Meta-modelling of carbon fluxes from crop and grassland multi-model outputs

Roland Hollós[1,2,3], Nándor Zrinyi[2,4], Zoltán Barcza[2,3], Gianni Bellocchi[5], Renáta Sándor[1], János Ruff[6], Nándor Fodor[1,7]

[1] Centre for Agricultural Research HUN-REN, Agricultural Institute, Martonvásár, 2462, Hungary
[2] Eötvös Loránd University, Department of Meteorology, Budapest, 1117, Hungary
[3] Czech Academy of Sciences, Global Change Research Institute, Brno, 603 00, Czech Republic
[4] Eötvös Loránd University, Doctoral School of Earth Sciences, Budapest, 1117, Hungary
[5] INRAE, VetAgro Sup, Unité Mixte de Recherche sur l'Ecosystème Prairial (UREP), Clermont-Ferrand, 63000, France
[6] University of Pécs, Institute of Mathematics and Informatics, Pécs, 7624, Hungary
[7] University of Debrecen, Faculty of Agricultural and Food Sciences and Environmental Management, Debrecen, 4032, Hungary

* *Correspondence: Nándor Fodor (fodor.nandor@atk.hu)*

Abstract. We evaluated five stacking-based meta-models - Multiple Linear Regression, Random Forest, XGBoost, and also Random Forest and XGBoost with environmental covariates (RF+ and XGB+, respectively) - against the multi-model median (MMM) and best individual process-based models for gross primary production (GPP), ecosystem respiration (RECO) and net ecosystem exchange (NEE) at two cropland and two grassland sites. We tested two validation strategies for GPP and RECO (70%/30% training/validation approach and the time-aware leave-one-year-out (LOYO) method), and three strategies for NEE (70%/30%, complete LOYO and independent validation using the meta-model based RECO-GPP). All meta-models were associated with improved RMSE, bias and correlation. Based on the LOYO validation strategy, average correlation increase was ~2% for GPP (0.2-4.4%), 9% for RECO (5-13.6%) and 8% for NEE (0.5-12.5%). In the case of the independent validation strategy correlation increase was ~40% for NEE (24.5-64%). Bias was nearly eliminated except at one cropland site. Wilcoxon signed-rank tests confirmed that improvements were statistically significant for 72% of model-site-variable combinations, with particularly robust results for grassland sites where all meta-models significantly outperformed MMM, while one cropland site with limited data showed no significant improvements. SHapley Additive exPlanations (SHAP) analysis of XGB+ showed that diverse individual models, not always the top performers, contributed most, and that temperature - especially for RECO in croplands and NEE in grasslands - was the dominant environmental driver, while precipitation had minor effects. These findings highlight the predictive and diagnostic advantages of meta-modeling approaches over MMM, with potential applications across agroecosystem, Earth system and environmental model ensembles.

1 Introduction

Biogeochemical processes are central to agricultural planning, underpinning concepts such as "climate-smart agriculture", "low-carbon agriculture" and "greenhouse gas-mitigating farming practices" within narratives/storylines (e.g. Thornton et al., 2018; Hou and Hou, 2019; Anuga et al., 2020). As governments and societies reshape economies, particularly in the context of decarbonisation (e.g. Sroufe and Watts, 2022),

biogeochemical modelling has emerged as an essential tool to support agricultural policies (Bellocchi, 2023). These models simulate the complex interactions between agriculture and ecosystem services, such as C sequestration, biodiversity conservation and water quality regulation, thereby empowering policymakers to integrate ecosystem value into agricultural decision-making and land-use strategies (e.g. Lambin and Meyfroidt, 2011). Moreover, by simulating a wide variety of ecosystem processes, the models facilitate the assessment of agricultural emissions and mitigation strategies, forming a scientific basis for climate policy and agricultural resilience (Li et al., 2006; Valin et al., 2013; Sándor et al., 2018; Lembaid et al., 2021, 2022; Gascuel-Odoux et al., 2022).

Many state-of-the-art biogeochemical models have a long development history with some of them spanning 50 years (Hidy et al., 2022). Still, inherent uncertainties in model structures, parameterisations and assumptions pose challenges for reliable predictions under diverse environmental and management conditions (e.g. Riccio et al., 2007; Bellocchi et al., 2010; Therond et al., 2011; Harrison et al., 2012; Brilli et al., 2017; Bilotto et al., 2021). This calls for improved mathematical tools and innovative solutions to strengthen the trust in the models.

Ensemble modelling has gained prominence as a robust method to address these uncertainties in biogeochemical models (e.g. Challinor et al., 2013; Snow et al., 2014; Calanca et al., 2016; Jones et al., 2017; Knutti et al., 2019). Each model adopts unique assumptions about processes like soil nutrient cycling, photosynthesis, allocation and crop phenology, leading to significant variability in predictions under similar scenarios. Such structural uncertainties (e.g. divergent representations of water stress or heat tolerance), along with parameter uncertainty can lead to inconsistent predictions. Ensemble approaches mitigate these discrepancies by synthesising outputs from multiple models, compensating biases and highlighting consistent trends (e.g. Bassu et al., 2014; Rosenzweig et al., 2014; Kollas et al., 2015; Li et al., 2015; Ruane et al., 2016, 2017; Sándor et al., 2017, 2020; Ehrhardt et al., 2018; Wallach et al., 2018). This increases the reliability of predictions and reduces the risks associated with over-reliance on individual models.

Although ensemble techniques are proven to improve modelling accuracy and increase the trust in biogeochemical models, problems still exist. For example, the ensemble models are hard to explain, and currently only the simplest methods (averaging or median calculation; Sándor et al., 2016) are the most widely used for ecosystem models. These simple model combination algorithms usually have many implicit presumptions that models often cannot meet (for example, equal variance and independence). In the field of machine learning (ML), which generates models directly from data, instead of following physical, biological and biogeochemical principles, much more developed techniques exist, which can be potentially exploited to improve the predictive power of the ensemble output. These ”ML-based” techniques usually have less and more explicitly stated conditions (Scowen et al., 2023). As an advanced, potentially promising approach, ML based meta-modelling extends the concept of ensemble technique by synthesising outputs even across multiple calibration scenarios. Recent advancements in ML offer innovative tools to refine data-intensive modelling in Earth sciences (e.g. Li et al., 2015; Jackson et al., 2017; Keskin et al., 2019; Reichstein et al., 2019; Bai et al., 2021; Chandel et al., 2024; Wang et al., 2024). By using ensemble learning – a subset of ML methods – predictive frameworks can integrate diverse model outputs with improved accuracy and interpretability (e.g. Hansen and Salamon, 1990; Opitz and Maclin, 1999; Dietterich, 2000; Hagedorn al., 2005; Palmer, 2019). Stacking, as a flexible ensemble learning technique, combines outputs from heterogeneous models into a *meta-model*, assigning weights to highlight their relative importance (e.g. Breiman, 1996; Van der Laan et al., 2007; Sagi and Rokach, 2018). This approach has potential for biogeochemical modelling, offering simplicity and adaptability with Multiple Linear Regression (e.g. Kutner et al., 2005), and enhanced prediction

accuracy through decision-tree-based methods like Random Forest (e.g. Breiman, 2001a,b; Liaw and Wiener, 2002) and XGBoost (e.g. Chen and Guestrin, 2016).

The integration of diverse environmental drivers with sophisticated analytical methods is central to advancing predictive modelling in environmental science. For instance, Hengl et al. (2017) demonstrated that spatial predictions of soil properties improved significantly when large stacks of remote-sensing and environmental covariates were incorporated into ensemble machine-learning frameworks. This approach is further supported by research such as Luo et al. (2009), who found that high-frequency driver data reduces parameter equifinality in ecosystem data assimilation, and Pappas et al. (2014), who developed efficient methods for gap-filling hydrometeorological observations. More recently, Sándor et al. (2023) showed that analysing residual correlations among crop and grassland model ensembles can reveal structural model deficiencies and improve simulation of C-N fluxes when combined with ensemble averaging. Collectively, these studies highlight the substantial value of leveraging diverse variables, advanced techniques, and residual-based ensemble diagnostics to build more robust and accurate predictive models. Building on these advancements, the present study introduces a meta-modelling framework that integrates biogeochemical model outputs and environmental variables, guided by residual correlation insights, to address structural model error and enhance predictive performance.

This study aims to improve the predictive accuracy of biogeochemical flux calculations, particularly C fluxes, in diverse agricultural systems (crop rotations and grassland systems). Building on an already established multi-model framework, a novel methodology is used to develop a meta-model that balances scientific rigour and practical feasibility. The results of this meta-model are compared with the multi-model ensemble median and other meta-models to demonstrate the improvement, interpretability and reliability of ensemble predictions (where multi-model median is considered as the baseline). This dual focus aims to improve methodologies for sustainable crop and grassland management and to inform agricultural policy to better address the challenges of climate resilience and sustainability.

The novelty of this study lies in the interpretability of the multi-model system and its linkage with environmental factors to improve the performance of the method. Additionally, the application of multiple meta-modelling approaches provides an abstract framework for improved modelling exercises, further supported by SHapley Additive exPlanations (SHAP) to ensure interpretability.

2 Materials and methods

2.1 Meta-modelling framework

The methods employed in this study arise from the contemporary landscape of crop and grassland modelling, specifically focusing on ensemble modelling. We adopt and extend the concept of meta-modelling by comparing it to the multi-model median (MMM) baseline estimator (e.g. Sándor et al., 2018).

While the term *meta-modelling* can be conveyed ambiguously, some clarification is needed. Widely used in mathematics, computer science and engineering, a meta-model is understood as a model of one or more models. Here, we define it as the use of model outputs from a multi-model ensemble - for two particular methods supplemented by environmental variables - as inputs to a higher-level statistical model. Unlike traditional surrogate

models that approximate the structure of a process-based model, our meta-model approach complements, rather than replaces, process-based models by exploiting the information embedded in multiple model outputs. The enhanced predictive power and the resulting trust in model outcomes are particularly valuable for informing policy decisions. This method goes beyond Bayesian model averaging, non-democratic model selection and some machine learning based output combination techniques, allowing for more flexibility in capturing nonlinear relationships.

Here we adopt *stacked generalization* (called simply as *stacking* hereinafter), as an ensemble method wherein a meta-model integrates predictions generated by multiple base models. Thus, throughout the study, stacking refers to an ensemble learning technique in which the multi-model framework is used to construct a new estimation for the target variables. This method relies on the observation that diverse models inherently capture different aspects of real-world systems like the plant-soil system. Consequently, training a meta-model to optimally integrate these varied predictions allows stacking to leverage the individual strengths of the base models, ultimately leading to improved generalisation across various tasks focusing on the plant-soil systems.

Linear models are the most prevalent meta-models in ensemble learning, largely due to their simplicity, computational efficiency, and ease of interpretation. Their additive structure makes it straightforward to quantify the relative influence of individual models within an ensemble, which is particularly valuable for diagnostic purposes. Notably, a precursor to formal meta-modelling can be seen in the Granger–Ramanathan averaging method (Granger and Ramanathan, 1984; Nand et al., 2025), which introduced the idea of optimally weighting model outputs through constrained linear regression. This approach laid the foundation for modern stacking techniques by demonstrating how combining forecasts in a regression framework could systematically improve predictive accuracy.

However, when base models exhibit substantial structural differences and higher structural errors, the relationship between predicted and observed outcomes may no longer be well-approximated by a linear model. In such scenarios, simple linear approaches like the Granger–Ramanathan averaging may prove insufficient. Furthermore, the often-implicit assumption of conditional independence among base-model predictions in linear meta-modelling is frequently violated, rendering multiple linear regression a suboptimal choice.

To address this limitation, ensemble-based machine-learning meta-models offer a compelling alternative. These models can effectively capture complex, non-linear relationships and do not rely on the conditional independence of base model predictions. Random Forests, for example, have been among the earliest and most effective ensemble methods used as meta-models, particularly in contexts where linear approaches underperform (Zhao and Cheng, 2022).

While effective, Random Forests can be computationally demanding during both training and inference. They also tend to overfit the training data, potentially compromising generalization on independent data. XGBoost offers greater flexibility, often delivering superior predictive accuracy with generally lower computational requirements. However, XGBoost typically requires extensive hyperparameter tuning to achieve optimal performance.

A key limitation of all these approaches is their static nature: the functional relationship learned between the base-models' outputs and the target variable remains fixed. This static feature limits the long-term usability of meta-models, as evolving environmental conditions can shift the relevance of different structural components within the base-models. Consequently, the optimal meta-model may vary over time, necessitating continuous retraining to maintain accuracy, which represents a major drawback.

With advancements in machine learning, today we can significantly extend the approach described above. By incorporating differentiating environmental factors — such as temperature or precipitation — as additional features, the meta-model can learn not only the functional relationship between the base-model outputs and the measurements, but also how this relationship varies with environmental conditions.

This approach is only possible with flexible meta-models. Simpler models, such as generalized linear models, rely on assumptions like conditional independence and uncorrelated features — assumptions that are often violated when additional environmental factors are introduced. As a result, these simpler models may struggle to capture the conditional dependencies required for enhanced predictive performance.

Such an *environment-aware meta-modelling* framework (i.e. XGBoost or Random Forest with additional meteorological drivers) improves both predictive accuracy and interpretability by revealing the specific conditions under which different models succeed or fail. Examining these meta-models allow researchers and modellers to understand their model's limitations in varying contexts. This knowledge guides further model development and facilitates the selection of the most appropriate model for given circumstances, promoting valuable synergy between modelling approaches.

To establish a clear terminological framework, here we summarize the key components of the study. *Base models* are the individual process-based ecosystem models (e.g. DNDC or APSIM; Sándor et al., 2024) that generate the initial flux estimates. *Baseline* (MMM) is the Multi-Model Median, defined as the unweighted central tendency of the base model outputs and used as the benchmark for evaluating all subsequent meta-models. *Stacking (Meta-Modelling framework)* is the ensemble-learning framework in which the outputs of the base models serve as predictors for a second-level learner. *Meta-models* are ML algorithms that implement this stacking framework.

2.2 Source of the model ensemble

The model ensemble is based on data from international initiatives, primarily the Agricultural Model Intercomparison and Improvement Project (https://agmip.org) and the Integrative Research Group of the Global Research Alliance on agricultural GHGs (https://globalresearchalliance.org/research/integrative). These exercises have shown that ensembles of process-based biogeochemical models can reliably estimate agricultural productivity, as well as C and N emissions (and stocks) of agricultural systems (Ehrhardt et al., 2018; Sándor et al., 2018, 2020, 2024). These studies, mostly funded by national agencies, also contribute to the assessment of C storage potential (e.g. Farina et al., 2021), aligning with the '4 per mille Soils for Food Security and Climate' initiative established at the 2015 United Nations Climate Change Conference (COP21) by the French Ministry of Agriculture (https://www.4p1000.org).

Here, we refer to a multi-model scheme, using daily model outputs as inputs for the meta-models. We are revisiting salient elements from previously published studies to build the framework for our meta-modelling analysis. Specifically, we delve into comparisons with the multi-model median from the study by Sándor et al. (2020) on C fluxes, which, in turn, was based on the protocol of Ehrhardt et al. (2018).

We used daily outputs from 23 crop and grassland simulation models/versions (Table 1; for details see Sándor et al., 2020; 2024). Model names (M01–M28) were anonymised for consistency, with M11 excluded due to missing C fluxes. These models were applied to four long-term field sites: two grazed grasslands (G3, G4) and two croplands (C1, C2) across the UK, France (two sites) and Canada (Table 2). The original Sándor et al. (2020) paper included

one additional cropland site in India, but this site was excluded from the present study due to its relatively poor temporal coverage and because the observations were made using a different methodology, resulting in the absence of GPP and RECO estimates. All other sites relied on the eddy-covariance technique, ensuring methodological consistency.

Table 1: C-flux outputs provided by different models (Sándor et al., 2024), denoted by symbols such as ✓ (present) and NA (not available). Models are marked as M01-M26, and they are kept anonymous as in the Sándor et al. (2020) study. In cropland sites, we had gross primary production (GPP) from six models, net ecosystem exchange (NEE) from seven models and ecosystem respiration (RECO) from 15 models. At grassland sites, we had GPP from 10 models, NEE from 10 models and RECO from 12 models.

Model type		Crop models											Grassland models								Both systems			
Model id		M01	M02	M04	M09	M12	M13	M18	M19	M20	M25	M26	M03	M06	M16	M21	M22	M23	M24	M28	M05	M07	M08	M14
Outputs	GPP	✓	NA	NA	✓	NA	NA	NA	✓	NA	NA	NA	NA	✓	✓	✓	✓	✓	✓	✓	✓	✓	NA	✓
	RECO	✓	✓	✓	✓	✓	✓	✓	✓	NA	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	NEE	✓	NA	NA	✓	NA	NA	NA	✓	✓	NA	NA	NA	✓	✓	✓	✓	✓	✓	✓	✓	✓	NA	✓

Table 2: Cropland and grassland sites, and temporal coverage of available data used for the analysis. Different crop rotations were used in the cropland sites, including cereals (spring and winter wheat [W], triticale [T], maize [M] and rice [R]), legumes (soybean [S]), rapeseeds (canola and mustard [C]), borages (phacelia, P) and fallow intercropping periods [I].

Sites, country (latitude, longitude, altitude)	Years of available data (simulation period)	Land use	Annual mean temperature (°C)	Annual mean precipitation (mm)
C1: Ottawa, Canada (45.29, -75.77, 94 m a.s.l.)	2007-2012	W/S/C/M/W/C	7.2±0.9	936±121
C2: Grignon, France (48.85, 1.95, 125 m a.s.l.)	2008-2012	C/M/W/T/P/M /W/I	10.9±0.7	571±35
G3: Laqueuille, France (45.64, 2.74, 1040 m a.s.l.)	2003-2012	Permanent grassland (cattle grazing)	13.7±0.4	910±96
G4: Easter Bush, UK (55.52, -3.33, 190 m a.s.l.)	2002-2010	Permanent grassland (ewe grazing)	7.8±0.8	1078±205

High quality data covering climate, soil, agricultural practices and C fluxes were gathered from Sándor et al. (2024). Observations at these sites include e measurements of net ecosystem CO_2 exchange (NEE) data, further divided into two main fluxes (e.g. Reichstein et al., 2005; Raj et al., 2016): gross primary production (GPP), representing photosynthetic production from atmospheric CO_2, and total ecosystem respiration (RECO), encompassing the total C respired by plants, soil organisms and, in the case of grasslands, grazing animals.

Initialisation and calibration procedures aligned models with vegetation, soil and atmospheric fluxes from the study sites, following the protocol described in Ehrhardt et al. (2018). This comprehensive exercise involved a multi-stage approach, granting modellers access to increasingly detailed data for running and evaluating their models, progressing from uncalibrated to fully calibrated simulations. In summary, the calibration included five ascending

levels, incorporating additional data for refinement: blind test (S1), utilising only site specific weather and management data for the simulation periods; initialisation (S2), incorporating additional historical climate and management data, extending to years preceding simulation periods, along with regional productivity for initialisation purposes; partial calibration (S3), including biomass production and phenology data; intermediate calibration (S4), integrating soil temperature, soil water content and mineral N data; full calibration (S5), using N_2O emissions, NEE, GPP, RECO and soil organic C and N data.

For the purposes of this study, emphasis was placed on the outputs from the partial calibration stage (S3) stage. S3 involves calibration using plant data exclusively, which enhances the practical implementability of models for end users and beneficiaries. This approach recognises the importance of ensuring models are not only scientifically robust but also practically useful. Its validity is supported by findings that additional calibration stages beyond S3 yielded only modest improvements (Sándor et al., 2023).

2.3 Meta-modelling approaches and validation strategies

The meta-model construction was done at the daily time resolution for GPP, RECO and NEE. Two main approaches were tested in this study for the meta-model construction. The first one focuses on the classic stacking ensemble method (referred here as CSEM), where the meta-models were constructed using multiple linear regression (MLR), random forest (RF) method, and XGBoost (XGB) using the model outputs described in Section 2.2. The second approach, that is introduced in this study as a *novel method* (the environment-aware meta-modelling method), uses the same stacking methods but extends the input dataset so that besides the individual, process-based model outputs, we use two additional environmental factors (daily mean temperature and precipitation) to predict the final outcome (i.e. the meta-model). This approach (i.e. inclusion of the meteorological drivers) is referred to here as extended generic stacking ensemble method (XGB+ and RF+). The inclusion of only temperature and precipitation is grounded in their central roles as primary climatic drivers of terrestrial C fluxes. Temperature governs essential biological processes such as photosynthesis and respiration (Lloyd and Taylor, 1994; Xu and Baldocchi, 2004), while precipitation affects soil moisture and plant water stress, both critical for GPP and RECO (Reichstein et al., 2005; Schwalm et al., 2010). Extensive research has shown that fluctuations in these two variables explain much of the interannual and seasonal variability in ecosystem C exchange (Jung et al., 2007; Beer et al., 2010). Moreover, temperature and precipitation are consistently measured across sites, widely available and commonly integrated into ecological and climate models. Their inclusion enhances predictive performance without adding complexity or compromising generalisability (e.g. Hijmans et al., 2005). Supporting this approach, Dorman et al. (2013) advocate for the use of core climatic variables over an excess of redundant predictors to maintain model robustness and transferability.

Two meta-model validation *strategies* were used in the study for both main approaches described above. First, all meta-models were constructed using 70% of the observations for training, with the remaining 30% used for validation. Additionally, we implemented the complete leave-one-year-out (LOYO) approach, in which the observations from a complete year were omitted during training and subsequently used for validation. All years were omitted consecutively, and the validation results based on the omitted years were averaged and provided the basis for the evaluation of the strategy. This alternative strategy was included since carbon-flux time series are autocorrelated, and time-aware validation methods like LOYO can provide additional insight into model

applicability. The 70%/30% split (referred as 70/30 hereinafter) evaluates within-regime reconstruction performance, while LOYO assesses the model's ability to generalize and extrapolate beyond the conditions represented in the training test. In the study the LOYO based methods are presented mainly given the fact that they are more generalizable. Statistical analysis is also provided for the 70%/30% split for comparability.

For the MLR method we used the *lm* function of base R. The RF method (*randomForest* package from CRAN; Liaw and Wiener, 2002) was used with 1000 trees. For each tree the number of predictors were 1/3 of the number of input data streams (GPP, RECO, NEE and meteorological drivers if applicable) to prevent overfitting. The subset of predictors was used randomly. The minimum size of the terminal nodes was five. Splitting the dataset into training and testing dataset was done randomly using the built-in random number generator of base R, without specifying a seed. For each site, the same site-specific random number sequence was used across all meta-model trainings, while sequences differed between sites due the varying lengths of their time series. The XGB method was implemented using the *XGBoost* package from CRAN (Chen and Guestrin, 2016). For hyperparameter optimization we used random search technique with 50 evaluations for XGBoost (the details of the procedure can be read in Appendix B). Figure 1 shows the overview of the stacking methods used in the study.

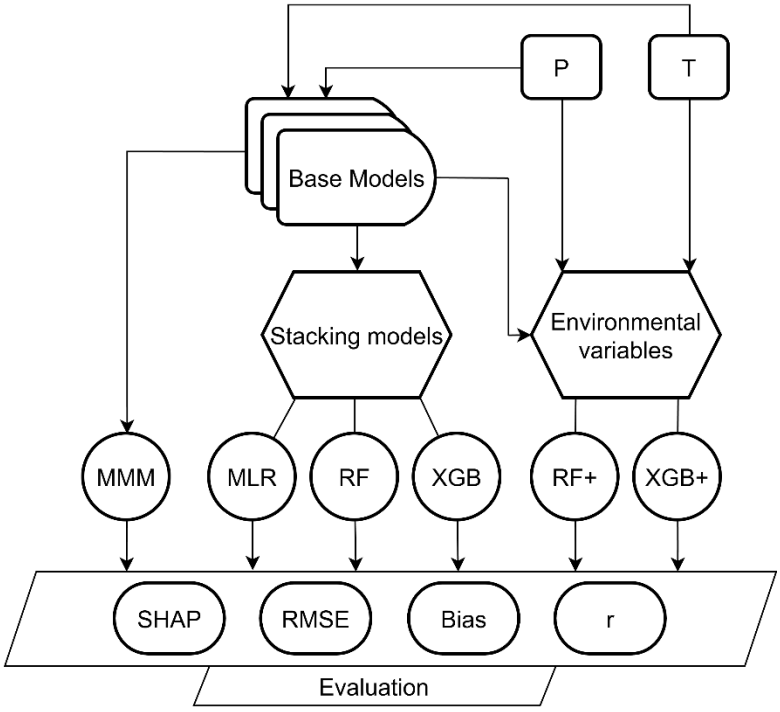


Figure 1: Schematic illustration of the applied stacking ensemble learning methodology. The workflow combines outputs from multiple simulations (base models) with multi-model median (MMM) and stacking models (MLR, RF, XGB). The XGB+ and RF+ models are extended versions that incorporate additional input features: temperature (T) and precipitation (P). The performance of these models and the ensembles are evaluated using four metrics: SHAP (for interpretability), RMSE (root mean square error), Bias and the Pearson correlation coefficient (r). The final output is the site-level carbon flux, which is the same variable type as the initial model outputs.

For XGBoost+ feature contributions were quantified using SHAP values (Shapley, 1953; Lundberg and Lee, 2017). SHAP values are grounded in cooperative game theory, representing the expected marginal contribution of a feature

across all possible feature coalitions. This formulation provides a theoretically consistent and locally accurate decomposition of model predictions, enabling a granular understanding of feature effects. In contrast to traditional importance measures such as the mean decrease in impurity (often referred to as the Gini importance; Breiman, 2001a, b), SHAP values address several critical limitations. The Gini metric, while computationally efficient, is known to exhibit biases toward features with a higher number of categories or greater variance (Strobl et al., 2007). Moreover, it does not provide insight into the directionality or context-dependent contribution of features to individual predictions. SHAP overcomes these limitations by providing additive, model-agnostic attributions that remain consistent across different models and capture complex, non-linear interactions between features (Lundberg et al., 2020). These properties make SHAP particularly well-suited for interpreting ensemble methods like RF and gradient-boosted decision trees, where feature interactions and non-linearities are prominent.

All evaluations were performed separately for each study site, and for the studied carbon fluxes (GPP, RECO, NEE). No cross-site meta-model was developed in this study, meaning that all training and validation were performed on a site-specific temporal basis. To assess how well models approximated site-level observations, we compared each meta-model and the MMM using three common performance metrics (Richter et al., 2012) across 26 site-stage-output combinations: the root mean square error (RMSE), the Pearson correlation coefficient (r), and the bias.

The results are organized according to the carbon flux type. First GPP is discussed, as it represents plant photosynthesis and is the most important driver of the plant production and yield. Most likely GPP is the process that is the simplest to simulate by process-based models. This is because the mechanism of photosynthesis is well-discovered, and practically it is separated from other autotrophic or heterotrophic processes (since this is the only flux that comes into the ecosystem). GPP is a relatively large flux (compared e.g. to NEE or net biome production), so model optimization is perhaps simpler and more interpretable. GPP can be biased for some cases due to improper representation of phenology, and lack of model optimization.

RECO follows the discussion that is typically harder to simulate than GPP. This is because RECO is the sum of autotrophic and heterotrophic respiration, where both components have their own intrinsic uncertainties. For example, heterotrophic respiration is typically problematic due to the complexity of the decomposition processes in the soil including the rhizosphere and the presence, complexity and activity of the microbial/fungal life forms. RECO can be erroneously simulated if one of the two major components is misrepresented (or even worse if, due to the compensation of errors, RECO looks good due to wrong reasons). Nevertheless, RECO is still a larger carbon flux, so its representation can be expected to be relatively well-adjusted.

Finally, we present NEE related results. This flux is probably the hardest to simulate, since it is the sensitive balance between RECO and GPP (by definition, NEE=RECO-GPP). Capturing the magnitude and variability of NEE requires the proper representation of plant phenology (start and end of season), and accurate representation of GPP and RECO. If the meta-model captures NEE with greater precision than the more traditional model ensembles, this represents a significant accomplishment.

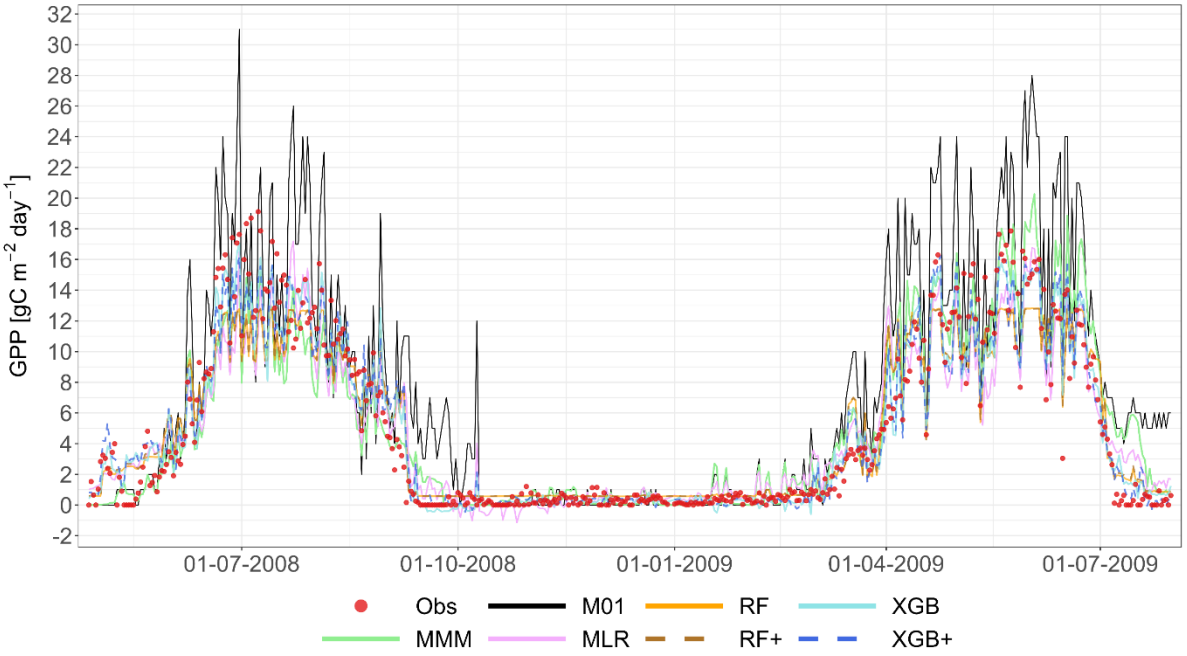
As the meta-model-based NEE is not calculated as RECO-GPP but rather it is modelled independently from GPP and RECO (due to methodological reasons), this introduces inconsistency. To address this issue, we also calculated NEE using the meta-model-based GPP and RECO, thus providing consistent results. This independent calculation is used as another level of validation for the LOYO based model construction. It is important to note that successful and unbiased simulation of NEE is a major step forward supporting e.g. atmospheric inversions or any other bottom-up carbon flux estimations.

To formally assess whether meta-models significantly outperformed the Multi-Model Median baseline, we conducted one-sided Wilcoxon signed-rank tests on paired daily squared residuals. For each site-variable-model combination, separate tests were performed for each Leave-One-Year-Out cross-validation fold, and p-values were aggregated using the median across test years to obtain a robust measure of typical performance. This approach avoids pseudoreplication that would arise from pooling all validation data together, while the median aggregation reduces sensitivity to occasional outlier years. The Wilcoxon test is particularly appropriate for this analysis as it makes no distributional assumptions about prediction errors, is robust to outliers and skewed distributions, and properly accounts for the paired nature of our comparisons (same observations predicted by different models). To address the multiple testing problem arising from testing 15 hypotheses per site (5 models × 3 variables), we applied the Benjamini-Hochberg procedure per site to control the false discovery rate at 5%.

3. Results

3.1. GPP

Fig. 2 shows a representative example of the observed and simulated time series of GPP at the Grignon cropland site (C2) for the 2008-2009 growing seasons (maize and after that winter wheat), and for two consecutive years (2004-2005) at the Easter Bush permanent grassland site (G4). All meta-modelling approaches are plotted based on the LOYO validation strategy alongside the best-performing individual model (M01 for Grignon and M22 for Easter Bush). Appendix A contains the complete simulated dataset for all sites, and for all years, and also for the 70/30 validation strategy.



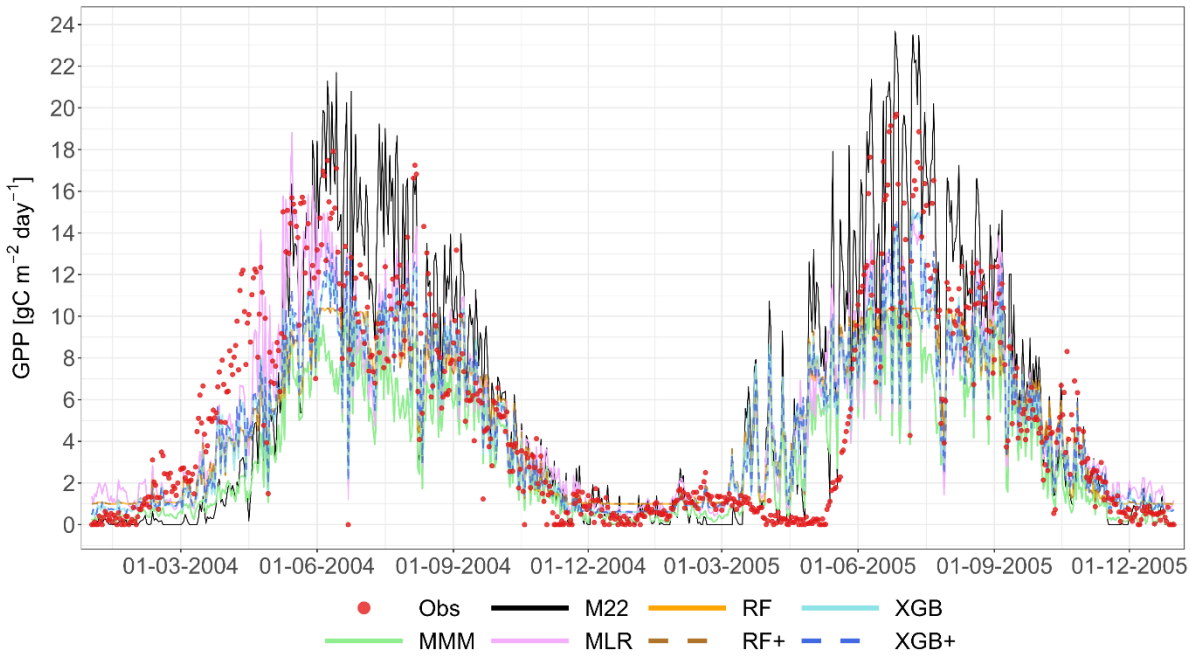


Figure 2: Performance of the multi-model median (MMM, green), the constructed meta-models and the best-performing individual models for simulating GPP based on the LOYO strategy. The top panel shows results for the Grignon cropland site (C2), which includes maize and winter wheat. The bottom panel shows two years of data for the Easter Bush grassland site (G4). Observations are marked by red dots. The meta-models include Multiple Linear Regression (MLR, purple), Random Forest (RF, orange), XGBoost (XGB, light blue), Random Forest+ (RF+, hatched brown) and XGBoost+ (XGB+, hatched dark blue). The best-performing individual models (M01 at C2, M22 at G4) are shown in black. Dates are provided in the format of dd-mm-yyyy.

The plots indicate a relatively large scatter in model results compared to observations. Notably, the best-performing individual model for the cropland time series exhibited considerable biases, generally overestimating GPP. In contrast, all meta-model approaches indicate reduced bias. To provide objective quality indicators, further quantitative analysis was performed.

Fig. 3 compares observed and modelled GPP for all four experimental sites and all meta-model construction methods used the LOYO strategy, including the MMM. The figure includes a scatterplot showing the performance of the best individual model, selected based on its RMSE score. RMSE was chosen as the primary selection metric because it correlates well with other performance metrics (Kobayashi and Salam, 2000; Robeson and Willmott, 2023). The figures were constructed using one validation year selected based on the performance of the XGB+ method.

Fig. 3 shows a gradual improvement in performance from top to bottom and tighter alignment of data points along the 1:1 line. For sites C1, C2 and G3, and to some extent for G4 as well, MMM already corrects much of the bias seen in the individual simulations.

At site C1, improvements across MLR, RF, RF+, XGB and XGB+ is associated with the restricted temporal coverage of the dataset (in LOYO strategy half of the dataset was retained). At the grassland sites G3 and G4, where longer time series was available providing a larger number of data points led to more pronounced improvement.

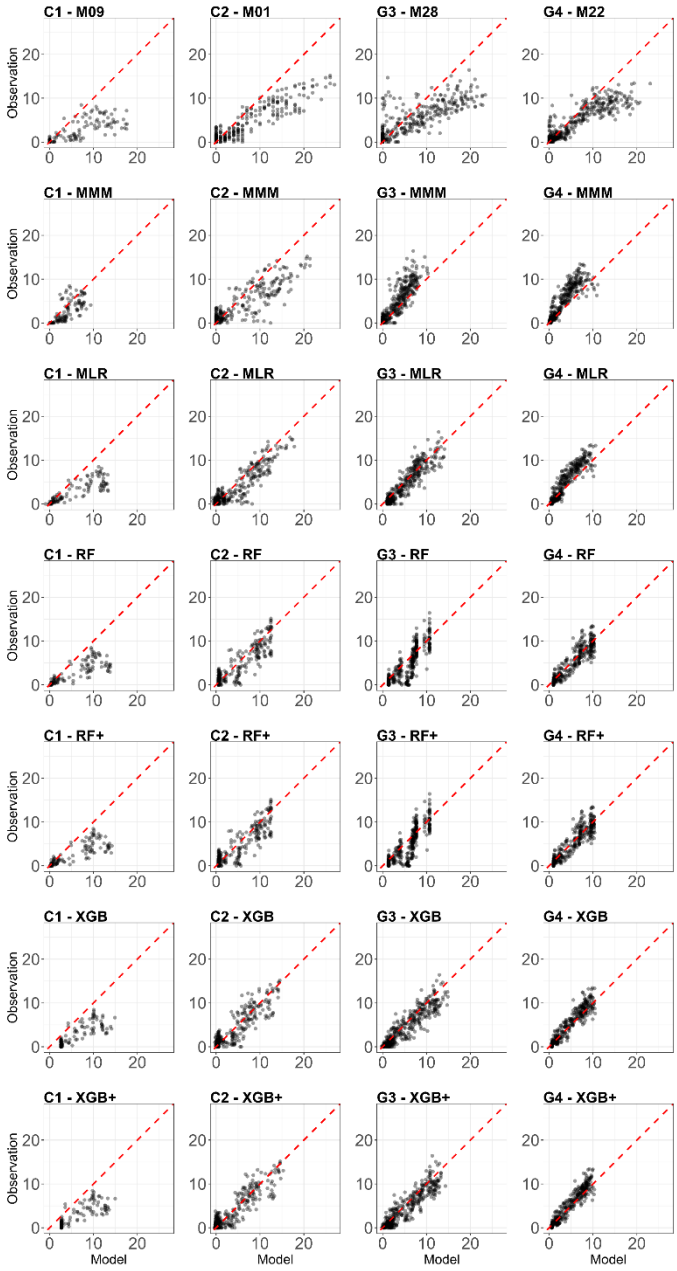


Figure 3: Comparison of the best individual model, the constructed meta-models and the traditional Multi-Model Median with the observations for all sites (from left to right C1, C2, G3 and G4) and for the entire time series for GPP based on the LOYO strategy. From top to bottom: best individual model, MMM, MLR, RF, RF+, XGB and XGB+. All units are in g C m^{-2} day^{-1}. Red dashed line represents the 1:1 relationship.

Table 3 presents statistical measures for all sites and modelling approaches evaluated, alongside the best-performing individual model, using data only from the validation subset.

Table 3: Statistical evaluation of the best performing individual model, the multi-model median (MMM) and the applied meta-models (MLR, RF, RF+, XGB and XGB+) with respect to three performance metrics for GPP: root mean square error (RMSE), bias and Pearson's correlation coefficient (r). Only validation data were used for the

calculation of the statistics. RMSE and BIAS are provided in g C m^{-2} day^{-1} units. The table present performance metrics for the 70/30 and the LOYO strategy as well.

Site	Metric	best base model	MMM	MLR	RF	RF+	XGB	XBG+
C1	RMSE 70/30	3.65	3.03	2.41	1.98	1.95	1.93	2.05
	RMSE LOYO	3.90	2.88	4.57	4.68	4.72	4.53	4.56
	BIAS 70/30	1.52	-0.61	-0.56	-0.51	0.22	-0.19	-0.28
	BIAS LOYO	1.94	-0.04	-0.25	-0.11	-0.07	0.79	0.81
	r 70/30	0.799	0.864	0.889	0.929	0.888	0.922	0.914
	r LOYO	0.862	0.795	0.877	0.867	0.863	0.835	0.837
C2	RMSE 70/30	4.22	2.88	2.55	1.93	2.09	1.92	1.90
	RMSE LOYO	4.14	2.86	2.62	2.71	2.72	2.80	2.81
	BIAS 70/30	1.02	-0.05	0.16	0.1	0.18	0.12	0.15
	BIAS LOYO	1.15	-0.03	0.02	0.16	0.17	0.09	0.00
	r 70/30	0.776	0.790	0.817	0.899	0.886	0.901	0.902
	r LOYO	0.839	0.850	0.852	0.828	0.828	0.825	0.818
G3	RMSE 70/30	4.48	2.96	2.12	1.79	2.28	1.80	1.78
	RMSE LOYO	3.47	2.87	2.26	2.55	2.55	2.31	2.30
	BIAS 70/30	1.22	-1.6	0.02	0.03	0.19	0.06	0.05
	BIAS LOYO	-1.74	-1.60	-0.03	-0.07	-0.07	-0.01	-0.01
	r 70/30	0.774	0.793	0.855	0.899	0.833	0.897	0.901
	r LOYO	0.724	0.825	0.862	0.806	0.805	0.861	0.861
G4	RMSE 70/30	3.02	2.96	2.37	1.84	2.51	1.80	1.82
	RMSE LOYO	2.80	2.95	2.81	2.64	2.63	2.67	2.71
	BIAS 70/30	0.11	-1.36	0.04	0.02	0.00	0.02	0.03
	BIAS LOYO	0.06	-1.37	-0.23	-0.03	-0.03	-0.07	-0.08
	r 70/30	0.754	0.809	0.844	0.911	0.824	0.914	0.912
	r LOYO	0.820	0.835	0.828	0.851	0.851	0.843	0.837

Table 3 highlights the improved performance of meta-models relative to both the best individual model and the MMM, which was previously published in Sándor et al. (2020). Notably, moving to more advanced meta-models results in reduction in both RMSE and absolute bias, which is more emphasized in case of the 70/30 validation strategy. For the G4 site RMSE reduction was marginal for LOYO, while for C1 RMSE increased as we moved to more advanced methods (again, this is attributed to the limited validation data and differing crop types). For the two grassland sites, the meta-models provide almost bias-free estimations, while for the crop sites the bias remains low but does not approach zero. The correlation coefficient typically increased by a maximum of 0.11 for the best-performing meta-model compared with the MMM in case of within-regime validation (70/30).

Wilcoxon signed-rank tests on daily prediction errors (Table 4) revealed strong and consistent improvements of meta-models over the Multi-Model Median baseline. For the grassland sites, all five meta-models performed significantly better than MMM for GPP, NEE, and RECO predictions. At cropland site C2, the tree-based models (RF, RF+, XGB, XGB+) showed significant improvements across all three variables, while MLR achieved significance only for RECO. At site C3, similar patterns emerged with tree-based models outperforming MMM for all variables. In contrast, site C1 showed no significant differences for any model-variable combination, likely due to high inter-annual variability with only two test years available. Overall, 43 out of 60 model-site-variable

405 combinations demonstrated significantly better performance than MMM, with XGBoost-based models showing the most consistent improvements across sites and variables.

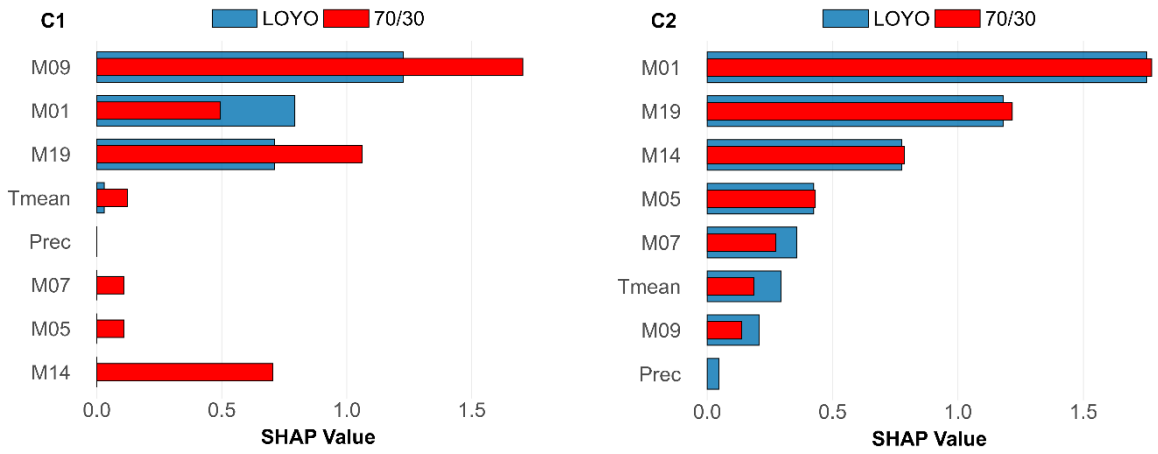
Table 4: Statistical significance of meta-model performance compared to the Multi-Model Median (MMM) based on one-sided Wilcoxon signed-rank tests on daily prediction errors for GPP. The alternative hypothesis tested
410 whether meta-models produce smaller squared residuals than MMM. Tests were performed separately for each Leave-One-Year-Out cross-validation fold, with p-values aggregated using the median across test years. The Benjamini-Hochberg procedure was applied per site to control the false discovery rate at 5%. Significance levels: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; ns = not significant.

Site	MLR	RF	RF+	XGB	XGB+
C1	ns	ns	ns	ns	ns
C2	ns	***	**	***	***
G3	***	**	**	***	***
G4	***	***	***	***	***

415

Given the novelty of the XGB+ method and its occasional better performance over all other approaches, here we analyse its functioning using SHAP values of the input data streams (individual models, plus temperature and precipitation) for both validation strategies. SHAP value analysis (Fig. 4) shows that certain models (e.g. M09, M01, M19 in case of LOYO, while M09, M19, M14 and M01 in case of 70/30 at C1; M01, M19 and M14 using LOYO
420 and 70/30 as well at C2; M16, M24, M14 and M05 using LOYO and 70/30 as well at G3; M16, M06, M22 in case of LOYO, while M22, M16 and M06 in case of 70/30at G4) dominate contributions to GPP prediction across sites. Temperature is also identified as a notable predictor (at C2 and G3 for both validation strategies). Precipitation consistently exhibits minimal impact on GPP prediction across all sites.

425



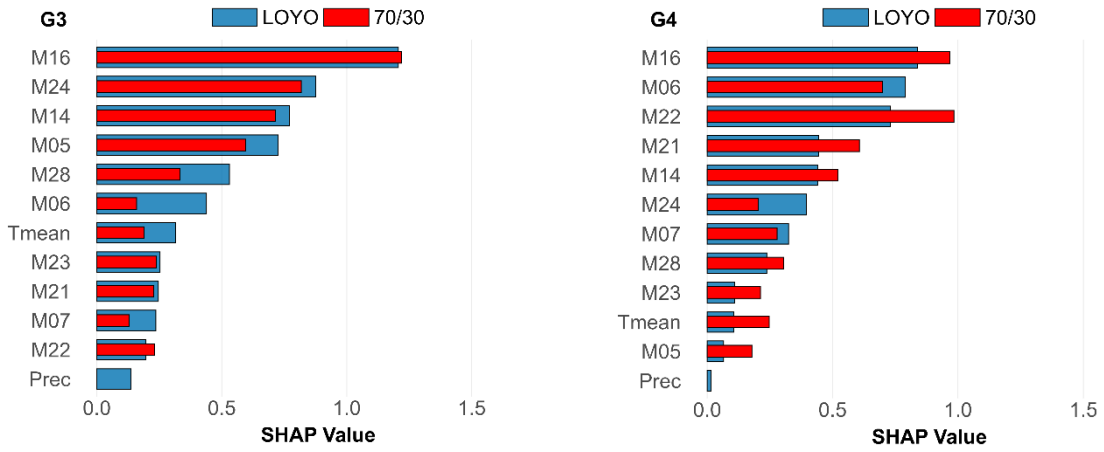


Figure 4: The SHAP values of the XGB+ meta-model for GPP based on the 70/30 strategy (red) and the LOYO strategy (blue). Larger values mean stronger contribution to the resulting GPP. Tmean stands for daily mean temperature, and Prec is daily precipitation.

3.2 RECO

Fig. 5 presents a two-year comparison of simulated and observed RECO for Grignon (C2) and Easter Bush (G4), again based on the LOYO strategy. The graphs display the results of the meta-models, the MMM and the best-performing individual models (M26 and M22). Appendix A contains the complete simulated dataset for all sites, and for all years, as well as for the 70/30 strategy.

At both sites, the MMM consistently underestimates respiration. This bias is especially pronounced at G4, where the MMM produces sharp decline of RECO around June-July, resulting in a poor fit to the observations. In contrast, the meta-models and the best individual model successfully follow the positive peak of the measured values at this site. Regarding the individual models, the best-performing model at C2 (M26) overestimates both maize and winter wheat respiration. At G4, the best individual model (M22) greatly overestimates the maximum values for both years. This visual analysis suggests that the MMM does not clearly outperform M22 or M26.

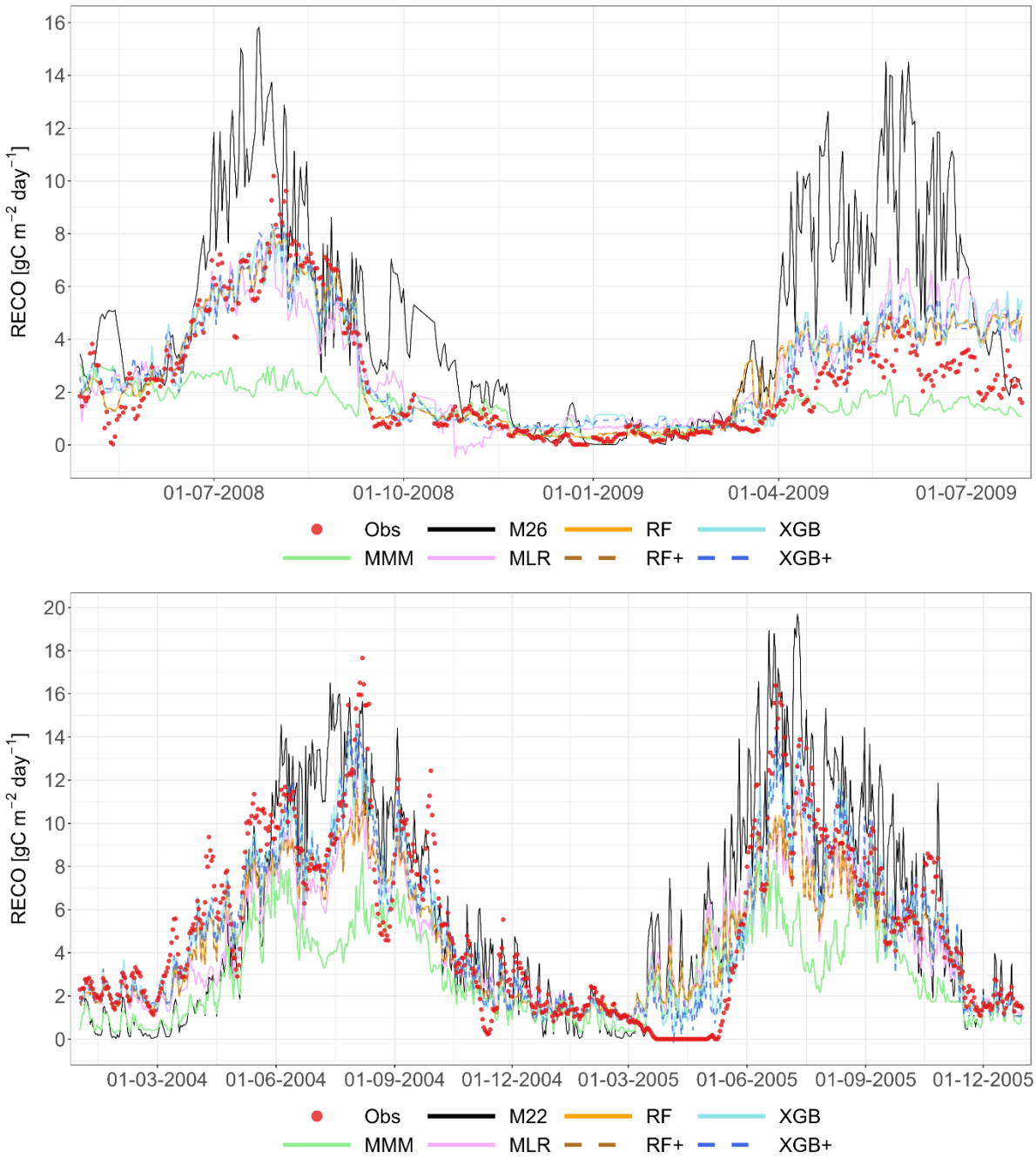


Figure 5: Performance of the multi-model median (MMM, green), the constructed meta-models and the best-performing individual models for simulating RECO (LOYO strategy). The top panel shows results for the Grignon cropland site (C2), which includes maize and winter wheat. The bottom panel shows two years of data for the Easter Bush grassland site (G4). Observations are marked by red circles. The meta-models include Multiple Linear Regression (MLR, purple), Random Forest (RF, orange), XGBoost (XGB, light blue), Random Forest+ (RF+, hatched brown) and XGBoost+ (XGB+, hatched dark blue). The best-performing individual models (M01 at C2, M22 at G4) are shown in black. Dates are provided in the format of dd-mm-yyyy.

A more quantitative analysis can be performed by inspecting Fig. 6, which presents scatterplots comparing simulated and observed RECO. The figures were constructed using one validation year selected based on the performance of the XGB+ method. The MLR consistently provides a "middle-ground" performance, with good performance at G4.

The scatter plots in Fig. 6 distinctly illustrate that the MMM has a negative bias, which means that it consistently underestimates the observed RECO. This bias is eliminated even by the simplest meta-model (MLR), and as more sophisticated meta-model techniques are employed, the alignment with the 1:1 line becomes significantly better.

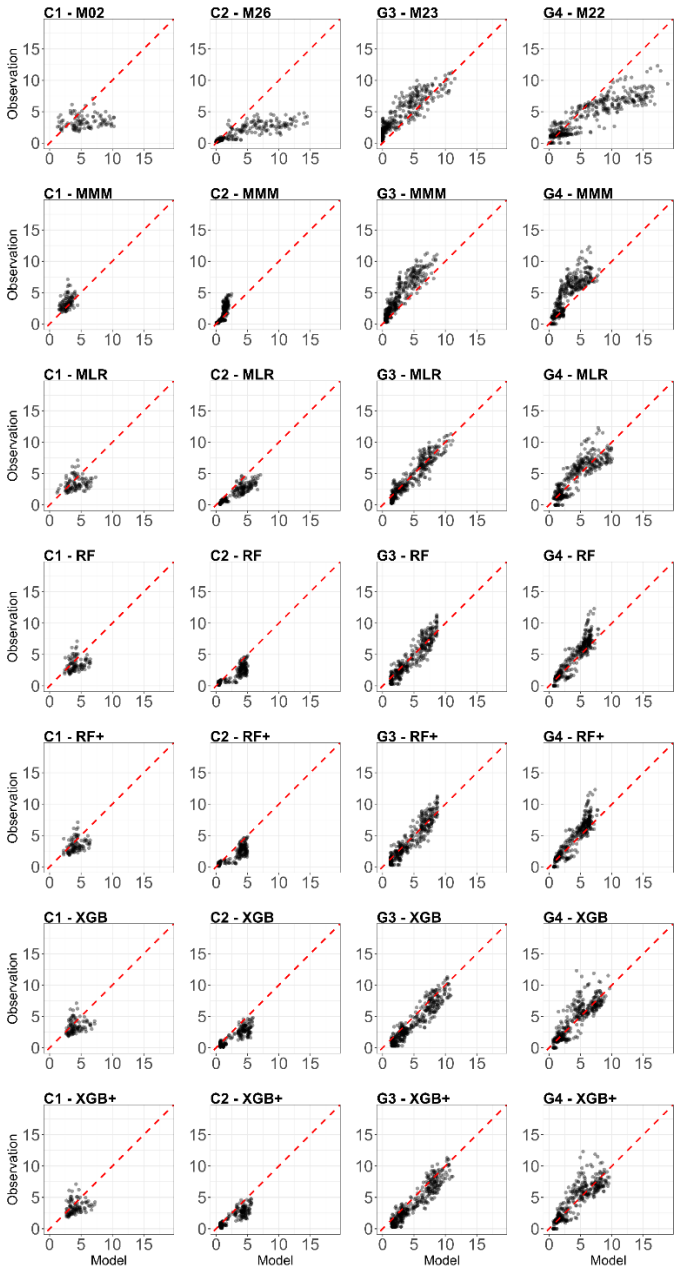


Figure 6: Comparison of the best individual model, the multi-model median and the constructed meta-models with observations for RECO based on the LOYO approach. Each row represents a different model type, and columns correspond to the sites (from left to right: C1, C2, G3, and G4). The top row shows the best individual models with their identifiers (M13 at C1, M01 at C2, M22 at G3 and G4). The remaining rows show the MMM, MLR, RF, RF+, XGB and XGB+. All units are in g C m^{-2} day^{-1}. The red dashed line represents the 1:1 relationship.

Table 5 provides statistics calculated based on the validation dataset both for 70/30 and LOYO strategy. A comparison of the MMM and the best individual models reveals some key differences. At C1 the MMM exhibits lower RMSE values, indicating better accuracy. The correlation coefficient is higher than the best individual model

at C1 (0.469 vs. 0.381 for 70/30, while 0.549 vs. 0.418 for LOYO) but slightly lower at C2. At the grassland sites (G3 and G4), the MMM shows slightly higher RMSE values but a better correlation coefficient than for the best individual models. The MMM at the C1 and C2 sites behaves differently compared to the other models. Overall, the meta-models demonstrate a distinct improvement over both the MMM and the best individual models (except for G4 with the LOYO strategy). Among the meta-models, RF, RF+, XGB and XGB+ are the top performers. At the grassland sites, the performance differences between RF, RF+, XGB, and XGB+ were small, with all three models demonstrating strong performance across all metrics under the LOYO strategy. Overall, correlation coefficients increased in a variable fashion (C2 was associated with the largest increase both under LOYO and 70/30, while the increase was marginal for the grassland sites. The meta-models typically show almost unbiased estimates, with the exception of site C1.

Table 5: Statistical evaluation of the best-performing individual model, the multi-model median (MMM) and the applied meta-models (MLR, RF, RF+, XGB and XGB+) for RECO. Three performance metrics are used: root mean square error (RMSE), bias and Pearson's correlation coefficient (r). Only validation data were used for the calculation of the statistics. RMSE and BIAS are provided in g C m^{-2} day^{-1} units.

Site	Metric	best base model	MMM	MLR	RF	RF+	XGB	XBG+
C1	RMSE 70/30	2.61	1.72	0.96	0.84	0.75	1.03	0.92
	RMSE LOYO	1.65	1.57	3.04	1.69	1.58	1.72	1.60
	BIAS 70/30	1.56	-1.29	-0.32	-0.23	0.17	-0.14	-0.17
	BIAS LOYO	-0.76	-1.10	-1.69	-0.23	-0.17	-0.13	-0.01
	r 70/30	0.381	0.469	0.707	0.777	0.818	0.607	0.707
	r LOYO	0.418	0.549	0.370	0.309	0.434	0.295	0.415
C2	RMSE 70/30	2.72	2.24	1.30	0.79	0.78	0.94	0.93
	RMSE LOYO	2.16	2.09	2.84	1.69	1.71	1.69	1.65
	BIAS 70/30	1.48	-1.09	0.01	0.01	0.00	0.03	0.05
	BIAS LOYO	-1.10	-1.13	-0.24	-0.07	-0.09	0.13	-0.01
	r 70/30	0.599	0.556	0.821	0.940	0.947	0.912	0.912
	r LOYO	0.617	0.685	0.727	0.775	0.768	0.781	0.778
G3	RMSE 70/30	2.42	2.72	1.56	1.27	1.57	1.19	1.19
	RMSE LOYO	2.16	2.49	1.71	1.66	1.66	1.96	1.94
	BIAS 70/30	0.44	-1.69	0.08	0.06	0.00	0.05	0.02
	BIAS LOYO	0.09	-1.75	0.06	0.00	0.00	0.23	0.25
	r 70/30	0.641	0.700	0.855	0.907	0.843	0.918	0.918
	r LOYO	0.830	0.820	0.888	0.879	0.879	0.862	0.866
G4	RMSE 70/30	2.36	2.52	1.97	1.36	1.78	1.40	1.35
	RMSE LOYO	2.26	2.50	2.38	2.20	2.20	2.50	2.44
	BIAS 70/30	0.51	-1.08	0.04	0.01	0.00	0.06	0.05
	BIAS LOYO	0.47	-1.13	0.02	-0.03	-0.03	-0.01	0.00
	r 70/30	0.714	0.720	0.798	0.912	0.851	0.905	0.912
	r LOYO	0.768	0.759	0.787	0.798	0.798	0.742	0.744

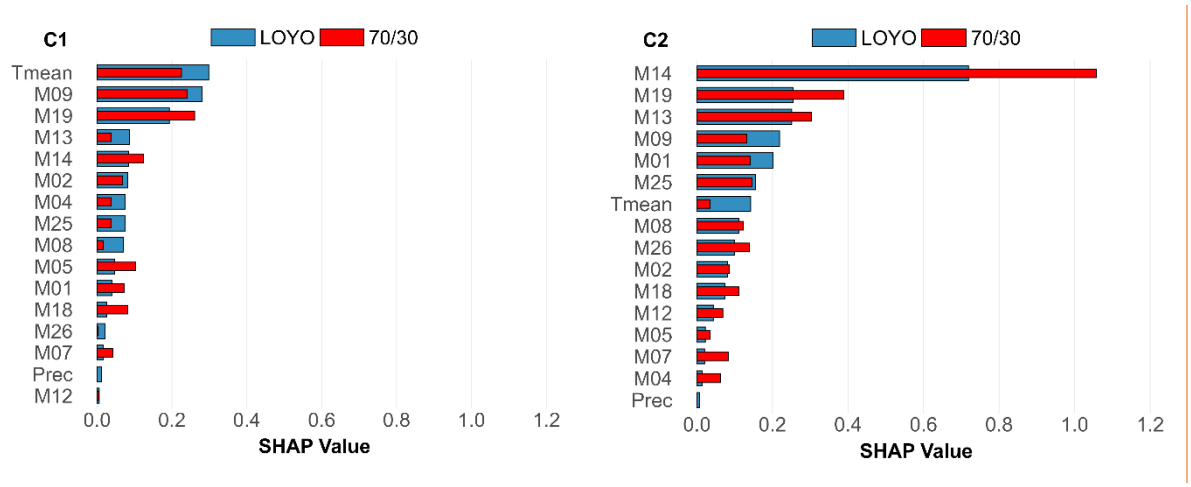
Wilcoxon signed-rank tests on daily prediction errors (Table 6) confirmed the performance improvements observed in the RMSE metrics. For RECO predictions, all five meta-models achieved statistical significance at both grassland

sites (G3 and G4) and at cropland site C2. Site C1 showed no significant differences for any model. This pattern demonstrates that the RMSE improvements for RECO are statistically robust and not merely artifacts of random variation, particularly for the grassland ecosystems where all meta-modelling approaches consistently outperformed the baseline. Notably, in some cases a meta-model may show higher RMSE than MMM yet still demonstrate statistical significance in the Wilcoxon test (e.g., MLR at site C2: RMSE of 2.84 vs. MMM's 2.09, yet significant at $p < 0.001$). This reflects the different aspects these metrics evaluate: RMSE is sensitive to outliers and can be inflated by poor performance in a single year, while the Wilcoxon test based on median p-values assesses whether the model typically performs better across most years, providing a more robust measure of consistent improvement.

Table 6: Statistical significance of meta-model performance compared to the Multi-Model Median (MMM) based on one-sided Wilcoxon signed-rank tests on daily prediction errors for RECO. The alternative hypothesis tested whether meta-models produce smaller squared residuals than MMM. Tests were performed separately for each Leave-One-Year-Out cross-validation fold, with p-values aggregated using the median across test years. The Benjamini-Hochberg procedure was applied per site to control the false discovery rate at 5%. Significance levels: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; ns = not significant

Site	MLR	RF	RF+	XGB	XGB+
C1	ns	ns	ns	ns	ns
C2	***	***	***	***	***
G3	***	***	***	***	***
G4	***	***	***	***	***

SHAP values (Fig. 7) indicate that model outputs are the primary contributors of RECO predictions. Among environmental predictors, mean temperature emerges as an important contributor, especially at site C1, where its influence nearly matches top models M19 and M09 for 70/30, and it is the most important for LOYO. Precipitation influence on RECO is negligible across all sites.



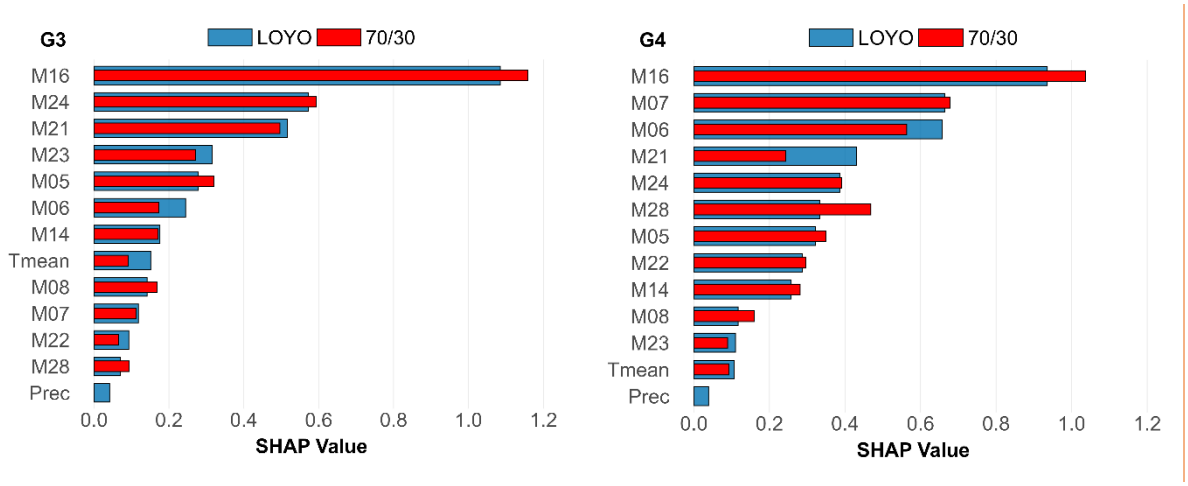
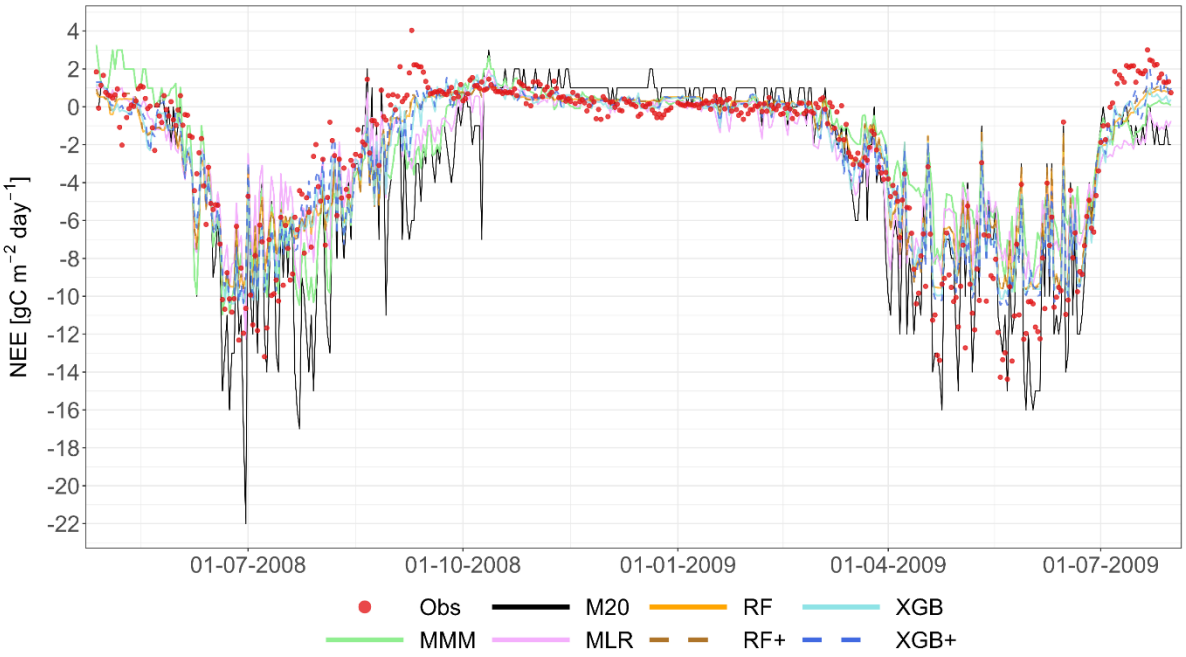


Figure 7: The SHAP values of the XGB+ meta-model for RECO based on the 70/30 strategy (red) and the LOYO strategy (blue). Larger values mean stronger contribution to the resulting RECO. Tmean stands for daily mean temperature, and Prec is daily precipitation.

3.3 NEE

Fig. 8 illustrates the simulated and observed NEE for Grignon (C2) and Easter Bush (G4) over two years based on the LOYO strategy. Appendix A contains the complete simulated dataset for all sites, and for all years, for the 70/30 strategy as well.

The MMM shows greater consistency with other meta-models at C2 compared to its performance for RECO in Fig. 5. For both sites, M20 (best model) tends to diverge most from the other models during observation peaks (this is emphasized for grasslands), whereas RF, RF+, XGB and XGB+ are consistently following a similar pattern, suggesting similar performance for C2.



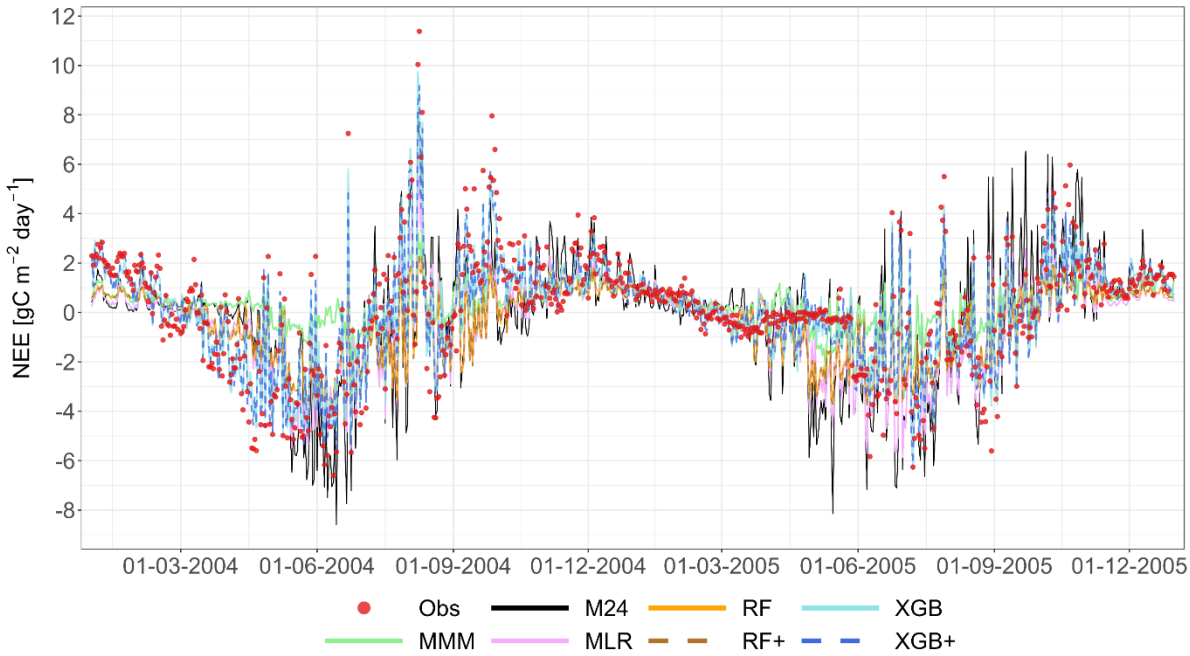


Figure 8: Performance of the multi-model median (MMM, green), the constructed meta-models and the best-performing individual models for simulating NEE based on the LOYO strategy. The top panel shows results for the Grignon cropland site (C2), which includes maize and winter wheat. The bottom panel shows two years of data for the Easter Bush grassland site (G4). Observations are marked by red circles. The meta-models include Multiple Linear Regression (MLR, purple), Random Forest (RF, orange), XGBoost (XGB, light blue), Random Forest+ (RF+, hatched brown) and XGBoost+ (XGB+, hatched dark blue). The best-performing individual models (M20 at C2, M24 at G4) are shown in black. Dates are provided in the format of dd-mm-yyyy.

To provide a more objective assessment, Fig. 9 presents scatterplots comparing the observations and the models. The figures were constructed using one validation year selected based on the performance of the XGB+ method. At the crop sites (C1 and C2), the meta-models, particularly RF, RF+, XGB and XGB+, show a strong linear relationship with the observed data, explaining a large portion of the variance.

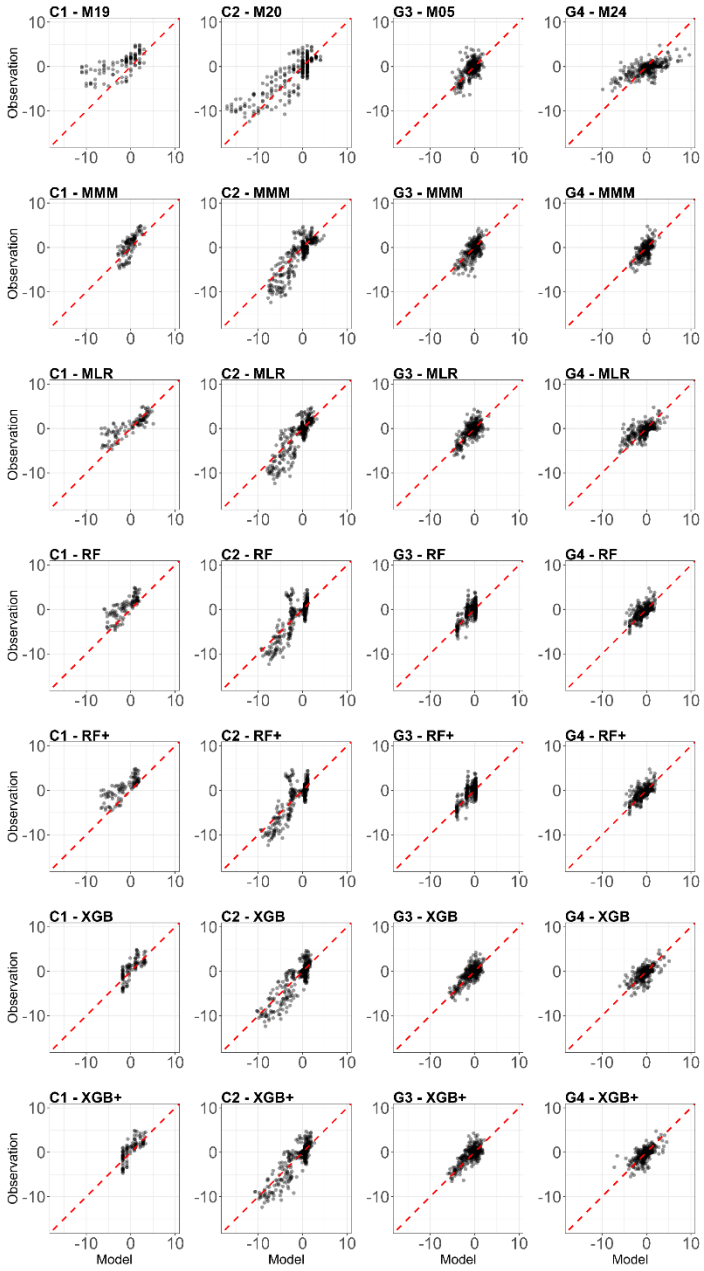


Figure 9: Comparison of the best individual model, the multi-model median and the constructed meta-models with observations for NEE using the LOYO approach. Each row represents a different model type, and columns correspond to the sites (from left to right: C1, C2, G3, and G4). The top row shows the best individual models with their identifiers (M20 at C1, M19 at C2, M22 at G3, M23 at G4). The remaining rows show the MMM, MLR, RF, RF+, XGB and XGB+. All units are in g C m^{-2} day^{-1}. The red dashed line represents the 1:1 relationship.

Table 7 shows statistics for all the models using three metrics based on the validation subset, for all three validation approaches (70/30, LOYO, and the NEE estimation using meta-model based RECO-GPP, abbreviated here as INDEP). Across all the four sites, the RF, RF+, XGB and XGB+ values differ little from each other in all three metric for 70/30 strategy. The MLR meta-model is on par with these four meta-models, except at C2, where its performance is worse across all three metrics, making it more comparable to the MMM for 70/30 (but not for LOYO and INDEP). The best-performing individual models typically have poor statistics compared to all other models.

Overall, the correlation coefficient increased in the range of 0.004-0.289 for the best-performing meta-model compared to MMM, with the most pronounced improvement observed at G3 using the INDEP strategy. Bias shows similarity to those presented for GPP and RECO, meaning that with the exception of C1 the meta-models provide results that are almost bias-free. At site C2, the INDEP approach for NEE achieves lower RMSE than direct LOYO (e.g., XGB+: 1.96 vs. 2.25) but introduces a small systematic bias (~0.3-0.4 g C m^{-2} d^{-1}), illustrating a precision-accuracy trade-off in physically consistent reconstruction. This trade-off suggests INDEP is preferable for applications prioritizing precision and physical consistency (e.g., gap-filling, carbon budgets), while direct stacking is better suited when unbiased estimates are essential (e.g., model validation, trend analysis).

Table 7: Statistical evaluation of the best-performing individual model, the multi-model median (MMM) and the applied meta-models (MLR, RF, XGB and XGB+) for NEE, based on the three validation strategies (70/30, LOYO and the RECO-GPP approach that is referred as INDEP). Three performance metrics are used: root mean square error (RMSE), bias and Pearson's correlation coefficient (r). Only validation data were used for the calculation of the statistics. RMSE and BIAS are provided in g C m^{-2} d^{-1} units.

Site	Metric	best base model	MMM	MLR	RF	XGB	RF+	XBG+
C1	RMSE 70/30	3.87	3.06	1.77	1.78	1.81	1.71	1.89
	RMSE LOYO	2.98	2.48	2.51	2.83	3.47	2.98	3.44
	RMSE INDEP	3.04	2.57	2.66	2.46	2.52	2.78	2.776
	BIAS 70/30	-0.48	0.1	0.38	0.46	0.66	0.00	0.52
	BIAS LOYO	-1.11	0.00	0.49	0.12	1.78	0.01	1.75
	BIAS INDEP	-1.15	-1.03	-1.18	-1.16	-1.21	-1.82	-1.866
	r 70/30	0.382	0.759	0.909	0.929	0.921	0.915	0.910
	r LOYO	0.772	0.766	0.864	0.816	0.683	0.830	0.770
	r INDEP	0.698	0.755	0.818	0.84	0.839	0.818	0.825
C2	RMSE 70/30	3.98	2.29	2.15	1.42	1.48	1.76	1.44
	RMSE LOYO	2.71	2.23	2.38	2.23	2.26	2.22	2.25
	RMSE INDEP	4.146	3.929	2.768	2.593	2.594	2.063	1.955
	BIAS 70/30	1.03	0.4	-0.1	-0.07	-0.02	0.00	0.02
	BIAS LOYO	0.06	0.48	-0.02	-0.06	0.01	-0.07	0.01
	BIAS INDEP	0.35	-1.23	0.26	0.38	0.36	0.39	0.284
	r 70/30	0.584	0.761	0.785	0.913	0.904	0.875	0.909
	r LOYO	0.703	0.789	0.775	0.793	0.791	0.791	0.788
	r INDEP	0.601	0.695	0.705	0.747	0.746	0.850	0.865
G3	RMSE 70/30	2.82	1.87	1.78	1.53	1.55	1.78	1.55
	RMSE LOYO	1.89	1.88	1.82	1.77	1.77	1.77	1.76
	RMSE INDEP	4.72	2.18	1.88	1.99	2.01	1.60	1.562
	BIAS 70/30	0.44	0.31	-0.03	0.01	0.02	0.00	-0.01
	BIAS LOYO	0.13	0.32	-0.02	-0.01	-0.03	-0.01	-0.03
	BIAS INDEP	-2.31	-0.15	-0.04	-0.08	-0.09	0.02	0.025
	r 70/30	0.299	0.584	0.623	0.741	0.734	0.625	0.734
	r LOYO	0.573	0.583	0.620	0.647	0.653	0.645	0.656
	r INDEP	0.270	0.452	0.581	0.52	0.515	0.726	0.741
G4	RMSE 70/30	2.61	2.06	1.64	1.39	1.39	1.53	1.36
	RMSE LOYO	2.04	1.95	1.80	1.67	1.84	1.67	1.81
	RMSE INDEP	2.13	2.13	1.83	1.92	1.92	1.91	1.868
	BIAS 70/30	0.48	0.67	-0.03	-0.01	0.03	0.00	0.02
	BIAS LOYO	0.20	0.67	0.05	0.03	0.13	0.03	0.12
	BIAS INDEP	0.33	0.24	0.10	-0.08	-0.08	0.01	0.005

	r 70/30	0.435	0.653	0.751	0.829	0.829	0.758	0.836
	r LOYO	0.718	0.680	0.737	0.763	0.702	0.764	0.705
	r INDEP	0.680	0.511	0.683	0.621	0.623	0.663	0.677

565

Wilcoxon signed-rank tests on daily prediction errors (Table 8) revealed significant improvements for NEE predictions at the grassland sites, where all five meta-models outperformed MMM at both G3 and G4. At cropland site C2, the four tree-based models showed significant improvements, while MLR did not achieve significance. Site
570 C1 showed no significant differences for any model. The statistical testing confirms that meta-models provide robust improvements in NEE prediction quality, particularly for grassland ecosystems. As with RECO, some cases show statistical significance despite minimal RMSE differences (e.g., RF at site G3), reflecting the complementary nature of these metrics: the Wilcoxon test evaluates typical daily performance across years, while RMSE is more sensitive to occasional large prediction errors.

575

Table 8: Statistical significance of meta-model performance compared to the Multi-Model Median (MMM) based on one-sided Wilcoxon signed-rank tests on daily prediction errors for NEE. The alternative hypothesis tested whether meta-models produce smaller squared residuals than MMM. Tests were performed separately for each Leave-One-Year-Out cross-validation fold, with p-values aggregated using the median across test years. The
580 Benjamini-Hochberg procedure was applied per site to control the false discovery rate at 5%. Significance levels: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; ns = not significant.

Site	MLR	RF	RF+	XGB	XGB+
C1	ns	ns	ns	ns	ns
C2	***	***	***	***	***
G3	*	**	**	***	***
G4	**	***	***	***	***

SHAP analysis presented in Fig. 10 reveals that a few key model outputs primarily determine NEE predictions for the XGB+ method, just like in the case of GPP and RECO. Meteorological variables, particularly temperature and
585 precipitation, contribute modestly but importantly, with a stronger effect noted at grassland sites G3 and G4, with larger contribution in case of LOYO.

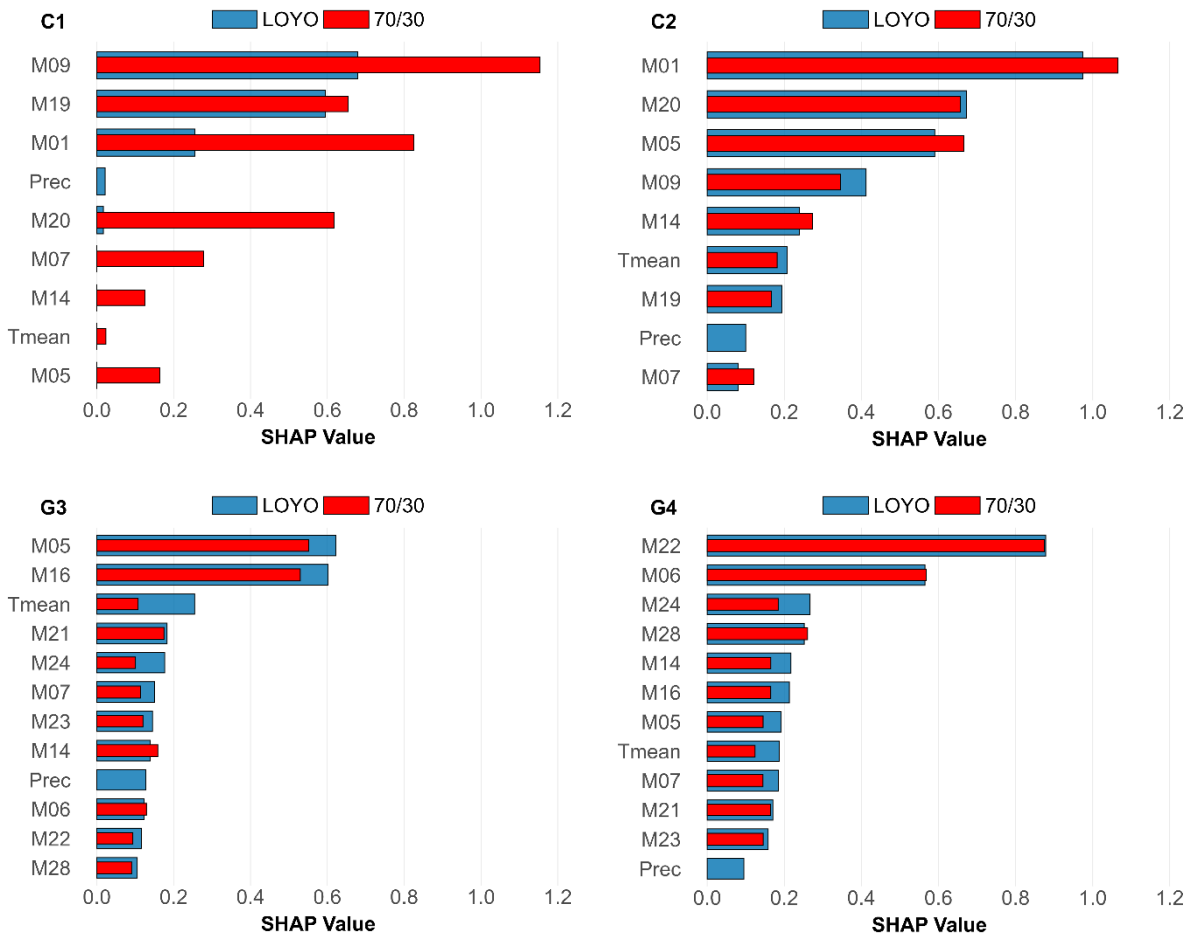


Figure 10: The SHAP values of the XGB+ meta-model for NEE based on the 70/30 strategy (red) and the LOYO strategy (blue). Larger values mean stronger contribution to the resulting NEE. Tmean stands for daily mean temperature, and Prec is daily precipitation.

4 Discussion

4.1 Synthesis of the meta-model approaches

The stacking-based meta-models (MLR, RF, RF+, XGB, XGB+) consistently outperformed both the best individual process-based models and the traditional MMM for sites with sufficient data coverage. By combining ensemble learning with key environmental covariates, RF+ and XGB+ emerged as a good candidate as well, in some cases achieving lower errors and higher correlations than all other approaches. Thus, while the MMM retained its role as a robust benchmark, its predictive power was consistently surpassed by the stacking-based methods.

This aligns with findings from Shahhosseini et al. (2020, 2021), who demonstrated that ensemble learning approaches can surpass the performance of individual models for crop yield prediction, and with Zhang et al. (2022), who reported similar benefits of ensemble learning in winter wheat yield modelling. The added value of explicitly incorporating drivers such as temperature into the meta-modelling framework also supports the conclusions of Mathieu and Aires (2018), who emphasised that agro-climatic indices like temperature and precipitation can significantly enhance model accuracy. Our results go beyond these earlier studies by demonstrating that such

improvements are really possible (even without incorporating meteorology) across multiple flux components (GPP, RECO, NEE) and agroecosystem types, including both croplands and grasslands.

Novelty of the study is the XGBoost and RF with environmental covariates (which means stacking with meteorological data) that represents a hybrid, regime-dependent integration framework rather than a purely statistical aggregation. This method seems to have potential that needs to be exploited in the future at sites with better temporal data coverage.

4.2 Cross-comparison of model contributions and environmental drivers for XGB+

Across all three fluxes (GPP, RECO, NEE), XGB+ also heavily relies on individual model outputs, confirming that the meta-model is successfully leveraging model consensus to improve predictive performance. However, the consistent appearance of temperature - particularly at the C1 site for RECO and at grassland sites for NEE - highlights that certain temperature-driven processes are not fully captured by the base models and need to be explicitly considered. Precipitation, while generally having less impact, may still play a role in certain site-specific cases (e.g. GPP at C1), suggesting localised interactions between water availability and C fluxes. Importantly, the climate characteristics of the study sites provide important context for interpreting these results. These locations lie within mid-latitude temperate zones with well-defined seasonality, and some (e.g. Ottawa) experience high continentality, marked by large annual temperature ranges and pronounced growing season transitions. In such climates, temperature becomes the primary constraint on biological activity during both dormancy and active periods, while precipitation is often more evenly distributed or limiting only during episodic droughts (Baldocchi, 2008; Koster et al., 2004; Sun et al., 2025). This climatic backdrop helps explain why temperature consistently emerges as a key predictor across sites and flux types, while precipitation plays a more site-specific and secondary role. These patterns align with broader findings from temperate ecosystems, where thermal constraints dominate respiration and photosynthesis processes, particularly outside of water-limited systems.

The dominant role of specific models in GPP predictions, supplemented by temperature's influence (Fig. 4), suggests that while ensemble model outputs capture much of the variability, temperature-dependent processes remain only partially resolved by the base models. This aligns with global studies such as Zhu et al. (2016) and Bellocchi et al. (2023), who identified temperature as a key driver of GPP dynamics, especially in temperate and high-latitude ecosystems. The minimal role of precipitation in GPP prediction indicates that water availability was not a limiting factor at the examined sites, consistent with ecosystem-specific findings reported by Reichstein et al. (2013), which showed temperature or radiation often dominate in temperate regions.

For RECO, the strong temperature effect (Fig. 7) corroborates well-established biological principles, such as the temperature sensitivity of respiration processes described by Lloyd and Taylor (1994). The negligible role of precipitation further suggests these sites are not subject to drought stress, echoing findings from Xu et al. (2025) who highlighted thermal thresholds as primary controls over ecosystem carbon fluxes.

NEE's XGB+ related sensitivity to meteorological variables at grassland sites (Fig. 10) confirms the known climate sensitivity of these systems (Baldocchi et al., 2018). The significant influence of temperature and soil water content on GPP and RECO in grasslands, as noted by Xia et al. (2024), explains the prominence of these variables in NEE

predictions. These findings underscore the importance of climate drivers in modulating carbon fluxes beyond what model ensembles alone capture.

Fig. 11 provides deeper insight into the functioning of XGB+ by ranking the influence of the top-performing process-based models (using SHAP values) for each flux and site. Several distinct patterns emerge. For RECO predictions, the influence of the highest-performing models was less prominent than for GPP, particularly at site G3 for both strategies. This indicates that, for RECO, the ensemble relied less on the best individual models. In contrast, for GPP, the top-performing models had a substantial impact for the crop sites, ranking first at both C1 and C2 for both the LOYO and 70/30 strategy. This suggests that XGB+ leveraged structurally diverse yet individually strong models to enhance predictive accuracy. For grassland sites the influence of the best models remains strong, especially for the 70/30 strategy. In case of NEE, the70/30 strategy exhibited moderate to low contributions from top-performing models - especially at grassland sites, pointing to a more balanced integration of individual models and ensemble diversity. This differs for the LOYO strategy where the top-performing individual models provide notable influence across all sites in the meta-model results.

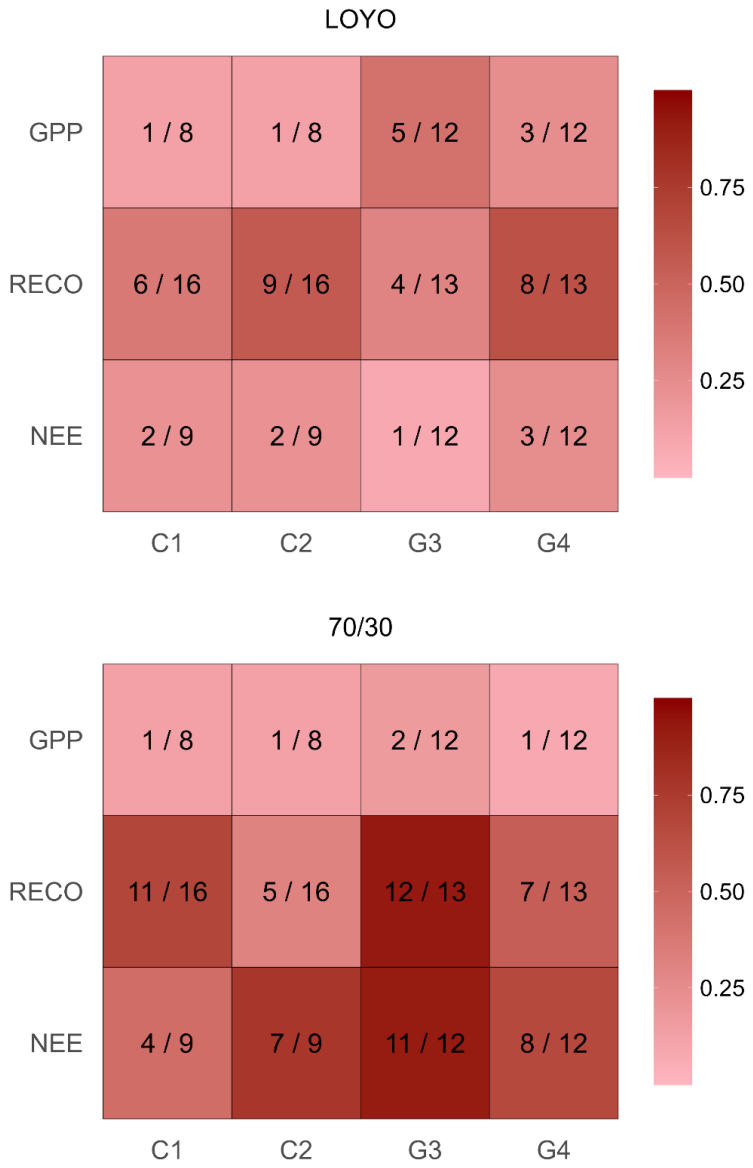


Figure 11: Ranking of the top-performing individual models (selected based on RMSE) across each site and model output based on their contribution to the XGB+ model. The first number in each cell indicates the ranking for the best model among the other models according to the SHAP value, while the second number is the total number of models. Red shades are visual representations of the first number relative to the second indicating how important or negligible the best model's influence was.

These findings reinforce a key principle of ensemble learning: the models with the highest standalone accuracy are not always the most influential within the ensemble. By contributing unique and non-redundant information, even structurally weaker models can significantly improve an ensemble's predictive ability. This finding, based on SHAP analysis, is consistent with the principle of ensemble diversity highlighted by Chergui and Kechadi (2022), which posits that combining a variety of models with different strengths and weaknesses leads to more robust and accurate predictions. However, our SHAP analysis adds a novel layer of interpretability by quantifying the relative influence of top-performing versus structurally diverse models within the ensemble, providing diagnostic insights that go beyond earlier ensemble studies (Shahhosseini et al., 2020).

Gains in performance were more pronounced for croplands than for grasslands, especially for GPP and NEE. This aligns with previous research (Bansal et al., 2024; Nand et al., 2025) which found that management-intensive crop systems, like maize and winter wheat, benefit more from dynamic environmental and management data due to their seasonal and management-driven variability. For instance, Nand et al. (2025) demonstrated that weighted multi-model averaging, specifically the Granger-Ramanathan B method for actual evapotranspiration (ETa), reduced RRMSE by 4–8.5% in croplands. The Granger-Ramanathan B method is a multi-model averaging approach that requires non-negative weights that sum to one, and it provides the closest match to measured values for daily ETa in maize simulations. Conversely, grasslands, characterised by more stable phenological cycles, show smaller improvements. Their flux variability is often more strongly influenced by biotic controls rather than environmental drivers alone (Reichstein et al., 2007; Stoy et al., 2013). This is because grasslands exhibit more stable, biologically mediated carbon fluxes, while the sharp seasonal transitions and management-driven dynamics of crop systems make them particularly responsive to external data.

Overall, the XGB+ meta-model achieved good predictive performance across diverse agro-environmental contexts, while also providing enhanced interpretability by quantifying the contributions of individual models and environmental drivers. These dual benefits - improved performance and diagnostic insight - position XGB+ as a powerful framework for advancing predictive modelling and guiding process-model refinement.

4.3 Implications and applications

The demonstrated improvements in predictive performance of stacking-based meta-models have direct implications for crop and grassland biogeochemical modelling. By consistently outperforming both the best-performing individual models and the MMM, these approaches provide a pathway for enhancing C-flux predictions across diverse agro-environmental contexts. This is particularly relevant for applications such as greenhouse gas inventorying, site-specific management planning, and scenario analysis under climate change, where reliability is critical.

Beyond performance gains, the interpretability of the meta-models offers valuable diagnostic insights for process modellers. The analysis of regression coefficients and feature contributions highlights the value of retaining structurally diverse models in ensembles, rather than limiting selection to the top-performing candidates. Moreover, identifying temperature as a dominant external contributor point to opportunities for refining process-based models, such as improving their representation of phenological and temperature response mechanisms. These insights can also inform calibration practices: by quantifying the relative influence of models and drivers, the meta-modelling framework can guide targeted recalibration and prioritisation of model improvement efforts. In this way, meta-model outputs can serve as both a predictive tool and a diagnostic instrument for iterative process model development.

Together, these findings have broader implications for how model ensembles should be constructed and applied. Our results highlight that the traditional practice of equal-weight multi-model averaging (like with the MMM) may not be sufficient for delivering credible predictions. As noted in Section 4.4, the approach depends on long, high-quality time series for training, which may limit application in some contexts. Consistent with the arguments of Mathieu and Aires (2018), Eyring et al. (2019), Shahhosseini et al. (2020) and Chergui and Kechadi (2022), who emphasised that structurally dependent and unevenly performing models require differentiated treatment, our findings demonstrate that stacking-based meta-models - combining adaptive weighting with key environmental drivers - consistently outperform both the best individual models and the MMM. This evidence calls for a paradigm shift in ensemble design: future model intercomparison and synthesis efforts should move toward performance-informed, diagnostic-driven weighting strategies that explicitly incorporate relevant covariates. Such approaches not only improve predictive accuracy but also generate actionable insights for refining underlying process-based models, ultimately accelerating progress toward more reliable and process-rich representations of agroecosystem C dynamics. This calls for coordinated action by the modelling community to adopt adaptive, performance-informed ensemble frameworks in future intercomparison efforts. Furthermore, our framework could be applied as a post-processing step to archived ensemble outputs from major multi-model initiatives (discussed in Section 5), maximising the value of existing model intercomparison investments without requiring new simulations.

4.4 Limitations and future research

While the proposed meta-modelling framework offers notable advantages, several limitations warrant acknowledgment.

The approach depends on sufficiently long and consistent time series for training and evaluation, which were not uniformly available across all sites and flux components, limiting generalisability in data-scarce regions (e.g. C1 site). This data dependence is highlighted in Section 4.3 as a key consideration for ensemble design.

While the results indicated that the models show high temporal fidelity, spatial extrapolation to unobserved sites was not the objective of this work. This needs further investigation in forthcoming studies.

Incorporating external covariates improves performance but increases complexity and introduces dependencies on auxiliary datasets that may not always be accessible or reliable. Our analysis also focused on a limited set of covariates (temperature and precipitation), leaving unexplored the potential benefits of integrating hydrological or soil-related variables. Residual analysis can potentially be used to identify site-specific environmental variables that could be included as additional covariates, of course considering the potential collinearity issues.

The interpretability of regression coefficients also comes with caveats: while they provide insight into model influence within the ensemble, they do not directly reveal mechanistic underpinnings, and the relationships captured remain largely empirical. Consequently, care must be taken when using these findings to inform process-level changes.

Future research should expand the set of environmental covariates to include soil moisture, global radiation, management practices, and hydrological variables, which are likely to improve representation of key drivers in both crops and grasslands. There is a great potential in including the same environmental variables in a more sophisticated way that can include average (or cumulative) conditions (like heat-sum) for longer time periods and also considering lagged effects using asynchronous climate data. Testing the framework across a broader range of management systems, crop types, and climatic zones would also help assess scalability and robustness. Moreover, integrating advanced interpretability tools (e.g., causal inference methods) could move beyond purely statistical associations toward more mechanistically grounded insights. Finally, co-developing meta-model frameworks with process-based modellers could establish a feedback loop in which ensemble diagnostics directly inform iterative improvements to individual models, accelerating convergence toward more reliable, process-rich representations.

5 Conclusions

International initiatives that foster collaborations between researchers working on agricultural and grassland models are creating new opportunities in the field of process-oriented modelling. By gathering outputs from several models with different representations of plant and soil processes and using standardised protocols, exploitation of the potential of the ensembles becomes possible. However, those multi-model ensemble techniques are still in their infancy, as typically simple multi-model means or medians are constructed as robust estimations that typically overperform the individual models. Some studies attempted to use more sophisticated methods like skill-based model selection and even machine learning, but the potential of the multi-model frameworks is still being explored. In this study, building on a previous multi-model exercise performed under the umbrella of international initiatives, new combinations of models are tested that we call meta-models.

The introduced meta-models significantly improved the accuracy of C-flux estimates in crop and grassland ecosystems compared to individual process-based models and traditional multi-model medians. Formal statistical testing confirmed that improvements were particularly robust for grassland sites, where all meta-models significantly outperformed the multi-model median across all variables, while cropland results were more variable, with one site showing no significant improvements due to limited data availability. By integrating structurally diverse models and incorporating key environmental variables - particularly temperature -, these meta-models deliver not only more reliable predictions but also diagnostic insights into the relative contributions of models and environmental drivers. This underscores the importance of maintaining model diversity in ensembles and highlights opportunities to refine process-based models, especially regarding temperature responses and phenological processes, while also demonstrating that meta-modelling effectiveness depends on sufficient training data, particularly in cropland systems with high inter-annual variability.

Perhaps the most important element of our modelling framework is the presentation of a method to goes beyond the static nature of meta-models. As the changing environmental conditions can alter the relevance of the base-models, the optimal meta-model may vary over time. Without the environmental variables, continuous retraining is needed

to maintain accuracy. Our solution is a step forward for more reliable and scientifically sound projections under changing climate.

Performance gains were more pronounced for grasslands under LOYO validation than for croplands, likely reflecting the stronger influence of management interventions and crop rotations that violate LOYO's temporal stationarity assumptions. Nevertheless, the approach relies on long, high-quality datasets and auxiliary covariates, and its empirical nature limits direct mechanistic interpretation. Furthermore, it is important to note that the performance of the meta-modelling framework remains inherently bounded by ensemble-level systematic biases, particularly at data-limited sites such as C1. These limitations were detailed in Section 4.4.

Importantly, this framework opens the door to re-analyzing outputs from major multi-model initiatives like AgMIP (Agricultural Model Intercomparison and Improvement Project; https://agmip.org) and MACSUR (Modelling European Agriculture with Climate Change for Food Security; https://www.faccejpi.net/en/faccejpi/actions/core-theme-1/knowledge-hub-macsur.htm). Rather than requiring new simulations, archived ensemble datasets from these projects could be post-processed using stacking meta-models to enhance predictive skill and extract new diagnostic insights - maximising the value of past investments in large-scale model intercomparison. We strongly encourage international modelling communities to pilot such stacking-based re-analyses, which offer a low-cost, high-impact opportunity to unlock new insights and improve ensemble predictions.

While promising, this approach requires long, high-quality datasets and auxiliary inputs and remains empirical in nature, calling for caution when inferring mechanistic causation. Future research should expand the set of environmental drivers (e.g., soil and hydrological variables), test scalability across broader agroecosystems, and apply advanced interpretability tools. Collaborative development with process-based modellers could translate these statistical gains into mechanistic improvements, ultimately leading to a new generation of hybrid ensemble frameworks for agricultural and grassland biogeochemistry. We encourage the international modelling communities to pilot such stacking-based re-analyses, leveraging their rich archives to unlock new insights and improve ensemble predictions without additional simulation costs.

Competing interest

The authors declare that they have no conflict of interest.

Code and Data availability

The exact versions of the R scripts used to produce the results presented in this paper are available from the GitHub repository: https://github.com/hollorol/metamodeling_of_c under the GPL-3 licence. The input data used to produce the results presented in this paper is archived on the Harvard Dataverse repository under doi:10.7910/DVN/5TO4HE.

Author contribution

RH: Methodology, Conceptualization, Visualization, Writing (original and revised draft preparation)

NZ: Software, Visualization, Writing (original and revised draft preparation)

ZB: Methodology, Writing (original and revised draft preparation)

GB: Methodology, Writing (original and revised draft preparation), Supervision

RS: Data curation, Validation

JR: Conceptualization, Formal analysis

NF: Funding acquisition, Resources, Project administration, Writing (original and revised draft preparation)

Acknowledgements

The present article was published under the auspices of the MACSUR (Modelling European Agriculture with Climate Change for Food Security) Science-Policy Knowledge Forum (MACSUR SciPol Pilot, June 2021-December 2022, and MACSUR SciPolNet, May 2024-April 2026), with the support of the INRAE metaprogramme "Climate change in agriculture and forests: Adaptation and mitigation" (CLIMAE) and INRAE's Public Policy-Support Directorate (DAPP). It falls within the thematic area of the French government IDEX-ISITE initiative (reference: 16-IDEX-0001; project CAP 20-25). This work has been partly implemented by the National Multidisciplinary Laboratory for Climate Change (RRF-2.3.1-21-2022-00014) project within the framework of Hungary's National Recovery and Resilience Plan supported by the Recovery and Resilience Facility of the European Union. This work was also supported by the National Research, Development and Innovation Office, Hungary [project ID: ADVANCED_24 150795]. Also supported by the FK 131813 project, implemented with support provided by the National Research, Development and Innovation Fund of Hungary, financed under the FK_19 funding scheme. Also supported by the TKP2021-NVA-29 project of the Hungarian National Research, Development and Innovation Fund, with the support provided by the Ministry of Culture and Innovation of Hungary. Also supported by the "Advanced methods of greenhouse gases emission reduction and sequestration in agriculture and forest landscape for climate change mitigation" (CZ.02.01.01/00/22_008/0004635) project. RS, RH and GB received mobility funding from the French-Hungarian bilateral partnership through the BALATON (N° 44703TF)/TéT (2019-2.1.11-TÉT-2019-00031) programme. Support was also provided by the TKP2021-NKTA-06 project that has been implemented with the support provided by the Ministry of Innovation and Technology of Hungary from the National Research, Development and Innovation Fund, financed under the [TKP2021-NKTA] funding scheme. Also supported by the AI4Impact programme of the Hungarian Research Network (HUN-REN). The authors are grateful to Gergő Szabó for his support.

References

Anuga, S.W., Chirinda, N., Nukpezah, D., Ahenkan, A., Andrieu, N., Gordon, C.: Towards low carbon agriculture: Systematic-narratives of climate-smart agriculture mitigation potential in Africa. Current Research in Environmental Sustainability 2, 100015. https://doi.org/10.1016/j.crsust.2020.100015 , 2020.

Asseng, S., Ewert, F., Rosenzweig, C., Jones, J.W., Hatfield, J.L., Ruane, A.C., Boote, K.J., Thorburn, P.J., Rötter, R.P., Cammarano, D., Brisson, N., Basso, B., Martre, P., Aggarwal, P.K., Angulo, C., Bertuzzi, P., Biernath, C., Challinor, A.J., Doltra, J., Gayler, S., Goldberg, R., Grant, R., Heng, L., Hooker, J., Hunt, L.A., Ingwersen, J., Izaurralde, R.C., Kersebaum, K.C., Müller, C., Naresh Kumar, S., Nendel, C., O'Leary, G., Olesen, J.E., Osborne, T.M., Palosuo, T., Priesack, E., Ripoche, D., Semenov, M.A., Shcherbak, I., Steduto, P., Stöckle, C., Stratonovitch, P., Streck, T., Supit, I., Tao, F., Travasso, M., Waha, K., Wallach, D., White, J.W., Williams, J.R., Wolf, J.: Uncertainty in simulating wheat yields under climate change. Nature Climate Change 3, 827-832. https://doi.org/10.1038/nclimate1916, 2013.

Bai, Y., Zhang, S., Bhattarai, N., Mallick, K., Liu, Q., Tang, L., Im, J., Guo, L., Zhang, J. : On the use of machine learning based ensemble approaches to improve evapotranspiration estimates from croplands across a wide environmental gradient. Agricultural and Forest Meteoorology 298–299, 108308. https://doi.org/10.1016/j.agrformet.2020.108308, 2021.

Baldocchi, D.: 'Breathing' of the terrestrial biosphere: Lessons learned from a global network of carbon dioxide flux measurement systems. Australian Journal of Botany 56, 1–26. https://doi.org/10.1071/BT07151, 2008.

Baldocchi, D., Chu, H., Reichstein, M.: Intercomparison of ten eddy covariance flux partitioning methods with a global dataset of CO_2 fluxes. Agricultural and Forest Meteorology 256-257, 223-233. https://doi.org/10.1016/j.agrformet.2017.05.015, 2018.

Bansal, Y., Lillis, D., Kechadi, MT.: A neural meta model for predicting winter wheat crop yield. Machine Learning 113: 3771–3788. https://doi.org/10.1007/s10994-023-06455-1, 2024.

Bassu, S., Brisson, N., Durand, J.L., Boote, K.J., Lizaso, J., Jones, J.W., Rosenzweig, C., Adam, M., Basso, B., Baron, C., Basso, B., Biernath, C., Boogaard, H., Conijn, S., Corbeels, M., Deryng, D., De Sanctis, G., Gayler, S., Grassini, P., Hatfield, J., Hoek, S., Izaurralde, C., Jongschaap, R., Kemanian, A.R., Kersebaum, K.C., Kim, S.-H., Kumar, N.S., Makowski, D., Müller, C., Nendel, C., Priesack, E., Pravia, M.V., Sau, F., Shcherbak, I., Tao, F., Teixeira, E., Timlin, D., Waha, K.: How do various maize crop models vary in their responses to climate change factors? Global Change Biology 20, 2301-2320. https://doi.org/10.1111/gcb.12520, 2014.

Beer, C., Reichstein, M., Tomelleri, E., Ciais, P., Jung, M., Carvalhais, N., Rödenbeck, C., Arain, M. A., Baldocchi, D., Bonan, G. B., Bondeau, A., Cescatti, A., Lasslop, G., Lindroth, A., Lomas, M., Luyssaert, S., Margolis, H., Oleson, K.W., Roupsard, O., Veenendaal, E., Viovy, N., Williams, C., Woodward, F.I., Papale, D.: Terrestrial gross carbon dioxide uptake: Global distribution and covariation with climate. Science 329, 834–838. https://doi.org/10.1126/science.1184984, 2010.

Bellocchi, G.: MACSUR SciPol Policy brief 5: Assessing Emissions and Mitigation Practices in Agriculture, towards an effective use of models for policy. Zenodo. https://doi.org/10.5281/zenodo.8038881, 2023.

Bellocchi, G., Barcza, Z., Hollós, R., Acutis, M., Bottyán, E., Doro, L., Hidy, D., Lellei-Kovács, E., Ma, S., Minet, J., Pacskó, V., Perego, A., Ruget, F., Seddaiu, G., Wu, L., Sándor, R.: Sensitivity of simulated soil water content, evapotranspiration, gross primary production and biomass to climate change factors in Euro-Mediterranean grasslands. Agricultural and Forest Meteorology 343, 109778. https://doi.org/10.1016/j.agrformet.2023.109778, 2023.

Bellocchi, G., Rivington, M., Donatelli, M., Matthews, K.: Validation of biophysical models: issues and methodologies. A review. Agronomy for Sustainable Development 30, 109-113. https://doi.org/10.1051/agro/2009001, 2010.

Bilotto, F., Harrison, M.T., Migliorati, M.D.A., Christie, K.M., Rowlings, D.W., Grace, P.R., Smith, A.P., Rawnsley, R.P., Thorburn, P.J., Eckard, R.J.: Can seasonal soil N mineralisation trends be leveraged to enhance pasture growth? Science of the Total Environment 772:145031. https://doi.org/10.1016/j.scitotenv.2021.145031, 2021.

Breiman, L.: Stacked regressions. Machine Learning 24, 123–140. https://doi.org/10.1007/BF00058655, 1996.

Breiman, L.: Stacked regressions. Machine Learning 24, 49–64. https://doi.org/10.1007/BF00117832, 2001a.

Breiman, L.: Random forests. Machine Learning 45, 5–32. https://doi.org/10.1023/A:1010933404324, 2001b.

Brilli, L., Bechini, L., Bindi, M., Carozzi, M., Cavalli, D., Conant, R., Dorich, C.D., Doro, L., Ehrhardt, F., Farina, R., Ferrise, R., Fitton, N., Francaviglia, R., Grace, P., Iocola, I., Klumpp, K., Léonard, J., Martin, R., Massad, R.S., Recous, S., Seddaiu, G., Sharp, J., Smith, P., Smith, W.N., Soussana, J-F., Bellocchi, G.: Review and analysis of strengths and weaknesses of agro-ecosystem models for simulating C and N fluxes. Science of the Total Environ. 598, 445-470. https://doi.org/10.1016/j.scitotenv.2017.03.208, 2017.

Calanca, P., Deléglise, C., Martin, R., Carrère, P., Mosimann, E.: Testing the ability of a simple grassland model to simulate the seasonal effects of drought on herbage growth. Field Crops Research 187, 12-23. https://doi.org/10.1016/j.fcr.2015.12.008, 2016.

Challinor, A.J., Smith, M.S., Thornton, P.: Use of agro-climate ensembles for quantifying uncertainty and informing adaptation. Agricultural and Forest Meteorology 170, 2-7. https://doi.org/10.1016/j.agrformet.2012.09.007, 2013.

Chandel, S., Kleber, M., Jahn, R., Vogel, C.: Soil science-informed machine learning. Geoderma 452, 117094. https://doi.org/10.1016/j.geoderma.2024.117094, 2024.

Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794. https://doi.org/10.1145/2939672.2939785, 2016.

Chergui, N., Kechadi, M.T.: Data analytics for crop management: a big data view. Journal of Big Data 9, 1-37. https://doi.org/10.1186/s40537-022-00668-2, 2022.

Dietterich, T.G.: Ensemble methods in machine learning. In: Multiple classifier systems. MCS 2000. Lecture Notes in Computer Science, vol 1857. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-45014-9_1, 2000.

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J.R.G., Gruber, B., Lafourcade, B., Leitão, P.J., Münkemüller, T., McClean, C., Osborne, P.E., Reineking, B., Schröder, B., Skidmore, A.K., Zurell, D., Lautenbach, S.: Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. Ecography 36, 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x, 2013.

Ehrhardt, F., Soussana, J.-F., Bellocchi, G., Grace, P., McAuliffe, R., Recous, S., Sándor, R., Smith, P., Snow, V., Migliorati, M.D.A., Basso, B., Bhatia, A., Brilli, L., Doltra, J., Dorich, C.D., Doro, L., Fitton, N., Giacomini, S.J., Grant, B., Harrison, M.T., Jones, S.K., Kirschbaum, M.U.F., Klumpp, K., Laville, P., Léonard, J., Liebig, M., Lieffering, M., Martin, R., Massad, R.S., Meier, E., Merbold, L., Moore, A.D., Myrgiotis, V., Newton, P., Pattey, E., Rolinski, S., Sharp, J., Smith, W.N., Wu, L., Zhang, Q.: Assessing uncertainties in crop and pasture ensemble model simulations of productivity and N2O emissions. Global Change Biology 24, e603-e616. https://doi.org/10.1111/gcb.13965, 2018.

Eyring, V., Cox, P.M., Flato, G.M., Friedlingstein, P., Hall, A., Hawkins, E., Hewitt, H.T., Joshi, M., Klein, S.A., Knutti, R., Meehl, G.A., O'Neill, B.C., Piani, C., Raper, S.C.B., Riahi, K., Roeckner, E., Sanderson, B.M., Wenzel, S.: Taking climate model evaluation to the next level. Nature Climate Change 9, 102–110. https://doi.org/10.1038/s41558-018-0355-y, 2019.

Farina, R., Sándor, R., Abdalla, M., Álvaro-Fuentes, J., Bechini, L., Bolinder, M.A., Brilli, L., Chenu, C., Clivot, H., De Antoni Migliorati, M., Di Bene, C., Dorich, C.D., Ehrhardt, F., Ferchaud, F., Fitton, N., Francaviglia, R., Franko, U., Giltrap, D.L., Grant, B.B., Guenet, B., Harrison, M.T., Kirschbaum, M.U.F., Kuka, K., Kulmala, L., Liski, J., McGrath, M.J., Meier, E., Menichetti, L., Moyano, F., Nendel, C., Recous, S., Reibold, N., Shepherd, A., Smith, W.N., Smith, P., Soussana, J.-F., Stella, T., Taghizadeh-Toosi, A., Tsutskikh, E., Bellocchi, G.: Ensemble modelling, uncertainty and robust predictions of organic carbon in long-term bare-fallow soils. Global Change Biology 27, 904-928. https://doi.org/10.1111/gcb.15441, 2021.

Fatichi, S., Pappas, C., Zscheischler, J., Leuzinger, S.: Modelling carbon sources and sinks in terrestrial vegetation. New Phytologist 221, 652-668. https://doi.org/10.1111/nph.15451, 2019.

Gascuel-Odoux, C., Lescourret, F., Dedieu, B., Detang-Dessendre, C., Faverdin, P., Hazard, L., Litrico-Chiarelli, I., Petit, S., Roques, L., Reboud, X., Tixier-Boichard, M., de Vries, H., Caquet, T.: A research agenda for scaling up agroecology in European countries. Agronomy for Sustainable Development 42, 53. https://doi.org/10.1007/s13593-022-00786-4, 2022.

Granger, C.W.J., Ramanathan R.: Improved methods of combining forecasts. Journal of Forecasting 3:197–204. https://doi.org/10.1002/for.3980030207, 1984.

Hagedorn, R., Doblas-Reyes, F.J., Palmer, T.N.: The rationale behind the success of multi-model ensembles in seasonal forecasting – I. Basic concept. Tellus A: Dynamic Meteorology and Oceanography 57, 219–233. https://doi.org/10.1111/j.1600-0870.2005.00103.x, 2005.

Hansen, L.K., Salamon, P.: Neural network ensembles. IEEE Transactions on Pattern Analysis and Machine Intelligence 12, 993–1001. https://doi.org/10.1109/34.58871, 1990.

Harrison, M.T., Evans, J.R., Moore, A.D.: Using a mathematical framework to examine physiological changes in winter wheat after livestock grazing: 1. Model derivation and coefficient calibration. Field Crops Research 136, 116-126. https://doi.org/10.1016/j.fcr.2012.06.015, 2012.

Hengl, T., de Jesus, J.M., Heuvelink, G.B., Gonzalez, M.R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., Guevara, M.A., Vargas, R., MacMillan, R.A., Batjes, N.H., Leenaars, J.G.B., Ribeiro, E., Wheeler, I., Mantel, S., Kempen, B.: SoilGrids250m: Global gridded soil information based on machine learning. PLoS ONE 12, e0169748. https://doi.org/10.1371/journal.pone.0169748, 2018.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A.: Very high-resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25, 1965-1978. https://doi.org/10.1002/joc.1276, 2005.

Hou, J., Hou, B.: Farmers' adoption of low-carbon agriculture in China: An extended theory of the planned behavior model. Sustainability 11, 1399. https://doi.org/10.3390/su11051399, 2019.

Jackson, R. B., Lajtha, K., Crow, S. E., Hugelius, G., Kramer, M. G., Rhodes, C.: The ecology of soil carbon: pools, vulnerabilities, and biotic and abiotic controls. Annual Review of Ecology, Evolution, and Systematics 48, 419-445. https://doi.org/10.1146/annurev-ecolsys-112414-054234

Janes-Bassett, V., Davies, J., Rowe, Ed C., Tipping, E., 2020. Simulating long-term carbon nitrogen and phosphorus biogeochemical cycling in agricultural environments. Science of the Total Environment 714, 136599. https://doi.org/10.1016/j.scitotenv.2020.136599, 2017.

Jones, J.W., Antle, J.M., Basso, B.O., Boote, K.J., Conant, R.T., Foster, I., Godfray, H.C.J., Herrero, M., Howitt, R.E., Janssen, S., Keating, B.A., Muñoz-Carpena, R., Porter, C.H., Rosenzweig, C., Wheeler, T.R.: Brief history of agricultural systems modelling. Agricultural Systems 155, 240-254. https://doi.org/10.1016/j.agsy.2016.05.014, 2017.

Jung, M., Vetter, M., Herold, M., Churkina, G., Reichstein, M., Zaehle, S., Ciais, P., Viovy, N., Bondeau, A., Chen, Y., Trusilova, K., Feser, F., Heimann, M.: Uncertainties of modelling gross primary productivity over Europe: A systematic study on the effects of using different drivers and terrestrial biosphere models. Global Biogeochemical Cycles 21, GB4021. https://doi.org/10.1029/2006GB002915, 2007.

Keskin, H., Grunwald, S., Basso, B.: Machine learning advances to predict crop yield for agricultural systems. Soil Science Society of America Journal 83, 1521-1531. https://doi.org/10.2136/sssaj2019.06.0203, 2019.

Knutti, R., Baumberger, C., Hirsch Hadorn, G.: Uncertainty quantification using multiple models - prospects and challenges. In: Beisbart C., Saam N.J. (eds.) Computer simulation validation: fundamental concepts, methodological frameworks, and philosophical perspectives. Springer: Cham, pp. 835–855, 2019.

Kobayashi, K., Salam, M.U.: Comparing simulated and measured values using mean squared deviation and its components. Agronomy Journal 92, 345-352. https://doi.org/10.2134/agronj2000.922345x, 2000.

Kollas, C., Kersebaum, K.C., Nendel, C., Manevski, K., Müller, C., Palosuo, T., Armas-Herrera, C.M., Beaudoin, N., Bindi, M., Charfeddine, M., Conradt, T., Constantin, J., Eitzinger, J., Ewert, F., Ferrise, R., Gaiser, T., Garcia de Cortazar-Atauri, I., Giglio, L., Hlavinka, P., Hoffmann, H., Hoffmann, M.P., Launay, M., Manderscheid, R., Mary, B., Mirschel, W., Moriondo, M., Olesen, J.E. Öztürk, I., Pacholski, A., Ripoche-Wachter, D., Roggero, P.P., Roncossek, S., Rötter, R.P., Ruget, F., Sharif, B., Trnkam, M., Ventrella, D., Waha, K., Wegehenkel, M., Weigel, H.-J., Wu, L.: Crop rotation modelling - A European model intercomparison. European Journal of Agronomy 70, 98–111. https://doi.org/10.1016/j.eja.2015.06.007, 2015.

Koster, R.D., Dirmeyer, P.A., Guo, Z., Bonan, G., Chan, E., Cox, P., Gordon, C.T., Kanae, S., Kowalczyk, E., Lawrence, D., Liu, P., Lu, C.-H., Malyshev, S., McAvaney, B., Mitchell, K., Mocko, D., Oki, T., Oleson, K., Pitman, A., Sud, Y. C., Taylor, C. M., Verseghy, D., Vasic, R., Xue, Y., Yamada, T.: Regions of strong coupling between soil moisture and precipitation. Science 305, 1138–1140. https://doi.org/10.1126/science.1100217, 2004.

Kutner, M.H., Nachtsheim, C.J., Neter, J. and Li, W.: Applied linear statistical models. 5th Edition, McGraw-Hill, Irwin, New York, 2005.

Lambin, E.F., Meyfroidt, P.: Global land use change, economic globalization, and the looming land scarcity. Proceedings of the National Academy of Sciences 108, 3465-3472. https://doi.org/10.1073/pnas.1100480108, 2011.

Lembaid, I., Moussadek, R., Mrabet, R., Bouhaouss, A.: Modeling soil organic carbon changes under alternative climatic scenarios and soil properties using DNDC model at a semi-arid Mediterranean environment. Climate 10, 23. https://doi.org/10.3390/cli10020023, 2022.

Lembaid, I., Moussadek, R., Mrabet, R., Doauik, A., Bouhaouss, A.: Modeling the effects of farming management practices on soil organic carbon stock under two tillage practices in a semi-arid region, Morocco. Heliyon. 7, e05889. https://doi.org/10.1016/j.heliyon.2020.e05889, 2021.

Li, T., Cui, L., Kuhnert, M., McLaren, T.I., Pandey, R., Liu, H., Wang, W., Xu, Z., Xia, A., Dalal, R.C., Dang, Y.P.: A comprehensive review of soil organic carbon estimates: Integrating remote sensing and machine learning technologies. Journal of Soils and Sediments 24, 3556-3571. https://doi.org/10.1007/s11368-024-03913-8, 2015.

Li, C., Farahbakhshazad, N., Jaynes, D.B., Dinnes, D.L., Salas, W., McLaughlin, D.: Modeling nitrate leaching with a biogeochemical model modified based on observations in a row-crop field in Iowa. Ecological Modelling 196, 116-130. https://doi.org/10.1016/j.ecolmodel.2006.02.007, 2006.

Li, T., Hasegawa, T., Yin, X., Zhu, Y., Boote, K., Adam, M., Bregalgio, S., Buis, S., Confalonieri, R., Fumoto T., Gaydon, D., Marcaida III, M., Nakagawa, H., Oriol, P., Ruane, A.C., Ruget, F., Balwinder-Singh, B., Singh, U., Tang, L., Tao, F., Wilkens, P., Yoshida, H., Zhang, Z., Bouman, B.: Uncertainties in predicting rice yield by current crop models under a wide range of climatic conditions. Global Change Biology 21, 1328–1341. https://doi.org/10.1111/gcb.12758, 2015.

Liaw, A., Wiener, M.: Classification and regression by randomForest. R News 2, 18-22. http://CRAN.R-project.org/doc/Rnews, 2002.

Lobell, D.B., Schlenker, W., Costa-Roberts, J.: Climate trends and global crop production since 1980. Science 333, 616-620. https://doi.org/10.1126/science.1204531, 2011.

Lloyd, J., Taylor, J.A.: On the temperature dependence of soil respiration. Functional Ecology 8, 315–323. https://doi.org/10.2307/2389824, 1994.

Lundberg, S.M., Erion, G., Lee, S.-I.: Consistent individualized feature attribution for tree ensembles. arXiv:1802.03888. https://doi.org/10.48550/arXiv.1802.03888, 2020.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, 4-9 December, pp. 4766-4777.

Luo, Y., Weng, E., Wu, X., Gao, C., Zhou, X., Zhang, L.: Parameter identifiability, constraint, and equifinality in data assimilation with ecosystem models. Ecological Applications 19, 571-574. https://doi.org/10.1890/08-0561.1, 2009.

Mangani, R., Tesfamariam, E., Engelbrecht, C.J., Bellocchi, G., Hassen, A., Mangani, T.: Potential impacts of extreme weather events in main maize (*Zea mays* L.) producing areas of South Africa under rainfed conditions. Regional Environmental Change 19, 1441-1452. https://doi.org/10.1007/s10113-019-01486-8, 2019.

Martre, P., Wallach, D., Asseng, S., Ewert, F., Jones, J.W., Rotter, R.P., Boote, K.J., Ruane, A.C., Thorburn, P.J., Cammarano, D., Hatfield, J.L., Rosenzweig, C., Aggarwal, P.K., Angulo, C., Basso, B., Bertuzzi, P., Biernath, C., Brisson, N., Challinor, A.J., Doltra, J., Gayler, S., Goldberg, R., Grant, R.F., Heng, L., Hooker, J., Hunt, L.A., Ingwersen, J., Izaurralde, R.C., Kersebaum, K.C., Müller, C., Kumar, S.N., Nendel, C., O'leary, G., Olesen, J.E., Osborne, T.M., Palosuo, T., Priesack, E., Ripoche, D., Semenov, M.A., Shcherback, I., Steduto, P., Stöckle, C.O., Stratonovitch, P., Streck, T., Supit, I., Tao, F., Travasso, M., Waha, K., White, J.W., Wolf, J., 2015. Multimodel ensembles of wheat growth: many models are better than one. Global Change Biology 21, 911-925.

Mathieu, J.A., Aires, F.: Assessment of the agro-climatic indices to improve crop yield forecasting. Agricultural and Forest Meteorology 253-254:15-30. https://doi.org/10.1016/j.agrformet.2018.01.031, 2018.

Nand, V., Qi, Z., Ma, L., Helmers, M.J., Madramootoo, C.A., Smith, W.N., Zhang, T., Weber, T.K.D., Pattey, E., Li, Z., Wang, J., Jin, V.L., Jiang, Q., Tenuta, M., Trout, T.J., Cheng, H., Harmel, R.D., Kimball, B.A., Thorp, K.R., Boote, K.J., Stockle, C., Suyker, A.E., Evett, S.R., Brauer, D.K., Coyle, G.G., Copeland, K.S., Marek, G.W., Colaizzi, P.D., Acutis, M., Alimagham, S.M., Archontoulis, S., Babacar, F., Barcza, Z., Basso, B., Bertuzzi, P., Constantin, J., De Antoni Migliorati, M., Dumont, B., Durand, J.L., Fodor, N., Gaiser, T., Garofalo, P., Gayler, S., Giglio, L., Grant, R., Guan, K., Hoogenboom, G., Kim, S.H., Kisekka, I., Lizaso, J., Masia, S., Meng, H., Mereu, V., Mukhtar, A., Perego, A., Peng, B., Priesack, E., Shelia, V., Snyder, R., Soltani, A., Spano, D., Srivastava, A., Thomson, A., Timlin, D., Trabucco, A., Webber, H., Willaume, M., Williams, K., van der Laan, M., Ventrella, D., Viswanathan, M., Xu, X., Zhou, W.: Evaluation of multimodel averaging approaches for ensembling evapotranspiration and yield simulations from maize models. Journal of Hydrology 661: 133631. https://doi.org/10.1016/j.jhydrol.2025.133631, 2025.

Olesen, J.E., Bindi, M.: Consequences of climate change for European agricultural productivity, land use and policy. European Journal of Agronomy 16, 239-262. https://doi.org/10.1016/S1161-0301(02)00004-7, 2002.

Opitz, D., Maclin, R.: Popular ensemble methods: An empirical study. Journal of Artificial Intelligence Research 11, 169–198. https://doi.org/10.1613/jair.614, 1999.

Ostle, N.J., Smith, P., Fisher, R., Woodward, F.I., Fisher, J.B., Smith, J.U., Galbraith, D., Levy, P., Meir, P., McNamara, N.P., Bardgett, R.D.: Integrating plant–soil interactions into global carbon cycle models. Journal of Ecology 97, 851–863. https://doi.org/10.1111/j.1365-2745.2009.01547.x, 2009.

Pappas, C., Papalexiou, S.M., Koutsoyiannis D.: A quick gap filling of missing hydrometeorological data. Journal of Geophysical Research Atmosphere 119, 9290–9300. https://doi.org/10.1002/2014JD021633, 2014.

Raj, R., Hamm, N.A.S., van de Tol, C., Stein, A.: Uncertainty analysis of gross primary production partitioned from net ecosystem exchange measurements. Biogeosciences 13, 1409-1422. https://doi.org/10.5194/bg-13-1409-2016, 2006.

Reichstein, M., Bahn, M., Ciais, P., Frank, D., Mahecha, M. D., Seneviratne, S. I., Zscheischler, J., Beer, C., Buchmann, N., Frank, D. C., Papale, D., Rammig, A., Smith, P., Thonicke, K., van der Velde, M., Vicca, S., Walz, A., Wattenbach, M.: Climate extremes and the carbon cycle. Nature 500, 287–295. https://doi.org/10.1038/nature12350, 2013.

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

Riccio, G., Giunta, G., Galmarini, S.: Seeking for the rational basis of the Median Model: the optimal combination of multi-model ensemble results. Atmospheric Chemistry and Physics 7, 6085-6098. https://doi.org/10.5194/acp-7-6085-2007, 2007.

Richter, K., Atzberger, C., Hank, T.B., Mauser, W.: Derivation of biophysical variables from Earth observation data: validation and statistical measures. Journal of Applied Remote Sensing 6, 063557. https://doi.org/10.1117/1.JRS.6.063557, 2012.

Reichstein, M., Ciais, P., Papale, D., Valentini, R., Running, S., Viovy, N., Cramer, W., Granier, A., Ogée, J., Allard, V., Aubinet, M., Bernhofer, C., Buchmann, N., Carrara, A., Grünwald, T., Heimann, M., Heinesch, B., Knohl, A., Kutsch, W., Loustau, D., Manca, G., Matteucci, G., Miglietta, F., Ourcival, J.M., Pilegaard, K., Pumpanen, J., Rambal, S., Schaphoff, S., Seufert, G., Soussana, J.-F., Sanz, M-J., Vesala, T., Zhao, M.: Reduction of ecosystem productivity and respiration during the European summer 2003 climate anomaly: A joint flux tower, remote sensing and modelling analysis. Global Change Biology 13, 634-651. https://doi.org/10.1111/j.1365-2486.2006.01224.x, 2007.

Reichstein, M., Falge, E., Baldocchi, D., Papale, D., Aubinet, M., Berbigier, P., Bernhofer, C., Buchmann, N., Gilmanov, T., Granier, A., Grünwald, T., Havránková, K., Ilvesniemi, H., Janous, D., Knohl, A., Laurila, T., Lohila, A., Loustau, D., Matteucci, G., Meyers, T., Miglietta, F., Ourcival, J.-M., Pumpanen, J., Rambal, S., Rotenberg, E., Sanz, M., Tenhunen, J., Seufert, G., Vaccari, F., Vesala, T., Yakir, D., Valentini, R.: On the separation of net ecosystem exchange into assimilation and ecosystem respiration: Review and improved algorithm. Global Change Biology 11, 1424–1439. https://doi.org/10.1111/j.1365-2486.2005.001002.x, 2005.

Robeson, S.M., Willmott, C.J.: Decomposition of the mean absolute error (MAE) into systematic and unsystematic components. PLoS ONE 18, e0279774. https://doi.org/10.1371/journal.pone.0279774, 2023.

Rosenzweig, C., Elliott, J., Deryng, D., Ruane, A.C., Müller, C., Arneth, A., Boote, K.J., Folberth, C., Glotter, M., Khabarov, N., Neumann, K., Piontek, F., Pugh, T.A.M., Schmid, E., Stehfest, E., Yang, H., Jones, J.W.: Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison. Proceedings of the National Academy of Sciences of the United States of America 111, 3268-3273. https://doi.org/10.1073/pnas.1222463110, 2014.

Ruane, A.C., Hudson, N.I., Asseng, S., Camarrano, D., Ewert, F., Martre, P., Boote, K.J., Thorburn, P.J., Aggarwal, P.K., Angulo, C., Basso, B., Bertuzzi, P., Biernath, C., Brisson, N., Challinor, A.J., Doltra, J., Gayler, S., Goldberg, R., Grant, R.F., Heng, L., Hooker, J., Hunt, L.A., Ingwersen, J., Izaurralde, R.C., Kersebaum, K.C., Kumar, S.N., Müller, C., Nendel, C., O'Leary, G., Olesen, J.E., Osborne, T.M., Palosuo, T., Priesack, E., Ripoche, D., Rötter, R.P., Semenov, M.A., Shcherbak, I., Steduto, P., Stöckle, C.O., Stratonovitch, P., Streck, T., Supit, I., Tao, F., Travasso, M., Waha, K., Wallach, D., White, J.W., Wolf, J.: Multi-wheat-model ensemble responses to interannual climate variability. Environmental Modelling & Software 81, 86-101. https://doi.org/10.1016/j.envsoft.2016.03.008, 2016.

Ruane, A.C., Rosenzweig, C., Asseng, S., Boote, K.J., Elliott, J., Ewert, F., Jones, J.W., Martre, P., McDermid, S.P., Müller, C., Snyder, A., Thorburn, P.J.: An AgMIP framework for improved agricultural representation in integrated assessment models. Environmental Research Letters 12, 125003. https://doi.org/10.1088/1748-9326/aa8da6, 2017.

Sagi, O., Rokach, L.: Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8, e1249. https://doi.org/10.1002/widm.1249, 2018.

Sándor, R., Barcza, Z., Acutis, M., Doro, L., Hidy, D., Köchy, M., Minet, J., Lellei-Kovács, E., Ma, S., Perego, A., Rolinski, S., Ruget, F., Sanna, M., Seddaiu, G., Wu, L., Bellocchi, G.: Multi-model simulation of soil temperature, soil water content and biomass in Euro-Mediterranean grasslands: Uncertainties and ensemble performance. European Journal of Agronomy 88, 22-40. https://doi.org/10.1016/j.eja.2016.06.006, 2017.

Sándor, R., Ehrhardt, F., Basso, B., Bellocchi, G., Bhatia, A., Brilli, L., Migliorati, M.D., Doltra, J., Dorich, C., Doro, L., Fitton, N., Giacomini, S.J., Grace, P., Grant, B., Harrison, M.T., Jones, S., Kirschbaum, M.U.F., Klumpp, K., Laville, P., Léonard, J., Liebig, M., Lieffering, M., Martin, R., McAuliffe, R., Meier, E., Merbold, L., Moore, A., Myrgiotis, V., Newton, P., Pattey, E., Recous, S., Rolinski, S., Sharp, J., Massad, R.S., Smith, P., Smith, W., Snow, V., Wu, L., Zhang, Q., Soussana, J.-F.: C and N models Intercomparison – benchmark and ensemble model estimates for grassland production. Advances in Animal Biosciences 7, 245-247. https://doi.org/10.1017/S2040470016000297, 2016.

Sándor, R., Ehrhardt, F., Brilli, L., Carozzi, M., Recous, S., Smith, P., Snow, V., Soussana, J.F., Dorich, C.D., Fuchs, K., Fitton, N., Gongadze, K., Klumpp, K., Liebig, M., Martin, R., Merbold, L., Newton, P.C.D., Rees, R.M., Rolinski, S., Bellocchi, G.: The use of biogeochemical models to evaluate mitigation of greenhouse gas emissions from managed grasslands. Science of the Total Environment 15, 292-306. https://doi.org/10.1016/j.scitotenv.2018.06.020, 2018.

Sándor, R., Ehrhardt, F., Grace, P., Recous, S., Smith, P., Snow, V., Soussana, J.-F., Basso, B., Bhatia, A., Brilli, L., Doltra, J., Dorich, C.D., Doro, L., Fitton, N., Grant, B., Harrison, M.T., Kirschbaum, M.U.F., Klumpp, K., Laville, P., Léonard, J., Martin, R., Massad, R.S., Moore, A., Myrgiotis, V., Pattey, E., Rolinski, R., Sharp, J., Skiba,

U., Smith, W., Wu, L., Zhang, Q., Bellocchi, G.: Ensemble modelling of carbon fluxes in grasslands and croplands. Field Crops Research 252, 107791. https://doi.org/10.1016/j.fcr.2020.107791, 2020.

Sándor, R., Ehrhardt, F., Grace, P., Recous, S., Smith, P., Snow, V., Soussana, J.-F., Basso, B., Bhatia, A., Brilli, L., Doltra, J., Dorich, C.D., Doro, L., Fitton, N., Grant, B., Harrison, M.T., Kirschbaum, M.U.F., Klumpp, K., Laville, P., Léonard, J., Martin, R., Massad, R.S., Moore, A., Myrgiotis, V., Pattey, E., Rolinski, R., Sharp, J., Skiba, U., Smith, W., Wu, L., Zhang, Q., Bellocchi, G.: Experimental and simulated data for crop and grassland production and carbon-nitrogen fluxes. Open Data Journal for Agricultural Research 24, 22-27. Available at: https://odjar.org/article/view/18594, 2024.

Sándor, R., Ehrhardt, F., Grace, P., Recous, S., Smith, P., Snow, V., Soussana, J.-F., Basso, B., Bhatia, A., Brilli, L., Doltra, J., Dorich, C.D., Doro, L., Fitton, N., Grant, B., Harrison, M.T., Skiba, U., Kirschbaum, M.U.F., Klumpp, K., Laville, P., Léonard, J., Martin, R., Massad, R.S., Moore, A., Myrgiotis, V., Pattey, E., Rolinski, R., Sharp, J., Smith, W., Wu, L., Zhang, Q., Bellocchi, G.: Residual correlation and ensemble modelling to improve crop and grassland models. Environmental Modelling & Software 161, 105625. https://doi.org/10.1016/j.envsoft.2023.105625, 2023.

Schwalm, C R., Williams, C.A., Schaefer, K., Arneth, A., Bonal, D., Buchmann, N., Chen, J., Law, B.E., Lindroth, A., Luyssaert, S., Reichstein, M., Richardson, A.D.: Assimilation exceeds respiration sensitivity to drought: A FLUXNET synthesis. Global Change Biology 16, 657–670. https://doi.org/10.1111/j.1365-2486.2009.01991.x, 2010.

Scowen, M., Athanasiadis, I. N., Bullock, J. M., Eigenbrod, F., Willcock, S.: The current and future uses of machine learning in ecosystem service research. Science of the Total Environment 799, 149263. https://doi.org/10.1016/j.scitotenv.2021.14926, 2021.

Shahhosseini, M., Hu, G., Archontoulis, S.V.: Forecasting corn yield with machine learning ensembles. Frontiers in Plant Science 11, 1120. https://doi.org/10.3389/fpls.2020.01120, 2020.

Shahhosseini, M., Hu, G., Huber, I., Archontoulis, S.V.: Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. Scientific Reports 11, 1606. https://doi.org/10.1038/s41598-020-80820-1, 2021.

Shapley, L.S.: A value for n-person games. In: Kuhn H.W., Tucker A.W. (eds.), Contributions to the theory of games II (Vol. 28, pp. 307–317). Princeton University Press. 1953.

Smith, P., House, J.I., Bustamante, M., Sobocká, J., Harper, R., Pan, G., West, P.C., Clark, J.M., Adhya, T., Rumpel, C., Paustian, K., Kuikman, P., Cotrufo, M.F., Elliott, J.A., McDowell, R., Griffiths, R.I., Asakawa, S., Bondeau, A., Jain, A.K., Meersmans, J., Pugh, T.A.M.: Global change pressures on soils from land use and management. Global Change Biology 22, 1008-1028. https://doi.org/10.1111/gcb.13068, 2016.

Snow, V., Rotz, C.A., Moore, A.D., Martin-Clouaire, R., Johnson, I.R., Hutchings, N.J., Eckard, R.J.: The challenges - and some solutions - to process-based modelling of grazed agricultural systems. Environmental Modelling & Software 62, 420-436. https://doi.org/10.1016/j.envsoft.2014.03.009, 2014.

Sroufe, R., Watts, A., 2022. Pathways to agricultural decarbonization: Climate change obstacles and opportunities in the US. Resources, Conservation and Recycling 182, 106276. https://doi.org/10.1016/j.resconrec.2022.106276

Stoy, P.C., Mauder, M., Foken, T., Marcolla, B., Boegh, E., Ibrom, A., Arain, M.A., Arneth, A., Aurela, M., Bernhofer, C., Cescatti, A., Dellwik, E., Duce, P., Gianelle, D., van Gorsel, E., Kiely, G., Knohl, A., Margolis, H., McCaughey, H., Merbold, L., Montagnani. L., Papale, D., Reichstein, M., Saunders, M., Serrano-Ortiz, P., Sottocornola, M., Spano, D., Vaccari, F., Varlagin, A.: A data-driven analysis of energy balance closure across FLUXNET research sites: The role of landscape scale heterogeneity. Agricultural and Forest Meteorology 171-172, 137-152. https://doi.org/10.1016/j.agrformet.2012.11.004, 2013.

Strobl, C., Boulesteix, A. L., Zeileis, A., Hothorn, T.: Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics 8, 25. https://doi.org/10.1186/1471-2105-8-25, 2007.

Sun, W., Zhou, S., Yu, B., Zhang, Y., Keenan, T.F., Fu, B.: Soil moisture-atmosphere interactions drive terrestrial carbon-water trade-offs. Communications Earth & Environment 6, 169. https://doi.org/10.1038/s43247-025-02145-z, 2025.

Therond, O., Hengsdijk, H., Casellas, E., Wallach, D., Adam, M., Belhouchette, H., Oomen, R., Russell, G., Ewert, F., Bergez, J.-E., Janssen, S., Wery, J., van Ittersum, M.K.: Using a cropping system model at regional scale: low-data approaches for crop management information and model calibration. Agriculture, Ecosystems & Environment 142, 85-94. https://doi.org/10.1016/j.agee.2010.05.007, 2011.

Thornton, P.K., Whitbread, A., Baedeker, T., Cairns, J., Claessens, L., Beethgen, W., Bunn, C., Friedmann, M., Giller, K.E., Herrero, M., Howden, M., Kilcline, K., Nangia, V., Ramirez-Villegas, J., Kumar, S., West, P.C., Keating, B.: A framework for priority-setting in climate smart agriculture research. Agricultural Systems 167, 161-175. https://doi.org/10.1016/j.agsy.2018.09.009, 2018.

Valin, H., Havlík, P., Mosnier, A., Herrero, M., Schmid, E., Obersteiner, M.:. Agricultural productivity and greenhouse gas emissions: trade-offs or synergies between mitigation and food security? Environmental Research Letters 8, 035019. https://doi.org/10.1088/1748-9326/8/3/035019, 2013

Van der Laan, M.J., Polley, E.C., Hubbard, A.E.: Super learner. Statistical Applications in Genetics and Molecular Biology 6, 25. https://doi.org/10.2202/1544-6115.1309, 2007.

Van der Velde, M., Tubiello, F.N., Vrieling, A., Bouraoui, F.: Impacts of extreme weather on wheat and maize in France: Evaluating regional crop simulations against observed data. Climatic Change 123, 699–711. https://doi.org/10.1007/s10584-011-0368-2, 2014.

Wallach, D., Martre, P., Liu, B., Asseng, S., Ewert, F., Thonburn, P.J., van Ittersum, M., Aggarwal, P.K., Ahmed, M., Basso, B., Biernath, C., Cammarano, D., Challinor, A.J., De Sanctis, G., Dumont, B., Rezaei, E.E., Fereres, E., Fitzgerald, G.J., Gao, Y., Garcia-Vila, M., Gayler, S., Girousse, C., Hoogenboom, G., Horan, H., Izaurralde, R.C., Jones, C.D., Kassie, B.T., Kersebaum, K.C., Klein, C., Koehler, A.-K., Maiorano, A., Minoli, S., Müller, C., Kumar, S.N., Nendel, C., O'Leary, G.J., Palosuo, T., Priesack, E., Ripoche, D., Rötten, R.P., Semenov, M.A., Stöckle, C., Stratonovitch, P., Streck, T., Supit, I., Fao, F., Wolf, J., Zhang, Z.: Multi-model ensembles improve predictions of crop-environment-management interactions. Global Change Biology 24, 5072-5083. https://doi.org/10.1111/gcb.14411, 2018.

Wang, Z., Liu, Z., Huang, M.: NDVI joint process-based models drive a learning ensemble model for accurately estimating cropland net primary productivity (NPP). Frontiers in Environmental Science 11, 1304400. https://doi.org/10.3389/fenvs.2023.1304400, 2024.

Wolpert, D.H., Macready, W.G.: Coevolutionary free lunches. IEEE Transactions on Evolutionary Computation 9, 721–35. https://doi.org/10.1109/TEVC.2005.856205, 2005.

Xia, J., Chen, T., Zhang, K., Wang, Y., Chen, G.: Impacts of climate extremes on carbon fluxes and their underlying mechanisms in a typical temperate grassland ecosystem. Science of the Total Environment 907, 167755. https://doi.org/10.1016/j.scitotenv.2023.167755, 2024.

Xu, L., Baldocchi, D.D.: Seasonal variation in carbon dioxide exchange over a Mediterranean annual grassland in California. Agricultural and Forest Meteorology 123, 79–96. https://doi.org/10.1016/j.agrformet.2003.10.004, 2004.

Xu, X., Xu, J., Li, B., Li, J., Nie, M.: Ecosystem carbon fluxes exhibit thermal response thresholds at which carbon–climate feedback changes. Global Ecology and Biogeography 34, e70030. https://doi.org/10.1111/geb.70030, 2025.

Zhang, J., Tian, H., Wang, P., Tansey, K., Zhang, S., Li, H.: Improving wheat yield estimates using data augmentation models and remotely sensed biophysical indices within deep neural networks in the Guanzhong Plain, PR China. Computers and Electronics in Agriculture 192, 106616. https://doi.org/10.1016/j.compag.2021.106616, 2022.

Zhu, Z., Piao, S., Myneni, R.B., Huang, M., Zeng, Z., Canadell, J. G., Ciais, P., Sitch, S., Friedlingstein, P., Arneth, A., Cao, C., Cheng, L., Kato, E., Koven, C., Li, Y., Lian, X., Liu, Y., Liu, R., Mao, J., Pan, Y., Peng, S., Peñuelas, J., Poulter, B., Pugh, T.A.M., Stocker, B. D., Viovy, N., Wang, X., Wang, Y., Xiao, Z., Yang, H., Zaehle, S., Zeng, N.: Greening of the Earth and its drivers. Nature Climate Change 6, 791–795. https://doi.org/10.1038/nclimate3004, 2016.

1315

1320

Appendix A

The following figures demonstrate the simulated fluxes of GPP, RECO and NEE for all sites and years included in the study, except those that are presented in the main text. Graphs are presented with a very short caption only containing: flux type, site name and type, covered years. See the main text for explanations.

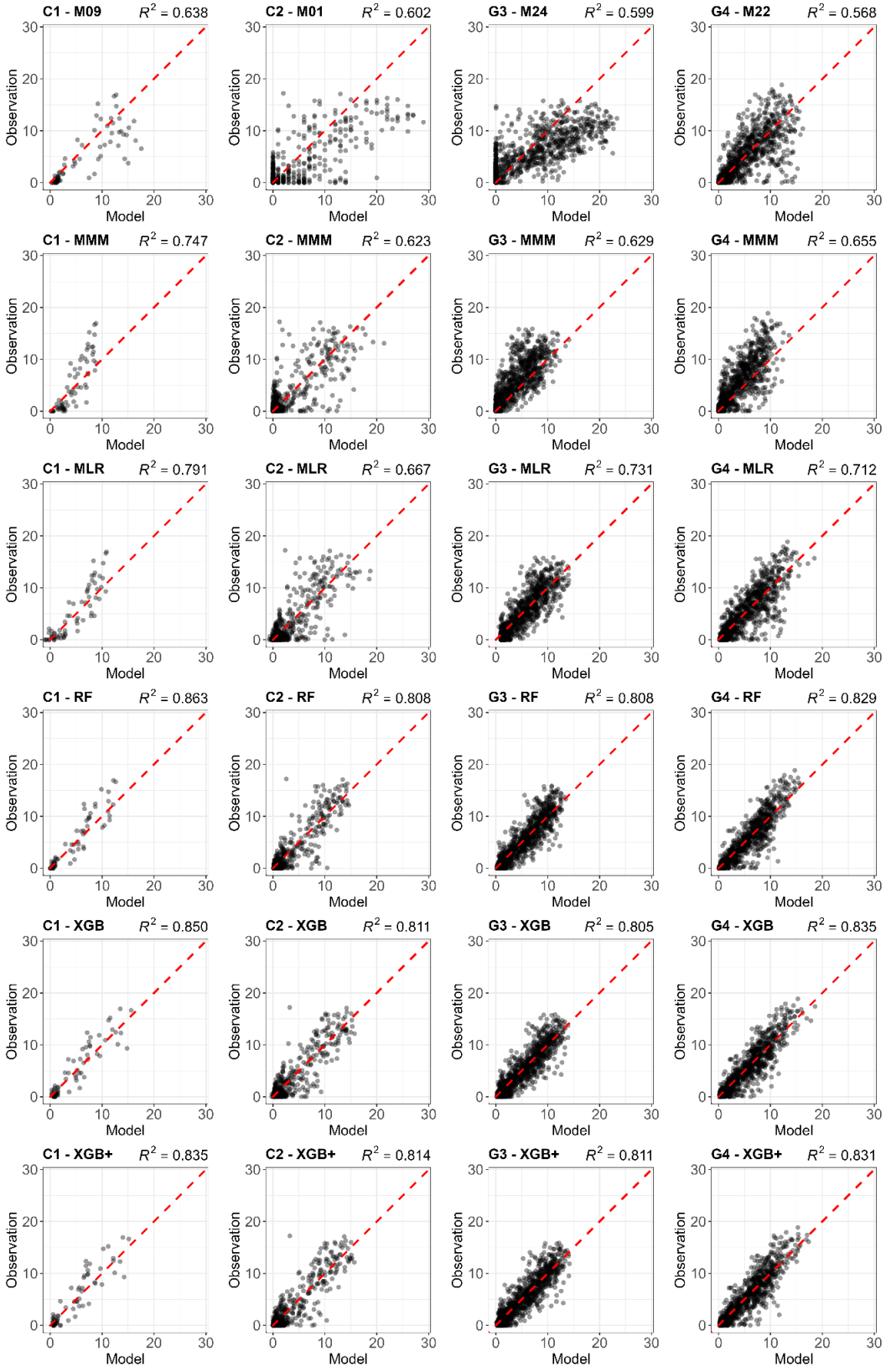
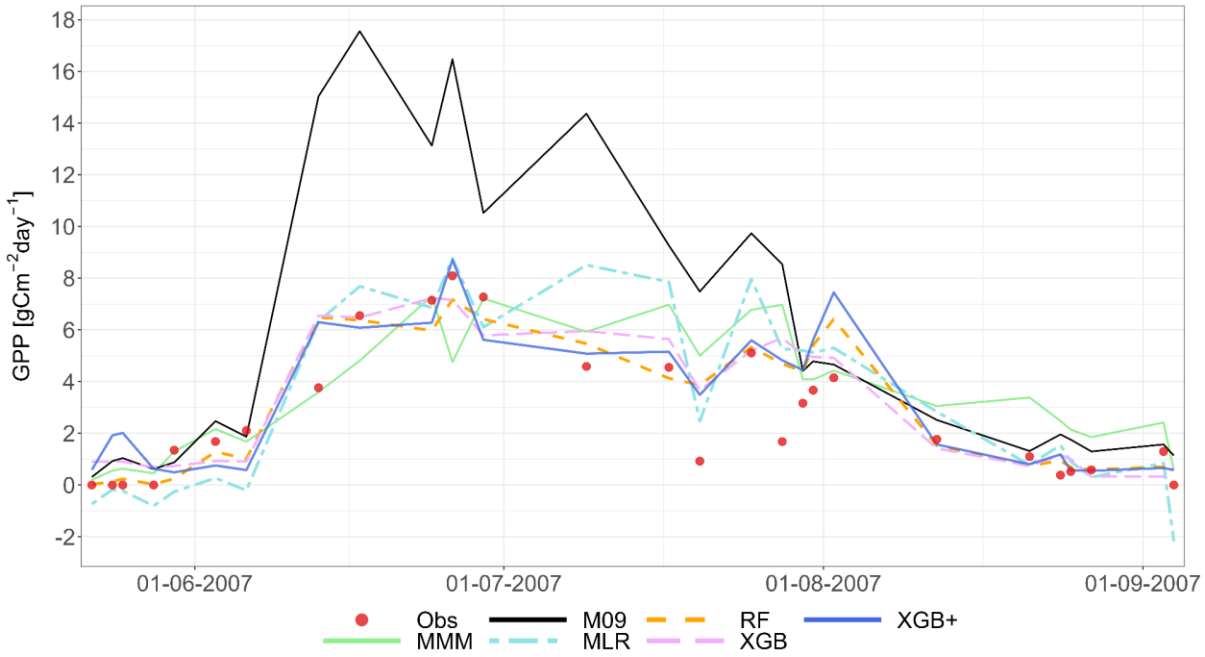


Figure A1: Comparison of the best individual model, the constructed meta-models and the traditional Multi-Model Median with the observations for all sites (from left to right C1, C2, G3 and G4) and for the entire time series for GPP based on the 70/30 approach. From top to bottom: best individual model with ID, MMM, MLR, RF, RF+, XGB and XGB+. All units are in g C m^{-2} day^{-1}. Red dashed line represents the 1:1 relationship.



1335

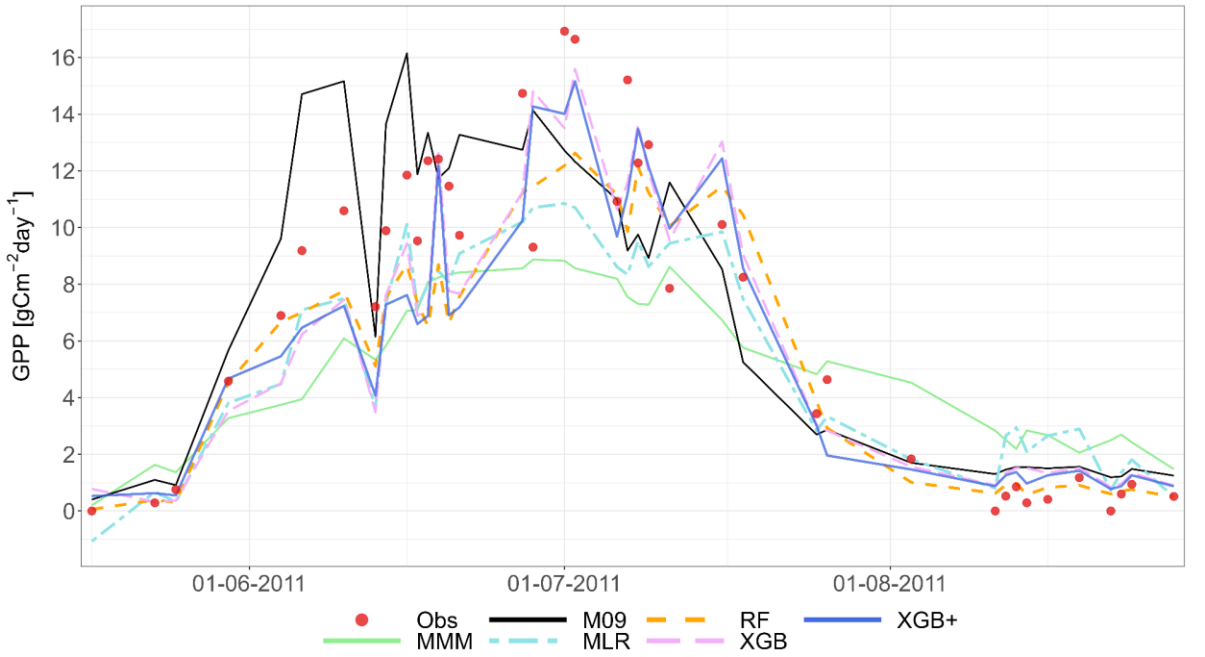
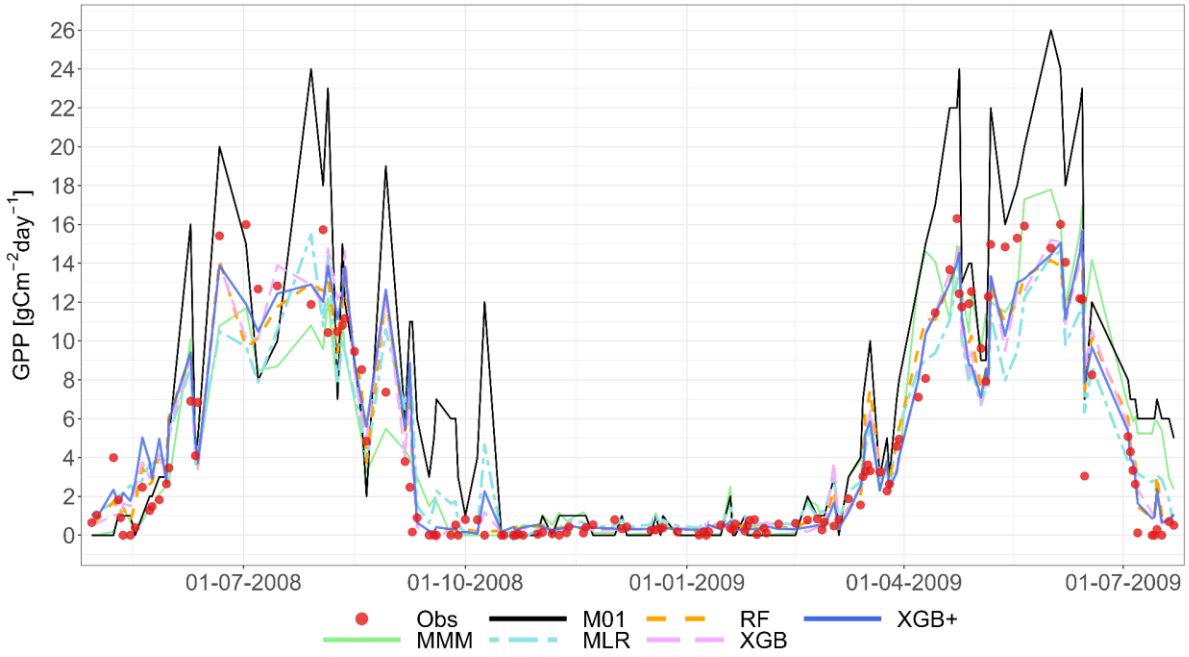


Figure A2: GPP, C1 - Ottawa (CA), cropland, 2007 and 2011, 70/30 strategy.

1340



1345

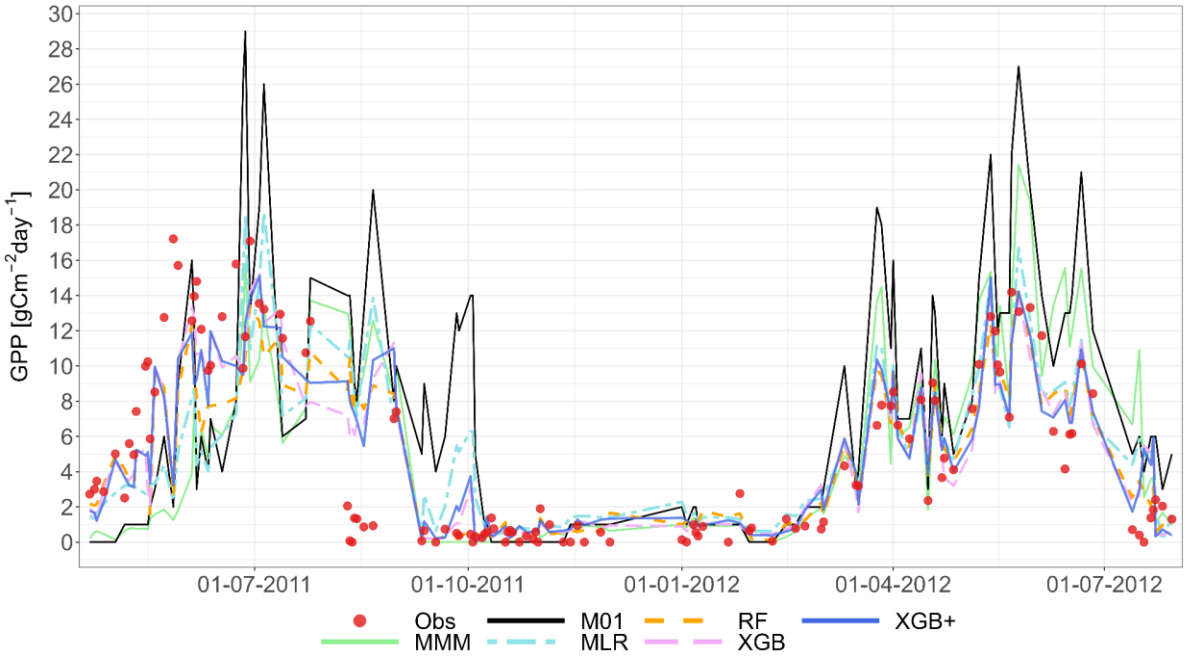
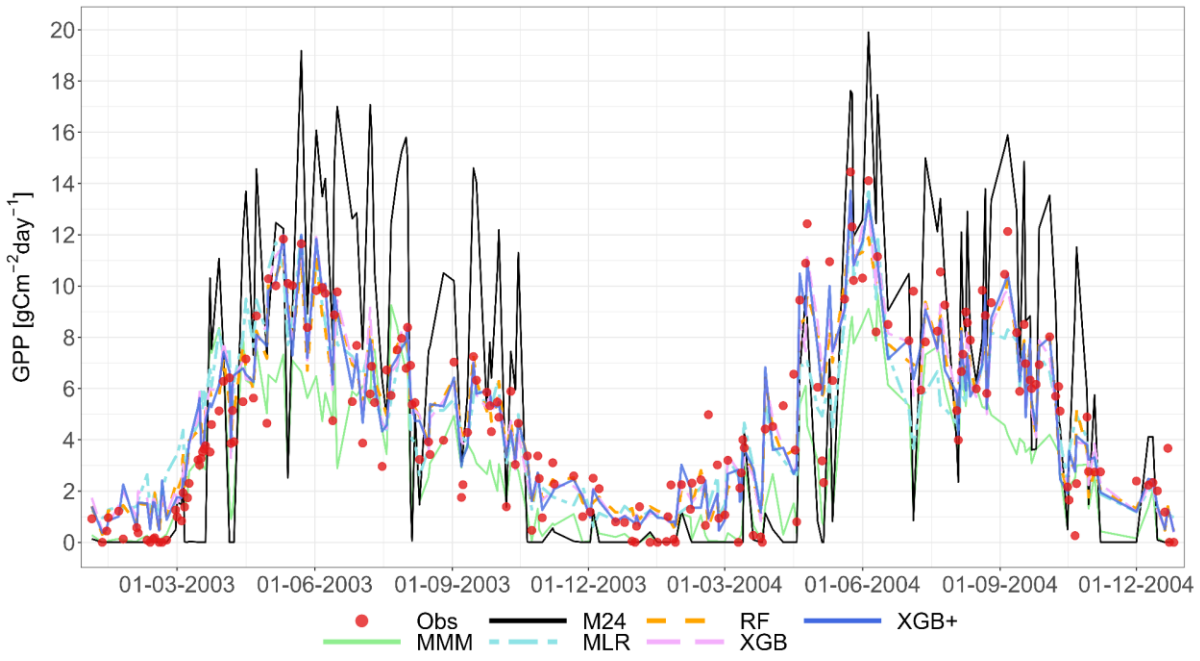
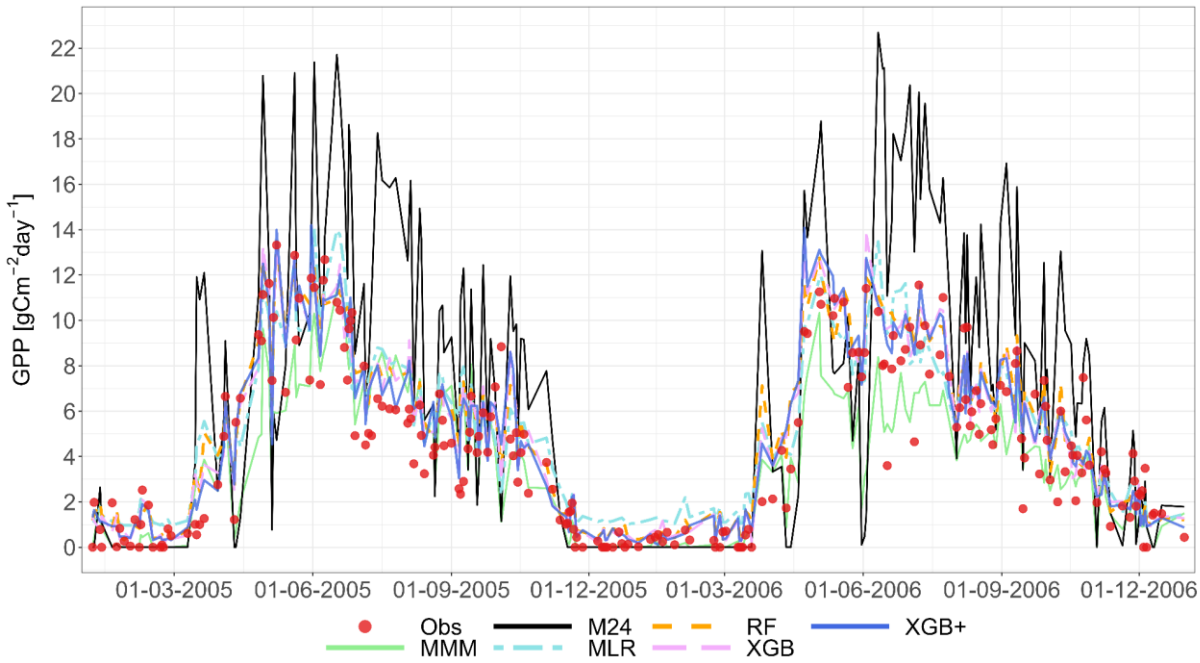
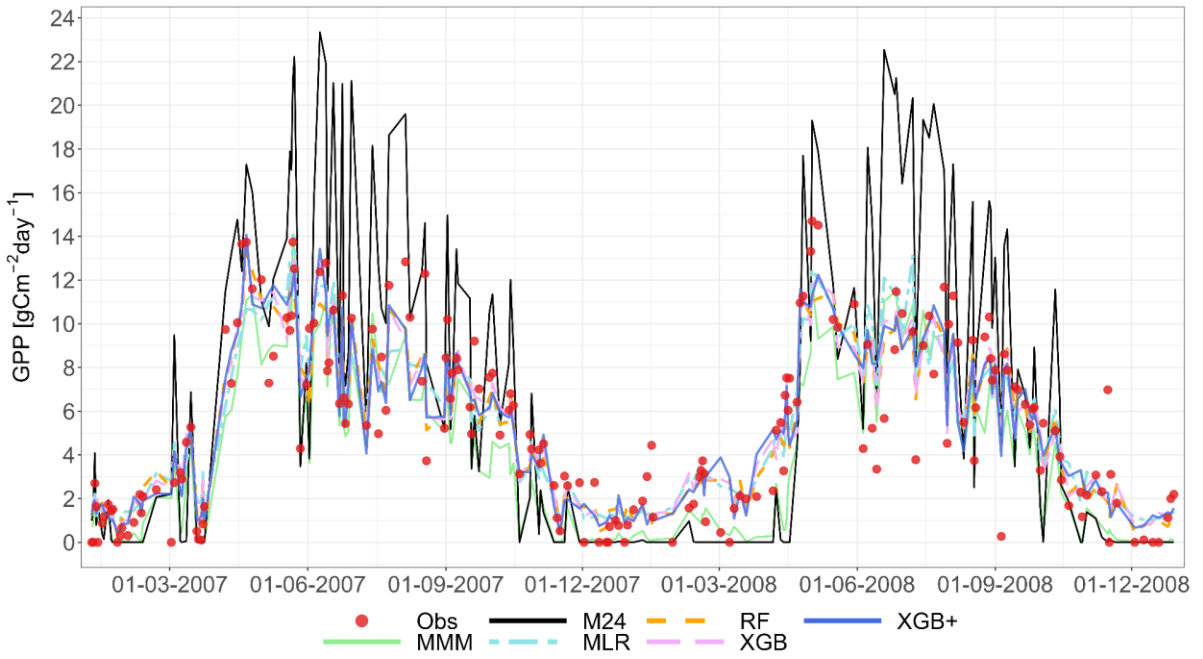


Figure A3: GPP, C2 - Grignon, (FR), cropland, 2008, 2009, 2011 and 2012, 70/30 strategy.


GPP [gCm^{-2}day^{-1}]
0
2
4
6
8
10
12
14
16
18
20
01-03-2003
01-06-2003
01-09-2003
01-12-2003
01-03-2004
01-06-2004
01-09-2004
01-12-2004
Obs
M24
RF
XGB+
MMM
MLR
XGB

1350


GPP [gCm^{-2}day^{-1}]
0
2
4
6
8
10
12
14
16
18
20
22
01-03-2005
01-06-2005
01-09-2005
01-12-2005
01-03-2006
01-06-2006
01-09-2006
01-12-2006
Obs
M24
RF
XGB+
MMM
MLR
XGB



1355

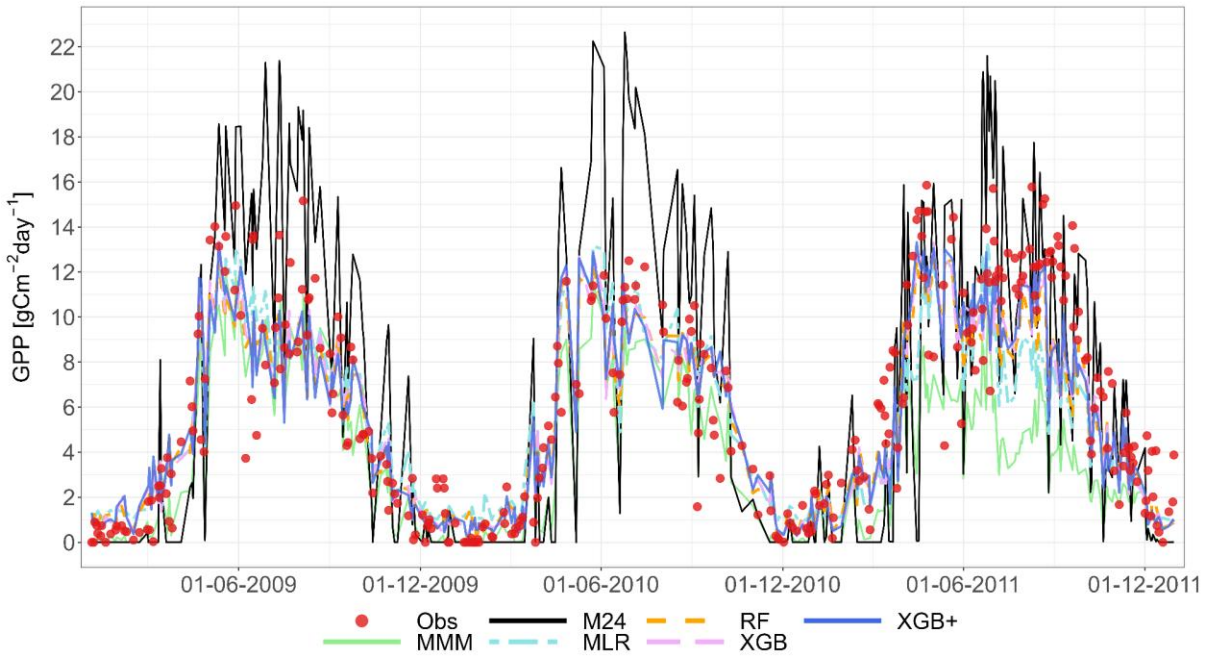
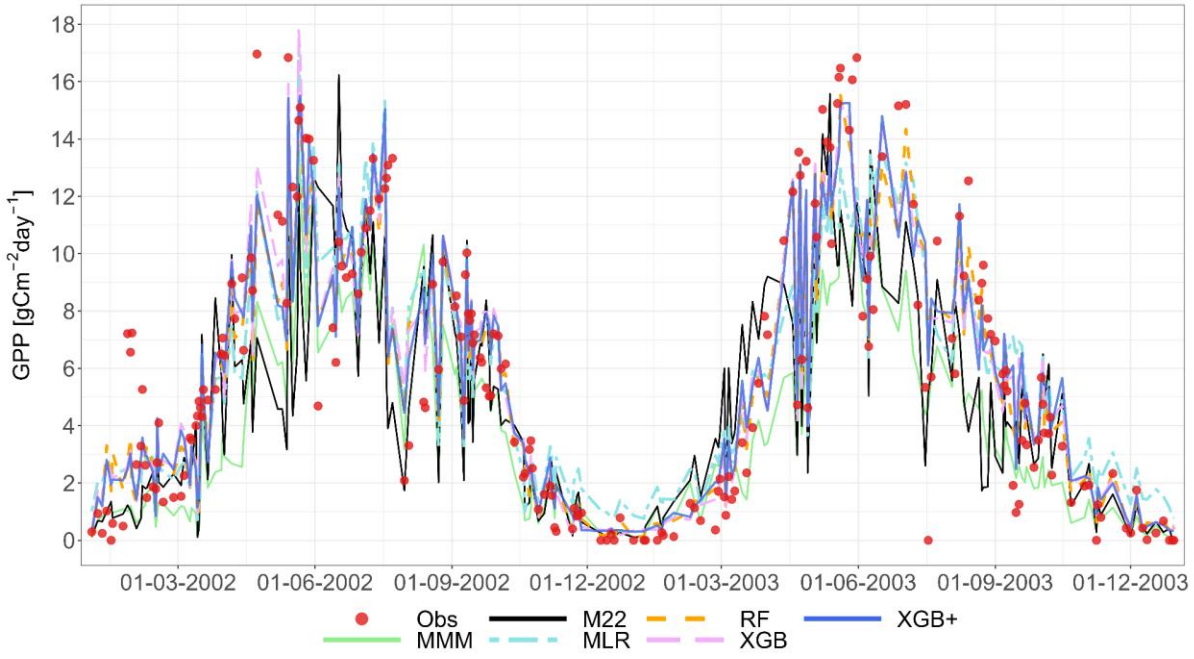
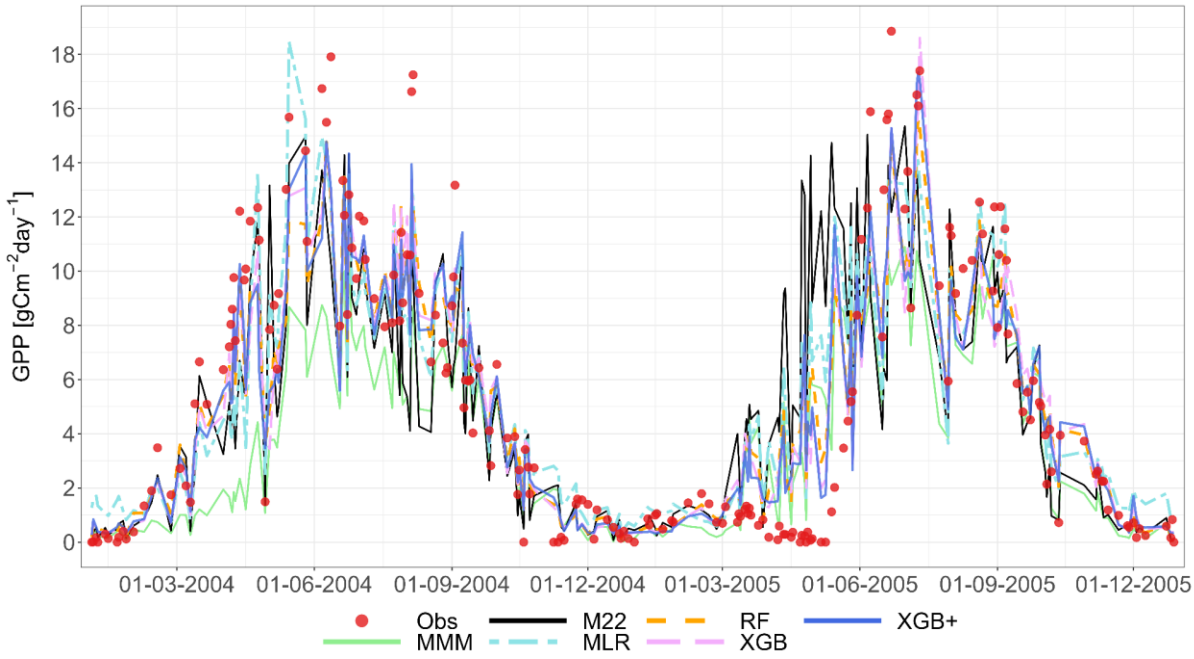


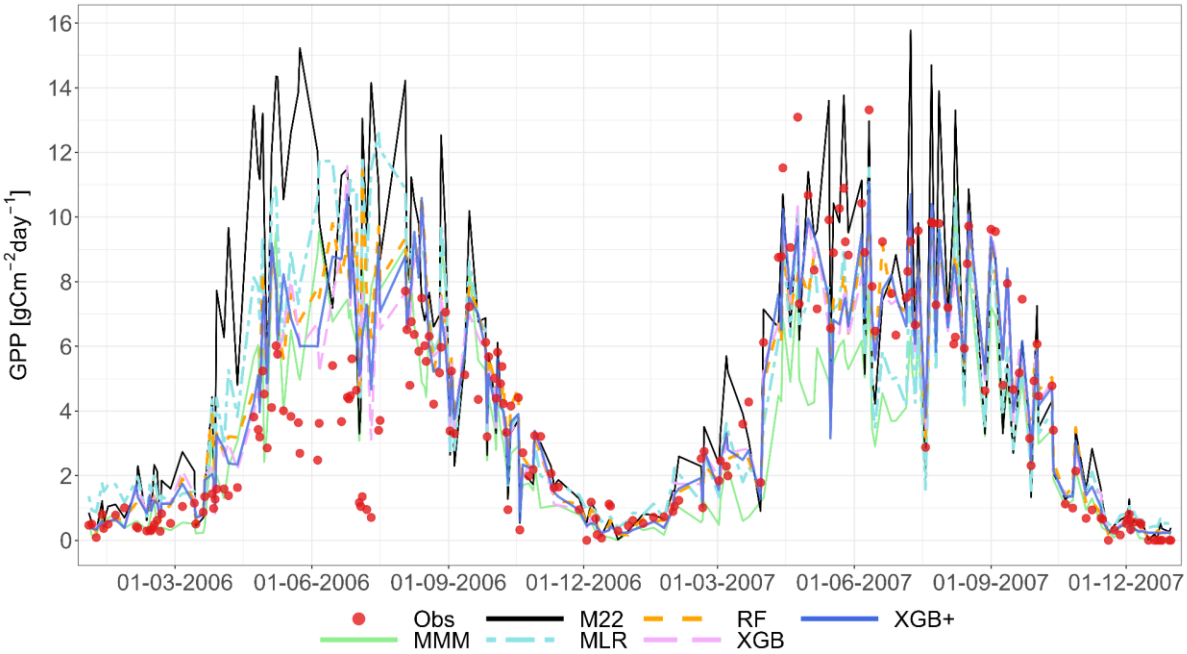
Figure A4: GPP, G3 - Laqueuille (FR), grassland, 2003-2011, 70/30 strategy.


GPP [gCm^{-2}day^{-1}]
0
2
4
6
8
10
12
14
16
18
01-03-2002
01-06-2002
01-09-2002
01-12-2002
01-03-2003
01-06-2003
01-09-2003
01-12-2003
Obs
M22
RF
XGB+
MMM
MLR
XGB

1360


GPP [gCm^{-2}day^{-1}]
0
2
4
6
8
10
12
14
16
18
01-03-2004
01-06-2004
01-09-2004
01-12-2004
01-03-2005
01-06-2005
01-09-2005
01-12-2005
Obs
M22
RF
XGB+
MMM
MLR
XGB

1365



1370

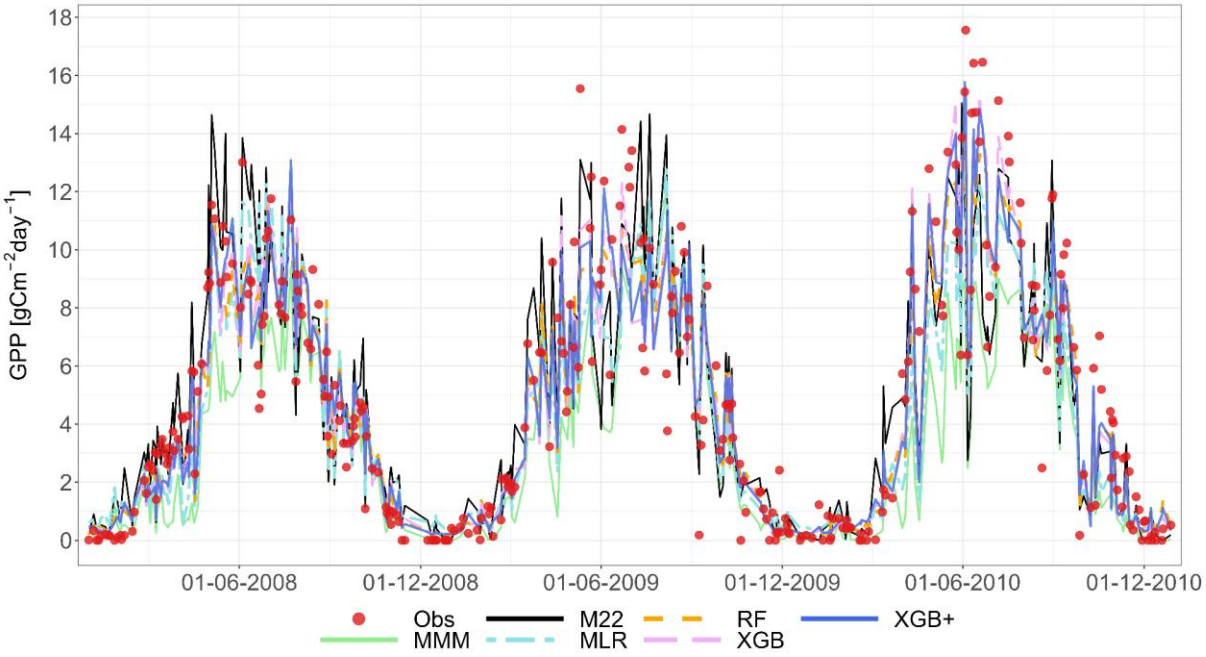


Figure A5: GPP, G4 - Easter Bush (UK), grassland, 70/30 strategy.

1375

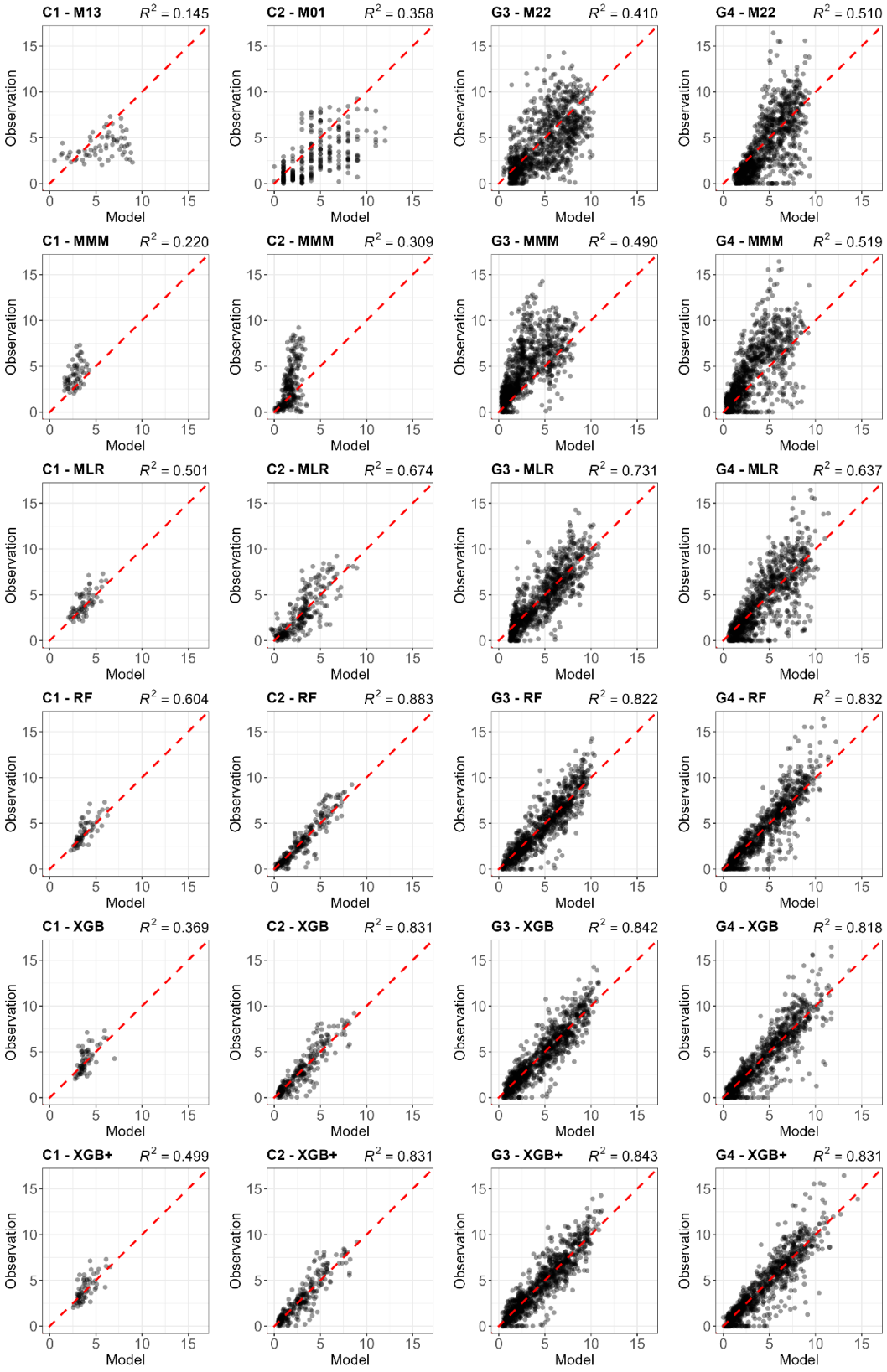
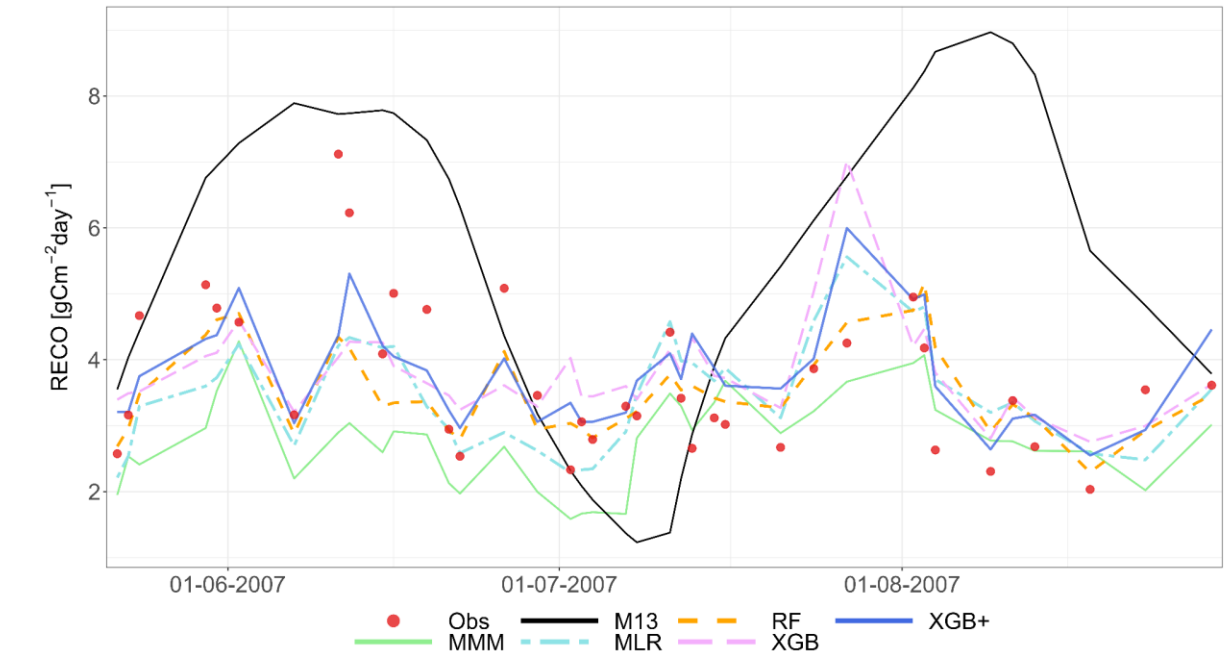


Figure A6: Comparison of the best individual model, the multi-model median and the constructed meta-models with observations for RECO based on the 70/30 approach. Each row represents a different model type, and columns correspond to the sites (from left to right: C1, C2, G3, and G4). The top row shows the best individual models with their identifiers (M13 at C1, M01 at C2, M22 at G3 and G4). The remaining rows show the MMM, MLR, RF, XGB and XGB+. All units are in g C m^{-2} day^{-1}. The red dashed line represents the 1:1 relationship.



1385

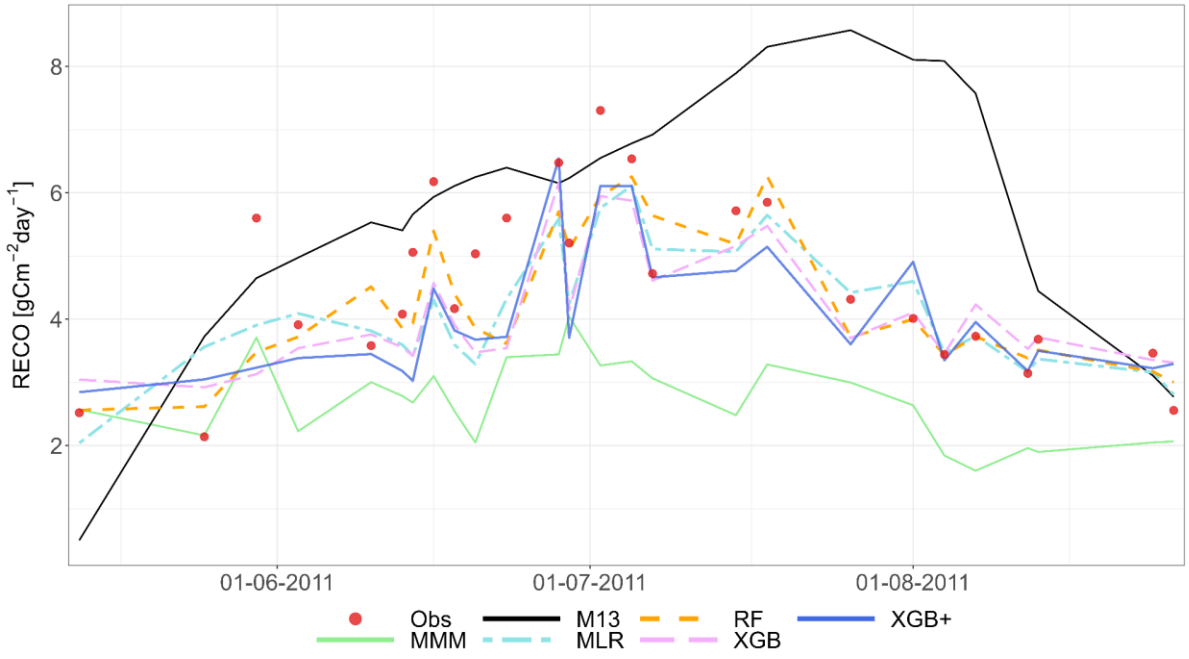


Figure A7: RECO, C1 - Ottawa (CA), cropland, 2007 and 2011, 70/30 strategy.

1390

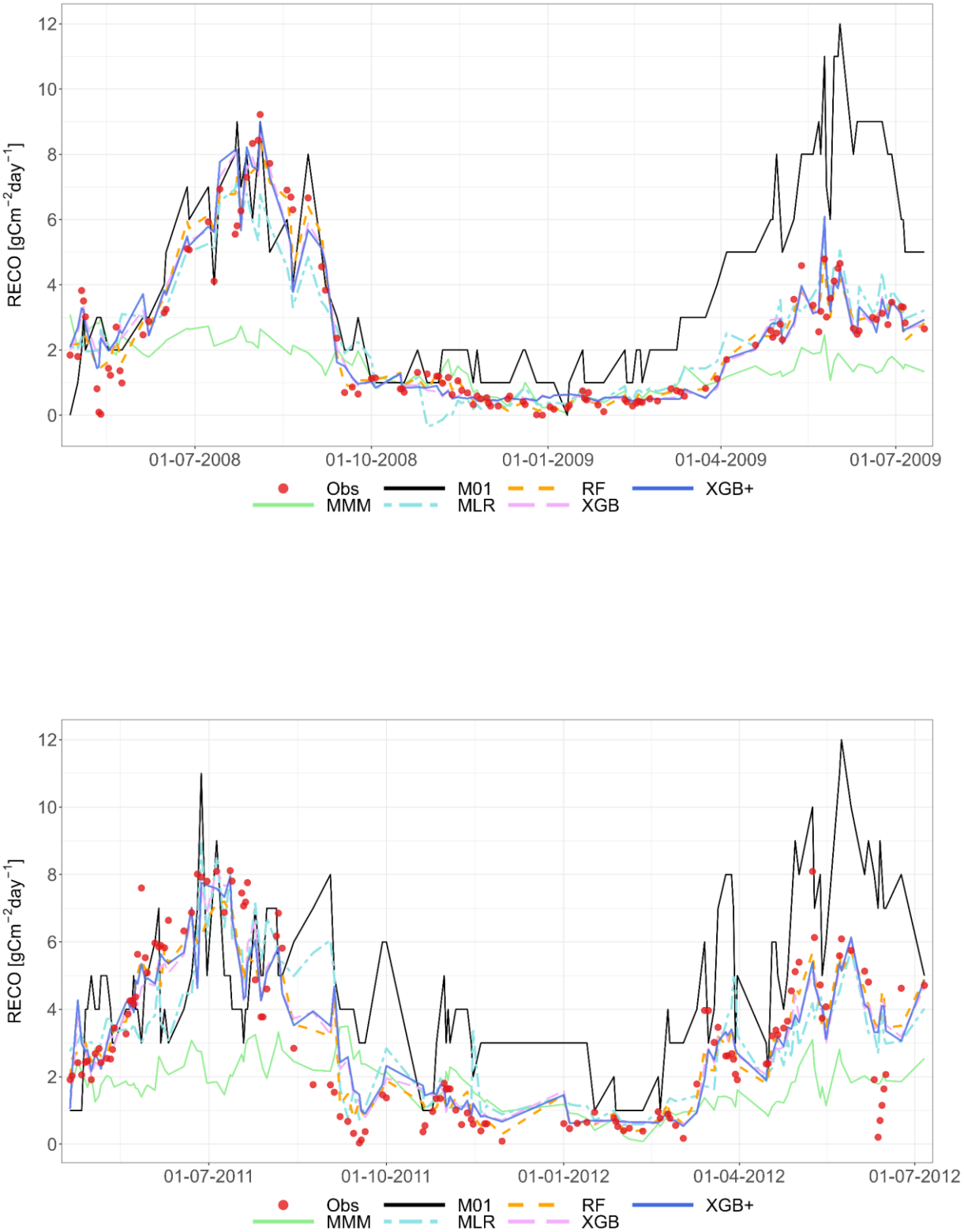
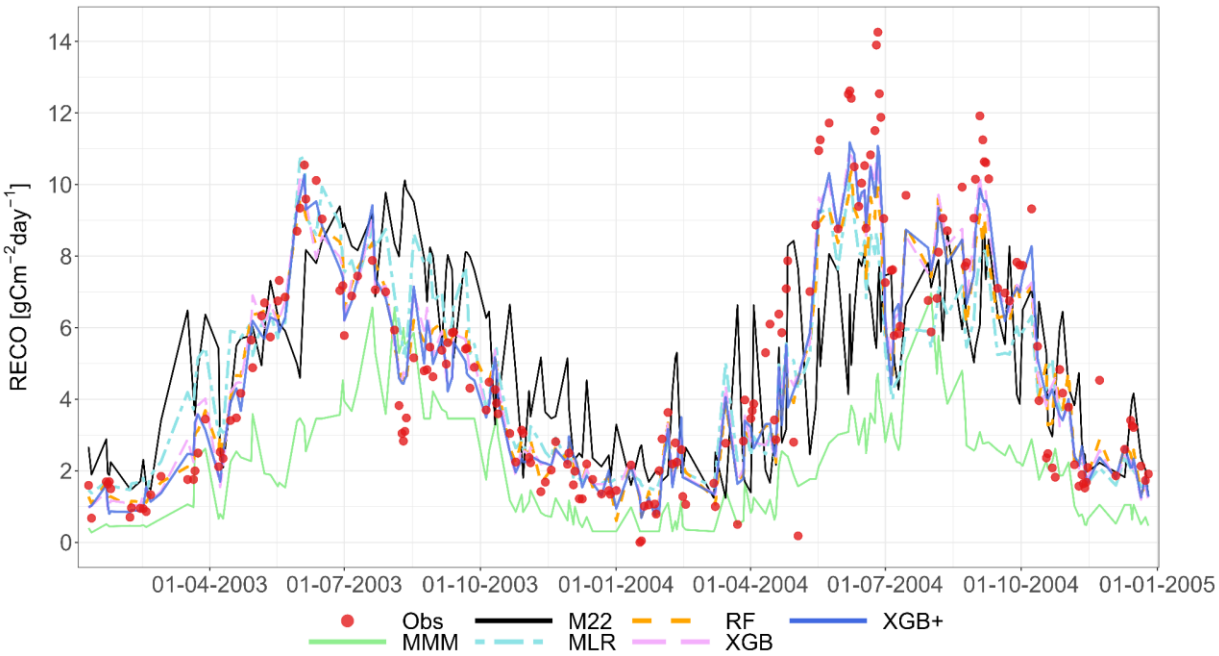
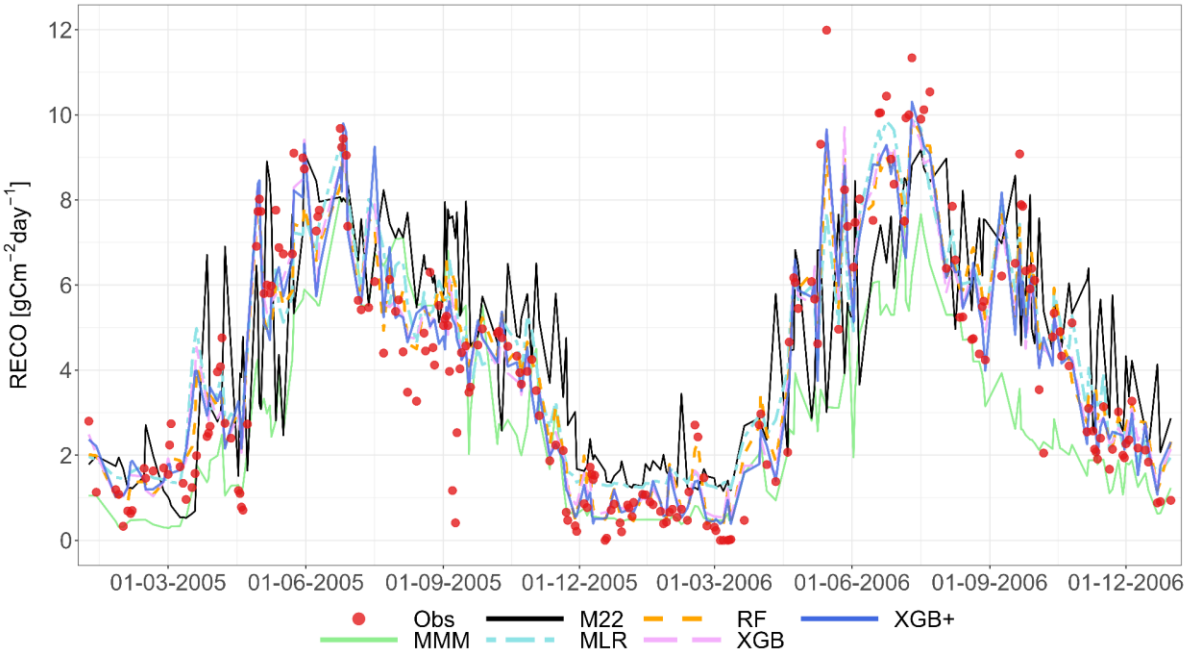
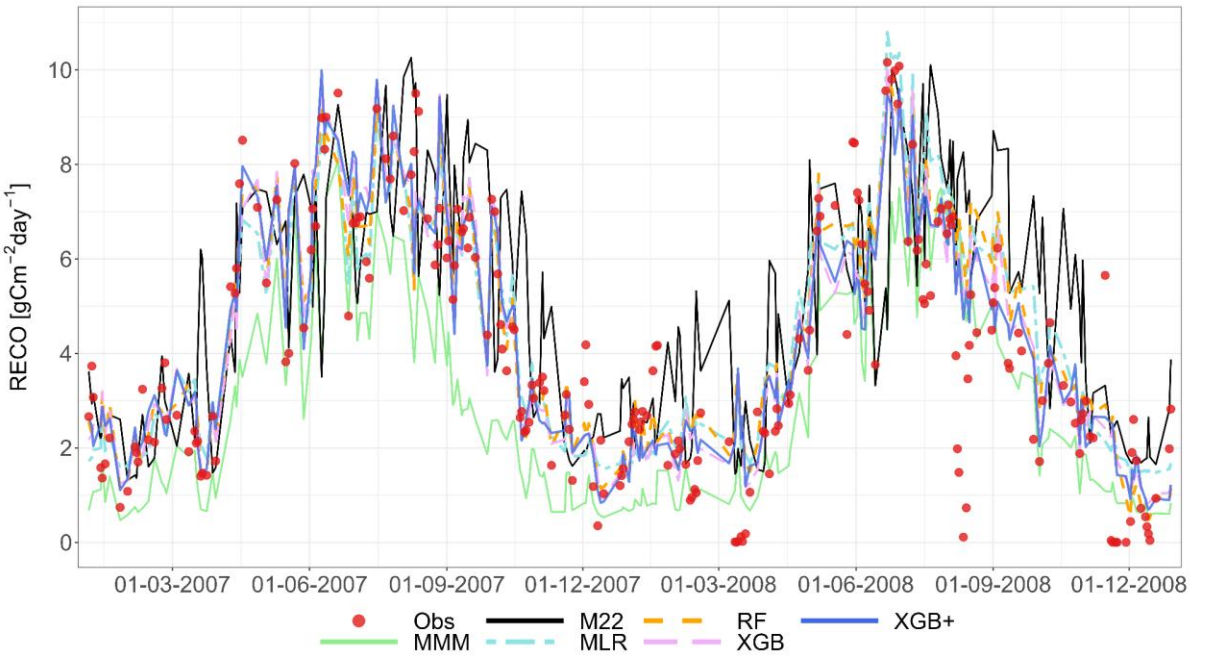


Figure A8: RECO, C2 - Grignon, (FR), cropland, 2007, 2008, 2011 and 2012, 70/30 strategy.


RECO [gCm^{-2}day^{-1}]
0
2
4
6
8
10
12
14
01-04-2003
01-07-2003
01-10-2003
01-01-2004
01-04-2004
01-07-2004
01-10-2004
01-01-2005
Obs
MMM
M22
MLR
RF
XGB
XGB+

1400


RECO [gCm^{-2}day^{-1}]
0
2
4
6
8
10
12
01-03-2005
01-06-2005
01-09-2005
01-12-2005
01-03-2006
01-06-2006
01-09-2006
01-12-2006
Obs
MMM
M22
MLR
RF
XGB
XGB+



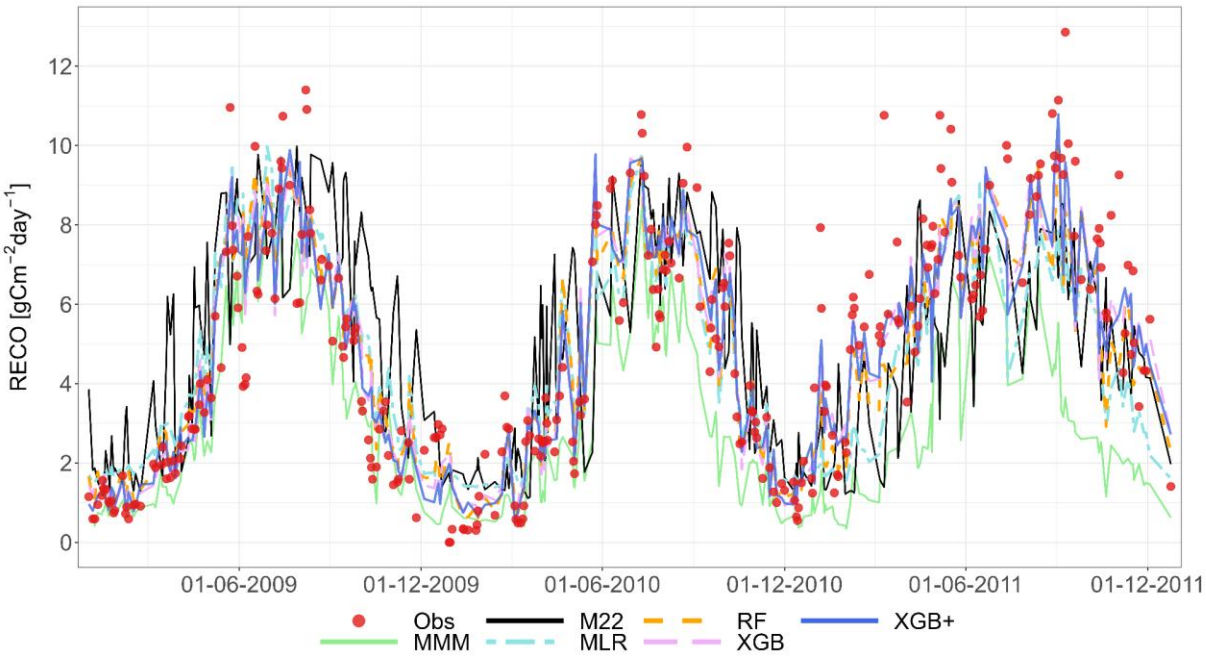
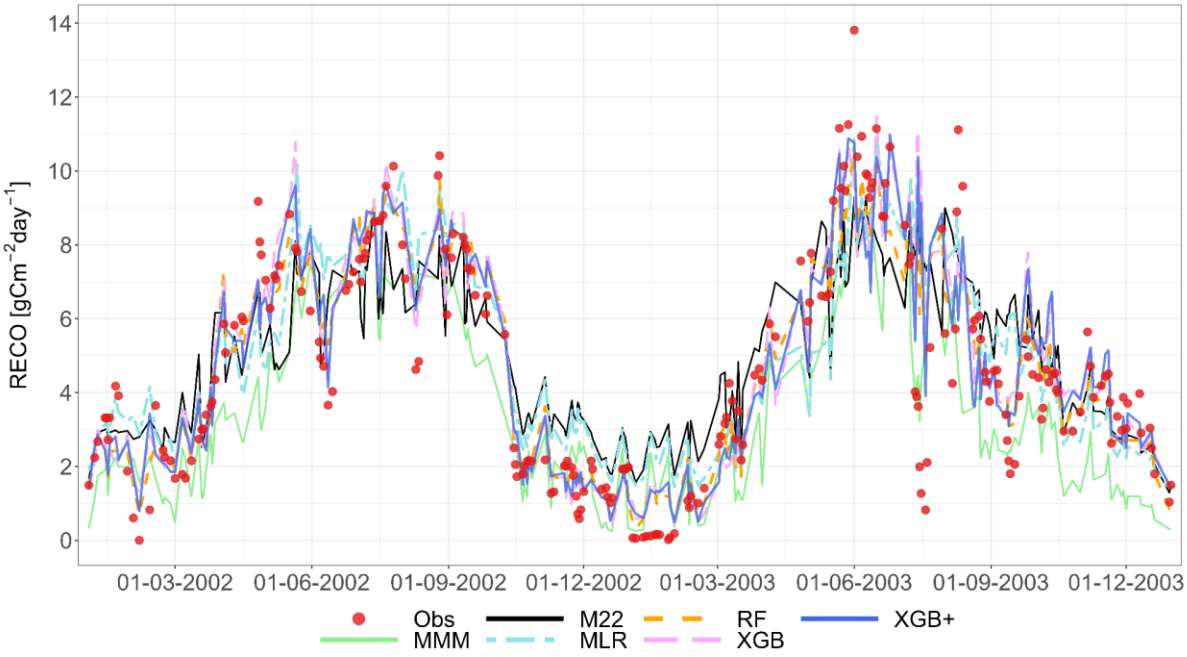
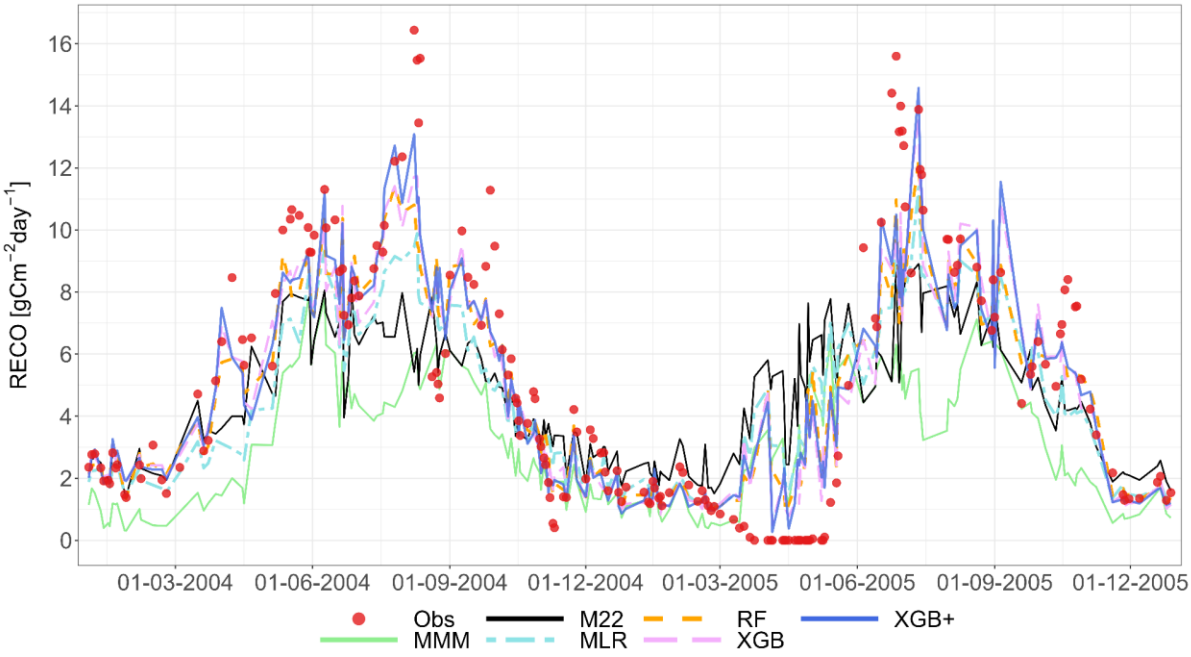
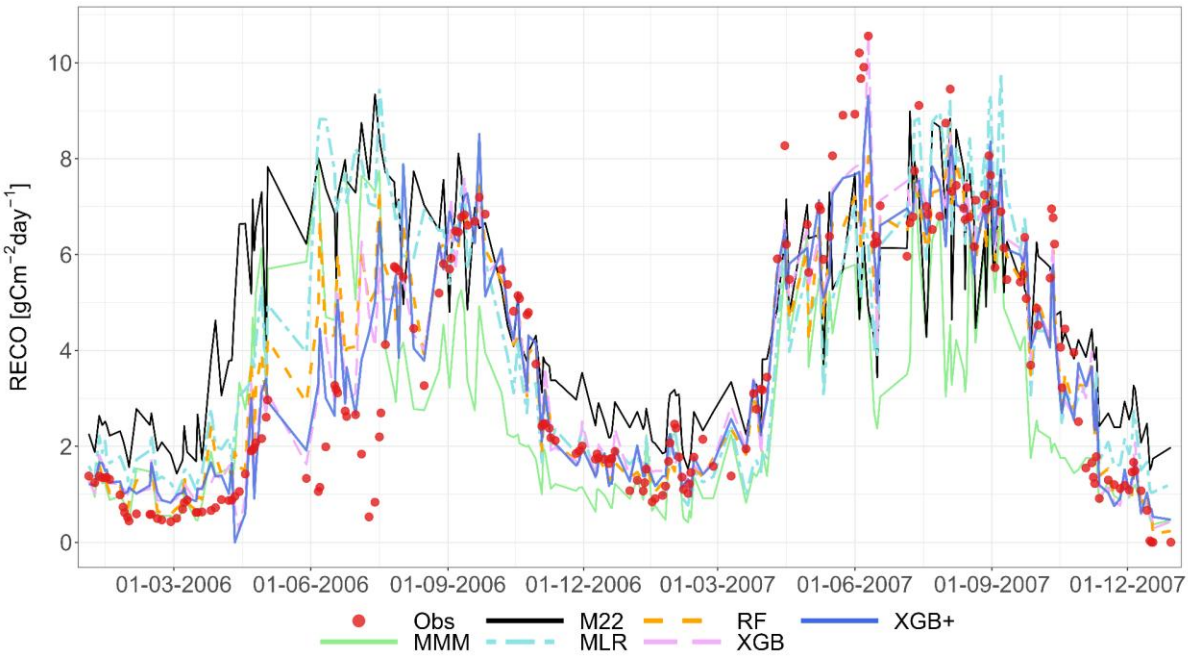


Figure A9: RECO, G3 - Laqueuille (FR), grassland, 2003-2011, 70/30 strategy.


RECO [gCm^{-2}day^{-1}]
0
2
4
6
8
10
12
14
01-03-2002
01-06-2002
01-09-2002
01-12-2002
01-03-2003
01-06-2003
01-09-2003
01-12-2003
Obs
MMM
M22
MLR
RF
XGB
XGB+

1415


RECO [gCm^{-2}day^{-1}]
0
2
4
6
8
10
12
14
16
01-03-2004
01-06-2004
01-09-2004
01-12-2004
01-03-2005
01-06-2005
01-09-2005
01-12-2005
Obs
MMM
M22
MLR
RF
XGB
XGB+



1420

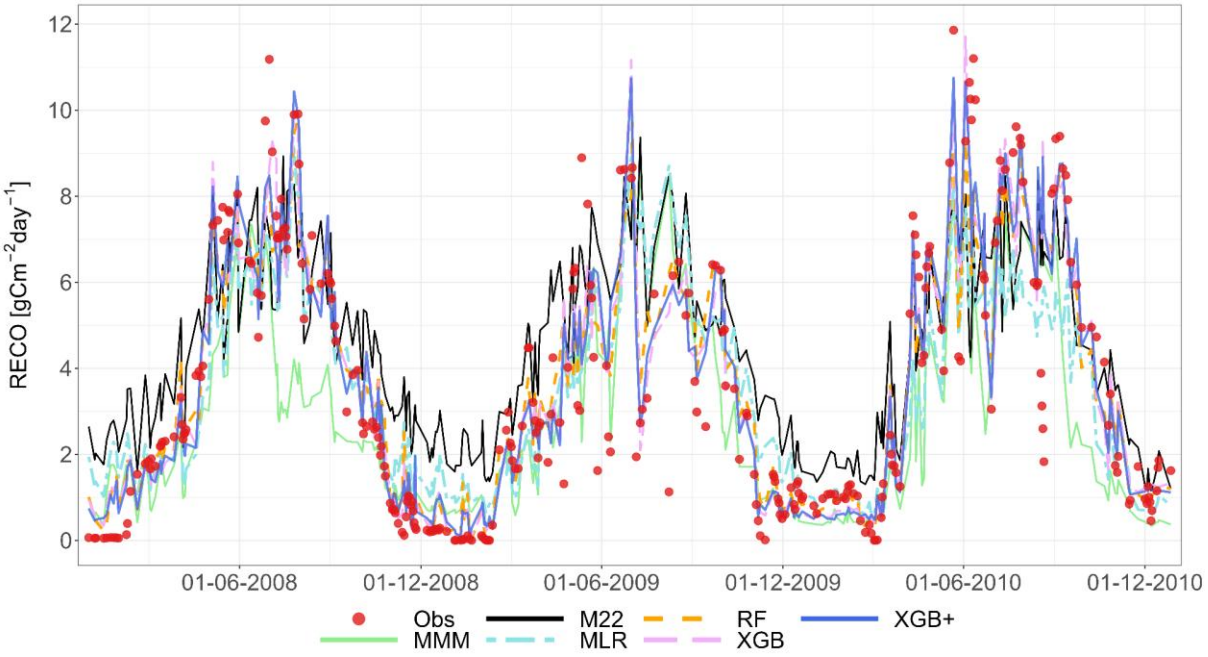


Figure A10: RECO, G4 - Easter Bush (UK), grassland, 70/30 strategy.

1425

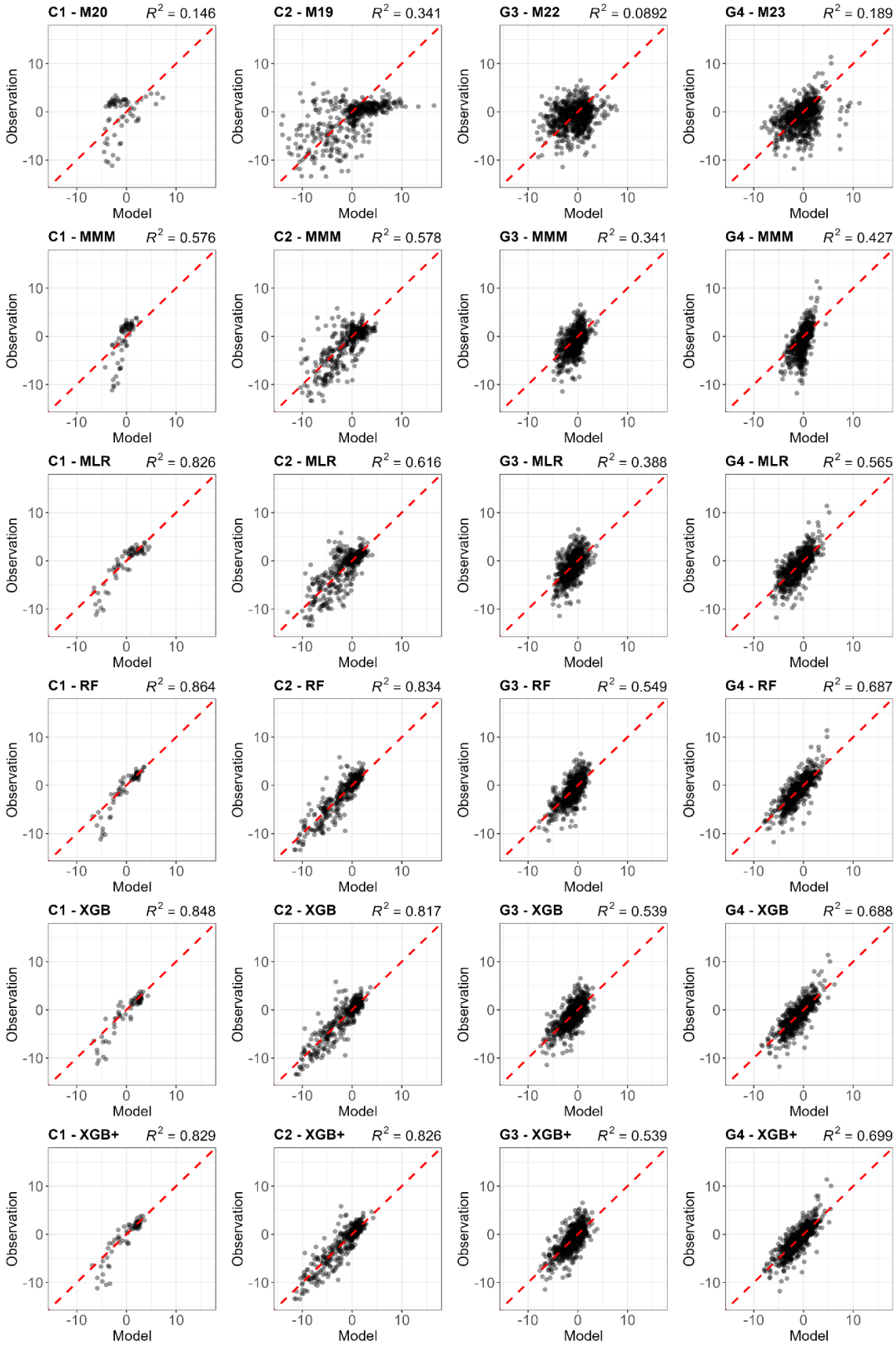
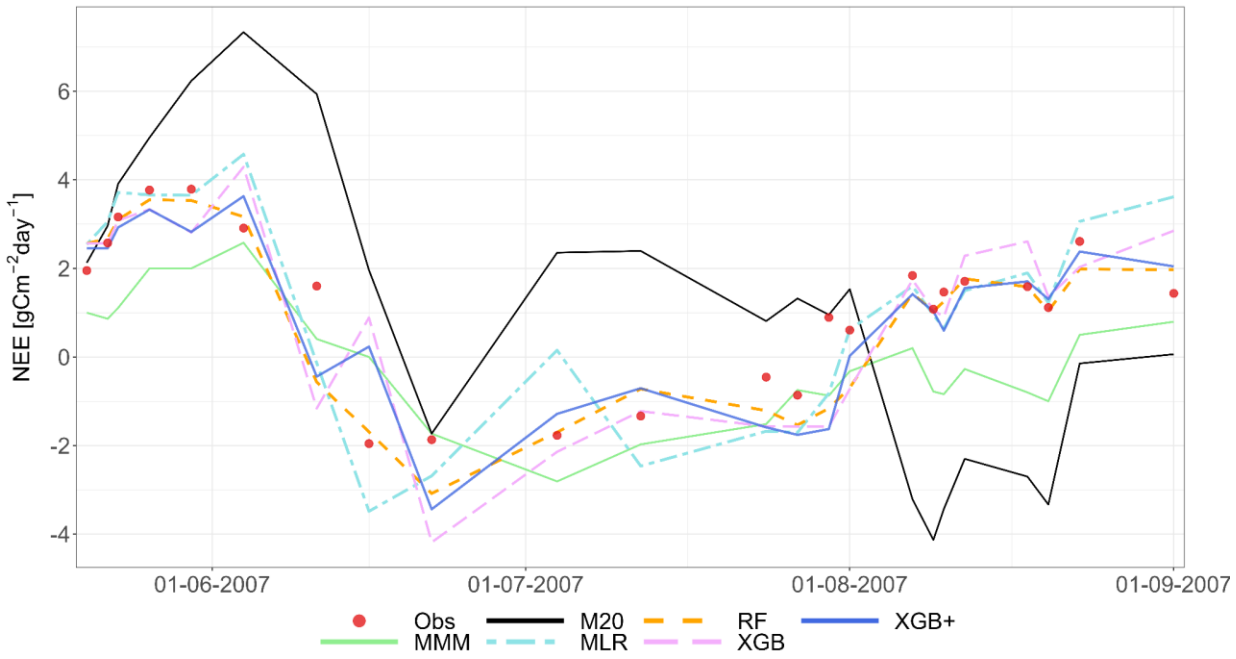


Figure A11: Comparison of the best individual model, the multi-model median and the constructed meta-models with observations for NEE using the 70/30 approach. Each row represents a different model type, and columns correspond to the sites (from left to right: C1, C2, G3, and G4). The top row shows the best individual models with their identifiers (M20 at C1, M19 at C2, M22 at G3, M23 at G4). The remaining rows show the MMM, MLR, RF, XGB and XGB+. All units are in g C m^{-2} day^{-1}. The red dashed line represents the 1:1 relationship.



1435

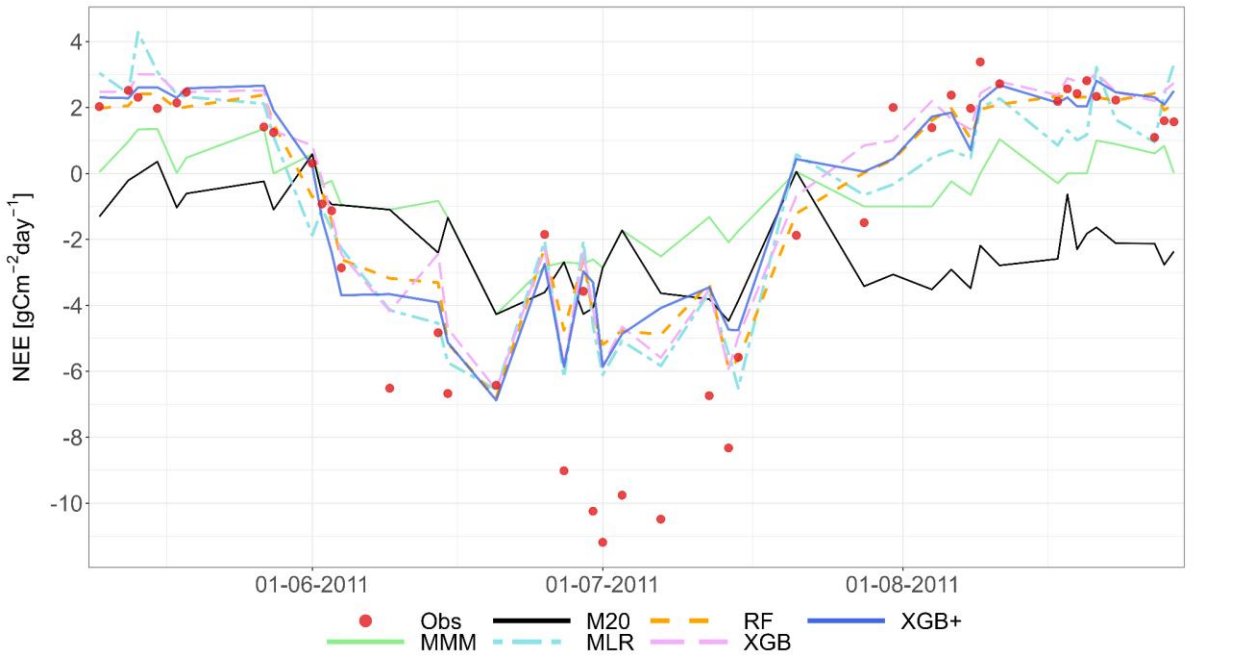
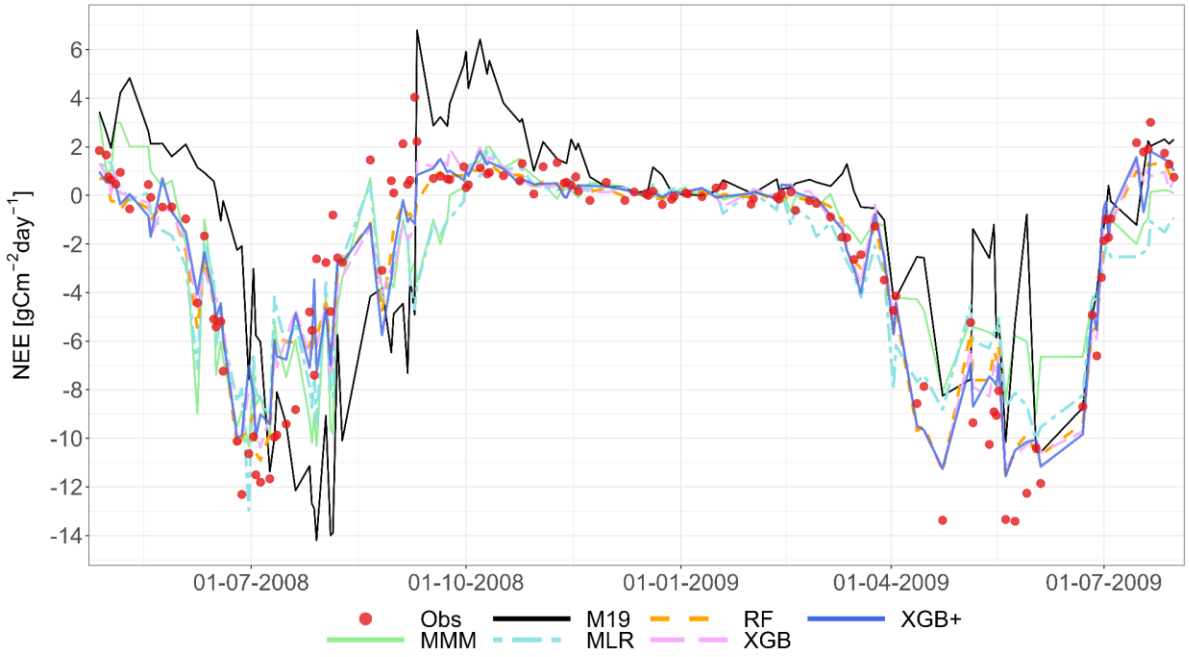
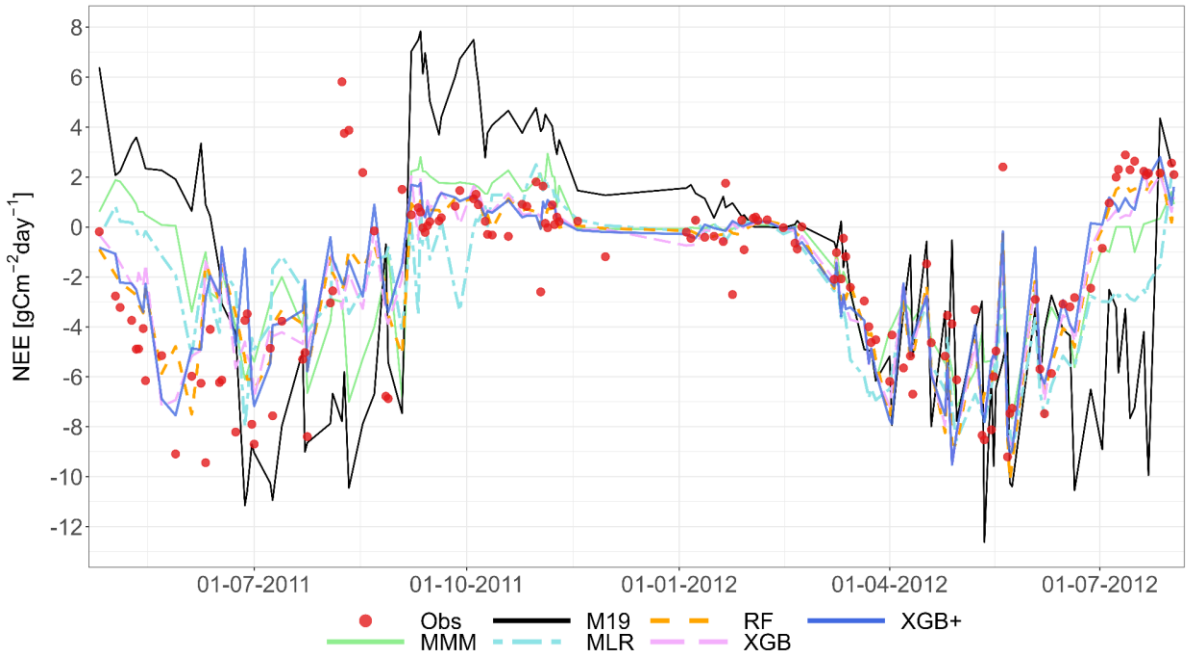


Figure A12: NEE, C1 - Ottawa (CA), cropland, 2007 and 2011, 70/30 strategy.

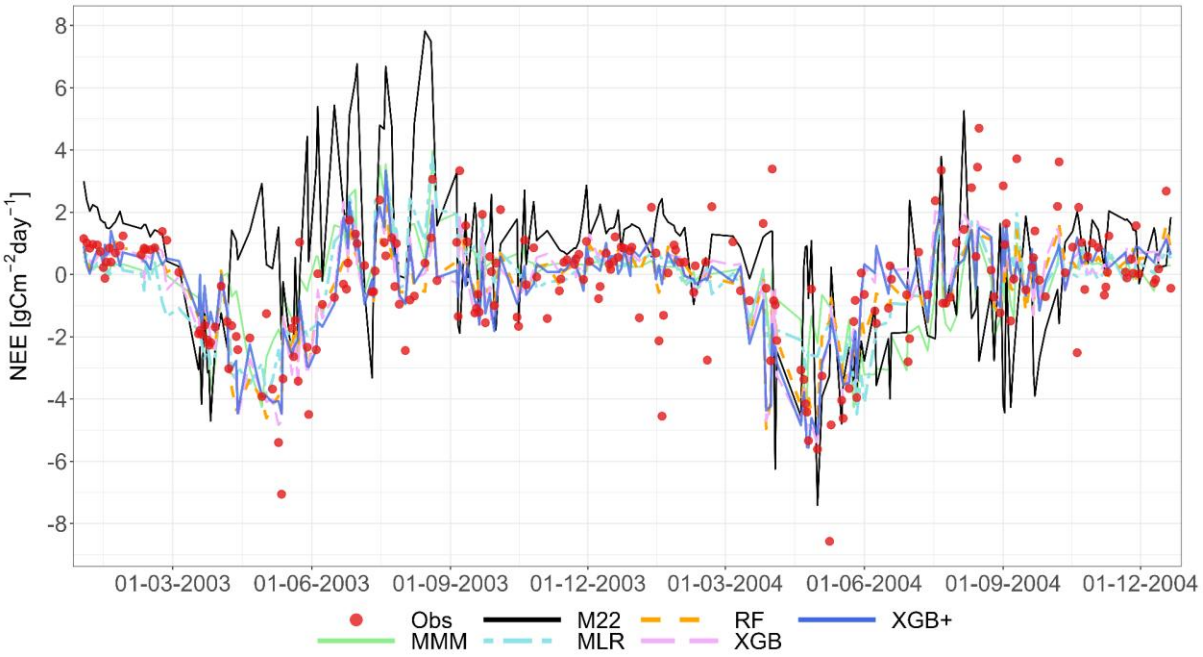
1440

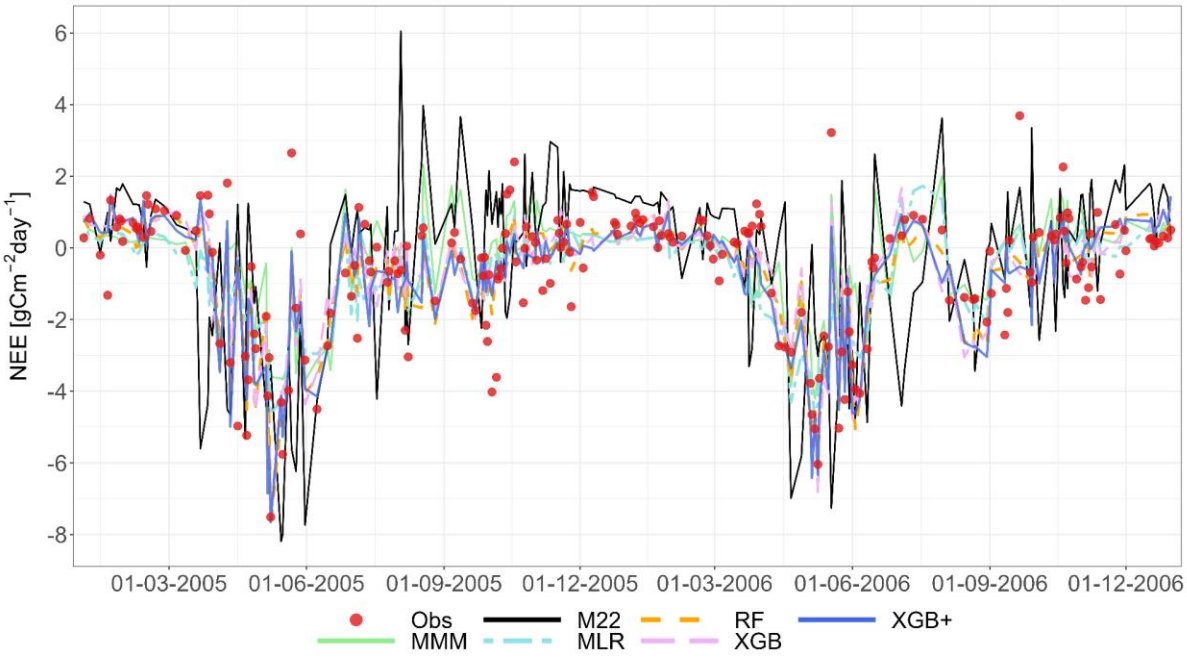


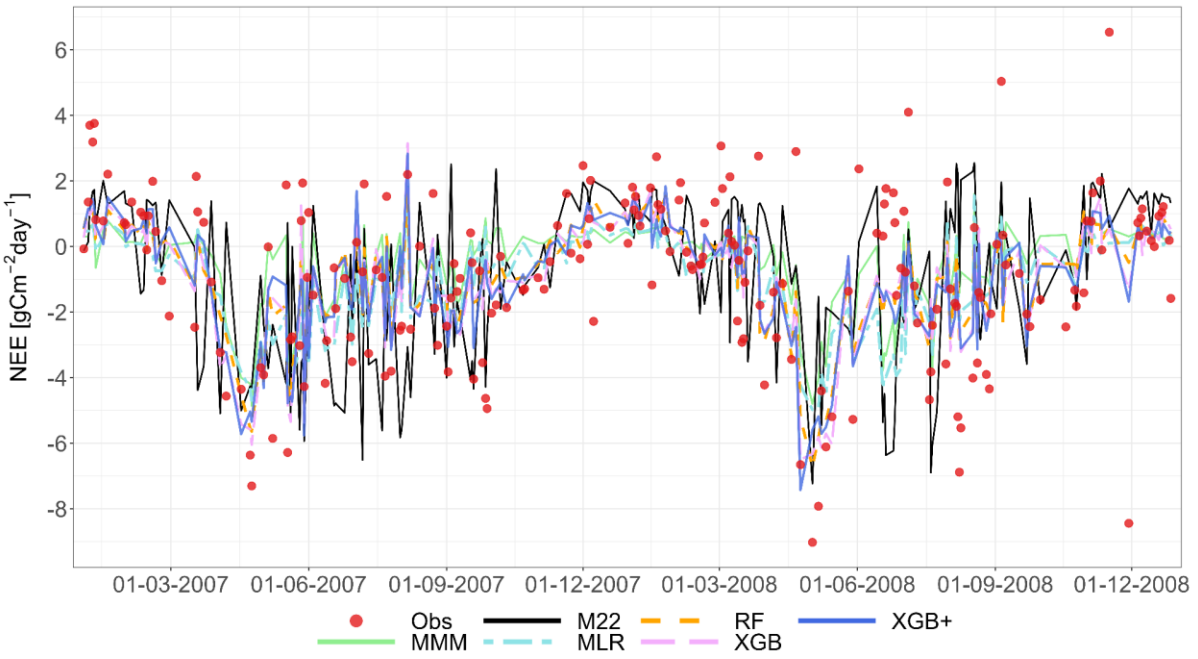


1445

Figure A13: NEE, C2 - Grignon, (FR), cropland, 2008, 2009, 2011 and 2012, 70/30 strategy.


NEE [gCm^{-2}day^{-1}]
8
6
4
2
0
-2
-4
-6
-8
01-03-2003
01-06-2003
01-09-2003
01-12-2003
01-03-2004
01-06-2004
01-09-2004
01-12-2004
Obs
M22
RF
XGB+
MMM
MLR
XGB


NEE [gCm^{-2}day^{-1}]
6
4
2
0
-2
-4
-6
-8
01-03-2005
01-06-2005
01-09-2005
01-12-2005
01-03-2006
01-06-2006
01-09-2006
01-12-2006
Obs
M22
RF
XGB+
MMM
MLR
XGB



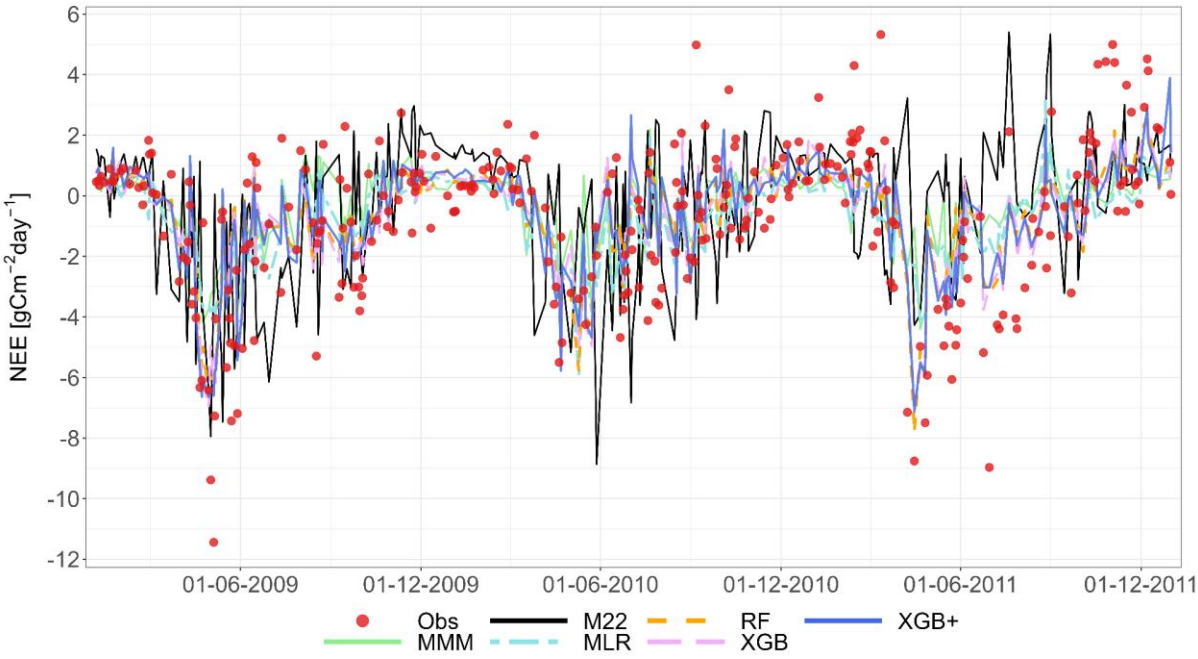
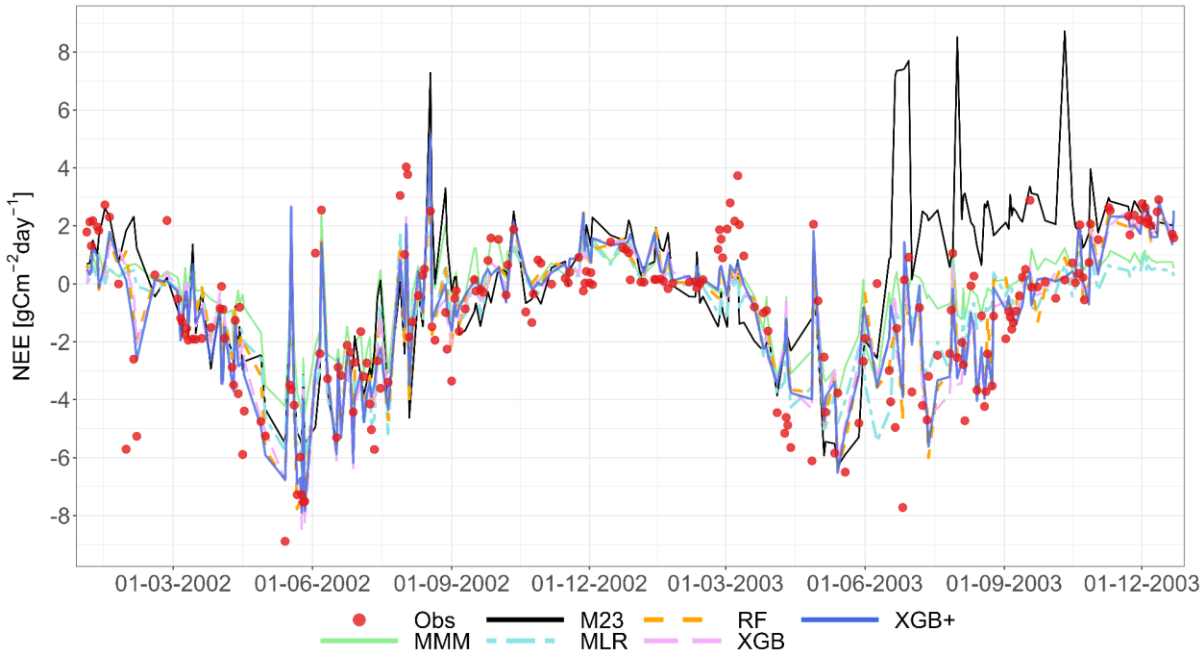
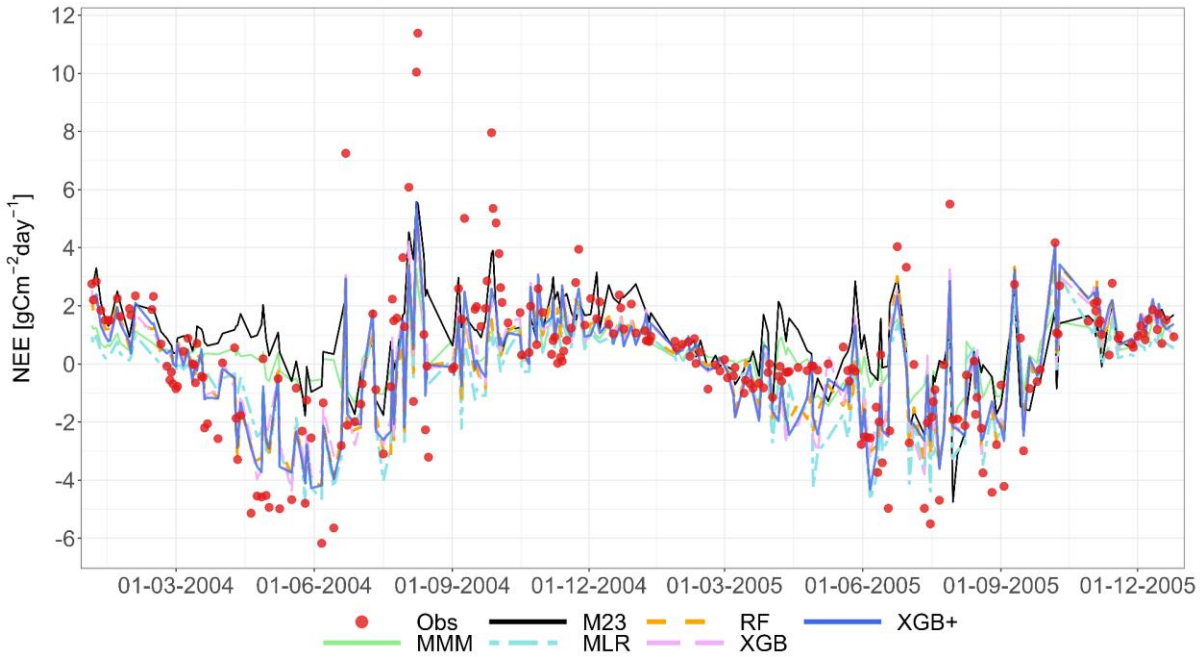
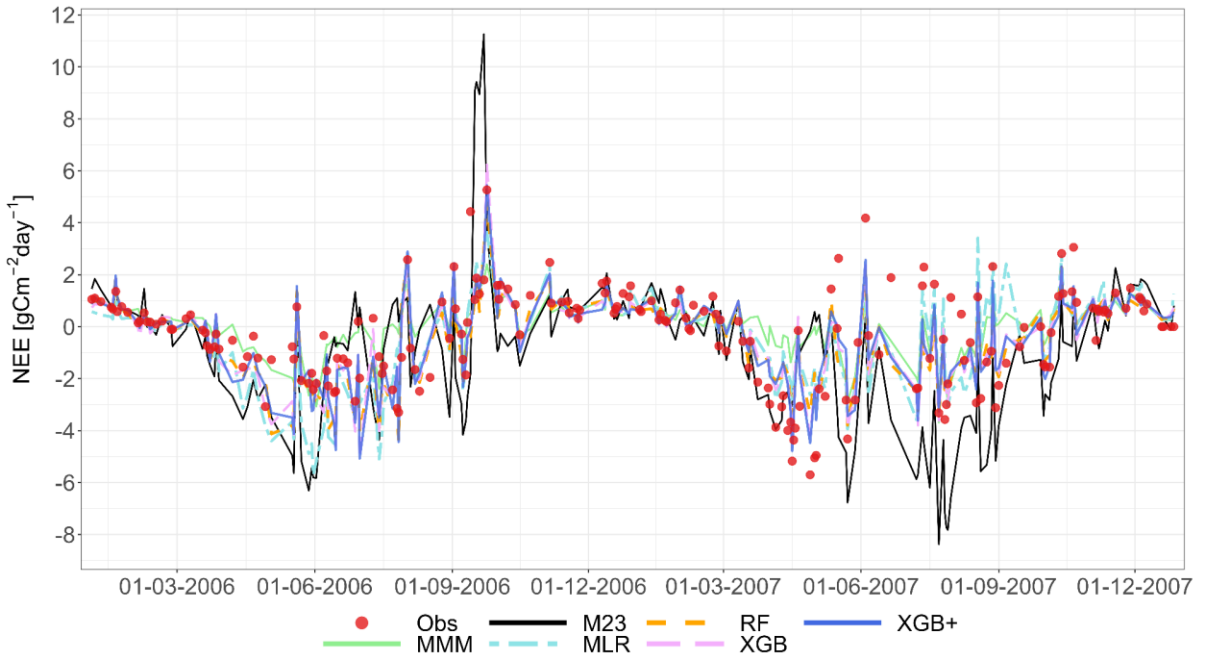


Figure A14: NEE, G3 - Laqueuille (FR), grassland, 2003-2011, 70/30 strategy.

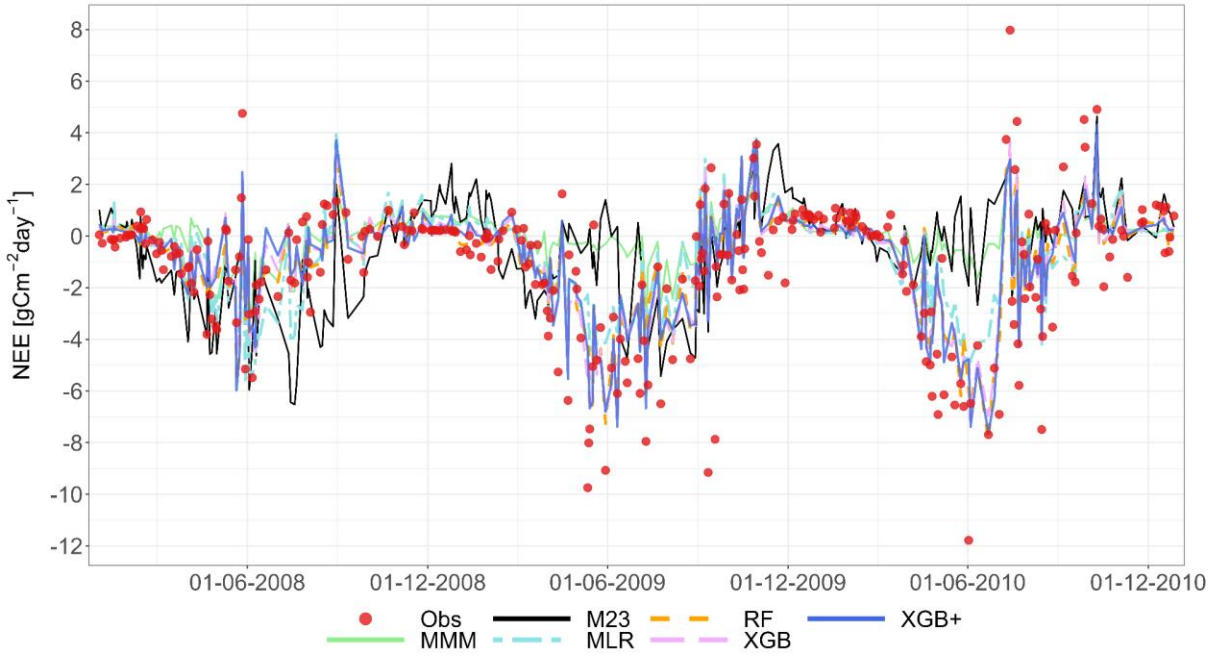

NEE [gCm^{-2}day^{-1}]
8
6
4
2
0
-2
-4
-6
-8
01-03-2002
01-06-2002
01-09-2002
01-12-2002
01-03-2003
01-06-2003
01-09-2003
01-12-2003
Obs
M23
RF
XGB+
MMM
MLR
XGB

1470

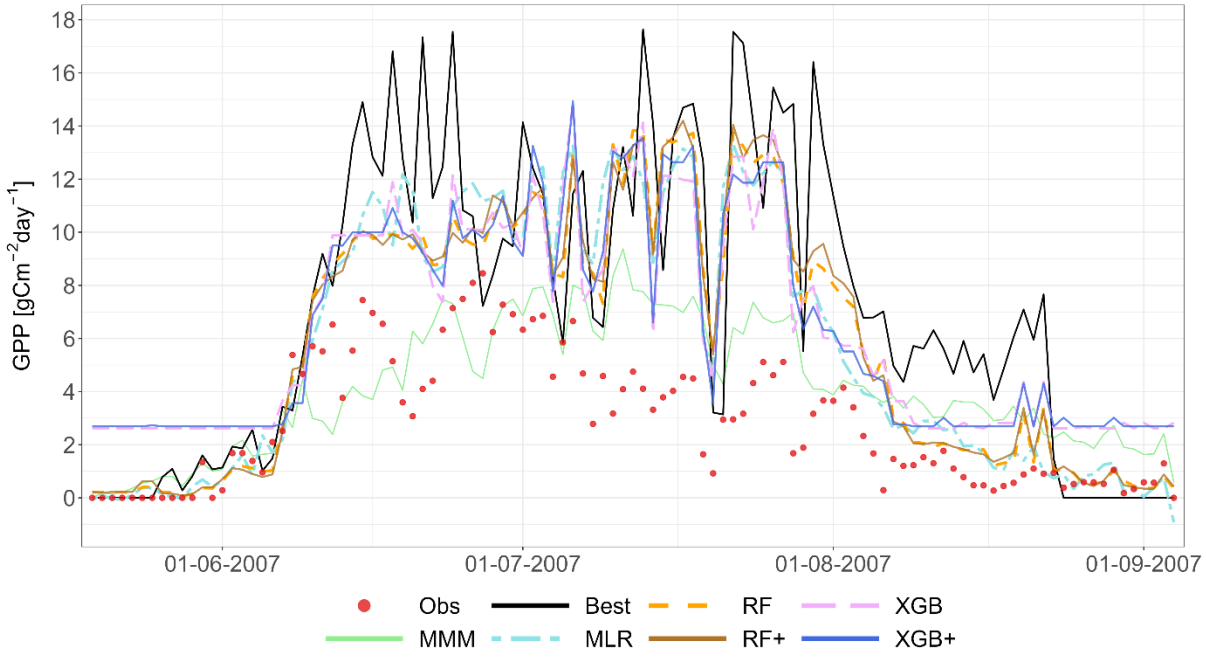

NEE [gCm^{-2}day^{-1}]
12
10
8
6
4
2
0
-2
-4
-6
01-03-2004
01-06-2004
01-09-2004
01-12-2004
01-03-2005
01-06-2005
01-09-2005
01-12-2005
Obs
M23
RF
XGB+
MMM
MLR
XGB



1475



1480 **Figure A15:** NEE, G4 - Easter Bush (UK), grassland, 70/30 strategy.



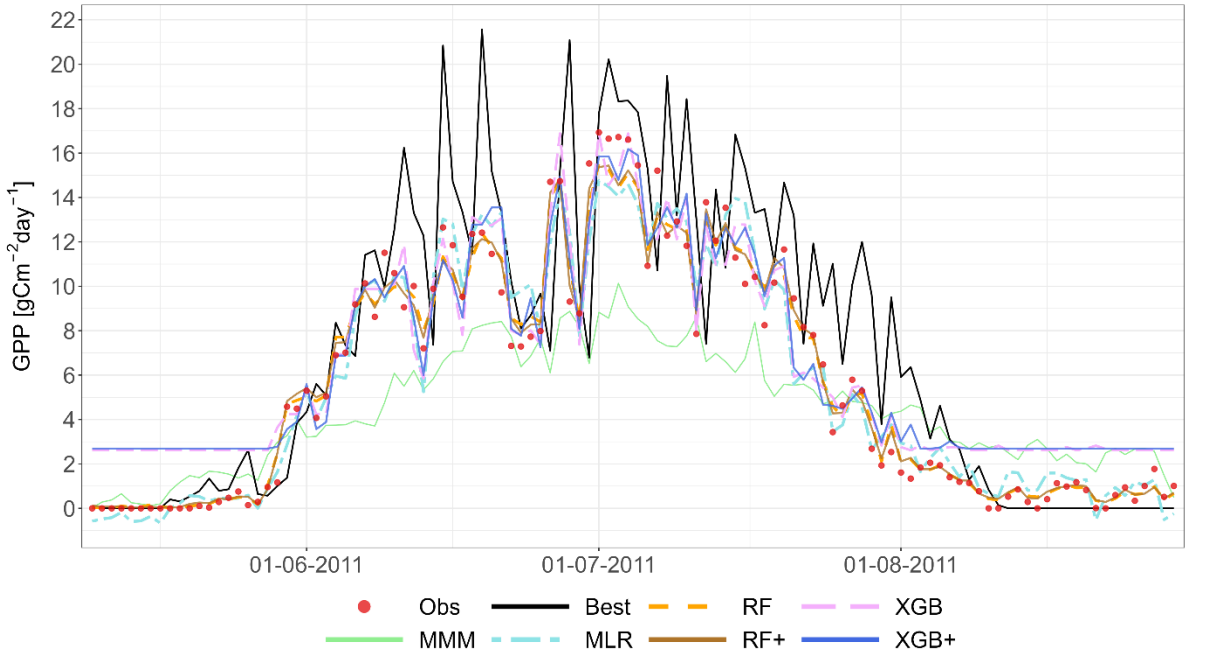


Figure A16: GPP, C1 - Ottawa (CA), cropland, 2007 and 2011, LOYO strategy.

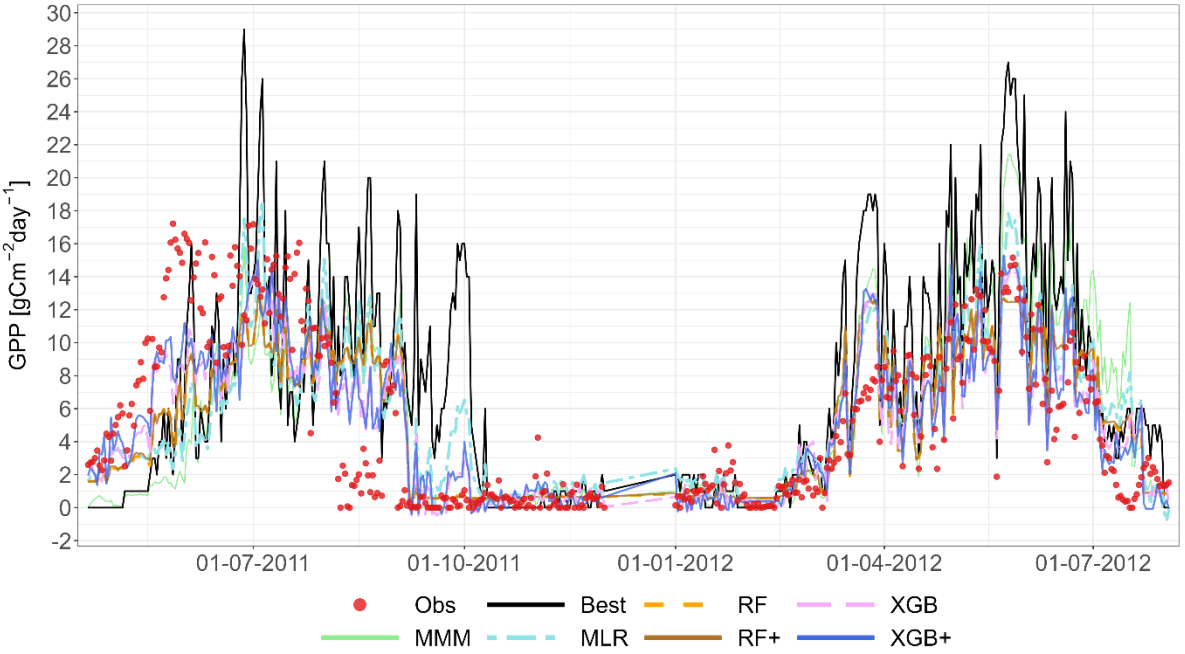
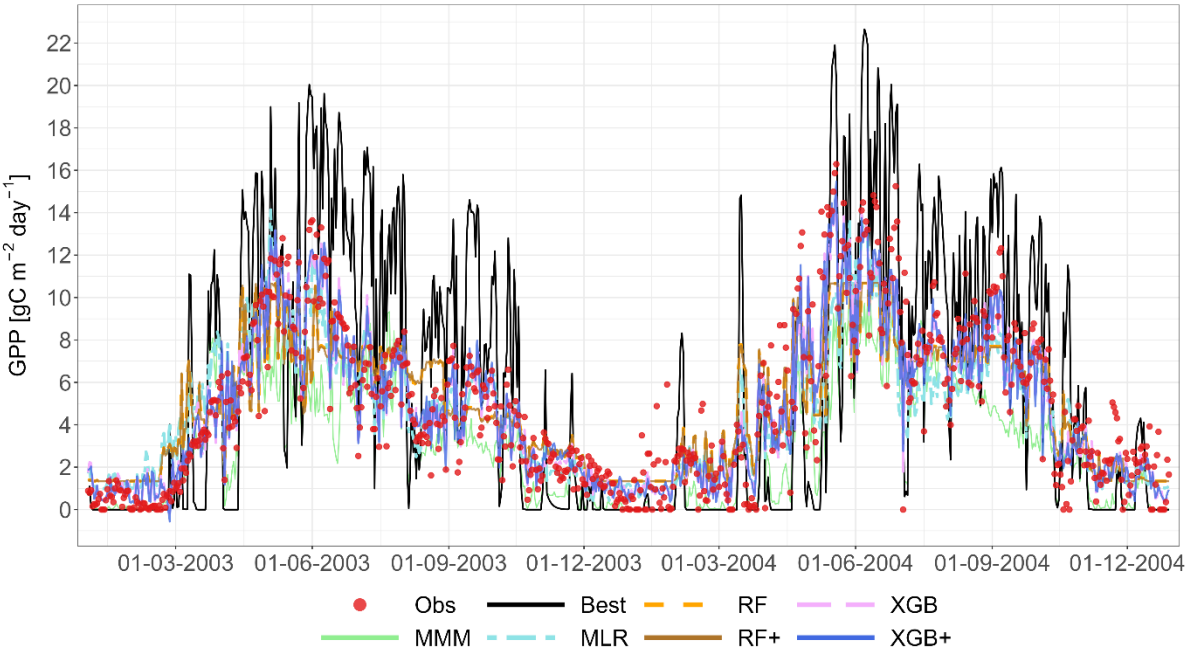
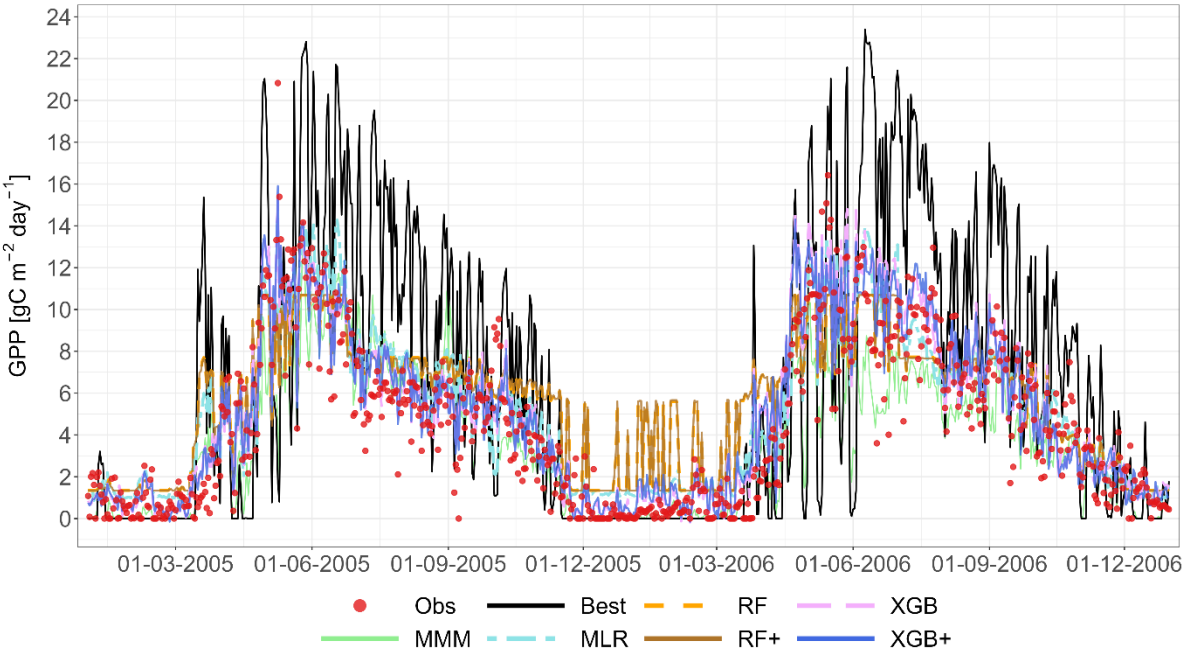


Figure A17: GPP, C2 - Grignon, (FR), cropland, 2011 and 2012, LOYO strategy.

1495


GPP [gC m^{-2} day^{-1}]
0
2
4
6
8
10
12
14
16
18
20
22
01-03-2003
01-06-2003
01-09-2003
01-12-2003
01-03-2004
01-06-2004
01-09-2004
01-12-2004
Obs
Best
RF
XGB
MMM
MLR
RF+
XGB+

1500


GPP [gC m^{-2} day^{-1}]
0
2
4
6
8
10
12
14
16
18
20
22
24
01-03-2005
01-06-2005
01-09-2005
01-12-2005
01-03-2006
01-06-2006
01-09-2006
01-12-2006
Obs
Best
RF
XGB
MMM
MLR
RF+
XGB+

1505

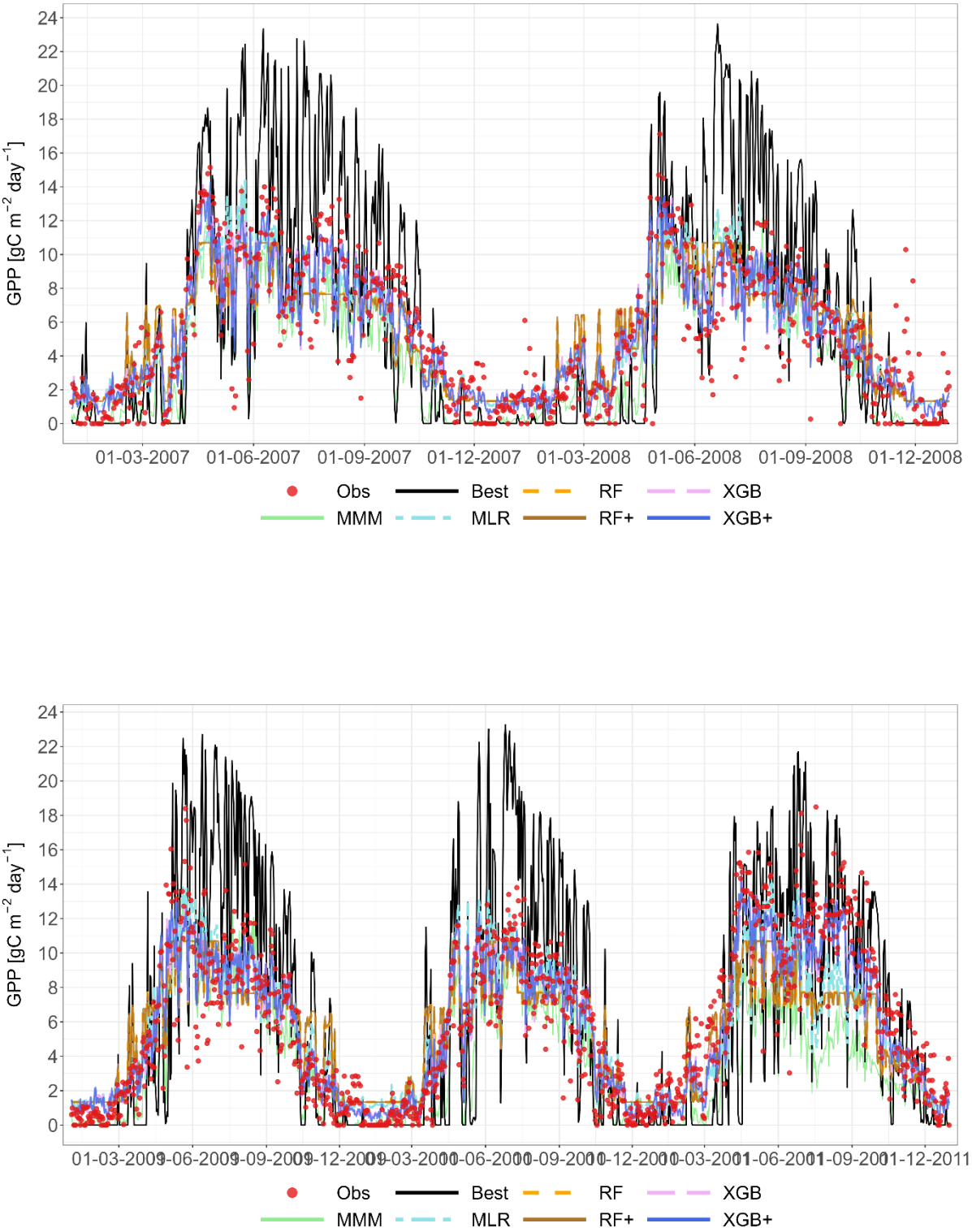
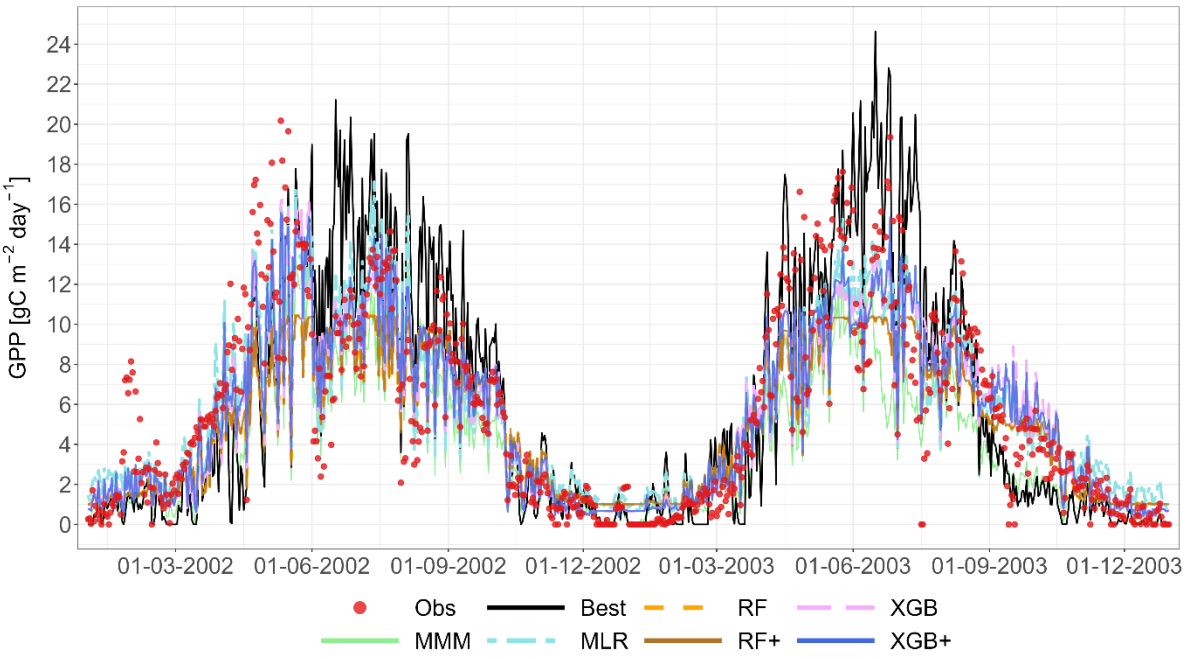
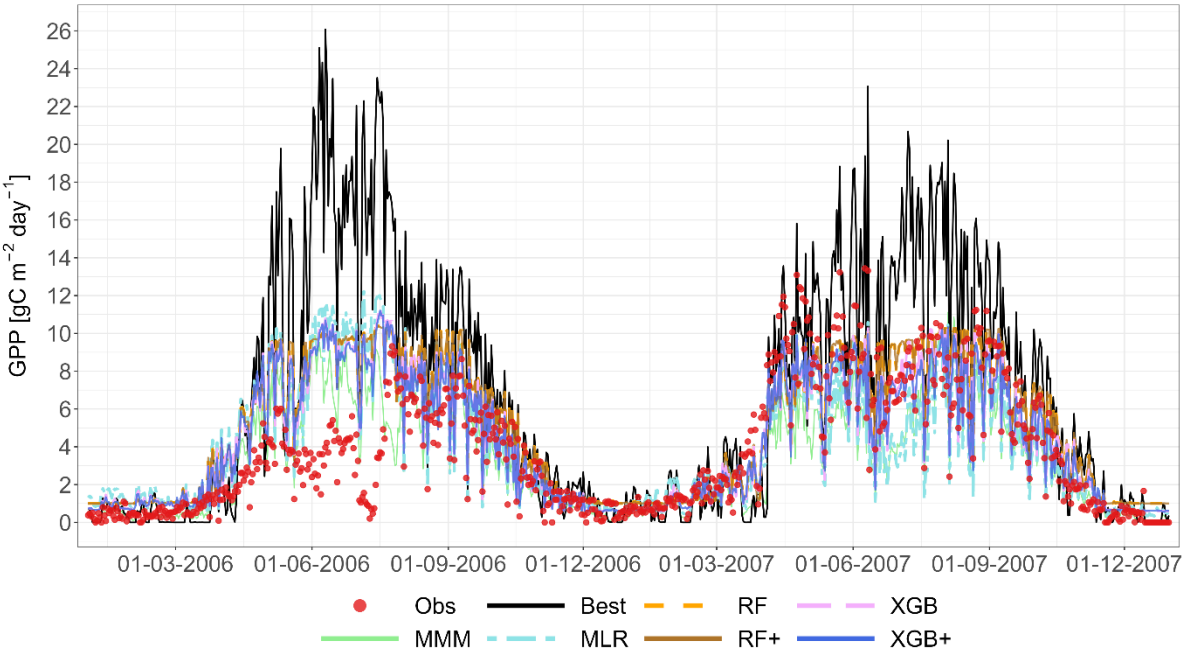
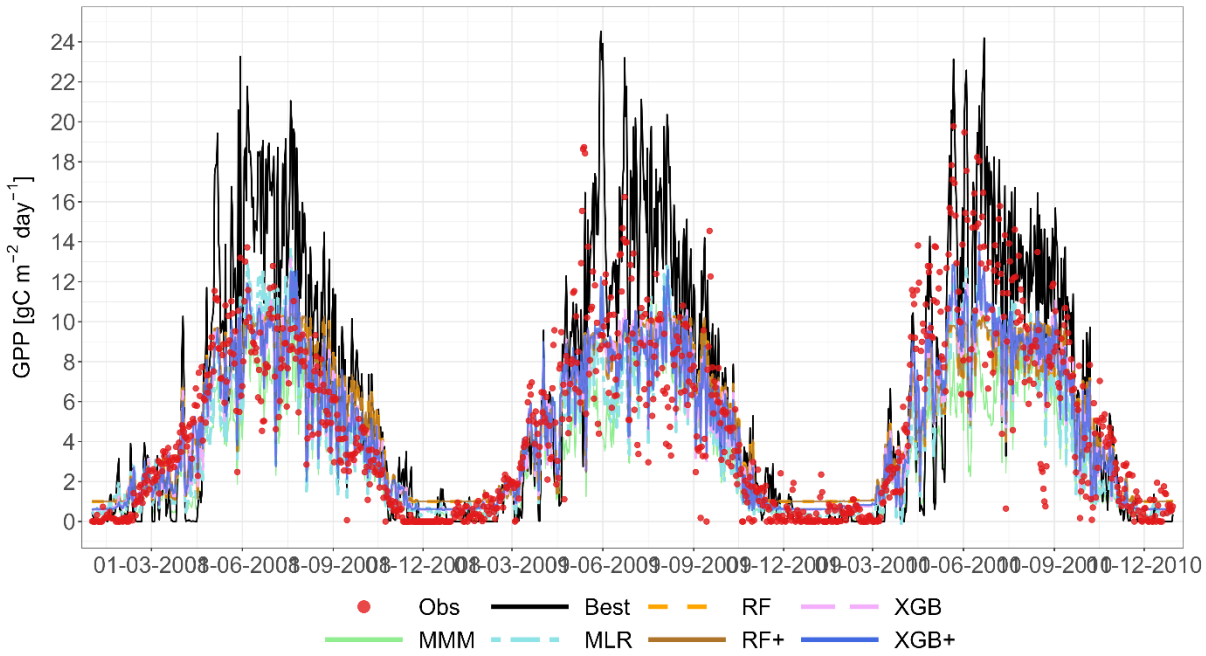


Figure A18: GPP, G3 - Laqueuille (FR), grassland, 2003-2011, LOYO strategy.


GPP [gC m^{-2} day^{-1}]
0
2
4
6
8
10
12
14
16
18
20
22
24
01-03-2002
01-06-2002
01-09-2002
01-12-2002
01-03-2003
01-06-2003
01-09-2003
01-12-2003
Obs
Best
RF
XGB
MMM
MLR
RF+
XGB+

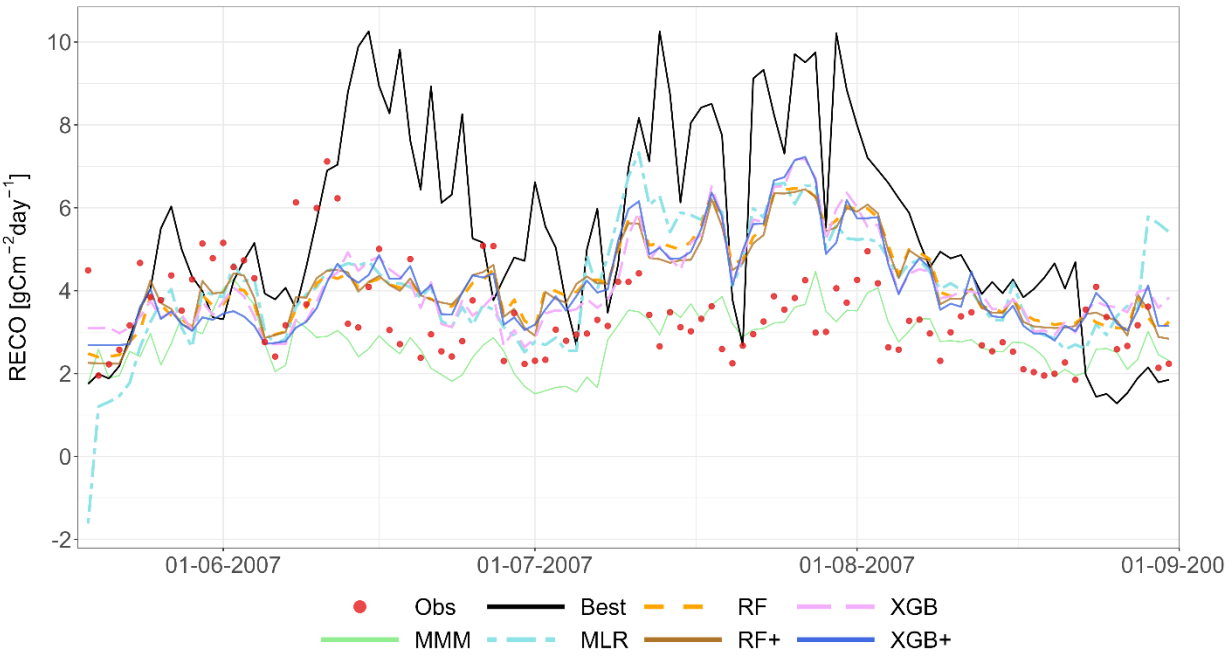
1520


GPP [gC m^{-2} day^{-1}]
0
2
4
6
8
10
12
14
16
18
20
22
24
26
01-03-2006
01-06-2006
01-09-2006
01-12-2006
01-03-2007
01-06-2007
01-09-2007
01-12-2007
Obs
Best
RF
XGB
MMM
MLR
RF+
XGB+



1525

Figure A19: GPP, G4 - Easter Bush (UK), grassland, LOYO strategy.



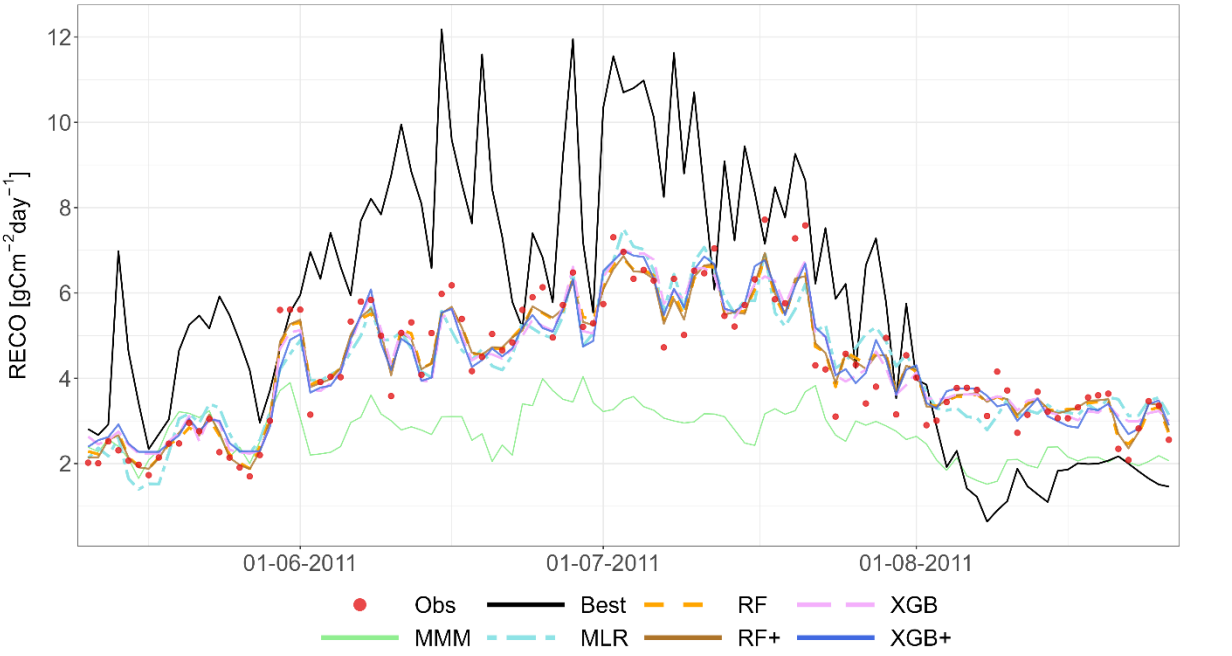


Figure A20: RECO, C1 - Ottawa (CA), cropland, 2007 and 2011, LOYO strategy.

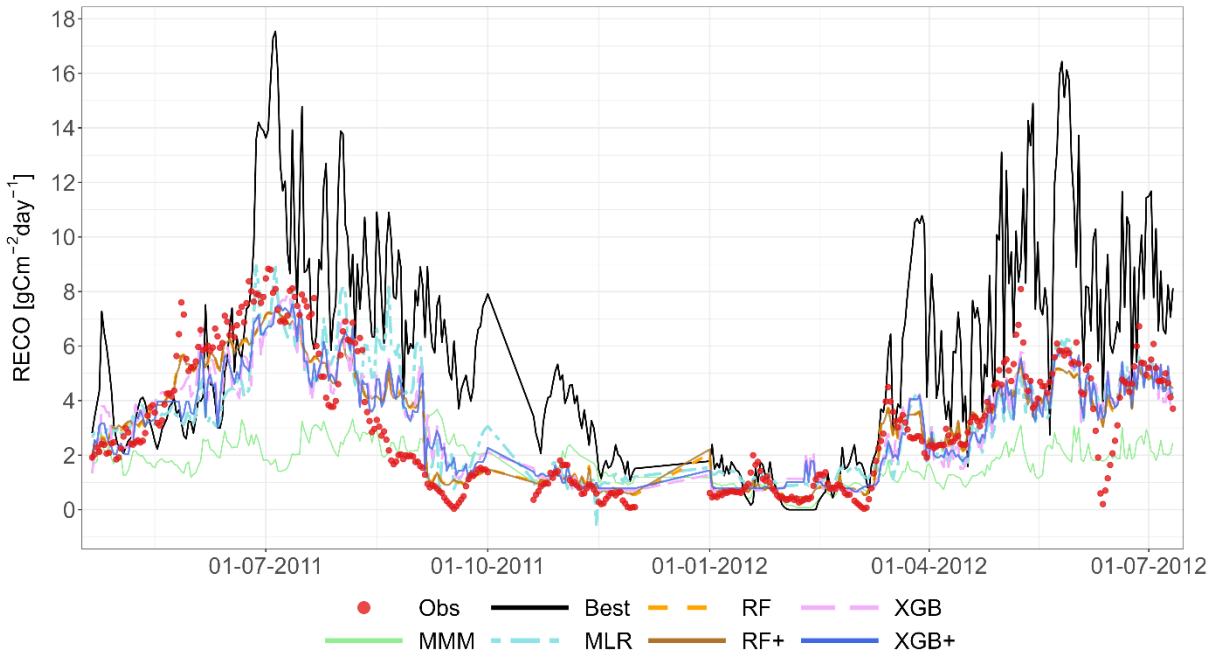
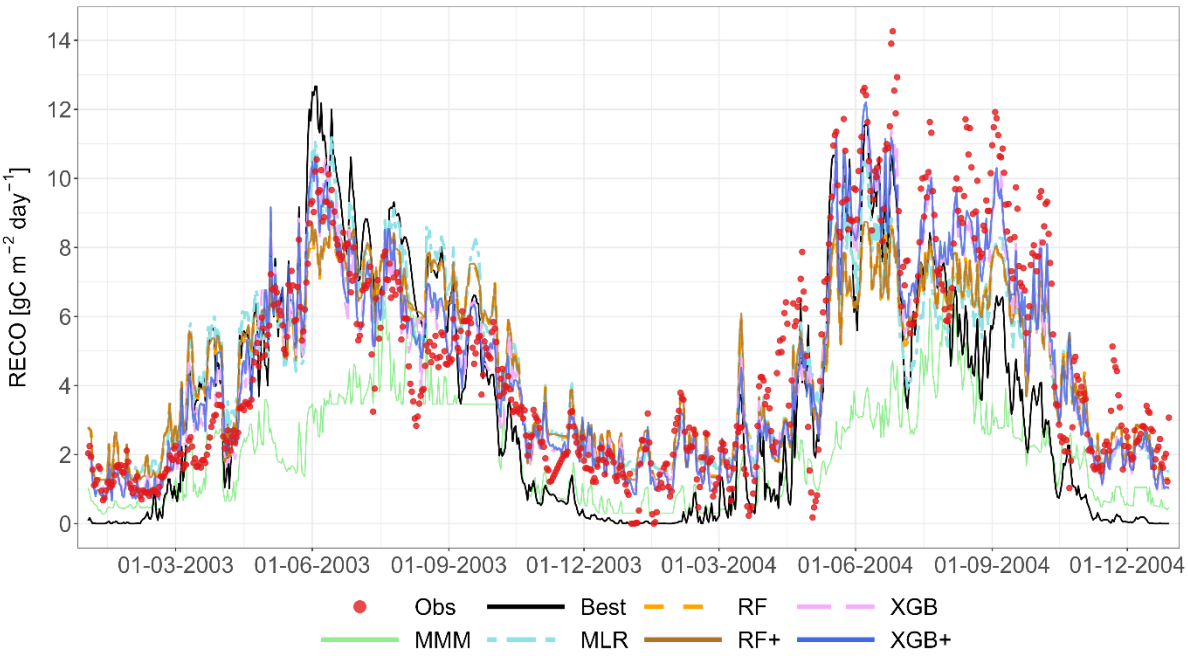
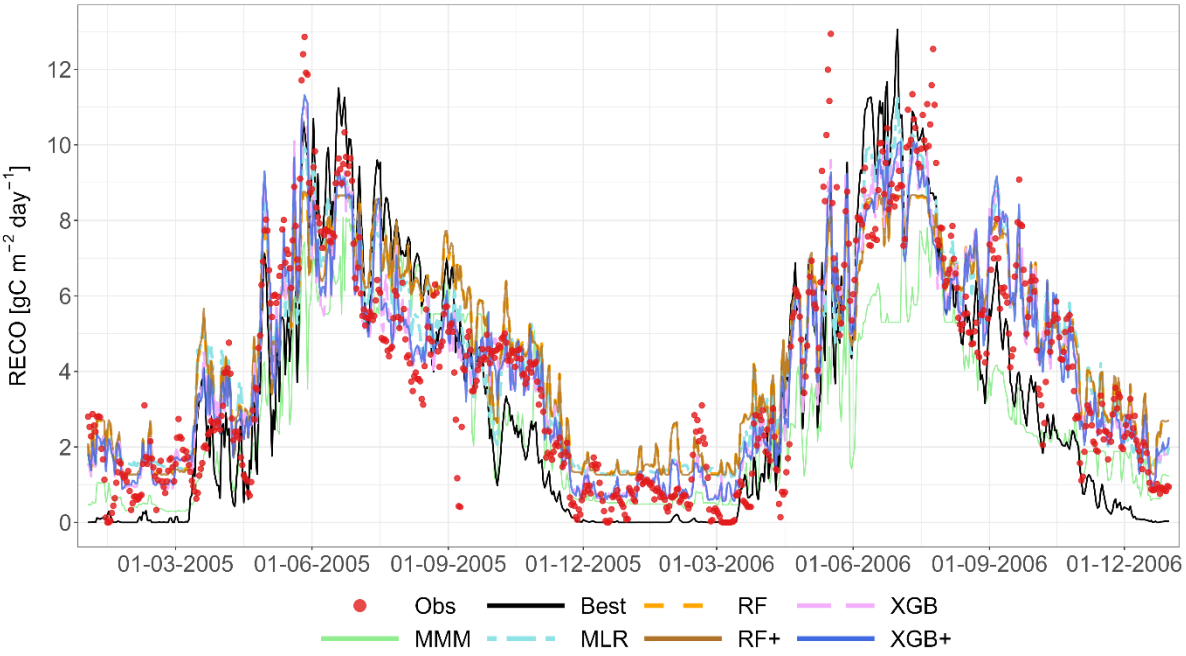


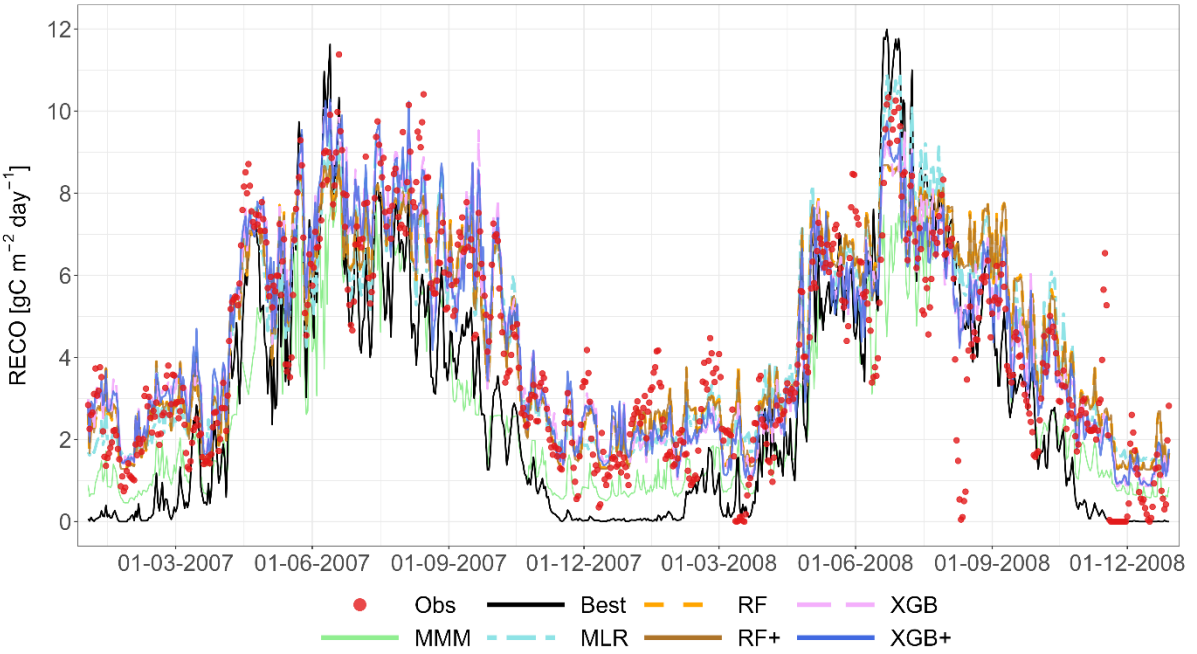
Figure A21: RECO, C2 - Grignon, (FR), cropland, 2011 and 2012, LOYO strategy.

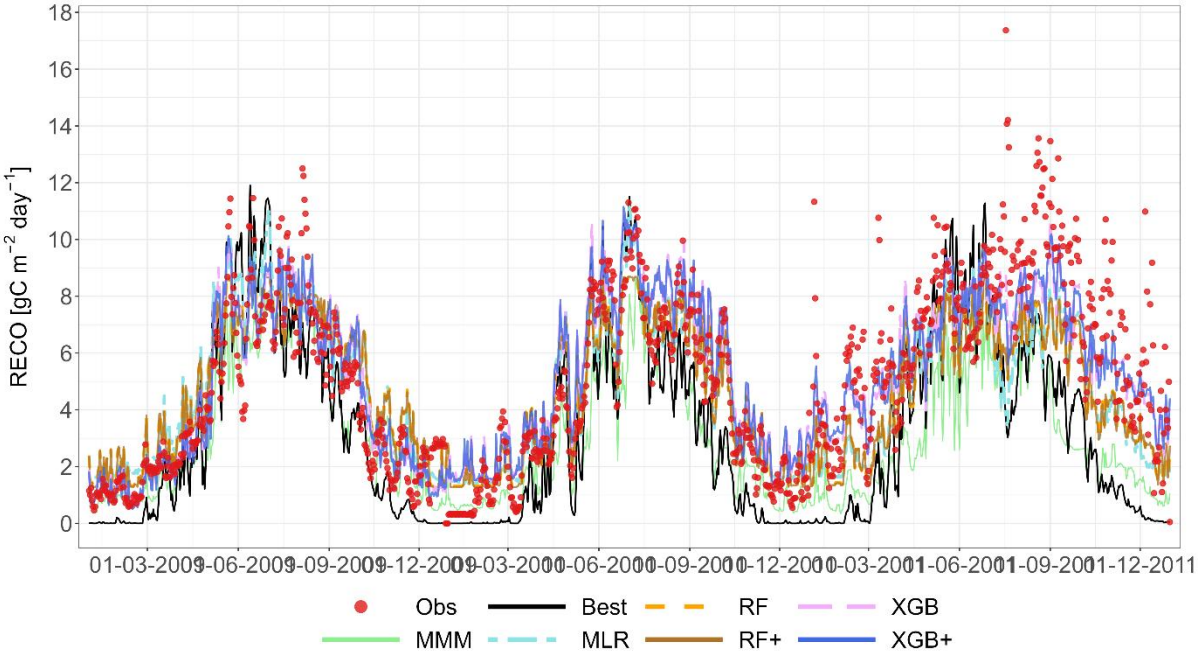

RECO [gC m^{-2} day^{-1}]
0
2
4
6
8
10
12
14
01-03-2003
01-06-2003
01-09-2003
01-12-2003
01-03-2004
01-06-2004
01-09-2004
01-12-2004
Obs
Best
RF
XGB
MMM
MLR
RF+
XGB+

1560


RECO [gC m^{-2} day^{-1}]
0
2
4
6
8
10
12
01-03-2005
01-06-2005
01-09-2005
01-12-2005
01-03-2006
01-06-2006
01-09-2006
01-12-2006
Obs
Best
RF
XGB
MMM
MLR
RF+
XGB+

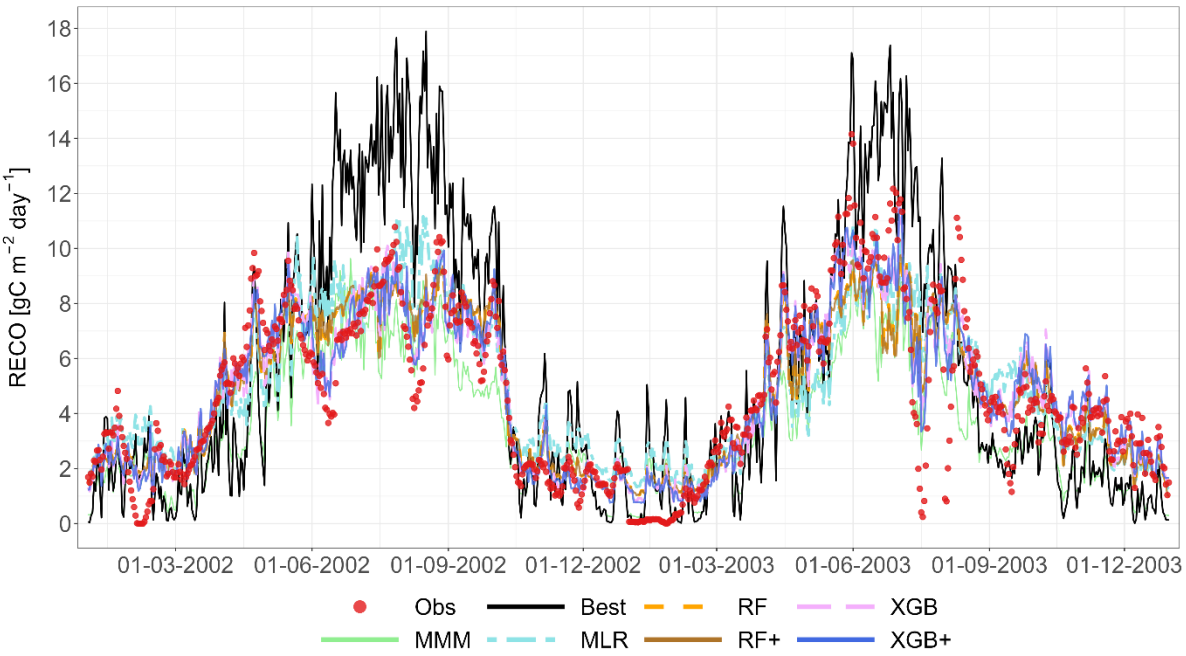
1565



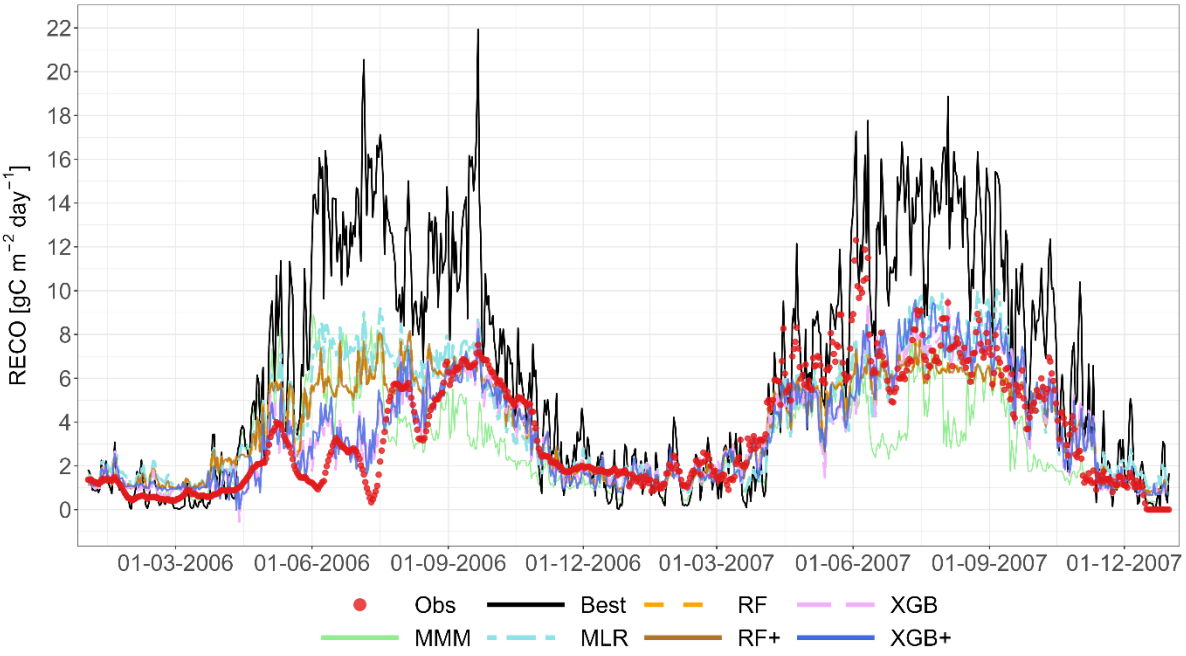


1570

Figure A22: RECO, G3 - Laqueuille (FR), grassland, 2003-2011, LOYO strategy.


RECO [gC m^{-2} day^{-1}]
0
2
4
6
8
10
12
14
16
18
01-03-2002
01-06-2002
01-09-2002
01-12-2002
01-03-2003
01-06-2003
01-09-2003
01-12-2003
Obs
Best
RF
XGB
MMM
MLR
RF+
XGB+

1575


RECO [gC m^{-2} day^{-1}]
0
2
4
6
8
10
12
14
16
18
20
22
01-03-2006
01-06-2006
01-09-2006
01-12-2006
01-03-2007
01-06-2007
01-09-2007
01-12-2007
Obs
Best
RF
XGB
MMM
MLR
RF+
XGB+

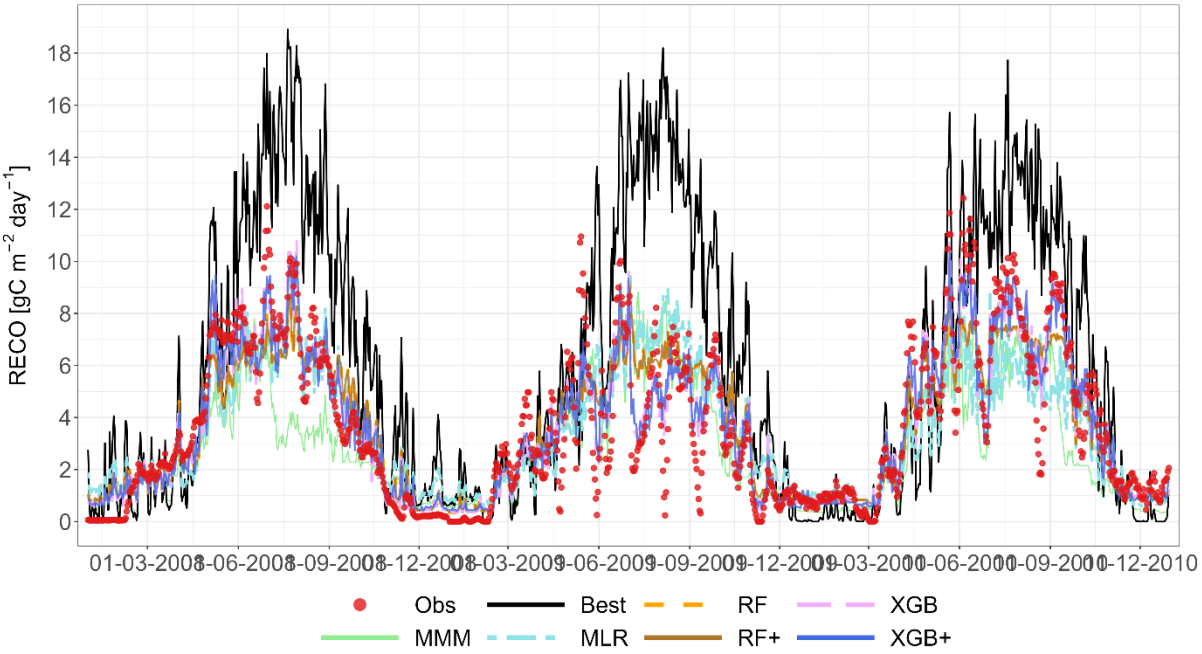


Figure A23: RECO, G4 - Easter Bush (UK), grassland, LOYO strategy.

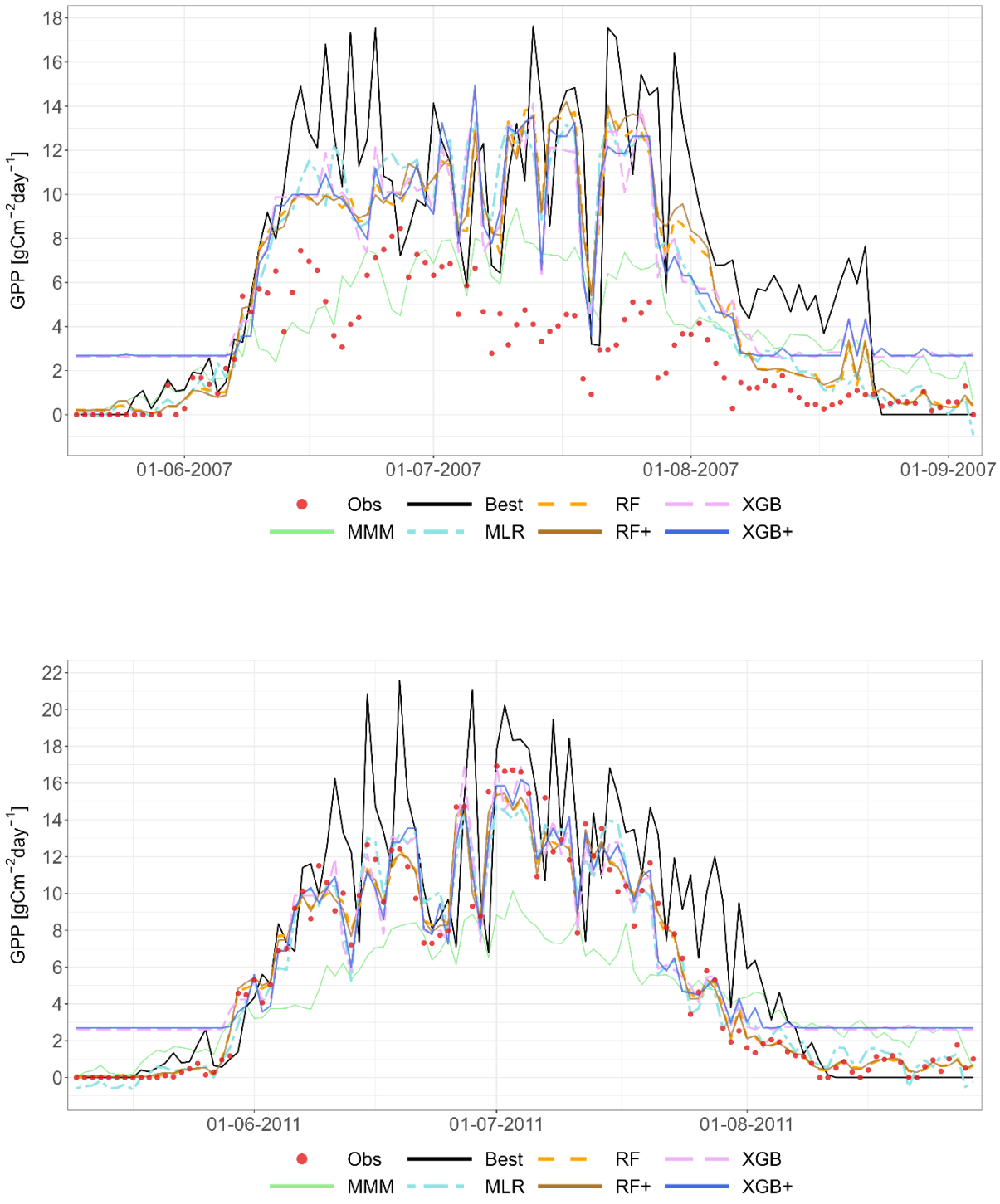


Figure A24: NEE, C1 - Ottawa (CA), cropland, 2007 and 2011, LOYO strategy.

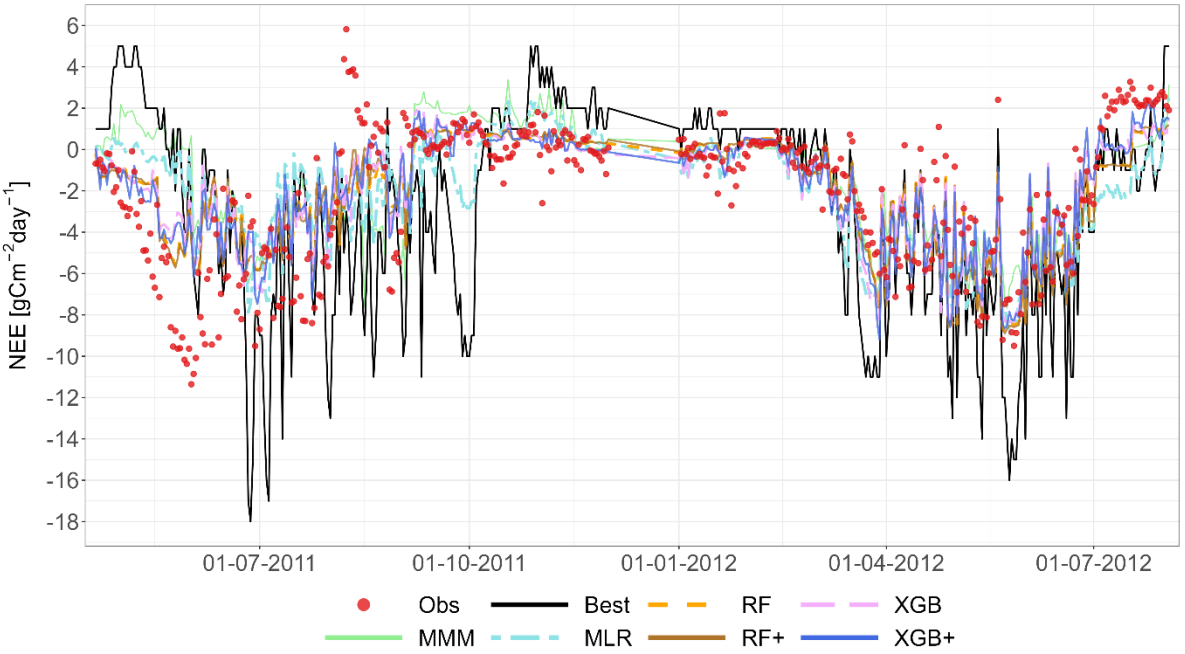
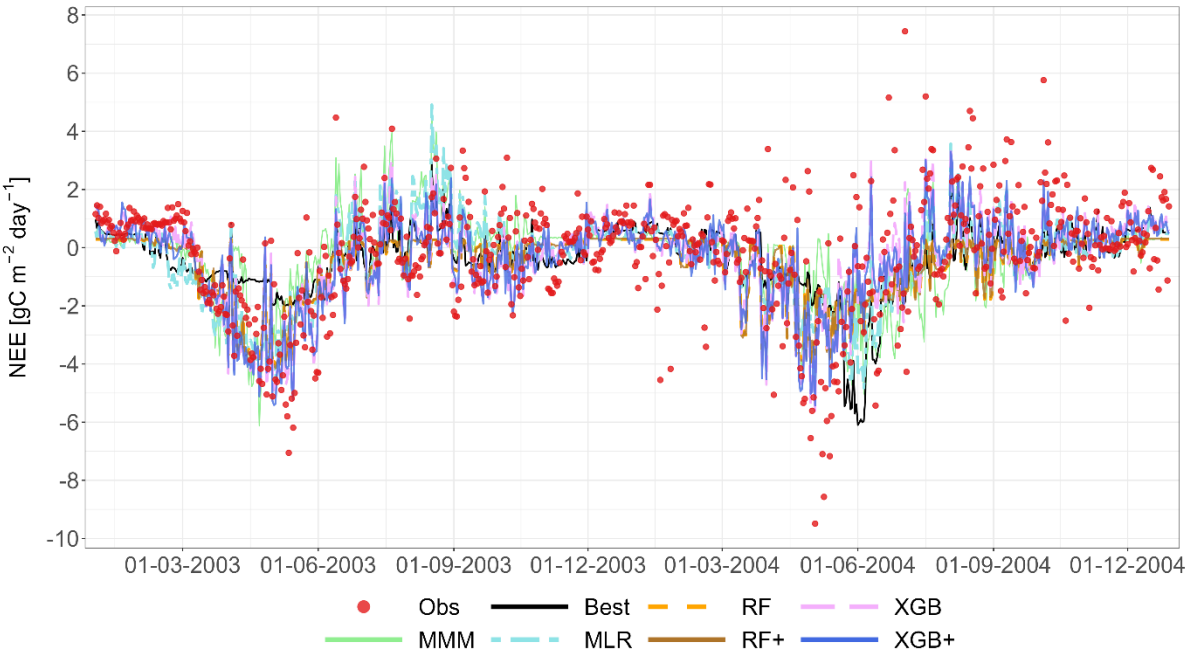
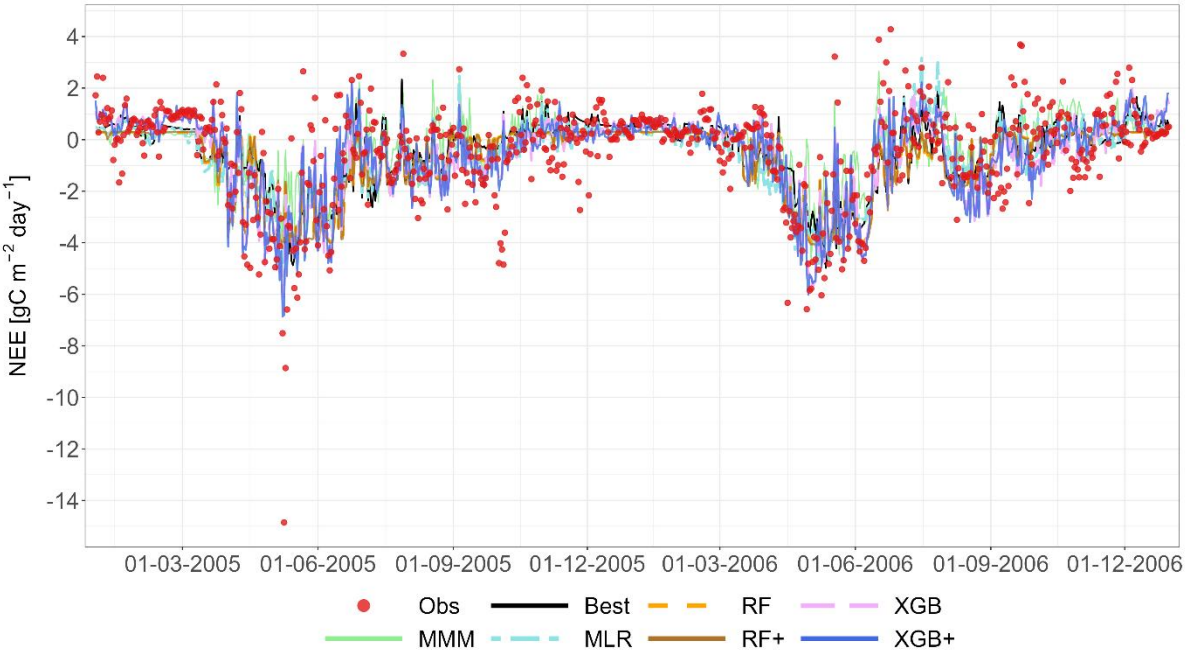


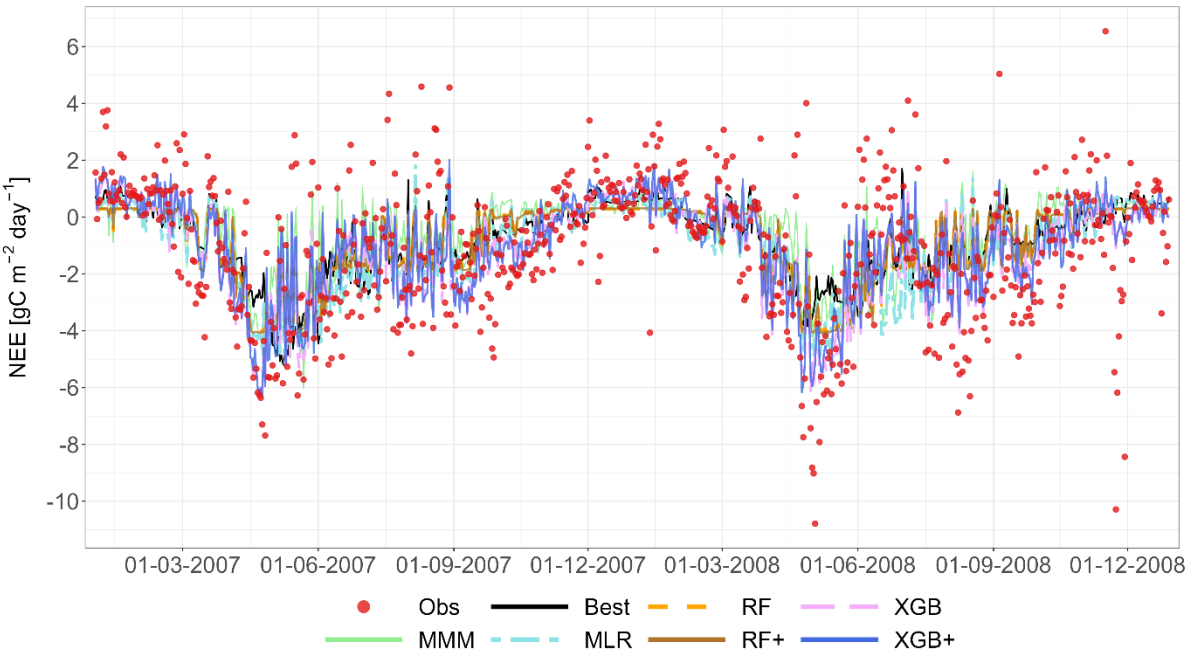
Figure A25: NEE, C2 - Grignon, (FR), cropland, 2011 and 2012, LOYO strategy.

1595


NEE [gC m^{-2} day^{-1}]
8
6
4
2
0
-2
-4
-6
-8
-10
01-03-2003
01-06-2003
01-09-2003
01-12-2003
01-03-2004
01-06-2004
01-09-2004
01-12-2004
Obs
Best
RF
XGB
MMM
MLR
RF+
XGB+


NEE [gC m^{-2} day^{-1}]
4
2
0
-2
-4
-6
-8
-10
-12
-14
01-03-2005
01-06-2005
01-09-2005
01-12-2005
01-03-2006
01-06-2006
01-09-2006
01-12-2006
Obs
Best
RF
XGB
MMM
MLR
RF+
XGB+

1600



1605

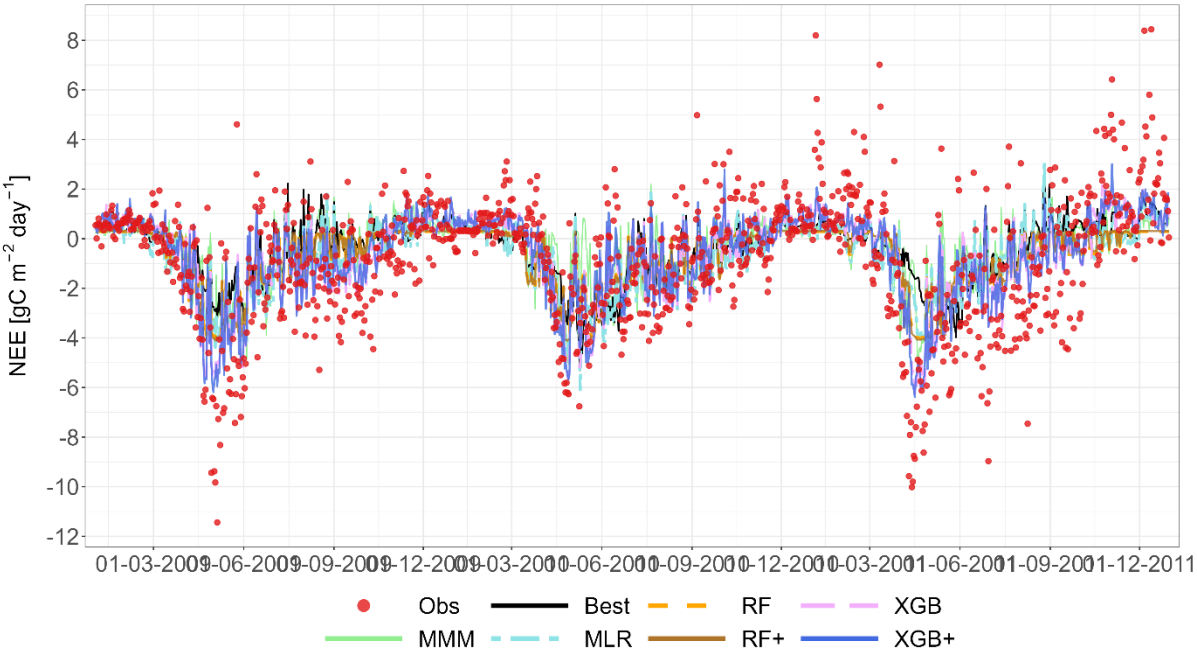
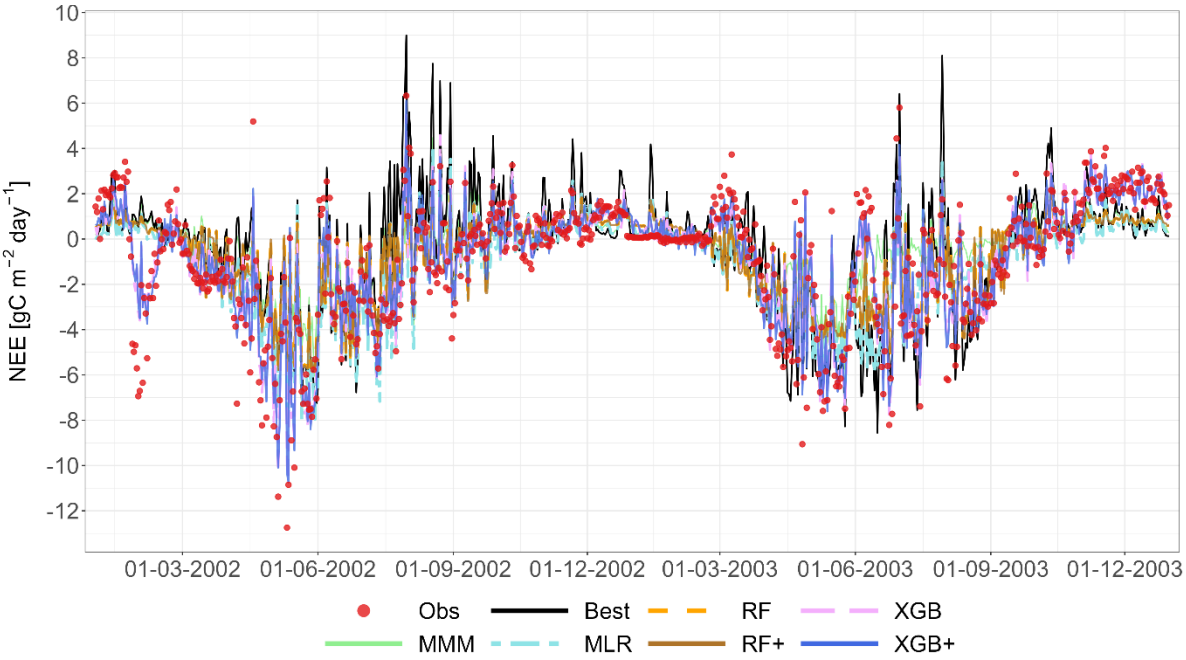
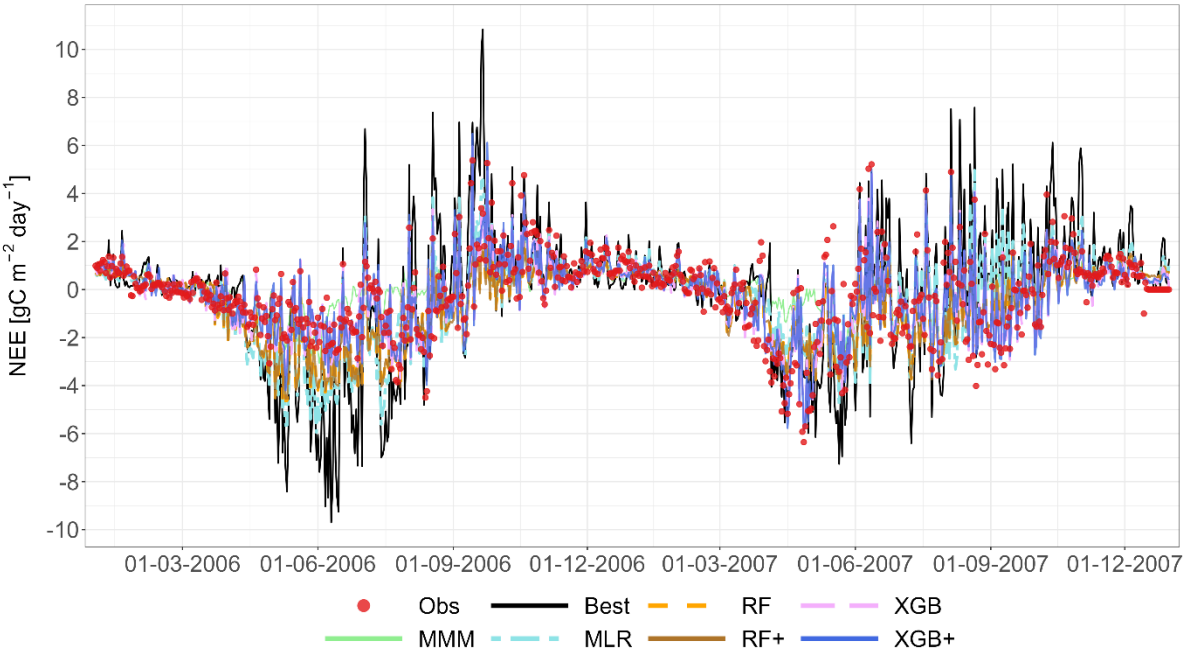


Figure A26: NEE, G3 - Laqueuille (FR), grassland, 2003-2011, LOYO strategy.


NEE [gC m^{-2} day^{-1}]
10
8
6
4
2
0
-2
-4
-6
-8
-10
-12
01-03-2002
01-06-2002
01-09-2002
01-12-2002
01-03-2003
01-06-2003
01-09-2003
01-12-2003
Obs
Best
RF
XGB
MMM
MLR
RF+
XGB+

1610


NEE [gC m^{-2} day^{-1}]
10
8
6
4
2
0
-2
-4
-6
-8
-10
01-03-2006
01-06-2006
01-09-2006
01-12-2006
01-03-2007
01-06-2007
01-09-2007
01-12-2007
Obs
Best
RF
XGB
MMM
MLR
RF+
XGB+

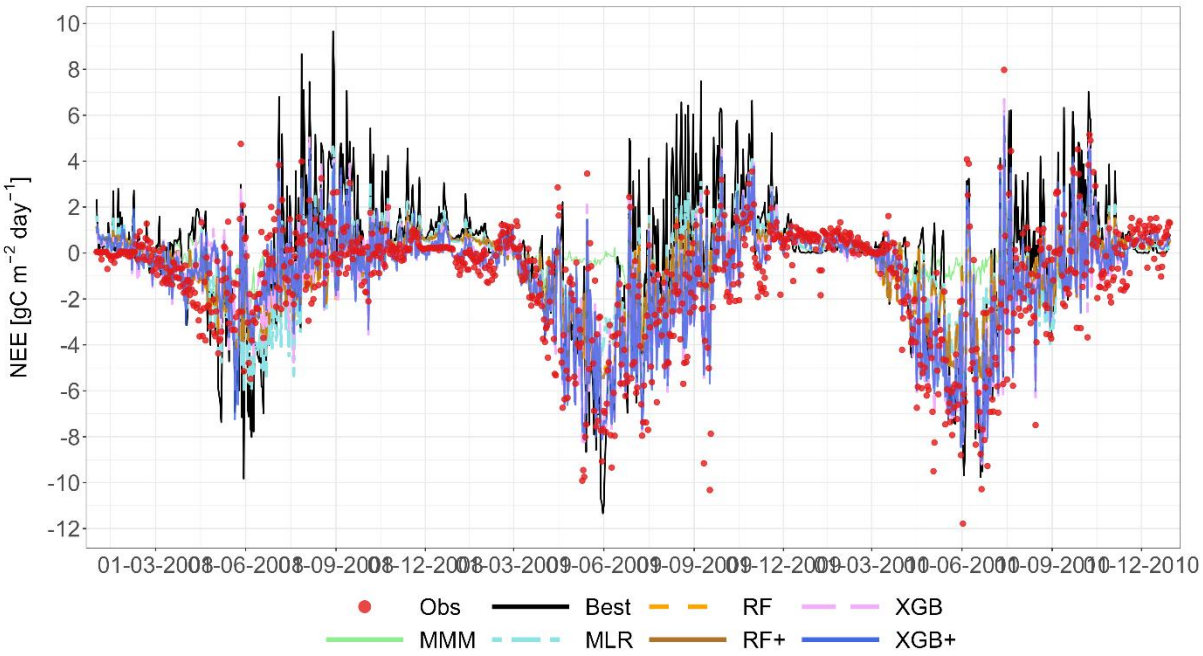


Figure A27: NEE, G4 - Easter Bush (UK), grassland, LOYO strategy.

Appendix B: XGBoost Hyperparameter Optimization Details

This appendix provides detailed information about the hyperparameter optimization procedure used for the XGBoost (XGB) and XGBoost-plus (XGB+) meta-models described in the main manuscript. Both model variants use identical hyperparameter ranges and optimization procedures, differing only in their input features: XGB uses only process-based model predictions, while XGB+ additionally incorporates meteorological variables.

B1. Optimization Methodology

Hyperparameter optimization was performed using random search with 50 evaluations per model. For each evaluation, seven hyperparameters were randomly sampled from predefined ranges (Table B1): learning rate (eta), maximum tree depth (max_depth), row and column sampling ratios (subsample and colsample_bytree), minimum child weight (min_child_weight), and L1 and L2 regularization terms (alpha and lambda).

Each parameter combination was evaluated using an internal 80/20 split of the training data, ensuring that test data were never used during hyperparameter selection. Models were trained for up to 1000 boosting rounds with early stopping: training terminated if validation RMSE did not improve for 20 consecutive rounds. The hyperparameter set achieving the lowest validation RMSE was selected, and the final model was retrained on the full training data using the optimal number of rounds determined during tuning.

This optimization procedure was performed independently for each combination of site (C1, C2, G3, G4), variable (GPP, NEE, RECO), and Leave-One-Year-Out cross-validation fold, resulting in 148 optimized models. The random search was parallelized across 15 CPU cores to improve computational efficiency.

Table B1. Hyperparameter search ranges for XGBoost models.

Parameter	Description	Search Range	Type
eta	Learning rate	0.01 – 0.3	Continuous
max_depth	Maximum tree depth	3 – 10	Integer
subsample	Row sampling ratio	0.6 – 1.0	Continuous
colsample_bytree	Column sampling ratio per tree	0.6 – 1.0	Continuous
min_child_weight	Minimum sum of instance weight in child	1 – 10	Integer
lambda	L2 regularization term	0.01 – 10.0	Continuous
alpha	L1 regularization term	0.0 – 10.0	Continuous

Note: All continuous parameters were sampled uniformly from the specified ranges; integer parameters were sampled uniformly from discrete values within the range.

B2. Selected Hyperparameter Values

Table B2 summarizes the hyperparameter values selected by the optimization process across all 148 trained models. The typical values (medians) generally fall in the middle of the search ranges, indicating effective exploration of the parameter space. Most models trained for the full 1000 rounds (typical value: 1500, after the internal tuning phase), suggesting that the early stopping criterion was rarely triggered and models benefited from extensive training.

Table B2. Selected hyperparameter values across all XGB and XGB+ models.

Parameter	Typical Value	Mean	Range [Min – Max]	Interquartile Range
eta	0.22	0.20	0.07 – 0.28	0.16 – 0.23
max_depth	3	3.2	2 – 6	2 – 4

subsample	0.68	0.70	0.53 – 0.89	0.54 – 0.81
colsample_bytree	0.48	0.58	0.40 – 0.93	0.44 – 0.70
min_child_weight	10	8.8	3 – 12	5 – 12
lambda	3.40	3.22	0.98 – 4.89	2.44 – 4.11
alpha	0.05	0.06	0.00 – 0.18	0.03 – 0.06
nrounds	1500	1390	57 – 1500	1500 – 1500

Note: Summary statistics based on 148 trained models: 4 sites × 3 variables (GPP, NEE, RECO) × 2-9 LOYO folds per site × 2 model types (XGB and XGB+). "Typical Value" represents the median across all models. The nrounds value is the actual number of boosting rounds used after early stopping (maximum was 1000).

B3. Variable-Specific Hyperparameter Patterns

Table B3 shows the selected hyperparameter values separately for each target variable. Notable differences emerge: GPP predictions favor shallower trees (typical depth = 2) with higher column sampling (0.75), NEE uses intermediate complexity (depth = 3), while RECO employs deeper trees (depth = 5) with lower column sampling (0.43). These patterns suggest that the optimization procedure successfully adapted model complexity to the characteristics of each target variable.

Table B3. Selected hyperparameter values by target variable.

Variable	Parameter	Typical Value	Range [Min – Max]
GPP (25 XGB + 25 XGB+)	eta	0.16	0.13 – 0.27
	max_depth	2	2 – 4
	subsample	0.75	0.62 – 0.89
	colsample_bytree	0.75	0.44 – 0.93
	min_child_weight	11	4 – 12
	lambda	3.81	1.67 – 4.45
	alpha	0.07	0.03 – 0.18
	nrounds	1500	144 – 1500
NEE (25 XGB + 25 XGB+)	eta	0.23	0.10 – 0.28
	max_depth	3	2 – 6
	subsample	0.68	0.53 – 0.84
	colsample_bytree	0.68	0.40 – 0.70
	min_child_weight	9	3 – 12
	lambda	2.44	0.98 – 4.11
	alpha	0.06	0.01 – 0.06
	nrounds	1500	57 – 1500
RECO (24 XGB + 24 XGB+)	eta	0.22	0.07 – 0.23
	max_depth	5	3 – 5
	subsample	0.64	0.53 – 0.81
	colsample_bytree	0.43	0.41 – 0.85
	min_child_weight	10	5 – 12
	lambda	3.40	1.61 – 4.89
	alpha	0.05	0.00 – 0.05
	nrounds	1500	215 – 1500

Note: Each site-variable-fold combination has 2 models (XGB and XGB+). Number of LOYO folds per site: C1=2, C2=5, G3=9, G4=9. Total expected: (2+5+9+9) × 3 variables × 2 models = 150 models. The actual count of 148 models indicates 2 RECO models were not successfully trained.

B4. Fixed XGBoost Parameters

In addition to the optimized hyperparameters, several XGBoost parameters were held constant across all models (Table B4). These included the objective function (squared error regression), evaluation metric (RMSE), tree construction method (exact greedy algorithm), and early stopping criterion (20 rounds without improvement).

Table B4. Fixed XGBoost parameters (not optimized).

Parameter	Value	Description
objective	reg:squarederror	Regression with squared error loss function
eval_metric	rmse	Root Mean Squared Error for evaluation
tree_method	exact	Exact greedy tree construction algorithm
nrounds (max)	1000	Maximum number of boosting rounds allowed
early_stopping_rounds	20	Stop training if validation RMSE does not improve for 20 rounds