Reducing geographical bias in global CO_2 flux forecasts.

Jasper Smit - Author

Gerbrand Koren - Thesis Supervisor

Pablo Mosteiro - co-reader

Data science / Geoscience

Utrecht University

Netherlands

01-07-2022

Abstract

Carbon fluxes play an important role in our climate model on Earth. These fluxes have also been shown to be related to global warming. However, these fluxes are only measured at specific locations. Therefore, to obtain a global prediction, these local observations need to be upscaled to a global data product. This has previously been done by combining half-hourly flux measurements with globally available meteorological data and using machine learning to predict the flux based on the measurements alone. When doing so, an important distinction between tropical and extra-tropical regions was not considered. This is important as vegetation cycles are very different in the tropics which has a large impact on the carbon fluxes. These differences are hard to detect for a model because there is a very large imbalance in data availability for the tropics versus the extratropics. In this paper, this distinction is examined and multiple methods are proposed to make the model spatially aware. These methods include a tropic boolean variable, the latitude and longitude coordinates of the measurement, and a separate model for tropics and extratropics. The results showed that a non-spatially aware model does indeed struggle to predict correct diurnal cycles. The predictions improved by introducing spatial variables to the model with the best performing approach being the two separate models. But with more data, the latitude longitude model might perform the best as the model can figure out the tropic to extratropical transition itself. This showed that current approaches are indeed lacking due to spatial bias, and this paper addresses multiple possible solutions.

Contents

1 Introduction **4**

2 Data **5**

2.1 Site-level data 5

2.2 Global data product 7

2.3 Global vs local measurements 8

2.4 Overview 9

3 Methodology **13**

3.1 Pre-processing 13

3.2 Modelling choices 14

3.2.1 Random forest 14

3.2.2 Parameter settings 14

3.2.3 Approaches 15

3.3 Assessing performance 16

4 Results **17**

4.1 Last year prediction results 17

4.2 Cross validation results 22

5 Conclusion **23**

6 Discussion **24**

7 Appendices **26**

References **30**

1 Introduction

Improving our understanding of the carbon cycle and exchanges between our ecosystems and atmosphere is important to getting a better understanding of our climate model on earth (Tramontana et al., 2016). Research has shown positive feedback between terrestrial carbon cycles and global warming (Luo, 2007). These are some of the reasons why attention has been given to model the carbon exchange between the atmosphere and terrestrial biosphere. In order to model this complex relationship, we need a way to collect data on the actual amount of carbon that is exchanged with the atmosphere; from now on this will be referred to as CO_2 fluxes. These fluxes are measured by so called eddy-covariance towers. These towers measure a very local exchange between terrestrial ecosystems and the atmosphere (Billesbach, 2011). They provide half-hourly measurements of multiple variables and, most importantly, of the net ecological exchange.

The local nature of this measurement is however a problem when a global system or phenomenon, such as climate change, is the subject at hand. This is because we would need a very large number of towers spread equally across the world to get a global measurement. Because this is not feasible, Bodesheim et al. (2018) proposed an approach using machine learning to upscale these local flux measurements to a global product using globally available meteorological data. They trained a random forest model on a set of eddy covariance towers coupled with global meteorological data that yielded promising results.

Something of importance that was however overlooked in this research, is the major role that the tropics play in determining the global atmospheric concentration of CO_2. Research has shown that the tropics play an important role through both deforestation and photosynthesis, since tropical forests account for almost 60% of global terrestrial photosynthesis (Malhi & Grace, 2000). The reason why tropical forests should be examined in more detail in relation to eddy covariance towers is that the sites in tropical regions are vastly underrepresented compared to the sites in North America and Europe. This may

lead to problems when looking at it from a machine learning perspective as the model has more data to learn from the extratropical sites. This over-representation might lead the model to only adapt to the relations present in the extratropical regions and ignore the tropical relations. This might lead to poor performance if the relations are indeed different in a tropical environment.

In this paper, multiple approaches will be examined to account for this bias and the difference between tropical and extratropical relations in the carbon cycle, the ultimate goal being to reduce the geographical bias in the global CO_2 flux data product.

2 Data

In this section of the paper the used data sources for the analysis will be discussed. To learn the relations between the fluxes (the target variable) and the available global meteorological data, we need both of these data sets for the same location. Target fluxes can be gathered from site data from the FLUXNET network. This data is discussed in Section 2.1. This same network provides a data product where globally available meteorological variables are already matched with the previously mentioned fluxes. This will be discussed in 2.2. And lastly to check the validity of the use of this data product, it will be compared against the locally measured variables at the site level in section 2.3.

2.1 Site-level data

At the site level, there are "eddy covariance" towers that provide half-hourly CO_2 flux measurements. These measurements only apply to a very small spatial extent, practically only for the very specific location of the tower itself. This makes it difficult to use on a large scale as a very large number of towers would be needed to measure these fluxes for a large spatial extent. For this paper, data from 42 FLUXNET and Ameriflux eddy covari-

ance towers are used, see Figure 1 and Table 7, which together give 406 years of observed data. These sites were manually selected with a few criteria in mind. As mentioned above, the goal is to balance the tropic performance and the extratropic performance of the model with the tropics being defined as between -23.5 degrees latitude and 23.5 degrees latitude as this is a common standard. We know that there is more extratropic data available. So as a first attempt to limit the extratropical bias introduced by data availability, the extratropical sites were undersampled. Meaning that a lower percentage of the available sites were selected compared to tropical sites. An attempt was made to get a good world representation from the extratropical sites available, selecting ones that have a long-running history and are known to produce high-quality data.

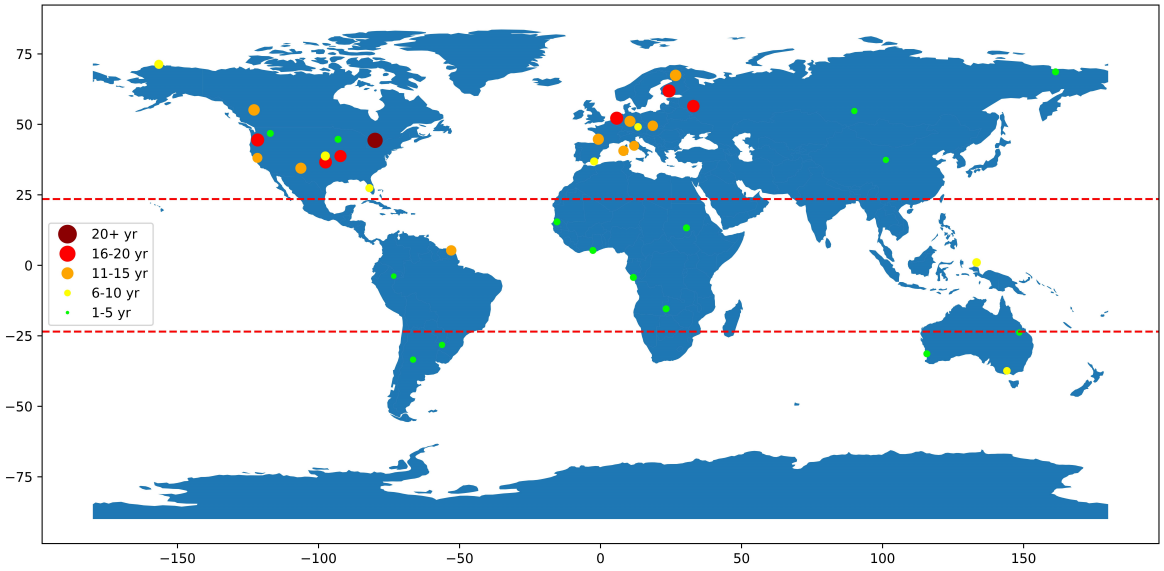

Figure 1: A map of the 42 sites used from FluxNet and Ameriflux across the world with 8 tropical sites and 34 extratropical sites. The color and size are indicative of how much valid years of data there are in the selected data set. The dashed lines indicate the selected tropical boundaries.

2.2 Global data product

For the global data product, the ERA-Interim data set (Dee, 2021) is used for two main reasons. The first one is that it is available in a global gridded format. This is crucial for this analysis as the end goal is to make global estimates for CO_2 fluxes, therefore it would not be logical to use data that is not available on a global scale. The second reason why this data set is chosen is a more practical reason. The FLUXNET data portal provides a so called "FULLSET Data Product" where the micrometeorological variables from ERA-Interim are already matched to the half hourly flux measurements as well as the locally measured meteorological variables. This allows for a more straightforward data-processing process, as we do not have to work with global gridded data products, as the matching was already done. The variables used are described in Table 1.

Table 1: ERA-Interim meteorological data used from the Fluxnet FULLSET data product.

Variable	**Unit**	**Description**
TA_F	deg C	Air temperature
SW_IN	W/m^2	Shortwave radiation incoming
LW_IN	W/m^2	Longwave radiation incoming
NETRAD	W/m^2	Net radiation
VPD	hPa	Vapor Pressure Deficit (High means dry)
PA	kPa	Atmospheric pressure
P	mm	Precipitation
WS	m/s^1	Wind speed
WD	Decimal degrees	Wind direction
RH	%	Relative humidity (High means wet)

2.3 Global vs local measurements

Before we can simply use this global data set, the validity of this global data set must be verified. We can do this by comparing the available locally measured variables to the matched globally available data. This was done using a correlation calculation, the results of which can be seen in Figure 2. This figure shows the locally measured variables with the suffix "_F" and the global variables with the suffix "_ERA". From the figure it can be seen that the local and global variables match very well with each other, indicated by the high correlation scores. The only real exception that we can see is the precipitation variable. This is most likely due to the fact that this is very local and is easily influenced by its direct environment. For example, the rain collector could be shielded from the rain or could be slightly tilted with the wind blowing the wrong direction all leading to incorrect measurement. Furthermore, rainfall can simply be a very local event. The global variable has an approximate resolution of 80 km (Dee, 2021), which is a large area to measure something as local as rainfall. For these reasons, the decision was made not to include the variable in the analysis.

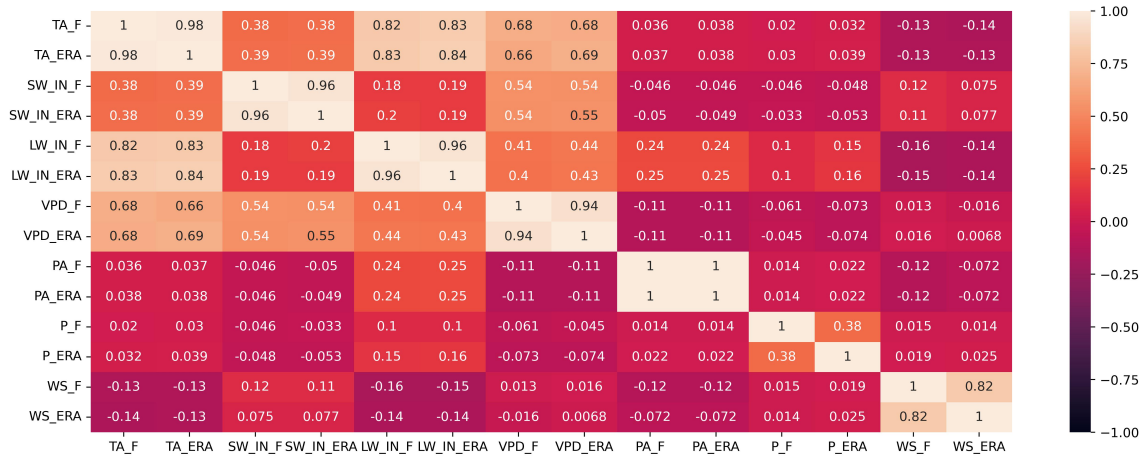

Figure 2: A correlation plot between the ERA-Interim variables ("_ERA" suffix) and the locally measured variables ("_F" suffix).

2.4 Overview

Now that it is known where the data comes from, some figures and statistics can be examined to get a better feel for the data. Table 2 shows the summary statistics for the extratropical regions and table 3 for the tropical regions. Some important points will be presented. The first one being the difference in number of records available. For the extratropics we have a total of 5.7 million measurements while for the tropics there are only 700 thousand measurements, and this is after we have undersampled the extratropics. The differences between various variable values will be made more clear with figures later. Something else to note are the -9999 values. These indicate incorrect measurements for that specific variable. The model that will be used in the analysis (Random Forest) can deal with these values as this model uses decision nodes and not a numerical formula to provide its estimations. Therefore it is able to essentially determine that these values are not correct and choose to ignore this variable or treat it differently. Simply removing these values would remove approximately 1.2 million rows of data, of which 0.2 million tropical rows (27%) and 1 million extratropical rows (11%). For completeness, the overview tables with these values removed can be seen in the Appendix in Table 8 and Table 9 to get a better sense of the distribution of the affected variables. Keeping these values in the data set is, of course, not the ideal scenario and will be discussed in Section 6, but it seemed the best option in this case.

	TA_ERA	SW_IN_ERA	LW_IN_ERA	VPD_ERA	PA_ERA	P_ERA	WS_ERA	WD	RH	NETRAD	NEE_VUT_REF
count	5698176.0	5698176.0	5698176.0	5698176.0	5698176.0	5698176.0	5698176.0	5698176.0	5698176.0	5698176.0	5698176.0
mean	9.3	155.1	310.0	5.1	96.3	0.0	2.6	-1534.0	-1426.1	-1774.8	-0.5
std	11.4	236.6	54.8	6.1	6.2	0.2	1.6	3820.8	3584.3	3914.5	6.0
min	-52.9	0.0	89.5	0.0	67.9	0.0	0.0	-9999.0	-9999.0	-9999.0	-79.3
25%	2.1	0.0	275.4	1.1	94.3	0.0	1.5	42.6	42.5	-76.9	-1.5
50%	10.2	0.9	312.6	2.8	98.4	0.0	2.3	171.0	71.0	-14.9	0.7
75%	17.3	254.4	346.9	6.6	100.4	0.0	3.4	251.6	89.0	81.8	2.2
max	45.4	1221.7	537.8	79.9	106.2	22.2	17.4	360.0	123.7	1015.9	50.0

Table 2: Summary statistics of selected variables from sites in extratropical regions.

	TA_ERA	SW_IN_ERA	LW_IN_ERA	VPD_ERA	PA_ERA	P_ERA	WS_ERA	WD	RH	NETRAD	NEE_VUT_REF
count	701328.0	701328.0	701328.0	701328.0	701328.0	701328.0	701328.0	701328.0	701328.0	701328.0	701328.0
mean	25.7	167.8	393.2	12.5	98.2	0.1	2.3	-2035.0	-1748.4	-1616.9	-0.6
std	3.9	260.4	36.5	11.7	3.4	0.5	1.1	4168.6	3871.1	3813.6	8.8
min	8.3	0.0	255.3	0.0	88.5	0.0	0.0	-9999.0	-9999.0	-9999.0	-75.7
25%	24.1	0.0	378.9	5.0	98.1	0.0	1.5	27.2	18.0	-62.7	-2.6
50%	25.6	0.0	408.3	6.5	99.4	0.0	2.2	98.7	69.9	-23.8	0.7
75%	27.2	279.6	414.6	21.7	100.7	0.0	3.0	182.4	88.1	162.6	4.6
max	43.3	1133.9	486.4	75.0	101.5	29.0	9.2	360.0	100.0	975.4	49.9

Table 3: Summary statistics of selected variables from sites in tropical regions.

After looking at the summary tables, it is important to understand the differences between tropical regions and extratropical regions. In Figure 3 the diurnal cycles of the NEE measurements are shown by month for the tropical and extratropical regions. The main thing that this figure demonstrates is the seasonal cycle that is evident in the extratropical regions but not in the tropical regions. This seasonal effect comes from vegetation growth that is experienced in most extratropical regions in the warmer months (spring and summer) and vegetation decay in the colder months (fall and winter). This confirms the idea that the tropics and extratropics should be looked at as different systems.

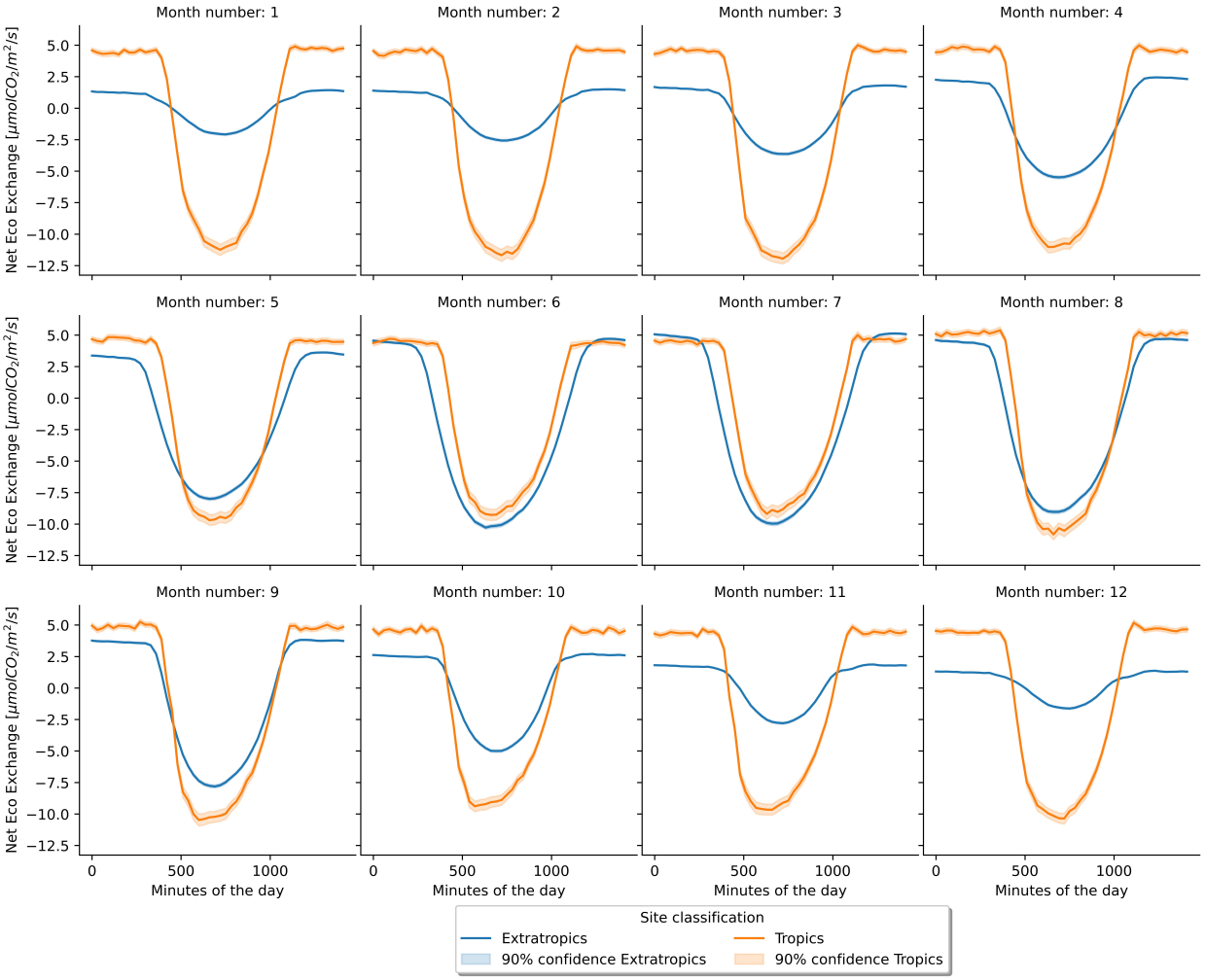

Figure 3: Diurnal cycles of Net Eco Exchange for tropic vs extratropical sites by month with a 90% confidence interval.

To further examine the differences, Figures 4 and 5 show the differences in air temperature and vapor pressure deficit, respectively, for a single diurnal cycle in different months. Looking at Figure 4, the seasonal cycles can also be identified in the extratropical regions through the temperature differences by month. This wide temperature range is not there in the tropical regions; the temperatures are higher and more consistent throughout the year. When we shift our attention to the comparison of the vapor pressure deficit in Figure 5 (remember that higher values indicate a dryer environment), we see something similar. Values tend to be higher in the tropics and are less spread throughout the year

compared to the extratropics. However, there is more of a spread than there was with temperature.

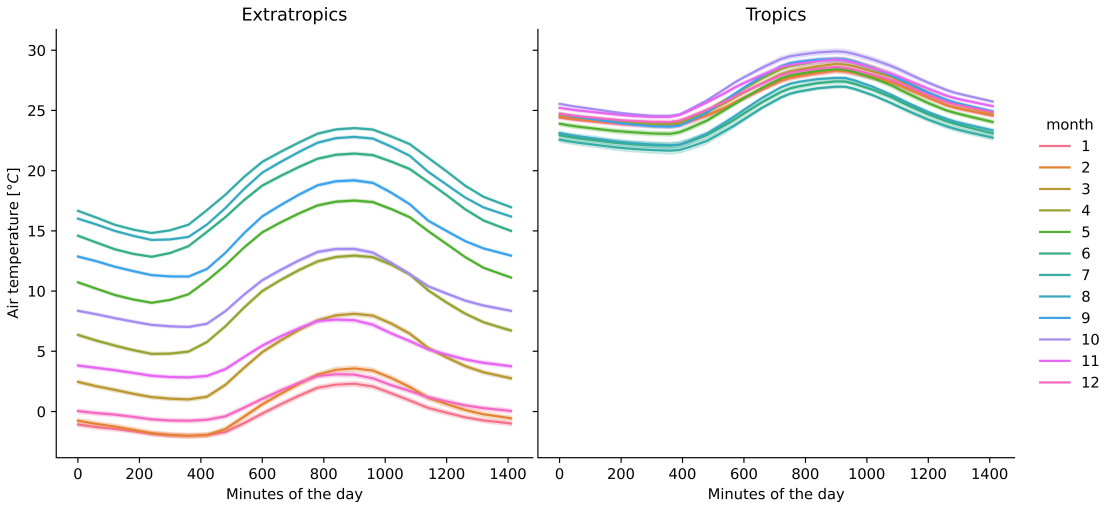

Figure 4: Air temperature per month for tropics and extratropics.

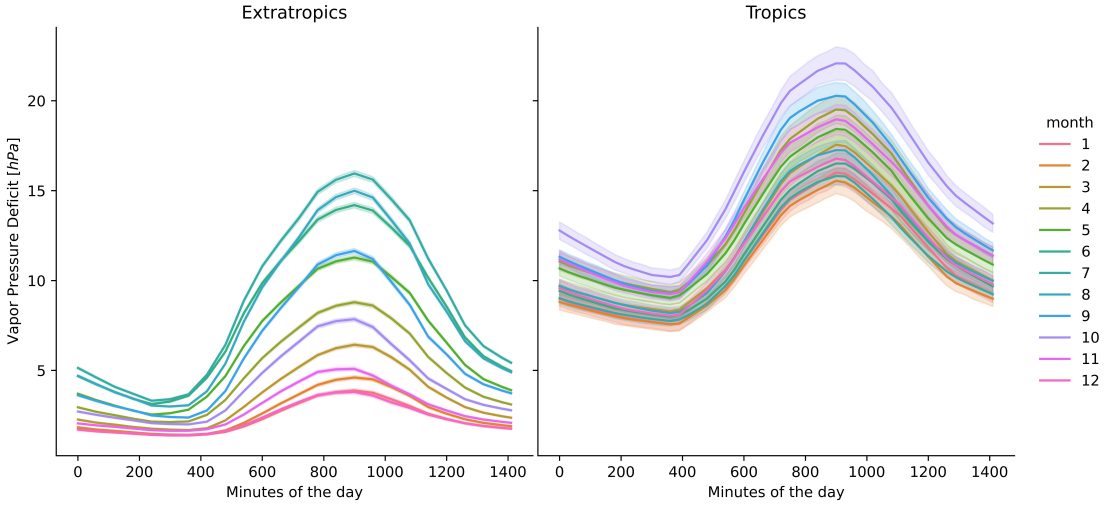

Figure 5: Vapor pressure deficit per month for tropics and extratropics.

3 Methodology

After discussing the data that will be used in the analysis in the previous section, the analysis itself will be discussed. This includes a few pre-processing steps, even though most of it has already been done through clever data selection and making use of the FULLSET data product. Then, the model used to predict the net ecological exchange will be discussed together with the different ways of assessing the performance of this model.

3.1 Pre-processing

During the data collection, a step was already taken in order to reduce the bias of the model towards extratropical regions by effectively undersampling these regions to limit the "class" imbalance. During the pre-processing phase, more measures were taken to possibly aid in reducing the bias during the modeling phase. The first was to retrieve the latitude and longitude coordinates of each site to include them in the data. This data can be used to make the model spatially aware, possibly aiding in determining the correct relations between the meteorological data and the net ecological exchange. Based on the coordinates, a new variable was created to indicate whether a site lies within the tropics or not. The tropics were defined as the band between the Tropic of Cancer and the Tropic of Capricorn between latitudes -23.5 and 23.5 as this is one of the most common definitions (Benbow & McIntosh, 2009). Furthermore, the timestamp provided was split into the separate features: "year", "month", "day", "minutes". Here, "minutes" simply refer to the amount of minutes that have passed in that specific day. Besides these simple additions, no other major modifications were done, other than removing all invalid measurements for the target variable (Net Ecological Exchange) and rows with missing values for any of the selected variables. This resulted in the removal of 718,896 (10.1%) measurement points from the data, of which 596,160 (9.4%) were from extratropical sites and 122,736 (14.9%) were from tropical sites.

3.2 Modelling choices

3.2.1 Random forest

In order to model this complex process, a model that allows for this complexity to be captured must be used. Therefore, a random forest model was used which consists of a set of randomized decision trees (Breiman, 2001). These groups of decision trees can be used both for classification and regression purposes. Previous work on which this thesis is largely based was also done with random forests (Bodesheim et al., 2018) and other work using flux data, such as the work of Tramontana et al. (2016) and Jung et al. (2020) demonstrating the viability of this model for cases like this. A major advantage of using a random forest is the use of bagging (Breiman, 2001). This technique refers to the process of aggregating predictions of multiple individual models (in this case the individual decision trees of which the forest is made) which have all been learned based on different subsets of data sampled out of the training set. Because each decision tree "grows" separately from the other trees, the predictions will be different. Then to get the final prediction, the individual trees are averaged in the case of regression.

3.2.2 Parameter settings

Random forest models allow multiple parameters to be adjusted for better performance or different desired behavior of the model. These parameters include, but are not limited to, the number of parallel trees that are used, the maximum depth of each individual tree, the minimum number of samples that are required at each decision point (split), and the number of features considered for each decision split. The most common way to determine the optimal values for the task at hand is to use cross validation (Refaeilzadeh, Tang, & Liu, 2016). With cross-validation, the available data is split into so-called folds (typically 5 or 10). The model is then trained on all folds except one, the fold kept out is used to evaluate the performance. This is done for all folds, and the performance metrics are

averaged across folds. This can give an indication of how well the model will perform on unseen data, without having to keep a separate test set to do this. This ties in to parameter tuning, as multiple combinations of parameters can be tested, and a cross-validation is done for each unique combination of parameters.

Through this process, it was found that 250 parallel trees was optimal when balancing computing power and results. Furthermore, a maximum tree depth was determined to be 10 to prevent overfitting to specific site characteristics. The model overfitting to site characteristics is hard to avoid but it is important as global predictions will largely be for places it has not seen before.

3.2.3 Approaches

In order to properly evaluate methods to reduce spatial bias in addition to the described data selection process, multiple different models were considered, trained, and compared. The selected models are the following:

- Model with no location data
- Model with the "tropic" variable
- Model with Latitude and Longitude as variable
- Two separate models, one for the tropics and one for extratropics

This list of models can essentially be read as the most "basic" model first and then increasing in complexity as the list goes with complexity referring to the information it is getting.

3.3 Assessing performance

To assess and compare the performance of these models, two approaches were taken. The first approach is to take the last year of available data for every selected site as a test set. This allows us to have a test set that is at least temporally new to the model, but it has seen data from this site before. It also should be kept in mind that this in relative terms reduces the tropical data set more than it does for the extratropical data set, as there is more data available for the extratropical sites. This difference is quite big as the relative test set size for the tropics is 25% of available data and 0.0003% for the extratropics which is 91 thousand times more.

The second approach is called "group k fold cross validation". This is a form of cross validation but the set is split up according to a "group" variable which in our case is the site name. It takes a value n and splits the data into approximate equal n parts that are non-overlapping. Therefore, the same group will not appear in two different folds. This deals with the problem of having seen a location before which is essential to consider with the use case (up-scaling to a global data product) where predictions will be made for unseen locations. Furthermore, this method also allows for more efficient computing as opposed to leaving out a single site per fold because the number of folds can be manually determined and for leave one site out not be biased, the model would have to effectively need to be run 42 times, one for each site and this is not feasible resource wise.

However, there are still issues with this method that must be considered when looking at the results in the following section. First, leaving out one tropical site again has a relative larger effect on the training set, as there are fewer tropical sites available. The second issue is due to the fact that some sites have much longer histories of data that the model relies on to learn relations from, for example, droughts that occurred in certain years. This leads to a very large spread in performance between the different folds, as the performance is largely dependent on what data it has seen and how closely it relates to the sites it has not seen before but is trying to predict. This is unfortunately something that

cannot be fixed with the limited availability of computational resources as you would need to get more sites with more data to become less dependent on specific sites being in or out of the training / test set. This will also be discussed in section 6.

4 Results

4.1 Last year prediction results

Measure [$\frac{\mu molCO_2}{m^2 s}$]	**No location**	**Tropical variable**	**Latitude / Longitude**	**2 Seperate models**
R^2 [-]	0.56	0.57	0.62	0.66
MAE	2.59	2.57	2.27	2.12
$RMSE$	4.17	4.14	3.89	3.67
Tropic $RMSE$	4.16	4.08	4.13	3.78
Extratropic $RMSE$	4.18	4.16	3.83	3.65

Table 4: Model performance (250 trees with maximum depth = 10) comparison for various metrics and split by tropic and extratropic for the last year of measurements test set.

Table 4 shows the R-squared and RMSE measure for both predictions overall and for the tropics and extratropics separately. Including a variable for the tropics seems to not really improve the model by a lot, but this seems different when looking at the actual values, which we will do later in Figure 6. The real improvements seem to come when including the latitude and longitude of the site into the model. A possible explanation for this is that there is a lot of variability in vegetation (main driver of carbon fluxes) in the extratropical regions, and now that the model is able to access the location of the data, it is able to distinguish between these different places with different vegetation levels. But this improvement might also be due to the fact that the model is now able to overfit to a

specific site which it has seen before which in turn might reduce the generalizability of the model. It is difficult to access the exact reason.

Measure [$\frac{\mu molCO_2}{m^2 s}$]	**No location**	**Tropical variable**	**Latitude / Longitude**	**2 Seperate models**
R^2 [-]	0.58	0.59	0.68	0.66
MAE	2.50	2.46	2.00	2.12
$RMSE$	4.10	4.06	3.55	3.67
Tropic $RMSE$	3.99	3.86	3.82	3.78
Extratropic $RMSE$	4.11	4.11	3.49	3.65

Table 5: Model performance (250 trees with maximum depth = 15 except for the two separate models) comparison for various metrics and split by tropic and extratropic for the last year of measurements test set.

For completeness, the same models were run with depth set at 15 to see how much this might help improve the predictions. The results can be seen in 5. This shows that the most difference can be found in the tropic RMSE for both the model without location data and the tropical variable. This could be due to the fact that tropical sites have a more complex relationship and need deeper trees to show this. We can better understand this by looking at the specific prediction values versus the observed values.

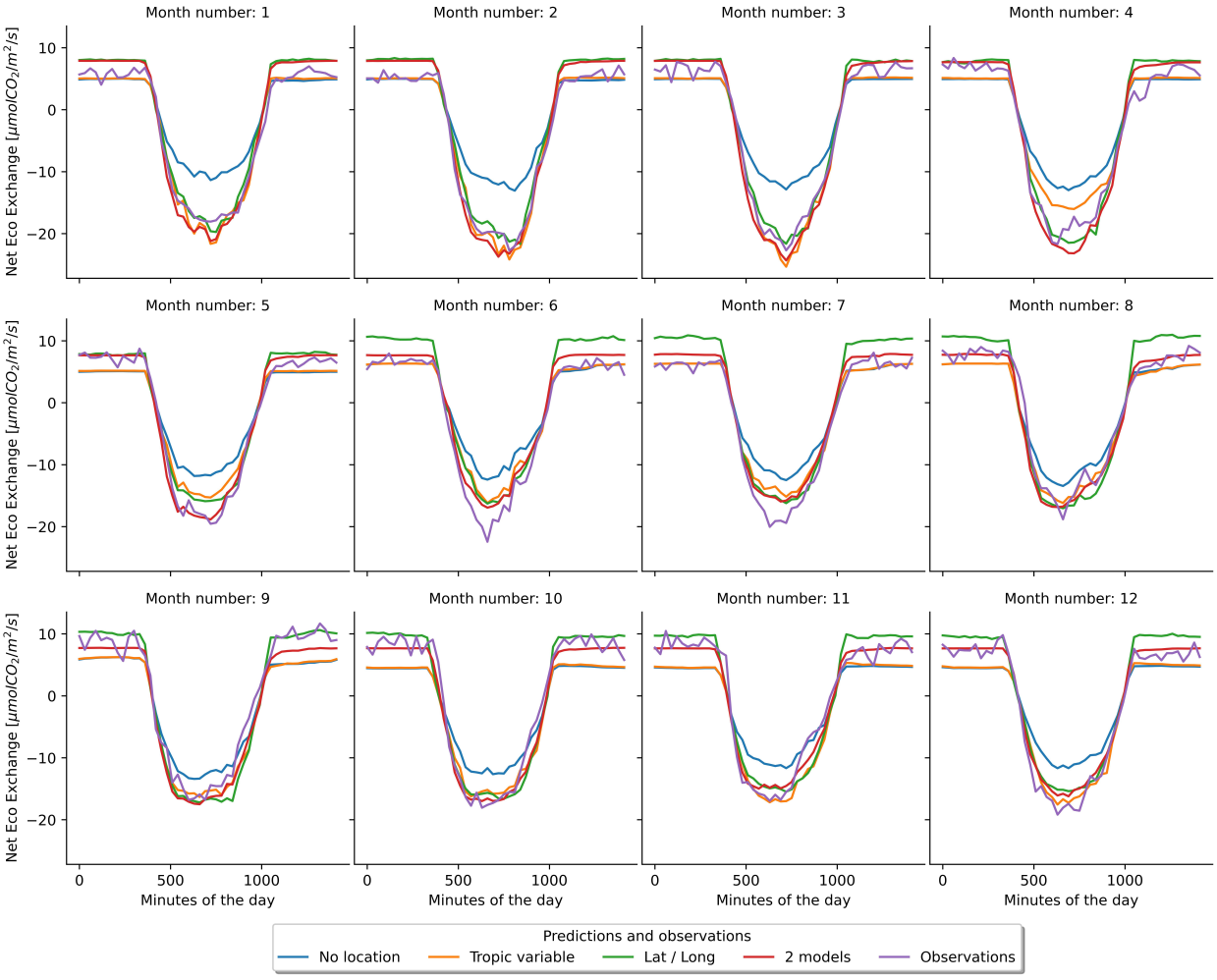

Figure 6: Observations and predictions of all 4 models (depth = 10) for the last year of data available for site PE-QFR (Lat = -3.8344, Long = -73.3190)

Looking at the prediction results shown in Figure 6, we can get a better understanding of how the prediction models differ from each other after having learned from the exact same measurements and predicting on the same test set. In this figure we see the predicted and observed Net Eco Exchange values (*NEE_VUT_REF*) for the last year of data available for the PE-QFR site (predictions for all sites can be seen in de Appendix). The plot is split up for each month of the year and each subplot represents a single diurnal (24 hour) cycle where values are averaged for each time step for that single month.

One of the first things to stand out is the blue line, representing the predictions of the model that received no location information as input. Comparing this line to the observations (purple line) we can see that it tends to predict too high for the Net Ecological Exchange values but it does get closer as we get later into the year. This is an indication that this model is attempting to apply a seasonal cycle to a tropical site (even though there practically is none) because it has seen this in most of the extratropical sites. The fact that this is still evident even when the extratropical sites are effectively undersampled shows that this is most likely an issue that occurred in the work of Bodesheim et al. (2018) as they did not account for the differences in seasonality between extratropical and tropical sites. The models that received some information regarding the location of the site do not show this seasonal cycle in their predictions and are all quite similar to each other.

However, during the night (the left and right portion of each individual plot) the models do show a difference. The model that has access to the latitude and longitude of the measurements (green) and the two separate models (red) systematically predict higher values during the night than the other two models. This seems to match better with the observed values during certain months and worse in others. But overall, the observed values do seem to be closer to the green and red lines (models that have access to location data).

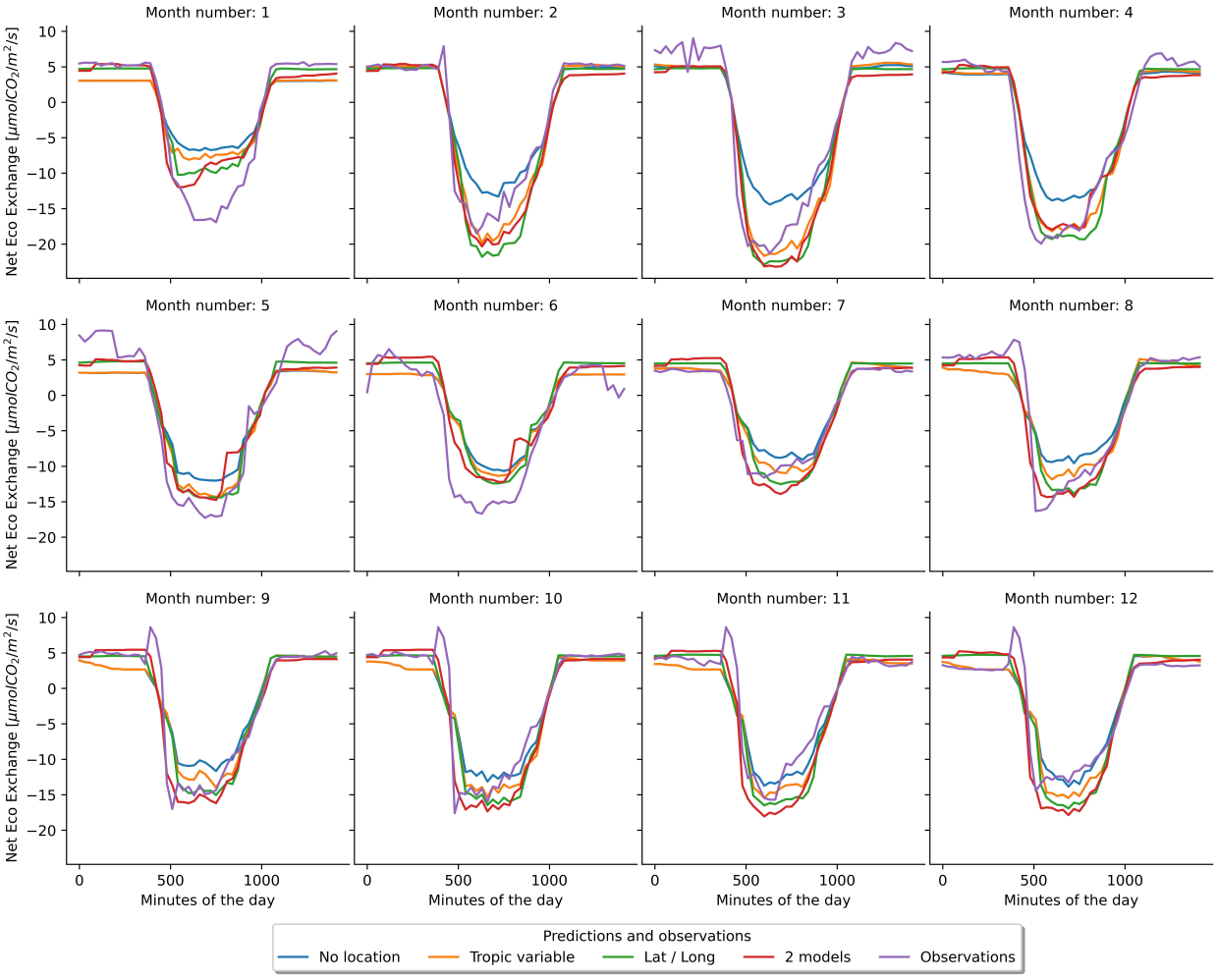

Figure 7: Observations and predictions of all 4 models (depth = 10) for the last year of data available for site GH-Ank (Lat = 5.26854, Long = -2.69421)

Figure 7 shows the same plot as in Figure 6, but now for the site: "GH-Ank". In this figure there are less clear systematic differences between the different models; however, we can still see predictions that are too low in the typical winter months for the northern hemisphere and that are not evident in tropical regions, as shown in Section 2.4. Furthermore, in January (month number 1) we can observe all models predicting too low values with the most simple model being the lowest and the predictions getting closer the more complex the model becomes.

4.2 Cross validation results

Measure [$\frac{\mu molCO_2}{m^2 s}$]	No location	Tropical variable	Latitude Longitude	Tropic model	Extratropic model
R^2 [-]	0.66	0.66	0.62	0.39	0.6
R^2 [-] - STD	0.05	0.05	0.11	0.58	0.15
$RMSE$	4.73	4.77	4.91	6.06	4.71
$RMSE$ - STD	0.71	0.7	0.57	1.96	0.51

Table 6: Model performance comparison for k-fold cross-validation (5 folds)

In Table 6 the results for the previously explained kfold cross validation with 5 folds are displayed. Due to the nature of kfold cross validation, looking at the individual predictions that are made within each fold are not representative for the overall generalisability of the model. This is because the predictions and the performance of a specific fold is very dependent on what data it received to train on and what data it has to predict. This variability from fold to fold is only amplified by the unequal distribution of tropic and extratropic sites available. So for example, a fold could see no tropical sites but receive all tropical sites to predict on. This would result in the tropical performance being terrible, while the extratropic results will be good. But if the tropic sites were spread across folds, one could observe much better results. These individual predictions would be needed to split the performance in extratropic performance and tropic performance, hence why this is not provided here. Providing those values might lead to improper conclusions. From the provided metrics, we can see that the standard deviations are very large relative to the mean performance values. Practically, the performance of all models fall within a single standard deviation of each other. This is most likely due to the simple fact that we do not have a lot of sites (only 42) and we only made 5 folds due to processing limitations.

This large fluctuation in performance does confirm the suspicion that it is difficult to predict the Net Eco Exchange for previously unseen locations. Most likely even

more so for locations that show different patterns than the majority of the data (extratropical).

5 Conclusion

This thesis had the objective of improving and elaborating on previous work done by Bodesheim et al. (2018) on the application of machine learning techniques to create a global gridded data product of CO_2 flux forecasts. The results shown and discussed in section 4 confirm the suspicion introduced in the introduction that it is problematic to assume the same relations between meteorological variables exist in tropical and extratropical regions, an example of this problem was shown in Figure 6. In this figure, the model that did not have access to location data, such as the model used by Bodesheim et al. (2018), is shown to try to fit a seasonal cycle to a tropical site, but these cycles are typically not present in tropical regions, as shown in Figure 3. Figure 6 showed that adding an indication of location to the model improved the ability to correctly predict the diurnal cycle. The overall performance metrics also improved when adding some indication of location, shown in Table 4. This table also showed that adding latitude and longitude coordinates as predictive variables, mostly improved extratropical performance which is likely due to the fact that the model can now discern between different extratropical locations that have a wide spread of vegetation and therewith a wide variety of diurnal cycles. The best performance can be seen when using two separate models, one for the tropics (between -23.5 and 23.5 degrees latitude in this thesis) and one for the extratropics. The question with two separate models like this then becomes where the exact cut-off should be to give the best solution. A solution to this question could be to use the latitude and longitude coordinates and give the model more data to learn from to then find the relation and the transition between tropics and extratropics itself, removing the need to make 2 separate models. Unfortunately not many conclusions can be drawn regarding the generalizability of the created models due to the fact that there is a lot of variability in performance

from site to site when they are not included in the training set, this is shown in Table 6. This variability makes it impossible to draw conclusions based on the current data set and methods regarding the generalizability of the model.

6 Discussion

Regarding the process of up-scaling flux forecasts to a global gridded data product, certain points are still left to be investigated which could not be addressed in this thesis. Ultimately, the findings in this thesis regarding tropic and extratropic performance should be validated using more sites and applying more computational techniques, such as leave-one-site-out cross-validation with all sites, to properly get an insight into the generalizability of the model. This is important because this is the main difficulty of the upscaling approach in general; there is no real way to verify how well this model predicts the fluxes for locations where there are no eddy covariance towers to measure the actual CO_2 fluxes.

A problem that is not easily addressed is the data imbalance between the tropics and the extratropics. Despite the fact that the available extratropical sites were heavily undersampled, there is still much more extratropical data. It might be possible that when the latitude and longitude coordinates are given to the model with access to all available sites, the model can find enough relations to differentiate between the different diurnal cycles (as suggested in the previous section). The lack of tropical sites also limited the analysis when using group k-fold cross-validation techniques, as this adversely affects the tropical data.

Furthermore, one has to be more conservative when removing dirty data that would otherwise simply be thrown out as discussed in section 2.4 where the choice was made to keep in -9999 values.

Another possible improvement for upscaling flux forecasts, is the inclusion of

more sophisticated climate information in the prediction model. An example of this can be the Köppen Climate Classification of a specific location. This could allow the model to learn more about underlying meteorological relations applicable to larger regions which should help with the generalisability of the model.

7 Appendices

Site name	From	Until
AR-SLu	2009-01-01	2011-12-31
AR-Vir	2010-01-01	2012-12-31
AU-Emr	2011-01-01	2013-12-31
AU-Gin	2011-01-01	2014-12-31
AU-Stp	2008-01-01	2014-12-31
AU-Wom	2010-01-01	2014-12-31
CA-Cbo	1994-01-01	2020-12-31
CA-LP1	2007-01-01	2020-12-31
CG-Tch	2006-01-01	2009-12-31
CN-HaM	2002-01-01	2004-12-31
CZ-BK1	2004-01-01	2014-12-31
CZ-BK2	2006-01-01	2012-12-31
DE-Hai	2000-01-01	2012-12-31
DE-Lkb	2009-01-01	2013-12-31
ES-Amo	2007-01-01	2012-12-31
FI-Hyy	1996-01-01	2014-12-31
FI-Sod	2001-01-01	2014-12-31
FR-LBr	1996-01-01	2008-12-31
GF-Guy	2004-01-01	2014-12-31
GH-Ank	2011-01-01	2014-12-31
IT-Noe	2004-01-01	2014-12-31
IT-Ro2	2002-01-01	2012-12-31
NL-Loo	1996-01-01	2014-12-31
PE-QFR	2018-01-01	2019-12-31
RU-Che	2002-01-01	2005-12-31
RU-Fyo	1998-01-01	2014-12-31

Table 7 continued from previous page

Site name	From	Until
RU-Ha1	2002-01-01	2004-12-31
SD-Dem	2005-01-01	2009-12-31
SN-Dhr	2010-01-01	2013-12-31
US-ARM	2003-01-01	2020-12-31
US-CF1	2017-01-01	2020-12-31
US-EDN	2018-01-01	2019-12-31
US-KLS	2012-01-01	2019-12-31
US-MOz	2004-01-01	2019-12-31
US-Me2	2002-01-01	2020-12-31
US-Mpj	2008-01-01	2020-12-31
US-NGB	2012-01-01	2019-12-31
US-ONA	2015-01-01	2020-12-31
US-Ro5	2017-01-01	2020-12-31
US-Snf	2018-01-01	2020-12-31
US-Tw1	2011-01-01	2020-12-31
ZM-Mon	2000-01-01	2009-12-31

Table 7: The used eddy-covariance sites in this thesis with starting and end dates of the used data.

	TA_ERA	SW_IN_ERA	LW_IN_ERA	VPD_ERA	PA_ERA	P_ERA	WS_ERA	WD	RH	NETRAD	NEE_VUT_REF
count	4039096.0	4039096.0	4039096.0	4039096.0	4039096.0	4039096.0	4039096.0	4039096.0	4039096.0	4039096.0	4039096.0
mean	10.5	163.8	313.7	5.4	96.1	0.0	2.6	190.6	71.7	87.8	-0.6
std	10.2	242.9	52.1	6.2	6.5	0.2	1.5	96.5	22.1	201.8	6.2
min	-41.0	0.0	103.0	0.0	67.9	0.0	0.0	0.0	0.0	-320.8	-79.3
25%	3.5	0.0	279.7	1.3	93.5	0.0	1.5	115.5	56.4	-45.3	-1.9
50%	11.1	4.9	315.3	3.1	98.4	0.0	2.3	202.6	75.9	-2.0	0.7
75%	17.8	276.9	348.8	7.1	100.4	0.0	3.4	263.2	90.6	154.2	2.4
max	43.7	1221.7	537.8	79.9	104.2	22.2	17.0	360.0	123.7	1015.9	50.0

Table 8: Summary statistics of selected variables from sites in extratropical regions excluding all rows with the value -9999.

	TA_ERA	SW_IN_ERA	LW_IN_ERA	VPD_ERA	PA_ERA	P_ERA	WS_ERA	WD	RH	NETRAD	NEE_VUT_REF
count	514878.0	514878.0	514878.0	514878.0	514878.0	514878.0	514878.0	514878.0	514878.0	514878.0	514878.0
mean	25.8	171.1	396.3	12.3	98.6	0.1	2.3	143.9	68.1	114.8	-0.6
std	3.8	263.2	34.7	11.7	2.9	0.6	1.1	85.5	27.6	225.4	9.2
min	8.3	0.0	261.8	0.0	88.5	0.0	0.0	0.0	3.2	-126.5	-75.7
25%	24.5	0.0	387.6	5.0	98.3	0.0	1.5	77.1	48.0	-43.0	-3.3
50%	25.6	0.2	409.4	6.2	99.4	0.0	2.2	125.9	78.2	-9.6	0.7
75%	26.8	285.2	414.9	19.3	100.7	0.0	3.0	201.0	90.1	245.9	5.1
max	43.3	1133.9	486.4	75.0	101.5	29.0	9.1	360.0	100.0	975.4	49.9

Table 9: Summary statistics of selected variables from sites in tropical regions excluding all rows with the value -9999.

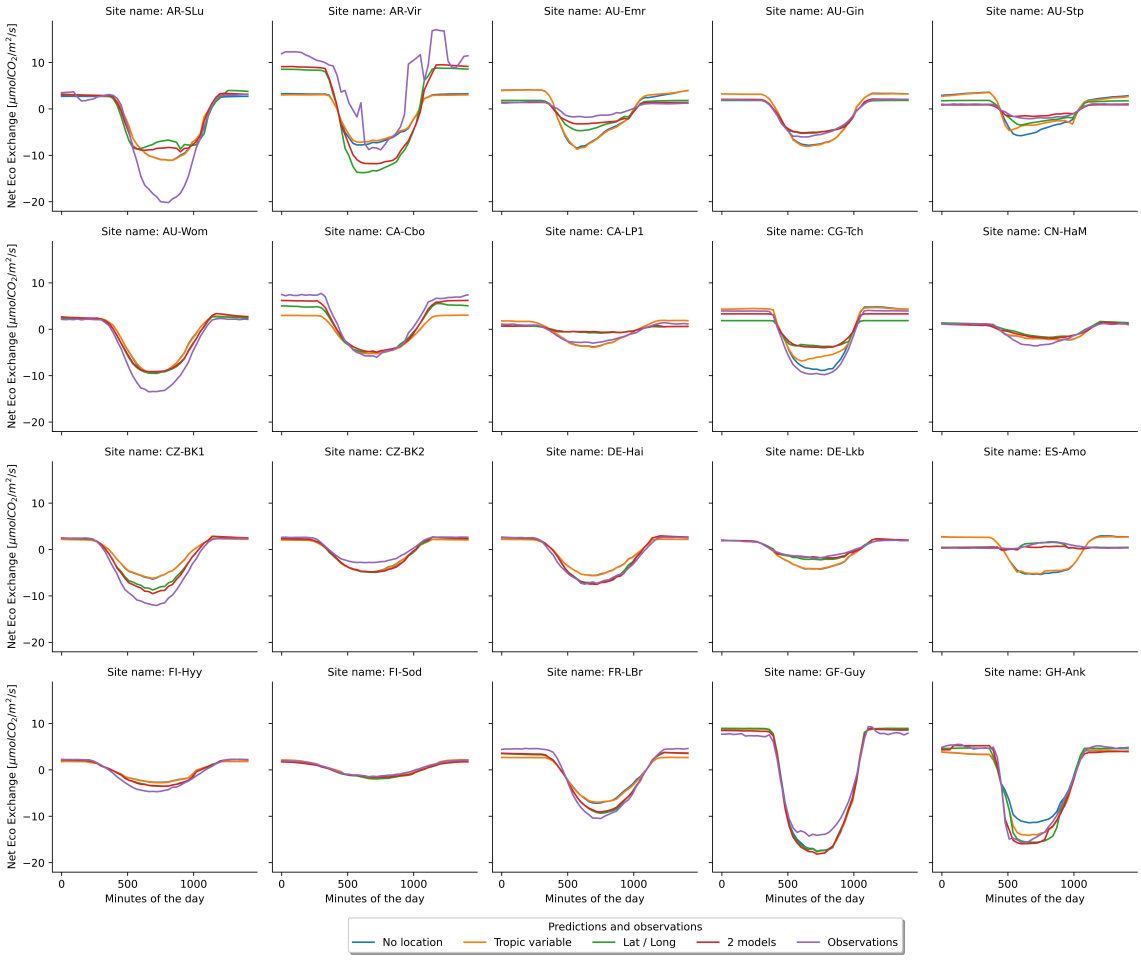

Figure 8: Part 1 of predictions for the last year of data available as test set for all models.

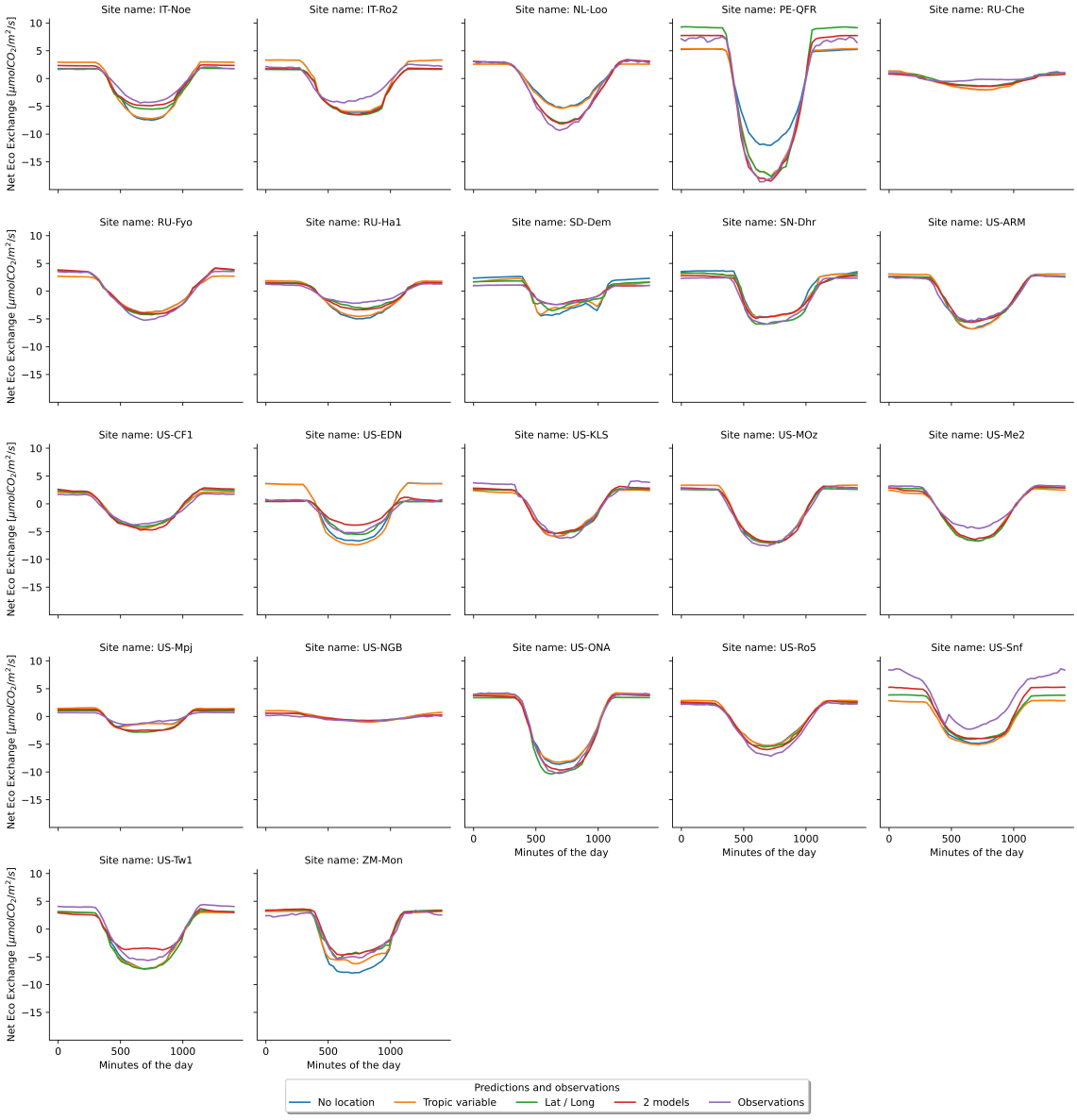

Figure 9: Part 2 of predictions for the last year of data available as test set for all models.

References

Benbow, M. E., & McIntosh, M. D. (2009, 1). Benthic invertebrate fauna, tropical stream ecosystems. *Encyclopedia of Inland Waters*, 216-231. doi: 10.1016/B978-012370626-3.00164-2

Billesbach, D. P. (2011). Estimating uncertainties in individual eddy covariance flux measurements: A comparison of methods and a proposed new method. *Agricultural and Forest Meteorology*, *151*, 394-405. doi: 10.1016/j.agrformet.2010.12.001

Bodesheim, P., Jung, M., Gans, F., Mahecha, M. D., & Reichstein, M. (2018, 7). Upscaled diurnal cycles of land-atmosphere fluxes: A new global half-hourly data product. *Earth System Science Data*, *10*, 1327-1365. doi: 10.5194/ESSD-10-1327-2018

Breiman, L. (2001). Random forests. , *45*, 5-32.

Dee, D. (2021, Sep). *Era-interim.* Retrieved from `https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era-interim`

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*, 1343-1365. Retrieved from `https://doi.org/10.5194/bg-17-1343-2020` doi: 10.5194/bg-17-1343-2020

Luo, Y. (2007). Terrestrial carbon-cycle feedback to climate warming. *Annual Review of Ecology, Evolution, and Systematics*, *38*, 683-712. Retrieved from `http://www.jstor.org/stable/30033876`

Malhi, Y., & Grace, J. (2000, 8). Tropical forests and atmospheric carbon dioxide. *Trends in Ecology Evolution*, *15*, 332-337. doi: 10.1016/S0169-5347(00)01906-6

Refaeilzadeh, P., Tang, L., & Liu, H. (2016). Cross-validation. *Encyclopedia of Database Systems*, 1-7. Retrieved from `https://link.springer.com/referenceworkentry/10.1007/978-1-4899-7993-3_565-2` doi: 10.1007/978-1-4899-7993-3_565-2

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*, 4291-4313. Retrieved from www.biogeosciences.net/13/4291/2016/ doi: 10.5194/bg-13-4291-2016