Evaluating Digital Tools for Sustainable Agriculture using Causal Inference

Ilias Tsoumas[*1,2] **Georgios Giannarakis**[*1] **Vasileios Sitokonstantinou**[1]
Alkiviadis Koukos[1] **Dimitra Loka**[3] **Nikolaos Bartsotas**[1]
Charalampos Kontoes[1] **Ioannis Athanasiadis**[2]
[1]BEYOND Centre, IAASARS, National Observatory of Athens
[2]Wageningen University and Research
[3]Hellenic Agricultural Organization ELGO DIMITRA
{i.tsoumas,giannarakis,vsito,akoukos,nbartsotas,kontoes}@noa.gr
{ilias.tsoumas, ioannis.athanasiadis}@wur.nl
dimitra.loka@elgo.gr

Abstract

In contrast to the rapid digitalization of several industries, agriculture suffers from low adoption of climate-smart farming tools. Even though AI-driven digital agriculture can offer high-performing predictive functionalities, it lacks tangible quantitative evidence on its benefits to the farmers. Field experiments can derive such evidence, but are often costly and time consuming. To this end, we propose an observational causal inference framework for the empirical evaluation of the impact of digital tools on target farm performance indicators. This way, we can increase farmers' trust by enhancing the transparency of the digital agriculture market, and in turn accelerate the adoption of technologies that aim to increase productivity and secure a sustainable and resilient agriculture against a changing climate. As a case study, we perform an empirical evaluation of a recommendation system for optimal cotton sowing, which was used by a farmers' cooperative during the growing season of 2021. We leverage agricultural knowledge to develop a causal graph of the farm system, we use the back-door criterion to identify the impact of recommendations on the yield and subsequently estimate it using several methods on observational data. The results show that a field sown according to our recommendations enjoyed a significant increase in yield (12% to 17%).

1 Introduction

The increasing global population and the changing climate are putting pressure on the agricultural sector, demanding the sustainable production of adequate quantities of nutritious food, feed and fiber. In this context, we need climate-smart agriculture [37, 22] to optimize crop management with zero waste, enhance resilience, increase production and reduce emissions [41]. Unfortunately, the agricultural sector experiences limited adoption of pertinent smart farming technologies [26] that could drive the required sustainable production. This might seem odd at first sight, given the recent surge of sophisticated digital tools that utilize Artificial Intelligence (AI) and big Earth data [53]; yet farmers are skeptical about their effectiveness as most lack quantitative evidence on their benefits [39, 36]. Traditionally, quantifying the impact of a service would require the design and execution of a randomized experiment [9]. Nevertheless, field experiments for the evaluation of digital agriculture tools are seldom done since they are inflexible, requiring follow-up experiments for any changes in the product, but also costly and time-consuming [57]. Thus, an observational causal inference

[*]Equal contribution.

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

framework [46] can fill this gap by emulating the experiment we would have liked to run [30]. Causal inference with observational data has been the subject of recent work across diverse disciplines, including ecology [3], public policy [23, 24], and Earth sciences [42, 47, 50]. In agriculture, it has been used to identify and estimate the effect of agricultural practices on various agro-environmental metrics [48, 18, 27]. According to Adelman (1992), the comprehensive evaluation of decision support systems has three facets: i) the subjective , ii) the technical and iii) the empirical evaluation [1]. While subjective and technical evaluation have been sufficiently practiced [61, 51], the empirical evaluation methods, and in particular with regards to the impact assessment of digital agriculture tools, have been seldom employed. Thus, we propose a framework for the empirical evaluation of digital agriculture recommendations with causal inference. In this context, we evaluate a recommendation system for the optimal sowing of cotton, given sowing time is of great importance for arable crops. Mistimed sowing can lead to suboptimal plant emergence and adversely affect the crop yield [32, 10, 7, 44, 49].

To the best of our knowledge, there are no works that evaluate the effectiveness of any type of decision support or recommendation system in the agricultural sector through causal reasoning and beyond their predictive accuracy [40, 45]. The contributions of this work are summarized as follows: i) the design of the first empirical evaluation framework for digital agriculture based on causal inference; ii) the implementation of it to assess the impact of a recommendation system, which was operationally used in a real-world case study; iii) the identification of the causal effect of sowing recommendations on yield, its subsequent estimation, and the evaluation of estimates using refutation tests.

2 Case Study

In this work, we implement the empirical evaluation of a knowledge-based recommendation system [2] for optimal cotton sowing, which aims to make farmers' production, and hence their profit, resilient against climate change. The recommendations are based on satisfying specific environmental conditions, as retrieved from the related literature, which would ensure successful cotton planting. The system is operationally deployed using high resolution weather forecasts. A.2 of the Appendix contains the design, implementation, algorithmic presentation and the technical evaluation of the system. We provided the recommendations in the form of daily maps, indicating unfavorable and favorable conditions, over the fields of the participating farmers. The sowing recommendation maps were served through the website of their cooperative, which farmers visited on a daily basis during the growing season of 2021. The cooperative collected and provided the required data for each field (i.e., geo-referenced boundaries, sowing & harvest date, seed variety, yield). We then combined this data with publicly available observations from heterogeneous sources (i.e., Sentinel-2 images, climate variables, soil maps) to engineer an observational dataset that enables the causal analysis.

3 Causal Evaluation Framework

Notation & Terminology. We encode the farm system in the form of a Directed Acyclic Graph (DAG) $G \equiv (V, E)$ where V is a set of vertices consisting of all relevant variables, and E is a set of directed edges connecting them [46]. The directed edge $A \to B$ indicates causation from A to B, in the sense that changing the value of A and holding everything else constant will change the value of B. We are using Pearl's do-operator to describe interventions, with $\mathbb{P}(Y = y|do(T = t))$ denoting the probability that $Y = y$ given that we intervene on the system by setting the value of T to t. We name the variable T, of which we aim to estimate the effect, as *treatment* and the variable Y, which we want to quantify the impact of T on, as *outcome*. The parents of a node are its *direct causes*, while a parent of both the treatment and outcome is referred to as a *common cause* or *confounder*. Our end goal is to account for exactly the variables $Z \subseteq V$ that will allow us to estimate the Average Treatment Effect (ATE) of the treatment on outcome, as shown in Eq. (1).

$$\text{ATE} = \mathbb{E}\big[Y|do(T = 1)\big] - \mathbb{E}\big[Y|do(T = 0)\big] \tag{1}$$

Problem Formulation & Causal Graph. We thus aim to develop a causal graph G whose vertices V capture the relevant actors of the system we study, and edges E indicate their relationships. The system recommendations should be part of the graph, along with cotton yield and the agro-environmental conditions that interfere in this physical process. Because the end goal is the evaluation of the recommendation system and its actual impact on yield, we designate as *treated* the fields that farmers sowed on a day that was seen as favorable by the system, and as *control* the fields that were

sown on a non-favorable day. We define a day as favorable when all environmental conditions are satisfied. Binarizing the treatment in that way allows for greater flexibility in estimator selection and easier interpretation. Beyond the recommendation system, multiple factors influence the decision to sow or not. This is precisely the challenge we aim to address by employing a graphical analysis and explicitly modeling the farm system structure. The ATE we aim to estimate captures the difference between what the average yield would have been if we intervened and forced farmers to follow the recommendation by sowing on a favorable day, and the average yield if we forced them to defy the recommendation by sowing on an unfavorable day. Given that confounding factors are controlled for, we henceforth refer to the ATE as the *(average) causal effect of following the recommendation* in the sense described above. Figure 1 displays the final causal graph G. We note that, in reality, it is impossible to account for all factors interacting in the system in order to claim that the estimated effect will not contain any bias. However, because the selection of variables is deeply rooted on well-understood agro-environmental interactions (detailed analysis of graph building in A.2 of Appendix), bias is expected to be minimized, in the sense that no important interactions are left unaccounted for. Furthermore, we extensively test the reliability of effect estimates through multiple refutation checks.

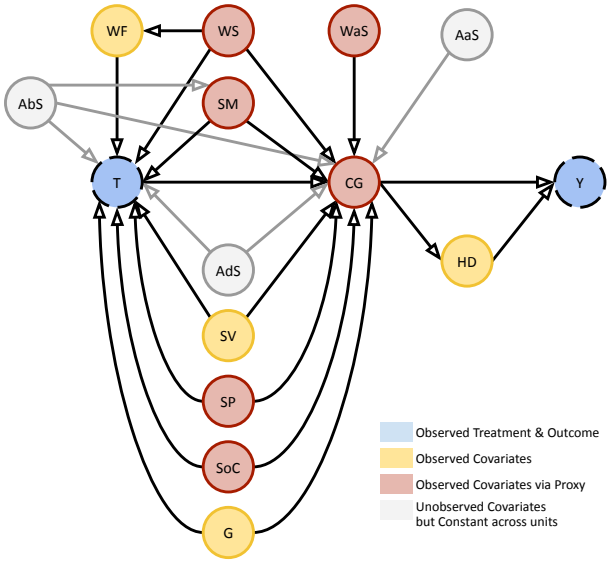

Figure 1: Graph of the farm system.

Id	Variable Description
T	Treatment
WF	Weather forecast
WS	Weather on sowing day
WaS	Weather after sowing
CG	Crop Growth
SM	Soil Moisture on sowing
SP	Topsoil physical properties
SoC	Topsoil organic carbon
SV	Seed Variety
G	Geometry of field
AdS	Practices during sowing
AbS	Practices before sowing
AaS	Practices after sowing
HD	Harvest Date
Y	Outcome (Yield)

Table 1: Variables identifier and description.

Identify, Estimate Effect & Refute Estimate. Because the calculation of causal effects requires access to counterfactual values that are by definition not observed [31], observational methods rely on identification techniques and assumptions that aim at reducing causal estimands such as $\mathbb{P}(Y = y|do(T = t))$ to statistical ones, such as $\mathbb{P}(Y = y|T = t)$. The back-door criterion is a popular identification method that solely relies on a graphical test to infer whether adjusting for a set of graph nodes $Z \subseteq V$ is sufficient for identifying $\mathbb{P}(Y = y|do(T = t))$ from observational data. After (if) we have obtained an adjustment set of variables Z satisfying the back-door criterion we can identify the causal effect of T on Y as $\mathbb{P}(y|do(t)) = \sum_z \mathbb{P}(y|t, z)\mathbb{P}(z)$.

In our study, ATE estimation is done with several methods of varying complexity. Linear regression and distance matching are selected as baseline estimation methods. The popular Inverse Propensity Score (IPS) weigthing is also used [56]. We finally apply modern machine learning methods, i.e., the baseline T-learner and the state-of-the-art X-learner [35].

Given the fact that ground truth estimates are not observed, we resort to performing robustness checks and sensitivity analyses of estimates, in line with recent research [52, 15]. We perform the following tests: i) Placebo treatment, where the treatment is randomly permuted and the estimated effect is expected to drop to 0; ii) Random Common Cause (RCC), where a random confounder is added to the dataset and the estimate is expected to remain unchanged; iii) Random Subset Removal (RSR), where a subset of data is randomly selected and removed and the effect is expected to remain the

same; iv) Unobserved Common Cause (UCC), where an unobserved confounder acts on the treatment and outcome without being added to the dataset, and the estimates should remain relatively stable.

4 Experiments and Results

Causal Effect Estimation				Refutations						
				Placebo		RCC		UCC	RRS	
Method	**ATE**	**CI**	**p-value**	**Effect***	**p-value**	**Effect***	**p-value**	**Effect***	**Effect***	**p-value**
Linear Regression	546	(211, 880)	0.0015	-25.74	0.39	546	0.49	85	543	0.45
Matching	448	(186, 760)	0.0060	50.82	0.39	432	0.40	116	438	0.48
IPS weighting	471	(138, 816)	0.0010	38.82	0.40	470	0.40	113	462	0.45
T-Learner (RF)	372	(215, 528)	0.0240	9.26	0.49	373	0.46	-	353	0.42
X-Learner (RF)	437	(300, 574)	0.0050	5.10	0.50	430	0.37	-	409	0.36

Table 2: ATE point estimates, 95% confidence intervals and p-values. Refutation tests fail if their p-value is less than 0.05. Numbers are in cotton kg/ha.

The sowing period lasted from early April to early May, the harvest took place in September, and yields ranged from $1,250$ to $6,960$ kg/ha. The dataset consists of 171 fields (51 treated and 120 control). Applying the back-door criterion on graph G (Figure 1), the following adjustment set of nodes $Z = \{\text{WS}_{\text{MIN, MAX}}, \text{SOC}, \text{SM}, \text{G}, \text{SP}_{\text{SILT, CLAY, SAND}}, \text{ABS}, \text{ADS}, \text{SV}_{\text{1-13}}\}$ was found sufficient for identifying the ATE. Variables in Z are numerical, including the one-hot encoded vectors of the categorical $\text{SV}_{\text{1-13}}$ variable of variety. AbS and AdS are constant and thus excluded from estimation methods.

Table 2 show the results of the ATE estimation per method, alongside 95% confidence intervals and p-values. Besides Linear Regression, confidence intervals and the resulting p-values are bootstrapped. Both the T-learner and X-learner use a Random Forest for modeling the outcome Y. All methods detect a significant ATE at 95% confidence level, with point estimates ranging from 372 to 546 kilograms of cotton per hectare. For context, the average observed yield is $3,145$ kg/ha. We thus infer that the causal effect of following the sowing recommendation on yield is significantly positive, driving a yield increase ranging from 12% to 17%. Furthermore, Table 2 illustrates that estimation methods are robust against refutation tests. Specifically, Placebo ATE estimates do not differ significantly from 0, while RCC and RSR estimates do not differ significantly from the already obtained ATE. For the UCC test, the mean ATE estimates are reduced yet remain positive, despite unobserved confounding of significant magnitude (more details in part A.3 of the Appendix).

The results indicate that the recommendation system's advice drove a net increase in yield that was both statistically significant and robust. Therefore, farmers are equiped with a provably valuable tool that optimizes the chances of a successful growing season with higher production, and lowers the likelihood of resorting to expensive actions and wasting resources, e.g., replanting the field.

5 Conclusion

In this study, we design, implement, and test a digital agriculture recommendation system for the optimal sowing of cotton. Using the collected data and leveraging domain knowledge, we evaluate the impact of system recommendations on yield. To do so, we utilize and propose causal inference as an ideal tool for empirically evaluating decision support systems. This idea can be upscaled to other digital agriculture tools as well as to different fields with well-established domain knowledge. This paradigm is in principle different to decision support systems that frequently use black-box algorithms to predict variables of interest, but are oblivious to the evaluation of their own impact. In that sense, this work comes to empower the farmer towards resilient agriculture, by introducing an AI framework for elaborating on the assumptions, reliability, and impact of a system that promises green and climate-smart advice.

Acknowledgments and Disclosure of Funding

This work has been supported by the EU Horizon 2020 Research and Innovation program through the following projects: EIFFEL (grant agreement No. 101003518), CALLISTO (grant agreement No.

101004152), e-shape (grant agreement No. 820852), EXCELSIOR (grant agreement No. 857510). It was also supported by the MICROSERVICES project (2019-2020 BiodivERsA joint call, under the BiodivClim ERA-Net COFUND programme, and with the GSRI, Greece - No. T12ERA5-00075).

References

[1] L. Adelman. *Evaluating decision support and expert systems*. Wiley-Interscience, 1992.

[2] C. C. Aggarwal. Knowledge-based recommender systems. In *Recommender systems*, pages 167–197. Springer, 2016.

[3] S. Arif and M. A. MacNeil. Utilizing causal diagrams across quasi-experimental approaches. *Ecosphere*, 13(4):e4009, 2022.

[4] C. Ballabio, P. Panagos, and L. Monatanarella. Mapping topsoil physical properties at european scale using the lucas database. *Geoderma*, 261:110–123, 2016.

[5] M. Bange and S. Milroy. Impact of short-term exposure to cold night temperatures on early development of cotton (gossypium hirsutum l.). *Australian journal of agricultural research*, 55 (6):655–664, 2004.

[6] M. P. Bange, S. J. Caton, and S. P. Milroy. Managing yields of high fruit retention in transgenic cotton (gossypium hirsutum l.) using sowing date. *Australian Journal of Agricultural Research*, 59(8):733–741, 2008.

[7] P. J. Bauer, O. L. May, and J. J. Camberato. Planting date and potassium fertility effects on cotton yield and fiber properties. *Journal of production agriculture*, 11(4):415–420, 1998.

[8] R. Boman and R. Lemon. Soil temperatures for cotton planting. *Texas Coop. Ext. Bull. SCS-2005–17*, 2005.

[9] R. F. Boruch. *Randomized experiments for planning and evaluation: A practical guide*, volume 44. Sage, 1997.

[10] J. M. Bradow and P. J. Bauer. Germination and seedling development. In *Physiology of cotton*, pages 48–56. Springer, 2010.

[11] B. J. Calder, L. W. Phillips, and A. M. Tybout. The concept of external validity. *Journal of consumer research*, 9(3):240–244, 1982.

[12] M. Casamitjana, M. C. Torres-Madroñero, J. Bernal-Riobo, and D. Varga. Soil moisture analysis by means of multispectral images according to land use and spatial resolution on andosols in the colombian andes. *Applied Sciences*, 10(16):5540, 2020.

[13] H. Chen, T. Harinen, J.-Y. Lee, M. Yung, and Z. Zhao. Causalml: Python package for causal machine learning. *arXiv preprint arXiv:2002.11631*, 2020.

[14] M. Christiansen and R. Thomas. Season-long effects of chilling treatments applied to germinating cottonseed. *Crop Science*, 9(5):672–673, 1969.

[15] C. Cinelli and C. Hazlett. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):39–67, 2020.

[16] N. Dalezios, C. Domenikiotis, A. Loukas, S. Tzortzios, and C. Kalaitzidis. Cotton yield estimation based on noaa/avhrr produced ndvi. *Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere*, 26(3):247–251, 2001.

[17] D. de Brogniez, C. Ballabio, A. Stevens, R. Jones, L. Montanarella, and B. van Wesemael. A map of the topsoil organic carbon content of europe generated by a generalized additive model. *European Journal of Soil Science*, 66(1):121–134, 2015.

[18] J. M. Deines, S. Wang, and D. B. Lobell. Satellites reveal a small positive yield effect from conservation tillage across the us corn belt. *Environmental Research Letters*, 14(12):124038, 2019.

[19] T. J. DiCiccio and B. Efron. Bootstrap confidence intervals. *Statistical science*, 11(3):189–228, 1996.

[20] H. Dong, W. Li, W. Tang, Z. Li, D. Zhang, and Y. Niu. Yield, quality and leaf senescence of cotton grown at varying planting dates and plant densities in the yellow river valley of china. *Field Crops Research*, 98(2-3):106–115, 2006.

[21] L. Eklundh and P. Jönsson. Timesat: A software package for time-series processing and assessment of vegetation dynamics. In *Remote sensing time series*, pages 141–158. Springer, 2015.

[22] Food and A. O. of the United Nations. *FAO's Strategic Framework 2022-31*. 2022. URL `https://www.fao.org/3/cb7099en/cb7099en.pdf`.

[23] D. Fougère and N. Jacquemet. Causal inference and impact evaluation. *Economie et Statistique/Economics and Statistics*, (510-511-512):181–200, 2019.

[24] D. Fougère and N. Jacquemet. Policy evaluation using causal inference methods. In *Handbook of Research Methods and Applications in Empirical Microeconomics*, pages 294–324. Edward Elgar Publishing, 2021.

[25] T. B. Freeland Jr, B. Pettigrew, P. Thaxton, and G. L. Andrews. Agrometeorology and cotton production. *World Meteorological Organization*, 2006.

[26] A. Gabriel and M. Gandorfer. Adoption of digital technologies in agriculture—an inventory in a european small-scale farming region. *Precision Agriculture*, pages 1–24, 2022.

[27] G. Giannarakis, V. Sitokonstantinou, R. S. Lorilla, and C. Kontoes. Towards assessing agricultural land suitability with causal machine learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1442–1452, 2022.

[28] J. M. Green. Border effects in cotton variety tests 1. *Agronomy Journal*, 48(3):116–118, 1956.

[29] J. L. Hatfield and J. H. Prueger. Temperature extremes: Effect on plant growth and development. *Weather and climate extremes*, 10:4–10, 2015.

[30] M. A. Hernán and J. M. Robins. Using big data to emulate a target trial when a randomized trial is not available. *American journal of epidemiology*, 183(8):758–764, 2016.

[31] P. W. Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

[32] J. Huang. Different sowing dates affected cotton yield and yield components. *International Journal of Plant Production*, 10(1):63–83, 2016.

[33] N. Huntington-Klein. *The effect: An introduction to research design and causality*. Chapman and Hall/CRC, 2021.

[34] G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[35] S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, Mar. 2019. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1804597116. URL `http://www.pnas.org/lookup/doi/10.1073/pnas.1804597116`.

[36] E. D. Lioutas, C. Charatsari, and M. De Rosa. Digitalization of agriculture: a way to solve the food problem or a trolley dilemma? *Technology in Society*, 67:101744, 2021.

[37] L. Lipper, P. Thornton, B. M. Campbell, T. Baedeker, A. Braimoh, M. Bwalya, P. Caron, A. Cattaneo, D. Garrity, K. Henry, et al. Climate-smart agriculture for food security. *Nature climate change*, 4(12):1068–1072, 2014.

[38] S. Liu, M. Remley, F. Bourland, R. Nichols, W. Stevens, A. P. Jones, and F. Fritschi. Early vigor of advanced breeding lines and modern cotton cultivars. *Crop Science*, 55(4):1729–1740, 2015.

[39] J. Lowenberg-DeBoer and B. Erickson. Setting the record straight on precision agriculture adoption. *Agronomy Journal*, 111(4):1552–1569, 2019.

[40] S. Luma-Osmani, F. Ismaili, B. Raufi, and X. Zenuni. Causal reasoning application in smart farming and ethics: A systematic review. *Annals of Emerging Technologies in Computing (AETiC), Print ISSN*, pages 2516–0281, 2020.

[41] J. Lynch, M. Cain, D. Frame, and R. Pierrehumbert. Agriculture's contribution to climate change and role in mitigation is distinct from predominantly fossil co2-emitting sectors. *Frontiers in sustainable food systems*, page 300, 2021.

[42] A. Massmann, P. Gentine, and J. Runge. Causal inference for process understanding in earth sciences. *arXiv preprint arXiv:2105.00912*, 2021.

[43] NCEP, NWS, NOAA and USDOC. Ncep gfs 0.25 degree global forecast grids historical archive, 2015. URL https://doi.org/10.5065/D65D8PWK.

[44] R. L. Nielsen, P. R. Thomison, G. A. Brown, A. L. Halter, J. Wells, and K. L. Wuethrich. Delayed planting effects on flowering and grain maturation of dent corn. *Agronomy Journal*, 94 (3):549–558, 2002.

[45] D. Pasquel, S. Roux, J. Richetti, D. Cammarano, B. Tisseyre, and J. A. Taylor. A review of methods to evaluate crop model performance at multiple and changing spatial scales. *Precision Agriculture*, pages 1–25, 2022.

[46] J. Pearl. *Causality*. Cambridge university press, 2009.

[47] A. Pérez-Suay and G. Camps-Valls. Causal inference in geoscience and remote sensing from observational data. *IEEE Transactions on Geoscience and Remote Sensing*, 57(3):1502–1513, 2018.

[48] S. S. Qian and R. D. Harmel. Applying statistical causal analyses to agricultural conservation: A case study examining p loss impacts. *JAWRA Journal of the American Water Resources Association*, 52(1):198–208, 2016.

[49] M. F. Richards, L. Maphosa, and A. L. Preston. Impact of sowing time on chickpea (cicer arietinum l.) biomass accumulation and yield. *Agronomy*, 12(1):160, 2022.

[50] J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Glymour, M. Kretschmer, M. D. Mahecha, J. Muñoz-Marí, et al. Inferring causation from time series in earth system sciences. *Nature communications*, 10(1):1–13, 2019.

[51] G. Schröder, M. Thiele, and W. Lehner. Setting goals and choosing metrics for recommender system evaluations. In *UCERSTI2 workshop at the 5th ACM conference on recommender systems, Chicago, USA*, volume 23, page 53, 2011.

[52] A. Sharma and E. Kiciman. Dowhy: An end-to-end library for causal inference. *arXiv preprint arXiv:2011.04216*, 2020.

[53] A. Sharma, A. Jain, P. Gupta, and V. Chowdary. Machine learning applications for precision agriculture: A comprehensive review. *IEEE Access*, 9:4843–4873, 2020.

[54] W. C. Skamarock, J. B. Klemp, J. Dudhia, D. O. Gill, Z. Liu, J. Berner, W. Wang, J. G. Powers, M. G. Duda, D. M. Barker, et al. A description of the advanced research wrf model version 4. *National Center for Atmospheric Research: Boulder, CO, USA*, 145:145, 2019.

[55] J. L. Snider, C. Pilon, and G. Virk. Seed characteristics and seedling vigor. In *Cotton Seed and Seedlings*, pages 9–23. The Cotton Foundation, 2020.

[56] E. A. Stuart. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1, 2010.

[57] J. Vaessen. Challenges in impact evaluation of development interventions: opportunities and limitations for randomized experiments. IOB Discussion Papers 2010.01, Universiteit Antwerpen, Institute of Development Policy (IOB), 2010. URL https://EconPapers.repec.org/RePEc:iob:dpaper:2010001.

[58] J. J. Varco. Soil climatic and edaphic effects on cotton germination and the final stand. In *Cotton Seed and Seedlings*, pages 43–53. The Cotton Foundation, 2020.

[59] G. Virk, J. L. Snider, and C. Pilon. Physiological contributors to early season whole-crop vigor in cotton. *Crop Science*, 59(6):2774–2783, 2019.

[60] D. Wanjura, E. Hudspeth Jr, and J. Bilbro Jr. Emergence time, seed quality, and planting depth effects on yield and survival of cotton (gossypium hirsutum l.). *Agronomy Journal*, 61(1):63–65, 1969.

[61] Z. Zhai, J. F. Martínez, V. Beltran, and N. L. Martínez. Decision support systems for agriculture 4.0: Survey and challenges. *Computers and Electronics in Agriculture*, 170:105256, 2020.

[62] D. Zhao, K. R. Reddy, V. G. Kakani, J. J. Read, and S. Koti. Canopy reflectance in cotton for growth assessment and lint yield prediction. *European Journal of Agronomy*, 26(3):335–344, 2007.

A Supplementary Material

A.1 Agricultural Recommendation System

In this work, we design, implement and evaluate a knowledge-based recommendation system [2] for optimal cotton sowing. The recommendations are based on satisfying specific environmental conditions, as retrieved from the related literature, which would ensure successful cotton planting. The system is operationally deployed using high resolution weather forecasts. Sec. 1 of the Appendix contains an algorithmic presentation of the system.

According to literature, the minimum daily-mean soil temperature for cotton germination is $16°$C [10]. Soil or ambient temperatures lower than $10°$C result in less vigorous and malformed seedlings [8]. As a general rule for cotton, agronomists recommend daily-mean soil temperatures higher than $18°$C for at least 10 days after sowing and daily-maximum ambient temperatures higher than $26°$C for at least 5 days after sowing. We summarize the conditions for optimal cotton sowing in Table 3 [25, 8]. Using these conditions and Numerical Weather Predictions (NWP) we implement a recommendation system that advises on whether any given day is a good day to sow or not.

Type of Temperature	**Statistic**	**Condition**	**Condition Priority**
soil (0-10 cm)	mean	>18°C	optimum
ambient (2 m)	max	>26°C	optimum
soil (0-10 cm)	mean	>15.56°C	mandatory
soil (0-10 cm)	min	>10°C	mandatory
ambient (2 m)	min	>10°C	mandatory

Table 3: Optimal conditions for sowing cotton. All conditions refer to the period from sowing day to 5 days after, except the first soil condition that refers to 10 days after.

Open-access high-resolution NWP forecasts are rarely available. For this reason, we implement the WRF-ARW model [54] with a grid resolution of 2 km. This enables us to reach a high spatio-temporal resolution for parameters that are crucial during the cotton seeding period, namely the soil and ambient temperature that are retrieved in hourly rate for the forthcoming 2.5 days. Ideally, 10-day predictions at a 2 km spatial resolution should be available every morning, as it is required by the conditions in Table 3. However, this would demand an enormous amount of computational power. To simulate the desired data, we combine the 2.5-day high resolution forecasts with the GFS [43] 15-day forecasts that are given on a 0.25 degrees (roughly 25 km) spatial resolution.

$$a_i = \frac{GFS_{day=i}}{GFS_{day=1}}, i \in \{3, ..., 10\} \tag{2}$$

$$ART_j = \begin{cases} WRF_{day=j} & , j \in \{1, 2\} \\ WRF_{day=1} \cdot a_j & , j \in \{3, ..., 10\} \end{cases} \tag{3}$$

Eq. (2) shows how we extract the 10-day weather trend factor using GFS forecasts. We calculate the percentage change between each forecast (for $day = 3$ to $day = 10$) and the corresponding next day ($day = 1$) forecast. Eq. (3) shows how we produce the artificial (ART) 10-day forecasts at 2 km spatial resolution. We keep the original WRF forecasts for the next two days and for the rest we apply the respective 10-day trend factor to the next day WRF forecast.

We generate ART forecasts in order to provide recommendations that can vary up to the field-level, which would have been impossible with GFS forecasts alone. This is depicted in Figure 2. In order to evaluate the quality of our ART forecasts, we compared them with measurements from the nearest operational weather station in the area of interest for the critical sowing period, from 15/4/2021 to 15/5/2021. We have limited our comparison to the maximum and minimum ambient temperatures, as there were not any soil temperature measurements available. It is worth noting that the nearest grid point of GFS to the station is only 0.87 km away, however the maximum distance can be up to 12 km away. On the other hand, the equivalent grid point of ART is 1.41 km away, which incidentally is the maximum possible distance between any location and the nearest ART point.

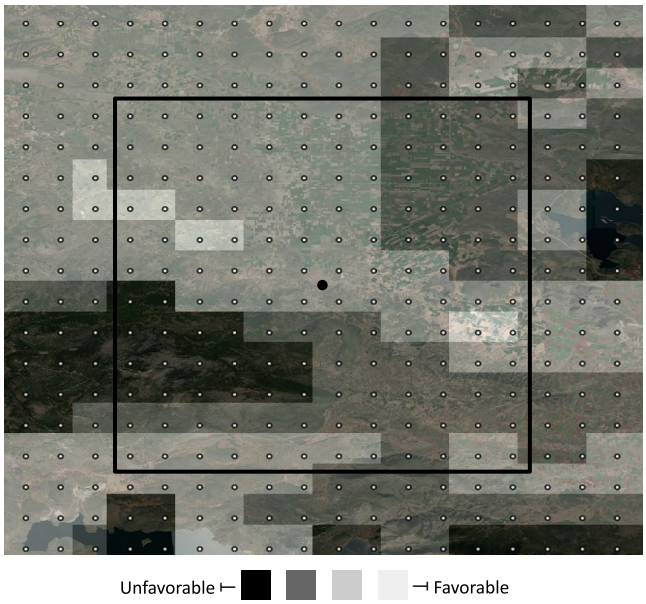

Figure 2: Optimal sowing map for a given day. The black circle at the center depicts the GFS grid point that represents the entire black-lined box. The white circles depict the 144 ART grid points for the same area.

Initially, we compared the next day forecasts of GFS against their ART (or WRF) equivalent. The comparison analysis revealed a Mean Absolute Error (MAE), between the two forecasts and the station for maximum ambient temperature, equal to $2.39°$C (GFS) versus $1.48°$C (ART), and for minimum ambient temperature $1.52°$C (GFS) versus $1.74°$C (ART). Overall, WRF appears to behave well and slightly better than GFS. This difference is expected to be greater for other locations in the grid, as for this particular case the station happened to be very close to the GFS grid point. Furthermore, we calculated the MAE and Root Mean Squared Error (RMSE) of all daily 5-day forecasts of ART against the ground station for a period of interest. For the maximum temperature we found $MAE = 2.41$, $RMSE = 3.11$, whereas for the minimum temperature we found $MAE = 2.75$, $RMSE = 3.70$. A graphical comparisons of ART forecasts against the ground station measurements is presented in the Figure 3)

A.2 Cotton Domain Knowledge and Graph Building

Cotton yield and quality are ultimately determined by the interaction between the genotype, environmental conditions and management practices throughout the growing season. Nevertheless, the first pivotal steps for a profitable yield are a successful seed germination and emergence which are greatly dependent on timely sowing [60, 7, 10].

Emergence and germination mediate the effect of T on Y; however, Crop Growth (CG) was not observed. We thus turned to the popular Normalized Difference Vegetation Index (NDVI) in order to obtain a reliable proxy of CG, and specifically used the trapezoidal rule across NDVI values from sowing to harvest [21]. Even though in the case of cotton, trapezoidal NDVI is not linearly correlated with yield [16, 62], it is correlated with early season Leaf Area Index (LAI) [62], which in turn is a good indicator of early season crop growth rate [59]. Furthermore, seed germination and seedling emergence are greatly dependent on soil moisture. Hence, soil moisture SM is a confounder for the relation $T \rightarrow Y$. As a SM proxy, we used the well-known Normalized Difference Water Index (NDWI) at sowing day which is highly correlated with soil moisture in bare soil [12].

Agricultural management practices before sowing (AbS) comprise tilling operations for preparing a good seedbed. Practices during sowing (AdS) include a sowing depth of $4 - 5$ cm and an average distance of 0.91 m between rows and 7.62 cm between seeds. After sowing practices (AaS) comprise basic fertilization, irrigation and pest management. It is reasonable to think that all aforementioned

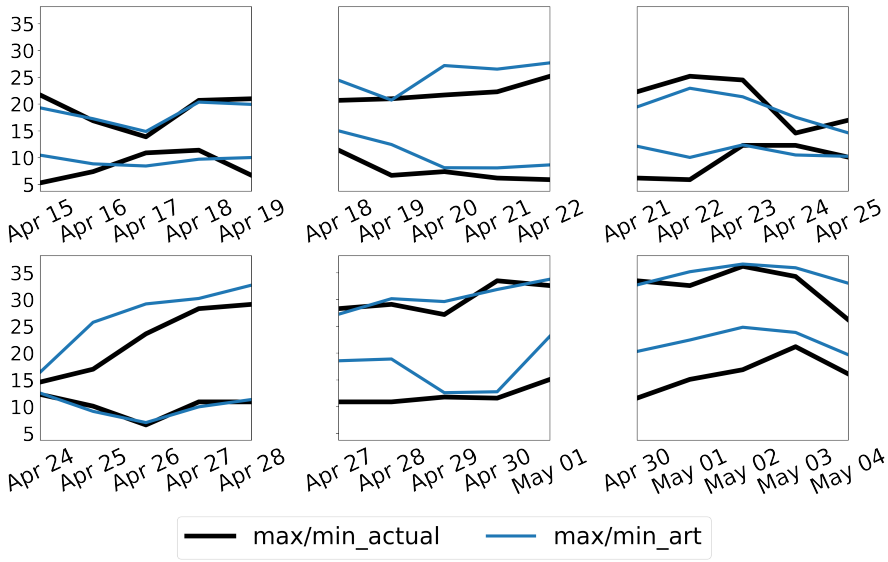

Figure 3: 5-day max/min temperature for ART forecasts and ground station measurements. Each plot shows a different 5-day prediction.

practices are a result of a common cause that we can define as Agricultural Knowledge (AK), capturing the skills and expertise of a farmer. We possess no quantitative information on the agricultural knowledge or the practices followed by each farmer. However, the farmer's cooperative is not large, and aims for consistent, high-quality produce. As a result, they have developed highly consolidated routines for interacting with their crops: this includes common practices, homogeneous fertilizer application, and jointly owned machinery. We thus note that even if we do not have numerical data on AbS, AdS, AaS, the cooperative directors do not observe significant differences across fields and for the purposes of our study these variables are considered to be constant.

At the same time, it is rational to assume that the agricultural knowledge (AK) of any farmer interacts with crops exclusively through management practices. Because of the aforementioned condition, the influence of AK on the system is nullified and we hence omit it from the graph. While we note that the above limit the external validity of our results [11], by assuming that agricultural practices are constant for all farmers and that AK only interacts with the system through them, we implicitly control for all of them [33].

Apart from soil moisture, soil and ambient temperatures at the time of sowing and for 5-10 days after, affect seed germination, seedling development and final yield [59, 8, 58]. Low temperatures result in reduced germination, slow growth and less vigorous seedlings that are more prone to diseases and sensitive to weed competition [14, 60, 10]. This knowledge is incorporated in the sowing recommendations, in the form of numerical rules, and consequently in the treatment T. We thus added in the graph the weather forecast WF (variables listed in Table 3) as a parent node of T. We also had access to the weather on the day of sowing WS (min & max ambient temperature in °C) from a nearby weather station, influencing WF, T, and CG.

Topsoil (0-20 cm) properties SP (% content of clay, silt and sand) and organic carbon content SoC (g C kg^{-1}) also affect cotton seed germination and seedling emergence due to differences in water holding capacity and consequently in soil temperature and aeration, drainage and seed-to-soil contact [58]. Data on SP and SoC were retrieved from the European Soil Data Centre (ESDAC) [4, 17]. Both variables were included in the graph as confounders of T and CG. Seed variety also determines seed germination, emergence and final yield [55]. Seed mass and vigor [38, 55] are related to the seed variety (SV); we hence added the latter as a confounder for T and Y. In this case, we had 13 different cotton SVs.

The geometrical properties of the field (perimeter to area ratio, G) were also considered, as border effects can play a minor role on crop growth, confounding the effect of T on Y [28]. Since temperature is the primary environmental factor controlling plant growth [5, 29], temperature fluctuations were

Id	Variable Description	Source
T	Treatment	Recommendation System
WF	Weather forecast	GFS, WRF
WS	Weather on sowing day	Nearest weather station
WaS	Weather after sowing	Nearest weather station
CG	Crop Growth	NDVI via Sentinel-2
SM	Soil Moisture on sowing	NDWI via Sentinel-2
SP	Topsoil physical properties	Map by ESDAC
SoC	Topsoil organic carbon	Map by ESDAC
SV	Seed Variety	Farmers' Cooperative
G	Geometry of field	Farmers' Cooperative
AdS	Practices during sowing	Farmers' Cooperative
AbS	Practices before sowing	Farmers' Cooperative
AaS	Practices after sowing	Farmers' Cooperative
HD	Harvest Date	Farmers' Cooperative
Y	Outcome (Yield)	Farmers' Cooperative

Table 4: Farm system variable identifier, description and source.

observed throughout the growing season from the nearest weather station, constituting a parent variable WaS (min & max ambient temperature in °C) of crop growth CG. Lastly, the Harvest Date (HD) mediates the effect of CG on Y, influencing both yield potential and quality [20, 6]. Table 4 summarizes the variables' description, abbreviation and source.

A.3 Implementation & Results Details

For the experiments, we are using the popular doWhy [52] and Causal ML [13] Python libraries.

Propensity modeling is a prerequisite of IPS weighting. We thus begin by discussing the propensity model that is fit. Given the relatively small dataset size, logistic regression is used on the scaled back-door adjustment set Z for classifying each field into the treatment/control group. We subsequently trim the dataset by removing all rows with extreme propensity scores (< 0.2 or > 0.8) to aid the overlap assumption [34]. The resulting distribution of propensity scores can be seen at Figure 4. The model scores 0.81 in accuracy, 0.64 in F1-score, and 0.88 in ROC-AUC. After trimming extreme propensity scores, a subset of 48 treated and 37 control units remains. There is decent overlap between the propensity score distributions of the treatment and control group, indicating that they are comparable and enabling reliable propensity-based ATE estimation.

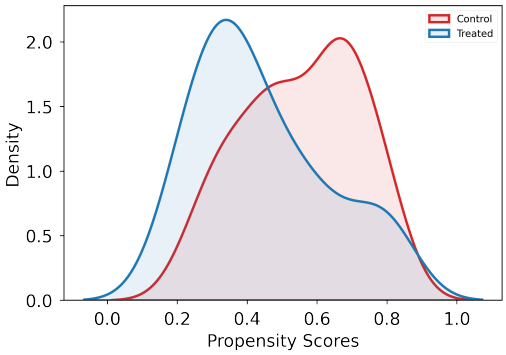

Figure 4: Distribution of propensity scores for the control and treatment group after trimming extreme scores.

Besides Linear Regression, other methods do not provide confidence intervals by default. For matching, IPS, and meta-learners confidence intervals and the resulting p-values are hence bootstrapped. Also, the Placebo, RCC and RSR refutations tests are bootstrapped to generate confidence intervals and p-values [19]. Confidence intervals and p-values are bootstrapped (1000 iterations).

The UCC refutation test returns a heatmap of new ATE estimates depending on the strength of injected unobserved confounding. Figure 5 contains heatmaps with the impact of unobserved confounding on the ATE estimate for the linear regression, matching, and IPS weighting methods which were deployed through doWhy. Observing the heatmaps, we note that the estimation methods are robust to a moderate amount of unobserved confounding, in the sense that the ATE values of the lower region of each heatmap (where the effect of the unobserved confounder does not dominate the treatment and outcome values) largely remains positive and comparable to the real ATE estimate. We note that as the strength of unobserved confounding increases, significant volatility in effect estimates is expected, as the effect is no longer fully identified. For more information about the implementation and the proper interpretation of test, see the doWhy library documentation [52].

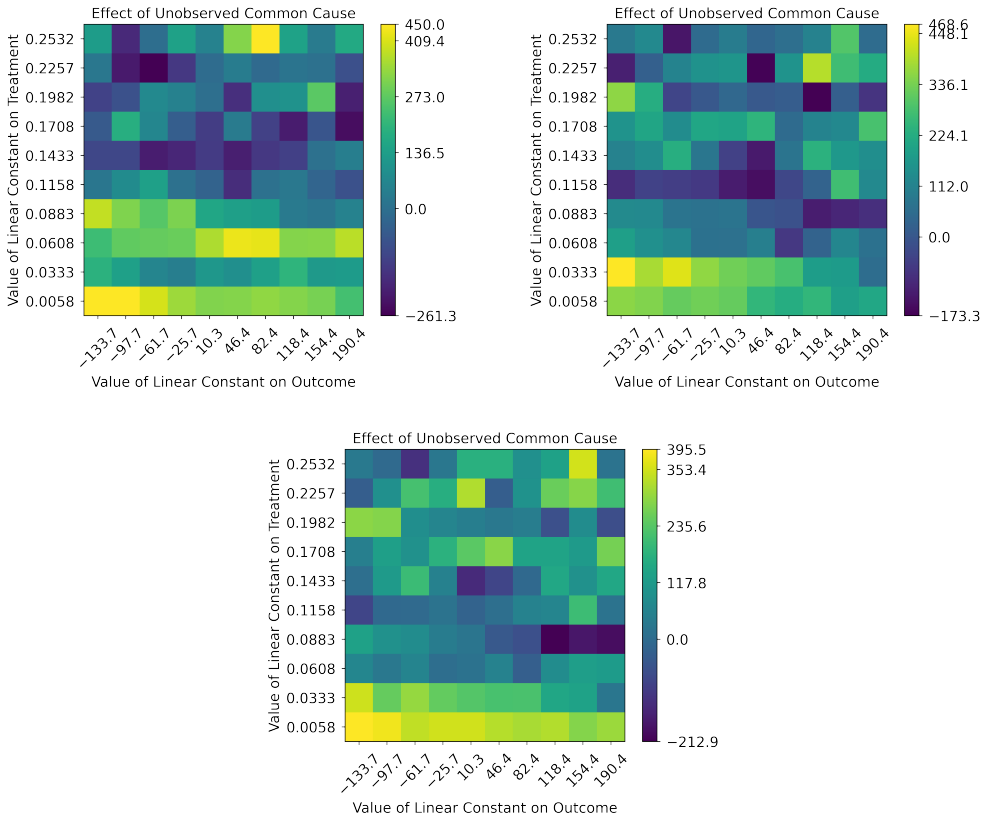

Figure 5: Unobserved Common Cause heatmap results for Linear Regression (Top Left), Matching (Top Right) and IPS weighting (Bottom) In the main paper, we report the average cell value of each heatmap (i.e, the average ATE across multiple combinations of unobserved confounding.)