Feasibility of Machine learning approaches to upscale Net Ecosystem Exchange in Complex terrain of a tropical region using Remote Sensing and reanalysis data

Written by: Koo, Wai Hang (4039980) for thesis of
MSc. Environmental and Resource Management – Environmental Science track,
Brandenburg University of Technology Cottbus-Senftenberg

Supervised by: Prof. Dr. rer. nat. Katja Trachte, Chair of Atmospheric processes,
Brandenburg University of Technology Cottbus-Senftenberg

Abstract

Net Ecosystem Exchange (NEE) is crucial for identifying carbon sinks or sources. However, direct measurement by Eddy Covariance technique often lacks coverage in some remote areas with complex terrain, especially with high topography and therefore meteorological differences. This study has successfully established an algorithm to train and test the Random Forest (RF) models from remotely available data, such as global reanalysis and satellite remote sensing products, to upscale 2 NEE measurements in a highly variable terrain with explainable behavior. RF is also shown to be useful in identifying drivers through Interpretable Machine Learning. However, it also highlights the problem with having only 2 point-sourced measurements as training, such as the lack of representation spatially, and the forced linearity of temporally stationary variables such as Land Use Classes. Nonetheless, this study suggests that upscaling can be done as a reasonable and explainable way to identify carbon sinks and sources in complex terrain even with limited measurements.

Declaration of Independent Work

The author declares that he / she has written the thesis at hand independently, without outside help and without the use of any other but the listed sources. Thoughts taken directly or indirectly from external sources (including electronic sources) are marked accordingly without exception. Sources used verbatim and contentual were quoted according to the recognized rules for scientific working. This thesis has not been submitted in the same or similar form, not even partially, within the scope of a different examination.

Thus far it also has not been publicized yet.

I herewith agree that the thesis will be examined for plagiarism with the help of a plagiarism-detection service.

_____   _____

Place, date   Author's signature

Table of Contents

Table of Figures

List of tables

Introduction

Net Ecosystem Exchange (NEE), also referred to as Ecosystem Carbon Flux, represents the net carbon flux within an ecosystem, primarily influenced by the uptake and release of Carbon Dioxide (CO_2) by plants through photosynthesis and respiration (Abdalla et al., 2013). The measurement of NEE could provide an assessment tool for identifying carbon sinks and sources. However, the direct measurement method using Eddy Covariance (EC) is limited to point measurements. Such measurement and the ensemble of stations do not provide sufficient spatial resolutions to represent areas with complex topography and land use classes. Therefore, a top-down model approach using satellite data and a global reanalysis model could be especially helpful for remote areas with complicated variability and therefore with fewer available measurements. However, reanalysis has limited spatial resolution and satellite imagery can sometimes come with questionable data quality. This could lead to an underrepresentation of remote regions with complex sub-grid terrains, which often have a higher capability of being Carbon sinks due to less anthropogenic influences and higher density of trees and leaves.

In response to these limitations, there is a growing interest in data-driven statistical models that do not require process-based understanding to model ecosystem dynamics, thus avoiding the complexity of parameterization and discretization. Therefore, Machine Learning (ML), which is often an extension of regular statistical algorithms, is gaining traction in ecosystem dynamics models due to its ability to harness the power of high-performance computing.

Random Forest is an algorithm within the subcategory of Supervised learning in ML, which means the model is trained by labelled input and tuned to predict the input as precisely as possible. The Random Forest algorithm is based on the classical decision tree algorithm, in which the prediction is done by splitting every leaf with certain criteria. However, with only one tree, overfitting dominates the prediction and thus is known to hinder accuracy, especially with a large set of dependent variables. Random Forest expands the classical algorithm by employing multiple trees and forms an ensemble of trees with different pathways of decision, thus named 'Forest' (Breiman, 2001). It aims to reduce overfitting problems caused by the intrinsic bias of a single tree. However, as with any Machine Learning algorithm, it is a black box model, in which the prediction process cannot be directly examined as in the parameters in traditional physical models. Yet, Random Forest has its advantages

against the traditional regression and physical-based models. Firstly, due to the input-mimicking nature of Random Forest, it is almost guaranteed that the output prediction will be at least a close replica of the original data. Secondly, while most of the Machine Learning algorithms depend highly on the hyperparameters, which are the parameters defining the regressor, Random Forest does not have a high dependency on these parameters. The only important one is the number of trees that formulate the ensemble (Probst & Boulesteix, 2017). Therefore, Hyperparameter tuning can be done at a minimum and will not have a huge impact on the performance. Thirdly, Random forests are known to be able to handle thousands of variables accurately (Leo Breiman, 2004), and therefore suitable for predicting complicated dynamics.

This study is developed based on a previous project (Koo, 2023), which aimed and successfully establish Linear regression models and Machine Learning (RF) model of temporal NEE in Southern Ecuador, thus investigating the drivers and the potential carbon sink or source. It was concluded that the RF model is more capable of predicting NEE with only remote sensing and global reanalysis model outputs. Therefore, it sheds light on the possibility of building a full 3-dimensions RF model of NEE in the area with mainly the remotely available data, such as satellite imageries and global reanalysis model. This study aims to expand the RF model and the data preparation algorithms into a spatial-temporal model from point-sourced measurements, otherwise known as upscaling, with the Random Forest algorithm that could be used for the prediction of NEE in remote areas with high topographic variability. The models will be examined to determine the feasibility of the algorithm in terms of NEE upscaling from minimal availability of data and high topographic variability.

Methodology

Study Site

The Andes Mountains lie on the western edge of South America, with Aconcagua (6961m) as the highest point and stretches from 10N to 53S (Garreaud, 2009) of South America's west Coast. The complex variability of topography, which causes the heterogeneous landscapes with high variability in meteorological conditions associated with diverse vegetation types, led to complications in NEE prediction due to the high topographic variability in the often sub-global-model-grid scale peaks and

valleys.

Two EC stations were installed in the San Francisco Valley, between the provincial capital of Loja and Zamora in Southern Ecuador. The valley cuts through the center of the Andes Mountain Ridge, with Rio Zamora passing through in the middle, creating a zonal direction valley. The two EC stations are set up along the land use gradient. The pasture station is situated on the north of the valley, while the Forest station is set up in a tropical rainforest area on the southern slope of the valley.

The targeted area is selected as a 0.1 x 0.1-degree grid, corresponding to roughly 10 x 10 km of the San Francisco Valley area. The meridional and Zonal bounds are at 79.12E and 79.02E, and 3.9S and 4.0S respectively. It covers River Zamora running from the southwest and a branch from the northwest, converging in the middle before running first east then south. The rivers carved the valleys into ~1700m a.s.l. between mountains peaks at ~2600m a.s.l. The valleys form with a Y-shape turned 90 degrees anticlockwise, before turning southward at the eastern edge of the targeted area. The whole area is a rather homogeneous tropical rainforest, with occasional deforested pastures and grasslands.

The temporal range of this study covers all months in the years 2019 and 2020, with ERA-sourced variables having a time step of 1 hour. FC as a part of EC variables has a time step of 30 minutes. The temporal resolution of different MODIS variables is listed in Table 1.
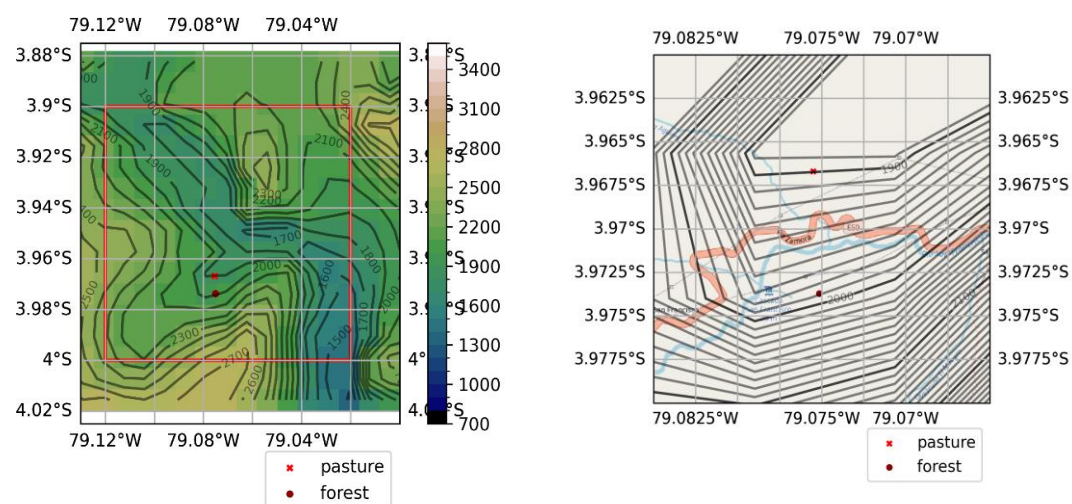
Figure 1. a.) Topography map of the Target Area (in red square) (left) and b.) detailed elevation map (right). Contours are the topographic elevation from GTOPO30, with 100m (a) and 10m (b) interval from GTOPO30 (Earth Resources Observation And Science (EROS) Center, 2017). Background of (b) are from Open Street Map licensed under CC BY-SA 2.0.
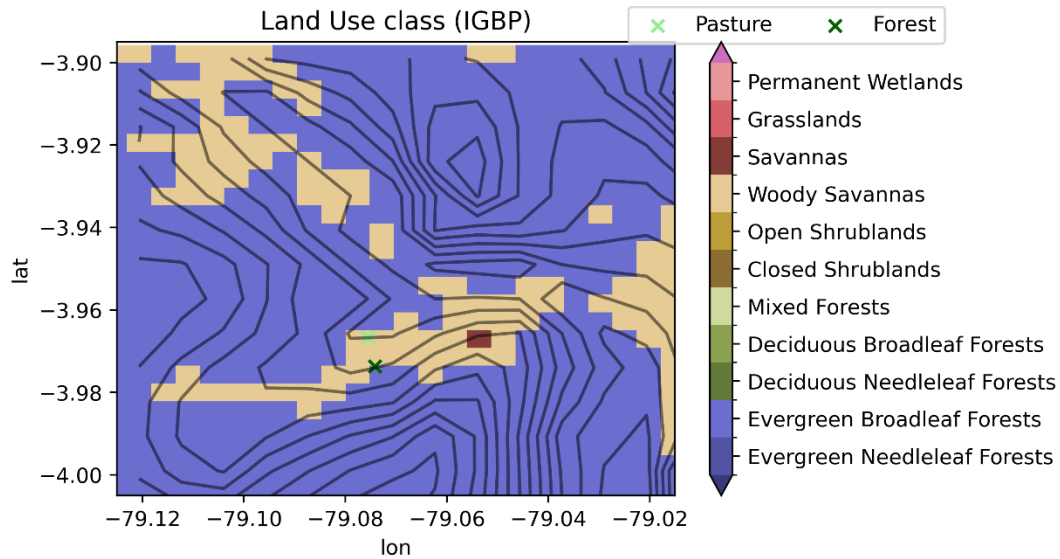
Figure 2. Land Use Class in the target area based on the definition of IGBP (Friedl & Sulla-Menashe, 2022). Contour represents the topographic height in 100m intervals from GTOPO30 (Earth Resources Observation And Science (EROS) Center, 2017)

Most of the target area are covered by Evergreen Broadleaf forests as defined by IGBP, which correspond to >60% tree cover by broadleaf trees with a canopy >2m. A smaller fraction of the area is woody savannas and savannas, which correspond to tree cover between 30% to 60 %, and 10% to 30% respectively (Friedl & Sulla-Menashe, 2022).

Climate of the region

The region exhibits significant spatial variability in climate due to variations in altitude and topography. Lower foothills and valley bottoms are often characterized by tropical rainforests, classified as Af in the Koppen-Geiger Climate Classification system. However, as elevation increases towards the central ridge and the target area, the climate transitions to a temperate classification (Cfb), featuring warm summers and no distinct dry season (Beck et al., 2023).
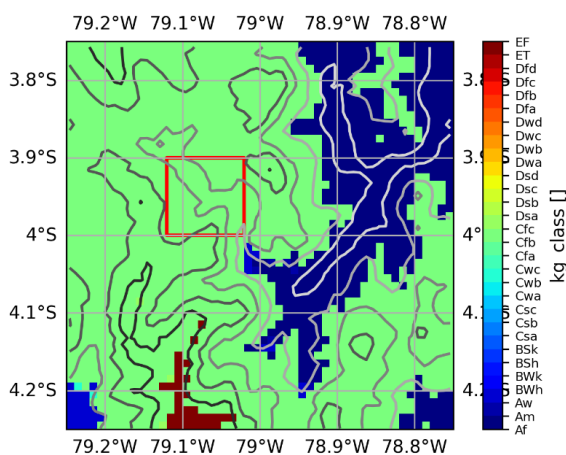
Figure 3. 0.1-degree Koppen-Geiger Climate Class map produced using data from (Beck et al., 2023). Contour represents the topographic height in 250m intervals from GTOPO30 (Earth Resources Observation And Science (EROS) Center, 2017). The red square shows the target area.
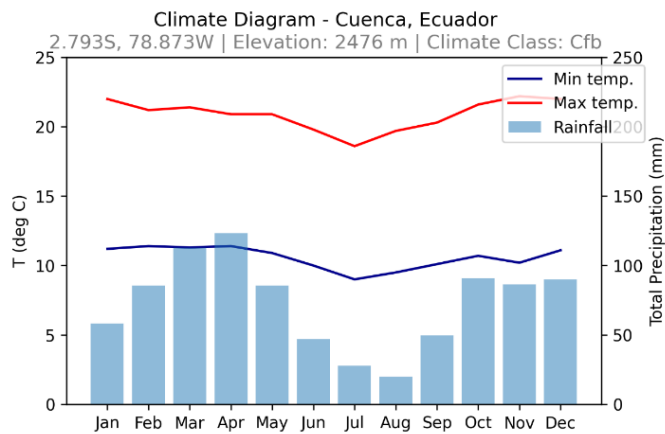
Figure 4. Climate Diagram of Cuenca, Ecuador (~120km North of the target region). Data are obtained from the World Weather Information Service of the World Meteorological Organization.

Figure 4 presents a climate diagram of Cuenca, Ecuador, situated approximately 120 kilometres north of the target region, providing additional insights into regional climate patterns.

Despite the region's tropical setting, temperature variation exhibits minimal seasonality due to its proximity to the equator. However, significant differences in rainfall are observed between dry and wet seasons. In the southern Andes, winter (JJA) experiences notably lower precipitation compared to summer (DJF). Local topography plays a crucial role in precipitation distribution, with higher elevations leading to air parcel lifting and condensation, contributing to enhanced rainfall.

The prevailing easterly trade winds, a result of the Hadley Cell, supply the region with moisture, particularly during the wet season. The convergence of trade winds in the Inter-tropical Convergence Zone (ITCZ) amplifies rainfall patterns, with meridional shifts in the ITCZ leading to seasonal variations in precipitation.

In summary, while the region exhibits a tropical climate, strong seasonality in precipitation is observed due to the meridional movement of the ITCZ. Winter months coincide with a northward shift of the westerly jet, resulting in reduced precipitation around the region (Garreaud, 2009).

Raw Data

To obtain a set of meaningful variables as independent variables, variables from multi-dimensional datasets are employed by several sources from either satellite sensing or global reanalysis, i.e., ERA5-land, MODIS v6.1, and GTOPO30. The dependent variable, i.e. NEE, is alternatively obtained from observation of the two EC stations. The independent variables are selected based on the theoretical relevance to NEE, which can be classified as either basic meteorological and vegetation parameters, radiation components, topographic altitude, or seasonality.

ERA data is provided by the ERA5-land Reanalysis model of ECWMF and is available in NetCDF4 format (Copernicus Climate Change Service, 2019) with a raw resolution of 0.1 x 0.1 degrees (~27 x 27 km at the equator). The variables applied include zonal and meridional wind field, air and dew point temperature, precipitation, and radiation budget components.

MODIS (Moderate Resolution Imaging Spectroradiometer)-derived parameters, on the other hand, are obtained from NASA Terra EOS AM-1 product on the NASA Earthdata database (Didan, 2021; Myneni et al., 2015). The satellite sensing data of NDVI, EVI, FaPAR, and LAI was obtained in HDF4-EOS format. The feature list is largely similar to the one for the previous study, which includes products of MOD15A2H and MOD13Q1, both are version v6.1. However, due to the expansion of the temporal and spatial dimensions of the targeted area, several additional variables are introduced in the dataset, which includes elevation from GTOPO30 (Earth Resources Observation And Science (EROS) Center, 2017), Land Use Classes (MODIS Product MCD12Q1) and vegetation cover percentages (MODIS Product MOD44B, including % Tree cover, % non-Tree vegetation, and % non-vegetation cover), and the day of years (from 1 to 365 (year 2019) or 366 (leap year such as 2020)) as an integer array.

Table 1 displays all used variables and their resolutions, abbreviations, and units.

Table 1. List of variables, the sources, resolution, and definitions. **Bolded** variables are altitude-correlated and suitable for DEM interpolation (see data preprocessing).

Type	Spatial resolution	Temporal resolution	Variables [unit in files]	Full Name (As in metadata)	Symbol	Unit (as presented)	Source(s)
ERA5 Land Reanalysis	0.1 x 0.1 degree (~9km x 9km) Single level →*1 x 1km by DEM interpolation (see data preprocessing)*	Hourly	**u10 [ms⁻¹]**	**10-meter U wind component**	**u10**	**m s⁻¹**	(Copernicus Climate Change Service, 2019)
			v10 [ms⁻¹]	**10-meter V wind component**	**u10**	**m s⁻¹**	
			d2m [K]	**2-meter dewpoint temperature**	**d2m**	**K**	
			t2m [K]	**2-meter temperature**	**t2m**	**K**	
			slhf [Jm⁻²]	Surface latent heat flux	slhf	J m⁻²	
			ssr [Jm⁻²]	Surface net solar radiation	ssr	J m⁻²	
			str [Jm⁻²]	Surface net thermal radiation	str	J m⁻²	
			sp [Pa]	**Surface pressure**	**sp**	**Pa**	
			sshf [Jm⁻²]	Surface sensible heat flux	sshf	J m⁻²	
			e [m h⁻¹]	Total Evaporation	tp	m	
			Swvl1 [m³ m⁻³]	**Volumetric soil water layer 1 (0 - 7cm)**	**SWC**	**m³ m⁻³**	
			fal	Forecast albedo	fal	/	
DEM (GTOPO30)	30 arc sec (~1km)	NA	Elevation	Elevation above sea level (a.s.l.)	height	m	(Earth Resources Observation And Science (EROS) Center, 2017)
EC	NA (Point Data)	Half-hourly (30 min)	FC [µmol CO2 m⁻²s⁻¹]	Carbon Flux / Net Ecosystem Exchange	NEE	gC m⁻²d⁻¹	/
MODIS	500m x 500m	8-days averaged	Fpar_500m	Fraction of Absorbed Photosynthetically Active Radiation 8-Day	faPAR	/	MODIS MOD15A2H v61 (Myneni et al., 2015)
			LAI_500m	Terra Leaf Area Index	LAI	/	
	250m x 250m	16 days	NDVI	250 m Normalized difference vegetation index	NDVI	/	MOD13A1 v61 (Didan, 2021)
			EVI	250 m monthly Enhanced Vegetation Index	EVI	/	
		Yearly	Percent_Tree_Cover		TC	%	MOD44B (DiMiceli et al., 2022)
			Percent_NonTree_Vegetation		nTC	%	
			Percent_NonVegetated		Nonveg	%	
	1km x 1km	Yearly	LC_Type1	Land_Cover_Type_1	LC	/	MCD12Q1 v61 (Friedl & Sulla-Menashe, 2022)
/	NA (Point Data)	Daily	Day_of_Year	Day of Year	DayofYear	/	/

Data Pre-treatment

Within the raw variables, there are several temporal and spatial resolutions. Temporal-wise, the variables can be grouped as half-hourly (NEE), Hourly (all ERA5-land), 8 days (LAI, faPAR), 16 days (NDVI, EVI), and Yearly (% TC, % nTC, % Nonveg, LC). Spatial-wise, it ranges from 250m (NDVI, EVI, % TC, % nTC, % Nonveg,), 500m (LAI, faPAR), ~1km (height, LC), to ~9km (ERA5-land variables). To establish a functional model input, the data must first be homogenized into a uniform grid system. Ultimately, to reach a meaningful homogenization, process the final (spatial) resolution has to be the coarsest of all raw variables. This brings up the problem with the coarse grid of ERA5-land. Even though we switched from the 0.25-degree (~27km) grid resolution from ERA5, which was used in the previous study, to 0.1 degree (~9km) from ERA5-land, the resolution is still an order of magnitude of 1 higher than the others. Therefore, it could be critical to perform some sorts of interpolations to refine the spatial resolution of the ERA variables, and thus the final model output. It is thus decided to have an extra interpolation procedure to help predictions in the often sub-grid features.

ERA variables Interpolation
ERA variables include several alleged elevations-correlated variables (t2m, d2m, sp, u10, v10, and SWC). These could be correlated to elevation with a simple bilinear regression, and therefore a finer Digital Elevation Model (DEM) can help bring down the resolution gap. Temperatures and surface pressure (t2m, d2m, and sp) are known to be linearly dependent on altitude by physical processes. For example, t2m is dependent on the altitude via the Lapse rate near the surface. Surface Pressure (sp) is bounded by the air column mass above a certain vertical position, and therefore altitude dependent. The zonal and meridional wind velocities (u10 and v10) are somewhat correlated to height due to laminar flow. Consequently, an interpolation can be done through linear regressions of such variables against elevation. However, since the base model of ERA5-land is IFS model documentation CY45R1, which does not use geometric height as the vertical axis, an approximation of the model height is derived from the barometric equation (eq. 1). With that, a relationship between the corresponding variables and height can be formed.

ERA5-land model surface heights are calculated from an extended set ranging between 79.2°E and -78.8°E longitude and 3.8°S and 4.2°S latitude using a derived version of the Barometric equation (eq. 1), which is deducted from the hydrostatic equation $\frac{dP}{dz} = -\rho g$ and the equation of states $P = \rho R T$. The calculations are done assuming a constant vertical temperature profile (lapse rate L= -0.0065Km[-1]), and the reference pressure (P$_0$ =

101325 Pa) and temperature (T_0=288.15K) are set according to the International Standard Atmosphere (ISA) (Johnson et al., 2002) with every timestep and the mean value of all the timesteps are the calculated model altitude.

$$H = \frac{T_0}{L}\left(\left(\frac{P}{P_0}\right)^{\frac{-LR}{g}} - 1\right)$$

(1)

The ERA5-land dataset and the calculated model height are then put against the respective variables in linear regression for each time step in ERA5-land, in which the y-intercept and the slope can be obtained. The y-intercepts and slopes can be paired into a regression equation with a general form as eq. 2. The interpolated variables are then calculated using the elevation from GTOPO30. The y-intercepts and slopes, as well as the correlation coefficient between the fitted linear relationship and the reanalysis, are shown in Table 2.

$$Var|_{height} = slope \times height + y\ intercept$$

(2)

, where Var is the variable in question.

Table 2. Linear Regression details (mean Y-intercepts and mean Slope in eq. 2), Correlation coefficients (R^2), and Pearson's coefficients (p-value) of the variables to the said Linear Relationships. Bald p-values indicate a p>0.05, which failed to be proven significant.

Variables	Mean Y-intercept	Mean slope	$\overline{R^2}$	p-value (95%CI)
t2m	297.289	-0.004245	0.9214	6.2247e-07
d2m	296.128	-0.004836	0.9570	3.2748e-07
sp	98891.385	-9.2125	0.9990	1.9665e-35
u10	1.0840	-0.0008888	0.6117	0.03407
v10	-0.1218	8.0392e-05	0.1861	**0.2357**
SWC	0.4033	-3.2309e-06	0.0421	**0.5038**

Table 2 shows that the assumption of linearity does not fit v10 and SWC due to the low R^2 value and a rejection of the null hypothesis of the significant linearity, resulting in an alternative interpolation method being introduced. For the sake of uniformity, u10 will also be interpolated by the same means as v10.

Alternatively, another regression approach can be used for u10, v10, and SWC, which would be using RF as an interpolation approach based on (Wang et al., 2021). Because these variables are assumed to be correlated with height, and it is not linear as shown in table 2, coordinates and model heights can be used to better correlate to wind fields and Soil Water Contents. Due to the high seasonality of the variables, additional timestamps (Day-Of-Year, DOY; and hour-of-day, hr) are added to help increase the temporal accuracy of the regression.

ERA-sourced variables without a clear logical pathway to establish correlations with topographic height, including the radiation budget components and evaporation ('fal', 'slhf', 'ssr', 'str', 'sshf', 'e') are resampled to the same resolution of the other interpolated variable, i.e., 14 x 14 grids of ~1 x 1 km resolution. Radiation budget components are then converted from the original accumulated form into the daily max value before being divided by a factor of 3600 x 24 = 86400, owing to them being reported as daily accumulations by the hour (Copernicus Climate Change Service, 2019). All of the processed variables are then merged and renamed ERAint.

EC Quality Control and pre-treatments [NEE]

As discussed in the preceding project, a quality control measure must be installed to ensure less noise would enter the model training process, thus creating bias with the data-driven prediction. Therefore, a further quality control threshold is passed to the half-hourly measured Carbon Flux (FC). Any half-hourly indices with a value higher than +50 or lower than -50 are removed before any aggregations. FC, with the unit [μmol CO2 m^{-2} s^{-1}], are converted to [g C m^{-2} $(30\ min)^{-1}$] by a scaling factor of 12 [gC mol^{-1}]*10^{-6}*1800 [30min/s]. The quality-controlled NEE measurements are then aggregated into hourly sums to match the time steps with ERAint.

One intrinsic problem of Eddy Covariance measurement is that data gaps are frequent. Therefore, the quality of the measured FC/NEE must be examined to determine any potential impacts due to the temporal imbalance of non-NA data concentration. Figure 4 shows that gaps exist in June – October 2019 and November 2020 for Forest. For Pasture, no major gaps exist. 6 months (forest) and 4 months (pasture) have data availability <50%. Overall, 42.57% of Forest and 62.23% of Pasture data are available within the 2 years.
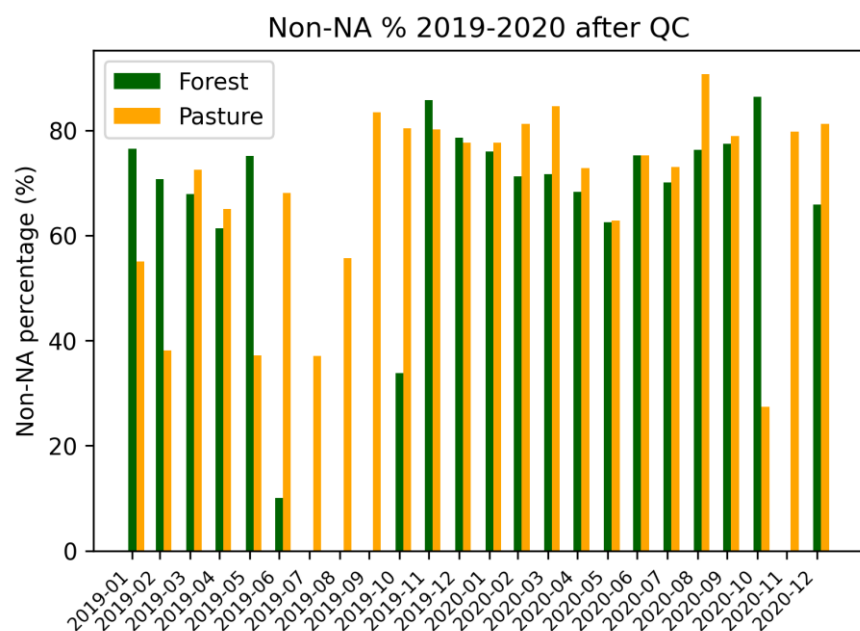
Figure 5. Monthly Data availability (in %) after quality control (QC) for the 2 EC measurements.

Data Processing and homogenization

Following DEM-interpolation, the variables are prepared to be merged and fit in a unified grid system based on the coarsest spatially-gridded dataset, which is ERAint's 14x14 grids at 30 arc seconds (1km) resolution. Temporally, the coarsest time steps are the yearly MODIS Land Use Classifications (LC), which is inappropriate in terms of achieving meaningful predictions. Therefore, it is decided that the time step will be unified to be daily, which means that any datasets with coarser temporal resolution (8-days-mean for LAI, faPAR; 16-days-mean for NDVI, EVI; Yearly for LC) are extrapolated using xarray's embedded dataset.interp() function, in which the method here selected is 'nearest', i.e., Nearest Neighbour Interpolation, in which the daily aggregate is equal to their closest temporal available data point. All files are first aggregated into hourly temporal steps, and the daily aggregated mean is then calculated.

Table 3. File management, spatial and temporal resolutions of different variables before homogenization.

File	Spatial resolution	Temporal resolution	Variables
ERAint	**30 arcsec (~1km x 1km) Single level**	Hourly	DEM interpolated: t2m, d2m, sp, u10, v10 Resampled: SWC, fal, slhf, ssr, str, sshf, e
MODIS250	250m x 250m	16-days averaged	NDVI, EVI
MODIS	500m x 500m	8-days averaged	LAI, faPAR
MODIS_extra	1km x 1km	Yearly	LC
MODIS_extra2	250m x 250m	Yearly	TC, nTC, Nonveg
/	/	Daily	DOY
NEE	NA (Point Data)	Half-hourly (0.5h)	NEE

All daily-aggregated files are then combined simply using the xr.merge() function in the x-array package in Python, with the parameter 'compat' set to 'override' to override the coordinate system. The results, named 'All_var.nc', is a data set with spatial resolution of the DEM (~1 x 1km) and daily temporal timesteps and is used as the input data for the following steps and model training.
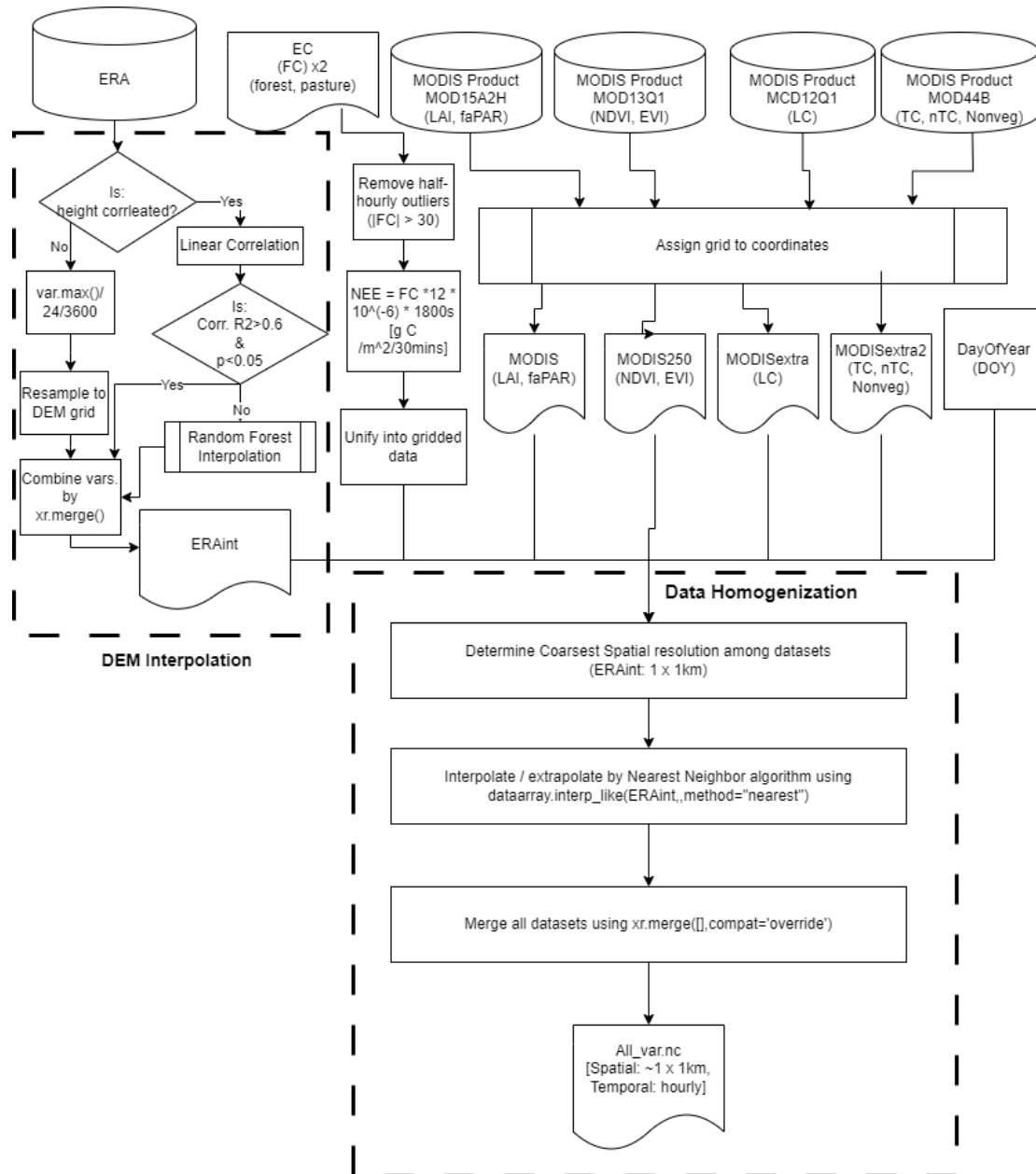
Figure 6. Full Schematic Diagram of the data preparation procedures, including DEM interpolation, merging, and data homogenization.

Feature Selection

After obtaining a homogeneous dataset, i.e., Allvar.nc, the variables are in principle ready to be used as fitting of an RF model. However, due to both the concern in the calculation cost, i.e., the time needed to fit and predict the independent variable NEE, and the quality of the final predictions, Feature Selection has been done to get a subset of variables capable of achieving higher model quality. Different algorithms could produce a different subset of variables and therefore should be tested side by side.

The most straightforward feature selection algorithm is to select variables based on the Pearson correlation Coefficient (R) between them and NEE. However, due to the complexity of Random Forest, variables with low R-values could also play an important role in the RF prediction. In this study, we opted to select variables with a Pearsons Correlation Coefficient (R) to daily NEE values higher than ±0.15, i.e., $|R| > 0.15$.

Another straightforward algorithm is making No Selection. This guarantees all important features would not be dismissed. However, not only does it have a higher calculative burden on the processor, but overfitting could also occur.

Besides the fundamental selection, two advanced feature selection algorithms are introduced. They are both sequential selectors which are greedy search algorithms, which try to find the most crucial variables by passing through a scoring system, which is in this case, Root Mean Squared Error (RMSE) from a recursive examination process. For BFE, in each iteration, a subset of the k input variables with k-1 variables are fitted to an RF regressor and then evaluated. The highest performance subset is then used as the input of the next iteration. For FFS, it starts with an empty set and includes the variables with the highest scoring. The (sub-)set of the variables is then fitted with another variable and the scores. In this study, we implement a modified algorithm with a floating mechanism, in which before the recursion each variable that is in the subset is excluded and removed if it improves the score (Pudil et al., 1994).

Table 4. Summary of all Feature Selection Algorithms used and examined in this study.

Feature Selection Algorithm	Short name	Criterion	Cross Validations (K-folds)	Selected variables
Correlation	COR	$\|R\| > 0.15$	No	't2m', 'sp', 'u10', 'height', 'ssr', 'e', 'LC'
No Selection	ALL	NA	No	All
Forward (Floating) Feature Selections	FFS	Inclusion with the best scenario	12	't2m', 'u10', 'height', 'fal', 'slhf', 'ssr', 'str', 'sshf', 'e', 'LAI', 'DayofYear'
Backward (Floating) Feature Elimination	BFE	Exclusion from the worst scenario	12	't2m', 'sp', 'u10', 'SWC', 'fal', 'slhf', 'ssr', 'str', 'e', 'ndvi', 'LAI', 'DayofYear'

Due to the complexity of both Sequential Feature Selection mechanisms, it is done with mlxtend.feature_selection (Raschka, 2018) with the scoring method set as negative RMSE as defined in scikit-learn (Pedregosa et al., 2012). The input is separated into 12 random folds. Each run one-fold will be designated as testing data and the others will be training, which is fitted and repeated with each fold as validation for once. The mean of the RMSE is then the criterion to determine the best (FFS) / worst (BFE) feature.

With the four approaches with variable subsets, four RF models (COR, ALL, FFS, and BFE, named after the feature selection algorithms used) can be fitted and examined independently.

Hyperparameter Tuning
An important step of establishing any Machine Learning approach is to get an optimal set of parameters determining, in the case of RF, the trees and forest structures. This allows the regressor to optimize the structure such that the forest can reach a maximum predicting power.

However, tuning could be theoretically limitless due to the sheer complexity of the list of possible parameters. As previously discussed, one advantage of RF is that the tuning is not as complicated if we consider only a few most important hyperparameters. A randomized Cross Validation Search, in which only a fixed amount (200) of randomized hyperparameter sets is tested with 12-fold Cross Validation, and the one with the highest after-CV-performance (in accuracy in sci-kit learn package) is selected as the best set of hyperparameters, which is used for all models.

Table 5. List of tuned hyperparameters with the testing ranges, name in the sci-kit-learn library, and the tuned values. The names in the RandomForest Package in R are included as references.

Hyperparameters name (scikit-learn)	Description	Range	Tuned Values
n_estimators	Number of trees in an ensemble	500:3000 by 50	2700
max_features	The number of features to consider when looking for the best split	['auto', 'sqrt']	'sqrt'
max_depth	The maximum depth of the tree.	10:110 by 10, None	None
min_samples_split	The minimum number of samples required to split an internal node	[2, 5, 10]	5
min_samples_leaf	The minimum number of samples required to be at a leaf node.	[1, 2, 4]	2
bootstrap	Whether bootstrap samples are used when building trees	[True, False]	True

Model fitting

Four models are fitted with the four variable subsets with the same procedure. The original dataset is converted to a Pandas data frame and sliced with the variable subset. The positions with NA values (spatially all grids except the 2 grids with station data, temporally the data gaps of either measurement) are dropped and the data frame is then reshaped in a vertical array. NEE are then separated before fitting.

The prediction is run with the same array with NEE dropped, which is then added back to the data frame with the coordinates. The data frame is then converted back to a 3-dimensional data array.

Cross Validations of the models

Apart from the implementation of IML as the means to interpret the model itself, the performance of such models must be validated and compared. Due to the sparsely available data points distributed spatially, in which the 2 grids correspond to the 2 measurements, 3-dimensional cross-validation is deemed not possible. Therefore, only the temporal dimension will be tested with a K-fold CV, where K is the number of runs and the number of entries to be separated into. As seen in Reitz et al. (2021), the acquired performance indicators vary with the number K. It is thus crucial to test with different K values to assess the performance fully.

Interpretable Machine Learning (IML)

Due to the black box nature and the sheer numbers involved in an ensemble, it is impossible to directly examine the fitted forest in a meaningful way. Therefore, interpretable Machine Learning approaches are implemented to estimate the logic behind them. In this study, we investigate the (permutation) feature importances, Accumulated Local Effects (ALE) plots, and SHapley Additive exPlanations (SHAP) values. Permutation Feature Importance is an increase of error after a random permutation of the feature value. The higher the feature importance, the higher it would be for prediction. ALE plot shows the impact of each variable on the final result, which helps find and identify the undermining relationship between the dependent variables to the prediction, disregarding the relations between features unlike the more popular Partial Dependence Plots (PDP) (Molnar, 2023).

Partial dependence plots (PDPs) and Accumulated Local Effect plots (ALE) are methods to interpret the global effects on the prediction by individual variables. Both works by visualizing the effect of each variable on the prediction. However, PDPs are known to be unreliable, especially when variables are correlated with each other, which is a violation of the assumption of independence for PDPs (Molnar, 2023). Therefore, ALE plots are prepared as a means to qualitatively examine individual variables' effect on prediction.

ALE is often described as a global ML interpretation method, referring to its ability to depict the average behavior of an ML model. Even global interpretations give valuable insight into how the model would perform on an average basis, local interpretation methods can provide an in-depth analysis of individual models and their predictions. SHAP values were calculated as a means to assert and observe how the prediction of individual models changes with different values of input in individual models (Lundberg, Scott & Lee, 2017). It can be quantitatively interpreted as the contribution of each input variable to the predictions (Molnar, 2023), therefore useful in terms of interpreting each of the predictions. The SHAP values and the plots are calculated and generated with the SHAP library (Lundberg, Scott M. et al., 2020).

Results

Selection by Simple Correlation (COR)

Figure 7. Temporal mean value of daily aggregated NEE predicted by model COR plotted with contour of height with 100m interval (left), the spatial spread of the temporal mean compared to the measured NEE (middle), and the spatial-temporal spread of the predicted NEE [gC m^{-2}d^{-1}]

Figure 8. Comparison of daily aggregated NEE predicted by model COR with the 2 measurements (forest for above, pasture for below) plotting with line plots (left) and violin plots (right).
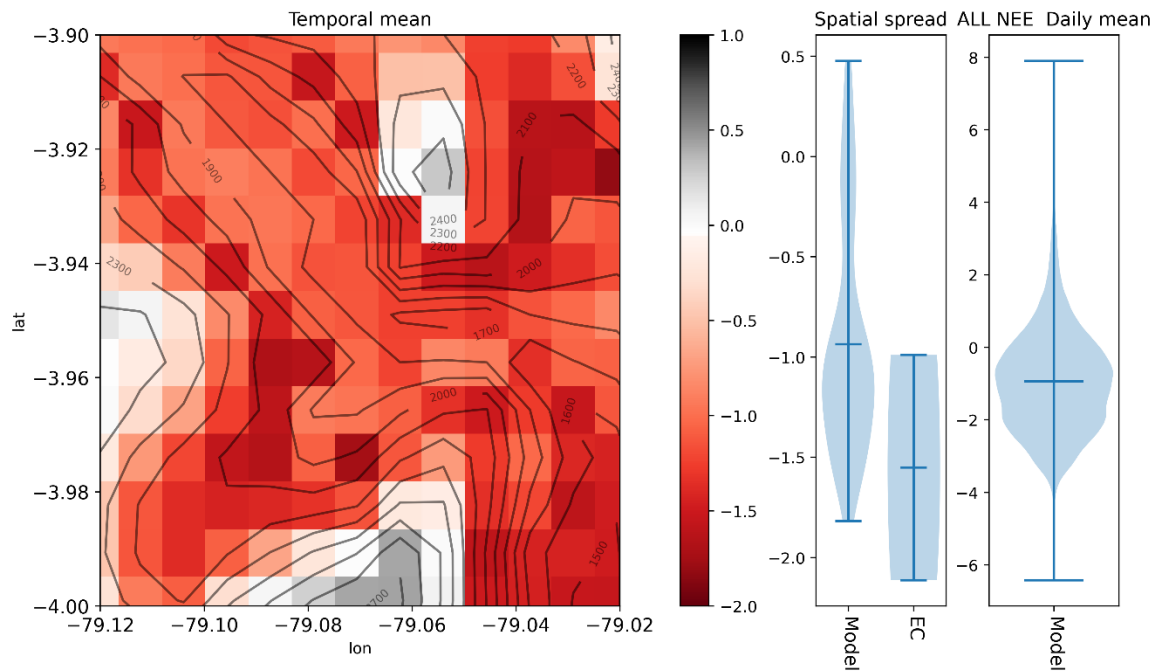
Figure 9. Temporal mean value of daily aggregated NEE predicted by model ALL plotted with contour of height with 100m interval (left), the spatial spread of the temporal mean compared to the measured NEE (middle), and the spatial-temporal spread of the predicted NEE [gC m^{-2}d^{-1}]
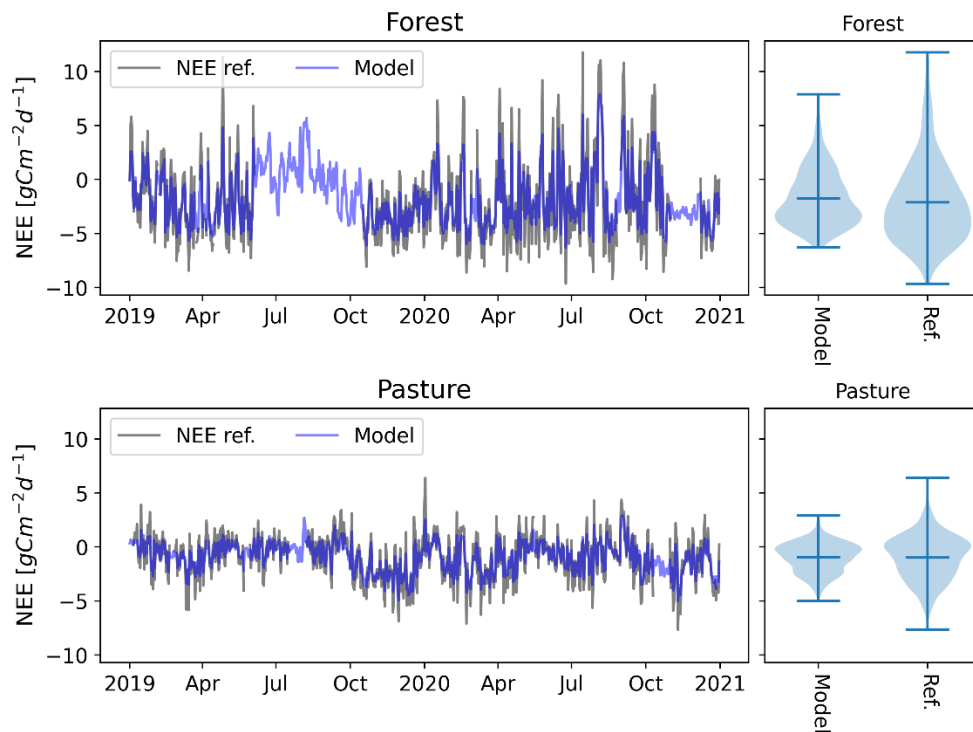
Figure 10. Comparison of daily aggregated NEE predicted by model ALL with the 2 measurements (forest for above, pasture for below) plotting with line plots (left) and violin plots (right).
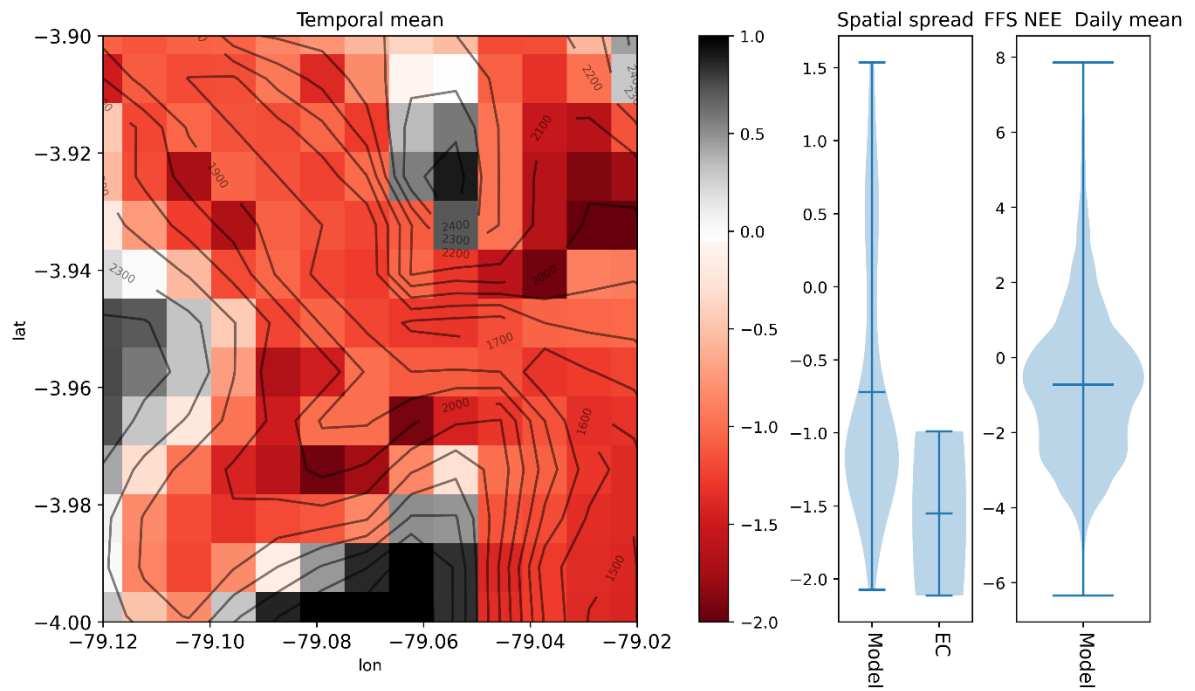
Figure 11. Temporal mean value of daily aggregated NEE predicted by model FFS plotted with contour of height with 100m interval (left), the spatial spread of the temporal mean compared to the measured NEE (middle), and the spatial-temporal spread of the predicted NEE [gC m^{-2}d^{-1}]
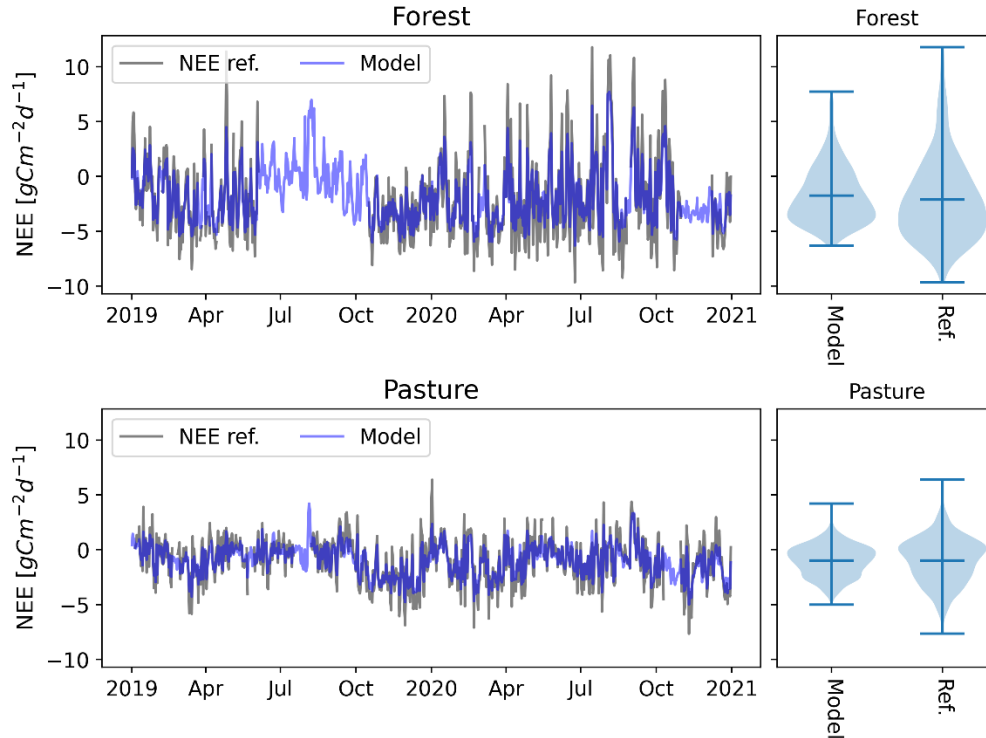
Figure 12. Comparison of daily aggregated NEE predicted by model FFS with the 2 measurements (forest for above, pasture for below) plotting with line plots (left) and violin plots (right).
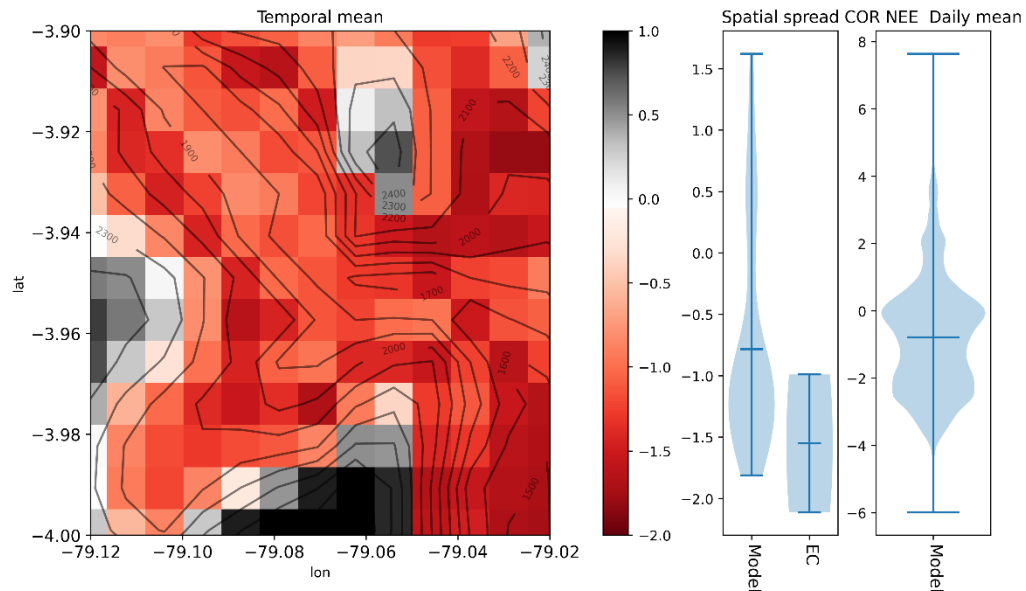
Backward Feature Elimination (BFE)

Figure 13. Temporal mean value of daily aggregated NEE predicted by model BFE plotted with contour of height with 100m interval (left), the spatial spread of the temporal mean compared to the measured NEE (middle), and the spatial-temporal spread of the predicted NEE [gC m^{-2}d^{-1}]
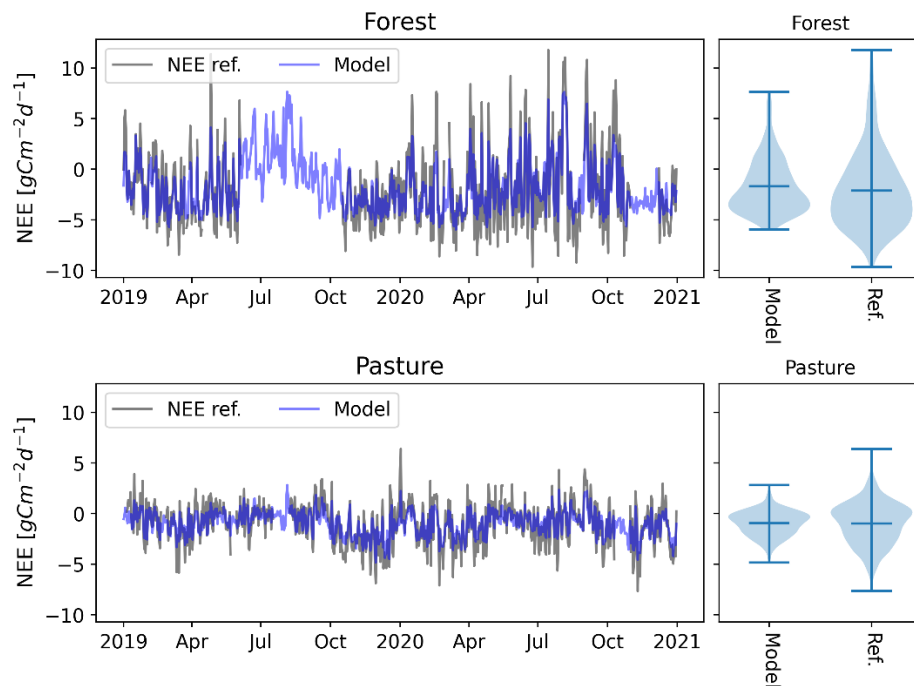

Figure 14. Comparison of daily aggregated NEE predicted by model BFE with the 2 measurements (forest for above, pasture for below) plotting with line plots (left) and violin plots (right).

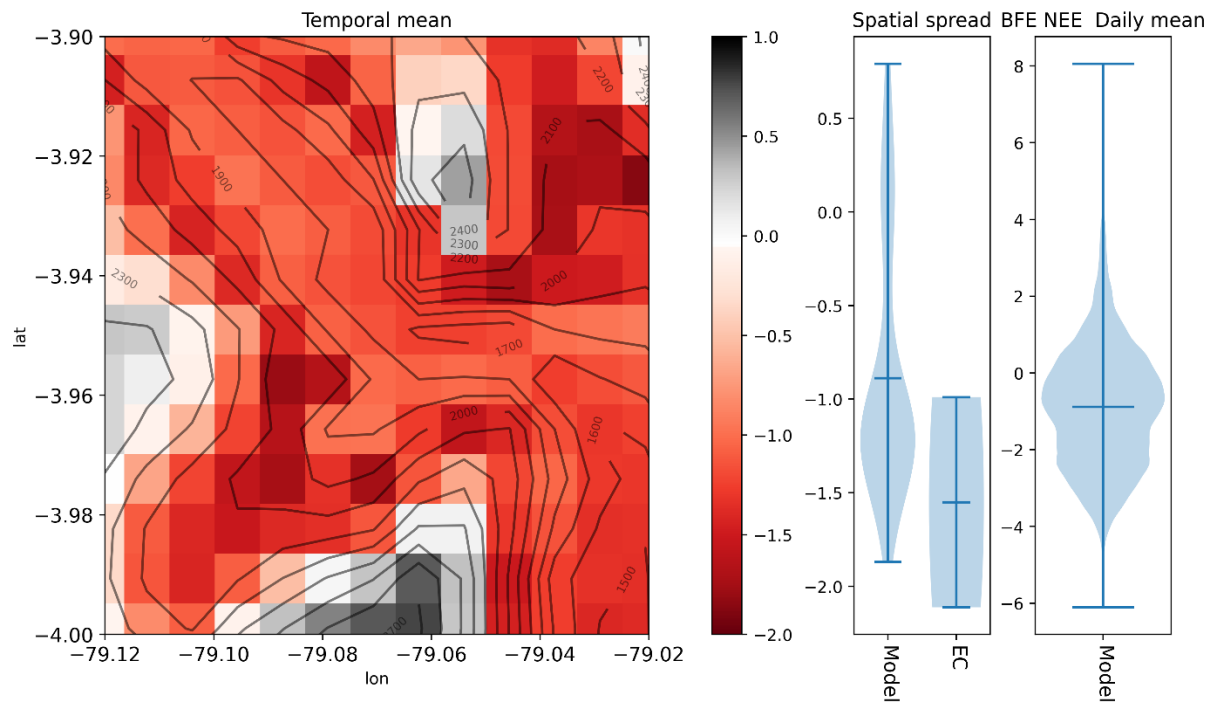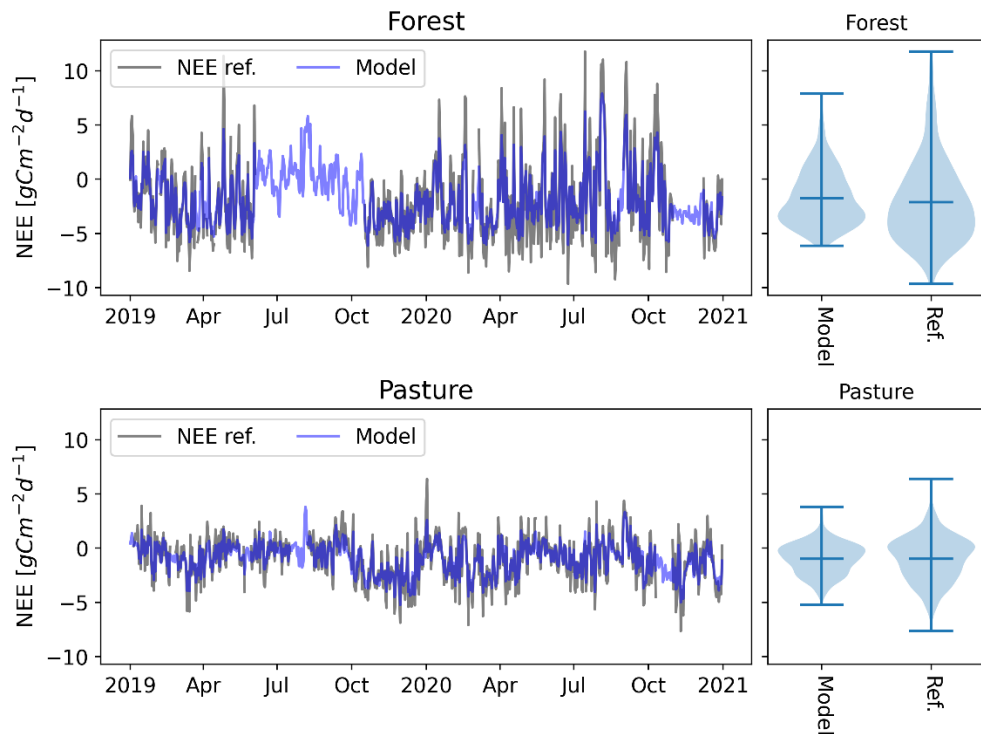Figures 7 to 14 show the out-of-bag model upscaling results using the four sets of variables with four selection algorithms. All temporal mean plots show that NEE is correlated to the topography shown as the contour. The general pattern shows a negative signal in the valley and a positive above on the peaks with surface height >2300m a.s.l.

However, different models return some slight differences with the temporal mean, i.e., the prediction of carbon sink (for negative NEE) or source (for positive NEE). In the first instance, COR and FFS return a stronger positive NEE on the southern slope above 2300m a.s.l. ALL predicted weaker positive NEE for the same area, while BFE has the same spatial pattern but an even weaker positive signal, with the 3 highest elevation grids falling between +0.5 to 0 $gCm^{-2}d^{-1}$. Similar patterns can be seen in the peak north and the slope west of the zonal valley. The positive signal >2300m a.s.l. are the highest in FFS and COR (NEE>0.5 $gCm^{-2}d^{-1}$), while ALL and BFE predicted a weaker positive at the peak ($0 \leq NEE \leq +0.5$, $gCm^{-2}d^{-1}$).

Another interesting aspect of inter-model variations is in the treatment of the valley. FFS shows stronger signals disregarding positive or negative. In the upper valleys (west of 79.06E), COR and ALL show a weaker negative signal with height changes, which can be seen in the slopes especially the southern side of the northwestern valley. ALL has a seeming noise generated in the same area. In the lower valley (east of 79.06E), COR and ALL generated stronger negative signals in the southeastern valley than BFE and FFS in the same area.

Throughout the target area, the strongest carbon sinks are predicted in the area with heights between 2000m and 2100m a.s.l. and at the lower valley below the attitude of 1600m a.s.l. in all models. All of the models are rather consistent in this aspect.

These variations of upscaled NEE with altitude have led to a wide spread of temporal mean NEE, and therefore the stronger the signals the model generated, the higher the spatial spread of temporal mean NEE observed. All models' predictions have some overlapping with the measurement, with FFS and BFE covering the most range of measured NEE. ALL have the least of their coverage to the measured NEE mean, with some positive outliers.

The temporal spread of prediction in the two measuring grids compared with the EC measurements (Figures 8,10, 12, and 14) shows similar patterns across all models, which means the models successfully replicated the patterns of the inputs. However, all models

generated smaller peaks than EC measurements. It can also be seen in the smaller spreads in the violin plots. Another interesting point is that all four models have a similar shape during the larger gaps in the forest (June – October 2019).

Table 6. Out-of-bag (OOB) model performance in terms of Root-mean-squared-error (RMSE).

Model	Input variables	Mean RMSE (pasture, forest) [gCm^{-2}d^{-1}]
COR	't2m', 'sp', 'u10', 'height', 'ssr', 'e', 'LC'	1.0773 (0.69429, 1.4021)
ALL	All	0.93339 (0.63646, 1.1898)
FFS	't2m', 'u10', 'height', 'fal', 'slhf', 'ssr', 'str', 'sshf', 'e', 'LAI', 'DayofYear'	0.91956 (0.68157, 1.2353)
BFE	't2m', 'sp', 'u10', 'SWC', 'fal', 'slhf', 'ssr', 'str', 'e', 'ndvi', 'LAI', 'DayofYear'	0.92714 (0.66365, 1.2194)

Table 6 shows the out-of-bag (OOB) performance of the four models, which means the RMSEs are calculated with the prediction and the measurement in their full lengths. It does not represent the predictability of the individual models, but rather to see how close the prediction is to the each of original measurements, i.e., the reproducibility of the input. FFS achieved the best in terms of the OOB RMSE, followed by COR and BFE. COR achieved relatively worse in OOB RMSE. In terms of the individual points, the Forest measurement had a higher RMSE than pasture in all models. COR had the highest RMSE in pasture and forest.

Compared to the other studies, the RMSEs of all of the models are lower than any of the combinations tested in the previous study (Koo, 2023), where the RF model reached an RMSE of 1.093 gCm^{-2}d^{-1}.

Feature Importance

Permutation Feature Importance is the measurement of error after feature permutation, in which a feature is more important if the permutation of that feature results in higher error (Breiman, 2001; Molnar, 2023). Generally, it provides a compact and global insight into the drivers the model found to link the independent variable, i.e., NEE in this case. The higher the permutation feature importance the variables had in each of the models, the stronger the effect it has on the prediction. This can be concluded as the variable with high importance was considered a stronger driver by the individual models.
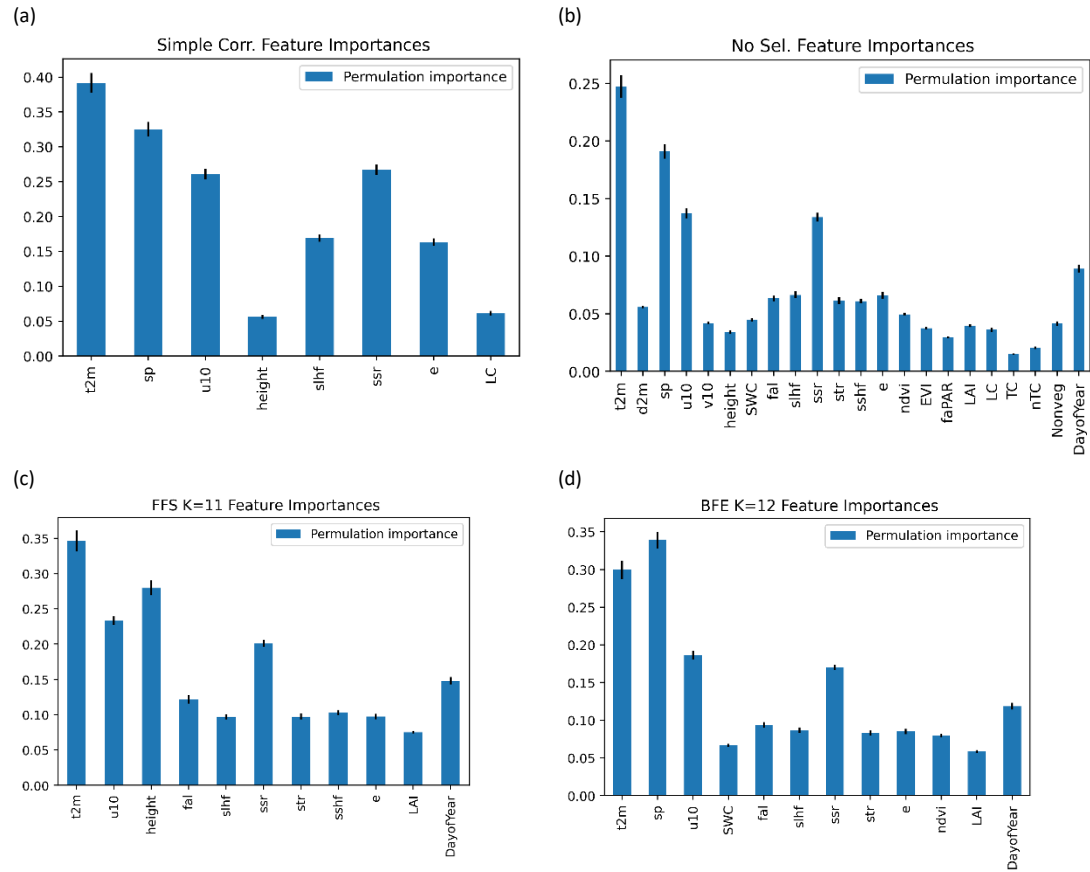
Figure 15. Feature Importance plotted to each selected variable from each selection algorithm: (a) COR, (b) ALL, (c) FFS, and (d) BFE.

All models show a similar set of variables with the highest importance, which is t2m (highest in COR, ALL, FFS, second highest in BFE), sp (highest in BFE, second highest in COR and ALL), and u10 (third in ALL and BFE, second in FFS, fourth in COR). Other important features include ssr, e (both in all 4 models), and DayofYear (ALL, FFS, BFE). Compared to the previous study, where the most important variables of forest and pasture were sp, slhf, and u10, and albedo, sp, and air temperature respectively (Koo, 2023), t2m and sp are important in both studies. On the other hand, albedo is not deemed important in the models in the current study.

Cross Validations

Cross-validation with different numbers of folds (K=[48, 45, 42,36, 2]) is done along the temporal axis. Table 7 shows the mean Root-Mean-Squared-Error (RMSE), mean Mean Absolute Error (MAE), and squared Peasons Correlation Coefficient (R^2) across all folds with all tested K values.

Table 7. Cross Validation results with different model and number of temporal folds (K=48, 45, 42, 36, 2) assessed with Root-Mean-Squared-Error (RMSE; gCm^{-2}d^{-1}), Mean Absolute Error (MAE; gCm^{-2}d^{-1}), and squared Pearsons Correlation Coefficient (R^2) and their standard deviation (sd). **Bold** indicates best performance within the same number of folds (K).

Folds (K)	Model	RMSE	RMSE SD	MAE	MAE SD	R^2	R^2 SD
48	COR	2.536061	0.977153	2.020719	0.744109	0.166771	0.501934
	ALL	2.505288	0.998961	1.987833	0.739335	0.112799	0.407516
	FFS	2.515045	0.994773	1.993163	0.758595	**0.169736**	0.568367
	BFE	**2.483198**	0.971525	**1.97832**	0.732518	0.121472	0.529584
45	COR	2.548789	0.959881	2.042734	0.747458	0.190202	0.636075
	ALL	2.527453	1.01801	2.015372	0.782343	0.127155	0.505226
	FFS	2.516468	0.99027	2.008973	0.765202	**0.211517**	0.913252
	BFE	**2.497732**	0.979465	**2.001078**	0.759934	0.209951	1.060968
42	COR	2.556094	0.944461	2.031035	0.719653	**0.100083**	0.39607
	ALL	2.519513	0.96905	1.993053	0.724996	0.047541	0.306654
	FFS	2.512993	0.950603	1.984756	0.714484	0.095347	0.502851
	BFE	**2.497017**	0.938057	**1.966539**	0.711704	0.0557	0.448263
36	COR	2.571448	0.931997	2.040215	0.726534	**0.140385**	0.403764
	ALL	2.527109	0.972596	1.991642	0.73252	0.072957	0.33626
	FFS	2.544378	0.967226	2.002874	0.730619	0.139576	0.494705
	BFE	**2.502435**	0.937583	**1.975062**	0.717178	0.083418	0.442346
2 (per year)	COR	2.630988	0.450128	2.022837	0.270953	-0.02469	0.013613
	ALL	2.630803	0.560712	2.012457	0.349109	-0.03584	0.094688
	FFS	2.682727	0.714091	2.082268	0.502635	**-0.01456**	0.197817
	BFE	**2.568816**	0.559101	**1.966343**	0.349552	-0.08877	0.091671

RMSEs in each model generally increased with decreasing K, i.e., increasing fold size. Across all models, BFE has the least RMSE in all K. SDs of RMSEs showed a decreasing trend with decreasing folds K, and across different models, BFE has the least sd of RMSE in K=42 and K=48. MAE increased with decreasing K, and across models BFE performed the best with the least MAE in all K. R^2 has the general decreasing trend with decreasing K but the K=42 has the highest R^2 value in all models, which is consistent with Reitz et al. (2021). However, the standard deviation shows a large spread across different folds, which means that across different folds the correlation ranged between a strong correlation (16% of the folds = 8/48 folds has a value >mean+1 sd=0.64 for BFE with K=48, assuming a normal distribution) and no correlations at all. Overall, due to the lowest RMSE and MAE, BFE is the best-performing model.

The cross-validation scores are comparable to other similar studies. The main reference for this study, which used RF models with sequential Feature Selection algorithms to upscale NEE in western Germany, reported a cv score of 48-fold RMSE for FFS, BFE, and

no selection (comparable to ALL) of 2.54, 2.55, and 2.91 ($gCm^{-2}d^{-1}$), MAE of 1.89, 1.78, and 2.02 ($gCm^{-2}d^{-1}$), and R^2 values of 0.31, 0.32, and 0.24 (Reitz et al., 2021). This study achieved a lower RMSE but higher MAE, while R^2 values in this study were significantly worse. Another study that uses Remote Sensing and meteorological data to derive a global NEE model with similar algorithms produced an RMSE of 1.64 and 0.68 ($gCm^{-2}d^{-1}$) in deciduous broadleaf forest and savannah with IGBP classification (Zhuravlev et al., 2022), which is significantly better than in this study. This could indicate room for improvement with either a coarser resolution or more input.

Upscaling results

All models had the predicted NEE value within the range of around -6 gC m^{-2} d^{-1} to 8 gC m^{-1} d^{-1}, where the median lay at around -1 gC m^{-1} d^{-1}. Across the year, in November, December, and March EC recorded the strongest negative NEE median as around -2.5 gC m^{-2} d^{-1}. The strongest positive NEE signal is measured in September. Across all models, the strongest negative occurred in the same months as EC, i.e., March or November. However, all models predicted either July or August having the highest positive NEE.

Throughout the target area, the largest differences lie between the peaks and the valleys, where there is likely to be a positive NEE on the peaks and a negative NEE in the valley. However, different models had a different cut-off height. ALL and BFE have a higher cut-off height of 2400m, which includes only 1 pixel in the northern peak and 4 pixels in the southern slope into positive predicted NEE. The cut-off height of COR and FFS is seemingly lower at around 2200m. Figure 16 shows the comparison between models and EC measurements with different heights. NEE is more positive on the peaks across all models, where more months are positive than the lower grounds. In the height of the EC towers (1800m ≤height ≤ 2200m), only Winter in 2019 saw positive NEE. All models predicted a similar trend in the valley with some occasional outliers, mostly for FFS and COR. On the peaks, the spreads between models are not only higher on average but are even higher in winter (JJA) months, with FFS almost consistently producing stronger positive signals than others. The only exception of all models converging is in October and November. For the lower valley grids with grid height <1800m, most of the months converged except COR. Large divergence can be noticed in February and March. Overall, most models converged in the lower height and the training height (1800m ≤height ≤ 2200m).

Due to the diverging predictions on the peaks, it can be concluded that the models are less confident about the peak in terms of predicting precision, especially in Winter. It could be a result of the lack of training data. Two EC stations are situated at 1900m and

2000m in their respective grids in GTOPO30 DEM. There is a chance that the model simply cannot fully assess the variables' relation to NEE due to a lack of training data in the higher elevation. Figure 16 proved this point by showing that, close to the grid height of the EC towers all models converged, while below or above the tower height yielded divergences between the models.
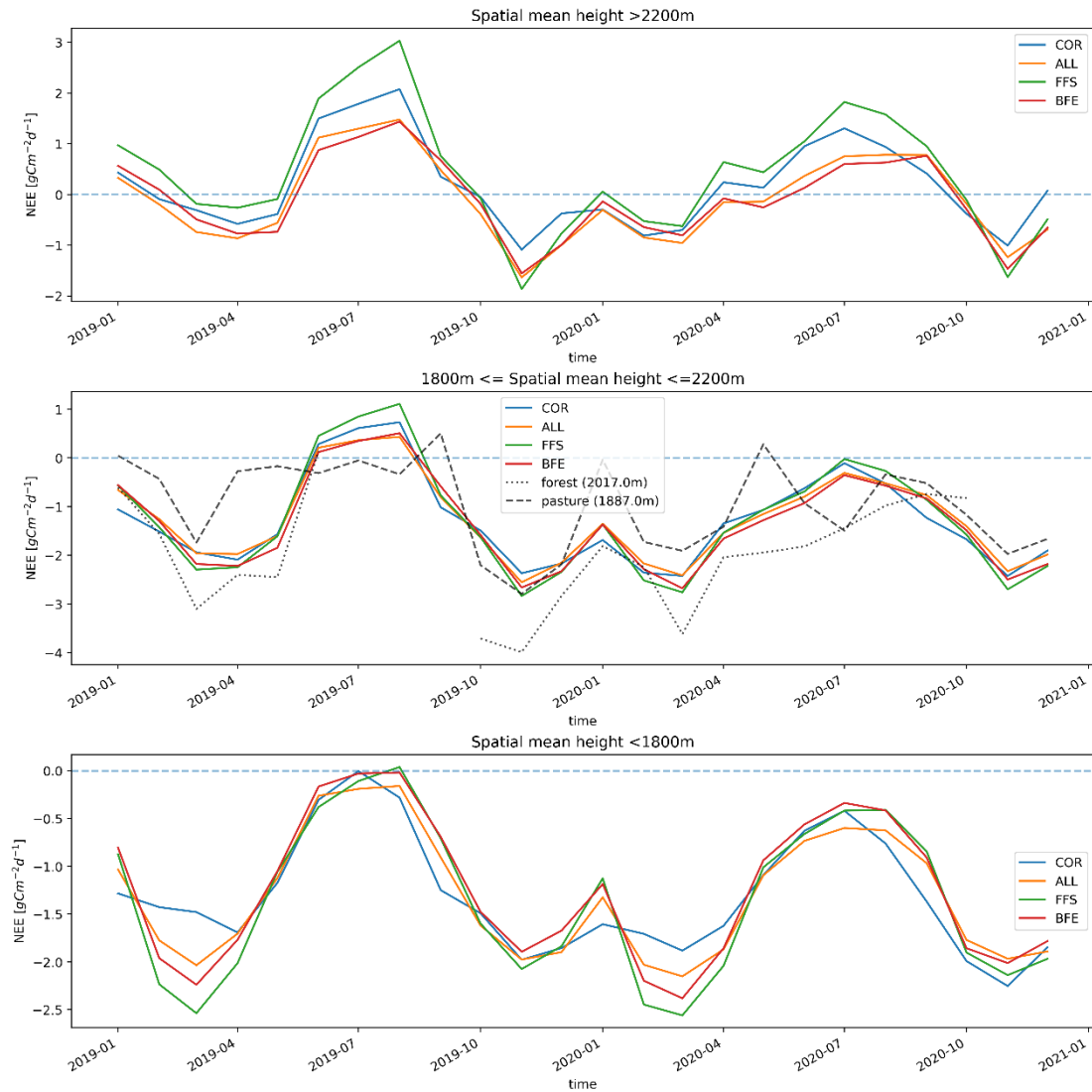
Figure 16. Spatial means of predicted and measured NEE with grid height >2200m a.s.l. (top), between 1800 and 2200m a.s.l. (middle) and ≤2200m a.s.l. (bottom).

Seasonality and temporal trend

The region has an aforementioned high seasonality with precipitation but not temperatures. Therefore, the seasonality of the upscaled NEE was inspected to identify any seasonal patterns between the higher and lower grounds. Figure 17 shows that the

models predicted the general seasonal trends which is similar to the EC measurement. However, all models predicted a larger-than-50% positive NEE during Winter (JJA) while EC stayed mostly negative. This period corresponds to the relatively colder months and the drier periods. However, it could also be the effect of the data gaps (June – October 2019) or simply the effect of higher grounds.
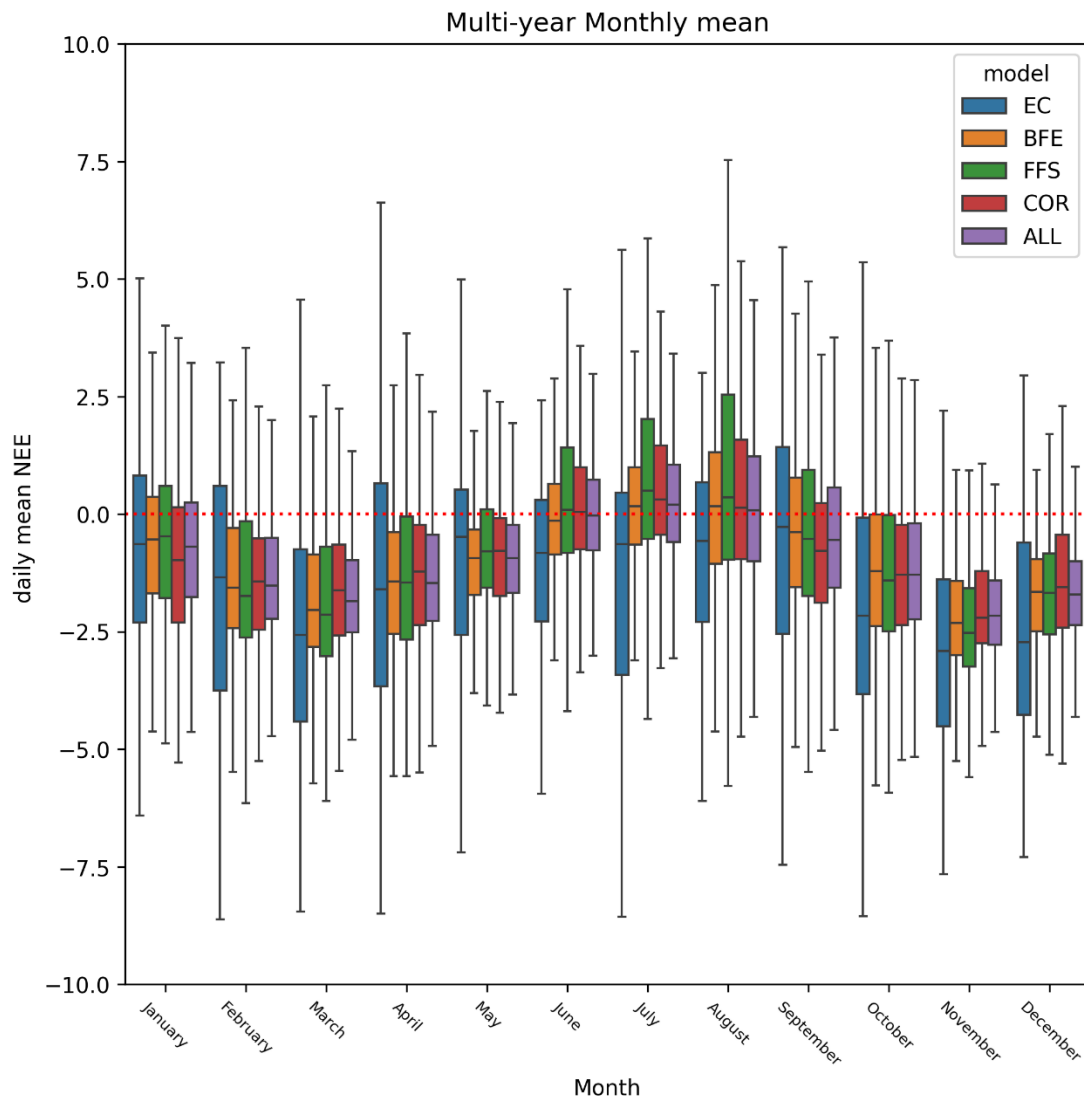
Figure 17. Monthly aggregated mean of predicted daily NEE (in unit gC m^{-2}d^{-1}) across the area from the models COR, ALL, FFS, and BFE compared to the mean of the two EC measurements. The boxes represent the lower value: 5 percentile, 25 percentile (Q1), median (Q2), and 75 percentile (Qe). Outliers (< median-1.5*IQR or > median +1.5*IQR, where IQR=Q3-Q1) that are outside of the whiskers are omitted.

Figure 18. Seasonal panel plots with columns in the models (from left: COR, ALL, FFS, and BFE) and rows in austral seasons (DJA: Summer; MAM: Autumn; JJA: Winter; SON: Spring). Greys represent positive and reds represent negative, with the colour bar unit as [gC m^{-2} d^{-1}]
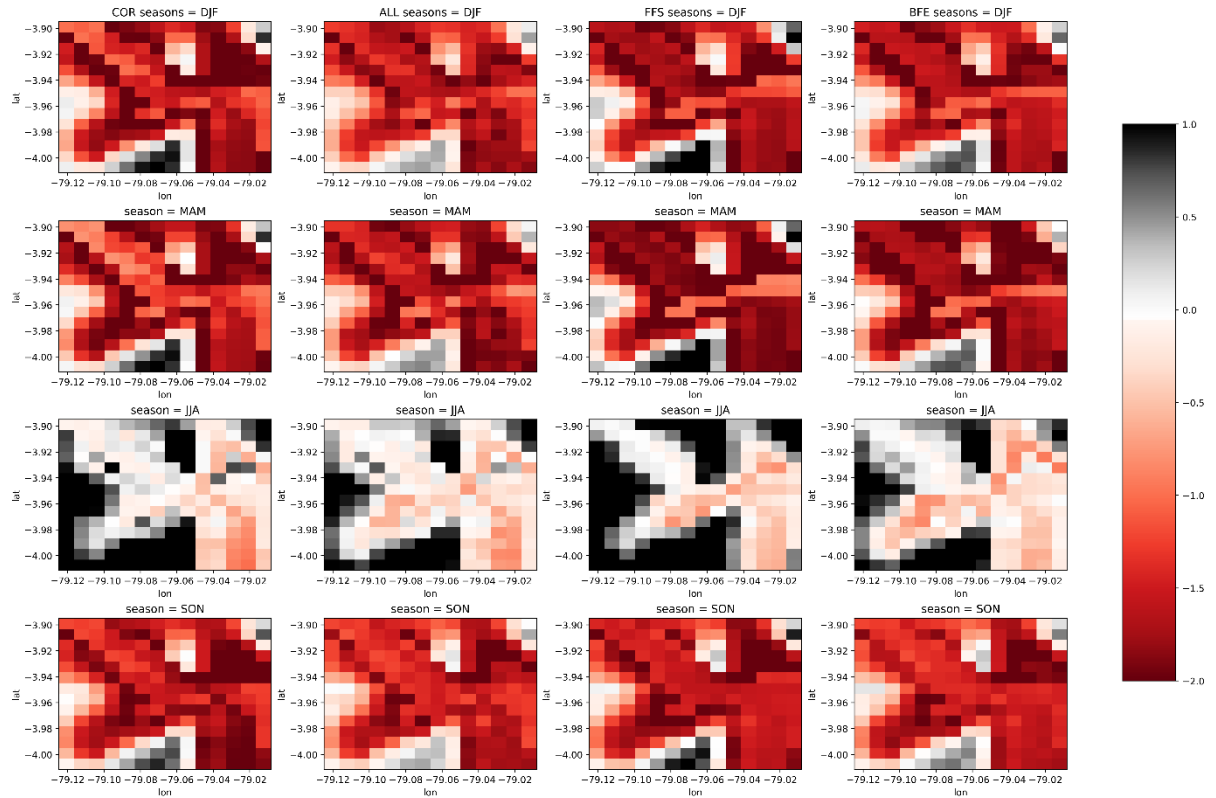
Figure 18 shows the seasonal patterns of all models. All of the models show a similar pattern across different models, in which a positive (for Winter) and weakly negative (for all other seasons) signal especially on the higher grounds and slopes were predicted. However, the magnitude differs across models, but not by much in different seasons. During austral Summer (DJF), all models show the slopes at the edge of the valley having a stronger negative NEE than the middle. However, during Winter and Spring (MAM) FFS and BFE show a significantly weaker negative signal as a line extending from the southwestern valley towards the east. COR and ALL have a similar pattern, albeit being distorted and not as clear. In Winter (JJA), almost all grids are >0 gC m^{-2} d^{-1} except the southeastern, lower valley, and near to the mentioned line were predicted a weaker negative, while the upper valley was close to zero. Across the models, COR had its stronger negative signal staying east of 79.06°E, while ALL extended the negative signal to all valleys (NW and SW valleys). FFS has the strongest negative signal between the southwestern valley and northeast as a straighter line, while BFE predicted the strongest negative to follow the 2100m contour line in the northeastern corner. In Spring, NEE showed a negative signal, although weaker than in Summer and autumn. COR predicted the strongest negative in the middle of the lower valley, the northeastern corner, and along the slope in the upper valley. ALL and FFS only showed a strong signal along the

slope and in the northeastern corner. BFE had the strongest negative NEE in the northeastern corner and the opening to the southwestern valley.

Best Model Diagnostics
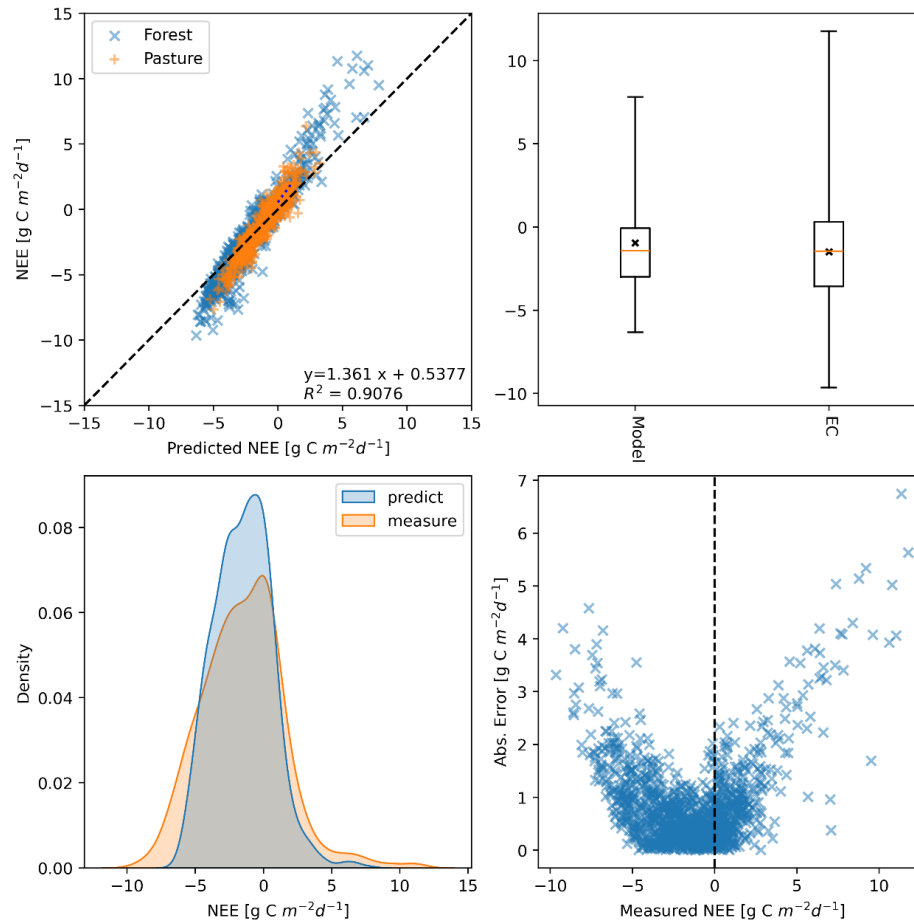
Figure 19. Diagnostic Assessment of prediction quality in comparison to EC tower measurements, including a scatter plot with measurement (forest and pasture) against vs prediction and the regression equation (top-left, a); boxplot with the prediction (model) and measurement (EC) with crosses depicting mean values (top-right, b); probability density distribution oof prediction and observation (bottom-left, c); absolute error with observation NEE (bottom-right, d). The vertical line at zero marks the border between CO2 sinks (left) and CO2 sources (right).

Although BFE performed the best in cross-validation, the model prediction quality indicates a difference between the predicted values and the measurements. Figure 19(a) shows that the slope and y-intercept between measurement and prediction is 1.355 and +0.5377 gC m^{-2} d^{-1}, which could be interpreted as the model having +0.5377 gC m^{-2} d^{-1} error close to 0 NEE. This can also be seen in the mean of the model and EC, which is -0.9682 and -1.489 gC m^{-2} d^{-1} in figure 19(b) respectively, being around 0.5 gC m^{-2} d^{-1} apart. Nonetheless, the median of the model and EC shown in figure 19(b) are close (-

1.340 and -1.450 gC m^{-2} d^{-1}). The spread of the model, measured here by the Inter-quartile range, is smaller than EC, which was also reported by Reitz et al. (2021). It is also the result of the rather high distribution of predicted NEE towards the range between -5 and 0, which is also noted in Reitz et al. (2021) as the range of measured NEE being the most frequent. Therefore, it is justified that the absolute error increases outside the range in figure 19(d).

Interpretable Machine Learning

Accumulated Local Effect Plots
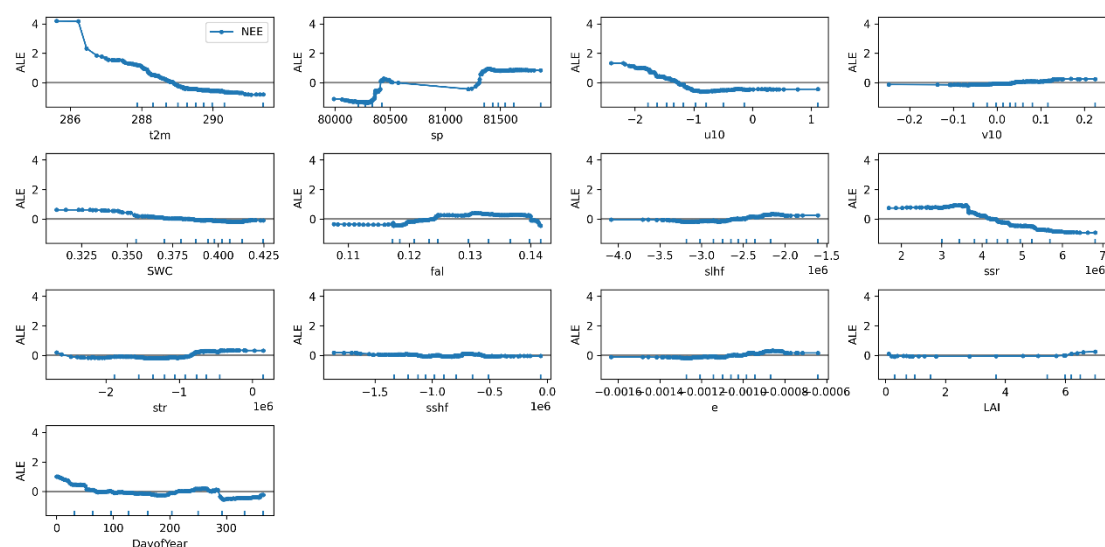
Figure 20. The accumulated Local Effect (ALE) Plots of BFE with ALE against 2-meter temperatures (t2m [K]), surface pressure (sp [pa]), 10-meter zonal wind (u10 [ms^{-1}]), 10-meter meridional wind (v10 [ms^{-1}]), surface water content (SWC [m^3 m^{-3}]), Forecast albedo (fal [/]), surface latent heat flux (slhf [Jm^{-2}]), surface net solar radiation (ssr [Jm^{-2}]), surface net thermal radiation (str [Jm^{-2}]), total evaporation (e, m), Terra Leaf Area Index (LAI), and Day of Year. Distributions of variables are shown as a rug on the x-axis. Produced with the Python package Alibi Explain (Klaise et al., 2021).

Figure 20 shows that the model is not linear, which conforms with the expectation of RF. Most of the variables selected by BFE do not have a strong variance to prediction, which can be seen as a flat and straight line with around 0 ALE. However, the shape can fluctuate randomly. SWC as an example has a flat and close to 0 ALE line when its values >0.350. It might seem that SWC<0.350 has a positive impact on predicting NEE values. Yet, the curve in that range is a mere extrapolation as indicated by the lack of distribution at the rug. Another example comes with ssr, in which the ALE curve has a clear cut at 3.5 Jm^{-2}, where ssr <3.5 Jm^{-2} has a positive impact on NEE values, and a

negative impact with >3.5 Jm^{-2}. These are hints of NEE being piecewise dependant on such variables.

The magnitude of ALE indicates the relative and average impact of each variable on the prediction. Figure 20 shows that the variables with the largest impact on the prediction are: t2m, sp, u10, ssr, and DayofYear, which conforms with the permutated feature importances (figure 15). T2m shows a relatively linear relationship in the range of distribution, with decreasing ALE with increasing t2m. Sp shows a positive monotonic relationship between sp and ALE. U10 shows 2-part relationships, with u10<-1ms^{-1} having an increasing ALE with increasing easterly wind speed. When u10>-1ms^{-1}, ALE does not change with u10, even when the wind is westerly (u10>0ms^{-1}). Day of Year decrease with in the first 3 months. Between April and October, the Day of the Year does not have a main effect on the prediction value. In November and December (day of Year >300), ALE dips into negative, indicating a negative NEE prediction tendency.

SHapley Additive exPlanations (SHAP)

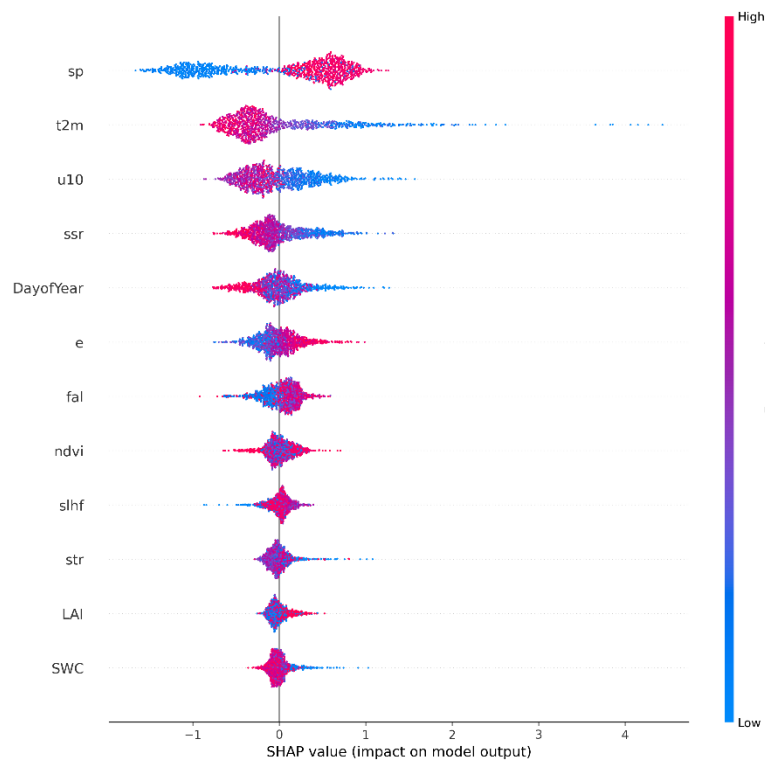Figure 21. SHAP Summary plots for model BFE, plotting the SHAP value of all variables in set BFE at 1212 indices within the training input's range. Positive (negative) SHAP indicates the prediction tendency to be higher (lower) than the expected (mean) value of NEE (-1.492 gC m^{-2} d^{-1}). Feature values are plotted with the colormap indicating the magnitude of the values within the training range.

Multiple points with similar SHAP values are jettisoned along the y-axis. SHAP values are computed and plotted with the SHAP library (Lundberg, Scott & Lee, 2017; Lundberg, Scott M. et al., 2020)

SHAP values help portray the effect of local predictions from different magnitudes of inputs. With the colors showing the input magnitude, the impact of individual variables can be easily identified. SWC, LAO, str, slhf, and NDVI show their minimal contribution in most instances, with some presumably extreme values having a minor influence. From Forecasted albedo (fal) and evaporation (e), the color started to separate, indicating a stronger influence from the input magnitude. Low albedo and evaporation values tend to decrease NEE, while medium (purple) and high fal values are more likely to increase the prediction value. Seasonality (shown by the variable DayofYear) shows only a slight tendency in high value (towards the end of the year) to have lower NEE and low DayofYear value (towards the start of the year) to have a higher NEE value. All other dates do not have a strong effect. High values of ssr, u10, and t2m have negative effects on prediction value, while low value has a relatively smaller chance of a boost to the prediction value.

Discussion

As seen in the results in cross-validation, BFE has the best performance in terms of RMSE and MAE. BFE was the best performing of all models in terms of predicting and not simply data replicating. FFS and ALL had high OOB performances in terms of spatial spread of temporal mean. Yet the relatively large range of spatial spread of FFS (figure 11), the low OOB RMSE scores for ALL (table 6), and high cross-validation errors for both (table 7) showed that FFS and ALL might be overfitted to the training measurement points. For ALL, that retrospectively justified the use of feature selection algorithms. It is also important to point out that a high amount of often inter-related variables do not necessarily improve the model. Instead, it can lead to overfitting as in ALL. However, the algorithms are not fool proved as seen in FFS. Therefore, interpretation should be done with extra caution. Cross-validation also helps filter these overfitting cases.

The most intuitive conclusion could be drawn in terms of the conformity of the prediction of the model on the topographic height of the grids. Figure 16 shows that upscaling is highly dependent on the topography in the area, where a general trend of lower height being a stronger carbon sink can be observed. However, the temporal

means showed that across the period of 2 years, the strongest carbon sinks were projected to be around 2100m. This could be a result of u10, which is slightly stronger in negative in around the same grids.

Drivers

The upscaling results of all models are at least somewhat conforming to the topographic features in the target region. Together with t2m being the feature in all models with the highest importance, it can be concluded that (air) temperature is a significant driver of NEE. Another important feature was u10, which is the indicator of the ITCZ position as mentioned in the beginning. In summer, ITCZ covers the area which allows the convergence bringing moisture and promoting convection (Garreaud, 2009). The area has a prevailing easterly most of the year. However, the magnitude of the wind velocity, shown here as vectorized u10, also has a seasonality with -1.5 ms^{-1} in Winter (JJA) and -0.753 ms^{-1} in Summer (DJF). Therefore, it makes sense to have u10 as one of the high-importance potential drivers. The high importance of surface pressure (sp) in all models could also be explained by the relative position of ITCZ across the year, and the topography spatially.

A prominent secondary driver is the net incoming solar radiation (ssr), which can be explained by the process involved. Ssr as the incoming shortwave radiation contains the Photosynthetic Active Radiation, which is driving the photosynthesis rate. However, faPAR was not listed top in the feature importance of the only model that selected it (ALL; figure 15b). Another possibility could be that ssr is the energy source of increasing canopy temperatures, and thus helped predict NEE through indirect effect via air temperature.
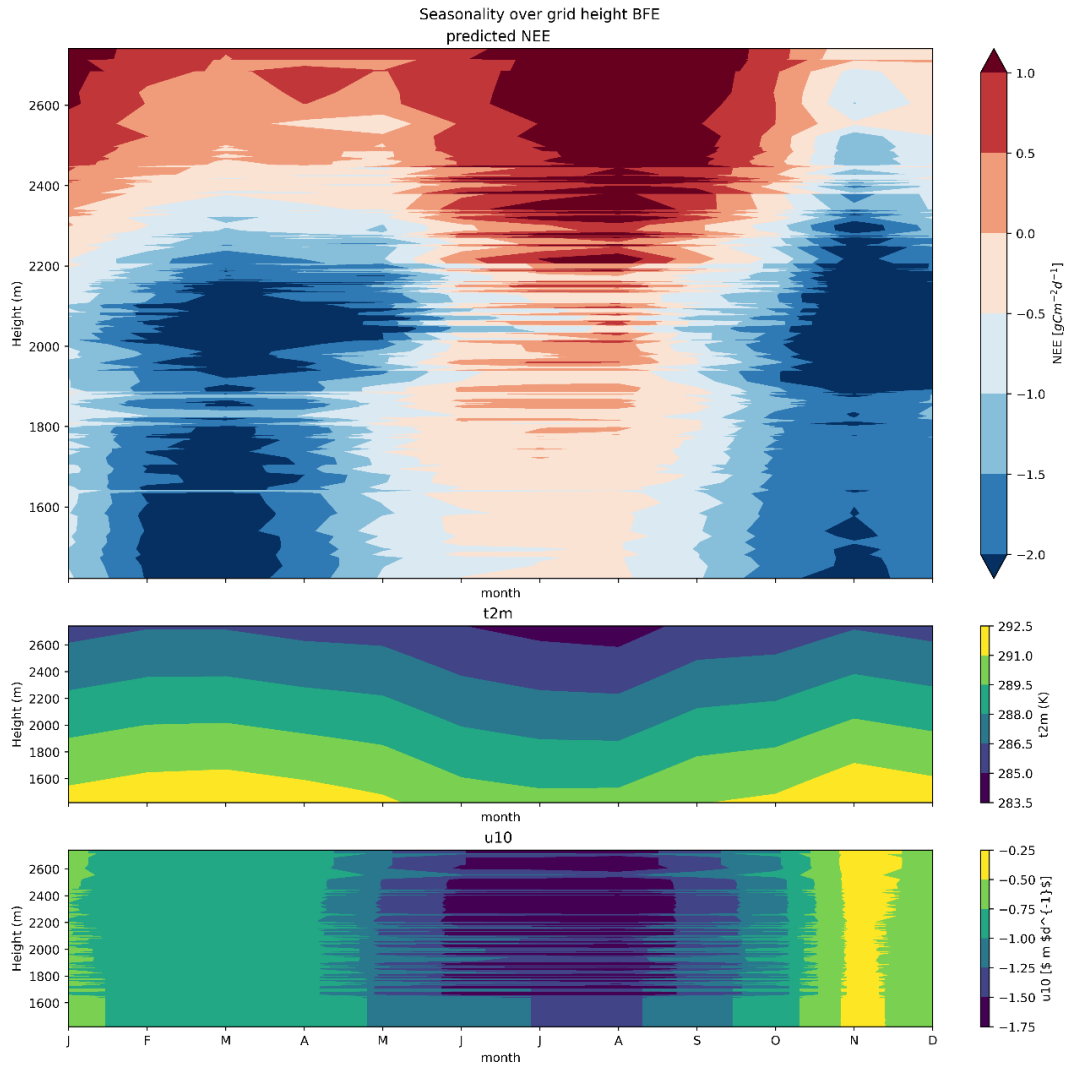
Figure 22. Heat map of monthly mean daily NEE predicted by BFE (top), 2-meter temperature (t2m, middle), and 10-meter zonal wind vector (u10, bottom) plotted against the grid height and month.

Figure 22 shows the predicted daily NEE followed loosely by the seasonality of t2m and u10, with a general trend of positive signal in the higher ground. Temporally, the strongest negative signal is seen around March and November, with the months (JJA) with the lowest t2m and highest easterly wind speed having the strongest positive signal extending to 1800m. The 0-line separating positive and negative signals also shifts in a year. Negative NEE extends from 2400m to 2600m until May, where it shrinks to below 1800m until August, when it starts raising to >2700m in October. Strips can also be observed at around 1800m, which could be from the impacts of the minor important variables or even noise from the EC measurements.

Figure 24. The accumulated Local Effect (ALE) Plots of COR with ALE against grid height (m a.s.l.) and LC (2= Evergreen Broadleaf Forests, 8=Woody Savannas according to IGBP). Distributions of variables are shown as rug on the x-axis. Produced with python package Alibi Explain (Klaise et al., 2021).

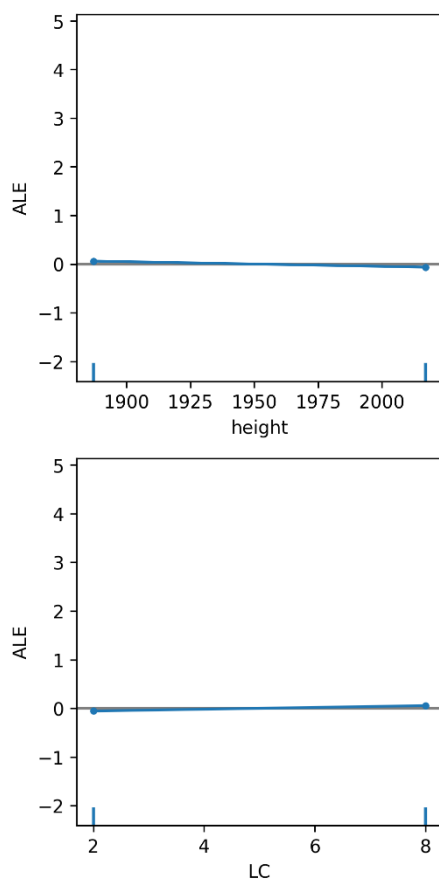BFE is the only model without selecting height as one of the features. The probable reason could be seen in the ALE plots of COR. Figure 23 shows that height was not capable of generating a good prediction, both due to the flat and near-zero ALE, and the fact that only 2 grids have a height in the training data set, which corresponds to the two EC towers. The ALE plot of height showed that with only 2 points, RF cannot generate any patterns other than a linear relationship.

Similarly, LC was not selected by BFE, possibly with the same rationale. It could be the case for all variables either with yearly temporal resolutions or static, spatial-only variables, which are likely resulting in involuntary linearity and thus further reducing its correlation with training NEE. This could be the cause for BFE to exclude all these yearly interval variables.

Overall, all the important variables showed a piecewise relation to NEE (t2m, u10, ssr, DayofYear), while the minor variables did not have clear if not no patterns at all. This means that there are thresholds of multiple conditions that change the resulting NEE values, which makes it almost impossible to establish a simple equation. However, ALE plots provide a general idea of how the equation could look according to the model.

SHAP values showed a major difference of u10, in which the medium value has the highest negative impact on the prediction value, mainly since the highest value of u10 was positive, indicating a westerly wind rather than the speed of easterlies. Surface pressure showed a rather clear cut of positive impact for high pressure, which defiles the explanation of NEE increasing with increasing topographic height. This justifies that the position of ITCZ has a major influence on the prediction through sp. Therefore, the topography-conforming pattern of upscaled NEE is likely temperature-guided as no other height-conforming variables were presented in the BFE variables set. The negative signal in winter (figure 18), which is predicted by all models, could be a result of the relatively warmer temperature in the

lower valley, which is likely to stay above ~288K, the inflection points as seen in the ALE plots (figure 20).

Limitations

Research related to establishing data-driven model for ecosystems carbon fluxes, e.g., Gross Primary Productivity (GPP) and NEE, have exploded in number in recent years. However, most of the models were designed with a large network of EC towers as input NEE measurement in mind. This is due to the relatively high interest in the more populated area, which leads to a higher density of available measuring towers. In the more complicated, remoted area like in this study, EC measurements are scattered and exiguous. In this study, only data from two EC towers were used, which is shown to be performing worse than the main referring study by Reitz et al. (2021). For example, in a similar diagnostic plot as figure 20a, the slope and y-intercept for the measurement against prediction were reported as 1.06 and 0.15, while in this study it is 1.361 and 0.5377 for the best-performing model. The huge performance difference could be attributed to the large amount of EC measurements (16 towers) in Reitz et al., (2021) in comparison to 2 in this study. The small sample in this study amplifies the impact of noise in the training data on the prediction, which is either from the EC measurements, the coarse resolution of the global reanalysis model, and the intrinsic low signal-to-noise ratio of satellite remote sensing products, or more likely to be a combination of these. Also, a long-term disturbance, such as a month-long local draught or even ENSO patterns, can significantly influence the prediction. Especially when the temporal range is limited to 2 years, it can be expected that extreme patterns could have a high impact on the model establishment. Therefore, it can be expected that the bias observed would decrease further if more years of input were added.

Another main limitation of this algorithm is the limited spatial representation of the two close-by EC towers with similar altitudes. This significantly limited the accuracy of prediction outside of the altitude range as seen in figure 16. This could be expected in terms of Land Use Classes (LC) as well since only 2 of the classes were represented. This was not seen simply as a result of the relatively small target area with a rather homogeneous LC distribution. This 'forced linearity' as observed in figure 22 would be a hint that more stations are required to truly establish a meaningful model.

Overfitting is another common problem with data-driven model training, in which the model is trained to replicate the training data. FFS had good OOB scores (table 6) but a rather bad RMSE and MAE score in cross-validation. More importantly, the minor variables in terms of feature importance (figure 15c) levitated comparing to e.g., BFE.

These are indications that FFS has possibly overfit. This highlights the importance of thoroughly verifying each model by cross-validation, feature importance analysis, and even interpretable machine learning.

Conclusion

This study has somewhat successfully developed an algorithm to upscale point-source Eddy Covariance measured NEE into 30 arc seconds (approximately 1km) gridded NEE. It showed that it is possible to upscale with a limited input. Nonetheless, with the limit, the algorithm successfully established a somewhat explainable upscaled NEE. Generally, the area was mostly carbon sink with negative daily NEE, with the peaks above the altitude of around 2400m a.s.l. being a net carbon source during the period. The area was also projected to have a strong dependency on seasonality and topography. On the higher grounds, it is predicted to be a strong carbon source in the austral winter (JJA) and a weak carbon source in all other seasons. In the valley, strong carbon sinks were predicted in all seasons except winter. In winter, a weaker carbon sink was expected in the lower valley and a neutral state in the higher valley to the northwestern valley.

The findings can be concluded that it is feasible for the data-driven algorithm to upscale point-source NEE into spatial-temporal prediction. However, throughout the study, it became apparent that even the best-performing model (BFE in terms of cross-validation) was also rather flawed. It is speculated to be mainly the result of a lack of input representing temporal, spatial, and variability of Land Use Classes. However, it is also clear that the general algorithm would work even in this complex topography, given by RMSE and MAE being in the same range as Reitz et al. (2021). That hints that improvements are needed to achieve better prediction quality. This study has attempted to tackle the intrinsic problem of lack of data, both temporally and spatially, with an interpolation of the coarsest input (ERA5-land variables) and feature selection processes. It is concluded that among the tested feature selection algorithms BFE generated a variable set that was capable of producing the best model result, which aligns with the findings from Reitz et al. (2021). Nonetheless, it was observed that it could bring extra problems when only 2 points were brought as input, including the forced linearity of the yearly variables.

Multiple interpretable machine-learning approaches were employed to see through the often-assumed black box of RF, which provided insight into a generalized and simplified form of the model equation. ALE showed that inflection points exist in almost all major

variables selected by BFE, which could be speculated as the reason for the hardship in establishing an accurate process-based model for NEE. SHAP values showed sp as a strong driver, indicating a strong influence from the position of ITCZ, hence dominating the temporal change of NEE in the region. In contrast, t2m was likely to be the main driver of the spatial variation of the upscaled NEE.

Overall, the algorithm can certainly be improved by adding more training data, either temporarily or spatially. This highlights the main intrinsic challenge of data-driven models like Machine Learning, which is their dependency on the quantity, quality, and representativity of the training inputs. However, improvement of these aspects may not be possible in a more complex terrain, especially in the less populated regions of the world. Nonetheless, this study reached a reasonable upscaling result despite the said challenges. Therefore, this study can be the justification for the feasibility of NEE upscaling and modelling in such areas with limited accessible measurement.

References

Abdalla, M., Saunders, M., Hastings, A., Williams, M., Smith, P., Osborne, B., Lanigan, G., & Jones, M. B. (2013). Simulating the impacts of land use in northwest europe on net ecosystem exchange (NEE): The role of arable ecosystems, grasslands and forest plantations in climate change mitigation. *Science of the Total Environment, 465*, 325-336. https://doi.org/10.1016/j.scitotenv.2012.12.030

Beck, H. E., McVicar, T. R., Vergopolan, N., Berg, A., Lutsko, N. J., Dufour, A., Zeng, Z., Jiang, X., van Dijk, Albert I. J. M., & Miralles, D. G. (2023). High-resolution (1 km) köppen-geiger maps for 1901–2099 based on constrained CMIP6 projections. *Scientific Data, 10*(1), 1-16. https://doi.org/10.1038/s41597-023-02549-6

Breiman, L. (2001). Random forests |. *Machine Learning, 45*, 5-32. https://doi.org/https://doi.org/10.1023/A:1010933404324

Copernicus Climate Change Service. (2019). *ERA5-land hourly data from 2001 to present*. ECMWF. https://doi.org/10.24381/cds.e2161bac

Didan, K. (2021). *MODIS/terra vegetation indices 16-day L3 global 500m SIN grid V061*. NASA EOSDIS Land Processes Distributed Active Archive Center. https://doi.org/10.5067/modis/mod13a1.061

DiMiceli, C., Sohlberg, R., & Townshend, J. (2022). *MODIS/terra vegetation continuous fields yearly L3 global 250m SIN grid V061*. NASA EOSDIS Land Processes Distributed Active Archive Center. https://doi.org/10.5067/modis/mod44b.061

Earth Resources Observation And Science (EROS) Center. (2017). *Global 30 arc-second elevation (GTOPO30)*. U.S. Geological Survey. https://doi.org/10.5066/f7df6pqs

Friedl, M., & Sulla-Menashe, D. (2022). *MODIS/Terra+Aqua land cover type yearly L3 global 500m SIN grid V061*. NASA EOSDIS Land Processes Distributed Active Archive Center. https://doi.org/10.5067/modis/mcd12q1.061

Garreaud, R. D. (2009). The andes climate and weather. *Advances in Geosciences, 22*(22), 3-11. https://doi.org/10.5194/adgeo-22-3-2009

Johnson, D. L., Roberts, B. C., Vaughan, W. W., & Parker, N. C. (2002). *Reference and standard atmosphere models.* Unpublished manuscript. Retrieved Mar 14, 2024, https://ntrs.nasa.gov/citations/20020092087

Klaise, J., Van Looveren, A., Vacanti, G., & Coca, A. (2021). Alibi explain: Algorithms for explaining machine learning models [Abstract]. *Journal of Machine Learning Research, 22*(181) 1-7. http://jmlr.org/papers/v22/21-0017.html

Koo, W. H. (2023). *Assessment of different statistical and machine learning model approaches of net ecosystem exchange (NEE) prediction using remote sensing and reanalysis data.* Unpublished manuscript.

Leo Breiman. (2004). Consistency for a simple model of random forests . *Statistics Department University of California at Berkeley, Technical Report 670*

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*(1), 56-67. https://doi.org/10.1038/s42256-019-0138-9

Lundberg, S., & Lee, S. (2017). *A unified approach to interpreting model predictions.* (). Ithaca: Cornell University Library, arXiv.org. https://doi.org/10.48550/arXiv.1705.07874 Retrieved from Publicly Available Content Database https://search.proquest.com/docview/2077007862

Molnar, C. (2023). *Interpretable machine learning (second edition) : A guide for making black box models explainable* (2nd ed.). Christoph Molnar.

Myneni, R., Knyazikhin, Y., & Park, T. (2015). *MOD15A2H MODIS/terra leaf area index/FPAR 8-day L4 global 500m SIN grid V006 [data set]*. NASA EOSDIS Land Processes DAAC. https://doi.org/10.5067/modis/mod15a2h.006

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2012). *Scikit-learn: Machine learning in python*

Probst, P., & Boulesteix, A. (2017). *To tune or not to tune the number of trees in random forest.* (). Ithaca: Cornell University Library, arXiv.org. https://doi.org/10.48550/arXiv.1705.05654 Retrieved from Publicly Available Content Database https://search.proquest.com/docview/2074241611

Pudil, P., Novovičová, J., & Kittler, J. (1994). Floating search methods in feature selection. *Pattern Recognition Letters, 15*(11), 1119-1125. https://doi.org/10.1016/0167-8655(94)90127-9

Raschka, S. (2018). MLxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack. *Journal of Open Source Software, 3*(24), 638. https://doi.org/10.21105/joss.00638

Reitz, O., Graf, A., Schmidt, M., Ketzler, G., & Leuchner, M. (2021). Upscaling net ecosystem exchange over heterogeneous landscapes with machine learning. *Journal of Geophysical Research. Biogeosciences, 126*(2), n/a. https://doi.org/10.1029/2020JG005814

Wang, A., Xu, L., Li, Y., Xing, J., Chen, X., Liu, K., Liang, Y., & Zhou, Z. (2021). Random-forest based adjusting method for wind forecast of WRF model. *Computers & Geosciences, 155*, 104842. https://doi.org/10.1016/j.cageo.2021.104842

Zhuravlev, R., Dara, A., Santos, André Luís Diniz dos, Demidov, O., & Burba, G. (2022). Globally scalable approach to estimate net ecosystem exchange based on remote sensing, meteorological data, and direct measurements of eddy covariance sites. *Remote Sensing (Basel, Switzerland), 14*(21), 5529. https://doi.org/10.3390/rs14215529
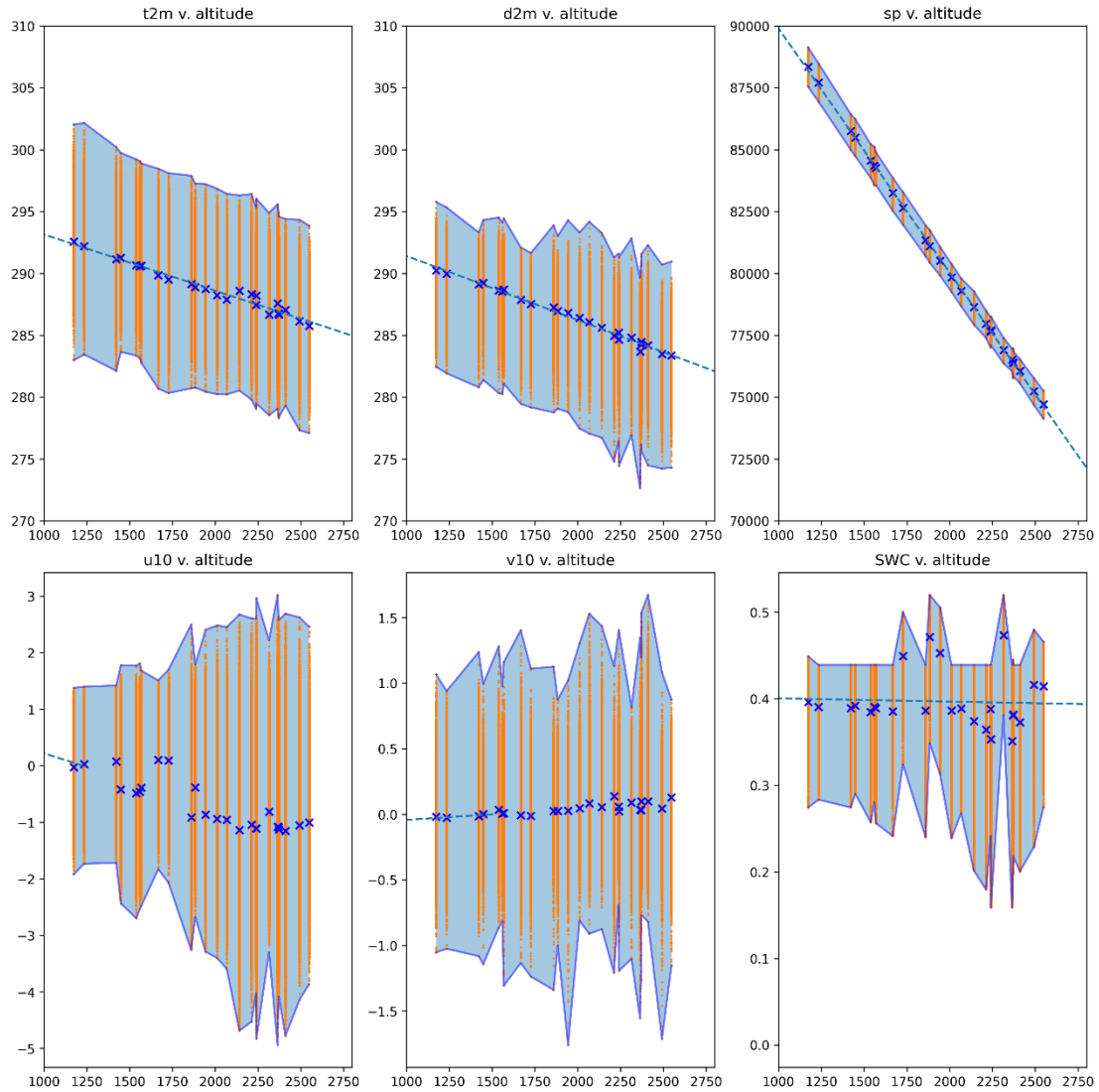
Appendix

Figure A1. Correlation of the alleged altitude-related ERA5-land variables (t2m, d2m, sp, u10, v10, and SWC) plotted against the calculated ERA5-land altitude. Blue crosses indicate the temporal mean values of each grid cell, while orange crosses indicate the individual distribution of each time step. Shaded areas show the distribution of all distribution concerning calculated grid cell altitudes.