DOUBLE MACHINE LEARNING FOR CAUSAL HYBRID MODELING - APPLICATIONS IN THE EARTH SCIENCES

A PREPRINT

Kai-Hendrik Cohrs
Image Processing Laboratory
Universitat de València
València, Spain
kai.cohrs@uv.es

Gherardo Varando
Image Processing Laboratory
Universitat de València
València, Spain
gherardo.varando@uv.es

Gustau Camps-Valls
Image Processing Laboratory
Universitat de València
València, Spain
gcamps@uv.es

Nuno Carvalhais
Max Planck Institute for Biogeochemistry
Jena, Germany
ELLIS Unit Jena
ncarvalhais@bgc-jena.mpg.de

Markus Reichstein
Max Planck Institute for Biogeochemistry
Jena, Germany
ELLIS Unit Jena
Markus.Reichstein@bgc-jena.mpg.de

February 22, 2024

ABSTRACT

Hybrid modeling integrates machine learning with scientific knowledge with the goal of enhancing 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. Nevertheless, they are not always sufficient to capture the complex and usually not completely known relationships in the real world. Support vector machines [7], random forests (RFs) [8], or neural networks (NNs) [9] 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 [10]. (ii) Often, ML models are not robust and fail to generalize out of the domain of the data used for training [11, 12]. (iii) They violate physical properties and laws of nature, such as conservation laws, symmetries, or equi- and invariances [13, 14]. 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 [15].

All these issues are gaining attention in ML and Earth system science literature. Explainable artificial intelligence (XAI) tackles questions on the explainability of black box models [16, 17], and research in generalization and extrapolation aims at ensuring robustness outside of the training domain [18–20]. A flourishing area of research is science-aware or knowledge-guided machine learning, which combines the knowledge-driven and data-driven worlds to overcome inconsistencies [21–24]. One example is physics-informed neural networks (PINNs) [25], 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 [22]. Finally, hybrid modeling replaces some components of mechanistic models with machine learning [26–28]. This constraint makes the models more interpretable and serves as a regularizer for better generalization to unseen data.

However, 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 common 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 [29], which becomes even more difficult and prohibitively expensive in deep learning [30, 31]. Ultimately, equifinality can jeopardize the interpretability of the results. Second, regularization techniques in machine learning can introduce bias on the physical parameters [27]. 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 [32]. For hybrid modeling, though, apart from equifinality, this can lead to different interpretations of the parameter of interest and thus needs to be done with care.

Let us illustrate the opportunities and challenges of hybrid modeling in a relevant geoscience problem that will accompany us throughout this work. Modeling the temperature dependence of ecosystem respiration is a fundamental step in better understanding biosphere evolution and responses under global warming scenarios [33–35]. The functional relationship between temperature and respiration has been classically represented via the Q_{10} respiration model:

$$R_{eco}(X, T_A) = R_b(X, T_A) \cdot Q_{10}^{(T_A - T_{A,ref})/10}, \tag{1}$$

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 [27]. Because of the optimization technique, we will refer to this method as *gradient-descent-based hybrid modeling (GD-based HM)*.

Equifinality in this problem can be shown by reformulating (1) for $c > 0$:

$$R_{eco}(X, T_A) = R_b(X, T_A) c^{(T_A - T_{A,ref})/10} \cdot \left(\frac{Q_{10}}{c}\right)^{(T_A - T_{A,ref})/10}. \tag{2}$$

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 evidently 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 [36] 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 [37, 38]. Similarly, a regularization term can make the problem identifiable. This has been formally proven for solving hybrid ODEs [39]. Regularization, however, is known to introduce bias on parameters of interest in semi-parametric modeling problems [40].

Finally, to illustrate the importance of the choice of input parameters, consider that in (1), the exclusion of seasonality variables leads to a Q_{10} that does not only describe the immediate temperature effects but also the stronger variability over the year. The parameter Q_{10} would hence be larger than it should be if we want it to model just the immediate temperature response. A negligent selection of input variables can strongly impact the final estimator of Q_{10}. *Being right for the wrong reasons* is thus a major problem if we want hybrid models to be interpretable.

The crucial point is that we want Q_{10} to describe a *causal effect*, as we believe that temperature is the direct driver of respiration. This is opposed to the usual machine learning scenario, where the mere correlation of variables is enough to achieve good predictive performance. When moving from correlation to causation, we can intervene on the variables, i.e., change the temperature values and hope for a realistic prediction [41, 42].

Many times 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 [43]. Furthermore, coupling causal discovery to identify the causal drivers in climate models before applying deep learning algorithms improved performance and interpretability [44, 45]. 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 mitigates equifinality, and is robust to biases of training and regularization. As a first step, we propose a method based on double machine learning (DML) [40]. DML is a causal effect estimation technique developed in econometrics, where it is common to investigate the effect of some proposed treatment [46, 47]. It has recently been used for effect estimation in the environmental sciences [48]. 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 do inference, as the estimators are shown to be approximately normally distributed, which yields confidence intervals [40].

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

$$Y = \theta(X) \cdot f(T) + g(X, W), \tag{3}$$

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:

1. The temperature sensitivity Q_{10} model for ecosystem respiration [49–51] and,
2. the light-use efficiency model for carbon flux partitioning [52].

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 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, as in (**??**). 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.

2 Case studies and Data

Carbon fluxes are crucial in the global carbon cycle, a key component of the Earth's climate system. Net ecosystem exchange (NEE) is the net carbon dioxide flux measured using the eddy covariance (EC) technique [53]. The data for our studies is half-hourly data from FLUXNET, a global network of EC towers that collect data on carbon dioxide and water vapor exchange between the atmosphere and the terrestrial biosphere [54]. Different biogeochemical processes contribute to the carbon balance of the land. In particular and as common, we split NEE as

$$NEE = -GPP + R_{\text{eco}}, \tag{4}$$

where gross primary production (GPP) describes the gross carbon uptake by the environment and ecosystem respiration (RECO) the carbon release of all organisms.

2.1 The Q_{10} model

The Q_{10} model for RECO (1) has been presented in the introduction to illustrate the problem of equifinality. Following the example of [27], we use data from the EC tower in Neustift, Austria, available in the FLUXNET2015 dataset [55]. 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 Appendix A.1.1).

Ecosystem respiration is a latent flux that can only be measured under controlled conditions like a sealed chamber. It is not directly observed at flux towers during the day. At night, however, we assume GPP to be zero as no photosynthesis occurs, and all carbon flow stems from respiration. From the available years, 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).

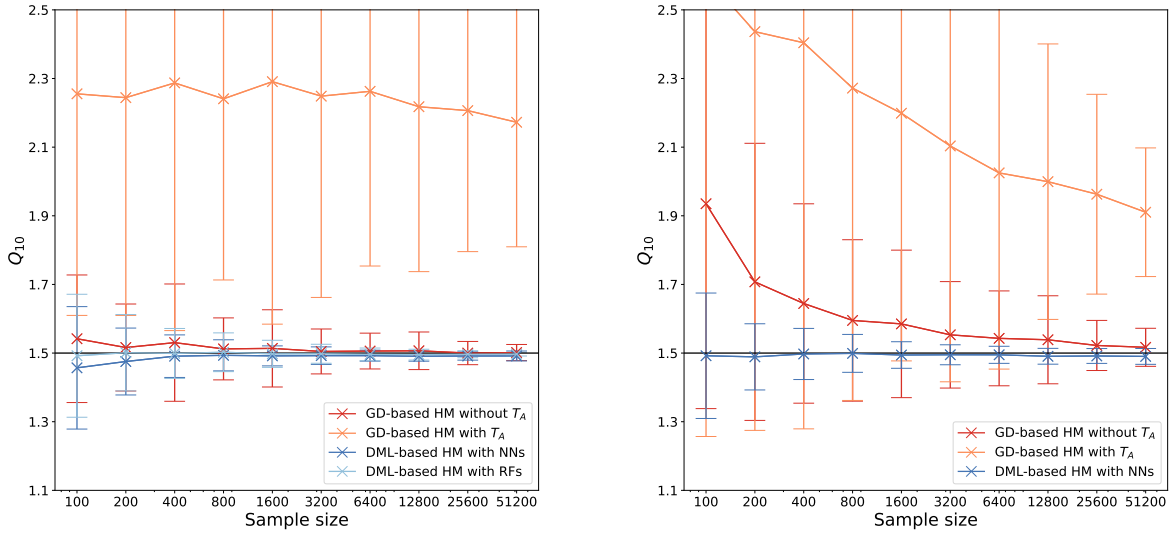
2.2 CO2 Flux partitioning

Direct measurements of GPP or RECO at the ecosystem level are impossible. 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 [56–60]. Several hybrid-modeling approaches have recently been proposed modeling both fluxes with NNs [37, 61, 62].

Separating a single signal into two additive signals is generally prone to equifinality issues. [61] 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

$$NEE = -LUE \cdot SW + R_{\text{eco}}, \tag{5}$$

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 light-use efficiency (LUE) parametrized by a NN. [37] 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.



(a) Without dropout.

(b) With dropout.

Figure 1: 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.

As a proof of concept, we evaluate the proposed method on synthetically generated observations (see Appendix C.1). We only used real observations for the real data and applied the hybrid modeling approach site-wise per year. We lay out the selection criteria for data and sites in Appendix A.2. We used data from 36 different FLUXNET2015 sites for both synthetic and real data. Details on the single sites can be found in Appendix A.3. For comparison, we use the respective partitioned R_{eco} and GPP fluxes obtained from the daytime and nighttime methods, already provided as part of the FLUXNET2015 dataset. Moreover, we compare the partitions to the results obtained with NNs from [61].

3 Double machine learning for hybrid modeling – a causal perspective

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.

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.

The Q_{10} estimation results are shown in Fig. 1. First, Fig. 1a 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 (1) 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.

4.1.2 Robustness against regularization bias.

Dropout is commonly used in deep learning for regularization [63] or uncertainty quantification [64]. Fig. 1b 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 harder 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.

4.1.3 Results on real data

As discussed in Section 2.1, we obtain measured respiration data using night-time 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. 2. 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 [65] 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 (4). 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

In real data, GPP does not follow a linear relationship with incoming radiation. Especially on a half-hourly scale, the assimilation of CO_2 saturates with increasing radiation [66]. For this reason, we transform SW before applying the DML scheme. In particular, we employ a rectangular hyperbola similar to the established daytime method for flux partitioning. The DML then retrieves a modulating factor of the LUE depending on various meteorological drivers. 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 [61] and by the established daytime and nighttime methods [56, 59]. 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. The results are reported in Table 1. Overall the consistency of the method based on DML lies in a similar range of values to the NN approach [61] 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 root-mean-square error (RMSE). Only for the GPP and NEE of the nighttime method, the values lie on average

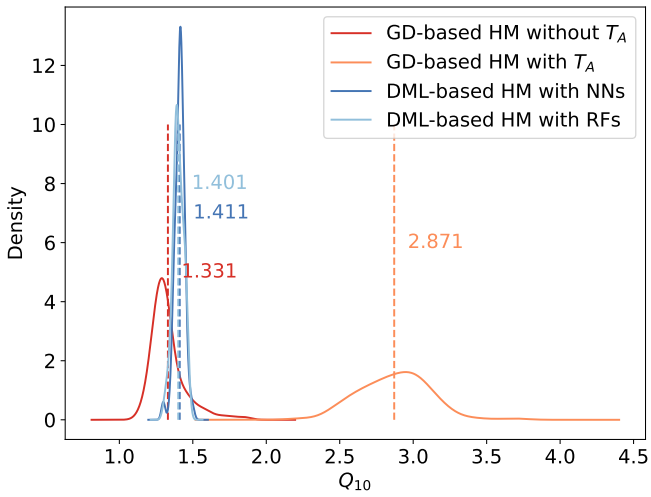


Figure 2: Estimation of Q_{10} on real data. Both DML-based HM find on average a Q_{10} value of 1.401 and 1.411 for RFs and neural networks (NNs), respectively. This agrees with values from the literature that find a Q_{10} value around 1.41 ± 0.1 [65]. The value for the GD-based HM is lower at 1.331 when leaving out T_A as a predictor. With T_A, problems of equifinality show up again.

Table 1: Cross consistency in terms of R^2, $RMSE$ and bias of retrieved GPP, RECO and estimated NEE between the established daytime and nighttime methods and the nonparametric methods based on DML. The reported statistics are median and in brackets 0.25/0.75 quantiles over all site-years.

Flux	*Methods*	R^{2*}	$RMSE^*(\frac{\mu \text{mol CO}_2}{m^2 s})$	$Bias(\frac{\mu \text{mol CO}_2}{m^2 s})$
RECO	*DT-DML*	0.62(0.41/0.74)	1.18(0.75/1.46)	0.00(−0.20/0.14)
	DT-ANN	0.69(0.50/0.81)	0.98(0.70/1.29)	0.02(−0.12/0.18)
	NT-DML	0.74(0.50/0.83)	0.89(0.57/1.15)	0.00(−0.11/0.10)
	NT-ANN	0.85(0.65/0.92)	0.68(0.47/0.84)	0.07(−0.02/0.16)
	DT-NT	0.73(0.63/0.83)	0.95(0.64/1.21)	0.00(−0.22/0.16)
	ANN-DML	0.63(0.34/0.77)	0.99(0.66/1.24)	−0.07(−0.22/0.10)
GPP	*DT-DML*	0.96(0.93/0.97)	1.25(0.74/1.49)	0.00(−0.16/0.11)
	DT-ANN	0.96(0.93/0.97)	1.22(0.76/1.52)	0.04(−0.04/0.17)
	NT-DML	0.90(0.84/0.92)	1.97(1.16/2.47)	−0.02(−0.13/0.10)
	NT-ANN	0.93(0.89/0.95)	1.53(0.90/2.02)	0.07(−0.02/0.18)
	DT-NT	0.89(0.82/0.92)	1.85(1.20/2.42)	0.02(−0.16/0.13)
	ANN-DML	0.95(0.92/0.97)	1.32(0.71/1.61)	−0.08(−0.23/0.08)
NEE	*DT-DML*	0.95(0.93/0.97)	1.07(0.71/1.29)	−0.02(−0.11/0.07)
	DT-ANN	0.94(0.91/0.96)	1.13(0.76/1.36)	−0.03(−0.12/0.03)
	NT-DML*	0.87(0.81/0.89)	1.92(1.15/2.36)	0.01(−0.02/0.06)
	NT-ANN*	0.93(0.90/0.94)	1.29(0.79/1.82)	0.00(−0.01/0.01)
	*DT-NT**	0.86(0.79/0.90)	1.68(1.12/2.25)	−0.03(−0.12/0.03)
	ANN-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.

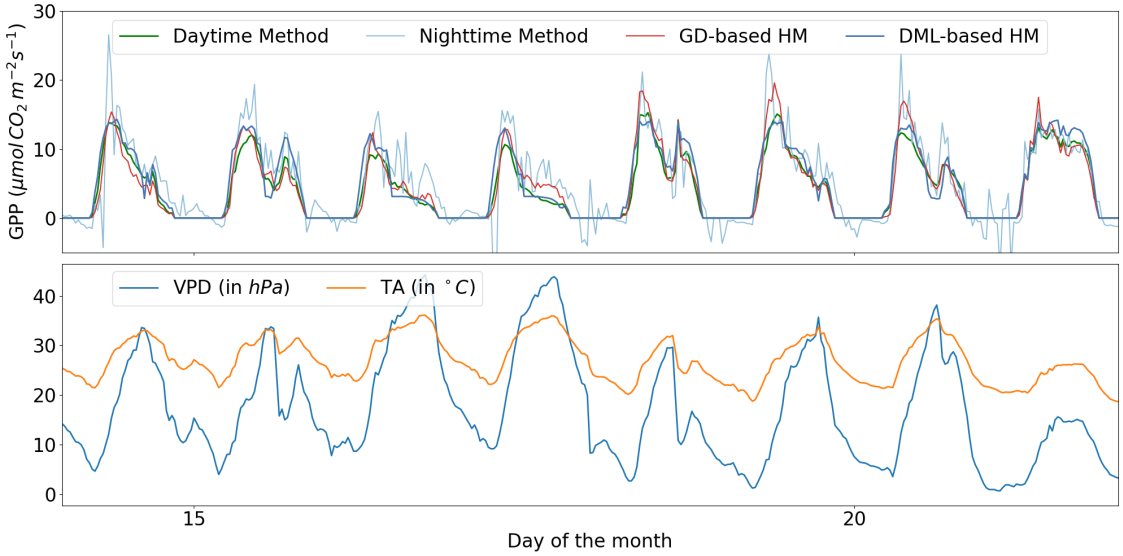


Figure 3: Retrieved GPP flux of daytime method, nighttime method and DML-based HM in July 2006 in France Le-Bray. The DML-based HM retrieved a similar flux to the daytime method that decreases with the increase of VPD.

slightly above with $1.97 \frac{\mu \,\mathrm{mol}\,\mathrm{CO_2}}{m^2 s}$ and $1.92 \frac{\mu \,\mathrm{mol}\,\mathrm{CO_2}}{m^2 s}$, respectively. The nighttime method fits respiration overnight and obtains GPP as the residuals between the estimated R_{eco} and measured 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, mostly in all cases, shows a slightly smaller bias compared to both standard methods than these methods between them. 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 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 Appendix C.1.

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. 3, 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 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 $15^\circ C$ and $30^\circ C$ (see Fig. 4). It is consistent over the four consecutive years the method was applied to 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) [67]. Fig. 5 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 [68]. 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 [68]. We find that the daytime and nighttime methods cannot capture this sudden

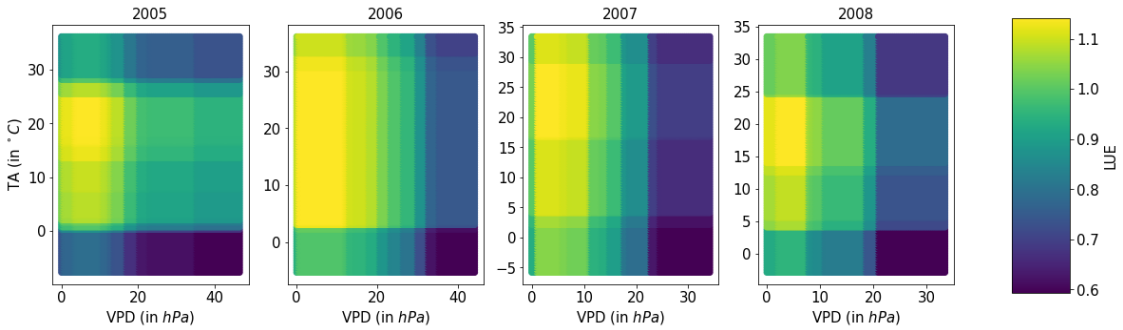


Figure 4: Functional behavior of the learned LUE in the years 2005 to 2008 over VPD and TA. The LUE shows a consistent functionality over the different years where an increase in VPD, which marks lower water availability, reduces productivity. This is also consistent with the functionality that the daytime method implements parametrically.

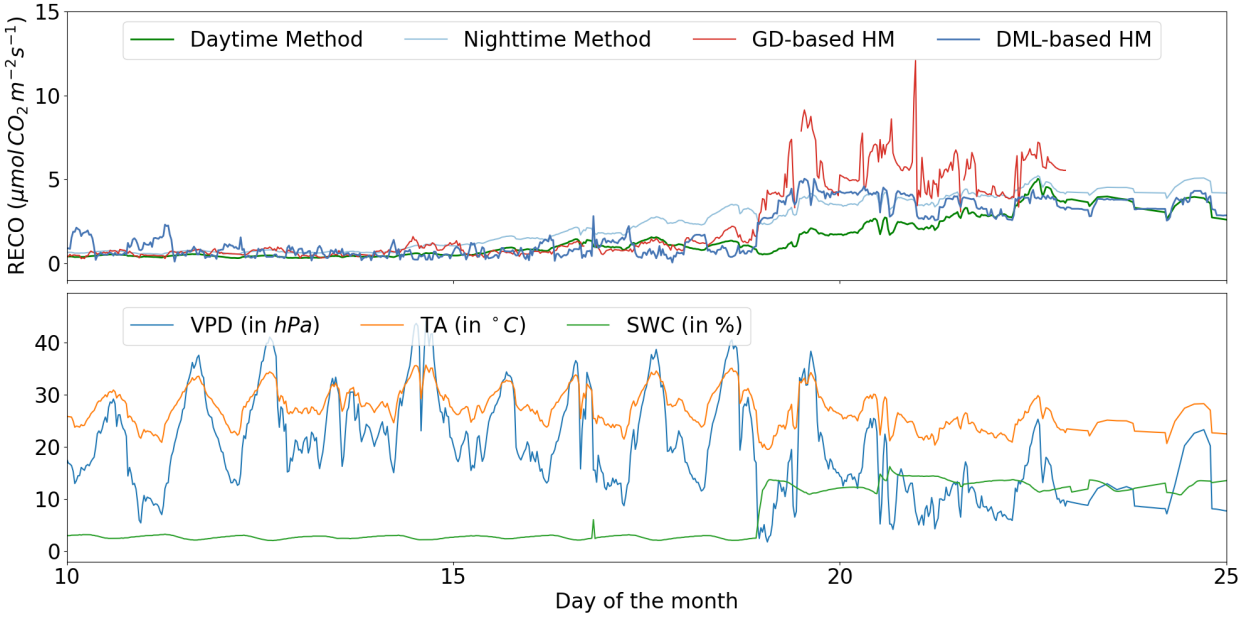


Figure 5: Retrieved R_{eco} flux of daytime, nighttime, and both hybrid modeling methods in July 2010 in Santa Rita in the US. The daytime and nighttime methods show slow adaption to the change in dynamics caused by a rain pulse event that followed a long drought. Both hybrid modeling approaches can retrieve the expected immediate increase in respiration. The estimate of the GD-based HM are lower and less noisy.

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.

5 Conclusions

Machine learning is entering all fields of science, such as physics, chemistry, biology, or environmental sciences, due to its ability to work with large and complex datasets to identify patterns and make predictions. It is becoming a complementary tool to enhance scientific research and discovery traditionally ruled by knowledge of first principles alone. Its limitations are evident: lack of transparency and interpretability, weak generalizability to unseen data, and violation of laws of nature. Domain-guided machine learning aims to incorporate expert knowledge to overcome these limitations. Hybrid modeling is one approach to do so which introduces machine learning models into scientific

equations to learn complex functions from data that we cannot derive from first principles alone. It turns out that this alone is not enough 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 proposed 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. 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. There are two main problems with the proposed method using DML. First, 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 [43]. Efforts would need to be put into parallelizing the fitting of the first-stage estimators to make this approach computationally less costly. The other 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. Even though one loses the advantageous properties of DML for causal effect estimation, it is thinkable that to reduce equifinality, one reduces predictors or flexibility based on causal knowledge to have the problem well-posed. Fit the model and fix part before introducing other predictors and increasing flexibility. This procedure we applied to introduce T_A into the base respiration R_b after estimating Q_{10} could have similarly been applied to the approach where Q_{10} was fitted using gradient descent. When possible, we still encourage using more evolved causal methods, such as DML. Combined with dropout, it even allows us to have a full probabilistic assessment of a model such as the Q_{10} model. A common technique for obtaining uncertainty estimates for NNs is dropout, which we introduced here as a regularization technique [64]. 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 normal 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.

We believe that the marriage of causality and hybrid modeling is the obvious next step toward obtaining more interpretable and trustworthy results in knowledge-guided machine learning. We have demonstrated how this can be done with causal effect estimation. We hope that more ways of enforcing causality in hybrid models can be explored in the future.

Acknowledgments

This work received support from the European Research Council (ERC) under the ERC Synergy Grant USMILE (grant agreement 855187). We express our gratitude to Gianluca Tramontana for generously providing his data and patiently answering all our queries.

References

[1] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," 2023.

[2] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu,

C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 2020.

[3] Y. Zhang, J. Qin, D. S. Park, W. Han, C.-C. Chiu, R. Pang, Q. V. Le, and Y. Wu, "Pushing the limits of semi-supervised learning for automatic speech recognition," 2022.

[4] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *IEEE Intelligent Systems*, vol. 24, no. 2, pp. 8–12, Mar. 2009. [Online]. Available: http://dx.doi.org/10.1109/MIS.2009.36

[5] Z. C. Lipton, "The mythos of model interpretability," *Queue*, vol. 16, no. 3, pp. 30:31–30:57, Jun. 2018. [Online]. Available: http://doi.acm.org/10.1145/3236386.3241340

[6] L. R. Kump, J. F. Kasting, and R. G. Crane, *The Earth System*, 3rd ed. Pearson, 2013.

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

[8] 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*, vol. 2016, pp. 1–33, 2016. [Online]. Available: http://www.biogeosciences-discuss.net/bg-2015-661/

[9] G. Camps-Valls, D. Tuia, X. X. Zhu, and M. E. Reichstein, *Deep learning for the Earth Sciences: A comprehensive approach to remote sensing, climate science and geosciences*. Wiley & Sons, 2021. [Online]. Available: https://github.com/DL4ES

[10] C. Rudin and J. 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*, vol. 1, no. 2, nov 22 2019, https://hdsr.mitpress.mit.edu/pub/f9kuryi8.

[11] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, "Dataset shift in machine learning," 2009.

[12] M. Sugiyama and M. Kawanabe, *Learning Under Covariate Shift*, 2012, pp. 19–19.

[13] 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*, vol. 566, pp. 195–204, Feb 2019.

[14] G. Marcus, "Deep learning: A critical appraisal," *arXiv preprint arXiv:1801.00631*, 2018.

[15] 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*. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press, 2021, vol. In Press. [Online]. Available: https://doi.org/10.1017/9781009157896

[16] R. Roscher, B. Bohn, M. Duarte, and J. Garcke, "Explainable machine learning for scientific insights and discoveries," *IEEE Access*, vol. PP, pp. 1–1, 02 2020.

[17] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, 2021. [Online]. Available: https://www.mdpi.com/1099-4300/23/1/18

[18] B. Neyshabur, S. Bhojanapalli, D. Mcallester, and N. Srebro, "Exploring generalization in deep learning," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/10ce03a1ed01077e3e289f3e53c72813-Paper.pdf

[19] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and P. Yu, "Generalizing to unseen domains: A survey on domain generalization," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2022.

[20] X. Shen and N. Meinshausen, "Engression: Extrapolation for nonlinear regression?" 2023.

[21] E. de Bezenac, A. Pajot, and P. Gallinari, "Deep learning for physical processes: Incorporating prior scientific knowledge," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=By4HsfWAZ

[22] G. Camps-Valls, D. Svendsen, L. Martino, J. Muñoz-Marí, V. Laparra, M. Campos-Taberner, and D. Luengo, "Physics-aware Gaussian processes in remote sensing," *Applied Soft Computing*, vol. 68, pp. 69–82, Jul 2018.

[23] J. Cortés-Andrés, G. Camps-Valls, S. Sippel, E. Székely, D. Sejdinovic, E. Diaz, A. Pérez-Suay, Z. Li, M. Mahecha, and M. Reichstein, "Physics-aware nonparametric regression models for Earth data analysis," *Environmental Research Letters*, vol. 17, no. 5, 2022.

[24] A. Karpatne, R. Kannan, and V. Kumar, *Knowledge Guided Machine Learning: Accelerating Discovery using Scientific Knowledge and Data*, 1st ed. Chapman and Hall/CRC, 2022.

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

[26] W. L. Zhao, P. Gentine, M. Reichstein, Y. Zhang, S. Zhou, Y. Wen, C. Lin, X. Li, and G. Y. Qiu, "Physics-constrained machine learning of evapotranspiration," *Geophysical Research Letters*, vol. 46, no. 24, pp. 14 496–14 507, 2019.

[27] M. Reichstein, B. Ahrens, B. Kraft, G. Camps-Valls, N. Carvalhais, F. Gans, P. Gentine, and A. Winkler, "Combining system modeling and machine learning into hybrid ecosystem modeling," in *Knowledge-Guided Machine Learning*, 2022.

[28] A. Koppa, D. Rains, P. Hulsman, R. Poyatos, and D. G. Miralles, "A deep learning-based hybrid model of global terrestrial evaporation," *Nature Communications*, vol. 13, no. 1, p. 1912, Apr 2022. [Online]. Available: https://doi.org/10.1038/s41467-022-29543-7

[29] J. Oberpriller, D. R. Cameron, M. C. Dietze, and F. Hartig, "Towards robust statistical inference for complex computer models," *Ecology Letters*, vol. 24, no. 6, pp. 1251–1261, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/ele.13728

[30] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, V. Makarenkov, and S. Nahavandi, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information Fusion*, vol. 76, pp. 243–297, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253521001081

[31] P. Izmailov, S. Vikram, M. D. Hoffman, and A. G. Wilson, "What are bayesian neural network posteriors really like?" in *International Conference on Machine Learning*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:233443782

[32] M. Kuhn and K. Johnson, *Applied Predictive Modeling*, 01 2013.

[33] M. U. Kirschbaum, "Will changes in soil organic carbon act as a positive or negative feedback on global warming?" *Biogeochemistry*, vol. 48, pp. 21–51, 2000.

[34] N. G. Smith and J. S. Dukes, "Plant respiration and photosynthesis in global-scale models: incorporating acclimation to temperature and CO2," *Global change biology*, vol. 19, no. 1, pp. 45–63, 2013.

[35] C. Huntingford, O. K. Atkin, A. Martinez-De La Torre, L. M. Mercado, M. A. Heskel, A. B. Harper, K. J. Bloomfield, O. S. Osullivan, P. B. Reich, K. R. Wythers *et al.*, "Implications of improved representations of plant respiration in a changing climate," *Nature Communications*, vol. 8, no. 1, p. 1602, 2017.

[36] G. Vardi, "On the implicit bias in deep-learning algorithms," *Commun. ACM*, vol. 66, no. 6, pp. 86–93, may 2023. [Online]. Available: https://doi.org/10.1145/3571070

[37] W. Zhan, X. Yang, Y. Ryu, B. Dechant, Y. Huang, Y. Goulas, M. Kang, and P. 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*, vol. 321, p. 108980, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168192322001708

[38] R. ElGhawi, B. Kraft, C. Reimers, M. Reichstein, M. Körner, P. Gentine, and A. J. Winkler, "Hybrid modeling of evapotranspiration: inferring stomatal and aerodynamic resistances using combined physics-based and machine learning," *Environmental Research Letters*, vol. 18, no. 3, p. 034039, mar 2023. [Online]. Available: https://dx.doi.org/10.1088/1748-9326/acbbe0

[39] Y. Yin, V. L. Guen, J. Dona, E. de Bézenac, I. Ayed, N. Thome, and P. Gallinari, "Augmenting physical models with deep networks for complex dynamics forecasting*," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2021, no. 12, p. 124012, dec 2021. [Online]. Available: https://dx.doi.org/10.1088/1742-5468/ac3ae5

[40] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins, "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, vol. 21, no. 1, pp. C1–C68, 01 2018. [Online]. Available: https://doi.org/10.1111/ectj.12097

[41] J. Pearl, *Causality: Models, Reasoning and Inference*, 2nd ed. USA: Cambridge University Press, 2009.

[42] J. Runge, A. Gerhardus, G. Varando, V. Eyring, and G. Camps-Valls, "Causal inference for time series," *Nature Reviews Earth & Environment*, vol. 10, p. 2553, 2023.

[43] S. Wang, S. Sankaran, and P. Perdikaris, "Respecting causality is all you need for training physics-informed neural networks," 2022.

[44] F. Iglesias-Suarez, P. Gentine, B. Solino-Fernandez, T. Beucler, M. Pritchard, J. Runge, and V. Eyring, "Causally-informed deep learning to improve climate models and projections," 2023.

[45] 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*, no. 2553, pp. 1–13, 2019.

[46] M. C. Knaus, M. Lechner, and A. Strittmatter, "Heterogeneous employment effects of job search programs," *Journal of Human Resources*, vol. 57, no. 2, pp. 597–636, mar 2020. [Online]. Available: https://doi.org/10.3368%2Fjhr.57.2.0718-9615r1

[47] J. M. Davis and S. B. Heller, "Using causal forests to predict treatment heterogeneity: An application to summer jobs," *The American Economic Review*, vol. 107, no. 5, pp. 546–550, 2017. [Online]. Available: http://www.jstor.org/stable/44250458

[48] Q. Sun, T. Zheng, X. Zheng, M. Cao, B. Zhang, and S. Jiang, "Causal interpretation for groundwater exploitation strategy in a coastal aquifer," *Science of The Total Environment*, vol. 867, p. 161443, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S004896972300058X

[49] S. Arrhenius, "Über die reaktionsgeschwindigkeit bei der inversion von rohrzucker durch säuren," *Zeitschrift für physikalische Chemie*, vol. 4, no. 1, pp. 226–248, 1889.

[50] J. H. Van't Hoff, R. A. Lehfeldt *et al.*, "Lectures on theoretical and physical chemistry," 1899.

[51] J. Lloyd and J. Taylor, "On the temperature dependence of soil respiration," *Functional ecology*, pp. 315–323, 1994.

[52] Y. Pei, J. Dong, Y. Zhang, W. Yuan, R. Doughty, J. Yang, D. Zhou, L. Zhang, and X. Xiao, "Evolution of light use efficiency models: Improvement, uncertainties, and implications," *Agricultural and Forest Meteorology*, vol. 317, p. 108905, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168192322000983

[53] G. 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.

[54] 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. 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*, vol. 82, no. 11, pp. 2415–2434, 2001.

[55] G. Pastorello, C. Trotta, E. Canfora, H. Chu, D. Christianson, Y.-W. Cheah, C. Poindexter, J. Chen, A. Elbashandy, M. Humphrey, P. Isaac, D. Polidori, M. Reichstein, A. Ribeca, C. van Ingen, N. Vuichard, L. Zhang, B. Amiro, C. Ammann, M. A. Arain, J. Ardö, T. Arkebauer, S. K. Arndt, N. Arriga, M. Aubinet, M. Aurela, D. Baldocchi, A. Barr, E. Beamesderfer, L. B. Marchesini, O. Bergeron, J. Beringer, C. Bernhofer, D. Berveiller, D. Billesbach, T. A. Black, P. D. Blanken, G. Bohrer, J. Boike, P. V. Bolstad, D. Bonal, J.-M. Bonnefond, D. R. Bowling, R. Bracho, J. Brodeur, C. Brümmer, N. Buchmann, B. Burban, S. P. Burns, P. Buysse, P. Cale, M. Cavagna, P. Cellier, S. Chen, I. Chini, T. R. Christensen, J. Cleverly, A. Collalti, C. Consalvo, B. D. Cook, D. Cook, C. Coursolle, E. Cremonese, P. S. Curtis, E. D'Andrea, H. da Rocha, X. Dai, K. J. Davis, B. D. Cinti, A. d. Grandcourt, A. D. Ligne, R. C. De Oliveira, N. Delpierre, A. R. Desai, C. M. Di Bella, P. d. Tommasi, H. Dolman, F. Domingo, G. Dong, S. Dore, P. Duce, E. Dufrêne, A. Dunn, J. Dušek, D. Eamus, U. Eichelmann, H. A. M. ElKhidir, W. Eugster, C. M. Ewenz, B. Ewers, D. Famulari, S. Fares, I. Feigenwinter, A. Feitz, R. Fensholt, G. Filippa, M. Fischer, J. Frank, M. Galvagno, M. Gharun, D. Gianelle, B. Gielen, B. Gioli, A. Gitelson, I. Goded, M. Goeckede, A. H. Goldstein, C. M. Gough, M. L. Goulden, A. Graf, A. Griebel, C. Gruening, T. Grünwald, A. Hammerle, S. Han, X. Han, B. U. Hansen, C. Hanson, J. Hatakka, Y. He, M. Hehn, B. Heinesch, N. Hinko-Najera, L. Hörtnagl, L. Hutley, A. Ibrom, H. Ikawa, M. Jackowicz-Korczynski, D. Janouš, W. Jans, R. Jassal, S. Jiang, T. Kato, M. Khomik, J. Klatt, A. Knohl, S. Knox, H. Kobayashi, G. Koerber, O. Kolle, Y. Kosugi, A. Kotani, A. Kowalski, B. Kruijt, J. Kurbatova, W. L. Kutsch, H. Kwon, S. Launiainen, T. Laurila, B. Law, R. Leuning, Y. Li, M. Liddell, J.-M. Limousin, M. Lion, A. J. Liska, A. Lohila, A. López-Ballesteros, E. López-Blanco, B. Loubet, D. Loustau, A. Lucas-Moffat, J. Lüers, S. Ma, C. Macfarlane, V. Magliulo, R. Maier, I. Mammarella, G. Manca, B. Marcolla, H. A. Margolis, S. Marras, W. Massman, M. Mastepanov, R. Matamala, J. H. Matthes, F. Mazzenga, H. McCaughey, I. McHugh, A. M. S. McMillan, L. Merbold, W. Meyer, T. Meyers, S. D. Miller, S. Minerbi, U. Moderow, R. K. Monson, L. Montagnani, C. E. Moore, E. Moors, V. Moreaux, C. Moureaux, J. W. Munger, T. Nakai, J. Neirynck, Z. Nesic, G. Nicolini, A. Noormets, M. Northwood,

M. Nosetto, Y. Nouvellon, K. Novick, W. Oechel, J. E. Olesen, J.-M. Ourcival, S. A. Papuga, F.-J. Parmentier, E. Paul-Limoges, M. Pavelka, M. Peichl, E. Pendall, R. P. Phillips, K. Pilegaard, N. Pirk, G. Posse, T. Powell, H. Prasse, S. M. Prober, S. Rambal, Ü. Rannik, N. Raz-Yaseef, C. Rebmann, D. Reed, V. R. d. Dios, N. Restrepo-Coupe, B. R. Reverter, M. Roland, S. Sabbatini, T. Sachs, S. R. Saleska, E. P. Sánchez-Cañete, Z. M. Sanchez-Mejia, H. P. Schmid, M. Schmidt, K. Schneider, F. Schrader, I. Schroder, R. L. Scott, P. Sedlák, P. Serrano-Ortíz, C. Shao, P. Shi, I. Shironya, L. Siebicke, L. Šigut, R. Silberstein, C. Sirca, D. Spano, R. Steinbrecher, R. M. Stevens, C. Sturtevant, A. Suyker, T. Tagesson, S. Takanashi, Y. Tang, N. Tapper, J. Thom, M. Tomassucci, J.-P. Tuovinen, S. Urbanski, R. Valentini, M. van der Molen, E. van Gorsel, K. van Huissteden, A. Varlagin, J. Verfaillie, T. Vesala, C. Vincke, D. Vitale, N. Vygodskaya, J. P. Walker, E. Walter-Shea, H. Wang, R. Weber, S. Westermann, C. Wille, S. Wofsy, G. Wohlfahrt, S. Wolf, W. Woodgate, Y. Li, R. Zampedri, J. Zhang, G. Zhou, D. Zona, D. Agarwal, S. Biraud, M. Torn, and D. Papale, "The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data," *Scientific Data*, vol. 7, no. 1, p. 225, Jul. 2020.

[56] M. Reichstein, E. Falge, D. Baldocchi, D. Papale, M. Aubinet, P. Berbigier, C. Bernhofer, N. Buchmann, T. Gilmanov, A. Granier, T. Grünwald, K. Havránková, H. Ilvesniemi, D. Janous, A. Knohl, T. Laurila, A. Lohila, D. Loustau, G. Matteucci, T. Meyers, F. Miglietta, J.-M. Ourcival, J. Pumpanen, S. Rambal, E. Rotenberg, M. Sanz, J. Tenhunen, G. Seufert, F. Vaccari, T. Vesala, D. Yakir, and R. Valentini, "On the separation of net ecosystem exchange into assimilation and ecosystem respiration: review and improved algorithm," *Global Change Biology*, vol. 11, no. 9, pp. 1424–1439, 2005. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2486.2005.001002.x

[57] A. M. Moffat, D. Papale, M. Reichstein, D. Y. Hollinger, A. D. Richardson, A. G. Barr, C. Beckstein, B. H. Braswell, G. Churkina, A. R. Desai, E. Falge, J. H. Gove, M. Heimann, D. Hui, A. J. Jarvis, J. Kattge, A. Noormets, and V. J. Stauch, "Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes," *Agricultural and Forest Meteorology*, vol. 147, no. 3, pp. 209–232, 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016819230700216X

[58] A. R. Desai, A. D. Richardson, A. M. Moffat, J. Kattge, D. Hollinger, A. G. Barr, E. Falge, A. Noormets, D. Papale, M. Reichstein, and V. J. Stauch, "Cross-site evaluation of eddy covariance GPP and RE decomposition techniques," *Agricultural and Forest Meteorology*, vol. 148, pp. 821–838, 2008.

[59] G. Lasslop, M. Reichstein, D. Papale, A. D. Richardson, A. Arneth, A. Barr, P. Stoy, and G. Wohlfahrt, "Separation of net ecosystem exchange into assimilation and respiration using a light response curve approach: critical issues and global evaluation," *Global Change Biology*, vol. 16, no. 1, pp. 187–208, 2010. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2486.2009.02041.x

[60] T. F. Keenan, M. Migliavacca, D. Papale, D. Baldocchi, M. Reichstein, M. Torn, and T. Wutzler, "Widespread inhibition of daytime ecosystem respiration," *Nature Ecology & Evolution*, vol. 3, no. 3, pp. 407–415, Mar 2019. [Online]. Available: https://doi.org/10.1038/s41559-019-0809-2

[61] G. Tramontana, M. Migliavacca, M. Jung, M. Reichstein, T. F. Keenan, G. Camps-Valls, J. Ogee, J. Verrelst, and D. Papale, "Partitioning net carbon dioxide fluxes into photosynthesis and respiration using neural networks," *Global Change Biology*, vol. 26, no. 9, pp. 5235–5253, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/gcb.15203

[62] V. T. Trifunov, M. Shadaydeh, J. Runge, M. Reichstein, and J. Denzler, "A data-driven approach to partitioning net ecosystem exchange using a deep state space model," *IEEE Access*, vol. 9, pp. 107 873–107 883, 2021.

[63] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014. [Online]. Available: http://jmlr.org/papers/v15/srivastava14a.html

[64] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 1050–1059. [Online]. Available: https://proceedings.mlr.press/v48/gal16.html

[65] M. D. Mahecha, M. Reichstein, N. Carvalhais, G. Lasslop, H. Lange, S. I. Seneviratne, R. Vargas, C. Ammann, M. A. Arain, A. Cescatti, I. A. Janssens, M. Migliavacca, L. Montagnani, and A. D. Richardson, "Global convergence in the temperature sensitivity of respiration at ecosystem level," *Science*, vol. 329, no. 5993, pp. 838–840, 2010. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.1189587

[66] E. Falge, D. Baldocchi, R. Olson, P. Anthoni, M. Aubinet, C. Bernhofer, G. Burba, R. Ceulemans, R. Clement, H. Dolman, A. Granier, P. Gross, T. Grünwald, D. Hollinger, N.-O. Jensen, G. Katul, P. Keronen, A. Kowalski, C. T. Lai, B. E. Law, T. Meyers, J. Moncrieff, E. Moors, J. Munger, K. Pilegaard, Ü. Rannik, C. Rebmann, A. Suyker, J. Tenhunen, K. Tu, S. Verma, T. Vesala, K. Wilson, and S. Wofsy, "Gap filling strategies for

defensible annual sums of net ecosystem exchange," *Agricultural and Forest Meteorology*, vol. 107, no. 1, pp. 43–69, 2001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168192300002252

[67] R. L. Scott, J. A. Biederman, E. P. Hamerlynck, and G. 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*, vol. 120, no. 12, pp. 2612–2624, 2015. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015JG003181

[68] F. S. Chapin, P. A. Matson, and H. A. Mooney, *Principles of terrestrial ecosystem ecology*, 2002nd ed. New York, NY: Springer, May 2013.

[69] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.

[70] S. Athey, J. Tibshirani, and S. Wager, "Generalized random forests," *The Annals of Statistics*, vol. 47, no. 2, pp. 1148 – 1178, 2019. [Online]. Available: https://doi.org/10.1214/18-AOS1709

[71] D. J. Foster and V. Syrgkanis, "Orthogonal statistical learning," 2020.

A Data

A.1 Synthetic data

A.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

$$Q_{10} = 1.5, \tag{6}$$
$$R_{\text{eco}}^{\text{syn}} = R_b^{\text{syn}} \cdot Q_{10}^{0.1 \cdot (T_A - 15)} \cdot (1 + \epsilon), \tag{7}$$
$$R_b^{\text{syn}} = 0.75 \cdot (\tilde{R}_b^{syn} - \min(\tilde{R}_b^{syn}) + 0.1 \cdot \pi), \tag{8}$$
$$\tilde{R}_b^{syn} = 0.01 \cdot SW_{\text{POT}}^{\text{SM}} - 0.005 \cdot SW_{\text{POT}}^{\text{SM,diff}}, \tag{9}$$

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 (8) 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.

A.1.2 LUE model

The code for generating the data is taken from the work of [27], 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:

$$GPP^{syn} = LUE^{syn} \cdot SW_{in}, \tag{10}$$
$$LUE^{syn} = 0.5 \cdot \exp\left(-0.1 \cdot (T_A - 20)\right)^2) \cdot \min(1, \exp(-0.1 \cdot (VPD - 10))). \tag{11}$$

Finally, we compute NEE following (4) with additional multiplicative heteroscedastic noise:

$$NEE^{syn} = (-GPP^{syn} + R_{\text{eco}}^{\text{syn}}) \cdot (1 + \sigma\varepsilon), \tag{12}$$

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

A.2 Fluxnet data selection

For the data selection of real data from FLUXNET2015 [55], we closely followed [61] to compare our method to the neural network approach that imposes similar structural equations. We choose the same set of sites and use the same quality criterion to select site-years. 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 [61], we use the NEE obtained from the 50th percentile of the CUT method [55].

A.3 FLUXNET sites

The 36 FLUXNET sites used for the flux partitioning experiments are shown in Table 2. The table further provides information on plant type, latitude, and longitude.

Table 2: FLUXNET sites used for flux partitioning experiments with DML.

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

B Method

B.1 Derivation of DML estimator for g

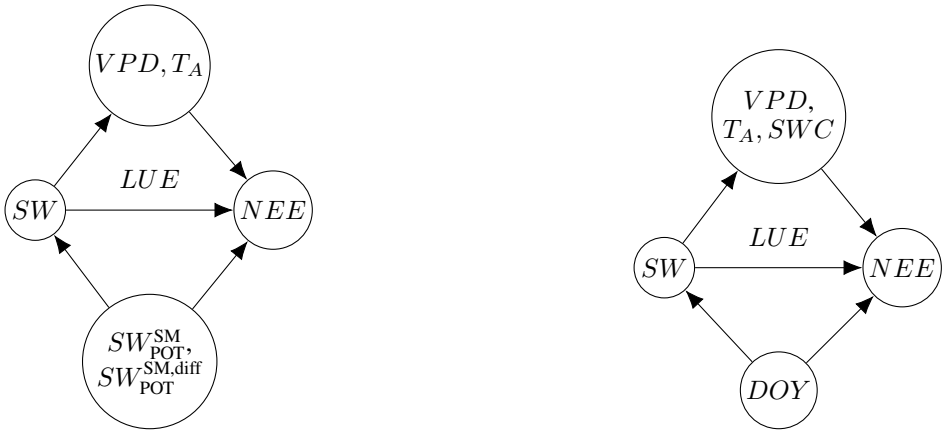
One way of obtaining an estimator for g instead of fitting it directly is by reusing all estimators of DML. It is easy to see that

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

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.

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

Figure 6: Assumed causal graphs for the estimation with the causal hybrid modeling approach in Q_{10} estimation.



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

Figure 7: Assumed causal graphs for the estimation with the causal hybrid modeling approach in flux partitioning.

B.2 Causal graph of the Q_{10} model

The causal graph we assume for the Q_{10} model is shown in Fig. 6. The smooth potential radiation cycle given by $SW^{\text{SM}}_{\text{POT}}$ and $SW^{\text{SM,diff}}_{\text{POT}}$ represent seasonality, and thus, they have 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. Furthermore, water availability also has a strong effect on respiration [68]. However, the temperature-sensitivity Q_{10} should only describe the immediate temperature effect. Effects of water should be modeled 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.

B.3 Causal graph of the LUE model

The causal graphs assumed for the LUE model are shown in Fig. 7. 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. Thus, both variables enter as mediators on the path to NEE. For the real data, we use the day of the year DOY for modeling 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 still enter as modifiers on the effect of light on NEE, affecting GPP.

Table 3: 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)

B.4 Details on the neural networks

The NNs used for the GD-based HM had 2 hidden layers with 16 units each. A $\tanh$ nonlinearity was applied at the end of each hidden layer. To obtain non-negative results for the base respiration, a final softplus function was applied to the output of the last layer. 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. The initial Q_{10} is sampled from a Gaussian with $\sigma = 0.1$ and $\mu = 1.5$. 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 [69] 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.

C Additional results

C.1 Retrievel of linear model

We generated synthetic data following [27], a partially linear LUE model with varying coefficients. As inputs, we used time series of measured meteorological forcings and added heteroscedastic noise over different noise levels (see Appendix A.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 3 and Table 4. We note that the DML approach gives theoretical guarantees for estimating GPP and not necessarily for R_{eco} [70, 71]. Our proposed method retrieves good 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 strong 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 harder 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 good estimates of the NEE signal.

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.

Table 4: 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)