A Comparison of Optimization Techniques for Large-scale Allocation of Soybean Crops

MATHILDE CHEN, UMR 518 MIA-PS, INRAE, Université Paris-Saclay, France, CIRAD, UMR PHIM, France, and PHIM, Univ Montpellier, CIRAD, INRAE, Institut Agro, IRD, France
GEORGE KATSIRELOS, UMR 518 MIA-PS, INRAE, Université Paris-Saclay, France
DAVID MAKOWSKI, UMR 518 MIA-PS, INRAE, Université Paris-Saclay, France
ALBERTO TONDA, UMR 518 MIA-PS, INRAE, Université Paris-Saclay, France and Institut des Systèmes Complexes Paris-Ile-de-France, France

The optimal allocation of crops to different parcels of land is a problem of paramount practical importance, not only to improve food and feed production, but also to address the challenges posed by climate change. However, this optimization problem is inherently complex due to the large number of agricultural sites available which generates a vast search space that renders traditional optimization techniques impractical. Moreover, as maximizing average production may generate solutions characterized by high year-by-year instability and lead to large and unrealistic cultivated areas, it is necessary to optimize crop allocation considering several objectives at the same time. In order to tackle this complex optimization problem, we propose a multi-objective approach, simultaneously maximizing the average production, minimizing the year-on-year production variance, and minimizing the total cultivated surface. The approach relies on an established multi-objective evolutionary algorithm, and employs a machine learning model able to predict crop production from weather and irrigation conditions, trained on historical data, making it possible to tackle allocation problems of large size. The proposed approach is compared to a quadratic programming algorithm tailored to the target problem. A case study 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 informative trade-offs between the three conflicting objectives considered, and identify realistic and meaningful crop allocations for supporting stakeholders' decisions.

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

Additional Key Words and Phrases: crop allocation, crop yield forecasting, machine learning, multi-objective optimization

1 INTRODUCTION

Production of agricultural crops for human and animal consumption is heavily influenced by location-dependent conditions, such as quality of the soil, weather patterns, or rainfall. The allocation of specific types of crops to the most appropriate areas for maximizing crop production is becoming a major issue in a context of growing negative impact of climate change [Raza et al. 2019]. The increased frequency of extreme weather events is causing greater year-to-year production instability and increasing the risk of food shortages. Soybean production, in

Authors' Contact Information: Mathilde Chen, UMR 518 MIA-PS, INRAE, Université Paris-Saclay, Palaiseau, France and CIRAD, UMR PHIM, F-34398 Montpellier, France and PHIM, Univ Montpellier, CIRAD, INRAE, Institut Agro, IRD, Montpellier, France, mathilde.chen@cirad.fr; George Katsirelos, UMR 518 MIA-PS, INRAE, Université Paris-Saclay, Palaiseau, France, georgios.katsirelos@inrae.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 and Institut des Systèmes Complexes Paris-Ile-de-France, Paris, France, alberto.tonda@inrae.fr.

ACM 2688-3007/2025/6-ART
https://doi.org/10.1145/3744254

particular, is a major source of concern; it is the world's main source of protein for animal feed, and its production is concentrated in South America, where it has a negative impact on natural areas and causes deforestation. In the European Union (EU), soybean self-sufficiency did not exceed 16% in 2021-2022 [EFSCM 2023]. To cover the 36.3 Megatons (Mt, i.e. 10^6 tons) of soybean consumed in the 27 Members States (average 2018-2022 [FAO 2024]), EU imports soybean from a few countries, especially from Brazil (45.4%), the United-States of America (40.7%), Ukraine (7.4%), and Canada (4.0%) (2023-2024, from the 'Dashboard for the monitoring of food supply and food security' available in the Food supply and security data portal [Pellizzoni et al. 2024]). Relocating soybean production to Europe could have a number of advantages for both the trade balance and the sustainability of agricultural production, but the feasibility of producing large quantities of soybean in Europe remains uncertain. More specifically, we do not know how to optimally distribute soybean in different European regions in order to achieve both a high and stable production 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 relying on classical optimization techniques [Ben-Ari and Makowski 2016]. However, such approaches have been implemented so far considering a small number of spatial units (i.e., a few dozens) while realistic crop allocations involve thousands of crop sites. The application of traditional algorithms has important limitations when the number crop allocation units considered is very large.

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 on soybean yields (i.e., the production by unit of cultivated land), irrigation, and climate, we estimate soybean annual productivity in the EU between 2000 and 2023. These projections are generated for each cell of a spatial grid of 0.5° resolution covering croplands in the EU, corresponding to a total of 3,509 cells over 23 years. Based on this dataset, a multi-objective optimization is performed to generate a set of solutions (i.e., land use allocations) approximating a Pareto front and combining three objectives: (1) maximizing the mean annual soybean production, (2) minimizing the inter-year production variance, and (3) minimizing the total cultivated soybean area.

Preliminary results of this research line have been reported in [Chen et al. 2024c]. However, the present work explores several methodological issues in greater depth. First, we optimize crop allocation for different combinations of the three objectives. Second, as all objective functions are either linear or quadratic, we also compare the MOEA results against an ad-hoc optimization algorithm able to return guaranteed Pareto-optimal points and designed specifically for this application. Finally, after obtaining a set of non-dominated candidate solutions in the space of the objectives, we analyze the trade-offs between the three objective functions, and identify four land use allocations offering new perspectives to improve the level of soybean self-sufficiency in Europe.

The paper is organized as follows. Section 2 introduces the scope of this work. The proposed approach is described in Section 3. The 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

2.1 Forecasting of crop yields

The production by unit of cultivated land is commonly named crop yield and is usually expressed in t/ha. Crop yield predictions at large scales play a significant role for commodity trading and implementation of food security policies [Hoefsloot et al. 2012] and they are essential to prevent food shortages in case of harvest losses. They are also used to make projections about the impact of climate change on crop production [Silva and Giller 2020].

Historically, crop yield predictions were obtained with process-based models which integrate biophysical mechanisms underlying plant growth and development [Bali and Singla 2021]. However, these models are not

always reliable, due to their large number of parameters and the difficulty of estimating them accurately [Leng and Hall 2020]. These models may provide inaccurate forecasts [Lobell and Burke 2010; Müller et al. 2021] which can lead to contradictory conclusions depending on the type of model used [Makowski et al. 2015].

In parallel, statistical linear regression models are simpler and less costly to implement compared to process-based models [Bali and Singla 2021]. For these reasons, they were frequently used in crop yield prediction as well [Basso and Liu 2019; Laudien et al. 2020]. However, standard regression models do not perform well when their inputs are highly correlated and do not always capture all the possible interactions between predictors [Basso and Liu 2019]. More flexible algorithms, such as random forest [Breiman 2001], gradient boosting [Hastie et al. 2009], or artificial neural networks [Goodfellow et al. 2016], are now commonly used to forecast yields of crops [Bali and Singla 2021; Van Klompenburg et al. 2020] and often show better performances compared to traditional statistical methods such as linear regression, in particular to predict soybean yields [Barbosa dos Santos et al. 2022; Guilpart et al. 2022; Kaul et al. 2005].

2.2 Optimization of crop allocation

Globally, the demand for food is expected to increase by 35% to 56% between 2010 and 2050 [Van Dijk et al. 2021] and so far the actual time trend of food demand has closely followed previous predictions. In the last decades, crop productivity has increased thanks to genetic improvements of cultivars 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 [Arata et al. 2020], albeit with negative environmental repercussions. Some studies suggest that there is still some potential to intensify production on existing land [Wu et al. 2018], but this could also lead to more adverse environmental effects. In addition, there is a considerable amount of uncertainty on the impact of climate change on crop production, whose effects are hard to anticipate [Müller et al. 2021].

An alternative approach is to allocate larger proportions of cropland on arable lands with high productivity in order to concentrate the crop production on relatively small areas, while preserving natural ecosystems (such as forest, permanent grasslands, or wetlands) as much as possible. Several studies used crop models to make projections in new areas where the crop is not yet grown and examined the consequences of different choices of crop allocation (e.g. [Jackson et al. 2019; Su et al. 2021]). A similar approach was used to simulate the yields in various climate change scenarios to investigate the suitability of production in the context of climate change (e.g. [Gao et al. 2021; Guilpart et al. 2022]). A number of studies have been carried out over the last few decades to find an optimal distribution of cultivated land at different scales, from the farm [Dogliotti et al. 2005] to the continent [Chen et al. 2024b; Makowski et al. 2000].

Most of these studies relied on linear programming models optimizing average crop production under various constraints without considering the year-by-year variability of the crop production. This aspect is important to consider when optimizing crop allocation because the concentration of crop production on restricted areas with high productivity may lead to a reduction of the resilience of the production systems. Contrary to the average production, the production variance cannot be expressed as a linear function but, instead, as a quadratic function of crop allocations. For this reason, it is more difficult to optimize crop allocation taking into account the production variance. Only one study quantified the trade-off between average value and variance of crop production using quadratic optimization [Ben-Ari and Makowski 2016]. Although useful, this approach only provides estimates at a low geographical resolution (i.e., geographical units covering groups of countries) and does not provide information on how to allocate crop lands locally. Therefore, an optimized soybean allocation strategy which simultaneously maximizes both the yield and stability of production while minimizing the total surface used is needed to identify suitable production areas at a finer scale.

2.3 Multi-objective optimization in agriculture

Multi-objective optimization is a branch of optimization dealing with problems featuring multiple conflicting objectives [Deb 2001; Deb et al. 2016]. 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 multi-objective optimization 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 *dominates* I' if:

$$\begin{aligned} F_j(I) \leq F_j(I') \qquad & \forall j \in \{1, ..., N_o\} \\ \exists m, F_m(I) < F_m(I') & \end{aligned} \tag{1}$$

where F_j are the fitness functions for the $j = \{1, ..., N_o\}$ different objectives. When I dominates I', we write $I \sqsubset I'$. If there exists no I' that dominates I, then I is a non-dominated solution. The set of all non-dominated solutions defines the *Pareto optimal set* of a MOO problem and the set of their valuations, $(F_1(I), \ldots, F_{N_o}(I))$ for all I in the Pareto optimal set defines its *Pareto front*. In many multi-objective evolutionary algorithms, domination is only considered with respect to other candidate solutions, usually the ones inside the population at the current generation, plus the ones stored in a dedicated archive.

Given two distinct sets of solutions, it is not always easy to compare them, which is the case for example, when no solution in either of the two sets dominates any solution in the other set. Such sets of solutions can be compared using measures such as their *hypervolume* [Zitzler et al. 2007]. Intuitively, the hypervolume of a set of solutions S is proportional to the number of solutions that are dominated by at least one solution $I \in S$. Formally, given a set of solutions S and a reference point r in the objective space that is dominated by all solutions, the hypervolume of S is

$$H_r(S) = \Lambda(\{I \mid I \sqsubset r \wedge \exists I' \in S.I' \sqsubset I\}) \tag{2}$$

where Λ denotes the Lebesgue measure. When comparing two different sets S, S', $H_r(S) \geq H_r(S')$ generally indicates that S is a better approximation of the true Pareto optimal set than S'.

MOEAs currently represent the state of the art in the multi-objective optimization domain. While the most recent research works in the field explored complex problems with 10 or more objectives [Deb and Jain 2014; Ishibuchi et al. 2008], for applications with up to three objectives the most established algorithm is arguably the Non-Sorted Genetic Algorithm II (NSGA-II) [Deb et al. 2002], which is used in this work. However, where many MOO applications have objective functions that are hard to evaluate (sometimes even hard to write explicitly), in our case (see section 3), the objectives are not only known and easy to write, but even *tractable*, meaning that a provably optimal solution can be computed in polynomial time. Indeed, when all objectives are linear, it is possible to compute the exact Pareto optimal set, i.e., the set of all solutions which are are not dominated, along with certificates that no other solutions exist [Löhne and Weißing 2017]. In this work, the objectives are linear and quadratic, and are both tractable but, as an approach which relies on purely linear objectives does not apply here, we should rely on more general continuous multi-objective optimization techniques [Eichfelder 2021].

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 [Chen et al. 2022], 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. [Linker 2020] 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 EvMOGA [Martínez-Iranzo et al. 2009] to find non-dominated solutions for the global problem. Finding optimal combinations of crops in greenhouses is the subject of [Márquez et al. 2011], 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 [Gil et al. 2006] algorithms. The work presented in [Karamian et al. 2023] 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 [Mouhrim et al. 2022]. In [Jain et al. 2021], 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 multi-objective algorithm; however, the optimization problem is framed so that only the total area dedicated to each species of plant is taken into account, and the exact locations of the crops in a candidate solution are not considered. 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.

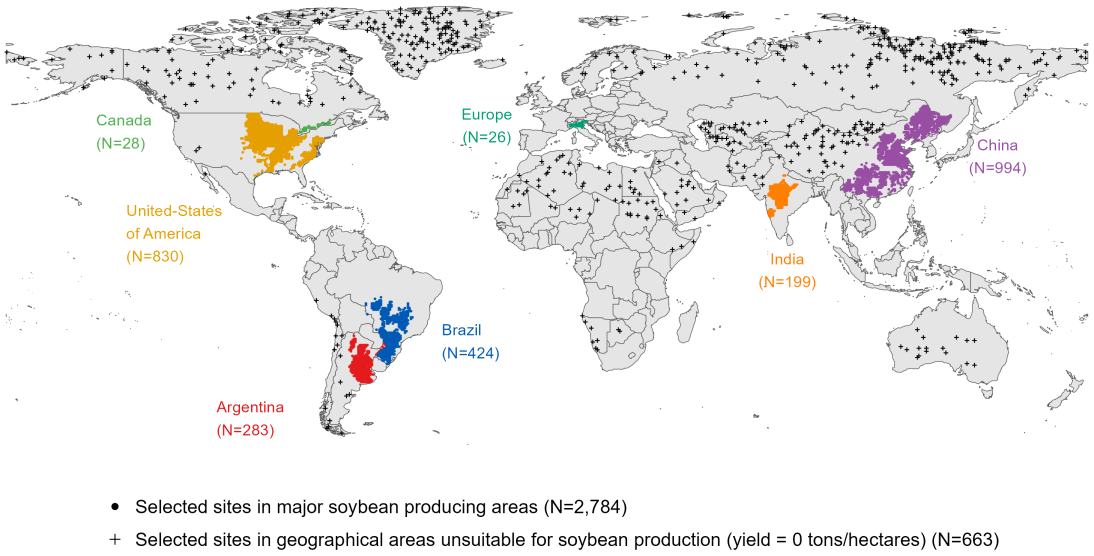

Fig. 1. Map of the grid-cells where soybean yield data were collected for the training set of the machine learning algorithm, later used to predict yield in the European Union's croplands. The yield dataset covers both sites selected in major soybean producing countries (colored dots) and zero-yield sites located in areas unsuitable for soybean cultivation (black crosses). Colored labels indicate, for each soybean-producing country, the number of sites included in the dataset.

3 IMPLEMENTATION TO OPTIMIZE SOYBEAN ALLOCATION

In this work, we use a MOEA to perform a multi-objective optimization of geographical allocation of soybean. Given a map of a large-scale territory, here EU's croplands, 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: (1) maximizing the average production over a given time period; (2) minimizing the between-year production variability; and (3) minimizing the total amount of land allocated to soybean in Europe (to allow for other land uses). As all objective

functions require predictions of soybean yields for all grid cells, 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

Using the random forest [Breiman 2001] model developed in [Chen et al. 2024a], we make projections of soybean productivity in Europe based on climate conditions and irrigation practices. The choice of the model was guided by previous works, comparing various machine learning and statistical models [Guilpart et al. 2022], exploiting different ways to aggregate climate features to predict soybean productivity [Chen et al. 2024a]. Readers can refer to [Chen et al. 2024a], which presents the full procedure applied for data pre-processing, model development, training, and evaluation.

Briefly, the model is trained on historical soybean yield gridded data of resolution of 0.5° and covering period from 1981 to 2016 [Iizumi and Sakai 2020] on grid-cells (hereafter referred to as "sites") located in the EU. The low number of sites available for this region (N=26) leads us to also include sites located in major soybean producing areas, including Argentina (N=283), Brazil (N=424), Canada (N=28), China (N=994), India (N=199), and United-States of America (N=830). To improve model prediction accuracy, additional sites located in geographical areas unsuitable for crop production, characterized by climate deemed inappropriate for soybean cultivation such as desert and artic areas, are included in the training dataset. These zero-yield sites are randomly drawn from six zones identified as environmentally improper for any crop production based on the Köppen-Geiger climate classification [Kottek et al. 2006]. The selection process is designed to ensure a balanced distribution of sites across climate zones. The number of zero-yield sites is determined so that they represented 20% of the full training dataset. In total, 3, 447 sites constitute the training data (Figure 1).

Before training, yield data are detrended in each site using splines to avoid any confusion with technological progress due to improved cultivars and technological progress. Using the ERA5-Land database [Muñoz-Sabater et al. 2021], we derive monthly averages of six climate variables for each combination of site and year (hereafter referred to as "site-year"). The variables considered in this study are: minimum and maximum temperatures (both in℃), precipitation (in mm), solar radiation (in MJ), reference evapotranspiration (mm/day), and vapor pressure deficit (kPA) during soybean growing season. Growing season of soybean is defined country-by-country according to the crop calendars provided by the Agricultural Market Information System.

Several models based on climate predictors are tested: a random forest model including monthly averages of climate data ($avg.m$), a random forest model based on seasonal averages (i.e., average over the whole growing season of soybean) ($avg.s$), two random forest models using two ($pca.m.2$) or three ($pca.m.3$) principal components as inputs, where the principal components captured the variation in monthly averages of climate data. Each model also includes irrigation fraction (in %) as a predictor.

For each model, performance in predicting annual soybean yield from climate and irrigation predictors is assessed using two metrics. The first metric is the Root of the Mean Squared Error (RMSE, in t/ha), computed as:

$$RMSE = \sqrt{\frac{1}{X \cdot Y} \sum_{x=1}^{X} \sum_{y=1}^{Y} (p_{xy} - o_{xy})^2} \tag{3}$$

where p_{xy} and o_{xy} represent the predicted and observed yields in the grid-cell x and the year y, respectively. 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.

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

$$R^2 = 1 - \frac{\sum_{x=1}^{X}\sum_{y=1}^{Y}(p_{xy} - o_{xy})^2}{\sum_{x=1}^{X}\sum_{y=1}^{Y}(o_{xy} - \bar{o})^2} \tag{4}$$

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 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 are computed for each model following two separate cross-validation procedures:

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

The R package `terra`[1] is used for raster data manipulation. The yield model training and testing are implemented with the R package `randomForest`[2]. The code used for model's development and assessment is freely available on the GitHub repository: https://github.com/MathildeChen/SOYBEAN_PRED_COMP.

	Cross-validation results					
	year-by-year		**group-by-group of sites**		**average**	
	R^2	**RMSE**	R^2	**RMSE**	R^2	**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. Performances of random forest models to predict soybean yield using different method to aggregate climate predictors. Higher values of R^2 and lower values of root mean square error (RMSE, in t/ha) indicate 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. All models also included irrigation fraction as predictor.

Model performances estimated from each cross-validation procedure and the results averaged 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 are also comparable across models. The model presenting the best performance on average (i.e., the highest R^2 and the lowest RMSE values) is the random forest using the scores associated with the first two principal components (*pca.m.*2). On average, this model shows a R^2 of 0.93 and a RMSE of 0.38 tons/hectare.

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

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 $[0.0, 1.0]$:

$$\mathbf{I} = \{I_1, I_2, ..., I_N\} \tag{5}$$

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 yearly crop production, (ii) minimizing the between-year production variability (production standard deviation), 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:

$$p_y = \sum_{x=1}^{X} I_x \cdot \mathbf{P}_{x,y} \tag{6}$$

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:

$$\bar{p} = \frac{1}{Y} \sum_{y=1}^{Y} p_y \tag{7}$$

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:

$$F_1(\mathbf{I}) = -\bar{p} = -\frac{1}{Y} \sum_{y=1}^{Y} \sum_{x=1}^{X} I_x \cdot \mathbf{P}_{x,y} \tag{8}$$

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

$$F_2(\mathbf{I}) = \sqrt{\frac{1}{Y} \sum_{y=1}^{Y} (\bar{p} - p_y)^2} \tag{9}$$

The third fitness function, to be minimized as well, is simply the total surface allocated to soybean in each candidate solution $\mathbf{I}$, expressed as:

$$F_3(\mathbf{I}) = \sum_{x=1}^{X} I_x \cdot \mathbf{S}_x \tag{10}$$

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 $p_M = 0.1$. These probabilities are the default values found in the package used for the experiments [Garrett 2012].

3.2.4 Evolutionary framework. The MOEA employed for the multi-objective optimization is the NSGA-II [Deb et al. 2002], 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}.

3.3 Multi-objective optimization with quadratic programming

As all objective functions considered for the target problem are either linear or quadratic, each of them can be optimized separately in polynomial time, by using efficient, established optimization techniques with guarantees of finding optimal points. In order to assess the efficiency of the MOEA proposed approach, we devise an ad-hoc algorithm combining linear and quadratic programming, able to decompose the problem and find provably non-dominated solutions.

The optimization of crop allocation for minimizing cropland area or maximizing average crop production can be performed with a linear program. Although the optimization of crop allocation for minimizing the between-year production standard deviation (σ) is not straightforward, the minimization of the between-year production variance (σ^2) can be performed with a quadratic program. Since the variance is polynomial and as squaring the objective does not change the ordering of possible solutions, optimizing σ can be done in polynomial time.

There exist algorithms which enumerate all points on the Pareto optimal set of a set of linear optimization problems [Löhne and Weißing 2017]. However, no such algorithm exists when the problem is quadratic. In this context, we design an algorithm based on scalarization, able to optimize linear combinations of our three objectives (i.e., crop land area, average production, production variance). It attempts to generate a set of points on the Pareto optimal set with good coverage by identifying areas of the front which are not well covered by the current set of points and focus on those. The drawback of relying on scalarization is that when the Pareto front is non-convex, there are regions of the front which it cannot discover, although variations exist that do not have this problem, such as the augmented Chebyshev scalarization function [Wierzbicki 1982]. In our application, as the front is indeed convex, scalarization based on a linear combination of the objectives is preferable due to its simplicity.

In the following, we describe the particular case of three objectives for simplicity, although the algorithm in its general formulation can be used with an arbitrary number of objectives. We first introduce some notation. Let $\vec{v}$ be a 3-dimensional vector, $||\vec{v}|| = 1$, and let $P_{\vec{v}}(\mathbf{I})$ be the problem min $\sum_{i \in 1..3} v_i F_i(\mathbf{I})$. In other words, $P(\vec{v})$ is the scalarization of the multi-objective problem where the weights of the objectives are given by $\vec{v}$. Note that for linear and quadratic objectives, $P_{\vec{v}}(\mathbf{I})$ is itself a quadratic problem, hence an $\mathbf{I}$ that minimizes it can be computed in polynomial time. We write $optP_{\vec{v}}(\mathbf{I})$ for the optimal value and $\mathbf{I}_{\vec{v}} = argoptP_{\vec{v}}(\mathbf{I})$ for the minimizer. If we evaluate each individual objective F_i on that minimizer, we get the vector $\vec{o}(\vec{v})$, i.e., $\vec{o}(\vec{v}) = (F_i(\mathbf{I}_{\vec{v}}))$. Because $P_{\vec{v}}(\mathbf{I})$ is tractable and we compute its minimizer $\mathbf{I}_{\vec{v}}$ exactly, $\vec{o}(\vec{v})$ is a point on the Pareto front of the three objectives, otherwise for any solution $\mathbf{I}' \sqsubset \mathbf{I}$, we have by definition $\sum_{i \in 1..3} v_i F_i(\mathbf{I}') < \sum_{i \in 1..3} v_i F_i(\mathbf{I}_{\vec{v}})$, a contradiction.

Consider now $\vec{v}_1$ and $\vec{v}_2$. Let $\vec{v}'$ be any point between $\vec{v}_1$ and $\vec{v}_2$, i.e., for some $t \in [0, 1]$, $\vec{v}' = t\vec{v}_1 + (1-t)\vec{v}_2$. Then, the optimum of $P_{\vec{v}'}$ evaluates between $\vec{o}_{\vec{v}_1}$ and $\vec{o}_{\vec{v}_2}$, that is, $\min(o_i(\vec{v}_1), o_i(\vec{v}_2)) \leq o_i(v'_i) \leq \max(o_i(\vec{v}_1), o_i(\vec{v}_2))$ for $i \in [1, 3]$. Moreover, given a set of points $\{\vec{v}_1, \ldots, \vec{v}_n\}$ (which does not necessarily correspond to the complete Pareto front), we say that $\vec{v}_i$ and $\vec{v}_j$ are adjacent if there exists no $\vec{v}_k$, $k \in [1, n] \setminus \{i, j\}$ which can be written as $\vec{v}_k = t\vec{v}_1 + (1-t)\vec{v}_2, t \in [0, 1]$.

The algorithm we use is based on the following observation. Given a collection of points $\vec{v}_1, \ldots, \vec{v}_n$, then evaluating $optP_{\vec{v}}(\mathbf{I})$ for $\vec{v}$ between two adjacent points, say $\vec{v}_1$, $\vec{v}_2$, will increase the hypervolume of the front. When there are two objectives, by the fact that $\vec{v}_1$ and $\vec{v}_2$ are on the Pareto front, $\vec{v}$ can not evaluate to better than both of $\vec{v}_1$ and $\vec{v}_2$ in either dimension. The most optimistic possibility is that $o_i(\vec{v}) = \min(o_i(\vec{v}_1), o_i(\vec{v}_2)) + \epsilon$ for some small positive ϵ. Under this optimistic scenario, the increase of the hypervolume of the front is exactly

$\prod_{i\in[1,2]}(\max(o_i(\vec{v}_1), o_i(\vec{v}_2)) - (\min(o_i(\vec{v}_1), o_i(\vec{v}_2)) + \epsilon)) \cong \prod_{i\in[1,2]}(\max(o_i(\vec{v}_1), o_i(\vec{v}_2)) - \min(o_i(\vec{v}_1), o_i(\vec{v}_2)))$. We write $OH(\vec{v}_1, \vec{v}_2)$ for this quantity.

The above observation leads us to a strategy for generating points on the front in two dimensions: start with the collection of points $\vec{v}_1 = (1, 0), \vec{v}_2 = (0, 1)$. Then, in order to generate a new point from an existing collection, we pick the pair of adjacent points $\vec{v}_i, \vec{v}_j$ that maximizes the optimistic estimation for the increase in hypervolume, evaluate the new point $\frac{1}{2}\vec{v}_i + \frac{1}{2}\vec{v}_j$, and iterate until we satisfy a stopping criterion. In our implementation, we stop when we have generated N_{2d} points, where N_{2d} is a parameter.

In the case of more than two objectives, we proceed as follows. We first generate points on the Pareto front of the first two objectives until a stopping criterion is met. In our implementation the stopping criterion is whether the number of points is at least N_{seed}, where N_{seed} is a parameter. At the end of this first step, we get the collection of two-dimensional points and for each pair of adjacent points $\vec{v}_i = (x, y), \vec{v}_j = (x', y')$, we generate two points $\vec{v}'_i = (x, y, 0)$ and $\vec{v}'_j = norm(x', y', 1)$ [3], then proceed to generate points on the front between $\vec{v}'_i$ and $\vec{v}'_j$. We extend our measure of optimistic increase of hypervolume OH to more than two dimensions as follows. Let $\vec{v}_1 = (x_1, y_1, z_1)$ and $\vec{v}_2 = (x_2, y_2, z_2)$. We define $OH(\vec{v}_1, \vec{v}_2)$ as OH over the projections to the second and dimension (in the general case: to the ith and $(i-1)$th dimension), i.e., we define $OH((x_1, y_1, z_1), (x_2, y_2, z_2)) \equiv OH((y_1, z_1), (y_2, z_2))$. We continue this until a stopping criterion is met, in our implementation until we have generated N_{ext} points, where N_{ext} is a parameter.

As we explained earlier, optima of the scalarized problem give non-dominated points, hence points on the Pareto front. Conversely, it has been shown [Mueller-Gritschneder et al. 2009] that for all three-dimensional points $\vec{v} = (x, y, z)$, there exist two dimensional points on the two-dimensional Pareto front, $\vec{v}_1 = (x_1, y_1)$, $\vec{v}_2 = (x_2, y_2)$ that that $\vec{v}$ lies between the three dimensional extensions of $\vec{v}_1$ and $\vec{v}_2$.

3.3.1 Parameters. When applied to the crop allocation problem, we generate $N_{2d} = 500$ points for two-dimensional fronts. For three-dimensional fronts, we generate $N_{seed} = 30$ points on the front of F_1 and F_3 and then, for each adjacent pair of points on that front, we extend to F_2 and generate $N_{ext} = 30$ points between them. The total number of points generated is 993, taking into account the original points $(0, 1)$ and $(1, 0)$ and the doubling when we move from two dimensions to three (because each point (x, y) on the two-dimensional front becomes two points $(x, y, 0)$ and $(x, y, 1)$ on the three dimensional front).

There are two reasons why we start with F_1 and F_3, both related to efficiency. First, recall that F_1 and F_3 are both linear, so generating the initial N_{seed} points is extremely fast, less than a second in total.

The second reason for using this order has to do with effectiveness of the points generated, i.e., with the improvement in hypervolume per generated point. Recall that we cannot optimize F_2 directly but only F_2^2, yet we compute $\vec{o}(\vec{v})$ using F_1, F_2, F_3. Since the magnitude of all three objectives is similar [4], squaring F_2 in the scalarization approach creates objective functions that heavily favor F_2. To see the problem, consider two points $\vec{v}_i = (x, y), \vec{v}_j = (x', y')$ on the two-dimensional front. The algorithm extends them to the third dimension, creating the points $\vec{v}'_i = (x, y, 0), \vec{v}'_j = norm(x', y', 1)$, which give optimization problems with objectives $\min xF_1 + yF_3$ and $\min x'F_1 + y'F_3 + F_2^2$, respectively. It then generates the point $\vec{v}_1 = \frac{1}{2}\vec{v}'_i + \frac{1}{2}\vec{v}'_j = norm((x+x')/2, (y+y')/2, 0.5)$, and it computes $\vec{o}(\vec{v}_1)$. However, because F_2^2 has much larger magnitude than F_1 and F_3, we have that $\vec{o}(\vec{v}_1) \cong \vec{o}(\vec{v}'_j)$, i.e., it makes little difference whether its weight is 1 or 0.5. As a result, $OH(\vec{v}_1, \vec{v}'_j) \cong 0$ and $OH(\vec{v}_1, \vec{v}'_i) \cong OH(\vec{v}'_i, \vec{v}'_j)$. This remains the case for a few more iterations until the weight of F_2^2 has been reduced sufficiently to compensate for the difference in magnitude. The result of this in the Pareto front generated is that some of the N_{ext} points generated at each iteration of the algorithm on three dimensions are not very informative.

[3] We write $norm(\vec{v})$ for the function that normalizes $\vec{v}$ to length 1, i.e., $norm(\vec{v}) = \frac{\vec{v}}{||\vec{v}||}$

[4] This is a property of the specific objective functions we use, not an inherent algorithmic property.

Despite this pathological behavior, we found in preliminary experiments that the algorithm generates a good coverage of the front. We can, however, do better by accounting for this effect. Specifically, when extending two-dimensional pairs of points $\vec{v}_i = (x, y), \vec{v}_j = (x', y')$ to three dimensions, we use 0.1 as the initial weight of F_2^2, hence we extend them to $\vec{v}'_i = (x, y, 0), \vec{v}'_j = norm(x', y', 0.1)$. This has the effect of essentially skipping a few of the unproductive steps described above and makes the algorithm have similar behavior as if we had set N_{ext} higher. It remains future work to find a method which automatically adjusts for large differences in magnitude between different objectives.

4 RESULTS OF THE OPTIMIZATION

The geographical area considered for the allocation of soybean corresponds to the 27 Members States of the EU (EU27). Following previous studies, the territory is divided into 0.5° grid-cells, for a total of 3,509 grid-cells. From the climate data reported in each grid-cell from 2000 to 2023, we compute soybean yield projections using the *pca.m.2* machine learning model trained on historical data as explained above. These projections are used in the MOEA to optimize the allocation of soybean cultivation to maximize average production, to minimize between-year variability of production as well as the total surface allocated over in the 3,509 grid-cells. Panels a. and b. of Figure 2 shows the mean and standard deviation, respectively, of projected yields in the EU27.

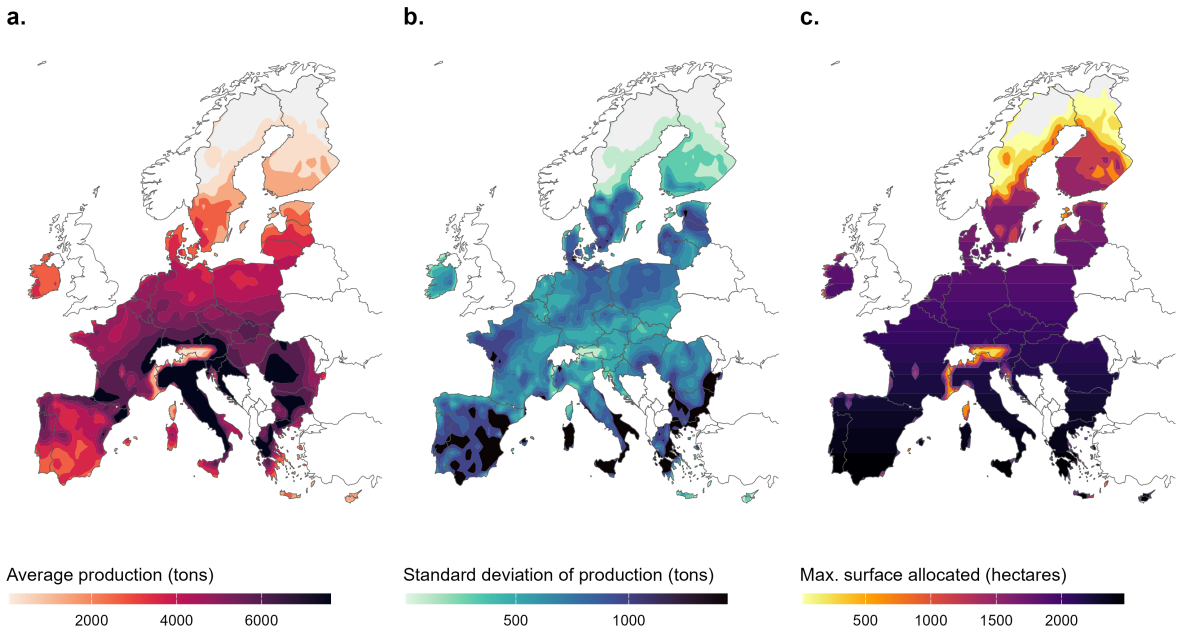

Fig. 2. Average values of projected soybean yields (panel a., on the left), between-year standard deviation of projected soybean yields (panel b., in the center), and maximum area that could be allocated to soybean (panel c., on the right) for each grid-cell in the 27 Member States of the European Union between 2000 and 2023. Soybean yield projections in each grid-cell are estimated by a random forest model trained on historical soybean yield data and including climate and irrigation fraction as predictors.

We make the hypothesis that the land allocated to soybean would be restricted in Europe, given that cropland is already used to grow other major crops such as maize or wheat. In addition, 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 constrain the area allocated to soybean in each grid-cell to be lower than the minimum value between (i) 5% of the total grid-cell area and (ii) 20% of the available cropland of each grid-cell. The threshold of 20% was selected to simulate a crop sequence where soybean is grown every five years. The resulting maximum soybean areas are presented in panel c. of Figure 2.

We compare six optimization settings, according to the set of optimization criteria considered (i.e., average production and surface; average production and variance; average production, variance, and surface) and the optimization approach used (NSGA-II or quadratic programming). In addition, we estimate the hypervolume of the final non-dominated fronts in each setting to ease the comparison between algorithms. The values of the hypervolumes are computed on normalized values for each objective, considering the highest and lowest values obtained for each objective between the two algorithms. Computations of the hypervolumes are implemented in Python, using the `pymoo` [Blank and Deb 2020] library[5].

After few preliminary runs, the MOEA is set with hyperparameters values of $\mu = 1,000$, $\lambda = 2,000$, and $E_{max} = 10^6$ for all experimental runs. MOEA is implemented in Python, using the `inspyred` [Garrett 2012] library for the evolutionary engine. The quadratic programming algorithm used as a comparison is implemented in C++, using IBM CPLEX version 12.7 to optimize the scalarized problems. All experiments are run on a 72-core Intel(R) Xeon(R) w9-3475X server with 128GB of RAM. In the following, we focus on comparing the quality of the fronts generated by the two algorithms rather than computational efficiency. The reason is that efficiency is hard to compare due to the differences in performance between the different programming languages employed in the study. A whole run of NSGA-II with the aforementioned hyperparameters lasts approximately 48 hours to ultimately return a non-dominated set of 1,000 candidate solutions. Obtaining the same number of solutions using the quadratic programming algorithm takes about 2 hours (obtaining one provably non-dominated point using quadratic programming takes about 7 seconds). Since in this particular case all objective functions are either linear or quadratic, the quadratic programming algorithm will always be more efficient.

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 fronts found by the six optimization settings are displayed in Figure 3. Plots on the left panel represent the non-dominated fronts for both optimization approaches, for comparison. Plots in the middle and right panels display the same fronts, but separately for NSGA-II and quadratic programming. For each setting, a candidate solution of the multi-objective optimization problem is 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 average and the between-year standard deviation (measuring the inter-annual variability) of soybean production over $2000 - 2023$ period in the EU27 (Table 2).

As shown in the panels a and b of Figure 3, the non-dominated front identified by the NSGA-II approach is similar to the one identified through quadratic programming when optimization is performed on production and either surface or between-year variability. In these experiments, the hypervolume of the NSGA-II front is 2 and 1% lower compared to the front found by quadratic programming, respectively. More differences are observed when the three objectives are simultaneously optimized (panel c of Figure 3). In that case, the candidate solutions identified by the quadratic programming are more spread out compared to those found by the NSGA-II approach.

[5]Pymoo, multi-objective optimization in Python, https://pymoo.org/

The hypervolume of the front found by quadratic programming is 10% higher compared to the one obtained with NSGA-II, possibly indicating issues with hyperparameter settings.

Among all candidate solutions, the average soybean production ranges from 0.00 to 9.48 Mt, with a mean of 5.55 Mt. This result shows that none of the candidate solution generate enough soybean production to satisfy 100% of the needs of EU27 in soybean (i.e. 36.3 Mt on average between 2018 and 2022, according to FAOSTATS data [FAOSTAT et al. 2024]). The share of candidates solutions yielding to at least 5.81 Mt of soybean in average (i.e., corresponding to the 16% self-sufficiency rate covered in the EU27 in 2021 [EFSCM 2023]) varies from 15% to 61% depending on the optimized criteria and approach. Among all candidate solutions, the between-year standard deviation ranges between 0.00 and 0.46 Mt and is on average equal to 0.22 Mt. Minimum and maximum surface allocated to soybean are 0.00 and 4.416 Mha, respectively. On average over all solutions, 2.21 Mha are allocated to soybean crops. It is interesting to note that the frequency of candidates solutions with the total surface allocated to soybean higher than the current total soybean area in the EU (i.e., 1 Mha on average between 2018 and 2022, according to FAOSTATS data [FAOSTAT et al. 2024]) varies from 42% to 88% depending on the optimization settings.

The candidate solutions identified by NSGA-II and quadratic programming show significant differences when the optimization is conducted on average production and soybean surface only or on the three objectives simultaneously (p-value estimated using Student's t-test on differences in average production, standard deviation, or total soybean surface between NSGA-II and quadratic programming <0.001). On the contrary, the candidate solutions for NSGA-II are non-significantly different, in terms of average production, standard deviation, and soybean surface, to those identified by the quadratic programming in the two-objective setting based on production and standard deviation (p>0.05).

In each of the six optimization settings, standard deviation of production and total allocated surface tend to be both higher in the solutions presenting the highest mean production. On the reverse, mean production tends to decrease in candidate solutions that are more stable (i.e. with lower between-year standard deviation) and with lower total allocated surface.

	Average production (Mt)		**Between-year variability (Mt)**		**Total surface allocated (Mha)**	
	NSGA-II	***Quadratic program -ming***	***NSGA-II***	***Quadratic program -ming***	***NSGA-II***	***Quadratic program -ming***
Among all simulations	**5.279 (2.757) [0.000 - 9.478]**	**5.372 (3.166) [0.000 - 9.478]**	**0.210 (0.134) [0.000 - 0.456]**	**0.256 (0.160) [0.000 - 0.460]**	**2.220 (1.265) [0.000 - 4.416]**	**2.177 (1.446) [0.000 - 4.416]**
optimization on						
Average production and surface	5.171 (2.815) [0.001 - 9.477]	6.329 (2.901) [0.000 - 9.478]	0.257 (0.137) [0.000 - 0.456]	0.316 (0.140) [0.000 - 0.460]	2.023 (1.248) [0.000 - 4.415]	2.575 (1.384) [0.000 - 4.416]
Average production and variability	5.658 (2.619) [0.080 - 9.477]	5.554 (2.698) [0.033 - 9.478]	0.192 (0.136) [0.000 - 0.455]	0.185 (0.137) [0.000 - 0.455]	2.622 (1.172) [0.040 - 4.416]	2.577 (1.196) [0.023 - 4.416]
Average production, variability, and surface	5.225 (2.766) [0.000 - 9.478]	2.817 (2.425) [0.000 - 9.478]	0.204 (0.131) [0.000 - 0.455]	0.113 (0.110) [0.000 - 0.458]	2.179 (1.270) [0.000 - 4.416]	1.049 (0.984) [0.000 - 4.416]

Table 2. Soybean average productions, between-year standard deviations (between-year variability), and allocated surfaces of the candidate solutions constituting the final non-dominated fronts found by the six optimization settings. Data are mean (standard deviation) and [minimum - maximum] values.

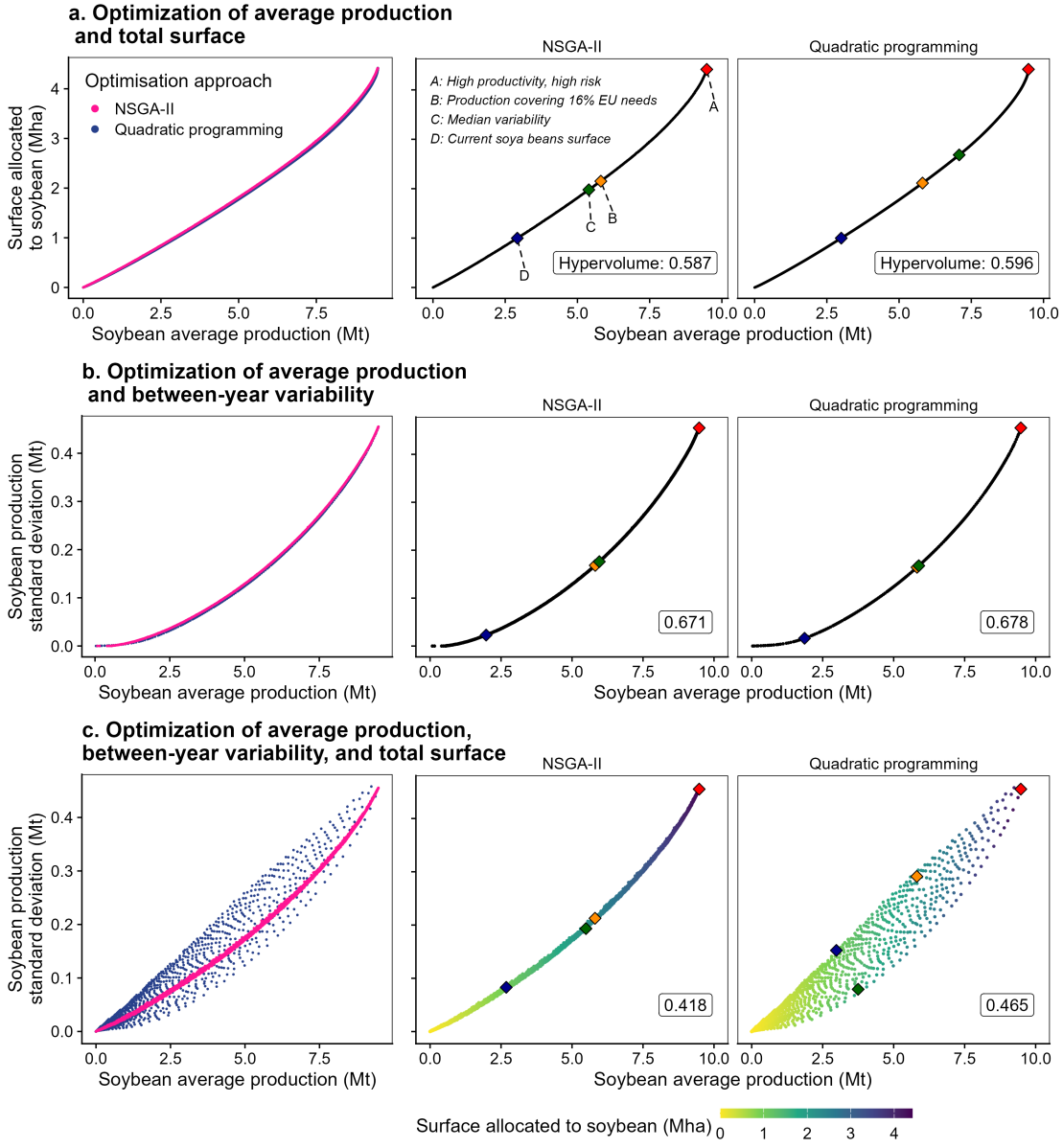

Fig. 3. Final non-dominated fronts, according to the optimized criteria and the optimization approach. Plots on the left panel represent the fronts non-dominated for both optimization approaches. Plots in the middle and right panels display the same fronts, but separately for NSGA-II and quadratic programming. For the panels a. and b., candidate solutions in the front are presented as a 2D projection in the space of the two objectives. In panel c., the color of the candidate solutions additionally describes the third objective (surface). Mean and standard deviation of production are expressed in Megatons (Mt, i.e. 10^6 tons), and total surface per 10^6 ha (Mha). For each individual front, the hypervolume computed on normalized values for each objective, (i.e., considering the highest and lowest value obtained for each objective between the two algorithms) is presented in a label in the right-bottom corner. The four contrasting scenarios detailed in this article are identified as colored diamonds.

4.2 Analysis of selected soybean allocations in the EU27

Among all candidate solutions, four relevant contrasting scenarios were identified: (1) the solution yielding to the highest average production, but also to the lowest production stability ("Scenario A - High production, high risk"), (2) the solution showing the highest stability (lowest standard deviation) among solutions with average production covering at least 16% of soybean needs of the EU27 (this self-sufficiency rate corresponds to 5.81 Mt of soybean, according to FAOSTATS data [FAOSTAT et al. 2024], "Scenario B - Production covering 16% EU needs"), (3) the solution presenting the highest average production and a median standard deviation of production ("Scenario C - Median variability"), (4) the solution characterized by a total allocated area equivalent to current surfaces in the EU (i.e., 1 Mha based on FAOSTATS data [FAOSTAT et al. 2024]) and with highest average production and stability ("Scenario D - Current surface").

Each scenario is represented on the final non-dominant front as colored diamonds (Figure 3). Average production, variability (standard deviation), and allocated surface of the candidates solution corresponding to each optimization setting are presented in the Table 3. The spatial allocations of soybean in the EU27 corresponding to the four scenarios are presented in Figures 4, 5, 6, and 7.

		Average production (Mt)		**Standard deviation (Mt)**		**Total surface (Mha)**	
Scenario	**Optimized criteria**	**NSGA-II**	**Quadratic programming**	**NSGA-II**	**Quadratic programming**	**NSGA-II**	**Quadratic programming**
A	PS	9.477221	9.47809	0.45519966	0.45519173	4.414558	4.4160737
	PV	9.477462	9.47809	0.45511268	0.45519173	4.4156933	4.4160737
	PVS	9.47809	9.47809	0.45519173	0.45519173	4.4160737	4.4160737
B	PS	5.810658	5.811706	0.29225767	0.30078274	2.1516086	2.1132721
	PV	5.817176	5.823068	0.16829714	0.16429107	2.6735434	2.6734359
	PVS	5.809639	5.815237	0.21265622	0.29126955	2.3323299	2.115791
C	PS	5.39488	7.081144	0.27164508	0.35864661	1.9776238	2.6837497
	PV	5.961742	5.88473	0.17609441	0.16757476	2.7342476	2.6995271
	PVS	5.489298	3.747736	0.19330679	0.07934193	2.2111077	1.6119288
D	PS	2.919167	2.999042	0.14928436	0.15322764	0.99809	0.9989066
	PV	1.974349	1.860643	0.02290593	0.01618338	0.9990555	0.9768853
	PVS	2.682715	2.982066	0.08304864	0.15236099	0.9951798	0.9926413

Table 3. Solutions in each optimization setting for four contrasting scenarios: Scenario A - High production, high risk; Scenario B - Production covering 16% EU27 needs; Scenario C - Median variability; Scenario D - Current soybean surface. Abbreviations: PS: average production and surface; PV: average production and variability; PVS: average production, variability, and surface.

4.2.1 Scenario A - High production, high risk. This scenario is characterised by values of average production, between-year standard deviation of production, and total allocated surface of 9.48 Mt, 0.46 Mt, and 4.42 Mha, respectively. The surface allocated to soybean in this scenario covers a big share of the European territory, from North-East of Spain to South of Scandinavian countries (Figure 4). Here, the allocation are driven by both maximum available areas for soybean (Figure 2c) and the potential average soybean yields. [Guilpart et al. 2022] aimed also to produce cropland distributions maximizing production. However, their approach was different and did not rely on a formal optimization technique. Instead, their cropland distributions were obtained simply by allocating soybeans to grid cells showing the highest average soybean yields in Europe without considering cropping areas. In [Guilpart et al. 2022], soybeans were thus mainly allocated to a broad band covering the central part of Europe, from West to East, i.e. Northern Spain, Central and Southern France, Northern Italy, and the Balkan regions. These regions correspond to those with the highest average yields in Europe according to Figure 2

(panel a). Here, in Figure 4, soybean is allocated to regions with high average yields, but also to regions with large cultivated areas, located in southern Europe (Figure 2, panel c). Since large cultivated areas can compensate for lower yields, it is logical to allocate soybeans to grid cells with large cultivated areas, even if these grid cells have slightly lower yield values. This explains why our cropland distributions are not identical to those of [Guilpart et al. 2022].

4.2.2 Scenario B - Production covering 16% *EU27 needs.* For this scenario, all candidates solutions yield to at least 5.81 Mt of soybean in average, but they have different levels of variability and allocated surface, depending on the optimized criteria and optimization approach.

When performing the optimization on production and surface ("PS" setting in Table 3), the quadratic programming leads to a solution with higher variability compared to the one found through NSGA-II (0.301 and 0.292 Mt, respectively), but with lowest surface (2.113 and 2.152 Mha, respectively). When performing the optimization on average production and variability ("PV" setting), both solutions present similar allocated surface but slightly different average productivity levels (5.817 and 5.823 Mt for NSGA-II and quadratic programming, respectively). Finally, when the three objectives are considered ("PVS" setting), the solution found by NSGA-II shows, compared to the one identified by quadratic programming, a lower variability (0.213 and 0.291 Mt, respectively) but higher allocated surface (2.332 and 2.116 Mha, respectively), for equivalent levels of average productivity (5.810 and 5.815 Mt, respectively).

For scenario B, the geographical allocation corresponding to each solution mentioned above are presented in Figure 5. For all optimization settings, soybean is preferably allocated in the areas with higher soybean productivity, such as Italy, South-Western France, Romania, and Northern Greece (Figure 2a). Once these areas are saturated with soybean, the crop is allocated to other regions depending on the optimization criteria and the optimization method.

For example, when the optimization is performed on average production and variability, solutions found by both NSGA-II and quadratic programming allocated higher surfaces to soybean in a area spanning from Northern Spain to Romania and from Greece to Southern Germany. The solution identified by NSGA-II additionally allocated soybean in Sardinia, North-Western France, North-Eastern Germany, Poland, Latvia, and Lithuania. By comparison, geographical allocations of soybean based on the optimization of average production and variability expand to Southern Spain, Ireland, and Scandinavian countries because these regions are characterized by more stable production. Finally, the geographical allocation based on the three criteria and using the NSGA-II algorithm is a mix between previously mentioned settings, but with more area allocated in Southern Europe and less in Northern Europe. Quadratic programming found similar candidate solution when allocating soybean crops by simultaneously optimizing average production and surface or by simultaneously optimizing average production, variability, and surface.

4.2.3 Scenario C - Median variability. In scenario C, the solutions identified by NSGA-II and quadratic programming show contrasting levels of average production, variability, and surface.

Similarly to scenario B, the solutions found by optimization on average production and surface for the scenario C is concentrated in the South-West and North-East of France, Italy, Greece, Belgium, Central Europe until East of Poland. Like in previous scenario, the allocated area is more concentrated when the optimization is performed on surface and average production (first line of Figure 6) or when using quadratic programming compared to NSGA-II.

4.2.4 Scenario D - Current surface. In the scenario corresponding to current total area cultivated with soybean, i.e. roughly 1.0 Mha, the average production would be reduced and concentrated mainly in Italy and Central Europe, as well as Southern Germany, with lower areas allocated to soybean in the rest of Europe (Figure 7). In

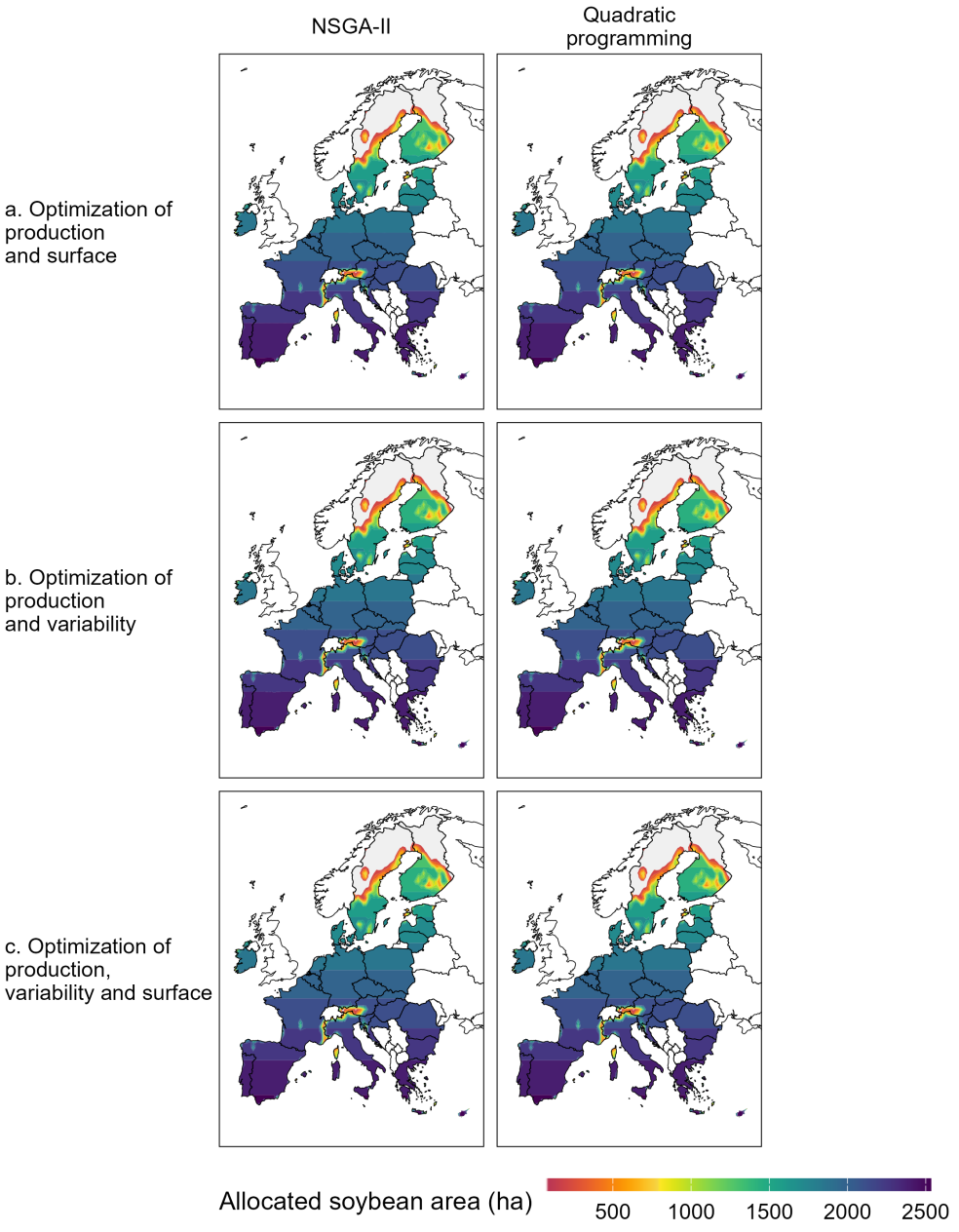

Fig. 4. Geographical allocation of soybean in Europe in the scenario A (high productivity, high risk), according to the optimization setting.

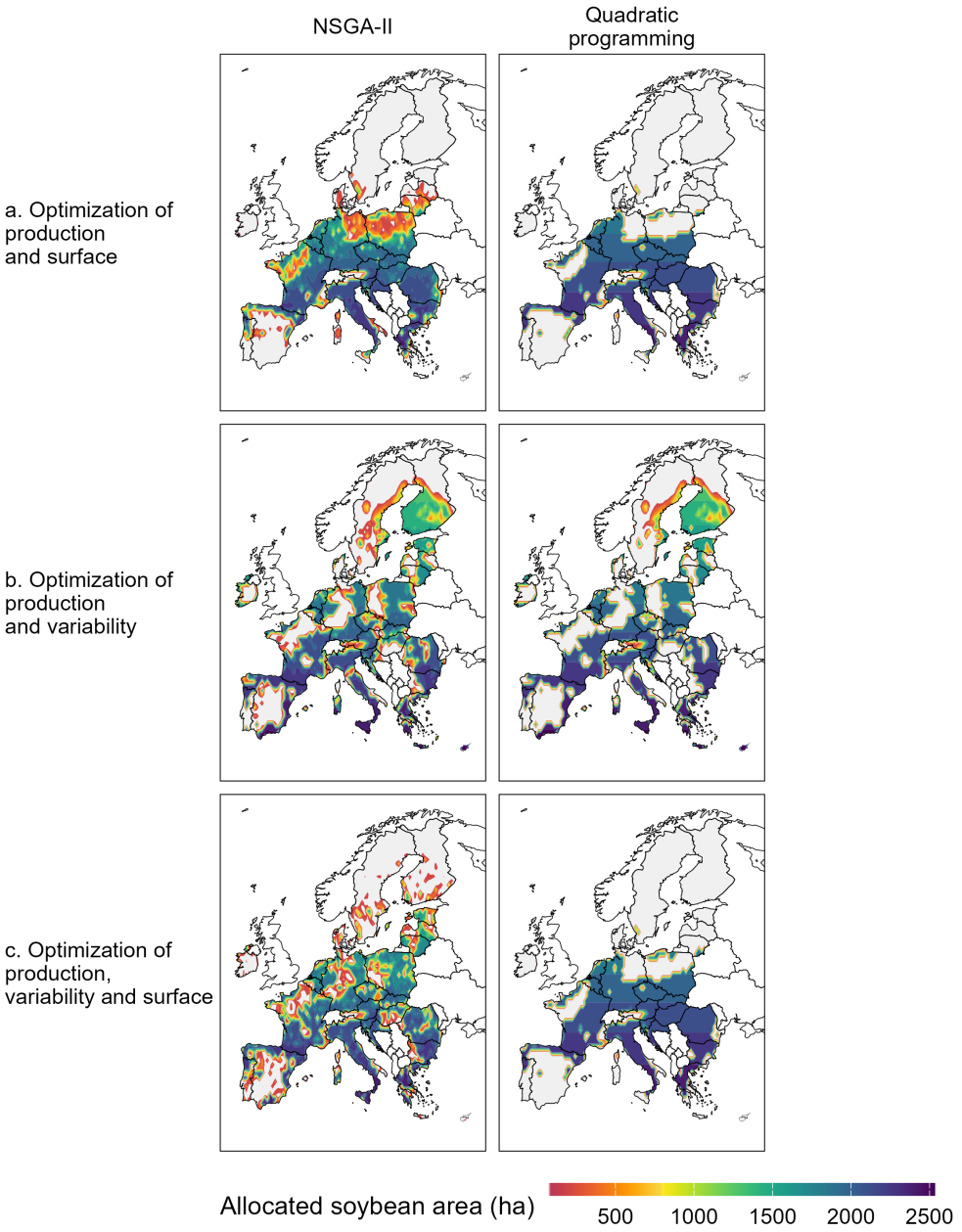

Fig. 5. Geographical allocation of soybean in Europe in the scenario B (production covering 16% of EU27's needs in soybean), according to the optimization setting.

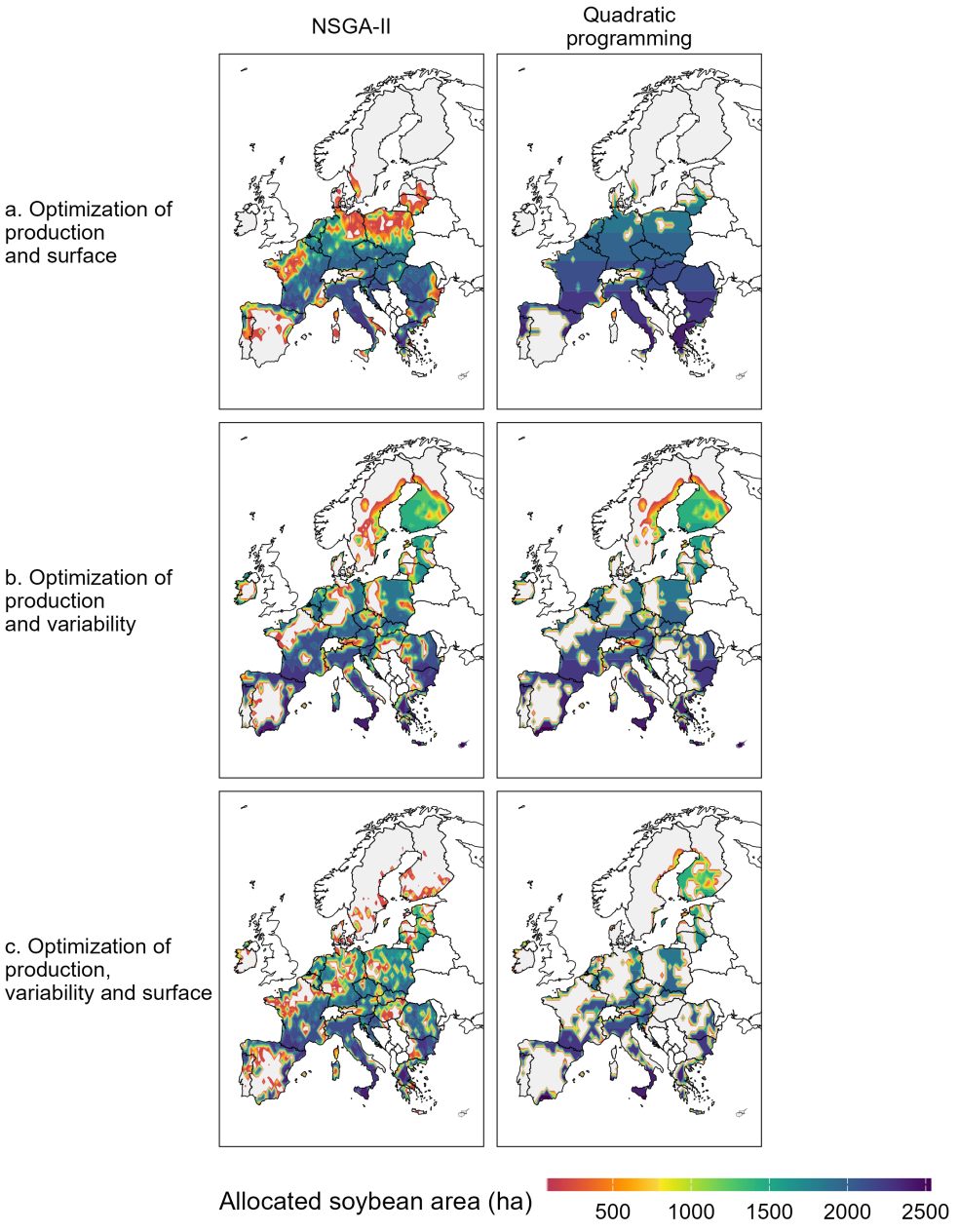

Fig. 6. Geographical allocation of soybean in Europe in the scenario C (median variability), according to the optimization setting.

this scenario, average production is halved compared to scenario C, but variability is reduced (Table 3). Patterns of geographical allocation are similar between the different types of optimization.

5 CONCLUSION AND FUTURE WORK

Expanding on the preliminary results previously reported in [Chen et al. 2024c; Guilpart et al. 2022], this work presents a multi-objective optimization approach to the problem of allocating a crop among thousands of sites located in a large area, aiming at simultaneously maximizing annual average production and stability, while minimizing the area allocated to the crop. The case study presented here concerns the allocation of soybean, a major crop for protein supply, over the European continent. We use a high-resolution grid, a machine learning model to predict annual yield for each site-year, and an established MOEA to generate a non-dominated set of candidate solutions. In order to assess its efficiency, the MOEA is tested against an ad-hoc quadratic optimization algorithm able to deliver non-dominated points with a guarantee of optimal results. An analysis of the Pareto fronts allowed us to identify contrasting and relevant crop allocation scenarios.

The comparison between MOEA and quadratic optimization leads to interesting results. The relatively simple scalarization approach matches the performance of NGSA-II with two objectives and is even better when considering three objectives, at least in terms of diversity of the front. This is not unexpected, given that the underlying objectives are tractable. However, even though quadratic optimization is in polynomial time, its scalability remains an open question, especially because the number of points that must be generated to achieve good coverage of the front increases exponentially with the number of objectives to optimise. This suggests that tackling an expanded geographical region with multiple species of crops may require several days of computation. However, our results show that it is useful to take advantage of tractable optimization problems when this is possible.

In the present study, soybean production is allocated based on three objectives (i.e., crop annual productivity, stability, and surface), all following a quadratic function (Figure 3). In the future, the optimization framework proposed here could be enhanced by incorporating additional optimization criteria related to environmental and economic factors, such as price and agronomic inputs. To effectively interpret the results, optimizing more than three criteria would require more advanced visualization methods.

Future research could focus on the simultaneous allocation of multiple crop species to propose scenarios for more diversified agronomic systems. Crop diversification, which involves increasing the spatial and temporal diversity of crops, can help preserve biodiversity, minimize topsoil erosion, and even enhance overall production [Beillouin et al. 2021]. For example, intercropping maize and soybean in the same field is suggested as a promising strategy to improve soybean self-sufficiency in the EU27 while maintaining domestic maize production [Chen et al. 2024b]. Additionally, further work is needed to determine if the approach developed here for soybean can be transferred to other crops. In the context of maize-soybean intercropping, it might be possible to derive the non-dominated front for maize from the front obtained for soybean. However, this approach could introduce bias in the optimization by ignoring optimal solutions that are more distant from the initial front. A specific optimization would thus probably be relevant in this case.

Multi-objective optimization is an promising tool to explore scenarios of crop allocation which are robust to different climate scenarios. Indeed, one solution could be optimal in current climate conditions, but not anymore in future conditions due to climate change. It will be thus useful to compare the solutions obtained when optimizing crop allocation in different climate scenarios [Guilpart et al. 2022]. Alternatively, it would be possible to consider the difference between current and future optima as a fourth criterion to optimize. This would allow us to find future crop allocations that are not too different from current crop allocations. In the context of climate change, it could also be interesting to consider the potential application of dynamic optimization, where the objective function evolves over time. This approach could provide a more realistic representation of the system

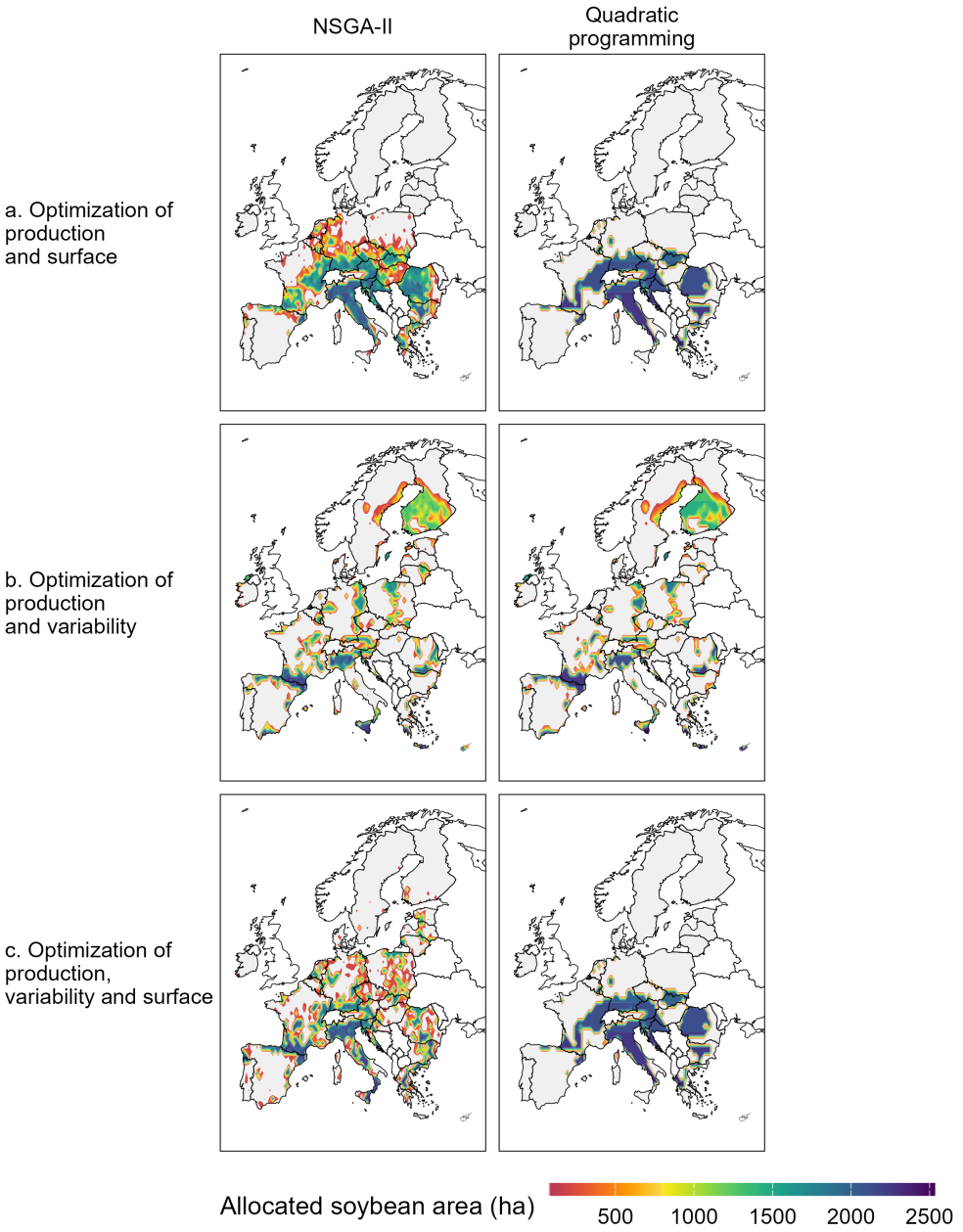

Fig. 7. Geographical allocation of soybean in Europe in the scenario D (current surfaces), according to the optimization setting

by accounting for temporal variations of climate. However, the feasibility of implementing dynamic optimization in our specific situation needs to be carefully evaluated. Additionally, it will be crucial to conduct a sensitivity analysis to understand the robustness of the optimal solutions under future climates. By altering the objective function, one can determine whether the current solution remains optimal and assess the magnitude of changes required to either maintain or shift the optimal solution. This analysis would offer valuable insights into the robustness of our optimization framework under varying climate conditions.

REFERENCES

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

Nishu Bali and Anshu Singla. 2021. Emerging Trends in Machine Learning to Predict Crop Yield and Study Its Influential Factors: A Survey. *Archives of Computational Methods in Engineering* 29, 1 (March 2021), 95–112. https://doi.org/10.1007/s11831-021-09569-8

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.

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

Damien Beillouin, Tamara Ben-Ari, Eric Malézieux, Verena Seufert, and David Makowski. 2021. Positive but variable effects of crop diversification on biodiversity and ecosystem services. *Global Change Biology* 27, 19 (July 2021), 4697–4710. https://doi.org/10.1111/gcb.15747

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

J. Blank and K. Deb. 2020. pymoo: Multi-Objective Optimization in Python. *IEEE Access* 8 (2020), 89497–89509.

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

Mathilde Chen, Nicolas Guilpart, and David Makowski. 2024a. Comparison of methods to aggregate climate data to predict crop yield: an application to soybean. *Environmental Research Letters* 19, 5 (May 2024), 054049.

Mathilde Chen, Nicolas Guilpart, and David Makowski. 2024b. High potential contribution of intercropping to soybean and maize self-sufficiency in Europe. *bioRxiv* (2024), 2024–12.

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

Mathilde Chen, David Makowski, and Alberto Tonda. 2024c. Multi-Objective Optimization for Large-scale Allocation of Soybean Crops. In *Proceedings of the Genetic and Evolutionary Computation Conference* (Melbourne, VIC, Australia) *(GECCO '24)*. Association for Computing Machinery, New York, NY, USA, 1174–1182. https://doi.org/10.1145/3638529.3654026

Kalyanmoy Deb. 2001. *Multi-objective optimization using evolutionary algorithms*. Wiley Interscience Series in Systems and Optimization, Vol. 16. John Wiley & Sons, Hoboken, NJ. Fundamental citation for MOEAs.

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

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.

Kalyanmoy Deb, Karthik Sindhya, and Jussi Hakanen. 2016. Multi-objective optimization. In *Decision sciences*. CRC Press, London, England, 161–200.

S Dogliotti, MK Van Ittersum, and WAH Rossing. 2005. A method for exploring sustainable development options at farm scale: a case study for vegetable farms in South Uruguay. *Agricultural systems* 86, 1 (2005), 29–51.

EFSCM. 2023. *State of Food Security in the EU (Autumn 2023). A qualitative assessment of food supply and food security within the framework of the EFSCM.* Technical Report.

Gabriele Eichfelder. 2021. Twenty years of continuous multiobjective optimization in the twenty-first century. *EURO Journal on Computational Optimization* 9 (2021), 100014. https://doi.org/10.1016/j.ejco.2021.100014

FAO. 2024. Food Balances (2010-). https://www.fao.org/faostat/en/#data/FBS "[Online; accessed 8-December-2024]".

R FAOSTAT et al. 2024. FAOSTAT database. *Food Agric. Organ. UN* (2024).

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 (March 2021), 295. https://doi.org/10.3390/land10030295

Aaron Garrett. 2012. inspyred. https://github.com/aarongarrett/inspyred. Accessed January 2024.

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

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

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.

Trevor Hastie, Robert Tibshirani, Jerome Friedman, Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. Boosting and additive trees. *The elements of statistical learning: data mining, inference, and prediction* (2009), 337–387.

Peter Hoefsloot, Amor VM Ines, Jos C van Dam, Gregory Duveiller, Francois Kayitakire, and James Hansen. 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

Toshichika Iizumi and Toru Sakai. 2020. The global dataset of historical yields for major crops 1981–2016. *Scientific Data* 7, 1 (2020), 97.

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, NY, US, 2419–2426. https://doi.org/10.1109/cec.2008.4631121

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.

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

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

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.

Markus Kottek, Jürgen Grieser, Christoph Beck, Bruno Rudolf, and Franz Rubel. 2006. World map of the Köppen-Geiger climate classification updated. (2006).

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.

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.

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

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.

Andreas Löhne and Benjamin Weißing. 2017. The vector linear program solver Bensolve – notes on theoretical background. *European Journal of Operational Research* 260 (2017), 807–813. Issue 3.

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.

David Makowski, Eligius MT Hendrix, Martin K van Ittersum, and Walter AH Rossing. 2000. A framework to study nearly optimal solutions of linear programming models developed for agricultural land use exploration. *Ecological Modelling* 131, 1 (2000), 65–77.

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

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* (Boston, Massachusetts) *(GECCO '22)*. Association for Computing Machinery, New York, NY, USA, 352–355. https://doi.org/10.1145/3520304.3528898

Daniel Mueller-Gritschneder, Helmut Graeb, and Ulf Schlichtmann. 2009. A Successive Approach to Compute the Bounded Pareto Front of Practical Multiobjective Optimization Problems. *SIAM Journal of Optimization* 20 (2009), 915–934. Issue 2.

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.

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.

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

Luigi Pellizzoni, Laura Centemeri, Maura Benegiamo, and Carla Panico. 2024. A new food security approach? Continuity and novelty in the European Union's turn to preparedness. *Agriculture and Human Values* (Oct. 2024), 1–17.

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.

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.

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.

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.

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.

Andrzej P. Wierzbicki. 1982. A mathematical basis for satisficing decision making. *Mathematical Modelling* 3, 5 (1982), 391–405. https://doi.org/10.1016/0270-0255(82)90038-0 Special IIASA Issue.

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.

Eckart Zitzler, Dimo Brockhoff, and Lothar Thiele. 2007. The hypervolume indicator revisited: On the design of Pareto-compliant indicators via weighted integration. In *4th International Conference on Evolutionary Multi-Criterion Optimization (EMO)*. Springer, 862–876.

Received 15 December 2024; revised 1 June 2025; accepted 4 June 2025