Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Environmental Sustainability

Operations research and machine learning to manage risk and optimize production practices in agriculture: good and bad experience

James Cock[1], Daniel Jiménez[1], Hugo Dorado[1,2] and Thomas Oberthür[3]

The potential for operations research (OR) with farmer-supplied data coupled with machine learning (ML) to improve crop management is explored through a series of case studies from developing countries. The information provided by the farmers ranged from solely yield to a description of the management of the crop and some details of the growth environment. The climate or weather conditions of the georeferenced farms were estimated from publicly available databases. Two principal analytical approaches were used. The first benchmarks crop performance against farmers' practices and the second establishes relatively homogenous environmental conditions (HECs) in which the variation in crop response is due to variation in management practices and not to spatiotemporal variation in biophysical factors. Both approaches depend on large amounts of data that can only realistically be obtained from records of on-farm experiences using an OR focus. ML effectively defined HECs for crops with limited prior knowledge on the biophysical factors that influence crop response. The definition of HECs facilitated the identification of either individual farmers who managed their crops well within individual HECs or combinations of management practices well suited to the specific spatiotemporal environmental conditions. This opens the way for farmers to learn better agricultural practices from others in the same HEC. Variation in yield and fertilizer response was associated with variation in the El Niño Southern Oscillation (ENSO) patterns up to 24 months before the harvest: this offers the opportunity for farmers to minimize risk, based on ENSO predictions, even when they have no information on how ENSO influences their weather patterns. Despite concerns about the quality of farmer data, the consistency of the analyses suggests that even relatively crude production data from individual farms analyzed with ML can provide useful guidelines for crop management. Limited variation in management on farmers' fields may limit the ability to identify optimal practices, however, this constraint can be partially obviated by superimposing varied management practices on farmers' fields. The use of OR combined with ML complements, rather than replaces, traditional research methodologies. Furthermore, the approach must be used carefully with emphasis on the dangers of extrapolation to circumstances that are not encompassed by the original datasets.

Addresses
[1] Alliance of Bioversity International and the International Center for Tropical Agriculture (CIAT), Cali, Colombia
[2] Mathematical and Statistical Methods – Biometrics & Plant Production Systems, Wageningen University, Droevendaalsesteeg 4, 6708 PB Wageningen, the Netherlands
[3] Business and Partnership Development, African Plant Nutrition Institute, Lot 660, Hay Moulay Rachid, Ben Guerir, Morocco

Corresponding author: Jiménez, Daniel (d.jimenez@cgiar.org)

Current Opinion in Environmental Sustainability 2023, **62**:101278

This review comes from a themed issue on **Open Issue 2023: Sustainability Science, Digitization and AI**

Edited by **Victor Galaz, Pauline Dube** and **William Solecki**

For complete overview of the section, please refer to the article collection, "Open Issue 2023: Sustainability Science, Digitization and AI"

Available online 3 April 2023

Received: 17 June 2022; Revised: 3 February 2023;
Accepted: 27 February 2023

https://doi.org/10.1016/j.cosust.2023.101278

Introduction

For millennia, farmers observed what occurs in their fields and have adopted those practices they deemed as advantageous [1,2]. In this process, farmers experimented with novel approaches to improve and manage their crops and fields in the face of both risk of short-term variations in production and uncertainty of longer-term sustainability of their production practices. This process was often haphazard and unsystematic, with some improvements to crops unwittingly selected as farmers domesticated them [3]. Nevertheless, much knowledge was consciously accumulated on how to grow crops. Some of this knowledge was recorded in writing (e.g. The Georgics by Virgil), however, much was handed down from generation to generation. The Georgics with references to fallow, and the orations to rain gods in many cultures, indicate farmers' concern with risk and sustainability. Campbell and Overton [2] in their history of Norfolk (UK)

farming over six centuries note that "medieval and early farmers exhibited considerable ingenuity, adaptability and innovation in their husbandry practices." This occurred before the era of modern science-based agriculture: hence, this innovation must be based on their own and shared observations and experiences. More recently, farmers in Britain in the 17th–19th century experimented and shared their knowledge to produce a revolution in agricultural productivity [4,5]. However, farmer–researchers face major problems. Arthur Young, in 18th century England, noted the large number of observations needed to establish the validity of conjectures inferred from isolated field observations [4]. Young's comments reflect the difficulty of drawing conclusions from simple observations due to the large number of variables that affect agricultural production, the heterogeneity of the agricultural landscape, and doubts as to whether a positive result was merely a chance event.

In the mid-19th century, the printing of *Organic Chemistry and Its Applications to Agriculture and Physiology* by Justus Von Liebig in 1840, followed by the founding of experimental stations such as Rothampsted in England in 1843 and the Morrill Land Grant College Act in the United States in 1862, ushered in an era in which scientific principles were applied to agriculture. In the spirit of Von Liebig, researchers identified limiting factors and searched for means to remove them one by one. In 1921, Fisher provided the means to ascertain whether a particular response was likely due to the treatment or solely to chance [6]. The scientific study of agriculture coupled with Fisher's statistical methods led to a factor-by-factor approach to agricultural research in which researchers modify a small number of factors or practices while controlling all the others [7,8]. The factor-by-factor model has proved extremely successful [9]. Nevertheless, the conditions on farmers' fields vary from place to place and over time: they differ substantially from the carefully controlled conditions of experiments. Spatial and temporal variations in crop and livestock production are far greater than the treatment effects on experimental plots and the stability of the treatment response is highly sensitive to the specific conditions of the trial [10]. Thus, for example, sites for experimental plots are normally chosen for their uniformity, often on flat, well-drained terrain with soils meticulously prepared before planting, and all operations done opportunely [11]: this is a far cry from the real world of the farmer. Hence, to extrapolate from well-managed controlled experiments in one specific set of circumstances to the unique conditions faced by the farmer is fraught with pitfalls [8,12,13]. Furthermore, in many crops and circumstances, agricultural decisions are based on limited information with little detailed trial data available for the biophysical and management environment with which the farmers have to contend [11,14,15].

Agriculture, influenced by the scientific factor-by-factor approach, moved toward modifying the cropping environment to suit modern cultivars that became available. This trend is clearly demonstrated by a series of universal cultivars grown across the world, illustrated by the sugarcane variety POJ 2878 that was planted in virtually every sugar-producing country in the world and the widespread planting of IR8, IR5, and IR20 in the lowland tropics at the beginning of the green revolution [16].

While the scientific approach dominated from the late 19th century, a parallel trend arose to mold farming to the environment with farmers being nature's partner [17]. This can be seen with the trend toward a multitude of varieties each specifically suited to individual conditions in wheat in the United States [18]. Furthermore, the increased concerns over environmental degradation, climate change, and heavy dependency on agrochemicals brought to the fore an agroecological approach beginning in the 1930s [17] and organic farming from the 1940s [19,20]. Nevertheless, the capacity of agroecology and organic farming to supply the world with the products it demands is doubted [21]. Furthermore, a large body of researchers questioned whether the unidirectional linear model of scientist to farmer was suitable for all circumstances and instigated the adage of farmers first with farmers actively involved in what became known as participatory research [22,23]. A major feature of the participatory approach was its capacity to ensure that technology was suited to the socioeconomic and environmental milieu of the farmers [24–27]. However, participatory research itself became a source of contention [25,26]: some proclaimed it as a universal panacea, while others judged it as biased, impressionistic, and unreliable [27].

The advent of precision agriculture opened the possibility of complementing the factor-by-factor and the participatory approach with technology tailor-made to the specific conditions faced by farmers in space and time. Precision agriculture, defined as *a management strategy that uses information technologies to bring data from multiple sources to bear on decisions associated with crop production* [28], is strikingly similar to Operations research (OR) that *looks at an organization's operations and uses mathematical or computer models, or other analytical approaches, to find better ways of doing them* [29–31].

OR, largely based on farmers' experiences, is increasingly used to optimize management practices. Two principal analytical approaches have evolved. The first epitomized by CropCheck benchmarks crop performance against farmers' practices [32,33]. The second approach is based on the concept of recommendation domains that are classified into relatively homogeneous groups for which the same recommended technology will be generally applicable [34,35]. In this second

approach, it is assumed that within groups with relatively homogenous environmental conditions (HECs), the variation in crop response is due to variation in management practices and not to spatiotemporal variation in biophysical factors that farmers cannot control.

Both the above approaches, with or without HECs, depend on large amounts of data and sophisticated analysis to determine which combinations of management practices are associated with improved farm performance. Such large datasets can only realistically be obtained from records of on-farm experiences using an OR focus. The quality of these data may not be as good as that obtained from controlled experiments. However, the loss of quality of the data is likely more than compensated by the increased number and range of factors considered within the datasets. Furthermore, errors in the individual data points, if the errors are not systematic, will be counteracted by the greater number of points.

The standard statistical methods developed for a small number of variables and dependent on assumptions of the nature of the response to them are not well adapted to analyzing these datasets. On the other hand, machine learning (ML) is appropriate to handle these often-complex and noisy datasets with no formal structure and little prior knowledge on the relationships between multiple factors [36,37].

In this paper, we describe, in a series of case studies based on previously published research, how OR can be used to provide better risk management and identify more sustainable, agricultural, and production systems by combining data from on-farm operations and other sources with ML. The emphasis is on agriculture in the developing world, where there is frequently limited knowledge on which crops to grow and how best to manage them. Furthermore, as it is necessary to reach a reasonable level of management at the management unit level before looking at within-field management [1], we highlight management at the field level.

Case studies

We chose cases with which we are familiar: unless one has been involved in the data collection and analysis, it is difficult to pinpoint difficulties encountered as these are normally glossed over in formal publications. The cases used are briefly described in Table 1. Furthermore, from each case, only the major lessons learned are presented. For further details of each case, the reader is referred to the original publications.

To structure the presentation, three main themes in which ML can play a major role in improving OR are

Table 1

Details of case studies.

Crop/*country*	Main data sources	Climate data	Soil data	HECs or equivalent	ML methods	Sample size	References
Lulo (*Colombia*)	Farmers, RASTA[a] public databases	Yes	Yes	Yes	Artificial neural networks	254 cropping events	[38]
Andean Blackberry (*Colombia*)	Farmers, RASTA[a] public databases	Yes	Yes	No	Artificial neural networks	488 cropping events	[39]
Coffee (*Colombia*)	Farmers, public databases	Yes	Yes	No	CaNasta Bayesian model	445 farms	[1,40]
Maize (*Colombia*)	Farmers, RASTA[a] public databases	Yes	Yes	No	Random Forest	800+ cropping events	[41]
Rice (*Colombia*)	Fedearroz farm survey data, public databases			No	Random Forests and Conditional Inference Forest	1615 cropping events	[42]
Plantain (*Colombia*)	Farmers' recall, public databases, RASTA[a]	Yes	Yes	Yes	Conditional Random Forest	752 cropping events	[43]
Wheat (*Ethiopia*)	Field trials, public databases	Yes	Yes	Yes	Random Forest	6585 trial observations at 179 sites	[44]
Maize (*Mexico*)	Farm survey, public databases	Yes	Yes	No	Random Forest	4000+ cropping events	[45]
Cacao (*Indonesia*)	Farmers, public databases	No	Partial	No	Bayesian neural networks	73 farms with harvests every 2 weeks over five years	[46,47]

[a] Rapid Soil and Terrain Assessment (https://cgspace.cgiar.org/handle/10568/69682).

highlighted: (i) choice of the right crop; (ii) improved management; and (iii) weather risk management.

The right crop in the right place

Growing the right crop is important as: (i) intensified production through the choice of suitable crops that are well managed reduces pressure on land and limits the destruction of natural habitats; (ii) choosing the right crop well adapted to a given location can minimize the need for soil amendments and pesticides, which may harm the environment.

Lulo (*Solanum quitoense*)

Expert opinion is normally used to define HECs. However, the information on where to grow lulo was limited. Information was collected on the performance of lulo on a series of farms, which were characterized for a range of biophysical conditions. ML algorithms, that make no assumptions on relationships between variables, were used to define the HECs for Lulo [48]. Once HECs were defined, mixed models, guided by expert opinion for categorical variables, explained more than 80% of the total variation in lulo yield, with HEC and farm explaining most of the variation. Higher yields were associated with appropriate environmental conditions (indicated by HEC) and the categorical variable of the individual farm. Although there was no information on individual management practices, those farms with 'good' management, defined by their consistently higher yields in any HEC, could readily be identified.

Andean Blackberry (*Rubus glaucus*)

Little information was available on the response of the Andean Blackberry to variations in the growing conditions and management. Data were collected from farms on the production of blackberries. The growing conditions on the farms were also defined using readily available databases coupled with on-farm observations, particularly for soil and terrain attributes [49]. Artificial neural networks identified geographic areas with higher yields than those that would be expected solely from the environmental conditions [49]. This suggested that the farmers in those geographical areas managed their crops effectively, opening the opportunity for farmers from other areas to learn from them. However, there was not sufficient information to determine which management factors led to the high yields. In addition, the analysis questioned the widely accepted optimum-temperature range for Andean Blackberry.

Coffee quality

The quality of coffee (*Coffea arabica* L.) is closely associated with both the management practices and the environmental conditions. Samples were obtained from coffee growers from georeferenced fields in two geographic regions to identify sites with a suitable climate for production of high-quality coffee and to guide the development of denomination of origin labels. The quality of the distinct samples was determined by cupping [1,40]. Cupping involves preparing coffee in a specific manner and tasting it: a final score is often given with higher scores denoting better quality of the beverage [50]. Initial efforts to use large multivariate models and regression models to establish relationships and associations between cupping score and the growing environment were not successful. The CaNasta Bayesian model [51] identified conditions associated with good coffee quality in Southwest Colombia. Interestingly, it was not possible to pinpoint single factors that defined coffee quality (the factor-by-factor approach did not work) but rather specific combinations of factors were decisive. This illustrates the causal nature of the association of coffee quality with site-specific environmental characteristics.

Crop management

Recommended crop management practices are frequently based on carefully managed experimental plots that give much greater yields than those obtained by farmers in apparently similar conditions [8]. Possibly due to this difference, farmers generally have more confidence in the results obtained by farmers with similar conditions to theirs [1,52].

Managing maize (*Zea mays*) in the North Coast of Colombia

Data were systematically collected from farmers on the performance and management of their maize crops. Data from a range of sources were used to characterize the growing conditions on the individual farms. Estimates of weather, rather than climate data, were recorded for each farm. Data were analyzed using a random forest ML algorithm assisted by expert guidance [53]. Random forests have the advantage that they grow without suffering from overtraining [54]. Five key management practices were associated with improved farmers' yields. In subsequent field trials, adoption of the key practices, in most cases, increased yields substantially. However, on a small number of farms, yields were low despite the use of key practices. In the initial data cleaning, those data points with low or zero yields were eliminated. Initial analysis indicated, surprisingly, that yields were greater on slopes than on flatlands, particularly when rains were heavier than usual. We suspected that the low yields and crop failures were often due to waterlogging on the flatter areas. Thus, by cleaning the data, we had failed to identify drainage or flood control as key practices. This indicates the danger of cleaning data without understanding if the cleaning processes are valid [55]. The analysis also indicated that some recommendations made by extension agents, such as phosphorus application rates, needed to be modified, while those data confirmed many of the other recommendations.

Rice (*Oryza sativa*) in Colombia

The Colombian Rice Growers Federation (Fedearroz) keeps records from surveys on management practices from individual farmers' fields. With yield as the dependent variable, ML algorithms identified factors that influenced yield. The analysis generated much information that confirmed the importance of distinct variables. Thus, for example, both planting date and the solar radiation in the grain-filling period were verified as important yield determinants. Apart from identifying the significance of well-known variables, the analysis highlighted the poor performance of some varieties previously thought to be well adapted to specific sites or regions [42,56].

Plantain (*Musa balbisiana*) in the Andes

Little reliable information was available on plantain production in the coffee-growing region of Colombia. A system for collecting data on-farm operations and characterization of the farms was developed [57]. Initial ML algorithms identified two HECs for plantain. Expert opinion indicated that each of these HECs encompasses an excessively large range of variation of the individual factors. Supervised ML increased the number of HECs for intercropped plantain to five and monocropped plantain to four [57]. A mixed model was used to analyze the variation in productivity. Incorporation of HECs in the mixed model increased the ability of the models to explain the variation in productivity. Further analysis indicated that growers adapt their management practices to the HECs. As the approach used observational information for two plantain cropping systems, the interpretation was challenged by confounded datasets and the lack of records within some of the HECs. This suggests that limited variation in farmer's management practices may limit the ability to discern how the crop responds to distinct management when using an OR approach.

Wheat (*Triticum aestivum*) in Ethiopia

Given the low agronomic effectiveness and poor economic efficiencies of blanket fertilizer applications, there has been a move toward site-specific plant nutrient management [58]. Data on the response of wheat to distinct fertilizer management were compiled from more than six thousand field trials covering a wide range of environmental conditions. Climate data for the individual trials were estimated [44]. ML algorithms identified three principal yield response groups, equivalent to HECs, based on soil and climate, with a small number of principal components defining each HEC [44]. The ML model provided the basis for higher-resolution nutrient recommendations than those currently used in Ethiopia. Hence, with the relevant climatic and edaphic information, the expected response to plant nutrients at a particular site can be determined and site-specific fertilizer applications can be programmed.

Maize in Chiapas Mexico

The International Centre for Maize and Wheat Improvement generated a wealth of data of more than four thousand cropping events, on the management practices used of maize farmers in the state of Chiapas [59]. The farmers grew either hybrid or landrace maize. Landraces are normally open-pollinated varieties that have been selected by farmers for generations and are well adapted to local conditions. Hybrid maize varieties are produced by controlled crosses of genetically distinct parents identified in a formal crop improvement program [60,61]. ML models identified large differences between the landrace and the hybrid maize systems, and the interaction between these systems and management variables [45]. Distinct management guidelines were established for each system. However, despite the large dataset, it was not possible to define advantageous site-specific management practices nor was it possible to relate these practices to specific weather conditions.

A ML algorithm (global best-harmony search) identified management practices that would increase grain yield by 1.8 t ha^{-1} [45]. The model did not indicate how economic yield could be optimized nor did it evaluate possible detrimental effects such as excessive use of inputs that may lead to environmental degradation.

Weather risk management

Farmers are aware of the risks involved in making management decisions with little knowledge of the weather patterns the crop will face once planted. Most recommendations made to farmers by extension workers are optimized for the climate or mean weather conditions of the site. However, the actual weather often varies markedly from that expected from studying the climate.

Cacao (*Theobroma cacao*) in Indonesia

Variation in weather conditions, differences in crop management, and soil conditions were considered the most likely causes of the cacao yield variation between farms and years from observations on fertilizer use by farmers in Indonesia [46]. However, the structure of the dataset was such that standard statistical analysis could not be used to further analyze the causes of the variation. The original dataset was augmented with publicly available data on the El Niño-Southern Oscillation (ENSO), which is known to influence weather patterns, and analyzed with ML algorithms [47]. Analysis indicated that 75% of the on-farm variation in yield was associated with variation in ENSO. Furthermore, the ENSO patterns up to 24 months before the harvest were associated with both the yield and the fertilizer response of the cacao. Thus, farmers may be able to tailor their fertilizer management based on the predicted response and hence reduce risk.

General discussion

Farmers want to know what will work for them on their farm at a particular moment in space and time [15]. The traditional factor-by-factor approach, with controlled experiments, is not capable of appraising the multiple interactions between many factors that influence crop performance under a particular spatiotemporal circumstance [8]. However, the vast number of cropping events managed by farmers often covers a large extent of the variation of those factors that affect crop development. This opens the opportunity to use OR to monitor on-farm cropping events and use the variation that occurs to answer questions that are of interest to farmers. For example, "Which crops are suitable for my farm?" or "How do my results compare to those of other farmers with similar conditions?" or "How can I better manage my crop in the face of uncertain weather patterns?" However, to answer these questions, there are two prerequisites. First, methodologies are required to obtain information on crop performance on individual farmer's fields coupled with data on how the crops are managed and a detailed description of the crop growth environment. Second, novel techniques that handle unstructured datasets and discover unexpected associations and interactions are needed [36,37,62]. Fortunately, options are now available for data collection on farms and techniques to compile large datasets and to handle and analyze these large, often-unstructured, datasets.

In those crops with a substantial knowledge base, many questions can be answered based on the accumulated experience of farmers, when this has been recorded, and many years of research. Nevertheless, in Australia with a well-developed farm industry, many farmers still grow crops in unsuitable areas [63]. Planting an unsuitable crop, or a poorly adapted cultivar, likely makes poor use of natural resources and probably requires the use of costly and often environmentally questionable inputs to sustain production. For both lulo and the Andean Blackberry, with simple georeferentiation of the individual fields, it was possible to infer the climate from publicly available databases. The analysis of both lulo and the Andean Blackberry indicates that collection of relatively crude production data from farms coupled with climate data, followed by ML, can provide useful guidelines on areas suitable for growing these two crops. In these two crops, ML was advantageous as no prior assumptions on the response of the crop to specific environmental factors were required [48,49]. Furthermore, the conditions suitable for the Andean Blackberry were refined and those for lulo confirmed.

The example of suitable areas for producing high-quality coffee clearly demonstrates the validity of data obtained from samples from farmers and the utility of ML to analyze the complex datasets generated: the specific combinations of factors that are associated with good-quality coffee were identified. We note that the analysis of individual factors one by one or with standard statistical procedures would not have uncovered these highly specific combinations of factors [1].

The attempts to define management practices associated with high yields of various crops produced mixed results and highlighted some potential pitfalls in the approach. In the case of maize on the North Coast of Colombia, five key factors for increased production were identified and shown to effectively increase production. However, excessive data cleaning failed to identify the danger of planting maize in low-lying flat areas with poor drainage at times when rainfall is heavy [53,55].

In both the Andean Blackberry and lulo, there was insufficient information on management practices to define which practices were associated with high yields under a given set of conditions. The lack of farm records and information on the management practices used by farmers is a major limitation to the use of OR: effective ML depends on the availability of data from on-farm operations. Owing to the lack of management data *per se*, individual farms were used as a proxy for management level. ML based on production information on farms, characterized for their growing conditions, identified those farmers that managed their crops well. The identification of farms with good management is, of itself, of great significance. First, those farmers who manage their crops well can be used by technical assistance services in farmer-to-farmer technology exchange and sharing programs. Second, once identified, those farmers with good management can be visited, their practices documented, and others can learn from their knowledge and experience.

The Colombian plantain example highlights a complication of using exclusively on-farm data. The growers planted distinct cultivars and used distinct practices in each HEC. Hence, it was not possible to determine whether growers had wisely selected the best cultivar and practices for their conditions as these were confounded with HEC. A further complication occurs in areas where standard practices are widespread: the limited variation in some practices makes it impossible to define the crop response to variation of those practices.

The possible yield increases from adopting improved practices in maize of more than 2.5 t ha^{-1} in Colombia and Mexico and wheat in Ethiopia by close to 1.8 t ha^{-1} are significant for farmers. These are much larger than those reported using a similar approach of farmer-supplied data for soybeans in the North Central United States [52]. The 'yield gap' with US soybeans is small, whereas that for maize and wheat in developing countries is large (www.yieldgap.org). This suggests that collection and analysis of information from commercial

farms are likely to be particularly effective in areas where the yield gap is large.

HECs increased the variation explained by the models in the case of lulo, plantain, and wheat, suggesting that they are a useful tool for separating out management and environmental effects. HECs or their equivalent have been used for this purpose in sugarcane in Colombia [64], oil palm in Indonesia [65], soybean in the United States [52], and wheat and barley in Canada [66]. These crops have all been the subject of considerable research and either expert opinion or standard statistical procedures were used to define the HECs. However, in many crops, especially minor crops, which may be of great significance locally, there is insufficient information to use either expert opinion or standard statistical techniques to define HECs: it is in these cases that farmer-supplied data coupled with ML provide a low-cost opportunity to define HECs for each crop.

Farmers are constantly aware of the risks of changing weather patterns that can either favor or prejudice their crops. As we have noted, most farmers base their crop management on the long-term climate rather than the weather. However, as succinctly put by science fiction author Robert Heinlein "Climate is what we expect, weather is what we get" [67]. Consequently, farmers' practices are often not coherent with the weather. Furthermore, to minimize risk of investing in a crop that later fails, they frequently use suboptimal levels of inputs such as fertilizers. Until recently, weather forecasting on a sufficiently long-term basis for farmers to make crop management decisions was not possible. Now, longer-term weather predictions or outlooks based on deep learning of the dynamics of phenomena such as the ENSO are becoming a reality [68]. However, those farmers working with crops that are underresearched, often have no idea how their crops respond to variations in the weather patterns and hence longer-term weather forecasts are not useful. The case of cacao and the ENSO phenomenon indicates how information on crop performance and its association with the dynamics of a driver of weather can be used to adjust the crop management to the expected patterns without analyzing the weather itself. The complexity of the influence of the ENSO phenomenon on the weather of a particular region and how variation in the weather influences the crop and its response to variations in management are exceedingly complex. It is only through ML that useful associations between the driver of weather, ENSO, and the crop response can be drawn [47]. Nevertheless, care must be taken with these types of relationship as they are location-specific and cannot be extrapolated to other areas or crops [69].

Optimal management practices vary according to the weather conditions [70]. We note that the HEC of a particular field may change as weather patterns change over time. Knowledge of temporal variation in HECs provides the opportunity to better predict which management practices are likely to be optimal and to determine the magnitude of yield gaps for the HEC corresponding to a particular year. This holds not only for the less-researched crops but also for all crops. The case of cacao highlights the potential payoff from modifying practices according to ENSO-induced weather phenomena. Although in 'normal' years fertilizer applications were profitable for most farmers, in years with specific ENSO patterns, fertilizer application for most farmers would be a waste of money.

There is a legitimate concern with the quality of data collected from commercial farms and multiple public sources. Nevertheless, the consistency of the findings reported here with a range of crops confirms the hypothesis that the data quality, with relatively large datasets, was sufficient to establish meaningful conclusions on how to improve crop management and suitability of the site for a particular crop.

Although ML offers great opportunities for optimization, it has several limitations. As noted in the Mexican wheat case, the question of what is to be optimized and the potential negative impact of optimization for one parameter on others must be considered [45]. Furthermore, models based on ML, unlike mechanistic models, cannot be applied to circumstances that differ substantially from the initial training sets [69,71]. In the case of plantain, the small variation of growing conditions and management practices recorded restricted the applicability of the knowledge generated to a minimal area. When there is little variation in management practices, this deficiency can be offset by superimposing distinct management practices on farmers' fields as illustrated by cacao. An opportunity exists to transfer knowledge gained from one HEC to another similar HEC. However, two HECs that are essentially the same in terms of weather patterns and soil traits may differ substantially in their socioeconomic environment and disease and pest pressures. In these cases, it would be dangerous to use knowledge from one HEC in another without considering the differences. At the same time, knowledge of the similarity between HECs can be extremely useful for transferring technology when differences are recognized. Thus, recognition of the similarity of the climatic conditions in Africa and those in Latin America coupled with the knowledge of the lack of parasitoids in Africa was key to the successful introduction of biological control of the cassava mealy bug in Africa [72]. We note that the initial discovery of the cassava mealy bug in the Americas, which led to the identification of biological control agents, was made by Dr Bellotti as he observed cassava fields in Paraguay.

Closing reflection

We use examples from our own work as the case studies. We took this approach as there are inherent dangers in drawing conclusions from published work without having an intimate knowledge of how the information was generated. Nevertheless, in the discussion, we do refer to other work that uses either OR or ML.

The examples we have given show the tremendous potential of using data from farms coupled with information from other sources in an OR approach to understand the response of crops to distinct management under a specific set of conditions. For those crops that have not been extensively researched, or those that are grown over a wide range of conditions, this understanding, and the use of that knowledge to better manage the crops, is only possible in a reasonable time frame and at a reasonable cost through an OR approach coupled with ML. We acknowledge that the approach depends on obtaining reliable data from farmers. The lack of farm records, which is common in developing country agriculture, is a major obstacle to successful implementation. Furthermore, automated field data collection and monitoring tools can complement farm records by providing provide real-time and accurate information on farming operations. Nevertheless, the consistency of the results suggests that even quite crude datasets with little information on farm management practices can be useful. At the same time, we are aware that care must be taken not to extrapolate beyond the range of variation in the original datasets or to apparently similar conditions that differ in parameters not included in the original analysis.

Finally, the approach we present is not a panacea and should not be thought of as a replacement for more traditional research approaches: rather it complements them. The approach described will never produce a new variety or cultivar, however, when new varieties are released, records of their performance on the farms with ML will help farmers decide which cultivar is particularly suited to their farm and how it should be managed.

Author contributions

James Cock, Thomas Oberthür: Conceptualization, Investigation, Writing – review & editing. **Daniel Jiménez:** Conceptualization, Methodology, Investigation, Funding acquisition, Writing – original draft, Writing – review & editing. **Hugo Dorado**: Methodology, Formal analysis, Data curation, Visualization, Writing – review & editing.

Data Availability

Data will be made available on request.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work has been independently peer-reviewed and was carried out with financial support from several programs, initiatives, and projects being implemented by CGIAR and partners, including AgriLAC Resiliente and the CGIAR Research Initiative on Digital Innovation – which aims to identify pathways to accelerate the transformation towards sustainable and inclusive agri-food systems by generating research-based evidence and innovative digital solutions. We would like to thank all funders who supported this research through their contributions to the CGIAR Trust Fund (https://www.cgiar.org/funders/). Special thanks to Corporación Biotec for their support in the case studies on Lulo and Andean Blackberry, and to the International Plant Nutrition Institute (IPNI) for their contribution to the work on cacao. We are also grateful to the many farmers who generously shared their data and knowledge and to the anonymous reviewers whose insightful comments significantly enhanced the quality of the manuscript.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. • Cock J, Oberthür T, Isaacs C, Läderach PR, Palma A, Carbonell J, Watts G, Amaya A, Collet L, Lema G, *et al*.: **Crop management based on field observations: case studies in sugarcane and coffee**. *Agric Syst* 2011, **104**:755-769.
The case studies indicate how machine learning based on observations of sample from farmers provided insights into combinations of factors that influence crop quality that it was not possible to define with a traditional factor by factor approach.

2. Campbell BMS, Overton M: **A new perspective on medieval and early modern agriculture: six centuries of Norfolk farming c.1250-c.1850**. *Past Present* 1993, **141**:38-105.

3. Harlan J: **Crops and Man**. American Society of Agronomy; 1992.

4. Pretty J: **Farmers' extension practice and technology adaptation: agricultural revolution in 17-19th century Britain**. *Agric Hum Values* 1991, **8**:132-148.

5. Overton M: Agricultural Revolution in England;The Transformation of the Agrarian Economy 1500-1850. Cambridge Studies in Historical Geography; 1996:1-6.

6. Stigler S: **Fisher in 1921**. *Stat Sci* 2005, **20**:32-49.

7. O'Neil C: **Arms race: going to college?"**. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Publishing Group; 2016.

8. Marchant B, Rudolph S, Roques S, Kindred D, Gillingham V, Welham S, Coleman C, Sylvester-Bradley R: **Establishing the precision and robustness of farmers' crop experiments**. *Field Crop Res* 2019, **230**:31-45.

9. The University of Reading: **Concepts Underlying the Design of Experiments**. Statistical Services Centre; 2000.

10. Lacoste M, Cook S, McNee M, Gale D, Ingram J, Bellon-Maurel V, MacMillan T, Sylvester-Bradley R, Kindred D, Bramley R, *et al*.: **On-farm experimentation to transform global agriculture**. *Nat Food* 2022, **3**:11-18.

11. Laajaj R, Macours K, Masso C, Thuita M, Vanlauwe B: **Reconciling yield gains in agronomic trials with returns under African smallholder conditions**. *Sci Rep* 2020, **10**:1-15.

12. Abate GT, Bernard T, de Brauw A, Minot N: **The impact of the use of new technologies on farmers' wheat yield in Ethiopia: evidence from a randomized control trial**. *Agric Econ* 2018, **49**:409-421.

13. Edreira JIR, Cassman KG, Hochman Z, Van Ittersum MK, Van Bussel L: **Beyond the Plot: Technology Extrapolation Domains for Scaling Out Agronomic Science**; 2018.

14. Whitsed R, Corner R, Cook S: **A model to predict ordinal suitability using sparse and uncertain data**. *Appl Geogr* 2012, **32**:401-408.

15. Coe RIC, Njoloma J, Sinclair F: **To control or not to control: how do we learn more about how agronomic innovations perform on farms?** *Exp Agric* 2019, **55**:303-309.

16. Evenson R: **Cycles in research productivity in sugarcane, wheat, and rice**. In *Proceedings of the Conference on Resource Allocation and Productivity in National and International Agricultural Research*. Edited by Dalrymple DG, Evenson RE, Kislev Y. Airlie House,Va. Chapter 8: 20; 1975.

17. Merrill MC: **Eco-agriculture: a review of its history and philosophy**. *Biol Agric Hortic* 1983, **1**:181-210.

18. Chai Y, Pardey PG, Silverstein KAT: **Scientific selection: a century of increasing crop varietal diversity in US wheat**. *Proc Natl Acad Sci* 2022, **119**:e2210773119.

19. Behera KK, Alam A, Vats S: **Organic Farming History and Techniques**; 2012.

20. Clarke R: **Organic Farming: History, Timeline, and Impact** . ⟨https//www.treehugger.com/organic-farming-history-timeline-and-impact-5189324⟩ (Accesed Jan 31, 2023).

21. Connor DJ: **What is the real productivity of organic farming systems?** *Outlook Agric* 2021, **50**:125-129.

22. Caron P, Biénabe E, Hainzelin E: **Making transition towards ecological intensification of agriculture a reality: the gaps in and the role of scientific knowledge**. *Curr Opin Environ Sustain* 2014, **8**:44-52.

23. Chambers R, Pacey A, Thrupp L: **Farmer First: Farmer Innovation and Agricultural Research**. Intermediate Technology Publications; 1989.

24. Ashby J, Lilja N: **Participatory research: does it work? Evidence from participatory plant breeding**. In *Proceedings of the 4th International Crop Science Congress. "New directions for a diverse planet."*; 2004:1–14.

25. Bentley JW: **Facts, fantasies, and failures of farmer participatory research**. *Agric Hum Values* 1994, **11**:140-150.

26. Bentley JW: **Folk experiments**. *Agric Hum Values* 2006, **23**:451-462.

27. Cornwall A, Jewkes R: **What is participatory research**. *Soc Sci Med* 1995, **41**:1667-1676.

28. National Research Council: **Precision Agriculture in the 21st Century**. National Academies Press; 1997.

29. Hillier FS, Lieberman GJ: **Introduction to Operations Research**. McGraw-Hill Education; 2010.

30. INFORMS: **The Institute for Operations Research and the Management Sciences. What are Operations Research and Analytics? What is Management Science?**; 2022 ⟨https//www.informs.org/Resource-Center/INFORMS-Student-Union/FAQs-About-OR-Analytics⟩ (Accesed Jun 16, 2022).

31. UIA: **Developing Operational Research. The Encyclopedia of World Problems and Human Potential**; 2022. ⟨http//encyclopedia.uia.org/en/strategy/209427⟩ (Accesed Jun 15, 2022).

32. Lacy J: **Cropcheck: farmer benchmarking participatory model to improve productivity**. *Agric Syst* 2011, **104**:562-571.

33. Araya F, Acevedo R, Cabello MC, Jaramillo C, Gonzalez I, Toro M: **CropCheck Chile:Sistema de Extension para el Sector AgroAlimentario**. Fundación Chile en el Programa Cropcheck; 2010.

34. Gauch HG, Zobel RW: **Identifying mega-environments and targeting genotypes**. *Crop Sci* 1997, **37**:311-326.

35. Byerlee D, Biggs S, Collinson M, Harrington L, Winkelmann D: **On-Farm Reserach to Develop Technologies Appropriate to Farmers**. Int Assoc Agric Econ 1981 Occas Pap Ser No 2; 1981.

36. Wolfert S, Ge L, Verdouw C, Bogaardt MJ: **Big data in smart farming ? A review**. *Agric Syst* 2017, **153**:69-80.

37. Kitchin R, Big Data: **new epistemologies and paradigm shifts**. *Big Data Soc* 2014, **1**:1-12.

38. Jiménez D, Cock J, Jarvis A, Garcia J, Satizábal HF, Damme P Van, Pérez-Uribe A, Barreto-Sanz M: **Interpretation of commercial production information: A case study of lulo (*Solanum quitoense*), an under-researched Andean fruit**. *Agric Syst* 2011, **104**:258-270.

39. Jiménez D, Cock J, Satizábal HF, Barreto SMA, Pérez-Uribe A, Jarvis A, Van Damme P: **Analysis of Andean blackberry (*Rubus glaucus*) production models obtained by means of artificial neural networks exploiting information collected by small-scale growers in Colombia and publicly available meteorological data**. *Comput Electron Agric* 2009, **69**:198-208.

40. Oberthür T, Läderach P, Posada H, Fisher MJ, Samper LF, Illera J, Collet L, Moreno E, Alarcón R, Villegas A, *et al*.: **Regional relationships between inherent coffee quality and growing environment for denomination of origin labels in Nariño and Cauca, Colombia**. *Food Policy* 2011, **36**:783-794.

41. Jiménez D, Delerce S, Dorado H, Cock J, Armando L, Agamez A, Jarvis A: **A scalable scheme to implement data-driven agriculture for small-scale farmers**. *Glob Food Sec* 2019, **23**:256-266.

42. Delerce S, Dorado H, Grillon A, Rebolledo MC, Prager SD, Patiño VH, Garcés Varón G, Jiménez D: **Assessing weather-yield relationships in rice at local scale using data mining approaches**. *PLoS One* 2016, **11**:e0161620.

43. Jiménez D, Dorado H, Cock J, Prager SD, Delerce S, Grillon A, Andrade Bejarano M, Benavides H, Jarvis A: **From Observation to Information: Data-Driven Understanding of on Farm Yield Variation**. *PLoS One* 2016, **11**.

44. Abera W, Tamene L, Tesfaye K, Jiménez D, Dorado H, Erkossa T, Kihara J, Ahmed JS, Amede T, Ramirez-Villegas J: **A data-mining approach for developing site-specific fertilizer response functions across the wheat-growing environments in Ethiopia**. *Exp Agric* 2022, **58**:1-16.

45. • Dorado H, Delerce S, Jimenez D, Cobos C: **Finding optimal farming practices to increase crop yield through global-best harmony search and predictive models, a data-driven approach**. In *Advances in Computational Intelligence. Lecture Notes in Computer Science*. Edited by Batyrshin I, Martínez-Villaseñor M, Ponce Espinosa H. Springer; 2018:15-29.

Machine learning algorithms are used to demonstrate how relatively noisy, crude farmer generated information can be used to predict optimal farming practices and provide guidelines on how farmers can substantially increase yields.

46. Hoffmann MP, Cock J, Samson M, Janetski N, Janetski K, Rötter RP, Fisher M, Oberthür T: **Fertilizer management in smallholder cocoa farms of Indonesia under variable climate and market prices**. *Agric Syst* 2020, **178**:102759.

47. •• Chapman R, Cock J, Samson M, Janetski N, Janetski K, Gusyana D, Dutta S, Oberthür T: **Crop response to El Niño-Southern Oscillation related weather variation to help farmers manage their crops**. *Sci Rep* 2021, **11**:1-8.

This research clearly indicates relationships established by machine learning it is possible to provide useful information to farmers on how to manage their crops in the face of variations in weather associated with the ENSO phenomenon.

48. Jiménez D, Cock J, Jarvis A, Garcia J, Satizábal HF, Van Damme P, Pérez-Uribe A, Barreto-Sanz MA: **Interpretation of commercial**

production information: a case study of lulo (*Solanum quitoense*), an under-researched Andean fruit. *Agric Syst* 2011, **104**:258-270.

49. Jiménez D, Cock J, Satizábal HF, Barreto SMA, Pérez-Uribe A, Jarvis A, Van Damme P: **Analysis of Andean blackberry (*Rubus glaucus*) production models obtained by means of artificial neural networks exploiting information collected by small-scale growers in Colombia and publicly available meteorological data**. *Comput Electron Agric* 2009, **69**:198-208.

50. SCAA Specialty Coffee Association of America: **SCAA Protocols: Cupping Specialty Coffee**; 2015. ⟨https//www.scaa.org/PDF/resources/cupping-protocols.pdf⟩ (Accesed Jan 31, 2023).

51. Whitsed R, Corner R, Cook S: **A model to predict ordinal suitability using sparse and uncertain data**. *Appl Geogr* 2012, **32**:401-408.

52. • Andrade JF, Mourtzinis S, Rattalino Edreira JI, Conley S, Gaska J, Kandel HJ, Lindsey LE, Naeve S, Nelson S, Singh M, *et al*.: **Field validation of a farmer-data approach to close soybean yield gaps in the US North Central Region**. *Agric Syst* 2022, **200**:103434.

This paper clearly demonstrates how data obtained from farmers when well analyzed with machine learning can provide guidelines on how yields can be improved.

53. Jiménez D, Delerce S, Dorado H, Cock J, Armando L, Agamez A, Jarvis A: **A scalable scheme to implement data-driven agriculture for small-scale farmers**. *Glob Food Secur* 2019, **23**:256-266.

54. Breiman L: **Random forest**. *Mach Learn* 1999, **45**:1-35.

55. •• Galaz V, Centeno MA, Callahan PW, Causevic A, Patterson T, Brass I, Baum S, Farber D, Fischer J, Garcia D, *et al*.: **Artificial intelligence, systemic risks**. *and sustainability, Technol Soc* 2021,101741.

The inherent risks of using machine learning, especially those related to risk and longer term sustainability issues are debated and provide a cautionary note to the use of these novel technologies.

56. Young A, Verhulst S: **Aclimate Colombia: Open Data to Improve Agricultural Resiliency**. Open Data's Impact ⟨http//odimpact.org/files/case-aclimate-colombia.pdf⟩; 2017.

57. •• Jiménez D, Dorado H, Cock J, Prager SD, Delerce S, Grillon A, Andrade Bejarano M, Benavides H, Jarvis A: **From observation to information: data-driven understanding of on farm yield variation**. *PLoS One* 2016, **11**:e0150015.

This article demonstrates how relatively crude data sets obtained from farmers it is possible to gain valuable insights into how management can be improved. At the same time it clearly illustrates some of the limitations of using solely farmers` experiences to improve crop management.

58. Dobermann A, Witt C, Dawe D: **Performance of site-specific nutrient management in intensive rice cropping systems of Asia**. *Better Crop Int* 2022, **16**:25-30.

59. Jimenez D, Ramirez J, Gardeazabal A, Lougee R: **Transforming Food Production and Supply with OR Analytics**; 2021.

60. Wies G, Navarrete-Segueda A, Ceccon E, Larsen J, Martinez-Ramos M: **What drives management decisions and grain yield variability in Mesoamerican maize cropping systems? Evidence from small-scale farmers in southern Mexico**. *Agric Syst* 2022, **198**:103370.

61. Trevisan RG, Martin NF, Fonteyne S, Verhulst N, Dorado Betancourt HA, Jimenez D, Gardeazabal A: **Multiyear maize management dataset collected in Chiapas, Mexico**. *Data Brief* 2022, **40**:107837.

62. Bronson K, Knezevic I: **Big Data in food and agriculture**. *Big Data Soc* 2016, **3**:1-5.

63. Guerin LJ, Guerin TF: **Constraints to the adoption of innovations in agricultural research and environmental management: a review**. *Aust J Exp Agric* 1994, **34**:549-571.

64. Palma AE, Luna CA, Carbonell J, BO: **A new method to classify natural conditions for sugarcane planting in a region in Colombia**. In *Proc Inter Am Sugar Cane Semin Crop Prod Mech*. Miami, Florida, USA; 9–11 September 1998:52–62 (En), 318–328 (Es).

65. Cock J, Kam SP, Cook S, Donough C, Lim YL, Jines-Leon A, Lim CH, Primananda S, Yen BT, Mohanaraj SN, *et al*.: **Learning from commercial crop performance: oil palm yield response to management under well-defined growing conditions**. *Agric Syst* 2016, **149**:99-111.

66. Lu W, Atkinson DE, Newlands NK: **ENSO climate risk: predicting crop yield variability and coherence using cluster-based PCA**. *Model Earth Syst Environ* 2017, **3**:1343-1359.

67. Heinlein RA: **Time Enough for Love**. Putnam; 1973.

68. Salman AG, Kanigoro B, Heryadi Y: **Weather forecasting using deep learning techniques**. In *Proceedings of the 2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*; 2015:281–285.

69. Meyer H, Pebesma E: **Predicting into unknown space? Estimating the area of applicability of spatial prediction models**. *Methods Ecol Evol* 2021, **12**:1620-1633.

70. Barrios-Perez C, Okada K, Varón GG, Ramirez-Villegas J, Rebolledo MC, Prager SD: **How does El Niño Southern Oscillation affect rice-producing environments in central Colombia?** *Agric Meteorol* 2021, **306**:108443.

71. Hastie T, Tibshirani R, Friedman J: **The Elements of Statistical Learning: Data Mining, Inference and Prediction**. Springer; 2009.

72. Neuenschwander P: **Biological control of the Cassava Mealybug in Africa: a review**. *Biol Control* 2001, **21**:214-229.