Multi-Objective Optimization for Large-scale Allocation of Soybean Crops

Mathilde Chen
UMR 518 MIA-PS
INRAE, Université Paris-Saclay
Palaiseau, France
CIRAD, UMR PHIM
F-34398 Montpellier, France
PHIM, Univ Montpellier, CIRAD, INRAE, Institut Agro, IRD
Montpellier, France
mathilde.chen@cirad.fr

David Makowski
UMR 518 MIA-PS
INRAE, Université Paris-Saclay
Palaiseau, France
david.makowski@inrae.fr

Alberto Tonda
UMR 518 MIA-PS
INRAE, Université Paris-Saclay
Palaiseau, France
Institut des Systèmes Complexes
Paris-Ile-de-France
Paris, France
alberto.tonda@inrae.fr

ABSTRACT

The optimal allocation of crops to different parcels of land is a problem of paramount practical importance, not only to improve production, but also to address the challenges posed by climate change. However, this optimization problem is inherently complex, characterized by a vast search space that renders traditional optimization techniques impractical without oversimplified assumptions. Compounding this challenge, climate change introduces conflicting objectives, as solutions aiming to just maximize total yield may be more susceptible to extreme weather events, and thus obtain more unpredictable year-by-year outcomes. In order to tackle this complex optimization problem, we propose a multi-objective approach, simultaneously maximizing the overall yield, minimizing the year-on-year yield variance, and minimizing the total cultivated surface. The approach exploits an established multi-objective evolutionary algorithm, and employs a machine learning model able to predict yield from weather and soil conditions, trained on historical data, making it possible to tackle allocation problems of large size. An experimental evaluation focusing on the allocation of soybean crops in the European continent for the years 2000-2023 shows that the proposed methodology is able to identify different trade-offs between the conflicting objectives, that an expert analysis later reveals to be realistic and meaningful for driving stakeholder decisions.

CCS CONCEPTS

• **Applied computing → Environmental sciences**; **Multi-criterion optimization and decision-making**.

KEYWORDS

crop allocation, crop yield forecasting, machine learning, multi-objective optimization

GECCO '24, July 14–18, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-0494-9/24/07…$15.00
https://doi.org/10.1145/3638529.3654026

ACM Reference Format:
Mathilde Chen, David Makowski, and Alberto Tonda. 2024. Multi-Objective Optimization for Large-scale Allocation of Soybean Crops. In *Genetic and Evolutionary Computation Conference (GECCO '24), July 14–18, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3638529.3654026

1 INTRODUCTION

The growth of crops for human and animal consumption is heavily influenced by location-dependent conditions, such as quality of the soil, weather patterns, or rainfall. The correct allocation of specific types of crops to the most appropriate areas for maximizing yield is a problem of clear practical importance, not only for commercial purposes, but also to potentially mitigate the growing negative impact of climate change [36]. The increased appearance of extreme weather had strong repercussions on the variance of crop yields, introducing high levels of production instability between years and a previously unseen unreliability in the food supply. Soybean production in particular is a major source of concern; it is the world's main source of protein for animal feed, but its production is concentrated in South America, where it is the source of major environmental impacts, notably due to deforestation. As Europe imports large quantities of soybean from South America, a substantial number of experts advise to relocate soybean production to Europe, raising the question of the feasibility and sustainability of growing this crop in the old continent. The use of optimization methods would make it possible to determine where to grow soybeans in Europe in order to obtain production that is both high on average and stable over time, while minimizing the area cultivated to allow other types of agricultural production.

Several approaches have been proposed for the optimization of crop allocations, typically resorting to classical optimization techniques from operational research [5]. However, the application of such algorithms has important limitations, either requiring an unrealistic linearization of the objective functions, or severely limiting the number of variables considered. Furthermore, previous approaches only take into account a single objective function, typically maximizing yield, while given the rising importance of climate change, it would be extremely valuable to also take into account other objectives, such as minimizing between-year variance in yield and minimizing land occupation.

In this work, we propose to apply a multi-objective evolutionary algorithm (MOEA) to the problem of crop allocation. Relying upon a machine learning model trained on historical data for the production of soybean, we generate time series of crop yields values from 2000 to 2023 in each cell of a spatial grid covering croplands in Europe. Based on this dataset, a multi-objective optimization is performed to generate a set of land use allocations approximating a Pareto front, combining three objectives: maximizing mean yearly soybean production, minimizing inter-year production variance, and minimizing the total cultivated soybean area. After obtaining a set of non-dominated candidate solutions in the space of the objectives, we perform an expert analysis of four scenarios, representing different trade-offs. These candidate solutions offer new perspectives to improve the level of food self-sufficiency in Europe.

The rest of the paper is organized as follows. Section 2 introduces the scope of this work. The proposed approach is described in Section 3. The experimental evaluation, presenting a case study on the allocation of soybean crops in the European continent, is detailed in Section 4. Finally, Section 5 delineates some conclusions and outlines future works.

2 BACKGROUND

This section summarizes the minimal information necessary to introduce the scope of the work.

2.1 Forecasting of crop yields

Crop yield predictions at large scales play a significant role for commodity trading and implementation of food security policies [14]. Yield predictions are essential to prevent food shortages in case of harvest losses or failures, and are frequently used to make projections about the impact of climate change on crops [37].

Historically, crop yield predictions were obtained with process-based models which intended to integrate biophysical mechanisms underlying plant growth and development [2]. However, these models are not always reliable, due to the major issues related to the estimation of their numerous parameters [26]. Previous research showed that these models may provide inaccurate forecasts [28, 32] which can lead to contradictory conclusions depending on the type of model used [29].

In parallel, statistical linear regression models are simpler and less costly to implement compared to process-based models [2]. For these reasons, they were frequently used in crop yield prediction as well [4, 25]. Standard regression models do not perform well when their inputs are highly correlated and do not always capture all the possible interactions between predictors [4]. Because of their greater flexibility, more modern machine learning algorithms are now commonly used to forecast yields of crops from climatic predictors [2, 40] and often show better performances compared to traditional statistical methods such as linear regression, in particular to predict soybean yields [3, 18, 24].

2.2 Optimization of crop allocation

Globally, the demand for food is expected to increase by 35% to 56% between 2010 and 2050 [39], and so far the actual trend has closely followed the predictions. In the last decades, crop yield (the production by unit of cultivated land) generally increased, due to genetic advancements and better crop management practices; the intensive use of fertilizers and pesticides, as well as improvements in mechanization of agriculture also had a strong positive impact on yield [1], albeit with negative repercussions. Some studies suggest that there is still some potential to intensify production on existing land [41], but this could also lead to more adverse environmental effects. In addition, there is a considerable amount of uncertainty on potential crop yield changes under climate change, whose effect are hard to model and predict [32].

A different approach to increasing yield is to allocate larger proportions of cropland on remaining arable lands, while preserving natural ecosystems (such as forest, permanent grasslands or wetlands) and without increasing global cropland areas. Several studies used machine learning or statistical models to make projections in new areas where the crop is not yet grown, and examine the consequences of different choices (e.g. [21, 38]). The same approach is used to simulate the yields in various climate change scenarios, to investigate the suitability of production in the context of global warming (e.g. [15, 18]).

In these studies, the geographical allocation of crops is generally based on the average or maximum productivity of the examined regions. An important limit of this approach is that it neglects the instability of the production, i.e., the year-by-year variability of the productivity due to contrasted weather conditions. Thus, the variability of production is a second aspect to be simultaneously considered when allocating crops, especially in the context of climate change. Only one study quantified the trade-off between these two dimensions for maize, rice, soybean, and wheat using quadratic optimization [5]. Although useful, this approach only provides estimates at a low geographical resolution (i.e. at a continental scale) and does not provide information on how to allocate crop lands locally. Thus, an optimized allocation of crops that simultaneously maximizes both overall amount and stability of agricultural production while limiting the total surface dedicated to production is needed to identify, at a finer scale, suitable areas to produce crops.

2.3 Multi-objective optimization in agriculture

Multi-objective optimization (MOO) is a branch of optimization dealing with problems featuring multiple conflicting objectives [9, 12]. Differently from single-objective optimization, where the goal is to find a single solution with the best possible value of the target cost function, the aim of MOO algorithms is to find a non-dominated front of candidate solutions, each one representing a different trade-off between the multiple objectives. More formally, in a minimization problem, a candidate solution I_i is considered non-dominated if there is no I_k such that:

$$F_j(I_k) \leq F_j(I_i) \quad \forall j \tag{1}$$

where F_j are the fitness functions for the j different objectives, and $I_{k \neq i}$ are all other candidate solutions considered as a comparison, usually the ones inside the population at the current generation, plus the ones stored in a dedicated archive.

Multi-objective evolutionary algorithms (MOEAs) currently represent the state of the art in the MOO domain. While the most recent research in the field is exploring complex problems with 10 or more objectives [10, 20], for applications with up to three objectives the most established algorithm is arguably the Non-Sorted Genetic Algorithm II (NSGA-II) [11], which is used in this work.

Not surprisingly, multi-objective approaches are popular choices for framing optimization problems in agriculture, where each candidate solution is often a trade-off between multiple conflicting needs. For example, in [8], the authors use NSGA-II to find optimal strategies for the management of rice fields, finding compromises between irrigation events, use of rainfall, and yield. [27] proposes the use of different crop and water use models to solve a multi-objective optimization problem of crop and irrigation allocation at the level of a single farm, solving sub-problems using a Particle Swarm Optimization algorithm and then calling the MOEA Ev-MOGA [30] to find non-dominated solutions for the global problem. Finding optimal combinations of crops in greenhouses is the subject of [34], where the multi-objective problem is framed as maximizing global yield while at the same time minimizing water use, employing both NSGA-II and the msPESA [17] algorithms. The work presented in [23] proposes to take into account multiple factors to evaluate agricultural policies in the Miandarband Iranian region, from production using crop models, to environmental impacts using Life Cycle Assessment tools, to societal impacts using expert models obtained interviewing local farmers, a three-objective optimization approach that was also applied to the evaluation of the production in insect farms [31]. In [22], the authors aim to find the mix of crops to plant over the Telangana Indian region that maximizes economic returns and minimizes the use of fertilizers, using a novel ad-hoc stochastic multiobjective algorithm; however, the optimization problem is framed so that the exact locations of the crops in a candidate solution are not considered, and only the total area dedicated to each species of plant is taken into account. To the best of the authors' knowledge, multi-objective optimization has never been used to address the problem of crop allocation over vast geographical areas, for example at the level of whole countries or continents, using fine-grained geographical locations.

3 PROPOSED APPROACH

In this work, we propose to use a MOEA to perform a multi-objective optimization of geographical allocation of soybean. Given a map of a large-scale territory, for example a country or a continent, divided into cells with a spatial grid, a candidate solution represents the amount of land allocated to soybean in each grid cell. The objectives include maximizing the total production over a given number of years; minimizing the between-year variability; and minimizing the total amount of land allocated. As all objective functions require a prediction of the amount of soybeans produced given specific weather conditions, we employ a machine learning model trained on historical data to provide time series of yield forecast in each grid cell.

3.1 Machine learning forecast of yield

We train a random forest model [6] to predict yearly yield of soybean in Europe. The choice of the model was guided by previous works, comparing and ranking various machine learning and statistical models [18], exploiting different ways to aggregate climate features to predict soybean productivity [7].

The model is trained on historical soybean yield grided data of resolution of 0.5° and covering period from 1981 to 2016 [19] on 2, 626 cells spread in major producing areas (i.e. Argentina, Brazil, Canada, China, India, Italy, and United-States). To improve model prediction accuracy, additional cells located in geographical areas unsuitable for crop production were randomly selected. The number of these grid-cells is determined so that they represented 20% of the full training dataset. In total, 3, 286 grid-cells constitute the training data (Figure 1).

For each site, yield data are detrended using splines to avoid any confusion with technological progress due to improved cultivars and technological progress. Using the ERA5-Land database [33], we derive monthly averages of six climate variables for each combination of sites and year (hereafter referred to as "site-year"). The variables considered in this study are: minimum and maximum temperatures (both in°C), precipitation (in mm), solar radiation (in MJ), reference evapotranspiration (mm/day), and vapor pressure deficit (kPA) during soybean growing season. Growing of soybean was defined country-by-country according to the crop calendars provided by the Agricultural Market Information System[1].

Several models based on climate predictors are tested. First, a random forest model including monthly averages of climate data is considered (*avg.m*). Second, we fit a random forest model based on seasonal averages (i.e., average over the whole growing season of soybean) (*avg.s*). Third, we apply principal component analysis to capture the largest variation in monthly averages of climate data, and we use the scores associated with the two ($pca.m.2$) or three ($pca.m.3$) principal components as climate predictors of yield in the last two random forest models considered. Each model also includes irrigation fraction (in %) as a predictor.

For each model, performance in predicting yearly yield of soybean from climate and irrigation predictors was assessed using the Root of the Mean Squared Error (RMSE, in tons/hectare), computed as:

$$RMSE = \sqrt{\frac{1}{X \cdot Y} \sum_{x=1}^{X} \sum_{y=1}^{Y} (p_{xy} - o_{xy})^2} \tag{2}$$

where p_{xy} and o_{xy} represent the predicted and observed yields in the grid-cell x and the year y; X and Y are the total number of grid-cells and years, respectively. The lower the RMSE, the lower the difference between predictions and observations, which corresponds to a better performance of the model.

A second performance metric used is the R^2 (unitless), computed as:

$$R^2 = 1 - \frac{\sum_{x=1}^{X} \sum_{y=1}^{Y} (p_{xy} - o_{xy})^2}{\sum_{x=1}^{X} \sum_{y=1}^{Y} (o_{xy} - \bar{o})^2} \tag{3}$$

where $\bar{o}$ is the mean value of observed yield over all years y and grid-cells x.

For R^2, a value of 1.0 corresponds to a perfect match of predictions to observed data, a value of 0.0 indicates that predictions are

[1]Available at: http://www.amis-outlook.org/amis-about/calendars/soybeancal/en/

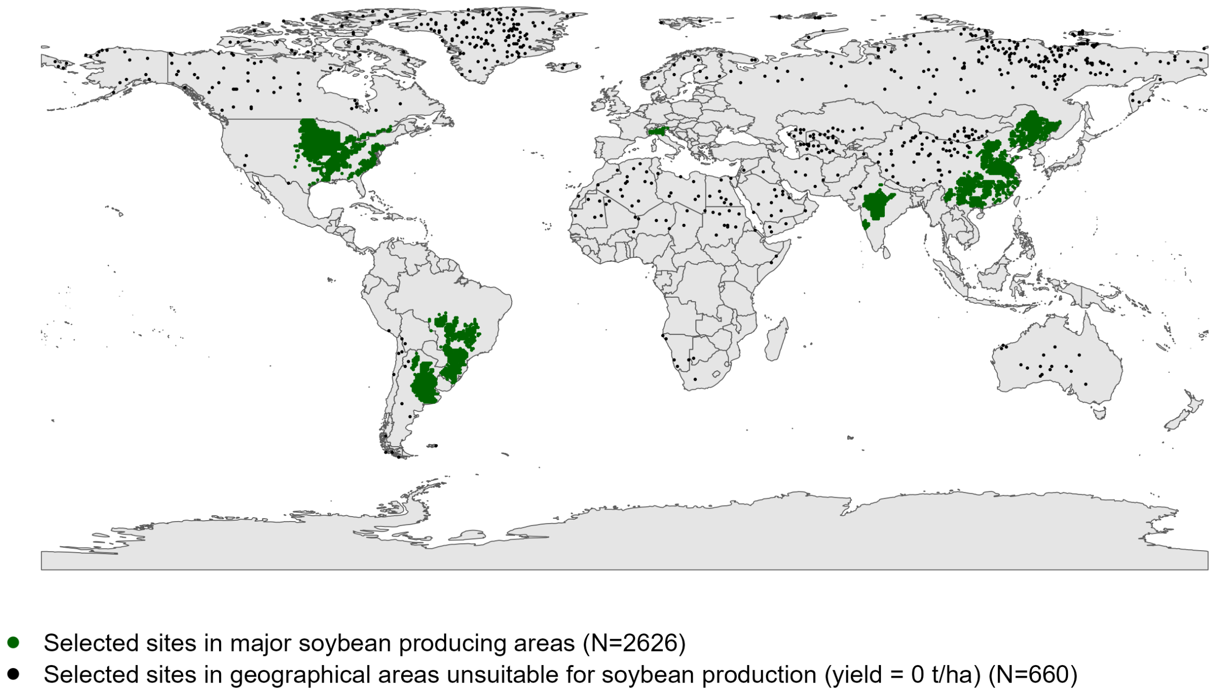

Figure 1: Map of the grid-cells where data was collected for the training set of the machine learning algorithm, later used to predict yield in the European area. The cells include both historical soybean fields with high yield (green dots), and areas that are considered unsuitable for soybean cultivation (black dots).

as accurate as the mean of observed data. In contrast, a value of R^2 lower than 0.0 occurs when the observed mean is a better predictor than the tested model.

R^2 and RMSE were computed for each model following two separate cross-validation procedures:

- First, a year-by-year cross-validation was performed, to assess model's capability in predicting yields in a new year, which is not included in the training dataset (temporal extrapolation).
- Secondly, a group-wise cross-validation was employed, wherein 10 groups of randomly selected sites were used to evaluate the model's ability to forecast yields in novel geographic regions not encompassed within the training dataset (spatial extrapolation).

Models' performances estimated from each cross-validation procedure and the mean over both procedures are presented in Table 1. Values of R^2 equal or higher than 0.90 highlight good performances of all tested models. RMSE values were also comparable across models. The model presenting the best performance on average (i.e., the highest R^2 and the lowest RMSE values) was the random forest using the scores associated with the two first principal components (*pca.m.2*). On average, this model shows a mean R^2 of 0.93 and a RMSE of 0.38 tons/hectare.

	Cross-validation results					
	year-by-year		**group-by-group of sites**		**average**	
	R2	**RMSE**	**R2**	**RMSE**	**R2**	**RMSE**
pca.m.3	0.91	0.41	0.94	0.34	0.92	0.38
pca.m.2	0.92	0.39	0.93	0.36	0.93	0.38
avg.m	0.90	0.44	0.94	0.33	0.92	0.38
avg.s	0.91	0.41	0.90	0.44	0.91	0.42

Table 1: Comparative performance of random forest models to predict soybean yield, using different features. Higher values of R2 correspond to higher predictive performance. Lower values of root mean square error (RMSE) correspond to higher predictive performance. Details on the cross-validation procedures used to compute R^2 and RMSE values can be found in the main text. Abbreviations: *pca.m.3*: **model based on the scores associated with the three first components derived from monthly averages of climate data;** *pca.m.2*: **model based on the scores associated with the two first components derived from monthly averages of climate data;** *avg.m*: **model based on monthly averages of climate data;** *avg.s*: **model based on seasonal averages of climate data.**

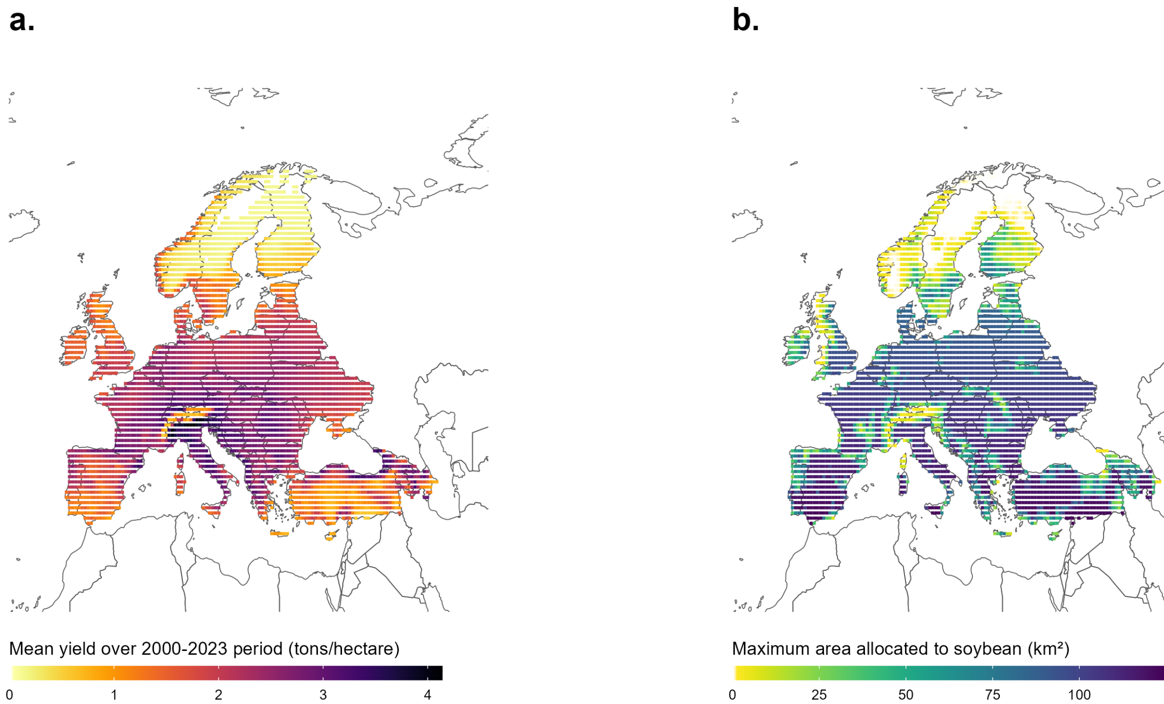

Figure 2: Mean annual amounts of predicted soybean yields for each grid-cell in Europe, between 2000 and 2023 (panel a., on the left); Maximum possible surface that could be allocated to soybean for each grid-cell in Europe (panel b., on the right).

3.2 Multi-objective optimization framework

3.2.1 Structure of a candidate solution. A candidate solution $\mathbf{I}$ in the problem is a vector of continuous values, each in $[0.0, 1.0]$:

$$\mathbf{I} = \{I_0, I_1, ..., I_N\} \tag{4}$$

where each value I_x represents the fraction of available soil surface allocated to soybean for grid-cell x. It is important to notice that I_x is only the fraction of the maximal surface that can actually be dedicated to soybean in grid-cell x, so a value of 1.0 does not correspond to the entire surface of the cell.

3.2.2 Fitness functions. We propose to treat the problem as a multi-objective optimization task where the three objectives considered are (i) maximizing the mean of the annual crop production, (ii) minimizing the between-year production variability, and (iii) minimizing the total surface allocated to soybean.

We start from a matrix $\mathbf{P}$ of crop production projections, where each row corresponds to a grid-cell, and each column corresponds to a year; thus, $\mathbf{P}_{x,y}$ contains the crop production projection for grid-cell x and year y, measured in tons (1 *ton* = 1,000 *kg*) of soybean produced. Given a candidate solution $\mathbf{I}$, the corresponding projected production p_y for a specific year y can be defined as:

$$p_y = \sum_{x=0}^{N} I_x \cdot \mathbf{P}_{x,y} \tag{5}$$

where $\mathbf{I}_x$ is the fraction of land surface allocated to the production of soybeans for grid-cell x.

The mean annual production for a candidate solution can then be computed as:

$$\bar{p} = \frac{1}{Y} \sum_{y=0}^{Y} p_y \tag{6}$$

with Y the total number of years considered. As the optimization task will be framed as a minimization problem, the first fitness function F_1 will be thus defined as:

$$F_1(\mathbf{I}) = -\bar{p} = -\frac{1}{Y} \sum_{y=0}^{Y} \sum_{x=0}^{N} I_x \cdot \mathbf{P}_{x,y} \tag{7}$$

The second fitness function, to be minimized, is the inter-year standard deviation of the crop production, defined as:

$$F_2(\mathbf{I}) = \sqrt{\frac{1}{Y} \sum_{y=0}^{Y} (\bar{p} - p_y)^2} \tag{8}$$

The third fitness function, to be minimized as well, is simply the total surface taken by the soybean fields represented by candidate solution $\mathbf{I}$, expressed as:

$$F_3(\mathbf{I}) = \sum_{x=0}^{N} I_x \cdot \mathbf{S}_x \tag{9}$$

where $\mathbf{S}$ is a vector containing the value of the maximum surface available for soybean crops associated to each grid-cell x, expressed in hectares (*ha*).

3.2.3 Genetic operators. The genetic operators used in the proposed approach are a classic one-point crossover and a Gaussian mutation with mean $\mu_M = 0.0$ and standard deviation $\sigma_M = 0.1$. The crossover is applied with probability $p_C = 0.8$ and the mutation is applied with element-wise probability $p_M = 0.1$.

3.2.4 Evolutionary framework. The MOEA employed for the multi-objective optimization is the Non-Sorting Genetic Algorithm II (NSGA-II) [11], which still represents the state of the art for multi-objective problems with up to three objectives. NSGA-II is set with a $(\mu + \lambda)$ replacement scheme, a tournament selection of size $\tau = 0.02 \cdot \mu$, and a stop condition triggered after a maximum number of function evaluations E_{max}.

4 CASE STUDY

The territory considered for the allocation of soybean corresponds to Continental Europe, including the 27 members of European Union (EU27), added with Albania, Armenia, Azerbaijan, Bosnia and Herzegovina, Belarus, Switzerland, United Kingdom, Georgia, Moldovia, Montenegro, Macedonia, Norway, Serbia, Kosovo, Turkiye, and Ukraine. Following a common domain methodology, the territory is divided into 0.5° grid-cells, for a total of $3,509$ grid-cells. For each grid-cell, we compute soybean yield projections from climate data from 2000 to 2023 using the $pca.m.2$ machine learning model trained on historical yield data. These projections are used in the MOEA to optimize the allocation of soybean cultivation to maximize mean production, to minimize between-year variability of production as well as the total surface allocated over in the $3,509$ grid-cells. Panel a. of Figure 2 shows the mean of projected yields in the area considered for our case study.

We make the hypothesis that the land allocated to soybean would be highly constrained in Europe, given that cropland is already used to grow other major crops and that soybean cannot be grown in place of natural areas (e.g. permanent pastures), in line with the Common Agricultural Policy of the European Union aiming at their protection. Thus, we limit soybean area for each grid-cell to the minimum between (i) 5% of the total grid-cell area or (ii) 20% of the available cropland of each grid-cell can be allocated to soybean crops. The threshold of 20% was selected to simulate a crop sequence where soybean is grown every four years. The maximum available area is represented in panel b. of Figure 2.

After a few preliminary runs, the MOEA used for the proposed approach is set with hyperparameters $\mu = 1,000$, $\lambda = 2,000$, and $E_{max} = 10^6$ for all experimental runs. The code for the experiments is implemented in Python and R, using the `inspyred` [16] library for the evolutionary engine, `scikit-learn` [35] for the machine learning part, and the R packages `terra`[2] and `randomForest`[3] for raster data manipulation and yield prediction model fitting. The data and the code necessary to reproduce the experiments are freely available on the GitHub repository: https://github.com/albertotonda/optimization-crop-allocation.

4.1 Regular conditions

The final non-dominated front found by the proposed approach is presented in Figure 3. A candidate solution of the multi-objective optimization problem is thus represented by an array of $3,509$ floating point values, representing the percentage of land allocated to soybean for each cell in the grid. Each candidate solution was characterized by the total surface allocated to soybean, the mean and the between-year variability (i.e., standard deviation of production) of soybean production over $2000 - 2023$ period in Europe.

Variability of production and total allocated surface tend to be both higher in solutions presenting the highest mean production. On the reverse, mean production seems to decrease in candidate solutions that are more stable (i.e. with lower between-year variability) and with lower total allocated surface. Among all candidate solutions, the overall production ranges from 2.5 to 24.5 Megatons (Mt, 1 Mt = 10^6 tons) of soybean, with a mean of 12.3 Mt. This result shows that none of the candidate solution induce enough soybean to fill 100% of the needs of EU27 in soybean (i.e. 44.5 Mt on average between 2017 and 2022, according to FAOSTATS data [13]). Among all solutions, the between-year variability ranges between 0.1 and 1.1 Mt and is on average 0.5 Mt. Minimum and maximum surface allocated to soybean are $11,733$ and $107,301$ km^2, respectively. It is interesting to note that the total surface allocated to soybean in candidates solutions is higher than the current total soybean area in Continental Europe (i.e., $25,804.2$ km^2 in 2021, according to FAOSTATS data [13]) in 80% of the candidate solutions. Finally, $51,494$ km^2 are on average allocated to soybean crops.

4.2 Experts analysis

Among all candidate solutions, agronomists identify four relevant scenarios. The first is the solution yielding to the highest production, but also to the lowest stability of production ("Scenario A - High production, high risk"). The second scenario considered is the solution showing the highest stability among those with production covering at least 50% of soybean needs of the EU27, i.e. 22.5 Mt of soybean, according to FAOSTATS data [13] ("Scenario B - Production covering 50% EU27 needs"). Third, the solution presenting the highest production and a median variability of production is selected ("Scenario C - Median variability"). Finally, the solution characterized by a total allocated area equivalent to current surfaces in Continental Europe based on FAOSTATS data [13] and with highest production and stability is included ("Scenario D - Current surface"). Each scenario is represented on the final non-dominant front as colored diamonds (Figure 3). Geographical allocation of soybean in Europe in each scenario is represented in Figure 4.

4.2.1 Scenario A - High production, high risk. In scenario A, mean production, variability of production, and total allocated surface reach 24.5 Mt, 1.1 Mt, and $107,300.8$ km^2, respectively. In this scenario, soybean covers almost all territory used for this case study, preferably from North-East of Spain to South of Scandinavian countries. In this scenario, the allocation seems to be driven by the maximum available area for soybean rather than potential yield in the cell (Figure 2b).

4.2.2 Scenario B - Production covering 50% EU27 needs. Geographical allocation is similar is scenarios A and B, although the latest one exhibits lower production (22.5 Mt), between-year variability (1.0 Mt), and total surface ($96,981.3$ km^2), compared to the former.

[2]https://cran.r-project.org/web/packages/terra/index.html
[3]https://cran.r-project.org/web/packages/randomForest/index.html

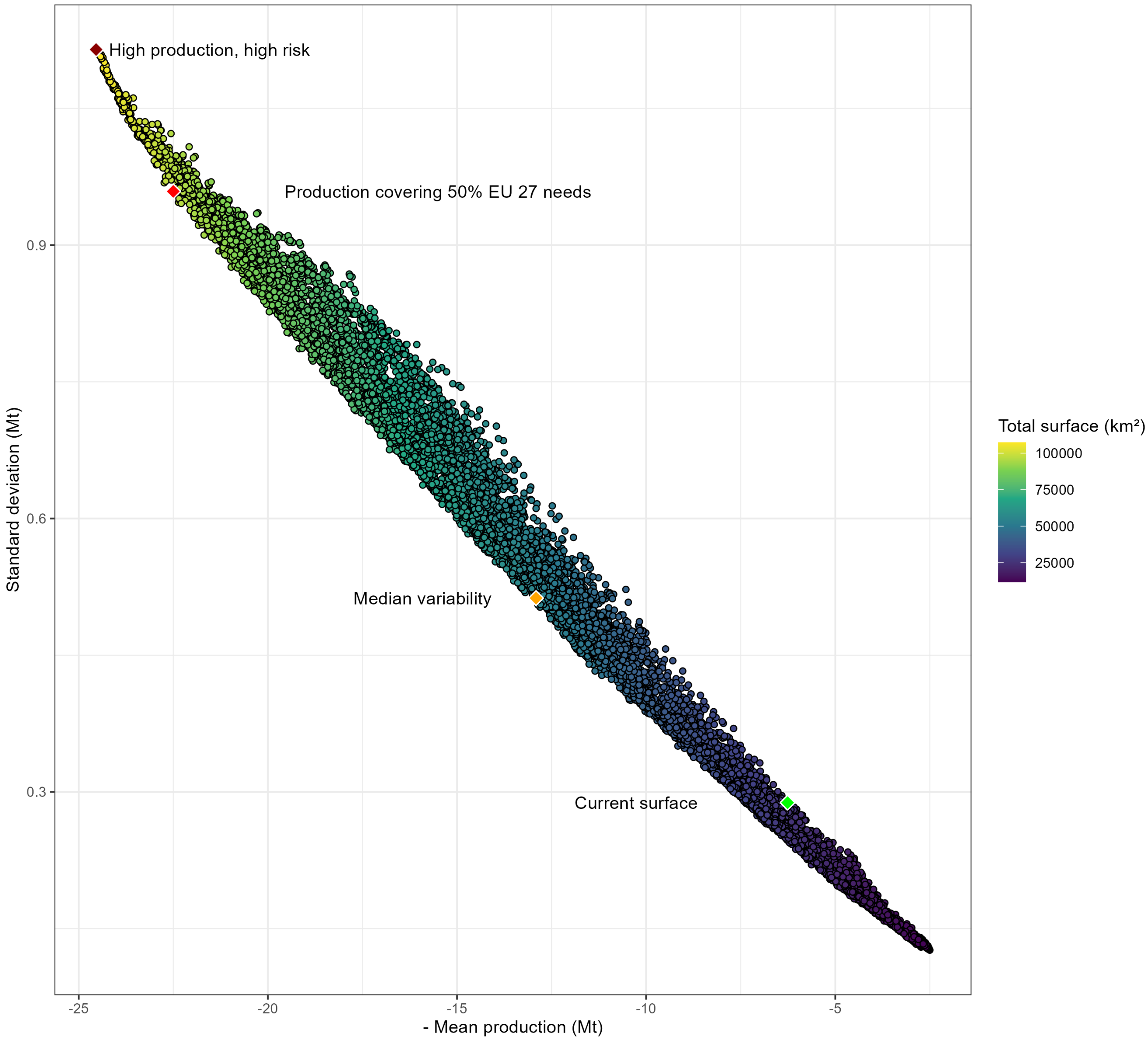

Figure 3: Final non-dominated front found by the proposed approach at the end of the first experimental run, under the hypothesis of regular climate conditions. Candidate solutions in the front are presented as a 2D projection in the space of the first two objectives, with the color of the candidate solutions describing the third objective; the x-axis is reversed for readability, with higher (better) values of the mean yield towards the left. Mean and standard deviation of production are expressed in Megatons (Mt, i.e. 10^6 tons), and total surface in km^2. Four scenarios chosen by expert agronomists and detailed in this article are identified as colored diamonds.

Also, scenario B shows slightly lower area allocated to soybean in Spain, the United Kingdom, and Turkiye, compared to scenario A.

4.2.3 Scenario C - Median variability. In scenario C, characterized by a median variability of production, soybean production is concentrated in the South-West and North-East of France, Italy, Belgium, Central Europe until East of Poland, and covers a part of Belarus, Ukraine and Turkiye. This third scenario produced 12.9 Mt in average, which corresponds to more than 25% of EU27 needs in soybean, while production variability and total allocated are maintained at 0.5 Mt and $54,595.6\ km^2$.

4.2.4 Scenario D - Current surface. If we consider the scenario corresponding to current area of soybean, i.e. roughly $25,000\ km^2$, we see that the production would be even more reduced and concentrated mainly in Italy and Germany, with lower areas in the

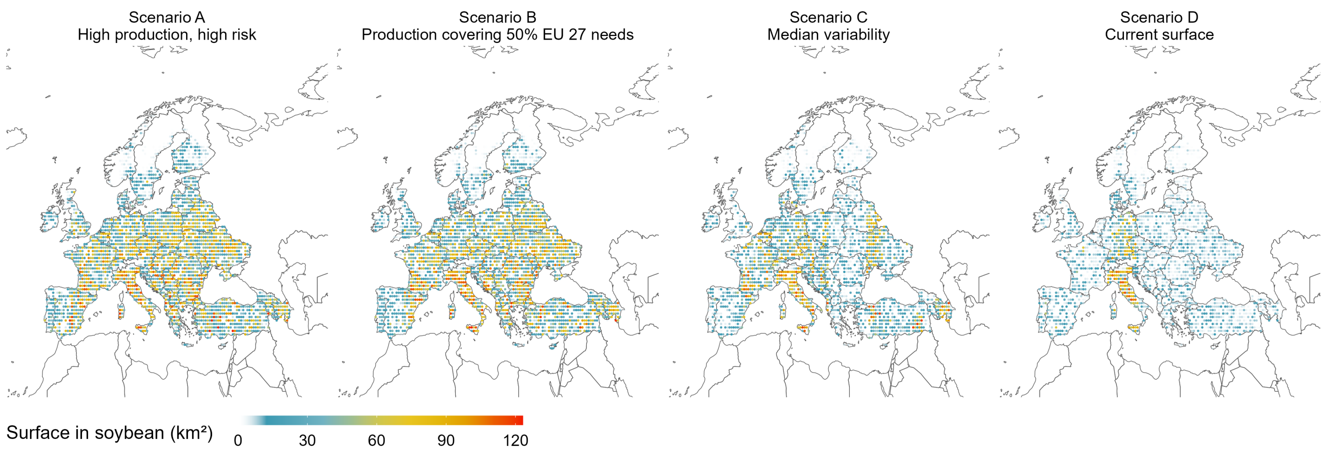

Figure 4: Geographical allocation of soybean in Europe in four scenarios identified by agronomists. Abbreviations: EU 27: 27 members of the European Union.

rest of Europe. In this scenario, production is halved compared to scenario D (mean and standard deviation in scenario D: 6.3 and 0.3 Mt, respectively).

5 CONCLUSIONS AND FUTURE WORKS

This work presents a first multi-objective optimization approach to the problem of allocating crops over large land areas, aiming at simultaneously maximizing annual yield, minimizing between-year variability, and minimizing the surface dedicated to crops. The case study presented deals with the allocation of soybean over the European continent, using a relatively high-resolution grid, a machine learning model to predict yield for each cell in the grid, and an established MOEA to generate a non-dominated front of candidate solutions. An analysis of selected trade-offs on the front obtained at the end of the experimental run shows that the results are sensible and describe realistic scenarios. Four scenarios are analyzed in terms of overall production, between-year variability, total allocated surface, and geographical allocation of soybean.

In the majority of candidate solutions identified by our approach, the total allocated surface is higher than the surface currently allocated to this crop in Continental Europe. This suggests that the area allocated to soybean should be increased to reach these levels of production, as suggested in previous publications [18]. In this case study, the crop frequency of soybean was set to one year every four years: It is possible that the production could be increased by more frequent cultivation. This would also make it possible to reduce the total area dedicated to this particular crop, which is interesting considering that an increase of surface could probably be achieved only at the expense of other crops (e.g. maize) or of natural areas. Another solution that could be considered to save land is intercropping, i.e., the simultaneous growth of multiple crops in the same field. Previous meta-analyses estimate that, compared to monocropping, maize-soybean intercropping might considerably increases the efficiency of land use [42].

Thus, future works will explore the optimization of the simultaneous allocation of multiple species of crops (for example, maize and soybeans), with more complex yield models, able to take into account synergies and competitions between plants, and more separate objectives for each type of crop.

REFERENCES

[1] Linda Arata, Enrico Fabrizi, and Paolo Sckokai. 2020. A worldwide analysis of trend in crop yields and yield variability: Evidence from FAO data. *Economic Modelling* 90 (2020), 190–208. https://doi.org/10.1016/j.econmod.2020.05.006

[2] Nishu Bali and Anshu Singla. 2022. Emerging trends in machine learning to predict crop yield and study its influential factors: A survey. *Archives of computational methods in engineering* (2022), 1–18.

[3] Valter Barbosa dos Santos, Aline Moreno Ferreira dos Santos, José Reinaldo da Silva Cabral de Moraes, Igor Cristian de Oliveira Vieira, and Glauco de Souza Rolim. 2022. Machine learning algorithms for soybean yield forecasting in the Brazilian Cerrado. *Journal of the Science of Food and Agriculture* 102, 9 (2022), 3665–3672.

[4] Bruno Basso and Lin Liu. 2019. Seasonal crop yield forecast: Methods, applications, and accuracies. *advances in agronomy* 154 (2019), 201–255.

[5] Tamara Ben-Ari and David Makowski. 2016. Analysis of the trade-off between high crop yield and low yield instability at the global scale. *Environmental Research Letters* 11, 10 (oct 2016), 104005. https://doi.org/10.1088/1748-9326/11/10/104005

[6] Leo Breiman. 2001. Random forests. *Machine learning* 45 (2001), 5–32. Issue 1. Fundamental citation for Random Forest.

[7] Mathilde Chen, Nicolas Guilpart, and David Makowski. 2024. Comparison of Methods To Aggregate Climate Data To Predict Crop Yield: An Application to Soybean. (2024).

[8] Mengting Chen, Raphael Linker, Conglin Wu, Hua Xie, Yuanlai Cui, Yufeng Luo, Xinwei Lv, and Shizong Zheng. 2022. Multi-objective optimization of rice irrigation modes using ACOP-Rice model and historical meteorological data. *Agricultural Water Management* 272 (Oct. 2022), 107823. https://doi.org/10.1016/j.agwat.2022.107823

[9] Kalyanmoy Deb. 2001. *Multi-objective optimization using evolutionary algorithms.* Vol. 16. John Wiley & Sons. Fundamental citation for MOEAs.

[10] Kalyanmoy Deb and Himanshu Jain. 2014. An Evolutionary Many-Objective Optimization Algorithm Using Reference-Point-Based Nondominated Sorting Approach, Part I: Solving Problems With Box Constraints. *IEEE Transactions on*

Evolutionary Computation 18 (8 2014), 577–601. Issue 4. https://doi.org/10.1109/TEVC.2013.2281535
[11] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and Tamt Meyarivan. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE transactions on evolutionary computation* 6, 2 (2002), 182–197.
[12] Kalyanmoy Deb, Karthik Sindhya, and Jussi Hakanen. 2016. Multi-objective optimization. In *Decision sciences*. CRC Press, 161–200.
[13] R FAOSTAT et al. 2024. FAOSTAT database. *Food Agric. Organ. UN* (2024).
[14] European Commission. Joint Research Centre. Institute for Environment, Sustainability., and CCAFS. 2012. *Combining crop models and remote sensing for yield prediction - concepts, applications and challenges for heterogeneous, smallholder environments: report of CCFAS JRC Workshop at Joint Research Centre, Ispra, Italy, June 13 14, 2012.* Publications Office. https://doi.org/10.2788/72447
[15] Yuan Gao, Anyu Zhang, Yaojie Yue, Jing'ai Wang, and Peng Su. 2021. Predicting Shifts in Land Suitability for Maize Cultivation Worldwide Due to Climate Change: A Modeling Approach. *Land* 10, 3 (2021). https://doi.org/10.3390/land10030295
[16] Aaron Garrett. 2012. inspyred. https://github.com/aarongarrett/inspyred. Accessed January 2024.
[17] C. Gil, A. Márquez, R. Baños, M. G. Montoya, and J. Gómez. 2006. A hybrid method for solving multi-objective global optimization problems. *Journal of Global Optimization* 38, 2 (Nov. 2006), 265–281. https://doi.org/10.1007/s10898-006-9105-1
[18] Nicolas Guilpart, Toshichika Iizumi, and David Makowski. 2022. Data-driven projections suggest large opportunities to improve Europe's soybean self-sufficiency under climate change. *Nature Food* 3, 4 (2022), 255–265.
[19] Toshichika Iizumi and Toru Sakai. 2020. The global dataset of historical yields for major crops 1981–2016. *Scientific Data* 7, 1 (2020), 97.
[20] Hisao Ishibuchi, Noritaka Tsukamoto, and Yusuke Nojima. 2008. Evolutionary many-objective optimization: A short review. In *2008 IEEE congress on evolutionary computation (IEEE world congress on computational intelligence)*. IEEE, 2419–2426.
[21] Nicole D Jackson, Megan Konar, Peter Debaere, and Lyndon Estes. 2019. Probabilistic global maps of crop-specific areas from 1961 to 2014. *Environmental Research Letters* 14, 9 (2019), 094023.
[22] Sonal Jain, Dharavath Ramesh, and Diptendu Bhattacharya. 2021. A multi-objective algorithm for crop pattern optimization in agriculture. *Applied Soft Computing* 112 (Nov. 2021), 107772. https://doi.org/10.1016/j.asoc.2021.107772
[23] Faranak Karamian, Ali Asghar Mirakzadeh, and Arash Azari. 2023. Application of multi-objective genetic algorithm for optimal combination of resources to achieve sustainable agriculture based on the water-energy-food nexus framework. *Science of The Total Environment* 860 (Feb. 2023), 160419. https://doi.org/10.1016/j.scitotenv.2022.160419
[24] Monisha Kaul, Robert L Hill, and Charles Walthall. 2005. Artificial neural networks for corn and soybean yield prediction. *Agricultural Systems* 85, 1 (2005), 1–18.
[25] Rahel Laudien, Bernhard Schauberger, David Makowski, and Christoph Gornott. 2020. Robustly forecasting maize yields in Tanzania based on climatic predictors. *Scientific reports* 10, 1 (2020), 19650.
[26] Guoyong Leng and Jim W Hall. 2020. Predicting spatial and temporal variability in crop yields: an inter-comparison of machine learning, regression and process-based models. *Environmental Research Letters* 15, 4 (2020), 044027.
[27] Raphael Linker. 2020. Unified framework for model-based optimal allocation of crop areas and water. *Agricultural Water Management* 228 (Feb. 2020), 105859. https://doi.org/10.1016/j.agwat.2019.105859
[28] David B Lobell and Marshall B Burke. 2010. On the use of statistical models to predict crop yield responses to climate change. *Agricultural and forest meteorology* 150, 11 (2010), 1443–1452.
[29] David Makowski, Senthold Asseng, Frank Ewert, Simona Bassu, Jean-Louis Durand, Tao Li, Pierre Martre, Myriam Adam, Pramod K Aggarwal, Carlos Angulo, et al. 2015. A statistical analysis of three ensembles of crop model responses to temperature and CO2 concentration. *Agricultural and Forest Meteorology* 214 (2015), 483–493.
[30] Miguel Martínez-Iranzo, Juan M. Herrero, Javier Sanchis, Xavier Blasco, and Sergio García-Nieto. 2009. Applied Pareto multi-objective optimization by stochastic solvers. *Engineering Applications of Artificial Intelligence* 22, 3 (April 2009), 455–465. https://doi.org/10.1016/j.engappai.2008.10.018
[31] Nisrine Mouhrim, Sergiy Smetana, Anita Bhatia, Alexander Mathys, Ashley Green, Daniela Peguero, and Alberto Tonda. 2022. Towards multi-objective optimization of sustainable insect production chains. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion (GECCO '22)*. ACM. https://doi.org/10.1145/3520304.3528898
[32] Christoph Müller, James Franke, Jonas Jägermeyr, Alex C Ruane, Joshua Elliott, Elisabeth Moyer, Jens Heinke, Pete D Falloon, Christian Folberth, Louis Francois, et al. 2021. Exploring uncertainties in global crop yield projections in a large ensemble of crop models and CMIP5 and CMIP6 climate scenarios. *Environmental Research Letters* 16, 3 (2021), 034040.
[33] Joaquín Muñoz-Sabater, Emanuel Dutra, Anna Agustí-Panareda, Clément Albergel, Gabriele Arduini, Gianpaolo Balsamo, Souhail Boussetta, Margarita Choulga, Shaun Harrigan, Hans Hersbach, et al. 2021. ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. *Earth system science data* 13, 9 (2021), 4349–4383.
[34] Antonio L. Márquez, Raúl Baños, Consolación Gil, María G. Montoya, Francisco Manzano-Agugliaro, and Francisco G. Montoya. 2011. Multi-objective crop planning using pareto-based evolutionary algorithms. *Agricultural Economics* 42, 6 (May 2011), 649–656. https://doi.org/10.1111/j.1574-0862.2011.00546.x
[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
[36] Ali Raza, Ali Razzaq, Sundas Mehmood, Xiling Zou, Xuekun Zhang, Yan Lv, and Jinsong Xu. 2019. Impact of Climate Change on Crops Adaptation and Strategies to Tackle Its Outcome: A Review. *Plants* 8, 2 (Jan. 2019), 34.
[37] João Vasco Silva and Ken E Giller. 2020. Grand challenges for the 21st century: what crop models can and can't (yet) do. *The Journal of Agricultural Science* 158, 10 (2020), 794–805.
[38] Yang Su, Benoit Gabrielle, Damien Beillouin, and David Makowski. 2021. High probability of yield gain through conservation agriculture in dry regions for major staple crops. *Scientific Reports* 11, 1 (2021), 3344.
[39] Michiel Van Dijk, Tom Morley, Marie Luise Rau, and Yashar Saghai. 2021. A meta-analysis of projected global food demand and population at risk of hunger for the period 2010–2050. *Nature Food* 2, 7 (2021), 494–501.
[40] Thomas Van Klompenburg, Ayalew Kassahun, and Cagatay Catal. 2020. Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture* 177 (2020), 105709.
[41] Wenbin Wu, Qiangyi Yu, Liangzhi You, Kevin Chen, Huajun Tang, and Jianguo Liu. 2018. Global cropping intensity gaps: Increasing food production without cropland expansion. *Land use policy* 76 (2018), 515–525.
[42] Zhan Xu, Chunjie Li, Chaochun Zhang, Yang Yu, Wopke van der Werf, and Fusuo Zhang. 2020. Intercropping maize and soybean increases efficiency of land and fertilizer nitrogen use; A meta-analysis. *Field Crops Research* 246 (2020), 107661.