Environmental Data Science (2025), 4: e23, 1–14
doi:10.1017/eds.2025.14

POSITION PAPER

Leveraging causality and explainability in digital agriculture

Ilias Tsoumas[1,2], Vasileios Sitokonstantinou[3], Georgios Giannarakis[1], Evagelia Lampiri[4], Christos Athanassiou[4], Gustau Camps-Valls[3], Charalampos Kontoes[1] and Ioannis N. Athanasiadis[2]

[1]BEYOND Centre, IAASARS, National Observatory of Athens, Athens, Greece
[2]Artificial Intelligence, Wageningen University & Research, Wageningen, The Netherlands
[3]Image Processing Laboratory (IPL), Universitat de València, València, Spain
[4]Department of Agriculture, Crop Production and Rural Environment, University of Thessaly, Volos, Greece
Corresponding authors: Ilias Tsoumas and Vasileios Sitokonstantinou; Emails: i.tsoumas@noa.gr; vasileios.sitokonstantinou@uv.es

Received: 08 November 2024; **Revised:** 15 January 2025; **Accepted:** 20 March 2025

Keywords: agriculture; causal inference; evidence-based decision making; pest management; causal machine learning

Abstract

Sustainable agricultural practices have become increasingly important due to growing environmental concerns and the urgent need to mitigate the climate crisis. Digital agriculture, through advanced data analysis frameworks, holds promise for promoting these practices. Pesticides are a common tool in agricultural pest control, which are key in ensuring food security but also significantly contribute to the climate crisis. To combat this, Integrated Pest Management (IPM) stands as a climate-smart alternative. We propose a causal and explainable framework for enhancing digital agriculture, using pest management and its sustainable alternative, IPM, as a key example to highlight the contributions of causality and explainability. Despite its potential, IPM faces low adoption rates due to farmers' skepticism about its effectiveness. To address this challenge, we introduce an advanced data analysis framework tailored to enhance IPM adoption. Our framework provides (i) robust pest population predictions across diverse environments with invariant and causal learning, (ii) explainable pest presence predictions using transparent models, (iii) actionable advice through counterfactual explanations for in-season IPM interventions, (iv) field-specific treatment effect estimations, and (v) assessments of the effectiveness of our advice using causal inference. By incorporating these features, our study illustrates the potential of causality and explainability concepts to enhance digital agriculture regarding promoting climate-smart and sustainable agricultural practices, focusing on the specific case of pest management. In this case, our framework aims to alleviate skepticism and encourage wider adoption of IPM practices among policymakers, agricultural consultants, and farmers.

Impact Statement

We present a new data analysis framework based on causality and explainability to help farmers adopt sustainable alternatives to traditional practices for agricultural management. The framework makes agricultural management more practical and trustworthy by providing clear, reliable predictions, advice tailored to specific fields, and impact assessment of recommended actions. In our example, this could lead to less reliance on harmful pesticides, helping to protect the environment and fight climate change. With this tool, farmers can make better-informed decisions that benefit their crops and the planet, promoting a healthier and more sustainable future.

1. Introduction

Digital agriculture integrates agricultural expertise with digital technologies, such as remote sensing, IoT, and data analytics, to effectively leverage diverse data sources like satellite imagery, weather forecasts, and soil health metrics. This approach promotes more sustainable, resilient, and profitable farming by enabling data-driven decisions across the agricultural value chain (Basso and Antle, 2020). This approach is essential for adapting agriculture to our rapidly changing climate and mitigating its impact on climate change (Balasundram et al., 2023). Artificial Intelligence (AI) serves digital agriculture as the means to transform the data into insights, estimations, forecasts, and recommendations that aim to support decision-making to balance agriculture's environmental, societal, and economic aspects. However, digital agriculture has remained largely confined to using almost solely correlation-based AI, which excels at predictive tasks but cannot go further. In this context, we propose exploiting two underutilized branches of AI by digital agriculture—causality and explainability. They can unlock capabilities beyond the continuous pursuit of prediction accuracy for enhancing digital agriculture, given that it considers agricultural knowledge and practice and integrates it into the modeling and inference parts (Sitokonstantinou et al., 2024). Thus, causality and explainability bring in digital agriculture domain-aware robust models, explainable predictions, counterfactual reasoning, and quantifying effects of advice, action, and policy.

Pest management is a quintessential example in this context, demonstrating the valuable contributions that causality and explainability offer. Conventional pest management has been shown to contribute to climate change. Raising temperatures, intensifying ultraviolet radiation, and reducing relative humidity are expected to increase pest outbreaks and undermine the efficacy of pest control methods like host-plant resistance, bio-pesticides, and synthetic pesticides (Sharma and Prabhakar, 2014; Skendžić et al., 2021). Despite climate experts' warnings, pesticide use in agriculture adversely affects public health (Boedeker et al., 2020) and contributes to the climate crisis. This impact includes: (i) greenhouse gas (GHG) emissions from pesticide production, packaging, and transportation (Audsley et al., 2009), (ii) compromised soil carbon sequestration (Xu et al., 2020), (iii) elevated GHG emissions from soil (Spokas and Wang, 2003; Marty et al., 2010; Heimpel et al., 2013), and (iv) contamination of adjacent soil and water ecosystems, resulting in biodiversity loss (Sharma et al., 2019).

Thus, a vicious cycle has been established between pesticides and climate change (Sharma et al., 2022). In response, the European Commission (EC) has taken action to reduce all chemical and high-risk pesticides by 50% by 2030. Achieving such reductions requires adopting integrated pest management (IPM), which promotes sustainable agriculture and agroecology. IPM consists of eight principles inspired by the Food and Agriculture Organization (FAO) description. The authors in Barzman et al. (2015) condense these principles into *prevention and suppression*, *monitoring*, *decision-making*, *non-chemical methods*, *pesticide selection*, *reduced pesticide use*, *anti-resistance strategies*, and *evaluation*.

Data-driven methods have played a crucial role in optimizing pest management decisions. Some studies employ supervised machine learning techniques, such as Random Forests and Artificial Neural Networks (ANNs), satellite Earth observations, and in-situ data for pest presence prediction (Aparecido et al., 2019; Zhang et al., 2019). Others extend their models to include weather data (Skawsang et al., 2019). Recurrent Neural Networks (RNNs) capture temporal features from weather data, effectively handling unobservable counterfactual outcomes (Xiao et al., 2019). Iost Filho et al. (2022) highlight the extraction of fine-scale information for Integrated Pest Management (IPM) using meteorological data, insect scouting records, machine learning, and remote sensing. Nanushi et al. (2022) propose an interpretable machine learning solution integrating numerical weather predictions, vegetation indices, and trap catch data for estimating *Helicoverpa armigera* presence in cotton fields. This approach enhances the decision-making aspect of IPM, shifting away from traditional threshold-based pesticide applications. The interpretability of these predictions enhances trust and allows for incorporating domain expertise in pest management decision-making.

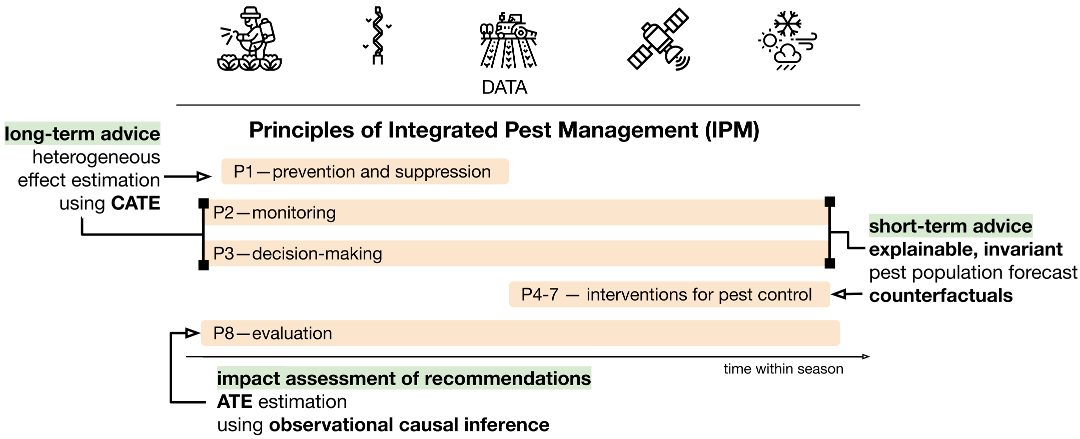

Figure 1. *Causal and explainable data analysis framework for enhanced IPM.*

2. Proposal

As Barzman et al. (2015) point out, threshold-based and "spray/don't spray" advice is not enough. There is a need for a new class of digital tools that consider the entire set of IPM principles to enhance decision-making truly. In this direction, we propose a data analysis framework for IPM based on causality and explainability. It consists of short-term actionable advice for in-season interventions and long-term advice for supporting strategic farm planning (Figure 1).

This way, we will upgrade the *monitoring* and *decision-making* IPM principles leading to actionable advice for direct pest control interventions and assisting the selection of practices relevant to other IPM principles, such as the *use of non-chemical methods* and *reduce pesticide dosage.* Additionally, the proposed framework will better inform farmers concerning the potential impact of practices that, in turn, will enhance the IPM principle of *prevention and suppression*, for example, crop rotation, day of sowing, and no-tillage. Furthermore, our framework employs observational causal inference to continuously assess the recommendations above and satisfy the IPM principle of *evaluation.*

In this study, we exploit the proposed framework, demonstrating its applicability and efficiency in a case study for pest management. While the case study is specific it represents the general case of pest management in several crops and conditions, and the typical availability of data for such case studies.

3. Data

Our approach relies on diverse data sources as a key leverage to capture a comprehensive picture of the past, present, and future agro-environmental conditions. This will enable us to improve the modeling and comprehension of pest dynamics.

3.1. Earth observations

We leverage biophysical and biochemical properties such as Leaf Area Index (LAI), Normalized Difference Vegetation Index (NDVI), chlorophyll content, as well as data on evapotranspiration and soil moisture. These factors play a crucial role in monitoring pest population dynamics. The data is derived from the Sentinel-1/2 and Terra/Aqua (MODIS) satellite missions that provide open access to optical multi-spectral and Synthetic Aperture Radar (SAR) images.

3.2. Terrain & soil characteristics

We incorporate data from open-access digital elevation models and information on topsoil physical properties and soil organic carbon content (de Brogniez et al., 2015; Ballabio et al., 2016). This allows us to include fixed or long-term characteristics specific to the area of interest.

https://doi.org/10.1017/eds.2025.14 Published online by Cambridge University Press

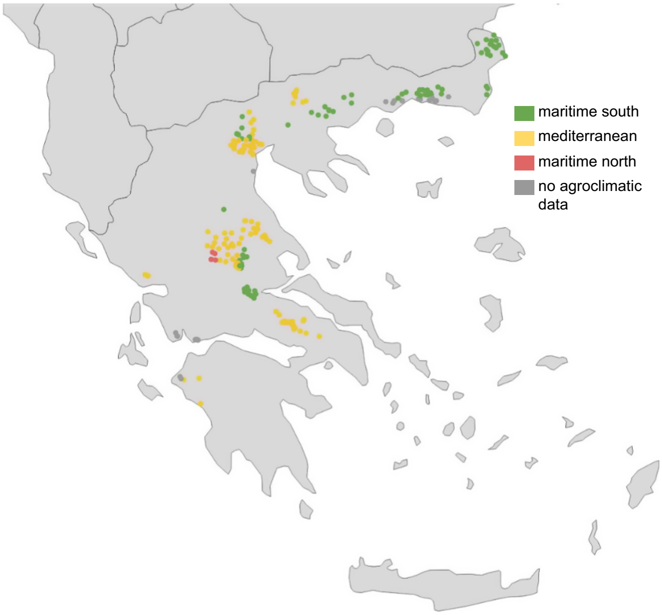

Figure 2. *Traps distribution in the Greek mainland for 2019–2022. Colors indicate the different agroclimatic zones in which traps from the dataset belong. These zones have been identified based on the study conducted by Ceglar et al. (2019).*

3.3. Numerical weather predictions (NWP) and reanalysis of environmental datasets

Any high spatial resolution weather forecast can be used. We utilize a custom configuration of WRF-ARW (Skamarock et al., 2019) at a spatial resolution of 2 km. Hourly predictions are made, and for each trap location (i.e., where we have measurements about pest abundance), we obtain daily values for air (2 m) and soil temperature (0 m), relative humidity (RH), accumulated precipitation (AP), dew point (DP), and wind speed (WS). These parameters have been widely used in related work and are extremely valuable for learning from past (reanalysis) and future (NWP) pest states.

3.4. In-field measurements

In-field measurements involve ground observations of pest abundance using pheromone traps specifically designed for monitoring the cotton bollworm, known by the scientific name *Helicoverpa armigera* (*H. armigera*). These traps contain the active ingredients Z-11-hexadecen-1-al and Z-9-hexadecenal. The traps are used from the beginning of the first generation until the end of the season, with regular replacement every 4 to 6 weeks. The company Corteva Agriscience Hellas has established a dense (in time and space) trap network (Figure 2) that covers almost all areas in the Greek mainland where cotton is cultivated. The traps are strategically positioned at suitable distances from each other to prevent interference and ensure accurate data collection. An agronomist examines the traps and counts the trapped insects at regular intervals every 3–5 days. Corteva Agriscience Hellas provides historical data consisting of 398 trap sequences and 8202 unique data points from 2019 to 2022 (Table 1). They also provide auxiliary data on pesticide application, potential crop damage from pests, the severity of the damage, trap replacements, and scouter comments.

4. Approach and methods

4.1. Causal graph for representing domain knowledge

We constructed a causal graph (Figure 3) based on domain knowledge and expertise, denoted as G, that represents the underlying causal relationships within the pest-farm ecosystem for the *H. armigera* case.

Table 1. *Summary of trap data*

Year	Traps	Measurements	Mean	std	Sprays	Sprayed fields %
2022	126	2507	19.73	4.22	30	18.25
2021	109	2245	20.30	1.79	17	11.01
2020	81	1693	20.54	4.77	12	8.64
2019	82	1757	21.29	6.43	21	21.95

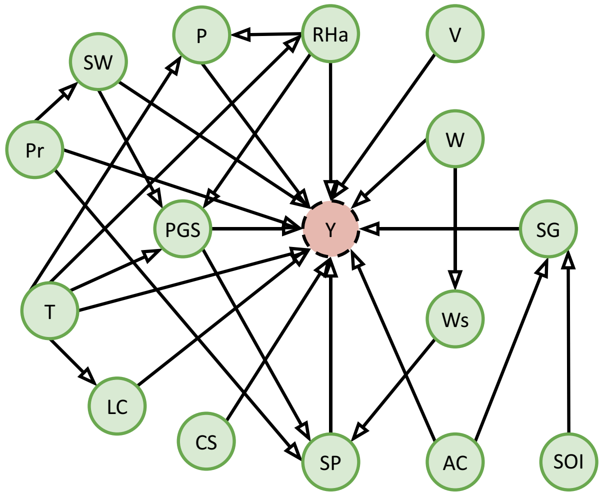

Figure 3. *Causal graph of a pest-farm ecosystem for Helicoverpa armigera case.*

The graph G comprises vertices V, which represent the variables in the system, and directed edges E, which symbolize the cause-and-effect relationships between these variables. Besides helping us articulate domain knowledge, the causal graph G will benefit the downstream technical analyses in various ways. For instance, G will be employed for effect identification via graphical tests (Pearl, 2009), where the structure of G is integral to discerning causal relationships. Conversely, in the case of estimating conditional average treatment effects within the potential outcomes framework, G will be utilized as a conceptual guide for considering causal structures during the control phase. In invariant causal prediction, the graph will facilitate the construction of an accurate list of invariant features using causal parents of the target outcome. Moreover, the structural knowledge captured in G could benefit invariant learning methods by guiding the environment E definition. This diverse and tailored incorporation of G is aimed at optimizing the utilization of domain knowledge by the specifications and objectives of each analytical technique.

Specifically, in the current case of the pest-farm ecosystem of *H. armigera*, various biotic and abiotic factors (Table 2) can influence the population dynamics Y of *H. armigera* (Sharma et al., 2012). Temperature T plays a crucial role, affecting the insect's growth, development, fecundity, and survival (Howe, 1967). The size SG of the first generation is related to the size of the second generation, and the Southern Oscillation Index SOI has a significant correlation with the size of the first spring generation (Maelzer and Zalucki, 1999, 2000). Additionally, the life cycle LC of *H. armigera* is temperature-dependent, with completion occurring between 17.5°C and 32.5°C (Mironidis and Savopoulou-Soultani, 2014). The presence of parasitoids and natural enemies in cotton cultivation is crucial to many IPM programs, including the control of *H. armigera* (Pereira et al., 2019). Many egg parasitoids of different families are known for their high parasitism P rates and their effectiveness in reducing the population of *H. armigera* (Noor-ul-Ane et al., 2015). Nevertheless, parasitism rates are influenced by temperature and

https://doi.org/10.1017/eds.2025.14 Published online by Cambridge University Press

Table 2. *Pest-farm ecosystem variables*

Id	Variable description
T	Temperature
SW	Soil water
RHa	Air relative humidity
SG	Size of generation
Pr	Precipitation
LC	Life cycle
P	Parasitism
V	Variety
Sp	Spraying
CS	Cropping system
AC	Adjacent crops
W	Wind
Ws	Spraying wind
SOI	South oscillation index
PGS	Plant growth stage
Y	Outcome (*H. armigera* population)

relative humidity (Kalyebi et al., 2005; Noor-ul-Ane et al., 2015). Moreover, the efficacy of spray application *Sp* also impacts population dynamics (Wardhaugh et al., 1980). The efficacy of *Sp* is significantly influenced by the plant growth stage *PGS*. During the seedling stage, limited leaf surface area reduces spray coverage, while the vegetative stage offers more extensive leaf area, enhancing spray interception. However, dense canopies at later stages may impede spray penetration. Plant physiology also varies, affecting the absorption and translocation of sprayed substances (Fishel and Ferrell, 2010).

Other environmental factors come into play as well. Precipitation *Pr* affects the population size, with heavy precipitation leading to a decrease in the population (Ge et al., 2003). It also increases soil water content *SW* which affects the emergence rate of *H. armigera* similar to air relative humidity *RHa* (Fajun et al., 2003). The presence of fruiting organs during the plant growth stage *PGS* is important for population dynamics, as it serves as the oviposition site for females (Fitt, 1989). Crop variety *V*, such as transgenic Bt cotton, can suppress the second generation of *H. armigera*, while both different cropping systems *CS* and adjacent crops *AC* can influence the population structure (Wardhaugh et al., 1980; Gao et al., 2010; Lu et al., 2013). Finally, wind *W* and wind direction play a significant role in the emergence of *H. armigera*, influencing the distance covered during migration from nearby locations. Additionally, wind conditions at the time of spraying W_s can also impact the effectiveness of the intervention. These various factors collectively shape the population dynamics of *H. armigera* in a complex and interconnected manner as defined through domain knowledge and depicted in the causal graph (Figure 3).

4.2. Invariant & causal learning for robust pest prediction

Our goal is to predict near-future pest populations (Y_{t+1}) using Earth observation (EO) and environmental data (X_t) along with weather forecasts (W_{t+1}) by learning the function $y_{t+1} = f(x_t, w_{t+1})$. Pest management recommendations heavily depend on these predictions. Conventional machine learning methods (Aparecido et al., 2019; Skawsang et al., 2019; Xiao et al., 2019; Zhang et al., 2019), which often assume that data points are independent and identically distributed (i.i.d.), struggle to generalize to unseen environments, capture spatiotemporal variability, and adapt to climate change. These methods are prone to learning spurious correlations, limiting their effectiveness in dynamic and non-i.i.d. scenarios.

To address these challenges, we turn to causal learning (Schölkopf and von Kügelgen, 2022), which leverages domain knowledge and is grounded in the principle of independent causal mechanisms. This

https://doi.org/10.1017/eds.2025.14 Published online by Cambridge University Press

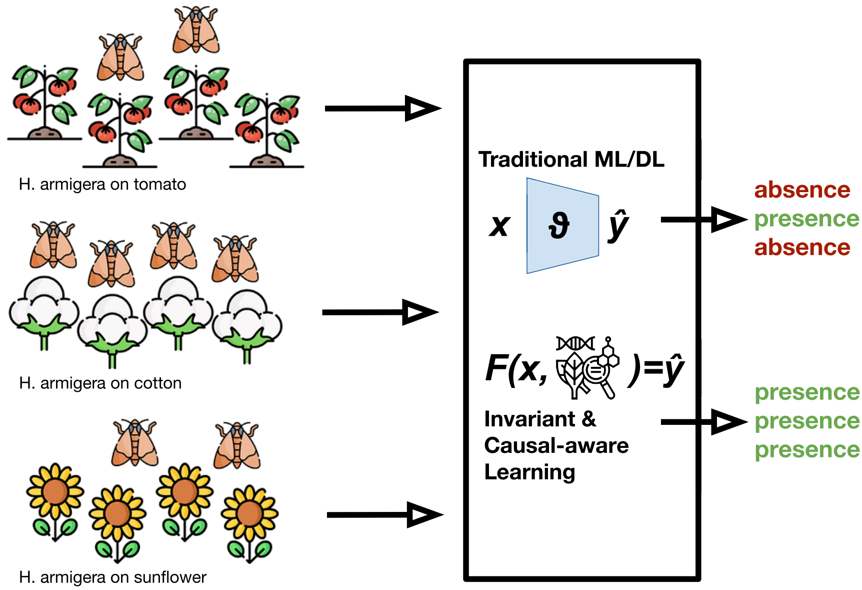

Figure 4. *Invariant learning for robust predictions. Stable and accurate predictions in diverse environments, such as when* H. armigera *feeds on different crops exhibiting variations in phenotype, agricultural management practices, and spatial distribution. Traditional ML methods risk capturing spurious correlations, such as associating pest abundance with a specific crop (e.g., cotton) due to its higher frequency in the dataset, leading to biased predictions based on the underlying crop rather than true pest presence.*

principle suggests that joint probabilities can be decomposed into separate mechanisms, each reflecting an underlying causal relationship that remains stable despite environmental changes. By incorporating this principle, our models can improve generalization and robustness across varying conditions.

We achieve this by integrating invariant learning with causality and categorizing dataset units into environments E as different agroclimatic zones or host crops (Figure 4). While E influences feature x_t, w_{t+1}, it does not directly affect the target Y_t. Utilizing Invariant Causal Prediction (ICP) (Heinze-Deml et al., 2018), Directed Acyclic Graphs (DAGs), and Invariant Risk Minimization (IRM) (Arjovsky et al., 2019), we can select causal features, identify potential causal relationships, and capture latent causal structures. These tools allow us to build models that are effective in current conditions and adaptable to future environmental changes.

4.3. Explainability & counterfactual reasoning for short-term advice

We define the problem as a binary classification of pest presence or absence at the next time step, using Earth observation (EO) data (X_t) and weather forecasts (W_{t+1}). The goal is to predict the pest population value at the next time step, Y_{t+1}, by learning the function $y_{t+1} = f(x_t, w_{t+1})$. To enhance the trustworthiness of our predictions, we employ Explainable Boosting Machines (EBM) (Nori et al., 2019). This glass-box model achieves high performance while providing inherent explanations at both global and local levels. EBM's additive nature allows for the sorting and visualization of feature contributions on a local scale for each one of predictions and a global level to summarize the general behavior of the model depending on features (Figure 5), which facilitates a better understanding of the primary drivers of the model and enhances trust in its outputs.

We propose generating counterfactual examples as recommended interventions to bolster trust further and provide actionable insights. Following the setup of (Mothilal et al., 2020), we search for minimal perturbations to the feature values (x_t, w_{t+1}) that would alter the prediction to the desired class using the

https://doi.org/10.1017/eds.2025.14 Published online by Cambridge University Press

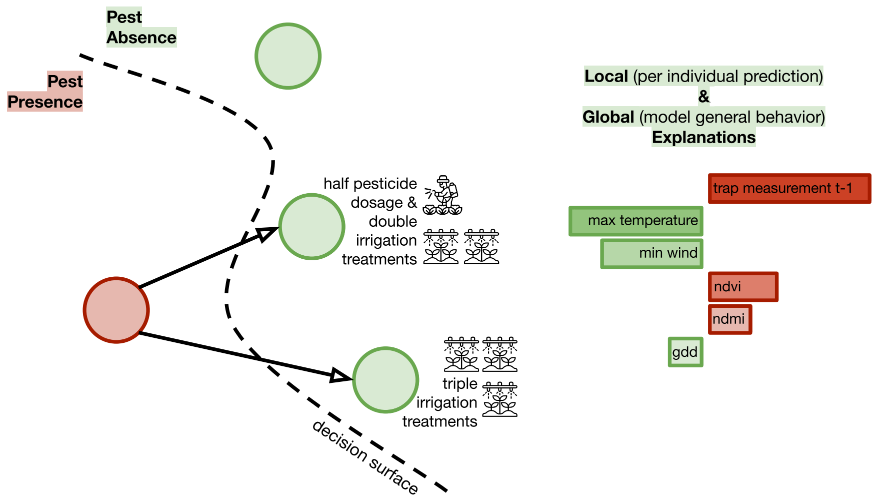

Figure 5. *Explainability for trustworthiness enhancement, on the right, with local and global explanations of each prediction and general model behavior, respectively, & Counterfactual explanations as agricultural actionable recommendations on the left.*

same model f. These counterfactual examples represent proposed actions that could be implemented in natural farm systems, ensuring practicality and feasibility (Wachter et al., 2017; Mothilal et al., 2020). The approach ensures that the generated counterfactuals are close to the original input but predicted in the desired class, providing feasible and actionable recommendations for IPM (Figure 5).

4.4. Heterogeneous treatment effects for long-term advice

We provide long-term pest prevention and suppression advice by assessing how agricultural practices (e.g., crop rotation, balanced fertilization, sowing dates) impact pest harmfulness and yield indices. Since different agro-environments may respond variably to the same practice, it is crucial to account for this heterogeneity. We estimate the conditional average treatment effect (CATE) following the potential outcomes framework (Rubin, 2005).

The CATE quantifies the difference in potential outcomes, represented as $\mathbb{E}[Y(T=1) - Y(T=0)|X]$, where $Y(T)$ denotes the value of a random variable Y (e.g., pest harmfulness and yield) if a unit is treated with treatment $T \in \{0,1\}$. By controlling for field characteristics X—which capture the heterogeneity across different agro-environmental conditions—we can better understand how specific practices affect outcomes in various contexts (Figure 6). This approach allows us to provide tailored and effective long-term IPM advice sensitive to each field's unique conditions (Giannarakis et al., 2022).

4.5. Causal inference for evaluating advice effectiveness

We employ causal inference techniques to assess the effectiveness of our pest control recommendations, adapting approaches recently introduced in agricultural contexts (Tsoumas et al., 2023). Specifically, in the case of pest management and with available panel data (Table 1), we utilize causal models such as difference-in-differences (DiDs) (Abadie, 2005), synthetic control (Arkhangelsky et al., 2021) and synthetic DiDs (Abadie, 2021) to quantify the treatment effect of adhering to our framework's recommendations (*treated units*) compared to those who did not (*control units*). Historical intervention data retrospectively annotated based on whether our framework recommended action, will serve as the basis for advice evaluation. Causal inference will be performed per-environment to ensure comparability between treatment and control groups, adhering to the parallel trends assumption (Lechner et al., 2011).

https://doi.org/10.1017/eds.2025.14 Published online by Cambridge University Press

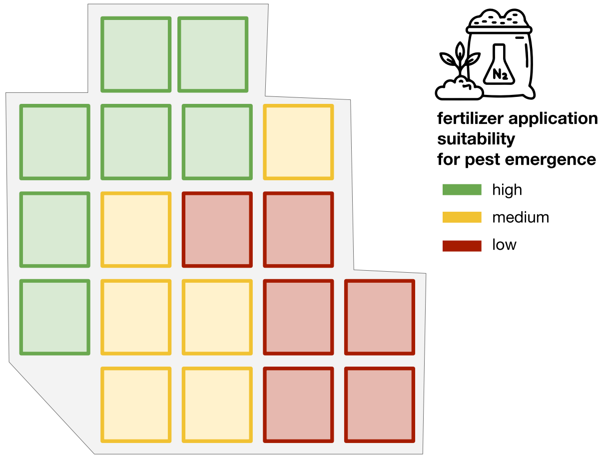

Figure 6. *Conditional Average Treatment Effect (CATE) is seen as long-term personalized guidance. By accounting for each land unit's unique characteristics, we can estimate a distinct treatment effect for each land unit. For example, how differences in land's characteristics can change the impact of fertilizer application on increasing the risk of pest emergence in the future.*

However, digital agriculture requires a two-level evaluation of interventions to disentangle the effectiveness resulting from the accuracy of the recommendation (for intervention) in terms of space–time from the inherent efficacy of the intervention. It is crucial to determine what effect, if any, is attributable to the space and time of application and what is due to the pesticide itself.

In this context, we conducted an initial analysis using the aforementioned panel data to quantify the impact of pesticide application on pest abundance in a real-world setting without expert system guidance, employing staggered DiDs with fixed effects (Eq. 4.1).

The staggered approach accounts for units receiving treatment at different periods. We include unit-fixed effects to control for each unit's time-invariant characteristics and time-fixed effects to capture overall time trends that affect all units in each period. The unit of analysis is the plot where the pest trap is located, with periods modeled at the weekly level. Here, Y_{it} represents the outcome variable, accumulated pest abundance, for each unit i at the time t, and treated_time_{it} is an indicator of whether the unit i receives treatment (pesticide application) in a period t (in a staggered manner across units). Specifically, β_0 is the intercept, β_1 is the treatment effect coefficient, α_i represents unit fixed effects, γ_t captures time fixed effects, and ε_{it} is the error term. Thus, β_1 provides the average causal effect of the treatment (pesticide application) on the outcome (accumulated pest abundance) for treated units (ATT), as presented in Table 3 for each cultivation period from 2019 to 2022.

$$Y_{it} = \beta_0 + \beta_1 \cdot \text{treated\_time}_{it} + \alpha_i + \gamma_t + \epsilon_{it} \tag{4.1}$$

For the years 2021 and 2022, we observe a statistically significant reduction in pest abundance, while for 2019 and 2020, we find the opposite effect. At first glance, this contradiction may seem unusual, but several reasonable explanations could account for it. Since the data come from real-world agricultural practice, it likely encapsulates some of the following issues: (i) Some interventions may have been applied incorrectly regarding timing and method, reducing or eliminating their efficacy in the pest-infested plots. This could lead to a biased estimate that the pest population increased after pesticide application (Figure 7). This occurs because the counterfactual is constructed by taking the growth trend from a plot without intervention, which might not experience the same infestation or pest pressure level. So, a mistreated plot that probably follows a steeper population increase, simply due to its higher infestation levels, can lead to this fallacy that pesticide application increases pest population. (ii) After discussions with the data provider (Corteva Agriscience Hellas), noise within the control group labels may be

https://doi.org/10.1017/eds.2025.14 Published online by Cambridge University Press

Table 3. *Results of staggered DiDs with controls for unobserved heterogeneity at the unit and time levels by including fixed effects*

Staggered DiDs estimates with fixed effects			
Year	ATT	CI	p-value
2019	35.6065	(30.569, 40.644)	0.000
2020	36.9961	(29.826, 44.166)	0.000
2021	−13.8687	(−20.803, −6.934)	0.000
2022	−8.5789	(−13.549, −3.609)	0.001

Note: It includes point estimates, 95% confidence intervals, and p-value. Numbers represent the increase/decrease of accumulated pest catchments at the trap level after the intervention.

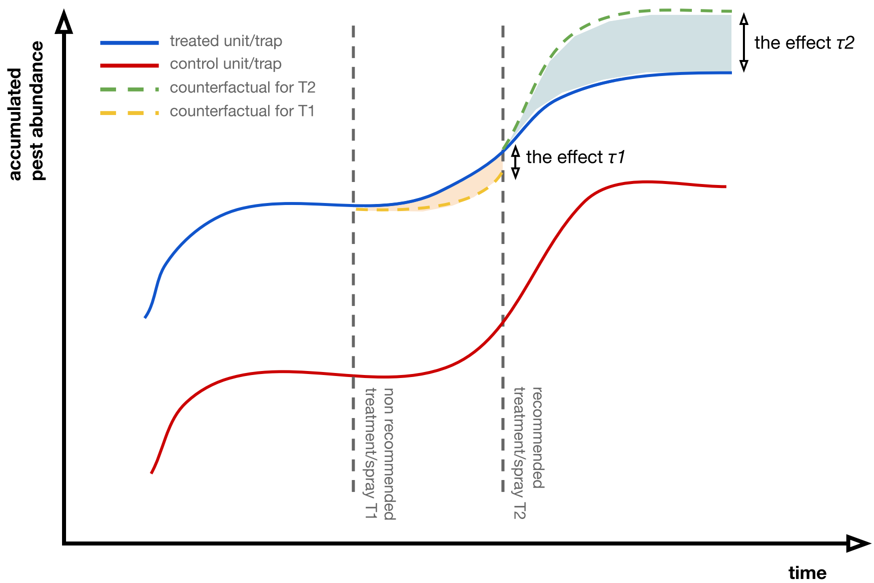

Figure 7. *A visual example of DiDS for assessing the real-world impact of pesticide application. It demonstrates how, even when the parallel trends assumption holds in both conditions, applying an intervention (i.e., spray) at a non-recommended time can lead to unexpected effects compared to applying the intervention at the recommended time.*

possible. The company is confident in the labels for treated plots, as they receive this information directly from farmers. However, they cannot be as certain about the control group. Some farmers may have applied pest control practices in their plots but chose not to report them for various reasons, such as using less expensive pesticides from competitor companies or participating in eco-schemes prohibiting pesticide use. Consequently, we face a scenario of positively labeled and unlabeled data, a common issue in machine learning. (iii) The assumption of parallel trends may not hold universally, or unobserved confounders may vary over time and between units.

In a more robust causal analysis, we can technically or conceptually address these issues. Technically, we could retrospectively employ a recommendation system or consult experts, as aforementioned, to annotate each time–space slot as favorable or unfavorable for intervention. On the other hand, we can conceptually accept reality and precisely define what causal effect we retrieve. In this case, the ATT in a real-world setting includes different application accuracy levels, farmer's skills, expert guidance, and proper timing. To address the second issue, we plan to use Positive-Unlabeled (PU) learning methods (Bekker and Davis, 2020) to train a classifier on covariates, as they are outlined in Section 3. Using the positively labeled (treated) units only as ground truth and PU learning for training, this classifier will help establish a control group consisting only of unlabeled units that are classified there with high confidence. Lastly, a formal investigation with statistical tests is required to retain only cases where the parallel trends assumption holds. Clear assumptions statements should also be made regarding the potential of unobserved confounders that may vary by time and unit. By leveraging these techniques, we aim to rigorously evaluate the impact of our recommendations on pest control outcomes and attribute the effects to the right factors, providing robust evidence for the effectiveness of our framework in diverse agricultural environments.

5. Conclusions

In conclusion, this article presents a new framework integrating causality and explainability into digital agriculture, with a focus on enhancing pest management practices. By leveraging advanced data analysis techniques, such as causal inference and invariant learning, our approach addresses the limitations of conventional correlation-based models, providing more robust and transparent decision-making tools. This framework not only supports real-time pest control interventions but also facilitates strategic long-term planning by offering insights into the heterogeneous effects of various agricultural practices.

Our study illustrates how incorporating explainability can bolster farmers' trust and adoption of sustainable practices like IPM. The framework's use of counterfactual reasoning and explainable predictions ensures that farmers receive actionable, field-specific recommendations that can adapt to different environmental conditions. Additionally, the causal analysis embedded within our methodology allows for ongoing evaluation of the framework's effectiveness, ensuring the recommendations are impactful and contribute positively to agricultural sustainability.

We consider that a successful application to pest management will highlight, in a tangible way, the broader potential of this framework to enhance digital agriculture to drive sustainable, evidence-based practices across agriculture. Therefore, we plan to implement the proposed ideas outlined in Section 4 using the data described in Section 3. In parallel, we are gathering additional in-situ data in collaboration with Corteva Agriscience Hellas to enrich our dataset for the same pest and crop, as well as independently for other crops and pests. Finally, we explore how this approach could be adapted to related areas.

Future research will aim to expand this framework beyond pest management, exploring its potential applications in other areas of digital agriculture, such as crop disease management and nutrient optimization. Additionally, integrating advanced machine learning models to account for real-time weather data and unforeseen environmental factors will further refine prediction accuracy. Developing user-friendly tools and interfaces that facilitate farmer interaction with these data-driven insights will be critical to fostering widespread adoption.

The growing demand for sustainable agriculture underlines the importance of integrating advanced data analysis frameworks like ours. By systematically quantifying and explaining agricultural interventions, this framework offers a promising pathway for enhancing the adoption of digital agriculture in alignment with global sustainability goals. This comprehensive, data-driven approach promises to make sustainable agricultural practices more practical, facilitating a transition to a resilient and environmentally conscious food system.

Open peer review. To view the open peer review materials for this article, please visit http://doi.org/10.1017/eds.2025.14.

Acknowledgments. We express our gratitude to Corteva Agriscience Hellas, particularly to Dr. George Zanakis, the Marketing & Development Manager, for their invaluable support, insights, trust, and provision of data.

Author contribution. Conceptualization: Ilias Tsoumas, Vasileios Sitokonstantinou; Methodology: Ilias Tsoumas, Vasileios Sitokonstantinou, Evagelia Lampiri; Software: Ilias Tsoumas; Formal analysis: Ilias Tsoumas; Investigation: Evagelia Lampiri;

Data curation: Ilias Tsoumas; Writing—Original Draft: Ilias Tsoumas, Vasileios Sitokonstantinou, Evagelia Lampiri; Writing—Review & Editing: all authors; Visualization: Ilias Tsoumas All authors approved the final submitted draft.

Competing interests. The authors declare none.

Data availability statement. Data availability is governed by the terms of the Memorandum of Understanding (MoU) between the National Observatory of Athens and Corteva Agriscience Hellas. Access to the data can be granted upon a formal written request to both entities. Corteva Agriscience Hellas reserves the right to make the final decision regarding access to the raw in-situ data.

Ethics statement. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

Funding statement. This research is funded by the European Union's Horizon Europe Programme under Grant Agreement No. 101135422 (UNIVERSWATER—Ilias Tsoumas, Charalampos Kontoes) and supported by the European Union Horizon Europe Research and Innovation program under Grant Agreement #101070496 (Smart Droplets—Ioannis N. Athanasiadis). Vasileios Sitokonstantinou and Gustau Camps-Valls acknowledge support from the Generalitat Valenciana and the Conselleria d'Innovació, Universitats, Ciència i Societat Digital, through the project "AI4CS: Artificial Intelligence for Complex Systems: Brain, Earth, Climate, Society" (CIPROM/2021/56). Additionally, Gustau Camps-Valls acknowledges the support of the European Research Council (ERC) under the ERC Synergy Grant USMILE (Grant Agreement 855187).

References

Abadie A (2005) Semiparametric difference-in-differences estimators. *The Review of Economic Studies 72*(1), 1–19.

Abadie A (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature 59*(2), 391–425.

Aparecido LE, **Rolim G**, **Moraes J**, **Costa C and Souza P** (2019) Machine learning algorithms for forecasting the incidence of coffee arabica pests and diseases. *International Journal of Biometeorology 64*(4), 671–688. https://doi.org/10.1007/s00484-019-01856-1

Arjovsky M, **Bottou L**, **Gulrajani I and Lopez-Paz D** (2019) Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Arkhangelsky D, **Athey S**, **Hirshberg DA**, **Imbens GW and Wager S** (2021) Synthetic difference-in-differences. *American Economic Review 111*(12), 4088–4118.

Audsley E, **Stacey K**, **Parsons DJ and Williams AG** (2009) *Estimation of the Greenhouse Gas Emissions from Agricultural Pesticide Manufacture and Use.* Cranfield University

Balasundram SK, **Shamshiri RR**, **Sridhara S and Rizan N** (2023) The role of digital agriculture in mitigating climate change and ensuring food security: An overview. *Sustainability 15*(6), 5325.

Ballabio C, **Panagos P and Monatanarella L** (2016) Mapping topsoil physical properties at European scale using the Lucas database. *Geoderma 261*, 110–123.

Barzman M, **Bàrberi P**, **Birch ANE**, **Boonekamp P**, **Dachbrodt-Saaydeh S**, **Graf B**, **Hommel B**, **Jensen JE**, **Kiss J**, **Kudsk P**, et al (2015) Eight principles of integrated pest management. *Agronomy for Sustainable Development 35*, 1199–1215.

Basso B and Antle J (2020) Digital agriculture to design sustainable agricultural systems. *Nature Sustainability 3*(4), 254–256.

Bekker J and Davis J (2020) Learning from positive and unlabeled data: A survey. *Machine Learning 109*(4), 719–760.

Boedeker W, **Watts M**, **Clausing P and Marquez E** (2020) The global distribution of acute unintentional pesticide poisoning: Estimations based on a systematic review. *BMC Public Health 20*(1), 1–19.

Ceglar A, **Zampieri M**, **Toreti A and Dentener F** (2019) Observed northward migration of agro-climate zones in europe will further accelerate under climate change. *Earth's Future 7*(9), 1088–1101.

de Brogniez D, **Ballabio C**, **Stevens A**, **Jones R**, **Montanarella L and vanWesemael B** (2015) 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.

Fajun C, **Baoping Z and Xiaoxi Z** (2003) Effects of soil mositure during pupual stage on population development of cotton bollowrm, helicoverpa armigera (hubner). *Acta Ecologica Sinica 23*(1), 112–121.

Fishel FM and Ferrell JA (2010) Managing pesticide drift: Pi232/pi232, 9/2010. *EDIS 2010*(7).

Fitt GP (1989) The ecology of heliothis species in relation to agroecosystems. *Annual Review of Entomology 34*(1), 17–53.

Gao Y, **Zhai B**, et al (2010) Active temperature selection of flying helicoverpa armigera (lepidoptera: Noctuidae) moths. *Acta Entomologica Sinica 53*(5), 540–548.

Ge F, **Liu X**, **Ding Y**, **Wang X and Zhao Y** (2003) Life-table of helicoverpa armigera in northern china and characters of population development in southern and northern china. *Ying Yong Sheng tai xue bao= The Journal of Applied Ecology 14*(2), 241–245.

Giannarakis G, **Sitokonstantinou V**, **Lorilla RS and Kontoes C** (2022) Towards assessing agricultural land suitability with causal machine learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA pp. 1441–1451 doi: 10.1109/CVPRW56347.2022.00150.

Heimpel GE, **Yang Y**, **Hill JD and Ragsdale DW** (2013) Environmental consequences of invasive species: Greenhouse gas emissions of insecticide use and the role of biological control in reducing emissions. *PLoS One 8*(8), e72293.

Heinze-Deml C, **Peters J and Meinshausen N** (2018) Invariant causal prediction for nonlinear models. *Journal of Causal Inference 6*(2), 20170016.

Howe R (1967) Temperature effects on embryonic development in insects. *Annual Review of Entomology 12*(1), 15–42.

Iost Filho FH, **de Bastos Pazini J**, **Alves TM**, **Koch RL and Yamamoto PT** (2022). How does the digital transformation of agriculture affect the implementation of integrated pest management?. *Frontiers in Sustainable Food Systems 6* https://www.frontiersin.org/journals/sustainable-food-systems/articles/10.3389/fsufs.2022.972213 doi 10.3389/fsufs.2022.972213 ISSN 2571-581X

Kalyebi A, **Sithanantham S**, **OverholtW**, **Hassan S and Mueke J** (2005). Parasitism, longevity and progeny production of six indigenous kenyan trichogrammatid egg parasitoids (hymenoptera: Trichogrammatidae) at different temperature and relative humidity regimes. *Biocontrol Science and Technology 15*(3), 255–270.

Lechner M, et al (2011) The estimation of causal effects by difference-in-difference methods. *Foundations and Trends® in Econometrics 4*(3), 165–224.

Lu Z-Z, **Zalucki MP**, **Perkins LE**, **Wang D-Y and Wu L-L** (2013) Towards a resistance management strategy for helicoverpa armigera in bt-cotton in northwestern china: An assessment of potential refuge crops. *Journal of Pest Science 86*, 695–703.

Maelzer D and Zalucki M (1999) Analysis of long-term light-trap data for helicoverpa spp.(lepidoptera: Noctuidae) in australia: The effect of climate and crop host plants. *Bulletin of Entomological Research 89*(5), 455–463.

Maelzer D and Zalucki M (2000) Long range forecasts of the numbers of helicoverpa punctigera and h. armigera (lepidoptera: Noctuidae) in australia using the southern oscillation index and the sea surface temperature. *Bulletin of Entomological Research 90*(2), 133–146.

Marty M, **Spurlock F and Barry T** (2010) Volatile organic compounds from pesticide application and contribution to tropospheric ozone. In *Hayes' Handbook of Pesticide Toxicology*. Elsevier, pp. 571–585.

Mironidis G and Savopoulou-Soultani M (2014) Development, survivorship, and reproduction of helicoverpa armigera (lepidoptera: Noctuidae) under constant and alternating temperatures. *Environmental Entomology 37*(1), 16–28.

Mothilal RK, **Sharma A and Tan C** (2020) Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, New York, NY, USA, pp. 607–617 https://doi.org/10.1145/3351095.3372850.

Nanushi O, **Sitokonstantinou V**, **Tsoumas I and Kontoes C** (2022) Pest presence prediction using interpretable machine learning. In *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, Nafplio, Greece pp. 1–5, doi: 10.1109/IVMSP54334.2022.9816284.

Noor-ul-Ane M, **Arif MJ**, **Gogi MD and Khan MA** (2015) Evaluation of different integrated pest management modules to control helicoverpa for adaptation to climate change. *International Journal of Agriculture and Biology 17*(3).

Nori H, **Jenkins S**, **Koch P and Caruana R** (2019) Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*.

Pearl J (2009) *Causality*. Cambridge University Press.

Pereira FP, **Reigada C**, **Diniz AJF and Parra JRP** (2019). Potential of two trichogrammatidae species for helicoverpa armigera control. *Neotropical Entomology 48*, 966–973.

Rubin DB (2005) Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association 100*(469), 322–331.

Schölkopf B and von Kügelgen J (2022) From statistical to causal learning. In *Proceedings of the International Congress of Mathematicians 1*.

Sharma A, **Kumar V**, **Shahzad B**, **Tanveer M**, **Sidhu GPS**, **Handa N**, **Kohli SK**, **Yadav P**, **Bali AS**, **Parihar RD**, et al (2019) Worldwide pesticide usage and its impacts on ecosystem. *SN Applied Sciences 1*, 1–16.

Sharma PK,**Kumar U**,**Vyas S**, **Sharma S**,&**Shrivastava S** (2012) Monitoring of helicoverpa armigera (hubner)(lepidoptera: Noctuidae) through pheromone traps in chickpea (cicer arietinum) crop and influence of some abiotic factors on insect population. *Journal of Environmental Science, Toxicology and Food Technology 1*(5), 44–46.

Sharma HC and Prabhakar CS (2014) Impact of climate change on pest management and food security. In *Integrated Pest Management*. Elsevier, pp. 23–36.

Sharma A, **Reeves M and Washburn C** (2022) *Pesticides and Climate Change: A Vicious Cycle* (tech. rep.). Pesticide Action Network North America.

Sitokonstantinou V, **Porras EDS**, **Bautista JC**, **Piles M**, **Athanasiadis I**, **Kerner H**, **Martini G**, **Sweet L-b**, **Tsoumas I**, **Zscheischler J**, et al (2024) Causal machine learning for sustainable agroecosystems*ar. Xiv preprint arXiv:2408.13155.*

Skamarock WC , **Klemp JB**, **Dudhia J**, **Gill DO**, **Liu Z**, **Berner J**,**Wang W**, **Powers JG**, **Duda MG**, **Barker DM**, et al (2019) *A Description of the Advanced Research WRF Model Version 4*, vol. *145*(145). Boulder, CO: National Center for Atmospheric Research, p. 550.

Skawsang S, **Nagai M**, **Tripathi N and Soni P** (2019) Predicting rice pest population occurrence with satellite-derived crop phenology, ground meteorological observation, and machine learning: A case study for the central plain of thailand. *Applied Sciences 9*, 4846. https://doi.org/10.3390/app9224846

Skendžić S, **Zovko M**, **Živković IP**, **Lešić V and Lemić D** (2021) The impact of climate change on agricultural insect pests. *Insects 12*(5), 440.

Spokas K and Wang D (2003) Stimulation of nitrous oxide production resulted from soil fumigation with chloropicrin. *Atmospheric Environment 37*(25), 3501–3507.

Tsoumas I, **Giannarakis G**, **Sitokonstantinou V**, **Koukos A**, **Loka D**, **Bartsotas N**, **Kontoes C and Athanasiadis I** (2023) Evaluating digital agriculture recommendations with causal inference. *Proceedings of the AAAI Conference on Artificial Intelligence 37*(12), 14514–14522.

Wachter S, **Mittelstadt B and Russell C** (2017) Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology 31*, 841.

Wardhaugh K, **Room P and Greenup L** (1980) The incidence of heliothis armigera (hübner) and h. punctigera wallengren (lepidoptera: Noctuidae) on cotton and other host-plants in the namoi valley of new south wales. *Bulletin of Entomological Research 70*(1), 113–131.

Xiao Q, **Li W**, **Kai Y**, **Chen P**, **Zhang J and Wang B** (2019) Occurrence prediction of pests and diseases in cotton on the basis of weather factors by long short term memory network. *BMC Bioinformatics 20*(Suppl 25), 688. https://doi.org/10.1186/s12859-019-3262-y

Xu S, **Sheng C and Tian C** (2020) Changing soil carbon: Influencing factors, sequestration strategy and research direction. *Carbon Balance and Management 15*, 1–9.

Zhang J, **Huang Y**, **Pu R**, **González-Moreno P**, **Yuan L**, **Wu K and Huang W** (2019) Monitoring plant diseases and pests through remote sensing technology: A review. *Computers and Electronics in Agriculture 165*, 104943. https://doi.org/10.1016/j.compag.2019.104943

Cite this article: Tsoumas I, Sitokonstantinou V, Giannarakis G, Lampiri E, Athanassiou C, Camps-Valls G, Kontoes C and Athanasiadis IN (2025). Leveraging causality and explainability in digital agriculture. *Environmental Data Science*, 4: e23. doi:10.1017/eds.2025.14

Evaluating Digital Agriculture Recommendations with Causal Inference

Ilias Tsoumas[*1,2], **Georgios Giannarakis**[*1], **Vasileios Sitokonstantinou**[1], **Alkiviadis Koukos**[1], **Dimitra Loka**[3], **Nikolaos S. Bartsotas**[1], **Charalampos Kontoes**[1], **Ioannis Athanasiadis**[2]

[1]BEYOND Centre, IAASARS, National Observatory of Athens
[2]Wageningen University & 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 smart farming tools. Even though recent advancements in AI-driven digital agriculture can offer high-performing predictive functionalities, they lack tangible quantitative evidence on their benefits to the farmers. Field experiments can derive such evidence, but are often costly, time consuming and hence limited in scope and scale of application. 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 (e.g., yield in this case). This way, we can increase farmers' trust via enhancing the transparency of the digital agriculture market, and in turn accelerate the adoption of technologies that aim to secure farmer income resilience and global agricultural sustainability against a changing climate. As a case study, we designed and implemented a recommendation system for the optimal sowing time of cotton based on numerical weather predictions, which was used by a farmers' cooperative during the growing season of 2021. We then leverage agricultural knowledge, collected yield data, and environmental information to develop a causal graph of the farm system. Using the back-door criterion, we identify the impact of sowing recommendations on the yield and subsequently estimate it using linear regression, matching, inverse propensity score weighting and meta-learners. The results revealed that a field sown according to our recommendations exhibited a statistically significant yield increase that ranged from 12% to 17%, depending on the method. The effect estimates were robust, as indicated by the agreement among the estimation methods and four successful refutation tests. We argue that this approach can be implemented for decision support systems of other fields, extending their evaluation beyond a performance assessment of internal functionalities.

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. Nowadays, many industries enjoy automated and effective decision-making via harnessing the data that digitalization generates. However, the agricultural sector experiences limited adoption of precision agriculture and smart farming technologies (Gabriel and Gandorfer 2022). This might seem odd at first sight, given the surge of sophisticated digital tools that utilize Artificial Intelligence (AI) techniques and combine remote sensing data with data from Internet of Things (IoT) sensors to offer agricultural information of great detail (Sharma et al. 2020; Nanushi et al. 2022; Choumos et al. 2022). Yet farmers are skeptical about the effectiveness and actual contribution of these tools to their revenues and daily work (Lowenberg-DeBoer and Erickson 2019; Lioutas, Charatsari, and De Rosa 2021).

Traditionally, quantifying the impact of a service would require the design and execution of a randomized experiment (Boruch 1997). Nevertheless, field experiments for the evaluation of digital agriculture tools are seldom done since they are costly and time-consuming, requiring specialized designs and follow-up experiments for any changes in the product (Vaessen 2010; Diggle and Chetwynd 2011). In addition, field experiments can jeopardize the crops and hence the farmers' livelihood; and if potential damages are not covered, no prudent farmer would want to participate. As a result, the providers of digital agriculture tools often resort to unproven promises that unavoidably create customer mistrust. An observational causal inference framework (Pearl 2009) can fill this gap by emulating the experiment we would have liked to run (Hernán and Robins 2016).

Related Work

Causal inference with observational data has been the subject of recent work across diverse disciplines, including ecology (Arif and MacNeil 2022), public policy (Fougère and Jacquemet 2019), and Earth sciences (Massmann, Gentine, and Runge 2021; Runge et al. 2019). In agriculture, it has been used to identify and estimate the effect of agricultural practices on various agro-environmental metrics (Qian and Harmel 2016; Deines, Wang, and Lobell 2019; Giannarakis et al. 2022a,b). Using causal inference to test digital agriculture can provide reliable insights of superior socioeconomic impact than of those inferred by naive descriptive studies, including transparent benefits for the farmers, increased reliability, and honest pricing.

[*]These authors contributed equally.

According to Adelman (1992), the comprehensive evaluation of decision support systems has three facets: i) the subjective evaluation that assesses the system from the perspective of the end-user, ii) the technical evaluation that assesses the performance of the system's internal functionalities and iii) the empirical evaluation that experimentally assesses the impact of the system (Adelman 1992). Subjective evaluation has been widely practiced for decision support (Zhai et al. 2020) and recommender systems (Pu, Chen, and Hu 2012), retrieving user feedback (e.g., on usability, accessibility etc.) through surveys and questionnaires. From the perspective of the AI-based expert systems, the technical evaluation is based on predictive, classification and ranking accuracy metrics (Schröder, Thiele, and Lehner 2011). Technical evaluation metrics extending beyond accuracy are used within the context of recommender systems (e.g., coverage, serendipity) (Ge, Delgado-Battenfeld, and Jannach 2010). These metrics capture the recommendation quality as perceived by the user, connecting the technical and subjective evaluation concepts.

Interestingly, empirical evaluation methods, and in particular with regards to the impact assessment of digital agriculture tools, have been seldom employed. From the perspective of agricultural economics, tangential questions have been studied (Müller 1974; Roberts et al. 2009; Schimmelpfennig and Ebel 2016; McFadden, Rosburg, and Njuki 2022), but without approaching the question from a farm system standpoint, hence not leveraging available structural knowledge and reaping its benefits (Cinelli, Forney, and Pearl 2020). Thus, we propose a framework for the empirical evaluation of digital agriculture recommendations with causal inference. In this context, we designed and implemented a recommendation system for the optimal sowing of cotton. The system was tested in a real-world case study by providing it to a local agricultural cooperative and monitoring the results.

For several arable crops (e.g., cotton, maize, chickpea) sowing time is of great importance. Mistimed sowing can lead to suboptimal plant emergence and adversely affect the crop yield (Huang 2016; Nielsen et al. 2002; Richards, Maphosa, and Preston 2022). The agro-climatic conditions for optimal sowing have been extensively studied (Freeland Jr et al. 2006; Boman and Lemon 2005). Soil temperature, soil moisture and ambient temperature, during the first days after sowing, play a crucial role in proper germination and emergence and ultimately determine the yield and its quality (Bradow and Bauer 2010; Bauer, May, and Camberato 1998). The University of California (UC IPM 2004) and the Texas Tech University (Barbato, Seshadri, and Mulligan 2011) have developed digital tools that provide daily recommendations for optimal cotton sowing, using air temperature forecasts from the National Oceanic and Atmospheric Administration (NOAA). The accuracy of the University of California's cotton planting tool was studied by comparing the temperature forecasts to ground truth measurements from two weather stations (Munier et al. 2004). The tool forecasted the correct planting conditions (Kerby, Keeley, and Johnson 1989) 75% of the time, showing that weather forecast models can provide solid sowing recommendations.

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 (Luma-Osmani et al. 2020; Pasquel et al. 2022). The contributions of this work are summarized as follows: i) the design and implementation of the first empirical evaluation framework for digital agriculture based on causal inference; ii) the development of a high-resolution, knowledge-based recommendation system for the optimal sowing of cotton using weather predictions, 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 with different methods (linear regression, matching, inverse propensity score weighting and meta-learners), and the use of refutation tests to evaluate the robustness of estimates. Due to its accurate weather forecasts and customized rules, we find that the system offered effective sowing recommendations. It increased the cotton yield of the farmers that followed the recommendations by a factor that ranged from 12% to 17%, depending on the estimation method used.

Agricultural Recommendation System

In this work, we design, implement and evaluate a knowledge based recommendation system (Aggarwal 2016) 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 (Bradow and Bauer 2010). Soil or ambient temperatures lower than 10°C result in less vigorous and malformed seedlings (Boman and Lemon 2005). 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 1 (Freeland Jr et al. 2006; Boman and Lemon 2005). 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 1: 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 (Skamarock et al. 2019) 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; soil and ambient temperature are retrieved at an 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 1. 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 (NCEP and USDOC 2015) 15-day forecasts that are given on a 0.25 degrees (roughly 25 km) spatial resolution.

$$a_i = \frac{GFS_{day=i}}{GFS_{day=1}}, i \in \{3, ..., 10\} \quad (1)$$

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

Eq. (1) 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. (2) 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.

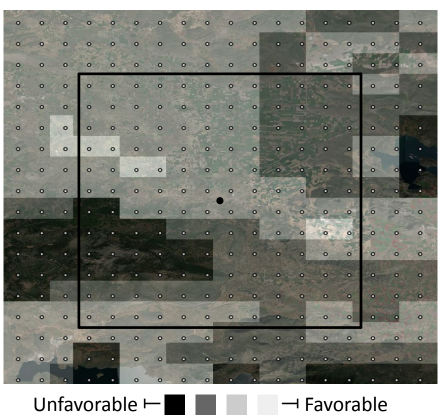

Figure 1: Optimal sowing map for a given day. The black circle at the center depicts the GFS grid point that represents the entire black-lined box. The white circles depict the 144 ART grid points for the same area.

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 1. 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 no 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.

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 graphical comparisons, see Appendix Figure 1). For the maximum temperature we found $MAE = 2.41$, $RMSE = 3.11$, whereas for the minimum temperature we found $MAE = 2.75$, $RMSE = 3.70$.

Real-World Case Study. We combined the ART weather forecasts and the conditions listed in Table 1 to produce a recommendation system in the form of daily maps over the fields of the farmers of the cooperative (Figure 1). The sowing recommendation maps were served through the website of the cooperative that farmers visited on a daily basis. The cooperative collected and provided for each field: its georeferenced boundaries, the sowing date, the seed variety, the harvest date, the precise final yield, and for a subset of the fields the yield of the previous year. We then combined this data with publicly available observations from heterogeneous sources, such as satellites (Sentinel-2), weather stations and GIS maps, to engineer an observational dataset that enables a causal analysis for studying the impact of the recommendation system on the yield.

Causal Evaluation Framework

Notation and 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 (Pearl 2009). The directed edge $A \rightarrow B$ 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. Following popular terminology, 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. (3).

$$\text{ATE} = \mathbb{E}\big[Y|do(T=1)\big] - \mathbb{E}\big[Y|do(T=0)\big] \quad (3)$$

Problem Formulation. 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, i.e., the corresponding 10-day ART forecasts satisfied the appropriate conditions, and as *control* the fields that were sown on a non-favorable day. Because the system outputs 4 levels of recommendation ranging from 0 (bad) to 3 (good), we define a day as favorable when all conditions are satisfied, i.e., when the system outputs the highest recommendation value that is 3. A day is then defined as non-favorable when the system outputs any other recommendation value. Binarizing the treatment in that way allows for greater flexibility in estimator selection and easier interpretation. Formally,

$$T = \begin{cases} 1 & \text{, if recommendation of sowing date} \in \{3\} \\ 0 & \text{, if recommendation of sowing date} \in \{0,1,2\} \end{cases}$$

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. Such an estimand is of primary significance for the farmers, but also for proving the reliability and therefore accelerating the adoption of smart farming tools. 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.

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 (Bauer, May, and Camberato 1998; Bradow and Bauer 2010).

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 (Eklundh and Jönsson 2015). Even though in the case of cotton, trapezoidal NDVI is not linearly correlated with yield (Dalezios et al. 2001; Zhao et al. 2007), it is correlated with early season Leaf Area Index (LAI) (Zhao et al. 2007), which in turn is a good indicator of early season crop growth rate (Virk, Snider, and Pilon 2019). Furthermore, seed germination and seedling emergence are greatly dependent on soil moisture. Hence, soil moisture SM is a confounder for the relation $T \rightarrow Y$ that we study. As a SM proxy, we used the Normalized Difference Water Index (NDWI) at sowing day which is highly correlated with soil moisture in bare soil (Casamitjana et al. 2020).

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 practices are a result of a common cause that we can define as Agricultural Knowledge (AK), capturing the skills and experience 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 (Calder, Phillips, and Tybout 1982), 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 (Huntington-Klein 2021).

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 (Virk, Snider, and Pilon 2019; Boman and Lemon 2005; Varco 2020). Low temperatures result in reduced germination, slow growth and less vigorous seedlings that are more prone to diseases and sensitive to weed competition (Bradow and Bauer 2010). 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 1) 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 (Varco 2020). Data on SP and SoC were retrieved from the European Soil Data Centre (ESDAC) (Ballabio, Panagos, and Monatanarella 2016; de Brogniez et al. 2015). Both variables were included in the graph as confounders of T and CG. Seed variety also determines seed germination, emergence and final yield (Snider, Pilon, and Virk 2020). Seed mass and vigor (Liu et al. 2015; Snider, Pilon, and Virk 2020) 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.

Id	Variable Description	Source
T	Treatment	Farmers' Cooperative, RS
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 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 2: Farm system variable identifier, description and source. RS = Recommendation System.

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 (Green 1956). Since temperature is the primary environmental factor controlling plant growth (Bange and Milroy 2004; Hatfield and Prueger 2015), temperature fluctuations were 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 (Dong et al. 2006; Bange, Caton, and Milroy 2008). Table 2 summarizes the variables' description, abbreviation and source.

Causal Graph. Figure 2 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, 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.

Effect Identification and Estimation. Because the calculation of causal effects requires access to counterfactual values that are by definition not observed (Holland 1986), observational methods rely on identification techniques and assumptions that aim at reducing causal estimands such as

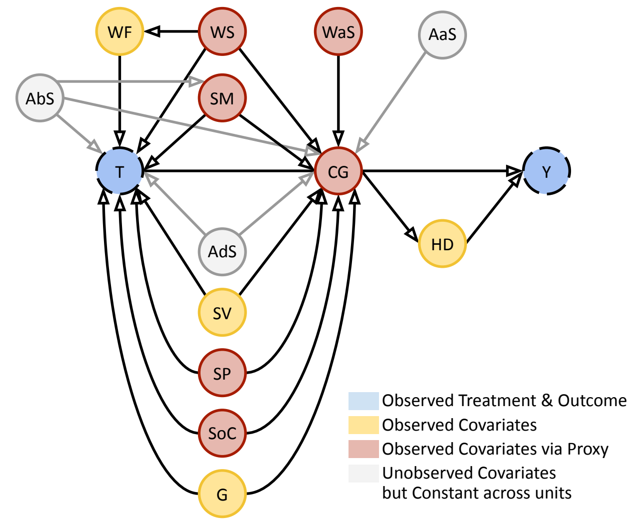

Figure 2: Graph of the farm system, encoding the causal relations between the relevant agro-environmental actors.

$\mathbb{P}(Y = y|do(T = t))$ 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. Formally, a set of variables Z satisfies the back-door criterion relative to an ordered pair of variables (T, Y) in a DAG G, if no node in Z is a descendant of T and Z blocks every path between T and Y that contains an arrow into T. After (if) we have obtained a back-door adjustment set to condition on, we can proceed with estimating the ATE of interest. The back-door criterion already provides a formula for the interventional distribution. Given a 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. To check covariate balance and as a method prerequisite, we model the propensity scores $\mathbb{P}(T = 1|Z = z)$, i.e., the probability of receiving treatment given features (Rosenbaum and Rubin 1983). Linear regression and distance matching are selected as baseline estimation methods. The popular Inverse Propensity Score (IPS) weigthing is also used (Stuart 2010). We finally apply modern machine learning methods, i.e., the baseline T-learner and the state-of-the-art X-learner (Künzel et al. 2019).

Refutation Methods. One of the biggest challenges in causal inference pertains to model evaluation. 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 (Sharma and Kiciman 2020; Cinelli and Hazlett 2020). 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. The Placebo, RCC and RSR tests are bootstrapped to generate confidence intervals and p-values (DiCiccio and Efron 1996). The UCC returns a heatmap of new ATE estimates depending on the strength of unobserved confounding.

Experiments, Results and Discussion

The sowing period lasted from early April to early May 2021, the harvest took place from mid to late September, and the yield per hectare ranged from $1,250$ to $6,960$ kg (Figure 2 of Appendix contains relevant histograms). The dataset consists of 171 fields (51 treated and 120 control). Variables that registered intra-field values (NDVI, NDWI) were averaged at the field-level. For the experiments, we are using the popular doWhy (Sharma and Kiciman 2020) and Causal ML (Chen et al. 2020) Python libraries. Our implementation and dataset are publicly available (Tsoumas 2023).

Applying the back-door criterion on graph G (Figure 2), the following adjustment set of nodes Z was found to be sufficient for identifying the ATE:

$$Z = \{\mathrm{WS}_{\mathrm{MIN,\,MAX}},\ \mathrm{SOC},\ \mathrm{SM},\ \mathrm{G},\ \mathrm{SP}_{\mathrm{SILT,\,CLAY,\,SAND}},\ \mathrm{ABS},\ \mathrm{ADS},\ \mathrm{SV}_{1\text{-}13}\} \quad (4)$$

All variables in Z are numerical, including the one-hot encoded vectors of the categorical $\mathrm{SV}_{1\text{-}13}$ variable of seed variety. AbS and AdS are constant, and thus excluded for estimation purposes. We scale the data by subtracting the mean from each variable and dividing by its standard deviation. The treatment T is binarized, with 1 indicating that a farmer sowed on a favorable day, and 0 indicating the opposite.

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 (Imbens and Rubin 2015). The resulting distribution of propensity scores can be seen at Figure 3. 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.

Table 3 and Figure 4 show the results of the ATE estimation per method, alongside the corresponding 95% confidence intervals and p-values. 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. Both the T-learner and X-learner use a Random Forest for modeling the outcome Y.

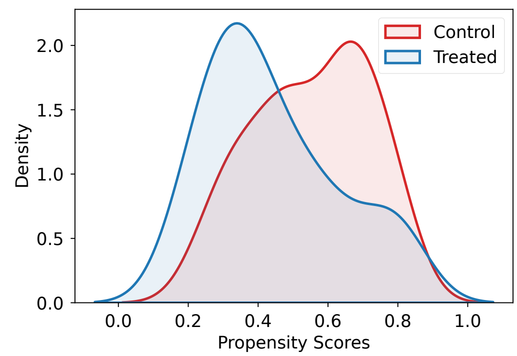

Figure 3: Distribution of propensity scores for the control and treatment group after trimming extreme scores.

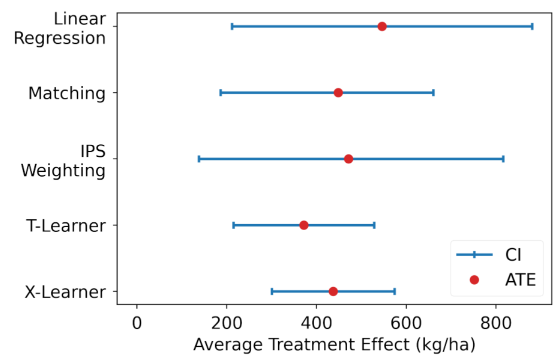

Figure 4: ATE point estimates and 95% confidence intervals for all estimation methods.

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%, depending on the estimation method used.

Of central importance are the refutation tests we run after having estimated the recommendation impact. Table 3 features analytic results for all method / refutation test combinations. All estimation methods are robust against performing the following data manipulations and re-estimating the ATE: randomly permuting the treatment (Placebo test), adding a confounder (RCC test), sampling a subset of data (RSR test) and creating unobserved confounding (UCC test). 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. Confidence intervals and p-values are bootstrapped (1000 iterations).

The results indicate that the recommendation system's advice drove a net increase in yield that was deemed both significant and robust from a statistical perspective. By utilizing

Causal Effect Estimation				Refutations						
				Placebo		RCC		UCC	RSR	
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 3: Results of Average Treatment Effect estimation. Includes point estimates, 95% confidence intervals, and four refutation tests. For the Placebo, RCC and RSR refutations, the new ATE estimate is reported (denoted as Effect*), alongside the respective p-value (< 0.05 indicates a failed test). The UCC column reports the mean ATE estimate of the corresponding heatmap (for full heatmaps and details see Sec. 2 of Appendix). Numbers are in cotton kg/ha, rounded to the nearest integer.

the theory of graphical causal models, the analysis transparently puts forward its assumptions and explicitly incorporates domain knowledge in it. Combined with accurate and performant systems, such analyses can benefit the reliability and adoption of digital agriculture as well as farmers' trust. The provision of information on the actual impact expected from a recommendation system may also enable a cost-benefit analysis on behalf of the farmer, by simply comparing the digital tool cost to the expected yield gain.

Even though the analysis is transparent, it is as good as the causal assumptions it makes and the DAG it develops. Our graph is consistent with agri-environmental knowledge on cotton, however there is always a possibility that bias exists, either due to a missed confounder or due to a missed interaction between observed variables. The robustness checks we performed were all successful; noting that when we add strong unobserved confounding, the UCC test estimates become volatile - an expected behavior to a certain degree.

Given the homogeneous management practices among farmers in our data, we remark that external validity of estimates is low, as results cannot be expected to generalize to other farms that might follow different routines. Nevertheless, it is not uncommon for farmers to follow similar practices in other regions or even entire countries. While the transfer of effect estimates warrants caution, the same does not hold for the proposed framework itself. Given relevant data and knowledge, a graph-based empirical evaluation of an agricultural recommendation system can normally proceed. If consensus among estimation methods in terms of ATE significance is reached, the tool is deemed beneficial; otherwise, more work is required. All in all, this system equips farmers with a provably valuable tool based on cotton knowledge and weather forecasts. It contributes to a successful growing season and lowers the likelihood of farmers resorting to expensive actions, e.g., replanting a field.

For the growing season of 2022, the recommendation system was deployed at national scale and extended to two other crops (maize, sunflower). These new pilot applications will allow us to practically test the external validity of our results across different seasons, crops and locations. Moreover, given the developed causal graph G, the crop growth (CG) variable that is sufficiently captured through its NDVI proxy, mediates the effect of T on Y. The front-door criterion (Pearl 2009) might thus provide an alternative identification method for the ATE, and we plan on exploring it in collaboration with domain experts. Finally, the more growing seasons the recommendation system has seen, the more data are obtained. Going beyond ATE estimation by learning Conditional ATEs and using causal machine learning methods for providing personalized effect estimates is another next step. Due to the rich and well-established domain knowledge, we finally believe that the potential of causal reasoning in agriculture extends far beyond effect identification. Fitting Structural Causal Models and performing counterfactual inference can enable a greater understanding of the farm system and supercharge decision support tools.

Most generally, the essential condition that allowed us to utilize causal inference for empirically evaluating an agricultural recommendation system is the ample, long-established domain knowledge that exists. Decision support systems are being used on multiple fields (Marakas 2003) such as medical decision making (Sutton et al. 2020) or forest and fire management (Segura, Ray, and Maroto 2014; Martell 2015). The aforementioned fields possess accumulated domain knowledge on the interactions a good system exploits; the same way we possess information on environmental conditions related to cotton planting. We thus expect graphical approaches to be valuable for the empirical evaluation of decision support systems of diverse domains.

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 the defence of the farmer, by introducing an AI framework for elaborating on the assumptions, reliability, and impact of a system before discussing service fees.

Acknowledgements

We thank the Orchomenos Agriculture Cooperative for collaboration and data provision, and Corteva Hellas for support. This research was funded by the EU H2020 program through the "EuroGEO Showcases: Applications Powered by Europe" (e-shape) project (Grant agreement ID: 820852). It was also funded by the General Secretariat for Research and Innovation (GSRI, Greece) under the Action ERANETs 2021A [Call ID: 037KE - A/A MIS 4888] (Project Number: T12ERA5-00075) and through the 2019-2020 BiodivERsA joint call for research proposals [BiodivClim ERA-Net COFUND programme, MICROSERVICES project].

References

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

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

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

Ballabio, C.; Panagos, P.; and Monatanarella, L. 2016. Mapping topsoil physical properties at European scale using the LUCAS database. *Geoderma*, 261: 110–123.

Bange, M.; and Milroy, S. 2004. 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.

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

Barbato, L.; Seshadri, S.; and Mulligan, K. 2011. Temperature Forecast and Cotton Seed Planting Interactive Map. https://agriculture.basf.us/crop-protection/products/seeds/cotton-planting-forecasting.html. Accessed: 2021-03-15.

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

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

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

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

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

Casamitjana, M.; Torres-Madroñero, M. C.; Bernal-Riobo, J.; and Varga, D. 2020. 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.

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

Choumos, G.; Koukos, A.; Sitokonstantinou, V.; and Kontoes, C. 2022. Towards Space-to-Ground Data Availability for Agriculture Monitoring. *arXiv preprint arXiv:2205.07721*.

Cinelli, C.; Forney, A.; and Pearl, J. 2020. A Crash Course in Good and Bad Controls. *SSRN Electronic Journal*.

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

Dalezios, N.; Domenikiotis, C.; Loukas, A.; Tzortzios, S.; and Kalaitzidis, C. 2001. Cotton yield estimation based on NOAA AVHRR produced NDVI. *Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere*.

de Brogniez, D.; Ballabio, C.; Stevens, A.; Jones, R.; Montanarella, L.; and van Wesemael, B. 2015. 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.

Deines, J. M.; Wang, S.; and Lobell, D. B. 2019. Satellites reveal a small positive yield effect from conservation tillage across the US Corn Belt. *Environmental Research Letters*, 14(12): 124038.

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

Diggle, P. J.; and Chetwynd, A. G. 2011. Experimental design: agricultural field experiments and clinical trials. In *Statistics and Scientific Method*. Oxford University Press.

Dong, H.; Li, W.; Tang, W.; Li, Z.; Zhang, D.; and Niu, Y. 2006. 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.

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

Fougère, D.; and Jacquemet, N. 2019. Causal inference and impact evaluation. *Economie et Statistique/Economics and Statistics*.

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

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

Ge, M.; Delgado-Battenfeld, C.; and Jannach, D. 2010. Beyond accuracy: evaluating recommender systems by coverage and serendipity. In *Proceedings of the fourth ACM conference on Recommender systems*, 257–260.

Giannarakis, G.; Sitokonstantinou, V.; Lorilla, R. S.; and Kontoes, C. 2022a. Personalizing Sustainable Agriculture with Causal Machine Learning. *arXiv preprint arXiv:2211.03179*.

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

Green, J. M. 1956. Border Effects in Cotton Variety Tests 1. *Agronomy Journal*, 48(3): 116–118.

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

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

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

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

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

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

Kerby, T.; Keeley, M.; and Johnson, S. 1989. Weather and seed quality variables to predict cotton seedling emergence. *Agronomy Journal*, 81(3): 415–419.

Künzel, S. R.; Sekhon, J. S.; Bickel, P. J.; and Yu, B. 2019. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10): 4156–4165.

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

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

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

Luma-Osmani, S.; Ismaili, F.; Raufi, B.; and Zenuni, X. 2020. Causal Reasoning Application in Smart Farming and Ethics: A Systematic Review. *Annals of Emerging Technologies in Computing (AETiC), Print ISSN*, 2516–0281.

Marakas, G. M. 2003. *Decision support systems in the 21st century*, volume 134. Prentice Hall Upper Saddle River.

Martell, D. L. 2015. A review of recent forest and wildland fire management decision support systems research. *Current Forestry Reports*, 1(2): 128–137.

Massmann, A.; Gentine, P.; and Runge, J. 2021. Causal inference for process understanding in Earth sciences. *arXiv preprint arXiv:2105.00912*.

McFadden, J. R.; Rosburg, A.; and Njuki, E. 2022. Information inputs and technical efficiency in midwest corn production: evidence from farmers' use of yield and soil maps. *American Journal of Agricultural Economics*, 104(2): 589–612.

Müller, J. 1974. On sources of measured technical efficiency: the impact of information. *American Journal of Agricultural Economics*, 56(4): 730–738.

Munier, D.; Goodell, P.; Strand, J.; et al. 2004. Accuracy of cotton-planting forecasts assessed in the San Joaquin Valley. *California agriculture*, 58(3): 164–168.

Nanushi, O.; Sitokonstantinou, V.; Tsoumas, I.; and Kontoes, C. 2022. Pest presence prediction using interpretable machine learning. *arXiv preprint arXiv:2205.07723*.

NCEP, N., NWS; and USDOC. 2015. NCEP GFS 0.25 Degree Global Forecast Grids Historical Archive. https://doi.org/10.5065/D65D8PWK. Accessed: 2022-05-20.

Nielsen, R. L.; Thomison, P. R.; Brown, G. A.; Halter, A. L.; Wells, J.; and Wuethrich, K. L. 2002. Delayed planting effects on flowering and grain maturation of dent corn. *Agronomy Journal*.

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

Pearl, J. 2009. *Causality*. Cambridge university press.

Pu, P.; Chen, L.; and Hu, R. 2012. Evaluating recommender systems from the user's perspective: survey of the state of the art. *User Modeling and User-Adapted Interaction*, 22(4): 317–355.

Qian, S. S.; and Harmel, R. D. 2016. 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.

Richards, M. F.; Maphosa, L.; and Preston, A. L. 2022. Impact of Sowing Time on Chickpea (Cicer arietinum L.) Biomass Accumulation and Yield. *Agronomy*, 12(1): 160.

Roberts, M. J.; Schimmelpfennig, D.; Livingston, M. J.; and Ashley, E. 2009. Estimating the value of an early-warning system. *Applied Economic Perspectives and Policy*, 31(2): 303–329.

Rosenbaum, P. R.; and Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1): 41–55.

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

Schimmelpfennig, D.; and Ebel, R. 2016. Sequential adoption and cost savings from precision agriculture. *Journal of Agricultural and Resource Economics*, 97–115.

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

Segura, M.; Ray, D.; and Maroto, C. 2014. Decision support systems for forest management: A comparative analysis and assessment. *Computers and Electronics in Agriculture*, 101: 55–67.

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

Sharma, A.; and Kiciman, E. 2020. DoWhy: An end-to-end library for causal inference. *arXiv preprint arXiv:2011.04216*.

Skamarock, W. C.; Klemp, J. B.; Dudhia, J.; Gill, D. O.; Liu, Z.; Berner, J.; Wang, W.; Powers, J. G.; Duda, M. G.; Barker, D. M.; et al. 2019. A description of the advanced research WRF model version 4. *NCAR: Boulder, CO, USA*, 145: 145.

Snider, J. L.; Pilon, C.; and Virk, G. 2020. Seed Characteristics and Seedling Vigor. In *Cotton Seed and Seedlings*, 9–23. The Cotton Foundation.

Stuart, E. A. 2010. 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.

Sutton, R. T.; Pincock, D.; Baumgart, D. C.; Sadowski, D. C.; Fedorak, R. N.; and Kroeker, K. I. 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, 3(1): 1–10.

Tsoumas, I. 2023. Code&Data. https://github.com/Agri-Hub/AAAI23-Eval-AgriRecommendations. Accessed: 2023-02-10.

UC IPM. 2004. Weather Products: Cotton Planting Forecast. ipm.ucanr.edu/WEATHER/cottonforecast.html. Accessed: 2022-05-10.

Vaessen, J. 2010. 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).

Varco, J. J. 2020. Soil Climatic and Edaphic Effects on Cotton Germination and the Final Stand. In *Cotton Seed and Seedlings*, 43–53. The Cotton Foundation.

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

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

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