Causal Modeling of Soil Processes for Improved Generalization

Anonymous Author(s)
Affiliation
Address
email

Abstract

Measuring and monitoring soil organic carbon is critical for agricultural productivity and for addressing critical environmental problems. Soil organic carbon not only enriches nutrition in soil, but also has a gamut of co-benefits such as improving water storage and limiting physical erosion. Despite a litany of work in soil organic carbon estimation, current approaches do not generalize well across soil conditions and management practices. We empirically show that explicit modeling of cause-and-effect relationships among the soil processes improves the out-of-distribution generalizability of prediction models. We provide a comparative analysis of soil organic carbon estimation models where the skeleton is estimated using causal discovery methods. Our framework provide an average improvement of 81% in test mean squared error and 52% in test mean absolute error.

1 Introduction

Soil organic carbon, the carbon component of organic compounds found in soil both as biomass and as sequestered compounds and necromass, has been called "natural insurance against climate change" [35]—with evidence associating increased soil organic matter with increased crop yields [30, 32, 37]. Climate change is increasing the variability in crop yields and increasing food insecurity [29, 19, 20]. This variability in yield is further exacerbated by conventional soil management practices unconcerned with soil organic carbon. Proper management of soil, including its organic carbon component, can mitigate shortages in food, water, energy and adverse repercussions of climate change [20]. Measuring and monitoring soil organic carbon can therefore have a positive impact in solving several environmental problems. This has led to increased interest by environmentalists, economists and soil scientists, as interdisciplinary collaborations, in improving public awareness and policy making [8, 18, 9, 17, 26, 7].

While the problem of studying soil organic carbon is well-motivated, forming hypotheses and designing experiments to estimate soil organic carbon can be challenging. Changes in soil organic carbon are not only dictated by weather events and management practices, but also by other soil processes such as plant nutrient uptake, soil organisms, soil texture, micro-nutrient content and soil disturbance. This makes soil a complex "living" porous medium [16, 20, 15]. Due to the complex nature of soil science, the exact relations among all soil processes is not yet known. There is no accepted universal method for studying soil organic carbon and the relation among soil processes [5]. Moreover, current models of soil organic carbon (RothC-26.3 [1], Century [2], DNDC [10], and some inter-model comparisons [12, 33, 14, 13]) are not consistent with the latest advances in understanding

Submitted to Tackling Climate Change with Machine Learning: workshop at NeurIPS 2022.

of soil processes and do not generalize to different soil conditions found globally as they are limited by their modeling assumptions[22].

Our goal is to create a method that can generalize well across regions, soil types and soil management practices. We aim to create a causal machine learning framework that can aid in standardization efforts for soil organic carbon measuring, reporting and verification. Although ML methods have demonstrated an improved ability to predict soil organic carbon [45, 39], the reliance of conventional ML on the i.i.d. assumption that training data represents the deployed environment limits their out-of-distribution generalizability [40, 39, 49, 36]. To improve this, either large (and diverse) data sets can be utilized or careful architectural modifications can help in creating a surrogate model or an emulator of the real-world physical system. These architectural modifications are also limited by domain experts' understanding of the physical system [44]. Using causal discovery methods is a way to overcome this limitation as these frameworks help explicitly model cause-and-effect relations among the soil processes to improve the out-of-distribution generalizability.

In this paper, we present an approach based on causal graphs to estimate soil organic carbon stocks. We provide a comparative analysis of soil organic carbon estimation using causal and non-causal approaches and show that causal approaches produce better results for soil organic carbon estimation on unseen fields (from different locations, with different soil properties, management practices and land use). We briefly discuss related literature in Section 2, followed by the problem formulation, data set and our methodology in Section 3. Results, discussion and future work is discussed in Section 4.

2 Background

Our approach combines recent advances in causal discovery and graph neural networks (GNNs). Causal discovery [46, 31, 21] is an approach for identifying cause-and-effect relations between variables of a system, using data under causal ignorability and sufficiency assumptions and leveraging partial *a priori* knowledge of relationships. Utilizing such causal discovery methods can quantify complex interactions of the different soil processes that govern soil organic carbon and its effects on soil quality, which cannot be directly measured but are emergent properties. Quantifying how soil organic carbon stocks are influenced by other soil process, and how soil organic carbon affects other soil processes and soil functions can move us closer to a universal or standardized modeling framework for all soil processes and for measuring soil organic carbon. Also, graph neural networks [34, 41, 48, 50] present approaches to work with non-Euclidean graph data and model complex relationships between entities. A survey of recent GNN advancements can be found here [47]. Typically, GNN based methods assume homogeneity of nodes. Direct application of GNN approaches is not straightforward when nodes and edges are heterogeneous; this is the case in our application, where both the nature of the node (nodes could represent soil processes, climate variables, management practices) and the type of data associated with each node (e.g., soil process nodes could constitute continuous geochemical composition changes while farming or management practices might be frequency of an operation being performed on the farm) differ markedly.

3 Materials and Methods

3.1 Data

We utilize an extensive and rich data set from the North Wyke Farm Platform `http://www.rothamsted.ac.uk/north-wyke-farm-platform`. This data is available for multiple fields with different land use types, which makes it appropriate for studying spatial out-of-distribution generalization, and limits bias due to causal ignorability and sufficiency assumptions. The North Wyke Farm Platform data consists of observations of three pasture-based livestock farming systems, each consisting of five component catchments of approximately 21 ha each. High resolution long term data including soil organic carbon, soil total nitrogen, pH as well as management practices are collected to study the sustainability of different types of land use (treatments) over time (2012 to present). In the baseline period (April 2011 to March 2013), all three farming systems were managed as permanent pastures, grazed by livestock and sheep. In April 2013, one system was resown with high sugar grasses, having a high water-soluble carbohydrate content with the aim of increasing livestock growth ("the red system"), a second system was resown with a high sugar grass-white clover mix ("the blue system") to reduce the requirement for inorganic nitrogen fertilizer application. The remaining ("the green system") continued as a permanent pasture for long-term monitoring. Appendix A.2 shows a map of the North Wyke Farms showing the layout of the individual farms and their management practices [51].

We create a train-test split to ascertain the generalizability of the proposed approach when fields are managed differently. For example, inversion ploughing is an important management practice because it results in the loss of organic carbon from agricultural soils[24]. Figures 1 and 2 show the distribution of soil organic carbon and number of times the fields were ploughed for our data. Note that the red and blue systems were ploughed whereas the green system was not. This is also seen as a consistent higher levels of carbon for the green fields than the red and blue fields. Training data consists of 7 red and 8 blue system fields. Test data consists of a total of 7 green system fields. Our data set comprises management practices (including the number of fertilizer applications, pesticide applications, plough events, etc.), total nitrogen, total organic carbon and soil pH for each field. More details on data preprocessing are included in Appendix A.1.

3.2 Problem Formulation

Complex interactions between soil organic carbon, soil processes and other exogenous factors (e.g., environmental and management practices) limit the generalization capabilities of conventional ML methods. In this paper, our aim is twofold, understanding how different management practices affect soil organic carbon and then estimating it in a way that it generalizes in out-of-distribution environments. Let $M = M_1, M_2, M_3, \ldots, M_n$ represent farm management practices, C, O represent soil organic carbon and other observed soil properties respectively, studied across k locations.

Our approach is first to learn the cause-and-effect relations among soil variables (in this study they are, soil organic carbon, nitrogen and pH) and the management practices followed in the farm. The learned causal graph of different soil processes can be represented as $G \in \mathbb{R}^{(|M|+|O|+|C|)\times(|M|+|O|+|C|)}$. Depending on the causal discovery method employed for learning the causal graph, edge indices and attributes can be derived to create a skeleton that can be utilized in GNN-based regression to estimate soil organic carbon at a location as a function of the other variables. The generalization power of a causal graph skeleton based upon a GNN model relies on the graph G that is used as prior knowledge for the prediction task. Here, for the regression task, instead of measuring the conditional expected response $\mathbb{E}(C|M, O)$, we evaluate $\mathbb{E}(C|M, O, G)$ which is influenced by not only $p(D)$ but also causal graph G, where $D = \{C, M, O\}$. Depending on the causal discovery method, additional assumptions can be made about the data [21].

3.2.1 Causal Discovery

In our experiments, a causal graph consists of nodes—representing variables or physical processes and directed edges represent causal relationships between nodes. While prior knowledge or trial-and-error guessing can be used to create causal graphs, we make use of established causal discovery algorithms to create the directed graphs. To generate causal graphs using the North Wyke Farm Platform data, we use the PC algorithm, a constraint-based method, [4] and two score-based methods, Greedy Equivalence Search (GES) [6, 3] and Greedy Interventional Equivalence Search (GIES) [11]. See Appendix A.3 for more details.

3.2.2 Causal Graph Neural Network

While causal graphs estimated from causal discovery methods are used to obtain skeletons for GNNs, we compare two paradigms, Edge-Conditioned Convolution Message Passing Neural Networks (ECC MPNNs) [28, 23] and GraphSAGE [25]. Comparing different message passing procedures allows us to study how added complexity in learning influences generalization in the prediction task.

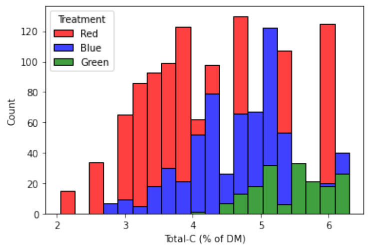

Figure 1: Soil organic carbon distribution

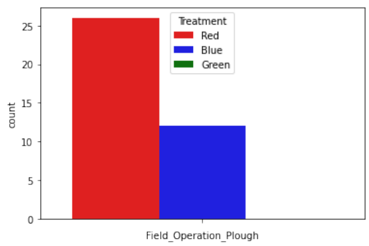

Figure 2: Plough event distribution for the three treatments

Model	Causal	MSE	MAE
PC + GraphSAGE	Yes	1.1002	1.0116
GES + GraphSAGE	Yes	0.9248	0.9193
GIES + GraphSAGE	Yes	0.7803	0.7904
PC + ECC MPNN	Yes	**0.2816**	**0.5258**
GES + ECC MPNN	Yes	0.2864	0.5302
GIES + ECC MPNN	Yes	0.2951	0.5385
Random Edges + GraphSAGE	No	2.9052	1.6686
XGBoost	No	4.5007	2.0860
MLP	No	3.8415	1.9254
Random Forest	No	2.7996	1.6263

Table 1: Comparison of soil organic carbon estimation approaches based on Mean Squared Error (MSE) and Mean Absolute Error (MAE) to show how well different approaches generalize for permanent pasture i.e. the "green system". We use high sugar grass pastures ("red" and "blue" systems) for training and the "green" system for testing. We compare GraphSAGE architecture and ECC MPNN architecture.

These methods adopt different neighborhood definitions to compute message passing signals. For a directed graph $G(V, E)$, where V is a finite set of nodes and E is a set of edges, we can define a neighborhood for a given node i as $N(i)$. For an ECC MPNN, $N(i)$ comprises all of the ancestors in the directed graph. In GraphSAGE, a neighborhood is defined as a function of varying search depths $k \in \{0, 1, ..., K\}$, wherein, the number of adjacent nodes are sub-sampled for message passing with a node i. Starting at $k = 0$, the neighboring feature vectors are aggregated at each search depth k and concatenated with a node's representation. The final representation is obtained as the aggregation at depth K. More details on how node embeddings are updated are mentioned in Appendix A.4.

4 Results and Discussion

Our experiments investigate the impact of soil processes and farming practices on soil organic carbon estimation. Through empirical evidence, we demonstrate the improvement in out-of-distribution generalization offered by coupling causal discovery methods with GNNs. To study generalization power, we use the test set Mean Squared Error (MSE) and Mean Absolute Error (MAE) as evaluation metrics. Results in Table 1 suggest that causal approaches outperform non-causal approaches for soil organic carbon estimation and generalize well to unseen locations. The causal graph generated by the PC method is more parsimonious than the score-based methods' graphs and offers the best prediction skill when used as skeleton for ECC MPNN model. The causal graphs resulting from the 3 causal discovery approches are in Appendix A.6 and the details of ML algorithms and hyperparameter tuning are in Appendix A.5.

Causal modeling has the potential to offer further improvements. Several new advancements utilize continuous optimization methods to learn causal graphs [46] and may offer better time complexity for higher dimensional data sets. In further experiments, more explicit incorporation of temporal heterogeneity can also be done via methods like GNN-RNN [42, 43] for soil organic carbon estimation and using approaches like amortized causal discovery [38]. While our method generalizes well across different soil types, management practices and land use, a more extensive study across global geographies can aid in evaluation over a broader range of soil conditions, crops, and weather patterns.

References

[1] K. Coleman and D. S. Jenkinson. “RothC-26.3 - A Model for the turnover of carbon in soil”. In: *Evaluation of Soil Organic Matter Models*. Ed. by David S. Powlson, Pete Smith, and Jo U. Smith. Berlin, Heidelberg: Springer Berlin Heidelberg, 1996, pp. 237–246. ISBN: 978-3-642-61094-3.

[2] W. J. Parton. “The CENTURY model”. In: *Evaluation of Soil Organic Matter Models*. Ed. by David S. Powlson, Pete Smith, and Jo U. Smith. Berlin, Heidelberg: Springer Berlin Heidelberg, 1996, pp. 283–291. ISBN: 978-3-642-61094-3.

[3] Christopher Meek. “Graphical Models: Selecting causal and statistical models”. PhD thesis. PhD thesis, Carnegie Mellon University, 1997.

[4] Peter Spirtes et al. *Causation, prediction, and search*. MIT press, 2000.

[5] Johan Bouma. “Land quality indicators of sustainable land management across scales”. In: *Agriculture, Ecosystems & Environment* 88.2 (2002), pp. 129–136.

[6] David Maxwell Chickering. “Optimal structure identification with greedy search”. In: *Journal of machine learning research* 3.Nov (2002), pp. 507–554.

[7] Rattan Lal. “Soil carbon sequestration to mitigate climate change”. In: *Geoderma* 123.1-2 (2004), pp. 1–22.

[8] Rattan Lal et al. *Managing soil carbon*. 2004.

[9] Robert E White. *Principles and practice of soil science: the soil as a natural resource*. John Wiley & Sons, 2005.

[10] Donna L. Giltrap, Changsheng Li, and Surinder Saggar. “DNDC: A process-based model of greenhouse gas fluxes from agricultural soils”. In: *Agriculture, Ecosystems & Environment* 136.3 (2010). Estimation of nitrous oxide emission from ecosystems and its mitigation technologies, pp. 292–300. ISSN: 0167-8809. DOI: https://doi.org/10.1016/j.agee.2009.06.014. URL: https://www.sciencedirect.com/science/article/pii/S0167880909001996.

[11] Alain Hauser and Peter Bühlmann. “Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs”. In: *The Journal of Machine Learning Research* 13.1 (2012), pp. 2409–2464.

[12] Taru Palosuo et al. “A multi-model comparison of soil carbon assessment of a coniferous forest stand”. In: *Environmental Modelling & Software* 35 (2012), pp. 38–49.

[13] WN Smith et al. “Crop residue removal effects on soil carbon: Measured and inter-model comparisons”. In: *Agriculture, ecosystems & environment* 161 (2012), pp. 27–38.

[14] Katherine EO Todd-Brown et al. “Causes of variation in soil carbon simulations from CMIP5 Earth system models and comparison with observations”. In: *Biogeosciences* 10.3 (2013), pp. 1717–1736.

[15] Kristin Ohlson. *The soil will save us: How scientists, farmers, and foodies are healing the soil to save the planet*. Rodale Books, 2014.

[16] Vincent de Paul Obade and Rattan Lal. “Soil quality evaluation under different land management practices”. In: *Environmental earth sciences* 72.11 (2014), pp. 4531–4549.

[17] TAPAS Bhattacharyya and DK Pal. “The Soil: A natural resource”. In: *Soil Science: An Introduction; Rattan, RK, Katyal, JC, Dwivedi, BS, Sarkar, AK, Tapas Bhattacharyya, JC, Tarafdar, SK, Eds* (2015), pp. 1–19.

[18] Generose Nziguheba et al. “Soil carbon: a critical natural resource-wide-scale goals, urgent actions.” In: *Soil carbon: Science, management and policy for multiple benefits*. CABI Wallingford UK, 2015, pp. 10–25.

[19] Deepak K Ray et al. “Climate variation explains a third of global crop yield variability”. In: *Nature communications* 6.1 (2015), pp. 1–9.

[20] Vincent de Paul Obade and Rattan Lal. “Towards a standard technique for soil quality assessment”. In: *Geoderma* 265 (2016), pp. 96–102.

[21] Peter Spirtes and Kun Zhang. “Causal discovery and inference: concepts and recent methodological advances”. In: *Applied informatics*. Vol. 3. 1. SpringerOpen. 2016, pp. 1–28.

[22] H. Vereecken et al. “Modeling soil processes: Review, key challenges, and new perspectives”. English. In: *Vadose Zone Journal* 15.5 (2016). ISSN: 1539-1663. DOI: 10.2136/vzj2015.09.0131.

[23] Justin Gilmer et al. "Neural message passing for quantum chemistry". In: *International conference on machine learning*. PMLR. 2017, pp. 1263–1272.

[24] Neal R. Haddaway et al. "How does tillage intensity affect soil organic carbon? A systematic review". In: *Environmental Evidence* 6.1 (Dec. 2017), p. 30. ISSN: 2047-2382. DOI: 10.1186/s13750-017-0108-9. URL: https://doi.org/10.1186/s13750-017-0108-9.

[25] Will Hamilton, Zhitao Ying, and Jure Leskovec. "Inductive representation learning on large graphs". In: *Advances in neural information processing systems* 30 (2017).

[26] Budiman Minasny et al. "Soil carbon 4 per mille". In: *Geoderma* 292 (2017), pp. 59–86.

[27] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[28] Martin Simonovsky and Nikos Komodakis. "Dynamic edge-conditioned filters in convolutional neural networks on graphs". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 3693–3702.

[29] Meetpal S Kukal and Suat Irmak. "Climate-driven crop yield and yield variability and climate change impacts on the US Great Plains agricultural production". In: *Scientific reports* 8.1 (2018), pp. 1–18.

[30] Per Schjønning et al. "Chapter Two - The Role of Soil Organic Matter for Maintaining Crop Yields: Evidence for a Renewed Conceptual Basis". In: ed. by Donald L. Sparks. Vol. 150. Advances in Agronomy. Academic Press, 2018, pp. 35–79. DOI: https://doi.org/10.1016/bs.agron.2018.03.001. URL: https://www.sciencedirect.com/science/article/pii/S0065211318300245.

[31] Clark Glymour, Kun Zhang, and Peter Spirtes. "Review of causal discovery methods based on graphical models". In: *Frontiers in genetics* 10 (2019), p. 524.

[32] E. E. Oldfield, M. A. Bradford, and S. A. Wood. "Global meta-analysis of the relationship between soil organic matter and crop yields". In: *SOIL* 5.1 (2019), pp. 15–32. DOI: 10.5194/soil-5-15-2019. URL: https://soil.copernicus.org/articles/5/15/2019/.

[33] Boris Ťupek et al. "Evaluating CENTURY and Yasso soil carbon models for CO2 emissions and organic carbon stocks of boreal forest soil with Bayesian multi-model inference". In: *European Journal of Soil Science* 70.4 (2019), pp. 847–858.

[34] Chuxu Zhang et al. "Heterogeneous graph neural network". In: *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2019, pp. 793–803.

[35] Nils Droste et al. "Soil carbon insures arable crop production against increasing adverse weather due to climate change". In: *Environmental Research Letters* 15.12 (2020), p. 124034.

[36] Mostafa Emadi et al. "Predicting and mapping of soil organic carbon using machine learning algorithms in Northern Iran". In: *Remote Sensing* 12.14 (2020), p. 2234.

[37] Rattan Lal. "Soil organic matter content and crop yield". In: *Journal of Soil and Water Conservation* 75.2 (2020), 27A–32A. ISSN: 0022-4561. DOI: 10.2489/jswc.75.2.27A. eprint: https://www.jswconline.org/content/75/2/27A.full.pdf. URL: https://www.jswconline.org/content/75/2/27A.

[38] Sindy Löwe et al. *Amortized Causal Discovery: Learning to Infer Causal Graphs from Time-Series Data*. 2020. DOI: 10.48550/ARXIV.2006.10833. URL: https://arxiv.org/abs/2006.10833.

[39] José Padarian, Budiman Minasny, and Alex B McBratney. "Machine learning and soil sciences: A review aided by machine learning tools". In: *Soil* 6.1 (2020), pp. 35–52.

[40] Alexandre MJ-C Wadoux, Budiman Minasny, and Alex B McBratney. "Machine learning for digital soil mapping: Applications, challenges and suggested solutions". In: *Earth-Science Reviews* 210 (2020), p. 103359.

[41] Jie Zhou et al. "Graph neural networks: A review of methods and applications". In: *AI Open* 1 (2020), pp. 57–81.

[42] Joshua Fan et al. "A GNN-RNN Approach for Harnessing Geospatial and Temporal Information: Application to Crop Yield Prediction". In: *NeurIPS 2021 Workshop on Tackling Climate Change with Machine Learning*. 2021. URL: https://www.climatechange.ai/papers/neurips2021/29.

[43] Pedro Gomes, Silvia Rossi, and Laura Toni. "Spatio-temporal Graph-RNN for Point Cloud Prediction". In: *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2021, pp. 3428–3432.

[44] Alexander Lavin et al. “Simulation intelligence: Towards a new generation of scientific methods”. In: *arXiv preprint arXiv:2112.03235* (2021).

[45] Thu Thuy Nguyen. “Predicting agricultural soil carbon using machine learning”. In: *Nature Reviews Earth & Environment* 2.12 (2021), pp. 825–825.

[46] Matthew J Vowels, Necati Cihan Camgoz, and Richard Bowden. “D’ya like DAGs? A survey on structure learning and causal discovery”. In: *ACM Computing Surveys (CSUR)* (2021).

[47] Zonghan Wu et al. “A Comprehensive Survey on Graph Neural Networks”. In: *IEEE Transactions on Neural Networks and Learning Systems* 32.1 (Jan. 2021), pp. 4–24. DOI: 10.1109/tnnls.2020.2978386. URL: https://doi.org/10.1109%2Ftnnls.2020.2978386.

[48] Ying Chen et al. “Identifying field and road modes of agricultural Machinery based on GNSS Recordings: A graph convolutional neural network approach”. In: *Computers and Electronics in Agriculture* 198 (2022), p. 107082.

[49] Sabine Grunwald. “Artificial intelligence and soil carbon modeling demystified: power, potentials, and perils”. In: *Carbon Footprints* 1.1 (2022), p. 5.

[50] Lukasz Tulczyjew et al. “Graph Neural Networks Extract High-Resolution Cultivated Land Maps From Sentinel-2 Image Series”. In: *IEEE Geoscience and Remote Sensing Letters* 19 (2022), pp. 1–5.

[51] *North Wyke Farm Map*. Accessed:9/16/2022.

A Appendix

A.1 Data preprocessing

Field names (22 fields) and management practices (54 practices) are one hot encoded. Numerical data (i.e., total Nitrogen, total carbon and soil pH) is scaled using a min-max scaling scheme. In addition, to understand the long-term cumulative effects of changes in management practices, we include lag variables that capture the number of times a management practice / farming operation was performed in the last 1.5 months, 6 months, 1 year, 2 years. Different data are available at varying cadence, so we merge them together at a daily level by averaging the values of features collected at a finer resolution.

A.2 North Wyke Data Farm

Layout of the North Wyke Data Farm showing color coded fields to represent the land use type. The red system is sown with high sugar grasses, having a high water-soluble carbohydrate content with the aim of increasing livestock growth, the blue system sown with a high sugar grass-white clover mix to reduce the requirement for inorganic nitrogen fertilizer application. The remaining fields continued as a permanent pasture for long-term monitoring (the green system).

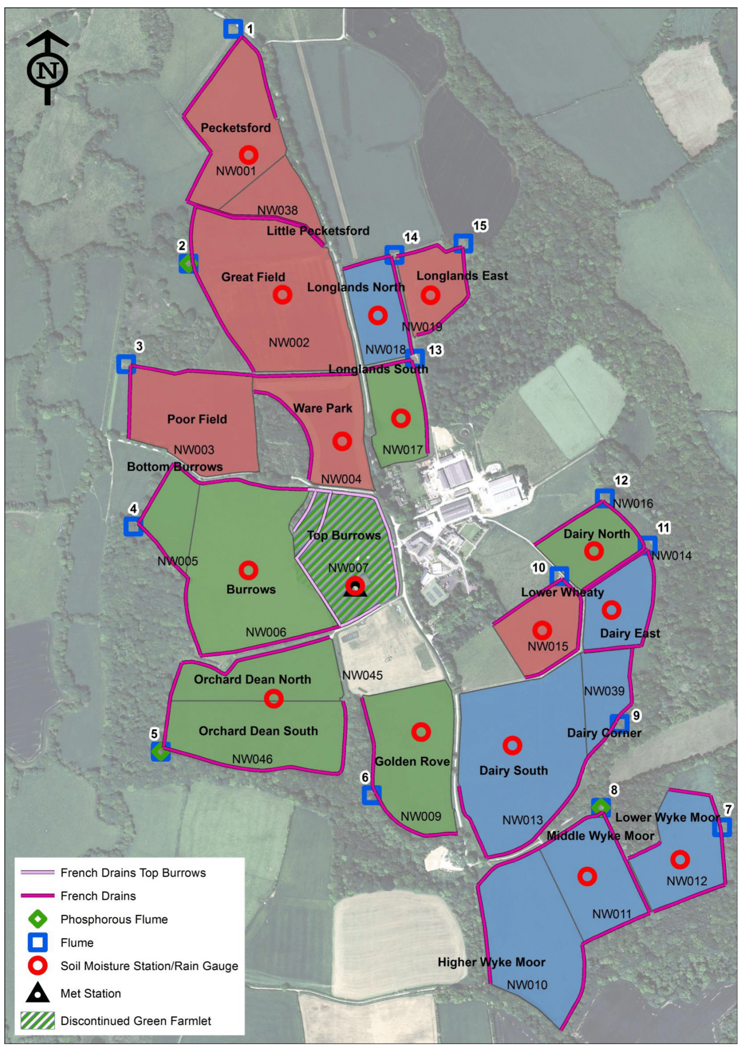

Figure 3: Layout of the North Wyke Farms

A.3 Causal discovery approaches: PC, GES and GIES

The PC algorithm relies on conditional independence testing to establish causal relations. An undirected graph is used as an initial skeleton and edges between independent variables are eliminated. Conditional independence of variables conditioned on a set S, $D_i \perp\!\!\!\perp D_j | S$, is evaluated to eliminate additional edges [46]. The score-based methods evaluate the fitness of causal graphs based on a scoring function to obtain an optimal graph [27]. In the case of GES, the Bayesian Information Criterion (BIC) is used as the scoring function. Starting with an empty graph, edges are added if they improve (lower) the score. The graph is then mapped to a corresponding Markovian equivalence class followed by elimination of edges that may provide further improvement, assessed using the BIC [31]. Similar to GES, GIES also utilizes a quasi-Bayesian score and searches for the causal graph that optimizes for the score. GIES is a generalization of GES that incorporates interventional data. Apart from adding edges (forward phase) and removing edges (backward phase) that improve score, GIES introduces a "turning phase" to improve estimation wherein at each iteration, an edge is turned to obtain an essential graph with same number of edges.

A.4 Details of GNN approaches

For ECC MPNN, at each layer l of the feed-forward neural network, the embedding signal can be computed as,

$$\mathbf{h}_i^l \leftarrow \frac{1}{|N(i)|} \sum_{j \in N(i)} F^l(E_{j,i}; W^l) h_j^{l-1} + b^l \tag{1}$$

where, W^l and b^l are the weight matrix and the bias term defined at layer l. In GraphSAGE, embeddings at search depth k for given node i can be computer as,

$$\mathbf{h}_i^k \leftarrow \sigma(W^k[\mathbf{h}_i^{k-1}, AGG(\{\mathbf{h}_u^{k-1}, \forall u \in N(i)\})]) \tag{2}$$

where, σ is a non-linear activation function and W^k is the weight matrix at depth k. $\mathbf{h}_{N(i)}^k = AGG(\{\mathbf{h}_u^{k-1}, \forall u \in N(i)\})$ is the signal aggregated over all sub-sampled neighbors at depth k. AGG can be any aggregator function including trainable neural network aggregator. Architectures for both methods include paradigm based convolution layer followed by linear layers and then ReLU activation. Figure 4 shows the neighborhood definition for the two GNN approaches considered.

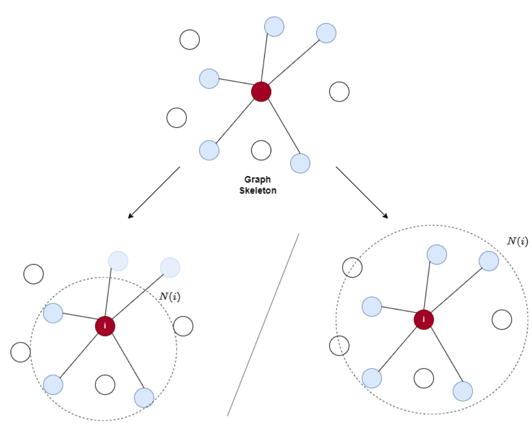

Figure 4: Neighborhood definition in Causal Graph Neural Networks, GraphSAGE and ECC MPNN. ECC MPNN defines all connected adjacent nodes as neighborhood while GraphSAGE aggregates information over different depths, starting from depth 0 (node features itself used as neighboring feature vector) till depth K, where, at each depth k, connected nodes as sub-sampled to be included in neighboring feature vector.

A.5 Architectural choices and hyperparameter tuning

In the GraphSAGE architecture, we choose K to be number of connected nodes, mean as the aggregator function, AGG. In ECC MPNN network, since we have only one type of edge feature in the form of directed edge existence, number of edge feature is set to 1. Model hyperparemeters are chosen from grid search - Adam optimization method for both GNNs, learning rate 0.0015 for GraphSAGE and 0.0020 for EC MPNN. The GraphSAGE architecture uses three sequential GraphSAGE convolution layers to learn embeddings. The architecture also comprises three feed-forward layers, used to generate estimates for the target variable from the embeddings. For the ECC MPNN model, two ECC convolution layers are stacked sequentially to estimate the embeddings. A final feed-forward layer is then used to learn the target variable estimates. We combine three causal discovery methods, PC, GES and GIES, with the two GNN architectures to obtain 6 variants of Causal GNNs. We compare these with four benchmarks, XGBoost (100 estimators, 20 max depth), Random Forest (100 estimators), MLP (random grid search based hyperparameter set) and Random Edges + SageGRAPH (50 random directed edges used as skeleton).

A.6 Causal graphs

In this section, we present the causal graphs produced by the three algorithms: PC algorithm [4] Greedy Equivalence Search (GES) [6] and Greedy Interventional Equivalence Search (GIES) [11]. The nodes of the graph represent the features considered that include one hot encoded fields (nodes named as Field_*field_name*), one hot encoded management practices (Field_Operation_*operation*), total Nitrogen (total-N), total carbon (total-C) and soil pH (pH). The existence of an edge represents a causal relationship and the direction of the arrow represents the direction of influence.

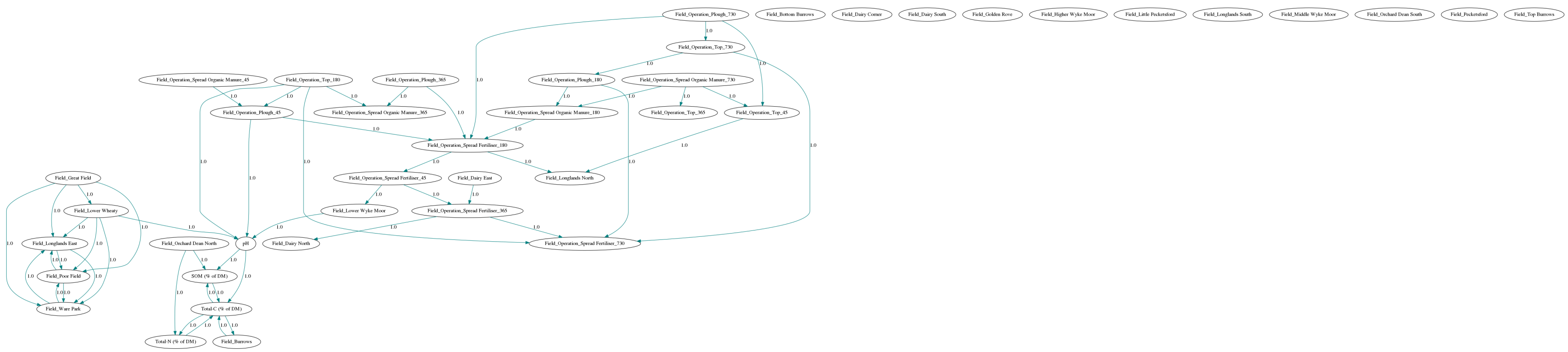

Figure 5: Causal graph generated by PC algorithm

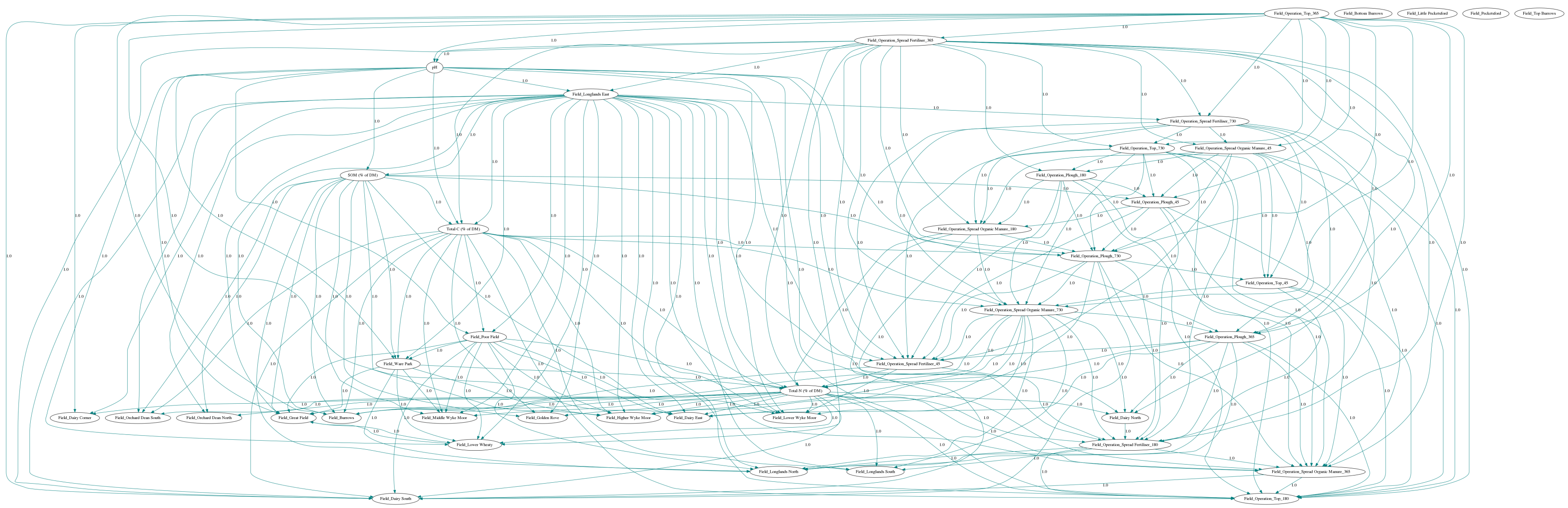

Figure 6: Causal graph generated by GES algorithm

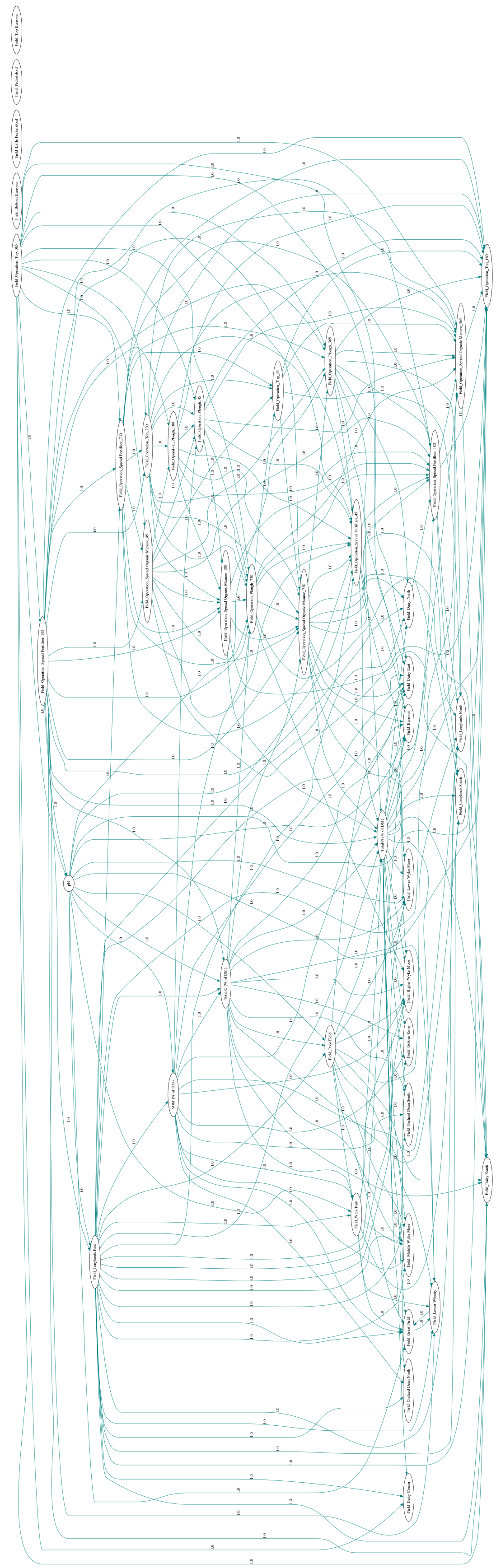

Figure 7: Causal graph generated by GIES algorithm