Causal hybrid modeling with double machine learning - Applications in carbon flux modeling

Causal hybrid modeling with double machine learning – Applications in carbon flux modeling

Kai-Hendrik Cohrs[1], **Gherardo Varando**[1], **Nuno Carvalhais**[2,3] **Markus Reichstein**[2,3] **& Gustau Camps-Valls**[1]

[1] Image Processing Laboratory (IPL), Universitat de València
[2] Max Planck Institute for Biogeochemistry, Jena, Germany.
[3] ELLIS Unit Jena

E-mail: kai.cohrs@uv.es

December 2023

Abstract. Hybrid modeling integrates machine learning with scientific knowledge to enhance interpretability, generalization, and adherence to natural laws. Nevertheless, equifinality and regularization biases pose challenges in hybrid modeling to achieve these purposes. This paper introduces a novel approach to estimating hybrid models via a causal inference framework, specifically employing Double Machine Learning (DML) to estimate causal effects. We showcase its use for the Earth sciences on two problems related to carbon dioxide fluxes. In the Q_{10} model, we demonstrate that DML-based hybrid modeling is superior in estimating causal parameters over end-to-end deep neural network (DNN) approaches, proving efficiency, robustness to bias from regularization methods, and circumventing equifinality. Our approach, applied to carbon flux partitioning, exhibits flexibility in accommodating heterogeneous causal effects. The study emphasizes the necessity of explicitly defining causal graphs and relationships, advocating for this as a general best practice. We encourage the continued exploration of causality in hybrid models for more interpretable and trustworthy results in knowledge-guided machine learning.

Keywords: Knowledge-guided machine learning, Hybrid modeling, Causal effect estimation, Double machine learning, Temperature sensitivity, Carbon flux partitioning

1. Introduction

Machine learning (ML), specifically deep learning (DL), has proven to be effective in identifying and modeling complex patterns from data sets. This led to unprecedented progress in fields such as computer vision [1], natural language processing [2], and speech recognition [3]. These data-driven models also increasingly complement or even substitute mechanistic methods in science [4, 5].

In the Earth sciences, for instance, the common way to understand and model the Earth's properties, structure, and processes is using knowledge of first principles, realized in mechanistic models based on functional equations [6]. These models allow principled predictions of how the system under study would behave under different conditions [7]. Nevertheless, they are not always sufficient to capture the complex and usually not completely known relationships in the real world.

Computational constraints and missing understanding have led to simplified or even missing representation of important processes in the current generation of climate models [8]. Structural limitations often necessitate parameterizations to approximate complex processes. Significant uncertainties include the representation of cloud feedbacks [9], resolving ocean components at varying resolutions [10], surface energy partitioning [11], representing key processes like vegetation response to CO_2 [12], and difficulties in representing functional structures across different biome types [13]. Addressing these challenges is essential for enhancing the accuracy and reliability of Earth system models in projecting future climate change and weather extremes.

Integration of machine learning (ML) with abundant Earth data presents a promising avenue to overcome the limitations of current Earth system models [14, 15]. Support vector machines [16], random forests (RFs) [17], or neural networks (NNs) [18] are highly flexible, make little prior assumptions on the functional form and can integrate the large datasets abundant in Earth and climate sciences.

The flexibility of ML models comes with some known downsides: (i) Many popular machine learning models are black boxes, meaning that we do not understand the internal reasoning behind the model's predictions [19]. (ii) Often, ML models are not robust and fail to generalize out of the domain of the data used for training [20, 21]. (iii) They violate physical properties and laws of nature, such as conservation laws, symmetries, or equi- and invariances [14, 22]. These are crucial matters in Earth and climate sciences, where a prime goal is to make realistic predictions on the Earth's system under a changing climate [23].

All these issues are gaining attention in ML and Earth system science literature. Research in generalization and extrapolation aims at ensuring robustness outside of the training domain [24–26]. Explainable artificial intelligence (XAI) tackles questions on the explainability of black box models [27–29], which find growing usage in remote sensing problems [30, 31]. At the same time, the general goal of explaining black boxes is being challenged by advocates for glass box models, i.e., inherently interpretable models [32, 33], and there is an ongoing debate on the evaluation and rigorousness of

XAI methods [34, 35].

A flourishing area of research is science-aware or knowledge-guided machine learning (KGML), which combines the knowledge-driven and data-driven worlds to overcome inconsistencies [36]. These methods increasingly find their way into various domains within Earth sciences [37–42]. One example is physics-informed neural networks (PINNs) [43], where an additional term is added to the loss for training that punishes deviations from physical laws encoded with ODEs or PDEs. Alternatively, ML models can be trained on a combination of data and simulations from physical models to improve consistency in the sparse observation regime [37].

Finally, hybrid modeling replaces some components of mechanistic models with machine learning [44–46]. This constraint makes the models more interpretable and serves as a regularizer for better generalization to unseen data. If we use deep learning models as the machine learning component, the only requirement for fitting these hybrid models is that the parametric components are differentiable [47]. Then, gradient-based optimization allows joint optimization of the neural network (NN) parameters and physical parameters of the mechanistic model and leads to seamless data integration. In the following, we will refer to this as *gradient-descent-based hybrid modeling (GD-based HM)*. It serves as a baseline for our proposed method.

There are persisting challenges in hybrid modeling. Firstly, these models are prone to *equifinality*, which denotes the existence of multiple models and sets of parameters that describe the data similarly well. Already in the standard mechanistic modeling, this is a well-known difficulty when not only model performance but also retrieving meaningful parameters is the goal. In this setting, robust inference already poses a challenge [48], which becomes even more difficult and prohibitively expensive in deep learning [49, 50]. Ultimately, equifinality can jeopardize the interpretability of the results. Second, regularization techniques in machine learning can introduce bias on the physical parameters [45]. Finally, given the flexibility of non-parametric models such as NNs, it is tempting to use different sets of variables for the model and choose the ones that lead to the best overall performance. For a pure prediction task, that is a sensible procedure [51]. For hybrid modeling, though, apart from equifinality, this can lead to bias or different interpretations of the parameter of interest in the causal sense. We might be *right for the wrong reasons* and imperil the desired interpretability of the hybrid model (see Box 1 for an illustrative example).

In many instances, physical equations encode actual cause-effect relationships. It is essential to capture the causal relationships between the variables to obtain interpretable and more accurate models. Respecting the causal direction of time has shown to be effective in training PINNs for chaotic systems where previous approaches failed [52]. Furthermore, coupling causal discovery to identify the causal drivers in climate models before applying deep learning algorithms improved performance and interpretability [53, 54]. Causally constrained recurrent NNs more accurately reflect underlying processes and were shown to enhance our understanding of methane in wetlands [55]. Ultimately, causality aims at *being right for the right reasons*.

Therefore, we believe it is time for a *causal hybrid modeling* framework, where we introduce an explicit physical prior by assuming a causal graph and framing the problem as a causal effect estimation problem within the hybrid modeling framework. We will show how this approach leads to well-defined problems, thus mitigating equifinality and being robust to biases of training and regularization. As a first step, we propose a method based on double machine learning (DML) [56]. DML is a causal effect estimation technique developed in econometrics, where it is common to investigate the effect of some proposed treatment on an outcome variable [57, 58]. It has recently been used for effect estimation in the environmental sciences [59]. We suggest that this causal effect estimation technique can be applied to a class of hybrid models where the effect of some input driver on the output is encoded. We coin this method *DML-based hybrid modeling (DML-based HM)*.

Apart from the causal perspective, DML has favorable properties over naive fitting approaches. Regularization of the estimators for the non-parametric part of the equation can introduce substantial bias in estimating the parametric part of the equation. Using DML, even for erroneous estimators, we can still obtain consistent estimators of the causal effect coefficient. This is particularly useful if the confounding effects are high-dimensional or are described by a complicated function that is hard to learn. Furthermore, it enables us to make inferences, as the estimators are shown to be approximately normally distributed, which yields confidence intervals [56].

Within the proposed framework based on DML, we can solve problems that can be transformed into a regression problem of the form

$$Y = \theta(X) \cdot f(T) + g(X, W), \tag{1}$$

where T is a one-dimensional input variable and X and W are further sets of predictors. We assume that f is a known transformation of T, and our hybrid modeling goal is to estimate the non-parametric functions θ and g. We will see relevant examples of problems that fall into this class. This includes, in particular, the problems where θ describes the effect of T on Y. This effect can be constant or depend on some other predictors X.

We demonstrate the advantages of DML-based HM in two examples around carbon fluxes:

(i) The temperature sensitivity Q_{10} model for ecosystem respiration [60–62] and,
(ii) the light-use efficiency model for carbon flux partitioning [63].

These two models are particularly relevant as they allow statements on the productivity and respiration of plants under changing conditions.

Our contributions are as follows: In the case of synthetic data for Q_{10}, DML retrieves the Q_{10} temperature sensitivity parameter more robustly and efficiently than the GD-based HM approach, especially in the low data regime and under regularization.

It retrieves Q_{10} values consistent with the literature on measured respiration data. We show how equifinality can yield misleading results and how causal prior knowledge can solve the problem without giving up flexibility. In the carbon flux partitioning problem, we show how the method can be extended to the non-linear heterogeneous case, where the hybrid modeling retrieves consistent fluxes and shows competitive performance to the current state-of-the-art neural network.

In essence, we introduce DML-based HM as a novel approach to fitting hybrid models and show that the obtained estimates are more efficient and robust than the ones from GD-based HM. We describe a path to better pose problems with equifinality, enforcing causal interpretability instead of hoping for it.

Box 1: Equifinality in hybrid modeling

Modeling the temperature dependence of ecosystem respiration R_{eco} is a fundamental step in better understanding biosphere evolution and responses under global warming scenarios [64–66]. The functional relationship between temperature and respiration has been classically represented via the Q_{10} respiration model:

$$R_{eco}(X, T_A) = R_b(X, T_A) \cdot Q_{10}^{(T_A - T_A^{ref})/10}, \tag{2}$$

where Q_{10} is the parameter describing temperature sensitivity, X is a set of meteorological drivers and R_b describes the base respiration. Including air temperature T_A as a driver of R_b is an optional choice if we are to believe that there are effects of temperature beyond the exponential dependency through Q_{10}. A common hybrid modeling approach amounts to using a NN as an estimator for R_b, treating Q_{10} as a trainable parameter, and fitting everything end-to-end with gradient descent, as it has been done in [45].
Equifinality in this problem can be shown by reformulating (2) for $c > 0$:

$$R_{eco}(X, T_A) = R_b(X, T_A) c^{(T_A - T_A^{ref})/10} \cdot \left(\frac{Q_{10}}{c}\right)^{(T_A - T_A^{ref})/10}. \tag{3}$$

Thus, a flexible enough function estimator (e.g. a NN) could learn $R_b(X, T_A) c^{(T_A - T_A^{ref})/10}$ and obtain $\frac{Q_{10}}{c}$ as the temperature sensitivity. In this case, we would obtain one of the solutions by chance and thus reach erroneous conclusions about the temperature sensitivity.
In this example, equifinality arises because the problem is mathematically ill-posed. It is less obvious, however, when introducing several non-parametric models in more complicated physical equations. In practice, we will obtain a distribution over the parameters mainly driven by inductive biases of the learning algorithm or the network architecture [67] and which are not guided by any physical knowledge. Additional explicit information can alleviate this problem. These include the introduction of additional losses or adding prior knowledge [68, 69]. Similarly, a regularization term can make the problem identifiable. This has been formally proven for solving hybrid ODEs [70]. Regularization, however, is known to introduce bias on parameters of interest in semi-parametric modeling problems [56].

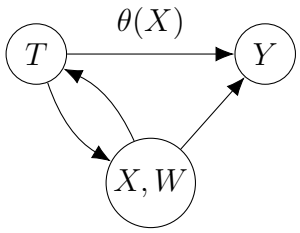


Figure 1: Causal graph of treatment effect estimation of T on Y. Sets X and W can enter both as confounders and mediators. Treatment effect θ can be heterogeneous and dependent on X or constant.

2. Double machine learning for hybrid modeling – a causal perspective

Our setting considers problems that can be expressed as in (1), which can be studied under a causal perspective, see Fig. 1. The parameter θ describes the direct effect of some treatment variable T on the outcome variable Y. Moreover, we have access to sets of predictors X and W that are confounding or mediating the effect of T on Y. Confounders are common causes of T and Y, while mediators are variables through which T indirectly affects Y. The inclusion of mediators has important implications for the interpretation of the results. When we estimate the effect of T on Y with mediators, we only obtain the direct effect by discounting the effects through these mediators. The variables in X can further enter as effect modifiers by modulating the effect θ of T on Y. Technically, we can use all mediators and confounders as effect modifiers when we include them all in X, leaving W empty, or treat θ as a constant effect by instead leaving X empty. At this point, we need to be careful with the choices of control variables X and W as we need to assume that all relevant confounders are observed and included. In particular, this means we need to be careful not to include mediators that have an unobserved common cause with Y or that we introduce a common effect of T and Y. Both cases would open a new path and substantially bias the estimation [71].

As per the DML framework, we must define an auxiliary equation that models the confounding and mediating effects of X and W on T. Assuming, without loss of generality, centered noise for both equations, we obtain

$$Y = \theta(X) \cdot f(T) + g(X, W) + \epsilon \qquad \mathbb{E}[\epsilon|X, W] = 0 \tag{4}$$

$$f(T) = m(X, W) + \eta \qquad \mathbb{E}[\eta|X, W] = 0 \tag{5}$$

$$\mathbb{E}[\eta \cdot \epsilon|X, W] = 0. \tag{6}$$

Sometimes, the original problem formulation must be manipulated to fit our setting. We will see examples of given transformations f, though the identity $f(T) = T$ could also be used when the relationship is assumed linear in T. The *causal effect* θ is modeled either as a constant coefficient or as a function of some covariates (heterogeneous effect).

We proceed according to the *partialling out* method in the DML framework [56]:

(i) Fit an estimator $\mathbb{E}[Y|X, W]$ of Y on X and W,

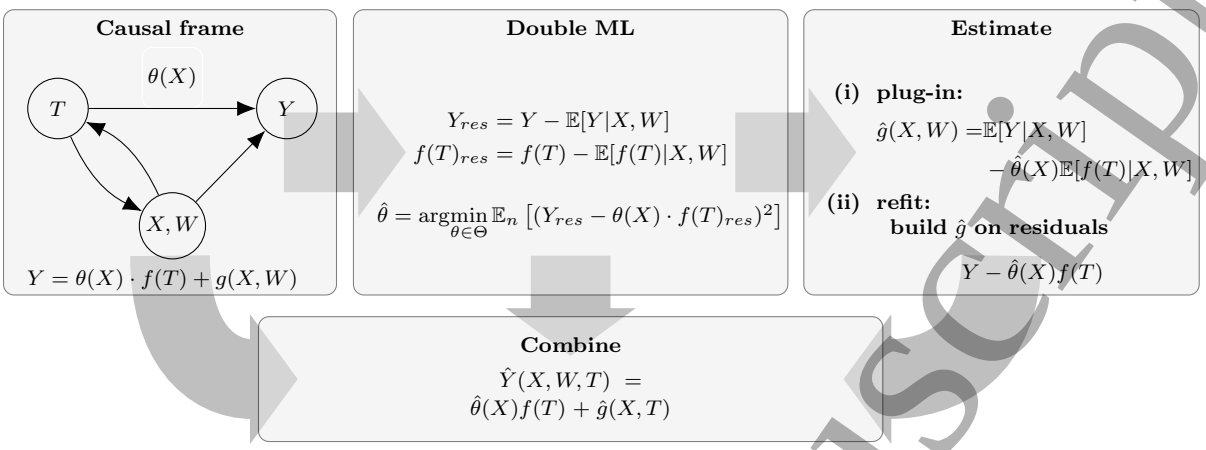


Figure 2: Schema of the proposed approach: (i) **Frame** the problem as a treatment effect estimation problem and assume causal graph. (ii) Build estimators of Y and $f(T)$ and deploy **DML** in the constant or heterogeneous treatment effect setting. (iii) **Estimate** g with plug-in estimator or via a final fitting on the residuals. And finally, (iv) **Combine** $\hat{\theta}$ and $\hat{g}$ into a causally interpretable hybrid model.

(ii) fit an estimator $\mathbb{E}[f(T)|X, W]$ of $f(T)$ on X and W,

(iii) compute their residuals as $Y_{res} = Y - \mathbb{E}[Y|X, W]$ and $f(T)_{res} = f(T) - \mathbb{E}[f(T)|X, W]$ and

(iv) estimate $\hat{\theta} = \arg\min_{\theta \in \Theta} \mathbb{E}_n \left[(Y_{res} - \theta(X) \cdot f(T)_{res})^2 \right]$.

We call the estimators in (i) and (ii) the first-stage estimators. The primary benefit of the DML framework is that it yields fast estimation rates and, under certain assumptions, asymptotic normality of θ. It is robust to errors in the first-stage estimators due to overfitting or regularization bias. This robustness stems from the observation that the moment equations corresponding to the final least squares loss in (iv) fulfill Neyman orthogonality with respect to the first-stage estimators [56]: The gradient with respect to the non-parametric estimators is zero in the optimum. This implies that small deviations away from the optimal non-parametric models still keep the true θ_0 as the optimal parameter of the score. This approach has been analyzed for a large set of model classes [56, 72–75]. For example, any combination of linear regression, decision trees, support vector machines, or NNs can be used to model the treatment and/or the outcome models. Similarly, any of these or a combination of models could be chosen to estimate the treatment effect. To maintain the theoretical guarantees of the DML framework, it is important to split the data and perform the first two fitting steps ((i),(ii)) on a different data subset than the last fitting step for the residuals (iv). By doing cross-fitting, data efficiency can be maintained.

If the only object of the analysis is the interpretable treatment effect θ, the task is completed by the above DML procedure. Nevertheless, as is usually the case in hybrid

modeling tasks, we are probably also interested in obtaining an estimator of g. For this, we have two options:

(i) Use $\hat{g}(X, W) = \mathbb{E}[Y|X, W] - \hat{\theta}(X) \cdot \mathbb{E}[f(T)|X, W]$ (plug-in) or

(ii) build an estimator on the residuals $Y - \hat{\theta}(X) \cdot f(T)$ (refit).

The plug-in estimator (i) uses all estimators fitted in the previous steps and can be obtained at no additional computational cost. A derivation of this estimator is given in Section Appendix A.1. On the downside, in contrast to θ, there are no theoretical guarantees on how well it describes g. Option (ii) adds a final supervised learning step, with the advantage being that we are not limited to using the X and W to estimate θ. Once θ has been estimated in a well-posed setting, we can now introduce, for example, T as a driver in the estimation of g. We can combine all estimators to obtain the fitted hybrid model for Eq. (1) (see Fig. 2 for a summary of the proposed procedure). By separating the problem into a causal inference and a standard supervised learning step, we have maintained its well-posedness. Next, we will explain how this technique can be effectively applied in two use cases around carbon fluxes.

3. Case studies

Carbon fluxes are crucial in the global carbon cycle, a key component of the Earth's climate system [76]. Net ecosystem exchange NEE is the net carbon dioxide flux measured using the eddy covariance (EC) technique [77]. The data for our studies is half-hourly data from FLUXNET, a global network of EC towers that collect data on carbon dioxide, energy fluxes, sensible heat fluxes, and water vapor exchange between the atmosphere and the terrestrial biosphere [78]. It offers comprehensive measurements of meteorological parameters and constitutes a crucial data source for ecosystem modeling and climate research.

Different biogeochemical processes contribute to the carbon balance of the land [79]. In particular and as common, we split NEE as

$$NEE = -GPP + R_{\text{eco}}, \tag{7}$$

where gross primary production GPP describes the gross carbon uptake by the environment and ecosystem respiration R_{eco} denotes the carbon release of all organisms.

3.1. The Q_{10} model

A common parametrization of R_{eco} is the Q_{10} respiration model [60–62]:

$$R_{eco}(X, T_A) = R_b(X) \cdot Q_{10}^{(T_A - T_A^{ref})/10}. \tag{8}$$

This model highlights temperature T_A as a principle driver of respiration, with Q_{10} denoting the temperature sensitivity parameter. Furthermore, R_b describes the base respiration, and X a set of meteorological drivers. Following the example of [45], we

use data from the EC tower in Neustift, Austria, available in the FLUXNET2015 dataset [80]. Based on this site, we extensively probe the DML-based HM in the controlled setting of synthetic data and showcase its potential on measured data. As the goal of this paper is not to provide a comprehensive analysis of global Q_{10} values, we limit ourselves to this site for our first use case.

Data Synthetic data is generated from a Q_{10} model with seasonally varying base respiration and measured air temperature T_A, and with true constant Q_{10} set to 1.5 (for details, see Section Appendix B.1.1). We provide additional experiments for Q_{10} values of 1.25 and 1.75 to showcase the robustness of the results.

Ecosystem respiration is a latent flux not directly observed at flux towers during the day. It can only be measured as nighttime NEE, as without photosynthesis, we assume GPP to be zero or under controlled conditions like a sealed chamber [79]. We use 2003 to 2007 for training and keep 2008 and 2009 for testing. Moreover, we consider only measured observations, which amount to approximately 10% of the nighttime data for training (4331 data points).

Applying DML-based HM Applying a log-transform to (8) and setting $f(T_A) = (T_A - T_A^{ref})/10$ yields

$$\log(R_{eco}(X, T_A)) = \log(R_b(X)) + f(T_A) \cdot \log(Q_{10}). \tag{9}$$

The resulting equation (9) describes a partially linear regression problem [81] equivalent to (1). Here, $\log(R_b(\cdot))$ represents the non-parametric function $g(\cdot)$ as we do not know the functional form of R_b. We aim to estimate the constant linear effect $\theta = \log(Q_{10})$ of the transformed temperature $f(T_A)$ on the log-transformed ecosystem respiration. In this work, we employ and compare both NNs and RFs as examples for first-stage estimators.

After obtaining the estimator $\hat{Q}_{10}$, we fit a NN on

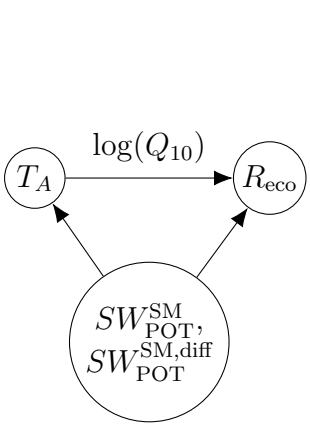
$$\frac{R_{eco}(X, T_A)}{\hat{Q}_{10}^{f(T_A)}} = NN(X, T_A). \tag{10}$$

We compare the *causal DML-based HM* to the *standard GD-based HM* as described in [45]. We fit

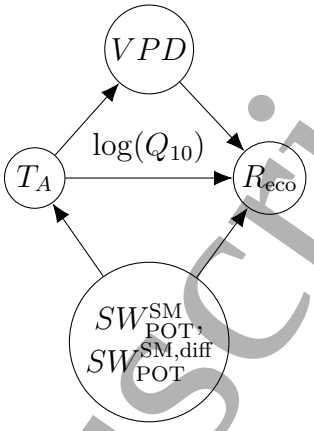
$$R_{eco}(X, T_A) = NN(X) \cdot Q_{10}^{(T_A - T_A^{ref})/10}, \tag{11}$$

with a NN representing the base respiration R_b. The weights of the NN are optimized together with Q_{10} using the Adam [82] optimizer.

We run the experiments with and without regularization for all involved NNs in both hybrid modeling approaches. For this, we use dropout at a rate of 0.2. This technique randomly drops nodes in a NN during training and was found to have a sparsifying effect on the model [83]. We apply dropout to all hidden units in the network. We



(a) Synthetic data.



(b) Real data.

Figure 3: Assumed causal graphs for the estimation with the causal hybrid modeling approach in Q_{10} estimation.

provide additional experiments with weight decay [84], another common regularization technique in deep learning at a rate of 0.1. To showcase the effect of equifinality, we also introduce T_A as an additional predictor in R_b. We will apply the same training procedure and NN architectures for both hybrid modeling approaches for comparability and to show robustness in the presence of biased estimators. We only drop the final nonlinearity for the first-stage estimators in the DML-based HM. Details on the NNs and their training can be found in Section Appendix B.3.

Causal graph of the Q_{10} model The causal graph we assume for the Q_{10} model is shown in Fig. 3. The smooth potential radiation cycle given by $SW^{\text{SM}}_{\text{POT}}$ and $SW^{\text{SM,diff}}_{\text{POT}}$ represent seasonality and, thus, has a confounding effect on temperature T_A and R_{eco}. For the real data, we add VPD to the graph, representing humidity and water availability. This variable enters as a mediator in the graph as temperature affects evaporation and how much water the air can hold [85]. Furthermore, water availability also has a strong effect on respiration [86]. However, the temperature-sensitivity Q_{10} should only describe the immediate temperature effect [85]. We model the effects of water in the base respiration factor R_b. Thus, assuming this graph, with our choices of variables, we estimate only the direct, immediate effect and not the one mediated through water or confounded by seasonality.

3.2. CO_2 Flux partitioning

3.2.1. Problem formulation Direct measurements of GPP or R_{eco} at the ecosystem level are difficult to obtain [79]. Alternatively, partitioning methods estimate these fluxes numerically from the measured NEE. Common approaches implement functional relationships based on physiology and estimate the fluxes using data-driven models [87–91]. Several hybrid-modeling approaches have recently been proposed modeling both

fluxes with NNs [38, 68, 92].

Separating a single signal into two additive signals is generally prone to equifinality issues. [38] tried to break the symmetry between fluxes in the partition by enforcing different sets of explanatory environmental covariates for the two fluxes and applying a simple hybrid model. In particular, the authors combined NNs with the light-use-efficiency model given by

$$NEE = -LUE \cdot SW + R_{\text{eco}}, \tag{12}$$

where LUE models the linear efficiency of the incoming shortwaves SW on the resulting GPP. In this form, GPP was modeled as the product of the incoming radiation and LUE parametrized by a NN. [68] showed that with different random initializations, this approach can lead to different resulting fluxes. The equifinality of the solution becomes particularly evident in extreme conditions. The authors can reduce variability through a multi-task learning approach. They introduce a second loss, forcing the network to learn to predict solar-induced chlorophyll fluorescence (SIF) from the separated GPP as both signals are known to be correlated under normal conditions.

Data As a proof of concept, we evaluate the proposed method on synthetically generated data (see Section Appendix C.4). We only used measured NEE for the real data and applied the hybrid modeling approach site-wise per year. For the data selection of real data from FLUXNET2015 [80], we closely followed [38] to compare our method to the neural network approach that imposes similar structural equations. We chose the same set of 36 different FLUXNET2015 sites (see Section Appendix B.2) and used the same quality criterion to select site-years, i.e., years of a specific site. This implies that fitting is done year-wise per site, and only measured data is used. To have enough high-quality data, only site-years for the analysis are selected where at least 80% of the meteorological data and 10% of each daytime and nighttime NEE were measured. As a target, similar to [38], we use the NEE obtained from the 50th percentile of the CUT method [80]. For comparison, we use the respective partitioned R_{eco} and GPP fluxes obtained from the daytime [90] and nighttime [87] methods, already provided as part of the FLUXNET2015 dataset. Moreover, we compare the partitions to the results obtained with NNs from [38].

Applying DML-based HM We want to fit the following flux partitioning equation

$$NEE = -LUE(X) \cdot f(SW) + R_{\text{eco}}(X, W), \tag{13}$$

where X and W are sets of meteorological drivers and f transforms the incoming radiation to allow for more flexible light-response curves, leading to a potentially non-linear light-use efficiency model. Here, $R_{\text{eco}}(\cdot)$ and $LUE(\cdot)$ represent $g(\cdot)$ and $\theta(\cdot)$ in the equivalent problem (1), respectively. This time, we use the estimator of R_{eco} obtained from the first-stage estimators. As a proof of concept, we apply this method with f

being the identity function for linearly generated data over different noise levels (see Section Appendix C.4).

For real data, the assumption of a linear relationship to SW is violated as GPP saturates with increasing light. We will thus first fit a transformation f of the light curve before applying the DML schema. In order to find f, we finally fit α and β in

$$NEE = -\frac{\alpha\beta\ SW}{\alpha SW + \beta} + \gamma. \tag{14}$$

with a moving window of 15 days, we always transform the 5 days in the center of the fitting interval. This procedure is motivated by the daytime flux partitioning method [90], which estimates a parameterized rectangular hyperbola over moving windows to obtain GPP. This heuristic allows us to find a flexible, smoothly changing light response curve. Other ways to obtain such a transformation can be envisaged. For the synthetic data, we use inputs according to how the data was generated, i.e., vapor pressure deficit VPD and temperature T_A for X and the seasonal cycle of potential radiation for W. On the real data, we use day of the year *doy*, VPD, temperature T_A, and soil water content SWC (for the sites where it is available) for X and leave W empty (For the assumed causal graphs, see Section 3.2.1). We use gradient boosting regressors [93], an ensemble method of multiple shallow decision trees for all involved fitting steps.

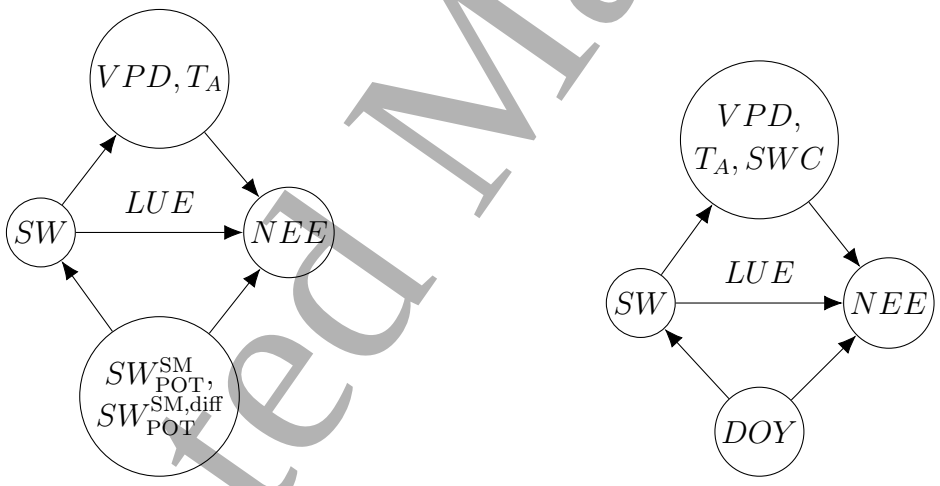
Causal graph of the LUE model The causal graphs assumed for the LUE model are shown in Fig. 4. As R_{eco} is modeled similarly to the Q_{10} model, we keep the same variables modeling the seasonal cycle. In addition to that, we include VPD and T_A, which were used to model GPP. The incoming radiation SW has an effect on the temperature as well as on water vapor [85]. Thus, both variables enter as mediators on the path to NEE. For the real data, we use the day of the year DOY to model the seasonality, which continues to be a confounder. In addition to the VPD and T_A, we add soil water content, which also enters as a mediator when available. Consequently, we estimate GPP as the direct effect of light on NEE, discounting the indirect effects through temperature, VPD, and SW, which we allocate to $RECO$. Note that in this setup, these three variables are still entered as modifiers on the effect of light on NEE, affecting GPP. Table 1 summarizes the variables used for the different setups of the use cases.

4. Results and Discussion

We show the applicability of our causal DML-based HM on two carbon flux modeling problems. We estimate the temperature sensitivity parameter in the Q_{10} model to showcase the robustness to regularization biases. We further illustrate the flexibility of the method to tackle the carbon flux partitioning problem.

Table 1: Summary of variables for the experiments. The variables denote: outcome variable Y, treatment T, control variables X and W, ecosystem Respiration R_{eco}, air temperature T_A, smooth cycle of shortwave radiation and its derivative $SW_{\text{POT}}^{\text{SM}}$ and $SW_{\text{POT}}^{\text{SM,diff}}$, vapor pressure deficit VPD, net ecosystem exchange NEE, day of the year DOY and soil water content SWC.

Use case	Data	Y	T	W	X
Q_{10} model	Synthetic	$\log(R_{\text{eco}})$	T_A	$SW_{\text{POT}}^{\text{SM}}$, $SW_{\text{POT}}^{\text{SM,diff}}$	-
	Measured	$\log(R_{\text{eco}})$	T_A	$SW_{\text{POT}}^{\text{SM}}$, $SW_{\text{POT}}^{\text{SM,diff}}$, VPD	-
CO_2 Flux partitioning	Synthetic	NEE	SW	$SW_{\text{POT}}^{\text{SM}}$, $SW_{\text{POT}}^{\text{SM,diff}}$	VPD, T_A
	Measured	NEE	SW	DOY	VPD, T_A, SWC

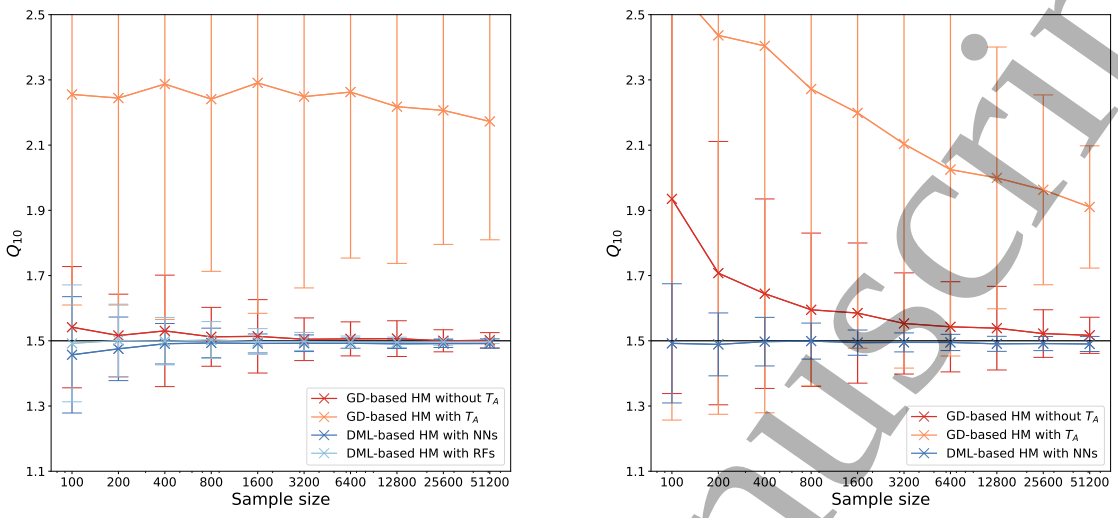


(a) Synthetic data. **(b)** Real data.

Figure 4: Assumed causal graphs for the estimation with the causal hybrid modeling approach in flux partitioning.

4.1. Q_{10} ecosystem respiration model.

4.1.1. Overall improved estimation capabilities. We simulated ecosystem respiration data from observations of FLUXNET. The true Q_{10} parameter was set to 1.5. We sample 100 datasets of varying sample sizes to see how the methods perform in different data regimes. We compare the GD-based HM approach using NNs to the proposed causal DML-based HM framework in two possible instantiations, either using RFs or NNs as first-stage estimators. Experiments are run with and without applying dropout regularization and introducing T_A as an additional predictor in base respiration.



(a) Without dropout. **(b)** With dropout.

Figure 5: Simulation study for Q_{10} estimation with the GD-based HM and the DML-based HM over 100 sampled datasets at different sample sizes. The plots show average and 95% CI for the estimated Q_{10} for different methods without (a) and with (b) dropout applied as a regularizer in the NN regression models. The true Q_{10} parameter has a value of 1.5. Introducing T_A as a predictor in R_b leads to equifinality problems. Dropout as a regularizer introduces bias on the estimation of Q_{10} in the GD-based HM case, while the causal hybrid modeling approach performs satisfactorily in the absence of equifinality.

The Q_{10} estimation results are shown in Fig. 5. First, Fig. 5a shows the results where no dropout was applied to the NNs. In this case, the estimates of the GD-based HM approach, where T_A is included as a predictor for R_b, show values that are, on average, between 2.1 and 2.3 over all sample sizes. They show a substantial mismatch to the true value of 1.5 and a wide spread at each sample size. This illustrates that equifinality expresses itself in the estimations as a wide range of values that hardly decreases with increasing sample size. We are not obtaining the full range of $\mathbb{R} > 0$ values, which is by (8) mathematically possible, but a range that is constraint alone by the initial Q_{10} value, the network's implicit biases and the first optimization steps of the gradient descent algorithm. This can make us mistake this for a valid inference of the method. Instead, methods that exclude T_A as a predictor find good estimators that converge with increasing data size. This is, in general, an encouraging result for all hybrid modeling approaches in this setup. Over the whole range, the GD-based HM shows wider spreads than the DML-based HM approaches, which converge notably faster with increasing data size. At low data, they also have lower bias than the GD-based HM approach. Remarkably, the random forest shows very little bias for solving this task over

the whole data regime. Experiments corresponding to Q_{10} values of 1.25 and 1.75 (see Section Appendix C.3) exhibit minor variations in magnitude, proportional to the effect parameter. However, they consistently affirm the findings obtained for $Q_{10} = 1.5$.

These results showcase the data efficiency of the DML-based approach. At the same time, it is currently computationally less efficient. The causal DML-based HM involves various fitting steps, which may seem uncomfortable compared to the usual end-to-end learning with NNs. One may think of ways also to make DML end-to-end possible. Here, one would apply NNs for all fitting steps and introduce a common loss over all optimization problems optimized with gradient descent. By weighting these losses adaptively, one can force this training to first fit the first stage estimators and then the treatment effect variable similar to what has been done in fitting PINNs respecting temporal and spatial causality [52]. Efforts would need to be put into parallelizing the fitting of the first-stage estimators to make this approach computationally less costly.

4.1.2. Robustness against regularization bias. Dropout is commonly used in deep learning for regularization [83] or uncertainty quantification [94]. Fig. 5b shows the Q_{10} estimations where dropout is applied to all NNs of the GD-based HM approach and the HM approach based on DML. With dropout, the GD-based HM approach has a more challenging time finding a good solution. It substantially overestimates the value of Q_{10} in the low data regime and only slowly gets more constrained and closer to the true value at the upper end of the used sample sizes. While the GD-based method got notably worse with the introduction of dropout, the DML shows robust results for the estimations over the full data range. On average, the Q_{10} estimations perform similarly to the experiments without dropout. In the low data regime, the bias in the estimation even decreased further. When fitting the GD-based HM with T_A, the regularization with dropout has a positive effect. The estimated values for Q_{10} are closer to the true value, and the spread reduces with more data points. The regularization through dropout restricts the space of solutions and reduces equifinality even though more data is necessary to overcome the stochasticity introduced through dropout. In Section Appendix C.1, we show additional results with weight decay [84], another common regularization technique. As it yields qualitatively similar results (see Fig. C1), we conclude that the presented findings are not only inherent to dropout. In Section Appendix C.2, we further test the robustness of the findings with $0.05, 0.1$, and 0.3 as additional dropout rates and find that the introduced bias in estimating Q_{10} is proportional to the magnitude of dropout. In all cases, the approach based on double machine learning remains advantageous.

In light of the results, DML, in combination with dropout, can be effectively used for a full probabilistic assessment of hybrid models with inference on the parameter of interest and the non-parametric part, as dropout is also a common technique for obtaining uncertainty estimates for NNs [94]. While the GD-based HM approach suffered from the application of dropout, the DML approach was robust. Moreover, the technique further yields confidence bands for the approximately normally distributed

estimators. By separating both estimations, we can obtain a distribution over the estimated Q_{10} and safely obtain uncertainty estimates for R_b using dropout.

4.1.3. Results on real data As discussed in Section 3.1, we obtain measured respiration data using nighttime NEE measurements. We apply GD-based HM and DML-based HM with NNs and RFs without dropout to the data. We used the full dataset of over 100 different random seeds. The obtained distributions of Q_{10} are shown in Fig. 6. The GD-based HM approach finds a mean value of 1.322, with a skewed distribution and estimated values ranging between 1 and 2. Including T_A as a predictor in the GD-based approach, the values lie in a completely different range between 2.5 and 3.5, with the mean being 2.816. The estimations based on DML yield a mean of 1.407 and 1.409 for the RFs and NNs, respectively, with similarly peaked distributions. The results of the DML estimate agree fairly well with the results of [95] that after controlling for seasonal confounding, find that Q_{10} takes values around 1.41 ± 0.1 independently of mean-annual temperature and biome.

4.2. CO_2 flux partitioning

We apply the causal DML-based HM to the problem of carbon flux partitioning as defined in (7). In this scenario, we model the effect as a heterogeneous treatment effect, a function of other predictors, parametrized with an ML model. We use gradient boosting estimators for all three estimators involved. Moreover, we show that the plug-in estimator for R_{eco} obtained by combining the first-stage estimators yields useful values without the need for an additional refit.

4.2.1. Consistent flux partitioning We use vapor pressure deficit VPD, air temperature T_A, and day of the year (for seasonality) as drivers over all sites. Where available, we also included soil water content. Since we do not have access to the real partial fluxes, we compare the retrieved fluxes to the ones obtained by the NN approach described in [38] and by the established daytime and nighttime methods [87, 90]. The daytime and nighttime methods are assumed to capture a simple cycle depending on a few meteorological drivers. New methods may deviate but should show a similar pattern overall. For the partitioned fluxes of two methods $(x_i)_{i=1}^N$ and $(y_i)_{i=1}^N$, we compute the R^2, the root-mean-square error (RMSE), given by $\sqrt{\frac{\sum_{i=1}^N (x_i - y_i)^2}{N}}$, and the bias as the difference between the sample means $\bar{x}$ and $\bar{y}$. The results are reported in Table 2.

Overall the consistency of the method based on DML lies in a similar range of values to the NN approach [38] when compared to the daytime and nighttime methods. The estimated data uncertainty of the used NEE measurements is $1.53 \frac{\mu\,\text{mol}\,\text{CO}_2}{m^2 s}$. For almost all compared fluxes, our method lies under this threshold in terms of RMSE. Only for the GPP and NEE of the nighttime method, the values lie on average slightly above with $1.97 \frac{\mu\,\text{mol}\,\text{CO}_2}{m^2 s}$ and $1.92 \frac{\mu\,\text{mol}\,\text{CO}_2}{m^2 s}$, respectively. The nighttime method fits respiration overnight and obtains GPP as the residuals between the estimated R_{eco} and measured

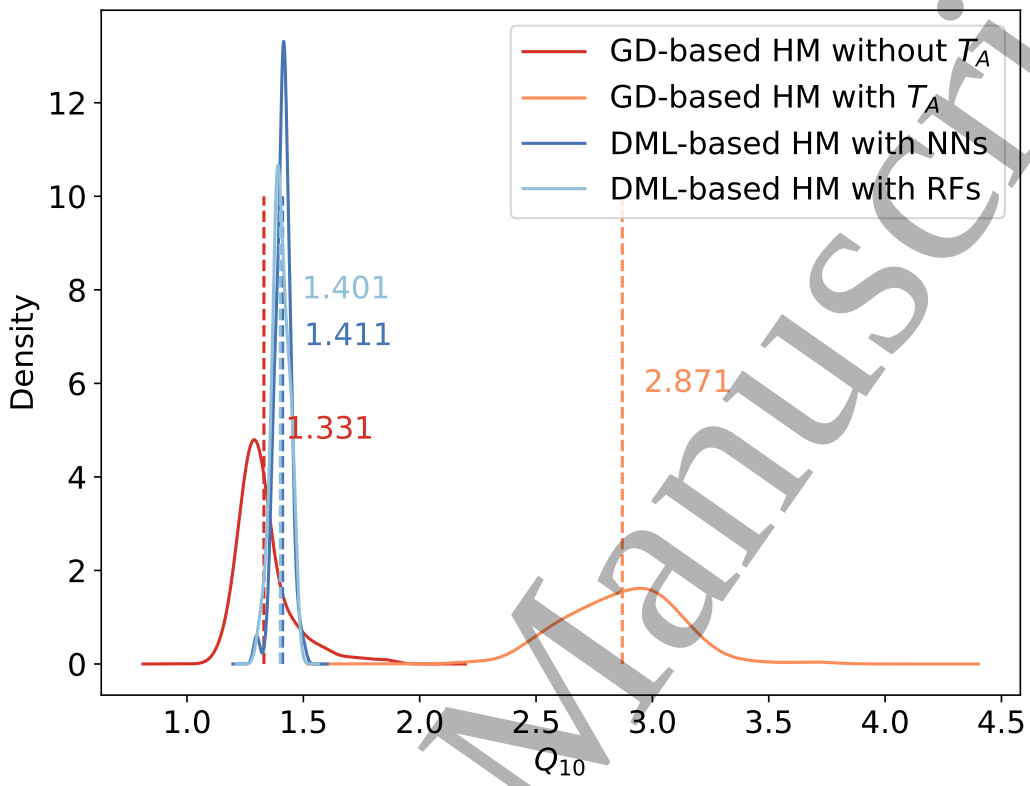


Figure 6: Estimation of Q_{10} on real data. Both DML-based HM find on average a Q_{10} value of 1.401 and 1.411 for RFs and NNs, respectively. This agrees with values from the literature that find a Q_{10} value around 1.41 ± 0.1 [95]. The value for the GD-based HM is lower at 1.331 when leaving out T_A as a predictor. With T_A, problems of equifinality show up again.

NEE. Thus, by construction, the NEE of the nighttime method corresponds to the measured NEE. Hence, both NEE and GPP of the nighttime method are higher in noise, and thus, a higher RMSE of our method is expected. When comparing the bias between methods, the causal DML-based HM shows a slightly smaller bias compared to both standard methods than these methods between them in almost all cases. Furthermore, it lies in a similar range to the GD-based HM.

Overall, our method shows higher similarity to the daytime method, which is expected due to the fitting of the rectangular hyperbola in the first step. The retrieved GPP is similar to the daytime method as the NN approach, and the obtained NEE is even closer. At the same time, the obtained R_{eco} shows a larger deviation even to the daytime method. This is because we used the plugin-in estimator for R_{eco} obtained from the first-stage DML estimators.

We could obtain a more sophisticated estimator by refitting another model on the residuals, as done in the case of the Q_{10} model, where we could also employ

Table 2: Cross consistency in terms of R^2, $RMSE$ and bias of retrieved GPP, $RECO$ and estimated NEE between the established daytime (DT) [90] and nighttime (NT) [87] methods and the GD-based HM with neural networks (NN) [38] and DML-based HM (DML), proposed in this work. The reported statistics are median and in brackets 0.25/0.75 quantiles over all site-years.

Flux	Methods	R^{2*}	$RMSE^*(\frac{\mu\,\mathrm{mol\,CO_2}}{m^2 s})$	$Bias(\frac{\mu\,\mathrm{mol\,CO_2}}{m^2 s})$
RECO	DT vs. DML	0.62(0.41/0.74)	1.18(0.75/1.46)	0.00(−0.20/0.14)
	DT vs. NN	0.69(0.50/0.81)	0.98(0.70/1.29)	0.02(−0.12/0.18)
	NT vs. DML	0.74(0.50/0.83)	0.89(0.57/1.15)	0.00(−0.11/0.10)
	NT vs. NN	0.85(0.65/0.92)	0.68(0.47/0.84)	0.07(−0.02/0.16)
	DT vs. NT	0.73(0.63/0.83)	0.95(0.64/1.21)	0.00(−0.22/0.16)
	NN vs. DML	0.63(0.34/0.77)	0.99(0.66/1.24)	−0.07(−0.22/0.10)
GPP	DT vs. DML	0.96(0.93/0.97)	1.25(0.74/1.49)	0.00(−0.16/0.11)
	DT vs. NN	0.96(0.93/0.97)	1.22(0.76/1.52)	0.04(−0.04/0.17)
	NT vs. DML	0.90(0.84/0.92)	1.97(1.16/2.47)	−0.02(−0.13/0.10)
	NT vs. NN	0.93(0.89/0.95)	1.53(0.90/2.02)	0.07(−0.02/0.18)
	DT vs. NT	0.89(0.82/0.92)	1.85(1.20/2.42)	0.02(−0.16/0.13)
	NN vs. DML	0.95(0.92/0.97)	1.32(0.71/1.61)	−0.08(−0.23/0.08)
NEE	DT vs. DML	0.95(0.93/0.97)	1.07(0.71/1.29)	−0.02(−0.11/0.07)
	DT vs. NN	0.94(0.91/0.96)	1.13(0.76/1.36)	−0.03(−0.12/0.03)
	NT* vs. DML	0.87(0.81/0.89)	1.92(1.15/2.36)	0.01(−0.02/0.06)
	NT* vs. NN	0.93(0.90/0.94)	1.29(0.79/1.82)	0.00(−0.01/0.01)
	DT vs. NT*	0.86(0.79/0.90)	1.68(1.12/2.25)	−0.03(−0.12/0.03)
	NN vs. DML	0.94(0.91/0.96)	1.27(0.77/1.52)	0.01(−0.02/0.05)

*The NT NEE value corresponds exactly to the measured NEE value.

SW as a predictor without experiencing equifinality. It would even allow using the previously estimated GPP as a predictor of R_{eco}. As an additional proof of concept, we apply the method to synthetic data with different levels of heteroscedastic noise. The method finds robust estimates even to high levels of noise. The results can be found in Section Appendix C.4.

4.2.2. Learned functionalities The consistency tables served as a sanity check that the methods produce reasonable estimations that contain similar trends over the day and year. The next questions are: Where do they produce similar outputs? When do the outputs differ? For this, we compare the retrieved fluxes on two different sites. In Fig. 7, we see the retrieved GPP flux over a few days in July 2006 in France Le Bray. We compare the DML-based HM to the GD-based HM, daytime and nighttime methods. The retrieved GPP of the daytime and hybrid modeling methods show similar

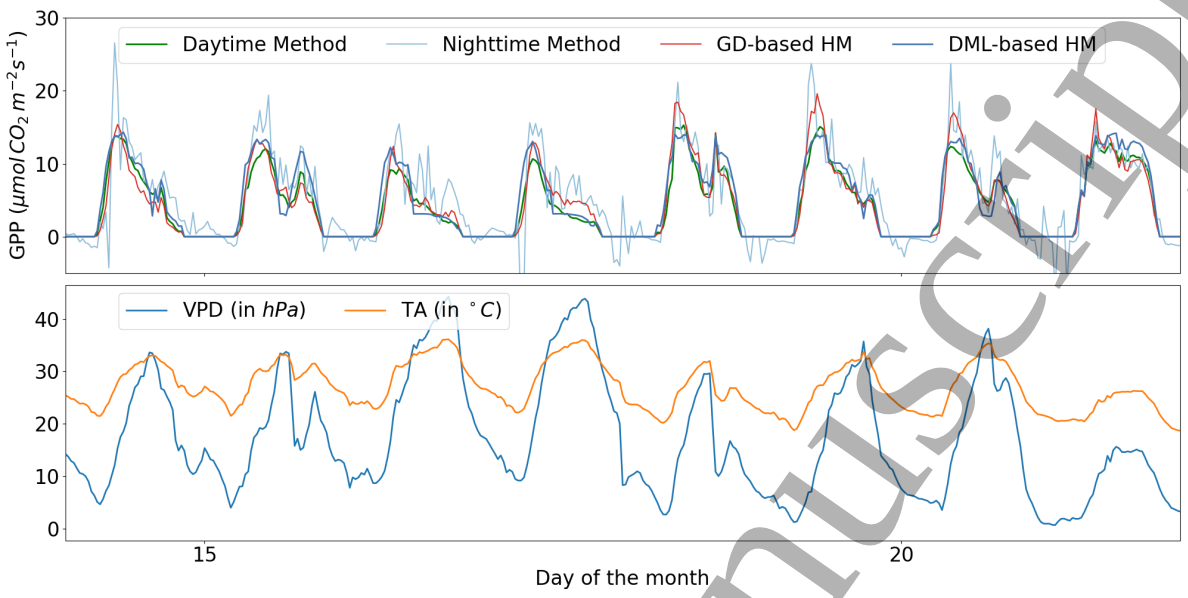


Figure 7: Retrieved GPP flux of daytime method, nighttime method and DML-based HM in July 2006 in France Le-Bray. The DML-based HM retrieved a similar flux to the daytime method that decreases with the increase of VPD.

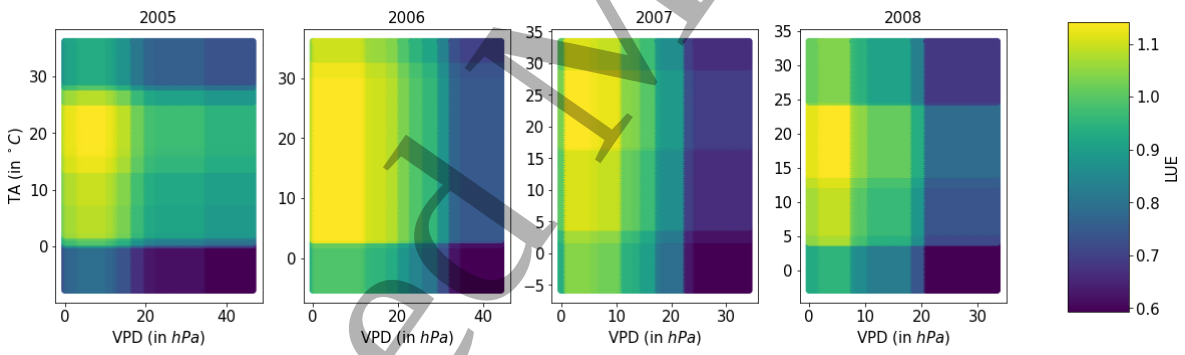


Figure 8: Functional behavior of the learned LUE in the years 2005 to 2008 over VPD and TA. The LUE shows a consistent functionality over the different years where an increase in VPD, which marks lower water availability, reduces productivity. This is also consistent with the functionality that the daytime method implements parametrically.

patterns. High VPD, which marks low water availability, reduces productivity. The daytime method implements this functionality parametrically. The LUE function of the DML-based HM approach learned a similar functionality that decreases with increasing VPD and has preferred temperatures roughly between $15°C$ and $30°C$ (see Fig. 8). It is consistent over the four consecutive years the method was applied to at this site. This demonstrates that the causal hybrid modeling approach can learn a similar functional relationship as the parametric daytime method in a non-parametric way. The nighttime method shows a noisier but qualitatively similar pattern.

To highlight the differences between the methods, we look at a grassland site in

Santa Rita (US) [96]. Fig. 9 shows the estimated R_{eco} over few days in July 2010. The selected time window was preceded by two months without rain, leading to low soil water content and, in turn, reduced respiration activity [86]. During the shown period, a rain event leads to a sudden increase in soil water content. Such an event is expected to lead to a sudden increase in respiration as it stimulates microbial activity [86]. We find that the daytime and nighttime methods cannot capture this sudden behavior as their estimation is based on window fitting and cannot detect sudden changes in dynamics. While R_{eco} estimated with the nighttime method increases even before the event, the daytime method yields slowly increasing respiration flux shortly after the event. Instead, the fluxes estimated with the non-parametric hybrid modeling approaches show an increase right at the event's time, demonstrating that they can adapt to sudden changes in dynamics. A difference between both hybrid modeling approaches shows that the GD-based HM estimates a stronger respiration pulse but yields a noisier estimate from the onset of the event.

Our approach offers unique advantages. While traditional daytime and nighttime methods are fully interpretable, they struggle to capture rapid dynamic changes due to their parametric nature. On the other hand, the end-to-end GD-based methods, such as the approach by [38], may lack interpretability due to non-identifiability or implicit functional constraints, relying on assumptions with unclear implications. In contrast, our causal interpretation-based approach offers a middle ground, providing reasonable estimates of fluxes while maintaining interpretability as it is grounded in causal assumptions. By identifying GPP as the causal effect of light on NEE, our method offers a clear and meaningful interpretation of the flux partitioning process. While it may not match the predictive performance and flexibility of pure deep learning, it offers a valuable alternative by combining interpretability with reasonable estimation accuracy.

The analysis we carried out merely serves as a proof of concept toward a causally meaningful flux partitioning method. To maintain comparability, we ran the experiments on the same sites and years with similar quality filters as [38]. For both DML-based HM as well as the GD-based HM approach with NNs, further research is necessary before they can be employed at scale in the data processing pipelines of FLUXNET sites. In particular, this would require a comprehensive analysis of the performance over all FLUXNET sites to disentangle the effects of geographical region, climate, vegetation, data quality, and data availability on the consistency of new flux partitioning methods. This should ideally be accompanied by simulations of sets of land surface models tailored for different land cover types to benchmark the adaptability of data-driven methods. This is beyond the scope of this work, which aims at introducing a causal approach to hybrid modeling. As for today, a benchmarking set and standardized evaluation pipeline are not available but could become key in the future when more data-driven flux partitioning models are developed. Understanding how these local factors influence the data-driven methods is crucial as the flux partitioning products serve as ground truth for downstream tasks such as upscaling from the site level to global fluxes

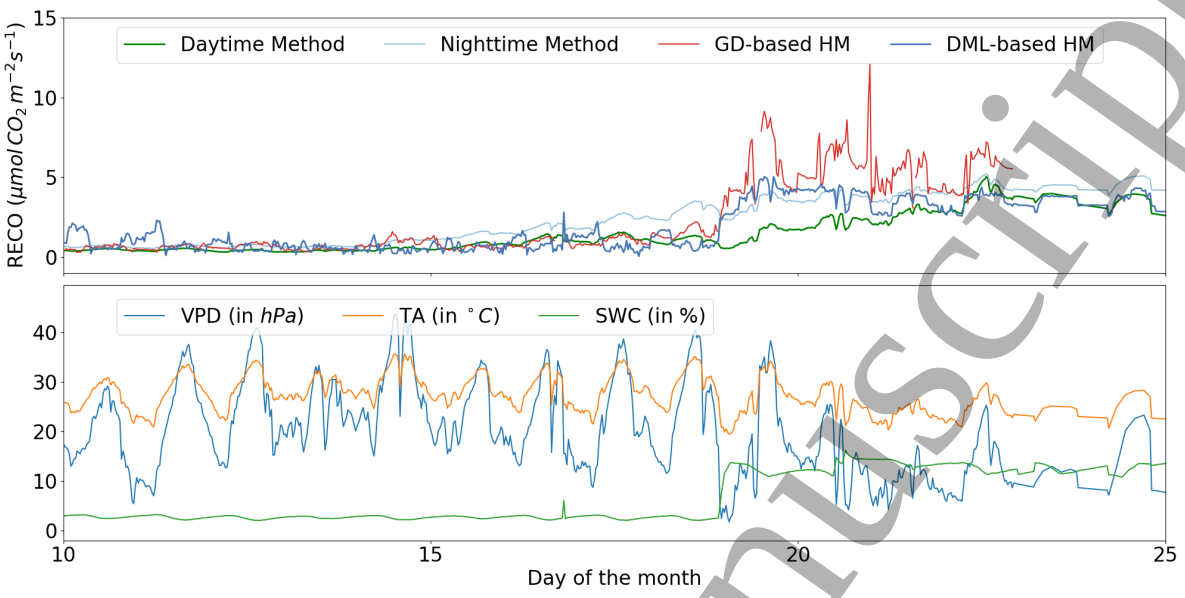


Figure 9: Retrieved R_{eco} flux of daytime, nighttime, and both hybrid modeling methods in July 2010 in Santa Rita in the US. The daytime and nighttime methods show slow adaption to the change in dynamics caused by a rain pulse event that followed a long drought. Both hybrid modeling approaches can retrieve the expected immediate increase in respiration. The estimates of the GD-based HM are lower and less noisy.

as aimed for in the FLUXCOM project [97].

5. Conclusions

Machine learning is becoming a complementary tool to enhance scientific research and discovery in all fields of science. Its limitations are evident: lack of transparency and interpretability, weak generalizability to unseen data, and violation of governing laws. Hybrid modeling aims to incorporate scientific knowledge to overcome these limitations. However, this alone is insufficient to obtain the interpretability we hope for. Spurious links between variables can lead to equifinality: many models describe the data similarly well. Therefore, we must also teach these hybrid models what seems evident to us: correlation is not causation. And it is causation that we want.

In this paper, we propose a first step in this direction. We split the fitting of hybrid modeling involving treatment effects into subsequent steps, where we first estimated the causal effect with DML and then estimated the remaining of the model. By separating different estimation steps and being explicit about the underlying causal graph and the causal effect, we were able to obtain a well-defined problem that, originally was ill-posed and, in practice, suffering from equifinality. We applied this technique to two problems of carbon flux estimation, namely, Q_{10} estimation in ecosystem respiration and carbon flux partitioning. We demonstrated the superiority of DML in retrieving parameters describing causal effects over end-to-end estimations with usual hybrid

modeling approaches using NNs. The estimation is shown to be efficient and robust and effectively reduces bias through regularization techniques such as dropout and weight decay. On real data, it could retrieve a value for Q_{10} consistent with the literature. We further showed the flexibility of the method by transforming the treatment and fitting a heterogeneous treatment effect of the LUE model for carbon flux partitioning as a non-parametric function. The retrieved fluxes were consistent with the ones of established methods, showed reasonable functional dependencies, and could improve on known limitations stemming from the window fitting of these methods.

We note that to apply the method effectively, assuming a causal graph and being explicit about the causal relationships of the involved variables is essential. This also includes thinking about unobserved confounders, mediators, and correlations between variables. We believe that this should be a general best practice. Our method encourages machine learners and practitioners to do so. A remaining problem is that even though we could show that it has broader applicability than the standard semi-linear regression problem, its relevance is still limited to hybrid models of a particular form containing parameters or non-parametric functions describing causal effects.

Integrating causality with hybrid modeling is crucial for achieving more interpretable and reliable outcomes in knowledge-driven machine learning. Our work has showcased this integration in two important problems in ecology through the application of causal effect estimation. Our causal hybrid modeling framework holds promise for enhancing interpretability and causal inference across diverse scientific fields that demand more insightful machine learning models. Looking ahead, we encourage further exploration and integration of causality concepts within hybrid modeling techniques.

Acknowledgments

This work was supported by the European Research Council (ERC) under the ERC Synergy Grant USMILE (grant agreement 855187). We thank Gianluca Tramontana for generously providing his data and patiently answering all our queries.

References

[1] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023.

[2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[3] Yu Zhang, James Qin, Daniel S. Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang, Quoc V. Le, and Yonghui Wu. Pushing the limits of semi-supervised learning for automatic speech recognition, 2022.

[4] Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, March 2009.

[5] Zachary C. Lipton. The mythos of model interpretability. *Queue*, 16(3):30:31–30:57, June 2018.

[6] Lee R. Kump, James F. Kasting, and Robert G. Crane. *The Earth System*. Pearson, 3rd edition, 2013.

[7] B. C. O'Neill, C. Tebaldi, D. P. van Vuuren, V. Eyring, P. Friedlingstein, G. Hurtt, R. Knutti, E. Kriegler, J.-F. Lamarque, J. Lowe, G. A. Meehl, R. Moss, K. Riahi, and B. M. Sanderson. The scenario model intercomparison project (ScenarioMIP) for CMIP6. *Geoscientific Model Development*, 9(9):3461–3482, 2016.

[8] V. Eyring, S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor. Overview of the coupled model intercomparison project phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958, 2016.

[9] Timothy A. Myers, Ryan C. Scott, Mark D. Zelinka, Stephen A. Klein, Joel R. Norris, and Peter M. Caldwell. Observational constraints on low cloud feedback reduce uncertainty of climate sensitivity. *Nature Climate Change*, 11(6):501–507, January 2021.

[10] Helene T. Hewitt, Malcolm Roberts, Pierre Mathiot, Arne Biastoch, Ed Blockley, Eric P. Chassignet, Baylor Fox-Kemper, Pat Hyder, David P. Marshall, Ekaterina Popova, Anne-Marie Treguier, Laure Zanna, Andrew Yool, Yongqiang Yu, Rebecca Beadling, Mike Bell, Till Kuhlbrodt, Thomas Arsouze, Alessio Bellucci, Fred Castruccio, Bolan Gan, Dian Putrasahan, Christopher D. Roberts, Luke Van Roekel, and Qiuying Zhang. Resolving and parameterising the ocean mesoscale in earth system models. *Current Climate Change Reports*, 6(4):137–152, Dec 2020.

[11] Kunxiaojia Yuan, Qing Zhu, William J. Riley, Fa Li, and Huayi Wu. Understanding and reducing the uncertainties of land surface energy flux partitioning within CMIP6 land models. *Agricultural and Forest Meteorology*, 319:108920, 2022.

[12] V. K. Arora, A. Katavouta, R. G. Williams, C. D. Jones, V. Brovkin, P. Friedlingstein, J. Schwinger, L. Bopp, O. Boucher, P. Cadule, M. A. Chamberlain, J. R. Christian, C. Delire, R. A. Fisher, T. Hajima, T. Ilyina, E. Joetzjer, M. Kawamiya, C. D. Koven, J. P. Krasting, R. M. Law, D. M. Lawrence, A. Lenton, K. Lindsay, J. Pongratz, T. Raddatz, R. Séférian, K. Tachiiri, J. F. Tjiputra, A. Wiltshire, T. Wu, and T. Ziehn. Carbon–concentration and carbon–climate feedbacks in CMIP6 models and their comparison to CMIP5 models. *Biogeosciences*, 17(16):4173–4222, 2020.

[13] Qing Zhu and Qianlai Zhuang. Parameterization and sensitivity analysis of a process-based terrestrial ecosystem model using adjoint method. *Journal of Advances in Modeling Earth Systems*, 6(2):315–331, 2014.

[14] M. Reichstein, G. Camps-Valls, B. Stevens, J. Denzler, N. Carvalhais, M. Jung, and Prabhat. Deep learning and process understanding for data-driven E arth system science. *Nature*, 566:195–204, Feb 2019.

[15] Gustau Camps-Valls, Devis Tuia, Xiao Xiang Zhu, and Markus Reichstein. *Bibliography*, pages 331–400. John Wiley & Sons, Ltd, 2021.

[16] G. Camps-Valls and L. Bruzzone. *Kernel methods for Remote Sensing Data Analysis*. Wiley & Sons, UK, Dec 2009.

[17] G. Tramontana, M. Jung, G. Camps-Valls, K. Ichii, B. Raduly, M. Reichstein, C. R. Schwalm, M. A. Arain, A. Cescatti, G. Kiely, L. Merbold, P. Serrano-Ortiz, S. Sickert, S. Wolf, and D. Papale. Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms. *Biogeosciences Discussions*, 2016:1–33, 2016.

[18] Gustau Camps-Valls, Devis Tuia, Xiao Xiang Zhu, and Markus Reichstein (Editors). *Deep learning for the Earth Sciences: A comprehensive approach to remote sensing, climate science and geosciences*. Wiley & Sons, 2021.

[19] Cynthia Rudin and Joanna Radin. Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. *Harvard Data Science Review*, 1(2), nov 22

2019. https://hdsr.mitpress.mit.edu/pub/f9kuryi8.

[20] Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. Dataset shift in machine learning. 2009.

[21] Masashi Sugiyama and Motoaki Kawanabe. *Learning Under Covariate Shift*, pages 19–19. 2012.

[22] Gary Marcus. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*, 2018.

[23] IPCC. *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, volume In Press. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2021.

[24] Behnam Neyshabur, Srinadh Bhojanapalli, David Mcallester, and Nati Srebro. Exploring generalization in deep learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[25] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, 2022.

[26] Xinwei Shen and Nicolai Meinshausen. Engression: Extrapolation for nonlinear regression?, 2023.

[27] Ribana Roscher, Bastian Bohn, Marco Duarte, and Jochen Garcke. Explainable machine learning for scientific insights and discoveries. *IEEE Access*, PP:1–1, 02 2020.

[28] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable AI: A review of machine learning interpretability methods. *Entropy*, 23(1), 2021.

[29] Gabrielle Ras, Ning Xie, Marcel Van Gerven, and Derek Doran. Explainable deep learning: A field guide for the uninitiated. *Journal of Artificial Intelligence Research*, 73:329–396, 2022.

[30] Antonios Mamalakis, Imme Ebert-Uphoff, and Elizabeth A. Barnes. *Explainable Artificial Intelligence in Meteorology and Climate Science: Model Fine-Tuning, Calibrating Trust and Learning New Science*, pages 315–339. Springer International Publishing, Cham, 2022.

[31] Adrian Höhl, Ivica Obadic, Miguel Ángel Fernández Torres, Hiba Najjar, Dario Oliveira, Zeynep Akata, Andreas Dengel, and Xiao Xiang Zhu. Opening the black-box: A systematic review on explainable ai in remote sensing, 2024.

[32] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, May 2019.

[33] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 16(none):1 – 85, 2022.

[34] Leon Sixt, Maximilian Granz, and Tim Landgraf. When explanations lie: Why many modified BP attributions fail. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9046–9057. PMLR, 13–18 Jul 2020.

[35] Timo Freiesleben and Gunnar König. Dear xai community, we need to talk! fundamental misconceptions in current xai research, 2023.

[36] Anuj Karpatne, Ramakrishnan Kannan, and Vipin Kumar. *Knowledge Guided Machine Learning: Accelerating Discovery using Scientific Knowledge and Data*. Chapman and Hall/CRC, 1 edition, 2022.

[37] Gustau Camps-Valls, Daniel Svendsen, Luca Martino, Jordi Muñoz-Marí, Valero Laparra, Manuel Campos-Taberner, and David Luengo. Physics-aware Gaussian processes in remote sensing. *Applied Soft Computing*, 68:69–82, Jul 2018.

[38] Gianluca Tramontana, Mirco Migliavacca, Martin Jung, Markus Reichstein, Trevor F. Keenan, Gustau Camps-Valls, Jerome Ogee, Jochem Verrelst, and Dario Papale. Partitioning net carbon dioxide fluxes into photosynthesis and respiration using neural networks. *Global Change Biology*, 26(9):5235–5253, 2020.

[39] Ankush Khandelwal, Shaoming Xu, Xiang Li, Xiaowei Jia, Michael Stienbach, Christopher Duffy, John Nieber, and Vipin Kumar. Physics guided machine learning methods for hydrology, 2020.

[40] Jordi Cortés-Andrés, Gustau Camps-Valls, Sebastian Sippel, Enikő Székely, Dino Sejdinovic, Emiliano Diaz, Adrián Pérez-Suay, Zhu Li, Miguel Mahecha, and Markus Reichstein. Physics-aware nonparametric regression models for Earth data analysis. *Environmental Research Letters*, 17(5), 2022.

[41] Licheng LIU, Wang Zhou, Kaiyu Guan, Bin Peng, Chongya Jiang, Jinyun Tang, Sheng Wang, Robert Grant, Symon Mezbahuddin, Xiaowei Jia, Shaoming Xu, Vipin Kumar, and Zhenong Jin. Knowledge-based Artificial Intelligence for Agroecosystem Carbon Budget and Crop Yield Estimation. *ESS Open Archive eprints*, 105:essoar.10509206, July 2023.

[42] Q. Zhu, F. Li, W. J. Riley, L. Xu, L. Zhao, K. Yuan, H. Wu, J. Gong, and J. Randerson. Building a machine learning surrogate model for wildfire activities within a global earth system model. *Geoscientific Model Development*, 15(5):1899–1911, 2022.

[43] M. Raissi, P. Perdikaris, and G.E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Jour. Comp. Phys.*, 378:686–707, 2019.

[44] Wen Li Zhao, Pierre Gentine, Markus Reichstein, Yao Zhang, Sha Zhou, Yeqiang Wen, Changjie Lin, Xi Li, and Guo Yu Qiu. Physics-constrained machine learning of evapotranspiration. *Geophysical Research Letters*, 46(24):14496–14507, 2019.

[45] Markus Reichstein, Bernhard Ahrens, Basil Kraft, Gustau Camps-Valls, Nuno Carvalhais, Fabian Gans, Pierre Gentine, and Alexander J. Winkler. Combining system modeling and machine learning into hybrid ecosystem modeling. In *Knowledge Guided Machine Learning*, page 26. Chapman and Hall/CRC, 1st edition edition, 2022.

[46] Akash Koppa, Dominik Rains, Petra Hulsman, Rafael Poyatos, and Diego G. Miralles. A deep learning-based hybrid model of global terrestrial evaporation. *Nature Communications*, 13(1):1912, Apr 2022.

[47] Chaopeng Shen, Alison P. Appling, Pierre Gentine, Toshiyuki Bandai, Hoshin Gupta, Alexandre Tartakovsky, Marco Baity-Jesi, Fabrizio Fenicia, Daniel Kifer, Li Li, Xiaofeng Liu, Wei Ren, Yi Zheng, Ciaran J. Harman, Martyn Clark, Matthew Farthing, Dapeng Feng, Praveen Kumar, Doaa Aboelyazeed, Farshid Rahmani, Yalan Song, Hylke E. Beck, Tadd Bindas, Dipankar Dwivedi, Kuai Fang, Marvin Höge, Chris Rackauckas, Binayak Mohanty, Tirthankar Roy, Chonggang Xu, and Kathryn Lawson. Differentiable modelling to unify machine learning and physical models for geosciences. *Nature Reviews Earth & Environment*, 4(8):552–567, Aug 2023.

[48] Johannes Oberpriller, David R. Cameron, Michael C. Dietze, and Florian Hartig. Towards robust statistical inference for complex computer models. *Ecology Letters*, 24(6):1251–1261, 2021.

[49] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021.

[50] Pavel Izmailov, Sharad Vikram, Matthew D. Hoffman, and Andrew Gordon Wilson. What are bayesian neural network posteriors really like? In *International Conference on Machine Learning*, 2021.

[51] Max Kuhn and Kjell Johnson. *Applied Predictive Modeling*. 01 2013.

[52] Sifan Wang, Shyam Sankaran, and Paris Perdikaris. Respecting causality is all you need for training physics-informed neural networks, 2022.

[53] Fernando Iglesias-Suarez, Pierre Gentine, Breixo Solino-Fernandez, Tom Beucler, Michael Pritchard, Jakob Runge, and Veronika Eyring. Causally-informed deep learning to improve climate models and projections, 2023.

[54] J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Clymour, M. Kretschmer, M. Mahecha, J. Muñoz-Marí, E. van Nes, J. Peters, R. Quax, M. Reichstein, M. Scheffer, B. Schölkopf, P. Spirtes, G. Sugihara, J. Sun, K. Zhang, and J. Zscheischler. Inferring causation from time series with perspectives in Earth system sciences. *Nature Communications*, (2553):1–13, 2019.

[55] Kunxiaojia Yuan, Qing Zhu, Fa Li, William J. Riley, Margaret Torn, Housen Chu, Gavin McNicol, Min Chen, Sara Knox, Kyle Delwiche, Huayi Wu, Dennis Baldocchi, Hongxu Ma, Ankur R. Desai, Jiquan Chen, Torsten Sachs, Masahito Ueyama, Oliver Sonnentag, Manuel Helbig, Eeva-Stiina Tuittila, Gerald Jurasinski, Franziska Koebsch, David Campbell, Hans Peter Schmid, Annalea Lohila, Mathias Goeckede, Mats B. Nilsson, Thomas Friborg, Joachim Jansen, Donatella Zona, Eugenie Euskirchen, Eric J. Ward, Gil Bohrer, Zhenong Jin, Licheng Liu, Hiroki Iwata, Jordan Goodrich, and Robert Jackson. Causality guided machine learning model on wetland ch4 emissions across global wetlands. *Agricultural and Forest Meteorology*, 324:109115, 2022.

[56] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018.

[57] Michael C. Knaus, Michael Lechner, and Anthony Strittmatter. Heterogeneous employment effects of job search programs. *Journal of Human Resources*, 57(2):597–636, mar 2020.

[58] Jonathan M.V. Davis and Sara B. Heller. Using causal forests to predict treatment heterogeneity: An application to summer jobs. *The American Economic Review*, 107(5):546–550, 2017.

[59] Qiguo Sun, Tianyuan Zheng, Xilai Zheng, Min Cao, Bo Zhang, and Shiqiang Jiang. Causal interpretation for groundwater exploitation strategy in a coastal aquifer. *Science of The Total Environment*, 867:161443, 2023.

[60] Svante Arrhenius. Über die reaktionsgeschwindigkeit bei der inversion von rohrzucker durch säuren. *Zeitschrift für physikalische Chemie*, 4(1):226–248, 1889.

[61] Jacobus Henricus Van't Hoff, Robert Alfred Lehfeldt, et al. Lectures on theoretical and physical chemistry. 1899.

[62] John Lloyd and JA Taylor. On the temperature dependence of soil respiration. *Functional ecology*, pages 315–323, 1994.

[63] Yanyan Pei, Jinwei Dong, Yao Zhang, Wenping Yuan, Russell Doughty, Jilin Yang, Decheng Zhou, Liangxia Zhang, and Xiangming Xiao. Evolution of light use efficiency models: Improvement, uncertainties, and implications. *Agricultural and Forest Meteorology*, 317:108905, 2022.

[64] Miko UF Kirschbaum. Will changes in soil organic carbon act as a positive or negative feedback on global warming? *Biogeochemistry*, 48:21–51, 2000.

[65] Nicholas G Smith and Jeffrey S Dukes. Plant respiration and photosynthesis in global-scale models: incorporating acclimation to temperature and CO2. *Global change biology*, 19(1):45–63, 2013.

[66] Chris Huntingford, Owen K Atkin, Alberto Martinez-De La Torre, Lina M Mercado, Mary A Heskel, Anna B Harper, Keith J Bloomfield, Odhran S O'sullivan, Peter B Reich, Kirk R Wythers, et al. Implications of improved representations of plant respiration in a changing climate. *Nature Communications*, 8(1):1602, 2017.

[67] Gal Vardi. On the implicit bias in deep-learning algorithms. *Commun. ACM*, 66(6):86–93, may 2023.

[68] Weiwei Zhan, Xi Yang, Youngryel Ryu, Benjamin Dechant, Yu Huang, Yves Goulas, Minseok Kang, and Pierre Gentine. Two for one: Partitioning CO2 fluxes and understanding the relationship between solar-induced chlorophyll fluorescence and gross primary productivity using machine learning. *Agricultural and Forest Meteorology*, 321:108980, 2022.

[69] Reda ElGhawi, Basil Kraft, Christian Reimers, Markus Reichstein, Marco Körner, Pierre Gentine, and Alexander J Winkler. Hybrid modeling of evapotranspiration: inferring stomatal and aerodynamic resistances using combined physics-based and machine learning. *Environmental Research Letters*, 18(3):034039, mar 2023.

[70] Yuan Yin, Vincent Le Guen, Jérémie Dona, Emmanuel de Bézenac, Ibrahim Ayed, Nicolas Thome, and Patrick Gallinari. Augmenting physical models with deep networks for complex dynamics forecasting*. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124012, dec 2021.

[71] Paul Hünermund, Beyers Louw, and Itamar Caspi. Double machine learning and automated

confounder selection: A cautionary tale. *Journal of Causal Inference*, 11(1):20220078, 2023.

[72] Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148 – 1178, 2019.

[73] X Nie and S Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 09 2020.

[74] Dylan J. Foster and Vasilis Syrgkanis. Orthogonal statistical learning, 2020.

[75] Denis Nekipelov, Vira Semenova, and Vasilis Syrgkanis. Regularized orthogonal machine learning for nonlinear semiparametric models, 2021.

[76] Gordon Bonan. *Ecological Climatology: Concepts and Applications*. Cambridge University Press, 3 edition, 2015.

[77] George Burba. *Eddy Covariance Method for Scientific, Industrial, Agricultural and Regulatory Applications: A Field Book on Measuring Ecosystem Gas Exchange and Areal Emission Rates*. 06 2013.

[78] D. Baldocchi, E. Falge, L. Gu, R. Olson, D. Hollinger, S. Running, P. Anthoni, C. Bernhofer, K. Davis, R. Evans, J. Fuentes, A. Goldstein, G. Katul, B. Law, X. Lee, Y. Malhi, T. Meyers, W. Munger, W. Oechel, Paw U,K.T., K. Pilegaard, H.P. Schmid, R. Valentini, S. Verma, T. Vesala, K. Wilson, and S. Wofsy. Fluxnet: A new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities. *Bulletin of the American Meteorological Society*, 82(11):2415–2434, 2001.

[79] E. Falge, J. Tenhunen, M. Aubinet, C. Bernhofer, R. Clement, A. Granier, A. Kowalski, E. Moors, K. Pilegaard, Ü. Rannik, and C. Rebmann. *A Model-Based Study of Carbon Fluxes at Ten European Forest Sites*, pages 151–177. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003.

[80] Gilberto Pastorello, Carlo Trotta, Eleonora Canfora, Housen Chu, Danielle Christianson, You-Wei Cheah, Cristina Poindexter, Jiquan Chen, Abdelrahman Elbashandy, Marty Humphrey, Peter Isaac, Diego Polidori, Markus Reichstein, Alessio Ribeca, Catharine van Ingen, Nicolas Vuichard, Leiming Zhang, Brian Amiro, Christof Ammann, M Altaf Arain, Jonas Ardö, Timothy Arkebauer, Stefan K Arndt, Nicola Arriga, Marc Aubinet, Mika Aurela, Dennis Baldocchi, Alan Barr, Eric Beamesderfer, Luca Belelli Marchesini, Onil Bergeron, Jason Beringer, Christian Bernhofer, Daniel Berveiller, Dave Billesbach, Thomas Andrew Black, Peter D Blanken, Gil Bohrer, Julia Boike, Paul V Bolstad, Damien Bonal, Jean-Marc Bonnefond, David R Bowling, Rosvel Bracho, Jason Brodeur, Christian Brümmer, Nina Buchmann, Benoit Burban, Sean P Burns, Pauline Buysse, Peter Cale, Mauro Cavagna, Pierre Cellier, Shiping Chen, Isaac Chini, Torben R Christensen, James Cleverly, Alessio Collalti, Claudia Consalvo, Bruce D Cook, David Cook, Carole Coursolle, Edoardo Cremonese, Peter S Curtis, Ettore D'Andrea, Humberto da Rocha, Xiaoqin Dai, Kenneth J Davis, Bruno De Cinti, Agnes de Grandcourt, Anne De Ligne, Raimundo C De Oliveira, Nicolas Delpierre, Ankur R Desai, Carlos Marcelo Di Bella, Paul di Tommasi, Han Dolman, Francisco Domingo, Gang Dong, Sabina Dore, Pierpaolo Duce, Eric Dufrêne, Allison Dunn, Jiří Dušek, Derek Eamus, Uwe Eichelmann, Hatim Abdalla M ElKhidir, Werner Eugster, Cacilia M Ewenz, Brent Ewers, Daniela Famulari, Silvano Fares, Iris Feigenwinter, Andrew Feitz, Rasmus Fensholt, Gianluca Filippa, Marc Fischer, John Frank, Marta Galvagno, Mana Gharun, Damiano Gianelle, Bert Gielen, Beniamino Gioli, Anatoly Gitelson, Ignacio Goded, Mathias Goeckede, Allen H Goldstein, Christopher M Gough, Michael L Goulden, Alexander Graf, Anne Griebel, Carsten Gruening, Thomas Grünwald, Albin Hammerle, Shijie Han, Xingguo Han, Birger Ulf Hansen, Chad Hanson, Juha Hatakka, Yongtao He, Markus Hehn, Bernard Heinesch, Nina Hinko-Najera, Lukas Hörtnagl, Lindsay Hutley, Andreas Ibrom, Hiroki Ikawa, Marcin Jackowicz-Korczynski, Dalibor Janouš, Wilma Jans, Rachhpal Jassal, Shicheng Jiang, Tomomichi Kato, Myroslava Khomik, Janina Klatt, Alexander Knohl, Sara Knox, Hideki Kobayashi, Georgia Koerber, Olaf Kolle, Yoshiko Kosugi, Ayumi Kotani, Andrew Kowalski, Bart Kruijt, Julia Kurbatova, Werner L Kutsch, Hyojung Kwon, Samuli Launiainen, Tuomas Laurila, Bev Law, Ray Leuning, Yingnian Li, Michael Liddell, Jean-Marc Limousin, Marryanna Lion, Adam J Liska, Annalea Lohila, Ana López-Ballesteros,

Efrén López-Blanco, Benjamin Loubet, Denis Loustau, Antje Lucas-Moffat, Johannes Lüers, Siyan Ma, Craig Macfarlane, Vincenzo Magliulo, Regine Maier, Ivan Mammarella, Giovanni Manca, Barbara Marcolla, Hank A Margolis, Serena Marras, William Massman, Mikhail Mastepanov, Roser Matamala, Jaclyn Hatala Matthes, Francesco Mazzenga, Harry McCaughey, Ian McHugh, Andrew M S McMillan, Lutz Merbold, Wayne Meyer, Tilden Meyers, Scott D Miller, Stefano Minerbi, Uta Moderow, Russell K Monson, Leonardo Montagnani, Caitlin E Moore, Eddy Moors, Virginie Moreaux, Christine Moureaux, J William Munger, Taro Nakai, Johan Neirynck, Zoran Nesic, Giacomo Nicolini, Asko Noormets, Matthew Northwood, Marcelo Nosetto, Yann Nouvellon, Kimberly Novick, Walter Oechel, Jørgen Eivind Olesen, Jean-Marc Ourcival, Shirley A Papuga, Frans-Jan Parmentier, Eugenie Paul-Limoges, Marian Pavelka, Matthias Peichl, Elise Pendall, Richard P Phillips, Kim Pilegaard, Norbert Pirk, Gabriela Posse, Thomas Powell, Heiko Prasse, Suzanne M Prober, Serge Rambal, Üllar Rannik, Naama Raz-Yaseef, Corinna Rebmann, David Reed, Victor Resco de Dios, Natalia Restrepo-Coupe, Borja R Reverter, Marilyn Roland, Simone Sabbatini, Torsten Sachs, Scott R Saleska, Enrique P Sánchez-Cañete, Zulia M Sanchez-Mejia, Hans Peter Schmid, Marius Schmidt, Karl Schneider, Frederik Schrader, Ivan Schroder, Russell L Scott, Pavel Sedlák, Penélope Serrano-Ortíz, Changliang Shao, Peili Shi, Ivan Shironya, Lukas Siebicke, Ladislav Šigut, Richard Silberstein, Costantino Sirca, Donatella Spano, Rainer Steinbrecher, Robert M Stevens, Cove Sturtevant, Andy Suyker, Torbern Tagesson, Satoru Takanashi, Yanhong Tang, Nigel Tapper, Jonathan Thom, Michele Tomassucci, Juha-Pekka Tuovinen, Shawn Urbanski, Riccardo Valentini, Michiel van der Molen, Eva van Gorsel, Ko van Huissteden, Andrej Varlagin, Joseph Verfaillie, Timo Vesala, Caroline Vincke, Domenico Vitale, Natalia Vygodskaya, Jeffrey P Walker, Elizabeth Walter-Shea, Huimin Wang, Robin Weber, Sebastian Westermann, Christian Wille, Steven Wofsy, Georg Wohlfahrt, Sebastian Wolf, William Woodgate, Yuelin Li, Roberto Zampedri, Junhui Zhang, Guoyi Zhou, Donatella Zona, Deb Agarwal, Sebastien Biraud, Margaret Torn, and Dario Papale. The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. *Scientific Data*, 7(1):225, July 2020.

[81] P. M. Robinson. Root-N-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988.

[82] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[83] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

[84] Anders Krogh and John A. Hertz. A simple weight decay can improve generalization. In *Proceedings of the 4th International Conference on Neural Information Processing Systems*, NIPS'91, page 950–957, San Francisco, CA, USA, 1991. Morgan Kaufmann Publishers Inc.

[85] Y. Luo and Xuhui Zhou. Soil respiration and the environment. *Soil Respiration and the Environment*, 01 2006.

[86] F Stuart Chapin, Pamela A Matson, and Harold A Mooney. *Principles of terrestrial ecosystem ecology*. Springer, New York, NY, 2002 edition, May 2013.

[87] Markus Reichstein, Eva Falge, Dennis Baldocchi, Dario Papale, Marc Aubinet, Paul Berbigier, Christian Bernhofer, Nina Buchmann, Tagir Gilmanov, André Granier, Thomas Grünwald, Katka Havránková, Hannu Ilvesniemi, Dalibor Janous, Alexander Knohl, Tuomas Laurila, Annalea Lohila, Denis Loustau, Giorgio Matteucci, Tilden Meyers, Franco Miglietta, Jean-Marc Ourcival, Jukka Pumpanen, Serge Rambal, Eyal Rotenberg, Maria Sanz, John Tenhunen, Günther Seufert, Francesco Vaccari, Timo Vesala, Dan Yakir, and Riccardo Valentini. On the separation of net ecosystem exchange into assimilation and ecosystem respiration: review and improved algorithm. *Global Change Biology*, 11(9):1424–1439, 2005.

[88] Antje M. Moffat, Dario Papale, Markus Reichstein, David Y. Hollinger, Andrew D. Richardson, Alan G. Barr, Clemens Beckstein, Bobby H. Braswell, Galina Churkina, Ankur R. Desai, Eva Falge, Jeffrey H. Gove, Martin Heimann, Dafeng Hui, Andrew J. Jarvis, Jens Kattge, Asko

Noormets, and Vanessa J. Stauch. Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes. *Agricultural and Forest Meteorology*, 147(3):209–232, 2007.

[89] Ankur Rashmikant Desai, Andrew D. Richardson, Antje Maria Moffat, Jens Kattge, D. Hollinger, Alan G. Barr, Eva Falge, Asko Noormets, Dario Papale, Markus Reichstein, and Vanessa J. Stauch. Cross-site evaluation of eddy covariance GPP and RE decomposition techniques. *Agricultural and Forest Meteorology*, 148:821–838, 2008.

[90] Gitta Lasslop, Markus Reichstein, Dario Papale, Andrew D. Richardson, Almut Arneth, Alan Barr, Paul Stoy, and Georg Wohlfahrt. Separation of net ecosystem exchange into assimilation and respiration using a light response curve approach: critical issues and global evaluation. *Global Change Biology*, 16(1):187–208, 2010.

[91] Trevor F. Keenan, Mirco Migliavacca, Dario Papale, Dennis Baldocchi, Markus Reichstein, Margaret Torn, and Thomas Wutzler. Widespread inhibition of daytime ecosystem respiration. *Nature Ecology & Evolution*, 3(3):407–415, Mar 2019.

[92] Violeta Teodora Trifunov, Maha Shadaydeh, Jakob Runge, Markus Reichstein, and Joachim Denzler. A data-driven approach to partitioning net ecosystem exchange using a deep state space model. *IEEE Access*, 9:107873–107883, 2021.

[93] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189 – 1232, 2001.

[94] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR.

[95] Miguel D. Mahecha, Markus Reichstein, Nuno Carvalhais, Gitta Lasslop, Holger Lange, Sonia I. Seneviratne, Rodrigo Vargas, Christof Ammann, M. Altaf Arain, Alessandro Cescatti, Ivan A. Janssens, Mirco Migliavacca, Leonardo Montagnani, and Andrew D. Richardson. Global convergence in the temperature sensitivity of respiration at ecosystem level. *Science*, 329(5993):838–840, 2010.

[96] Russell L. Scott, Joel A. Biederman, Erik P. Hamerlynck, and Greg A. Barron-Gafford. The carbon balance pivot point of southwestern u.s. semiarid ecosystems: Insights from the 21st century drought. *Journal of Geophysical Research: Biogeosciences*, 120(12):2612–2624, 2015.

[97] M. Jung, C. Schwalm, M. Migliavacca, S. Walther, G. Camps-Valls, S. Koirala, P. Anthoni, S. Besnard, P. Bodesheim, N. Carvalhais, F. Chevallier, F. Gans, D. S. Goll, V. Haverd, P. Köhler, K. Ichii, A. K. Jain, J. Liu, D. Lombardozzi, J. E. M. S. Nabel, J. A. Nelson, M. O'Sullivan, M. Pallandt, D. Papale, W. Peters, J. Pongratz, C. Rödenbeck, S. Sitch, G. Tramontana, A. Walker, U. Weber, and M. Reichstein. Scaling carbon fluxes from eddy covariance sites to globe: synthesis and evaluation of the fluxcom approach. *Biogeosciences*, 17(5):1343–1365, 2020.

Appendix A. Method

Appendix A.1. Derivation of DML estimator for g

One way of obtaining an estimator for g instead of fitting it directly is by reusing all estimators of DML. It is easy to see that

$$\begin{aligned} g(X,W) &= \mathbb{E}[g(X,W)|X,W] \\ &= \mathbb{E}[Y - \theta(X)f(T) - \epsilon|X,W] \\ &= \mathbb{E}[Y|X,W] - \mathbb{E}[\theta(X)f(T)|X,W] - \underbrace{\mathbb{E}[\epsilon|X,W]}_{=0} \\ &= \mathbb{E}[Y|X,W] - \theta(X)\mathbb{E}[f(T)|X,W] \\ &\approx \mathbb{E}[Y|X,W] - \hat{\theta}(X)\mathbb{E}[f(T)|X,W], \end{aligned}$$

where $\mathbb{E}[Y|X,W]$ represents the estimator of Y on X and W and $\mathbb{E}[f(T)|X,W]$ the estimator of $f(T)$ on X and W. From here, one can use an ensemble of the first-stage estimators over all folds to obtain the estimator of $\mathbb{E}[Y|X,W]$ and the estimator of $\mathbb{E}[f(T)|X,W]$. The estimator $\hat{\theta}(X)$ is a single estimator obtained as the result of DML.

Appendix B. Data

Appendix B.1. Synthetic data

Appendix B.1.1. Q_{10} model We use measured air temperature T_A and potential incoming radiation SW_{POT} for the synthetic data. Further, we compute

$$\text{for } Q_{10} \in \{1.5, 1.25, 1.75\}, \tag{B.1}$$

$$R_{\text{eco}}^{\text{syn}} = R_b^{\text{syn}} \cdot Q_{10}^{0.1\cdot(T_A-15)} \cdot (1+\epsilon), \tag{B.2}$$

$$R_b^{\text{syn}} = 0.75 \cdot (\tilde{R}_b^{syn} - \min(\tilde{R}_b^{syn}) + 0.1 \cdot \pi), \tag{B.3}$$

$$\tilde{R}_b^{syn} = 0.01 \cdot SW_{\text{POT}}^{\text{SM}} - 0.005 \cdot SW_{\text{POT}}^{\text{SM,diff}}, \tag{B.4}$$

where R_b^{syn} describes the base respiration, which we compute with a smooth daily radiation cycle. The smooth incoming potential radiation $SW_{\text{POT}}^{\text{SM}}$ and its smoothed difference quotient $SW_{\text{POT}}^{\text{SM,diff}}$ are computed by averaging moving windows of 10 days over the incoming potential radiation SW_{POT}. We apply the computations in (B.3) to ensure that R_b^{syn} is always positive. We sample ϵ from a centered truncated normal distribution with 0.2 standard deviation in the interval $[-0.95, 0.95]$ to obtain heteroscedastic noise over the observations.

Appendix B.1.2. LUE model The code for generating the data is taken from the work of [45], where the authors approach the partitioning of fluxes with neural networks on a synthetic dataset. $R_{\text{eco}}^{\text{syn}}$ is computed similarly as in the study on Q_{10}. While, for

generating GPP, we use the light-use efficiency model with LUE being a function of VPD and temperature T_A:

$$GPP^{syn} = LUE^{syn} \cdot SW_{in}, \tag{B.5}$$

$$LUE^{syn} = 0.5 \cdot \exp\left(-0.1 \cdot (T_A - 20))^2\right) \cdot \min(1, \exp(-0.1 \cdot (VPD - 10))). \tag{B.6}$$

Finally, we compute NEE following (7) with additional multiplicative heteroscedastic noise:

$$NEE^{syn} = (-GPP^{syn} + R_{\text{eco}}^{\text{syn}}) \cdot (1 + \sigma\varepsilon), \tag{B.7}$$

where noise $\varepsilon \sim \mathcal{N}(0, 1)$ is sampled from a standard Gaussian distribution and σ varies in $\{0, 0.05, 0.1, 0.2, 0.4, 0.7, 1.0, 2.0\}$.

Appendix B.2. FLUXNET sites

The 36 FLUXNET sites used for the flux partitioning experiments are shown in Table B1. The table further provides information on plant type, latitude, and longitude.

Appendix B.3. Details on the neural networks

The NNs used for the GD-based HM had two hidden layers with 16 units each. A tanh nonlinearity was applied at the end of each hidden layer. A final softplus function was applied to the output of the last layer to obtain non-negative results for the base respiration. This function is a smooth approximation of the $ReLU$ function. For the case of regularization, dropout was applied to the outputs of the hidden layers at a rate of 0.2. To probe other instances of regularization, we also used weight decay with hyperparameter 0.1 instead of dropout. The initial Q_{10} is sampled from a Gaussian with $\sigma = 0.1$ and $\mu = 1.5$ (or 1.25, 1.75 for the respective experiments). For the DML-based HM approach, we used the same network architecture without final softplus for the first-stage estimators. For the estimation of R_b after obtaining Q_{10}, we used the same network again, but this time we included the softplus nonlinearity. We used stochastic gradient descent with the Adam optimizer [82] for the training. We apply exponential learning rate decay as a scheduler with a decay rate of 0.95 over 500 steps. We trained the first stage estimators of the DML over 2000 iterations each. For the GD-based HM and the final g estimator in the causal DML-based HM, we trained over 10000 iterations. To avoid overfitting, 20% of the data is always kept as validation data for model selection.

Appendix C. Additional results

Appendix C.1. Regularization with weight decay

We reran the same setup with weight decay to show that the findings also apply to other regularization techniques beyond dropout. We find qualitatively similar results, where

Table B1: FLUXNET sites used for flux partitioning experiments with DML.

ID	Site code	IGBP	Lat	Lon	Years available
1	AU-Cpr	SAV	-34,00	140,59	2010–2014
2	AU-DaP	GRA	-14,06	131,32	2007–2013
3	AU-Dry	SAV	-15,26	132,37	2008–2014
4	AU-How	WSA	-12,49	131,15	2001–2014
5	AU-Stp	GRA	-17,15	133,35	2008–2014
6	BE-Lon	CRO	50,55	4,75	2004–2014
7	BE-Vie	MF	50,31	6,00	1996–2014
8	CA-Qfo	ENF	49,69	-74,34	2003–2010
9	DE-Geb	CRO	51,10	10,91	2001–2014
10	DE-Gri	GRA	50,95	13,51	2004–2014
11	DE-Kli	CRO	50,89	13,52	2004–2014
12	DE-Obe	ENF	50,79	13,72	2008–2014
13	DE-Tha	ENF	50,96	13,57	1996–2014
14	DK-Sor	DBF	55,49	11,64	1996–2014
15	FI-Hyy	ENF	61,85	24,29	1996–2014
16	FR-LBr	ENF	44,72	-0,77	1996–2008
17	GF-Guy	EBF	5,28	-52,92	2004–2014
18	IT-BCi	CRO	40,52	14,96	2004–2014
19	IT-Cp2	EBF	41,70	12,36	2012–2014
20	IT-Cpz	EBF	41,71	12,38	1997–2009
21	IT-MBo	GRA	46,01	11,05	2003–2013
22	IT-Noe	CSH	40,61	8,15	2004–2014
23	IT-Ro1	DBF	42,41	11,93	2000–2008
24	IT-SRo	ENF	43,73	10,28	1999–2012
25	NL-Loo	ENF	52,17	5,74	1996–2014
26	RU-Fyo	ENF	56,46	32,92	1998–2014
27	US-ARM	CRO	36,61	-97,49	2003–2012
28	US-GLE	ENF	41,37	-106,24	2004–2014
29	US-MMS	DBF	39,32	-86,41	1999–2014
30	US-NR1	ENF	40,03	-105,55	1999–2014
31	US-SRG	GRA	31,79	-110.83	2008–2014
32	US-SRM	WSA	31,82	-110,87	2004–2014
33	US-UMB	DBF	45,56	-84,71	2000–2014
34	US-Whs	OSH	31,74	-110,05	2007–2014
35	US-Wkg	GRA	31,74	-109,94	2004–2014
36	ZA-Kru	SAV	-25,02	31,50	2000–2013

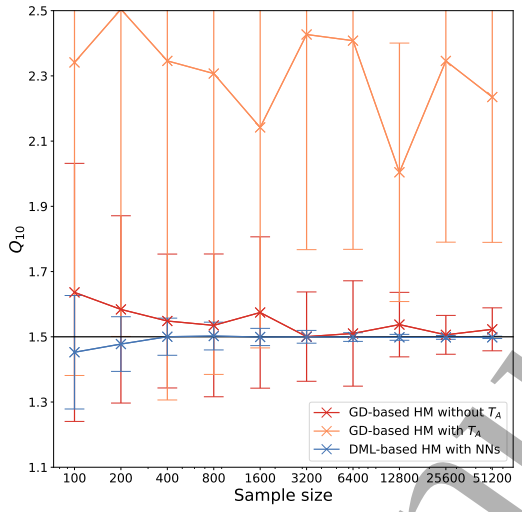


Figure C1: Additional simulation study for Q_{10} estimation with the GD-based HM and the DML-based HM similar to Fig. 5. with weight decay. Here, weight decay with a rate of 0.1 has been applied as regularization.

the DML-based HM converges robustly to the right Q_{10} values where the GD-based HM converges much slower and remains biased (see Fig. C1).

Appendix C.2. Additional dropout rates

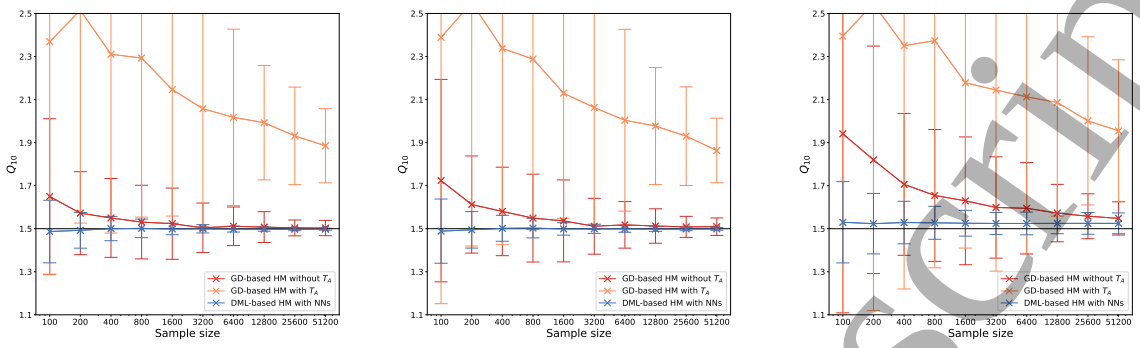
We ran the experiments with dropout rates of 0.05, 0.1, and 0.3. With increasing dropout rates, all methods have increasing errors in estimating Q_{10}. At a rate of 0.3, the estimation with double machine learning has a constant bias as the first-stage estimators do not converge sufficiently fast. Still, it stays robust with little data and outperforms the baseline over the whole data range.

Appendix C.3. Additional Q_{10} values

We ran the experiments with and without dropout with 1.25 and 1.75 as two additional Q_{10} values. We find that these setups affirm the observations for $Q_{10} = 1.5$. The errors in estimating the Q_{10} values grow and shrink proportionally to the magnitude of Q_{10}. This is to be expected as we deploy multiplicative noise, and thus, with higher Q_{10}, the magnitude of respiration and, hence, the absolute noise level grows (see Section Appendix C).

Appendix C.4. Retrievel of linear model

We generated synthetic data following [45], a partially linear LUE model with varying coefficients. We used time series of measured meteorological forcings as inputs and



(a) Dropout rate of 0.05. **(b)** Dropout rate of 0.1. **(c)** Dropout rate of 0.3.

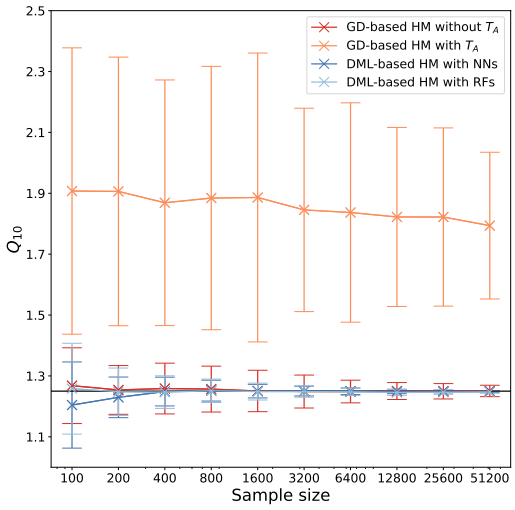
Figure C2: Additional simulation study for Q_{10} estimation with the GD-based HM and the DML-based HM similar to Fig. 5. We applied varying dropout rates to probe the robustness of the method. An increasing dropout rate leads to an increase in bias. The approach based on double machine learning consistently outperforms the neural network baseline.

added heteroscedastic noise over different noise levels (see Section Appendix B.1.2 for details).

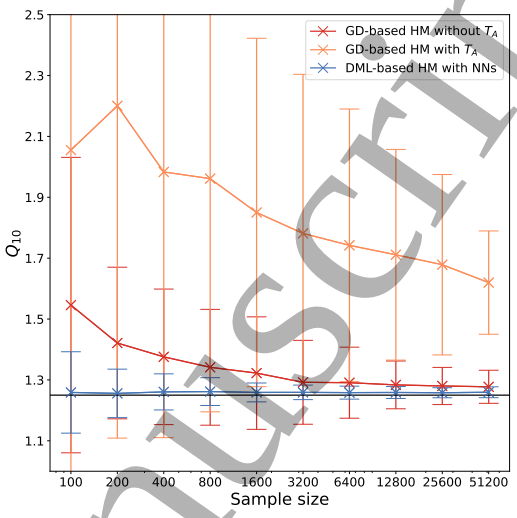
To test the robustness of the approach to noise, we perform experiments with an increasing level of heteroscedastic noise. The R^2 and RMSE of the retrieved fluxes are reported in Table C1 and Table C2. We note that the DML approach gives theoretical guarantees for estimating GPP and not necessarily for R_{eco} [72, 74]. Our proposed method retrieves reasonable estimates of GPP with a medium R^2 of 0.997 in the no-noise scenario. Even a heteroscedastic noise level of 0.4 does not yield any substantial drop in performance. Beyond that, the method is still robust as it retrieves the correct GPP at a noise level of 1.00 with a median value of 0.922. In flux partitioning, retrieving R_{eco} can be more challenging as it has a smaller magnitude than GPP, implying a smaller signal-to-noise ratio. Moreover, even though there is no guarantee on the used plugin-in estimator for R_{eco}, which we obtain by recycling the estimators of the DML approach, we still find it to yield useful results. The retrieved fluxes have a median R^2 over all site-years of 0.94. As expected, the effect of the noise on the retrieval of R_{eco} is stronger, but up to a σ of 0.4, the results are not strongly affected. When we combine both models, we obtain a model of NEE. Even with strong noise, this estimator retrieves reasonable estimates of the NEE signal.

Appendix D. Reproducibility

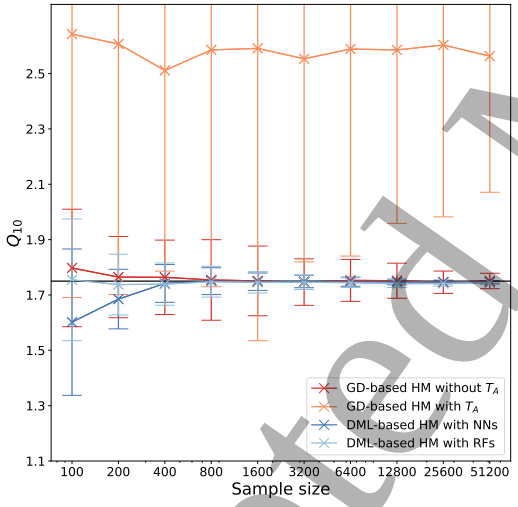
The data used to carry out experiments is available at https://fluxnet.org/data/fluxnet2015-dataset/. All code is being made available at https://github.com/KaiHCohrs/hybrid-q10-model-chm and https://github.com/KaiHCohrs/dml-4-fluxes-chm.



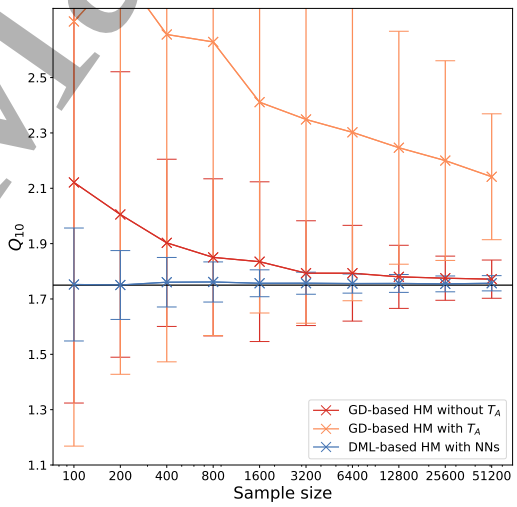
(a) Q_{10} of 1.25 without dropout.



(b) Q_{10} of 1.25 with dropout.



(c) Q_{10} of 1.75 without dropout.



(d) Q_{10} of 1.75 with dropout.

Figure C3: Additional simulation study for Q_{10} estimation with the GD-based HM and the DML-based HM similar to Fig. 5 with different values for Q_{10} In a) and b) Q_{10} was set to 1.25 and in c) and d) to 1.75. The findings are qualitatively similar to the case of 1.5. The magnitude of the errors grows with the magnitude of Q_{10}.

Table C1: Coefficient of determination R^2 for generated data on all 36 flux sites with different heteroscedastic noise levels between the GPP, $RECO$ and NEE obtained with the DML approach and the respective ground truth. For NEE, the noise-free value is stated. The reported statistics are the median and in brackets, the 0.25 and 0.75 quantiles over all site-years.

σ	GPP	R_{eco}	NEE_{clean}
0.00	0.997(0.994/0.998)	0.940(0.923/0.960)	0.978(0.973/0.983)
0.05	0.997(0.994/0.998)	0.940(0.923/0.959)	0.978(0.973/0.983)
0.10	0.997(0.993/0.998)	0.939(0.922/0.958)	0.978(0.973/0.982)
0.20	0.996(0.991/0.998)	0.936(0.917/0.956)	0.977(0.972/0.982)
0.40	0.993(0.985/0.996)	0.931(0.911/0.947)	0.975(0.969/0.979)
0.70	0.986(0.961/0.991)	0.914(0.888/0.929)	0.970(0.963/0.975)
1.00	0.977(0.930/0.985)	0.887(0.846/0.910)	0.964(0.955/0.970)
2.00	0.922(0.707/0.952)	0.751(0.617/0.813)	0.937(0.910/0.948)

Table C2: The RMSE (in $\frac{\mu\,\text{mol}\,CO_2}{m^2 s}$) for generated data on all 36 flux sites with different heteroscedastic noise levels between the GPP, $RECO$ and NEE obtained with the DML approach and the respective ground truth. For NEE, the noise-free and noisy values are stated. The reported statistics are the median and, in brackets, the 0.25 and 0.75 quantiles over all site-years.

σ	GPP	R_{eco}	NEE_{clean}	NEE_{noisy}
0.00	0.320(0.227/0.454)	0.861(0.770/1.104)	0.872(0.768/1.079)	0.872(0.768/ 1.079)
0.05	0.330(0.234/0.467)	0.864(0.771/1.109)	0.873(0.770/1.083)	1.029(0.827/ 1.311)
0.10	0.359(0.243/0.491)	0.878(0.778/1.136)	0.880(0.770/1.097)	1.197(0.949/ 1.615)
0.20	0.401(0.284/0.600)	0.921(0.794/1.184)	0.898(0.781/1.128)	1.701(1.346/ 2.573)
0.40	0.515(0.386/0.772)	0.973(0.825/1.335)	0.941(0.808/1.219)	2.977(2.349/ 4.850)
0.70	0.758(0.543/1.152)	1.139(0.895/1.577)	1.025(0.862/1.358)	5.101(3.965/ 8.434)
1.00	1.005(0.715/1.589)	1.285(0.971/1.872)	1.147(0.927/1.467)	7.162(5.583/11.949)
2.00	1.804(1.268/2.972)	1.880(1.361/3.058)	1.500(1.196/2.186)	14.316(11.104/23.889)