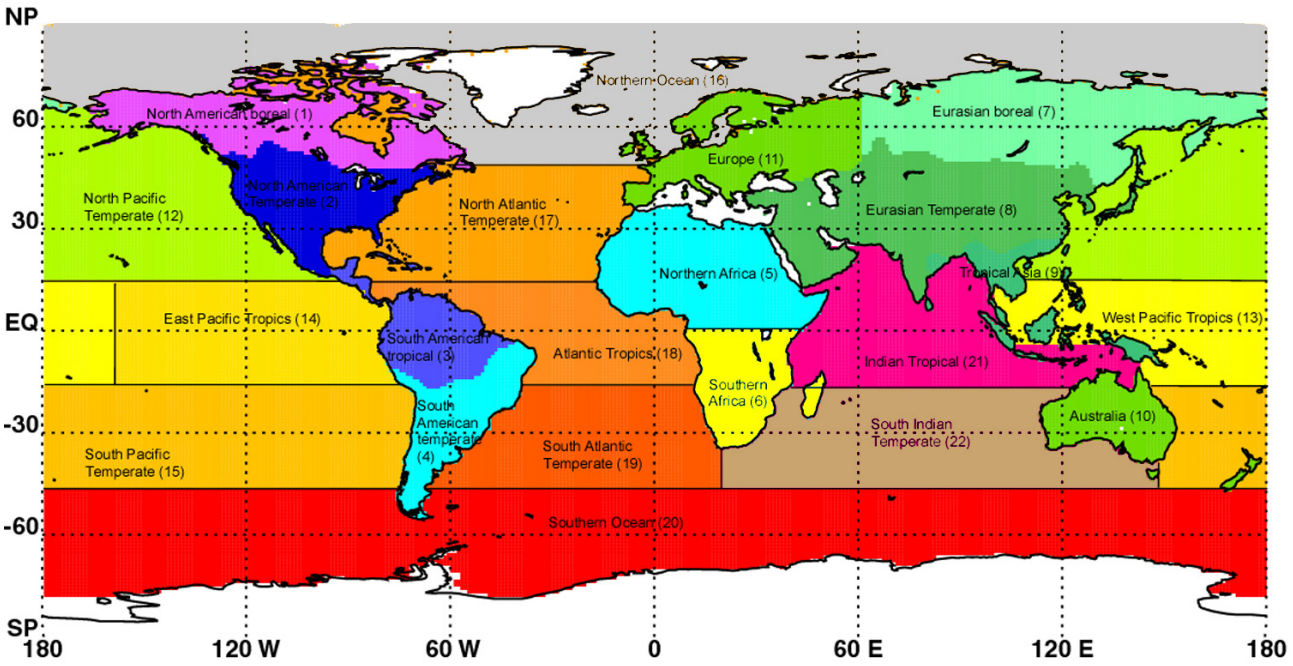

Figure 6.2: The TransCom regions, including the names of the regions. Taken from `http://transcom.project.asu.edu/transcom03_protocol_basisMap.php`.

by 1 degree gridbox. However, the sparse observation density of CO_2 mole fractions does not support this. In the current version of CTE, it is therefore assumed that the correlation of the scaling factor of biosphere fluxes within an ecoregion decreases exponentially over distance. Because of the low observation density in the southern hemisphere, this is only applied in the northern hemisphere (van der Laan-Luijkx et al., 2017). Table 6.1 shows an overview of the used covariances in biosphere fluxes in CTE. In the current CTE version, the scaling factors are calculated for every week.

Table 6.1: Scaling factor correlation (van der Laan-Luijkx et al., 2017) for the 11 land TransCom regions used in CTE (Gurney et al., 2002).

TransCom region	Covariance
North America boreal	within ecoregions
North America temperate	within ecoregions
South America tropical	across ecoregions
South America temperate	across ecoregions
Northern Africa	across ecoregions
Southern Africa	across ecoregions
Eurasia boreal	within ecoregions
Eurasia temperate	within ecoregions
Tropical Asia	across ecoregions
Australia	across ecoregions
Europe	within ecoregions

6.2 Set-up of the data-based model

This section will describe how the residuals of the resolved global carbon fluxes are predicted, using data-based models. Similar to Part I of this thesis, gradient boosted trees are used to predict the target variable from explanatory variables. If the reader is not familiar with gradient boosted trees, he/she is hereby encouraged to read section 2.1.1 first. This

Table 6.2: Overview of the models trained in this research.

Name	Split based on	Amount of models
Hemisphere	latitude	3
TransCom	TransCom region	11
Olson	Olson region	110
All	-	1

section describes first how the target data for the data-based models is obtained. Secondly, as different regions might have varying causes for errors in the prior model, different data-selection procedures are applied to train models. This is elaborated upon in the next section. Finally, the technical details of the data-based models and the quality assessment of the models is elaborated upon.

Obtaining the target data for the data-based model

The difference between prior and the posterior biosphere fluxes is used as target variable for the data-based models. The difference, or residuals, is calculated according to

$$Residuals = NEE_{opt} - NEE_{prior} \tag{6.2a}$$

$$Residuals = \lambda * NEE_{prior} - NEE_{prior}, \tag{6.2b}$$

where NEE_{opt} is the posterior NEE, calculated according to $\lambda * NEE_{prior}$, and NEE_{prior} is the NEE as simulated by the SiBCASA model. λ is the scaling factor, calculated by CTE. In this case, the posterior NEE fluxes are considered the true biosphere fluxes.

Data averaging

In order to reduce the computational costs of the data-based models, the average state of the ecoregion is used as input data. To clarify, ecoregions are defined as all grid-cells in that share the same Olson region (Olson et al., 2001) withing the TransCom continental region (e.g. European broadleaved forests). By doing so, also one ecoregion-aggregated NEE flux is calculated per ecoregion per month. For further analysis, the ecoregion fluxes are aggregated to TransCom regions.

Data-selection procedures

As stated before, the biosphere model might have varying causes for errors for different regions. For example, in SiBCASA, forests are deemed to be in a steady state. However, in regions where wood production is important, forests are managed as such that their productivity is as high as possible. As mature trees are logged, these forests are not in steady state and NEE is under-estimated by SiBCASA. In order to assess the quality of the data-based models for different regions, different models are trained for different regions. This is done in four distinct ways: 1) Based on latitude: a model is created for the northern hemisphere (NH, latitude $>$ 23N), for the equator (Eq, latitude between 23N and 23S) and for the southern hemisphere (SH, latitude $>$ 23S). These models are referred to as *Hemisphere*. 2) Based on TransCom region: For every land TransCom region, a different model is trained; these models are referred to as *TransCom*.; 3) Based on ecoregion: For every ecoregion, a different model is trained. These models are referred to as *Olson*.; 4) No split: a model is trained for all data-points. This model is referred to as *All*. An overview of the models trained is shown in Table 6.2. Note that the models are only applied in the regions for which they were trained.

Not all ecoregions are equally important for the carbon cycle, as for example deserts do not exchange a lot of carbon. In order to reduce the computational costs, only the most important ecoregions per TransCom region are used for ecoregion-specific model training and validation. NEE values can be both positive and negative, averaging out around 0. Therefore, the most important regions are defined here as the regions that contribute to more than 10% of the absolute NEE within their respective TransCom region. This reduced the amount of data by about 70%. As an example, the contribution of all ecoregions within boreal North America is shown in Figure 6.3. Only ecoregions that fall outside the orange box are used in for ecoregion-specific model training and validation.

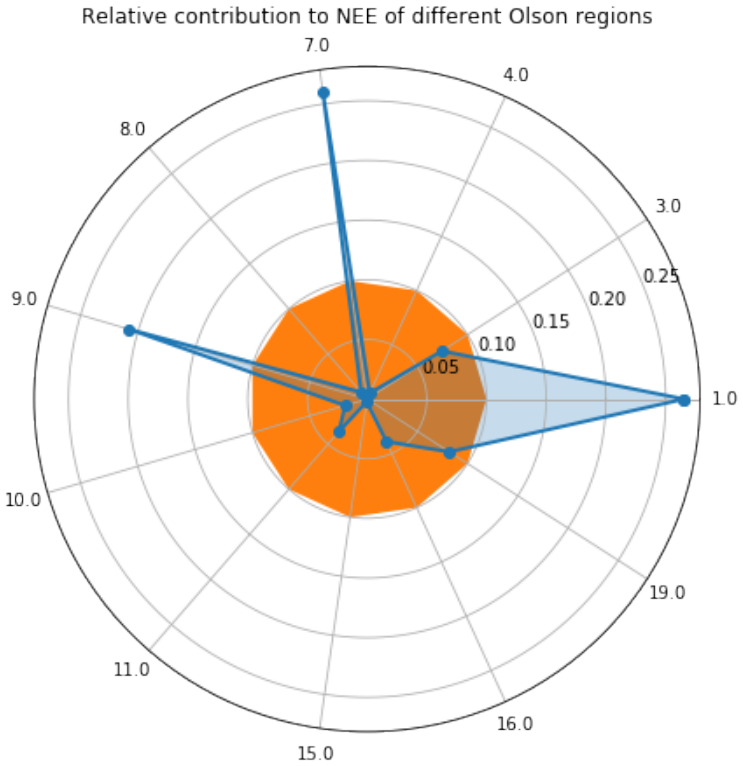

Figure 6.3: Radarplot of the relative contribution of different ecoregions to NEE in the boreal zone of North America

Hyperparameters

In order to maximise predictive quality of the data-based models, but reduce the chance of over-fitting, the hyperparameters of the model are tuned. Contrary to Section 2.1.1, where an exhaustive grid search was used, in this part of this thesis, a Bayesian approach to grid search is used. This is done using the Sequential Model-Based Optimization for General Algorithm Configuration (SMAC) (Hutter et al., 2011). SMAC first exhaustively searches a list of varying parameter configurations and uses a Bayesian approach to select the best combination of parameters. This is more cost-effective than an exhaustive grid search, as not all possible configurations are tested. Besides, it is more precise, as parameters are adjusted slightly based on an expected increase in performance. This is contrary to an exhaustive grid-search, where potential hyperparameters are taken from a prescribed list with a limited amount of freedom.

Variables used

In order to select the variables that add the most information on the residuals, recursive feature elimination is used in combination with the AIC (see also Section 2.1.1). To assess the most important features explaining residuals in all ecoregions, one set of features is selected based on all data. Because large numbers of variables bias the AIC, a bias corrected version of the AIC, the AIC_C is used. The AIC_C is calculated according to

$$AIC_C = (2 \cdot K) + n \cdot log\left(\frac{\sum_{i=1}^{n}(y^i_{measured} - y^i_{predicted})^2}{n}\right) + \frac{2K(K+1)}{n-K-1}, \tag{6.3}$$

where K is the number of variables used to train the model, n is the sample size, $y_{measured}$ are the target values, the NEE residuals according to CTE in this research and $y_{predicted}$ is the predicted NEE residual, according to the data-based model. The AIC_C is used for every model if the initial n/K is larger than 50 (Burnham and Anderson, 2004).

Because gradient boosted trees are subject to a slight randomness, RFE is conducted until the same variables result in the model with the lowest AIC five times. These variables are deemed the best variables for explaining

variance in the residuals and are used as input data for the models.

6.2.1 Model assessment

All models are trained on the years 2001 to 2014. The years 2015 and 2016 are used as testing data: this data is used for validation and quality assessment. By doing so, also the extrapolating quality of the models is assessed. For clarity, in the remainder of this section, the predicted fluxes by SiBCASA are indicated by *prior*, the predicted fluxes by CTE are indicated by *posterior*.

Because for the global carbon budget the total NEE is the most important, first the difference in the simulated global carbon budget between the prior, posterior and predicted fluxes is assessed.

Secondly, the improvement of the predictive quality of the machine-learned models with respect to the prior model is assessed. This first done using a Taylor-diagram (Taylor, 2001), indistinct of region. Secondly, in order to assess the predicitve quality of the models in different TransCom regions, also the RMSE per region is assessed. Thirdly, the seasonal cycle as predicted by the machine learned models is assessed, and the effect the simulated seasonal cycle has on the predicted NEE. Because it is more important to resolve the regions with high NEE better, only the most model statistics (Taylor plot and RMSE) of the most important regions as described in Section 6.2 are assessed.

6.3 Data used in this section

As target data for the data-based models, the residuals as calculated by CTE are used. The data used in this research is global data, consisting of 14538 land points on a 1 by 1 grid with monthly temporal resolution for 16 years of data.

As input data for the model, all variables that are generated and used by the SiBCASA model are used. The variables are shown in Appendix 12.2. To generate this data, SiBCASA is run on a 1 by 1 degree grid with a monthly temporal resolution. Although CT and SiBCASA are both run from 2000 up to and including 2016, the first year is deemed to be model-spin up and is therefore discarded, resulting in 16 years of data.

6.4 Workflow

An overview of the steps taken in this part of this thesis, including training and validation, using different methods, is shown in Figure 6.4.

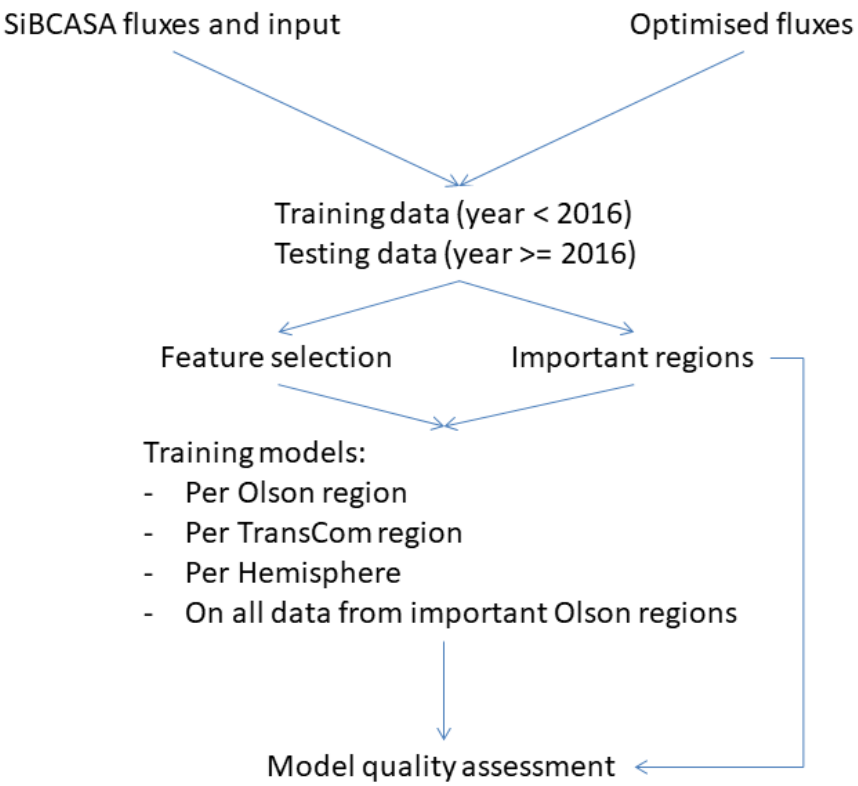

Figure 6.4: Basic flow diagram of the steps taken in this research.

Chapter 7

Results

This section describes the results of the data-based models predicting the posterior NEE. In this section, first the most important features that explain the residuals of CTE are assessed. Secondly, differences in the carbon budget simulated by the prior, posterior and machine learned model are assessed per TransCom region. Thirdly, the improvement of the predictive quality of the data-based models, compared to the prior model is assessed. This is done in three ways: by interpreting the Taylor diagram (Taylor, 2001) for global NEE of the most important ecoregions; by means of the RMSE improvement per TransCom region for the most important ecoregions and finally by assessing the capability of the prior and machine-learned models to simulate the seasonal cycle.

7.1 Important Features explaining variance in the residuals of CTE

In order to decrease computational costs of the data-based models and gain insight in the variables explaining the variance in the residuals of the CTE model, the most important variables are assessed. This is done based on a recursive feature elimination and the Akaike Information Criterion (AIC). The variables that resulted in the model with the lowest AIC are presented in Table 7.1. Additionally, the importance of the variable in the model is listed. The variable importance is based on the gain (see also Section 2.1.1). None of the variables listed has a correlation factor higher than 0.26 with the residuals.

Note that the visible and near infra-red (NIR) radiation and the percentage of sand in the soil are driver data for the model and are not simulated by the prior model.

Table 7.1: Most important variables in explaining the residuals in NEE according to CTE. The gain is a measure of the importance of the variable.

Long name	Short (SiBCASA) name	Gain
Sensible heat flux	hfss	45.97
Turbulent C flux	c_flux	41.47
Net Ecosystem Exchange	NEE_1	24.10
Weighted fractionation	wkiecps	22.95
Visible diffuse radiation	radvdc	22.95
NIR radiation	radndc	21.74
Percentage of sand in the soil	perc_sand	20.61
Ventilation mass flux	ventmf	17.35
Carbon flux from storage to wood	wood_frac	14.60
Shortwave incoming radiation	sw_dwn	14.24
Runoff	runoff	13.76
Carbon assimilation by C4 plants	assn_sum_c4	11.97
Snow depth on the ground	snow_depth	11.79
Chi squared between observed and simulated LAI	lai_chi_sqr	10.73

It appears that the sensible heat flux is the variable that explains most of the variance in the residuals, followed closely by the turbulent carbon flux. The Net ecosystem exchange is the third most important variable, but note that the gain is much higher for the two most important variables.

All variables, except for the fraction of sand in the soil, show a seasonal cycle, as do the residuals. Due to the black-box nature of the gradient boosted trees, it is impossible to assess if the variables that show a seasonal cycle explain the variance of the residuals because they follow a similar seasonal cycle as the residuals, or because the prior model over-simplifies important processes that are connected to the respective variables, resulting in a bias. A few observations are made however:

It is found that soils with a higher percentage of sand on average have smaller residuals than soils less sand. This could indicate a bias in SiBCASA for soils with less sand.

It has also been found that soils with a higher snow depth (>40cm) tend to have a lower prior NEE, and thus positive residuals, than soils with less snow. This could be due to CTE not taking snow depth into account in calculating the NEE, thereby over-estimating NEE at gridcells that are covered in snow. On the other hand, it could also indicate that SiBCASA tends to overestimate snow depth, thereby underestimating NEE. This has not been tested. However, the underestimation of NEE by SiBCASA in areas with less snow could also be due to the seasonality in the residuals.

The weighted fractionation is the fractionation of the heavier ^{13}C isotope, compared to the abundant ^{12}C isotope, which is dependent on the partial CO_2 pressure both in and outside the leaf (Farquhar et al., 1989). This variable was implemented in SiBCASA by Van der Velde et al. (2013), in order to study the global biospheric carbon sink in SiBCASA and does not directly affect the carbon fluxes in the biosphere. As the weighted fractionation depends on the simulated carbon concentrations both inside and outside of the leaf, it could be an indicator of errors in the calculation of these parameters. Because the carbon that is assimilated by plants is diffused through the stomata of the leaves. The speed of this diffusion depends on the carbon concentrations within and outside of the leaves. Therefore, the simulated carbon concentrations affect the NEE and hence the residuals.

The Chi squared between the observed and simulated LAI (χ^2_{LAI}) is used to assess if the measured LAI is systematically different than the simulated LAI. Large values for the χ^2_{LAI} indicate a large difference, whilst small values indicate no significant difference. The χ^2_{LAI} is generally the highest in spring, indicating that there is a flaw in the simulation of the LAI in spring in SiBCASA. Also in spring, the residuals are the largest, indicating that the NEE is underestimated by the model. In autumn however, the χ^2_{LAI} is very small, but the residuals are at a minimum, indicating that the simulated NEE is overestimated. This indicates that, even if the LAI is simulated correctly by the model, the NEE is overestimated.

The other variables listed in Table 7.1 are expected mainly to be important due to their seasonality.

7.2 Global carbon budget simulations

Net Ecosystem Exchange (NEE) is a key driver of the atmospheric CO_2 concentrations. As CO_2 is a long-lived trace gas, it is transported all over the globe. Being a long-lived gas, it is transported globally and biosphere fluxes at one continent can influence the CO_2 concentration at another. Therefore, it is important to assess the total global carbon uptake. In this section, also a distinction is made between different TransCom regions, in order to assess potential differences in quality.

The total posterior global carbon sink in 2016 is 3.35 PgC. The prior model simulated a global carbon sink of 3.39 PgC. Although the total global NEE as simulated by the prior is very close to the posterior flux, the NEE in different ecoregions is very different (Table 7.2). The similarity between the prior and posterior flux is due to compensating errors, for example the over-estimation of the NEE of Tropical South America and the under-estimation of the NEE in Australia

The prior model overestimates the carbon uptake in tropical South America by about 0.75 Pg. The machine learned models correct this over-estimation, although the correction by the machine-learned product trained on all data is small. The machine learned models provide systematic better estimates for Boreal North America, Tropical South Asia, Northern Africa, Boreal Eurasia and Australia. For Temperate Eurasia and Europe, three of the four machine-learned products estimate a carbon flux that is more different from the posterior flux than the prior. This is also because the difference between the posterior and the prior is quite small (0.08 PgC for both TransCom regions), making an improvement more difficult.

From the table, it appears that the machine-learned models trained per Olson region simulate the NEE to be more different from the posterior flux than the prior in 5 out of the 11 regions (Temperate North America, Temperate South America, Temperate Eurasia, Tropical Asia and Europe). The models trained per TransCom region and per hemisphere simulate the flux to be closer to the posterior in all but two regions. This indicates that the data-based models that are trained on a large part of the data, but also distinguish between different regions. This suggests that the variance in the residuals is not explained the same way for different regions.

Table 7.2: posterior, machine learned and prior carbon fluxes in 2016 per TransCom region in PgC/year. Also the total carbon uptake is shown. Note that negative fluxes indicate uptake. Machine-learned flux values in TransCom regions that are more different from the posterior flux than the prior are shown in bold.

Region	posterior	ML:TransCom	ML:Hemisphere	ML:Olson	ML:All	Prior
Boreal North America	-0.22	-0.20	-0.26	-0.12	-0.26	-0.06
Temperate North America	-0.36	-0.20	-0.16	**0.03**	-0.23	-0.04
Tropical South America	-0.26	-0.19	-0.31	-0.20	-0.80	-1.00
Temperate South America	-0.36	-0.26	-0.17	**-0.13**	-0.19	-0.16
North Africa	-0.76	-0.48	-0.56	-0.58	-0.47	-0.46
South Africa	-0.29	**-0.41**	-0.27	-0.31	**-0.34**	-0.33
Boreal Eurasia	-0.59	-0.74	-0.80	-0.83	-0.84	-0.33
Temperate Eurasia	-0.17	-0.20	**-0.06**	**-0.37**	**-0.02**	-0.09
Tropical Asia	-0.24	**-0.40**	-0.26	**-0.37**	-0.27	-0.34
Australia	0.15	0.12	0.14	0.03	-0.02	-0.43
Europe	-0.24	-0.25	**-0.38**	**-0.09**	**-0.40**	-0.16
Total	-3.35	-3.21	-3.09	-2.94	-3.83	-3.39

7.3 Model improvement

One of the objectives of this research is to decrease the errors in CTE by increasing the quality of the prior model. In order to decrease the error between the prior and posterior fluxes, the prior model (prior) is corrected by the machine learned model (prior + ML). In this section, the predictive quality of these models is compared to the predictive quality of the prior model.

7.3.1 Model comparison

The machine-learned models increase the model quality with respect to the prior model; the correlation between the prior+ML models and the posterior model is higher than the correlation between the prior and the posterior. Besides, the RMSE is lower for the prior+ML models than for the prior alone (Figure 7.1).

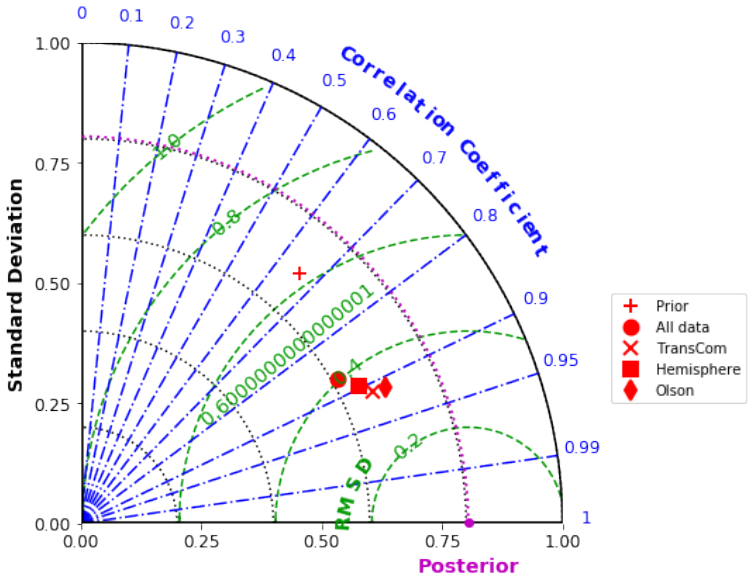

Figure 7.1: Taylor diagram of the prior and machine learned models in red. The posterior model is indicated by the magenta dot. The RMSE is shown in green as the RMSD. The blue lines show the Pearson correlation. The black dotted lines indicate the standard deviation and the magenta dots indicate the standard deviation of the posterior model.

From the figure, it appears that the machine learned fluxes have about the same correlation of 0.9 with the posterior flux, whereas the prior has a correlation of about 0.65. The model trained on all data shows the lowest correlation of the machine-learned models, of about 0.85. This indicates that the residuals of the prior model cannot be generalised globally, and a distinction needs to be made between regions to improve the model quality. This indicates that the variance in residuals explained differently for different regions. The Hemisphere, Transcom and Olson model have about the same correlation factor with the Posterior flux.

Also the RMSE (RMSD in Figure 7.1) is similar for all machine learned models. The model trained on all data shows the highest RMSE, again indicating that the residuals cannot be generalised to a global scale and that a distinction between regions is needed to improve the model quality. The Hemisphere model has a slightly larger RMSE than the TransCom and Olson model. The RMSE is assessed per TransCom region and in more detail below.

Figure 7.1 shows that the standard deviation of the machine-learned models trained on Olson regions has a standard deviation that is the most similar to the standard deviation of the posterior flux. The low standard deviations by all the machine learned models and the prior could indicate that the seasonal cycle of NEE is not well resolved. This is assessed below.

7.3.2 Root Mean Square Error per TransCom region

The prior model has a larger root mean square error (RMSE) for every TransCom region than the machine learned models (Figure 7.2). This indicates that the predictive quality of the machine-learned models is higher than the predictive quality of the prior model. For the TransCom regions in the southern hemisphere (i.e. Tropical and Temperate South America, North and South Africa and Australia), the RMSE is decreased the most. The machine-learned model that is trained with all data decreases the RMSE the least, which also shows in Figure 7.1. This model is indicated by the red bars in Figure 7.2. The other machine-learned models show about the same RMSE.

It is noteworhy that the machine-learned models perform about the same as the prior for TransCom region Temperate North America. For this region, the machine-learned models trained per Olson region are performing worse than the prior (Table 7.2. Nevertheless, in Figure 7.2, the models trained per Olson region are shown to reduce the RMSE the most in Temperate North America. This shows that only RMSE does not indicate a good model.

Note that this figure is created using only data from the important ecoregions. When the less important ecoregions are included in the analysis, the results are similar, albeit less clearly defined.

7.3.3 Seasonal cycle

As stated in section 7.3.1, the prior and the machine-learned NEE have a lower standard deviation than the posterior. This could be explained by an under-estimation of the amplitude of the seasonal cycle of NEE, which could in turn result in a bias in the carbon cycle. Therefore, the simulated seasonal cycle by the models is assessed. Both for the sake of brevity and because these regions show the most representative results, in this section, only the seasonal cycle in the TransCom regions South Africa and tropical Asia is elaborated upon (Figure 7.3).

For south Africa, the prior model simulates the seasonal cycle of the NEE to be out of phase with the posterior fluxes. All machine learned models correct the seasonal cycle of NEE to be in phase with the posterior fluxes. Because the seasonal cycle is only out of phase and the mean and amplitude of the seasonal cycle are simulated correctly by the prior model, the total NEE flux over 2016 and 2017 that is simulated by the prior is similar to the posterior fluxes. The machine-learned model trained on all data underestimates the amplitude of the seasonal cycle. As the average NEE is estimated correctly by this model, the simulated total NEE is similar to the posterior (Figure 7.3).

On the other hand, for tropical Asia, the machine learned models have more trouble capturing the variability of the NEE signal. Both the prior model and the machine-learned models underestimate the variability in the NEE (Figure 7.3). Due to the low seasonality near the equator, the NEE does not show a clearly defined seasonal cycle (Saigusa et al., 2008). Because the posterior and machine-learned fluxes fluctuate around the average prior flux, the total NEE over this region is similar for the prior, machine-learned and posterior models.

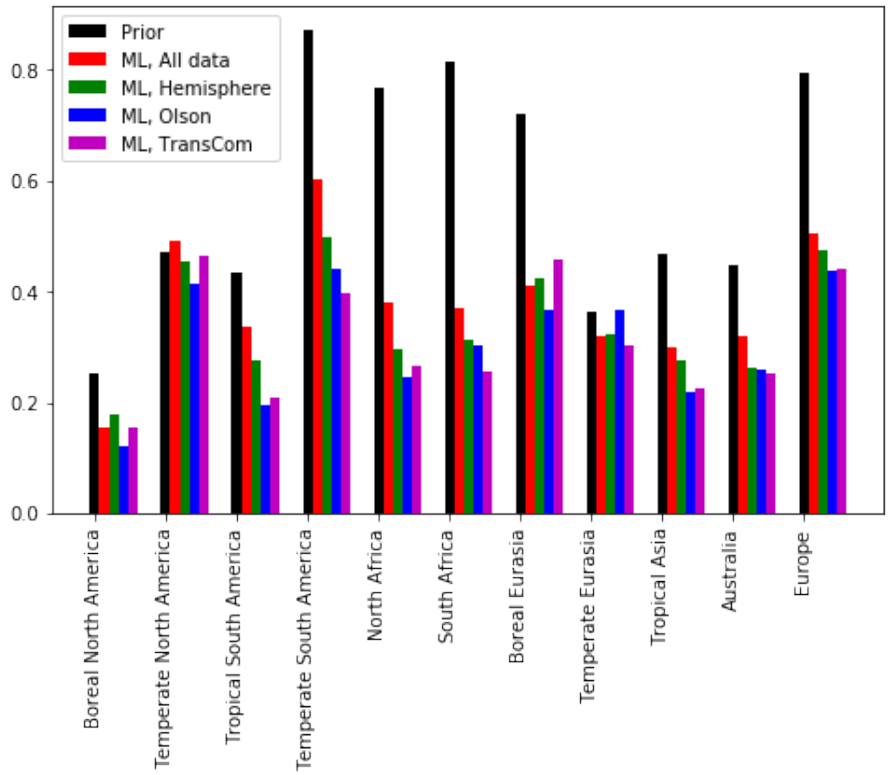

Figure 7.2: Root Mean Square Error of the models per TransCom region. The black bars indicate the prior model, the coloured bars indicate the machine learned models. The red bars are the models trained on all data, the models that are indicated by the green bars are trained with data per hemisphere. The blue and magenta bars are models trained per Olson and TransCom region, respectively.

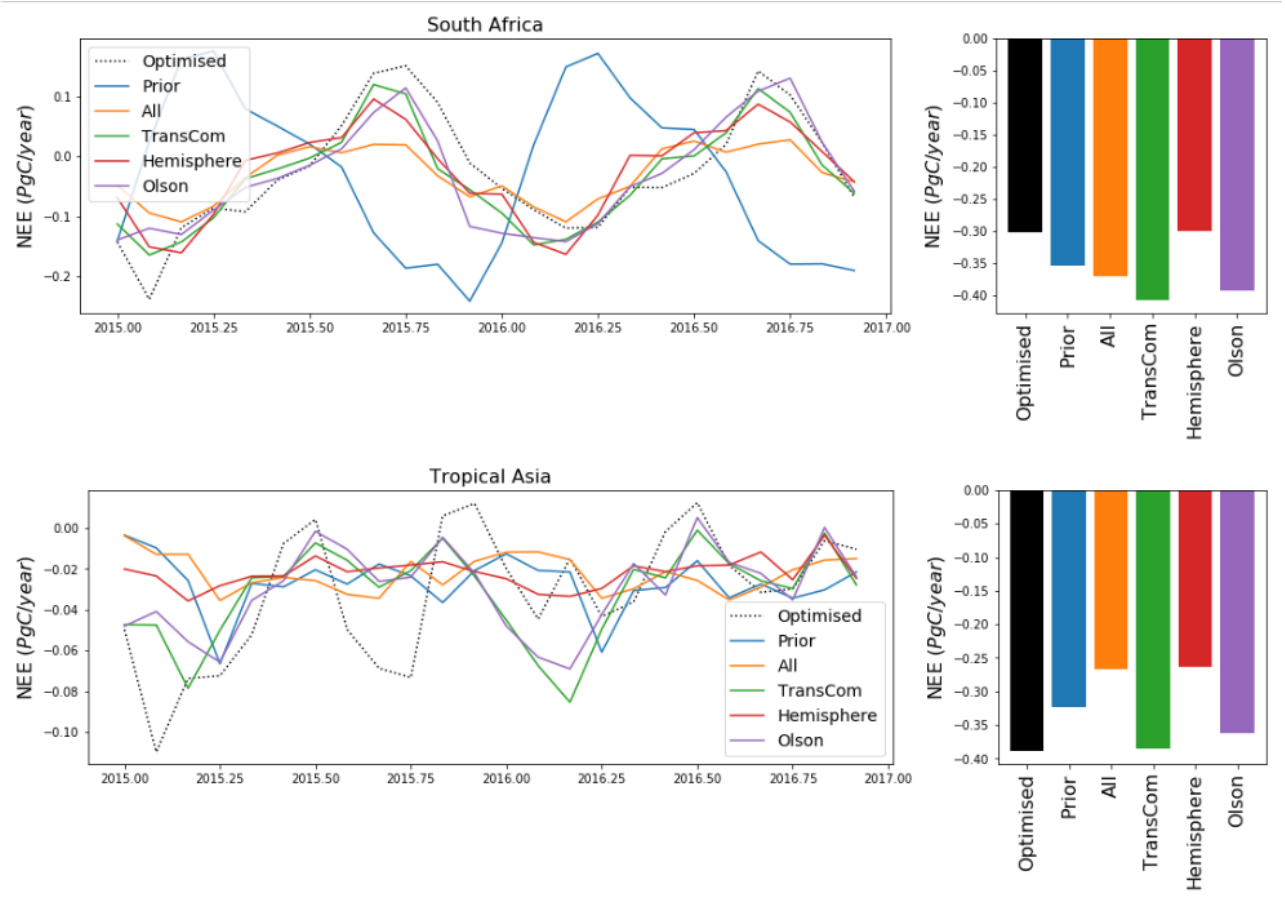

Figure 7.3: Left: Monthly predictions of the total NEE over south Africa (above) and Tropical Asia (below) in 2015 and 2016 per model. Right: The average NEE per year over 2015 and 2016 per model.

Chapter 8

Discussion and recommendations

In the previous chapters, data-based models were used to post-process output of the process-based model SiBCASA, which is used as prior for the inverse model CarbonTracker Europe (CTE). The post-processing is done in a similar fashion to model output statistics (MOS). Improving the prior model results in optimised fluxes that are less biased. Although an improvement of the model statistics was found for the corrected carbon flux by the data-based models, some improvements to the applied approach are possible. In this section, the methods and results are discussed. In doing so, also some recommendations for future research are given.

8.1 Discussion and recommendations of the methods

8.1.1 SiBCASA output

The CTE inversion uses SiBCASA net ecosystem exchange (NEE) as prior. The optimised fluxes are calculated based on the NEE flux that is calculated by this model. In this thesis, the output fluxes of CTE2018 are used as training data. However, the prior fluxes that are used in this thesis are not the same as the prior fluxes that are used by CTE2018, as a different SiBCASA run is used. The run used for this thesis ($SiBCASA_{Thesis}$) differs from the run used in CTE ($SiBCASA_{CTE}$) in that in ($SiBCASA_{Thesis}$), no fires are simulated. As burned area can regrow, the NEE is higher in the years after a fire. ($SiBCASA_{Thesis}$) was used due to the vast amount of output variables, that potentially could serve as explanatory variables in the data-based models. Due to a shortage of time, ($SiBCASA_{Thesis}$) is not used to create new results.

Although the SiBCASA run is different from the one used in CTE, the SiBCASA run used in this thesis could have been used as input for CTE. Therefore, the difference poses no problems for the methods and approach used in this thesis. Moreover,the model output of the correct SiBCASA run is expected to correlate better with the optimised fluxes than the SiBCASA run used in this thesis. Therefore, the results found in this thesis are expected to be conservative. This suggests that, if in a follow-up research the same SiBCASA model output is used for the inversion as input for the data-based models, the results of this thesis could be improved.

8.1.2 Extrapolated years

In this thesis, the years 2015 and 2016 are predicted, based on the preceding years from 2001 onward. 2015-2016 was a very strong El Niño season. Therefore, the results in this thesis show the predictions of CTE in a anomalous year. In order to study the performance of the model in non-anomalous years, it could add value to also predict other, non-anomalous years. As 2015 and 2016 are anomalous years, which are harder to predict, the results found in this research are expected to be conservative.

8.1.3 Correlating variables

In this thesis, the most important variables that explain variance in the residuals of CTE are used as input data in data-based models. However, some of these variables correlate with each other: for example the NEE and the turbulent carbon flux have a correlation factor of 0.99. Correlating variables do not add additional information on the system, as a split of the data on one of these variables could also be achieved by splitting on the other. Therefore one of the correlating variables could have been left out of the set of parameters. Because the variables used were found to result in the best model, according to the Akaike Information Criterion (AIC), the correlating variables are not left out in this thesis. Besides, the number of variables was already low. For a follow-up research, it could be chosen to remove correlating variables from the set of chosen variables. Although this is not expected to improve model statistics or results, this speeds up the training of the data-based model slightly.

8.1.4 Predicting NEE instead of the residuals

As the residuals of the carbon fluxes are predicted, the results found provide information on the error within SiBCASA, the prior model of CTE. Instead of the residuals however, also the optimised NEE could be predicted, using the output data of SiBCASA. Although this would abandon the objective of the research, which is improving the prior model through MOS, it has the benefit that the NEE is more evenly distributed than the residuals, as the residuals are often very close to 0. Because of this, the models trained might be biased towards 0, resulting in under-performing data-based models. This could in a future research be resolved by normalising the data, for example by taking the reciprocal of or the logarithm of the residuals as target data, or by predicting the NEE instead of the residual. Due to a lack of time, this has not been investigated.

8.1.5 Ecoregion average

As the optimised fluxes show some dipoles and other uncertainties, the data is averaged per ecoregion. This also reduced the computational costs of the data-based models. However, in this averaging, some of the spatial information is lost. Within an ecoregion, for example the leaf-area index might vary, resulting in varying carbon fluxes and varying residuals. This could also affect the feature importance. Therefore, for future research, not averaging the data might improve results, although also computational complexity is increased as the amount of data is increased 130-fold.

Predicting the NEE residual per gridcell has another advantage. In the CTE framework, the fluxes per gridcell are optimised. To use the data-based models as improved prior, the data-based models therefore need to be applied to 1 by 1 gridcells. For this, it is possible to use the ecoregion averaged correction, although a correction per gridcell would constrain the NEE better.

8.1.6 Feature engineering

Improvements to the predictive quality of the models could be achieved by feature engineering. As the aim of this research was, amongst others, to find the most important SiBCASA features that explain the variance in the residuals of NEE, this has not been done. Besides, it is expected that most of the variance explained is due to the seasonality of the variables. Therefore, feature engineering is not expected to increase the predictive quality of the models significantly.

8.2 Discussion of the results

In this thesis, it has been found that the prior fluxes of CTE can be improved by using machine-learned models to conduct model output statistics. Most notably, the models that are trained per TransCom region and per tropical are found to perform best. This indicates that the residuals stem from different processes in different regions. The good performance for the models per tropical indicates that a major distinction that needs to be made is the distinction between northern and southern hemisphere. This could be due to the gridded state vector, that is applied in the northern hemisphere.

On the other hand, the models that are trained per Olson region and on all data performed the worse. The reason for this is different for these two models. First, for the models trained per Olson region, it is expected that the bad performance is due to the small amount of data that is used in the training of the models. Therefore, the models are not generalised. More generalised models could be achieved by training with longer time periods. Secondly, for the models that are trained on all data, the opposite is expected to be the case. Due to the vast amount of data that is used, the model is biased towards 0. This is also shown in the under-estimation of the seasonal cycle and can be seen as a form of over-fitting.

As carbon dioxide is a long-lived trace gas, it is transported through the atmosphere. In CTE, this is resolved by the global transport model TM5. In order to assess whether the machine-learned fluxes are more suitable as prior, the fluxes need to be run through TM5. The simulated CO_2 mole fractions need then to be compared to the observed. Due to temporal constraints, this has not been done.

Chapter 9

Conclusion

Inverse models, such as CarbonTracker Europe (CTE) simulate carbon fluxes based on estimated carbon fluxes (the prior) and observations. The better the prior is, the smaller the uncertainty in the simulated carbon fluxes (the posterior) by CTE are. In this thesis, the prior flux is adjusted by finding relations between the difference between the prior and posterior model (the residuals) and the output of the prior model using machine learning. These relations are then used to improve the prior model for 2015 and 2016. The results show that the prior carbon flux estimates are improved by the data-based models. The two main findings of this thesis are:

- The variables that explain the residuals in the predicted NEE the best are the sensible heat flux and the turbulent C flux. The variables that result in the best model and therefore are used in the model, show a strong seasonality, similar to the residuals. This is with the exception of the fraction of sand in the soil.
- The carbon flux estimates are improved by data-based models. Four types of data-based models were trained, all of which improved the prior simulation of the carbon budget in TransCom regions. The models trained per TransCom region and models that are trained per Tropic were found to perform best, as they have a small root mean square error (RMSE) and simulate a carbon budget that is similar to the posterior. Although the models trained per Olson region have the smallest RMSE, they did only improve carbon budget simulations in 6 out of 11 TransCom regions. The model trained on all data reduced the RMSE the least. Most notably, the seasonal cycle was under-estimated by the model trained on all data.

All in all, data-based models appear to be a valid tool to conduct model output statistics on a process-based model. Potentially, the adjusted process-based model can be used as a less biased prior in an inverse model, resulting in posterior carbon fluxes that are less biased.

Chapter 10

Synthesis: The role of machine learning in the carbon cycle

In this thesis, parts of the carbon cycle are simulated using data-based models. This is done in two ways:

1. By using data-based models to make a prediction of gross primary production (GPP) based on measurements, satellite observations and meteorological data.
2. By using data-based models to perform model-output statistics (MOS) on a process based model, in order to improve the prior estimate for a carbon inversion.

This chapter reflects on the place of machine-learning and data-based model in modelling the carbon cycle.

Currently, data-based models are used as benchmark for global GPP estimates (Beer et al., 2010; Bodesheim et al., 2018). Due to their unbiased nature and high performance, process based models are scaled to these benchmarks. In this thesis, data-based models are used as a predictive model to predict global and local GPP. It has been found that the models have a high performance, but also that the performance is very dependent on the quality of the input data.

Besides the quality of the input data, data-based models have two additional drawbacks that limit their use to resolve the carbon cycle. Firstly, data-based models are trained on the past state of the system. In a changing climate, response to drivers of GPP might change. Therefore, it comes with risk to use a data-based model as a predictive model. Secondly, due to the high black-box calibre of most data-based models, the only new information that can be gained from most data-based models are the estimated flux and the feature importance. In the current state, data-based models do provide a reliable estimate of carbon fluxes, but do not add to our knowledge on the carbon cycle and how the carbon cycle might respond to a changing climate.

These drawbacks are of minor importance if data-based models are used to perform MOS on a process-based model. MOS is based on knowledge of the performance of a model and has not as an objective to improve the knowledge of the system, but to increase the (predictive) quality of the model. Therefore, the black-box calibre of the data-based model is not a major issue. Moreover, data-based models may help to locate and understand flaws in the prior model, through the means of assessing the feature importance. Therefore, data-based models could prove valuable in improving prior models by doing MOS.

For using data-based models to perform MOS on process-based models, it has been found that data-based models are more similar to the posterior, potentially reducing the uncertainty in global carbon fluxes. However, a data-inversion is still needed, as the improved prior model is not of high enough quality to replace the data-inversion.

In conclusion, the choice whether to use a data-based model to resolve the carbon cycle is dependent on the goal of the research. Due to the unbiased, theory-independent and highly non-linear predictions of data-based models, they can be used to both estimate carbon fluxes directly or improve process-based models with a high predictive quality. However, because of the high black-box calibre, data-based models cannot be used to improve our knowledge on the carbon cycle and how it will respond to a changing climate.

Chapter 11

Acknowledgements

This research was supervised by dr.ir. Liesbeth Florentie and prof.dr. Wouter Peters. I would like to thank them for their support, contributions and their critical questions. Besides, I would like to thank Wouter Peters for providing me the possibility to conduct a research that had my personal interest, and Liesbeth Florentie for guiding me through this. I would also like to thank Erik van Schaik and Ingrid van der Laan-Luijx for their help on CarbonTracker and SiBCASA.

I would also like to thank the members of the thesis-ring, who have helped me to improve my text. In special, I would like to thank Luuk Bersee, which whom I have held most of my breaks and who once drove the chair from my computer to the coffee machine because he thought I needed a break.

This research is based on the previous research and hard work by thousands of scholars globally. Although it is impossible to name them all, I would like to thank all of them for their hard work and the knowledge they provide us. In special, I would like to thank Christian Beer and his colleagues for laying the foundation for using data-based models in ecosystem research.

This work used eddy covariance data acquired and shared by the FLUXNET community, including these networks: AmeriFlux, AfriFlux, AsiaFlux, CarboAfrica, CarboEuropeIP, CarboItaly, CarboMont, ChinaFlux, Fluxnet-Canada, GreenGrass, ICOS, KoFlux, LBA, NECC, OzFlux-TERN, TCOS-Siberia, and USCCC. The ERA-Interim reanalysis data are provided by ECMWF and processed by LSCE. The FLUXNET eddy covariance data processing and harmonization was carried out by the European Fluxes Database Cluster, AmeriFlux Management Project, and Fluxdata project of FLUXNET, with the support of CDIAC and ICOS Ecosystem Thematic Center, and the OzFlux, ChinaFlux and AsiaFlux offices.

Bibliography

(2018). The data portal serving the fluxnet community.

Ahlström, A., Raupach, M. R., Schurgers, G., Smith, B., Arneth, A., Jung, M., Reichstein, M., Canadell, J. G., Friedlingstein, P., Jain, A. K., et al. (2015). The dominant role of semi-arid ecosystems in the trend and variability of the land co2 sink. *Science*, 348(6237):895–899.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer.

Arida, M. (2002). South atlantic anomaly.

Badgley, G., Anderegg, L. D., Berry, J. A., and Field, C. B. (2018). Terrestrial gross primary production: Using nirv to scale from site to globe.

Badgley, G., Field, C. B., and Berry, J. A. (2017). Canopy near-infrared reflectance and terrestrial photosynthesis. *Science advances*, 3(3):e1602244.

Baker, N. R. (2008). Chlorophyll fluorescence: a probe of photosynthesis in vivo. *Annu. Rev. Plant Biol.*, 59:89–113.

Baldocchi, D., Falge, E., Gu, L., Olson, R., Hollinger, D., Running, S., Anthoni, P., Bernhofer, C., Davis, K., Evans, R., et al. (2001). Fluxnet: A new tool to study the temporal and spatial variability of ecosystem–scale carbon dioxide, water vapor, and energy flux densities. *Bulletin of the American Meteorological Society*, 82(11):2415–2434.

Baldocchi, D. D. (2003). Assessing the eddy covariance technique for evaluating carbon dioxide exchange rates of ecosystems: past, present and future. *Global change biology*, 9(4):479–492.

Barr, A., Richardson, A., Hollinger, D., Papale, D., Arain, M., Black, T., Bohrer, G., Dragoni, D., Fischer, M., Gu, L., et al. (2013). Use of change-point detection for friction–velocity threshold evaluation in eddy-covariance studies. *Agricultural and Forest Meteorology*, 171:31–45.

Beer, C., Reichstein, M., Tomelleri, E., Ciais, P., Jung, M., Carvalhais, N., Rödenbeck, C., Arain, M. A., Baldocchi, D., Bonan, G. B., et al. (2010). Terrestrial gross carbon dioxide uptake: global distribution and covariation with climate. *Science*, 329(5993):834–838.

Bodesheim, P., Jung, M., Gans, F., Mahecha, M. D., and Reichstein, M. (2018). Upscaled diurnal cycles of land-atmosphere fluxes: a new global half-hourly data product. *Earth System Science Data Discussions*.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., and Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research*, 33(2):261–304.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM.

Cox, P. M., Betts, R. A., Jones, C. D., Spall, S. A., and Totterdell, I. J. (2000). Acceleration of global warming due to carbon-cycle feedbacks in a coupled climate model. *Nature*, 408(6809):184.

Cramer, W., Kicklighter, D. W., Bondeau, A., Iii, B. M., Churkina, G., Nemry, B., Ruimy, A., Schloss, A. L., and Model, P. O. T. P. N. (1999). Comparing global models of terrestrial net primary productivity (npp): overview and key results. *Global change biology*, 5(S1):1–15.

De'Ath, G. (2007). Boosted trees for ecological modeling and prediction. *Ecology*, 88(1):243–251.

Dee, D. P., Uppala, S. M., Simmons, A., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., Bauer, d. P., et al. (2011). The era-interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the royal meteorological society*, 137(656):553–597.

Dlugokencky, E. and Tans, P. (2015). Esrl global monitoring division–global greenhouse gas reference network. *Trends in Atmospheric Carbon Dioxide, Global-Globally Averaged Marine Surface Annual Mean Data*.

Dorogush, A. V., Gulin, A., Gusev, G., Kazeev, N., Prokhorenkova, L. O., and Vorobev, A. (2017). Fighting biases with dynamic boosting. *arXiv preprint arXiv:1706.09516*.

Duveiller, G. and Cescatti, A. (2016). Spatially downscaling sun-induced chlorophyll fluorescence leads to an improved temporal correlation with gross primary productivity. *Remote Sensing of Environment*, 182:72–89.

ESRL (2005). Carbontracker documentation ct2017 release.

Etheridge, D. M., Steele, L., Langenfelds, R., Francey, R., Barnola, J.-M., and Morgan, V. (1996). Natural and anthropogenic changes in atmospheric co2 over the last 1000 years from air in antarctic ice and firn. *Journal of Geophysical Research: Atmospheres*, 101(D2):4115–4128.

Farquhar, G. D., Ehleringer, J. R., and Hubick, K. T. (1989). Carbon isotope discrimination and photosynthesis. *Annual review of plant biology*, 40(1):503–537.

Frank, E., Wang, Y., Inglis, S., Holmes, G., and Witten, I. H. (1998). Using model trees for classification. *Machine Learning*, 32(1):63–76.

Friedl, M. A., Sulla-Menashe, D., Tan, B., Schneider, A., Ramankutty, N., Sibley, A., and Huang, X. (2010). Modis collection 5 global land cover: Algorithm refinements and characterization of new datasets. *Remote sensing of Environment*, 114(1):168–182.

Friedlingstein, P., Andrew, R. M., Rogelj, J., Peters, G., Canadell, J. G., Knutti, R., Luderer, G., Raupach, M. R., Schaeffer, M., van Vuuren, D. P., et al. (2014). Persistent growth of co 2 emissions and implications for reaching climate targets. *Nature geoscience*, 7(10):709.

Friedlingstein, P., Cox, P., Betts, R., Bopp, L., von Bloh, W., Brovkin, V., Cadule, P., Doney, S., Eby, M., Fung, I., et al. (2006). Climate–carbon cycle feedback analysis: results from the c4mip model intercomparison. *Journal of climate*, 19(14):3337–3353.

Friend, A. D., Arneth, A., Kiang, N. Y., Lomas, M., Ogee, J., Rödenbeck, C., Running, S. W., SANTAREN, J.-D., Sitch, S., Viovy, N., et al. (2007). Fluxnet and modelling the global carbon cycle. *Global Change Biology*, 13(3):610–633.

Giglio, L., Randerson, J. T., and van der Werf, G. R. (2013). Analysis of daily, monthly, and annual burned area using the fourth-generation global fire emissions database (gfed4). *Journal of Geophysical Research: Biogeosciences*, 118(1):317–328.

Gurney, K. R., Law, R. M., Denning, A. S., Rayner, P. J., Baker, D., Bousquet, P., Bruhwiler, L., Chen, Y.-H., Ciais, P., Fan, S., et al. (2002). Towards robust regional estimates of co 2 sources and sinks using atmospheric transport models. *Nature*, 415(6872):626.

Heimann, M. and Reichstein, M. (2008). Terrestrial ecosystem carbon dynamics and climate feedbacks. *Nature*, 451(7176):289.

Huston, M. A. and Wolverton, S. (2009). The global distribution of net primary production: resolving the paradox. *Ecological monographs*, 79(3):343–377.

Hutter, F., Hoos, H. H., and Leyton-Brown, K. (2011). Sequential model-based optimization for general algorithm configuration. In *International Conference on Learning and Intelligent Optimization*, pages 507–523. Springer.

Jackson, R. B., Canadell, J. G., Le Quéré, C., Andrew, R. M., Korsbakken, J. I., Peters, G. P., and Nakicenovic, N. (2015). Reaching peak emissions. *Nature Climate Change*, 6(1):7.

Joiner, J., Yoshida, Y., Vasilkov, A., Schaefer, K., Jung, M., Guanter, L., Zhang, Y., Garrity, S., Middleton, E., Huemmrich, K., et al. (2014). 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.

Joiner, J., Yoshida, Y., Zhang, Y., Duveiller, G., Jung, M., Lyapustin, A., Wang, Y., and Tucker, C. (2018). Estimation of terrestrial global gross primary production (gpp) with satellite data-driven models and eddy covariance flux data. *Remote Sensing*, 10(9):1346.

Jones, E., Oliphant, T., Peterson, P., et al. (2018). Signal processing (scipy.signal).

Jung, M., Reichstein, M., and Bondeau, A. (2009). Towards global empirical upscaling of fluxnet eddy covariance observations: validation of a model tree ensemble approach using a biosphere model. *Biogeosciences*, 6(10):2001–2013.

Jung, M., Reichstein, M., Margolis, H. A., Cescatti, A., Richardson, A. D., Arain, M. A., Arneth, A., Bernhofer, C., Bonal, D., Chen, J., et al. (2011). Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations. *Journal of Geophysical Research: Biogeosciences*, 116(G3).

Jung, M., Vetter, M., Herold, M., Churkina, G., Reichstein, M., Zaehle, S., Ciais, P., Viovy, N., Bondeau, A., Chen, Y., et al. (2007). Uncertainties of modeling gross primary productivity over europe: A systematic study on the effects of using different drivers and terrestrial biosphere models. *Global Biogeochemical Cycles*, 21(4).

Knorr, W. (2009). Is the airborne fraction of anthropogenic co2 emissions increasing? *Geophysical Research Letters*, 36(21).

Knyazikhin, Y., Martonchik, J., Myneni, R. B., Diner, D., and Running, S. W. (1998). Synergistic algorithm for estimating vegetation canopy leaf area index and fraction of absorbed photosynthetically active radiation from modis and misr data. *Journal of Geophysical Research: Atmospheres*, 103(D24):32257–32275.

Koffi, E., Rayner, P., Scholze, M., and Beer, C. (2012). Atmospheric constraints on gross primary productivity and net ecosystem productivity: Results from a carbon-cycle data assimilation system. *Global Biogeochemical Cycles*, 26(1).

Kooreman, M., Stammes, P., Tuinder, O., Boersma, F., van Schaik, E., and Botia, S. (2011). Gome-2 sun-induced fluorescence of terrestrial ecosystems retrieval. 12:2825–2830.

Koren, G., van Schaik, E., Araújo, A. C., Boersma, K. F., Gärtner, A., Killaars, L., Kooreman, M. L., Kruijt, B., van der Laan-Luijkx, I. T., von Randow, C., et al. (2018). Widespread reduction in sun-induced fluorescence from the amazon during the 2015/2016 el niño. *Phil. Trans. R. Soc. B*, 373(1760):20170408.

Krol, M., Houweling, S., Bregman, B., Broek, M. v., Segers, A., Velthoven, P. v., Peters, W., Dentener, F., and Bergamaschi, P. (2005). The two-way nested global chemistry-transport zoom model tm5: algorithm and applications. *Atmospheric Chemistry and Physics*, 5(2):417–432.

Krstajic, D., Buturovic, L. J., Leahy, D. E., and Thomas, S. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of cheminformatics*, 6(1):10.

Landsberg, J. and Waring, R. (1997). A generalised model of forest productivity using simplified concepts of radiation-use efficiency, carbon balance and partitioning. *Forest ecology and management*, 95(3):209–228.

Lasslop, G., Reichstein, M., Papale, D., Richardson, A. D., Arneth, A., Barr, A., Stoy, P., and Wohlfahrt, G. (2010). Separation of net ecosystem exchange into assimilation and respiration using a light response curve approach: critical issues and global evaluation. *Global Change Biology*, 16(1):187–208.

Lee, J.-E., Frankenberg, C., van der Tol, C., Berry, J. A., Guanter, L., Boyce, C. K., Fisher, J. B., Morrow, E., Worden, J. R., Asefi, S., et al. (2013). Forest productivity and water stress in amazonia: Observations from gosat chlorophyll fluorescence. *Proceedings of the Royal Society of London B: Biological Sciences*, 280(1761):20130171.

Li, J., Cheng, K., Wang, S., Morstatter, F., Robert, T., Tang, J., and Liu, H. (2016). Feature selection: A data perspective. *arXiv:1601.07996*.

Li, X., Xiao, J., He, B., Altaf Arain, M., Beringer, J., Desai, A. R., Emmel, C., Hollinger, D. Y., Krasnova, A., Mammarella, I., et al. (2018). Solar-induced chlorophyll fluorescence is strongly correlated with terrestrial photosynthesis for a wide variety of biomes: First global analysis based on oco-2 and flux tower observations. *Global change biology*.

Loescher, H. W., Law, B., Mahrt, L., Hollinger, D., Campbell, J., and Wofsy, S. C. (2006). Uncertainties in, and interpretation of, carbon flux estimates using the eddy covariance technique. *Journal of Geophysical Research: Atmospheres*, 111(D21).

Lundberg, S. M. and Lee, S.-I. (2017). Consistent feature attribution for tree ensembles. *arXiv preprint arXiv:1706.06060*.

Maccherone, B. (2005). Modis leaf area index/fpar.

Mao, J., Thornton, P. E., Shi, X., Zhao, M., and Post, W. M. (2012). Remote sensing evaluation of clm4 gpp for the period 2000–09. *Journal of Climate*, 25(15):5327–5342.

Maxwell, K. and Johnson, G. N. (2000). Chlorophyll fluorescenceâĂŤa practical guide. *Journal of experimental botany*, 51(345):659–668.

Meier-Fleischer, K. and Böttinger, M. (2018). poisson grid fill.

Melillo, J. M., McGuire, A. D., Kicklighter, D. W., Moore, B., Vorosmarty, C. J., and Schloss, A. L. (1993). Global climate change and terrestrial net primary production. *Nature*, 363(6426):234.

Miyata, A., Papale, D., Agarwal, D., Baldocchi, D., van Gorsel, E., Yu, G. R., Cleugh, H., Kim, J., Nary, L., Torn, M., Humphrey, M., Vargas, R., and Wolf, S. (2018). Fluxnet. `http://fluxnet.fluxdata.org/`. Accessed: 2019-02-03.

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., and Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, 50(3):885–900.

Myneni, R. B., Hoffman, S., Knyazikhin, Y., Privette, J., Glassy, J., Tian, Y., Wang, Y., Song, X., Zhang, Y., Smith, G., et al. (2002). Global products of vegetation leaf area and fraction absorbed par from year one of modis data. *Remote sensing of environment*, 83(1-2):214–231.

Olofsson, P. and Eklundh, L. (2007). Estimation of absorbed par across scandinavia from satellite measurements. part ii: Modeling and evaluating the fractional absorption. *Remote Sensing of Environment*, 110(2):240–251.

Olson, D. M., Dinerstein, E., Wikramanayake, E. D., Burgess, N. D., Powell, G. V., Underwood, E. C., D'amico, J. A., Itoua, I., Strand, H. E., Morrison, J. C., et al. (2001). Terrestrial ecoregions of the world: A new map of life on eartha new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. *BioScience*, 51(11):933–938.

Papale, D., Reichstein, M., Aubinet, M., Canfora, E., Bernhofer, C., Kutsch, W., Longdoz, B., Rambal, S., Valentini, R., Vesala, T., et al. (2006). Towards a standardized processing of net ecosystem exchange measured with eddy covariance technique: algorithms and uncertainty estimation. *Biogeosciences*, 3(4):571–583.

Parazoo, N. C., Bowman, K., Fisher, J. B., Frankenberg, C., Jones, D. B., Cescatti, A., Pérez-Priego, Ó., Wohlfahrt, G., and Montagnani, L. (2014). Terrestrial gross primary production inferred from satellite fluorescence and vegetation models. *Global change biology*, 20(10):3103–3121.

Peters, W. (2018). Carbontracker europe.

Peters, W., Jacobson, A. R., Sweeney, C., Andrews, A. E., Conway, T. J., Masarie, K., Miller, J. B., Bruhwiler, L. M., Pétron, G., Hirsch, A. I., et al. (2007). An atmospheric perspective on north american carbon dioxide exchange: Carbontracker. *Proceedings of the National Academy of Sciences*, 104(48):18925–18930.

Peters, W., Krol, M. C., Werf, G. R. V. D., Houweling, S., D., J. C., Schaefer, K., Masarie, K. A., Jacobson, A. R., Miller, J. B., Cho, C. H., and et al. (2009). Seven years of recent european net terrestrial carbon dioxide exchange constrained by atmospheric observations.

Peters, W., Miller, J., Whitaker, J., Denning, A., Hirsch, A., Krol, M., Zupanski, D., Bruhwiler, L., and Tans, P. (2005). An ensemble data assimilation system to estimate co2 surface fluxes from atmospheric trace gas observations. *Journal of Geophysical Research: Atmospheres*, 110(D24).

Piao, S., Ciais, P., Friedlingstein, P., de Noblet-Ducoudré, N., Cadule, P., Viovy, N., and Wang, T. (2009). Spatiotemporal patterns of terrestrial carbon cycle during the 20th century. *Global Biogeochemical Cycles*, 23(4).

Posada, D. and Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic biology*, 53(5):793–808.

Potter, C. S., Randerson, J. T., Field, C. B., Matson, P. A., Vitousek, P. M., Mooney, H. A., and Klooster, S. A. (1993). Terrestrial ecosystem production: a process model based on global satellite and surface data. *Global Biogeochemical Cycles*, 7(4):811–841.

Poulter, B., Frank, D., Ciais, P., Myneni, R. B., Andela, N., Bi, J., Broquet, G., Canadell, J. G., Chevallier, F., Liu, Y. Y., et al. (2014). Contribution of semi-arid ecosystems to interannual variability of the global carbon cycle. *Nature*, 509(7502):600.

Raulier, F., Bernier, P. Y., and Ung, C.-H. (2000). Modeling the influence of temperature on monthly gross primary productivity of sugar maple stands. *Tree Physiology*, 20(5-6):333–345.

Reichstein, M., Bahn, M., Ciais, P., Frank, D., Mahecha, M. D., Seneviratne, S. I., Zscheischler, J., Beer, C., Buchmann, N., Frank, D. C., et al. (2013). Climate extremes and the carbon cycle. *Nature*, 500(7462):287.

Reichstein, M., Falge, E., Baldocchi, D., Papale, D., Aubinet, M., Berbigier, P., Bernhofer, C., Buchmann, N., Gilmanov, T., Granier, A., et al. (2005). On the separation of net ecosystem exchange into assimilation and ecosystem respiration: review and improved algorithm. *Global Change Biology*, 11(9):1424–1439.

Saigusa, N., Yamamoto, S., Hirata, R., Ohtani, Y., Ide, R., Asanuma, J., Gamo, M., Hirano, T., Kondo, H., Kosugi, Y., et al. (2008). Temporal and spatial variations in the seasonal patterns of co2 flux in boreal, temperate, and tropical forests in east asia. *Agricultural and Forest Meteorology*, 148(5):700–713.

Sanders, A. F., Verstraeten, W. W., Kooreman, M. L., Van Leth, T. C., Beringer, J., and Joiner, J. (2016). Spaceborne sun-induced vegetation fluorescence time series from 2007 to 2015 evaluated with australian flux tower measurements. *Remote Sensing*, 8(11):895.

Schaefer, K., Collatz, G. J., Tans, P., Denning, A. S., Baker, I., Berry, J., Prihodko, L., Suits, N., and Philpott, A. (2008). Combined simple biosphere/carnegie-ames-stanford approach terrestrial carbon cycle model. *Journal of Geophysical Research: Biogeosciences*, 113(G3).

Schaik, E. (2016). Retrieving sun-induced fluorescence from the global ozone monitoring experiment 2. *personal communication*.

Schlesinger, W. H. and Bernhardt, E. S. (2013). Chapter 11 - the global carbon cycle. In Schlesinger, W. H. and Bernhardt, E. S., editors, *Biogeochemistry (Third Edition)*, pages 419 – 444. Academic Press, Boston, third edition edition.

Schneider, S. H. (1989). Global warming: are we entering the greenhouse century?

Sellers, P., Mintz, Y., Sud, Y. e. a., and Dalcher, A. (1986). A simple biosphere model (sib) for use within general circulation models. *Journal of the Atmospheric Sciences*, 43(6):505–531.

Sellers, P., Randall, D., Collatz, G., Berry, J., Field, C., Dazlich, D., Zhang, C., Collelo, G., and Bounoua, L. (1996a). A revised land surface parameterization (sib2) for atmospheric gcms. part i: Model formulation. *Journal of climate*, 9(4):676–705.

Sellers, P. J., Tucker, C. J., Collatz, G. J., Los, S. O., Justice, C. O., Dazlich, D. A., and Randall, D. A. (1996b). A revised land surface parameterization (sib2) for atmospheric gcms. part ii: The generation of global fields of terrestrial biophysical parameters from satellite data. *Journal of climate*, 9(4):706–737.

Simmons, A., Poli, P., Dee, D., Berrisford, P., Hersbach, H., Kobayashi, S., and Peubey, C. (2014). Estimating low-frequency variability and trends in atmospheric temperature using era-interim. *Quarterly Journal of the Royal Meteorological Society*, 140(679):329–353.

Sitch, S., Huntingford, C., Gedney, N., Levy, P., Lomas, M., Piao, S., Betts, R., Ciais, P., Cox, P., Friedlingstein, P., et al. (2008). Evaluation of the terrestrial carbon cycle, future plant geography and climate-carbon cycle feedbacks using five dynamic global vegetation models (dgvms). *Global Change Biology*, 14(9):2015–2039.

Still, C. J., Berry, J. A., Collatz, G. J., and DeFries, R. S. (2003). Global distribution of c3 and c4 vegetation: carbon cycle implications. *Global Biogeochemical Cycles*, 17(1):6–1.

Stuart, A. and Taeb, A. (2018). Data assimilation and inverse problems. *Draft*.

Sun, Y., Fu, R., Dickinson, R., Joiner, J., Frankenberg, C., Gu, L., Xia, Y., and Fernando, N. (2015). Drought onset mechanisms revealed by satellite solar-induced chlorophyll fluorescence: Insights from two contrasting extreme events. *Journal of Geophysical Research: Biogeosciences*, 120(11):2427–2440.

Swalin, A. (2018). Catboost vs. light gbm vs. xgboost âĂŞ towards data science.

Tait, L. W. and Schiel, D. R. (2013). Impacts of temperature on primary productivity and respiration in naturally structured macroalgal assemblages. *PLoS One*, 8(9):e74413.

Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres*, 106(D7):7183–7192.

Tol, C., Berry, J., Campbell, P., and Rascher, U. (2014). Models of fluorescence and photosynthesis for interpreting measurements of solar-induced chlorophyll fluorescence. *Journal of Geophysical Research: Biogeosciences*, 119(12):2312–2327.

Turner, B., Meyer, W. B., Skole, D. L., et al. (1994). Global land-use/land-cover change: towards an integrated study. *Ambio. Stockholm*, 23(1):91–95.

van der Laan-Luijkx, I. T., van der Velde, I. R., van der Veen, E., Tsuruta, A., Stanislawska, K., Babenhauserheide, A., Zhang, H. F., Liu, Y., He, W., Chen, H., Masarie, K. A., Krol, M. C., and Peters, W. (2017). The carbontracker data assimilation shell (ctdas) v1.0: implementation and global carbon balance 2001–2015. *Geoscientific Model Development*, 10(7):2785–2800.

Van der Velde, I., Miller, J., Schaefer, K., Masarie, K., Denning, S., White, J., Tans, P., Krol, M., and Peters, W. (2013). Biosphere model simulations of interannual variability in terrestrial 13c/12c exchange. *Global Biogeochemical Cycles*, 27(3):637–649.

Van der Velde, I., Miller, J., Schaefer, K., Van Der Werf, G., Krol, M., and Peters, W. (2014). Terrestrial cycling of 13 co 2 by photosynthesis, respiration, and biomass burning in sibcasa. *Biogeosciences*, 11(23):6553–6571.

Vuichard, N. and Papale, D. (2015). Filling the gaps in meteorological continuous data measured at fluxnet sites with era-interim reanalysis. *Earth System Science Data*, 7(2):157–171.

Wang, S., Huang, C., Zhang, L., Lin, Y., Cen, Y., and Wu, T. (2016). Monitoring and assessing the 2012 drought in the great plains: analyzing satellite-retrieved solar-induced chlorophyll fluorescence, drought indices, and gross primary production. *Remote Sensing*, 8(2):61.

Welp, L. R., Keeling, R. F., Meijer, H. A., Bollenbacher, A. F., Piper, S. C., Yoshimura, K., Francey, R. J., Allison, C. E., and Wahlen, M. (2011). Interannual variability in the oxygen isotopes of atmospheric co 2 driven by el niño. *Nature*, 477(7366):579.

Xiao, X., Hollinger, D., Aber, J., Goltz, M., Davidson, E. A., Zhang, Q., and Moore III, B. (2004a). Satellite-based modeling of gross primary production in an evergreen needleleaf forest. *Remote sensing of environment*, 89(4):519–534.

Xiao, X., Zhang, Q., Braswell, B., Urbanski, S., Boles, S., Wofsy, S., Moore III, B., and Ojima, D. (2004b). Modeling gross primary production of temperate deciduous broadleaf forest using satellite images and climate data. *Remote Sensing of Environment*, 91(2):256–270.

Yang, J., Tian, H., Pan, S., Chen, G., Zhang, B., and Dangal, S. (2018). Amazon droughts and forest responses: Largely reduced forest photosynthesis but slightly increased canopy greenness during the extreme drought of 2015/2016. *Global change biology*.

Yebra, M., Van Dijk, A. I., Leuning, R., and Guerschman, J. P. (2015). Global vegetation gross primary production estimation using satellite-derived light-use efficiency and canopy conductance. *Remote sensing of environment*, 163:206–216.

Yoshida, Y., Joiner, J., Tucker, C., Berry, J., Lee, J.-E., Walker, G., Reichle, R., Koster, R., Lyapustin, A., and Wang, Y. (2015). The 2010 russian drought impact on satellite measurements of solar-induced chlorophyll fluorescence: Insights from modeling and comparisons with parameters derived from satellite reflectances. *Remote Sensing of Environment*, 166:163–177.

Yuan, W., Liu, S., Yu, G., Bonnefond, J.-M., Chen, J., Davis, K., Desai, A. R., Goldstein, A. H., Gianelle, D., Rossi, F., et al. (2010). Global estimates of evapotranspiration and gross primary production based on modis and global meteorology data. *Remote Sensing of Environment*, 114(7):1416–1431.

Zhang, Y., Joiner, J., Gentine, P., and Zhou, S. (2018). Reduced solar-induced chlorophyll fluorescence from gome-2 during amazon drought caused by dataset artifacts. *Global change biology*, 24(6):2229–2230.

Zhang, Y., Xiao, X., Wu, X., Zhou, S., Zhang, G., Qin, Y., and Dong, J. (2017). A global moderate resolution dataset of gross primary production of vegetation for 2000–2016. *Scientific data*, 4:170165.

Zhao, M., Heinsch, F. A., Nemani, R. R., and Running, S. W. (2005). Improvements of the modis terrestrial gross and net primary production global data set. *Remote sensing of Environment*, 95(2):164–176.

Chapter 12

Appendix

12.1 FLUXNET towers

Table 12.1: All FLUXNET towers used in this research, including location

sitename	countryid	latitude	longitude	name
Virasoro	Argentina	-28.239500	-56.188600	AR-Vir
San Luis	Argentina	-33.464800	66.459800	AR-SLu
Neustift/Stubai Valley	Austria	47.116669	11.317500	AT-Neu
Daly River Pasture	Australia	-14.063300	131.318100	AU-DaP
Daly River Savanna	Australia	-14.159283	131.388000	AU-DaS
Emerald	Australia	-23.858700	148.474600	AU-Emr
Fogg Dam	Australia	-12.545200	131.307200	AU-Fog
Great Western Woodlands, Western Australia, Au...	Australia	-30.191300	120.654100	AU-GWW
Howard Springs	Australia	-12.495200	131.150050	AU-How
Loxton	Australia	-34.470400	140.655100	AU-Lox
Red Dirt Melon Farm, Northern Territory	Australia	-14.563600	132.477600	AU-RDF
Robson Creek, Queensland, Australia	Australia	-17.117500	145.630100	AU-Rob
Tumbarumba	Australia	-35.656600	148.151600	AU-Tum
Wallaby Creek	Australia	-37.429000	145.187250	AU-Wac
Brasschaat (De Inslag Forest)	Belgium	51.309167	4.520556	BE-Bra
Lonzee	Belgium	50.551586	4.746130	BE-Lon
Vielsalm	Belgium	50.305068	5.998052	BE-Vie
Santarem-Km67-Primary Forest	Brazil	-2.856667	-54.958889	BR-Sa1
Santarem-Km83-Logged Forest	Brazil	-3.018029	-54.971435	BR-Sa3
ON-Groundhog River Mixedwood	Canada	48.217300	-82.155500	CA-Gro
MB-Northern Old Black Spruce	Canada	55.880000	-98.481000	CA-Man
UCI 1850	Canada	55.879167	-98.483889	CA-NS1
UCI 1930	Canada	55.905833	-98.524722	CA-NS2
UCI 1964	Canada	55.911667	-98.382222	CA-NS3
UCI 1964wet	Canada	55.911667	-98.382222	CA-NS4
UCI 1981	Canada	55.863056	-98.485000	CA-NS5
UCI 1989	Canada	55.916667	-98.964444	CA-NS6
UCI 1998	Canada	56.635833	-99.948333	CA-NS7
SK-Old Aspen	Canada	53.628890	-106.197790	CA-Oas
SK-Southern Old Black Spruce	Canada	53.987170	-105.117790	CA-Obs
QC-Eastern Old Black Spruce (EOBS)	Canada	49.692470	-74.342040	CA-Qfo
SK-1977 Fire	Canada	54.484950	-105.817350	CA-SF1
SK-1989 Fire	Canada	54.253920	-105.877500	CA-SF2
SK-1998 Fire	Canada	54.091560	-106.005260	CA-SF3
ON-Turkey Point 2002 White Pine	Canada	42.660936	-80.559519	CA-TP1
ON-Turkey Point 1989 White Pine	Canada	42.774419	-80.458775	CA-TP2
ON-Turkey Point 1974 White Pine	Canada	42.706811	-80.348314	CA-TP3
ON-Turkey Point 1939 White Pine	Canada	42.709778	-80.357400	CA-TP4

Continued on next page

Table 12.1: All FLUXNET towers used in this research, including location

sitename	countryid	latitude	longitude	name
Chamau grassland	Switzerland	47.210222	8.410444	CH-Cha
Fruebuel grassland	Switzerland	47.115833	8.537778	CH-Fru
Laegeren	Switzerland	47.478083	8.365000	CH-Lae
Oensingen1 grass	Switzerland	47.285833	7.731944	CH-Oe1
Oensingen2 crop	Switzerland	47.286306	7.734333	CH-Oe2
Davos- Seehorn forest	Switzerland	46.815333	9.855917	CH-Dav
Changbaishan	China	42.402500	128.095833	CN-Cha
Changlin	China	44.593400	123.509200	CN-Cng
Damxung	China	30.850000	91.083333	CN-Dan
Dinghushan	China	23.166667	112.533333	CN-Din
Duolun-grassland	China	42.046667	116.283611	CN-Du2
Haibei Alpine Tibet Site	China	37.370000	101.180000	CN-HaM
Siziwang Grazed (SZWG)	China	41.790200	111.897100	CN-Sw2
Qianyanzhou	China	26.733333	115.066667	CN-Qia
Bily Kriz- Beskidy Mountains	Czech Republic	49.502129	18.536860	CZ-BK1
Gebesee	Germany	51.100100	10.914300	DE-Geb
Grillenburg- grass station	Germany	50.949469	13.512525	DE-Gri
Hainich	Germany	51.079167	10.453000	DE-Hai
Klingenberg	Germany	50.892881	13.522506	DE-Kli
Leinefelde	Germany	51.328217	10.367800	DE-Lnf
Oberbarenburg	Germany	50.783617	13.719631	DE-Obe
Tharandt- Anchor Station	Germany	50.963611	13.566944	DE-Tha
Foulum	Denmark	56.484200	9.587220	DK-Fou
Llano de los Juanes	Spain	36.926594	-2.752115	ES-LJu
Hyytiala	Finland	61.847500	24.295000	FI-Hyy
Jokioinen agricultural field	Finland	60.898600	23.513450	FI-Jok
Sodankyla	Finland	67.361861	26.637833	FI-Sod
Le Bray (after 6/28/1998)	France	44.717110	-0.769300	FR-LBr
Fontainebleau	France	48.476402	2.780142	FR-Fon
Grignon (after 6/5/2005)	France	48.844220	1.951910	FR-Gri
Puechabon	France	43.741390	3.595833	FR-Pue
Guyaflux	French Guiana	5.278772	-52.924862	GF-Guy
Renon/Ritten (Bolzano)	Italy	46.586860	11.433690	IT-Ren
Borgo Cioffi	Italy	40.523750	14.957444	IT-BCi
Zerbolo-Parco Ticino- Canarazzo	Italy	45.200872	9.061039	IT-PT1
Castelporziano	Italy	41.705249	12.376106	IT-Cpz
Collelongo- Selva Piana	Italy	41.849360	13.588140	IT-Col
Tonzi Ranch	United States	38.431600	-120.966000	US-Ton
Lavarone (after 3/2002)	Italy	45.956200	11.281320	IT-Lav
Monte Bondone	Italy	46.014678	11.045831	IT-MBo
Roccarespampani1	Italy	42.408120	11.930010	IT-Ro1
Roccarespampani2	Italy	42.390260	11.920930	IT-Ro2
San Rossore	Italy	43.727861	10.284444	IT-SRo
Horstermeer	Netherlands	52.240350	5.071301	NL-Hor
Loobos	Netherlands	52.166581	5.743556	NL-Loo
Sardinilla Pasture	Panama	9.313780	-79.631430	PA-SPs
Sardinilla Plantation	Panama	9.318140	-79.634600	PA-SPn
Cherskii	Russia	68.613040	161.341430	RU-Che

Continued on next page

Table 12.1: All FLUXNET towers used in this research, including location

sitename	countryid	latitude	longitude	name
Chokurdakh / Kytalyk	Russia	70.829139	147.494278	RU-Cok
Fedorovskoje-drained spruce stand	Russia	56.461528	32.922083	RU-Fyo
Samoylov Island- Lena Delta	Russia	72.373800	126.495800	RU-Sam
Ubs Nur- Hakasija - grassland	Russia	54.725170	90.002150	RU-Ha1
Demokeya	Sudan	13.282900	30.478300	SD-Dem
Stordalen Forest- Mountain Birch	Sweden	68.354149	19.050333	SE-St1
Ivotuk	United States	68.486472	-155.748000	US-Ivo
Atqasuk	United States	70.469611	-157.408944	US-Atq
Walnut Gulch Kendall Grasslands	United States	31.736527	-109.941880	US-Wkg
Walnut Gulch Lucky Hills Shrubland	United States	31.743833	-110.052222	US-Whs
Blodgett Forest	United States	38.895250	-120.632750	US-Blo
GLEES	United States	41.364400	-106.239400	US-GLE
Goodwin Creek	United States	34.254700	-89.873500	US-Goo
Harvard Forest EMS Tower (HFR1)	United States	42.537756	-72.171478	US-Ha1
Fermi National Accelerator Laboratory - (Prair...	United States	41.840617	-88.241033	US-IB2
Kennedy Space Center (slash pine)	United States	28.458304	-80.670903	US-KS1
Kennedy Space Center (scrub oak)	United States	28.608577	-80.671534	US-KS2
ARM Southern Great Plains site	United States	36.605800	-97.488800	US-ARM
Little Washita Watershed	United States	34.960400	-97.978895	US-LWW
Lost Creek	United States	46.082680	-89.979190	US-Los
Mead - irrigated continuous maize site	United States	41.165056	-96.476638	US-Ne1
Mead - irrigated maize-soybean rotation site	United States	41.164871	-96.470100	US-Ne2
Mead - rainfed maize-soybean rotation site	United States	41.179667	-96.439646	US-Ne3
Metolius Eyerly Burn	United States	44.579400	-121.500000	US-Me1
Metolius Intermediate Pine	United States	44.452300	-121.557400	US-Me2
Metolius Second Young Pine	United States	44.315400	-121.607800	US-Me3
Metolius Old Pine	United States	44.499200	-121.622400	US-Me4
Metolius First Young Pine	United States	44.437189	-121.566756	US-Me5
Morgan Monroe State Forest	United States	39.323150	-86.413139	US-MMS
Niwot Ridge (LTER NWT1)	United States	40.032878	-105.546403	US-NR1
Sylvania Wilderness Area	United States	46.242017	-89.347650	US-Syv
Oak Openings	United States	41.554540	-83.843760	US-Oho
Park Falls	United States	45.945878	-90.272304	US-PFa
Santa Rita Mesquite	United States	31.821430	-110.866110	US-SRM
Univ. of Mich. Biological Station	United States	45.559840	-84.713820	US-UMB
Vaira Ranch	United States	38.406667	-120.950733	US-Var
Willow Creek	United States	45.805927	-90.079859	US-WCr
Mongo	Zambia	-15.437778	23.252778	ZM-Mon
Intermediate hardwood (IHW)	United States	46.730472	-91.232944	US-Wi1
Intermediate red pine (IRP)	United States	46.686889	-91.152833	US-Wi2
Mature hardwood (MHW)	United States	46.634722	-91.098667	US-Wi3
Mature red pine (MRP)	United States	46.739333	-91.166250	US-Wi4
Mixed young jack pine (MYJP)	United States	46.653083	-91.085806	US-Wi5
Pine barrens #1 (PB1)	United States	46.624889	-91.298222	US-Wi6
Wisconsin Red pine clearcut (RPCC)	United States	46.649111	-91.069278	US-Wi7
Young hardwood clearcut (YHW)	United States	46.722333	-91.252417	US-Wi8
Young Jack pine (YJP)	United States	46.618778	-91.081444	US-Wi9
Young red pine (YRP)	United States	46.618778	-91.081444	US-Wi0

Continued on next page

Table 12.1: All FLUXNET towers used in this research, including location

sitename	countryid	latitude	longitude	name
Tchizalamou	Congo - Kinshasa	-4.289167	11.656417	CG-Tch
Amoladeras	Spain	36.833608	-2.252318	ES-Amo
Bily Kriz- grassland	Czech Republic	49.494430	18.542850	CZ-BK2
CZECHWET	Czech Republic	49.024650	14.770350	CZ-wet
Sardinia/Arca di Noe	Italy	40.606130	8.151460	IT-Noe
ARM Southern Great Plains burn site	United States	35.549740	-98.040230	US-ARb
ARM Southern Great Plains control site	United States	35.546490	-98.040060	US-ARc
Ankasa	Ghana	5.268543	-2.694206	GH-Ank
Skukuza	South Africa	-25.019700	31.496900	ZA-Kru
Santa Rita Creosote	United States	31.908312	-110.839480	US-SRC
Corral Pocket	United States	38.090000	-109.390000	US-Cop
Haibei Shrubland	China	37.665278	101.331111	CN-Ha2
Metolius New Young Pine	United States	44.323200	-121.604300	US-Me6
Univ. of Mich. Biological Station Disturbance	United States	45.562500	-84.697500	US-UMd
Olentangy River Wetland Research Park	United States	40.020100	-83.018300	US-ORv
ARM USDA UNL OSU Woodward Switchgrass 1	United States	36.426700	-99.420000	US-AR1
ARM USDA UNL OSU Woodward Switchgrass 2	United States	36.635800	-99.597500	US-AR2
Mayberry Wetland	United States	38.049800	-121.765100	US-Myb
Twitchell Island	United States	38.105500	-121.652100	US-Twt
Adelaide River	Australia	-13.076900	131.117800	AU-Ade
Dry River	Australia	-15.258800	132.370600	AU-Dry
Sturt Plains	Australia	-17.150800	133.350300	AU-Stp
Wombat	Australia	-37.422200	144.094400	AU-Wom
Seto Mixed Forest Site	Japan	35.250000	137.066700	JP-SMF
Moshiri Birch Forest Site	Japan	44.384200	142.318600	JP-MBF
Duolun Degraded Meadow	China	42.055100	116.281000	CN-Du3
Nuuk Fen	Denmark	64.130833	-51.386111	DK-NuF
Zackenberg Fen	Denmark	74.481433	-20.554517	DK-ZaF
Zackenberg Heath	Denmark	74.473200	-20.550300	DK-ZaH
Anklam	Germany	53.866167	13.683417	DE-Akm
Selhausen	Germany	50.870623	6.449653	DE-Seh
Spreewald	Germany	51.892250	14.033690	DE-Spw
Zarnekow	Germany	53.875943	12.889010	DE-Zrk
Enghave	Denmark	55.690528	12.191750	DK-Eng
Laguna Seca	Spain	37.097936	-2.965833	ES-LgS
Lanjaron-Salvage logging	Spain	36.969502	-3.475819	ES-Ln2
Lettosuo	Finland	60.641833	23.959700	FI-Let
LompolojÃďnkkÃď	Finland	67.997200	24.209183	FI-Lom
Castel d'Asso1	Italy	42.380411	12.026561	IT-CA1
Castel d'Asso2	Italy	42.377219	12.026039	IT-CA2
Castel d'Asso 3	Italy	42.380000	12.022200	IT-CA3
Castelporziano2	Italy	41.704266	12.357293	IT-Cp2
Lavarone2	Italy	45.954200	11.285300	IT-La2
Torgnon	Italy	45.844440	7.578055	IT-Tor
Bayelva, Spitsbergen	Norway	78.921631	11.831085	NO-Blv
Seida/Vorkuta	Russia	67.054680	62.940468	RU-Vrk
Lackenberg	Germany	49.099617	13.304667	DE-Lkb
Rollesbroich	Germany	50.621914	6.304126	DE-RuR

Continued on next page

Table 12.1: All FLUXNET towers used in this research, including location

sitename	countryid	latitude	longitude	name
Selhausen Juelich	Germany	50.865912	6.447169	DE-RuS
Schechenfilz Nord	Germany	47.806389	11.327500	DE-SfN
Ispra ABC-IS	Italy	45.812643	8.633579	IT-Isp
San Rossore 2	Italy	43.732026	10.290954	IT-SR2
Adventdalen	Norway	78.186000	15.923000	NO-Adv
Spasskaya Pad larch	Russia	62.255000	129.168000	RU-SkP
Dahra	Senegal	15.402780	-15.432220	SN-Dhr
Alice Springs	Australia	-22.283000	133.249000	AU-ASM
Calperum	Australia	-34.002060	140.589120	AU-Cpr
Cumberland Plains	Australia	-33.613297	150.722466	AU-Cum
Gingin	Australia	-31.375000	115.650000	AU-Gin
Riggs Creek	Australia	-36.656000	145.576000	AU-Rig
Ti Tree East	Australia	-22.287000	133.640000	AU-TTE
Whroo	Australia	-36.673200	145.029400	AU-Whr
ON-Turkey Point Deciduous	Canada	42.635312	-80.557561	CA-TPD
Australia Yanco site	Australia	-34.988282	146.291606	AU-Ync
Pasoh Forest Reserve	Malaysia	2.973000	102.306200	MY-PSO
Tiksi	Russia	71.594267	128.887817	RU-Tks
Curtice Walter-Berger cropland	United States	41.628500	-83.347100	US-CRT
Poker Flat Research Range Black Spruce Forest	United States	65.123700	-147.487600	US-Prr
Santa Rita Grassland	United States	31.789400	-110.827700	US-SRG
Saratoga	United States	41.396600	-106.802400	US-Sta
Twitchell Wetland West Pond	United States	38.107400	-121.646900	US-Tw1
Twitchell Corn	United States	38.104700	-121.643300	US-Tw2
Twitchell Alfalfa	United States	38.115900	-121.646700	US-Tw3
Twitchell East End Wetland	United States	38.103000	-121.641400	US-Tw4
Winous Point North Marsh	United States	41.464600	-82.996200	US-WPT
GLEES Brooklyn Tower	United States	41.365800	-106.239700	US-GBT

12.2 SiBCASA variables

Table 12.2: SiBCASA output variables, used as input variables in the machine learned models in this thesis

LongName	ShortName
ventilation mass flux	ventmf
friction velocity	ustar
lower boundary temperature	gt
canopy temperature	tcan
evaporation	ev
snow depth on ground	snow_depth
snow on canopy	snow_can
runoff	runoff
leaf conductance	gl
stomatal conductance	gs
aerodynamic resistance	ra
bulk canopy boundary layer resistance	rb

Continued on next page

Table 12.2: SiBCASA output variables, used as input variables in the machine learned models in this thesis

LongName	ShortName
canopy resistance	rc
ground-canopy air space resistance	rd
bulk snow density	snow_bulk
active layer depth	d_active
frozen layer	d_freeze
fractional snow coverage	snow_area
mass of snow on ground	snow_mass
evapotranspiration	evt
sensible heat flux	hfss
latent heat flux	fws
canopy heat storage flux	chf
soil heat storage flux	shf
canopy sensible heat flux	hfc
soil surface sensible heat flux	hfg
canopy latent heat flux,transpiration	hfect
canopy latent heat flux, intercepted	hfeci
soil surface latent heat flux	hfegs
soil surface latent heat flux,intercepted	hfegi
canopy air space vapor pressure	ea
canopy air space temperature	ta
mixing layer vapor pressure	em
canopy air space relative humidity	hura
transfer fraction from storage pool to leaf pool	leaf_frac
transfer fraction from storage pool to root pool	root_frac
transfer fraction from storage pool to wood pool	wood_frac
number of snow layers	snow_nsl
ground liquid water store	wslg
canopy liquid water store	wslcan
total soil moisture	mrtsoil
canopy pco2	pco2ap
leaf surface pco2	pco2s
leaf internal pco2	pco2i
chloroplast pco2	pco2c
driver data-windspeed	spdmsib
driver data-pressure	pssib
driver data-largescale precipitation	dlsprsib
driver data-cumulus precipitation	dcuprsib
driver data-temp	tssib
driver data-potential temperature	tsib3
driver data-mixed lyr h2o mixing ratio	sh_sib
driver data-visible beam radiation	radvbc
driver data-visible diffuse ratiation	radvdc
driver data-nir beam radiation	radnbc
driver data-nir diffuse radiation	radndc
driver downwelling shortwave radiation	sw_dwn
driver data-longwave downward radiation	dlwbotsib
total live biomass	carb_live
total dead biomass	carb_dead

Continued on next page

Table 12.2: SiBCASA output variables, used as input variables in the machine learned models in this thesis

LongName	ShortName
total litter carbon	carb_litter
total soil carbon	carb_soil
above ground wood biomass	carb_awood
total carbon	carb_tot
turbulent flux out of canopy	c_flux
canopy net photosynthesis	assimn
gross primary productivity/photosynthesis	gpp
resp_tot - gpp	NEE_1
conductance-based carbon flux	NEE_2
net primary productivity	npp
ground respiration	resp_grnd
total respiration	resp_tot
autotrophic respiration	resp_auto
heterotrophic respiration	resp_het
humidity stress factor	rstfac1
water stress factor	rstfac2
temperature stress factor	rstfac3
boundary layer co2	pco2m
leaf area index from	zlt
leaf area index from prognostic leaf	LAI
chi squared stat between sim/obs	lai_chi_sqr
lai	lai_err
absorbed fraction of	fpar
resp_tot_13c - gpp_13c	NEE_13C
total 13c respiration	resp_tot13c
auto 13c respiration	resp_aut13c
het 13c respiration	resp_het13c
d13c	d13c_assimn
d13c canopy	d13cca
kiecps_c3	kiecps_c3
kiecps_c4	kiecps_c4
kiecps_tot	kiecps_tot
gpp c3	gpp_c3
gpp c4	gpp_c4
d13c c3	d13c_c3
d13c c4	d13c_c4
frac c3	frac_c3
frac c4	frac_c4
fire emission	fire
burned area frac	burned_area
fire 13c emission	fire_13c
net assimilated 13c c3	assn_13c_c3
net assimilated 12c c3	assn_12c_c3
net assimilated 13c c4	assn_13c_c4
net assimilated 12c c4	assn_12c_c4
frac times assimilation	kie*assn
sum assimilation	assn_sum
weighted fractionation	wkiecps

Continued on next page

Table 12.2: SiBCASA output variables, used as input variables in the machine learned models in this thesis

LongName	ShortName
c3 frac times assimilation	kie*assn_c3
c3 sum assimilation	assn_sum_c3
c3 weighted fractionation	wkiecps_c3
c4 frac times assimilation	kie*assn_c4
c4 sum assimilation	assn_sum_c4
c4 weighted fractionation	wkiecps_c4
net assimilated 13c	assn_13c
net assimilated 12c	assn_12c
canopy respiration	resp_can
13c canopy respiration	resp_can13c
d13c atmosphere reference	d13cm
assimilated 13c	ass_13c
assimilated 12c	ass_12c
fraction of water available to plants	pawfrac
vm	vm