# Towards causal predictions of site-level treatment effects for 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 land and resources available to implement conservation actions, efforts must be effectively targeted to individual places. This demands predictions of how individual sites respond to alternative interventions. Meta-learner algorithms for predicting individual level treatment effects (ITEs) have been pioneered in marketing and medicine, but they have not been tested in ecology. We present a first application of meta-learner algorithms to ecology by comparing the performance of algorithms popular in other disciplines (S-, T-, and X-Learners) across a broad set of sampling and modelling conditions that are common to ecological observational studies. We conducted 4,050 virtual studies that measure the effect of forest management on soil carbon. These varied in sampling approach and meta-learner algorithm. The X-Learner algorithm that adjusts for selection bias yields the most accurate predictions of ITEs. Our findings pave the way for ecologists to leverage machine learning techniques for more effective and targeted management of ecosystems in the future.

## Keywords

conditional average treatment effect, treatment effect heterogeneity, 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:**
Rebecca Spake, School of Geography and Environmental Science, Building 44, Highfield, University of Southampton, Hampshire, SO17 1BJ, UK. Email: r.spake@soton.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 ecological and social 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, or 'treatments', from hereon. Currently, applied ecologists do not 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. However, when treatment effects vary in magnitude and direction across sites, 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 explicitly design workflows to 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 to patients and targeted marketing to customers. Prediction of ITEs is possible with individual sampling-unit-level data on outcomes of interest, information on treatments that units have been subjected to, and other covariates that also predict those outcomes. Spake et al. (2025) recently argued why and how applied ecology can capitalise on these rich advances, potentially allowing for effective conservation over large extents. A range of ITE prediction methods, known as meta-learners, have been developed (Box 1), differing in how they adjust for confounding covariates and selection bias. Simulations have recently been used to assess their predictive performance 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., 2019a; Okasa, 2022a). Given that ITEs defy direct observation because we can never truly observe the outcomes of individual units 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). Importantly, no single meta-learner consistently outperforms others across all data conditions (Knaus et al., 2021; Okasa, 2022a). 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., 2019a). 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 estimating ITEs. Specifically, we simulate soil carbon mass (our outcome variable) for Swedish National Forest Inventory plots (our 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 a systematic understanding of how treatment assignment, sampling, and choice of meta-learners interact to influence ITE accuracy within real-world situations. We hope that this provides guidance and an incentive for ecologists to implement meta-learner approaches.

## Methods

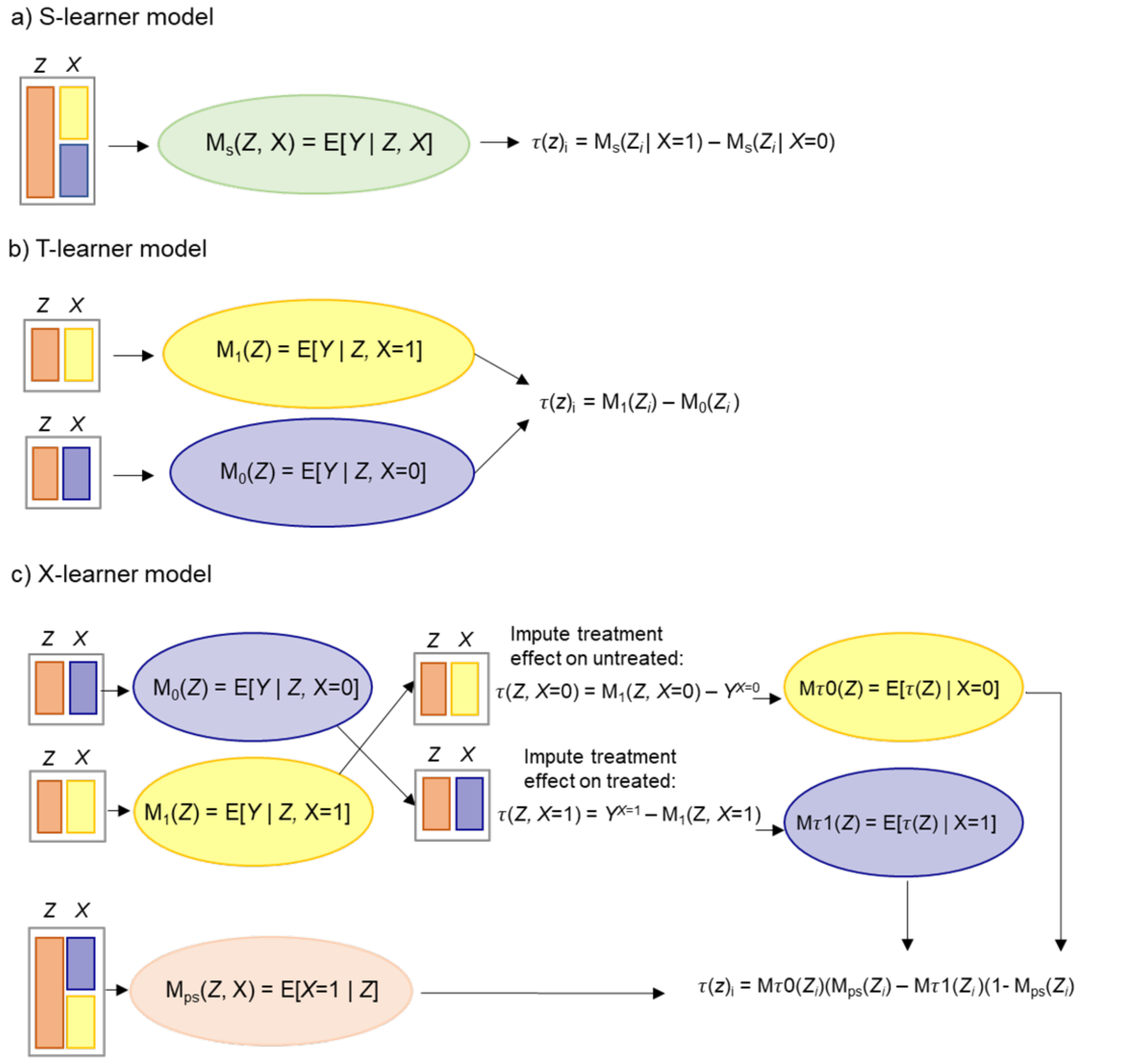
We implemented a virtual ecology approach that: i) simulated local forest dynamics under two alternative management interventions that were assigned in various ways to forest plots across a large, environmentally heterogeneous extent (treatment assignment mechanism). These forest plots were then: ii) 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.

Our outcome variable of interest was soil organic carbon, 'soil carbon' from hereon. We selected this metric, because maintaining or increasing soil carbon has many benefits for climate change mitigation, adaptation and biodiversity conservation, such as enhanced soil fertility and water-holding capacity, increasing productivity, and support to belowground biodiversity (Mayer et al., 2020). As the manner and intensity in which forests are managed can influence soil carbon, it is important that forest management planning considers soil carbon (Mayer et al., 2020; Mazziotta et al., 2022a). Many studies have therefore used plot-level data over large scales to statistically model how soil carbon varies as a function of multiple interacting drivers, including current and future climate, topography, and soil chemical, physical and biological properties), as well as management-related variables such as canopy-dominant species, age class and land use history in order to infer how soil carbon will change as a function of alternative management regimes and under future climates (e.g., (e.g., Chen et al., 2022; S. Lee et al., 2020; Liu et al., 2023; Mazziotta et al., 2022a; Vayreda et al., 2012). However, these approaches have adopted outcome prediction models that do not satisfy causal assumptions.

## 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. Consider the causal effect of clearcutting and replanting 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 clearfelling, 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 by adjustment by confounding covariates in the modeling process (Kimmel et al., 2021).

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



The single-model learners (S-learner, panel a in figure), are the simplest algorithms, and 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 protected area status (set-aside or managed), simultaneously with climatic and habitat 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, panel b in figure) 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 habitat 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-learner; (Künzel et al., 2019b), panel c in figure), 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 individuals 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.

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

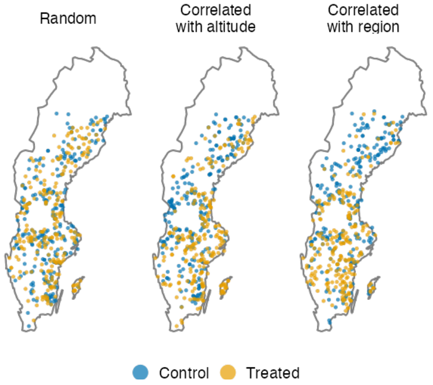
Forest stand dynamics were simulated for National Forest Inventory (NFI) plots across Sweden using the Heureka simulation system (Wikström, Edenius, Elfving, Eriksson, Lämås, et al., 2011). Heureka is widely used in both forestry and research to forecast the outcomes of alternate management scenarios across Sweden (Mazziotta et al., 2022b; Moor et al., 2022). Heureka comprises a set of empirical growth and yield models that simulate stand development in five-year time steps, combining tree species-specific models of tree establishment, growth and mortality. Soil carbon was modelled across cohorts using a dynamic decomposition Q-model, following the mass loss of litter over time for different litter compartments. The theoretical framework is presented in Ågren & Bosatta (1998) and the general implementation is described in Ågren & Hyvönen (2003).

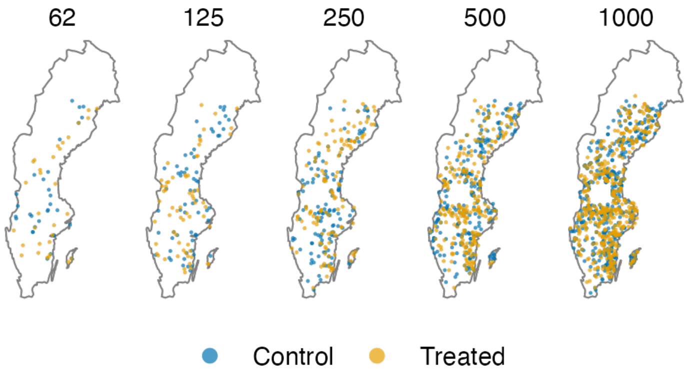
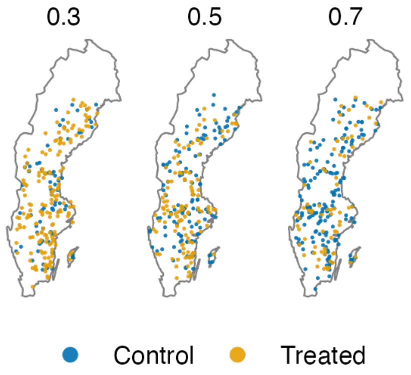
We simulated two scenarios of forest dynamics and management for the period 2010–2100, which were: set-aside ('control' from hereon), where stands were set-aside and left to develop without any intervention, and 'business as usual clearcutting forestry without thinning' ('treatment' from hereon). These stands were clearcut and replanted at the stand age of 60-120 years, at lower ages in southern, high-productive stands and higher in northern, low-productive stands, at typical clearcutting stand stockings. Simulations were 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. For each NFI plot and time step, a large number of management actions can be simulated, which together comprise a management schedule. We restricted Heureka's simulation output to plots for which both our control and treatment management scenarios were simulated. We limited our sample 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 that could be parameterised in statistical models. 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 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. Since the application of the clear-cutting treatment occurred at different time periods for different plots, we limited our study to only include plots which reached their maximum soil carbon at the 12th time step since the start of the simulation and plots with a single "peak" in maximum soil carbon - indicating they were only clear cut once during the simulation (yielding 1,806 plots). We did this because data on time since management 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.

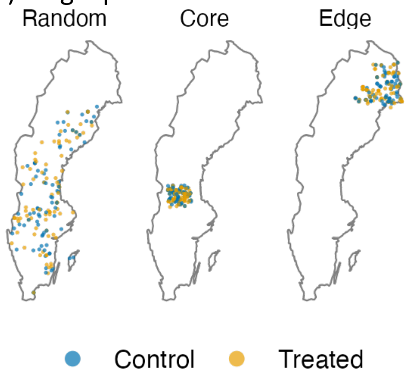
### *Virtual ecology approach*

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 (inducing varying degrees of selection bias), plot sampling (varying sample size and imbalance across treatment levels) and modelling (meta-learner algorithm choice, covariate omission and test data location) to quantify their influence on ITE accuracy. Each 'virtual study' consisted of different combinations of each of the six study features.

**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 selection bias, where non-random treatment assignment with respect to one or more covariates can induce differences in covariate overlap between the treatment groups, creating regions in areas of covariate space without appropriate comparators, i.e., where only treated, or only control, subjects are present. For example, unmanaged forest stands are often located at higher altitudes than managed stands due to differences in site accessibility and productivity. This constitutes a violation of the 'positivity' (or common support) assumption that every unit has a non-zero probability of being in either treatment group. When confounding variables are observable, they can be accounted for with various modelling strategies (e.g., X-learner algorithms adjust for propensity scores). If unobserved, then propensity score adjustment could introduce more bias. | We used three treatment assignment mechanisms to induce varying levels of selection bias: i) random (no selection bias); ii) 'correlated with altitude', where the likelihood of treatment decreased with altitude; and 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 to average treatment effect prediction. ITE estimation requires more parameter estimates (e.g., propensity scores), and two-model meta-learners (e.g., T- and X-learners) reduce sample sizes to | Five total sample sizes were used: 62, 125, 250, 500 or 1000 NFI plots sampled from across Sweden. |

| | | |
|---|---|---|
| | their treatment groups. This may represent a ‘cost’ to using more complex meta-learners. |  |
| **Sample imbalance** | Treatments can differ in their sample sizes in observational studies. For example, in monitoring programmes that aim for representative sampling of geographic units (e.g., nationally), sampling units might sample in proportion to the actual probability of treatment assignment in nature. Treatment levels with unbalanced sample sizes might be 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 sample size imbalance were implemented, with either 30%, 50% or 70% of the plots sampled that had been assigned as control in the previous step.<br> |
| *Modelling conditions* | | |

| | | |
|---|---|---|
| **Spatial overlap of test and training data** | 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. Studies vary in the location of the test data used - 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). | Three different spatial locations of test data were simulated: i) 'randomly selected' plots, sampled from NFI plots across Sweden; 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. 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). | Three meta-learner algorithms were implemented: i) Single model; ii) Two model and iii) Cross model |
| **Covariate omission** | Ecologists often make a strong (usually implicit) assumption that all important variables are observed (i.e., measurable) and therefore available for inclusion in our models, e.g., for predicting an outcome or for generating propensity scores. In reality, important covariates might be omitted due to a lack of availability, or due to 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 or exclusion of the covariate 'initial soil carbon' when training our base learners. |

### *Treatment assignment conditions: Selection bias*

To emulate observational studies, NFI plots were assigned to one of the two treatment levels: control (set-aside) and treatment (clearcut and replanted). Observational studies face the challenge of selection bias, where sampling units are non-randomly assigned to treatment levels. For example, unmanaged forest stands are often located at higher altitude than managed stands due to differences in site accessibility and productivity (Lindenmayer & Laurance, 2012), or due to different economic priorities of different administrative units of land. A range of approaches have been developed in applied statistics to deal with confounding and limit bias in treatment effect estimation (Kimmel et al., 2021). Several approaches use propensity score estimation, to summarise multiple confounders and quantify the degree of covariate imbalance across treatment levels. The propensity score is the probability of treatment assignment, conditional on observed baseline covariates (Austin, 2011).

We implemented different treatment assignment procedures in order to induce varying degrees of selection bias and thus confounding of treatment level with environmental covariates (and thus propensity scores), that are typical to observational studies (SI Table 1). Plots were assigned to either treated (BAU) or control (set aside) according to one of three treatment assignment procedures, the first inducing no bias: (1) 'random', to mimic randomised control designs (rare in ecology); and two that imposed systematic selection bias: (2) 'correlated with altitude', where plots located at lower altitudes were more likely to be assigned to treated (BAU) than control (set aside) conditions, and (3) 'correlated with region', where the probability of treatment assignment systematically varied with latitude. For random treatment assignment, 50% of the available plots were randomly selected as not treated using the `slice_sample` function from the R package `dplyr` (v 1.1.2). For the 'correlated with altitude' assignment, plots were assigned as above but the sampling weights were equal to altitude (SI Table 1). For the 'correlated with region' assignment, we assigned plots a sampling weight according to the region of Sweden where they were located. We aggregated regions 1 and 2.1, and regions 4 and 5 from the Swedish National Forest Inventory to group them into latitudinal bands. Sampling weights ranged from 0.1 to 0.4, increasing from North to South. We used the `weight_by` argument in `slice_sample` to randomly assign 50% of plots as not treated according to the sampling weights.

### *Sampling conditions: sample size and imbalance*

Two sampling conditions were varied, including the (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. Five different total sample sizes were used: 62, 125, 250, 500 or 1000 and three levels of sample size imbalance were implemented, with either 30%, 50% or 70% of the plots sampled that had been assigned as control in the previous step. Hence, we selected the control plots randomly using `slice_sample` where `n` was equal to the total sample size multiplied by the level of sample size imbalance. Treated plots were randomly sampled using `slice_sample` where `n` was equal to the total sample size minus the number of control plots which had been selected.

### *Modelling conditions for ITE prediction: meta-learner algorithm, covariate omission and test data location*

We set out to evaluate how modelling decisions influence ITE predictions. To emulate 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; S. Lee et al., 2020; Liu et al., 2023; Mazziotta et al., 2022a; Vayreda et al., 2012).

Three modelling conditions were varied between simulation runs. (1) The meta-learner algorithm used to predict ITEs; we compare S-, T- and X-learner algorithms. (2) The omission of a covariate from the suite of covariates used to predict ITEs. We chose to vary the inclusion of covariate 'initial soil carbon' (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). (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 gets withheld from the training dataset. Test data can differ from training data in both geographic and covariate space, with implications for predictive performance (Roberts et al., 2017).

We used the functions `S_RF`, `T_RF` and `X_RF` from the `causalToolbox` package (v 0.0.2.4) (Künzel et al., 2019a) to fit S-Learner, T-Learner or X-Learner algorithms, which use random forest models as base learners. 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, 2022a). 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., 2019a) 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.* (2019a) and are the default settings for the meta-learner functions in the `causalToolbox` package.

Simulated soil carbon after 20 time steps was predicted as a function of the covariates listed in SI Table 1. We included all the covariates listed in SI Table 1 (except when omitting initial soil carbon). We then used the models to estimate the unit-specific treatment effects (ITE values) for NFI plots in the test dataset with the `EstimateCate` function from `causalToolbox`.

We evaluated three different variants of spatial location of test data. 162 test plots were selected for ITE prediction for every simulation run from three possible pools (all excluding plots used in training datasets), i) 'randomly selected' plots, sampled from NFI plots that were widely distributed across Sweden; ii) 'core' plots, selected from a distinct area in the centre of Sweden. These test plots were located at the centre of training data's multi-dimensional covariate space (SI Figure 3); and iii) 'edge' plots located in the cooler and dryer north east. These test plots were located at the spatial periphery of the training data, as well as the periphery of multi-dimensional covariate space (SI Figure 3). 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 estimate performance in our test datasets (Yarkoni & Westfall, 2017): root mean square error (RMSE) and $R^2$ using the package `yardstick` (v 1.2.0) (Kuhn et al., 2023) for each of the 4,050 virtual studies (which each had 162 test plots). $R^2$ 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. The true ITE for each individual NFI plot was calculated as the difference between simulated soil carbon values under treatment (BAU) and control (set aside) management regimes at year 2100. Predicted ITE values were obtained from the `EstimateCate` function as described above. 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 models were fitted for each meta-learner, so we could compare the relative importance of study features for each meta-learner. We used the `tidymodels` (v 1.1.0) family of `R` packages with the `ranger` R package (v 0.15.1) (Probst et al., 2019) to build a random forest model and tune hyperparameters. We computed variable importance and constructed variable importance plots using the `vip` R package (v 0.4.0) (Greenwell & Boehmke, 2020) 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 (a typical method of ITE prediction in ecology) performed consistently poorer than T-Learner and X-Learner; across all study features, it 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).

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). The S-Learner and T-Learner algorithms had comparably high 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).

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).

Omitting an important variable from the models led to a reduction in predictive accuracy 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).

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 lowest when treatment assignment was random i.e, when there was no selection bias. For the two biassed treatment assignment procedures, RMSE values were higher when treatment assignment was correlated with region than when correlated with altitude (Figure 1).

Treatment imbalance affected ITE predictive accuracy distinctly for each meta-learner algorithm. The highest RMSE values (and lowest accuracy) of predictions from each meta-learner occurred at different degrees of treatment imbalance: 0.7 for the single model, 0.3 for the T-Learner and 0.5 for the X-Learner algorithms (Figure 1).

Regarding the location of test plots, we found that predictive accuracy was the lowest when training data were obtained from plots sampled randomly across the whole extent of Sweden, irrespective of meta-learner algorithm. For T-Learner and X-Learner algorithms, the RMSEs were similar for edge and core locations. 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).

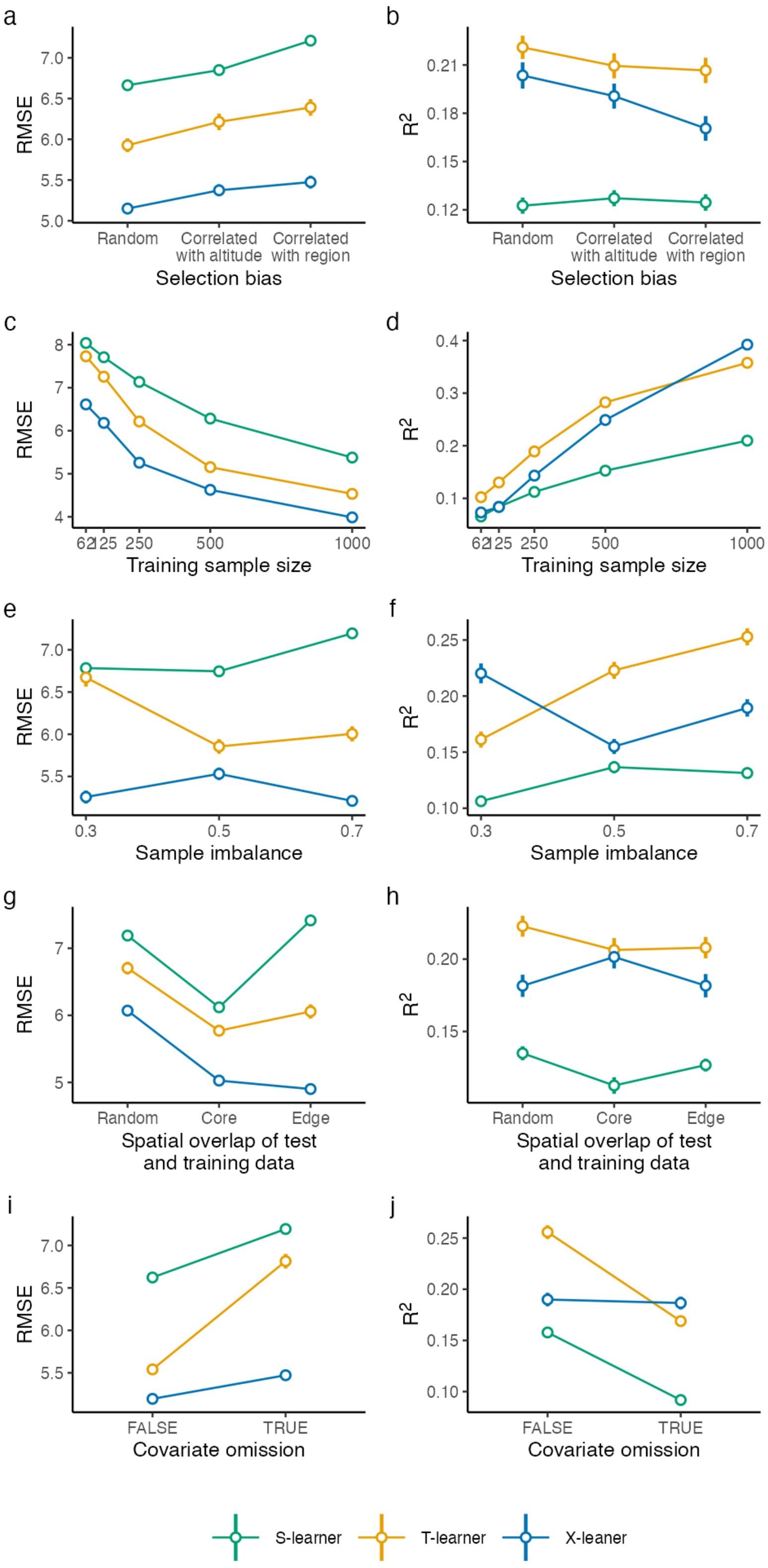


**Figure 1. Performance of three meta-learners (different colours) at predicting ITEs representing the effect of forest management on soil carbon, obtained from different combinations of study features (see Table 1).** Average RMSE (left) R$^2$ values (right) and their standard errors are shown. Note 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.

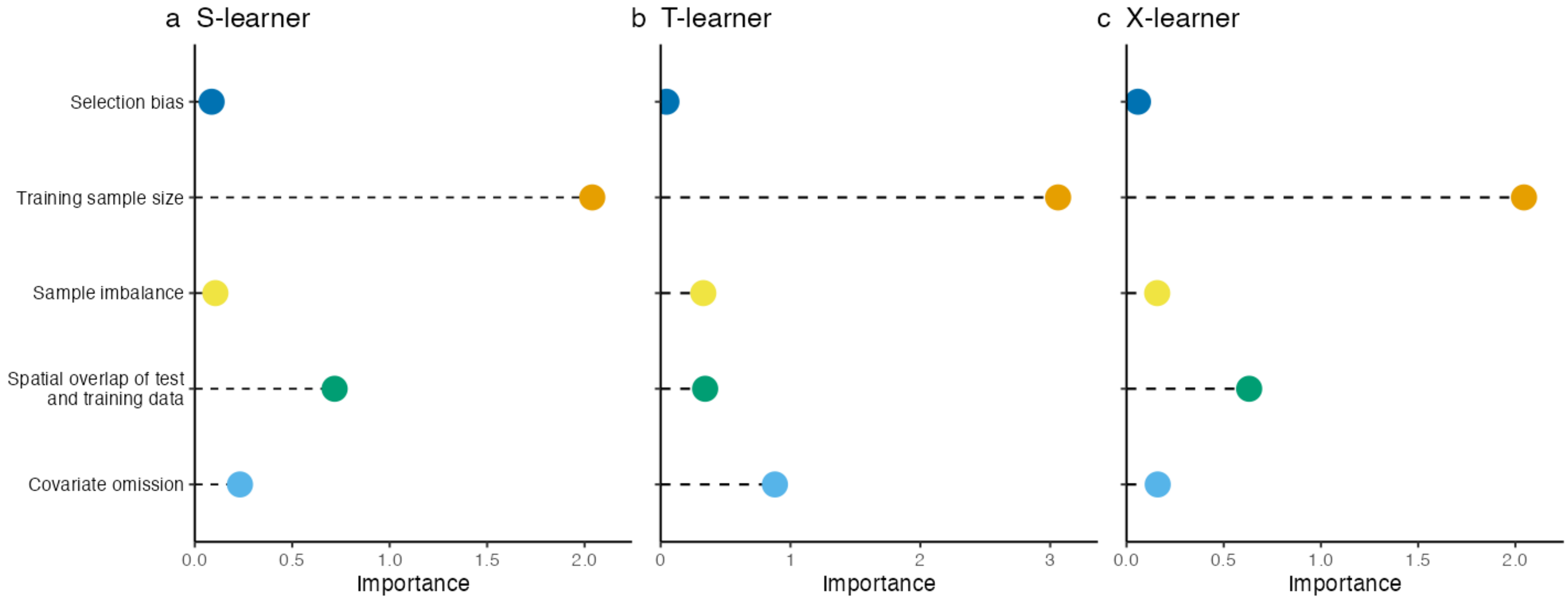


**Figure 2. Variable Importance Plots showing the relative importance of study features at influencing RMSE, shown for models employing different meta-learner algorithms (left to right).**

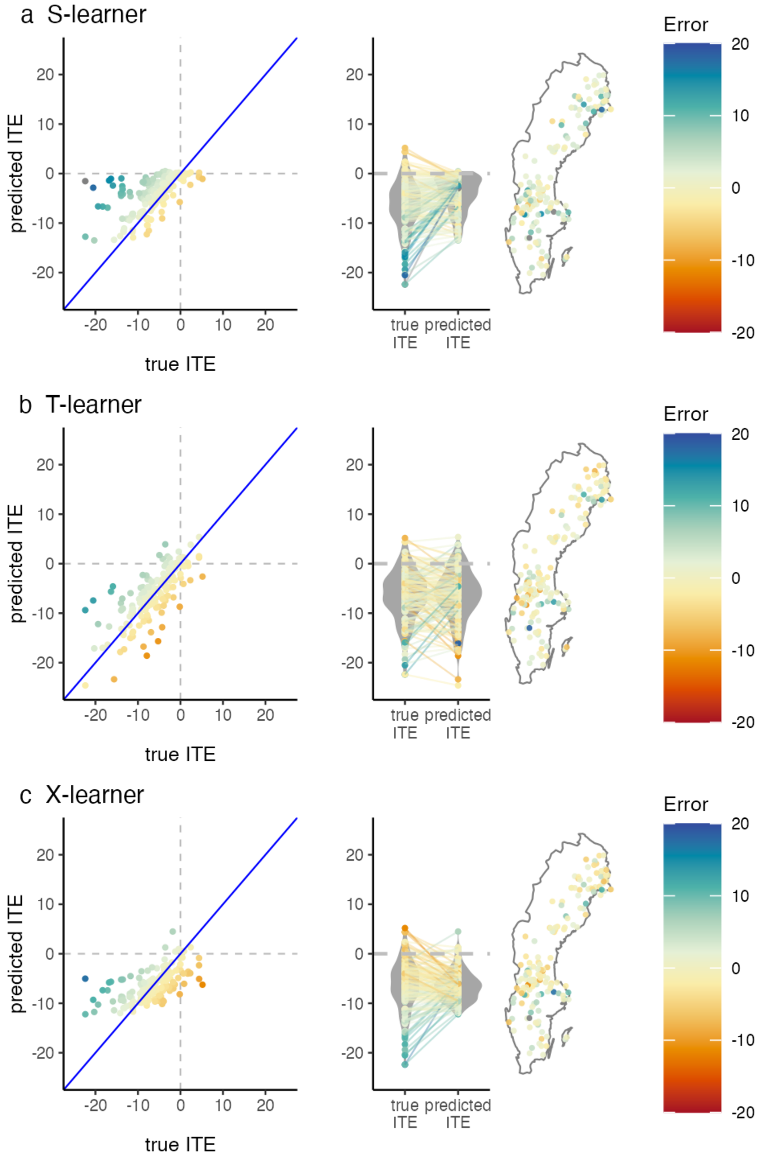


**Figure 3. Individual treatment effect (ITE) predictions obtained using three different meta-learner algorithms (top to bottom).** The first column of plots show the predicted ITE values against the true ITEs for individual National Forest Inventory plots in a test dataset spanning all of Sweden. Predictions 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 conditions were: treatment assignment = random, sample imbalance = 0.5, sample size = 1000, location of test plots = random, important variable omitted = FALSE.

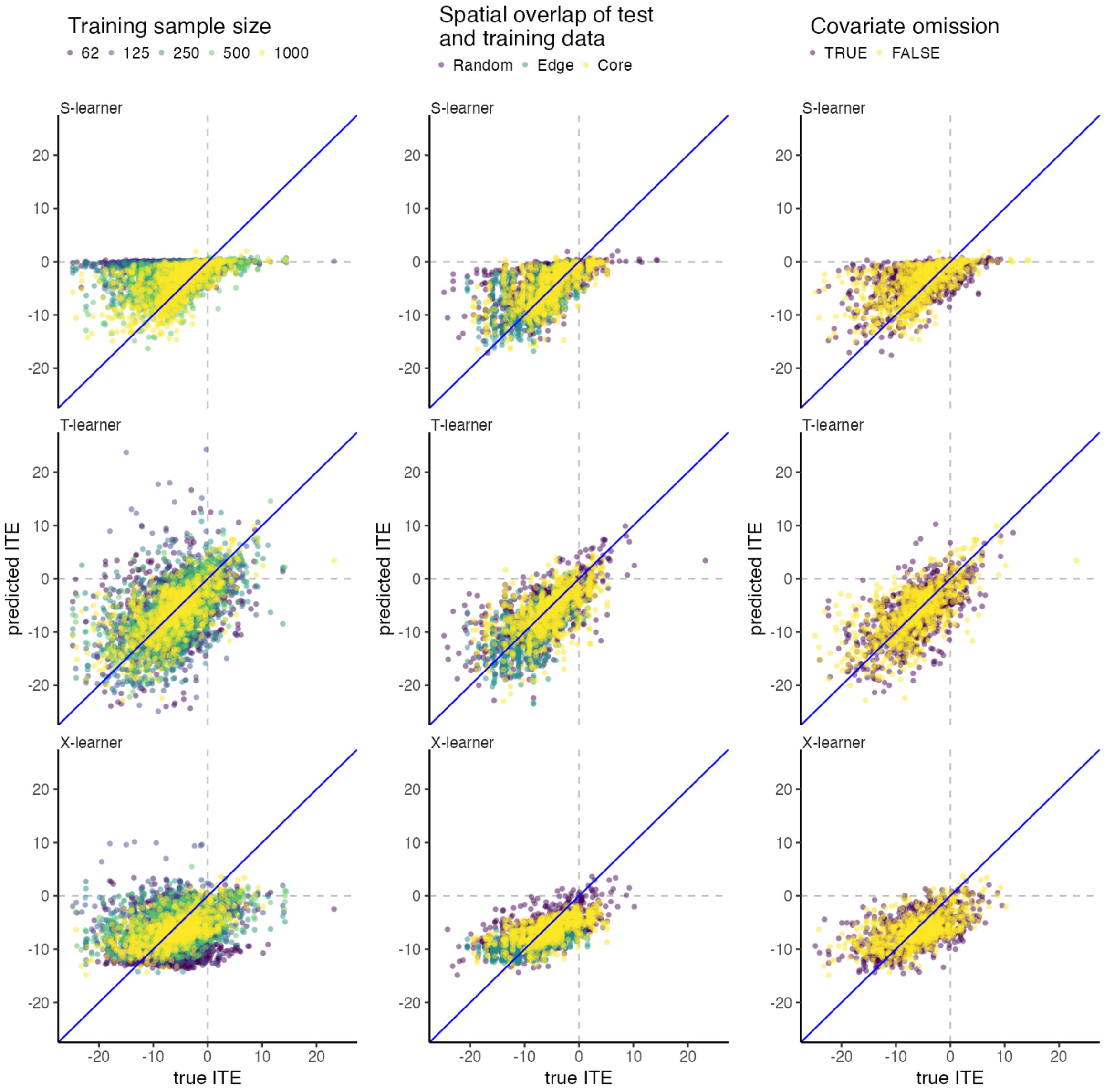


**Figure 4. Accuracy of individual treatment effect (ITE) predictions under different study features.** Each panel shows the predicted ITE against the true ITE, with each point corresponding to a single test data unit (National Forest Inventory plot). Predictions with zero error lie on the diagonal blue line. Each column of panels displays model predictions when one of three study features are varied (indicated by colour of points): sample size (left), location of test data (middle) and variable omission (right). Rows display results from three different meta-learner algorithms. Other than the three study features varied in each figure, the remaining study features were kept at: treatment assignment = random, sample imbalance = 0.5, sample size = 1000, location of test plots = random, important variable omitted = FALSE.

## Discussion

Here we present a first application in ecology of meta-learners for ITE prediction, and compare the relative performance of three meta-learners across a broad set of sampling and modelling conditions that are common to ecological observational studies. Our primary finding is the relative performance of the different meta-learners (in terms of the error) depend on the specific data setting. In general, the T-Learners and X-Learners that are increasingly applied in human-centred disciplines, such as marketing and medicine, produce consistently more accurate ITE predictions than the S-Learner approach that is typical to ecology. We provide guidance 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 performed consistently poorly; ITEs were underestimated and tended to be shrunk towards zero, a finding consistent with simulation studies from other disciplines (Künzel et al., 2019a; Okasa, 2022a). Since the treatment indicator is treated like any other covariate and plays no special role in the S-Learner, using the S-Learner in settings where treatment status is not a strong predictor of the outcome variable can be problematic. 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, 2022a). 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., 2019a; Salditt et al., 2024b)).

Since a S-Learner fits a single regression, it does not account for potentially varying distributions of the covariates across treatment levels, i.e., as a result of selection bias. Using S-Learners, the ITE cannot be interpreted as a causal prediction unless assumptions have been met through design, i.e., randomised treatment assignment. In our simulations, the treatment assignment mechanism (whether randomised or correlated with altitude or latitude) 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 we imposed with our treatment assignment mechanisms (treatment assignment was correlated with region or altitude, Table1). In reality, a greater degree of covariate imbalance could occur. Nevertheless, even when treatment assignment was random, and covariate distributions across control and treatment groups were thus 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 the outcome surface complexity is very different across the two groups. In other words, when the ITE function is more complex than either of outcome prediction functions.
What 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, to predict outcomes (species abundance), outcome-predictor relationships are often interpreted causally *posthoc*, 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.

Two-model approaches (T- and X-Learners) performed consistently better than the single-model approach, except for the lowest sample size category ($R^2$ Figure 1d). In contrast to the S-Learner, the two-model approaches do not suffer from the regularisation on treatment status, because the outcomes are modelled separately for each group. The comparable error of the approaches at the lowest sample sizes 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.

While the single-model approach performed consistently poorer than T-Learner and X-Learner approaches at moderate to larger sample size, the differences in the relative performance of the T-Learner and X-Leaner meta-learners are more nuanced, and become more pronounced the greater the sample imbalance, and when an important covariate was omitted. 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., 2024b). We found that X-Learner ITE predictive accuracy was less sensitive to sample imbalance than T-Learners. With sample 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 sample imbalance depends on which treatment group was smaller. This might suggest a more complex functional form was necessary to predict the outcome in the treated group, necessitating larger sample sizes. Curth *et al.* (2024) note that when prediction the potential outcomes separately for each treatment groups, 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.

The X-Learner was developed to overcome limitations of the S-Learner and the T-Learner, and to perform well for imbalanced samples, and whether the ITE is simple or complex in form. 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 then the S- and T- Learners. The intermediate treatment effects and propensity scores that are computed in the X-Learner approach 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.

## *Conclusions and future directions*

Further simulation studies are needed to inform on rules of thumb regarding the choice of meta-learner algorithm for different applied empirical settings, with different data structures inherent to observational studies (Okasa, 2022a). 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 a selection bias wherein 'treatments' (e.g., protected area status) are confounded with other covariates (e.g., slope), 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 species-rich areas (Mentges et al., 2021).

Here we have compared the performance of just three popular meta-learners, and employed random forests as their base learners. Numerous other meta-learners have been proposed in the literature, including the 'doubly-robust' DR-Learner (Kennedy, 2023) and the 'residualisation' R-Learner (Nie & Wager, 2021), which are extensions of the X-Learner. All meta-learners can use other machine learning methods 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., 2019a; Okasa, 2022a).

Future research could evaluate the effect of choice of covariates to use to construct the propensity scores. Literature suggests that scientific causal knowledge should help determine if a variable might be a confounder as opposed to leaving this to a purely statistical exercise that optimise model fit (Caron et al., 2022b). Propensity score models have long been known to be highly sensitive to model misspecification (J. Lee et al., 2024). In practice, and as was done in this study, the set of covariates that are used in models for outcome prediction are typically the same as those used to estimate the propensity scores (e.g., Künzel et al., 2019a; Salditt et al., 2024b). A sound understanding of how covariate choice for propensity score models influences 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 forest dynamics model. 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. Curth *et al.* (2021) found that simulation runs in which more covariates have large nonzero effects yielded higher heterogeneity in ITEs.

## 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].

## 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, Edenius, Elfving, Eriksson, Tomas, 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.

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. *Forest Ecology and Management*, *174*(1), 25–37. https://doi.org/10.1016/S0378-1127(02)00025-7

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, *46*(3), 399–424. https://doi.org/10.1080/00273171.2011.568786

Bernard, S., Heutte, L., & Adam, S. (2009). Influence of hyperparameters on random forest accuracy. In J. A. Benediktsson, J. Kittler, & F. Roli (Eds.), *Multiple Classifier Systems* (pp. 171–180). Springer. https://doi.org/10.1007/978-3-642-02326-2_18

Caron, A., Baio, G., & Manolopoulou, I. (2021). *Shrinkage Bayesian causal forests for heterogeneous* *treatment effects estimation* (No. arXiv:2102.06573). arXiv. https://doi.org/10.48550/arXiv.2102.06573

Caron, A., Baio, G., & Manolopoulou, I. (2022a). Estimating Individual Treatment Effects using Non-Parametric Regression Models: A Review. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *185*(3), 1115–1149. https://doi.org/10.1111/rssa.12824

Caron, A., Baio, G., & Manolopoulou, I. (2022b). Estimating individual treatment effects using non-parametric regression models: A review. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *185*(3), 1115–1149. https://doi.org/10.1111/rssa.12824

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. *Forest Ecology and Management*, *503*, 119820. https://doi.org/10.1016/j.foreco.2021.119820

Cheung, M., Dimitrova, A., & Benmarhnia, T. (2024). *An Overview of Modern Machine Learning Methods for Effect Measure Modification Analyses in High-Dimensional Settings* (No. arXiv:2401.15257). arXiv. https://doi.org/10.48550/arXiv.2401.15257

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. *Clinical Pharmacology & Therapeutics*, *115*(4), 710–719. https://doi.org/10.1002/cpt.3159

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

Fernández-Loría, C., & Provost, F. (2022). Causal Classification: Treatment Effect Estimation vs. Outcome Prediction. *Journal of Machine Learning Research*, *23*(59), 1–35.

Greenwell, B. M., & Boehmke, B. C. (2020). Variable Importance Plots—An Introduction to the vip Package. *The R Journal*, *12*(1), 343–366. https://doi.org/10.32614/RJ-2020-013

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

Holl, K. D., & Brancalion, P. H. S. (2020). Tree planting is not a simple solution. *Science*, *368*(6491), 580–581. https://doi.org/10.1126/science.aba8232

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960. https://doi.org/10.1080/01621459.1986.10478354

Jackson, E. E., Snäll, T., Gardner, E., Bullock, J. M., & Spake, R. (2024). *Towards causal predictions of site-level treatment effects in applied ecology* (Version 1.0.0) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.13269917

Kennedy, E. H. (2023). Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, *17*(2), 3008–3049. https://doi.org/10.1214/23-EJS2157

Kimmel, K., Dee, L. E., Avolio, M. L., & Ferraro, P. J. (2021). Causal assumptions and causal inference in ecological experiments. *Trends in Ecology & Evolution*, *36*(12), 1141–1152. https://doi.org/10.1016/j.tree.2021.08.008

Knaus, M. C., Lechner, M., & Strittmatter, A. (2021). Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence. *The Econometrics Journal*, *24*(1), 134–161. https://doi.org/10.1093/ectj/utaa014

Kuhn, M., Vaughan, D., & Hvitfeldt, E. (2023). *yardstick: Tidy Characterizations of Model Performance* (Version 1.2.0) [Computer software]. https://github.com/tidymodels/yardstick

Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019a). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, *116*(10), 4156–4165. https://doi.org/10.1073/pnas.1804597116

Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019b). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, *116*(10), 4156–4165. https://doi.org/10.1073/pnas.1804597116

Lee, J., Huling, J. D., & Chen, G. (2024). An effective framework for estimating individualized treatment rules. *Advances in Neural Information Processing Systems*, *37*, 8411–8476. https://openreview.net/pdf?id=G7L65B2P0y

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*(12), Article 12.

https://doi.org/10.3390/f11121318

Lindenmayer, D. B., & Laurance, W. F. (2012). A history of hubris – Cautionary lessons in ecologically sustainable forest management. *Biological Conservation*, *151*(1), 11–16. https://doi.org/10.1016/j.biocon.2011.10.032

Liu, Y., Trancoso, R., Ma, Q., Ciais, P., Gouvêa, L. P., Yue, C., Assis, J., & Blanco, J. A. (2023). Carbon density in boreal forests responds non-linearly to temperature: An example from the Greater Khingan Mountains, northeast China. *Agricultural and Forest Meteorology*, *338*, 109519. https://doi.org/10.1016/j.agrformet.2023.109519

Mayer, M., Prescott, C. E., Abaker, W. E. A., Augusto, L., Cécillon, L., Ferreira, G. W. D., James, J., Jandl, R., Katzensteiner, K., Laclau, J.-P., Laganière, J., Nouvellon, Y., Paré, D., Stanturf, J. A., Vanguelova, E. I., & Vesterdal, L. (2020). Tamm Review: Influence of forest management activities on soil organic carbon stocks: A knowledge synthesis. *Forest Ecology and Management*, *466*, 118127. https://doi.org/10.1016/j.foreco.2020.118127

Mazziotta, A., Lundström, J., Forsell, N., Moor, H., Eggers, J., Subramanian, N., Aquilué, N., Morán-Ordóñez, A., Brotons, L., & Snäll, T. (2022a). More future synergies and less trade-offs between forest ecosystem services with natural climate solutions instead of bioeconomy solutions. *Global Change Biology*, *28*(21), 6333–6348. https://doi.org/10.1111/gcb.16364

Mazziotta, A., Lundström, J., Forsell, N., Moor, H., Eggers, J., Subramanian, N., Aquilué, N., Morán-Ordóñez, A., Brotons, L., & Snäll, T. (2022b). More future synergies and less trade-offs between forest ecosystem services with natural climate solutions instead of bioeconomy solutions. *Global Change Biology*, *28*(21), 6333–6348. https://doi.org/10.1111/gcb.16364

Mentges, A., Blowes, S. A., Hodapp, D., Hillebrand, H., & Chase, J. M. (2021). Effects of site-selection bias on estimates of biodiversity change. *Conservation Biology*, *35*(2), 688–698. https://doi.org/10.1111/cobi.13610

Moor, H., Eggers, J., Fabritius, H., Forsell, N., Henckel, L., Bradter, U., Mazziotta, A., Nordén, J., & Snäll, T. (2022). Rebuilding green infrastructure in boreal production forest given future global wood demand. *Journal of Applied Ecology*, *59*(6), 1659–1669. https://doi.org/10.1111/1365-2664.14175

Moyano, J., Dimarco, R. D., Paritsis, J., Peterson, T., Peltzer, D. A., Crawford, K. M., McCary, M. A., Davis, K. T., Pauchard, A., & Nuñez, M. A. (2024). Unintended consequences of planting native and non-native trees in treeless ecosystems to mitigate climate change. *Journal of Ecology*, *112*(11), 2480–2491. https://doi.org/10.1111/1365-2745.14300

Nie, X., & Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, *108*(2), 299–319. https://doi.org/10.1093/biomet/asaa076

Okasa, G. (2022a). *Meta-learners for estimation of causal effects: Finite sample cross-fit performance* (No. arXiv:2201.12692). arXiv. https://doi.org/10.48550/arXiv.2201.12692

Okasa, G. (2022b). *Meta-Learners for Estimation of Causal Effects: Finite Sample Cross-Fit Performance* (No. arXiv:2201.12692). arXiv. https://doi.org/10.48550/arXiv.2201.12692

Pearl, J. (2018). *Theoretical impediments to machine learning with seven sparks from the causal revolution* (No. arXiv:1801.04016). arXiv. https://doi.org/10.48550/arXiv.1801.04016

Probst, P., Wright, M. N., & Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery*, *9*(3), e1301. https://doi.org/10.1002/widm.1301

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), 913–929. https://doi.org/10.1111/ecog.02881

Salditt, M., Eckes, T., & Nestler, S. (2024a). A Tutorial Introduction to Heterogeneous Treatment Effect Estimation with Meta-learners. *Administration and Policy in Mental Health*, *51*(5), 650–673. https://doi.org/10.1007/s10488-023-01303-9

Salditt, M., Eckes, T., & Nestler, S. (2024b). A tutorial introduction to heterogeneous treatment effect estimation with meta-learners. *Administration and Policy in Mental Health and Mental Health Services Research*, *51*(5), 650–673. https://doi.org/10.1007/s10488-023-01303-9

Spake, R., Bellamy, C., Graham, L. J., Watts, K., Wilson, T., Norton, L. R., Wood, C. M., Schmucki, R., Bullock, J. M., & Eigenbrod, F. (2019). An analytical framework for spatially targeted management of natural capital. *Nature Sustainability*, *2*(2), 90–97. https://doi.org/10.1038/s41893-019-0223-4

Spake, R., Jackson, E. E., Bullock, J. M., Gardner, E., Tipton, E., Grainger, M. J., & Doncaster, C. P.

(2025). Precision ecology for targeted conservation action. *Nature Ecology & Evolution*. https://doi.org/10.1038/s41559-025-02733-4

Spake, R., O'Dea, R. E., Nakagawa, S., Doncaster, C. P., Ryo, M., Callaghan, C. T., & Bullock, J. M. (2022). Improving quantitative synthesis to achieve generality in ecology. *Nature Ecology & Evolution*, *6*(12), Article 12. https://doi.org/10.1038/s41559-022-01891-z

Tipton, E., & Mamakos, M. (2023). *Designing randomized experiments to predict unit-specific treatment effects* (No. arXiv:2310.18500). arXiv. https://doi.org/10.48550/arXiv.2310.18500

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 in Ecology and Evolution*, *10*(2), 225–232. https://doi.org/10.1111/2041-210X.13107

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. *Global Change Biology*, *18*(3), 1028–1041. https://doi.org/10.1111/j.1365-2486.2011.02606.x

Wikström, P., Edenius, L., Elfving, B., Eriksson, L. O., Tomas, L., Sonesson, J., Öhman, K., Wallerman, J., Waller, C., & Klintebäck, F. (2011). The Heureka forestry decision support system: An overview. *Mathematical and Computational Forestry & Natural-Resource Sciences (MCFNS)*, *3*(2), 87–95.

Wikström, P., Edenius, L., Elfving, B. O., Eriksson, L. O., Lämås, T., Johan, S., Öhman, K., Wallerman, J., Waller, C., & Klintebäck, F. (2011). The Heureka Forestry Decision Support System: An Overview. *Mathematical and Computational Forestry and Natural-Resource Sciences*, *3*(2). https://res.slu.se/id/publ/36761

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Yates, K. L., Bouchet, P. J., Caley, M. J., Mengersen, K., Randin, C. F., Parnell, S., Fielding, A. H., Bamford, A. J., Ban, S., Barbosa, A. M., Dormann, C. F., Elith, J., Embling, C. B., Ervin, G. N., Fisher, R., Gould, S., Graf, R. F., Gregr, E. J., Halpin, P. N., … Sequeira, A. M. M. (2018). Outstanding

*Preprint*

# Supporting Information

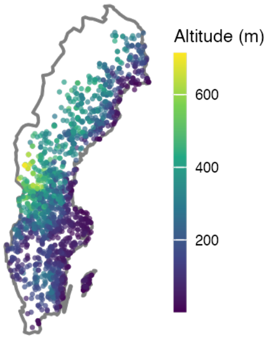
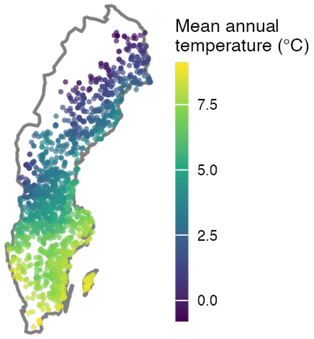
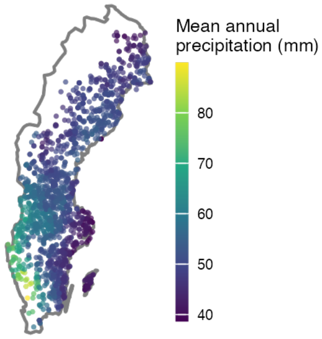
## Towards causal predictions of site-level treatment effects for applied ecology

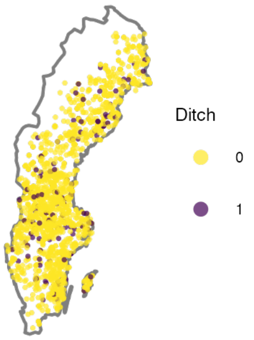
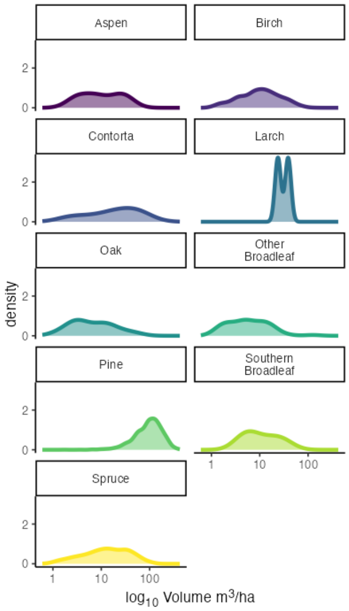
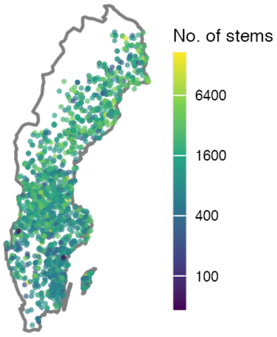
**Eleanor E. Jackson[1,2], Tord Snäll[3,4], Emma Gardner[5], James M. Bullock[5] & Rebecca Spake[1,6]**

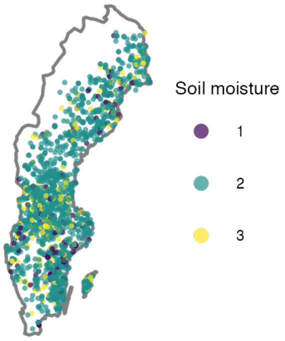
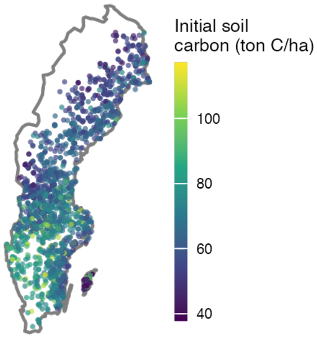
1 School of Biological Sciences, University of Reading, UK
2 Department of Biology, University of Oxford, UK
3 Skogforsk, Uppsala Science Park, Uppsala, Sweden
4 SLU Swedish Species Information Centre, Sweden
5 UK Centre for Ecology & Hydrology, Wallingford, UK
6 School of Geography and Environmental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

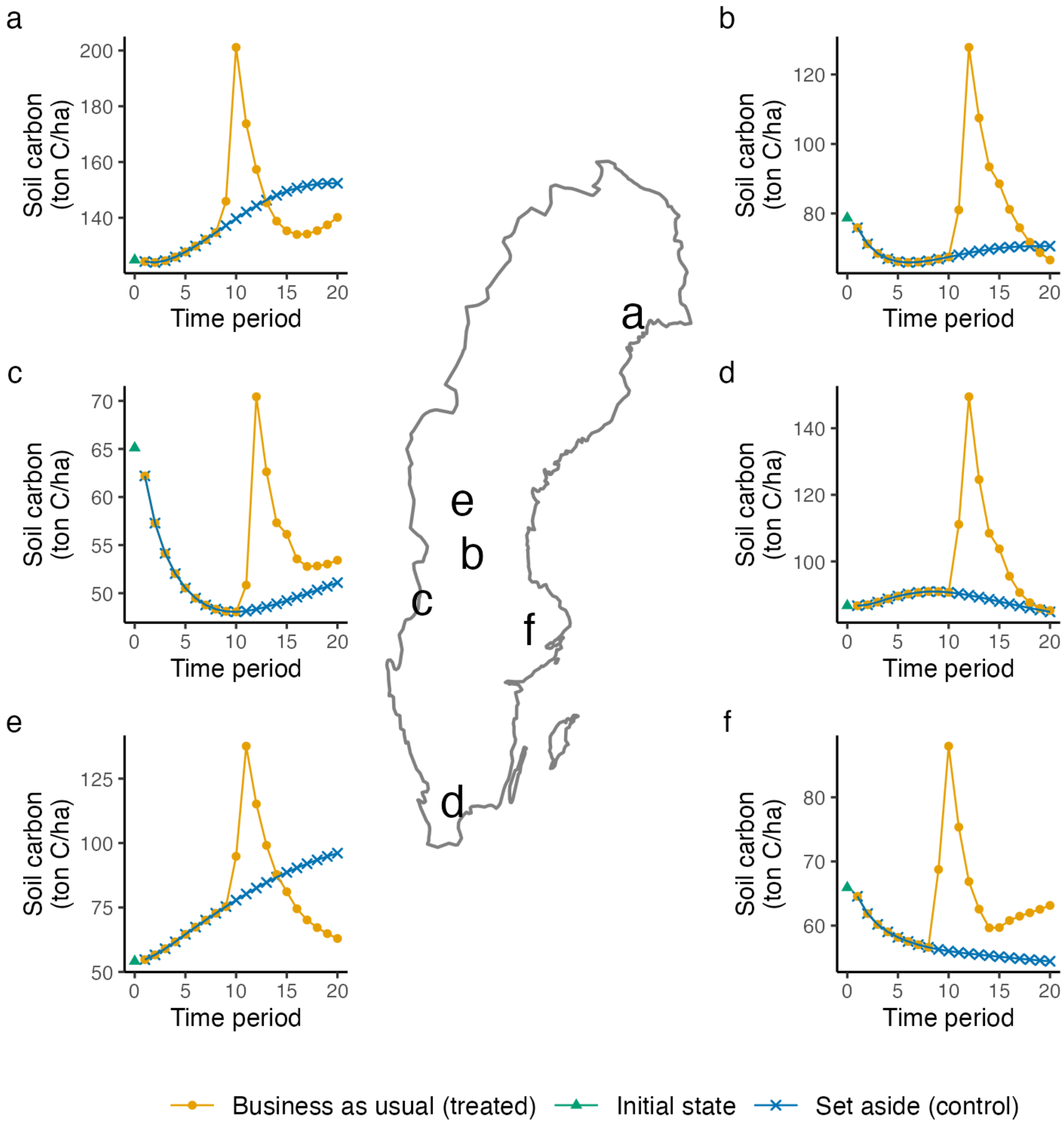
**Corresponding Author:**
Rebecca Spake, School of Geography and Environmental Science, Building 44, Highfield, University of Southampton, Hampshire, SO17 1BJ, UK. Email: r.spake@soton.ac.uk

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

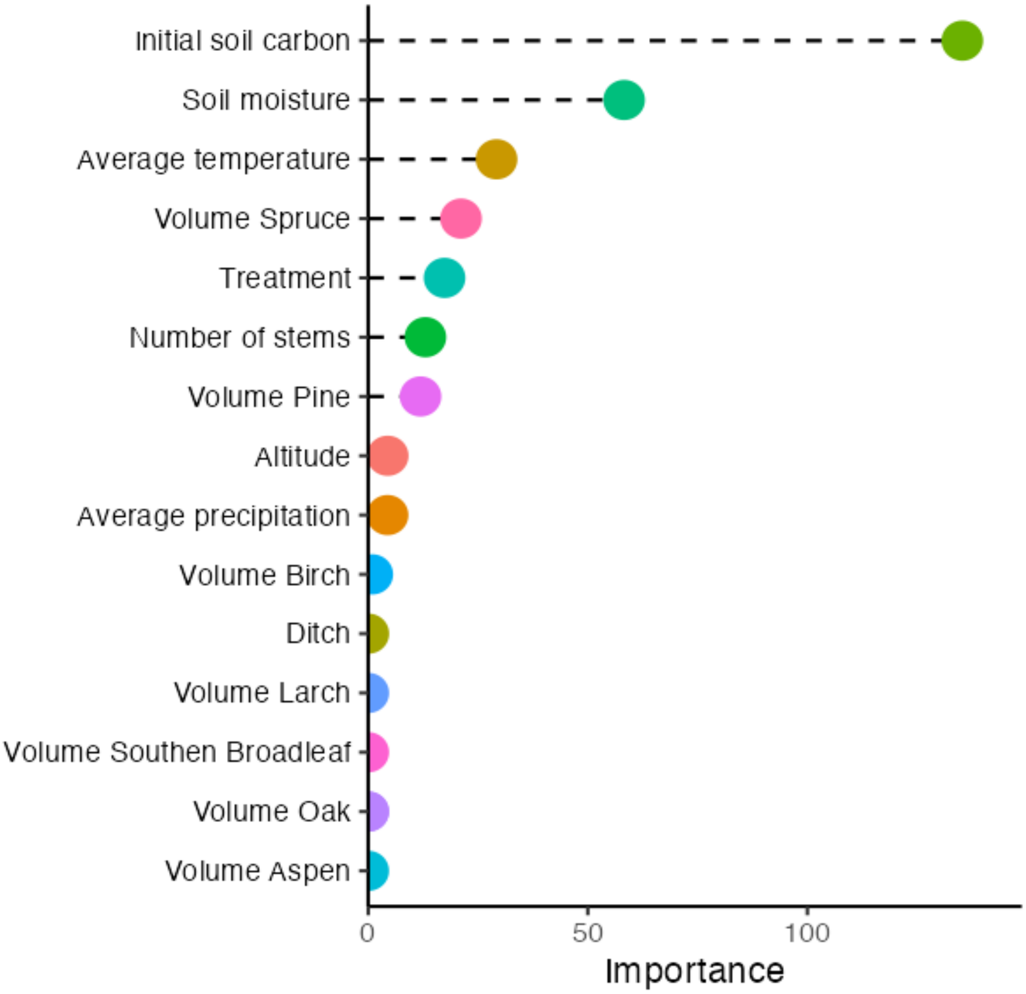
| Covariate | Details |
|---|---|
| Altitude<br><br> | Height above sea level, sourced from the Swedish National Forest Inventory. |
| Temperature<br><br> | Mean annual temperature, sourced from CRU TS (Climatic Research Unit gridded Time Series) (v. 4.07) (Harris et al., 2020). Plots were matched to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

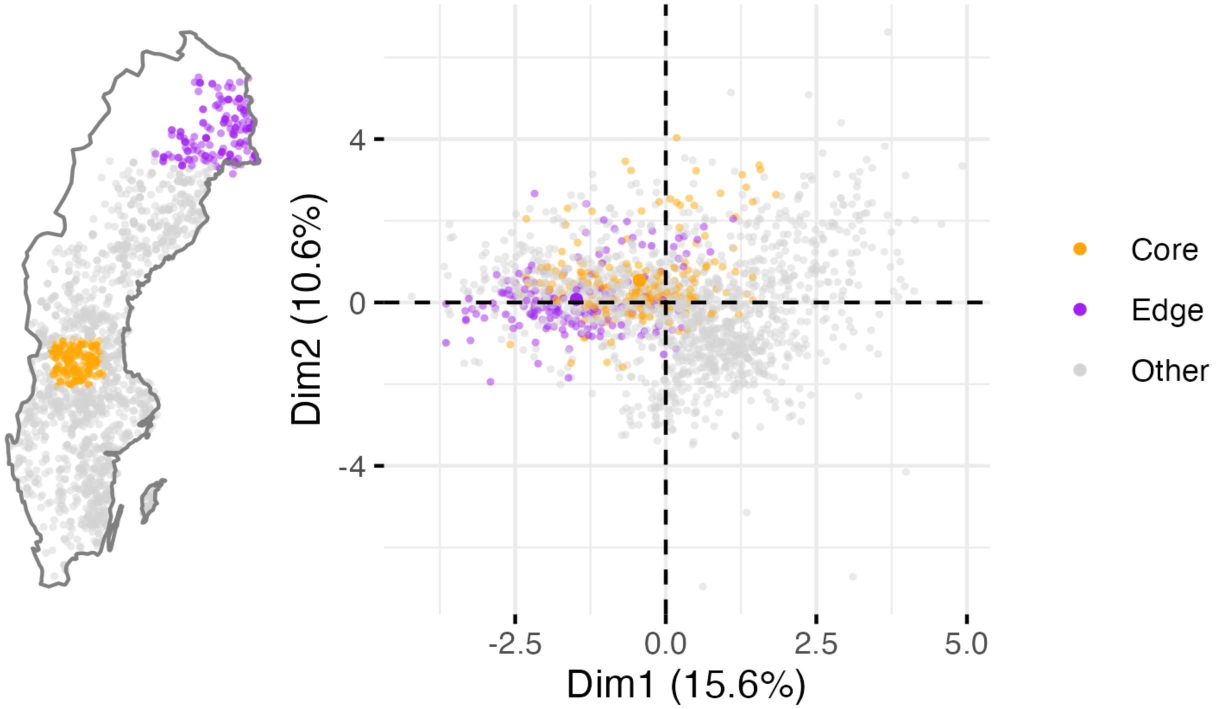
| | |
|---|---|
| Soil moisture<br><br> | An ordinal variable ranging from 1 (dry) to 3 (mesic-moist), sourced from the Swedish National Forest Inventory. (Note this only contains plots included in our study - we did not include plots with soil moisture 4 (moist) or 5 (wet) plots see methods.) |
| Initial 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. |



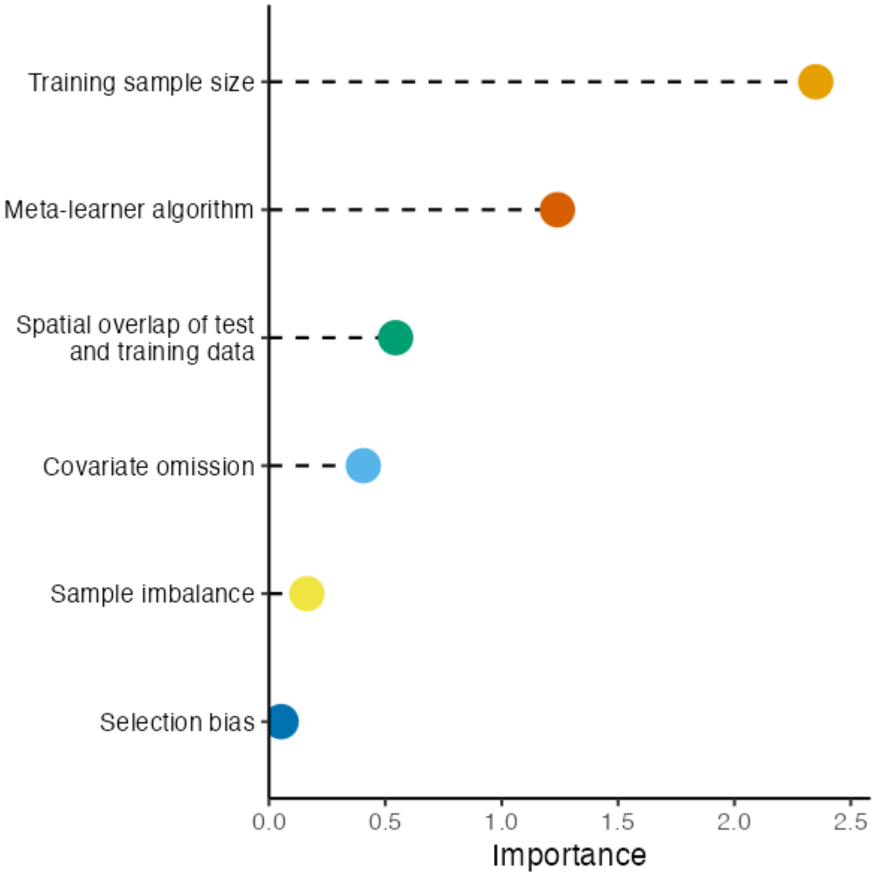
**SI Figure 1. Simulated soil carbon over time for six National Forest Inventory Plots.** Figure labels (a - f) are plotted on a map of Sweden to indicate their location. For the first year, soil carbon is plotted at its "initial state" (green triangle), this is the value measured during National Forest 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. 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 challenges in the transferability of ecological models. *Trends in Ecology & Evolution*, *33*(10), 790–802. https://doi.org/10.1016/j.tree.2018.08.001

Zurell, D., Berger, U., Cabral, J. S., Jeltsch, F., Meynard, C. N., Münkemüller, T., Nehrbass, N., Pagel, J., Reineking, B., Schröder, B., & Grimm, V. (2010). The virtual ecologist approach: Simulating data and observers. *Oikos*, *119*(4), 622–635. https://doi.org/10.1111/j.1600-0706.2009.18284.x