Received: 26 April 2021 | Accepted: 1 July 2021

DOI: 10.1111/gcb.15817

Global Change Biology WILEY

TECHNICAL ADVANCE

Predicting ecosystem responses by data-driven reciprocal modelling

Florian Schneider | Christopher Poeplau | Axel Don

Thünen Institute of Climate-Smart Agriculture, Braunschweig, Germany

Correspondence
Florian Schneider, Thünen Institute of Climate-Smart Agriculture, Bundesallee 65, 38116 Braunschweig, Germany.
Email: florian.schneider@thuenen.de

Funding information
European Union's Horizon 2020, Grant/Award Number: 862695; German Federal Ministry of Education and Research, Grant/Award Number: 031B0515E

Abstract

Treatment effects are traditionally quantified in controlled experiments. However, experimental control is often achieved at the expense of representativeness. Here, we present a data-driven reciprocal modelling framework to quantify the individual effects of environmental treatments under field conditions. The framework requires a representative survey data set describing the treatment (A or B), its responding target variable and other environmental properties that cause variability of the target within the region or population studied. A machine learning model is trained to predict the target only based on observations in group A. This model is then applied to group B, with predictions restricted to the model's space of applicability. The resulting residuals represent case-specific effect size estimates and thus provide a quantification of treatment effects. This paper illustrates the new concept of such data-driven reciprocal modelling to estimate spatially explicit effects of land-use change on organic carbon stocks in European agricultural soils. For many environmental treatments, the proposed concept can provide accurate effect size estimates that are more representative than could feasibly ever be achieved with controlled experiments.

KEYWORDS
association, causal inference, causation, correlation, land-use change, machine learning, soil organic carbon, statistical modelling

1 | INTRODUCTION

The quantification of cause-and-effect relationships is key to understanding and managing the environment in which we live. Traditionally, such relationships are studied in very small subsets of our environment, at mesocosm and microcosm scale, or under simplified model conditions. In controlled experiments, the effect size of a given treatment is evaluated against a corresponding untreated reference. The setting can be an ecological field trial in which the treatment might be plant diversity and the studied response is ecosystem functioning (paired samples). It could also be a pharmaceutical trial, where a test population is randomly split and one group receives a drug (treatment) while the other receives a placebo (reference), and these unpaired groups are compared. Such controlled experimental designs offer great potential for the accurate quantification of treatment effects. In fact, the more controlled (simplified) the experiment is, the more accurately cause-and-effect relationships can be determined. However, as experimental control is best achieved by simplifying environmental conditions, many cause-and-effect descriptions lack representativeness. Carefully and very accurately quantified effects observed in a single site or population might appear very different at other sites or in populations with altered environmental conditions.

Survey programs offer an alternative, data-driven approach for estimating treatment effects. While controlled experiments focus on accurately quantifying treatment effects at the expense of representativeness, the opposite is the case for surveys. Surveys aim to document target variables under the complete range of field conditions, but at the expense of easily being able to unravel causality. With increasing complexity of the studied environment, it is harder to distinguish causality from correlation (Huston, 1997). In the past two decades, machine learning methods have progressively been used to examine associations between target variables and environmental properties. Machine learning has allowed the identification and reproduction of complex associations, which has greatly advanced the ability to predict accurately using unseen data. However, even advanced machine learning methods are unable to distinguish between random associations and causal relationships (Wadoux et al., 2020). What machine learning can do, however, is link target values observed in a treatment group to environmental properties. This link could then be used to predict potential treatment effects for environmental observations that have not been treated (Bastin et al., 2019; Schneider & Don, 2019). The approach of training a model in one group and then applying this model to another group with a different treatment is referred to below as data-driven reciprocal modelling. In ecology and environmental science, only relatively few case studies to date have implemented data-driven reciprocal modelling methods, for example to examine the effect of human activity on the number of trees (Bastin et al., 2019) and soil compactness (Schneider & Don, 2019). However, these pioneering studies tend to insufficiently address two key issues: (i) explanatory variables used for model training should be unaffected by the examined treatment (otherwise treatment effects can be underestimated) and (ii) machine learning models should only be applied within their training space, not beyond it.

In the following, we propose a good practice approach for data-driven reciprocal modelling to estimate complex treatment effects under field conditions. We illustrate this powerful method using the LUCAS soil data set describing the state of European topsoil in the year 2015 (Orgiazzi et al., 2018; Panagos et al., 2020) and examine the effect of agricultural land-use change on soil organic carbon (SOC) stocks. Specifically, we address the question of how much atmospheric carbon (C) European agricultural soils could sequester if today's croplands were converted to grasslands. However, the method presented can also be applied to answer numerous other questions—all that is required is a representative data set that relates values of a chosen target variable (here, SOC stock) to values of the studied treatment (here, land use) along with values of other potential explanatory variables (here, soil, climate, geology and management variables).

2 | METHODS

2.1 | General concept

Data-driven reciprocal modelling can be used to quantify treatment effects in situ, directly in the environment in which the treatment naturally occurs. Instead of controlling the environment, the proposed method aims to describe it as a whole. This is achieved by means of a representative survey that documents values of the target variable along with the examined treatment, and those environmental features that potentially also cause variation in the target variable within the studied region or population. The target variable depends on its associated treatment and environment:

$$\textit{Target variable} = f\,(\textit{Treatment}, \textit{Environment}),$$

where *Treatment* is a dichotomous factor with two classes: each observation is either of group A or group B. To quantify the response to a change from group B to A, $\textit{Response}_{B\rightarrow A}$, data-driven reciprocal modelling requires (i) at least one observation in group B, (ii) sufficient observations in group A to statistically describe the target variable as a function of *Environment* and (iii) the *Environment* of group B to match or be a subset of the *Environment* of group A. If these criteria are met, data-driven reciprocal modelling can be implemented as follows (Figure 1):

Step 1. In group A, a data-driven statistical model is trained to predict the target variable as a function of environmental properties:

$$\textit{Target variable}_{\text{predicted}} = f\,(\textit{Environment}).$$

This model should reproduce associations between the *Target variable* and the *Environment* as accurately as possible (minimal underfitting) while still being generalizable to new observations within the boundaries of the training space (minimal overfitting). In the case of large and complex data sets, this task can often be achieved most efficiently with machine learning, choosing a model or model ensemble with an associated tuning approach that yields optimal performance. As a constraint, predictions should only include environmental variables that, based on expert knowledge, remain unaffected by the examined *Treatment*. This constraint is important in order to avoid underestimation of effect sizes (Step 3).

Step 2. Next, the method requires a comparison of the environmental properties of groups A and B. The observations of group B that are not comparable with the environmental properties of group A (training data) have to be deleted because they are located outside the model's space of applicability—they would be unknown to the model. The selection can be automated, for example as described by Meyer and Pebesma (2021), who judged the ability of any data-driven model to predict an unseen observation based on the Euclidean distance of this observation to the nearest training data point in the predictor space.

Step 3. Finally, the model that was trained on group A is applied to the remaining observations of group B. The resulting residuals represent case-specific effect size estimates for a change from group B to A:

$$\textit{Response}_{B\rightarrow A} = \textit{Target variable}_{\text{predicted}} - \textit{Target variable}_{\text{observed}}.$$

Step 4. In an optional post hoc analysis, the variation among individual effect sizes can be analysed further by relating the effect sizes from Step 3 to environmental properties. For example, this can

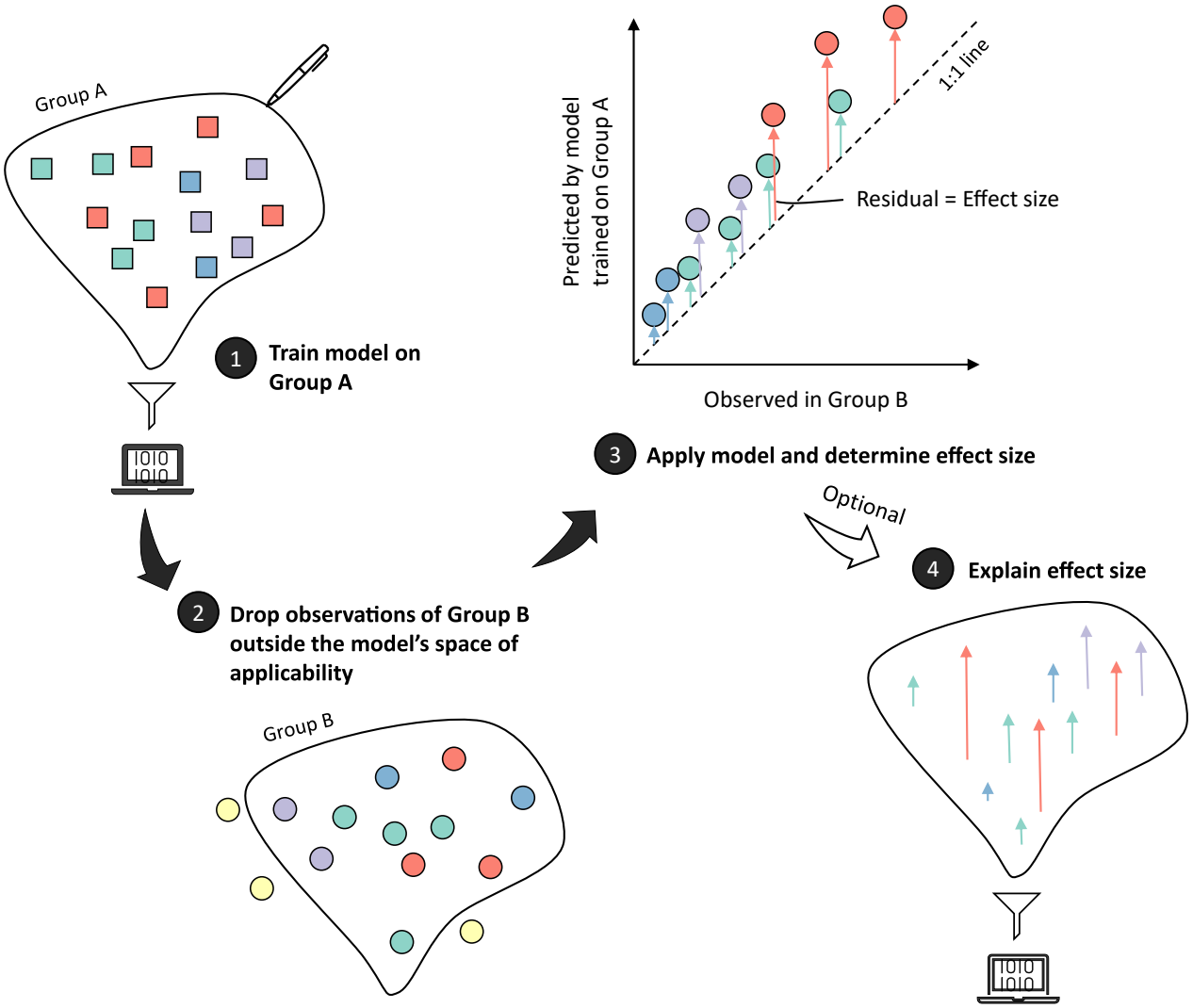

FIGURE 1 General concept for quantifying treatment effects by data-driven reciprocal modelling. Step 1: In group A (here, squares; in case study, grassland), train model to predict the target variable as a function of treatment-independent environmental properties (here, colour); the range of environmental properties used to train the model define its space of applicability. Step 2: In group B (here, circles; in case study, cropland), drop observations with properties outside the model's space of applicability. Step 3: Apply the model to the remaining observations of group B and determine residuals to obtain modelled responses to a change from B to A (here, arrow direction +length). Step 4: In an optional post hoc analysis, relate the different responses from Step 3 to environmental properties; in this final step, all potential predictors can be used, including those that were possibly set aside is Step 1 due to the treatment independency constraint

be done by training a second machine learning model, which predicts individual effect sizes. This model can then be interpreted, for example by means of partial dependence plots (Molnar, 2019). The constraint concerning environmental variable selection defined in Step 1 can be dropped in this final step, and for model training all potential predictors (based on expert knowledge) can be used, including those that might depend on the *Treatment*.

2.2 | Case study

Minor alterations in global SOC stocks can cause major changes in the concentration of atmospheric carbon dioxide (Minasny et al., 2017). In mineral soils, land-use change typically exhibits the strongest anthropogenic effect on SOC stocks (Paustian et al., 2016). However, site-specific effects of land-use change on SOC stocks may depend on environmental properties such as climate and soil texture (Don et al., 2011; Poeplau et al., 2011). In the present case study, we applied the new data-driven reciprocal modelling framework to provide spatially explicit, quantitative estimates for the effect of agricultural land-use change on SOC stocks in the European Union (EU-27) and the United Kingdom. The agricultural area of the EU-27 and the United Kingdom covers $1.73*10^6$ km^2, of which about two thirds are used as cropland and the remaining area is grassland (Eurostat, 2020). In the LUCAS 2015 Topsoil Survey, soils of these land-use types were sampled representatively to a depth of 20 cm

and analysed for SOC content, inorganic C content, nutrient contents (available phosphorus, available potassium, total nitrogen), texture and pH (Orgiazzi et al., 2018; Panagos et al., 2020). These data were provided by the European Soil Data Centre (https://esdac.jrc.ec.europa.eu/). Bulk density values of LUCAS sites were estimated with a pedotransfer function previously recommended for European agricultural soils by Hollis et al. (2012). The bulk density values obtained were used to convert SOC contents to stocks (Lugato et al., 2014):

$$SOC_{stock} = SOC_{content} * BD * Depth * \left(1 - \left(\frac{Coarse}{100}\right)\right),$$

where SOC_{stock} is given in Mg ha^{-1}, $SOC_{content}$ in per cent, BD is bulk density in g cm^{-3}, *Depth* equals 20 cm and *Coarse* represents the percentage of coarse fragments. Additionally, predictors were compiled representing climatic, geological and anthropogenic influences on SOC at the LUCAS sites. The bioclimatic variables of WorldClim (Fick & Hijmans, 2017), elevation, slope and aspect, as readily provided in the LUCAS 2015 Topsoil Database (Jones et al., 2020), were used to characterize climate and geology. The potential anthropogenic influence on SOC was characterized using variables from the agriculture (agr) database of the Statistical Office of the European Union (EUROSTAT). In total, 19 EUROSTAT variables were compiled, providing information about different farming systems, crop rotations and fertilization practices at the level of administrative regions (NUTS). Where EUROSTAT reported values for different years and/or different administrative levels, for a given variable and LUCAS point, further analyses were restricted to values of the smallest administrative level available and the mean value of time series. All EUROSTAT variables included were reported at NUTS 2 or NUTS 3 level. Finally, mean annual average values of the normalized difference vegetation index (NDVI) were extracted from Landsat images recorded between 2000 and 2019 at 30-m spatial resolution, as pre-processed by Hengl (2021), in order to characterize average photosynthetic activities of the LUCAS sites.

Of the initial 21,859 LUCAS sites, the following cases were omitted:

- land use that is neither cropland (annual or perennial) nor grassland (with or without sparse tree/shrub cover): 8136 sites (37% of total sites)
- organic soil with >8.7% SOC (different SOC dynamics and less area relevant): 1971 sites (9% of total sites)
- SOC or nitrogen (N) contents below the detection limit: 73 sites
- > 5% $CaCO_3$ (analytical uncertainties in SOC measurements): 4514 sites (21% of total sites; see Supplementary Material S1)
- observations where values were missing either for SOC or for at least one of the chosen predictors after filling missing soil texture values with data from earlier LUCAS sampling campaigns: 65 sites.

Ultimately this left 9925 LUCAS sites, of which 6155 were used as cropland and 3770 as grassland. The Supplementary Material S1 contains maps and background information characterizing this data set in detail.

Step 1. To estimate how much atmospheric C agricultural soils could theoretically sequester if today's cropland (group B) were converted to grassland (group A), a Random Forest model was trained to predict the SOC stocks of grassland. The following predictors were used for model training: soil texture, pH, C:N ratio, carbonate, soil groups, elevation, slope, aspect and climate data. These particular predictors were chosen because, based on expert knowledge, (i) they represent typical explanatory variables for SOC (Schneider et al., 2021) and (ii) land use can be assumed to have no causal effect on them, that is these predictors are independent of the examined land-use treatment. The resulting model was interpreted by computing permutation importance and illustrating associations of the five most important variables with SOC in partial dependence plots (Molnar, 2019).

Step 2. Second, the predictor space of the grassland model was examined. This space can be thought of as a point cloud with *n*-dimensions, where each dimension represents one predictor variable and each point in the cloud represents one grassland observation. The location of each point in this training space is defined by its respective value in each of the *n* predictor variables (here $n = 13$ continuous variables + one dummy coded soil group variable). After adding the cropland data to this space, (i) each predictor variable was standardized, that is mean-centred values were divided by their respective standard deviation, and (ii) the standardized values were multiplied by their respective variable importance in the grassland SOC model, as suggested by Meyer and Pebesma (2021). This was done to (i) make the predictor variables comparable and (ii) weight the predictors according to their importance in the SOC model. For each cropland observation, the Euclidean distance to its nearest grassland neighbour was calculated, and the result was divided by the average of all pairwise distances in the grassland data. This yielded a standardized index to characterize the dissimilarity of cropland and grassland observations. In the present study, the CAST::aoa function (Meyer, 2020) in R was used to calculate dissimilarity indices and define the associated space of applicability for the grassland model by applying the function's default threshold value of 0.95 (Meyer, 2020; Meyer & Pebesma, 2021). Cropland observations beyond the space of applicability of the grassland model were omitted.

Step 3. Finally, the SOC model, which was trained on grassland (group A), was used to predict SOC in the remaining cropland sites (group B). The resulting residuals represented the required estimates for the site-specific SOC response to converting today's cropland to grassland.

Step 4. In order to explain site-specific differences between these estimated effect sizes (residuals), a new Random Forest model was trained using all the available predictors for SOC, including those predictors that depended on the examined land-use treatment: variables characterizing agricultural activity and NDVI. After calculating permutation importance, selected key variables explaining SOC responses to land-use change were illustrated in partial dependence plots (Molnar, 2019).

Data analysis was performed using R v4.0.2 (R Core Team, 2020) in Rstudio v1.3.959 (RStudio Team, 2020) and built on tidyverse packages (Wickham et al., 2019). Random Forests were implemented using the ranger package (Wright & Ziegler, 2017), with mtry values chosen according to the square root of the number of predictor variables (Hastie et al., 2009). Model performance was evaluated using fivefold random cross-validation. When calculating permutation importance, collinearity was considered unproblematic since none of the continuous predictor pairs showed Spearman rank coefficients $|r_s| > 0.8$. Average values presented in the text are accompanied by their 95% bias-corrected and accelerated (BCa) confidence intervals based on 10^4 bootstrapped resamples. A commented version of the code used to implement data-driven reciprocal modelling for this case study is publicly accessible at https://doi.org/10.5281/zenodo.5171793. The same repository contains detailed instructions for compiling all necessary raw data and reproducing the presented findings from scratch.

3 | RESULTS AND DISCUSSION

The main result of the study was the new method of data-driven reciprocal modelling, as outlined in the Methods section. This method can be applied to a wide range of research questions for which large survey-based data are available. It helps to estimate treatment effects without being restricted to controlled experiments that include a control treatment and keeping all other environmental parameters the same, the *ceteris paribus* principle. In environmental research, this principle is often not achievable or comes at the expense of representativeness when being restricted to a few controlled field experiments. Here, we used the example of a soil survey and how land-use affects soil C stocks as a case study to illustrate the data-driven reciprocal modelling.

3.1 | Step 1. Train model on group A (here, grassland)

Random Forest predicted SOC stocks of grassland with moderate accuracy (R^2 = 0.59; RMSE = 17.4 Mg ha^{-1}; bias = −0.5 Mg ha^{-1}). The five most influential predictors described climatic conditions, soil texture and the C:N ratio of soils (Figure 2). Specifically, the model described SOC stocks increasing with precipitation in the driest month, attributing to grassland with more than 40 mm of such precipitation about 10 Mg ha^{-1} more SOC than grassland in a drier climate. Modelled SOC stocks decreased with increasing maximum temperatures and/or increasing coarse fragment fraction, while they increased with higher C:N ratios and/or higher clay content. Overall, the Random Forest model reflected data patterns that were in good agreement with previous studies explaining the variability of SOC stocks in European agricultural soils (Rial et al., 2017).

3.2 | Step 2. Stick to the model's space of applicability

In total, 276 (4%) of cropland sites were located outside the model's space of applicability, and were therefore excluded from further analyses (Figure 3; Figure 4 crosses). Affected sites were located in regions with strong cropland dominance, and were mostly characterized by a Mediterranean climate with maximum temperatures above 25°C in the hottest month and/or alkaline, clay-rich soil. These conditions were largely unknown to the model, because in Europe, grassland use is rare under such conditions.

3.3 | Step 3. Apply model to group B (here, cropland) to estimate effect sizes

In the remaining croplands, the model predicted SOC stocks to be 12.1 Mg C ha^{-1} (95% CI, 11.8 to 12.5) higher on average than measured values (Figure 4). This number illustrates the average SOC accrual if cropland were converted to grassland. Individual effect sizes differed drastically, but their spatial distribution revealed pronounced regional trends. On a country level, the potential SOC accrual from converting cropland to grassland was predicted to be highest in Belgium (mean 27.2 Mg C ha^{-1}; 95% CI, 24.7 to 29.3) and lowest in Estonia (mean 5.2 Mg C ha^{-1}; 95% CI, 0.8 to 8.3). Overall,

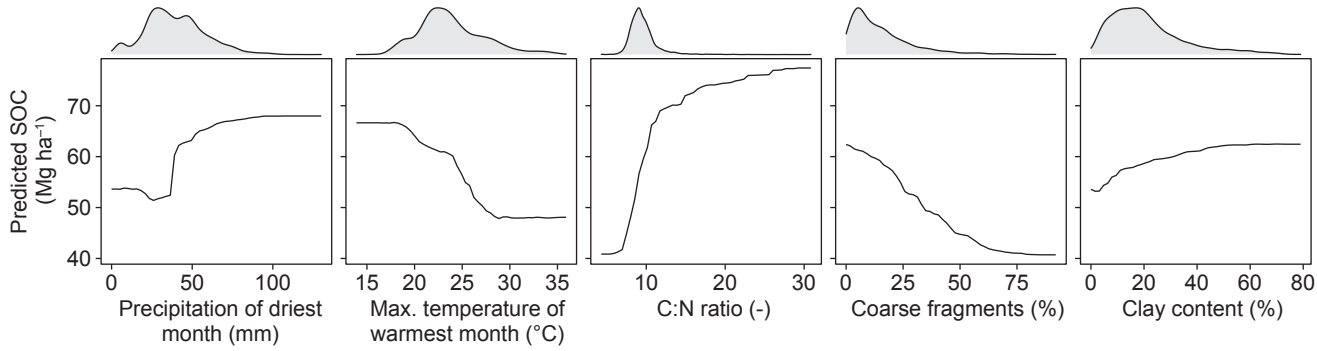

FIGURE 2 Partial dependence plots illustrating the modelled response of soil organic carbon (SOC) under grassland to important predictors. Density plots above show the distribution of these predictor variables in the training data (n = 3770)

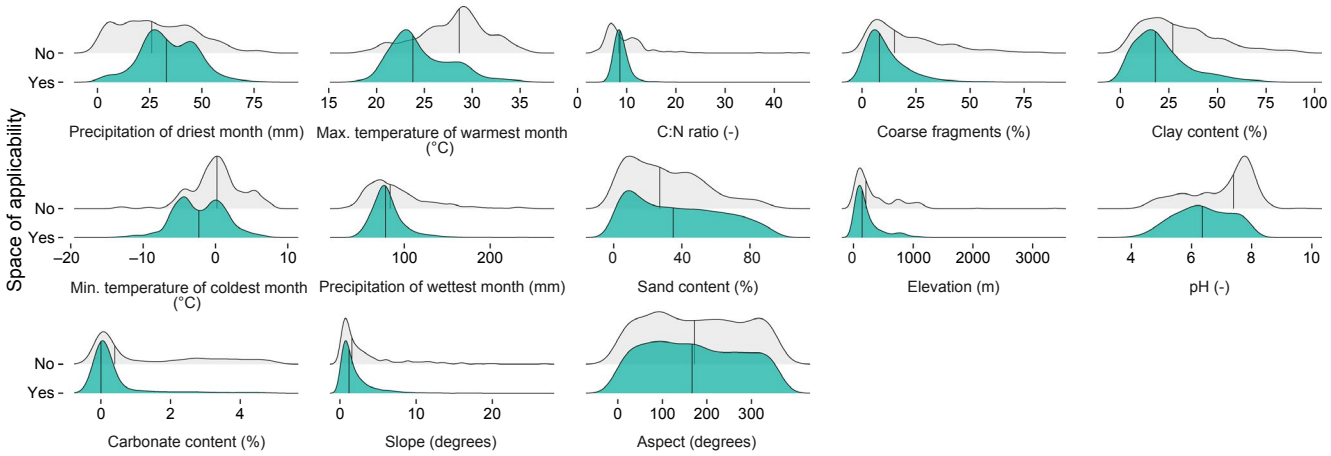

FIGURE 3 Density plots characterizing croplands within the model's space of applicability (green; $n = 5879$) and croplands outside of this space (grey; $n = 276$). All continuous predictors used for model training are shown, sorted by decreasing importance (from top left to bottom right). Vertical lines illustrate median values

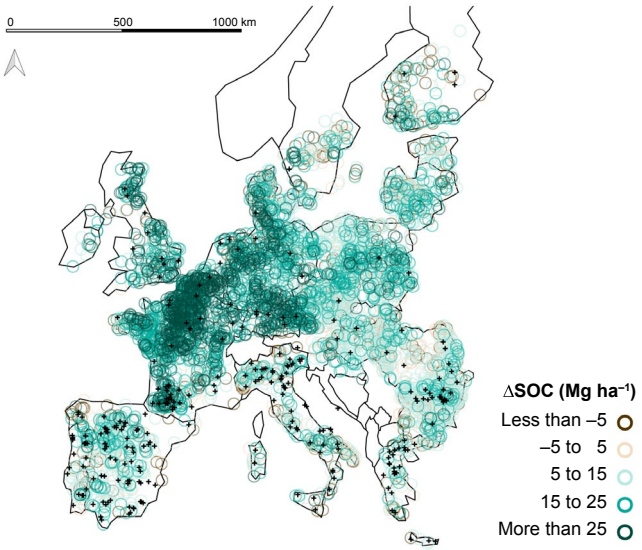

FIGURE 4 Estimated change in soil organic carbon (SOC) stocks by converting European cropland to grassland. Positive values denote SOC accrual (green colours), negative values denote SOC losses (brown colours). Crosses represent cropland sites with environmental properties for which no reliable estimate could be made since they were located outside the model's space of applicability

the average effect sizes of land-use change on SOC were within the range of values that have previously been observed in long-term field experiments (Poeplau & Don, 2013).

3.4 | Step 4: Explain effect sizes

The effects of land-use change on SOC stock were strongly related to climate, mineral N fertilization, C:N ratios of soils and mean annual NDVI values of the land surface (Figure 5). Cropland with mean annual precipitation below 40 mm in the driest month was predicted to sequester about 6 Mg C ha^{-1} less by grassland conversion than cropland with precipitation above this threshold value. This can largely be explained by the relatively low SOC stocks of grassland in dry areas (Figure 2). Additionally, in the dry areas of the Mediterranean, the widespread occurrence of silvoarable systems, such as olive trees paired with ley or cereals, might make croplands particularly SOC rich relative to adjacent grassland (Eichhorn et al., 2006). Converting such relatively productive silvoarable systems in the Mediterranean to grassland might even result in SOC losses (Figure 2, brown colours). This interpretation is underlined by the partial dependence of ΔSOC on the NDVI values of cropland: ΔSOC decreased with increasing mean annual NDVI values of cropland (Figure 5). Some Mediterranean croplands showed even higher mean annual NDVI values than neighbouring grasslands, that is cropland was 'greener' than grassland (Figure 6, green colours). However, on average this differed across Europe, with cropland showing lower mean annual average NDVI values than neighbouring grassland, particularly in the agricultural area between the Paris basin in France and Belgium (Figure 6). The same region was characterized by high mineral N fertilization, which was related to larger values of ΔSOC. Associations between soil properties and ΔSOC were minor, with the exception of the C:N ratio of cropland soil, underlining the close linkage between C and N cycles.

3.5 | Pros and cons of data-driven reciprocal modelling

Data-driven reciprocal modelling offers a new tool for case-specific estimates of effect sizes that take local environmental conditions into account. In our case study, the spatial patterns of predicted land-use change effects could be explained with comprehensible driver variables, confirming that most effect size estimates were plausible. However, it is important to note that the case-specific effect size estimates obtained via reciprocal modelling tend to be

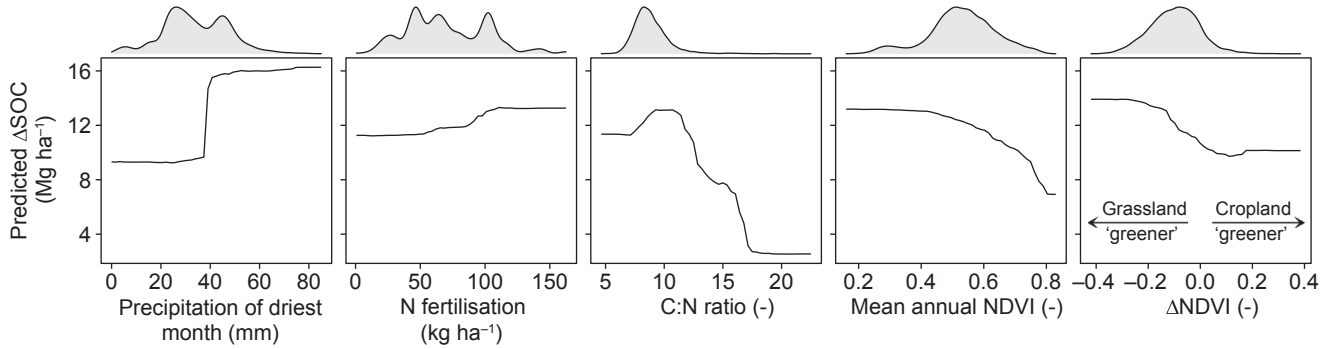

FIGURE 5 Partial dependence plots illustrating the modelled response of soil organic carbon to land-use change (ΔSOC) as a function of important predictors. Density plots above show the distribution of these predictor variables in the training data (n = 5879). In the rightmost panel, ΔNDVI represents the local difference in mean annual NDVI between cropland and grassland

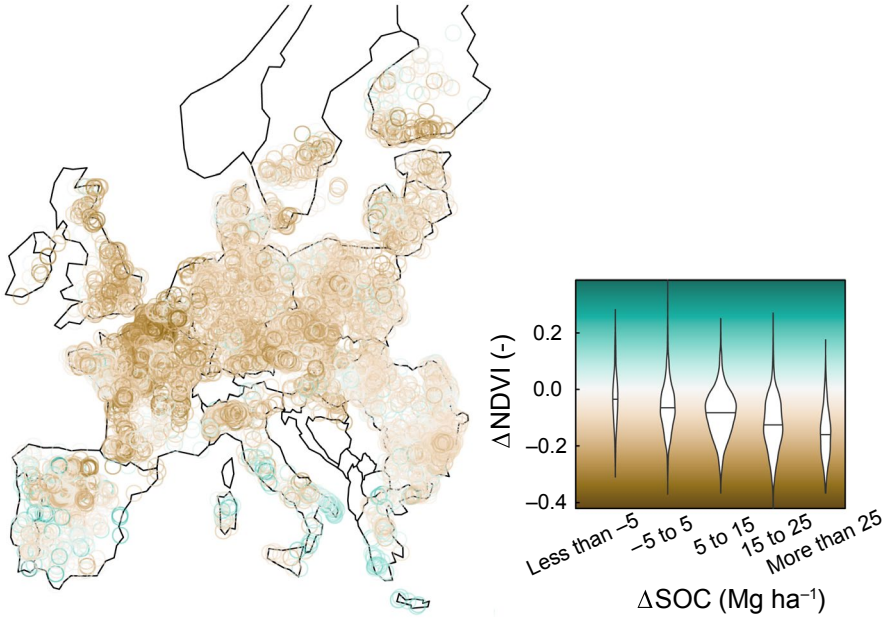

FIGURE 6 Difference in mean annual NDVI between each cropland and its five nearest grassland neighbours (ΔNDVI). Brown colours illustrate cropland being 'browner' than grassland (negative ΔNDVI), green illustrates the opposite. The lower the ΔNDVI, the higher the predicted SOC accrual in cropland (ΔSOC)

less accurate than what can theoretically be achieved with separate controlled experiments for each new case (Table 1). Results from reciprocal modelling are subject to both random and systematic errors. Most data-driven algorithms, including the most popular implementations of Random Forest, are designed to perform well on the majority of observations, which typically results in predictions being pulled towards the mean of the target variable. High values tend to be underestimated and low values overestimated. Such behaviour induces large errors among effect sizes for extreme observations. For example, in this study, this was the case for soils with extremely high SOC contents close to the defined boundary to organic soils. As these soils represented only a minor portion of the training data, Random Forest models significantly underestimated their SOC stocks, resulting in systematically underestimated SOC accrual for converting relatively SOC-rich cropland to grassland. If a given research question requires the adequate prediction of extreme effect sizes, then quantile prediction methods could be tested. However, this would compromise the prediction accuracy of the majority of observations. Therefore, a decision has to be made as to whether it is more important to optimize model accuracy for the majority or for extremes. The interpretation of individual effect sizes should be adapted accordingly: if the model is optimized to perform well on the majority, as in this study, then outliers and tails of the effect size distribution are prone to large errors and should therefore not be interpreted. Other sources of error in effect size estimates obtained via data-driven reciprocal modelling relate to poor data quality. In particular, undiscovered confounding factors due to missing training data can be another source of systematic error. For example, in the present study, SOC of European agricultural soils was assumed to be at steady state, that is SOC was assumed to neither decline nor increase under current management. This was a simplification that was certainly not true at all sites. For example, in Germany, about 10% of agricultural land showed changing land use within the past 20 years (German Agricultural Soil Inventory, unpublished data). In regions with widespread land-use changes within this period, it is likely that the present framework will have locally underestimated

TABLE 1 Pros and cons of determining effect sizes with controlled experiments, conventional mechanistic modelling and novel data-driven reciprocal modelling

	Controlled experiments - bottom-up approach -	Mechanistic modelling - extended bottom-up approach -	Data-driven reciprocal modelling - top-down approach -
Accuracy	High	High under the typically quite narrow range of conditions for which it was calibrated. Low elsewhere	Medium
Minimum limit of detection	Low	Medium. Can be improved by better mathematical models	High. Can be improved by more and better data, to a small extent also by model tuning
Response in space and time	None (constant value)	As implemented (restricted to known processes, which can be mathematically described)	Based on associations in the training data (not restricted to known processes)
Scalability	Low	Medium. Every process influencing the target must be described mathematically (requires detailed process understanding and time). Numerous assumptions	High. Can be implemented at any spatio-temporal scale on an endless number of different treatments. No process understanding required to build the model—algorithm does it within seconds to minutes. No or few assumptions
Confounding factors	Addressed by randomization (if possible), or simplifying environmental conditions	Addressed by manual model extensions (requires in-depth process understanding)	Addressed by data collection
Quantifying effects of previously unseen, novel treatments	Possible	Possible if underlying processes can be described mathematically	Not possible
Ethical concerns	Treatments may inflict harm on the environment (e.g. toxicology studies in the field)	Less of an issue	Problematic historic treatments, for example oil spills, do not have to be repeated to estimate their effects elsewhere—we can learn from the past. But in social contexts, the method should be handled with extra care as it tends to neglect minorities

potential SOC accrual if such sites are not filtered out for model training. The remaining sources of error in data-driven reciprocal modelling tend to be random rather than systematic. Random errors typically evolve around non-matching temporal and/or spatial resolutions in training data sets. For example, in this study, the target variable (SOC) was recorded at single geographical points, while the associated agricultural activity data were only available as average values at the level of administrative regions. Random errors also frequently occur in the gathering and processing of data—these errors can be an additional reason for individual outliers.

A major bottleneck in the accurate prediction of treatment effects with reciprocal modelling is data availability. In the future, the quantity and quality of data, as well as computational power to train data-driven models, will continue to increase, which will allow further improvements to the accuracy of data-driven effect size estimates. Controlled experiments will always be indispensable as they contribute to process understanding and build the foundation for mechanistic models. Good mechanistic models that correctly describe all relevant processes will always provide better effect size estimates than data-driven reciprocal modelling. In practice, however, many phenomena are too complex to be accurately described with mechanistic models. Data-driven reciprocal modelling allows effect size to be estimated merely by analysing associations, without understanding causation. They provide the best possible quantitative estimate of effects without understanding every detail of their cause while achieving a maximum of representativeness.

ACKNOWLEDGEMENTS

This study was part of the EJPSoil project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 862695. Funding was also received via the BonaRes project Soil3 (grant number 031B0515E) of the German Federal Ministry of Education and Research (BMBF). The LUCAS 2015 topsoil data set used in this work was made available by the European Commission through the European Soil Data Centre managed by the Joint Research Centre (JRC), http://esdac.

jrc.ec.europa.eu/. Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

There were no conflicts of interests.

DATA AVAILABILITY STATEMENT

The R code used for data-driven reciprocal modelling in our case study is available on GitHub (https://github.com/FlorianSchne/reciprocalModelling) and Zenodo (https://doi.org/10.5281/zenodo.5171793). Data were derived from the following resources available in the public domain: European Soil Data Centre (https://esdac.jrc.ec.europa.eu/), European Statistical Office (https://ec.europa.eu/eurostat) and Open Data Science Europe (https://maps.opendatascience.eu/).

ORCID

Florian Schneider https://orcid.org/0000-0003-3036-6284
Christopher Poeplau https://orcid.org/0000-0003-3108-8810
Axel Don https://orcid.org/0000-0001-7046-3332

REFERENCES

Bastin, J.-F., Finegold, Y., Garcia, C., Mollicone, D., Rezende, M., Routh, D., Zohner, C. M., & Crowther, T. W. (2019). The global tree restoration potential. *Science*, *365*(6448), 76–79. https://doi.org/10.1126/science.aax0848

Don, A., Schumacher, J., & Freibauer, A. (2011). Impact of tropical land-use change on soil organic carbon stocks—A meta-analysis. *Global Change Biology*, *17*(4), 1658–1670. https://doi.org/10.1111/j.1365-2486.2010.02336.x

Eichhorn, M. P., Paris, P., Herzog, F., Incoll, L. D., Liagre, F., Mantzanas, K., Mayus, M., Moreno, G., Papanastasis, V. P., Pilbeam, D. J., Pisanelli, A., & Dupraz, C. (2006). Silvoarable systems in Europe—past, present and future prospects. *Agroforestry Systems*, *67*(1), 29–50. https://doi.org/10.1007/s10457-005-1111-7

Eurostat. (2020). *Main farm land use by NUTS 2 regions [ef_lus_main]*. Author. https://ec.europa.eu/eurostat/web/products-datasets/-/ef_lus_main

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, *37*(12), 4302–4315. https://doi.org/10.1002/joc.5086

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media.

Hengl, T. (2021). NDVI Landsat (quarterly). NDVI time-series, derived from the Landsat quarterly temporal composites. https://data.opendatascience.eu/geonetwork/srv/eng/catalog.search#/metadata/b69476b8-4d6c-4381-a719-649574cb917e

Hollis, J. M., Hannam, J., & Bellamy, P. H. (2012). Empirically-derived pedotransfer functions for predicting bulk density in European soils. *European Journal of Soil Science*, *63*(1), 96–109. https://doi.org/10.1111/j.1365-2389.2011.01412.x

Huston, M. A. (1997). Hidden treatments in ecological experiments: Re-evaluating the ecosystem function of biodiversity. *Oecologia*, *110*(4), 449–460. https://doi.org/10.1007/s004420050180

Jones, A., Fernández-Ugalde, O., & Scarpa, S. (2020). *LUCAS 2015 topsoil survey: Presentation of dataset and results. EUR: Vol. 30332*. Publications Office of the European Union. https://doi.org/10.2760/616084

Lugato, E., Panagos, P., Bampa, F., Jones, A., & Montanarella, L. (2014). A new baseline of organic carbon stock in European agricultural soils using a modelling approach. *Global Change Biology*, *20*(1), 313–326. https://doi.org/10.1111/gcb.12292

Meyer, H. (2020). *CAST: 'Caret' applications for spatial-temporal models*. https://CRAN.R-project.org/package=CAST

Meyer, H., & Pebesma, E. (2021). Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution*. https://doi.org/10.1111/2041-210X.13650

Minasny, B., Malone, B. P., McBratney, A. B., Angers, D. A., Arrouays, D., Chambers, A., Chaplot, V., Chen, Z.-S., Cheng, K., Das, B. S., Field, D. J., Gimona, A., Hedley, C. B., Hong, S. Y., Mandal, B., Marchant, B. P., Martin, M., McConkey, B. G., Mulder, V. L., … Winowiecki, L. (2017). Soil carbon 4 per mille. *Geoderma*, *292*, 59–86. https://doi.org/10.1016/j.geoderma.2017.01.002

Molnar, C. (2019). *Interpretable machine learning. A guide for making black box models explainable*. https://christophm.github.io/interpretable-ml-book/

Orgiazzi, A., Ballabio, C., Panagos, P., Jones, A., & Fernández-Ugalde, O. (2018). LUCAS soil, the largest expandable soil dataset for Europe: A review. *European Journal of Soil Science*, *69*(1), 140–153. https://doi.org/10.1111/ejss.12499

Panagos, P., Ballabio, C., Scarpa, S., Borrelli, P., Lugato, E., & Montanarella, L. (2020). *Soil related indicators to support agro-environmental policies: Soil erosion soil carbon soil nutrients and fertility. EUR: Vol. 30090*. Publications Office of the European Union. https://doi.org/10.2760/011194

Paustian, K., Lehmann, J., Ogle, S., Reay, D., Robertson, G. P., & Smith, P. (2016). Climate-smart soils. *Nature*, *532*(7597), 49–57. https://doi.org/10.1038/nature17174

Poeplau, C., & Don, A. (2013). Sensitivity of soil organic carbon stocks and fractions to different land-use changes across Europe. *Geoderma*, *192*, 189–201. https://doi.org/10.1016/j.geoderma.2012.08.003

Poeplau, C., Don, A., Vesterdal, L., Leifeld, J., Wesemael, B. A., Schumacher, J., & Gensior, A. (2011). Temporal dynamics of soil organic carbon after land-use change in the temperate zone—Carbon response functions as a model approach. *Global Change Biology*, *17*(7), 2415–2427. https://doi.org/10.1111/j.1365-2486.2011.02408.x

R Core Team. (2020). *R: A language and environment for statistical computing*. https://www.R-project.org/

Rial, M., Martínez Cortizas, A., & Rodríguez-Lado, L. (2017). Understanding the spatial distribution of factors controlling topsoil organic carbon content in European soils. *The Science of the Total Environment*, *609*, 1411–1422. https://doi.org/10.1016/j.scitotenv.2017.08.012

RStudio Team. (2020). *RStudio: Integrated development environment for R*. http://www.rstudio.com/

Schneider, F., Amelung, W., & Don, A. (2021). Origin of carbon in agricultural soil profiles deduced from depth gradients of C:N ratios, carbon fractions, $\delta^{13}C$ and $\delta^{15}N$ values. *Plant and Soil*, *460*(1–2), 123–148. https://doi.org/10.1007/s11104-020-04769-w

Schneider, F., & Don, A. (2019). Root-restricting layers in German agricultural soils. Part I: Extent and cause. *Plant and Soil*, *442*(1–2), 433–451. https://doi.org/10.1007/s11104-019-04185-9

Wadoux, A.-M.-J.-C., Samuel-Rosa, A., Poggio, L., & Mulder, V. L. (2020). A note on knowledge discovery and machine learning in digital soil mapping. *European Journal of Soil Science*, *71*(2), 133–136. https://doi.org/10.1111/ejss.12909

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wright, M. N., & Ziegler, A. (2017). Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, *77*(1). https://doi.org/10.18637/jss.v077.i01

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the Supporting Information section.

How to cite this article: Schneider, F., Poeplau, C., & Don, A. (2021). Predicting ecosystem responses by data-driven reciprocal modelling. *Global Change Biology*, 27, 5670–5679. https://doi.org/10.1111/gcb.15817