# Towards causal predictions of site-specific management effects in applied ecology

**Eleanor E. Jackson[1,2], Tord Snäll[3,4], Emma Gardner[5], James M. Bullock[5] & Rebecca Spake[1,6]**

## Abstract

With limited resources and the urgent need to reverse biodiversity loss, conservation efforts must be targeted to where they will be most effective. Targeting actions necessitates new approaches to causal prediction of sited-level responses to alternative interventions. We present the first application of 'meta-learner algorithms' to predict 'individual treatment effects' (ITEs) representing the effects of site-level management actions. We compare the performance of three algorithms that differ in how they handle selection biases typical to observational data: S-, T-, and X-Learners, across 4,050 virtual studies predicting the effect of forest management on soil carbon, the ITEs. The X-learner, an algorithm which adjusts for selection bias, consistently yielded the most accurate ITE predictions across studies varying in sample size and imbalanced sample sizes of treatment and control groups. Our study illustrates how ecologists can begin to select and apply causal prediction methods to inform targeted conservation action for ecological systems, and makes suggestions for further road-testing of these approaches.

## Keywords

conditional average treatment effect, national forest inventory, treatment effect heterogeneity, effect modification, uplift modelling

1 School of Biological Sciences, University of Reading, UK
2 Department of Biology, University of Oxford, UK
3 Skogforsk, Uppsala Science Park, Uppsala, Sweden
4 SLU Swedish Species Information Centre, Sweden
5 UK Centre for Ecology & Hydrology, Wallingford, UK
6 School of Geography and Environmental Science, University of Southampton, UK

**Corresponding Author:**
Eleanor E. Jackson, Department of Biology, University of Oxford, Life & Mind Building, South Parks Road, Oxford OX1 3EL UK. Email: eleanor.jackson@biology.ox.ac.uk

## Introduction

Tackling climate change and biodiversity loss requires effective conservation and restoration globally (IPBES 2019). However, on-the-ground action is often undermined by context dependency; variation in responses to interventions across individual sites due to complex interactions among socio-environmental drivers (Spake et al. 2019). Scientists, policymakers and practitioners have long recognised the need to target interventions to sites where they will be most effective, and to develop place-based action plans tailored to individual sites. For example, while tree planting has the potential to sequester vast amounts of carbon dioxide and restore biodiversity, outcomes depend strongly on both where and how trees are planted (Holl & Brancalion 2020; Moyano et al. 2024). Targeted and tailored management requires causal predictions of how individual sites will respond to alternative interventions; 'treatments' from hereon. It is argued that applied ecologists do not currently generate valid causal predictions of treatment effects at the level of individual sites (Spake et al. 2025). Instead, ecologists tend to estimate average treatment effects across all sites, or coarse subgroupings of sites defined by e.g., biome or ecosystem type. However, when treatment effects vary in magnitude and direction across sites within coarse groupings, average treatment effects are neither actionable nor desirable (Spake et al. 2022).

In contrast to ecology, human-centred disciplines such as medicine, econometrics and marketing generate individual-level causal predictions (Tipton & Mamakos 2023). These disciplines are undergoing a 'causal revolution' (Pearl 2018), and have pioneered individual treatment effect (ITE) prediction by exploiting large datasets, causal inference and computational advances to underpin personalised medicine and targeted marketing. A range of ITE prediction methods, known as meta-learners, have been developed (Box 1), differing in how they adjust for biases inherent to observational datasets, (i.e., where assignment to treatment is non-random). Spake et al. (2025) recently argued that novel methods for individualised prediction could have potential for improving predictions of treatment effects in applied ecology. However, we first need to understand how the accuracy of individualised predictions can vary in the face of biases inherent to ecological datasets. While the accuracy of average causal estimates rests on well-known, strong assumptions satisfied through either sampling design and/or statistical adjustment, predictive accuracy of individual quantities relies on factors such as covariate coverage. The performance accuracy of ITE prediction approaches are virtually unknown for ecology, and there is a need to identify how different approaches perform when confronted with the realities of ecological observational data (Spake et al. 2025).

Observational data in ecology are inherently susceptible to different forms of sample-selection bias that have potential to violate the causal assumptions on which ITE prediction rests. Given that ITEs defy direct observation (since we can never truly observe the outcomes for an individual unit under multiple treatments (Holland 1986)), simulation using synthetic data, where potential outcomes under treatment and control are known, is essential for evaluating meta-learner accuracy (Curth et al. 2021; Spake et al. 2025). Simulations have recently been used to assess the predictive performance of meta-learners across data scenarios typical of human-centred observational studies, varying parameters such as sample size, selection bias, and sparsity in covariate space (Künzel et al. 2019; Okasa 2022). Importantly, no single meta-learner consistently outperforms others across all conditions (Knaus et al. 2021; Okasa 2022).

However, more complex types of meta-learner algorithms that model both outcomes and treatment assignment generally yield more accurate predictions, especially with large datasets (Künzel et al. 2019). The performance of meta-learners on ecological data remains untested.

Here, for the first time, we compare the performance of different meta-learner algorithms for an ecological dataset. We take a 'virtual ecologist' approach (Zurell et al. 2010) and simulate the process of applying a treatment to several sites, collecting data and predicting ITEs. Specifically, we simulate soil carbon mass (our outcome variable) for Swedish National Forest Inventory plots (sites) under different management regimes (treatments), predicting 20 five-year time steps into the future. We systematically vary data properties to mimic conditions that are commonly encountered in observational studies in ecology, to gain understanding of how treatment assignment, sampling processes for both training and test datasets, and choice of meta-learners interact to influence ITE accuracy within real-world situations. We expect that: meta-learners which explicitly model both treatment assignment and outcomes will outperform simpler approaches (hypothesis 1), and that learners that model covariate differences among treatment groups will perform better when selection biases are present in observational datasets (hypothesis 2). However, we expect the accuracy improvement to diminish at smaller sample sizes (hypothesis 3), where the additional modelling may not be supported by the available data. We are aiming to incentivise ecologists to consider meta-learner approaches for targeted conservation action.

Our outcome variable of interest was soil organic carbon (hereafter, 'soil carbon'), due to its central importance in forest management. Maintaining or increasing soil carbon has many benefits for climate change mitigation, adaptation and biodiversity conservation (Mayer et al. 2020). Because forest management practices can influence soil carbon stocks, it is widely recognised that soil carbon should be an explicit consideration in forest planning (Mayer et al. 2020; Mazziotta et al. 2022a). A large body of empirical research has used large-scale datasets with plot-level measurements of soil carbon to statistically model its variation in response to interacting environmental drivers (e.g., climate, topography, soil properties) and management-related variables (e.g., canopy-dominant species, age class, and land-use history), aiming to predict how soil carbon responds to different management regimes under current and future climate scenarios (e.g., Chen et al. 2022; Lee et al. 2020; Liu et al. 2023; Mazziotta et al. 2022a; Vayreda et al. 2012). However, these studies do not measure treatment effects as they have relied on outcome prediction models that do not adjust for biases inherent to observational datasets (akin to the S-Learner approach).

## Methods

We implemented a virtual ecology approach that: i) simulated local forest dynamics under two alternative management interventions that were assigned in various ways to forest plots across a large, environmentally heterogeneous extent (treatment assignment mechanism). These forest plots were then: ii) sampled in different ways by virtual observers to generate alternative 'training' datasets (sampling conditions); and iii) subjected to different statistical modelling decisions to predict individual treatment effects among these alternative 'test' datasets (modelling conditions). The simulated data is our "truth" - observed soil carbon mass for individual sites both with and without a treatment. We “virtually” applied the treatment, sampled the population at simulation time step 20 (year 2100), and used the meta-learner approach to predict ITEs. We implemented different conditions on each virtual study and compared our ITE predictions with the true ITEs to gain a systematic understanding of how treatment assignment, sampling, and modelling conditions (Table 1) interact to influence ITE accuracy within real-world situations.

### *Study system and dataset: soil carbon response to alternative management interventions in Swedish forests*

We used existing model outputs describing forest stand dynamics for National Forest Inventory (NFI) plots across Sweden, simulated using the Heureka simulation system (Wikström *et al.* 2011). Heureka is widely used in both forestry and research to forecast the outcomes of alternative management scenarios across Sweden (Mazziotta *et al.* 2022b; Moor *et al.* 2022). It comprises a set of empirical growth and yield models that simulate stand development in five-year time steps, combining species-specific models of tree establishment, growth and mortality. Heureka models soil carbon across cohorts using a dynamic decomposition Q-model, representing the mass loss of litter over time for different litter compartments (e.g., coarse and fine roots, branches, stems and leaves). The theoretical framework is presented in Ågren & Bosatta (1998) and the general implementation is described in Ågren & Hyvönen (2003). Soil carbon predictions from Heureka have been validated with empirical data in Sweden (Ågren & Hyvönen 2003; Hyvönen *et al.* 2002) and Finland (Peltoniemi *et al.* 2004).

We used simulations of forest dynamics under two management levels for the period 2010–2100, which were: set-aside (‘control’ from hereon), where stands were left to develop without any intervention, and ‘business as usual’ even-aged forestry but excluding thinning to focus on the final clearcutting event (‘treatment’ from hereon). These latter stands were clearcut and replanted at typical clearcutting stand stockings with stand ages ranging 60-120 years, lower in southern, high-productivity stands and higher in northern, low-productivity stands. Simulations had been initialised with observed NFI input data recorded on 26,193 circular plots with a radius of 10 m from 2016 to 2020, representing the 23.5 Mha of productive forest land in Sweden. We restricted Heureka’s simulation output to NFI plots situated in pine-dominated stands (≥50% standing volume) that were between 30 and 50 years of age in 2016-2020 in order to limit variation in historical management. We removed plots containing peatland (as recorded in the NFI) and plots with high soil moisture (soil moist code four and five as recorded in the NFI) due to their low prevalence, yielding 2,580 plots, and assumed that climate averages from the period 1983–1992 remained constant for the whole simulation period, i.e., the standard setting of Heureka and assuming no climate change over the period. Since the application of the clearcutting treatment occurred at different time periods for different plots, we limited our study to only include plots with a single “peak” in maximum soil carbon - indicating they were only clearcut once during the simulation (yielding 1,806 plots). To achieve this, we filtered for plots that had reached their maximum soil carbon by the 12th time step since the start of the simulation. We did this because data on time since a management action are not usually available in observational studies employing NFI datasets. Soil carbon simulations for NFI plots under control and treatment conditions for each time step are shown in SI Figure 1.

## Box 1 Meta-learner algorithms for predicting Individual Treatment Effects (ITEs)

Meta-learners are rooted in the potential outcomes framework, in which treatment effects are derived from counterfactual comparisons of outcomes that would result from alternative treatment levels (Splawa-Neyman *et al.* 1990). Consider the causal effect of clearcutting and replanting trees on soil carbon in the *i*th forest plot. In this example, the causal treatment can take only two values, a set-aside and unmanaged control plot (*Xi* = 0, where *X* corresponds to the treatment) or a treated stand that has been clear-felled and replanted (*Xi* = 1). When the plot is a control, its soil carbon outcome is $Y_i^{X=0}$. When the same plot is treated, its outcome is $Y_i^{X=1}$. Both $Y_i^{X=0}$ and $Y_i^{X=1}$ are called potential outcomes because either one is potentially observable. The difference between these potential outcomes is the plot-level causal effect of clearcutting, i.e., the individual treatment effect for plot *i*. For prediction of treatment effects to be causally valid, potential selection biases (non-random assignment of treatments to sample units) must be addressed and several causal assumptions met, either through sampling design or adjustment by confounding covariates in the modelling process (Kimmel *et al.* 2021).

Meta-learners can be broadly distinguished into two groups; the simpler 'conditional mean regression' algorithms, and more complex 'pseudo-outcome' algorithms. Here, we briefly describe three popular algorithms: S-, T- and X-learners, depicted in Figure B1. See (Caron *et al.* 2022; Cheung *et al.* 2024; Okasa 2022; Salditt *et al.* 2024; Spake *et al.* 2025), for reviews.

Single-model learners (S-learner, Figure B1 a), are the simplest algorithms. S-learners train a single model to predict outcomes as a function of covariates *Z*, handling the variable indicating treatment assignment (*X*) like any other covariate. Single-model learners are typical in ecological modelling, for example, soil carbon might be modelled as a function of management status ('treated' vs unmanaged 'control'), simultaneously with climatic and topographical covariates (Z). The same model ($M_s$) is used to predict outcomes for individual sampling units *i*, forcing control (X = 0) and treatment (X = 1) conditions. For a given site, the ITE is given as the difference in predicted values of *Y* between the treatment and control, while holding all other covariates at their individual site-level values.

Two-model learners (T-learner, Figure B1b) predict ITEs by first splitting the data into two sub-datasets, one for control and one for the treatment groups, and two separate models ($M_0$ and $M_1$) using all covariates (except for treatment assignment) are used to predict the outcomes separately for control and treated units, respectively. For the soil carbon example, two separate models would be fitted as a function of climatic and topographical covariates for sites in each of the set-aside and managed datasets, and the ITE calculated as the difference between these predictions.

Cross-model learners (X-learners, Figure B1c) (Künzel *et al.* 2019), build on the two-model approach. They additionally account for potential differences in covariate distributions between treatment and control groups, arising from possible selection biases, by adjusting predictions by a parameter known as the propensity score: the probability of a unit being assigned to a particular treatment level given a set of observed covariates *Z*. X-learners are thus designed for observational studies where positivity assumptions might otherwise be violated (i.e., when certain units in a study population have zero chance of receiving the treatment). Like T-learners, two outcome models are initially fitted ($M_0$ and $M_1$) to predict outcomes Y separately for treatment and control datasets, respectively. A propensity score model ($M_{ps}$) is also fitted, predicting the treatment probability (X = 1) given Z. The intermediate outcome models ($M_0$ and $M_1$) and the propensity score model ($M_{ps}$) are often referred to as 'nuisance functions' in the machine learning literature. Intermediate treatment effects ($\boldsymbol{\tau}$) are imputed from $M_0$ and $M_1$ using predicted outcomes (*Y)* and covariates (*Z)* for treated and control datasets (hence the crossing over, panel c in figure). A second pair of models is fitted to predict these intermediate treatment effects ($M\boldsymbol{\tau}_0$ and $M\boldsymbol{\tau}_1$). Finally, the predicted treatment effects are adjusted by the propensity scores to predict ITEs. The adjustment puts more weight on treatment effects that have been estimated more precisely, i.e., the ones coming from the larger treated or control sample, respectively.

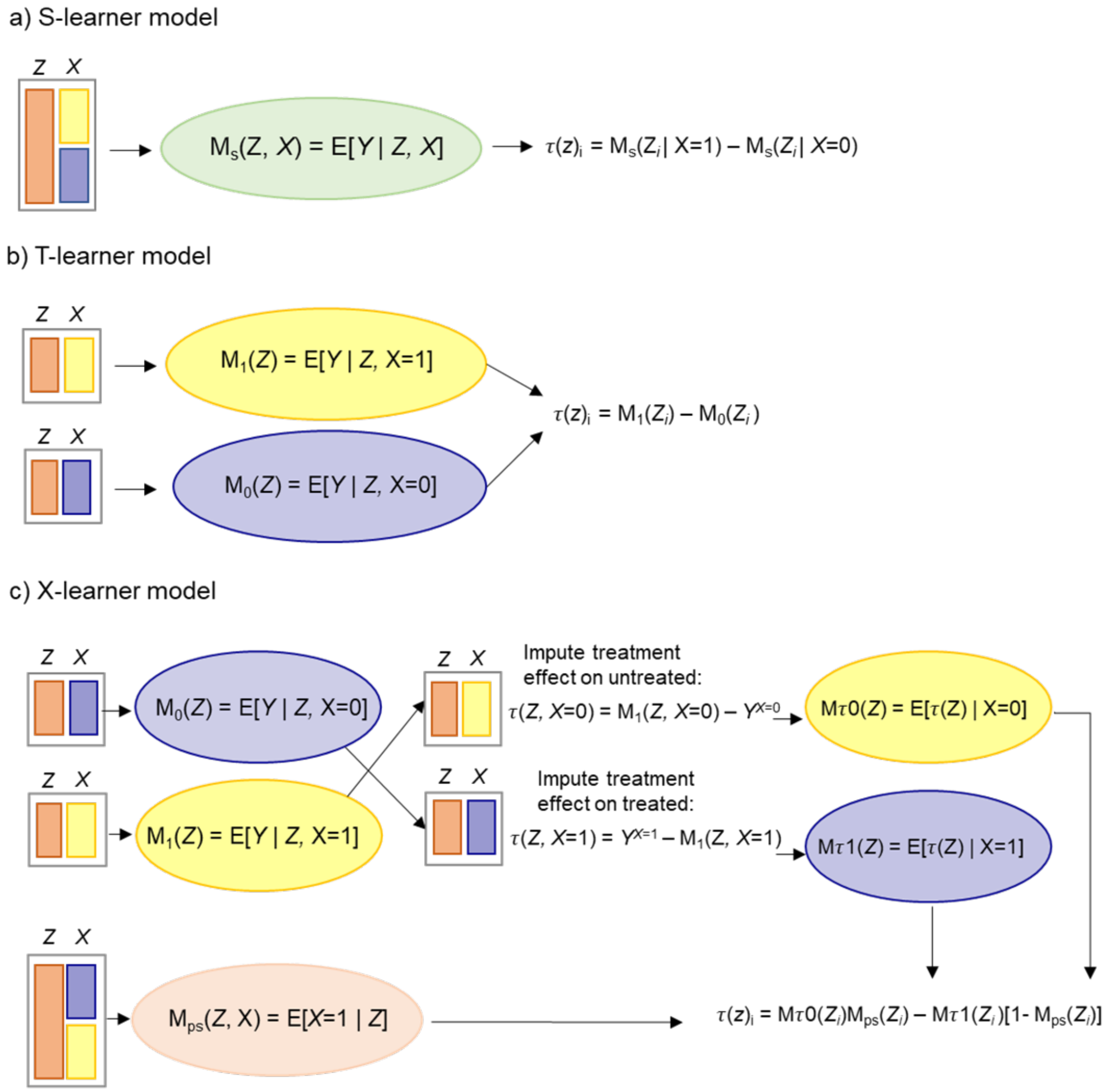
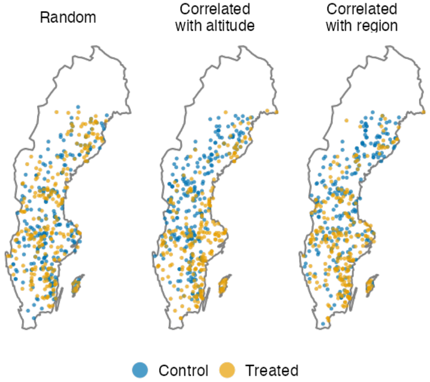
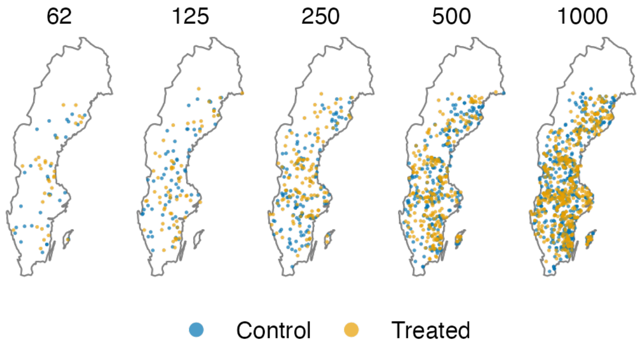
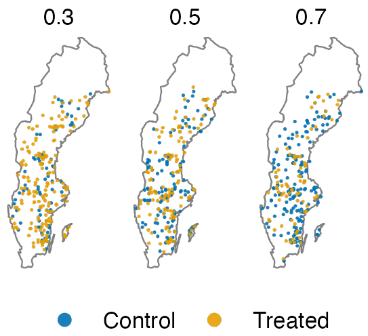
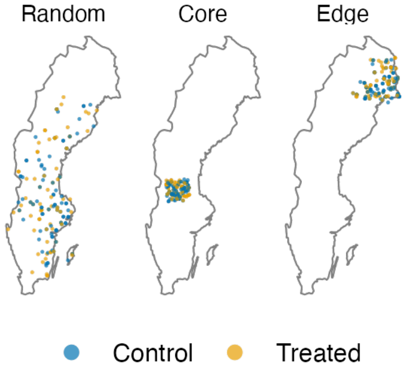

a) S-learner model
Z X
$M_s(Z, X) = E[Y \mid Z, X]$
$\tau(z)_i = M_s(Z_i \mid X{=}1) - M_s(Z_i \mid X{=}0)$
b) T-learner model
Z X
$M_1(Z) = E[Y \mid Z, X{=}1]$
Z X
$M_0(Z) = E[Y \mid Z, X{=}0]$
$\tau(z)_i = M_1(Z_i) - M_0(Z_i)$
c) X-learner model
Z X
$M_0(Z) = E[Y \mid Z, X{=}0]$
Z X
$M_1(Z) = E[Y \mid Z, X{=}1]$
Z X
Impute treatment effect on untreated:
$\tau(Z, X{=}0) = M_1(Z, X{=}0) - Y^{X=0}$
$M\tau 0(Z) = E[\tau(Z) \mid X{=}0]$
Z X
Impute treatment effect on treated:
$\tau(Z, X{=}1) = Y^{X=1} - M_1(Z, X{=}1)$
$M\tau 1(Z) = E[\tau(Z) \mid X{=}1]$
Z X
$M_{ps}(Z, X) = E[X{=}1 \mid Z]$
$\tau(z)_i = M\tau 0(Z_i)M_{ps}(Z_i) - M\tau 1(Z_i)[1 - M_{ps}(Z_i)]$

**Table 1.** Description of study features that were varied to evaluate their influence on individual treatment effect predictions.

| **Study feature** | **Rationale** | **In this study** |
|---|---|---|
| *Treatment assignment mechanism* | | |
| **Selection bias** | Observational studies face the challenge of non-random treatment assignment with respect to one or more covariates, which can induce differences in covariate overlap between the treatment groups. This constitutes a violation of the 'positivity' assumption that every unit has a non-zero probability of being in either treatment group. | Three treatment assignment mechanisms:<br>i) random (no selection bias);<br>ii) 'correlated with altitude', where the likelihood of treatment (i.e. stand management) decreased with altitude;<br>iii) 'correlated with region', where the likelihood of treatment varied systematically across NFI 'regions', corresponding to broad administrative areas from the North to the South.<br><br> Random<br>Correlated with altitude<br>Correlated with region<br>● Control<br>● Treated |
| *Sampling conditions* | | |

| | | |
|---|---|---|
| **Training sample size** | ITE prediction can be more data-hungry than conventional approaches and requires more parameter estimates (e.g., propensity scores), plus two-model meta-learners (e.g., T- and X-learners) reduce sample sizes to their treatment groups. This may represent a 'cost' to using more complex meta-learners. | Five total sample sizes: 62, 125, 250, 500 or 1000 NFI plots sampled from across Sweden.<br><br> |
| **Treatment imbalance** | Treatments can differ in their sample sizes in observational studies. and treatment levels with unbalanced sample sizes might be especially problematic when sample size is low overall (across both treatment levels), and when treatment effects are complicated (highly dependent on covariate values). | Three levels of treatment imbalance, with either 30%, 50% or 70% of the plots sampled that had been assigned as control in the previous step.<br><br> |
| *Modelling conditions* | | |

| | | |
|---|---|---|
| **Spatial overlap of test and training data** | Studies vary in the location of the test data withheld for model evaluation - analysts might choose test data that is a random subset of the entire available dataset, or choose to use test data that is geographically distinct from the training dataset (Valavi *et al.* 2019). The location of the test data in geographic space will likely influence the location of test data in covariate space, with implications for predictive performance (Yates *et al.* 2018). | We simulated sampling our test data from three different spatial locations: i) 'random' plots, sampled via stratified random sampling, where strata were latitudinal deciles; ii) 'core' plots, selected from a distinct area in the centre of Sweden, and iii) 'edge' plots in the North.<br><br> |
| **Meta-learner algorithm** | The choice of meta-learner algorithm can have major implications for the performance of ITE predictions (see Box 1). In ecology, convention is to fit a single model, including the treatment variable as any other covariate. Other disciplines including medicine and marketing have found improvements in accuracy with using Two model and Cross model algorithms (See Box 1). | We used the functions `S_RF`, `T_RF` and `X_RF` from the `causalToolbox` package (v 0.0.2.4) (Künzel *et al.* 2019) to fit S-Learner, T-Learner or X-Learner algorithms, which use random forest models as base learners. |
| **Covariate omission** | Important covariates might be omitted due to a lack of data or an incomplete understanding of the system. The exclusion of an important variable could result in biased ITE predictions due to unobserved confounding and/or the misspecification of propensity score models. | We systematically varied the inclusion of covariate 'initial soil carbon' (SI Table 1), because this covariate is not always available in empirical studies (i.e., not available for multiple years), and because it was consistently the most important variable according to variable importance scores (SI Figure 2). |

### *Virtual ecology approach*

Empirical studies often use a snapshot of real data to statistically model plot-level soil carbon as a function of environmental variables across broad extents. To emulate this, we took a 'snapshot' of the simulated soil carbon data, at simulation year 2100, i.e., the twentieth time step. In our virtual ecologist approach, we systematically varied six 'study features' (described in Table 1) related to the treatment assignment mechanism, sampling conditions and modelling conditions to quantify their influence on ITE accuracy. Each 'virtual study' consisted of different combinations of each of the six study features.

To emulate observational studies, we assigned NFI plots to one of the two treatment levels: control and treatment. We implemented different treatment assignment procedures in order to induce varying degrees of selection bias and thus confound treatment level with environmental covariates (and propensity scores), that are typical to observational studies (Table 1, SI Table 1).

After treatment assignment, we sampled plots for each 'virtual study'. We varied two types of sampling conditions: (1) 'total sample size', the number of NFI plots sampled across both treatment levels, and (2) 'sample size imbalance', the degree of imbalance between the sample sizes of control and treatment groups (Table 1).

To evaluate how modelling decisions influence ITE predictions we emulated observational studies that infer effects of forest management on soil carbon using statistical models. We selected covariates that are typically used for this purpose, including variables related to climate, topography, forest stand structure, soil conditions and management (described in SI Table 1) (e.g., Chen *et al.* 2022; Lee *et al.* 2020; Liu *et al.* 2023; Vayreda *et al.* 2012; Mazziotta *et al.* 2022b).

We varied three modelling conditions between each 'virtual study'. (1) The meta-learner algorithm used to predict ITEs; we compare the popular S-, T- and X-learner algorithms, (2) the omission of a covariate from the suite of covariates used to predict ITEs and (3) the spatial location of the 'test' dataset. The performance of predictive ecological models (e.g., species distribution models) is typically evaluated using a test or validation dataset that is withheld from the training dataset.

We made predictions for each virtual study using either an S-, T- or X- learner (Table 1). We chose to use random forest models as base models within our meta-learner frameworks since they are a popular choice in empirical studies using meta-learners, and there are a variety of software implementations available to researchers which are fast and reliable (Okasa 2022). While tuning hyperparameters for random forest models can greatly increase the accuracy of predictions (Bernard *et al.* 2009), tuning the base random forest models in meta-learner algorithms is difficult (Künzel *et al.* 2019), and since we were not interested in the selection of hyperparameters in the context of this study we used fixed hyperparameters for each algorithm. The hyperparameters were chosen in a simulation study by Künzel *et al.* (2019) and are the default settings for the meta-learner functions in the `causalToolbox` package.

We included all the covariates listed in SI Table 1 (except when omitting initial soil carbon) and used the meta-learner algorithms to predict the unit-specific treatment effects (ITEs) representing the effect of forest management on soil carbon after 20 time steps (100 years) for NFI plots in the test dataset.

We evaluated three different variants of spatial location of test data. We selected 162 test plots for ITE prediction for every simulation run from three possible pools (all excluding plots used in training datasets), (1) 'randomly selected' plots, sampled from NFI plots that were widely distributed across Sweden; (2) 'core' plots, selected from a distinct area in the center of Sweden, which were located at the center of training data's multi-dimensional covariate space (SI Figure 3); and (3) 'edge' plots located in the cooler and dryer north east. 'Edge' plots were located at the spatial periphery of the training data, as well as the periphery of multi-dimensional covariate space (SI Figure 3).
We performed a virtual study for every unique combination of our six study features (n = 810 unique combinations) with five replications for each, yielding 4,050 virtual studies in total.

## *Evaluating ITE estimate performance*

We computed two related measures of ITE prediction accuracy in our test datasets for each of the 4,050 virtual studies (which each had 162 test plots) (Yarkoni & Westfall 2017): root mean square error (RMSE) and $R^2$. $R^2$ is the squared correlation between the true ITE and the ITE estimate and RMSE is the square root of the mean squared error. To compute these values, we used measures of the true ITE values and predicted ITE values for test NFI plots in each virtual study. We calculated the true ITE for each individual NFI plot as the difference between simulated soil carbon values under treatment and control management regimes at year 2100. RMSE provides an absolute measure of the average distance that the predicted ITE values fall from the true ITE values in the units of the response variable (soil carbon), with low RMSE indicating ITE predictions that more closely matched the true ITE values for a test dataset, on average. $R^2$ measures the degree of consistency or correlation between true and predicted ITE values, and not of accuracy. $R^2$ values can be low when one or both of the ITE datasets (true or predicted) has low variation, e.g., if predictions are shrunk to a common value such as zero or the mean. To help interpret variation in RMSE and $R^2$ values with study features, we visualised the correspondence between true and predicted ITE values using scatterplots.

To quantify how the ITE prediction accuracy varies with study features, we modelled RMSE as a function of selection bias, sampling conditions (size and imbalance) and modelling conditions (covariate omission and test data location) given by each virtual study (Table 1). Separate random forest models were fitted for each meta-learner, so we could compare the relative importance of study features for each meta-learner. We computed variable importance which measures the increase in prediction error when that variable's values are randomly shuffled in the out-of-bag (OOB) data, and constructed variable importance plots to visualise the relationship between each study feature and the model's accuracy.

## Results

The predictive accuracy of plot-level ITEs measuring the effects of forest management on soil carbon varied considerably, depending on the choice of meta-learner, study features and their interactions. The S-Learner generally performed poorly compared to the T- and X-Learner (Figure 2; agreeing with hypothesis 1). Across all study features, the S-Learner achieved the lowest ITE accuracy (highest RMSE values, Figure 1) and yielded ITEs that correlated weakest with the distribution of true ITEs (lowest $R^2$ values, Figure 1). The S-Learner's poor performance can be attributed to its tendency to yield ITEs that are shrunk towards zero to a greater degree than for T-Learner and X-Learner algorithms (Figure 3).

We found that treatment imbalance affected ITE predictive accuracy distinctly for each meta-learner algorithm; importance scores were small for the S- and X-Learner algorithms but moderate for the T-learner. To visualise these interactive effects, we plotted RMSE and $R^2$ for each study feature (Table 1) and faceted by the degree of treatment imbalance (Figure 2).

### *Treatment assignment mechanism*

For all meta-learner algorithms, the imposition of selection bias (treatment assignment) was the least important study variable explaining predictive accuracy (Figure 2). RMSE was higher when treatment assignment was non-random, when correlated with region than when correlated with altitude (Figure 1).

### *Sampling conditions*

Sample size was the most influential of the tested study features in terms of determining predictive accuracy across the meta-learner algorithms (Figure 2). Larger sample sizes resulted in models with both greater accuracy and consistency with the true ITEs (Figures 1a and b, respectively). While we expected the X-learner's superior accuracy to diminish at low sample sizes due to additional modelling requirements (hypothesis 2), the S-Learner and T-Learner algorithms had comparably higher RMSE values at low sample sizes, with their differences in accuracy increasing with sample size (increasing divergence between orange and green lines in Figure 1a).

T-Learner performance was most sensitive to treatment imbalance (Figure 2); the T-Learner performed poorly, similar to the S-Learner, at 0.3 treatment imbalance (greater proportion untreated), but more similar to the X-Learner at 0.5 (no imbalance) and 0.7 (greater proportion treated). The S-Learner outperforms (lower RMSE) the T-Learner at small sample sizes only when treatment imbalance is 0.3 (proportion of untreated plots; Figure 2c).

### *Modelling conditions*

Regarding the location of test plots, we found that predictive accuracy was the lowest when training data were obtained from plots sampled randomly across Sweden, irrespective of meta-learner algorithm. The highest $R^2$ values were obtained for randomly stratified test plots (Figure 2). For T-Learner and X-Learner algorithms, the RMSEs were similar for edge and core locations. For the S-Learners, predictive accuracy was substantially lower when test plots were located at the edge of the country (which also corresponded to the edge of multivariate space, SI Figure 3), at 0.3 treatment imbalance.
Omitting an important variable from the models led to a reduction in predictive accuracy (RMSE) for all meta-learner algorithms, although it most strongly influenced performance of the T-Learner (Figure 2b), which showed the greatest reduction in accuracy (i.e., increase in RMSE) with variable omission (Figure 1g).

While consistently more accurate than the S-Learner models, the relative performance of T-Learner and X-Learner models varied as a function of study features and performance metric (Figure 1). The X-Learner yielded the lowest RMSE values, and therefore the highest accuracy, across all simulation runs. However, the T-Learners showed larger variation in their ITE predictions (more comparable to that seen in the observed data) compared to the X-Learner (Figure 3) and showed the higher $R^2$ values in most circumstances (Figure 1).

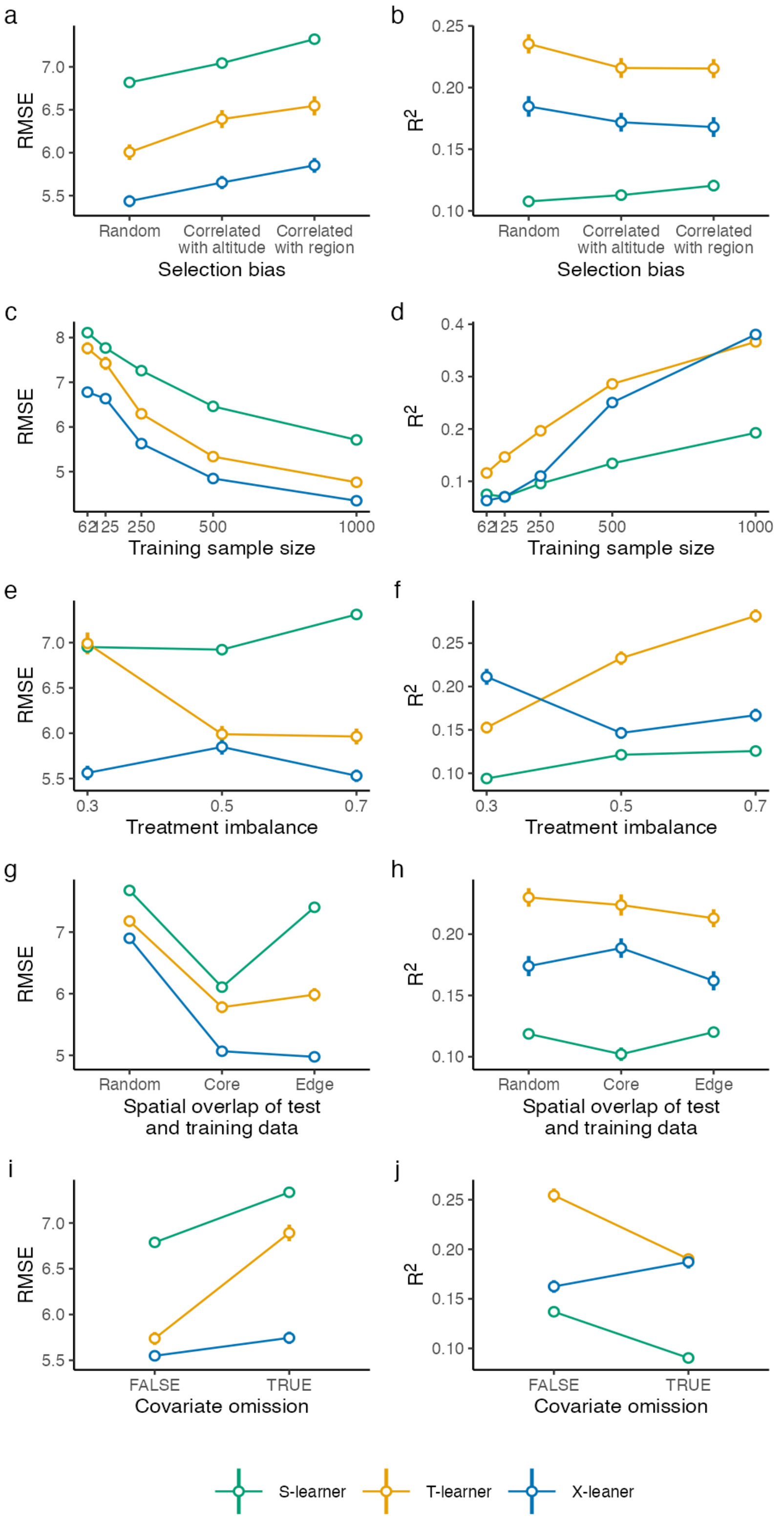


**Figure 1. Performance of three meta-learners (different colours) at predicting ITEs representing the effect of forest management on soil carbon, obtained from different combinations of study features (see Table 1).** Average RMSE (left) $R^2$ values (right) and their standard errors are shown. Note that in most cases the standard error bars are too small to be displayed. Each panel contains data from the full complement of the 4,050 virtual studies (see text). Mean R2 and RMSE are calculated using the true (simulated) and predicted ITEs from the test datasets in each study.

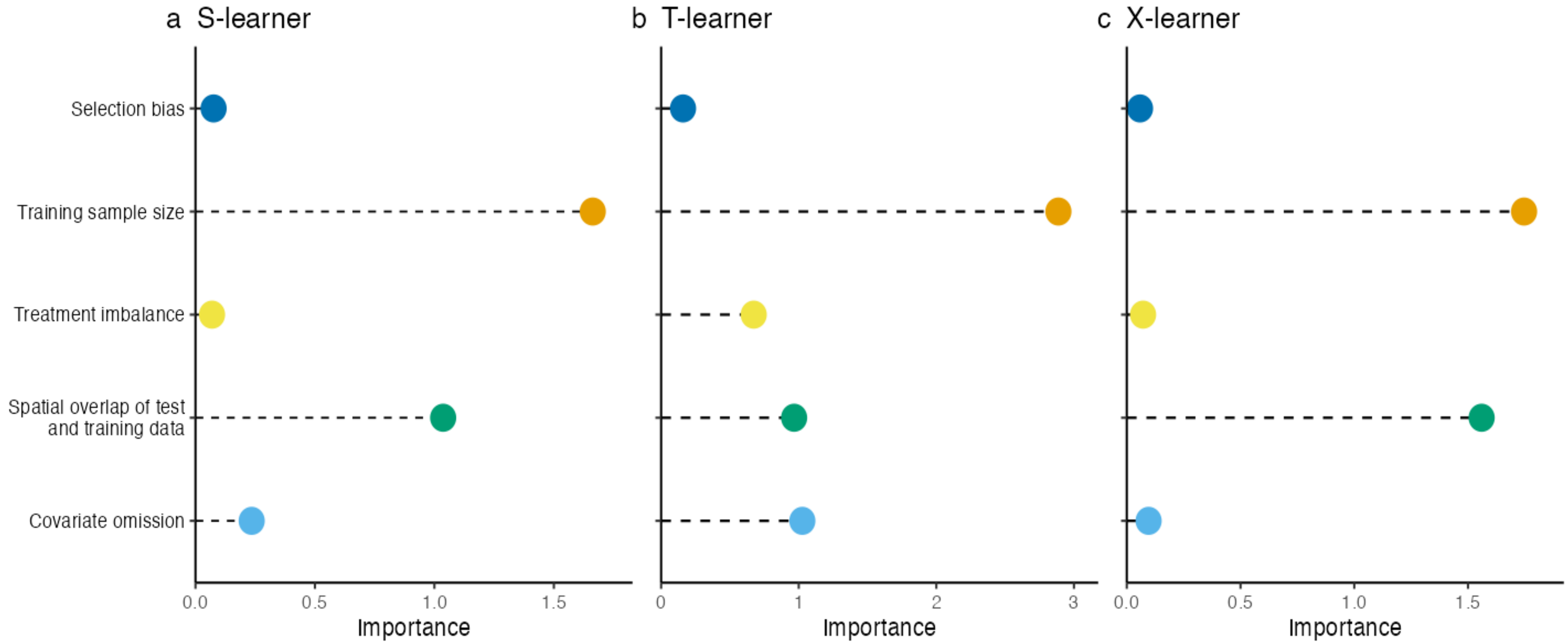


**Figure 2. Variable Importance Plots showing the relative importance of study features at influencing RMSE, shown for models employing different meta-learner algorithms (left to right).**

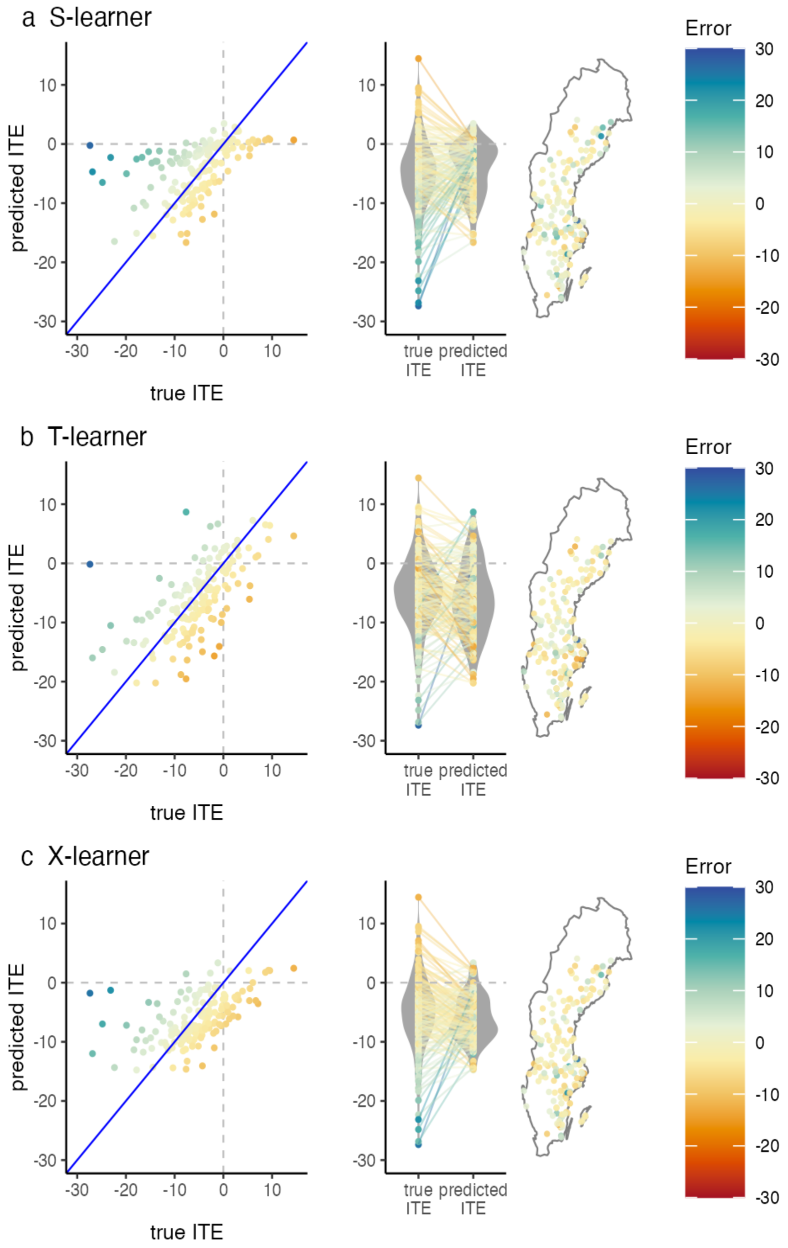


**Figure 3. Individual treatment effect (ITE) predictions obtained using three different meta-learner algorithms (top to bottom).** The first column of plots show the predicted ITE values against the true ITEs for individual National Forest Inventory plots in a test dataset spanning all of Sweden. Predictions with zero error lie on the diagonal blue line. The second column is an alternative way of visualising the same data. A line is drawn between the true and predicted values of ITE for each test NFI plot and the distribution of true and predicted ITEs are indicated by the half-violins. The final column shows the location of the test NFI plots used in Sweden. Study features were identical for the three virtual studies, varying only the meta-learner algorithm. Study features were: selection bias = random, treatment imbalance = 0.5, sample size = 1000, spatial overlap of test and training data = random, important variable omitted = FALSE.

## Discussion

Here we conducted a first application in ecology of meta-learners for ITE prediction, and compared the relative performance of three meta-learners across a broad set of sampling and modelling conditions that are common to ecological observational studies. We showed that the relative performance of different meta-learner algorithms depend on the underlying data, sampling and modelling conditions. In general, the T-Learners and X-Learners that are increasingly applied in human-centred disciplines, such as marketing and medicine, produced consistently better ITE predictions compared to the S-Learner approach that is typical to ecology. Further simulation studies are needed to establish rules of thumb regarding the choice of meta-learner algorithm for different applied empirical settings, with different data structures inherent to observational studies (Okasa 2022). Here we make suggestions below for applied ecologists wishing to make causal predictions, and for quantitative ecologists wishing to further explore how the accuracy of such predictions might vary with data conditions.

### *S-Learners perform poorly*

ITEs from S-Learners were underestimated and tended to be shrunk towards zero, consistent with previous simulation studies (hypothesis 1; Künzel *et al.* 2019; Okasa 2022). Since the treatment indicator is handled like any other covariate and plays no special role in the S-Learner, machine learning models may completely ignore treatment status during model-fitting. A tree ensemble method, like random forest, selects splitting variables randomly at each node in each tree, so treatment status may not be chosen in some trees. The likelihood of treatment being excluded from splitting rules increases with the number of covariates, as the model has more variables to choose from (Caron *et al.* 2021, 2022). Even when the treatment status remains in the model, the S-Learner may bias ITEs towards zero depending on the amount of regularisation (settings that prevent overfitting in predictive machine learning models (Künzel *et al.* 2019; Salditt *et al.* 2024)). Using the S-Learner in settings where treatment status is not a strong predictor of the outcome variable is likely to be problematic.

Since S-Learners fit a single regression, they do not account for potentially varying distributions of covariates across treatment levels; ITEs thus cannot be interpreted as causal unless assumptions have been met through design, i.e., randomised treatment assignment. In our simulations, the treatment assignment mechanism had weak effects on ITE prediction accuracy (low importance value in Figure 1e and limited effect in 2a). This is likely due to the relatively weak selection bias that resulted from our treatment assignment mechanisms (treatment assignment was random, or correlated with region or altitude, Table1). In other systems, a greater degree of covariate imbalance could occur that might influence the relative performance of meta-learners. Nevertheless, even when treatment assignment was random, and covariate distributions across control and treatment groups were similar, S-Learner performance remained consistently poor. This is likely because the S-Learner is restrictive in the way it models variation in ITEs as a function of covariates; In general, the S-Learner will perform poorly when predictor-response relationships differ across the control and treatment groups.

Two-model approaches (T- and X-Learners) performed generally better than the single-model approach (Figure 1). In contrast to the S-Learner, two-model approaches do not suffer from regularisation on treatment status, because the outcomes are modelled separately for each group. The comparable error of the approaches with the S-learner at the lowest sample sizes (Figure 1d) is likely evidence of a causal bias-variance trade-off, wherein the splitting of data in the two-model approaches yields a larger sampling variance, which may lead to more errors than the (biased) single model prediction approach that ignores counterfactuals (Fernández-Loría & Provost 2022). For greater sample sizes, two-model approaches offer greater predictive accuracy of ITEs, although this might not hold true if treatment effects are 'simple', e.g., by varying linearly with a small number of covariates. The two-model approaches performed similarly well at large sample sizes, in the absence of selection biases and variable omission.

## *The relative performance of two-model learners depends on treatment imbalance and covariate omission*

Neither the T- or X-learner performed uniformly best across all study features. Differences in their performance were nuanced, and became more pronounced the greater the treatment imbalance (Figure 1e and 1f), and when an important covariate was omitted (Figure 1i, 1j). Previous simulations from human-centred disciplines have shown that their relative performance can depend on both the size and the complexity of the treatment effects (Salditt *et al.* 2024). We found that X-Learner ITE predictive accuracy was less sensitive to treatment imbalance than T-Learners (Figure 1e,1f). With treatment imbalance, and only few data points available in one of the treatment groups, the T-learner may yield biased predictions if the individual model overfits the data in the small group, and so that differences in the two functions are (partly) due to random noise. Interestingly, the highest RMSE values (and lowest accuracy) using the T-Learner occurred when 30% of samples were treated, but had a lower error when samples were balanced or when 70% of samples were treated (Figure 1). Therefore, the effect of treatment imbalance depends on the direction of the imbalance, which might suggest a more complex functional form was required to predict the outcome in the treated group, necessitating larger sample sizes. Curth *et al.* (2024) note that when predicting the potential outcomes separately for each treatment group, prediction errors can either accumulate or cancel out across the two predictions, so that in finite samples, the model with the best fit in terms of the potential outcome is not necessarily the model with the best fit on the ITE.

We expected the X-Learner to be less sensitive to selection bias than the T-Learner (hypothesis 2). However, there was little effect of selection bias for both learners, as indicated by the near-parallel lines in Figure 1a and 1b. This is likely because covariate overlap was largely maintained, despite biasing treatment assignment by altitude and region. We note that the consequences of non-random treatment assignment will vary across studies depending on the importance of confounding covariates.

The X-Learner was developed to perform well for imbalanced treatment groups. By using the information of the control group to predict the ITE for the treatment group and vice versa (the 'crossing'), and adjusting for structural differences through propensity score weighting, X-Learners can remove some of the bias induced by regularisation and overfitting with the S-Learner and the T-Learner approaches. Yet, the X-Learner requires the estimation of more parameters than the S- and T- Learners: the computed intermediate treatment effects and propensity scores (Box 1) are ‘nuisance parameters’, and error in their estimation can propagate into the final error of the ITE. Although the sample splitting and cross fitting implemented in the cross-model approach can serve to reduce overfitting bias, in smaller samples, less data available for estimation might lead to lower accuracy due to errors in learning the ITE function itself. Consistent with expectation, the superior accuracy of the X-learner diminished at smaller sample sizes (hypothesis 3; Figure 1d).

Here we varied the degree of confounding (correlating treatment status with altitude and latitude), yet other degrees and types of confounding occur in real ecological datasets. For example, analysts using citizen scientist data might be confronted with selection bias wherein ‘treatments’ (e.g., protected area status) are confounded with other covariates (e.g., forest age), but also the additional challenge that sample site selection can vary with the outcome variable of interest (e.g., species richness), wherein citizens favour sampling in areas that are species-rich or pleasant to visit (Mair *et al.* 2017; Mentges *et al.* 2021). When predicting outcomes relevant to ecological communities, defining the observational unit for ITE can be challenging, and risks of carryover and spillover effects among sample plots will likely be important when there are landscape-level drivers (Spake *et al.* 2025).

## *Future research*

We have compared the performance of three popular meta-learners, and employed random forests as their base learners. Future research could use virtual ecology approaches to test the accuracy other meta-learners that have been proposed, including the ‘doubly-robust’ DR-Learner (Kennedy 2023) and the ‘residualisation’ R-Learner (Nie & Wager 2021), which are extensions of the X-Learner. Methods for continuous treatment variables are also under development, including disentangled representation networks (Hu *et al.* 2024). We note that meta-learners can use other machine learning methods as base-learners, including e.g., gradient boosted trees or neural networks. For other meta-learner algorithms, we direct readers to the numerous published reviews (Knaus *et al.* 2021; Künzel *et al.* 2019; Okasa 2022).

Future research could evaluate the impact of covariate choice for nuisance parameter estimation, specifically the propensity scores. Domain knowledge should guide the selection of confounders, rather than relying solely on statistical criteria to optimise model fit (Caron *et al.* 2022). Propensity score models have long been known to be highly sensitive to covariate choice (Lee *et al.* 2024). In practice, and in this study, covariate choice for outcome prediction models is typically the same for propensity score models (e.g., Künzel *et al.* 2019; Salditt *et al.* 2024). A sound understanding of how covariate choice for propensity score models influence ITE prediction accuracy would help to inform the choice of model for different data settings.

In addition to testing the implementation of alternative meta-learners and base learners, future research could vary the complexity of the treatment effect. Our virtual ecologist approach used potential outcomes for each NFI plot that were simulated using the Heureka forestry decision support system. Future work could use alternative virtual ecology approaches that generate treatment effect heterogeneity through different mechanisms in different ecological systems. Indeed, the distribution of individual treatment effects across sampling units is shifted and shaped by baseline differences, and variability in the direction and magnitude of treatment effects across individual sampling units.

### *Conclusions*

We have shown that two-model approaches are likely to provide more accurate site-level treatment effect predictions than single model approaches that are conventionally used in ecology. Why does the poor performance of single-model approaches matter for ecological studies? With ecological datasets spanning increasingly large spatial extents, this raises an important question as to whether single-model approaches - typical to ecology to e.g., predict variation in species abundance and distributions - might be underestimating site-level 'treatment' effects of interest that are often smaller in size relative to those of environmental drivers such as temperature that vary considerably across large extents. While these studies are often explicit in their aims, i.e., to predict outcomes (e.g., species abundance), or changes in biodiversity (with 'time' being the 'treatment' of interest), outcome-predictor relationships are often interpreted causally *post hoc*, with studies increasingly attempting to make ITE-like predictions at the site-level. Our findings highlight that attempts to predict ITEs using single-model approaches will likely yield biased predictions of ITEs, even if covariate distributions are equal across treatment levels (e.g., even if statistical matching is implemented), and when ITEs are small relative to other covariates. We therefore recommend two-model approaches for making causal predictions. Further work is needed to establish the utility of precision ecology approaches across different kinds of ecological systems.

## Acknowledgements

E.E.J. and R.S. were funded by the Natural Environment Research Council (NERC) for the project 'Transferable Ecology for a changing world' [grant number NE/X009998/1]. J.M.B. was funded by NERC through 'Trustworthy and Accountable Decision-Support Frameworks for Biodiversity' [grant number NE/X002233/1] and T.S. was funded by the Swedish Research Council (2019-03795).

## Data accessibility statement

Climate data were sourced from CRU TS (Climatic Research Unit gridded Time Series) (v. 4.07) (Harris *et al.* 2020). A subset of data simulated by Heureka (Wikström *et al.* 2011) (only the NFI plots and environmental variables which were used to generate the results in this paper) with metadata, and all code used to conduct the analysis and produce figures are annotated and archived in the Zenodo public repository (Jackson *et al.* 2024) 10.5281/zenodo.13269917. Code is additionally available in a GitHub repository https://github.com/ee-jackson/tree.

## References

Ågren, G.I. & Bosatta, E. (1998). *Theoretical ecosystem ecology: understanding element cycles*. Cambridge University Press, Cambridge ; New York, NY.

Ågren, G.I. & Hyvönen, R. (2003). Changes in carbon stores in Swedish forest soils due to increased biomass harvest and increased temperatures analysed with a semi-empirical model. *For. Ecol. Manag.*, 174, 25–37.

Bernard, S., Heutte, L. & Adam, S. (2009). Influence of hyperparameters on random forest accuracy. In: *Multiple Classifier Systems* (eds. Benediktsson, J.A., Kittler, J. & Roli, F.). Springer, Berlin, Heidelberg, pp. 171–180.

Caron, A., Baio, G. & Manolopoulou, I. (2021). Shrinkage Bayesian causal forests for heterogeneous treatment effects estimation.

Caron, A., Baio, G. & Manolopoulou, I. (2022). Estimating individual treatment effects using non-parametric regression models: A review. *J. R. Stat. Soc. Ser. A Stat. Soc.*, 185, 1115–1149.

Chen, X., Hisano, M., Taylor, A.R. & Chen, H.Y.H. (2022). The effects of functional diversity and identity (acquisitive versus conservative strategies) on soil carbon stocks are dependent on environmental contexts. *For. Ecol. Manag.*, 503, 119820.

Cheung, M., Dimitrova, A. & Benmarhnia, T. (2024). An Overview of Modern Machine Learning Methods for Effect Measure Modification Analyses in High-Dimensional Settings.

Curth, A., Peck, R.W., McKinney, E., Weatherall, J. & van der Schaar, M. (2024). Using machine learning to individualize treatment effect estimation: challenges and opportunities. *Clin. Pharmacol. Ther.*, 115, 710–719.

Curth, A., Svensson, D. & Weatherall, J. (2021). Really doing great at estimating CATE? A critical look at ml benchmarking practices in treatment effect estimation. In: *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*. Presented at the 35th Conference on Neural Information Processing Systems (NeurIPS 2021).

Fernández-Loría, C. & Provost, F. (2022). Causal Classification: Treatment Effect Estimation vs. Outcome Prediction. *J. Mach. Learn. Res.*, 23, 1–35.

Harris, I., Osborn, T.J., Jones, P. & Lister, D. (2020). Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. *Sci. Data*, 7, 109.

Holl, K.D. & Brancalion, P.H.S. (2020). Tree planting is not a simple solution. *Science*, 368, 580–581.

Holland, P.W. (1986). Statistics and causal inference. *J. Am. Stat. Assoc.*, 81, 945–960.

Hu, M., Chu, Z. & Li, S. (2024). DTRNet: Precisely correcting selection bias in individual-level continuous treatment effect estimation by reweighted disentangled representation. *Trans. Mach. Learn. Res.*

Hyvönen, R., Berg, M.P. & Ågren, G.I. (2002). Modelling carbon dynamics in coniferous forest soils in a temperature gradient. *Plant Soil*, 242, 33–39.

IPBES. (2019). *Global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services*. [object Object].

Jackson, E.E., Snäll, T., Gardner, E., Bullock, J.M. & Spake, R. (2024). Data and code associated with: Towards causal predictions of site-level treatment effects for applied ecology.

Kennedy, E.H. (2023). Towards optimal doubly robust estimation of heterogeneous causal effects.

*Electron. J. Stat.*, 17, 3008–3049.

Kimmel, K., Dee, L.E., Avolio, M.L. & Ferraro, P.J. (2021). Causal assumptions and causal inference in ecological experiments. *Trends Ecol. Evol.*, 36, 1141–1152.

Knaus, M.C., Lechner, M. & Strittmatter, A. (2021). Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence. *Econom. J.*, 24, 134–161.

Künzel, S.R., Sekhon, J.S., Bickel, P.J. & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proc. Natl. Acad. Sci.*, 116, 4156–4165.

Lee, J., Huling, J.D. & Chen, G. (2024). An effective framework for estimating individualized treatment rules. In: *Advances in Neural Information Processing Systems*. Presented at the 38th Conference on Neural Information Processing Systems (NeurIPS 2024), pp. 8411–8476.

Lee, S., Lee, S., Shin, J., Yim, J. & Kang, J. (2020). Assessing the Carbon Storage of Soil and Litter from National Forest Inventory Data in South Korea. *Forests*, 11, 1318.

Liu, Y., Trancoso, R., Ma, Q., Ciais, P., Gouvêa, L.P., Yue, C., *et al.* (2023). Carbon density in boreal forests responds non-linearly to temperature: An example from the Greater Khingan Mountains, northeast China. *Agric. For. Meteorol.*, 338, 109519.

Mair, L., Harrison, P.J., Jönsson, M., Löbel, S., Nordén, J., Siitonen, J., *et al.* (2017). Evaluating citizen science data for forecasting species responses to national forest management. *Ecol. Evol.*, 7, 368–378.

Mayer, M., Prescott, C.E., Abaker, W.E.A., Augusto, L., Cécillon, L., Ferreira, G.W.D., *et al.* (2020). Tamm Review: Influence of forest management activities on soil organic carbon stocks: A knowledge synthesis. *For. Ecol. Manag.*, 466, 118127.

Mazziotta, A., Lundström, J., Forsell, N., Moor, H., Eggers, J., Subramanian, N., *et al.* (2022a). More future synergies and less trade-offs between forest ecosystem services with natural climate solutions instead of bioeconomy solutions. *Glob. Change Biol.*, 28, 6333–6348.

Mazziotta, A., Lundström, J., Forsell, N., Moor, H., Eggers, J., Subramanian, N., *et al.* (2022b). More future synergies and less trade-offs between forest ecosystem services with natural climate solutions instead of bioeconomy solutions. *Glob. Change Biol.*, 28, 6333–6348.

Mentges, A., Blowes, S.A., Hodapp, D., Hillebrand, H. & Chase, J.M. (2021). Effects of site-selection bias on estimates of biodiversity change. *Conserv. Biol.*, 35, 688–698.

Moor, H., Eggers, J., Fabritius, H., Forsell, N., Henckel, L., Bradter, U., *et al.* (2022). Rebuilding green infrastructure in boreal production forest given future global wood demand. *J. Appl. Ecol.*, 59, 1659–1669.

Moyano, J., Dimarco, R.D., Paritsis, J., Peterson, T., Peltzer, D.A., Crawford, K.M., *et al.* (2024). Unintended consequences of planting native and non-native trees in treeless ecosystems to mitigate climate change. *J. Ecol.*, 112, 2480–2491.

Nie, X. & Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108, 299–319.

Okasa, G. (2022). Meta-learners for estimation of causal effects: Finite sample cross-fit performance.

Pearl, J. (2018). Theoretical impediments to machine learning with seven sparks from the causal revolution.

Peltoniemi, M., Mäkipää, R., Liski, J. & Tamminen, P. (2004). Changes in soil carbon with stand age – an evaluation of a modelling method with empirical data. *Glob. Change Biol.*, 10, 2078–2091.

Salditt, M., Eckes, T. & Nestler, S. (2024). A tutorial introduction to heterogeneous treatment effect estimation with meta-learners. *Adm. Policy Ment. Health Ment. Health Serv. Res.*, 51, 650–673.

Spake, R., Bellamy, C., Graham, L.J., Watts, K., Wilson, T., Norton, L.R., *et al.* (2019). An analytical framework for spatially targeted management of natural capital. *Nat. Sustain.*, 2, 90–97.

Spake, R., Jackson, E.E., Bullock, J.M., Gardner, E., Tipton, E., Grainger, M.J., *et al.* (2025). Precision ecology for targeted conservation action. *Nat. Ecol. Evol.*, 9, 1102–1111.

Spake, R., O'Dea, R.E., Nakagawa, S., Doncaster, C.P., Ryo, M., Callaghan, C.T., *et al.* (2022). Improving quantitative synthesis to achieve generality in ecology. *Nat. Ecol. Evol.*, 6, 1818–1828.

Splawa-Neyman, J., Dabrowska, D.M. & Speed, T.P. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Stat. Sci.*, 5, 465–472.

Tipton, E. & Mamakos, M. (2023). Designing randomized experiments to predict unit-specific treatment effects.

Valavi, R., Elith, J., Lahoz-Monfort, J.J. & Guillera-Arroita, G. (2019). blockCV: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods Ecol. Evol.*, 10, 225–232.

Vayreda, J., Martinez-Vilalta, J., Gracia, M. & Retana, J. (2012). Recent climate changes interact with stand structure and management to determine changes in tree carbon stocks in Spanish forests. *Glob. Change Biol.*, 18, 1028–1041.

Wikström, P., Edenius, L., Elfving, B.O., Eriksson, L.O., Lämås, T., Johan, S., *et al.* (2011). The Heureka Forestry Decision Support System: An Overview. *Math. Comput. For. Nat.-Resour. Sci.*, 3.

Yarkoni, T. & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspect. Psychol. Sci.*, 12, 1100–1122.

Yates, K.L., Bouchet, P.J., Caley, M.J., Mengersen, K., Randin, C.F., Parnell, S., *et al.* (2018). Outstanding challenges in the transferability of ecological models. *Trends Ecol. Evol.*, 33, 790–802.

Zurell, D., Berger, U., Cabral, J.S., Jeltsch, F., Meynard, C.N., Münkemüller, T., *et al.* (2010). The virtual ecologist approach: simulating data and observers. *Oikos*, 119, 622–635.

*Preprint*

# Supporting Information

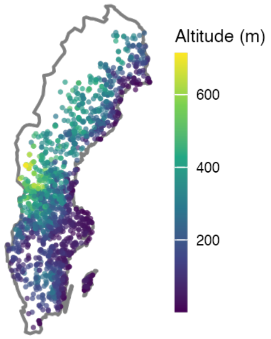
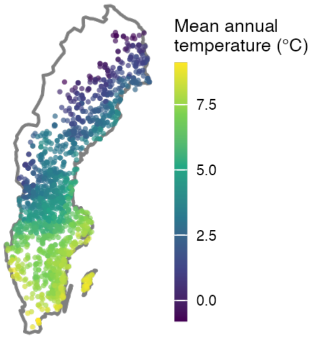
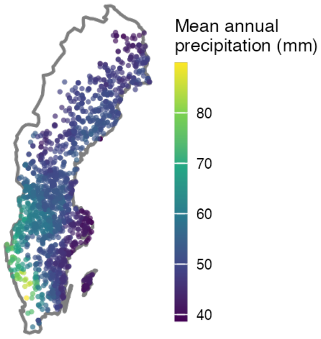
## Towards causal predictions of site-specific management effects in applied ecology

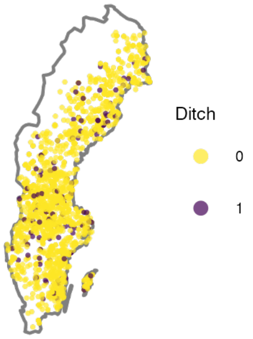
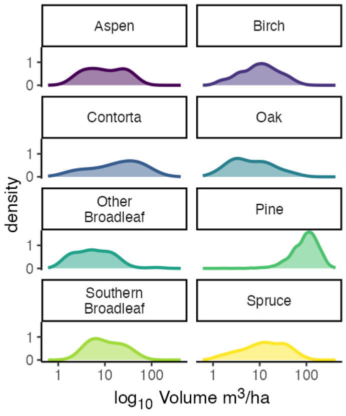
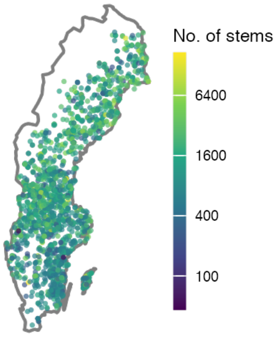
**Eleanor E. Jackson[1,2], Tord Snäll[3,4], Emma Gardner[5],**
**James M. Bullock[5] & Rebecca Spake[1,6]**

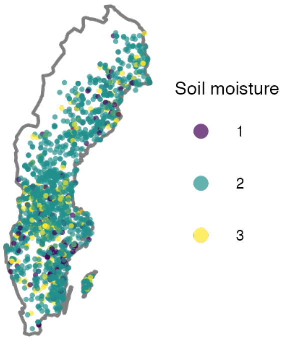
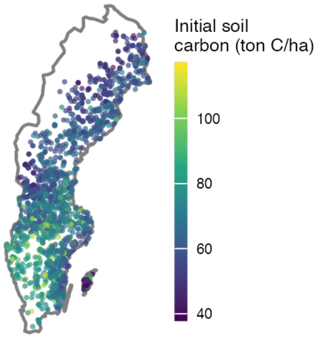
1 School of Biological Sciences, University of Reading, UK
2 Department of Biology, University of Oxford, UK
3 Skogforsk, Uppsala Science Park, Uppsala, Sweden
4 SLU Swedish Species Information Centre, Sweden
5 UK Centre for Ecology & Hydrology, Wallingford, UK
6 School of Geography and Environmental Science, University of Southampton, UK

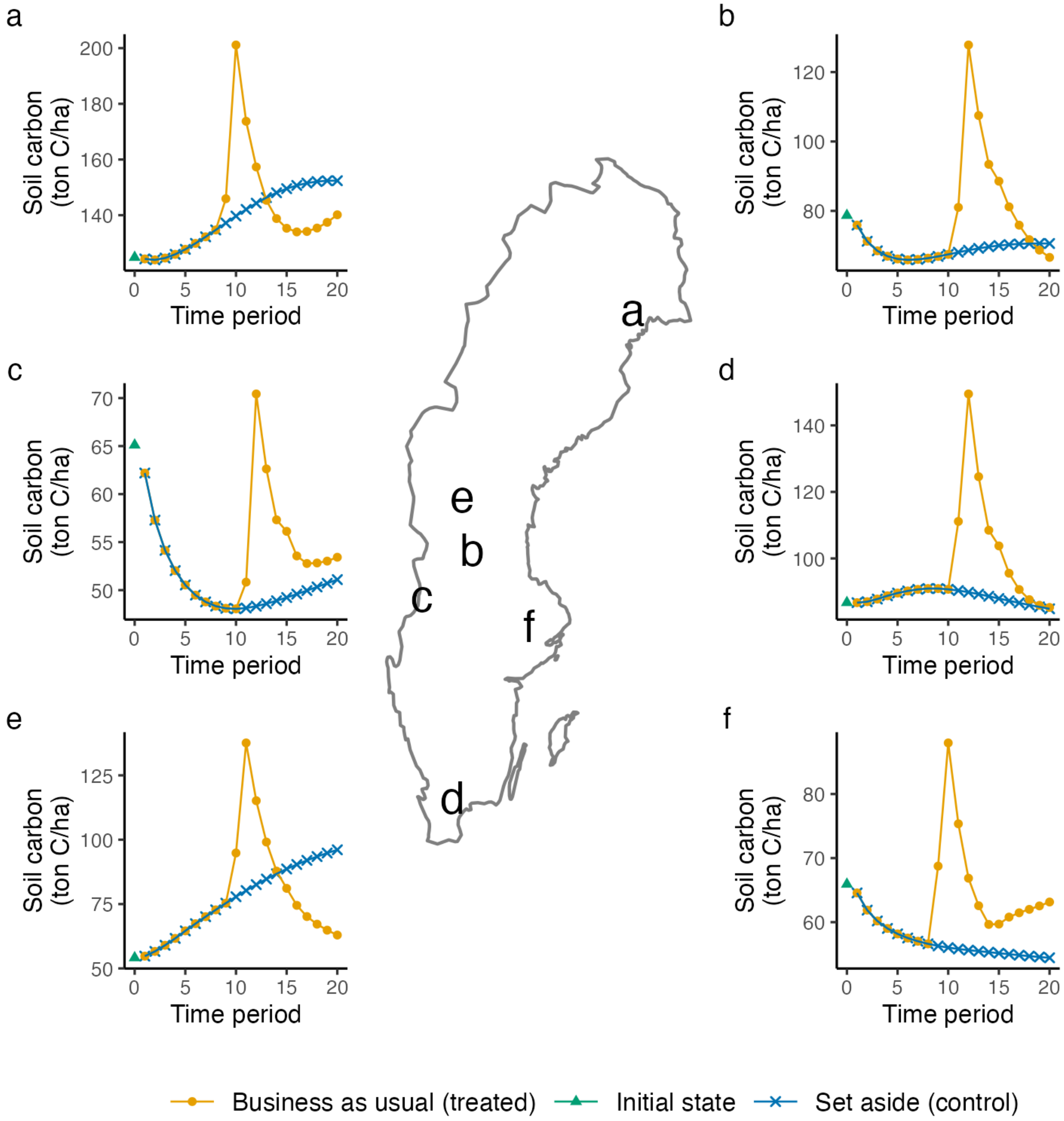
**Corresponding Author:**
Eleanor E. Jackson, Department of Biology, University of Oxford, Life & Mind Building, South Parks Road, Oxford OX1 3EL UK. Email: eleanor.jackson@biology.ox.ac.uk

**SI Table 1. Environmental covariates selected for use in statistical models predicting soil carbon.**

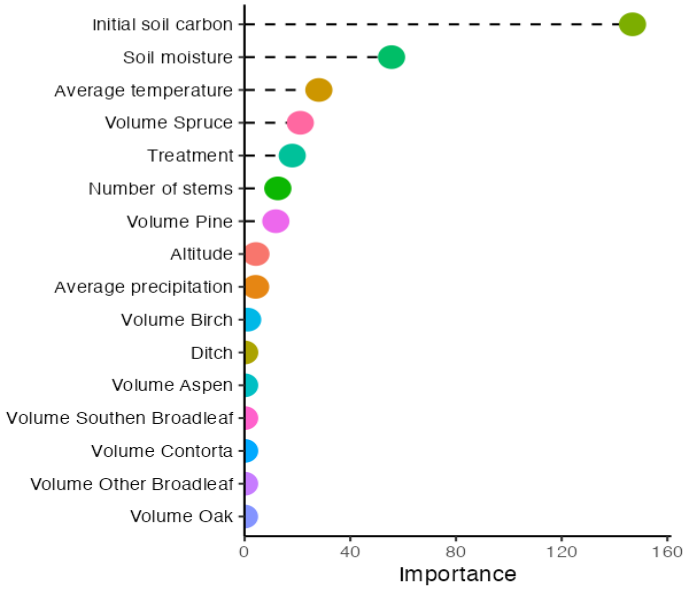
| **Covariate** | **Details** |
|---|---|
| Altitude<br> | Height above sea level, sourced from the Swedish National Forest Inventory. |
| Temperature<br> | Mean annual temperature, sourced from CRU TS (Climatic Research Unit gridded Time Series) (v. 4.07) (Harris et al., 2020). Plots were matched to the nearest climate station and mean annual temperature was averaged across a 5-year period prior to NFI sampling. |
| Rainfall<br> Mean annual precipitation (mm)<br>80<br>70<br>60<br>50<br>40 | Mean annual precipitation, sourced from CRU TS (Climatic Research Unit gridded Time Series) (v. 4.07) (Harris et al., 2020). Plots were matched to the nearest climate station and mean annual precipitation was averaged across a 5-year period prior to NFI sampling. |

| | |
|---|---|
| Ditch<br><br> | A binary variable indicating if the site has been ditched to aid water drainage, where 0 is no ditching and 1 is ditched. Sourced from the Swedish National Forest Inventory. |
| Volume of tree species<br><br> | Absolute volume of tree species within the plot, as recorded by the Swedish National Forest Inventory. (Note this only contains plots included in our study - we limited our sample to NFI plots situated in pine-dominated stands (≥50% standing volume), see methods). |
| Number of stems<br><br> | Total number of stems within the plot (DBH >= 4 cm) sourced from the Swedish National Forest Inventory. |

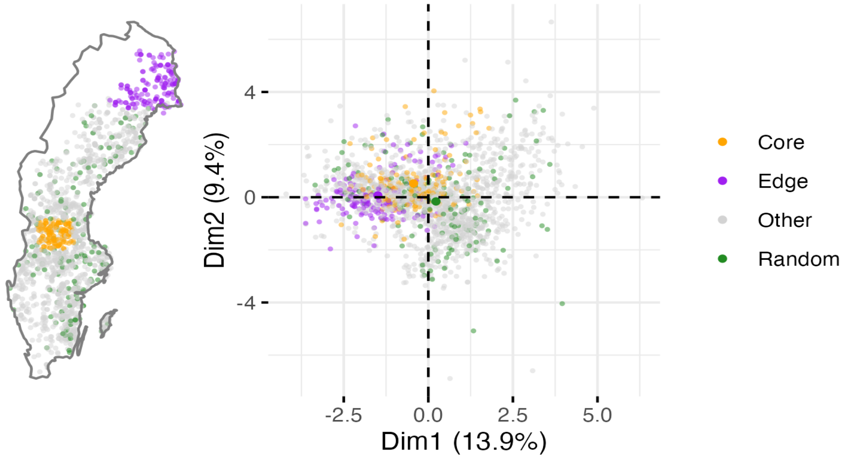
| | |
|---|---|
| Soil moisture<br><br> | An ordinal variable ranging from 1 (dry) to 3 (mesic-moist), sourced from the Swedish National Forest Inventory. (Note this only contains plots included in our study - we did not include plots with soil moisture 4 (moist) or 5 (wet) plots see methods.) |
| Initial soil carbon<br><br> | Total amount of soil organic carbon as measured in the Swedish National Forest Inventory at *t* = 0, and corresponding to input data for Heureka's soil carbon model. |



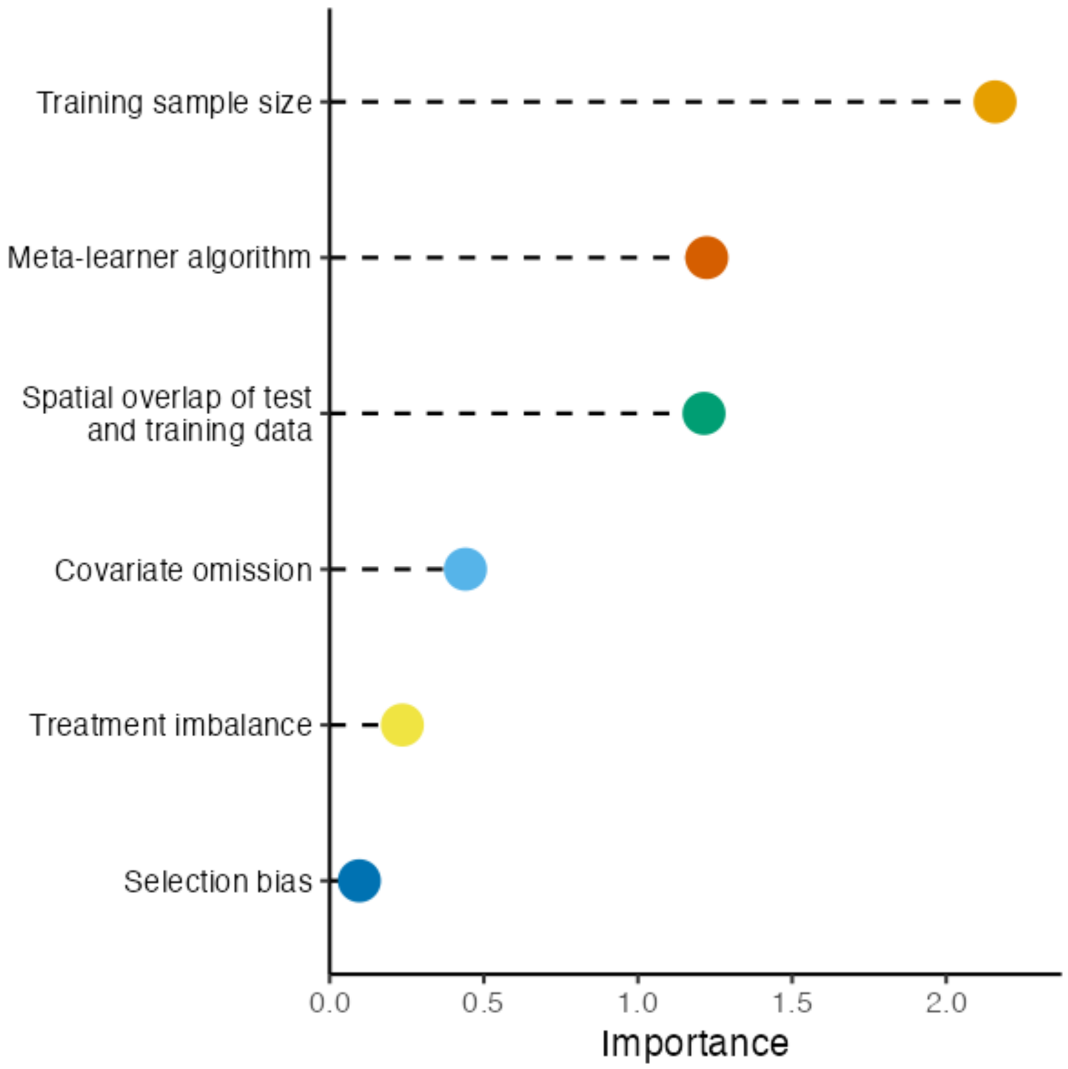
**SI Figure 1. Simulated soil carbon over time for six National Forest Inventory Plots.** Figure labels (a - f) are plotted on a map of Sweden to indicate their location. For the first year soil carbon is plotted at its “initial state” (green triangle), this is the value measured during National Forest Inventory surveys in the given year. Subsequent values were simulated by Heureka under either “business as usual” management where trees are clear cut and replanted (orange circle, “treated” in this study), or “set aside” (blue cross, “untreated”). Given the same management intervention, soil carbon after 20 years can be the same (d), higher (c, f), or lower (a, b, e) than if the same plot was set aside, demonstrating variation in the unit specific treatment effect.



**SI Figure 2. Importance of variables for predicting soil carbon after 20 simulated time steps.** Soil carbon at period 20 was modelled as a function of environmental variables (main text Table 2) and treatment (set aside or business as usual) in a random forest.



**SI Figure 3. Location of test data.** The left panel depicts a map of Sweden with the National Forest Inventory plots indicated by points. The right panel is a principal component analysis visualisation where points which are closer in space are plots with similar environmental covariates (listed in main text Table 2). Edge plots (purple) were selected to be at the periphery of the training data's multi-dimensional covariate space and core plots (orange) were selected to be at the centre of covariate space whilst being geographically distinct. Random plots (green) were sampled by stratified random sampling, grouped according to latitude. All other plots (grey) were not used in test datasets. See methods.



**SI Figure 4. Importance of study features for predicting RMSE.** RMSE for each virtual study (n = 4,050) was modelled in a random forest as a function of meta-learner algorithm, selection bias, sampling and modelling conditions (main text Table 1).

## References

Harris, I., Osborn, T. J., Jones, P., & Lister, D. (2020). Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. *Scientific Data*, *7*(1), 109. https://doi.org/10.1038/s41597-020-0453-3