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

Eleanor E. Jackson, eleanor.jackson@biology.ox.ac.uk
School of Biological Sciences, University of Reading, Reading, UK
Department of Biology, University of Oxford, Oxford, UK
https://orcid.org/0000-0002-9884-2972

Tord Snäll, tord.snall@skogforsk.se
The Forestry Research Institute of Sweden (Skogforsk), SE-75183, Uppsala, Sweden
SLU Swedish Species Information Centre, Swedish University of Agricultural Sciences, Uppsala, Sweden
https://orcid.org/0000-0001-5856-5539

Emma Gardner, emmgar@ceh.ac.uk
UK Centre for Ecology & Hydrology, Wallingford, UK
https://orcid.org/0000-0002-1669-7151

James M. Bullock, jmbul@ceh.ac.uk
UK Centre for Ecology & Hydrology, Wallingford, UK
https://orcid.org/0000-0003-0529-4020

Rebecca Spake, r.spake@soton.ac.uk
School of Biological Sciences, University of Reading, Reading, UK
School of Geography and Environmental Science, University of Southampton, UK
https://orcid.org/0000-0003-4671-2225

Authors' contributions: R.S. conceived the idea which was developed with input from all authors. T.S. provided the simulation data from Heureka and its interpretation. E.E.J. and R.S. developed the methodology. E.E.J. led modelling work and analysed output data with input from R.S. E.E.J. and R.S. led the writing. All authors contributed substantially to results interpretation and to manuscript drafts.

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. 2025) 10.5281/zenodo.13269917.

Competing Interests: We have no competing interests.

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

Corresponding author: Rebecca Spake, r.spake@soton.ac.uk, School of Geography and Environmental Science, University of Southampton, UK

Abstract

Targeted environmental management requires knowing where interventions will be most effective. Individual treatment effects (ITEs), which predict how each site would respond to alternative interventions, could help direct limited conservation resources to where management is expected to deliver the greatest benefit. Causal machine learning methods such as meta-learners can predict ITEs, yet most evidence on their performance comes from fields with much larger datasets than are typical in ecology, and with different evaluation criteria. We provide the first decision-relevant test of meta-learners for ecological ITE prediction, using forest-management simulations in which true site-level ITEs are known. We evaluated four meta-learners (S-, T-, X- and DR-learners) across 21,600 virtual observational studies varying in sample size, treatment imbalance, treatment assignment, spatial overlap between training and test data, and covariate omission. We assessed performance using separate metrics for two tasks: regional prioritisation, where accurate site ranking is critical, and local decision-making, where accurate effect-size estimation is required. The X-learner performed best on average, but relative performance varied with study conditions and evaluation metric. Our results show how ecological use of meta-learners can be guided by decision context, sample size and causal assumptions.

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). Targeted and tailored management therefore requires causal predictions of how individual sites will respond to alternative interventions, hereafter 'individual treatment effects' (ITEs; Spake *et al.* 2025). Average treatment effects can tell us whether an intervention works overall, but they are less actionable when effects are highly variable. To provide ITE predictions, there is increasing interest in the use of causal machine learning (ML) methods developed in human-centred fields such as medicine and marketing, where they inform personalised medicine and targeted advertising (Spake *et al.* 2025; Tipton & Mamakos 2023). Whereas traditional ML predicts outcomes for individual units, causal ML quantifies how outcomes change in response to treatment. These methods could support targeted environmental management, but their performance under ecological data conditions remains largely untested.

Estimating any causal effect is difficult because treatment effects cannot be observed directly. For any unit of observation, we observe only the outcome under the treatment it received, not the counterfactual outcome it would have had under another treatment: the 'fundamental problem of causal inference' (Holland 1986). Causal interpretation also requires assumptions such as no unmeasured confounding, meaning no unobserved factor drives both treatment assignment and the outcome. ITE prediction is harder than ATE estimation, because effects must be recovered across covariate space rather than averaged across all sites. It requires sufficient overlap, or positivity, throughout the covariate space, and enough data to learn potentially complex heterogeneity in treatment effects. Causal ML can flexibly model nonlinear relationships between covariates, treatments and outcomes (Feuerriegel *et al.* 2024), but this flexibility increases data demands. These demands are likely to be especially challenging in ecological datasets, where management is often non-random, spatially structured and applied to relatively small samples.

A family of causal ML methods, known as meta-learners, has been developed to predict ITEs (Curth & Schaar 2021; Kennedy 2023; Künzel *et al.* 2019). These methods differ in how they handle treatment assignment, group size imbalances, and model misspecification risk, and therefore in how well they perform under non-random treatment assignment, treatment imbalance and limited overlap. Spake *et al.* (2025) and Arif (2025) recently argued that causal ML methods hold promise for informing environmental management, but their

performance under ecological data conditions remains unclear. Importantly, their utility depends on the decision that the ITE is used to inform: the same set of predicted ITEs may be adequate for ranking sites for intervention, but inadequate for estimating the effect expected at a particular site.

Individual treatment effects could inform environmental management decisions at two scales. First, at the regional scale, a decision-maker with a limited resource budget must choose which sites to prioritise for intervention. This is a ranking problem: what matters is whether ITE predictions correctly order sites by their true treatment effects, not necessarily whether the predicted magnitudes are exact. Second, at the local scale, a manager asks what will happen if they intervene at a particular site. This is a magnitude problem: what matters is whether ITE predictions are accurate in the units of the outcome. Importantly, these criteria could diverge: a meta-learner may rank sites well while inaccurately predicting the size of their effects or estimate effect magnitudes reasonably while failing to identify the highest-priority sites.

Several recent studies have assessed the performance of meta-learners across data-generating processes typical of human-centred observational studies, varying sample size, treatment imbalance and the complexity of the treatment-effect function (e.g., whether linear or nonlinear; Künzel *et al.* 2019; Okasa 2022; Zhang *et al.* 2025). General guidance is emerging; for example, meta-learners that model both outcomes and treatment assignment can perform well under some forms of imbalance or confounding, although no single meta-learner consistently outperforms others across all scenarios (Knaus *et al.* 2021; Okasa 2022; Singh 2026). Recent simulation studies continue to emphasise that relative meta-learner performance is context-dependent, varying with data-generating process, sample size and evaluation metric (Curth *et al.* 2021). The transferability of these insights to ecology is limited for two reasons. First, ecological studies are characterised by smaller sample sizes. Second, performance has primarily been assessed through estimation errors: the difference between true and predicted ITEs for each unit. This is relevant to understanding how well predictions might inform local decisions, but not sufficient for evaluating their utility for regional targeting decisions that are important in conservation planning. The performance of meta-learners for ecological prioritisation is therefore unknown. Here, we translate meta-learners to ecology and evaluate their performance under conditions realistic to ecological studies, using decision-relevant criteria.

To evaluate whether meta-learners can predict ITEs well, we used simulated forest-management outcomes for Swedish National Forest Inventory plots. Because we simulated soil carbon development under both set-aside and business-as-usual management for each

site, we could calculate true ITEs while emulating observational studies in which each site is observed under only one assigned management condition. We varied study conditions, chosen to reflect ecological challenges: sample size, treatment imbalance, non-random treatment assignment, spatial differences between training and test data, and omission of a confounder. We focus on soil carbon because it is important in decision-making for climate change mitigation. In this context, we defined the target intervention (or “treatment”) as excluding a site from harvest; the aim is to protect the sites where business-as-usual clearcutting would otherwise lead to the greatest soil-carbon loss.

We expected that meta-learners that explicitly model both treatment assignment and outcomes would rank and estimate ITEs more accurately than simpler meta-learners; that this advantage would be greatest under non-random treatment assignment and confounder omission; and that this advantage would erode at smaller sample sizes, where additional modelling is less well supported by the data. By evaluating meta-learners against decision-relevant criteria, we aim to give applied ecologists a clearer basis for considering these methods in targeted environmental management decisions.

Methods

We implemented a ‘virtual ecologist’ (Zurell et al. 2010) approach consisting of four key steps: (1) we simulated local forest dynamics under treatment (protection) and control (business-as-usual clearcutting) conditions across a large number of environmentally heterogeneous sites. By simulating both potential outcomes for each site – observed soil carbon mass for individual sites both with and without a treatment – we generated ‘true’ ITEs for each forest site, against which to compare the accuracy of predictions. (2) We varied the treatment assignment mechanism, systematically varying the probability of treatment assignment with environmental gradients. (3) These forest sites were then sampled in different ways by virtual observers to generate alternative ‘training’ datasets. Each unique combination of data-generating and sampling conditions (treatment assignment, sample size, treatment imbalance, covariate omission, and test-data location; table 1) constitutes a distinct experiment; within each experiment. (4) We then applied and compared four meta-learners to predict individual treatment effects for the withheld 'test' plots. We replicated each experiment 50 times to average over stochastic variation in sampling, treatment assignment, and model fitting, and compared predicted ITEs with the true ITEs to understand how data-generating, sampling, and modelling conditions (table 1) interact to influence meta-learner performance.

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

We used model simulation 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.* 2022a; 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 simulates 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 alternatives for 2010–2100: business-as-usual even-aged forestry excluding thinning, hereafter the control, and set-aside, hereafter the treatment. Set-aside stands were left to develop without intervention, whereas business-as-usual stands were clearcut and replanted at typical clearcutting stand stockings, with clearcutting ages ranging from 60–120 years: lower in southern high-productivity stands and higher in northern low-productivity stands. We defined ITEs as soil carbon under set-aside minus soil carbon under business-as-usual management; positive ITEs therefore indicate sites where protection retained more soil carbon than clearcutting.

Simulations were initialised with observed NFI data from 26,193 circular plots of 10 m radius recorded between 2016 and 2020, representing 23.5 Mha of productive forest land in Sweden. We used data from only pine-dominated stands, defined as ≥50% pine standing volume, that were 30–50 years old in 2016–2020, to limit variation in historical management. We further removed plots containing peatland and plots with soil-moisture classes four and five, due to their very low prevalence. This yielded 2,580 plots. Climate averages from 1983–1992 were held constant over the simulation period, corresponding to the standard Heureka setting and assuming no climate change.

Because the clearcutting treatment occurred at different times for different plots, we retained only plots with a single peak in maximum soil carbon resulting from a single clearcutting event during the observation period. Specifically, we filtered to plots that had reached

maximum soil carbon by the 12th time step after the start of the simulation, yielding 1,806 plots. This restriction reflects the fact that information on time since management action is often unavailable in observational studies. Soil carbon simulations for example NFI plots under control and treatment conditions are shown in figure S1.

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, corresponding to the twentieth time step. In our virtual ecologist approach, we systematically varied study conditions related to treatment assignment, sampling conditions and modelling conditions to quantify their influence on individual treatment effect, ITE, prediction performance. Each virtual study consisted of a particular combination of these conditions.

We implemented two treatment-assignment mechanisms. Under random assignment, treatment was assigned independently of covariates. Under non-random assignment, observed treatment status was correlated with soil moisture: plots on mesic-moist soils (soil-moisture class three) were assigned twice the probability of being observed under set-aside/protection as drier plots. This assignment rule was imposed to create a controlled scenario in which treatment assignment was correlated with a pre-treatment covariate that also influenced treatment-effect variation. It was not intended to reconstruct actual Swedish harvest decisions. Assignment was implemented using sampling weights in `dplyr::slice_sample()` (table 1).

After treatment assignment, we sampled plots for each virtual study. We varied two sampling conditions: total training sample size, with 250, 500 or 1000 NFI plots sampled across both treatment levels; and treatment imbalance, with 30%, 50% or 70% of sampled plots assigned to the control group.

We also varied three modelling conditions: (1) the meta-learner algorithm, comparing S-, T-, X-, and DR-learners; (2) covariate omission, fitting models with or without the omission of soil moisture; and (3) the spatial relationship between the training and test data, using three test-location scenarios. In the stratified-random scenario, test plots were distributed across Sweden and sampled using latitudinal deciles as strata. In the core scenario, test plots were selected from a distinct area in central Sweden, near the centre of the training data's multidimensional covariate space. In the edge scenario, test plots were selected from the

cooler, drier north-east, at the spatial and covariate-space periphery of the training data (figure S2).

Each unique combination of treatment-assignment mechanism, training sample size, treatment imbalance, test-location scenario and covariate-omission condition constituted an experimental condition. The full factorial design comprised 108 conditions: two treatment-assignment mechanisms, three training sample sizes, three treatment-imbalance levels, three test-location scenarios and two covariate-omission conditions. Each condition was replicated 50 times, and all four meta-learners were applied to the same replicated datasets, yielding 21,600 virtual studies in total: 108 experiments × 50 replicates × 4 meta-learners.

Causal assumptions and covariate selection

The estimand for each virtual study was the individual treatment effect function, or ITE surface, describing how the effect of set-aside management on soil carbon varied with pre-treatment covariates. We defined the ITE as soil carbon under set-aside/protection minus soil carbon under business-as-usual management. Positive ITEs therefore indicate sites where protection retained more soil carbon than management, while negative ITEs indicate sites where business-as-usual management resulted in higher soil carbon stocks than set-aside, which could occur if post-harvest regeneration or productivity effects outweighed carbon losses from harvesting over the simulation period. These predicted ITE surfaces were evaluated both as effect-size estimates and as prioritisation rules for identifying sites with the largest expected carbon-stock benefit from protection.

Before applying meta-learners, we specified the assumed causal structure of the system using directed acyclic graphs, DAGs (Feuerriegel *et al.* 2024). This step determines which variables can validly be adjusted for when estimating treatment effects (Correia *et al.* 2026). Valid causal identification requires adjusting for confounders (which influence both treatment assignment and the outcome), while avoiding adjustment for colliders (variables simultaneously influenced by treatment and outcome) or post-treatment mediators (variables that lie on the causal path between the treatment and the outcome (Montgomery et al. 2018). These requirements apply regardless of the estimator used; the meta-learners compared here differ in how they use covariates but rely on the same underlying causal assumptions.

We specified two DAGs: one for random treatment assignment and one for non-random assignment correlated with soil moisture (figure S5). Both DAGs assumed no colliders. to emulate observational studies, the DAGs were drawn using domain knowledge from

previous studies of soil carbon variation in relation to climate, topography, forest stand structure, soil conditions and management (e.g., Chen *et al.* 2022; Lee *et al.* 2020; Liu *et al.* 2023; Mazziotta *et al.* 2022b; Vayreda *et al.* 2012), rather than from consulting with the internal workings of the Heureka model, which simulates stand development under the two management regimes assuming constant climate conditions, and the resulting soil-carbon trajectories.

All covariates used for training were measured or defined before the assigned management condition in the virtual observational dataset. These included fixed or baseline variables such as climate averages, altitude, soil moisture, initial soil carbon and initial stand structure. We excluded post-treatment variables generated during the Heureka simulation, such as tree volume after harvest or later stand development, because these variables may lie on the causal pathway from management to soil carbon. We therefore estimated the total effect of set-aside management on soil carbon, including any indirect effects mediated through subsequent changes in stand structure and growth. Conditioning on post-treatment mediators would instead estimate a controlled or partial effect and would remove part of the management effect of interest (Montgomery *et al.* 2018).

We included all covariates listed in table S1, except in the covariate-omission scenario where soil moisture was excluded. Soil moisture was central to the confounding contrast. Under random assignment, soil moisture is an outcome predictor but not a confounder; its omission represents the loss of an informative predictor. Under soil-moisture-correlated assignment, soil moisture contributes to both observed treatment assignment and treatment-effect variation, so omitting soil moisture creates an unmeasured-confounding scenario in the virtual observational study.

Meta-learner implementation

We compared four meta-learner algorithms: general frameworks for estimating heterogeneous treatment effects using standard supervised-learning algorithms as base models. Two types of meta-learners are distinguished as plug-in and two-step approaches (Curth & Schaar 2021). We briefly describe their frameworks here, and direct readers to Künzel et al. (2019), Curth and van der Schaar (2021), and Kennedy (2020) for full details.

Plug-in meta-learners first estimate potential outcome functions and then obtain ITEs by differencing predicted outcomes under treatment and control. The S-learner fits one outcome model with treatment included as a covariate, whereas the T-learner fits separate outcome models for treated and control units (NFI plots). These approaches are simple and flexible, but do not explicitly model the treatment-assignment process, and the T-learner can be

sensitive to treatment imbalance because each outcome model is trained on only one treatment group.

The X- and DR-learners are two-step approaches that use an initial set of nuisance models to construct a second-stage treatment-effect estimate. The X-learner first imputes treatment effects within each treatment group using outcome models, then models these imputed effects as functions of covariates and combines them using the propensity score; it was designed to improve performance under treatment imbalance. The DR-learner (Kennedy 2023) combines outcome and propensity-score models to construct a doubly robust pseudo-outcome, which is then regressed on covariates to estimate ITEs. It is designed to reduce bias under non-random treatment assignment and is consistent if either the outcome model or the propensity model is correctly specified.

Thus, the four meta-learners span simpler outcome-regression approaches and methods that explicitly use information about the treatment-assignment process. The X- and DR-learners are expected to be more robust to treatment imbalance and non-random treatment assignment than plug-in meta-learners for observational datasets, although their performance still depends on the quality of the fitted outcome and propensity models.

Our implementation was adapted from the R code accompanying a meta-learner tutorial by Salditt *et al.* (2024). However, we modified the workflow for pseudo-outcome meta-learners. For the X- and DR-learners, we did not use out-of-fold or out-of-bag nuisance predictions when constructing second-stage pseudo-outcomes. Instead, nuisance functions were estimated using the full training sample within each virtual study. While sample-splitting and cross-fitting can improve performance in larger samples, full-sample estimation is recommended for smaller samples (Okasa 2022). Because our training samples were small, and treatment imbalance further reduced the effective sample size within treatment groups, we used full-sample nuisance estimation throughout.

We did not tune random forest hyperparameters, instead keeping settings fixed across all meta-learners and study conditions. Tuning hyperparameters separately for each meta-learner and condition would confound differences among meta-learners or study conditions with differences in tuning. In addition, reliable tuning requires additional validation or cross-validation data, which is difficult in the small-sample settings considered here (Künzel *et al.* 2019). Our conclusions therefore concern relative meta-learner performance under a common modelling workflow, rather than absolute performance under optimally tuned models.

Base random forest models were fitted using 500 trees. At each node, four predictor variables were randomly considered as candidates for splitting (the square root of the total number of predictors, 16), observations were drawn with replacement, and the minimum terminal node size was set to five observations. These hyperparameters were selected as sensible default settings typically used in the literature (Probst *et al.* 2019).

For each virtual study, each meta-learner was used to predict unit-specific treatment effects representing the effect of forest management on soil carbon after 20-time steps, or 100 years, for NFI plots in the test dataset.

Evaluating ITE prediction performance

For each of the 21,600 virtual studies, each with 162 test plots, we calculated true ITEs as the difference between simulated soil carbon under set-aside treatment and business-as-usual control management at year 2100; positive ITEs therefore indicate sites where protection retained more soil carbon than clearcutting. We computed four complementary, decision-relevant error metrics, each oriented so that lower values indicate better performance (described in table 2). Targeting error and imprecision quantify errors relevant to regional prioritisation, where the aim is to identify the subset of sites with the greatest expected carbon stock increase from protection. Ranking error captures overall disagreement in the ordering of ITEs across all test plots, while estimation error captures numerical accuracy of plot-level ITE estimates.

To summarise overall meta-learner performance, we computed mean values of each metric at each study-condition level, along with standard deviations across virtual studies. To examine whether study conditions interacted in their effects on performance, we also computed mean performance within cells defined by combinations of conditions, such as treatment assignment, covariate omission, treatment imbalance and meta-learner. To help interpret differences among metrics and meta-learners, we visualised the correspondence between true and predicted ITE values across covariate values in the test datasets.

Results

Across the full set of virtual studies, meta-learner performance differed among metrics and algorithms (figure 1). The X-learner performed best on average across three of the four metrics, with the lowest targeting imprecision, ranking error and estimation error, while targeting error was almost identical for the X- and DR-learners. Mean targeting error was

lowest for the DR-learner, 0.54 ± 0.26 SD, closely followed by the X-learner, 0.54 ± 0.24, T-learner, 0.60 ± 0.24, and S-learner, 0.66 ± 0.18. The X-learner had the lowest targeting imprecision, 0.45 ± 0.09, followed by the DR-learner, 0.48 ± 0.11, T-learner, 0.51 ± 0.11 and S-learner, 0.55 ± 0.10.

Ranking error was also lowest for the X-learner, 0.28 ± 0.09, followed by the DR-learner, 0.33 ± 0.15, T-learner, 0.39 ± 0.14, and S-learner, 0.40 ± 0.16. Estimation error showed a similar pattern, with lowest RMSE for the X-learner, 4.12 ± 0.78, followed by the DR-learner, 4.35 ± 0.95, T-learner, 4.66 ± 0.93, and S-learner, 5.59 ± 1.34. These average differences indicate that the X-learner was the strongest overall meta-learner in this simulation study, although the DR-learner was highly competitive, particularly for targeting error. The large and overlapping standard deviations also show substantial variation among virtual studies, suggesting that the best-performing meta-learner depended on study conditions.

The four performance metrics captured related but non-identical aspects of ITE prediction performance (figure S3). Targeting error, targeting imprecision and ranking error were positively associated, but the strength of these relationships varied among metric pairs and meta-learners. The strongest association was between targeting imprecision and ranking error, both overall, $\rho=0.75$, and within individual meta-learners. Targeting error was only moderately associated with targeting imprecision, $\rho=0.37$, and ranking error, $\rho=0.37$. Associations involving estimation error were generally weaker for the targeting metrics, $\rho=0.21$ with targeting error and $\rho=0.27$ with targeting imprecision, but stronger with ranking error, $\rho=0.44$. These patterns show a distinction between accurate plot-level effect-size estimation, accurate ordering of ITEs across all sites, and accurate prioritisation of high-benefit sites. We therefore retained all four metrics in subsequent analyses and assessed whether study-condition interactions affected each metric differently.

Training sample size showed the clearest marginal association with performance. Across meta-learners, targeting error, targeting imprecision, ranking error and estimation error generally decreased with sample size. We therefore restricted subsequent interaction analyses to the largest training sample size, n = 1000, allowing us to examine how assignment mechanism, treatment imbalance, spatial test-location and covariate omission modified meta-learner performance under the least data-limited setting. These interaction analyses were used to determine whether the average superiority of the X-learner persisted across study conditions, or whether particular meta-learners were favoured under specific combinations of bias, imbalance, spatial extrapolation and covariate omission.

Interactions among study conditions

In the n=1000 subset, interactions among treatment assignment, covariate omission and treatment imbalance were clearest for ranking error and estimation error (figure 2; full cell means across all study-condition combinations are reported in table S2). The relative performance of the X- and DR-learners depended strongly on the direction of treatment imbalance. When 30% of plots were controls, meaning the business-as-usual/control arm was under-represented, the DR-learner was often competitive with, or better than, the X-learner, particularly for ranking error and targeting imprecision. In contrast, when 70% of plots were controls, and the set-aside/treatment arm was under-represented, the X-learner showed the clearest advantage, with the lowest ranking error under both random and non-random assignment and under both covariate-omission scenarios. Thus, the effect of imbalance depended not only on the degree of imbalance, but also on which treatment arm was under-represented.

To interpret this asymmetry, we compared the complexity of the two potential-outcome surfaces by modelling soil carbon under business-as-usual control and set-aside treatment separately as functions of covariates (figure S4). The set-aside/treatment surface was less consistently predictable than the business-as-usual/control surface, indicating that it was the harder response surface to learn. Imbalance was therefore more consequential when the smaller treatment group corresponded to the more complex outcome surface, that is, when 70% of plots were controls and the less-learnable set-aside/treatment group was under-represented.

Estimation error showed a different pattern. The S-learner had consistently higher RMSE than the other meta-learners across nearly all combinations of assignment mechanism, covariate omission and treatment imbalance. The ITE surface plots (figure 3) suggest that this occurred because the S-learner produced a compressed treatment-effect surface, underestimating heterogeneity across covariate space. This was most evident for covariates where the true ITE surface increased steeply, such as stem density and spruce volume: the S-learner remained comparatively flat, whereas the T-, X- and DR-learners better tracked the shape of the true surface. However, all learners tended to underpredict the highest true ITEs, indicating that prediction error was largest in the upper tail of high-benefit sites.

The unmeasured-confounding scenario (with non-random assignment plus soil moisture omitted) did not produce a uniform increase in error across learners or metrics. Instead, omission effects were mixed and depended on meta-learner, imbalance level and metric. For example, under non-random assignment, omitting soil moisture often reduced the S-

learner's ranking error, targeting imprecision and estimation error, while targeting error showed a less consistent response. For the X-learner, omission generally increased ranking error and estimation error slightly under high imbalance, but this effect was small relative to differences among learners. These patterns indicate that covariate omission consequences depended on how each learner represented the treatment-effect surface and which part of the error surface was emphasised by the metric.

The ITE surface plots help explain why the error metrics diverged. RMSE was sensitive to broad shrinkage in predicted ITE magnitudes, especially for the S-learner, whereas targeting metrics were most sensitive to whether learners recovered the upper tail of high-benefit sites. Ranking error captured broader ordering across the full test set. Thus, the different metrics emphasised different parts of the prediction error surface: magnitude error, overall rank-order error and errors in identifying the highest-priority sites for protection.

Discussion

This study evaluated the performance of four meta-learners for predicting ITEs in an ecological management setting, using decision-relevant criteria for both regional prioritisation of sites and local, site-specific effect-size prediction. Across the simulated conditions considered here, meta-learner performance depended not only on the algorithm, but also on the decision objective, sample size, treatment imbalance, treatment-assignment mechanism, covariate omission and spatial structure of the test data. The X-learner was the most consistently strong meta-learner in our case study, performing well for both prioritisation and estimation, while the S-learner generally performed worst, particularly for estimating treatment-effect magnitudes. However, these patterns should be interpreted as simulation-specific rather than universal recommendations, because our conclusions are based on one ecological data-generating process, one treatment-effect surface, four meta-learners and random forests as base learners. Below, we make suggestions for applied ecologists considering using meta-learners, and for quantitative ecologists seeking to test their performance under alternative ecological data-generating processes.

A key result is that the apparent "best" meta-learner depended on the performance metric (figure 1) and its interaction with study conditions (figure 2). Metrics designed for regional targeting, such as targeting error and targeting imprecision, emphasised whether meta-learners correctly identified the sites where protection would retain the most soil carbon. In contrast, estimation error measured whether predicted ITE magnitudes were numerically close to the truth across all test sites. These criteria were related but not interchangeable (as

shown by their moderate correlations, figure S3). Evaluating meta-learners using only one metric therefore risks recommending a method that is well suited to one decision but poorly suited to another.

For ecologists considering using meta-learners, our results suggest several practical cautions. Sites nominated as high priority by a learner should be checked for support in the training data, because targeting decisions focus on the upper tail of the predicted effect distribution and these sites may occupy sparsely sampled regions of covariate space where predictions are less reliable (figure 3). Analysts should therefore inspect whether top-ranked sites fall within the covariate range and treatment-overlap region represented in the training data. Moreover, algorithmic adjustment cannot solve unmeasured confounding. When important drivers of both treatment assignment and outcomes are omitted, ITE predictions may be biased.

Although the X-learner is designed to perform well under treatment imbalance, our results show that imbalance should not be interpreted only as a treated:control ratio. The degree of its superior performance over other meta-learners depended on which treatment arm was under-represented. In our case study, the set-aside/treatment outcome surface was harder to learn than the business-as-usual/control surface, with lower predictability and greater residual variation (figure S4). Imbalance was therefore more damaging when the smaller group corresponded to this more complex response surface, that is, when 70% of plots were controls and the set-aside/treatment group was under-represented. For applied ecologists, this suggests that treatment allocation should be interpreted alongside treatment-specific outcome complexity: before choosing a meta-learner, analysts should inspect both treatment imbalance and whether outcome–covariate relationships appear more variable or nonlinear in one treatment arm than the other.

We summarise these practical implications in the simulation-informed decision guide in figure 4, which is intended as a heuristic for the conditions evaluated here rather than a universal rule for method selection.

For researchers seeking to test the performance of causal ML in the future, our study highlights the value of virtual ecology approaches in which true treatment effects are known. Simulated datasets with known treatment effects allow researchers to separate different sources of error, including errors in effect-size estimation, ranking, top-site targeting and extrapolation across space. A limitation of this approach is that the meta-learners are modelling outputs from another model. Heureka provides the counterfactual outcomes needed to evaluate true ITEs, but its simulated outcomes are possibly smoother and less

noisy than empirical ecological data, measured with error. Meta-learner performance may therefore be optimistic relative to real applications involving measurement error, unmodelled environmental variation, stochastic disturbance, changing climate or species interactions.

For any benchmark dataset, an algorithm might overperform on simulated data due to specificities of the data-generating process, requiring comparison of ITE accuracy across alternative processes (Curth *et al.* 2021). Future benchmarks should therefore expand beyond the conditions evaluated here by varying the complexity of treatment-effect surfaces, the strength and number of confounders, collider bias, the degree of covariate overlap, and the structure of spatial dependence. Where prediction errors are spatially clustered, future applications should assess spatial autocorrelation in ITE errors and use spatially explicit validation where possible. Future simulation studies should test meta-learners under data-generating processes with greater environmental and temporal variation, including changing climate. Other ecological systems may pose challenges that are less prominent in many human-centred benchmarks, such as dispersal, spillovers, species movement, disturbance spread or neighbourhood effects that could violate assumptions such as ‘stable unit treatment value assumption’ (SUTVA), that one unit’s response to treatment is unaffected by other units’ assignments. Testing meta-learners under such ecological data-generating processes will be important before broader guidance can be developed.

We have compared the performance of four popular meta-learners and employed random forests as their base learners with fixed hyperparameterisation. Future research could use virtual ecology approaches to test the accuracy of alternative causal ML approaches that have been proposed (Arif 2025) and hyperparametrisations that may rank meta-learner performance differently. 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). Moreover, we note that methods for continuous treatment variables are also under development, including disentangled representation networks (Hu *et al.* 2024).

Conclusions

Overall, our results show that meta-learners can be useful tools for ecological ITE prediction, but that their evaluation should be tied explicitly to a management decision. In this case study, the X-learner was the most reliable across the metrics and conditions considered, yet one meta-learner was not universally best. Ecologists should select and evaluate meta-

learners according to the decision they aim to support, the data properties, and the causal assumptions. Our findings are based on a single data-generating process, treatment type, range of treatment effect sizes relative to environmental drivers, and we employ only one type of base meta-learner. Further simulation studies are needed to test causal ML performance under ecological data-generating processes that pose different challenges to causal prediction.

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] and project SustMultBiomass. Project SustMultBiomass is supported under the umbrella of CSA project ForestValue2 by Vinnova (2021-05011). ForestValue2 is funded by the European Union under Grant Agreement No. 101094340. The authors would like to acknowledge the use of the Reading Academic Computing Cluster (RACC) facility at the University of Reading. The authors would like to thank four anonymous reviewers for comments which greatly improved the manuscript. For the purpose of Open Access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript (AAM) version arising from this submission.

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.

Arif, S. (2025). Estimating causal effects with machine learning: A guide for ecologists. *Methods Ecol. Evol.*, 2041–210X.70191.

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.

Correia, H.E., Dee, L.E., Byrnes, J.E.K., Fieberg, J.R., Fortin, M.-J., Glymour, C., *et al.* (2026). Best practices for moving from correlation to causation in ecological research. *Nat. Commun.*, 17, 1981.

Curth, A. & Schaar, M. van der. (2021). Nonparametric Estimation of Heterogeneous Treatment Effects: From Theory to Learning Algorithms. In: *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*. Presented at the International Conference on Artificial Intelligence and Statistics, PMLR, pp. 1810–1818.

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).

Feuerriegel, S., Frauen, D., Melnychuk, V., Schweisthal, J., Hess, K., Curth, A., *et al.* (2024). Causal machine learning for predicting treatment outcomes. *Nat. Med.*, 30, 958–968.

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.

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. (2025). 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.

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, 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.

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.

Montgomery, J.M., Nyhan, B. & Torres, M. (2018). How conditioning on posttreatment variables can ruin your experiment and what to do about it. *Am. J. Polit. Sci.*, 62, 760–775.

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.

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

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.

Probst, P., Wright, M.N. & Boulesteix, A. (2019). Hyperparameters and tuning strategies for random forest. *WIREs Data Min. Knowl. Discov.*, 9, e1301.

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.

Singh, A. (2026). A large-scale empirical comparison of meta-learners and causal forests for heterogeneous treatment effect estimation in marketing uplift modeling.

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.

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.

Yadlowsky, S., Fleming, S., Shah, N., Brunskill, E. & Wager, S. (2025). Evaluating treatment prioritization rules via rank-weighted average treatment effects. *J. Am. Stat. Assoc.*, 120, 38–51.

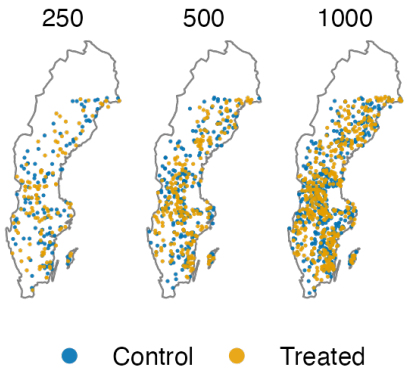
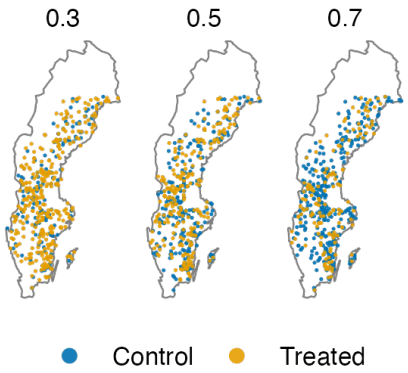
Yates, K.L., Bouchet, P.J., Caley, M.J., Mengersen, K., Randin, C.F., Parnell, S., *et al.* (2018). Outstanding challenges in the transferability of ecological models. *Trends Ecol. Evol.*, 33, 790–802.

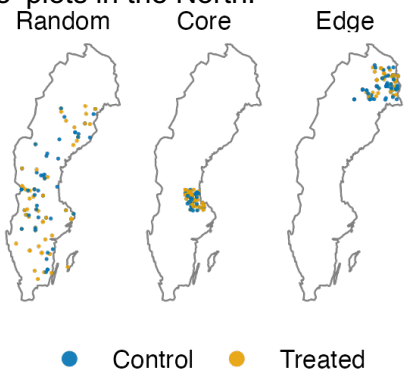
Zhang, J., Jin, Y. & Wang, X. (2025). Comparing meta-learners for estimating heterogeneous treatment effects and conducting sensitivity analyses. *Math. Comput. Appl.*, 30, 139.

Tables

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

Study condition	Rationale	In this study
Treatment assignment mechanism		
Treatment assignment	Observational studies face the challenge of non-random treatment assignment with respect to one or more covariates, which can reduce covariate overlap between treatment groups and require extrapolation into poorly supported regions of covariate space. Limited overlap can increase model dependence and misspecification risk but does not by itself demonstrate a formal violation of the positivity assumption. When the covariate driving treatment assignment also influences the outcome or treatment effect, it acts as a confounder; if it is not adjusted for, ITE estimates can be biased. This distinguishes design-based uncertainty, arising from confounding, from model-based uncertainty, arising from omitted predictors or mis specified response surfaces.	Two treatment-assignment mechanisms: 1) random, where treatment was assigned independently of covariates; 2) non-random, where assignment was correlated with soil moisture. In the non-random scenario, plots on mesic-moist soils (soil-moisture class three) were assigned twice the probability of being observed under set-aside/protection as plots in drier soil-moisture classes. 
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.	Three sample sizes: 250, 500 or 1000 NFI plots sampled from across Sweden. 
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. 
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 likely influences the location of test data in covariate space, with implications for predictive performance (Yates *et al.* 2018) (figure S2).	We simulated sampling our test data from three different spatial locations: 1) 'random' plots, sampled via stratified random sampling, where strata were latitudinal deciles; 2) 'core' plots, selected from a distinct area in the centre of Sweden, and 3) 'edge' plots in the North. 
Meta-learner algorithm	Meta-learners differ in how they use treatment-group information, outcome models and propensity scores to estimate ITEs. These differences will affect performance variably under treatment imbalance, non-random assignment and covariate omission.	Four meta-learners: S-learner and T-learner as plug-in meta-learners; X-learner and DR-learner as two-step meta-learners. Random forests were used as base learners for all methods. Implementation was adapted from Salditt et al. (2024); full algorithmic details are given in (Curth & Schaar 2021; Kennedy 2023; Künzel *et al.* 2019).
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 a covariate: 'soil moisture' (table S1). Soil moisture is a confounder when treatment assignment is non-random. If omitted from model training, it becomes an unmeasured confounder.

1

Table 2. Decision-relevant metrics used to evaluate ITE prediction errors. All four metrics were oriented as errors, so lower values indicate better performance. Because RMSE is measured in soil-carbon units and the test-location scenarios differ in the distribution of true ITEs, RMSE values should not be interpreted as directly comparable measures of prediction difficulty across test locations. Instead, RMSE should be interpreted within each test-location scenario, or alongside the distribution of true ITEs in that test set.

Performance metric	**How it works**	**Implementation**	**Interpretation**
Targeting error Informs regional prioritisation Asks: how much targeting value is lost when sites are prioritised using predicted rather than true ITEs?	A Target Operating Characteristic (TOC) curve is constructed by ranking test plots in decreasing order of predicted ITE. At each fraction q of plots selected, the mean ITE among the top-q plots minus the overall mean ITE is calculated. AUTOC summarises the area under this curve, giving greatest weight to the top of the ranking. Note: because AUTOC is asymmetric, the direction in which sites are ranked must correspond to the management decision.	We ranked sites in decreasing order of predicted ITE, so that the highest-priority sites were those predicted to retain the most additional soil carbon under set-aside management relative to business-as-usual management (clearcutting). We computed log-weighted AUTOC following Yadlowsky et al. (2025), using log(1/q) weighting, recommended when a small fraction of sites have large ITEs. AUTOC values for each meta-learner were normalised by the 'oracle' AUTOC obtained from ranking sites by true ITEs. Targeting error was calculated as 1−normalised AUTOC.	Targeting error measures the loss in targeting value relative to an oracle ranking, where sites are ordered by their true ITEs. Lower values indicate better targeting. A value of 0 indicates oracle-level prioritisation; values close to 1 indicate performance similar to random ranking; values >1 indicate worse-than-random targeting.

Targeting imprecision Informs regional prioritisation Asks: how well does the meta-learner identify the true top-priority sites?	Compares the set of top sites selected by predicted ITEs with the top set selected according to the true ITEs.	We identified the top 20% of test plots according to predicted ITEs and calculated the fraction that were also in the true top 20%. For 162 test plots, this corresponded to the 33 highest-ranked plots. Targeting imprecision was calculated as 1− this percentage.	Lower values indicate better recovery of the true priority set. A value of 0 means all predicted top-priority sites were also truly top-priority sites; a value of 1 means none were.
Ranking error Informs regional prioritisation. Asks: how well does the meta-learner recover the true ordering of treatment effects across all sites?	Measures rank agreement between predicted and true ITEs across the full test set, weighting all sites equally rather than emphasising only the top of the ranking.	We calculated Spearman's rank correlation, ρ, between predicted and true ITEs, then converted it to as 1−ρ	Lower values indicate better overall rank agreement. A value of 0 indicates perfect rank agreement; values close to 1 indicate little rank agreement; values >1 indicate negative rank agreement.
Estimation error Informs local management. Asks: How accurately does the meta-learner estimate the treatment effect at an individual site?	Measures the numerical distance between predicted and true ITEs in the units of the response variable.	We calculated the root mean square error, RMSE, between predicted and true ITEs.	Lower values indicate more accurate plot-level ITE estimates in the test dataset, on average. RMSE is measured in soil-carbon units.

1

Figures

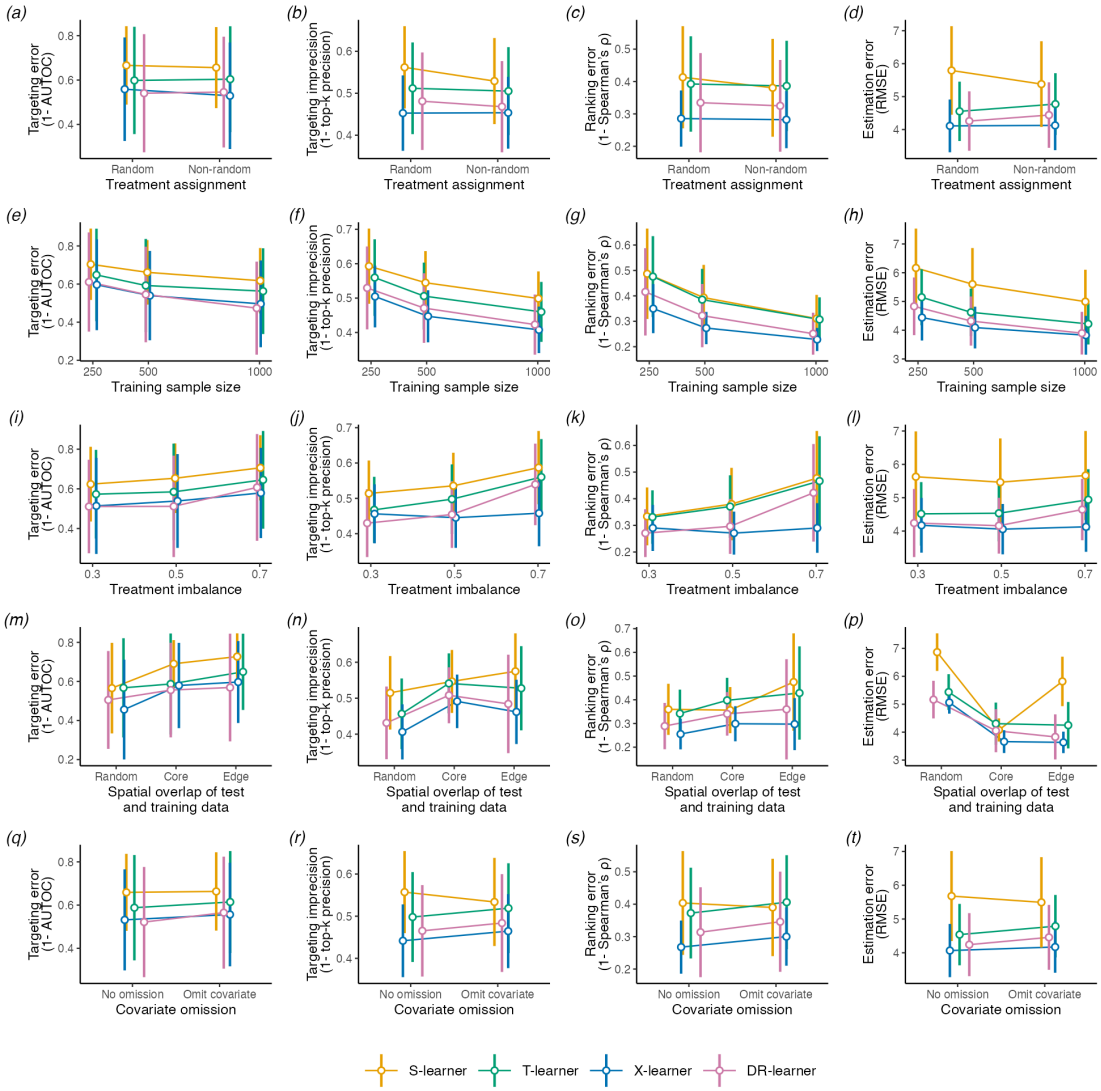


Figure 1. Overall error of four meta-learners for predicting individual treatment effects of forest management on soil carbon. Error is summarised across the full complement of virtual studies under different study conditions: treatment-assignment mechanism, training sample size, treatment imbalance, spatial relationship between training and test data, and covariate omission. Colours denote meta-learners: S-learner, T-learner, X-learner and DR-learner. Points show mean error and vertical bars show ±1 SD across virtual studies. Four complementary error metrics are shown (described in table 1). Lower values indicate better performance for all four metrics, although RMSE is measured in soil-carbon units and is not directly comparable in scale to the other three unitless metrics.

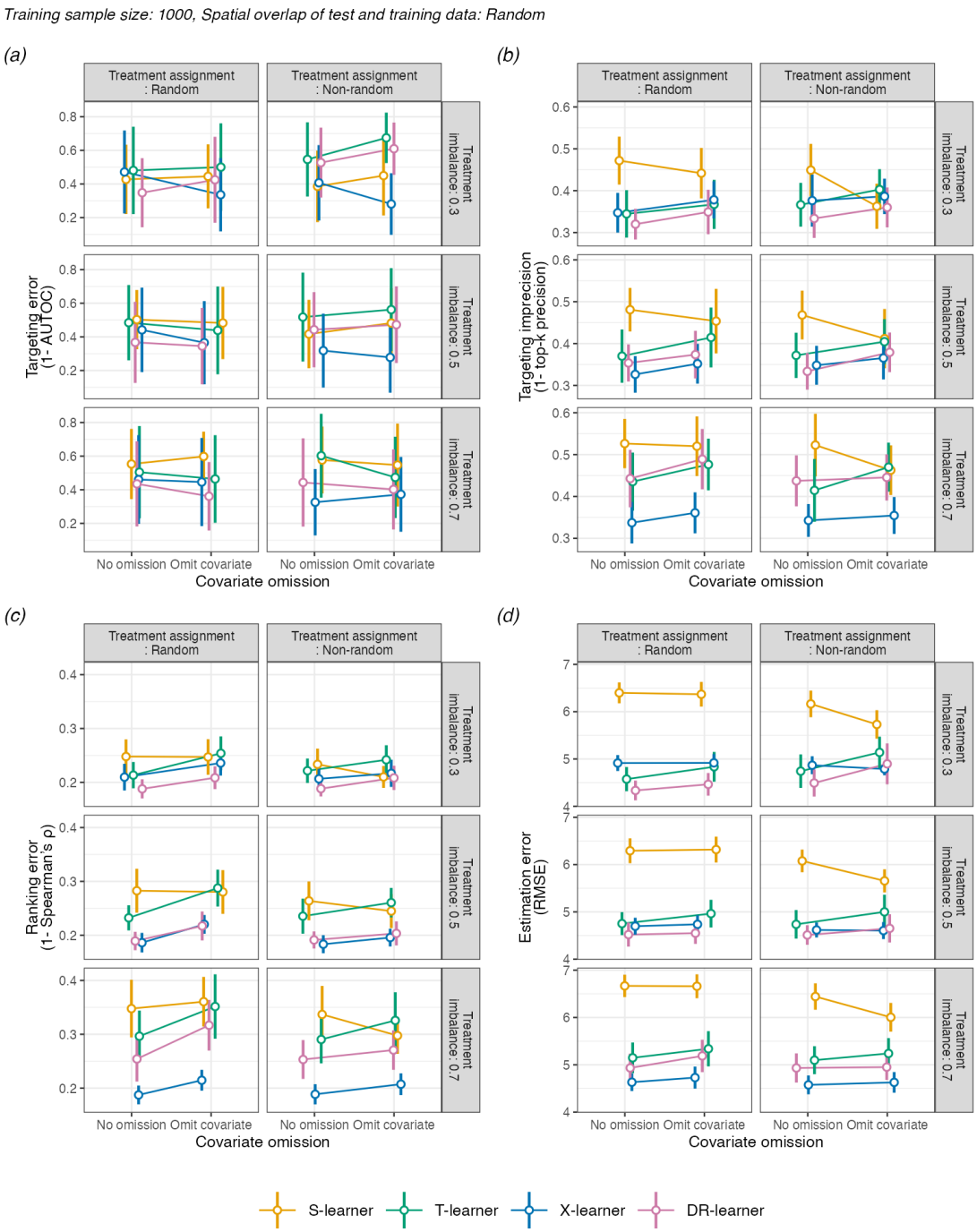


Figure 2. Performance of four meta-learners (different colours) at predicting ITEs representing the effect of forest management on soil carbon under different combinations of study conditions (see table 1). Shown are the mean and standard deviations across virtual studies that vary in study conditions (table 1) for four types of error: (a) Targeting error, defined as AUTOC, (b) Targeting precision, defined as top-k precision for the highest-priority sites (c) Ranking error, defined as Spearman's rho, and (d) Estimation error defined as root mean squared error (RMSE). Each panel contains data a subset of the virtual studies with training sample size of 1000 and test plot location of stratified: from left to

right, treatment imbalance is 0.3, 0.5, 0.7. Error metrics are calculated using the true (simulated) and predicted ITEs from the test datasets in each study.

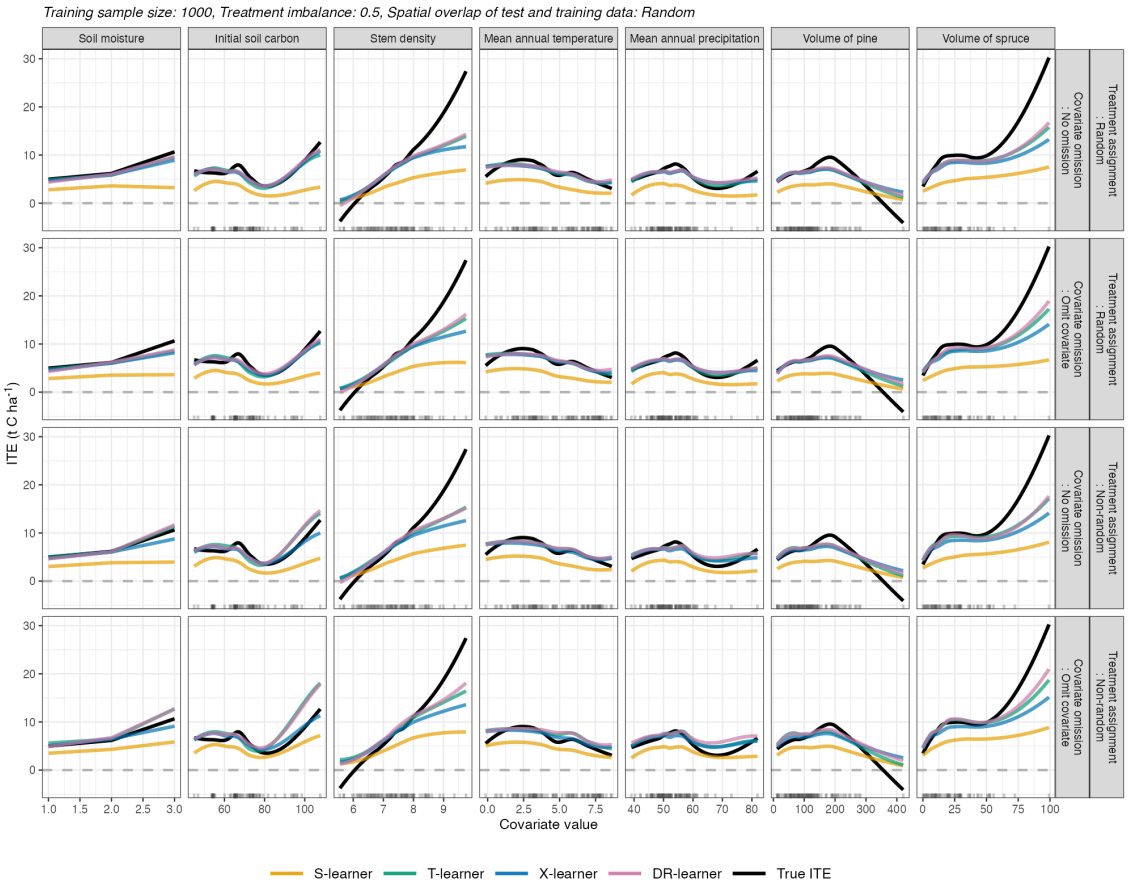


Figure 3. True and predicted individual treatment effect, ITE, surfaces across covariate space for each meta-learner under contrasting treatment-assignment and covariate-omission scenarios. Each panel shows how the ITE, defined as soil carbon under set-aside/protection minus soil carbon under business-as-usual management, varies with a covariate in the test data. Positive ITEs therefore indicate sites where protection retained more soil carbon than business-as-usual management. The black line shows the true ITE surface from the simulation, and coloured lines show predicted ITE surfaces from the S-, T-, X- and DR-learners. Lines are LOESS smooths of predicted and true ITEs against each covariate across virtual-study test datasets; soil moisture, a three-level ordinal variable, is shown as means at each level. Columns correspond to covariates, and rows correspond to combinations of treatment-assignment mechanism and covariate-omission scenario. Departure of a coloured line from the black line indicates systematic prediction error by that meta-learner in that region of covariate space. Rugs along the x-axis show the distribution of test-plot covariate values, indicating where the data are dense or sparse. Results are shown for training sample size n = 1000 and balanced treatment allocation.

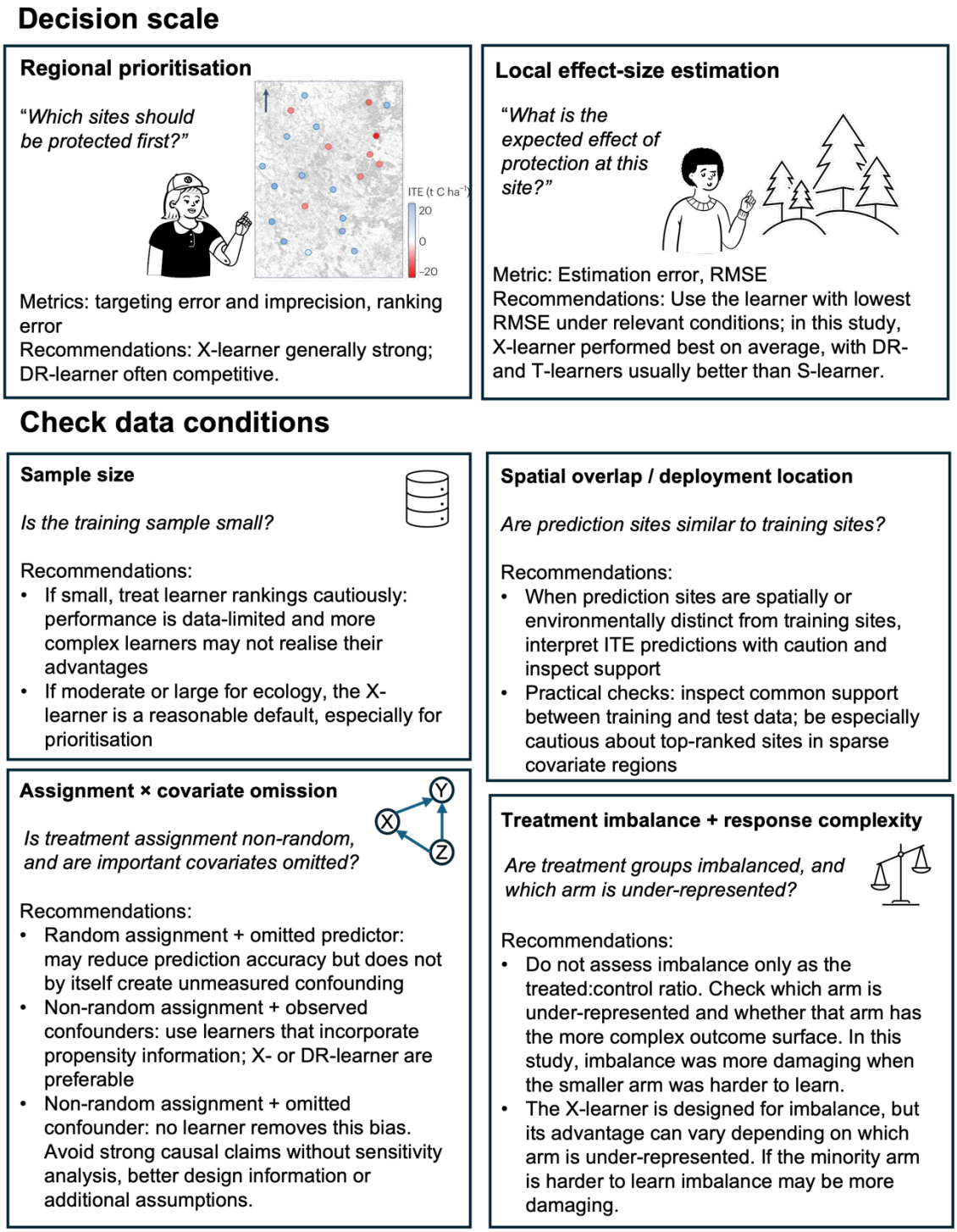


Figure 4. Simulation-informed guide for choosing meta-learners for ecological ITE prediction. The guide summarises recommendations from the simulated conditions evaluated in this study. The first choice is the management decision: regional prioritisation requires metrics that assess whether the highest-benefit sites are correctly identified, whereas local decision-making requires accurate site-level effect-size estimates (error descriptions in Table 2). Recommendations are based on the simulated conditions evaluated here and should be interpreted as study-specific heuristics, not universal rules for meta-learner selection.

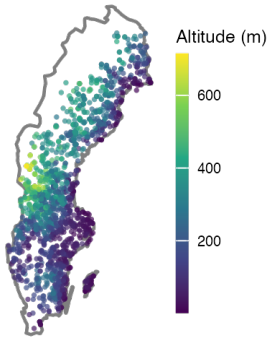
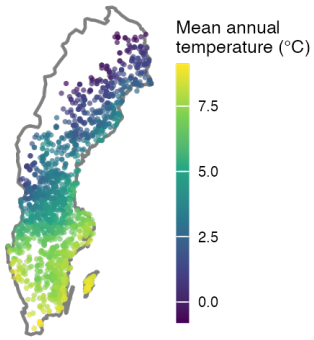
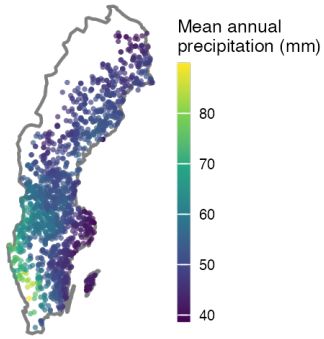
Supporting Information

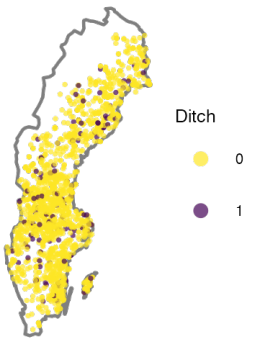
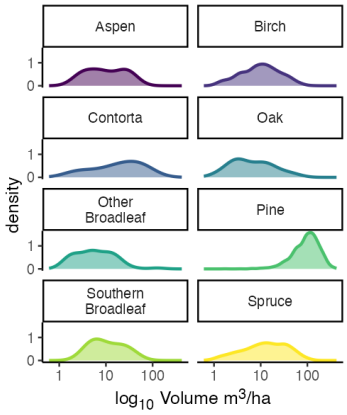
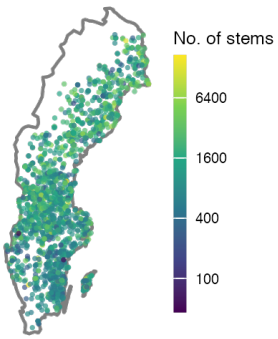
Towards causal predictions of site-level treatment effects in applied ecology

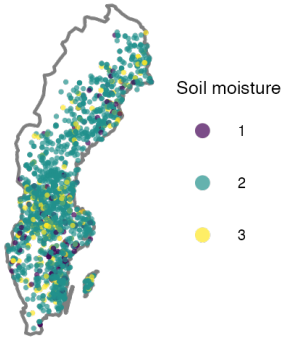
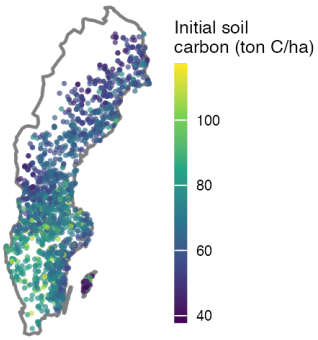
Eleanor E. Jackson[1,2], Tord Snäll[3,4], Emma Gardner[5], James M. Bullock[5], Rebecca Spake[1,6]

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

Table S1. Environmental covariates selected for use in statistical models predicting soil carbon. All covariate values were taken at time (t) = 0.

Covariate	**Details**
Altitude 	Height above sea level, sourced from the Swedish National Forest Inventory.
Temperature 	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 	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.

Ditching 	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 	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).
Stem density 	Total number of stems within the plot (DBH >= 4 cm) sourced from the Swedish National Forest Inventory.

Soil moisture 	An ordinal variable including 1 (dry), 2 (mesic) and 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 	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.

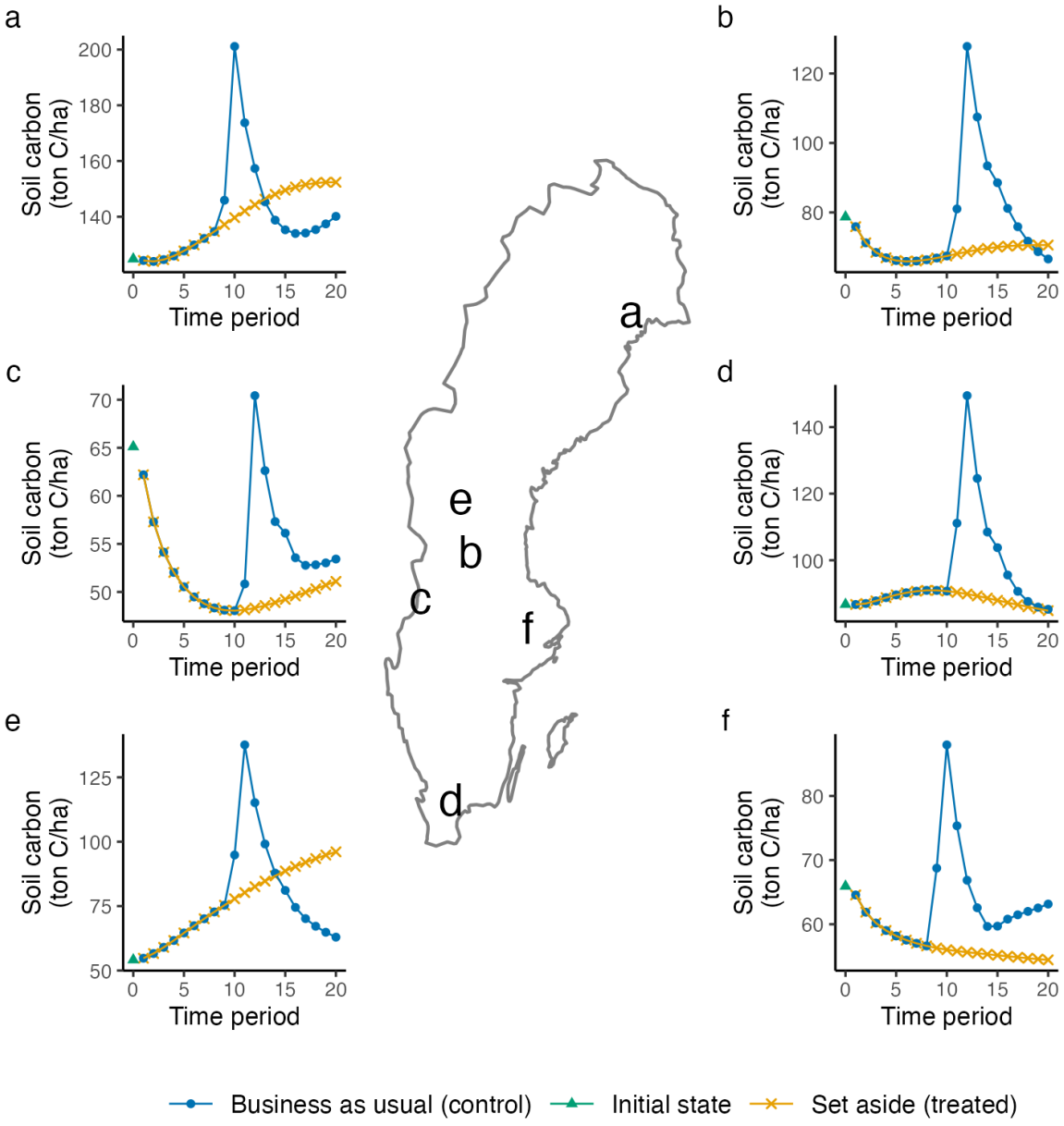


Figure S1. Simulated soil carbon over time for six National Forest Inventory Plots. Figure labels (a - f) are plotted on a map of Sweden to indicate their location. For the first year soil carbon is plotted at its "initial state" (green triangle), this is the value measured during National Forest Inventory surveys in the given year. Subsequent values were simulated by Heureka under either "business as usual" management where trees are clear cut and replanted (blue circle, "control" in this study), or "set aside" (orange cross, "treated"). 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.

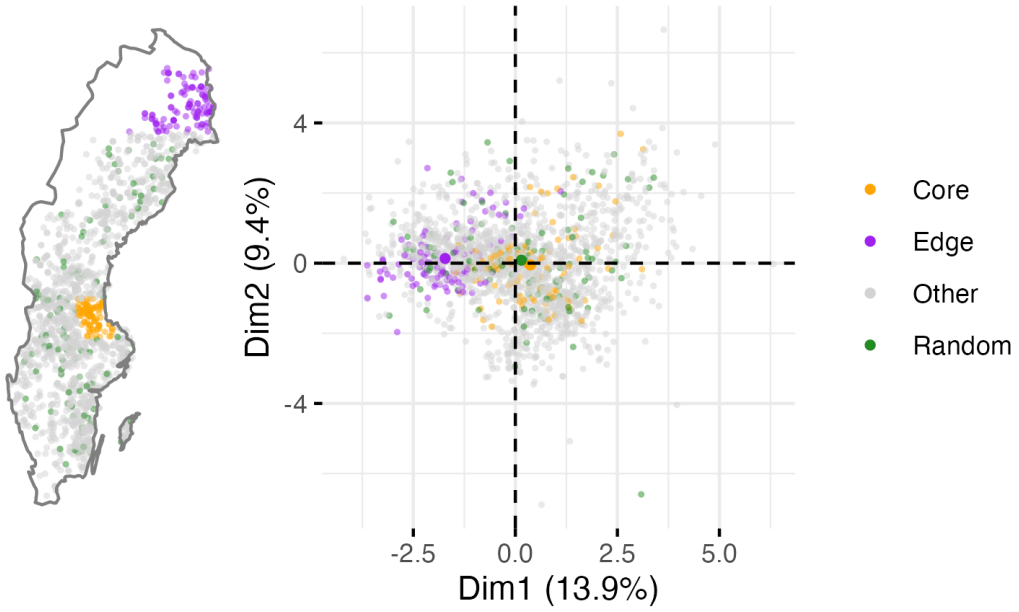


Figure S2. Location of test data. The left panel depicts a map of Sweden with the National Forest Inventory plots indicated by points. The right panel is a principal component analysis visualisation where points which are closer in space are plots with similar environmental covariates (listed in main 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.

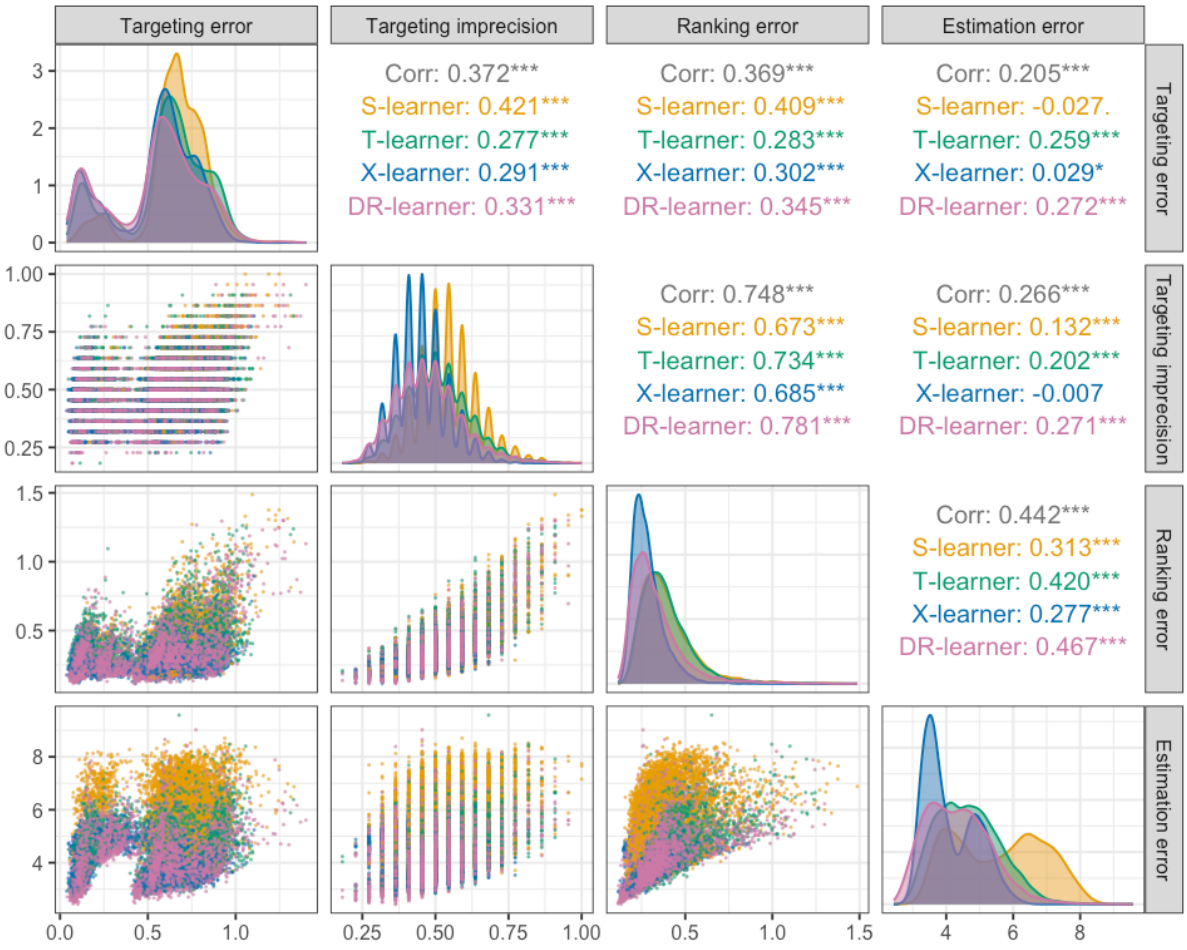


Figure S3. Pairwise relationships among ITE prediction performance metrics across virtual studies. Lower panels show individual virtual studies coloured by meta-learner; diagonal panels show learner-specific metric distributions; and upper panels show Spearman rank correlations between metric pairs. Metrics are targeting value, targeting imprecision, ranking performance and estimation error (described in table 2). Data are pooled across all combinations of treatment assignment, treatment imbalance, sample size, covariate omission, test-location scenario and replicate.

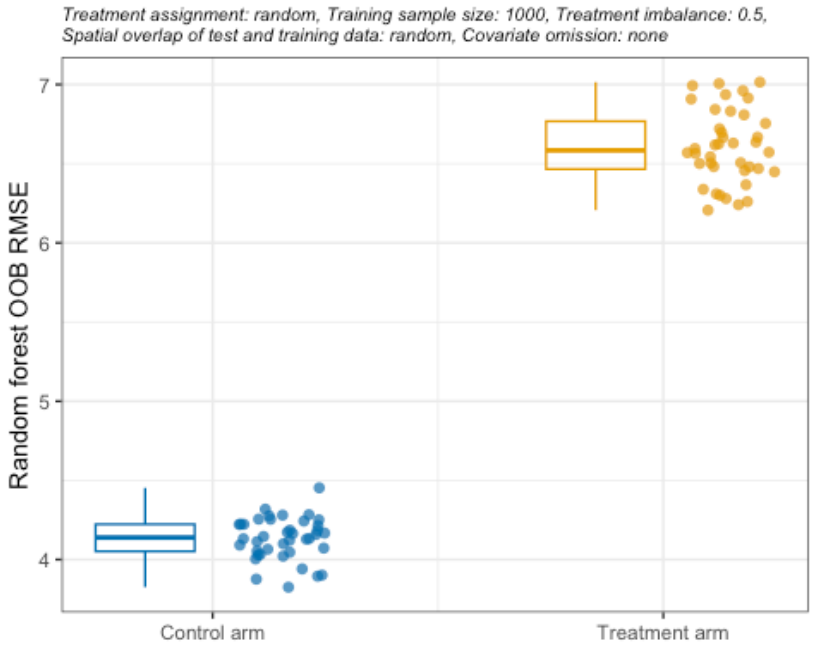


Figure S4. Treatment-specific outcome surfaces differed in learnability. Random forest out-of-bag, OOB, RMSE was estimated separately for the observed control and treatment arms across replicated virtual studies under one representative study condition: random treatment assignment, training sample size n=1000n = 1000n=1000, balanced treatment allocation, random spatial overlap of test and training data, and no covariate omission. Each point represents one replicated virtual study; boxplots summarise variation across replicates. Higher OOB RMSE indicates poorer predictability from the covariates. The treatment arm, set-aside/protection, had consistently higher OOB RMSE than the control arm, business-as-usual management, indicating that the treatment-specific outcome surface was harder to learn. This helps explain why treatment imbalance was more consequential when the treatment arm was under-represented.

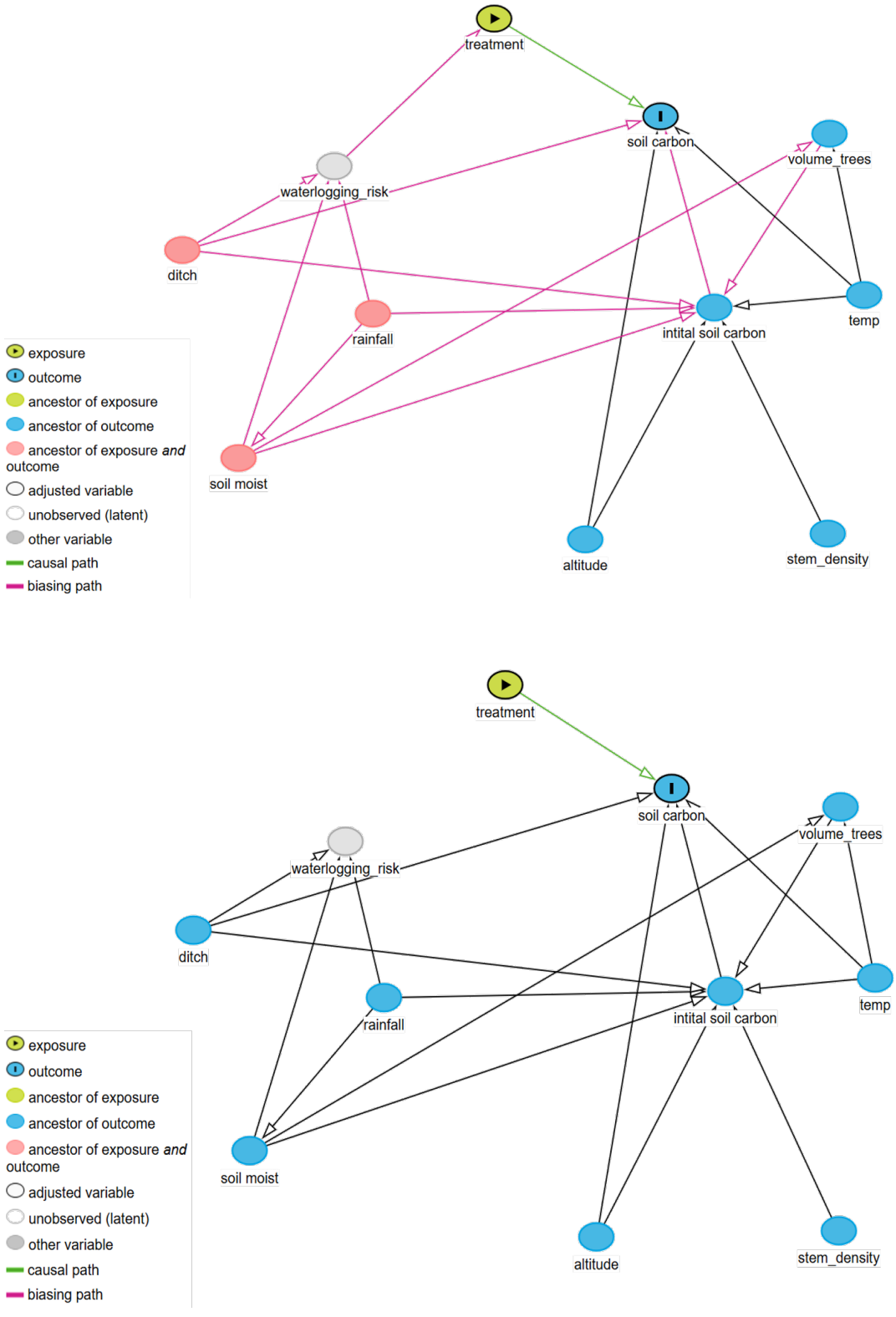


Figure S5. Directed acyclic graphs (DAGs) depicting causal relationships between the variables when treatment assignment is correlated with waterlogging risk (top) and when treatment assignment was random (bottom). DAGs were produced using dagitty.net.

Table S2. Mean, SD and 95% confidence interval of each performance metric for every study condition and meta-learner. Available as a separate supplementary file: table-S2.csv

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