university of groningen / faculty of mathematics and natural sciences / artificial intelligence

Improving Data Assimilation Approach for Estimating CO_2 Surface Fluxes Using ML

Master Thesis

H.M. Roothaert (s2929244)

November 25, 2022

Internal Supervisor: Dr. Celestine P. Lawrence (Artificial Intelligence, University of Groningen)
External Supervisors: Prof. Dr. Wouter Peters (University of Wageningen, The Netherlands)
Auke M. van der Woude (, MSc.) (University of Wageningen, The Netherlands)

Artificial Intelligence
University of Groningen, The Netherlands

Abstract

The environment is changing due to anthropogenic carbon emissions, and so is the carbon cycle regulating the exchange of CO_2 (i.e. fluxes) between the Earth's surface and the atmosphere. Measuring these changes is difficult, as it would require enormously dense observation networks to capture the strongly heterogeneous underlying flux-landscape. Through a combination of carbon exchange (CE) models and data assimilation (DA), the CarbonTracker data assimilation shell (CTDAS) generates a flux-landscape estimate which optimally matches the available observations. The current implementation of this DA approach is static; flux-landscape estimates produced in the past are not used for estimating new flux-landscapes. However, preliminary research has shown that seasonal, currently unused, patterns are present within the estimates of the DA approach. We propose three different methods for utilizing these patterns: a simple monthly mean model, a seasonal autoregressive integrated moving average (SARIMA) model, and a seasonal autoregressive integrated moving average with exogenous factors (SARIMAX) model. Preliminary results strongly indicate that the monthly mean model provides a substantial improvement over the current DA implementation once incorporated within CTDAS. In contrast, the SARIMA and SARIMAX models struggle to capture the non-stationary seasonal patterns.

Contents

Chapter 1: Introduction

Global warming is causing the glaciers to melt, seawater levels to rise and the weather to become more extreme, meaning more intense periods of rainfall and longer droughts. This is caused by the greenhouse effect, where greenhouse gases absorb infrared radiation and radiate it back to the Earth's surface. Some of the most prominent greenhouse gasses are carbon dioxide, methane, and water vapor. Although the increase in the concentrations of these gasses is not always caused by human activities, a slight change could disturb the current balance. Take for instance water vapor. On itself, a harmless gas that is omnipresent within the atmosphere. However, the higher the temperature of the atmosphere, the more water vapor it can hold without it falling down in some form of precipitation. The higher the water vapor concentration, the more infrared radiation is absorbed by it, resulting in a positive feedback loop where both temperature and water vapor concentration in the atmosphere keep increasing.
Examples like these show how a small change in the atmosphere can cause a snowball effect which would be difficult to stop. As such, the scientific community has urged policymakers to reduce the emissions of greenhouse gasses and eventually transition into a carbon-neutral economy (Intergovernmental Panel on Climate Change, 2021). The signing of the *Paris Agreement* (2015) is an answer to this call, as 195 countries expressed their commitment to a 50% reduction of greenhouse gas emissions by 2030 and neutrality by the second half of the century.
However, to determine the effectiveness of policies and track whether the parties that signed the Paris agreement are on track to reaching their goals, accurate models and validation methods on the emissions and distribution of greenhouse gasses are needed. The Global Carbon Project (GCP) was established to "work with the international science community to establish a common and mutually agreed on knowledge base to support policy debate and action to slow down and ultimately stop the increase of greenhouse gases in the atmosphere." (Poruschi, Dhakal, & Canadell, 2010). They aim to achieve this by focusing on the effect of anthropogenic activities on the global biogeochemical cycles which govern the emission of the three main greenhouse gasses: CO_2, CH_4 and N_2O.

1.1 Global Carbon Budget

One of the better-known biochemical cycles is the carbon cycle. The GCP aims to gain additional insights into the carbon cycle and share them across the scientific community using the annually updated Global Carbon Budget (GCB; (Le Quéré et al., 2015; Friedlingstein et al., 2020, 2021)). As the focus of the GCB is on the effect of human activities on the carbon cycle, their findings are expressed in the perturbation of the carbon cycle caused by anthropogenic activities as shown in Figure 1.1. Note how the transfer of carbon between the earth's surface and the atmosphere can be categorized into four different carbon sinks and sources: fossil fuel emissions ($\boldsymbol{F}^{\texttt{fossil}}$), forest fire emissions ($\boldsymbol{F}^{\texttt{fire}}$), biosphere uptake ($\boldsymbol{F}^{\texttt{bio}}$), and ocean uptake ($\boldsymbol{F}^{\texttt{ocean}}$). Measuring each of these components individually on a global scale is infeasible. Instead, carbon exchange (CE) models are used to estimate how these carbon sinks and sources interact with the atmosphere. The problem is that the used CE models are not perfect due to parameterizations and assumptions underlying

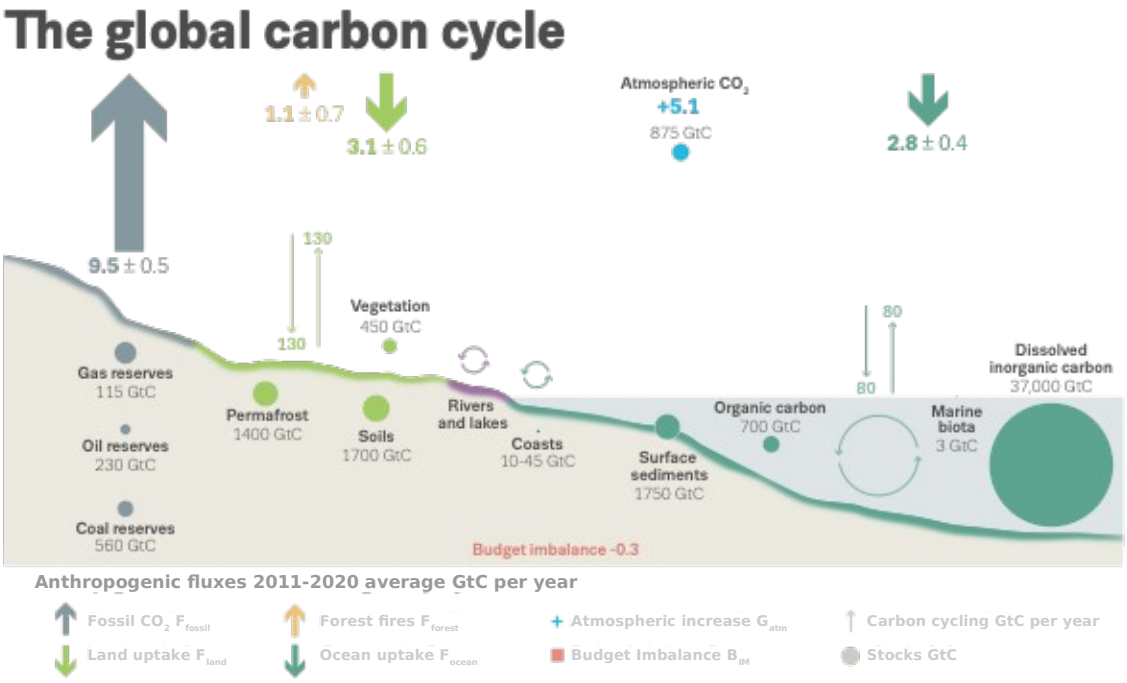


Figure 1.1: Schematic representation of the overall perturbation of the global carbon cycle caused by anthropogenic activities. See the legend for the corresponding arrows and units. Source: Friedlingstein et al. (2021)

the models. Intrinsic biases result in a *budget imbalance* ($B^{\mathtt{im}}$), where the measured atmospheric concentrations do not match the sum of the CE models. For the remainder of this thesis, $B^{\mathtt{im}}$ is defined as

$$B^{\mathtt{im}} = \boldsymbol{F}^{\mathtt{fire}} + \boldsymbol{F}^{\mathtt{fossil}} + \boldsymbol{F}^{\mathtt{bio}} + \boldsymbol{F}^{\mathtt{ocean}} - G^{\mathtt{atm}}, \tag{1}$$

where $G^{\mathtt{atm}}$ is the measured gain in atmospheric CO_2 (for more information, see Appendix A.1). One of the aims of the GCB is to increase our knowledge of the carbon cycle and thereby minimize $B^{\mathtt{im}}$.

One way of minimizing $B^{\mathtt{im}}$ is to scale the surface fluxes through a process called atmospheric inversion. A system handling such inversions consists of 4 main components:

1. *Prior surface fluxes*: These are the non-optimized fluxes, generally originating from CE models.

2. *Observation set*: This is a set of measured atmospheric concentrations which is used to validate the outcome of the atmospheric inversion.

3. *Transport model*: A transport model transports the surface fluxes to the atmospheric concentrations such that they can be compared to the observations.

4. *Optimization*: The optimization step combines all of the other three components by scaling the prior fluxes such that after transporting the fluxes, they optimally match the observations.

The exact implementation of an atmospheric inversion system depends mainly on the used optimization technique. However, an intuitive interpretation of the process is to view it as a way of finding a set of hyperparameters that, once applied to the CE models, result in a model which optimally matches the available observations.

One might think that the concept of atmospheric inversion is simply a matter of fitting models to observation data. Solely trying to find the best fit to observation data could result in an overfitted model, which is unable to make predictions on regions and/or time points not covered within the observation set. Connecting conclusions to such an inversion model would therefore be difficult. While in a sense this is indeed the case, the fitting procedure is in some regions highly constrained by measurements (Europe, US) and in some regions much less (Tropics, Siberia) (see Figure 1.2). The regions with a dense measurement network will be mainly informed by observations in a DA system, whereas regions with fewer measurements, such as the tropics, are mainly informed by the prior. The resulting reanalysis of multiple decades of surface CO_2 fluxes is thus a mixture of data-driven and prior-driven results. Knowing that these reanalyses are not perfect, and depend on the transport model used, as well as the prior fluxes and the prescribed error structure, conclusions about the carbon cycle are generally taken not from a single, but multiple inverse models (Jones et al., 2021a; Schuh et al., 2019; Gaubert et al., 2019; Friedlingstein et al., 2020, 2021; Hauck et al., 2020; Petetin et al., 2021). By combining CO_2 flux estimations of multiple models, the ensemble becomes more robust to overfitting and thus a trend/bias in the ensemble is a stronger indication of a potential bias than a trend/bias in a single model. The usage of such an ensemble, in combination with the omission of some of the data (specifically aircraft data, being a sensitive indicator of model vertical transport) while fitting the models and using the latter for validation of the results, adds to the overall validity of the ensemble. Moreover, the inverse method is the only available view on the GCB that uses atmospheric observations of CO_2 increases to constrain fluxes. This integral constraint posed by mass-conservation of all exchange, as well as highly accurate long-term measurements, is arguably the best knowledge available on the forcing of climate by CO_2. Hence, atmospheric inversion is a powerful tool for refining the GCB.

1.2 CarbonTracker Europe

One of these inversion projects is the CarbonTracker Europe (CTE) project, developed by Wageningen University - Meteorology and Air Quality Department (Peters et al., 2010). Originally forked from CarbonTracker in 2008 (Peters et al., 2007), CTE has released an annual/biennial version since 2015, where the focus lies on improving the weekly estimates of global CO_2 surface fluxes. These optimized fluxes have been used to refine the GCB since 2015 (Le Quéré et al., 2015). Given the expertise of Prof. Dr. Wouter Peters and Auke M. van der Woude, who are both actively working on

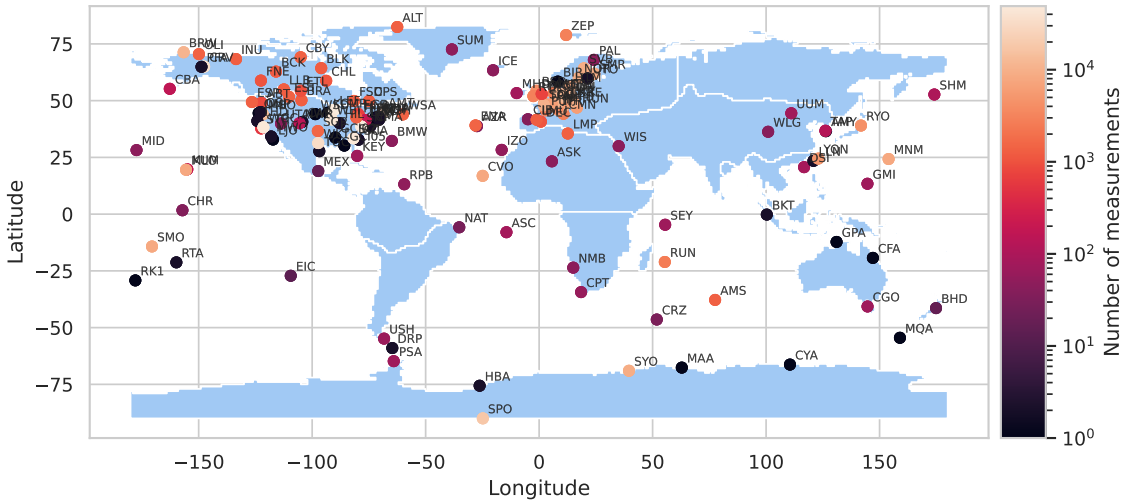


Figure 1.2: Location of the CO_2 measuring stations with the ObsPack dataset (Cox et al., 2021). The color indicates the number of measurements taken at each station between 2000 and 2021, ranging from 1 measurement at the BIR station in Europe to 49051 measurements taken at the SNP station in the North American temperate region. The combined number of measurements across all 147 measuring stations is 617090. The light-blue area indicates the 11 distinct TransCom regions defined by (Gurney et al., 2003), which are discussed in more detail in Appendix A.2

CTE, this thesis revolves around improving the functionality of CTE. Before discussing how CTE could potentially be improved, an explanation of the current system is given.

1.2.1 The ensemble Kalman filter

The algorithm used for minimizing the budget imbalance as defined in Equation 1 within CTE, is an implementation of the square-root ensemble Kalman filter (Whitaker & Hamill, 2002), further referred to as the ensemble Kalman filter (EKF). Similar to a regular Kalman filter, the EKF uses data assimilation (DA) to estimate some hidden state $\boldsymbol{\lambda} \in \mathbb{R}^s$, where $s \in \mathbb{N}$ is the number of elements within the state vector. The exact definition of $\boldsymbol{\lambda}$ is given in Appendix A.2. However, for understanding its function within the EKF, it suffices to interpret $\boldsymbol{\lambda}$ as a representation of the hidden biases within the CE models. As a result, an arbitrary flux landscape $\boldsymbol{F} \in \mathbb{R}^{360\times180}$ relates to a state vector $\boldsymbol{\lambda}$ as

$$\boldsymbol{F} = \boldsymbol{F}^{\texttt{prior}} \odot \mathcal{K}(\boldsymbol{\lambda}), \tag{2}$$

where $\odot$ is the Hadamard product, $\boldsymbol{F}^{\texttt{prior}} = \boldsymbol{F}^{\texttt{bio}} + \boldsymbol{F}^{\texttt{ocean}} + \boldsymbol{F}^{\texttt{fossil}} + \boldsymbol{F}^{\texttt{fire}} \in \mathbb{R}^{360\times180}$, and $\mathcal{K}: \mathbb{R}^s \to \mathbb{R}^{360\times180}$ is an operator which maps the elements within $\boldsymbol{\lambda}$ to a 1×1 degree global grid.

In essence, the EKF is a function that updates some *background* state vector $\boldsymbol{\lambda}^b$ into an *analyzed* state vector $\boldsymbol{\lambda}^a$. This $\boldsymbol{\lambda}^a$ results in surface fluxes $\boldsymbol{F}^a$ which are optimally consistent with some set of observations $\boldsymbol{y}^\circ \in \mathbb{R}^m$, where m is the number of observations. Or in other words, a set of fluxes

that causes the B^{im} defined in Equation 1 to be minimized. The remainder of this section will only discuss the three main components of the EKF: the cost function, the optimization strategy, and its state and covariance forecast model. For an in-depth explanation, see Peters et al. (2005).

Cost function Obtaining a $\boldsymbol{\lambda}^a$ such that B^{im} is minimized is done by minimizing a cost function $J(\boldsymbol{\lambda})$. This cost function has to capture two separate components: A penalty for the mismatch between the observations and the model estimates ($J^{\text{obs}}(\boldsymbol{\lambda})$) and a penalty for adjustments made to the original state ($J^{\text{state}}(\boldsymbol{\lambda})$). The need for J^{obs} speaks for itself; The difference between the model estimates and observations should be minimized. The need for J^{state} is less straightforward. This term is included to prevent the EKF from finding values for the $\boldsymbol{\lambda}$ which are physically unlikely, and is, therefore, used as a regularization term of the cost function. By including a term that penalizes large differences between $\boldsymbol{\lambda}^b$ and $\boldsymbol{\lambda}^a$, the optimized model produced by the EKF is guided towards the prior model as closely as possible. In turn, this makes it unlikely that the EKF moves too far away from a model which is considered to be physically plausible.

Within the most recent iteration of CTE, CTE2018, both components have been captured using the following formulas:

$$J^{\text{obs}}(\boldsymbol{\lambda}) = (\boldsymbol{y}^\circ - \mathcal{H}(\boldsymbol{\lambda}))^T \boldsymbol{R}^{-1} (\boldsymbol{y}^\circ - \mathcal{H}(\boldsymbol{\lambda})) \tag{3}$$

$$J^{\text{state}}(\boldsymbol{\lambda}) = (\boldsymbol{\lambda} - \boldsymbol{\lambda}^b)^T \boldsymbol{P}^{-1} (\boldsymbol{\lambda} - \boldsymbol{\lambda}^b), \tag{4}$$

where $\boldsymbol{R} \in \mathbb{R}^{m \times m}$ and $\boldsymbol{P} \in \mathbb{R}^{s \times s}$ are covariance matrices of $\boldsymbol{y}^\circ$ and $\boldsymbol{\lambda}$ respectively and $\mathcal{H}(\boldsymbol{\lambda}) : \mathbb{R}^s \to \mathbb{R}^m$ is the transport operator of the state vector, defined as

$$\mathcal{H}(\boldsymbol{\lambda}) = \mathcal{T}(\boldsymbol{F}^{\text{prior}} \odot \mathcal{K}(\boldsymbol{\lambda})) = \mathcal{T}(\boldsymbol{F}). \tag{5}$$

$\mathcal{T} : \mathbb{R}^{360 \times 180} \to \mathbb{R}^m$ is the transport model previously mentioned in Section 1.1, which transports m tracers to atmospheric concentrations. This step is needed to determine the difference between $\boldsymbol{y}^\circ$ and an arbitrary flux landscape $\boldsymbol{F}$ and in extension $\boldsymbol{\lambda}$. We note that in the current EKF, the transport model is the most expensive computationally as it requires close to 6 weeks of time on 50 CPUs to evaluate all $\boldsymbol{\lambda}$'s. This cost precludes many other minimization methods to be applied since they require multiple and fast state evaluations of (iterative) state solutions.

Combining Equation 3 and 4 yields the final cost function:

$$J(\boldsymbol{\lambda}) = J^{\text{obs}}(\boldsymbol{\lambda}) + J^{\text{state}}(\boldsymbol{\lambda}) = (\boldsymbol{y}^\circ - \mathcal{H}(\boldsymbol{\lambda}))^T \boldsymbol{R}^{-1} (\boldsymbol{y}^\circ - \mathcal{H}(\boldsymbol{\lambda})) + (\boldsymbol{\lambda} - \boldsymbol{\lambda}^b)^T \boldsymbol{P}^{-1} (\boldsymbol{\lambda} - \boldsymbol{\lambda}^b) \tag{6}$$

Optimization As explained above, B^{im} is minimized by minimizing the cost function J defined in Equation 6. By setting the derivative J' to 0, a minimum is found. Using calculus, the state vector $\boldsymbol{\lambda}$ and its covariance $\boldsymbol{P}$ for which $J' = 0$ can be shown (Tarantola (2005), as cited in Peters et al. (2005)) to be:

$$\boldsymbol{\lambda}_t^a = \boldsymbol{\lambda}_t^b + \boldsymbol{K}(\boldsymbol{y}_t^\circ - \mathcal{H}(\boldsymbol{\lambda}_t^b)) \tag{7}$$

$$\boldsymbol{P}_t^a = (\boldsymbol{I} - \boldsymbol{K}\boldsymbol{H})\boldsymbol{P}_t^b \tag{8}$$

in which t is the subscript for time, $\boldsymbol{H} \in \mathbb{R}^{m \times s}$ is the linearized matrix form of observation operator $\mathcal{H}$, and $\boldsymbol{K} \in \mathbb{R}^{s \times m}$ is the Kalman gain defined as:

$$\boldsymbol{K} = (\boldsymbol{P}_t^b \boldsymbol{H}^T)(\boldsymbol{H}\boldsymbol{P}_t^b \boldsymbol{H}^T + \boldsymbol{R})^{-1}. \tag{9}$$

State and covariance forecast model One powerful feature of a regular Kalman filter is its ability to combine observations with a state transition model to create an optimal estimate of some hidden state. This transition model $\mathcal{M}$ represents how the hidden state $\boldsymbol{\lambda}$, along with its covariance matrix $\boldsymbol{P}$, changes from t to $t+1$:

$$\boldsymbol{\lambda}^b{}_{t+1} = \mathcal{M}(\boldsymbol{\lambda}_t^a) \tag{10}$$

$$\boldsymbol{P}_{t+1}^b = \boldsymbol{M}\boldsymbol{P}_t^a \boldsymbol{M}^T + \boldsymbol{Q}, \tag{11}$$

where $\boldsymbol{M} \in \mathbb{R}^{s \times s}$ is linearized matrix of $\mathcal{M}$ and $\boldsymbol{Q} \in \mathbb{R}^{s \times s}$ represents the noise introduced by an imperfect transition model.

However, such a transition model does not yet exist. As stated at the start of this section, $\boldsymbol{\lambda}$ represents the biases within the combined CE models. The issue is that these biases at time t are independent of the biases at $t-1$. Instead, the biases depend on environmental conditions such as temperature and precipitation within the biosphere. These conditions often vary on a daily basis, while $\boldsymbol{\lambda}$ is determined on a weekly basis. This makes defining an accurate transition model based solely on $\boldsymbol{\lambda}_{t-1}$ impossible.

The original implementation of the EKF simply uses the identity matrix as a transition function, i.e.

$$\mathcal{M} = \boldsymbol{I} \tag{12}$$

since it is reasonable to assume that the environmental conditions at $t-1$ do not differ substantially from the conditions at t. Within CTE2018, this simple model has been replaced by a smoother over three time steps:

$$\mathcal{M}^{\texttt{smoothed}}(\boldsymbol{\lambda}^a, t) = (\boldsymbol{\lambda}^a{}_{t-1} + \boldsymbol{\lambda}^a{}_t + 1)/3 \tag{13}$$

where t is the subscript for the time in weeks.

1.2.2 Possible improvement of the ensemble Kalman filter

While Equation 13 is an improvement over Equation 12, the transition function remains relatively uninformed. Only information from the two previous time steps is used. There is however evidence for a pattern, and thus information, within the set of all previously analyzed state vectors which is currently not used. One of the most striking examples is the seasonal pattern within the scaling factor of two eco-regions with some of the biggest carbon fluxes on this planet; the northern taiga within the North American boreal region and crops within the Eurasia temperate region. Figure 1.3 shows the average scaling factor, along with the standard deviation, of these regions aggregated by month. The figure clearly shows that the analyzed scaling factor is consistently below 1 during the winter months of the northern hemisphere, meaning that the combined CE models overestimate the carbon flux within this area during this period.

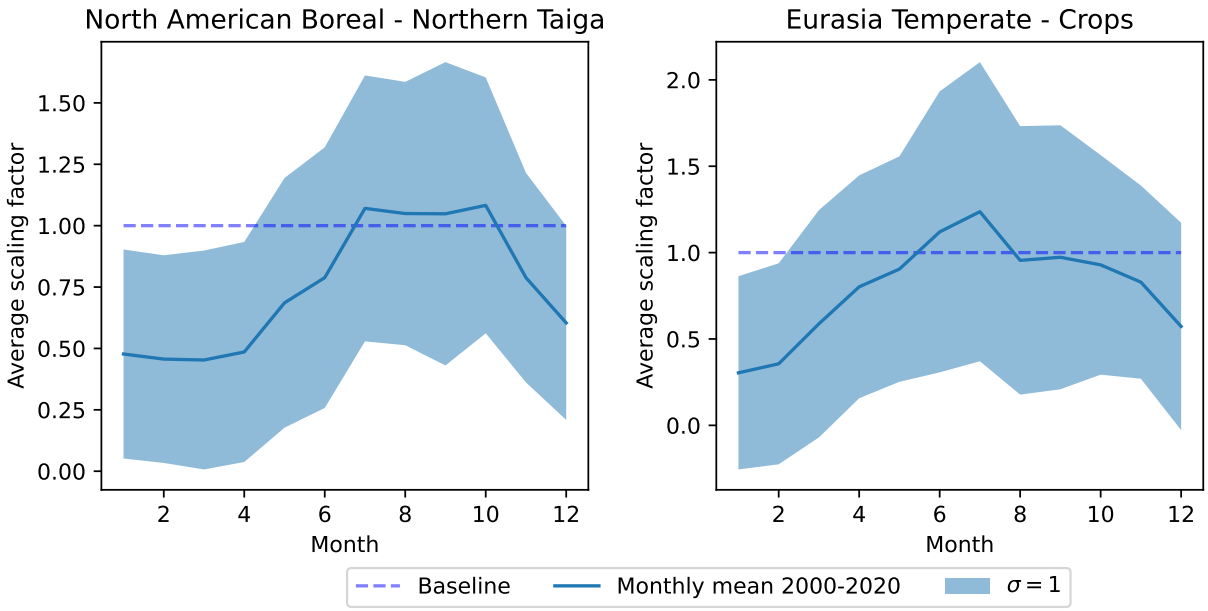


Figure 1.3: Both figures show an example of the mean of an analyzed scaling factor aggregated over a TransCom region and PFT. The σ is based on the average scaling factor per week, meaning each month has an N ranging between 85 and 94.

Integrating information such as the example depicted above within $\boldsymbol{\lambda}^b$ would result in a $\boldsymbol{\lambda}^b$ which is closer to the true biases within the CE models. This affects the cost function depicted in Equation 6 in two ways. The most intuitive effect is that fewer adjustments are needed to go from $\boldsymbol{\lambda}^b$ to $\boldsymbol{\lambda}^a$, reducing $J^{\mathtt{state}}(\boldsymbol{\lambda}^a)$. But more importantly, $J^{\mathtt{obs}}$ is reduced as well. A bias within $\boldsymbol{\lambda}^b_t$ could be propagated through the solution and affect $\boldsymbol{\lambda}^a$ at times $t+5$, $t+6$, or even $t+20$. This is a known issue of the time-stepping approach as CO_2 is conserved within the atmosphere. Another potential negative effect of biases within $\boldsymbol{\lambda}^b$, is that the prior covariance $\boldsymbol{P}^b$ needs to be inflated to ensure that the analyzed state vector $\boldsymbol{\lambda}^a$ is within the expected error of $\boldsymbol{P}^b$ or, alternatively, $\boldsymbol{R}$ needs to be increased to keep $\boldsymbol{y}^\circ - \mathcal{H}(\boldsymbol{\lambda}^a)$ within the expected mole fraction uncertainty. An unbiased $\boldsymbol{\lambda}^b$ would therefore result in a $\boldsymbol{\lambda}^a$ providing a closer resemblance to the observations $\boldsymbol{y}^\circ$ and thus reduce $J^{\mathtt{obs}}(\boldsymbol{\lambda}^a)$.

This forms the theoretical basis for how an improved model for $\boldsymbol{\lambda}^b$ would improve the functionality of the EKF. As a result, more confidence can be put into the modeled fluxes, allowing for more informed policy-making on curbing the effect of anthropogenic CO_2 emissions.

1.2.3 Used data

Each CE model used for determining the budget imbalance $B^{\mathtt{im}}$ defined in equation 1 is determined using a different model. However, the characteristics of the carbon fluxes estimated by these models

vary greatly. Fossil fuel emissions are for instance relatively constant with a strong recurring seasonal pattern, while the carbon fluxes within the biosphere are strongly influenced by meteorological conditions such as droughts (Smith et al., 2020). This influences the role each CE model has within the atmospheric inversion process.

Biosphere fluxes The prior biosphere fluxes ($\boldsymbol{F}^{\texttt{bio}}$) are taken from the Simple Biosphere model 4 (SiB4; (Haynes et al., 2019)). The estimates on biosphere exchange are based on meteorological data and therefore, the resulting fluxes vary greatly depending on the meteorological conditions. Because of this variability, the certainty attached to these fluxes is considerably less compared to the other CE models. This is the reason why the majority of elements within the state vector are associated with fluxes from the biosphere model (9805 out of 9835).

Ocean fluxes Ocean fluxes ($\boldsymbol{F}^{\texttt{ocean}}$) are taken from Jacobson et al. (2007b), which is based on a global ocean carbon flux inversion. Instead of using atmospheric measurements, carbon measurements taken from the ocean are used for this inversion. Compared to the biosphere fluxes, the ocean fluxes are relatively small and constant. As such, the confidence within these fluxes is substantially higher. However, as more than 70% of the earth's surface is covered by oceans, ocean carbon fluxes play a major role in the carbon cycle. Therefore, a few elements within the state vector (30 out of 9835) are used to scale these fluxes.

Forest fire emissions Forest fire emissions ($\boldsymbol{F}^{\texttt{fire}}$) are taken from Global Fire Assimilation System (GFAS). The GFAS fire emissions are based on satellite retrievals of heat and are generated using Copernicus Atmosphere Monitoring Service Information (CAMS) 2022. The carbon fluxes from this model are considered to be reliable enough to not require any further optimization. As such, no elements within the state vector are assigned to scale the fluxes from this model.

Fossil fuel emissions Fossil fuel emissions ($\boldsymbol{F}^{\texttt{fossil}}$) are taken from Gridded Fossil Emissions Dataset (GridFED; (Jones et al., 2021b)). GridFED includes gridded fossil fuel emission estimates based on the reported energy consumption statistics of countries. Fossil fuel emissions show a strong and predictable seasonal pattern, which allows for higher confidence to be put in this model. Therefore, this model is also not optimized within the atmospheric inversion process and no elements within the state vector are associated with carbon fluxes originating from this model.

Observations In order to estimate the gain in atmospheric CO_2 concentrations ($G^{\texttt{atm}}$), a set of real observations ($\boldsymbol{y}^{\circ}$) is used. These observations are taken from the Observation Package (ObsPack; (Cox et al., 2021)) dataset, which combines measurements processed by 463 laboratories across the globe. The distribution of these observations is shown in Figure 1.2. Satellite measurements are not included, as these measurements are associated with higher uncertainty.

1.3 Background literature

Several examples can be found where atmospheric inversion systems have been optimized. However, these efforts focused either on setting up a new DA system (Peters et al., 2005, 2007), increasing its resolution (van der Laan-Luijkx et al., 2017), or improving the data used by the system (Chevallier et al., 2009). To the author's knowledge, no literature is available on the subject of finding a transition model for the inversion system. There is however great interest in combining the field of earth systems science and machine learning in order to gain new insight into the often chaotic behavior of earth systems (Reichstein et al., 2019; Reichstein, Camps-Valls, Tuia, & Xiang Zhu, 2021b). Furthermore, the recently published book by Camps-Valls, Xiang Zhu, Tuia, and Reichstein (2021) provides plenty of inspiration on how to use ML, with most emphasis being put on DL, for problems often encountered within earth sciences. The field of atmospheric inversion is unfortunately not that well represented within that book. This is a shame as, over the past few years, sufficient data has become available to analyze the behavior of these inversion techniques and potentially correct their errors. As such, additional literature is provided that focuses on improving atmospheric inversions using ML.

Bastrikov et al. (2018) compared a genetic algorithm (GA) to a gradient descent approach to find the optimal parameter setting of seven PFTs. The main evaluation metric was the RMSE reduction between the prior and the posterior flux estimate. It was found that the GA performed just as well as, if not better, than the gradient descent approach while having similar computational costs. Therefore, it might be possible to use a GA for fitting the state vector as well. One downside of this approach would be the requirement of multiple forward runs of the flux transport, which is one of the most expensive parts of the atmospheric inversion process. For example, to estimate 20 years of $\boldsymbol{F}^{\mathtt{bio}}$ and $\boldsymbol{F}^{\mathtt{ocean}}$ with CTDAS requires 6 weeks of wall-clock time using 48 processors on the national supercomputer, with the atmospheric transport model TM5 consuming close to 90% of the resources.

Jung et al. (2020) discusses the validation of the FLUXCOM approach, which uses an ensemble of 9 different ML techniques to scale FLUXNET fluxes using atmospheric inversion-based techniques. The technical details of the machine learning implementation are discussed in another paper (Tramontana et al., 2016). What is discussed in this paper is how models using different approaches can be compared to each other, in a sense creating an objective benchmark for novel approaches. Several metrics are used for this comparison, including the inter-annual variability of the net ecosystem exchange (NEE), seasonal variation of the NEE, and spatial distribution of the mean annual gross primary production (GPP). Each of these metrics could be used alongside the more commonly used validation methods of masking a set of observations and comparing the fit of the estimated fluxes to aircraft measurements.

Another paper by Bonavita and Laloyaux (2020) used an artificial neural network (ANN) to infer and correct model error of numerical weather prediction (NWP) and climate prediction conducted with general circulation models. They state that one of the assumptions made by the Kalman filter and variations of the Kalman filter (i.e. the EKF), is that the model error is equal to 0. This effectively makes any Kalman filter-based approach blind to the presence of systematic model errors (Dee (2005), as cited in Bonavita and Laloyaux (2020)). Such systematic model errors are often referred

to as biases and can vary in space, time, and prevalent meteorological conditions. Correcting these biases is therefore important for the EKF to function optimally. One of the main evaluation criteria of the new transition function should therefore be its ability to reduce the prior model bias. A similar line of reasoning has been applied by Crespi, Petitta, Marson, Viel, and Grigis (2021), who adjusted monthly quantities from the seasonal forecasting system SEAS5 of the European Centre for Medium-Range Weather Forecasts (ECMWF). Furthermore, (Tramontana et al., 2016) used the model bias as one of their evaluation metrics as well.

Other applications of ML in earth sciences worth mentioning are the correction of O_3 forecasts using a gradient boosting machine based on a random decision forest (Petetin et al., 2021), the usage ANNs to estimate turbulent flows within large-eddy simulations (Stoffer et al., 2021) or partitioning FLUXNET fluxes into respiration and photosynthesis fluxes (Tramontana et al., 2020), or how differential equations bound by physical constraints can be solved using deep learning (Raissi, Perdikaris, & Karniadakis, 2017).

1.4 Thesis structure and goals

As mentioned within Section 1.2.2, this thesis revolves around improving the transition model of the EKF currently used within the CTE project. A clear seasonal trend provides the first evidence that this transition model can be improved by better utilizing previously analyzed state vectors. Further improvements could be achieved by including auxiliary meteorological information with ML to fully utilize the dependencies that exist between meteorological conditions and biases within the biosphere model. However, two hurdles need to be addressed before a ML implementation can be designed. The first one is the lack of a reliable evaluation metric for the proposed transition models. This problem is the main focal point within the first part of the thesis. The second problem is how to structure the data such that any correlations within the state vector and auxiliary meteorological data can easily be extracted by a ML implementation. Each of these hurdles is discussed in a separate part, along with a set of research questions.

Setting the baseline The first part of this thesis revolves around setting a baseline. Before such a baseline can be set, it needs to be clear how each model will be evaluated. It is unfeasible to perform a full inversion run for every model evaluation, and thus a less complex performance measure is needed. One approach would be to train the model on a set of previously analyzed state vectors and try to make a forecast model of this time series. However, the relation between the state vector and resulting atmospheric carbon concentrations after transporting the state vector is highly non-linear. A small change in the state vector could drastically affect the resulting flux landscape and in turn the estimated atmospheric concentrations. It is therefore uncertain whether evaluating a forecast model on its ability to forecast a state vector would provide a representative measure of its ability to minimize the bias within the prior fluxes. Therefore, the first question this thesis aims to answer is:

«*In which evaluation space is the performance of the transition model most likely to generalize to the performance within a full inversion run?*»

Finding an answer to this question requires several transition models which can be compared to each other. Before moving on to more complicated models, some baseline models are needed. This provides an opportunity to test the effect of utilizing the seasonal patterns within the state vector shown in Figure 1.3. This can be achieved rather easily by using a monthly average of the analyzed state vector as a background state vector. This results in a sub-question:

«*Would using the seasonal pattern found within the analyzed state vector (see Figure 1.3) result in a transition function producing a background state vector closer to the true state vector?*»

Comparing ML models Once a baseline has been created, it is possible to compare more complex models to this baseline. Instead of immediately delving into large and powerful deep-learning methods, Part II focuses on a stochastic time-series modeling approach. Various arguments are given in Part II as to why this stochastic modeling approach suffices as a starting point, but the main argument is to see whether the usage dependencies on various time scales could provide an improvement over the simpler transition models introduced in Part I. More precisely, Part II aims to answer the question:

«*Does the utilization of additional temporal dependencies (i.e. dependencies between time-steps) within the state vector result in a reduction of the mean bias of the prior biosphere and ocean flux with respect to the monthly average model?*»

The transitional transition model only used previous states to estimate the next state. There is however evidence that the biosphere, and thus also the biases within the biosphere, are affected by meteorological conditions such as droughts (Smith et al., 2020; van der Laan-Luijkx et al., 2015). Instead of using the temporal dependencies, it might be more effective to directly use the environmental conditions as predictor variables. These environmental conditions are shown to affect the biosphere model (Smith et al., 2020; van der Laan-Luijkx et al., 2015), and are therefore likely to affect the biases within the biosphere model as well. Therefore, the second question investigated in Part II is:

«*Could the utilization of predictor variables (i.e. temperature, precipitation), in combination with the temporal dependencies, result in a reduction in the mean bias of the prior biosphere and ocean flux with respect to the monthly average model?*»

Before these two questions can be answered, the data needs to be pre-processed. One of the main problems is the noise within the analyzed state vectors. As such, Part II also discusses the aggregation methods used to minimize the noise within the analyzed state vector, while also reducing its dimensionality. Additionally, the environmental variables are processed according to findings derived from a literature study.

Additional information As the target audience of this thesis consists of both atmospheric science and machine learning researchers, some jargon might not be as familiar to some as it is to others. Therefore, this thesis has a glossary explaining most of the used jargon, as well as all acronyms and mathematical objects. Finally, all code used for this project can be found at the GitHub repository of this project. Furthermore, slides from my colloquium might provide a different approach to explaining the used atmospheric inversion system. Next to being uploaded to the GitHub repository, the slides can also be downloaded through this link.

Part I
Setting the Baseline

As stated in Section 1.4, this thesis revolves around improving the transition model $\mathcal{M}$ currently used within the CTE project. This project uses the CarbonTracker data assimilation shell (CTDAS) as the implementation of the inversion process. Section 1.2.1 discussed some of the shortcomings and showed an indication of how the current implementation could be improved. This part of the thesis focuses on building a simple alternative to the current approach and its evaluation. By doing so, a baseline is constructed to which the more complex alternatives discussed in Part II can be compared.

This part starts with defining the used transition models in Section 2, followed by the experimental setup in Section 3. Section 4 discusses the results of the used transition models. However, while analyzing the results, several problems became apparent. Therefore the iterative process which has led to the final evaluation method is depicted in this section as well. The final section discusses the conclusions which can be drawn from the results and sets the baseline evaluation methods used for Part II.

Chapter 2: Methods

Before delving into the more complex machine learning methods for creating transition models, simpler methods are needed for constructing a baseline. The chosen set of simple models consists of the `prior`, `smoothed`, `monthly`, and `analyzed` models. These models are a set of state vectors generated using different methods. Each method is discussed below.

Prior The *prior* model is the result of not scaling any of the prior fluxes $\boldsymbol{F}^{\texttt{prior}}$. In other words, the transition function $\mathcal{M}^{\texttt{prior}}$ (see Equation 10) is as simple as

$$\mathcal{M}^{\texttt{prior}} = \mathbf{1}, \tag{14}$$

where $\mathbf{1} \in \mathbb{N}^{9835}$ is a vector of ones. Therefore, the *prior* model provides a 'worst case' comparison, as any effective scaling of the fluxes should improve the resulting flux landscape.

Smoothed The *smoothed* model ($\mathcal{M}^{\texttt{smoothed}}$) is the transition model currently used within the contribution of CTE to the GCP 2020 (Friedlingstein et al., 2020) and has been defined in Equation 13. It essentially smooths the analyzed state vector of the previous two time steps to produce the new background state vector.

Monthly To test the effect using the longer temporal patterns within the state vector shown in Figure 1.3, the *monthly* model $\mathcal{M}^{\texttt{monthly}}$ uses the monthly average of the analyzed state vector as the background state vector. To provide a baseline for the evaluation of more complex models, the implementation of this model has been kept relatively simple. Instead of a dynamic model, $\mathcal{M}^{\texttt{monthly}}$ is a static model which uses the first 19 years of analyzed state vectors ($\boldsymbol{\Lambda}^{\texttt{train}} = [\boldsymbol{\lambda}^a_0, \boldsymbol{\lambda}^a_1, ..., \boldsymbol{\lambda}^a_{t_0}] \in \mathbb{R}^{992\times 9835}$, where t_0 is the last week of 2018) for determining the average per month. The final two years ($\boldsymbol{\Lambda}^{test} = [\boldsymbol{\lambda}^a_{t_0}, \boldsymbol{\lambda}^a_{t_0+1}, ..., \boldsymbol{\lambda}^a_{t_1}] \in \mathbb{R}^{104\times 9835}$, where t_1 is the last week of 2020) are used for validation. This results in a model

$$\mathcal{M}^{\texttt{monthly}}(t) = \frac{1}{|\boldsymbol{\Lambda}^{\texttt{train}}_m|} \cdot \sum_{\boldsymbol{\lambda}^a \in \boldsymbol{\Lambda}^{\texttt{train}}_m} \boldsymbol{\lambda}^a \Bigg|_{\texttt{month}(t)=m}, \tag{15}$$

where t is the time in weeks, $m \in [Jan, Feb, ..., Dec]$ is one of the 12 months and $\boldsymbol{\Lambda}^{\texttt{train}}_m$ is the set of all $\boldsymbol{\lambda}^a_t \in \boldsymbol{\Lambda}^{\texttt{train}}$ where $\texttt{month}(t) = m$. $|\boldsymbol{\Lambda}^{\texttt{train}}_m|$ is the number of elements within $\boldsymbol{\Lambda}^{\texttt{train}}_m$. This function might look intimidating, but all it does is group the $\boldsymbol{\Lambda}^{\texttt{train}}$ by month and determines the average afterward. Do note that this implementation limits the possibility of validating the model as only the state vectors from the final two years can be used for an independent evaluation.

Analyzed The `analyzed` model $\mathcal{M}^{\texttt{analyzed}}$ simply returns the optimized state vector of the next time step:

$$\mathcal{M}^{\texttt{analyzed}}(\boldsymbol{\lambda}_t^a) = \boldsymbol{\lambda}_{t+1}^a \tag{16}$$

Note that this is only possible after having done a full inversion run. For this specific model, the $\mathcal{M}^{\texttt{smoothed}}$ transition was used to correct biases within the Simple Biosphere model 4 (SiB4) biosphere model.

This full inversion run serves as the target data under testing conditions in which the fluxes cannot be transported to atmospheric concentrations. Intuitively, a model trained on previously analyzed state vectors should perform worse than $\mathcal{M}^{\texttt{analyzed}}$ as those models are not fitted to observations while $\mathcal{M}^{\texttt{analyzed}}$ is.

Chapter 3: Experimental Setup

The used data consists of two separate datasets: 21 years of weekly $\boldsymbol{\lambda}^a$ ($\boldsymbol{\Lambda}^a$) and 2 years of observations ($\boldsymbol{y}^\circ$).

Weekly analyzed state vectors The set $\boldsymbol{\Lambda}^a$ consists of 1096 $\boldsymbol{\lambda}^a$s from the period 2000 to 2020 and was generated by a full inversion run of the CTE contribution to the GCP 2020 (Friedlingstein et al., 2020). This set has been divided into a training set ($\boldsymbol{\Lambda}^{\texttt{train}}$) and a testing set ($\boldsymbol{\Lambda}^{\texttt{test}}$) by putting the final two years aside, resulting in the sets $\boldsymbol{\Lambda}^{train} \in \mathbb{R}^{992\times 9835}$ and $\boldsymbol{\Lambda}^{\texttt{test}} \in \mathbb{R}^{104\times 9835}$ respectively.

Observations The set of observations $\boldsymbol{y}^\circ$ is a collection of 608659 measurements taken in 2019 and 2020 across the set of measuring stations M, where $|M| = 146$ (Cox et al., 2021). The distribution of the measuring stations and observations is shown in Figure 1.2. These observations were generated using various measuring techniques, flask and in situ, and altitudes, ranging from surface measurements to aircraft measurements taken at $1.3 * 10^4$m. Note that as the prior state vector prediction step occurs before the EKF optimization step, the prior state vector can also be evaluated based on its fit to observations later used by the EKF.

Chapter 4: Results

While constructing the baseline to which all newly implemented methods could be compared, several issues arose regarding the generalisability of the results. Since the different models can be compared to each other at different points within the inversion pipeline, the main question that needs to be answered is whether a good model fit at an early stage (i.e. modeled state vector vs. optimized state vector) generalizes to an improved inversion. Performing a complete inversion run for every proposed model is infeasible due to the costs of running this model (close to 6 weeks of time on 50 CPUs to evaluate all $\boldsymbol{\lambda}$'s.) Therefore it would be ideal if the model could be evaluated earlier within the pipeline, minimizing the computational costs.

Ideally, the model can be evaluated at the lowest level possible: *modeled state vector versus optimized state vector*. However, this would require the assumption that approximating the optimized state vector will result in a prior flux closely resembling the 'true' flux. This assumption cannot be made without providing additional evidence. Hence, intermediate steps are included such that any findings from the *modeled state vector versus observations* comparison can be tracked along the inversion pipeline. If the same qualitative properties of the comparison hold across the various steps, more confidence can be put in an evaluation procedure at those steps. A flowchart of the decision process and the chosen intermediate steps is shown in Figure 4.1.

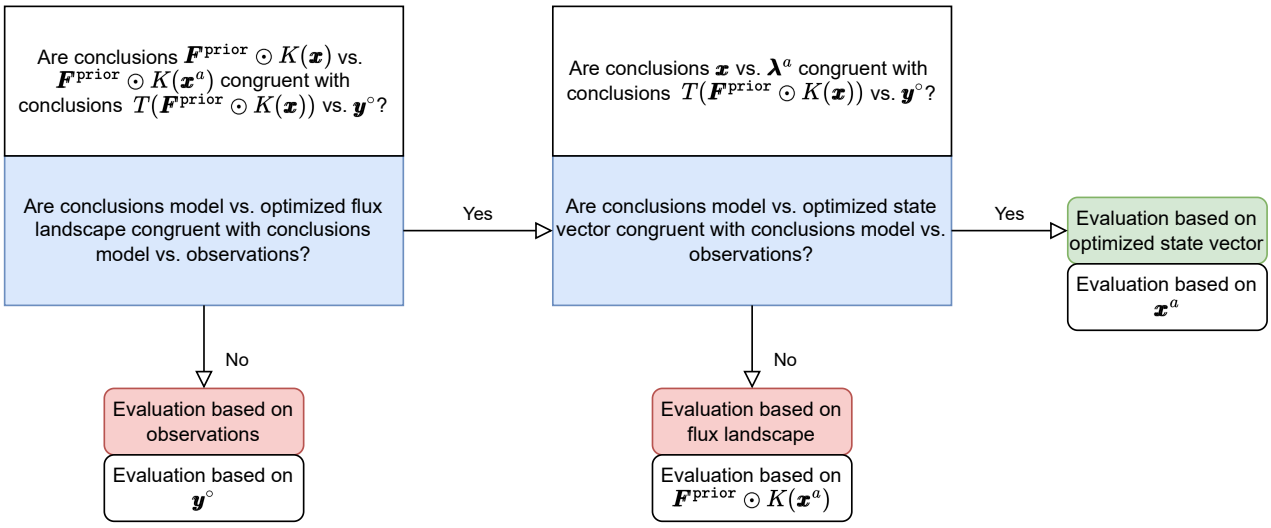


Figure 4.1: Flowchart of the decision process resulting in the final evaluation. By analyzing the results at intermediate steps, the point within the inversion pipeline can be determined at which optimization would result in an improved prior model. By verifying at each step if the conclusions from the comparison between $\mathcal{T}(\boldsymbol{F}^{\text{prior}} \odot \mathcal{K}(\boldsymbol{\lambda}))$ and $\boldsymbol{y}^\circ$ still hold, a cutoff point can be determined at which findings from the model no longer generalize to an improved flux model. If the conclusions from $\mathcal{T}(\boldsymbol{F}^{\text{prior}} \odot \mathcal{K}(\boldsymbol{\lambda}))$ versus $\boldsymbol{y}^\circ$ also hold for $\boldsymbol{\lambda}$ versus $\boldsymbol{\lambda}^a$, it is reasonable to evaluate intermediate models based on how well they fit $\boldsymbol{\lambda}^a$.

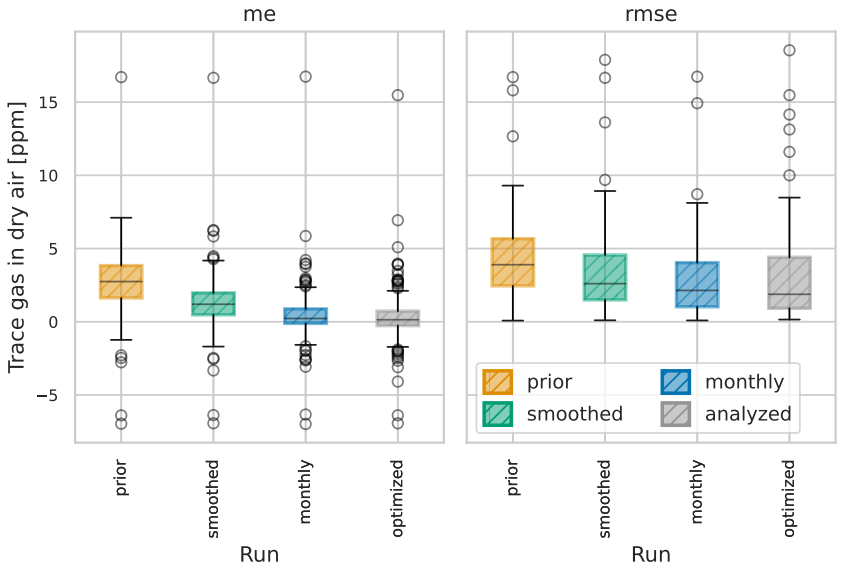


Figure 4.2: Boxplots showing the distribution of root-mean-square-error and mean error of individual measuring stations over the period 2019-2020.

4.1 Model versus observations

The first step is to evaluate the models discussed in Section 2 based on how well the transported state vector $\mathcal{T}(\boldsymbol{F}^{\texttt{prior}} \odot \mathcal{K}(\boldsymbol{\lambda}))$ matches the observations $\boldsymbol{y}^{\circ}$. This evaluation is based on how well the model is able to capture the concentrations observed at each site shown in Figure 1.2. Both the root-mean-square error (RMSE) and the mean error (ME) were determined at these sites. The RMSE serves as a measure of the variance of the error, while the ME represents the average bias at each measuring station. Ideally, both the RMSE and the ME are similar to each other and close to 0 as this would imply the model can consistently make accurate predictions about the measured CO_2 concentrations. The results are shown in Figure 4.2.

Figure 4.2 shows that the `prior` model performs the worst with the mean of the ME being 2.65 [ppm] and the mean RMSE being 4.27 [ppm]. The `smoothed` model provides a considerable improvement with a mean ME and RMSE of 1.27 [ppm] and 3.40 [ppm] respectively. The final two models, `monthly` and `optimized`, show similar results. Their mean ME values are 0.41 and 0.34, whereas their mean RMSE values are 2.86 and 3.13 respectively. This ranking is considered to be the main qualitative property of the *modeled state vector versus observations* comparison and will serve as a reference for the remaining comparisons.

4.2 Model versus optimized flux landscape

As the transportation of the state vector is costly, there would be a preference for skipping this step during intermediate evaluations. The earliest point within the inversion pipeline would be at the point of the analyzed flux landscape $\boldsymbol{F}^a = \boldsymbol{F}^{\texttt{prior}} \odot \mathcal{K}(\boldsymbol{\lambda}^a)$(see Equation 2). One potential problem is that this comparison is based on $\boldsymbol{\lambda}^a$. This $\boldsymbol{\lambda}^a$ is a product of the EKF and can therefore contain biases. Furthermore, some areas of the resulting flux landscape are noisy due to those areas being poorly constrained by observations. As a result, it is possible that the results from the comparison between $\boldsymbol{F}^{\texttt{prior}} \odot \mathcal{K}(\boldsymbol{\lambda})$ and $\boldsymbol{F}^{\texttt{prior}} \odot \mathcal{K}(\boldsymbol{\lambda}^a)$ do not match the qualitative results found in Section 4.1.

If the comparison between models in flux space shows the same qualitative results depicted in Figure 4.2 (i.e. `monthly` is better than `smoothed`, which is better than `prior`, which closely resembles `optimized`), it would entail that it is reasonable to evaluate intermediate models in flux space instead of directly comparing them to observations. Figure 4.3 shows that this is indeed the case. Note that $\boldsymbol{F}^{\texttt{monthly}}$ is consistently closest to $\boldsymbol{F}^{\texttt{analysed}}$, with the mean difference of only -0.07 $[\text{PgC} \cdot \text{yr}^{-1}]$. The differences between $\boldsymbol{F}^{\texttt{smoothed}}$ and $\boldsymbol{F}^{\texttt{prior}}$, and $\boldsymbol{F}^{\texttt{analysed}}$ are considerably higher with the mean differences being 1.14 $[\text{PgC} \cdot \text{yr}^{-1}]$ and 3.11 $[\text{PgC} \cdot \text{yr}^{-1}]$ respectively. As such, the qualitative results of the analysis in observation space are also present in the analysis in flux space. Another interesting feature to note is the seasonal pattern of a large difference between the models in the winter/spring weeks (i.e. weeks 0-20, 45 to 70, and 95-104) and a small difference within the summer/fall weeks (i.e. weeks 20-45 and 70-95).

4.3 Model versus optimized state vector

While evaluating within flux space is less complex compared to evaluating within observation space, the models discussed within Section 2 produce a set of state vectors. If it is possible to compare two sets of state vectors and get similar qualitative results as the evaluation within the observation space, intermediate evaluations would be even simpler. However, to be able to do so requires the assumption that findings from the state space generalize to flux space and eventually observation space. The previous section showed that the latter step, from flux space to observation space holds relatively well, as the ranking of the models in flux space was identical to the ranking in observation space. However, going from state space to flux space brings new challenges.

A challenge that arises is the difference in the size of the fluxes associated with the elements within the state vector. Some elements within the state vector apply to small regions with a small flux such as a single 1×1 degree grid cell of a desert at the Iberian Peninsula or apply to regions with a large flux, such as all tropical trees within the South American tropical TransCom region. To get a feel of the scale at which the values might differ, Figure 4.4 shows the distribution of the total ecosystem respiration (TER) associated with each element within the state vector. Here, the TER is used as a proxy for the scale of the total fluxes one could expect within a region over an entire year. The figure shows that for the majority of the elements within the state vector, the TER is close to 1 $[\text{mmol} \cdot \text{m}^{-2} \cdot \text{s}^{-1}]$, while for some the TER is as high as $3.587 \cdot 10^3$ $[\text{mmol} \cdot \text{m}^{-2} \cdot \text{s}^{-1}]$. As a result,

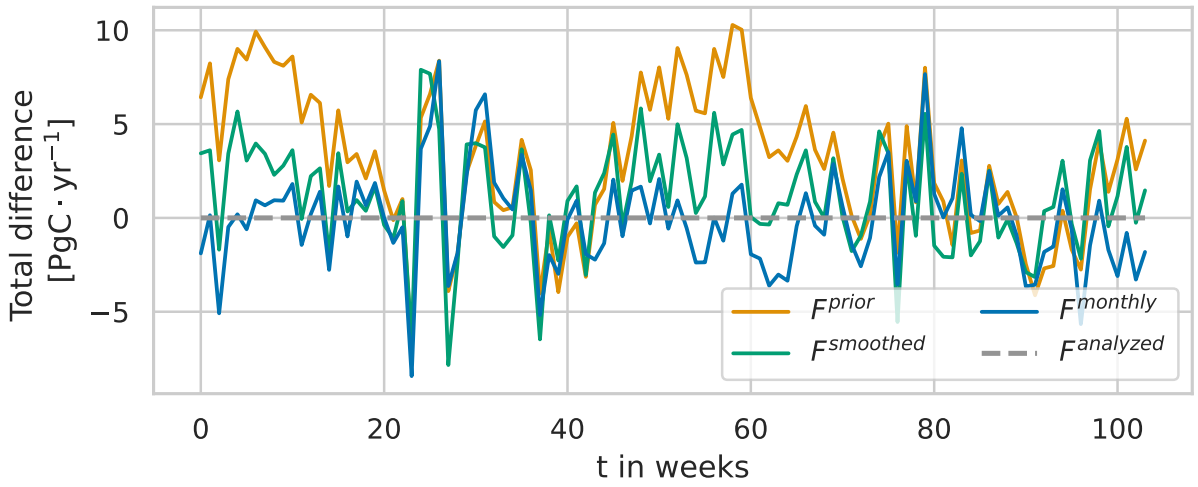


Figure 4.3: Time series of the weekly global difference between optimized inversion model $\boldsymbol{F}^{\texttt{analysed}}$ and prior models $\boldsymbol{F}^{\texttt{prior}}$, $\boldsymbol{F}^{\texttt{monthly}}$ and $\boldsymbol{F}^{\texttt{smoothed}}$ between 2019 and 2020. Results were obtained by subtracting $\boldsymbol{F}^{\texttt{analysed}}$ from the models $\boldsymbol{F}^{\texttt{prior}}$, $\boldsymbol{F}^{\texttt{monthly}}$ and $\boldsymbol{F}^{\texttt{smoothed}}$ and taking the global sum of fluxes for each week. The dotted line serves as the target given the aim of approximating $\boldsymbol{F}^{\texttt{analysed}}$.

a small error on an element within the state vector associated with a large TER could have a large effect on flux space. And vice versa, a large error on an element with a small TER would have a minimal effect in flux space.

The comparison between the different models in state vector space is shown in Figure 4.5. What it shows, is a similar pattern to the one observed in Figure 4.3, where $\boldsymbol{\lambda}^{\texttt{prior}}$ consistently has the highest difference, with a mean difference of 0.141, followed by $\boldsymbol{\lambda}^{\texttt{smoothed}}$ and $\boldsymbol{\lambda}^{\texttt{monthly}}$ with mean differences of 0.046 and -0.036 respectively. Also, the seasonal pattern of a large difference within the winter/spring weeks and a small difference within the summer/fall weeks is visible.

There is however one key difference between the analysis in observation space and the one in state vector space. This is best explained by table 4.1, where the averages from Figures 4.2, 4.3 and 4.5 have been summarized. Within flux space, the same substantial difference between the `smoothed` and `prior` model found in observation space is present, as well as the difference between `smoothed` and `monthly`. However, this improvement from `smoothed` to `monthly` is not visible in state vector space as the absolute mean differences are almost equal, being 0.046 and 0.036 respectively. This is a clear example where qualitative properties of the model present in the observation space, are not represented similarly in the state vector space. As such, any optimization effort executed in state vector space is not guaranteed to propagate to flux and observation space.

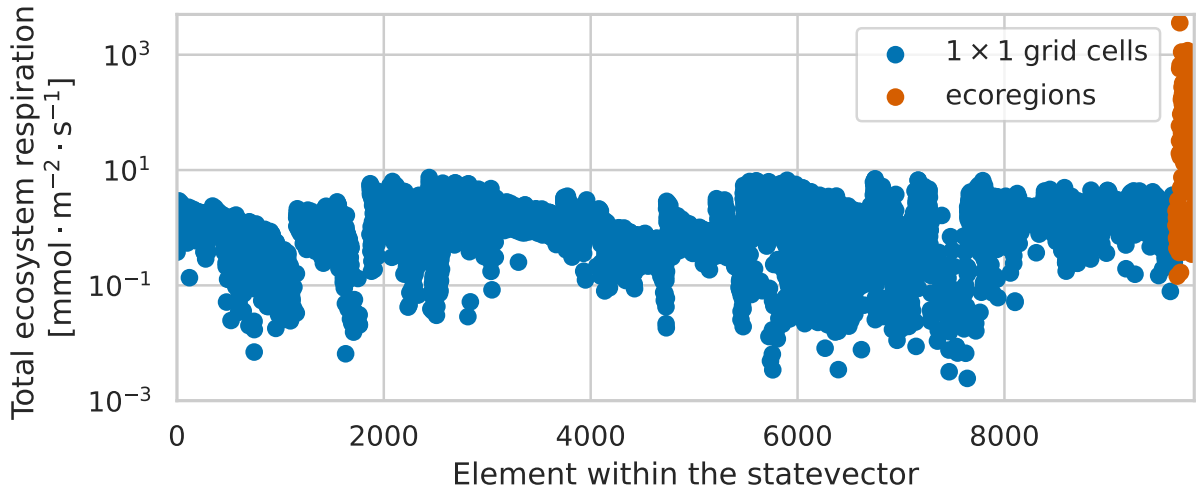


Figure 4.4: Distribution of the TER values associated with the elements of the state vector. Within this distribution, elements representing 1×1 degree grid cells are indicated in blue, whereas elements representing entire ecoregions are indicated in red. Notice how the elements associated with entire ecoregions scale fluxes ranging from 0.1 to 3000 $[\text{mmol} \cdot \text{m}^{-2} \cdot \text{s}^{-1}]$ while the elements associated with gridded cells scale fluxes ranging from 0.0001 to 10 $[\text{mmol} \cdot \text{m}^{-2} \cdot \text{s}^{-1}]$

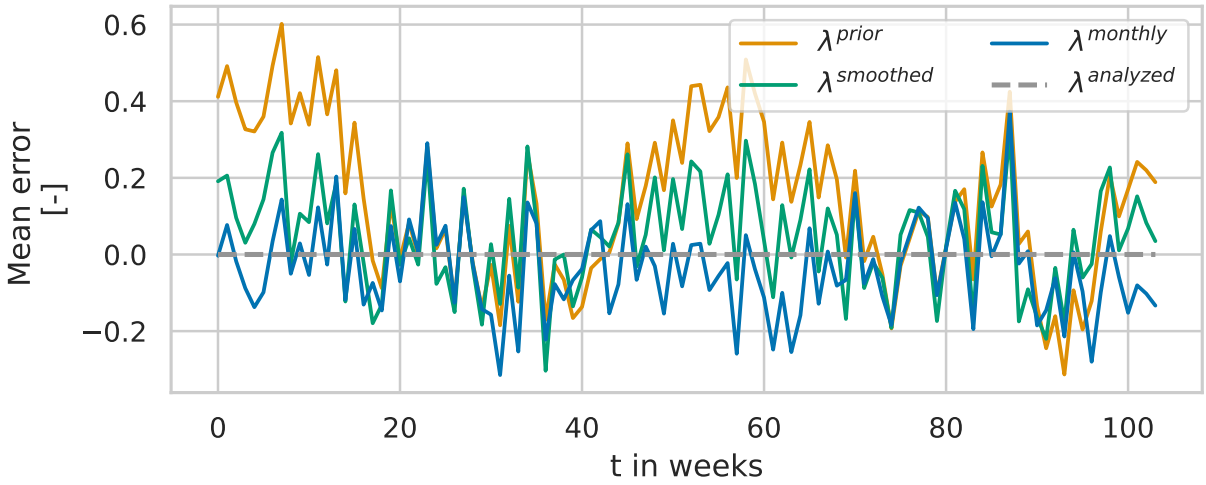


Figure 4.5: Time series of the mean difference between the elements within prior state vectors $\boldsymbol{\lambda}^{\texttt{prior}}$, $\boldsymbol{\lambda}^{\texttt{monthly}}$ and $\boldsymbol{\lambda}^{\texttt{smoothed}}$, and optimized state vector $\boldsymbol{\lambda}^{\texttt{optimized}}$. Results were obtained by subtracting $\boldsymbol{\lambda}^{\texttt{optimized}}$ from the models $\boldsymbol{\lambda}^{\texttt{prior}}$, $\boldsymbol{\lambda}^{\texttt{monthly}}$ and $\boldsymbol{\lambda}^{\texttt{smoothed}}$ and taking the average of the resulting vector. The dotted line serves as the target given the aim of approximating $\boldsymbol{\lambda}^{\texttt{optimized}}$.

Model	**Observation space** mean($\mathcal{H}(\boldsymbol{\lambda}^{\texttt{model}}) - \boldsymbol{y}^{\circ}$) $N = 146$, [ppm]	**Flux space** mean($\boldsymbol{F}^{\texttt{model}} - \boldsymbol{F}^{\texttt{analyzed}}$) $N = 104$, [PgC $\cdot$ yr^{-1}]	**state vector space** mean($\boldsymbol{\lambda}^{\texttt{model}} - \boldsymbol{\lambda}^{\texttt{analyzed}}$) $N = 104$, [$-$]
`prior`	2.65	3.11	0.141
`smoothed`	1.28	1.14	0.046
`monthly`	0.42	-0.07	-0.036
`optimized`	0.35	0.00	0.000

Table 4.1: Summary of the mean differences between the models in observation, flux, and state vector space and their respective target variable.

Chapter 5: Discussion and Recommendations

Within this research, the `monthly` model is proposed as an alternative to the transition model currently used within CTDAS. Figure 4.2 shows that the `monthly` model substantially reduces the mean error between the transported state vector and the observations, compared to the `smoothed` model. Furthermore, the root-mean-square error between the transported state vector and observations is reduced as well, although this reduction is less prominently visible.

By evaluating the model at different steps within the inversion pipeline, it has been shown that the step at which the model is evaluated influences the conclusions that can be drawn from the evaluation. Therefore, it is important to carefully think about which step is best suited for evaluation purposes. The two conflicting goals are; matching the observations as closely as possible while maintaining a sufficiently simple evaluation process. This trade-off has resulted in the decision process depicted in Figure 5.1.

Note that the used reasoning only indicates that the qualitative qualities of the results in the observations space are also present in the results of the flux space. As only four models were evaluated, no hard conclusions can be drawn from their comparison. To be able to do so, would require a deeper analysis of the dependencies between the various evaluation spaces. This is, especially for the dependencies between the observation and flux space, a complex task. The transport model $\boldsymbol{H}$ used for going from flux to observation space is highly non-linear and the available observations vary greatly in their frequency, accuracy, and spatial distribution.

5.1 Discussion on the evaluation

One step that could be taken to improve the comparison between the different evaluation spaces, is the inclusion of a variance measure of the error. This has not been included in the current analysis of the flux space as not all $1x1$ degree grid cells which define the flux space contribute equally to the total error between the model and $\boldsymbol{F}^{\texttt{analysed}}$. Cells near the equator cover a larger surface area and are therefore more likely to be associated with a larger flux, compared to cells near the poles. If all cells were to have equal weight in the determination of the variance, the error or the cells near the poles would artificially reduce the variance, while cells near the equator would increase the variance. Therefore, some normalization would be needed to provide an unbiased variance measure.

For the final evaluation of the model, additional performance measures more often used within the evaluation of inversion methods should be included (Jung et al., 2020; Tramontana et al., 2016). These include the inter-annual variability of the net ecosystem exchange (NEE), seasonal variation of the NEE, and spatial distribution of the mean annual gross primary production (GPP). To maintain an unbiased judgment, these evaluation methods are reserved for only the final evaluation.

5.2 Discussion of the methods

In future iterations of the project, the `monthly` model can be made a dynamic model with just only a few minor adjustments. Instead of using the first 19 years of state vectors for determining the

average for the final two years, a running average can be used, which is updated after a new state vector has been analyzed.

Besides improving the analysis by including the variance, adding additional models would add to the validity of the comparison as well. One simple, but possibly effective model, would be a combination of the `smoothed` and `monthly` model. This model would be more resilient to abnormal weather patterns as shorter temporal dependencies are included in the determination of the background state vector. It is reasonable to assume that if there was a heatwave that affected the state vector at t, the same heatwave could affect the state vector at $t+1$ in a similar fashion. This effect is currently not captured within the `monthly` model. Even though the `monthly` model is performing well as it is, including these shorter temporal dependencies might bring it the resulting background state vector even closer to the observations after transportation to atmospheric concentrations.

A final model worth investigating is another variant of the `monthly` model in which the weight of the analyzed state vectors used for determining the average is varied according to the recency of the analyzed state vector. Due to anthropogenic greenhouse gas emissions, the climate is rapidly changing. This is likely to affect the biases within the biosphere and ocean carbon flux models and in extension the analyzed state vector. By applying a decaying term to earlier analyzed state vectors, the average is less affected by state vectors that were fitted to an environment that is no longer representative of the current environment. Furthermore, the quantity of atmospheric CO_2 measurements increases over time (see supplementary Figure B.1). Especially with the new CO2M satellite data becoming available (Sierk, Bézy, Löscher, & Meijer, 2019), this trend is likely to continue. More observations result in a more constrained and hence more accurate analyzed state vector, providing another argument on why later analyzed state vectors should receive a higher weight when determining the monthly average.

As the main purpose of this part of the thesis is to construct a baseline evaluation method, the improvements proposed above have not been implemented. However, as the `monthly` model provides a substantial improvement over the currently used `smoothed` model, future research could focus on improving the `monthly` model discussed in this section while considering the suggestions made.

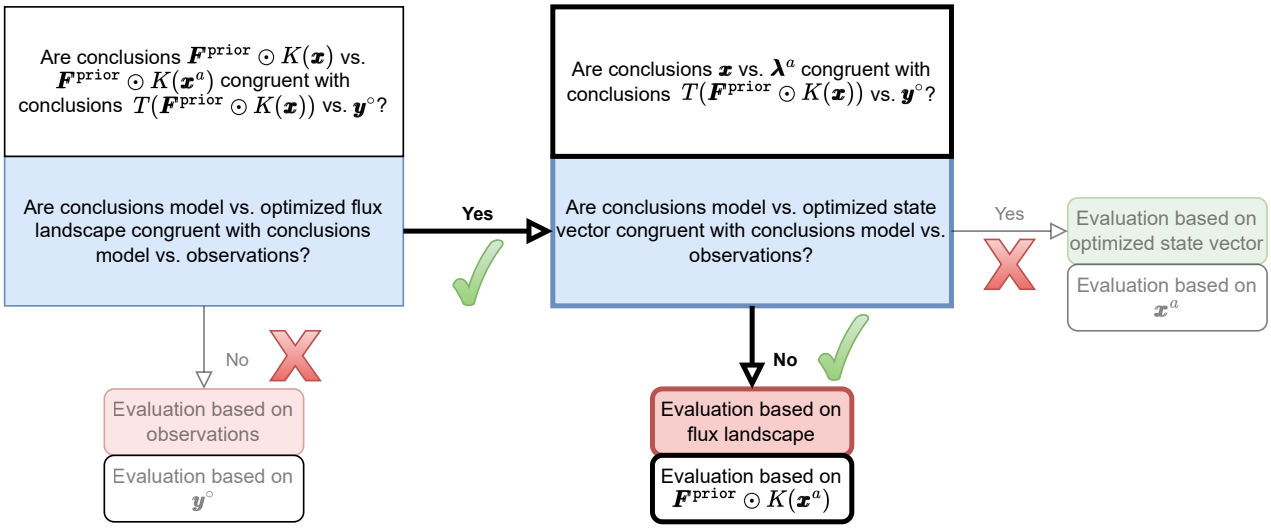


Figure 5.1: Filled in diagram of the decision process depicted in Figure 4.1. As the results of the evaluations of *model vs. optimized flux landscape* and *model vs. observations* are sufficiently similar, evaluation in observation space is not needed. The results of the comparisons of *model vs. optimized state vector* and *model vs. observations* did however differ on the ranking of the models. Hence, evaluation in state vector space is not a viable option. This leads to the conclusion of evaluating intermediate models in flux space.

Chapter 6: Conclusion

Based on the findings presented above, the research questions set out in Section 1.4 can be answered.

«*Would using the seasonal pattern found within the analyzed state vector (see Figure 1.3) result in a transition function producing a background state vector closer to the true state vector?*»

The comparison between the `smoothed` and `monthly` model in Figure 4.2 shows a substantial reduction in mean error. As the `monthly` model was explicitly designed to include the seasonal pattern referenced in the question as an alternative to the `smoothed` model which does not use this seasonal pattern, it can indeed be concluded that the usage of the seasonal pattern improved the ability of the transition model to match the true state vector.

«*In which evaluation space is the performance of the transition model most likely to generalize to the performance within a full inversion run?*»

The evaluation of various models in different evaluation spaces has shown that it is possible to evaluate the state vectors produced by transition models in flux space. The qualitative results from the evaluation in observation space match the qualitative results from the evaluation in flux space, while the evaluation in state vector space showed different qualitative results. As a result, an evaluation in state vector space is unlikely to generalize to observation space. Due to the limited number of models included in the analysis, a hard conclusion on the generalisability of the results in observation space to flux space cannot be made. However, these hard conclusions are not needed. The main purpose of the evaluation is to find an indication of the effectiveness of evaluating state vector models in an evaluation space less complex than the observation space. As the same qualitative properties are present in both the observation and flux space, there is sufficient evidence to support the evaluation of intermediate models in flux space.

Part II
Comparing ML Models

The monthly mean method presented in Part I already provides an improvement over the current CTE2018 implementation. However, it is still not known how much training data is needed for the monthly model to perform optimally. Furthermore, the monthly mean method is relatively ill-informed. Including additional predictor variables in combination with ML could improve the forecast function even further. This part of the thesis focuses on stochastic time-series models, which can utilize linear relationships between the predictor variable and target variable.

When building these forecast models, it is important to keep in mind the purpose of the model:

1. The goal is to develop a state transition model for the state vector, as defined in Equation 10. Hence, the output of the model should be easily translated into a state vector.

2. For an effective integration within CTDAS, the model should reach a reasonable performance using a limited amount of training data. If too much training data is needed, the transition model can only be applied during the final few years of the inversion run, limiting its usability.

3. Part I has shown that evaluation in state vector space could result in sub-optimal model selection. Therefore, the final model evaluation should be done in flux space, or, if time and resources allow it, observation space.

4. The original set of state vectors contains a substantial amount of noise. Some pre-processing of the data is needed to limit this noise, making it easier to extract relations between the predictor and the target variable.

While keeping the points listed above in mind, the following questions are investigated in the remainder of this part of the thesis:

«*Does the utilization of additional temporal dependencies (i.e. dependencies between time-steps) result in a reduction of the mean bias of the prior biosphere and ocean flux with respect to the monthly average model?*»

The current monthly model is a simple model which determines 12 independent parameters for every element in the state vector; one average scaling factor for every month of the year. This implementation is sub-optimal, as it is not able to capture weekly and yearly variability.

Heat anomalies are among the main factors driving anomalies in NEE (Rödenbeck, Zaehle, Keeling, & Heimann, 2018). A plausible explanation is the response of vegetation to extreme heat. When plants experience heat-induced stress, photosynthesis is limited to conserve water (Peters et al., 2018). The exact moment when plants enter this water-conserving mode is difficult to model. As such, heat anomalies are likely to be correlated to biases within the biosphere model. Considering that extreme temperature anomalies often last for several weeks (i.e. droughts), it is reasonable

to assume weekly temporal dependencies exist within the biases of the biosphere model. Longer temporal dependencies are also likely to exist, considering that the effects of severe droughts can last for multiple years (Yu et al., 2022; Kannenberg, Schwalm, & Anderegg, 2020). Perhaps even more interesting, is the finding that the used SiB4 model consistently under- or overestimates the NEE of some PFTs in a yearly recurring pattern (based on personal correspondence with dr. LMJ Kooijmans-de Vries). Hence, both weekly and yearly variability within the scaling factors could provide valuable information for a scaling factor forecast model.

Instead of trying to capture the temporal dependencies within the scaling factor based solely on previous scaling factors, it might be easier to use the environmental conditions causing these temporal dependencies by using the environmental conditions as predictor variables. Next to heat, other environmental conditions such as precipitation and solar radiation intensity could influence the heat stress experienced by plants as well. This line of reasoning results in the second and final question this part of the thesis aims to answer:

«*Could the utilization of predictor variables (i.e. temperature, precipitation), in combination with the temporal dependencies, result in a reduction in the mean bias of the prior biosphere and ocean flux with respect to the monthly average model?*»

Before the proposed questions can be answered, the process used to denoise the data is discussed. Afterward, the models used to answer the questions are explained, followed by the experimental setup. The four points listed at the start of this introduction provide a central theme during these sections. This part continues with the results of the conducted experiments and a discussion of these results. The final section draws conclusions from the results and aims to answer the questions proposed in this introduction.

Chapter 7: Methods

7.1 Data pre-processing

The first step in the analysis is defining, and gaining insights from, the target data. The goal of the transition model is to find a $\boldsymbol{\lambda}_t^b$ which resembles $\boldsymbol{\lambda}_t^a$ as closely as possible. Hence, the target variable is as simple as $\boldsymbol{\lambda}^a$. Basic intuition tells us that the scaling factors within $\boldsymbol{\lambda}^a$ should not be negative, as that would also reverse the diurnal cycle present within the biosphere, and should neither be much greater than 3, as that would imply a severe underestimation of the flux from the original CE models. Any value outside of this range is more likely to be noise from the EKF instead of being the result of a real bias within $\boldsymbol{F}^{\texttt{bio}}$ and $\boldsymbol{F}^{\texttt{ocean}}$. However, Figure 7.1 shows that a substantial amount of scaling factors within $\boldsymbol{\lambda}^a$ are outside of this range. Hence, some noise reduction is needed before $\boldsymbol{\lambda}^a$ can be used as a target variable.

The noisiest data points are located in the part of $\boldsymbol{\lambda}$ which represents the *North-American temperate*, *North-American boreal*, *Eurasia temperate*, *Eurasia boreal* and *Europe* TransCom regions as shown in figure 7.2. Each element within this part of $\boldsymbol{\lambda}$ represents a single 1×1 degree grid point and these individual grid points are not very well constrained by observations. Hence, individual grid points can be scaled up or down to unrealistic values. However, these gridded regions are also the regions where most of the observations are taken from (see figure 1.2). Therefore, the overall corrections made within these regions are most likely to be a result of an actual bias within the flux model. It is therefore likely that the corrections made on a higher scale are informative. A noise reduction method is needed which is able to capture the information present within these lower spatial resolution corrections.

7.1.1 The effective scaling factor

Unfortunately, it is not possible to simply take an average of all the scaling factors within an ecoregion. The problem is that the fluxes associated with each element vary. Therefore a weighted average is needed based on the prior flux. The most intuitive way of doing so is by determining the effective scaling factor for each ecoregion. This is achieved by dividing the sum of all optimized fluxes within an ecoregion by the sum of all prior fluxes. See Figure C.3 for a diagram explaining the procedure in more detail. The new effective scaling factor vector l is defined as follows:

$$l_r = \frac{\sum_{e \in E_r} \lambda_e \cdot \mathcal{K}^{-1}(\boldsymbol{F}^{\texttt{prior}})_e}{\sum_{e \in E_r} \mathcal{K}^{-1}(\boldsymbol{F}^{\texttt{prior}})_e}, \tag{17}$$

where r is one of the gridded ecoregions (See Table D.4 for a comprehensive list), E_r is the set of all elements within the state vector which lie within ecoregion r, and $\mathcal{K}^{-1} : \mathbb{R}^{360\times180} \to \mathbb{R}^{9835}$ is the inverse of $\mathcal{K}$ which maps a matrix of fluxes to elements within the state vector.

One issue inherent to the nature of the operation is the tendency for effective scaling factors associated with prior fluxes close to 0 to become unstable. The found values ranged from -6000 to $+4000$. As the prior flux associated with these scaling factors are close to 0 (see Figure 7.4), the

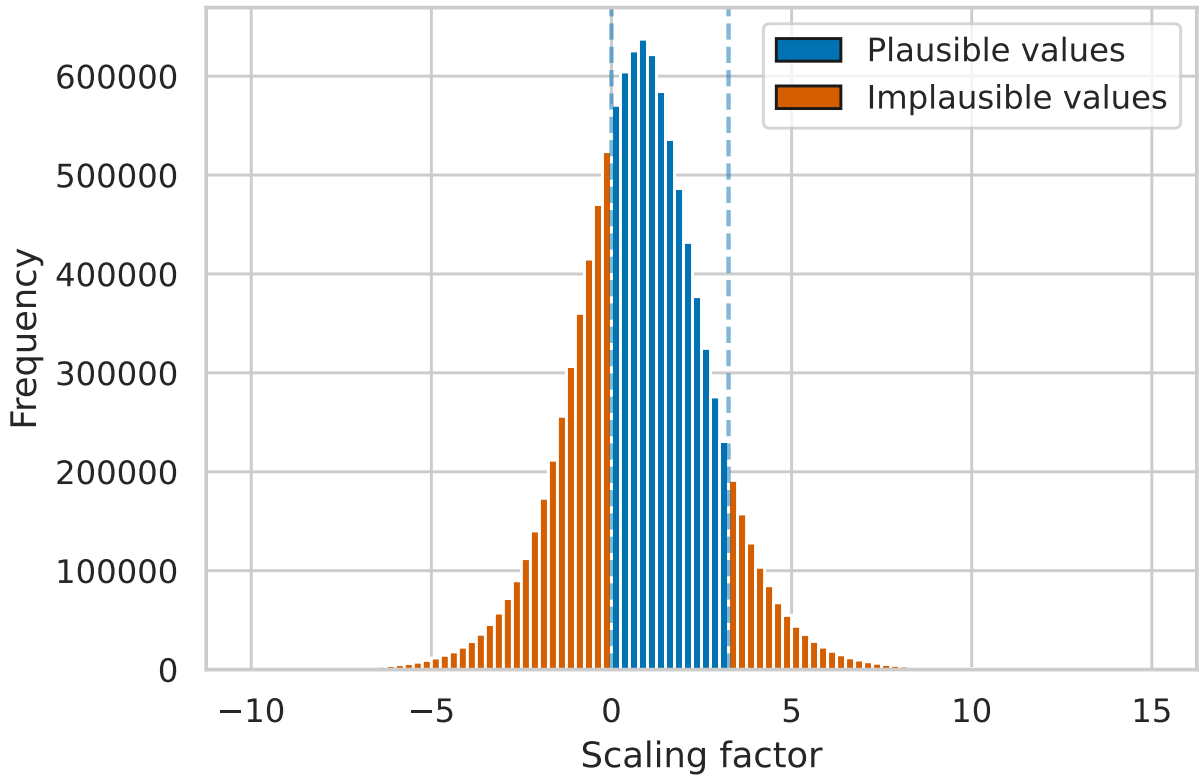


Figure 7.1: A histogram of all scaling factors within the set of all analyzed state-vectors ($\mathbf{\Lambda}^a \in \mathbb{R}^{9835\times1096}$). Values between 0 and 3 are considered to be plausible indicators of a true bias and are shown in blue. Values outside of this range are considered to be implausible and are likely to contain some noise. These values are indicated in red. In total, 56.6% of all scaling factors fall within the range of plausible values

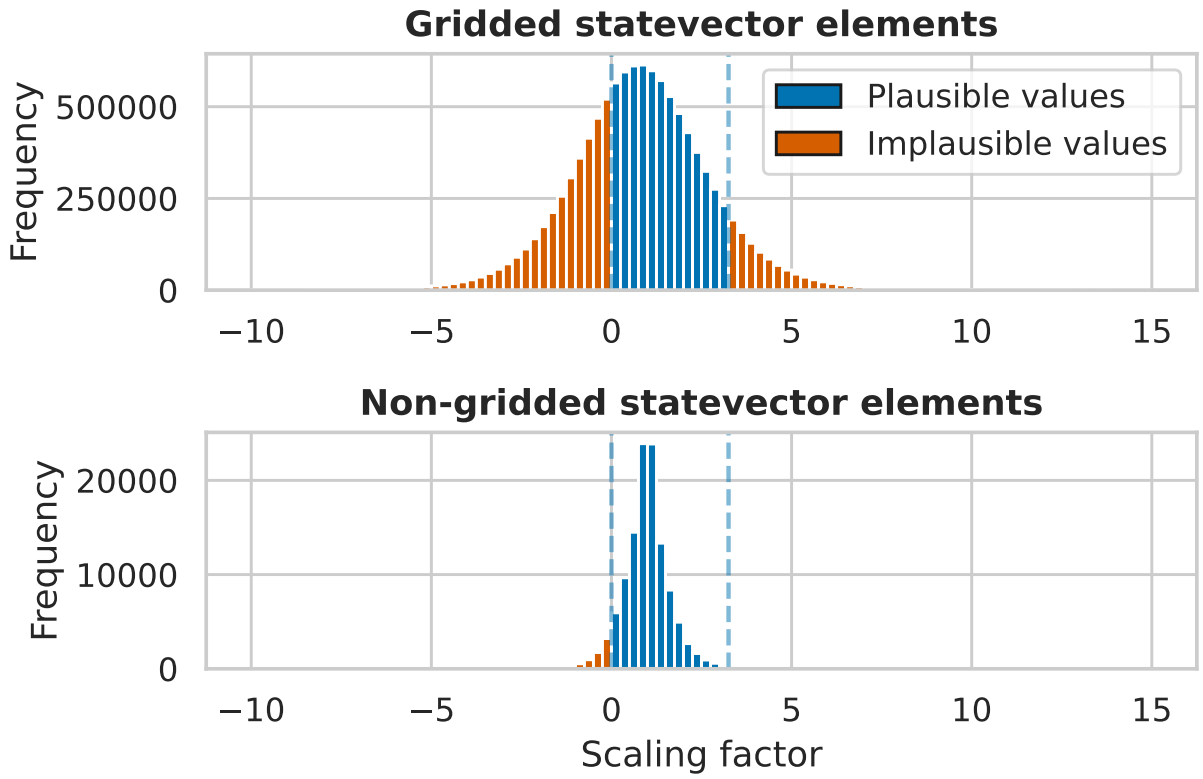


Figure 7.2: The distribution of the scaling factors associated with the gridded part of the state-vector (i.e. TransCom regions North American Boreal, North American Temperate, Eurasia Boreal, Eurasia Temperate, and Europe) compared to the distribution scaling factors associated with the non-gridded part of the state-vector elements (i.e. TransCom regions South American Tropical, South American Temperate, Northern Africa, Southern Africa, Tropical Asia, Australia, and the Oceans). Of the scaling factors associated with the gridded section of the state vector, 56.2% fall within the range of plausible values, compared to 93.2% of the scaling factors associated with the non-gridded part of the state vector.

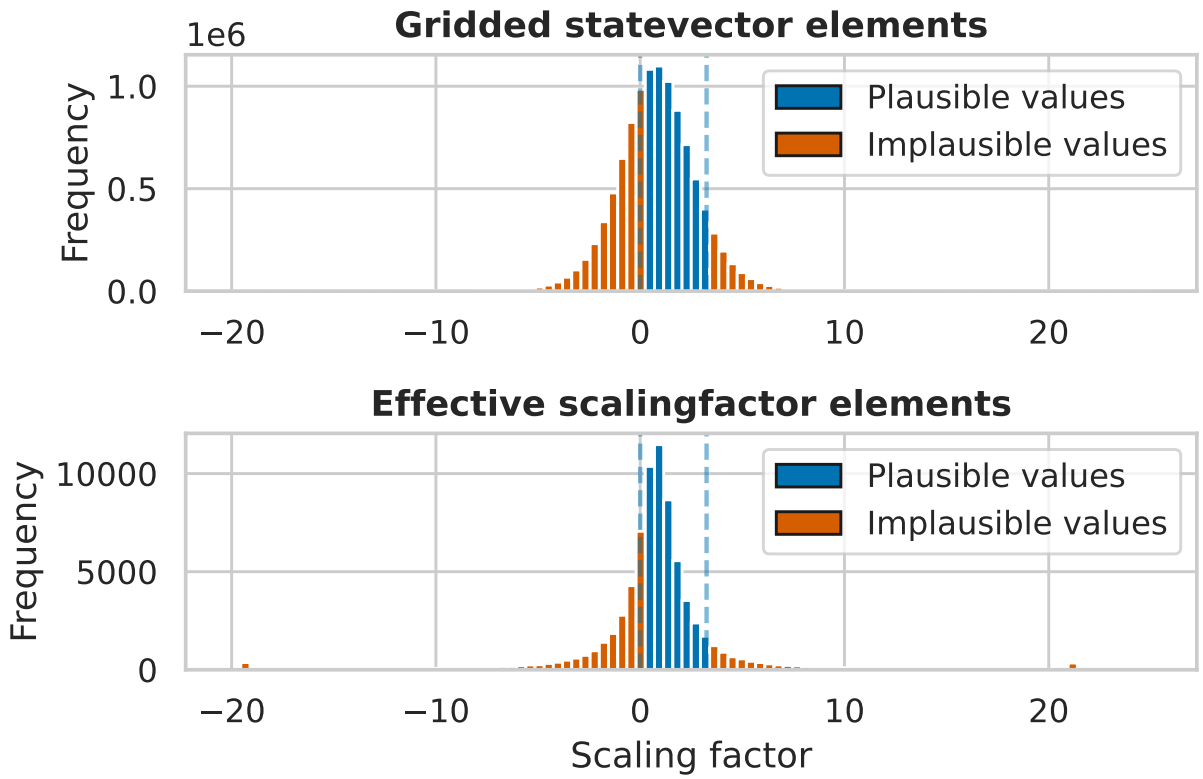


Figure 7.3: The distribution of the scaling factors from the gridded section of the state-vector compared to the effective scaling factor capped at $\pm 4\sigma$. Using the effective scaling factors increased the percentage of realistic scaling factors from 56.2% to 62.4%.

effect these scaling factors have on the eventual fit to the 'true' flux landscape is limited. Hence, this instability in the scaling factors is not of immediate concern. In the current implementation, the greater outliers are controlled for by capping the scaling factors at the somewhat arbitrary boundary of $\texttt{mean}(\boldsymbol{l}) \pm 4\sigma$. This cap had a very limited effect on the resulting distribution of effective scaling factors, which is shown in figure 7.3. Only the small peaks at ~ -19 and ~ 21 result from this cutoff point at 4σ. It might, however, be interesting to see how this instability affects the ability of ML methods to capture the trend within the aggregated scaling factors. This idea is explored further in Section 10.

7.1.2 Resulting distribution

While some improvement is achieved, a substantial part (37.6%) of the effective scaling factors remain 'implausible'. Additional data analysis shown in Figure 7.5 shows that the effect of using the effective scaling factors is stronger in some TransCom regions than in other TransCom regions.

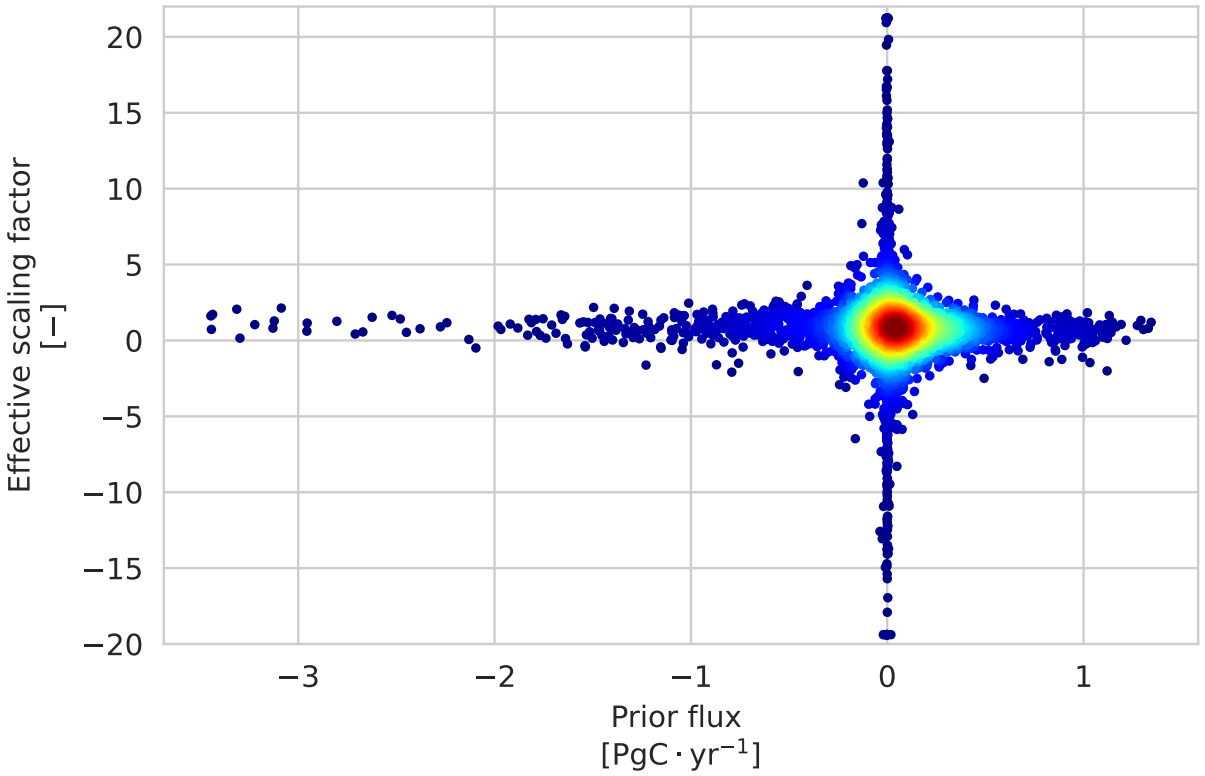


Figure 7.4: Correlation between the prior flux and the effective scaling factor. The color gives an indication of the density of the distribution and is generated using Gaussian kernel density estimation. The figure shows how most of the scaling factors outside of the range of [0, 3] are centered around a prior flux of 0

More concretely, Table D.3 shows that there is a decrease in the proportion of plausible scaling factors in the *Eurasia Temperate* TransCom region, while the proportion of effective scaling factors increases substantially in the *Europe* TransCom region. This difference in the proportion of realistic scaling factors could partially be explained by the constraints put on the scaling factor within each region by observations. As Figure 1.2 shows, the *Europe* TransCom region is highly constrained, while the *Eurasia Temperate* TransCom region is not.

Even though North America is well constrained as well, Figure 7.5 and Table D.3 show that the ratio of plausible effective scaling factors is only marginally higher than the ratio of plausible unmodified scaling factors. The reason for this large difference is uncertain, but it could be due to a different surface area distribution of the ecoregions. As explained before, the effective scaling factor can 'explode' if prior flux within the ecoregion is close to 0. Smaller ecoregions will have smaller fluxes and might therefore be more prone to have an unstable effective scaling factor.

7.1.3 Data used for the remainder of the thesis

Now, any ML implementation is only as good as the data upon which it is trained. As the quality of the target data seems to vary across TransCom regions, it might very well be possible that the ability of a forecast model to capture the trend within the effective scaling factors varies as well. Therefore, the remainder of the thesis focuses on the TransCom regions in which we have the most confidence in the scaling factors. These are the *North American Boreal*, *North American Temperate*, *Eurasia Boreal*, *Eurasia Temperate* and *Europe* TransCom regions. This means that all scaling factors associated with the ocean fluxes are excluded from the forecast model. Furthermore, each of the included TransCom regions is evaluated separately, making it possible to check whether the prior intuitions on the quality of the target data translate to better model performance.

7.2 Feature selection

The second research question focuses on the utilization of environmental conditions as predictor variables for the target variable which is the effective scaling factor. The list of environmental conditions which could be used is long. The European Centre for Medium-Range Weather Forecasts (ECMWF) models over 5000 variables, of which over 185 are of sufficient temporal and spatial resolution (weekly estimates on a global 1×1 degree grid). Finding a correlation to at least some elements within the set of all effective scaling factors is therefore an almost trivial task, simply because of the sheer amount of available variables. Instead, the focus should be on meaningful relations between the environmental conditions and the effective scaling factor.

Based on the expert judgment of A. van der Woude, 18 distinct variables were selected. A full list is provided in table D.5. These variables were directly taken from the ECMWF, but have been aggregated such that a weekly value per ecoregion remained. This was done in various stages. As the effect of the environmental conditions on the biases within the biosphere is unknown, it is also unknown whether the biases are influenced by the max value of a certain variable (a peak in the wind speed could result in a bias), the min value (a bias could be introduced by a sudden drop

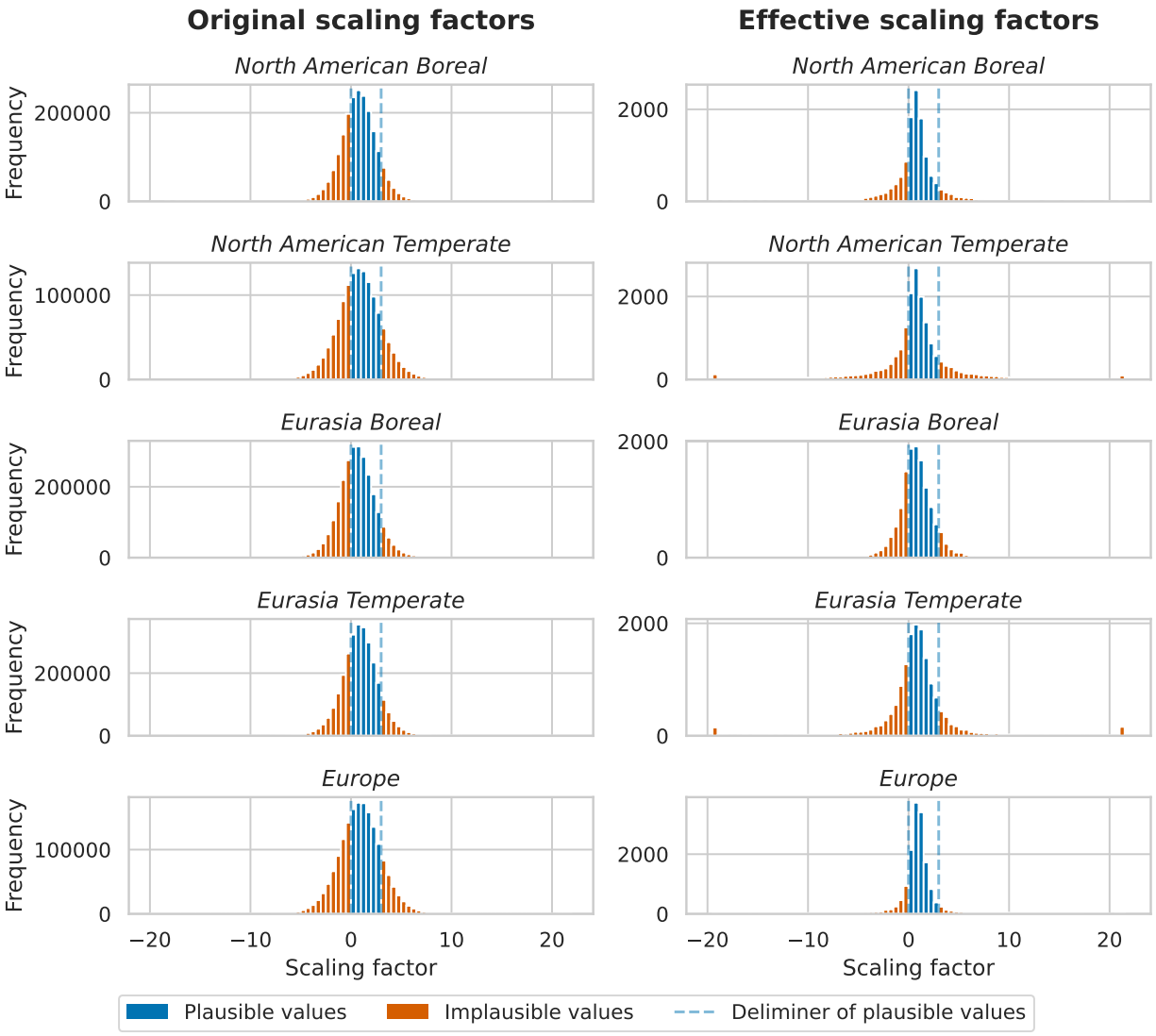


Figure 7.5: The distribution of the scaling factors from the gridded section of the state-vector compared to the effective scaling factor capped at $\pm 4\sigma$ grouped by TransCom region. The effect of using the effective scaling factors compared to the raw scaling factors varies greatly depending on the TransCom region

in temperatures), the average value (the average solar radiation over some period of time could be relevant), or the sum (the biosphere could be influenced by the absolute quantity of precipitation). Therefore, some variables have been aggregated in several different ways, again based on the expert judgment of A. van der Woude. However, our intuition did lead us to believe that these aggregation methods are mainly relevant for lower spatial resolutions. A maximum temperature within a 1×1 degree grid cell is interesting, but this single maximum value is not a proper representative of the entire ecoregion. As such, these different aggregation methods were only used to reach a global landscape of 1×1 weekly variable values. Combining the different variables with the different aggregation methods resulted in 36 potential predictor variables.

To reach the desired ecoregion resolution, all variables have been aggregated using a weighted average based on the total ecosystem respiration (TER). This TER can be interpreted as the combined carbon 'exhaled' by the ecosystem and is closely related to the overall NEE within an ecosystem (Lasslop et al., 2010, see figure 8). As opposed to NEE, the TER is less influenced by seasonality and thus remains relatively constant throughout the year. This makes averaging substantially easier. Therefore, all environmental variables are aggregated based on the average TER over the period 2000-2020. One downside is that the TER only applies to terrestrial ecoregions. However, the fluxes from the water PFT scale well with total surface area and thus the fluxes from the water ecoregions are aggregated using an average weighted by the surface area of each grid cell.

One final step remains for determining the environmental conditions that could be used as predictor variables. Background literature suggests that the main driving force behind anomalies in net ecosystem exchange (NEE), which is analogous to the net exchange of carbon between the atmosphere and the biosphere, are anomalies in temperature (Rödenbeck et al., 2018). These anomalies in NEE are difficult to capture within a biosphere model and as a result, are likely to be the cause of biases within this model. The assumption is therefore that the anomalies in temperature are among the factors determining the corrections made by the EKF, meaning that they affect the effective scaling factor. This line of reasoning resulted in not the direct variable value being used as a predictor variable, but the variable expressed in anomaly space.

In accordance with Rödenbeck et al. (2018), the anomaly is determined based on a rolling monthly mean over 11 years. The eventual implementation of monthly anomaly is similar to the monthly mean model, with the precise definition being

$$\boldsymbol{v}_t^{\Delta} = \boldsymbol{v}_t - \left(\frac{1}{|\boldsymbol{V}_{t,m}^{\texttt{rolling}}|} \cdot \sum_{\boldsymbol{v} \in \boldsymbol{V}_{t,m}^{\texttt{rolling}}} \boldsymbol{v}\right)\Bigg|_{\texttt{month}(t)=m}, \tag{18}$$

where t is time in weeks, $\boldsymbol{v}_t \in \mathrm{R}^r$ is one of the variables listen in Table D.5 at time t, with each of the r elements is linked to one of the ecoregions listed in table D.4 and $\boldsymbol{v}_t^{\Delta}$ the monthly anomaly at time t. $\texttt{month}(t)$ is an operator extracting the month of t, where $m \in [Jan, Feb, ..., Dec]$ is one of the 12 months of the year. $\boldsymbol{V}_t^{\texttt{rolling}}$ is a rolling window containing at most 11 years of variable data ending at t; $\boldsymbol{V}_{t,m}^{\texttt{rolling}}$ is the set of elements in $\boldsymbol{V}_t^{\texttt{rolling}}$ where $\texttt{month}(t) = m$.

Now, for the initial test on whether the inclusion of environmental conditions as predictor variables could result in a reduction of the mean bias within a full inversion run, we started with only

using the `t2m_AVG` monthly anomaly. This variable is the average temperature at 2 meters above the earth's surface and is the variable used in Rödenbeck et al. (2018). The usage of other variables is discussed in Section 9.2.

7.3 Used models

This thesis tests the performance of three separate models on the task of forecasting the effective scaling factor $\boldsymbol{l}$. The monthly mean model adjusted to the new data representation serves as a benchmark, as it has been shown in Part I that this model is currently the best available forecast model. The other models are the seasonal autoregressive integrated moving average (SARIMA) and the seasonal autoregressive integrated moving average with exogenous factors (SARIMAX) models. These models have been selected based on how well they are able to either mitigate a potential issue within the problem description or how well they are able to utilize potential sources of information. An overview of these criteria, along with their priority and an explanation, is given in Table 7.1.

Furthermore, the chosen models will only be trained on the ecoregions within the *North American boreal, North American temperate, Eurasia Boreal, Eurasia Temperate*, and *Europe* TransCom regions, as there are the regions with most confidence can be put in the effective scaling factor.

Note that in total 68 ecoregions are included in the analysis, which means that $\boldsymbol{l} \in \mathbb{R}^{68}$ (see Table D.4 for a full overview).

7.3.1 Monthly average

As already mentioned in Part I, the monthly average model has been developed as an intuitive and simple method for utilizing the seasonal pattern within the state vector depicted in Figure 1.3. Next to its simplicity, it ranks remarkably well on the top 4 criteria listed in Table 7.1 as shown in Table D.6.

The description of the monthly average model has not changed much compared to the model introduced in section 2, Equation 15. The only difference is the data used to 'train' the model. Instead of the state-vector $\boldsymbol{\lambda}$, the effective scaling factor $\boldsymbol{l}$ introduced in Section 7.1.1 is used, resulting in the final definition

$$\mathcal{M}^{\texttt{monthly}}(t) = \frac{1}{|\boldsymbol{L}_m^{\texttt{train}}|} \cdot \sum_{\boldsymbol{l}^a \in \boldsymbol{L}_m^{\texttt{train}}} \boldsymbol{l}^a \Bigg|_{\texttt{month}(t)=m}, \tag{19}$$

where t is the time in weeks, $m \in [Jan, Feb, ..., Dec]$ is one of the 12 months and $\boldsymbol{L}_m^{\texttt{train}}$ is the set of all $\boldsymbol{l}_t^a \in \boldsymbol{L}^{\texttt{train}}$ where $\texttt{month}(t) = m$. Again, all this model really does is group all previous scaling factors from every month and determines the average.

Table 7.1: Overview of the various problems and potential information sources encountered when creating an ML model of the transition function of the state-vector within CTDAS. Each point is listed according to a priority, where the higher the priority, the more important it is for the problem/information source to be mitigated/utilized.

Priority	Problem/ information source	Problem description
1	Limited data availability	Only a single time series is available. In theory, it would be possible to train on multiple state-vector series of different inversion runs. In practice, however, such inversion runs are computationally very expensive, requiring close to 6 weeks on 50 CPUs for a single run. Furthermore, it is uncertain how well a transition model trained on one state-vector series would generalize to another state-vector series. As such, it is best to work under the premise of having only a single series of state-vectors available
2	Integration within CTDAS	As stated within the introduction, our interest is not in just developing a transition model $\mathcal{M}$ for the EKF, but in improving the CTDAS as a whole. For this purpose, it is not only important for the ML implementation to be trainable on a limited amount of data, but also that it can continue learning while CTDAS continues analyzing new weeks. Either some form of online learning is needed, or the cost of retraining the ML model needs to be minimal
3	Noise within data	One of the most pressing issues with the available data, is its quality. As mentioned in Section 7.1, the data contains a substantial amount of noise, even after pre-processing. The ML algorithm should therefore be relatively robust against noise
4	Temporal dependencies	The introduction of Part II mentioned the temporal patterns found within the state-vector and touched upon some physical processes justifying the assumption of temporal dependencies within the state-vector. These temporal patterns could prove to be useful for training an ML-model
5	Exogenous variables	Also mentioned in the introduction of Part II, are the potential conditions determining the state-vector. Temperature anomalies result in anomalies in NEE, which are notoriously hard to capture within biosphere models. Other abnormal environmental conditions could also affect the state vector; thus, these conditions could contain valuable information when training a transition function.
6	Spatial dependencies	A downside of inversion methods are the introduced dipoles, where an error in one region is compensated by surrounding regions (Jacobson et al., 2007b). This means that it is likely that a spatial correlation exists between the scaling factors of various regions.

7.3.2 SARIMA

The SARIMA model was initially selected as it is a relatively simple model that can be trained on a single and relatively short time-series (Box, Jenkins, Reinsel, & Ljung, 1976). Furthermore, it provides more freedom to utilize more temporal dependencies compared to the monthly mean model. A downside is that it does not allow for a multivariate output, meaning that one SARIMA model needs to be trained for every included ecoregion. Furthermore, the bare SARIMA is unable to utilize exogenous variables. For a complete overview of how well it matches the selected criteria, see Table D.7.

The seasonal autoregressive integrated moving average (SARIMA) model is a multiplicative model of a auto-regressive integrated moving average (ARIMA) model with a seasonal component. Box et al. (1976) defines the SARIMA models using the notation $(p,d,q) \times (P,D,Q)_s$, where p is the order of the regular auto-regression (AR) term, d the order of the integration (I) term, q the order of the moving average (MA) term, P, D, and Q are their respective seasonal equivalents, and s represents the seasonality of the model. Furthermore, Box et al. (1976) suggests using the Box-Jenkins method for finding the values of these hyperparameters. This method consists of three steps:

1. **Model selection**: This step requires a thorough analysis of the data. By looking at the correlations within the target time series, an initial estimate is provided for the seasonality (S), auto-regression (AR), integration (I), and moving average (MA) terms of the model.

2. **Coefficient estimation**: Once an appropriate model is selected, the coefficients within the models are estimated.

3. **Model evaluation**: The model is evaluated on the training data, not the testing data. The main evaluation metric is the distribution of the residuals. In essence, the residuals should be independent of each other. If they are not, step 1 has to be redone.

Note that only the training data $\boldsymbol{L}^a \in \mathbb{R}^{888\times 68}$ has been used for the hyperparameter selection. The full Python implementation of the Box-Jenkins method can be found in the `SARIMA.ipynb` notebook provided in the GitHub repository of this thesis.

Model selection Step one of model selection is checking whether the series of effective scaling factors had a unit root. This is used to determine whether the mean of the series is stationary. An augmented Dicky-Fuller test confirmed that none of the effective scaling factors of the 68 included ecoregions contained a unit root. As such, both the d and D terms are set to 0. This effectively makes the SARIMA model a SARMA model, but for the sake of consistency, we will keep referring to the model as the SARIMA model.

The second step is determining the s, p, q, P, and Q terms. This is done using the autocorrelation function (ACF) and partial autocorrelation function (PACF). Interpreting the resulting graphs is not as straightforward as with most graphs, so before analyzing the graphs, some background information is given.

Autocorrelation function: The ACF shows the direct correlation between a time series and several lagged versions of the same time series. This is used to determine the q and Q terms since in a MA model without AR and I term, every time point is determined by the average of some underlying variable at the same point and the q prior points. Therefore, peaks at $lag = 1$ indicate that $q = 1$, peaks at $lag = 52$ indicate $Q = 1$, peaks at $lag = 1, 52, 53$ indicate $q = Q = 1$, etc.

Partial autocorrelation function: The PACF shows the correlation between a value in the time series and several lagged versions of the same time series. It is a partial function as it accounts for any correlations between the current lag and any smaller lags. Therefore, it is used to determine the p and P terms. Peaks at $lag = 1$ indicate that $p = 1$, peaks at $lag = 52$ indicate $P = 1$, peaks at $lag = 1, 52, 53$ indicate $p = P = 1$, etc.

Computing both the ACF and PACF for all lags over the entire training dataset is expensive. Therefore only the first 110 lags of all effective scaling factors of the ecoregions within the *Europe* TransCom region have been analyzed (see Figure B.2). The *Europe* TransCom region has been selected as this the effective scaling factors from this TransCom region has been determined to be most reliable (see Section 7.1.2) and 110 lags cover two cycles of the expected seasonal pattern.

The first thing to notice is the small peak in both the autocorrelation function (ACF) and partial autocorrelation function (PACF) at a lag of 52. This confirms our suspicion that there is a seasonal effect and thus the s can be set to 52 with a P and Q terms of *at least* 1.

Now, the problem is that a general set of hyperparameters should be found which fit all of the analyzed series of effective scaling factors. Figure B.2 shows that there is a substantial difference in the lags which are significantly correlated between the series of effective scaling factors. In general, $p = q = 2$ seems to be a good fit for some time series (e.g. figures B.2e and B.2l), while for others only a model with $P = Q = 1$ seems to be a proper fit (e.g. figures B.2a, B.2h, and B.2k, and some seem to be best described by a combination of the two (e.g. figures B.2b and B.2i). Therefore the $(2,0,2) \times (1,0,1)_{52}$ model seems to be the simplest model which can be applied to most series.

Parameter estimation After finding an initial guess for the hyperparameters of the SARIMA model, the values of the coefficients need to be found. The used SARIMA implementation (Seabold & Perktold, 2010) uses maximum likelihood estimation (MLE) for this task. The found coefficients show a substantial variation in their 95% confidence interval, but clearly show that the white noise term remains one of the most dominant factors determining the made forecasts (see Figure B.3). However, a substantial amount of coefficients are found to be significantly greater or smaller than 0, implying that at least some of the variance within the training data is captured.

Model evaluation While some variance is captured within the SARIMA models, the white noise components remain a dominant factor within the forecasts. This warrants additional analysis of the residuals of the model to determine whether there are correlations between lags that have been missed and should still be included. If these correlations are missed, this would result in a dependency between the residuals. Therefore, a Ljung-Box test is used to test for dependencies within the

Table 7.2: The results of the Ljung-Box test on the residuals of the SARIMA models trained on the 15 ecoregions within the *Europe* TransCom region. The null hypothesis of the residuals being independent of each other can only be rejected for ecoregion 195.0 at a significance level of $p < 0.05$.

Ecoregion	Ljung-Box statistic $lag = 110$	p	Ecoregion	Ljung-Box statistic $lag = 110$	p	Ecoregion	Ljung-Box statistic $lag = 110$	p
191.0	63.9	1.00	196.0	82.1	0.98	201.0	123.4	0.18
192.0	72.7	1.00	197.0	71.5	1.00	202.0	133.3	0.07
193.0	67.6	1.00	198.0	103.5	0.66	204.0	120.5	0.23
194.0	83.8	0.97	199.0	66.2	1.00	206.0	84.0	0.97
195.0	137.1	**0.04***	200.0	93.6	0.87	209.0	56.2	1.00

***** - *significant at the level of* $p < 0.05$

residuals across the same 110 lags used earlier.

Table 7.2 shows that the residuals of almost all ecoregions contain no significant correlation within any of the tested lags and can thus be considered white noise. Only ecoregion 195.0 contains a significant correlation. Table D.4 shows that this ecoregion is associated with the tropical forest PFT and table D.1 shows that this PFT makes up only 0.1% of the total surface area of the *Europe* TransCom region and thus contributes very little to the overall flux landscape of the entire TransCom region. Furthermore, the listed p-values in Table 7.2 did not account for multiple comparisons. Therefore, it is reasonable to assume that providing extra coefficients to the $(2,0,2) \times (1,0,1)_{52}$ model will not result in a meaningful improvement on the TransCom region scale. Hence, the $(2,0,2) \times (1,0,1)_{52}$ model is selected to be the final model.

7.3.3 SARIMAX

The final proposed model is the SARIMAX model. The only difference between this model and the SARIMA model, is that it can use exogenous variables for additional information. It does so by adding a coefficient for each used exogenous variable, capturing any linear correlation between the exogenous variables and the target variable. As explained in Section 7.2, only the monthly temperature anomaly is used as such an exogenous variable. Note that the Future iterations could also look into other variables, further discussed in Section 9.2.

Chapter 8: Experimental Setup

As explained in the previous section, the data used for this part of the thesis is the set of all weekly effective analyzed state vectors ($\boldsymbol{L}^a \in \mathbb{R}^{68\times1096}$) of the years 2000 to 2020. This set is divided into 17 years of training data ($\boldsymbol{L}^{\texttt{train}} \in \mathbb{R}^{68\times888}$) and 4 years of testing data ($\boldsymbol{L}^{\texttt{test}} \in \mathbb{R}^{68\times208}$) according to an approximate $80-20\%$ split. To test how much data is needed for each model to converge to its optimal performance, the training data is subdivided into 17 chunks, one for each year of available training data. This process is depicted in Figure 8.1.

The results from Part II indicate that the models should be evaluated in flux space. Furthermore, Section 7.1 showed substantial differences within the distribution of effective scaling factors between TransCom regions. Combine this with the goal of minimizing the budget imbalance described in Equation 1 and the main performance measure of modeled effective scaling factors $\boldsymbol{L}$ can be set to the mean error (ME) in flux space, where each TransCom region (tc) is evaluated separately:

$$\texttt{ME}(\boldsymbol{L})_{\texttt{tc}} = \frac{1}{|T|} \cdot \sum_{t\in T} \sum_{r\in R_{\texttt{tc}}} (l_{r,t} - l^a_{r,t}) \cdot f^{\texttt{prior}}_{r,t}, \tag{20}$$

where T is the set of all weeks in the testing dataset, $R_{\texttt{tc}}$ is the set of all ecoregions within TransCom region tc, $\boldsymbol{l}^a$ is the effective analyzed scaling factor, and $f^{\texttt{prior}}_{r,t}$ is the total flux of ecoregion r at time t. The mean absolute error (MAE), root-mean-square error (RMSE), mean absolute percentage error (MAPE), and coefficient of determination (R^2) are also included as secondary performance measures and are determined using the same structure.

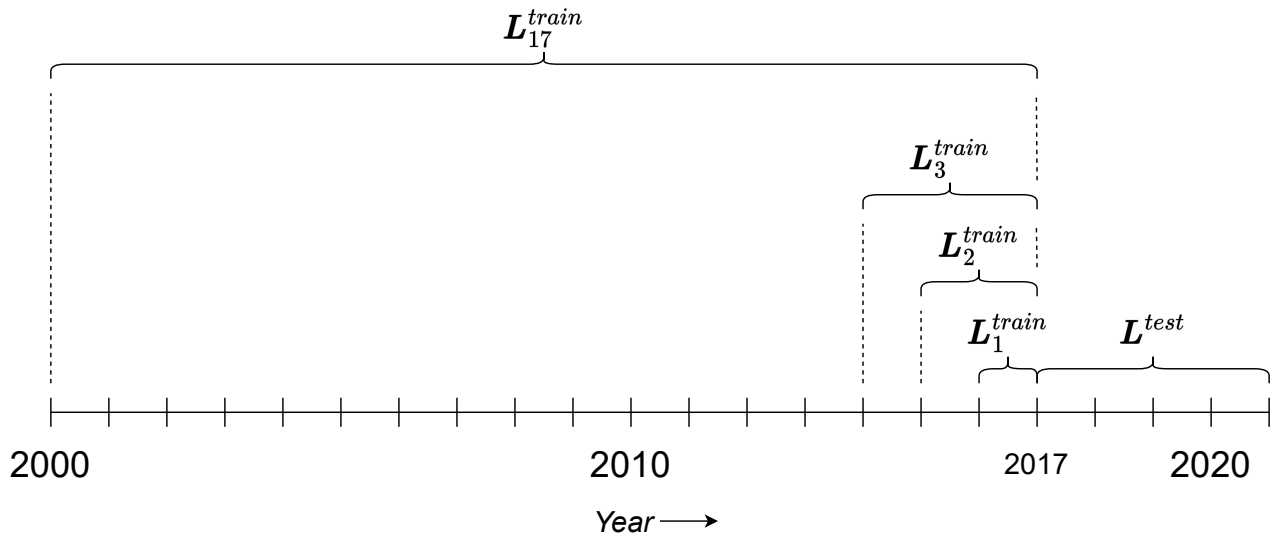


Figure 8.1: Illustration of the procedure used to subdivide the training data into smaller sets. Each set is defined as $\boldsymbol{L}_n^{\texttt{train}}$, where $n \in N$ is an integer representing the number of years of effective scaling factors present within the set, and $N = \{n \in \mathbb{N} \mid n \leq 17\}$. Note that $\boldsymbol{L}_n^{\texttt{train}} \subset \boldsymbol{L}_{n+1}^{\texttt{train}} \; \forall n < 17$.

Chapter 9: Results

This part discusses whether the implementation of the SARIMA and SARIMAX models, which in the remainder of the section will be referred to as the SARIMA(X) models, could provide an improvement over the current best transition model, being the monthly mean model. Due to a substantial amount of noise within the original state vector, these SARIMA(X) models, along with the monthly mean model to which they are compared, are not trained on the direct state vector, but on an aggregated version referred to as the effective scaling factor per ecoregion. Furthermore, 38 exogenous variables have been selected that could be used to provide additional information to the SARIMAX model, of which for now only the `t2m_AVG` variable is used as background literature lists this variable most likely to be correlated with anomalies in NEE.

This results section first discusses the performance of the SARIMA(X) models with respect to the new benchmark of the monthly mean model, mainly focusing on the ME as the priority is minimizing the budget imbalance after a full inversion run. Afterward, the available data is more thoroughly analyzed to find explanations on why the SARIMA(X) models behave as they do. Finally, some observations are made hinting at how well the discussed models could be implemented within CTDAS.

9.1 Performance of the SARIMA(X) models

The monthly mean model results in a larger reduction in ME than both the SARIMA and SARIMAX models. Figure 9.1 shows that across all TransCom regions, the ME of the monthly mean model is closest to 0 of all models, with the ME within the *North American Temperate* and *North American Boreal* TransCom regions approaching 0. The ME of the SARIMA(X) models are only marginally better than the prior model. As the ME is the main performance measure used for determining the bias within the flux-landscape, it provides a strong argument that the additional temporal dependencies utilized by the SARIMA(X) models did not result in a bias reduction. A possible explanation of why the SARIMA(X) models are unable to utilize the temporal dependencies can be found in the variability within the state vector elements.

9.2 A deeper analysis of the available data

Figure 9.2 shows that the effective scaling factor shows seasonal heteroskedasticity, which could potentially limit the effectiveness of the SARIMA(X) models. Both the SARIMA and the SARIMAX models work under the assumption that the time series upon which they are trained is produced by a stationary process (Box et al., 1976). This implies that both the mean scaling factor and the variability of the scaling factor should be independent of time. In other words, the series should have a constant mean and be homoscedastic. As the variance within the time series shows a seasonal variance, the series is heteroscedastic, violating the assumption of time series being stationary. Due to this heteroscedasticity, the SARIMA(X) models might not perform optimally, explaining why these models are unable to reduce the ME compared to the prior model in flux space. Reducing this

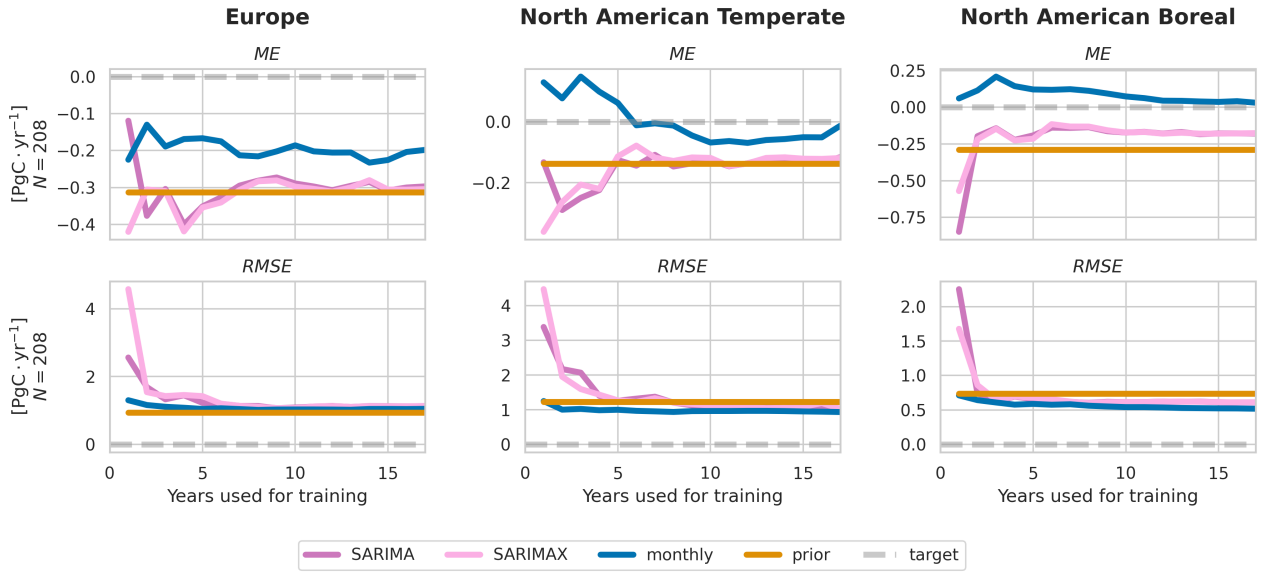


Figure 9.1: The performance of the monthly mean, SARIMA, and SARIMAX models on 4 years of test data (2017-2020) compared to the prior flux model, evaluated in flux space. The x-axis represents the number of years used for training the models, where 1 training year entails the model had been trained only on data from 2016 and 17 training years entails a model trained on the data from 2000 to 2016. The y-axis is either the mean error (ME) or root-mean-square error (RMSE), determined by the weekly difference between the estimated and optimized flux within the *Europe*, *North America Temperate*, and *North America Boreal* TransCom regions ($N = 4 * 52 = 208$). As the y-axes are not aligned, the 'target'-line is added as a visual aid representing the values a well-trained model should approach. The performance on the *Eurasia Temperate* and *Eurasia Boreal* TransCom regions are placed in Figures B.4 and B.5in the appendix, along with other performance measures as this data did not give any additional insights.
Notice how the SARIMA(X) models have very similar performance, with the monthly model performing substantially better with respect to ME but not RMSE. Furthermore, the ME of the monthly model is lower than the prior model after only using a single year of training data and flattens out after using approximately 5 years.

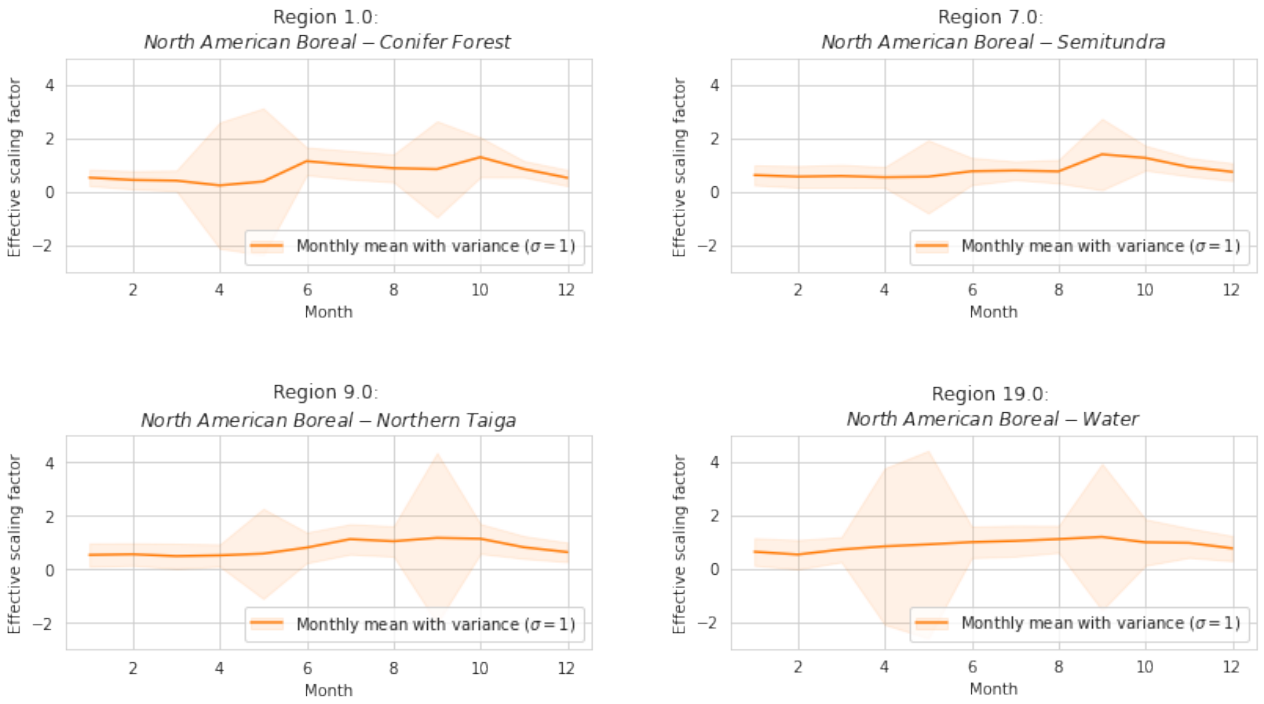


Figure 9.2: Four examples of seasonal heteroskedasticity within the effective scaling factor of ecoregions. The chosen examples are the 4 largest regions within the North American Boreal TransCom region (see Table D.1). The figures show the monthly mean of the effective scaling factor of the first 17 years of training data. The difference in variance is shown by a single standard deviation of the mean, where the variance is substantially bigger in the 4th, 5th, and 9th month compared to all other months.

heteroscedastic behavior of the effective scaling factor should be prioritized in any effort to improve the performance of the SARIMA(X) models. Before the heteroscedastic behavior can be minimized, the origin of this behavior needs to be investigated. This could be related to the correlation between the effective scaling factors and the prior flux.

The extreme values of the effective scaling factor are highly correlated with the prior flux to which the effective scaling factor applies. This is not the case for the unmodified state vector. Figure 9.3 shows that the distribution of extreme scaling factors (i.e those smaller than 0 and greater than 3) is, compared to the unmodified scaling factor, disproportionately centered around a prior flux of 0. This can be explained by the aggregation method used to determine the effective scaling factor, which is essentially a weighted average based on the prior flux. This means that the sum of the weighted scaling factors is divided by the sum of all prior fluxes. However, if this sum approaches 0, the resulting weighted average goes to plus or minus infinity. The prior flux can be 0 during some spring and autumn months when most ecosystems change from being a carbon sink to a carbon source and vice-versa (Baldocchi et al., 2001). While the effect of most extreme cases has been

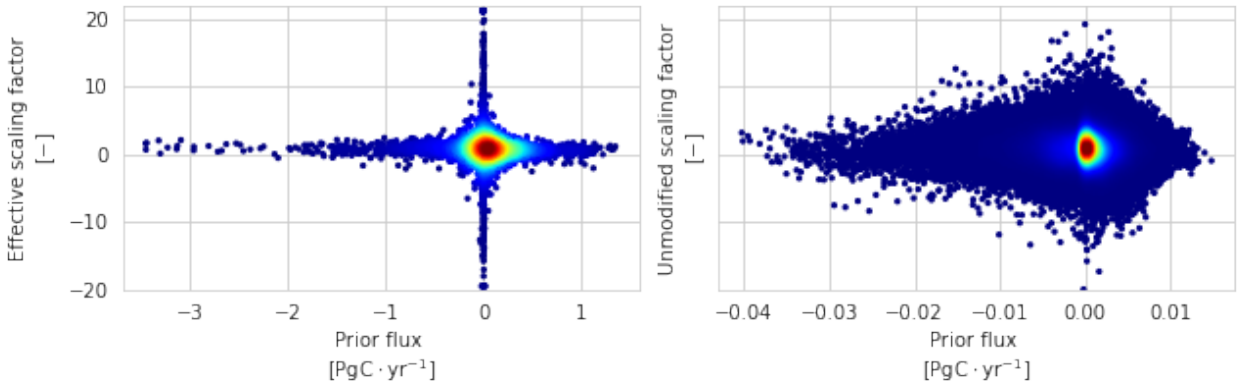


Figure 9.3: The correlation between the prior flux and the first year (2000) of effective analyzed scaling factor ($\boldsymbol{l}^a$; left) and the unmodified analyzed scaling factor ($\boldsymbol{\lambda}^a$; right) of the North American Boreal, North American Temperate, Eurasia Boreal, Eurasia Temperate, and Europe TransCom regions. The hue of the data points represents the density of data points within the plot and is determined using Gaussian kernel density estimation from the SciPy package within Python (Virtanen et al., 2020). The plot shows that the variance within the effective scaling factor (left) is highly correlated with the prior flux, with all extreme effective scaling factors being located near a flux prior flux of 0. This is not the case for the unmodified scaling factor (right), where the extremes are more evenly spread across the prior fluxes.

mitigated by introducing a cap on the effective scaling factor of $\pm 4\sigma$ from the mean, most of the introduced variance remains. This implies that a different aggregation method should be used in future iterations of the project. One option would be to use the absolute prior flux to create a weighted average, but this is only a viable option if no (or barely any) di-poles exist within the to-be-aggregated data. However, the found heteroscedasticity does not explain why the SARIMA and SARIMAX models have such similar performance. This explanation has to be sought in the correlation between the temperature anomalies and the effective scaling factor.

Figure 9.4 shows that there is no linear relationship between the effective scaling factor and the monthly temperature anomaly. The difference between the SARIMA and the SARIMAX models is that the SARIMAX model is able to utilize multiple exogenous variables. Just like a linear regression, the SARIMAX model aims to find a single coefficient that optimally matches the linear correlation between the exogenous variable and the target variable. This means that the correlation between the exogenous variable and the predictor variable has to be linear. Figure 9.4 shows that if any correlation exists between the monthly average temperature anomaly and the effective scaling factor, this correlation does not appear to be linear. As a result, the monthly temperature anomaly will not contribute to the ability of the SARIMAX models to forecast the next effective scaling factor. If the monthly temperature anomaly does not have this linear correlation with the scaling factor, it would be best to first test the other available environmental conditions for such a linear correlation before using them as exogenous variables to the SARIMAX model.

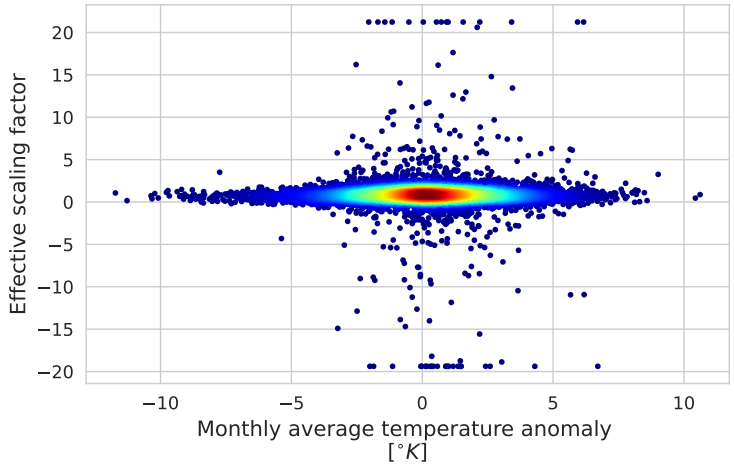


Figure 9.4: The correlation between the effective scaling factor and the mean temperature anomaly. Only the ecoregions which cover more than 10% of their respective TransCom region are taken into account (see Table D.1). Again, the color is used the give an indication of the density and is determined using the Gaussian kernel density estimation function from the SciPy package (Virtanen et al., 2020). The figure shows that there is no linear correlation between the effective scaling factor and the monthly temperature anomaly. A test of the Pearson's correlation coefficient (r) between the effective scaling factor and the monthly temperature anomaly within each of the tested ecoregions further confirms this finding as the r values ranged between -0.02 and 0.04

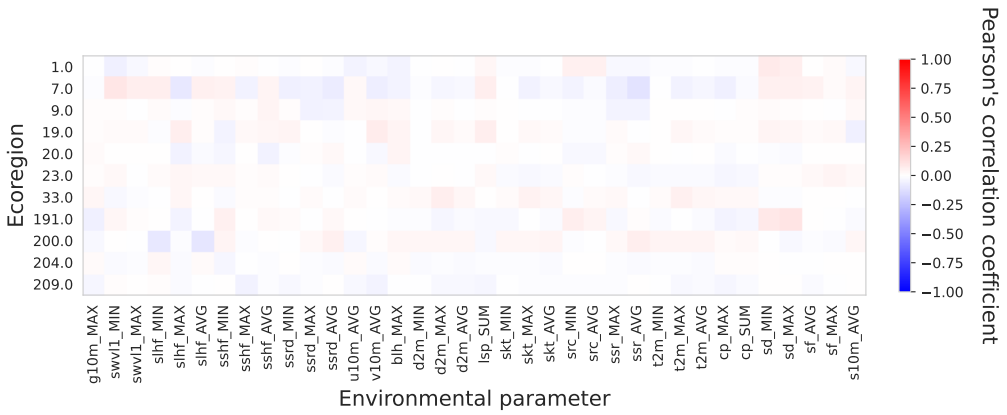


Figure 9.5: Pearson's correlation coefficient between all other environmental conditions gathered for their potential of being a predator of the scaling factor within ecoregions which cover at least 10% of the surface of their TransCom region. For a translation of the index of the ecoregion to the actual name of the ecoregion, see Table D.4. The largest linear correlation is between the `ssr_AVG` and the effective scaling factor of ecoregion 7.0 (*North American Boreal - Semitundra*), where $r = -0.11$. While not entirely equal to 0, the correlation remains weak at best.

Evidently, this sought linear correlation between the effective scaling factor and an exogenous variable does not exist within any of the included environmental conditions. Figure 9.5 shows that all correlation coefficients between the environmental variables and the effective scaling factor are close to 0. This implies that any effort to include other variables into the SARIMAX model will therefore not result in an improved forecast model. However, there are still some positive observations that can be made from the conducted experiments.

9.3 The fit within CTDAS

The final aspect of the evaluation of the conducted experiments focuses on how well the proposed models would fit within CTDAS. This is judged based on how much data is needed to train the model. As running the EKF within CTDAS is computationally expensive, it is infeasible to do a full inversion run, train the transition model on the produced state vectors, and afterward do a new inversion with the trained model. The to-be-implemented model should be able to learn during the inversion run, where it would be beneficial to have a transition model that can improve the performance of the entire inversion system as early on in the run as possible. The amount of data needed to train each of the proposed models can be found back in Figure 9.1.

Figure 9.1 shows that the monthly mean model already achieves a substantial reduction in the ME over the prior model using only a single year of training data. The absolute ME of the monthly mean model using only a single year of training data is comparable to the ME of the prior flux model within

the *North American Temperate* TransCom region (both around 0.13 [PgC $\cdot$ yr^{-1}]), substantially better in the *Europe* TransCom region (0.22 to 0.33 [PgC $\cdot$ yr^{-1}]), and almost fully compensated in the *North American Boreal* TransCom region (0.06 to 0.29 [PgC $\cdot$ yr^{-1}]). The RMSE of the monthly mean model remains relatively constant and only improves marginally as more training data is used. This implies that the monthly mean model could provide an improvement to CTDAS using only a single year of training data.

Furthermore, Figure 9.1 also shows that all models converge after approx 5 years of training data. Around using 5 years of training data, the absolute mean error of the monthly mean model converges to around 0.17, 0.06, and 0.12 [PgC $\cdot$ yr^{-1}] for the *Europe*, *North American Temperate*, and *North American Boreal* TransCom regions respectively. While the SARIMA(X) models do improve with more used training data and also converge around 5 years, the absolute ME remains comparable to the prior model in the *Europe* and *North American Temperate* TransCom regions and is only substantially reduced in the *North American Boreal* TransCom region (0.19 to 0.29 [PgC $\cdot$ yr^{-1}]). As a result, the optimal performance of all models can be expected after training on approximately 5 years of scaling factors. This also gives a hint on how a potential rolling mean model would suffice with a window of 5 years.

Chapter 10: Discussion

This part focused on the implementation of the SARIMA and SARIMAX models. The results indicate that these linear forecast models did not outperform the monthly mean model with respect to the ME in flux space, which has been determined in part I to be the main evaluation criteria as it most likely to resemble the budget imbalance during a full inversion run. A possible explanation for why the SARIMA(X) models are unable to improve upon the monthly model can be found in the heteroscedasticity of the effective scaling factor. The aggregation method used to determine the effective scaling factor per ecoregion introduces additional variance if the prior flux to which the scaling factor is associated approaches 0. While these scaling factors are not as interesting as they have a limited effect on the combined flux landscape, this introduced variance could complicate the extracting of the temporal patterns within the effective scaling factor.

The reason why the SARIMAX model has almost identical ME and RMSE to the SARIMA model, can be found in the correlation between the effective scaling factor and the monthly temperature anomaly that was used as the predictor variable. This correlation, if it exists at all, is (highly) non-linear. SARIMAX can only utilize linear relationships and thus the inclusion of the monthly temperature anomaly did not provide any useful information to the SARIMAX model. None of the variables which were selected to potentially correlate with the biases in the biosphere model had a substantial linear correlation to the effective scaling factor. Several potential reasons can be given for why these linear relations were not found. The first one again relates to the additional variance introduced by the aggregation method used to determine the effective scaling factor. As this additional variance is not related to the environmental conditions, it could obscure a small linear relation that does exist. The second potential reason relates to how the effect of anomalies in environmental conditions on the biases within the biosphere last longer than a single week. It has been shown that severe droughts can affect the NEE within an ecosystem for several years after this drought has occurred (Kannenberg et al., 2020). Therefore an environmental anomaly at t could also affect the following week's fluxes, or potentially all fluxes of the following year. This temporal dependency is not captured in a linear correlation. A third reason could be that a correlation does exist, but that this correlation is (highly) non-linear. This would make sense as a linear relation between a model error and an environmental condition is relatively easy to find. If it is easy to find, it is more likely that the researchers who build the biosphere model focus their efforts to correct this model error. As such, it is likely that any correlation between the biases within the biosphere and environmental conditions is difficult to find, and thus (highly) non-linear.

The final subsection of the results, Section 9.3, focused on the integration of the developed models within CTDAS. As the ME of the SARIMA(X) models in flux space is comparable to the ME of the prior model, neither model is the current implementation a viable candidate to replace the currently implemented smoother transition model. There are however a few possible improvements to the current implementation which could potentially improve the performance of the SARIMA(X) models.

10.1 Potential improvements

As already mentioned in the results section, the used aggregation method includes additional variance which could limit the effectiveness of the SARIMA(X) models. A relatively simple 'fix' would be to use a weighted average based on the absolute prior flux. This method has a similar behavior as the currently implemented effective scaling factor, except that no division is needed where the denominator approaches 0. A possible downside is that this method will produce non-sense scaling factors if the set of the to-be-aggregated fluxes contains both positive and negative values. The extent to how much this would be an issue should be investigated.

Instead of reducing the seasonal heteroskedasticity within the effective scaling factor, another approach would be to use a model more resistant to this heteroscedasticity. The autoregressive conditional heteroscedasticity (ARCH)-model has been specifically designed to capture the conditional variability, with the generalized autoregressive conditional heteroscedasticity (GARCH)-model providing a more flexible lag-structure (Bollerslev, 1986). By subtracting the GARCH-model from an ARMA model, the conditional variability is in a sense filtered out of the time series before it is passed down to the ARMA model. This principle could be simplified even further by dividing the 0-centered mean-stationary process by the variability of a rolling window (Stockhammar & Öller, 2012).

A totally different approach would be to change the target data of the ML model from a scaling factor to the flux corrections made by the EKF. Suppose we refer to these corrections as $\boldsymbol{\mu}$, which is of the same size as the original scaling vector $\boldsymbol{\lambda}$ with s elements (i.e. $\boldsymbol{\mu}, \boldsymbol{\lambda} \in \mathbb{R}^s$). This $\boldsymbol{\mu}$ is then defined as

$$\boldsymbol{\mu} = \boldsymbol{\lambda} \odot \boldsymbol{f}^{prior} - \boldsymbol{f}^{prior} = \boldsymbol{f}^{prior} \odot (\boldsymbol{\lambda} - 1), \tag{21}$$

where $\boldsymbol{f}^{prior} \in \mathbb{R}^s$ is a vector of the prior fluxes associated with each element in the scaling vector and $\odot$ is element-wise multiplication operator (i.e. the Hadamard product). This approach would completely mitigate the additional variance introduced by prior fluxes being close to 0 as $\boldsymbol{\mu}$ can be aggregated by simply taking the sum of all the elements within an ecoregion, no divisions are needed. There is a risk of the background scaling vector containing exceptionally large values as this division is needed to move back from $\boldsymbol{\mu}$ to $\boldsymbol{\lambda}$ since

$$\boldsymbol{\lambda} = \boldsymbol{\mu} \oslash \boldsymbol{f}^{prior} + 1, \tag{22}$$

where $\oslash$ is element-wise division operator (i.e. the Hadamard division). How this affects the functioning of the EKF should be investigated.

Not only will using the $\boldsymbol{\mu}$ instead of the λ mitigate the issue of heteroscedasticity within the effective scaling factor, but it also solves the issue of evaluating the model in flux space, while training the model on scaling factors. This poses an issue, as the error measure as used for MLE in the coefficient determination of the SARIMA(X) models is now disconnected from the final model evaluation metric of ME in flux space. The results of the SARIMA(X) models in scaling vector space (referring to the $\boldsymbol{\lambda}$ definition of the scaling vector) are shown in figures B.6 and B.7. These figures show that the ME, MAE, RMSE and R^2 of the SARIMA(X) models are comparable to those of the monthly mean model if more than 3 years of training data are used. This gives an indication

that the fit of the SARIMA(X) models with respect to the target data is comparable to the fit of the monthly mean model. However, as this fit in scaling vector space is not guaranteed to result in a good fit in flux space, the resulting flux landscape is substantially different.

Another suggestion for improving the SARIMA(X) approach would be to, just like the monthly model, make a forecast model of a monthly scaling vector instead of a weekly scaling vector. This would most likely reduce the noise within the target data even further as more data points are aggregated making it easier to extract the seasonal patterns. It should however be noted that this does limit the possibilities of including the shorter temporal dependencies. An additional benefit is that the number of steps between the seasonal dependencies is reduced from 52, sometimes 53, steps to 12. This would especially be useful for future gradient descent based recurrent neural network (RNN) implementations, which often suffer more from the vanishing gradient problem the further the temporal dependencies are separated from each other (Hochreiter, 1998).

A final suggestion for the improvement of the SARIMAX model, would be to perform an extensive analysis of the relationship between extreme environmental conditions and biases within the biosphere and use this analysis to find a representation of the environmental condition which results in a linear correlation between this variable and the scaling factor. We know that the scaling factors are driven by environmental conditions, and thus there should be some correlation between the two. One approach would be to use principal component analysis (PCA) to compress multiple environmental variables into a lower dimensionality. This could be effective as major dependencies exist between the environmental conditions. An increase in solar radiation results in a higher temperature, but a high soil moisture content could reduce this increase in temperature. As such, a linear combination of various environmental conditions could provide a set of variables better capable of capturing the relation between extreme conditions and the scaling factor.

10.2 Alternative ML models

The monthly average, SARIMA and SARIMAX models are intended to be a first attempt at capturing the transition model of the EKF. The decision to start with the implementation of the SARIMA(X) models has been based mainly on the two criteria with the highest priority, the ability to be 'trained' on limited data and easy integration within CTDAS. Once training data is shaped such that the SARIMA(X) models can provide a meaningful improvement to the prior model, more powerful models can be considered as well. Next to the addition of using the GARCH model, other potential models could focus more on multi-variate outputs.

In the current setting, a single model is trained for every ecoregion. This training is done independently from each other, i.e. the spatial relation between the ecoregions is completely disregarded. However, one of the known issues with inversion methods is the prominence of dipoles, where an error in one region is compensated by surrounding regions (Jacobson et al., 2007b). This spatial information might therefore be valuable when determining the transition function. Some models which could be considered are regression forests, an artificial neural network (ANN), a recurrent neural network (RNN), and the PROPHET model. Each of these models is discussed in more detail below.

Regression forests RFs are used more often in earth science systems and atmospheric inversions (Tramontana et al., 2016) and are considered a classic and reliable choice. In a direct comparison, RFs are shown to outperform a ARIMA model on a time-series forecasting task with less, but arguably less noisy, data (Kane, Price, Scotch, & Rabinowitz, 2014). A more powerful gradient boosting machine, the XGBoost algorithm (Chen & Guestrin, 2016), has shown to perform well in correcting ocean CO_2 sink models using observations and additional predictor variables (Bennington, Gloege, & McKinley, 2022).

Artificial neural networks ANNs are considered to be more powerful, but results are more difficult to interpret than those of a RF. This is considered to be a substantial downside of ANNs within the field of earth system science, where great value is given to the explainability of models (Reichstein, Camps-Valls, Tuia, & Xiang Zhu, 2021a). Nonetheless, ANNs are used in the field of earth system science in which varying degrees of success are reported. While they can be a great tool to be used alongside traditional non-ML based methods for partitioning CO_2 fluxes into photosynthesis and respiration (Tramontana et al., 2020), computational costs remain a hurdle to overcome (Stoffer et al., 2021). However, ANNs could provide a marginal improvement over RFs on the task of time-series forecasting, but are less viable for variable selection (Ahmad, Mourshed, & Rezgui, 2017).

Recurrent neural networks ANNs are not designed for time-series analysis. A RNN is. A variant has been applied to solar radiation forecasting with reasonable success (Faisal et al., 2022). An argument often used against RNNs is their inability to mitigate the vanishing gradient problem. This could pose problems when trying to capture the seasonal dependencies within the scaling factors. However, in direct comparison on a forecasting task with less, but arguably less noisy, data, a simple RNN greatly outperformed an ARMA model (Güldal & Tongal, 2010). A downside of RNNs is that, as already mentioned in the proposed improvements section, their ability to capture the temporal dependencies within the time series diminishes as the lag between the dependencies increases due to the vanishing gradient problem. This limits their effectiveness in capturing the seasonal dependencies present within the scaling vector. As the window in which the temporal dependencies are expected is relatively well known, feeding the lagged scaling factors into an ANN might be more effective (based on correspondence with Prof. Dr. H. Jaeger).

PROPHET Facebook originally developed PROPHET to make time series forecasting accessible for non-experts. By using easily interpretable parameters and providing automatically generated visuals of the trend within the time series, the process of fitting the model is considerably easier than most conventional forecasting methods (Taylor & Letham, 2017). The authors claim the system is particularly well suited for data with a strong seasonal trend, missing observations or large outliers, and historical trend changes such as product launches or logging changes. Furthermore, the PROPHET framework is equipped with tools for multi-variate predictions. Therefore it seems to be a viable alternative to the SARIMA model. Some literature provides evidence that the PROPHET

model can outperform a SARIMA model (McCoy, Pellegrini, & Perlis, 2018), although it should be mentioned that the quality of this literature is often questionable. Nonetheless, it might be worth including in a future comparison due to the ease of implementation.

10.3 Final remarks on the integration within the CarbonTracker data assimilation shell

Most of the discussion points thus far have focused on improving the scaling vector forecast model under the experimental condition of a series of analyzed scaling vectors being available. While this works as a proof of concept, it is still important to keep in mind that this experimental setting is a substantial simplification of the actual problem, namely the missing state transition function of the EKF which has to be learned during the inversion run. Therefore, it is essential that the model can provide an improvement over the prior model as early as possible. If the implemented model only provides an improvement in the ME after 20 years, only a single year remains in which the model is actually useful.

The monthly model did provide a substantial improvement of the prior flux model with respect to the ME using only a single year of training data. After using 5 years of training data, the mean error evens out, suggesting that a moving average of 5 years would suffice as a suitable transition model. The possibility of using a moving average connects well to one final point of discussion.

As mentioned, the series of scaling vectors used within the experimental setup was generated using a stationary transition model. This entails that the procedure generating the scaling vectors remained the same for all scaling vectors. However, if the transition model needs to be trained during the inversion run, the procedure generating the background scaling vectors changes, and this could in turn affect the procedure generating the analyzed scaling vectors. This transforms the series of analyzed scaling vectors from being generated by a stationary process to one generated by a potentially non-stationary process.

This is a problem, as most ML models are trained under the assumption that the used training data is representative of the target data. If the procedure generating the analyzed scaling vectors indeed substantially varies as time progresses, the model tries to make a forecast based on learned patterns that no longer apply. It is known that ARMA-like models are unable to account for these changes (X. Wang et al., 2021). While some variants exist that are able to learn in an online manner (Liu, Hoi, Zhao, & Sun, 2016), it remains uncertain whether such models would also be able to capture the seasonal patterns within the scaling vector. Such online learning methods are most often based on gradient descent, which means that the gradient of a seasonal correlation would need to be propagated for 52 timesteps. This is difficult to do given the noise within the target data and the limited amount of available training data, especially at the beginning of the inversion run. Furthermore, properly testing the online learning methods will be difficult given the substantial costs of doing an inversion run.

Chapter 11: Conclusion

Now it is finally time to summarize our findings and return to the research questions asked in the introduction of the part, starting with the first question:

«*Does the utilization of additional temporal dependencies (i.e. dependencies between time-steps) result in a reduction of the mean bias of the prior biosphere and ocean flux with respect to the monthly average model?*»

During our noise-reducing efforts, it has been decided to not include the ocean fluxes within our optimization efforts. The main focus was put on the gridded ecoregions where we have the most confidence in the corrections made by the EKF. Nonetheless, the SARIMA model used to investigate the effect of including additional temporal dependencies did not reduce the ME, which is analogous to the mean bias, with respect to the monthly mean model. Several potential improvements to the current implementation of the SARIMA model have been proposed, but based on the conducted experiments, it can be concluded that the utilization of additional temporal dependencies did not result in a reduction of the mean bias.

Next to testing whether the utilization of additional temporal dependencies would result in a reduction of the mean bias, the second posed question concerns the inclusion of environmental conditions as predictor variables:

«*Could the utilization of predictor variables (i.e. temperature, precipitation), in combination with the temporal dependencies, result in a reduction in the mean bias of the prior biosphere and ocean flux with respect to the monthly average model?*»

Again, the ocean fluxes have not been taken into account during this part of the thesis as these have been put aside during the noise-reduction process. To test whether a meaningful predictor variable could be used as a predictor variable for an ML model, initial tests were conducted using a SARIMAX trained on not only the temporal dependencies used by the SARIMA model, but also on the monthly temperature anomaly. This additional variable did not result in a substantial reduction in the ME compared to the SARIMA model. A deeper analysis of the correlations between the environmental variables and the effective scaling factor showed that no linear correlation between the two seems to exist, which is a requirement for the SARIMAX to be able to use this information.

Not only did the conducted experiments answer the asked research questions, but the potential integration within the CarbonTracker data assimilation shell (CTDAS) has been taken into account as well. While some uncertainties remain on how the proposed models would behave during a full inversion run, the results shown within this part of the thesis show that the monthly model can achieve a substantial reduction in the ME using only a single year of training data. Combine this with the finding from Part I that the monthly mean model is substantially better than the currently implemented smoother model, it is likely that the integration of the monthly mean model within CTDAS would result in a reduction of the overall budget imbalance.

Acknowledgements

Special thanks to Auke M. van der Woude for the time and effort he put into supervising this project on an almost daily basis. Also thanks to Prof. Dr. Wouter Peters and Dr. Celestine P. Lawrence for their valuable feedback.

Furthermore, I would like to acknowledge the help of Samuel Upton and Anne-Wil van den Berg during our discussions on the integration of machine learning in the field of atmospheric science, and the help of Joost van Hofslot and Joris Oonk during the preparation of my colloquium. Slides to this colloquium are found on the GitHub repository of this project or can be downloaded through this link.

This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative.

References

Ahmad, M. W., Mourshed, M., & Rezgui, Y. (2017). Trees vs neurons: Comparison between random forest and ann for high-resolution prediction of building energy consumption. *Energy and Buildings*, *147*, 77-89. Retrieved from https://www.sciencedirect.com/science/article/pii/S0378778816313937 doi: https://doi.org/10.1016/j.enbuild.2017.04.038

Baldocchi, D., Falge, E., Gu, L., Olson, R., Hollinger, D., Running, S., ... Wofsy, S. (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. Retrieved from https://journals.ametsoc.org/view/journals/bams/82/11/1520-0477_2001_082_2415_fantts_2_3_co_2.xml doi: 10.1175/1520-0477(2001)082<2415:FANTTS>2.3.CO;2

Bastrikov, V., MacBean, N., Bacour, C., Santaren, D., Kuppel, S., & Peylin, P. (2018). Land surface model parameter optimisation using in situ flux data: comparison of gradient-based versus random search algorithms (a case study using orchidee v1.9.5.2). *Geoscientific Model Development*, *11*(12), 4739–4754. Retrieved from https://gmd.copernicus.org/articles/11/4739/2018/ doi: 10.5194/gmd-11-4739-2018

Bennington, V., Gloege, L., & McKinley, G. A. (2022). Variability in the global ocean carbon sink from 1959 to 2020 by correcting models with observations. *Geophysical Research Letters*, *49*(14), e2022GL098632. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2022GL098632 (e2022GL098632 2022GL098632) doi: https://doi.org/10.1029/2022GL098632

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*(3), 307-327. Retrieved from https://www.sciencedirect.com/science/article/pii/0304407686900631 doi: https://doi.org/10.1016/0304-4076(86)90063-1

Bonavita, M., & Laloyaux, P. (2020). Machine learning for model error inference and correction. *Journal of Advances in Modeling Earth Systems*, *12*(12), e2020MS002232. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020MS002232 (e2020MS002232 10.1029/2020MS002232) doi: https://doi.org/10.1029/2020MS002232

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (1976). *Time series analysis: forecasting and control*. John Wiley & Sons.

Camps-Valls, G., Xiang Zhu, X., Tuia, D., & Reichstein, M. (2021). Introduction. In *Deep learning for the earth sciences* (p. 1-11). John Wiley & Sons, Ltd. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119646181.ch1 doi: https://doi.org/10.1002/9781119646181.ch1

Chen, T., & Guestrin, C. (2016). Xgboost: {A} scalable tree boosting system. *CoRR*, *abs/1603.02754*. Retrieved from http://arxiv.org/abs/1603.02754

Chevallier, F., Maksyutov, S., Bousquet, P., Bréon, F.-M., Saito, R., Yoshida, Y., & Yokota, T. (2009). On the accuracy of the co2 surface fluxes to be estimated from the gosat observations. *Geophysical Research Letters*, *36*(19). Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2009GL040108 doi: https://doi.org/10

.1029/2009GL040108

Cox, A., Vermeulen, A., Manning, A., Beyersdorf, A., Zahn, A., Manning, A., ... Loh, Z. (2021). *Multi-laboratory compilation of atmospheric carbon dioxide data for the period 1957-2019; obspack_co2_1_globalviewplus_v6.1_2021-03-01.* NOAA Global Monitoring Laboratory. Retrieved from http://www.esrl.noaa.gov/gmd/ccgg/obspack/data.php?id=obspack_co2_1_GLOBALVIEWplus_v6.1_2021-03-01 doi: 10.25925/20201204

Crespi, A., Petitta, M., Marson, P., Viel, C., & Grigis, L. (2021). Verification and bias adjustment of ecmwf seas5 seasonal forecasts over europe for climate service applications. *Climate*, *9*(12). Retrieved from https://www.mdpi.com/2225-1154/9/12/181 doi: 10.3390/cli9120181

Dee, D. P. (2005). Bias and data assimilation. *Quarterly Journal of the Royal Meteorological Society*, *131*(613), 3323-3343. Retrieved from https://rmets.onlinelibrary.wiley.com/doi/abs/10.1256/qj.05.137 doi: https://doi.org/10.1256/qj.05.137

Faisal, A. F., Rahman, A., Habib, M. T. M., Siddique, A. H., Hasan, M., & Khan, M. M. (2022). Neural networks based multivariate time series forecasting of solar radiation using meteorological data of different cities of bangladesh. *Results in Engineering*, *13*, 100365. Retrieved from https://www.sciencedirect.com/science/article/pii/S2590123022000354 doi: https://doi.org/10.1016/j.rineng.2022.100365

Friedlingstein, P., Jones, M. W., O'Sullivan, M., Andrew, R. M., Bakker, D. C. E., Hauck, J., ... Zeng, J. (2021). Global carbon budget 2021. *Earth System Science Data Discussions*, *2021*, 1–191. Retrieved from https://essd.copernicus.org/preprints/essd-2021-386/ doi: 10.5194/essd-2021-386

Friedlingstein, P., O'Sullivan, M., Jones, M. W., Andrew, R. M., Hauck, J., Olsen, A., ... Zaehle, S. (2020). Global carbon budget 2020. *Earth System Science Data*, *12*(4), 3269–3340. Retrieved from https://essd.copernicus.org/articles/12/3269/2020/ doi: 10.5194/essd-12-3269-2020

Gaubert, B., Stephens, B. B., Basu, S., Chevallier, F., Deng, F., Kort, E. A., ... Yin, Y. (2019). Global atmospheric co_2 inverse models converging on neutral tropical land exchange, but disagreeing on fossil fuel and atmospheric growth rate. *Biogeosciences*, *16*(1), 117–134. Retrieved from https://bg.copernicus.org/articles/16/117/2019/ doi: 10.5194/bg-16-117-2019

Güldal, V., & Tongal, H. (2010). Comparison of recurrent neural network, adaptive neuro-fuzzy inference system and stochastic models in eğirdir lake level forecasting. *Water resources management*, *24*(1), 105–128. doi: 10.1007/s11269-009-9439-9

Gurney, K. R., Law, R. M., Denning, A. S., Rayner, P. J., Baker, D., Bousquet, P., ... Yuen, C.-W. (2003). Transcom 3 co2 inversion intercomparison: 1. annual mean control results and sensitivity to transport and prior flux information. *Tellus B: Chemical and Physical Meteorology*, *55*(2), 555-579. Retrieved from https://doi.org/10.3402/tellusb.v55i2.16728 doi: 10.3402/tellusb.v55i2.16728

Hauck, J., Zeising, M., Le Quéré, C., Gruber, N., Bakker, D. C. E., Bopp, L., ... Séférian, R. (2020). Consistency and challenges in the ocean carbon sink estimate for the global carbon budget. *Frontiers in Marine Science*, *7*, 852. Retrieved from https://www.frontiersin.org/article/10.3389/fmars.2020.571720 doi: 10.3389/fmars.2020.571720

Haynes, K. D., Baker, I. T., Denning, A. S., Wolf, S., Wohlfahrt, G., Kiely, G., ... Haynes, J. M. (2019). Representing grasslands using dynamic prognostic phenology based on biological growth stages: Part 2. carbon cycling. *Journal of Advances in Modeling Earth Systems*, *11*(12), 4440-4465. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018MS001541 doi: https://doi.org/10.1029/2018MS001541

Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *06*(02), 107-116. Retrieved from https://doi.org/10.1142/S0218488598000094 doi: 10.1142/S0218488598000094

Huijnen, V., Williams, J., van Weele, M., Noije, T., Krol, M., Dentener, F., ... Pätz, H. (2010, 10). The global chemistry transport model tm5: Description and evaluation of the tropospheric chemistry version 3.0. *Geoscientific Model Development Discussions*, *3*. doi: 10.5194/gmd-3-445-2010

Intergovernmental Panel on Climate Change. (2021). *Sixth Assessment Report* (Vol. 2021) (No. August). Retrieved 2022-03-24, from https://www.ipcc.ch/report/ar6/wg1/downloads/report/IPCC_AR6_WGI_Headline_Statements.pdf

Jacobson, A. R., Mikaloff Fletcher, S. E., Gruber, N., Sarmiento, J. L., & Gloor, M. (2007a). A joint atmosphere-ocean inversion for surface fluxes of carbon dioxide: 1. methods and global-scale fluxes. *Global Biogeochemical Cycles*, *21*(1). Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2005GB002556 doi: https://doi.org/10.1029/2005GB002556

Jacobson, A. R., Mikaloff Fletcher, S. E., Gruber, N., Sarmiento, J. L., & Gloor, M. (2007b). A joint atmosphere-ocean inversion for surface fluxes of carbon dioxide: 2. regional results. *Global Biogeochemical Cycles*, *21*(1). Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2006GB002703 doi: https://doi.org/10.1029/2006GB002703

Jones, M. W., Andrew, R. M., Peters, G. P., Janssens-Maenhout, G., De-Gol, A. J., Ciais, P., ... Le Quéré, C. (2021a). Gridded fossil co 2 emissions and related o 2 combustion consistent with national inventories 1959–2018. *Scientific Data*, *8*(1), 1–23. doi: 10.1038/s41597-020-00779-6

Jones, M. W., Andrew, R. M., Peters, G. P., Janssens-Maenhout, G., De-Gol, A. J., Ciais, P., ... Le Quéré, C. (2021b). Gridded fossil co2 emissions and related o2 combustion consistent with national inventories 1959–2018. *Scientific Data*, *8*, 2052-4463. Retrieved from https://doi.org/10.1038/s41597-020-00779-6 doi: 10.1038/s41597-020-00779-6

Jung, M., Schwalm, C., Migliavacca, M., Walther, S., Camps-Valls, G., Koirala, S., ... Reichstein, M. (2020). Scaling carbon fluxes from eddy covariance sites to globe: synthesis and evaluation of the fluxcom approach. *Biogeosciences*, *17*(5), 1343–1365. Retrieved from https://bg.copernicus.org/articles/17/1343/2020/ doi: 10.5194/bg-17-1343-2020

Kane, M. J., Price, N., Scotch, M., & Rabinowitz, P. (2014, Aug 13). Comparison of arima and random forest time series models for prediction of avian influenza h5n1 outbreaks. *BMC Bioinformatics*, *15*(1), 276. Retrieved from https://doi.org/10.1186/1471-2105-15-276 doi: 10.1186/1471-2105-15-276

Kannenberg, S. A., Schwalm, C. R., & Anderegg, W. R. L. (2020). Ghosts of the past: how drought legacy effects shape forest functioning and carbon cycling. *Ecology Letters*, *23*(5), 891-901. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/ele.13485 doi: https://doi.org/10.1111/ele.13485

Keeling, C. D. (1960). The concentration and isotopic abundances of carbon dioxide in the atmosphere. *Tellus*, *12*(2), 200-203. Retrieved from https://doi.org/10.3402/tellusa.v12i2.9366 doi: 10.3402/tellusa.v12i2.9366

Keeling, C. D., Bacastow, R. B., Bainbridge, A. E., Jr., C. A. E., Guenther, P. R., Waterman, L. S., & Chin, J. F. S. (1976). Atmospheric carbon dioxide variations at mauna loa observatory, hawaii. *Tellus*, *28*(6), 538-551. Retrieved from https://doi.org/10.3402/tellusa.v28i6.11322 doi: 10.3402/tellusa.v28i6.11322

Lasslop, G., Reichstein, M., Papale, D., Richardson, A. D., Arneth, A., Barr, A., ... 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. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2486.2009.02041.x doi: https://doi.org/10.1111/j.1365-2486.2009.02041.x

Le Quéré, C., Moriarty, R., Andrew, R. M., Canadell, J. G., Sitch, S., Korsbakken, J. I., ... Zeng, N. (2015). Global carbon budget 2015. *Earth System Science Data*, *7*(2), 349–396. Retrieved from https://essd.copernicus.org/articles/7/349/2015/ doi: 10.5194/essd-7-349-2015

Liu, C., Hoi, S. C., Zhao, P., & Sun, J. (2016). Online arima algorithms for time series prediction. In *Thirtieth aaai conference on artificial intelligence.*

McCoy, J., Thomas H., Pellegrini, A. M., & Perlis, R. H. (2018, 11). Assessment of Time-Series Machine Learning Methods for Forecasting Hospital Discharge Volume. *JAMA Network Open*, *1*(7), e184087-e184087. Retrieved from https://doi.org/10.1001/jamanetworkopen.2018.4087 doi: 10.1001/jamanetworkopen.2018.4087

Olson, J. (1992). World ecosystems (we1. 4), digital raster data on a 10-minute geographic 1080x2160 grid. *Global Ecosystems Database, Ver. 1.0: Disk A, NGDC.*.

Paris agreement [UN Treaty]. (2015, 12 12). Retrieved from https://treaties.un.org/pages/ViewDetails.aspx?src=TREATY&mtdsg_no=XXVII-7-d&chapter=27&clang=_en

Peters, W., Jacobson, A. R., Sweeney, C., Andrews, A. E., Conway, T. J., Masarie, K., ... Tans, P. P. (2007). An atmospheric perspective on north american carbon dioxide exchange: Carbontracker. *Proceedings of the National Academy of Sciences*, *104*(48), 18925-18930. Retrieved from https://www.pnas.org/doi/abs/10.1073/pnas.0708986104 doi: 10.1073/pnas.0708986104

Peters, W., Krol, M. C., van der Werf, G. R., Houweling, S., Jones, C. D., Hughes, J., ... Tans, P. P. (2010). Seven years of recent european net terrestrial carbon dioxide exchange constrained by atmospheric observations. *Global Change Biology*, *16*(4), 1317-1337. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2486.2009.02078.x doi: https://doi.org/10.1111/j.1365-2486.2009.02078.x

Peters, W., Miller, J. B., Whitaker, J., Denning, A. S., Hirsch, A., Krol, M. C., ... Tans, P. P. (2005).

An ensemble data assimilation system to estimate co2 surface fluxes from atmospheric trace gas observations. *Journal of Geophysical Research: Atmospheres*, *110*(D24). Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2005JD006157 doi: https://doi.org/10.1029/2005JD006157

Peters, W., van der Velde, I. R., Van Schaik, E., Miller, J. B., Ciais, P., Duarte, H. F., ... White, J. W. C. (2018). Increased water-use efficiency and reduced co_2 uptake by plants during droughts at a continental scale. *Nature geoscience*, *11*(10), 744–748. doi: 10.1038/s41561-018-0212-7

Petetin, H., Bowdalo, D., Bretonnière, P.-A., Guevara, M., Jorba, O., Armengol, J. M., ... Pérez Garcia-Pando, C. (2021). Model output statistics (mos) applied to cams o_3 forecasts: trade-offs between continuous and categorical skill scores. *Atmospheric Chemistry and Physics Discussions*, *2021*, 1–36. Retrieved from https://acp.copernicus.org/preprints/acp-2021-864/ doi: 10.5194/acp-2021-864

Poruschi, L., Dhakal, S., & Canadell, J. (2010). *Gcp. 2010. ten years of advancing knowledge on the global carbon cycle and its management.* Tsukuba: Global Carbon Project. Retrieved from https://www.globalcarbonproject.org/global/pdf/GCP_10years_med_res.pdf

Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2017). Physics informed deep learning (part i): Data-driven solutions of nonlinear partial differential equations. *arXiv preprint arXiv:1711.10561*.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven earth system science. *Nature*, *566*(7743), 195–204. doi: 10.1038/s41586-019-0912-1

Reichstein, M., Camps-Valls, G., Tuia, D., & Xiang Zhu, X. (2021a). Introduction. In *Deep learning for the earth sciences* (p. 328-330). John Wiley & Sons, Ltd. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119646181.ch1 doi: https://doi.org/10.1002/9781119646181.ch1

Reichstein, M., Camps-Valls, G., Tuia, D., & Xiang Zhu, X. (2021b). Outlook. In *Deep learning for the earth sciences* (p. 328-330). John Wiley & Sons, Ltd. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119646181.ch23 doi: https://doi.org/10.1002/9781119646181.ch23

Rödenbeck, C., Zaehle, S., Keeling, R., & Heimann, M. (2018). How does the terrestrial carbon exchange respond to inter-annual climatic variations? a quantification based on atmospheric co_2 data. *Biogeosciences*, *15*(8), 2481–2498. Retrieved from https://bg.copernicus.org/articles/15/2481/2018/ doi: 10.5194/bg-15-2481-2018

Savitzky, A., & Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, *36*(8), 1627-1639. Retrieved from https://doi.org/10.1021/ac60214a047 doi: 10.1021/ac60214a047

Schuh, A. E., Jacobson, A. R., Basu, S., Weir, B., Baker, D., Bowman, K., ... Palmer, P. I. (2019). Quantifying the impact of atmospheric transport uncertainty on co2 surface flux estimates. *Global Biogeochemical Cycles*, *33*(4), 484-500.

Seabold, S., & Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python.

In *9th python in science conference.*

Sierk, B., Bézy, J.-L., Löscher, A., & Meijer, Y. (2019). The European CO2 Monitoring Mission: observing anthropogenic greenhouse gas emissions from space. In Z. Sodnik, N. Karafolas, & B. Cugny (Eds.), *International conference on space optics — icso 2018* (Vol. 11180, pp. 237 – 250). SPIE. Retrieved from https://doi.org/10.1117/12.2535941 doi: 10.1117/12.2535941

Smith, N. E., Kooijmans, L. M. J., Koren, G., van Schaik, E., van der Woude, A. M., Wanders, N., ... Peters, W. (2020). Spring enhancement and summer reduction in carbon uptake during the 2018 drought in northwestern europe. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1810), 20190509. Retrieved from https://royalsocietypublishing.org/doi/abs/10.1098/rstb.2019.0509 doi: 10.1098/rstb.2019.0509

Stockhammar, P., & Öller, L.-E. (2012). A simple heteroscedasticity removing filter. *Communications in Statistics-Theory and Methods*, *41*(2), 281–299.

Stoffer, R., van Leeuwen, C. M., Podareanu, D., Codreanu, V., Veerman, M. A., Janssens, M., ... van Heerwaarden, C. C. (2021). Development of a large-eddy simulation subgrid model based on artificial neural networks: a case study of turbulent channel flow. *Geoscientific Model Development*, *14*(6), 3769–3788. Retrieved from https://gmd.copernicus.org/articles/14/3769/2021/ doi: 10.5194/gmd-14-3769-2021

Tarantola, A. (2005). *Inverse problem theory and methods for model parameter estimation*. Society for Industrial and Applied Mathematics.

Taylor, S. J., & Letham, B. (2017). *Prophet: forecasting at scale.* Retrieved 2022-10-06, from https://research.facebook.com/blog/2017/2/prophet-forecasting-at-scale/

Tramontana, G., Jung, M., Schwalm, C. R., Ichii, K., Camps-Valls, G., Ráduly, B., ... Papale, D. (2016). Predicting carbon dioxide and energy fluxes across global fluxnet sites with regression algorithms. *Biogeosciences*, *13*(14), 4291–4313. Retrieved from https://bg.copernicus.org/articles/13/4291/2016/ doi: 10.5194/bg-13-4291-2016

Tramontana, G., Migliavacca, M., Jung, M., Reichstein, M., Keenan, T. F., Camps-Valls, G., ... Papale, D. (2020). Partitioning net carbon dioxide fluxes into photosynthesis and respiration using neural networks. *Global Change Biology*, *26*(9), 5235-5253. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/gcb.15203 doi: 10.1111/gcb.15203

van der Laan-Luijkx, I. T., van der Velde, I. R., Krol, M. C., Gatti, L. V., Domingues, L. G., Correia, C. S. C., ... Peters, W. (2015). Response of the amazon carbon balance to the 2010 drought derived with carbontracker south america. *Global Biogeochemical Cycles*, *29*(7), 1092-1108. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2014GB005082 doi: https://doi.org/10.1002/2014GB005082

van der Laan-Luijkx, I. T., van der Velde, I. R., van der Veen, E., Tsuruta, A., Stanislawska, K., Babenhauserheide, A., ... 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. Retrieved from https://gmd.copernicus.org/articles/10/2785/2017/ doi: 10.5194/gmd-10-2785-2017

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... SciPy 1.0 Contributors (2020). Scipy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods*, *17*, 261–272. Retrieved from https://rdcu.be/b08Wh doi: 10.1038/s41592-019-0686-2

Wang, S., Li, C., & Lim, A. (2019). *Why are the arima and sarima not sufficient.* arXiv. Retrieved from https://arxiv.org/abs/1904.07632 doi: 10.48550/ARXIV.1904.07632

Wang, X., Zhang, H., Zhang, Y., Wang, M., Song, J., Lai, T., & Khushi, M. (2021). Learning non-stationary time-series with dynamic pattern extractions. *CoRR*, *abs/2111.10559*. Retrieved from https://arxiv.org/abs/2111.10559

Whitaker, J. S., & Hamill, T. M. (2002). Ensemble data assimilation without perturbed observations. *Monthly Weather Review*, *130*(7), 1913 - 1924. Retrieved from https://journals.ametsoc.org/view/journals/mwre/130/7/1520-0493_2002_130_1913_edawpo_2.0.co_2.xml doi: 10.1175/1520-0493(2002)130<1913:EDAWPO>2.0.CO;2

Yu, X., Orth, R., Reichstein, M., Bahn, M., Klosterhalfen, A., Knohl, A., ... Bastos, A. (2022). Contrasting drought legacy effects on gross primary productivity in a mixed versus pure beech forest. *Biogeosciences*, *19*(17), 4315–4329. Retrieved from https://bg.copernicus.org/articles/19/4315/2022/ doi: 10.5194/bg-19-4315-2022

Acronyms

ACF	autocorrelation function
ANN	artificial neural network
AR	auto-regression
ARCH	autoregressive conditional heteroscedasticity
ARIMA	auto-regressive integrated moving average
ARMA	auto-regressive integrated moving average
CAMS	Copernicus Atmosphere Monitoring Service Information
CE	carbon exchange
CO2M	Copernicus Anthropogenic Carbon Dioxide Monitoring mission
CTDAS	CarbonTracker data assimilation shell
CTE	CarbonTracker Europe
DA	data assimilation
DL	deep learning
DoF	degrees of freedom
ECMWF	European Centre for Medium-Range Weather Forecasts
EKF	ensemble Kalman filter
GA	genetic algorithm
GARCH	generalized autoregressive conditional heteroscedasticity
GCB	Global Carbon Budget
GCP	Global Carbon Project
GFAS	Global Fire Assimilation System
GPP	gross primary production
GridFED	Gridded Fossil Emissions Dataset
I	integration
MA	moving average
MAE	mean absolute error
MAPE	mean absolute percentage error
ME	mean error

ML	machine learning
MLE	maximum likelihood estimation
NEE	net ecosystem exchange
ObsPack	Observation Package
PACF	partial autocorrelation function
PCA	principal component analysis
PFT	plant functional type
R^2	coefficient of determination
RF	regression forest
RMSE	root-mean-square error
RNN	recurrent neural network
SARIMA	seasonal autoregressive integrated moving average
SARIMAX	seasonal autoregressive integrated moving average with exogenous factors
SARMA	seasonal autoregressive moving average
SiB4	Simple Biosphere model 4
TER	total ecosystem respiration

Glossary

atmospheric inversion	The process of optimizing surface fluxes based on atmospheric measurements.
Box-Jenkins method	A process described by Box et al. (1976) on how to find the hyperparameters of a SARIMA model.
data assimilation	The practice of combining different sources of information to estimate possible states of a system as it evolves in time.
diurnal cycle	A pattern which recurs every 24 hours.
eco-region	All cells within a TransCom region which have the same PFT.
endogenous variable	A variable which is influenced by the model. A change in the model would also affect this variable. Antonym of an exogenous variable.
exogenous variable	A variable that is determined outside of the model. It is not influenced by the inner mechanics of the model. Antonym of an endogenous variable.
Hadamard division	The element-wise (or point-wise) division of vectors and matrices.
Hadamard product	The element-wise (or point-wise) multiplication of vectors and matrices.
heteroscedastic	A series which varies in its variance, i.e. not homoscedastic,
homoscedastic	A term describing the variance within a time series. Homoscedasticity implies that the variance within the time series does not change over time. For example, if an arbitrary time series $\boldsymbol{x}$ is considered homoscedastic, the variance within an arbitrary subsection $\boldsymbol{x}$ is the same as the variance within a different subsection of $\boldsymbol{x}$. If this is not the case, time series $\boldsymbol{x}$ is considered to be heteroscedastic.

inversion run	A full atmospheric inversion run. Within the context of CTDAS, this entails using the EKF to find optimal values for the state vector $\boldsymbol{\lambda}$.
Paris agreement	A treaty adopted by 195 countries that commit to combating and mitigating the effects of climate change.
plant functional type	A classification method used for categorizing plants according to their physical, phylogenetic, and phenological characteristics. See Olson (1992) for more information.
predictor variable	A variable that can be used to predict the target variable. In an ML pipeline, it is intuitive to see the predictor variables as the 'input' variables and the target variable as the 'output' variable.
SARIMA(X)	A conjunction of the SARIMA and SARIMAX models. It is used to refer to them both at the same time.
stationary	A term used for specifying that the method of generating a series of numbers does not vary over time. A stationary mean entails that the mean of an arbitrary subsection of an arbitrary time series x is the same as the mean of a different subsection of the same time series x. As a result of a time series being stationary, the time series has a constant mean and is homoscedastic.
target variable	A variable that an ML-model tries to model based on a (set of) predictor variable(s).
TransCom region	One of the 11 land regions and 11 ocean regions which together cover the entire globe as defined by Gurney et al. (2003).
transport model	A model which translates surface fluxes to atmospheric concentrations.

vanishing gradient problem	The vanishing gradient problem is one of the main hurdles preventing very deep gradient decent-based neural networks from learning dependencies between input and output data. As the gradient of the error is determined at the final output layer, multiple integration steps are needed for the gradient to reach the layers closer to the input layer. At each step, the gradient is diminished slightly. In the context of RNNs, the vanishing gradient problem materializes in the difficulty to capture temporal dependencies between time steps separated too far from each other.

Mathematical Symbols

$B^{\mathtt{im}}$	Budget imbalance.
$G^{\mathtt{atm}}$	Gain in atmospheric CO_2 concentrations.
J	Cost function.
$J^{\mathtt{obs}}$	The part of the cost function minimizes the difference with respect to the observations.
$J^{\mathtt{state}}$	The part of the cost function minimizes the difference to the state vector.
$\boldsymbol{\lambda}^a$	The analyzed state vector. Defined in Equation 7.
λ_e	The e-th element from state vector $\boldsymbol{\lambda}$.
$\mathcal{H}(\boldsymbol{\lambda})$	Observation operator of the state vector. Also referred to as the transport model.
$\mathcal{K}(\boldsymbol{\lambda})$	An operator mapping the state vector to a 1×1 degree grid.
$\mathcal{M}$	State transition model.
$\mathcal{M}^{\mathtt{analyzed}}(\boldsymbol{\lambda}_t^a)$	Transition function based on a full inversion run using the SiB4 biosphere model. This model serves as the target model under testing conditions where fluxes cannot be transported to atmospheric concentrations.
$\mathcal{M}^{\mathtt{monthly}}(\boldsymbol{\lambda}^a, t)$	Introduced monthly mean transition model.
$\mathcal{M}^{\mathtt{prior}}$	Prior transition model.
$\mathcal{M}^{\mathtt{smoothed}}(\boldsymbol{\lambda}^a, t)$	Current smoother implementation of the transition model.
$\mathcal{T}(\boldsymbol{F})$	Observation operator on a flux landscape.
$\boldsymbol{F}$	An arbitrary flux landscape.
$\boldsymbol{F}^{\mathtt{analyzed}}$	Flux landscape optimized by a full inversion run of the EKF which used the SiB4 model as the prior biosphere model. Within the context of this thesis, it is also used as F^a.
$\boldsymbol{F}^{\mathtt{bio}}$	The modeled biosphere fluxes.
$\boldsymbol{F}^{\mathtt{fire}}$	The modeled forest fire emissions fluxes.
$\boldsymbol{F}^{\mathtt{fossil}}$	The modeled fossil fuel emissions.
$\boldsymbol{F}^{\mathtt{monthly}}$	Flux landscape estimate of forecast model $\mathcal{M}^{\mathtt{monthly}}$.
$\boldsymbol{F}^{\mathtt{ocean}}$	The modeled biosphere fluxes.
$\boldsymbol{F}^{\mathtt{prior}}$	The prior flux landscape.
$\boldsymbol{F}^{\mathtt{smoothed}}$	Flux landscape estimate of forecast model $\mathcal{M}^{\mathtt{smoothed}}$.

$\boldsymbol{H}$	Linearized matrix form of observation operator $\mathcal{H}$.
$\boldsymbol{K}$	Kalman gain.
$\boldsymbol{L}$	Complete set of all effective scaling factors.
$\boldsymbol{M}$	Linearized matrix of state transition function $\mathcal{M}$.
$\boldsymbol{P}$	covariance matrix $\boldsymbol{\lambda}$.
$\boldsymbol{Q}$	Noise term introduced to the covariance matrix $\boldsymbol{P}$ as a result of an imperfect transition model $\mathcal{M}$.
$\boldsymbol{R}$	covariance matrix $\boldsymbol{y}^{\circ}$.
$\boldsymbol{V}$	Matrix of values of an environmental condition.
$\boldsymbol{\Lambda}$	The set of all analyzed state vectors.
$\boldsymbol{\lambda}$	The state vector. Consists of several scaling factors (λ).
$\boldsymbol{\lambda}^{b}$	The background state vector. Defined in Equation 10.
$\boldsymbol{l}$	Vector of effective scaling factors.
$\boldsymbol{v}$	Exogenous variable.
$\boldsymbol{y}^{\circ}$	Observations.
l_r	The effective scaling factor of the region r.

Appendix A: Additional Information

A.1 Atmospheric gain

The atmospheric gain is a measure of the increase in CO_2 within the atmosphere. Due to anthropogenic activities, the atmospheric gain has been increasing since the industrial revolution, which has been measured using flask measurements at various observation sites. One of these sites is the Mauna Loa site in Hawaii. This is also one of the sites where Keeling (1960) started tracking the CO_2 and noticed a yearly increase in concentrations. He picked this site in particular as the atmosphere at its location is a proper representative of the atmosphere in the northern hemisphere. He hypothesized that this was caused by anthropogenic activities, which was confirmed several years later (Keeling et al., 1976). His research made the scientific community aware of the possible negative effects of burning fossil fuels and provides the basis for modern research into greenhouse gasses and global warming. As such, the measurements from the Mauna Loa measuring station are often used as a reference for atmospheric CO_2 concentrations. Figure A.1 shows how well CTE is able to match the observations of the Mauna Loa observation site. The Manua Loa measurement site, along with 353 other measurement sites, lies at the basis of how the results from CTE and other atmospheric inversion methods are validated.

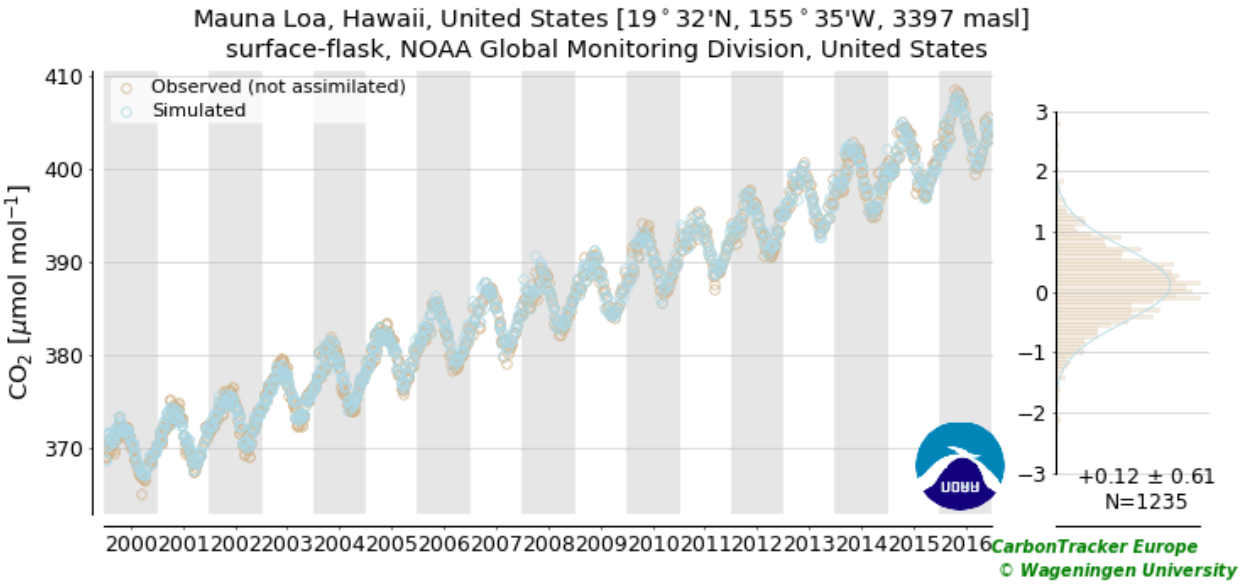


Figure A.1: The plot on the left shows the time series of CO2 mole fractions at the Mauna Loa CarbonTracker observation site. In the top panel, measured mole fractions (open brown circles) are plotted along with CarbonTracker simulated values (light blue open circles). The plot on the right shows the distribution of the residuals between the observations and the model, which should be unbiased (i.e., have a mean of zero) and distributed normally.

A.2 The state vector

The essence of bias correction capabilities of the EKF lies within the state vector $\boldsymbol{\lambda}$. In CTE, each element in $\boldsymbol{\lambda}$ represents a scaling factor for a gridded area on the globe. The aim of the EKF is to find values for these scaling factors such that the combined surface flux, after being transported through the transport model, provides an optimal fit to the observations. How this is done, is discussed in Section 1.2.1. This section provides additional information on how $\boldsymbol{\lambda}$ and its covariance structure is derived.

When defining $\boldsymbol{\lambda}$, all land surface of the earth is divided into 11 TransCom regions (Gurney et al. (2003), see Figure A.2). Each TransCom region is assigned a set of parameters within $\boldsymbol{\lambda}$. How many is determined by how well that TransCom region is constrained by observations. As previously mentioned, some areas (e.g. US, and Europe) are highly constrained by measurements, while others areas (Tropics, Northern Africa) are much less (see Figure 1.2). Therefore, TransCom regions that are on the northern hemisphere have a gridded state vector, where all elements in the state vector work on one 1x1 degree grid-cell. It is assumed that cells within the same plant functional type (PFT) co-vary based on distance, where grid cells closer to each other have a higher expected covariance than cells further away (van der Laan-Luijkx et al., 2017). On the southern hemisphere, one element is added to the state vector for each of the 19 PFTs taken into consideration. An overview of the used PFTs and their respective predominance within Europe is shown in Table D.1. Finally, the 11 ocean TransCom regions are divided into 30 large basins encompassing large-scale ocean circulation features (Jacobson, Mikaloff Fletcher, Gruber, Sarmiento, & Gloor, 2007a).

Due to the correlation length between elements with the same PFT within the gridded TransCom regions, the effective degrees of freedom (DoF) within these regions are greatly reduced. The exact number of DoF added by each TransCom region is determined by applying singular value decomposition to the set of all analyzed state vectors $\boldsymbol{\Lambda}^a$. See Peters et al. (2005) for more details. As a result, the DoFs present within the state vector are reduced to 1077.7, compared to the 9835 elements present within the state vector. For a complete overview, see Table D.2 or the diagram shown in Figure C.1.

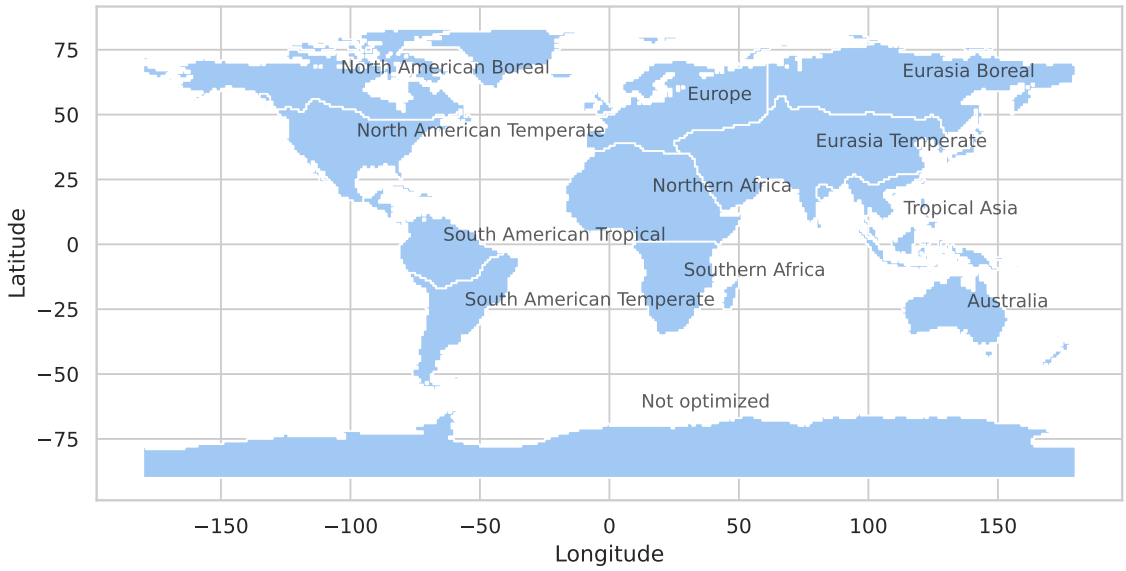


Figure A.2: The geographical boundaries of the land-based TransCom regions as defined in Gurney et al. (2003)

Appendix B: Supplementary Figures

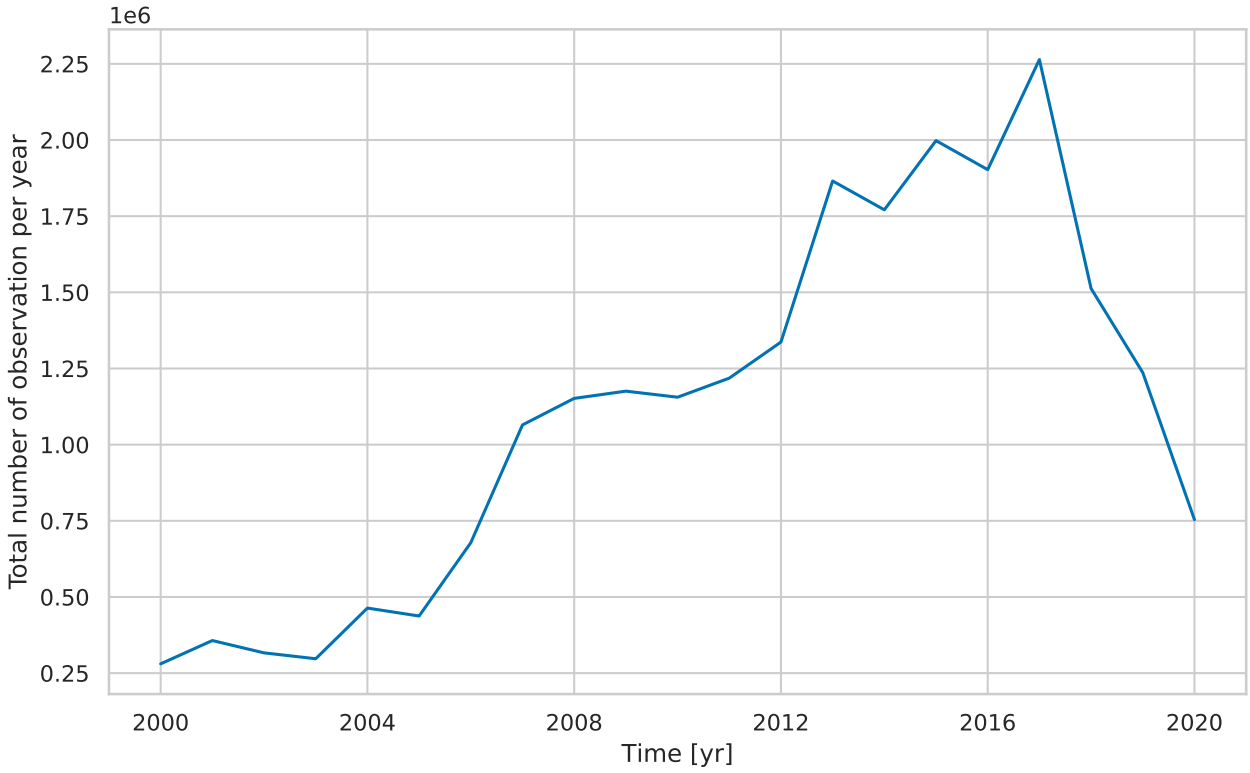
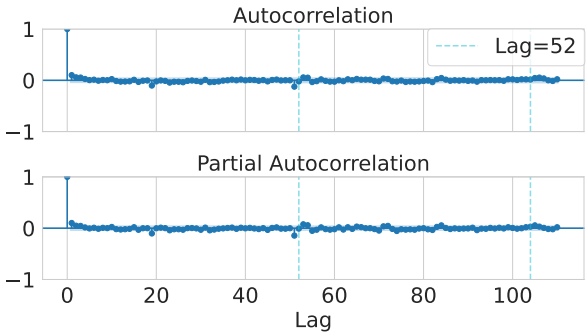
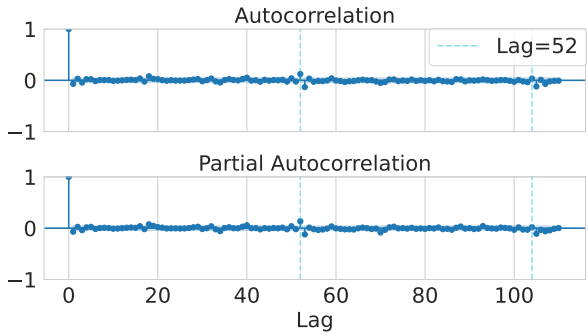


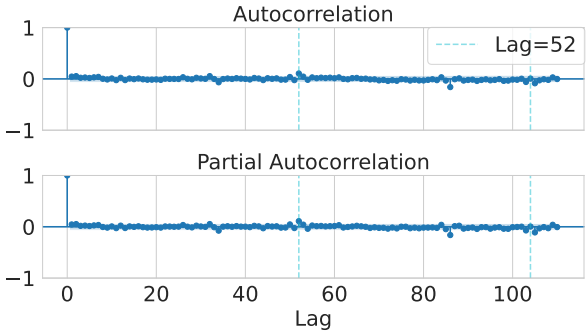
Figure B.1: Total number of observations per year. Observations originate from the sixth release of the GLOBALVIEWplus (GV+) cooperative data product (Cox et al., 2021)



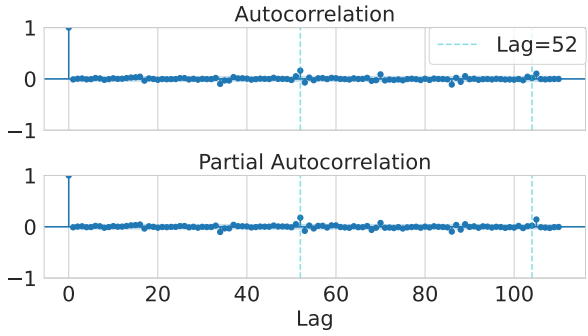
(a) Ecoregion 191.0



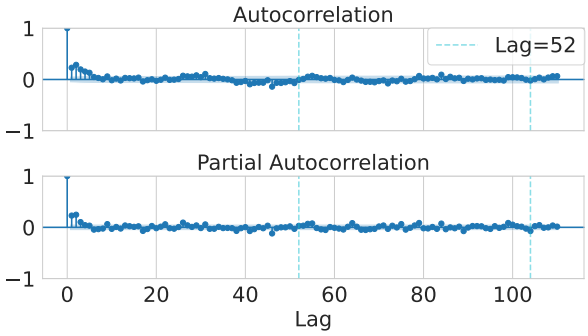
(b) Ecoregion 192.0



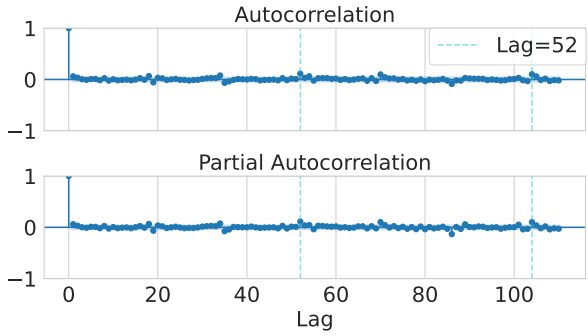
(c) Ecoregion 193.0



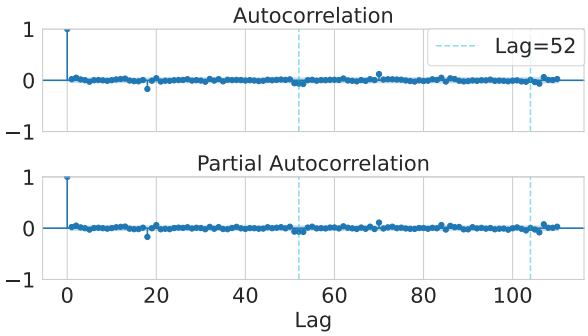
(d) Ecoregion 194.0



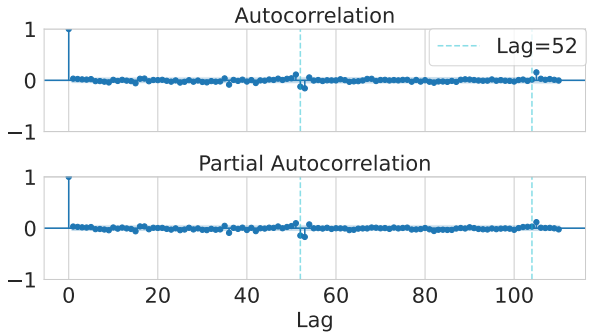
(e) Ecoregion 195.0



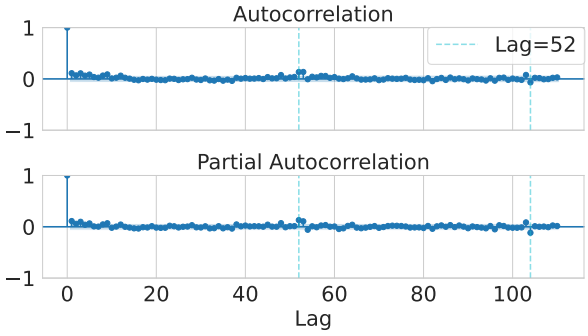
(f) Ecoregion 196.0



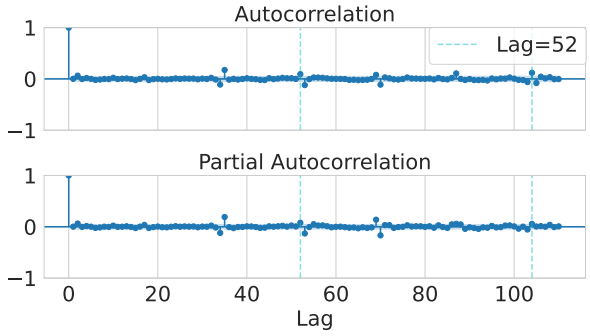
(g) Ecoregion 197.0



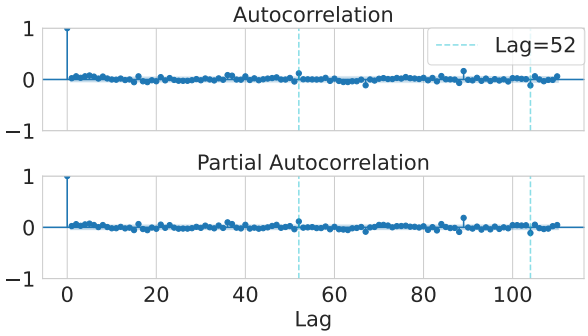
(h) Ecoregion 198.0



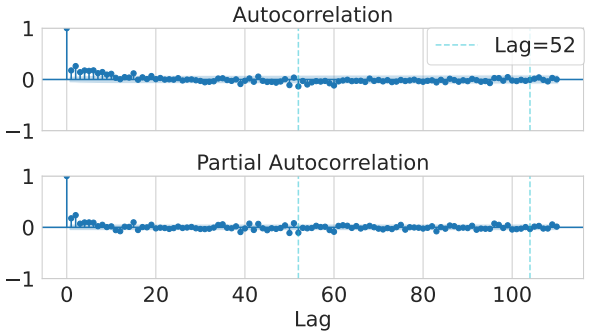
(i) Ecoregion 199.0



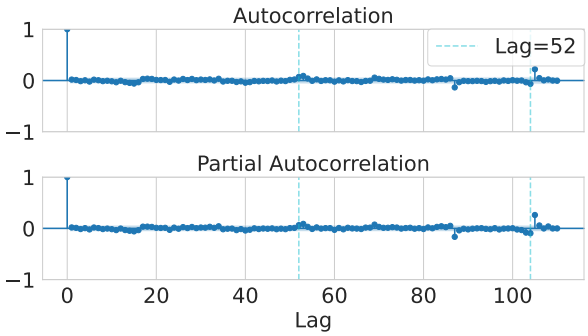
(j) Ecoregion 200.0



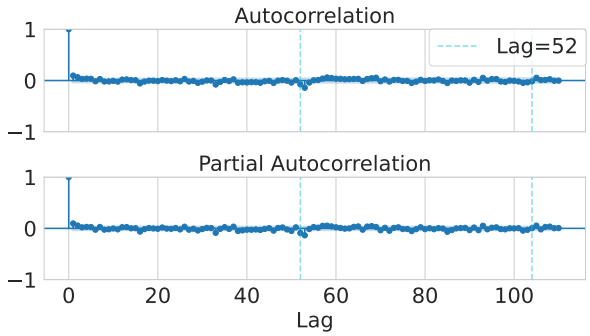
(k) Ecoregion 201.0



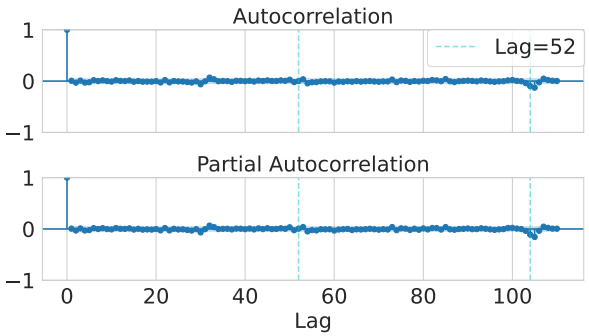
(l) Ecoregion 202.0



(m) Ecoregion 204.0

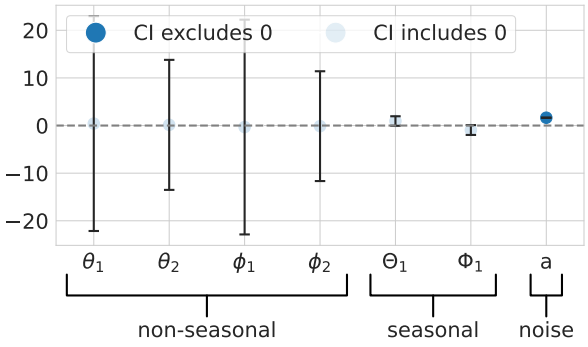


(n) Ecoregion 206.0

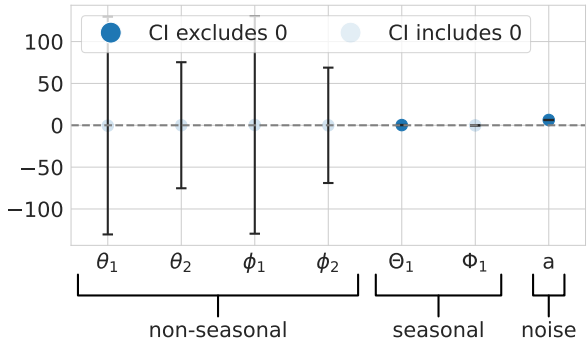


(o) Ecoregion 209.0

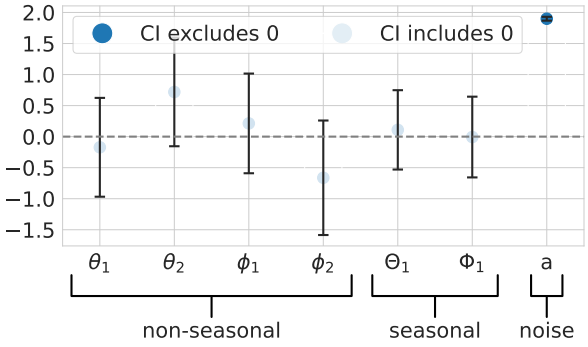
Figure B.2: The autocorrelation and partial autocorrelation of the effective scaling factor of all ecoregions within the *Europe* TransCom region. While there is a lot of variance between the correlations which are deemed significant between the ecoregions, almost all of them show a clear spike in significance at a lag of 52. Figures B.2e and B.2l also show a clear autocorrelation and partial autocorrelation at lag=1 and lag=2.



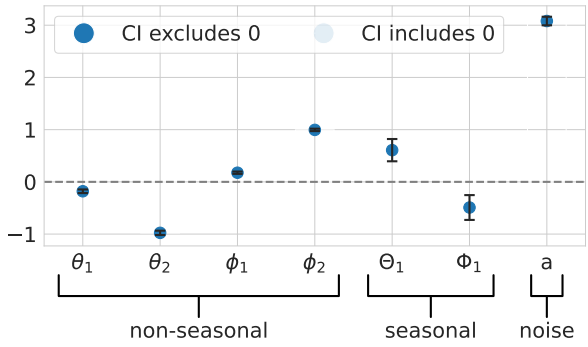
(a) Ecoregion 191.0



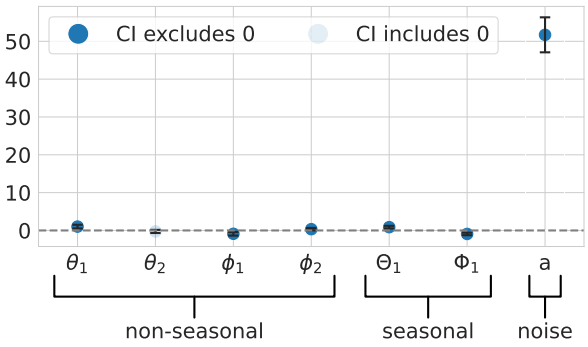
(b) Ecoregion 192.0



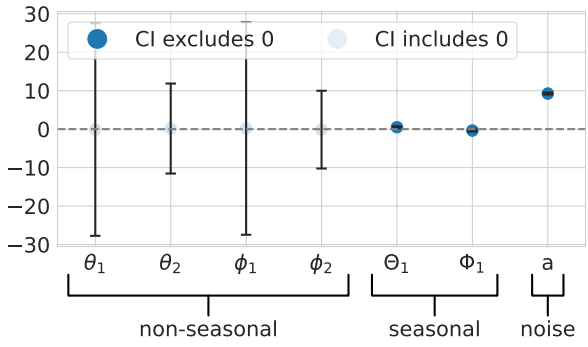
(c) Ecoregion 193.0



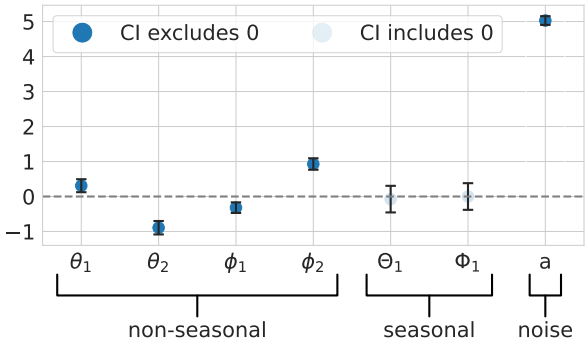
(d) Ecoregion 194.0



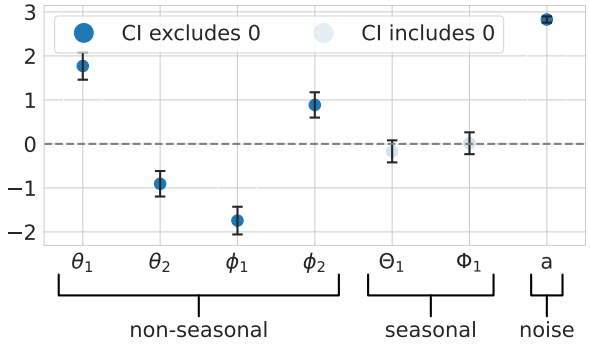
(e) Ecoregion 195.0



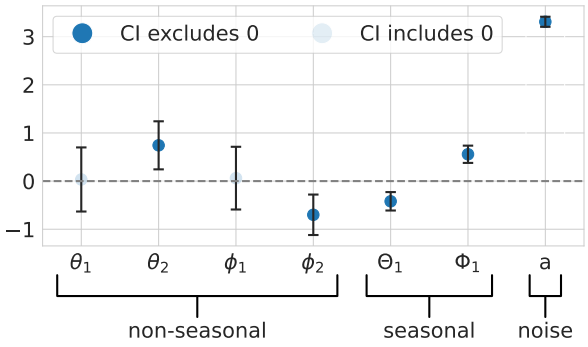
(f) Ecoregion 196.0



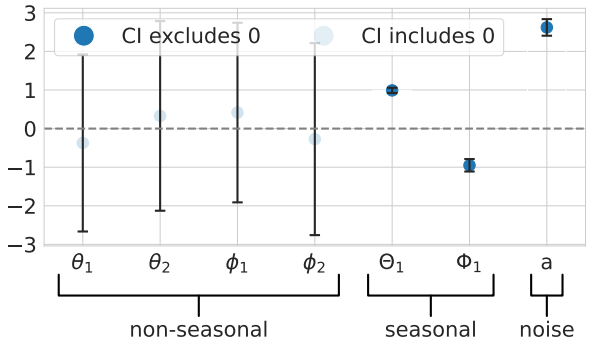
(g) Ecoregion 197.0



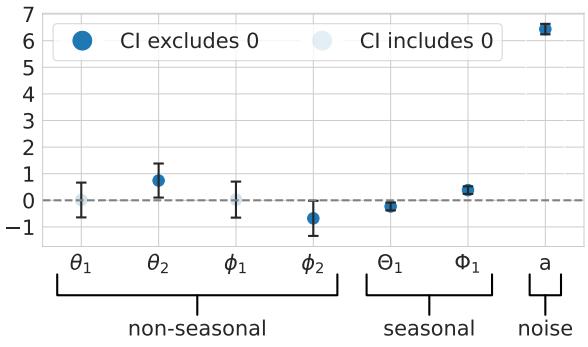
(h) Ecoregion 198.0



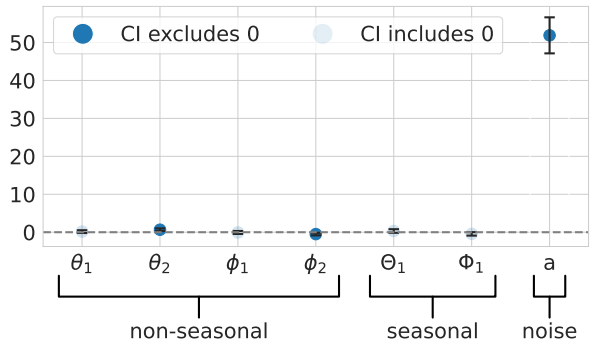
(i) Ecoregion 199.0



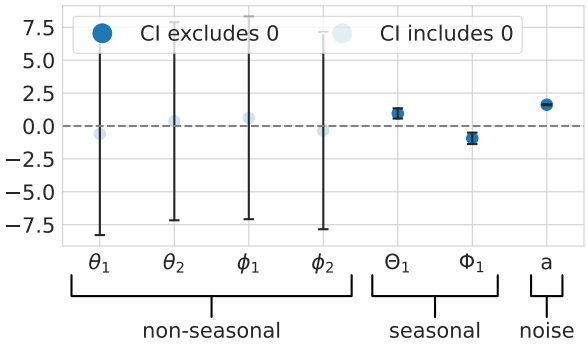
(j) Ecoregion 200.0



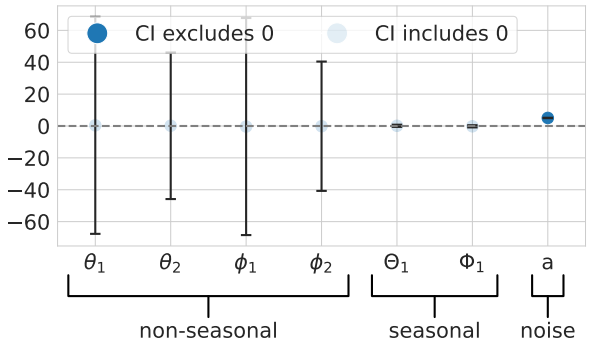
(k) Ecoregion 201.0



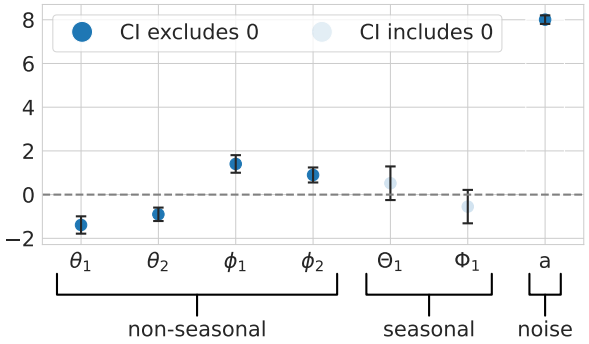
(l) Ecoregion 202.0



(m) Ecoregion 204.0



(n) Ecoregion 206.0



(o) Ecoregion 209.0

Figure B.3: The 95% confidence interval of the coefficients of the SARIMA models within the *Europe* TransCom region. The coefficients are named using the convention of Box et al. (1976). Note that the intercept coefficient is left out. Linking these coefficients back to the $(2,0,2) \times (1,0,1)_{52}$ model definition, the first four coefficients relate to the non-seasonal $(2,0,2)$ part of the model, the 5th and 6th coefficients relate to the seasonal $(1,0,1)_{52}$ part, and 7th and the final coefficient is the noise component that remained after model fitting. Overall, a clear difference is shown between the coefficients which are significantly different from 0 between the models for each ecoregion. Figures B.3j and B.3m show that no non-seasonal coefficients are significant, while figures B.3g and B.3o show that no seasonal coefficients are significant. Figure B.3d shows that all coefficients are significant, while Figure B.3c shows that no coefficients except the noise coefficient are significant. The only thing all plots have in common is that the noise coefficient is significantly greater than 0.

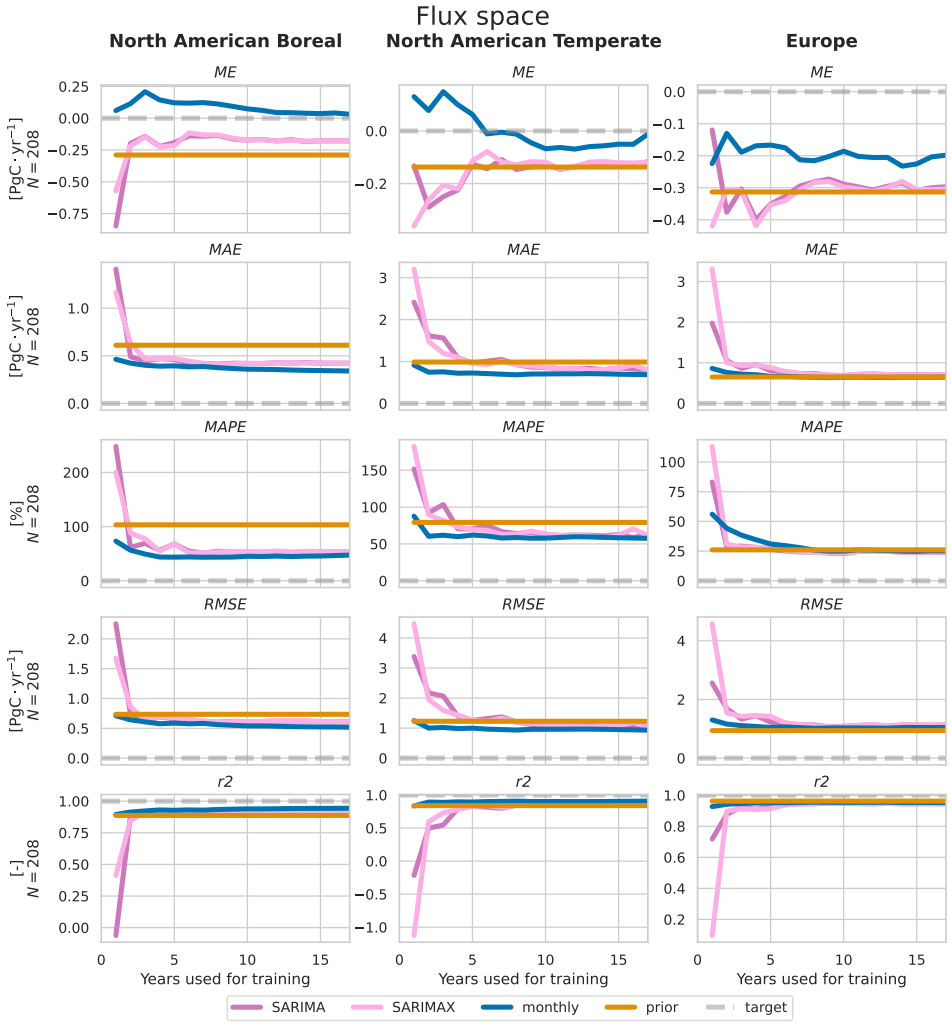


Figure B.4: The performance of the monthly mean, SARIMA, and SARIMAX models on 4 years of test data (2017-2020) compared to the prior flux model, evaluated in flux space. The x-axis represents the number of years used for training the models, where 1 training year entails the model had been trained only on data from 2016 and 17 training years entails a model trained on the data from 2000 to 2017. The y-axis is either the mean error (ME), mean absolute error (MAE), mean absolute percentage error (MAPE), root-mean-square error (RMSE) or coefficient of determination (R^2), determined by the difference between the estimated and optimized flux within the North American boreal, North American temperate and Europe TransCom regions on a weekly basis (N=4*52=208). As the y-axes are not aligned, the 'target'-line is added as a visual aid representing the values a well-trained model should approach.

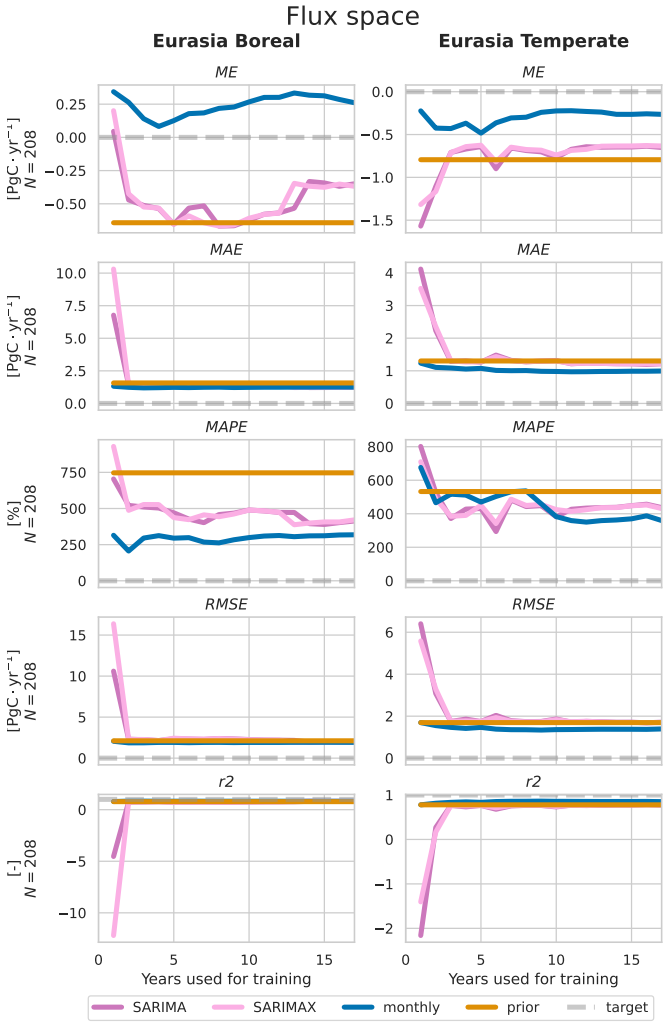


Figure B.5: The performance of the monthly mean, SARIMA, and SARIMAX models on 4 years of test data (2017-2020) compared to the prior flux model, evaluated in flux space. The x-axis represents the number of years used for training the models, where 1 training year entails the model had been trained only on data from 2016 and 17 training years entails a model trained on the data from 2000 to 2017. The y-axis is either the mean error (ME), mean absolute error (MAE), mean absolute percentage error (MAPE), root-mean-square error (RMSE) or coefficient of determination (R^2), determined by the difference between the estimated and optimized flux within the Eurasia boreal and Eurasia temperate TransCom regions on a weekly basis (N=4*52=208). As the y-axes are not aligned, the 'target'-line is added as a visual aid representing the values a well-trained model should approach.

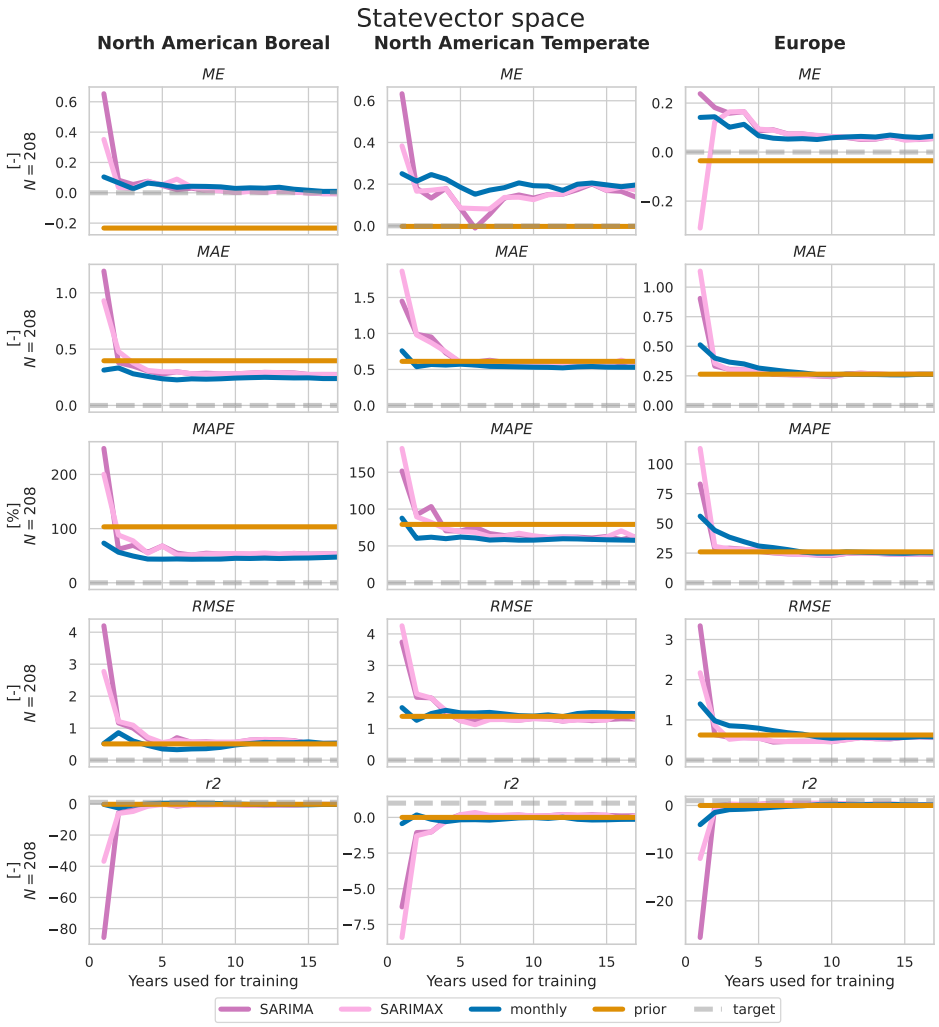


Figure B.6: The performance of the monthly mean, SARIMA, and SARIMAX models on 4 years of test data (2017-2020) compared to the prior scaling factor model of only ones, evaluated in scaling factor space. The x-axis represents the number of years used for training the models, where 1 training year entails the model had been trained only on data from 2016 and 17 training years entails a model trained on the data from 2000 to 2017. The y-axis is either the mean error (ME), mean absolute error (MAE), mean absolute percentage error (MAPE), root-mean-square error (RMSE) or coefficient of determination (R^2), determined by the difference between the estimated and optimized scaling factor within the North American boreal, North American temperate and Europe TransCom regions on a weekly basis (N=4*52=208). As the y-axes are not aligned, the 'target'-line is added as a visual aid representing the values a well-trained model should approach.

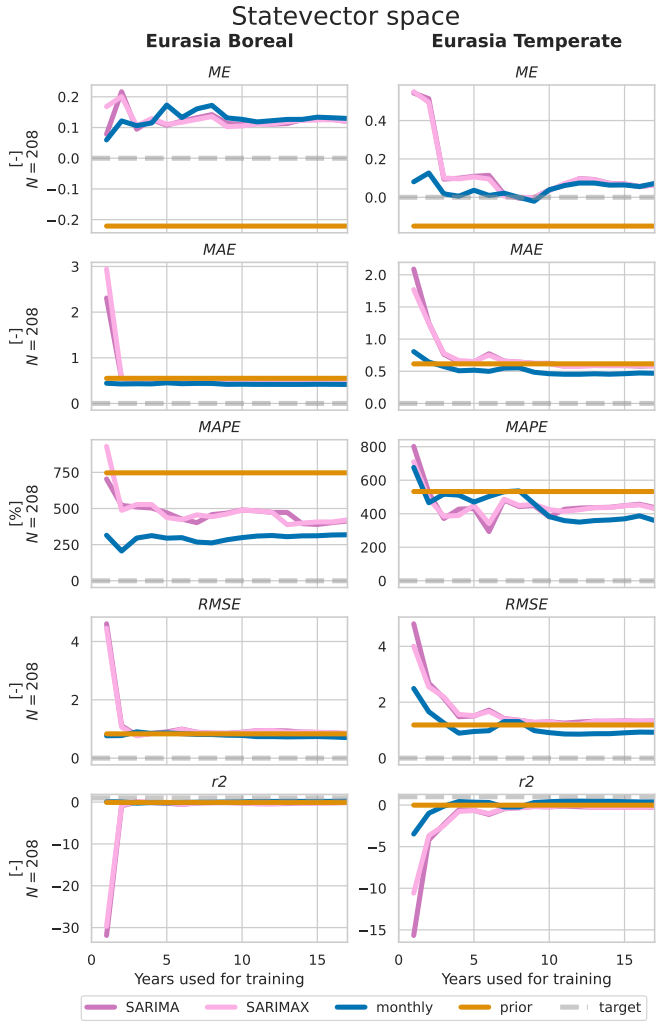
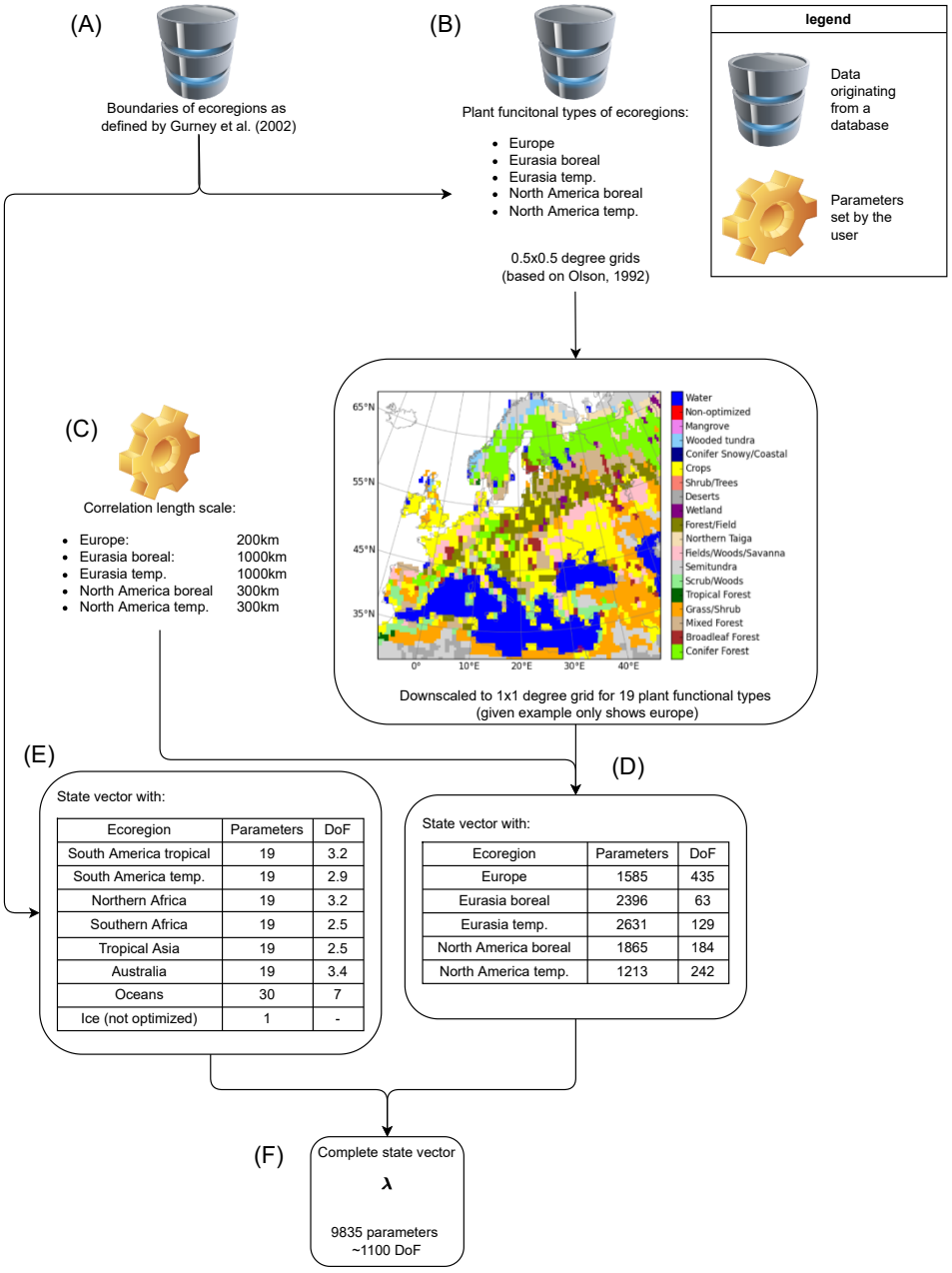


Figure B.7: The performance of the monthly mean, SARIMA, and SARIMAX models on 4 years of test data (2017-2020) compared to the prior scaling factor model of only ones, evaluated in scaling factor space. The x-axis represents the number of years used for training the models, where 1 training year entails the model had been trained only on data from 2016 and 17 training years entails a model trained on the data from 2000 to 2017. The y-axis is either the mean error (ME), mean absolute error (MAE), mean absolute percentage error (MAPE), root-mean-square error (RMSE) or coefficient of determination (R^2), determined by the difference between the estimated and optimized scaling factor within the Eurasia boreal and Eurasia temperate TransCom regions on a weekly basis (N=4*52=208). As the y-axes are not aligned, the 'target'-line is added as a visual aid representing the values a well-trained model should approach.

Appendix C: Supplementary Diagrams



Ecoregion	Parameters	DoF
South America tropical	19	3.2
South America temp.	19	2.9
Northern Africa	19	3.2
Southern Africa	19	2.5
Tropical Asia	19	2.5
Australia	19	3.4
Oceans	30	7
Ice (not optimized)	1	-

Ecoregion	Parameters	DoF
Europe	1585	435
Eurasia boreal	2396	63
Eurasia temp.	2631	129
North America boreal	1865	184
North America temp.	1213	242

Figure C.1: This diagram shows the breakdown of the state vector $\boldsymbol{\lambda}$, which scales the fluxes originating from both $\boldsymbol{F}^{\text{bio}}$ and $\boldsymbol{F}^{\text{ocean}}$. (A) Each entry in the state vector scales some area within the ecoregions defined by Gurney et al. (2003). As some regions are more constrained by observations, more parameters can be used to optimize the fit to the observations. (B) These areas are therefore gridded on a 0.5×0.5 resolution, where each cell is labeled with the predominant plant functional type (PFT) as defined by Olson (1992). (C) Cells with the same PFT are correlated based on a chosen correlation length in order to reduce the effective degrees of freedom (DoF) within the parameters. (D) The chosen correlation lengths within CTDAS result in each gridded ecoregion adding 63 to 435 DoF to the final state vector. (E) The ecoregions less constrained by observations are only given 19 additional entries in $\boldsymbol{\lambda}$, one for each plant functional type (PFT). All oceans are divided into 30 ocean basins according to the division used in Jacobson et al. (2007a). (F) Combining all components results in a state vector of 9835 parameters and $\sim$ 1100 DoF

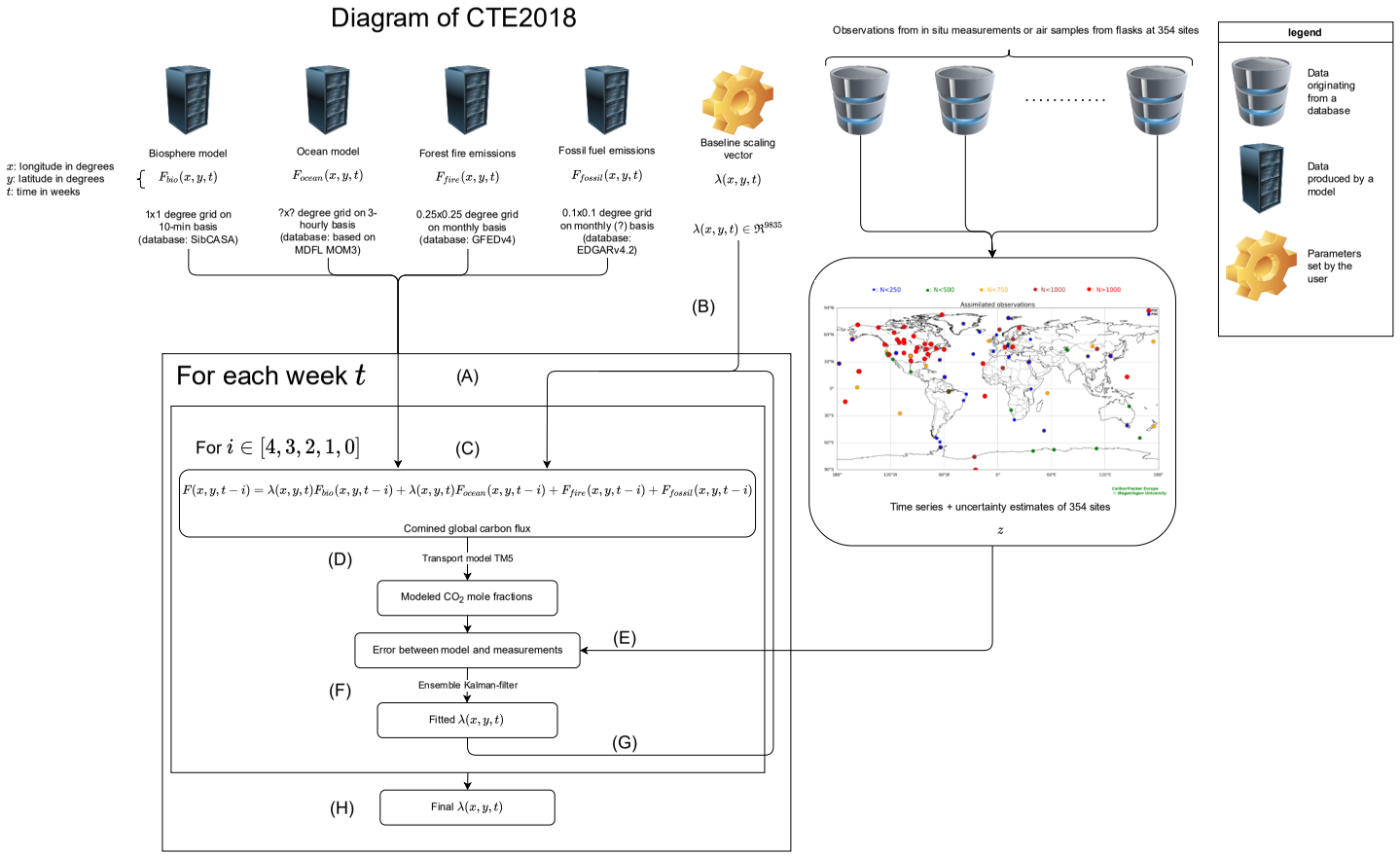


Figure C.2: (A) Posterior fluxes are determined on a weekly basis using data assimilation. (B) A baseline state vector is used to correct for first indications of biases within $\boldsymbol{F}^{\text{bio}}$ and $\boldsymbol{F}^{\text{ocean}}$. This is currently done using Equation 10. (C) The prior fluxes are multiplied with the baseline state vector, starting with the prior fluxes of $t-4$. (D) The resulting fluxes are transported to atmospheric concentrations using the TM5 model (Huijnen et al., 2010). (E) The estimated concentrations are matched to the observations, determining the fit of the current state vector. (F) The fit is optimized using the EKF. (G) The optimized state vector is used as the baseline for the next iteration. (F) The final state vector is determined after being fitted to 5 weeks of data.

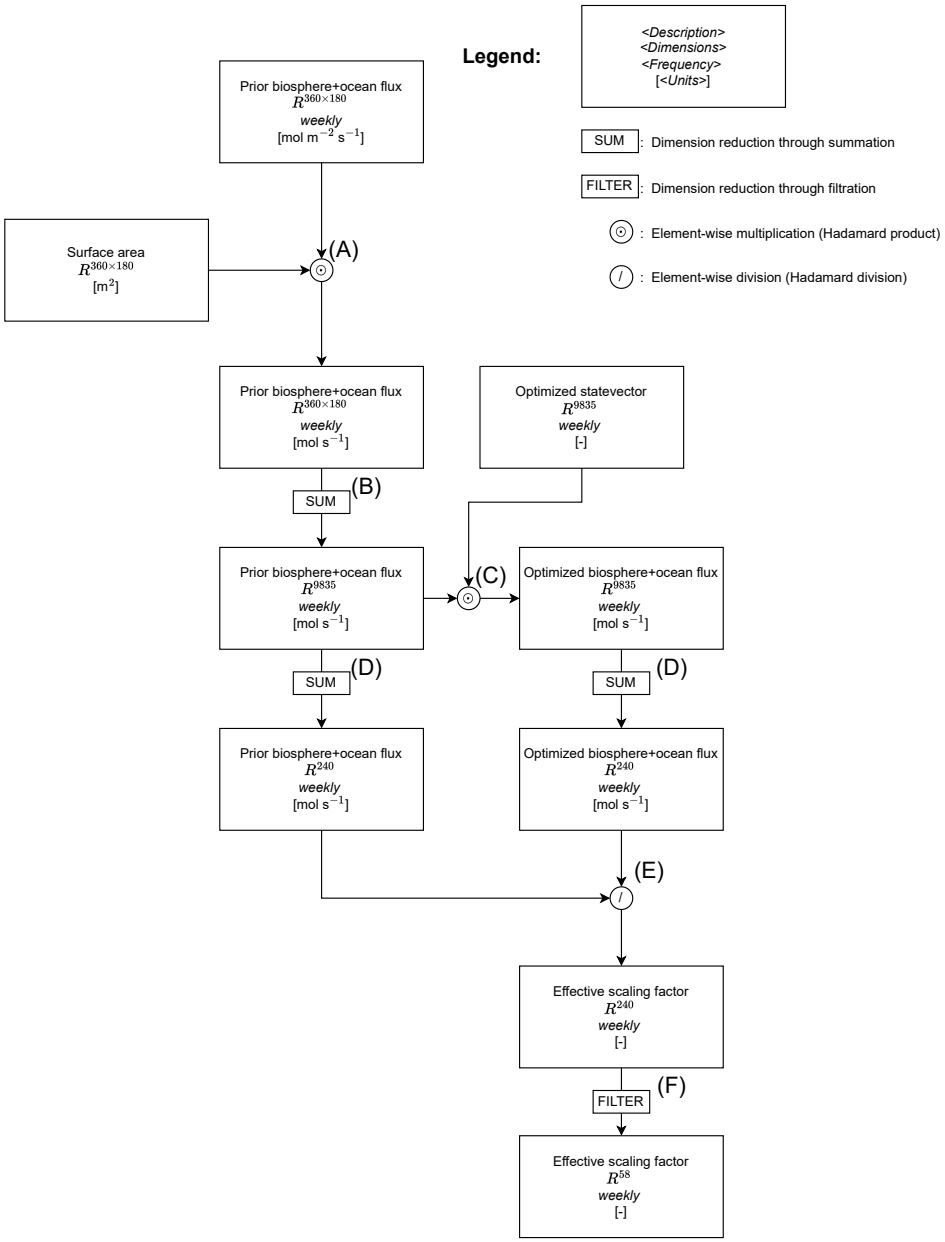


Figure C.3: Description of the aggregation procedure used to derive the target dataset. Both the biosphere fluxes (Haynes et al., 2019) and ocean fluxes (Jacobson et al., 2007b) are reported having the units $[\text{mol} \cdot \text{m}^{-2} \cdot \text{s}^{-1}]$, meaning they are dependent on the surface area of the cells. This dependency is removed by multiplying all fluxes with the surface area of the cell to which they apply (A). To move from grid-cell space to state vector space, fluxes from all cells associated with the same state vector element are summed (B). By applying element-wise multiplication between the fluxes associated with each state vector element and the optimized state vector, the optimized biosphere and ocean fluxes are determined (C). As explained in section 7.1, the gridded state vector elements are summed for each ecoregion to reduce the noise and create a single flux per ecoregion (D). The effective scaling factor is determined by applying an element-wise division of the optimized fluxes over the prior fluxes (E). The final step is to filter out the least constrained TransCom regions (i.e. South American Tropical, South American Temperate, Northern Africa, Southern Africa, Tropical Asia, Australia, and the oceans) (F). What is left, is the effective scaling factor of the ecoregions within the North American Boreal, North American Temperate, Eurasia Boreal, Eurasia Temperate, and Europe TransCom regions.

Appendix D: Supplementary Tables

Table D.1: plant functional type (PFT) categorization derived from Olson (1992), along with their respective area converge within the TransCom regions: *North American Boreal*, *North American Temperate* and *Europe*

		TransCom region surface area [%]		
category	PFT (Olson V 1.3a)	North American Boreal	North American Temperate	Europe
1	Conifer Forest	22.9	14.0	14.0
2	Broadleaf Forest	0.0	2.4	2.5
3	Mixed Forest	5.9	8.1	9.1
4	Grass/Shrub	0.5	21.9	8.1
5	Tropical Forest	0.0	0.5	0.1
6	Scrub/Woods	0.0	3.6	2.8
7	Semitundra	33.6	7.6	4.9
8	Fields/Woods/Savanna	0.3	8.9	6.6
9	Northern Taiga	16.4	0.0	2.2
10	Forest/Field	0.6	10.8	11.6
11	Wetland	3.2	0.6	0.8
12	Deserts	0.0	0.2	0.1
13	Shrub/Tree/Suc	0.0	0.1	0.0
14	Crops	0.0	17.2	22.7
15	Conifer Snowy/Coastal	0.4	0.6	0.0
16	Wooded tundra	3.6	0.1	1.6
17	Mangrove	0.0	0.0	0.0
18	Ice and Polar desert	0.0	0.0	0.0
19	Water	12.6	3.4	12.9
99	**All**	**100.0**	**100.0**	**100.0**

Note: *the number of PTFs has been reduced to 19 from the original 29. This was done mainly by filling the unused categories 16, 17, and 18, and by grouping similar (from an atmospheric inversion perspective) categories 23-26+29*

Table D.2: Used TransCom regions and the number of parameters and the degrees of freedom added to $\boldsymbol{\lambda}$ per TransCom region

index	TransCom region	# parameters	Correlation length [km]	DoF
1	North American Boreal	1865	300	184
2	North American Temperate	1213	300	242
3	South American Tropical	19	-	3.2
4	South American Temperate	19	-	2.9
5	Northern Africa	19	-	3.2
6	Southern Africa	19	-	2.5
7	Eurasia Boreal	2396	1000	63
8	Eurasia Temperate	2631	1000	129
9	Tropical Asia	19	-	2.5
10	Australia	19	-	3.4
11	Europe	1585	200	435
12	Ocean	30	-	7
	All	**9835**	**-**	**1077**.7

Table D.3: Overview of the percentages of plausible scaling factors within each gridded TransCom region, both for the original scaling factors and for the effective scaling factors per ecoregion. For a description of the aggregation procedure, see Figure C.3

TransCom region	% plausible original scaling factors	% plausible effective scaling factors
North American Boreal	58.7	66.1
North American Temperate	51.0	54.5
Eurasia Boreal	55.3	61.6
Eurasia Temperate	59.8	56.5
Europe	52.4	74.3

Table D.4: Look-up table of the indices of the ecoregions included in Part II of this thesis

	TransCom region				
PFT (Olson V 1.3a)	North American Boreal	North American Temperate	Eurasia Boreal	Eurasia Temperate	Europe
Conifer Forest	1.0	20.0	115.0	134.0	191.0
Broadleaf Forest	-	21.0	116.0	135.0	192.0
Mixed Forest	3.0	22.0	117.0	136.0	193.0
Grass/Shrub	4.0	23.0	118.0	137.0	194.0
Tropical Forest	-	24.0	-	138.0	195.0
Scrub/Woods	-	25.0	-	139.0	196.0
Semitundra	7.0	26.0	121.0	140.0	197.0
Fields/Woods/Savanna	8.0	27.0	122.0	141.0	198.0
Northern Taiga	9.0	-	123.0	-	199.0
Forest/Field	10.0	29.0	124.0	143.0	200.0
Wetland	11.0	30.0	125.0	144.0	201.0
Deserts	-	31.0	-	145.0	202.0
Shrub/Tree/Suc	-	32.0	-	146.0	-
Crops	-	33.0	128.0	147.0	204.0
Conifer Snowy/Coastal	15.0	34.0	-	-	-
Wooded tundra	16.0	35.0	130.0	-	205.0
Mangrove	-	-	-	-	-
Ice and Polar desert	-	-	-	-	-
Water	19.0	38.0	-	152.0	209.0

Note *The PFTs which do not have an index within some TransCom regions are not the dominant PFT in any of the* 1×1 *degree grid cells of this TransCom region*

Table D.5: Overview of the selected environmental conditions which could potentially influence the state vector. This selection is based on the expert judgment of A. van der Woude. The ECMWF index can be used to find additional information on the variable. If reading a digital version, the index has a hyperlink to the relevant ECMWF webpage.

Abbreviation	Full name	ECMWF parameter ID	Used aggregation methods
`blh`	Boundary layer height	159	MAX
`cp`	Convective precipitation	143	MAX, SUM
`d2m`	2 metre dewpoint temperature	168	MIN, MAX, AVG
`g10m`	10 metre wind gust	49	MAX
`lsp`	Large-scale precipitation	142	SUM
`s10m`*	10 meter wind speed	N.A.	MIN, MAX, AVG
`sd`	Snow depth	141	MIN, MAX
`sf`	Snowfall	144	AVG, MAX
`skt`	Skin temperature	235	MIN, MAX, AVG
`slhf`	Surface latent heat flux	147	MIN, MAX, AVG
`src`	Skin reservoir content	198	MIN, AVG
`sshf`	Surface sensible heat flux	146	MIN, MAX, AVG
`ssr`	Surface net solar radiation	176	MAX, AVG
`ssrd`	Surface solar radiation downwards	169	MIN, MAX, AVG
`swvl1`	Volumetric soil water layer 1	39	MIN, MAX
`t2m`	2 metre temperature	167	MIN, MAX, AVG
`u10m`	10 metre U wind component	165	AVG
`v10m`	10 metre V wind component	166	AVG

***** - *The `s10m` has been calculated by* $\sqrt{u10m^2 \odot v10m^2}$, where $\odot$ is an element-wise multiplication.

Table D.6: Overview of how the monthly average model mitigates the encountered problems and whether the model is able to utilize potential sources of information.

Priority	Problem/ information source	Mitigated/ utilised	Explanation
1	Limited data availability	✓	The monthly mean can be determined using only a single state vector. Only a limited amount of state vector is expected to be needed
2	Integration within CT-DAS	✓	Determining the average is very cost efficient and could potentially be done every iteration of the EKF. The most efficient implementation would consist of determining the average once a year.
3	Noise within data	✓	As the average is based on multiple several weeks within each month, several data points are available for determining the mean. The odds of the same noise-induced anomaly occurring often enough to distort the mean value is as a result limited. Under such circumstances, a mean, or rolling-mean, model is a simple and efficient way of smoothing the signal and reducing noise, provided no extreme anomalies exist (Savitzky & Golay, 1964). If those extreme anomalies do exist, a (rolling-)median model could be considered instead.
4	Temporal dependencies	~	By design, the monthly mean model captures the monthly trends observed within Figure 1.3. However, the shorter temporal dependencies spanning only a few weeks (e.g. those induced by heatwaves) mentioned in the introduction of Part II are almost entirely disregarded. The model would also be sub-optimal if the mean scaling factor within eco-regions is non-stationary. De- and/or reforestation could for instance affect the biases within the biosphere model, affecting the mean scaling factor over time.
5	Exogenous variables	✗	No, the monthly mean model is incapable of utilizing exogenous variables. A potential variant of the monthly mean model could somehow add weight to each scaling factor based on a certainty value correlated to some environmental conditions, but the viability of this approach should be tested first.
6	Spatial dependencies	✗	No, no spatial dependencies are utilised. Also, no remotely viable method for including these dependencies within the monthly-mean model comes to mind.

Table D.7: Overview of how the SARIMA model mitigates the encountered problems and whether the model can utilize potential sources of information.

Priority	Problem/ information source	Mitigated/ utilised	Explanation
1	Limited data availability	✓	SARIMA is a version of the ARMA model, which is well equipped to train on a single and short time-series (?, ?), and is considered to be a classical method for stochastic process modeling (Güldal & Tongal, 2010).
2	Integration within CT-DAS	≈	The SARIMA algorithm is relatively simple and can easily be trained yearly, provided that the number of target variables is limited. The algorithm provides a forecast model for a single target variable. As such, a separate model is needed for every target variable. Making a model for every element within the state vector might therefore be infeasible. However, SARIMA can be used on an aggregated version of the state vector (i.e. one aggregated by eco-region).
3	Noise within data	≈	SARIMA can accurately capture wide-sense stationary stochastic processes (S. Wang, Li, & Lim, 2019), meaning that the mean and the correlation function of the process should be constant. As such, the SARIMA algorithm is robust to noise, if this noise is stationary. If the noise varies over time or is correlated to external variables, this robustness to noise could deteriorate.
4	Temporal dependencies	✓	The SARIMA model uses seasonal and shorter temporal dependencies by design.
5	Exogenous variables	✗	No, the bare SARIMA model is not able to utilize exogenous variables
6	Spatial dependencies	✗	No, the bare SARIMA model is not able to utilize spatial dependencies

Table D.8: Overview of how the SARIMAX model mitigates the encountered problems and whether the model can utilize potential sources of information.

Priority	Problem/ information source	Mitigated/ utilised	Explanation
1	Limited data availability	✓	same as SARIMA, see Table D.7
2	Integration within CT-DAS	≈	same as SARIMA, see Table D.7
3	Noise within data	≈	same as SARIMA, see Table D.7
4	Temporal dependencies	✓	same as SARIMA, see Table D.7
5	Exogenous variables	≈	The SARIMAX algorithm does allow for the utilization of exogenous variables. However, adding too many could impede the integration within CTDAS as it substantially increases the computational costs. The used variable should therefore be selected carefully.
6	Spatial dependencies	✗	No, the bare SARIMAX model is not able to utilize spatial dependencies