Groundwater–CO$_2$ Emissions Relationship in Dutch Peatlands Derived by Machine Learning Using Airborne and Ground-Based Eddy Covariance Data

Laura M. van der Poel[1], Laurent V. Bataille[1], Bart Kruijt[1], Wietse Franssen[1], Wilma Jans[1], Jan Biermann[1], Anne Rietman[1], Alex J. V. Buzacott[2], Ype van der Velde[3], Ruben Boelens[4], and Ronald W. A. Hutjes[1]

[1]Earth Systems & Global Change Group, Wageningen University, 6708 PB Wageningen, Netherlands
[2]Institute for Atmospheric and Earth System Research, University of Helsinki, 00014 Helsinki, Finland
[3]Faculty of Science, Earth and Climate, Vrije Universiteit Amsterdam, 1081 HV Amsterdam, Netherlands
[4]HydroLogic, 3811 HN Amersfoort, Netherlands

Correspondence: Laura M. van der Poel (lvdp@ign.ku.dk)

Abstract. Peatlands worldwide have been transformed from carbon sinks to carbon sources due to years of intensive agriculture requiring low water tables. In the Netherlands, carbon dioxide (CO$_2$) emissions from drained peatlands mount up to 5.6 Mton annually and, according the Dutch climate agreement, should be reduced by 1 Mton in 2030. It is generally accepted that mitigation measures should include raising the water level, and the exact influence of water table depth has been increasingly
5 studied in recent years. Most studies do this by comparing annual Eddy Covariance (EC) site-specific CO$_2$ budgets to mean annual effective water table depths (WTDe). However, here we apply a different approach: we integrate measurements from 16 EC towers with EC measurements from 141 flights by a low-flying research aircraft, in an interpretable machine learning framework. We make use of the different strengths of tower and airborne data, temporal continuity and spatial heterogeneity, respectively. We apply time frequency wavelet analysis and a footprint model to relate the measured fluxes to the underlying
10 surface. Using spatio-temporal data, we train and optimize a boosted regression tree (BRT) machine learning algorithm and use Shapley values and various simulations to interpret the model's outputs. We find that emissions increase with 4.6 tonnes CO$_2$ ha^{-1} yr^{-1} (90% CI: 4.0-5.4) for every 10 cm WTDe up to a WTDe of 0.8 meter. For more drained conditions, emissions decrease again, following an optimum-based curve. Furthermore, we find that this effect is stronger in winter than in summer and that it varies between sites. This study shows the added value of using ML with different types of instantaneous data, and
15 holds potential for future applications.

1 Introduction

Despite covering only 3% of the earth's land surface, peatlands store around 25% of all terrestrial carbon and play a crucial role in the global carbon cycle (Yu et al., 2010). They are the most carbon-dense ecosystems of the terrestrial biosphere and have a true potential for climate change mitigation (Loisel et al., 2021; Leifeld and Menichetti, 2018). In natural, waterlogged fens
20 and bogs, uptake of carbon dioxide (CO$_2$) through vegetation and subsequent sequestration in peat soils, abundantly exceeds

the emission of methane (CH_4) (Frolking and Roulet, 2007). However, peat soils have been exploited and drained worldwide for fuel extraction and agricultural practices. They widely transformed from carbon sinks to carbon sources due to increased peat decomposition following higher oxygen availability, and are currently responsible of large CO_2 emissions.

The Netherlands has a long history of peat extraction and intensively draining peatlands for agriculture and livestock farming (Erkens et al., 2016; van den Akker et al., 2008). This has led to increased carbon dioxide emissions, currently accounting for ∼3% of all Dutch emissions (5.6 Mton annually), and land subsidence, which in turn increases the need for further drainage (Ruyssenaars et al., 2022; Kwakernaak et al., 2010). To counter this spiral, the Dutch government set a specific mitigation target for peat meadows: annual emissions must be reduced by 1 Mton by 2030 (Klimaatakkoord, 2019). It is generally accepted that counter measures to reduce such emissions should include raising the water table, as water table seems the predominant control on greenhouse gas emissions from managed peatlands (Evans et al., 2021). However, the exact impact of higher groundwater levels on CO_2 fluxes is not yet entirely established, and has been increasingly studied in the past years (Tiemeyer et al., 2020; Boonman et al., 2022; Fritz et al., 2017; Kruijt et al., 2023; Evans et al., 2021; Aben et al., 2024).

Studies investigating agricultural systems generally require correcting the annual net ecosystem exchange (NEE) for lateral movement of carbon associated with manure applications and harvests, and relate the resulting net ecosystem carbon balance (NECB) to mean annual groundwater level (Aben et al., 2024; Evans et al., 2021; Kruijt et al., 2023; Boonman et al., 2022). In unmanaged systems, annual NEE is equivalent to NECB. Averaging out daily and seasonal variation, the goal is to isolate the underlying effect of groundwater level. This way, multiple studies found linear relationships with similar slopes: between 3 and 4.5 t CO_2 ha^{-1} yr^{-1} extra emissions per 10 cm increase in WTDe (Jurasinski et al., 2016; Evans et al., 2021; Boonman et al., 2022; Fritz et al., 2017; Kruijt et al., 2023). Tiemeyer et al. (2020) fitted the Gompertz function to a set of annual balances, which shows a sharper increase in emissions at shallow water levels, but then saturates at around 0.4 m. Recently, multiple studies have applied this function (Koch et al., 2023; Friedrich et al., 2024; Nijman et al., 2024).

While studies on annual budgets provide valuable insights into the underlying groundwater-CO_2 relationship and differences between sites, some limitations emerge. First, to obtain an NECB estimate for any location, year-round data at that specific location is required, which is not always achievable and generally requires some trustworthy gap-filling. Second, whilst carbon import and export can add up to significant amounts, these numbers are often hard to obtain and generally unavailable at landscape level. Third, the site comparisons are robust only when comparing sites with markedly different average water table depths. The datasets used have been based on site specific observations, with well-defined, fairly homogeneous soil and vegetation characteristics and well-known water table management. However, these factors generally vary widely on the regional scale. Last, the annual estimates discard possibly important information from intra-annual variability and relationships with other factors than groundwater. In the current study, we aim to by-pass these limitations by alternatively exploring the short time scale at the regional level, to further unravel the influence of water table depth and other key drivers on CO_2 emissions from agricultural peatlands in the Netherlands.

We do this by incorporating flux measurements from a low-flying aircraft. Airborne measurements bear high spatial heterogeneity, since every measured flux originates from an area called the footprint spanning several kilometers, and an entire region can be covered by an appropriate flight pattern. However, a limitation of the airborne measurements is that they are

generally limited to daytime conditions. This limitation is particularly critical for CO_2 studies, given the different contributions of Gross Primary Productivity (GPP) and Ecosystem Respiration (Reco) to NEE. Here, we are specifically interested in the peat decomposition component of heterotrophic respiration, unrelated to GPP. Measurements by Eddy Covariance towers, on the other hand, are continuous and include nighttime fluxes, enabling NEE partitioning, but are limited by their fixed location.

60 Therefore, we consider complementary use of tower and airborne flux estimates essential to assess CO_2 fluxes at a regional scale.

Airborne measurements (integrated flux signals from their respective footprints) can be related to the underlying surface by environmental response functions using either more classical statistical methods (Hutjes et al., 2010), or by artificial intelligence approaches (Metzger et al., 2013; Serafimovich et al., 2018). Both depend on overlaying the footprint of all flux measurements

65 over maps of vegetation, land use and soils and/or direct satellite derived products. Here, we integrate tower and airborne data using the 'LTFM' approach initially developed by Metzger et al. (2013), which includes four principal steps: Low level flights, Time frequency wavelet analysis, Footprint modeling, and Machine learning. We aim to apply this ML approach to identify key predictors of NEE and understand their influence on CO_2 dynamics in Dutch peatlands.

Today, peat soils in the Netherlands can be found in the 'Groene Hart' and in the northwest, in the provinces of Friesland and

70 Overijssel. These areas share similarities, such as being mostly dominated by pastures with ditches, but also differ in certain aspects, such as average drainage depth, Friesland being the most intensively drained with lowest water tables. In the current study, we have three distinct flight tracks: above the Groene Hart area, southern Friesland, and the western part of Overijssel. These three regions are also equipped with flux towers, thus we measure both airborne and tower CO_2 fluxes from all three.

In our aim to understand the influence of key predictors, we inspect potential differences and similarities between these

75 areas. We do this by building several models: one model based on all data together, and models per area separately, trained only on the data of the respective area. We expect the model with all data will perform best, since in machine learning data quantity is often a determining factor for performance. Furthermore, we expect that although characteristics of the areas vary, the underlying processes do not, hence 'one model fits all', and area-specific models can be used to predict for other areas. To achieve physical interpretability of our ML-approach, we use the SHAP framework and model simulations, fully exploring the

80 identified relationships. We hypothesize that, alongside drivers of the diurnal cycle, water table depth plays an important role at the regional scale in all areas, and that the machine learning approach can help us to better understand drivers of CO_2 fluxes. We aim to finally create a robust, interpretable ML model that can be used in the Netherlands to predict CO_2 emissions from drained fen meadows at a regional scale.

2 Methods

85 In this study, we merge airborne flux measurements with ground-based Eddy Covariance (EC) measurements, to make use of their spatial and temporal strengths, respectively. Both are part of the intensive monitoring network implemented by the Netherlands Research Programme on Greenhouse Gas Dynamics in Peatlands and Organic Soils (NOBV, https://www.nobveenweiden.nl/en/), that was established following the Dutch mitigation target of 2019. The goal of the NOBV is to further understand greenhouse

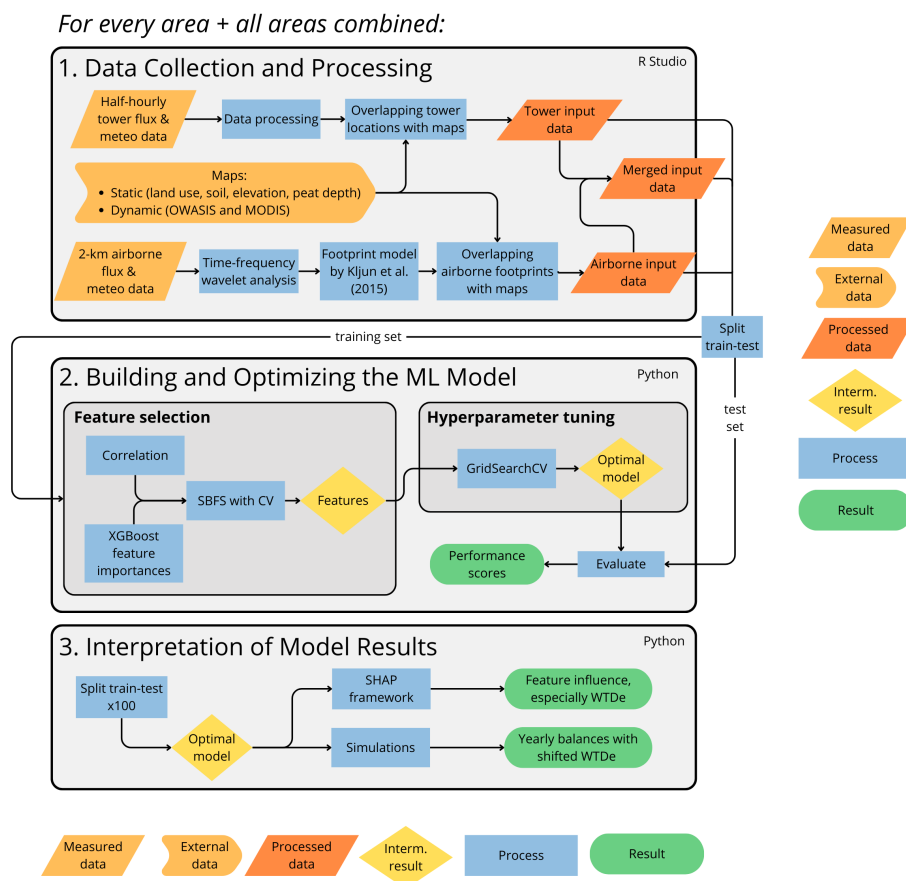*For every area + all areas combined:*

Figure 1. Methodological steps for the current study. We divide the methods in three main parts: (1) Data Collection and Processing; (2) Building and Optimizing the ML Model; (3) Interpretation of Model Results. We carried out all steps for the three study areas separately, as well as for all areas combined.

gas emissions from drained fen meadows, their drivers, and the efficiency of proposed mitigation measures. Here, we test a machine learning approach on the combination of airborne and tower data to assess the most important drivers of CO_2 emissions on a regional level, and to quantify the influence of water table depth. A preliminary analysis we conducted based on a subset of the data with this method showed promising results for combining airborne and tower data, corresponding to existing estimates.

In Fig. 1, a visual overview of the methodology is shown, which can be roughly split up into three parts: data collection and data processing; building and optimizing several machine learning models; and interpreting the results with Shapley values and simulations. In this section, we describe these three parts of our approach.
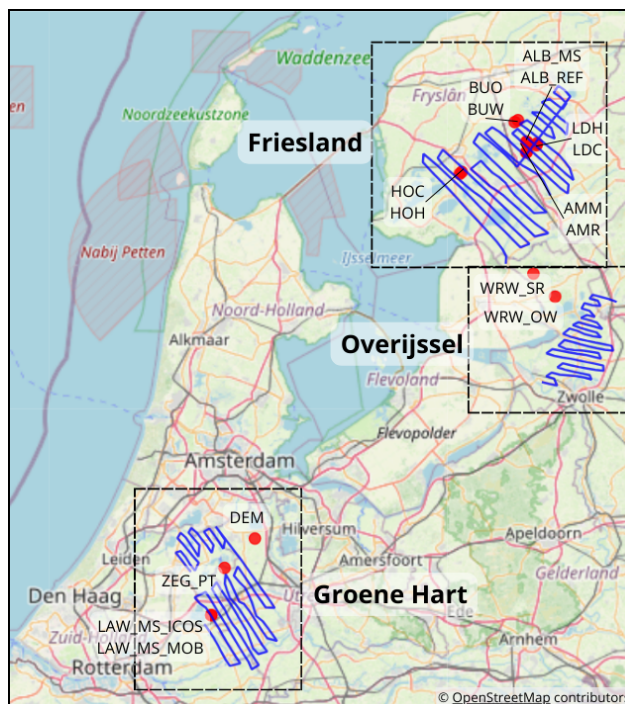
Figure 2. The EC tower network used in this study, with three flight tracks over the study areas.

2.1 Data Collection and Processing

2.1.1 Study area

The study area comprises the three main peat soil areas in the Netherlands: the Groene Hart in the south west of the country,
100   southwest Friesland, and the 'Kop van Overijssel' south of that (see Fig. 2). These peat areas have entirely been formed during
the Holocene, reaching their maximum extent (about 50% of present Netherlands) around 4000 years ago. Between 2000 and
1000 years ago, large tracts eroded away by a rising and repeatedly intruding sea. Since medieval times, peat has been extracted
by humans, and the land has been drained for agricultural purposes. Peat mining continued at a large scale until the late 19th
century, while drainage continues to this day (Erkens et al., 2016; van Asselen et al., 2020). Most remaining peatlands are fens,
105   and the very few that can be characterised as bogs are not subject of this study. The fen meadows are primarily used as pastures
for dairy farming and currently cover around 7% (ca 290.000 ha) of the Dutch land surface (Arets et al., 2021). Water table
depth in the study area ranges from surface level to 150 cm below surface level, with most deeply drained soils in Friesland.
The climate is temperate and humid, and the Dutch Meteorological Institution states mean annual temperatures between 9.5
and 11.5°C and annual precipitation between 670 and 1100 mm (KNMI, 2024).

2.1.2 Airborne Flux Measurements

The aircraft used is a SkyArrow 650 TCNS, a light weight environmental research aircraft with a push propeller. Weather permitting, i.e. with good visibility and no rain, surveys were done twice a week between March 2020 and December 2023, alternating between the three areas described above. However, between July 2020 and February 2021, and between November 2022 and June 2023, the aircraft did not fly due to technical issues. In total, we used data from 141 flights. Parallel flight-tracks of 2-3 km apart were designed perpendicular to the prevailing wind direction and landscape gradients, to get complete spatial coverage of the area of interest. Figure 3 shows a typical flight trajectory. Mean flying altitude was 60 meters, so built-up areas had to be avoided. The flight transects covered all major soil and land use classes, although the footprints were mostly dominated by pastures on peat soils - as are the areas.

The aircraft was equipped with an open-path gas analyzer for CO_2 and latent heat fluxes, a thermocouple for sensible heat, both depending on a BAT-probe for 3D wind speed and momentum flux. In addition, net radiation, incoming and reflected photosynthetic active radiation (PAR), air and surface temperature were measured. Most instruments sampled at either 50 or 20Hz. Post-flight processing started with de-spiking raw data. Next, 50 Hz air pressure and temperature measured by the BAT probe were converted to 3D wind fields and corrected for all aircraft motions. Then, covariances between wind and CO_2 concentrations were calculated. Conventionally, covariances are calculated over time and then spatially integrated over a fixed window, for example of 2 km long. However, varying conditions can require different window lengths (Sun et al., 2018). Furthermore, this block averaging method potentially suffers from spectral losses and reduced statistical precision in lower frequencies, as is the case with tower-based measurements using typical averaging times (Paleri et al., 2022). As in Metzger et al. (2013), we used wavelet cross-scalograms calculated over the entire flight to find covariances in the frequency domain: smaller-scale, local turbulent fluxes at high frequencies, and larger-scale mesoscale contributions at low frequencies, large wavelengths. Wavelengths larger than the boundary layer height were discarded. The fluxes at these two scales are then summed in 2 km windows to get the flux over all scales. Further processing was done by following the framework of Foken et al. (2004). In addition, the most important meteorological scalar variables were also averaged over a 2 km spatial window. For more detail on the aircraft and its equipment, see Vellinga et al. (2013).

To specify the spatial origin of the airborne measurements, the flux footprint model by Kljun et al. (2015) was used. This two-dimensional source weight function is a parameterization of the backward Lagrangian model and describes the spatial extent, position, and distribution of the contributing surface area. The function can be applied on wide-ranging boundary layer conditions, and has been widely used by studies dealing with airborne flux observations (Hannun et al., 2020; Sun et al., 2023; Vaughan et al., 2021; Serafimovich et al., 2018). The input parameters include Obukhov length (L), standard deviation of lateral wind velocity (σ_v), measurement height (z_m), friction velocity (u_*), which were measured directly by sensors on the aircraft, and planetary boundary layer height (h), which was extracted from the ERA5 product. Roughness length (z_0) is implicitly included in the footprint model by the fraction $u(z_m)/u_*$ (Kljun et al., 2015).

For every 2 km window, five overlapping sub-footprints were calculated and overlaid with various maps, described below. A typical flight with all sub-footprints is shown in Fig. 3. For maps with continuous values, weighted averages were computed,
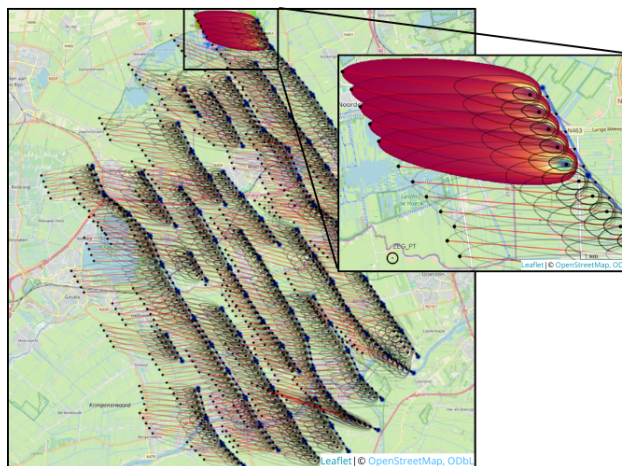
Figure 3. All sub-footprints of a typical flight over the Groene Hart. Only the first ones are colored by contribution. Zoomed in, the difference in airborne and tower (ZEG_PT) footprint sizes is visible.

whereas for categorical maps, the fraction of each category in the footprint was calculated. To obtain the final contributions to the flux measurements, the sub-footprints were combined and normalized. Footprints with more than 15% of built up area were excluded (discarding 17% of airborne data), as well as footprints where the dominant land use class is anything else than 'grasslands', 'fens and bogs', or 'summer crops' (discarding 3%). Finally, we had around 10,400 airborne data entries.

2.1.3 Eddy Covariance towers

The NOBV implemented an intensive monitoring network of EC towers (see Fig. 2). They are distributed over the three main fen meadow areas mentioned previously and cover a representative range of soil types and water levels. For site descriptions and specifics, as well as the processing of the raw tower data, see the report by Kruijt et al. (2023); below, we use the same site abbreviations.

We used four measurement sites in the Groene Hart (LAW_MS_MOB, LAW_MS_ICOS, DEM, ZEG_PT), ten in Friesland (ALB_RF, ALB_MS, AMM, AMR, BUW, BUO, HOC, HOH, LDC, LDH), and two in Overijssel (WRW_SR, WRW_OW). Although most of these sites are on agriculturally used land, the two sites in Overijssel are in a relatively wet nature area. We used data covering the period from 16-05-2020 to 31-10-2022, although it should be noted that sites vary in data availability. No attempts at gap filling were made.

At these towers, fluxes of CO_2 (and CH_4, evaporation and sensible heat, but not considered here) are measured with the EC method, alongside weather station measurements of PAR, four component radiation (shortwave and longwave, incoming and reflected), air temperature, humidity, rainfall, soil moisture and soil temperature. No attempts at gap filling were made; outliers defined as 0.5% highest and lowest values for CO_2 flux were removed, resulting in 66,400 half hourly data records in total. In Fig. 4, a part of the measurement period in Friesland is shown, comparing monthly airborne data to monthly daytime tower
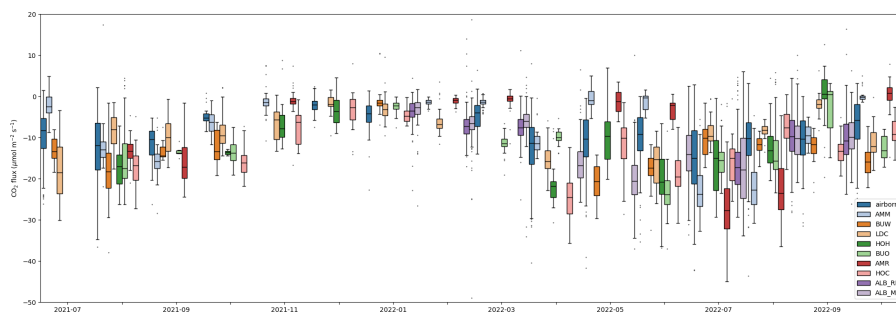
Figure 4. Airborne and daytime tower measurements between 10:00 and 16:00 of CO_2 fluxes in Friesland, binned per month.

Table 1. Number of datapoints per datatype and area, rounded to the nearest hundred. The 'merged' datasets consists of the airborne and tower datasets.

Datatype	Groene Hart	Friesland	Overijssel	All
Tower	21 100	30 300	15 000	66 400
Airborne	4 100	3 700	2 600	10 400
Merged	25 200	34 000	17 600	86 800

data across multiple sites. Although the airborne and tower data can never be compared directly due to intrinsic differences in footprint sizes, the seasonal trend and magnitude of the NEE fluxes is similar.

165   Similarly to the airborne data, the tower data was overlaid with various maps, but without using the footprint model by Kljun et al. (2015). We expected the differences between individual 30 min tower footprints to be negligible compared to the airborne footprints, so we set a fixed circular footprint for every tower (see Fig. 3). The radius was based on the average 80% of the footprint distance, which was given in the dataset. This footprint was used to extract corresponding spatial data with high spatial resolution (<200 m). For maps with spatial resolution as low as 250 m, the average value within a radius of 500 m
170   from the measurement point was taken (including >12 grid cells), because the value from a single grid cell could erroneously deviate from surrounding grid cells. Table 1 shows the final amount of data.

2.1.4 Supplementary Surface Data

Additional surface data, potentially explaining CO_2 flux variations, was gathered from various sources, including both static and dynamic maps (see Table 2). Static maps generally had a high spatial resolution (5-25m) and included land use, soil, peat
175   depth, and elevation. Water-related information was retrieved from the operational product OWASIS by Hydrologic (2019), which consisted of three daily maps: ground water level with respect to sea level (m BSL), soil moisture in the root zone (mm), and air-filled pore space in the unsaturated zone (mm). This product is at national scale with a spatial resolution of 250 m, and is based on the national hydrological model (LHM, Ligtenberg et al. (2021)). It is made operational by including evaporation

Table 2. Overview of supplementary data sources.

Variable	Spatial resolution	Temporal resolution	Source
Land use*	5 m	-	Landelijk Grondgebruik Nederland, LGN2020
Soil*	5 m	-	Bodem Data, BOFEK2020
Elevation	25 m	-	Algemeen Hoogtebestand Nederland, AHN3
Peat depth	100 m	-	Bodem Data
Groundwater level Air-filled pore space Soil moisture	250 m	Daily	OWASIS product from Hydrologic, personal communication (retrieved January 9, 2024)
NDVI EVI	250 m	8 days	MODIS: MOD13Q1 Terra and Aqua

*Variables with an asterisk are categorical.

and precipitation data and it is the only water-information product covering the entire study area. It performs well in showing
180 trends, but the absolute pixel-based accuracy is up for discussion, thus we consider a wider area than a single pixel. Lastly, two
vegetation indices were retrieved by remote sensing from MODIS: the Normalized Difference Vegetation Index (NDVI) and
the Enhanced Vegetation Index (EVI). Satellites Aqua and Terra, each with 16-day revisit time, were combined and linearly
interpolated to obtain daily values.

The categories of the land use and soil maps were reclassified to obtain a smaller but representative number of variables (see
185 Appendix A). Using the collected information, some additional covariates were calculated, such as effective water table depth
(WTDe) based on groundwater level and elevation, the percentage of all peat classes together present in the footprint (AllPeat),
peat on sand, and peat on peat. Combining peat depth with WTDe, the peat exposed to air ('exposed peat depth') in cm was
calculated.

2.2 Building and Optimizing the Machine Learning Model

190 For every area, as well as for all areas combined, we built a machine learning model based on the combination of tower
and airborne data. We used Boosted Regression Trees (BRT), as they are increasingly used in environmental studies, and are
furthermore recommended by studies analyzing airborne flux measurements (Metzger et al., 2013; Serafimovich et al., 2018;
Vaughan et al., 2021). The package XGBoost (eXtreme Gradient Boosting) was used, due to the high predictive performance
and computing speed (Nielsen, 2016).

195 For all sets of input data, we made a train-test division. Commonly, this is done in a random manner, but in our case, because
of time-dependency of the data and the potential for data-leakage, this is not recommended. Thus, we created the test set
by randomly selecting separate flight-legs, and by selecting one every four weeks of tower data, ensuring that the final data
distribution was around 80% train and 20% test. We expected this to be a good trade-off between avoiding data-leakage and

9

keeping enough data for training. Furthermore, all features in the training set were normalized. The features in the test set were
200   scaled accordingly, based on the statistical properties of the training set.

For each model, the model was tuned in two steps: *a)* reducing the number of explanatory variables: 'feature selection', and
b) optimizing the model's settings: 'hyperparameter tuning'. Feature selection was done in a hybrid manner, as is frequently
done by studies that use XGBoost (Ogunleye and Wang, 2019; Prabha et al., 2021; Sang et al., 2020; Wang and Ni, 2019). First,
we analyzed Pearson correlations to roughly select features correlated to CO_2, and to reject inter-correlated features. Second,
205   feature importances embedded in XGBoost were computed. These first two steps served as a pre-selection of features for the
third, more extensive method: Sequential Backward Floating Selection (SBFS). SBFS includes an extra 'floating' element
compared to the more standard and widely used Sequential Backward Selection (SBS), and is known to give good results
(Chandrashekar and Sahin, 2014; Rodríguez-Pérez and Bajorath, 2020). SBFS is more time costly than SBS but simultaneously
reduces the risk of missing important feature combinations due to early dropping of a specific feature. SBFS was run with 5-
210   fold cross validation, the model used was a XGBoost tree with n_estimators = 1000, learning_rate = 0.05, max_depth = 6,
subsample = 1, and the scoring metric for the algorithm was RMSE. As there are multiple features describing water dynamics,
several options were run separately, excluding inter-correlated features in the same run. To avoid unequal representation in
different folds of the cross validation, all datasets were shuffled beforehand. To evaluate which subset of features is optimal,
the R^2, MSE, bias and variance of each model proposed by SBFS were computed, and model parsimony was taken into account.

215   After optimization of the feature subset, the following hyperparameters were tuned: number of trees (n_estimators); max-
imum depth (tree complexity, max_depth); learning rate (learning_rate); minimum sum of instance weights in a leaf node
(min_child_weight); the ratio of columns when constructing each tree (colsample_bytree); and ratio of instances in every
boosting iteration (subsample). We performed a grid search with 5-fold cross validation (GridSearchCV from Scikit Learn) on
the training set, the parameter grid is shown in Appendix B. Again, R^2 was used as scoring metric. In addition, we set a mono-
220   tone constraint on the model to find a negative relationship between PAR and CO_2 flux, such that increasing PAR leads to more
negative fluxes. Finally, the model with optimized parameters was evaluated using the test set. Here, based on performance
metrics, model parsimony, and usability, the final models were selected for each area. To compare the models' power and found
relationships from different areas with each other, we also tested models between areas, using the same set of features.

2.3   Interpretation of Model Results

225   As we assume the underlying processes are the same, we expect that a model trained on one area can be extrapolated to another.
However, we expect the Overijssel model to be different, because while the aircraft covers agricultural land, we do not have
any agricultural tower sites in that area, such as in the Groene Hart and in Friesland.

Beyond optimizing the model for the best CO_2 predictions, we wanted to understand the model's functionalities and decisions
to infer knowledge on the underlying processes. Here, we used two approaches: the explainable AI tool SHapley Additive
230   exPlanations (SHAP) and annual simulations at our sites. For uncertainty quantification of these interpretations, we applied
bootstrapping: we repeated the partitioning in train and test sets one hundred times through a randomized parameter in the

splitting algorithm. We created one hundred models based on these train sets, and used the outputs from those models to assess the uncertainty.

The unified SHAP framework was developed to address the difficult interpretation of 'black box' machine learning models (Rodríguez-Pérez and Bajorath, 2020). The method relies on Shapley values, which determine the individual contribution of each feature to the final model outcome considering the collective contribution of all other features. Sequentially, each feature undergoes a process wherein its contribution to the model is negated by assigning a random value to it, thereby resulting in no added predictive power. By comparing model outputs with and without the contribution of a specific feature, the influence of this feature on the model is isolated (Lundberg and Lee, 2017). To consolidate how the model understands the effect of groundwater level, we try to fit regression lines on its SHAP values.

We assembled input data for the model at every site for the years 2020, 2021 and 2022. As our sites contained gaps in meteorological data, we used publicly available hourly data from Dutch Meteorological Institution (KNMI) and interpolated in time to obtain half-hourly values. We used station Cabauw for the Groene Hart area and station Hoogeveen for the Overijssel and Friesland areas. We let the model predict every half-hour flux and constructed annual NEE balances based on the actual WTDe level, as well as for hypothetical situations where the WTDe is altered by ± 10 cm.

3 Results

3.1 Optimized model settings

In this section we present the model optimization results. As expected, we found strong inter-correlations between meteorological variables (PAR, temperature, relative humidity), as well as between water-related variables (correlation matrices not shown here). Appendix C shows that CO_2 flux is most strongly related to PAR, which is also identified by XGBoost feature importance. Furthermore, temperature and EVI score high in all areas, whereas relative humidity scores high in all areas but Overijssel. The importance of the water-related features varies throughout areas. PeatDepth scores high in Friesland and Groene Hart, but lower in Overijssel and when all areas are combined.

For some soil classes such as pV and hV (see Appendix A for abbreviations) we find a surprisingly high correlation with CO_2 or high XGBoost Feature Importance. The data distribution of these features is different in the airborne and tower dataset: there are multiple towers where 100% of the footprint is (always) hV, whereas the airborne footprints contain a varying range of values for hV, mostly close to 0. Therefore, these features with strongly skewed distributions in the merged dataset were not taken into further consideration for construction of the models.

Based on the correlation matrices and two ranking plots, a pre-selection of features was made to use in the sequential backward floating selection. All models have known important drivers: PAR, RH, Tsfc and EVI. EVI was selected over NDVI because of better scores, and because of its correction for aerosol influence (Huete et al., 2002). Additionally, we included information on water and peat, but since we have various (correlated) features representing this information, we ran SBFS separately for each set of non-correlated features. The sets of features that we added to the drivers named above are: peat depth

Table 3. Hyperparameter results and corresponding scores, before (bef.) and after (aft.) hyperparameter tuning. Here, for each model, the same features are used: PAR, Tsfc, RH, EVI, PeatDepth and WTDe.

Area	colsample by tree	learning rate	max depth	min child weight	number estimators	sub-sample	R^2 bef.	R^2 aft.	RMSE bef.	RMSE aft.
All	0.9	0.005	11	4	3000	0.9	0.63	0.66	5.48	5.28
Friesland	0.8	0.005	11	4	3000	0.9	0.63	0.67	5.62	5.32
GrHart	0.8	0.005	11	1	2000	0.8	0.61	0.66	5.29	5.10
Overijssel	0.9	0.001	11	6	4000	0.8	0.17	0.58	6.90	5.30

combined with one water related feature (water table depth, soil moisture and air-filled pore space), and exposed peat depth,
265   resulting in four separate SBFS runs.

SBFS showed that as long as the feature set contains information on the daily and seasonal cycles, information about peat, and some water-related feature, the scores are very similar. The models for all areas, GrHart and Friesland all performed well, with little difference in scores. Overijssel's model performed less well. For every model, we selected two to three best performing feature subsets, and these continued to next step in model optimization: hyperparameter tuning. However, also
270   with optimized model settings, there was a minimal difference in performance with slightly different feature subsets. Hence, we selected six robust features, and tuned hyperparameters for every model with these features, to better enable comparisons between areas. This way, different results cannot be a consequence of different features used. The final features are: PAR, Tsfc, RH, EVI, PeatDepth and WTDe, and the optimized hyperparameters with corresponding scores are shown in Table 3.

Figure 5 shows how the models trained on different areas perform when tested on another area. Although the *Friesland*
275   model is the 'best model' in terms of R^2 and bias, we see that the model based on all data performs better overall when applied on the test sets of individual areas. Generally, models trained on specific areas have worse scores when predicting for other regions. Simulating the Overijssel data by models from other areas, and vice versa simulating the other areas with the Overijssel model results in the lowest R^2 scores, and the highest biases. For further analyses, we use the model based on all regions.

3.2   Shapley Explanations

280   To check the physical consistency of the trained model against prior knowledge and to understand how the model operates we analyze Shapley values for each of the selected features. Figure 6 shows an overview of all Shapley values for the model of all areas. A positive SHAP value indicates that the feature value has a positive contribution to the flux, i.e. increasing emissions or decreasing uptake, and a negative Shapley value vice versa. Increasing PAR and EVI drive more uptake and/or less emissions, whereas increasing temperature and deeper water table depth have the opposite effect. The influence of RH and PeatDepth is
285   not immediately clear from this beeswarm plot.

We delve further into the Shapley values for the features, through scatterplots shown in Fig. 7. Figure (a) shows the SHAP values of PAR and reflects the well-known light response curve, as learned by the model. As PAR increases, the contribution
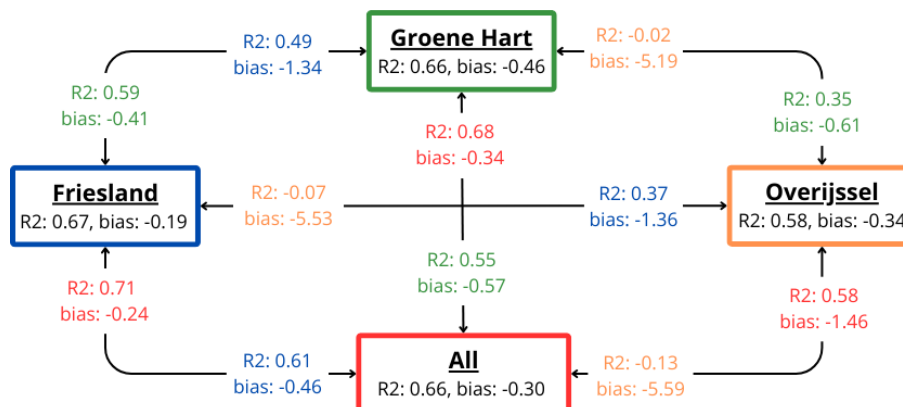
Figure 5. Models trained on data from different areas and corresponding performance scores when tested on the other areas. The same color is used for predictions from the same model. Here, the same features are used: PAR, Tsfc, RH, EVI, PeatDepth and WTDe, and the optimized hyperparameters are shown in Table 3.
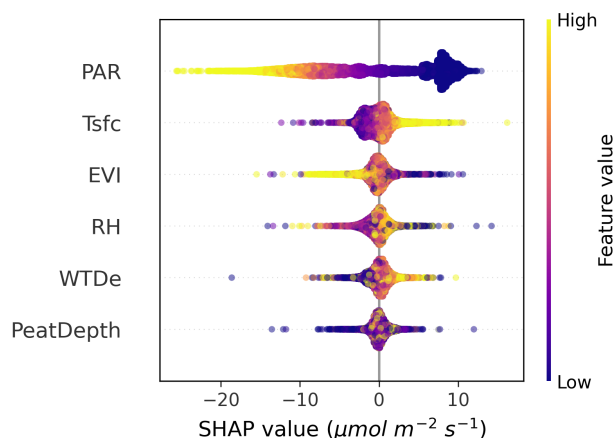

Figure 6. SHAP values for all features in the final model. The thickness indicates the amount of data points. For example, there are a lot of values with low PAR (nighttime at towers), that have a SHAP value between 5 and 10.

to the predicted flux becomes more negative, especially at higher temperatures. Conversely, with low PAR, SHAP values are positive and higher temperatures cause the flux to increase in the positive direction. Surface temperature drives nocturnal emissions (when PAR is 0), and drives day-time emissions once above about 15 °C. Optimal conditions for uptake are at RH between 40% and 80%, whereas drier conditions, correlating with highest PAR values, drive emissions, as well as wetter conditions that correlate with nighttimes. As WTDe increases, so does its SHAP value, but at deeper water levels this seems to level off, or even reverse. Below, we examine this in more detail. EVI values above about 0.55 have a clear stimulating effect of CO_2 uptake and the more so with higher PAR. EVI <0.55 does not influence CO_2 exchange, especially with PAR=0, indicating
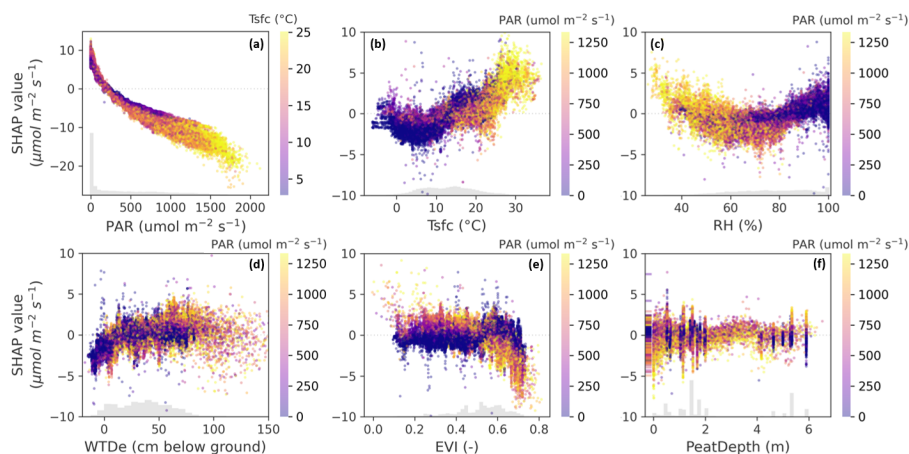
290

Figure 7. SHAP values of **a)** PAR, **b)** Tsfc, **c)** RH, **d)** WTDe, **e)** EVI and **f)** PeatDepth. The plots are colored by the values of another feature (Tsfc in (a), PAR in (b-f)), which in some cases correlates with the depicted one, due to diurnal or seasonal covariance. Nonetheless, the SHAP values represent the effect of only the selected feature on the x-axis. The color gives insight on the conditions in which this effect is happening. The vertical lines in plot (f) originate from the towers, as a static peat depth map was used. The amount of scatter indicates the robustness of the found relationships.

no clear effect of vegetation on nighttime emissions. This hints on emissions being driven mostly by heterotrophic respiration. Finally, peat depth shows no apparent relation with CO_2 exchange, even though its importance is large enough to be selected into the final model.

As we are specifically interested in WTDe, Fig. 8 shows fitted linear, parabolic and Gompertz lines on the bootstrapped Shapley values for WTDe. We tried fitting several other functions: bell-shaped functions, including and excluding a plateau in the middle; piecewise functions; sigmoid; logistic; shepherd, and more, but the depicted three performed best on our data, where WTDe ranges from -0.2 to 1.5 m. The linear regression stops at the peak of the parabola, at WTDe of 0.8 m. Similarly, the Gompertz function seems to truly flatten at this depth. However, the characteristic horizontal part at the beginning of the Gompertz curve is not visible in our data. Despite having different shapes, all three regression lines have approximately the same coefficient of determination (see Table 4). Up to 50 cm, the increase in emissions is coherent.

3.3 Simulations

For all our sites, we assembled half-hourly input data for 2020-2022. Inspecting the annual course of WTDe at the Overijssel sites, we found there is a systematic underestimation of the water table depth, i.e. OWASIS gives values associated to much drier conditions than the WTDe measurements. As a result, the average annual WTDe for Overijssel was the deepest of all our sites, while these sites are in wet nature areas. Hence, we discarded the simulations for the Overijssel sites. The WTDe at other sites was acceptable, although extremely deep summer water tables were often underestimated.
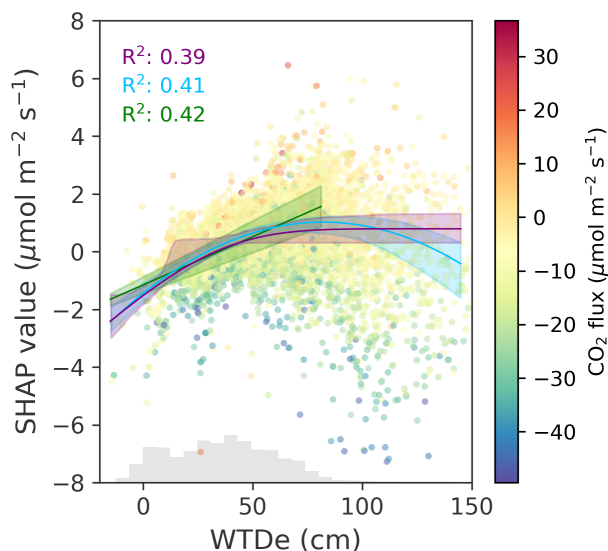
Figure 8. SHAP values with fitted linear, parabolic and Gompertz functions, colored respectively green, light blue and purple. Linear regression line stops at the peak of the parabola. The lines are drawn using the medians of 100x fitted parameters on the bootstrapped shapley values. The shaded areas represent the 90% confidence interval, based on the 5% and 95% of predictions at every WTDe-value. The R^2 scores are the means of 100 regression lines. See Table 4 for the fitted parameters.

Table 4. Fitted parameters for the linear, parabolic and Gompertz curves. The values represent the medians of the 100 fits based on boot-strapped models. The parameters in black correspond to CO_2 fluxes in μmol CO_2 m^{-2} s^{-1}, whereas the parameters in grey correspond to fluxes in ton CO_2 ha^{-1} yr^{-1}. x is WTDe in cm. As the slope in t CO_2 ha^{-1} yr^{-1} is one of our primary results, we present its confidence interval here: 4.64 (3.95, 5.39) * 10^{-1} per cm WTDe, while confidence intervals for all fitted parameters can be found in Appendix D.

Reg. type	R^2	Function	Parameters
lin. reg. up to $\pm$ 80 cm	0.42	$y = ax + b$	a: 3.34 * 10^{-2}, b: -1.15 *a: 4.64 * 10^{-1}, b: -16.0*
parabola	0.41	$y = A \cdot (x - x_0)^2 + B \cdot (x - x_0) + C$	A: -3.65 * 10^{-4}, B: 5.62 * 10^{-2}, C: -1.14, x_0: 5.18 *A: -5.07 * 10^{-3}, B: 7.81 * 10^{-1}, C: -15.9*
Gompertz	0.39	$y = S_{max} + S_{diff} \cdot e(-a \cdot e^{(-b \cdot x)})$	S_{max}: 0.79, S_{diff}: -10.2, a: 1.48, b: -1.65 * 10^{-2} *S_{max}: 11.0, S_{diff}: -142*

Italic: converted to t CO_2 ha^{-1} yr^{-1}.

Letting the model predict the fluxes and calculating the annual balances for the sites in Friesland and the Groene Hart resulted in Fig. 9. For every site-year, there are three dots: one based on the actual WTDe values, one where we subtracted 10 cm, and one where we added 10 cm to every value of WTDe. Therefore, the triplets represent the sensitivity of the site's annual
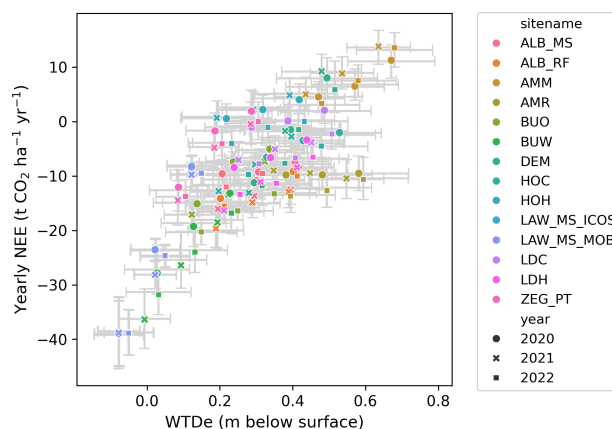
Figure 9. Simulated annual CO_2 balances for the sites in Friesland and the Groene Hart. The triplets represent simulations for WTDe - 10 cm, actual WTDe, and WTDe + 10 cm. The uncertainty intervals represent standard deviations, where the vertical intervals are based on 100x annual balances based on the bootstrapped data and corresponding models.

NEE balance to 10 cm change in WTDe. The sensitivity varies from site to site, but all sites combined, there is a curvilinear increase. Fitting a linear regression line results in a slope of 5.3 t CO_2 per ha^{-1} yr^{-1} per 10 cm increase in WTDe. However, closer examination shows the effect varies per site and per year, and levels off at deeper water levels. We did not extend the simulated effects for deeper water table depths as those scenarios would become unrealistic without altering the other features.

We grouped the sensitivity of predicted NEE fluxes to 10 cm WTDe increase by month, shown in Fig. 10. In gray, the range of sensitivity to 10 cm is depicted that is found in the literature. Although the mean of almost every month falls within the range and does not change sign, there is substantial scatter above and below. Furthermore, there is monthly variation in those means, being the lowest in summer, and highest in winter months. We tested whether this variation is statistically significant with a Welch's t-test, which accounts for the unequal variances across groups. Most month-to-month comparisons show significantly different means, but there is no distinct pattern (see Appendix E1). Grouping per season instead of per month shows that March-April-May (MAM) and June-July-August (JJA) do not differ significantly, but all other season comparisons do, suggesting different effects of 10 cm WTDe increase in autumn, winter, and in spring/summer (see Appendices E2 and E3). If we compare sensitivities to raising WTDe by 10 cm (rewetting), all seasons are significantly different (Appendix E3). Additionally, there is a large variation between sites (see Appendix E4).
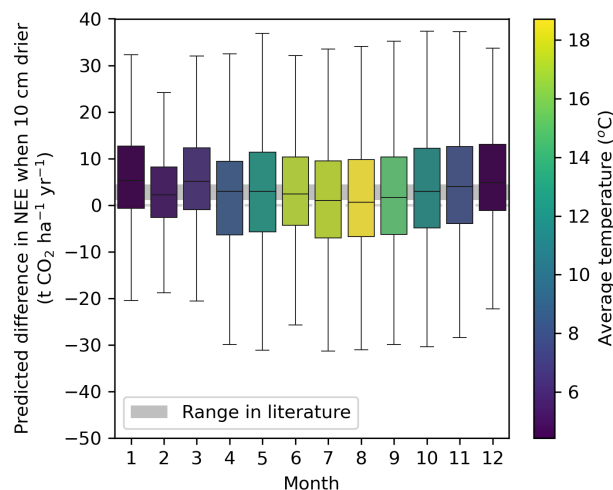
Figure 10. Monthly binned differences in NEE predictions when WTDe increased with 10 cm. The color represents the average temperature in our data for that specific month.

4 Discussion

4.1 Constructed models and their performance

330 Our final model explains 66% of the observed variance, and is able to provide further insights in key drivers of NEE. The found relationships with PAR, temperature, relative humidity and EVI are in line with physically known processes. The effects of water table depth will be discussed in detail in Sect. 4.3,

The obtained R^2 seems acceptable given the complex interactions analyzed and random noise levels typical for Eddy Covariance observations. Compared to studies also modeling CO_2 fluxes but by traditional methods, this R^2 is in the same range (Jung

335 et al., 2011; Zulueta et al., 2011; Dou et al., 2018). Also Zhou et al. (2023), who combined satellite data and EC measurements with a Random Forest model found an R^2 of 0.6. However, other similar studies using machine learning approaches tend to have higher R^2 (Metzger et al., 2013; Serafimovich et al., 2018; Vaughan et al., 2021), but these all focused on simulating heat fluxes, arguably a simpler process to analyze. Moreover, it appears that none of these studies used separate data subsets for learning and evaluation, as we did in the current study. Evaluating the model on the same dataset it was trained on, would

340 increase the R^2 for our best model from 0.66 to 0.89.

Although the Overijssel model performed acceptable, extrapolating to another area gives worse results than taking the mean of fluxes in respective area (negative R^2 scores). This does not indicate that the Overijssel area itself is fundamentally distinct. Instead, it is a consequence of the WTDe data not reflecting the wet site conditions well. This might also explain why the *All* model performs better on GrHart and Friesland than when tested on its 'own' test set, which includes Overijssel data. We

345 trained a model on the combination of only the Friesland and GrHart data, and it obtained an R^2 of 0.67 and bias of -0.26:

17

slightly better results than when the Overijssel data was included. We find that extrapolating between Friesland and GrHart gives reasonable results. Still, we find best results when using the model trained on all data. Overall, we think that these machine learning models, especially the model including GrHart and Friesland data, perform adequately and were well able to find relationships with spatio-temporal variables.

350 It has been suggested that peat depth exposed to air is a more direct indicator for peat decomposition and thus for CO_2 emissions than water table depth (Aben et al., 2024). In our study, we do not find this. The same holds for air-filled pore space, although this property is arguably more complex to model because of peat swelling and shrinking and varying porosity. However, it should be noted that the differences in performance when water table depth is replaced with another water-related feature are minimal, indicating a robust underlying process or relationship that the model is able to find.

355 We assumed our method of partitioning the data - selecting flight legs and weeks of measurements for the test set and use the rest for training - avoided data leakage. To further examine this, we also applied two other partitioning strategies: one more stringent, selecting entire flights and sites for the test set, and one less stringent, selecting a random 20% as test set. The former resulted in worse models, but this depended on which sites were left out, as some provide more data and insights than others. Nonetheless, this partitioning strategy should be further examined in future research, because it may better reduce data leakage.
360 The models based on randomly selected training data all performed slightly better than our the models we use in this study, indicating that our current strategy avoids some data leakage.

In the final optimized hyperparameters, we see that models differ mostly in minimum child weight. Models with a high value for this parameter are more conservative, and end up with fewer splits. We believe the final model based on all areas achieves a good balance between capturing complex patterns, and avoiding overfitting.

365 ## 4.2 Combining airborne & tower data: pros and cons

As discussed, airborne and tower data have different qualities: airborne data is spatially exhaustive but temporally limited, and tower data is temporally continuous but with a limited spatial extent. On a given day, the airborne data provides a gradient of seasonally varying landscape features (e.g. WTDe and EVI), as opposed to point values at towers, and represents the entire area, making the input data more diverse. Because of their complementarity, we were able to develop an ML model that includes
370 information on the daily cycle as well as extends beyond the tower locations. Nevertheless, practicalities when combining tower and airborne data may lead to some spurious correlations between features and target, due to specifically two aspects. Firstly, tower data includes night-time measurements while airborne data is only collected during the day, resulting in differences for weather-related features (PAR, RH, Tsfc) and almost all positive CO_2 fluxes stemming from tower data. Secondly, land use and soil classes are fixed for the tower data and variable for the aircraft data. Together, this led us to omit features from
375 merged models that had too distinct distributions in airborne and tower data, such as the soil class hV. Although we prevented artifacts by excluding these features, we also omitted possibly valuable information. We created a categorical variable with the most prevalent land use or soil class in the footprint, but this did not improve model performance. Hence, our results underpin what has been previously found: the relationship with WTDe holds regardless of the land use or soil class (Evans et al., 2021; Tiemeyer et al., 2020).

380   Another significant difference between the airborne and tower datasets in our study is their size: the former has 10,400 records, the latter 66,400. Consequently, the merged dataset consists mainly of tower measurements, and if we construct a model only based on tower data, we see very similar performance of the two models. In a preliminary study we conducted, the addition of airborne data significantly improved the ML model (R^2 increased from 0.47 to 0.61). However, in that case we had restricted datasets available: 7,900 records for the tower data, originating from only two towers and spanning 6 months, and
385   2,600 records for the airborne data (spanning 18 months). Under similar circumstances, when the tower dataset lacks spatial and temporal coverage, we believe the inclusion of airborne data can improve the model substantially. However, in the current study, as the tower network had been seriously extended resulting in much higher spatial coverage, this was not directly visible in an improved R^2. We examined the airborne data's added value in three ways.

      First, we tested excluding several towers from the training process, to determine if we could replicate the added value
390   of airborne data as in the previous study, and we compared resulting tower and merged models. The outcomes were highly variable, because each tower has different qualities, lengths of measurements, presence of gaps, etc., but removal of certain towers showed similar results as in our previous study. Second, we trained one model with equal amount of airborne and tower instances, which attained an R^2 of 0.62. Training the model only airborne data on the other hand, resulted in an R^2 of 0.37. This suggests that it is not only the high amount of data that is beneficial for the model, but that there is also intrinsic
395   value in adding tower measurements to airborne. Third, we let all models predict for airborne, tower and merged datasets. The difference between the merged and tower model shows when they are tested on the airborne data: the variance in airborne data was explained for 34% and 10% by the merged and tower models, respectively. The airborne data represents regional fluxes originating from across the entire area. Hence, to model these regional fluxes, extending beyond the locations of the measurement sites, it is necessary to include airborne measurements in the model. Future research could investigate where
400   the trade-off is: when do airborne measurements provide significant complementary benefits to tower data? With *x* number of towers of time *y*, how many flights should be done to obtain enough information? The costs of tower and airborne maintenance should be included. This research could shed light on the most efficient measurement strategies in areas with limited access to resources or with inaccessible terrains.

4.3   Influence of water table depth

405   ### 4.3.1   Found relationship compared to previous estimates

      Here, we compare our findings to groundwater–CO_2 emission relations found by previous studies, shown in Fig. 11 (parameters can be found in Appendix G). Note that although the literature relations, SHAP results and site simulations share the same units (t CO_2 ha^{-1} yr^{-1}), they represent different types of results. Regression lines of Boonman et al. (2022); Aben et al. (2024); Kruijt et al. (2023); Evans et al. (2021) are based on NECB estimates, and thus include corrections for annual carbon import
410   and export through manure and harvest. In addition, they are based on multi-site comparisons thus indicate mostly spatial dependencies. The regressions based on SHAP values do not distinguish between temporal or spatial influence, and depict the effect of changing WTDe on NEE flux as understood by the model, rather than actual CO_2 budgets. The annual NEE estimates
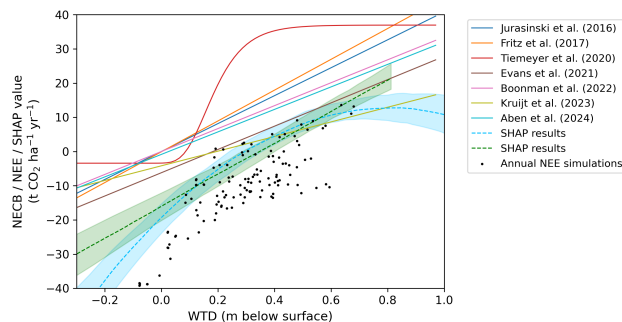
Figure 11. Comparison between current findings and literature studies on CO_2 flux vs water table depth. Three types of results are shown: regression lines from the literature based on annual NEE or NECB estimates; the dashed lines represent the linear and parabolic fits on the SHAP values; the simulated annual NEE totals are visualized in black, in three-fold per site (WTDe −10 cm, actual WTDe, and WTDe + 10 cm). The shaded areas represent the 90% confidence intervals of the SHAP regressions, based on the 5% and 95% of predictions at every WTDe-value using the bootstrapped models. SHAP regressions are based on direct WTDe values, whereas literature studies and site simulations use annual averages. All fitted regressions can be found in Appendix G.

on the other hand represent the sum of year-round simulated CO_2 fluxes with varying WTDe, neglecting carbon import and export as they are not part of the ML model.

415   Both our NEE estimates and SHAP regression lines show more negative values than those reported by other studies. Firstly, for the annual NEE estimates, this can be partially explained by the difference between NEE and NECB budgets. On highly productive grasslands the export of harvest can be significant, as for example shown by Kruijt et al. (2023) where both NEE and NECB are reported for the same pastures as in the current study. Generally, the carbon in harvested biomass is released back to the atmosphere within a year, mostly in close-by areas. However, these emissions are not measured by the EC towers,
420 though the aircraft might measure them when flying over e.g. barns. Furthermore, simulated fluxes are underestimated due to the negative model bias, which mounts up to 4.2 t CO_2 ha^{-1} annually. This bias may partly result from the lack of data on mowing events: the model continues to predict uptake as usual after mowing, whereas in reality, uptake stops immediately after grass removal (see Appendix F). Secondly, for SHAP values, the reference population mostly determines where SHAP=0, hence the change in impact on the model (i.e. the slope) is more meaningful than the exact level of impact (i.e. the intercept).

425   Apart from the negative offset, we find very similar relations to current estimates: per 10 cm WTDe, emissions increase annually with 4.6 t CO_2 ha^{-1} based on SHAP and 5.3 t CO_2 ha^{-1} based on annual NEE estimates. Since the annual NEE estimates are primarily based on farms with distinctive water table management - likely not representative of the entire area - we argue that the SHAP slope, which is based on the full dataset, is a better estimate. Considering the different approaches used, clarifying the negative offset, the correspondence with the literature is remarkable. It suggests that the underlying WTD-
430 CO_2 relation persists, also when harvest and manure are disregarded. This may be because we investigate the regional scale, where local fluctuations balance each other out. Another explanation might be that harvesting biomass has mostly short-term

impacts on the system (see e.g. Appendix (F). By the combination of aircraft measurements, instantaneous data rather than annual totals, and machine learning, we were able to extract this fundamental groundwater–CO_2 emissions relationship.

4.3.2 Non-linearity of the found relationship

435 We did not find a linear relationship over the entire range of WTDe in our data (see Fig. 8, 9 and 11). Emissions undoubtedly rise with deeper WTDe, but deeper than 0.8 m, they cease to increase. Based on the SHAP explanations, the optimum-based curve explains the data slightly better than the Gompertz curve, indicating a decrease rather than a stabilization of the effect at deeper water levels. We lack sufficient data at these deeper water levels to make a concluding statement, but given that our datapoints represent instantaneous NEE fluxes instead of annual estimates, it is entirely plausible that a curve with an

440 optimum indeed better represents the underlying process. For example, in conditions with WTDe >0.8 m, moisture conditions can be sub-optimal for peat decomposition. Nijman et al. (2024) and Campbell et al. (2021) found similar response curves comparing nighttime ecosystem respiration to respectively water-filled pore space and volumetric soil moisture content. In these studies, as well as in the current study, instantaneous measurements are compared, as opposed to the studies discussed above. Furthermore, the relationship between WTDe and CO_2 flux with deeper water levels is less direct, since the actual soil

445 moisture in the unsaturated zone can vary substantially, which potentially explains the larger scatter in Fig. 8 - in addition to less available training data.

As opposed to Tiemeyer et al. (2020), who finds saturation around a WTDe of 0.4 m, we find the increase to persist until 0.8 m, as was also found by Boonman et al. (2022) and Aben et al. (2024) for average summer WTDe. In the study by Tiemeyer et al. (2020), the amount of data for water levels deeper than 0.4 m is limited. Because we include fluxes on a short time

450 scale and we cover a wider range of WTDe values by using the aircraft, we have a larger dataset to support the observation of increasing emissions up to 0.8 m. Additionally, we do not find an initial plateau at shallow water levels, contrary to what is currently assumed. Although we included wet sites, such as those in Overijssel, the water information product we used was not able to capture these wet conditions and gave incorrect values. Hence, our model was not trained properly on wet sites, and we cannot substantiate the lack of a plateau.

455 4.3.3 Intra-annual and spatial variability

Based on the above discussion, our results are mostly in line with previous studies, with two notable exceptions: the optimum-based shape instead of the linear or Gompertz functions, and the increase of emissions until 0.8 m instead of 0.4 m, also found by (Boonman et al., 2022; Aben et al., 2024). A third insight is the intra-annual variation of sensitivity to WTDe change, as shown in Fig. 10 and in Appendix E. Further drainage in summer might not impact emissions as much, whereas drainage in

460 winter has a bigger impact. This might be a reflectance of the monthly varying WTDe: in summer, with very deep groundwater levels, the available oxygen for decomposition is not linearly depending on the groundwater and also determined by the soil structure and its capillary action. As a result, at very deep water levels, the soil in the unsaturated zone can be relatively wet and decrease the effect of further drainage. On the other hand, with more shallow water levels in winter, the presence of roots in the unsaturated zone leads to soil desiccation, promoting oxygen availability down to the water table. Hence, when the water

465   table is lowered, this has a more direct effect on CO_2 emissions. Furthermore, we find that raising the WTDe by 10 cm has a significantly different effect across all seasons. Our methodology enables future research on intra-annual occurrences such as season-based mitigation measures and extreme weather events.

In the current study, we applied the simulations on the site locations, which show high variability in responses. These site simulations were partially motivated by the goal of comparing predictions to measurements, as in Appendix F. However, since

470   our model is trained on data covering the entire region — enabled by the use of aircraft data — the simulations can be applied to any location in the area, which can be of interest to policy makers. Although our input data is too coarse to predict the emissions at farm-level, the model can provide insights at municipal or provincial scale in carbon flux dynamics over the years.

4.4 Implications for mitigation strategies

Our findings suggest that to reduce CO_2 emissions, the optimal water management would be to set the water level as high as

475   possible. To reduce greenhouse gas emissions in general and mitigate climate change, the trade-off with methane should be taken into account (Buzacott et al., 2024). The slope we found lies at the upper limit of current estimates, and our results show that emissions continue to increase up to a WTDe of 0.8 m, as opposed to 0.4 m in Tiemeyer et al. (2020). Together, these results suggest higher emissions than those that would otherwise be calculated based on previous studies. This would entail that effective mitigation strategies are even more crucial, as the potential for carbon emissions from drained soils may have been

480   underestimated. As such, mitigation measures should not only take into account average annual water table depth, but also the different system behavior throughout the year. Measures should focus on rewetting during the summer and specific attention should go to not lowering the water table during winter. In general, intra-annual variation of responses should be considered.

4.5 Recommendations

4.5.1 Incorporation of additional data

485   A general remark on the findings in this study is that the water level data is from a company, that develops and maintains the OWASIS information products together with water boards and knowledge institutes. Comparing OWASIS WTDe to measured WTDe at tower locations, we found that values in summer were often underestimated. Potential sources of errors in the OWASIS data are heterogeneous infiltration capacity within pixel cells, or limitations of remote sensing in capturing deeper soil layers. Despite these uncertainties and underestimations, we consider the OWASIS product suitable for the purposes of

490   this study, as it covers the entire study area, and we have a satisfactory amount of data to balance out pixel-errors.

Still, future research should prioritize including a high-quality soil water product of high(er) spatial resolution. Due to the use of aircraft data, the information on water should be spatially distributed and measurements at tower sites do not suffice. The Netherlands has manifold regional hydrological products that can be used for future regional studies (NHI, 2025). In addition, there are numerous remote sensing products that could be used or combined for water-related information, possibly

495   including ground truth data, as for example was done by Koch et al. (2023) for Denmark, who developed a WTD map for Danish peatlands with a spatial resolution of 10 m. Considering the likelihood that farmers might apply mitigation measures

such as changing the water level management or adding a clay layer on the peat, corresponding data would be valuable for subsequent studies. Furthermore, a vegetation index with higher spatial and temporal resolution could be incorporated, such as from the Dutch groenmonitor.nl website, ideally enabling identification of mowing events. In the current study, we did not

500 directly incorporate mowing or harvest events.

4.5.2 Methodological advancements

As mentioned previously, future research should investigate the optimal combination of tower and aircraft data. This way, strategies for the modeling of remote or data-sparse locations can be developed. Herein, a stricter train-test algorithm can be applied. The SHAP framework has proven highly effective in revealing the processes understood by the ML model. By making

505 the reference population more representative, the base value might become more meaningful. However, the SHAP values do not per definition reflect causal relationships. As an extension to the interpretation by SHAP, therefore, we recommend to include more causality-based and/or physics-based components. Here, we discuss some promising approaches, and our first attempts in applying them. However, they are subject of future plans.

To start, information theory (IT) has already been used in studies to examine causalities of CO_2 fluxes (Farahani and Good-

510 well, 2024; Yuan et al., 2022; Arora et al., 2019). It is a mathematical approach to study the amount of information in a dataset or process based on Shannon entropy: a measure of uncertainty in a system, quantifying its unpredictability. In the current study, Shannon entropy can shed light on the information in tower vs. airborne data. We computed the Shannon entropy of CO_2 fluxes in our tower, airborne and merged datasets (tower + airborne), as well as an entropy-based metric based on Farahani and Goodwell (2024) (see Appendix H for the formula and results). The airborne data has the highest Shannen entropy, indicating

515 the highest level of information in the dataset, followed by the merged data. The results on the metric suggest that the merged model captures the variability in the data better compared to the airborne or tower models. However, the differences are relatively small. Additionally, the results show the tower model is overly complex or noisy, whereas the airborne and merged models slightly under-represent the variability in the data, meaning they smooth over some of the finer details or variability in the CO_2 fluxes (Farahani and Goodwell, 2024). At the regional scale, we believe the latter is preferable.

520 A second option is to alter the loss function in a deep learning model, which is better modifiable than in XGBoost. For example, transfer entropy can be minimized through the loss as is done by Yuan et al. (2022), or directly physically inspired functions or models can be implemented, as successfully done by Liu et al. (2024). New, innovative approaches such as double ML based on causality offer great opportunities for further exploring the greenhouse gas dynamics in drained peatlands (Cohrs et al., 2024).

525 ## 5 Conclusions

In this study, we applied Boosted Regression Trees to learn the relationship between CO_2 flux and landscape characteristics from drained peatlands in the Netherlands. We investigated data from the three main fen meadow areas in the country (Groene Hart, Friesland and Overijssel) and finally constructed a model based on all areas combined. We merged CO_2 flux data from

both airborne and tower measurements, and, to our knowledge, this study is the first to use this combination as input for a

530 machine learning model. The models were optimized with feature selection and hyperparameter tuning, and we accounted for data-leakage by splitting the train and test set based on flight-legs and week numbers. Subsequently, we used the SHAP framework and simulations to assess the influence of most important and relevant environmental drivers.

The method works and the models perform reasonably well with R^2 scores between 0.58 and 0.67. We found that extrapolating the model from one region to another performs adequately, but that the model including all regions is best for this

535 purpose. As long as the feature subset contained information on the daily and seasonal cycles, information about peat, and some water-related feature, the scores of the models were very similar. Hence, the final, robust features that explained most of the variance in CO_2 fluxes are PAR (photosynthetically active radiation), temperature, relative humidity, EVI, peat depth and effective water table depth (WTDe).

Based on the SHAP values, we find an increase of CO_2 emissions until a water table depth of around 0.8 m below the surface.

540 These emissions increase with 4.6 t CO_2 ha^{-1} yr^{-1} per 10 cm WTDe on average (90% CI: 4.0-5.4 t CO_2 ha^{-1} yr^{-1}), which is in agreement with other estimates, albeit at the higher end of the range found in the literature. Together, these results suggest higher emissions related to WTDe than previous studies. Furthermore, we find that an optimum-based function describes the influence of WTDe best within our WTDe range. However, further research using instantaneous measurements on a short time scale (thus including data at deeper water levels) should point out whether the emissions decrease or stabilize after 0.8 m. We

545 find intra-annual and spatial variation in the response of CO_2 flux to 10 cm drying and rewetting. These aspects should be taken into account when developing mitigation measures.

Future research should prioritize including data on water table depth with higher spatial resolution that better captures wet and extremely drained conditions. Causality-based approaches and physics-guided ML models form a promising direction for future studies. Lastly, a comprehensive comparative study on the synergies between airborne and tower data could contribute

550 to establishing efficient, cost-effective measurement strategies.

Data availability. The simulated annual NEE totals and corresponding groundwater levels are available upon request. The input data for the ML models is not yet publicly available due to ongoing research by Bataille et al..

Appendix A: Reclassification tables

A1   Reclassification land use classes

New class	New code	LGN2020 classes and codes
Bare soil	bSl	beach (31); drift sands (35)
Build up Areas	Bld	urban build areas (18); rural build areas (19); forest in build areas (20); forest in rural build areas (22); grass in build areas (23); bare soil in build area (24); main (rail)roads (25); build area (26); grass in rural build area (28)
coniferous Forest	cFr	con. forest (12); forest in bogs (40)
deciduous Forest	dFr	orchard (9); dec. forest (11); forest in fens (43); tree nursery (61); fruit nursery (62)
Fens and Bogs	FnB	bogs (20); other marshland (41); reed lands (42); fenmeadow area (45)
Grasslands	Grs	nature grasslands (45); grass (1); other land use in rural area (27); salf marsh (30); coastal grass (46); other grass (47);
Greenhouses	Ghs	greenhouse horticulture (8)
Heath	Hth	open dune vegetation (32); closed dune vegetation (33); dune heath (34); heath (36); sparse grassy heath (37); dense grassy heath (38)
open Water	Wat	fresh water bodies (16); salt water bodies (17)
Shrubs	Shr	low shrubs in bogs (321); low shrubs in fens (322); low shrubs (323); high shrubs in bogs (331); high shrubs in fens (332); high shrubs (332)
Spring Crops	SpC	cereals (5); flower bulbs (10)
Summer Crops	SuC	maize (2); potatoes (3); sugar beets (4); other crops (6)

Figure A1. Reclassification of land use classes.

555 **A2   Reclassification soil classes**

New class	New code	Codes from soil map BOFEK2020
Moerige gronden	W	1: vWz,7: zEZ23,29: zWp,31: zEZ21,40: zEZ30,41: zWz,44: Wg,53: kWp,63: bEZ21,65: bEZ23,72: bEZ30,82: AWg,88: EK19,104: AWv,123: EK79,124: Wo,126: EZg23,131: EZg21,143: iWz,149: EZ50A,152: iWp,174: EK76,179: EL5,187: kWz,191: EZg30,226: vWp,239: uWz
Zandgronden	zandG	2: Hd30,16: Zn21,21: Zn10A,22: Zn23,23: Zn40A,24: Zn50A,32: Y30,33: Hd21,34: Zn50Ab,35: Zn30,36: Zn30A,37: Hn21,43: pZn23,45: Hn23,46: Zn30Ab,73: cHn21,84: cHd21,91: cZd21,94: cHn23,96: cY21,99: cHd23,107: cHd30,116: MOb75,121: MOo05,122: cY23,134: pZg23,135: pZg21,137: Hn30,140: cHn30,147: MOb72,151: cZd30,157: pZn30,159: cY30,171: cZd23,193: pZg30,201: pZn21,204: Zd23,209: Zb21,213: Hd23,218: Sn13A,224: tZd21,225: Zb23,227: Y21,229: Zd21,230: Sn14A,240: Y23,256: Zb20A,259: Zd20A,260: Zb30,261: Zd30,262: Zd30A,263: Y23b,264: Zd20Ab,265: Zb30A,268: tZd23,272: tZd30,282: pZg20A,296: MOo02,300: MOb12,301: MOb15
Zeeklei	zeeK	3: Mn25A,27: Mn35A,30: Mn45A,39: Mn12A,48: Mn56C,54: Mn15A,55: Mo10A,56: Mv41C,57: Mv61C,59: gMn15C,60: gMn88C,61: Mn82A,79: Mn25C,80: Mn85C,85: Mn86C,87: Mn15C,89: Mo20A,106: Mo80A,110: gMn25C,113: gMn53C,125: Mn22A,139: Mo80C,148: Mn52C,154: Mn56A,156: gMn83C,160: Mv81A,161: Mv51A,166: Mn82C,167: Mn86A,170: pMn85A,172: gMn58C,173: pMn55A,176: pMo50,177: pMo80,180: pMv81,194: MK,210: MA,222: pMn86C,228: gMn85C,236: kMn48C,241: kMn63C,251: kMn43C,271: gMn52C,274: gMn82C,275: pMn55C,276: pMn56C,277: pMn82A,278: pMn82C,279: pMn85C,280: pMv51,286: pMn52C,294: KT,295: kMn68C,297: MZz,298: pMn52A,299: Mo50C,302: MZk
Rivierklei	rivK	4: Rn46A,5: Rn47C,6: Rn15A,8: Rn15C,9: Rn44C,10: Rn67C,11: Rn52A,12: Rd90C,13: Rn42C,15: Rn14C,17: Rn62C,18: Rn66A,19: Rn45A,20: Rn45C,49: KRn1,50: pRn59,51: Rd10A,52: KRn2,81: FG,83: bRn46C,128: Rn82A,129: Rn94C,130: Rn95A,136: Ro40A,141: Rv01A,142: Rv01C,145: KRd7,162: Rn95C,165: Ro60C,178: KRn8,181: pRn56,182: pRn86,192: FK,195: KX,208: Rd10C,220: Rd90A,266: Ro40C,267: Ro60A,270: KRd1,285: pKRn1,287: pKRn2,289: pRn89
Vlierveengronden	V	14: Vp,153: Vc,158: Vd,163: Vk,169: Vr,202: Vs,203: Vo,269: Vb
Overige veengronden	overigV	25: zVc,26: zVz,28: zVp,38: zVs,64: aVp,77: aVc,86: aVs,144: iVz,273: hEV,290: iVp,291: iVc,292: iVs
Weideveengronden	pV	42: pVc,189: pVs,190: pVz,281: pVb,283: pVk,284: pVr,288: pVd
Koopveengronden	hV	47: hVk,132: hVc,138: hVd,146: hVr,175: hVb,247: hVs
Gedefinieerde associaties	gedA	58: AP,66: AZW7A,67: AZW6A,75: AO,76: AS,78: AZW8A,95: AQ,100: AR,101: AZW0A,102: AVo,103: AZW1A,105: AZ1,108: AHa,109: AHc,111: AHl,112: AM,115: AZW5A,117: AHk,118: AK,127: AFk,155: AFz,183: ABz,188: ABk,231: ABv,232: ABl,233: AEm9A,234: AEm8,235: AEp7A,237: AEm9,238: AGm9C,242: AAK,243: AAP,244: AHs,245: AHv,246: AHt,248: AVk,249: ALu,250: AHz,252: AD,254: AEk9,255: AMm,257: AEm5,258: AEp6A
Madeveengronden, ondergrond zand zonder humuspodzol	aVz	62: aVz
Leem en brikgronden	leem	68: BZd24,69: BLb6,70: BKh25,71: BKh26,74: BLd6,90: BZd23,92: BLh5,93: BLh6,97: BLn5,98: BLn6,114: BKd25,119: BLd5,120: BKd26,150: Ld6,168: pLn5,186: Lh6,196: Ldd6,197: Ld5,198: Ldh6,205: Ldh5,206: Lnd6,207: Lnh6,211: Lh5,212: Ln5,214: KK,215: KM,216: KS,219: Ln6,221: Lnd5,223: Ldd5
Koopveengronden, ondergrond zand	hVz	133: hVz
Waardveengronden	kV	164: kVc,184: kVs,199: kVk,200: kVr,217: kVd,293: kVb
Waardveengronden, ondergrond zand	kVz	185: kVz
Vlierveengronden, ondergrond zand zonder humuspodzol	V	253: Vz

Figure A2. Reclassification of soil classes.

Appendix B: Hyperparameter tuning grid

Hyperparameter	Values
n_estimators	1000, 2000, 3000, 4000, 6000, 7000
max_depth	5, 6, 8, 9, 11
learning_rate	0.1, 0.005, 0.001
subsample	0.8, 0.9, 1
colsample_bytree	0.8, 0.9, 1
min_child_weight	1, 2, 4, 6

Table B1. Parameter grid for hyperparameter tuning.
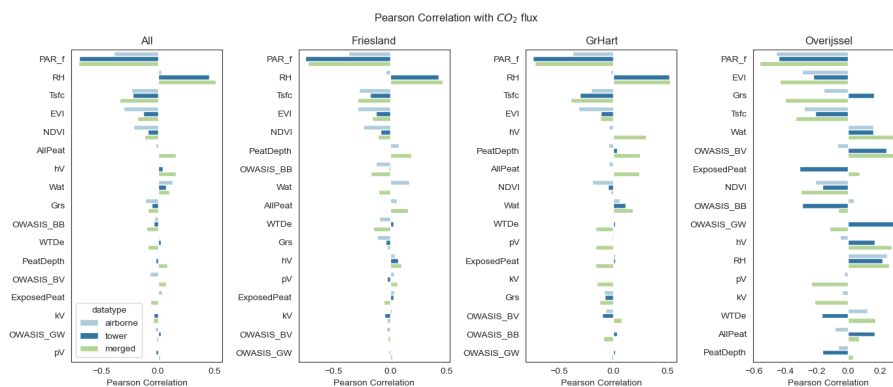
Appendix C: Model optimization results

Figure C1. Pearson correlation between features and CO_2 flux for airborne, tower and merged datasets per area.

Appendix D: Confidence intervals of SHAP regression lines

Table D1. Fitted parameters on the SHAP regression lines: LR indicates linear regression, P indicates parabola, G indicates Gompertz. The formulas are shown in Table 4. Values between brackets indicate 90% confidence intervals, based on the 5th and 95th percentile of all bootstrapped regression lines,

Reg.	Fitted parameters			
	a	b		
LR	3.34e-02 (2.84e-02, 3.88e-02)	-1.15e+00 (-1.45e+00, -8.72e-01)		
	4.64e-01 (3.95e-01, 5.39e-01)	*-1.60e+01 (-2.02e+01, -1.21e+01)*		
	A	B	C	x0
P	-3.65e-04 (-4.41e-04, -2.81e-04)	5.63e-02 (1.12e-02, 9.77e-02)	-1.14e+00 (-5.44e+00, 6.88e-01)	5.18e+00
	-5.07e-03 (-6.12e-03, -3.91e-03)	*7.81e-01 (1.55e-01, 1.36e+00)*	*-1.59e+01 (-7.56e+01, 9.56e+00)*	*(-4.95e+01, 6.49e+01)*
	S_{max}	S_{diff}	a	b
G	7.90e-01 (-2.77e+00, 1.29e+00)	-1.02e+01 (-5.62e+01, 4.10e+00)	1.48e+00 (2.09e-02, 3.21e+00)	-1.65e-02
	1.10e+01 (-3.85e+01, 1.80e+01)	*-1.42e+02 (-7.80e+02, 5.70e+01)*		*(-3.62e-01, 3.08e-02)*

Italic: converted to t CO_2 ha^{-1} yr^{-1}.

Appendix E: Sensitivity of NEE when WTDe $\pm$ 10 cm

560   ### E1   Welch's t-test on monthly values
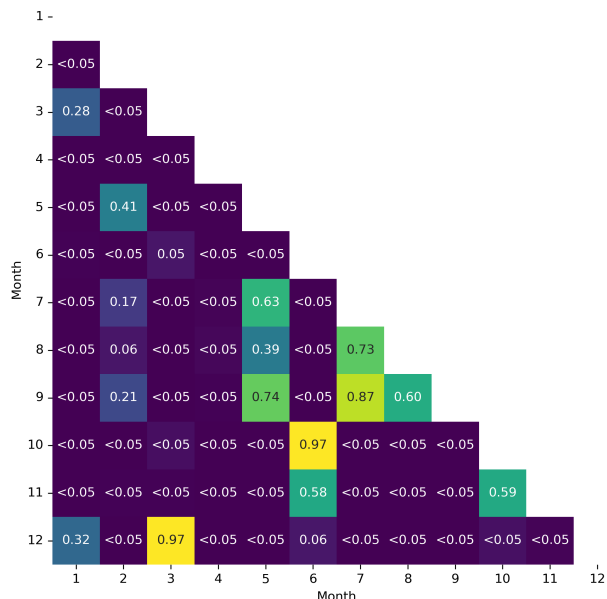
Figure E1. Results of p-values for month-to-month comparison of responses to 10 cm drying in WTDe, using Welch's t-test.
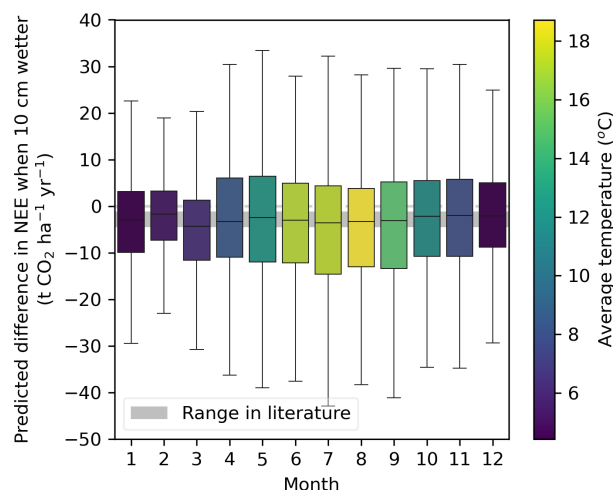
Figure E2. Monthly binned differences in NEE predictions when WTDe was raised with 10 cm. The color represents the average temperature in our data for that specific month.
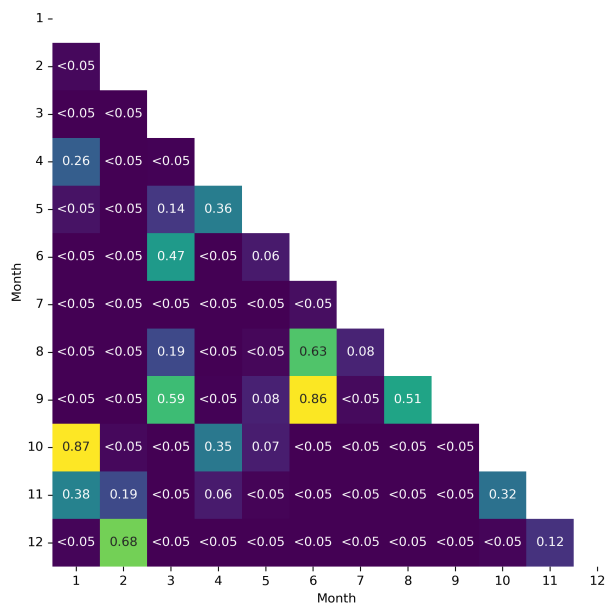
Figure E3. Results of p-values for month-to-month comparison of responses to 10 cm rewetting in WTDe, using Welch's t-test.

E2 Seasonal difference in fluxes after change in WTDe
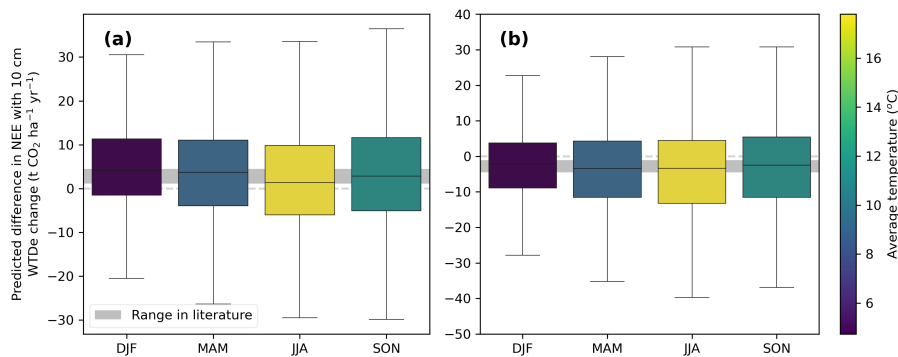
Figure E4. Responses of lowering of WTDe by 10 cm of all sites combined, binned per season. **(a)** shows the results for 10 cm drier, **(b)** for 10 cm wetter

E3 Welch's t-test on seasonal values
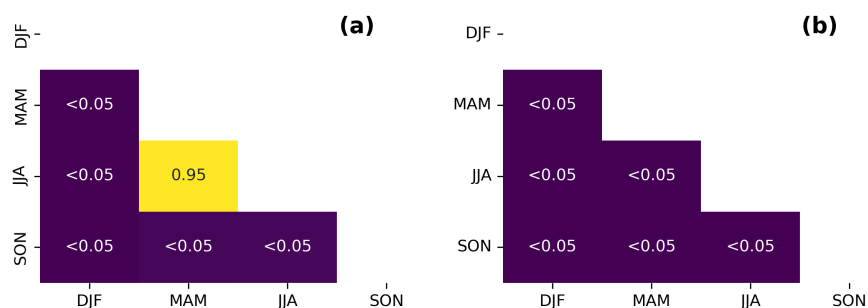
Figure E5. Results of p-values for season-to-season comparison of responses to 10 cm change in WTDe, using Welch's t-test. **(a)** shows the results for 10 cm drier, **(b)** for 10 cm wetter.

E4   Sensitivity per site
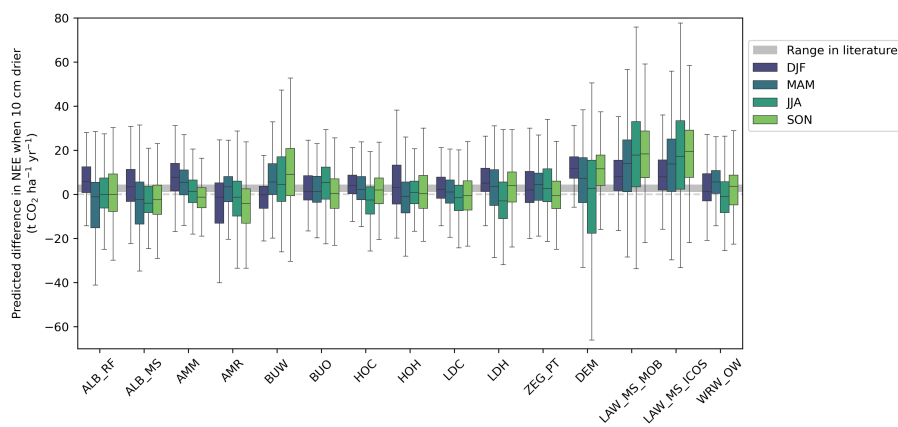
Figure E6. Responses of sites to lowering of WTDe by 10 cm, binned per season.
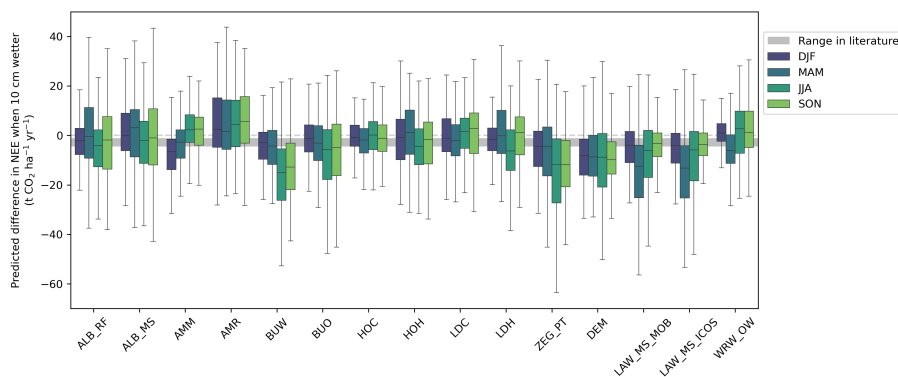
Figure E7. Responses of sites to raising of WTDe by 10 cm, binned per season.

Appendix F: Model predictions and mowing events
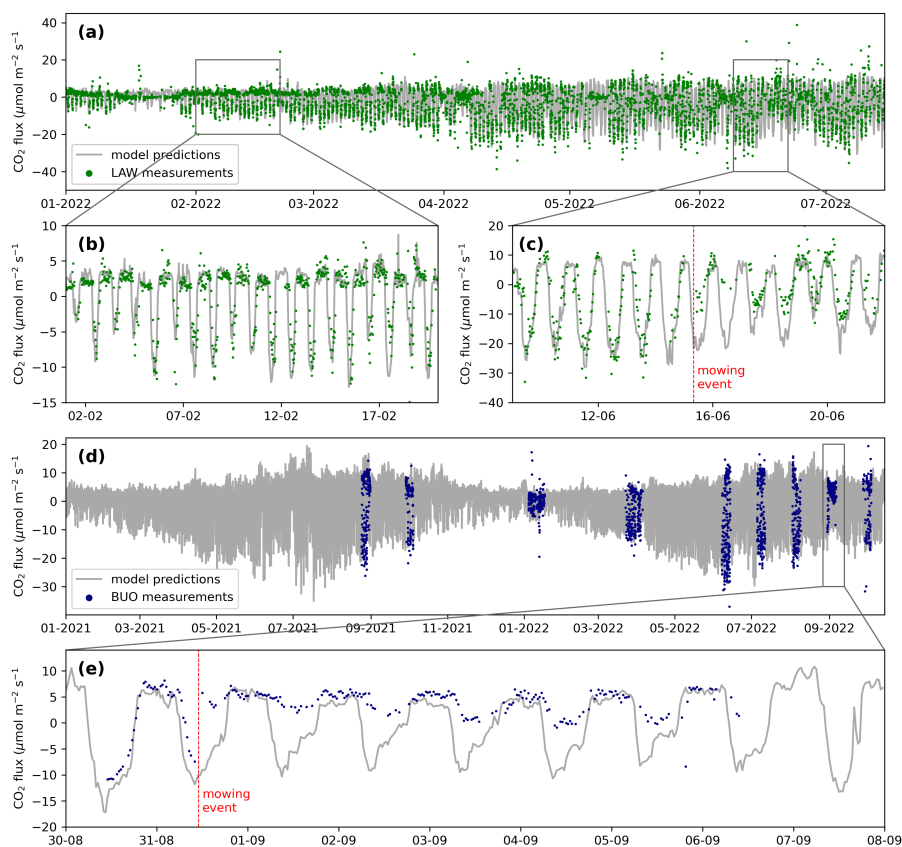
Figure F1. Predictions and measurements of CO_2 flux at LAW **(a–c)** and BUO **(d, e)**. In **(c)** and **(e)**, mowing events are highlighted with a red line. In **(e)**, uptake immediately stops after mowing, whereas in **(c)**, there is still uptake after mowing, which decreases the day after. We hypothesize this is due to the mowing of different parcels at LAW on different days, possibly by the neighboring farmer.

33

Appendix G: Relationships found in literature

565

Table G1. Literature studies and corresponding found relationships between emissions in ton CO_2 ha^{-1} yr^{-1} and water table depth below surface level in meters.

Study	Found relationship
Boonman et al. (2022)	NECB = 33.5 WTDe
Fritz et al. (2017)	NECB = 45.0 WTDe - 0.07
Tiemeyer et al. (2020)	NECB = -3.4 + 40.4 $e^{-7.5e^{-13WTDe}}$
Evans et al. (2021)	NECB = 34.0 WTDe - 6.2
Jurasinski et al. (2016)	NECB = 40.8 WTDe
Aben et al. (2024)	NECB = 32.9 WTDe - 0.8
Kruijt et al. (2023)	NECB = 78.5 WTDe - 15.3
SHAP results (current)	NEE = 46.6 WTDe - 16.0
SHAP results (current)	NEE = -51.4(WTDe - 0.05)^2 + 77.8(WTDe - 0.05) - 15.8

Appendix H: Shannon Entropy

Table H1. Results of Shannon entropy in our datasets for all regions combined.

	$H(CO_2flx)$ in training data	$H(CO_2flx)$ in test data (obs)	$H(CO_2flx)$ in modelled data	A_H
Airborne	4.657	5.165	5.506	0.127
Tower	4.055	4.115	5.198	-0.024
Merged	4.199	4.270	5.282	0.019

$$A_H = 1 - \frac{H(CO_{2mod})}{H(CO_{2obs})} \tag{H1}$$

Where H is Shannon Entropy, CO_2 mod and obs are modelled and observed CO_2 fluxes. The best model perfectly reflects the entropy in the observed data, such that $A_H = 0$.

processed the airborne dataset (WCS); LB, BK and JB processed the tower dataset. WF set up the initial footprint analysis script and managed the field-related software. Further processing and analysis was done by LvdP. WJ was responsible for all EC measurement sites, supported by JB. AR carried out a sub-study on Friesland data. RB is responsible for the OWASIS information product. LvdP led the article-writing

575   and AB, YvdV, RH, LB and RB contributed to revisions.

Competing interests.   The contact author has declared that neither of the authors has any competing interests.

References

585   Aben, R. C. H., Craats, D. V. D., Boonman, J., Peeters, S. H., Vriend, B., Boonman, C. C. F., Velde, Y. V. D., Erkens, G., and Berg, M. V. D.: Using automated transparent chambers to quantify CO 2 emissions and potential emission reduction by water infiltration systems in drained coastal peatlands in the Netherlands, EGUsphere, 403, https://doi.org/10.5194/egusphere-2024-403, 2024.

Arets, E. J. M. M., Kolk, J. W. H. V. D., Hengeveld, G. M., Lesschen, J. P., Kramer, H., Kuikman, P. J., and Schelhaas, M. J.: Greenhouse gas reporting for the LULUCF sector in the Netherlands, 2021.

590   Arora, B., Wainwright, H. M., Dwivedi, D., Vaughn, L. J., Curtis, J. B., Torn, M. S., Dafflon, B., and Hubbard, S. S.: Evaluating temporal controls on greenhouse gas (GHG) fluxes in an Arctic tundra environment: An entropy-based approach, Science of the Total Environment, 649, 284–299, https://doi.org/10.1016/j.scitotenv.2018.08.251, 2019.

Boonman, J., Hefting, M. M., Huissteden, C. J. V., Berg, M. V. D., Huissteden, J. V., Erkens, G., Melman, R., and Velde, Y. V. D.: Cutting peatland CO2 emissions with water management practices, Biogeosciences, 19, 5707–5727, https://doi.org/10.5194/bg-19-5707-2022, 595   2022.

Buzacott, A. J., Kruijt, B., Bataille, L., van Giersbergen, Q., Heuts, T. S., Fritz, C., Nouta, R., Erkens, G., Boonman, J., van den Berg, M., van Huissteden, J., and van der Velde, Y.: Drivers and Annual Totals of Methane Emissions From Dutch Peatlands, Global Change Biology, 30, https://doi.org/10.1111/gcb.17590, 2024.

Campbell, D. I., Glover-Clark, G. L., Goodrich, J. P., Morcom, C. P., Schipper, L. A., and Wall, A. M.: Large differences in CO2 emissions 600   from two dairy farms on a drained peatland driven by contrasting respiration rates during seasonal dry conditions, Science of the Total Environment, 760, https://doi.org/10.1016/j.scitotenv.2020.143410, 2021.

Chandrashekar, G. and Sahin, F.: A survey on feature selection methods, Comput. Electr. Eng., 40, 16–28, https://doi.org/10.1016/j.compeleceng.2013.11.024, 2014.

Cohrs, K. H., Varando, G., Carvalhais, N., Reichstein, M., and Camps-Valls, G.: Causal hybrid modeling with double machine learn- 605   ing—applications in carbon flux modeling, Machine Learning: Science and Technology, 5, https://doi.org/10.1088/2632-2153/ad5a60, 2024.

Dou, X., Yang, Y., and Luo, J.: Estimating Forest Carbon Fluxes Using Machine Learning Techniques Based on Eddy Covariance Measurements, Sustainability, 10, https://doi.org/10.3390/su10010203, 2018.

Erkens, G., Meulen, M. J. V. D., and Middelkoop, H.: Double trouble: Subsidence and CO2 respiration due to 1,000 years of Dutch coastal 610   peatlands cultivation, Hydrogeology Journal, 24, 551–568, https://doi.org/10.1007/s10040-016-1380-4, 2016.

Evans, C. D., Peacock, M., Baird, A. J., Artz, R. R., Burden, A., Callaghan, N., Chapman, P. J., Cooper, H. M., Coyle, M., Craig, E., Cumming, A., Dixon, S., Gauci, V., Grayson, R. P., Helfter, C., Heppell, C. M., Holden, J., Jones, D. L., Kaduk, J., Levy, P., Matthews, R., McNamara, N. P., Misselbrook, T., Oakley, S., Page, S. E., Rayment, M., Ridley, L. M., Stanley, K. M., Williamson, J. L., Worrall, F., and Morrison, R.: Overriding water table control on managed peatland greenhouse gas emissions, Nature, 593, 548–552, https://doi.org/10.1038/s41586- 615   021-03523-1, 2021.

Farahani, M. A. and Goodwell, A. E.: Causal Drivers of Land-Atmosphere Carbon Fluxes From Machine Learning Models and Data, Journal of Geophysical Research: Biogeosciences, 129, https://doi.org/10.1029/2023JG007815, 2024.

Foken, T., Göockede, M., Mauder, M., Mahrt, L., Amiro, B., and Munger, W.: Post-field data quality control, in: Handbook of micrometeorology: a guide for surface flux measurement and analysis, pp. 181–208, Springer, 2004.

620 Friedrich, S., Gerner, A., Tarantik, M., Chiogna, G., and Disse, M.: Assessment of rewetting scenarios under varying climate conditions in a partially restored raised bog in Bavaria, Germany, Journal of Hydrology: Regional Studies, 52, 101 695, https://doi.org/https://doi.org/10.1016/j.ejrh.2024.101695, 2024.

Fritz, C., Geurts, J., Weideveld, S., Temmink, R., Bosma, N., Wichern, F., Smolders, A., and Lamers, L.: Meten is weten bij bodemdalingmitigatie. Effect van peilbeheer en teeltkeuze op CO2-emissies en veenoxidatie, Bodem, p. 20, 2017.

625 Frolking, S. and Roulet, N. T.: Holocene radiative forcing impact of northern peatland carbon accumulation and methane emissions, Global Change Biology, 13, 1079–1088, https://doi.org/10.1111/j.1365-2486.2007.01339.x, 2007.

Hannun, R. A., Wolfe, G. M., Kawa, S. R., Hanisco, T. F., Newman, P. A., Alfieri, J. G., Barrick, J., Clark, K. L., DiGangi, J. P., Diskinm Glenn S, King, J., Kustas, W. P., Mitra, B., Noormets, A., Nowak, J. B., Thornhill, K. L., and Vargas, R.: Spatial heterogeneity in CO2, CH4, and energy fluxes: insights from airborne eddy covariance measurements over the Mid-Atlantic region, Environ. Res. Lett., 630 https://doi.org/10.1088/1748-9326/ab7391, 2020.

Huete, A., Didan, K., Miura, T., Rodriguez, E. P., Gao, X., and Ferreira, L. G.: Overview of the radiometric and biophysical performance of the MODIS vegetation indices, Remote Sensing of Environment, 83, 195–213, www.elsevier.com/locate/rse, 2002.

Hutjes, R. W. A., Vellinga, O. S., Gioli, B., and Miglietta, F.: Dis-aggregation of airborne flux measurements using footprint analysis, Agric. For. Meteorol., 150, 966–983, https://doi.org/10.1016/j.agrformet.2010.03.004, 2010.

635 Hydrologic: OWASIS: essentiële informatie voor het waterbeheer nu beschikbaar, https://www.hydrologic.nl/owasis-essentiele-informatie-voor-het-waterbeheer/, 2019.

Jung, M., Reichstein, M., Margolis, H. A., Cescatti, A., Richardson, A. D., Arain, M. A., Arneth, A., Bernhofer, C., Bonal, D., Chen, J., Gianelle, D., Gobron, N., Kiely, G., Kutsch, W., Lasslop, G., Law, B. E., Lindroth, A., Merbold, L., Vaccari, F., Williams, C., and Ur, A. W.: Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, 640 and meteorological observations, J. Geophys. Res., 116, https://doi.org/10.1029/2010JG001566, 2011.

Jurasinski, G., Günther, A., Huth, V., Couwenberg, J., and Glatzel, S.: Greenhouse gas emissions, Schweizerbart Science Publishers, ISBN 9783510652839, "http://www.schweizerbart.de//publications/detail/isbn/9783510652839/Wichtmann_Schroder_Joosten_Paludicu",, 2016.

Klimaatakkoord: Landbouw en landgebruik | Klimaatakkoord, https://www.klimaatakkoord.nl/landbouw-en-landgebruik, 2019.

645 Kljun, N., Calanca, P., Rotach, M. W., and Schmid, H. P.: A simple two-dimensional parameterisation for Flux Footprint Prediction (FFP), Geosci. Model Dev., 8, 3695–3713, https://doi.org/10.5194/GMD-8-3695-2015, 2015.

KNMI: KNMI Climate Data, https://www.knmi.nl/klimaat, 2024.

Koch, J., Elsgaard, L., Greve, M. H., Gyldenkærne, S., Hermansen, C., Levin, G., Wu, S., and Stisen, S.: Water-table-driven greenhouse gas emission estimates guide peatland restoration at national scale, Biogeosciences, 20, 2387–2403, https://doi.org/10.5194/bg-20-2387-2023, 650 2023.

Kruijt, B., Buzacott, A., van Giersbergen, Q., Bataille, L., Biermann, J., Berghuis, H., Heuts, T., Jans, W., Nouta, R., and van Huissteden Hefting F Hoogland, J. M.: CO2 emissions from peatlands in The Netherlands: drivers of variability in Eddy covariance fluxes, 2023.

Kwakernaak, C., van den Akker, J., Veenendaal, E., van Huissteden, K., and Kroon, P.: Mogelijkheden voor mitigatie en adaptatie Veenweiden en klimaat, Bodem, pp. 6–8, 2010.

655 Leifeld, J. and Menichetti, L.: The underappreciated potential of peatlands in global climate change mitigation strategies, Nature Communications, 9, https://doi.org/10.1038/s41467-018-03406-6, 2018.

Ligtenberg, J., W., B., A. van, D., D., K., M., P., J., V., and van der Vliet, N.: Validatie en toetsing LHM 4.1 Samenvattend hoofdrapport, PR4313.10, 2021.

Liu, L., Zhou, W., Guan, K., Peng, B., Xu, S., Tang, J., Zhu, Q., Till, J., Jia, X., Jiang, C., Wang, S., Qin, Z., Kong, H., Grant, R., Mezbahuddin,

660   S., Kumar, V., and Jin, Z.: Knowledge-guided machine learning can improve carbon cycle quantification in agroecosystems, Nature Communications, 15, https://doi.org/10.1038/s41467-023-43860-5, 2024.

Loisel, J., Gallego-Sala, A. V., Amesbury, M. J., Magnan, G., Anshari, G., Beilman, D. W., Benavides, J. C., Blewett, J., Camill, P., Charman, D. J., Chawchai, S., Hedgpeth, A., Kleinen, T., Korhola, A., Large, D., Mansilla, C. A., Müller, J., van Bellen, S., West, J. B., Yu, Z., Bubier, J. L., Garneau, M., Moore, T., Sannel, A. B., Page, S., Väliranta, M., Bechtold, M., Brovkin, V., Cole, L. E., Chanton, J. P., Christensen,

665   T. R., Davies, M. A., De Vleeschouwer, F., Finkelstein, S. A., Frolking, S., Gałka, M., Gandois, L., Girkin, N., Harris, L. I., Heinemeyer, A., Hoyt, A. M., Jones, M. C., Joos, F., Juutinen, S., Kaiser, K., Lacourse, T., Lamentowicz, M., Larmola, T., Leifeld, J., Lohila, A., Milner, A. M., Minkkinen, K., Moss, P., Naafs, B. D., Nichols, J., O'Donnell, J., Payne, R., Philben, M., Piilo, S., Quillet, A., Ratnayake, A. S., Roland, T. P., Sjögersten, S., Sonnentag, O., Swindles, G. T., Swinnen, W., Talbot, J., Treat, C., Valach, A. C., and Wu, J.: Expert assessment of future vulnerability of the global peatland carbon sink, Nat. Clim. Chang., 11, 70–77, https://doi.org/10.1038/s41558-020-

670   00944-0, 2021.

Lundberg, S. M. and Lee, S.-I.: A unified approach to interpreting model predictions, Advances in neural information processing systems, 30, 2017.

Metzger, S., Junkermann, W., Mauder, M., Butterbach-Bahl, K., Trancón Y Widemann, B., Neidl, F., Schäfer, K., Wieneke, S., Zheng, X. H., Schmid, H. P., and Foken, T.: Spatially explicit regionalization of airborne flux measurements using environmental response functions,

675   Biogeosciences, 10, 2193–2217, https://doi.org/10.5194/bg-10-2193-2013, 2013.

NHI: Netherlands Hydrological Instrumentarium Models, https://nhi.nu/modellen/, accessed: 2025-01-20, 2025.

Nielsen, D.: Tree Boosting With XGBoost, Ph.D. thesis, Norwegian University of Science and Technology, ISSN 17585880, https://doi.org/10.1111/j.1758-5899.2011.00096.x, 2016.

Nijman, T. P., van Giersbergen, Q., Heuts, T. S., Nouta, R., Boonman, C. C., Velthuis, M., Kruijt, B., Aben, R. C., and Fritz, C.:

680   Drainage effects on carbon budgets of degraded peatlands in the north of the Netherlands, Science of the Total Environment, 935, https://doi.org/10.1016/j.scitotenv.2024.172882, 2024.

Ogunleye, A. and Wang, Q.-G.: XGBoost model for chronic kidney disease diagnosis, IEEE/ACM transactions on computational biology and bioinformatics, 17, 2131–2140, 2019.

Paleri, S., Desai, A. R., Metzger, S., Durden, D., Butterworth, B. J., Mauder, M., Kohnert, K., and Serafimovich, A.: Space-Scale Re-

685   solved Surface Fluxes Across a Heterogeneous, Mid-Latitude Forested Landscape, Journal of Geophysical Research: Atmospheres, 127, https://doi.org/10.1029/2022JD037138, 2022.

Prabha, A., Yadav, J., Rani, A., and Singh, V.: Design of intelligent diabetes mellitus detection system using hybrid feature selection based XGBoost classifier, Comput. Biol. Med., 136, 104 664, https://doi.org/10.1016/j.compbiomed.2021.104664, 2021.

Rodríguez-Pérez, R. and Bajorath, J.: Interpretation of machine learning models using shapley values: application to compound potency and

690   multi-target activity predictions, J. Comput. Aided. Mol. Des., 34, 1013–1026, https://doi.org/10.1007/s10822-020-00314-0, 2020.

Ruyssenaars, P., van der Net, L., Coenen, P., Rienstra, J., Zijlema, P., Arets, E., Baas, K., Dröge, R., Geilenkirchen, G., t Hoen, M., et al.: Greenhouse gas emissions in the Netherlands 1990–2020, Rijksinstituut voor Volksgezondheid en Milieu RIVM, 2022.

Sang, X., Xiao, W., Zheng, H., Yang, Y., and Liu, T.: HMMPred : Accurate Prediction of DNA-Binding Proteins Based on HMM Profiles and XGBoost Feature Selection, Computational and Mathematical Methods in Medicine, 2020, 2020.

695 Serafimovich, A., Metzger, S., Hartmann, J., Kohnert, K., Zona, D., and Sachs, T.: Upscaling surface energy fluxes over the North Slope of Alaska using airborne eddy-covariance measurements and environmental response functions, Atmos. Chem. Phys., 18, 10 007–10 023, https://doi.org/10.5194/ACP-18-10007-2018, 2018.

Sun, Y., Jia, L., Chen, Q., and Zheng, C.: Optimizing window length for turbulent heat flux calculations from airborne eddy covariance measurements under near neutral to unstable atmospheric stability conditions, Remote Sensing, 10, https://doi.org/10.3390/rs10050670,
700   2018.

Sun, Y., Jia, L., Chen, Q., Lin, X., Sude, B., Quan, Z., and Hutjes, R. W.: Construction of a spatially gridded heat flux map based on airborne flux Measurements using remote sensing and machine learning methods, Agricultural and Forest Meteorology, 334, https://doi.org/10.1016/j.agrformet.2023.109424, 2023.

Tiemeyer, B., Freibauer, A., Borraz, E. A., Augustin, J., Bechtold, M., Beetz, S., Beyer, C., Ebli, M., Eickenscheidt, T., Fiedler, S., Förster,
705   C., Gensior, A., Giebels, M., Glatzel, S., Heinichen, J., Hoffmann, M., Höper, H., Jurasinski, G., Laggner, A., Leiber-Sauheitl, K., Peichl-Brak, M., and Drösler, M.: A new methodology for organic soils in national greenhouse gas inventories: Data synthesis, derivation and application, Ecol. Indic., 109, 105 838, https://doi.org/10.1016/j.ecolind.2019.105838, 2020.

van Asselen, S., Erkens, G., and de Graaf, F.: Monitoring shallow subsidence in cultivated peatlands, Proceedings of the International Association of Hydrological Sciences, 382, 189–194, 2020.

710 van den Akker, J., Kuikman, P., de Vries, F., Hoving, I., Pleijter, M., Hendriks, R., Wolleswinkel, R., Simões, R., and Kwakernaak, C.: Emission of CO2 from Agriculutral Peat Soils in the Netherlands and Ways to Limit this Emission, Proc. 13th Int. Peat Congr. After Wise Use–The Futur. Peatlands, 1, 61–64, 2008.

Vaughan, A. R., Lee, J. D., Metzger, S., Durden, D., Lewis, A. C., Shaw, M. D., Drysdale, W. S., Purvis, R. M., Davison, B., and Nicholas Hewitt, C.: Spatially and temporally resolved measurements of NOxfluxes by airborne eddy covariance over Greater London, Atmos.
715   Chem. Phys., 21, 15 283–15 298, https://doi.org/10.5194/acp-21-15283-2021, 2021.

Vellinga, O. S., Dobosy, R. J., Dumas, E. J., Gioli, B., Elbers, J. A., and Hutjes, R. W.: Calibration and Quality Assurance of Flux Observations from a Small Research Aircraft, J. Atmos. Ocean. Technol., 30, 161–181, https://doi.org/10.1175/JTECH-D-11-00138.1, 2013.

Wang, Y. and Ni, X. S.: A XGBoost risk model via feature selection and Bayesian hyper-parameter optimization, arXiv Prepr. arXiv1901.08433, 2019.

720 Yu, Z., Loisel, J., Brosseau, D. P., Beilman, D. W., and Hunt, S. J.: Global peatland dynamics since the Last Glacial Maximum, Geophysical research letters, 37, 2010.

Yuan, K., Zhu, Q., Li, F., Riley, W. J., Torn, M., Chu, H., McNicol, G., Chen, M., Knox, S., Delwiche, K., Wu, H., Baldocchi, D., Ma, H., Desai, A. R., Chen, J., Sachs, T., Ueyama, M., Sonnentag, O., Helbig, M., Tuittila, E. S., Jurasinski, G., Koebsch, F., Campbell, D., Schmid, H. P., Lohila, A., Goeckede, M., Nilsson, M. B., Friborg, T., Jansen, J., Zona, D., Euskirchen, E., Ward, E. J., Bohrer, G., Jin,
725   Z., Liu, L., Iwata, H., Goodrich, J., and Jackson, R.: Causality guided machine learning model on wetland CH4 emissions across global wetlands, Agricultural and Forest Meteorology, 324, https://doi.org/10.1016/j.agrformet.2022.109115, 2022.

Zhou, Y., Sachs, T., Li, Z., Pang, Y., Xu, J., Kalhori, A., Wille, C., Peng, X., Fu, X., Wu, Y., and Wu, L.: Long-term effects of rewetting and drought on GPP in a temperate peatland based on satellite remote sensing data, Science of the Total Environment, 882, https://doi.org/10.1016/j.scitotenv.2023.163395, 2023.

730 Zulueta, R. C., Oechel, W. C., Loescher, H. W., Lawrence, W. T., and Paw U, K. T.: Aircraft-derived regional scale CO2 fluxes from vegetated drained thaw-lake basins and interstitial tundra on the Arctic Coastal Plain of Alaska, Glob. Chang. Biol., 17, 2781–2802, https://doi.org/10.1111/j.1365-2486.2011.02433.x, 2011.