

A Service of

ZBW

Leibniz-Informationszentrum Wirtschaft Leibniz Information Centre for Economics

Troost, Christian; Parussis-Krech, Julia; Mejaíl, Matías; Berger, Thomas

Article — Published Version Boosting the Scalability of Farm-Level Models: Efficient Surrogate Modeling of Compositional Simulation Output

Computational Economics

Provided in Cooperation with: Springer Nature

Suggested Citation: Troost, Christian; Parussis-Krech, Julia; Mejaíl, Matías; Berger, Thomas (2022) : Boosting the Scalability of Farm-Level Models: Efficient Surrogate Modeling of Compositional Simulation Output, Computational Economics, ISSN 1572-9974, Springer US, New York, NY, Vol. 62, Iss. 3, pp. 721-759, https://doi.org/10.1007/s10614-022-10276-0

This Version is available at: https://hdl.handle.net/10419/306653

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.



WWW.ECONSTOR.EU

https://creativecommons.org/licenses/by/4.0/

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.





Boosting the Scalability of Farm-Level Models: Efficient Surrogate Modeling of Compositional Simulation Output

Christian Troost¹ \cdot Julia Parussis-Krech¹ \cdot Matías Mejaíl¹ \cdot Thomas Berger¹

Accepted: 9 May 2022 / Published online: 30 August 2022 $\ensuremath{\mathbb{C}}$ The Author(s) 2022

Abstract

Surrogate modeling can overcome computational and data-privacy constraints of micro-scale economic models and support their incorporation into large-scale simulations and interactive simulation experiments. We compare four data-driven methods to reproduce the aggregated crop area response simulated by farm-level modeling in response to price variation. We use the isometric log-ratio transformation to accommodate the compositional nature of the output and sequential sampling with stability analysis for efficient model selection. Extreme gradient boosting outperforms multivariate adaptive regressions splines, random forest regression, and classical multinomial-logistic regression and achieves high goodness-of-fit from moderately sized samples. Explicitly including ratio terms between price input variables considerably improved prediction, even for highly automatic machine learning methods that should in principle be able to detect such input variable interaction automatically. The presented methodology provides a solid basis for the use of surrogate modeling to support the incorporation of micro-scale models into large-scale integrated simulations and interactive simulation experiments with stakeholders.

Keywords Meta-modeling · Agent-based modeling · Mathematical programming · Farm-level simulation · Fractional response

1 Introduction

Data-driven surrogate models are representations of the response surface of complex simulation models. They are estimated from a sample of input-output mappings simulated with the original model (van der Hoog, 2019; Kleijnen, 2017; Asher et al., 2015).

Christian Troost christian.troost@uni-hohenheim.de

¹ Hans-Ruthenberg-Institute, University of Hohenheim, Wollgrasweg 43, 70599 Stuttgart, Germany

Data-driven surrogate modeling can considerably enhance the applicability of micro-scale (farm-level) agricultural economic models, such as whole-farm and agent-based models. Farm-level models (FLM) simulate the management decisions of a collection of individual farm holdings allowing for detail and complexity, *inter alia* in representing heterogeneous farming conditions, preferences, risk behavior and technology. They are well-established tools to simulate fundamental changes in farm production and resource use choices in response to regime-shifting drivers, such as climate change, technological innovation or policy intervention (Kremmydas et al., 2018; Reidsma et al., 2018; Troost & Berger, 2015; Berger & Troost, 2014; van Wijk et al., 2014).

Recently, there has been an increased interest in linking such detailed micro-scale behavioral models into macro-scale integrated assessment models (Brown et al., 2021; Müller et al., 2020; Lippe et al., 2019; Müller-Hansen et al., 2017; van Wijk, 2014). This would, on the one hand, enhance global assessments with high-resolution models of complex and heterogeneous decision-making—overcoming overly restrictive behavioral assumptions for which current aggregate macro-models have been criticized. On the other hand, it would address the major constraint of micro-scale models: Even when implemented for all farm holdings in a study area, they are usually confined to smaller spatial extents and cannot simulate larger-scale feedback to local decisions. In economic terms, this implies a strict small-country assumption applied not only to product and resource markets, but also to global environmental effects. Neglecting such global feedback may result in biased simulation outcomes and misleading conclusions (Müller et al., 2020; Müller-Hansen et al., 2017).

Data-driven surrogate models can help overcome three challenges associated with the incorporation of micro-scale models into large-scale assessments: (i) Computational time demands of micro-scale models for a substantial number of farm agents are often prohibitive for frequent iterative reevaluations required for equilibrium search in a macro-scale model. (ii) Integrating micro-scale models that have been developed by different specialized research groups for different small-scale locations into a large-scale framework is often hampered by heterogeneous use of modeling methods and software between research groups. (iii) Micro-scale models often employ sensitive, privacy-constrained micro data that cannot simply be shared for direct model reuse by other research groups (Troost & Berger, 2015).

A statistical surrogate model that has been estimated from representative output of a farm-level model can stand in for (emulate) the original farm-level model, where the use of the latter is not possible: A surrogate is typically computationally much cheaper than the original model, can serve as a unified interface to the response of heterogeneous model implementations, and isolates the response surface from privacy-constrained microdata. In addition, surrogate models also facilitate model optimization, uncertainty and sensitivity analysis, calibration and interactive model exploration of micro-scale models (Mössinger et al., 2022; van der Hoog, 2019; Lamperti et al., 2018; Kleijnen, 2017; Baustert & Benetto, 2017; Asher et al., 2015).

The usefulness of surrogate modeling of farm-level models has been exemplified in Happe et al. (2006), Domínguez et al. (2009), Lengers et al. (2014) and Seidel & Britz (2019). While these previous applications used classical econometric models, van der Hoog (2019) and Storm et al. (2020) suggest the use of machine-learning methods for surrogate model estimation as demonstrated e.g by Lamperti et al. (2018) for financial economic applications.

In this article, we systematically evaluate the capacity of four different surrogate modeling approaches including machine learning methods to capture the aggregate crop area response to price variation as simulated by a farm-level model. We address two important methodological gaps in the nascent literature on surrogate modeling of micro-scale agricultural economic models:

Firstly, we demonstrate a sequential sampling and evaluation design to ensure robust and efficient estimation of surrogate models from a limited sample of agricultural economic model outputs. We combine a quasi-random low discrepancy sequence with convergence and stability assessments to address the inevitable tradeoff between a comprehensive coverage of the variation in original model output and the computational cost of original model evaluations implied by larger sample sizes.

Secondly, part of the micro-scale model output is not continuous, but compositional: Land use and crop areas, which are essential indicators for many environmental and economic assessments, must sum up to the total available land in the area. At the same time, response to external drivers is highly complex and characterized by variable interactions and nonlinearities. To capture this compositional response, we combine the isometric log-ratio transformation *(ilr)* suggested by Egozcue et al. (2003) with multivariate adaptive regression splines (MARS) (Friedman, 1991), random forest regression (RF) (Breiman, 2001), and extreme gradient boosting (XGB) (Chen & Guestrin, 2016). We compare these methods with multinomial-logistic regression (MNL) as one of the classical statistical methods for categorical and compositional data analysis.

2 Data and Methods

Figure 1 summarizes the basic concept underlying the use of surrogate models to interpolate or predict the simulation output of micro-scale models: A farm-level model (FLM) provides a theory-based representation of farm-economic decisions, which allows for simulation analysis of structural changes and regime shifts in the socioeconomic and biophysical conditions of farming. Simulating a sample of model input-output combinations provides a data base for estimating a surrogate model (SM). This surrogate model predicts the FLM output for input combinations that have not been simulated yet, based on the estimated statistical relationships between FLM model inputs and outputs in the sample. In this way, the SM allows for a quick



Fig. 1 The basic concept: Using a surrogate model to interpolate or emulate output of a farm-level model

interpolation of FLM results for uncertainty analysis, iteration with larger-scale models or interactive result exploration.

We first describe the FLM that we use as test case in this article and the model input and output variables that we varied. We then explain our evaluation criteria and the experimental design we employed for comparing different surrogate modeling methods. In the last part of this section, we present the four different surrogate modeling approaches we examined.

2.1 The Farm-Level Model

Our analysis used the MPMAS Central Swabian Jura (MPMAS-CSJ) model that was developed by Troost & Berger (2015) to simulate agricultural adaptation to climate change in the Central Swabian Jura, a low mountainous area in Southwest Germany. MPMAS-CSJ simulates the production decisions of all full-time farmers in the study area by solving a mixed integer programming (MIP) problem for each farm agent in order to allocate the production factors (land, labor and capital) such that they maximize expected farm income while respecting individual-specific resource constraints and production options of each farm agent as well as sales and input prices, and the technical and agronomic constraints governing agricultural production.¹ MPMAS-CSJ represents about 530 farm holdings with about 36,000 ha of agricultural land of which about 22,000 ha are arable. It covers the full range of activities of the mixed-crop-livestock farming systems in the area such as crop production, dairy farming, bull fattening, pig production, and biogas production including the necessary investments in machinery and buildings and participation in agri-environmental measures. The spatial resolution of input data provided to the model is 1 ha, but agents can allocate crop production in arbitrarily small units. For an in-depth description of model equations, empirical parameterization, validation and uncertainty testing, please refer to Troost & Berger (2015) and Troost et al. $(2015)^2$

In MPMAS-CSJ, the interplay of agronomic constraints, soil-specific yields, crop utilization options (sale, forage, fermenter feedstock), agri-environmental support schemes, flexible allocation of machinery to field work, and flexible feed composition in animal nutrition leads to complex, heterogeneous and interdependent profitability profiles for the different crops: On the one hand, production options may complement each other, for example, as compatible or even mutually beneficial members of crop rotations, by providing complementary nutrients for animal feeding, by requiring field work at different points of time in the year, or by requiring the same type of agricultural equipment. On the other hand, they may compete with each other if they are incompatible in rotation, are substitutes in animal nutrition, coincide in peak labor demand or require very different agricultural equipment. This means

¹ Although the model has been implemented using the agent-based software package MPMAS (Schreinemachers & Berger, 2011), the model application in the present article does not include any agent-agent interactions. In keeping with MPMAS conventions, we use the term farm agent to refer to the model representation of an individual real-world farm holding throughout this article.

² The electronic supplement to Troost & Berger (2015) provides a full documentation of all equations in MPMAS-CSJ.

that changes in prices, yields, timing or cost of one crop may potentially affect the relative profitability of other crops, and that these effects may considerably differ between farm holdings and at different prices levels. The response surface representing the effects of price expectations on crop area choice was hence expected to be strongly nonlinear and characterized by structural breaks and segmentation.

2.2 Input and Output Variables

In the present article, we focus on the simulation of regional totals of crop areas on arable land as the output of interest, as would be relevant for integration into a large-scale assessment model. Regional totals result from the aggregation of simulated choices of the individual farm holdings.³ Nine crop categories were considered for arable production: winter wheat, spring malting barley, winter barley, spring fodder barley, winter rapeseed, silage maize, green winter wheat (for silage), clover/grass production on arable land, and fallow. Accordingly, for each simulation run we obtained J = 9 output variables indicating the amount of area that was simulated to fall into the respective crop category (j = 1, ..., J) in that run.

With respect to model inputs, we focused on variations in product and input price expectations used by farm agents for production planning. Since we simulated farm agent decisions for one average year for the present analysis, our MPMAS-CSJ application does not endogenously simulate formation of expectations: price expectations are set exogenously.⁴ We vary price expectations over the full range of potential price combinations for crops, animal products and important inputs and apply these to the baseline climate scenario of Troost & Berger (2015).⁵ Variation in the price for good g was expressed as coefficients (pc_g) relative to the 2000-2009 price average (\bar{P}_g) . For the simulations, we extended the ranges observed between 2000-2009 by about 20-30% at both ends, to capture potentially more extreme price relations in the future (cf. Table 1). As our simulations are intended to provide the basis for the estimation of surrogate models, we are interested in efficiently attributing changes in model outputs to changes in model inputs and their interaction. To avoid confounding of input factor effects, we varied price expectations for the individual items independently from each other, ignoring the existing correlations. This means that the resulting distribution of model outcomes should not be interpreted as a probability distribution for the crop choice. (A probability

³ MPMAS-CSJ simulates the individual crop choices as part of a wider, more comprehensive farm management decision. Since the structural scenario simulated with the MPMAS-CSJ model does not allow grassland to arable land conversion, the total arable area per farm agent is a fixed quantity and we could focus our analysis on this part of the MPMAS-CSJ output vector.

⁴ In principle, expected and actual prices differ: a price observed in one year does not necessarily mean farmers will use this price for planning for the following season. Rather price expectations used in planning are typically understood to reflect a combination of past price observations and currently obtainable information about future supply and demand. For the present analysis that focuses on land use decisions for one simulated year representing average conditions (rather than, for example, actual income or developments over a series of several years), only expected prices are relevant. Whenever we discuss "prices" in the present article, we refer to expected prices unless explicitly mentioned.

⁵ The baseline climate scenario used yields and field work timings as observed in the area around 2010.

Item	Symbol		Range
Beef and young cattle	pc_{beef}		[0.7, 1.3]
Fertilizer	pc_{fert}		[0.5, 2.0]
Fuel and energy	pc_{fuel}		[0.7, 1.5]
Milk	pc_{milk}		[0.7, 1.2]
Pork and pigs	pc_{pork}		[0.7, 1.3]
Wheat	pc_{wh}		[0.5, 2.0]
Ready-mix animal feed	pc_{fodd}		[0.7, 1.5]
Grain maize	pc_{mg}		[0.7, 1.6]
Malting barley	pc_{mb}		[0.5, 2]
Fodder barley	pc_{fb}		[0.7, 2]
Rapeseed	pc_{wr}		[0.7, 1.5]
Parameter		Symbol	Values
Scaling parameter for the amount of labor required for fresh grass harvest		fgl	[1, 3]
Demand for excess heat of biogas electricity production		chp	{yes, no}
Scaling parameter for the maximum manure application		man	[1, 1.5]
Supply of brewery byproducts for feeding		bbp	{yes, no}
Births scaling factor in the past (affects current household size)		bfp	{1, 1.05}
KTBL climatic region for time slots of field work		clr	{4, 5}
Probability that a male child is interested in taking over the farm		scm	$\{0.5, 1\}$
Probability to be able to hire a machinery service provider per day with suitable weather		pth	[0.5, 2]
Scaling parameter for the maximum wheat yield		ywh	[1, 1.1]
Price for hiring machinery services (location between min and max reported)		wfh	[0, 1]
Chosen starting population		pop	set of 3

Table 1 Price coefficients used for price variation in the experimental design and unfixed parameters representing uncertainty in MPMAS-CSJ

cf. Troost & Berger (2015) for a more in-depth explanation of uncertain parameters

distribution for crop areas could be generated from the output sample by weighting each simulation run with the joint probability of input factor values employed in that run or, alternatively, by later applying the estimated surrogate model to a sample of correlated price factors.)

In addition to price variations, we also included variations of uncertain model parameters into our analysis: Farm-level models, such as MPMAS-CSJ, that are based on theoretical and empirical process knowledge are typically subject to considerable parameter and input uncertainty (Buysse et al., 2007; Troost & Berger, 2015). Good modeling practice requires to clearly communicate the resulting uncertainty to readers and analyze it in order to assess the robustness of results and, in the long run, improve process understanding (Jakeman et al., 2006). As Berger & Troost (2014) argue, one should therefore refrain from identifying a single parameter

combination that best fits observation data. For MPMAS-CSJ, Troost & Berger (2015) reduced the model parameter space only, where clearly superior settings could be determined in specially designed calibration experiments that tested parameter combinations against three structurally different observation years. They then used an elementary effects screening (Campolongo et al., 2007) to determine 11 of the 19 unfixed parameters that caused the greatest variance in simulated differences between climate and price scenarios (cf. Table 1). As a consequence, MPMAS-CSJ had to be solved repeatedly to cover the space spanned by these 11 parameters and model results were communicated as ranges over this parameter space.

With 11 price expectation and 11 uncertain model parameters we obtained L = 22 input variables overall. Accordingly, the model input-output dataset that formed the basis for our surrogate model estimation consisted of J + L = 31 variables (or columns), with each row corresponding to one simulation run. The number of simulation runs and hence rows and the choice of input factor settings for each row is discussed in the following section.

2.3 Experimental Design and Evaluation Criteria

We assumed that the main objective of using a surrogate model is to predict simulation model output outside of the sample of input combinations simulated with the original FLM.

2.3.1 Sequential Experimental Design

As the form of the multivariate model output distribution is typically unknown and not following common parametric distributions, the sample size (N_{estim}) necessary to estimate robust and unbiased surrogate models cannot be calculated a priori (Lee et al., 2015). At the same time, the number of original model evaluations needed to generate the sample is an important determinant of the computational burden for creating a surrogate model $t_{gen,SM}$ (Eq. 1), which consists of the time needed to estimate the surrogate model from a given FLM input-output dataset $t_{estim,SM}(N_{estim})$, but also the time needed to create this sample, which is a product of N_{estim} and the runtime of the original model $t_{sim,FLM}$.

$$t_{gen,SM} = t_{estim,SM}(N_{estim}) + t_{sim,FLM} \cdot N_{estim}$$
(1)

Hence, on the one hand, modelers will in practice try to keep N_{estim} as small as possible. On the other hand, a smaller sample will inevitably lead to higher sampling error and stronger confounding of input factor effects, and increase the danger of overfitting and unstable surrogate model estimates. Specifying a value for N_{estim} that is generally sufficient for all FLM and SM combinations is hardly imaginable, as complexity of FLM responses will differ considerably from application to application, and hence a suitable N_{estim} has to be identified for each FLM-SM application specifically.

A pragmatic procedure is to gradually increase the number of FLM simulation runs in batches. One can then estimate the surrogate model on the batches simulated so far and subsequently measure the marginal improvements of expected predictive performance and stability of estimated models in a separate validation sample (crossvalidation). One can then stop if a sufficient predictive accuracy and model stability has been reached, no further improvement can be seen or a computational resource constraint on the total number of FLM evaluations has been reached.

In this article, we implemented such an incremental procedure based on a Sobol' sequence over L = 22 input factors.⁶ Sobol' sequences are S-optimal experimental designs that like Latin hypercube samples (LHS, e.g. Salle & Yıldızoğlu, 2014) ensure representative coverage of a parameter space when computational requirements limit the number of simulation runs. Their advantage over LHS is that they are more easily extended or reduced in size because the location of design points in the multidimensional parameter space remains the same for different sequence lengths. This means that extending a Sobol' sample just requires simulating the additional repetitions, whereas a larger LHS requires a complete resimulation of all repetitions of the new sample size (Tarantola et al., 2012).

Our incremental procedure started by using the first 500 elements of the Sobol' sequence as a training sample (TS) and the following 200 as a cross-method validation sample (VS1). (Note: We use the term training sample here to denote the full sample size provided to each surrogate modeling method at a certain iteration. Nonetheless, each surrogate modeling method may treat part of this full training sample as an intra-method training sample and part of it as an intra-method validation sample.)

We successively increased the TS size in nine steps (over 750, 1000, 1500, 2000, 2500, 3000, resp. 4000 design points) until including the first 5000 design points of the Sobol' sequence. The 200 design points following each TS were used as cross-method validation sample VS1. At each iteration, we used VS1 to assess the convergence of the predictive performance as well as the stability in prediction and surrogate model structure compared with previous estimations.

2.3.2 Expected Predictive Performance

As described above, the outcome for each simulation run k consisted of a vector of J = 9 crop categories, into which the total arable area was classified. The accuracy of the surrogate model prediction for the distribution of the area over these nine categories was measured by the share of correctly classified simulated area for this run k:

⁶ The Sobol' sequence of dimension 22 and length 10,000 was created using the runif.sobol function of the *R* package *fOptions* (Wuertz et al., 2021). As common in the sensitivity analysis literature, e.g. Saltelli et al., 2004, we use the term 'input factors' when we refer to these 22 parameters for which the Sobol' sequence was formed.

$$Scc_{k} = 1 - \frac{\sum_{j} |AreaSM_{j,k} - AreaFLM_{j,k}|}{2 \cdot \sum_{j} AreaFLM_{j,k}}$$
(2)

with $AreaSM_{j,k}$ the area in crop category *j* predicted by the surrogate model and $AreaFLM_{j,k}$ the actual area for crop category *j* simulated by the original FLM.⁷

We then assessed the predictive performance for a full input-output sample of K simulation runs by calculating the average and also the worst-case (minimum) *Scc* over the whole sample.

$$avgScc = \frac{\sum_{k}^{K} Scc_{k}}{K}$$
(3)

$$minScc = min\{Scc_1, \dots, Scc_K\}$$
(4)

2.3.3 Stability of Predictions and Feature Importance

It is important to not only assess the development of predictive accuracy of the model, but also its stability: Does the prediction for an input combination k vary whether the SM has been estimated from a larger or smaller sample? The same type of model with the same functional form and hyperparameter settings, once estimated from a small training sample and once estimated from a slightly larger training sample, might predict very different outcomes and show different deviations from perfect prediction, even if both versions have a similar goodness-of-fit (as long as the fit is not perfect). And even if the predicted outcome is stable: Is the influence attributed to a certain input factor stable? Surrogate model predictions or input factor importance rankings that fluctuate strongly over training sample sizes indicate a strong influence of training sample size and hence likely overfitting. They do not project confidence in the use of the surrogate model for sensitivity analysis.

To assess stability of predictions, we did a pairwise comparison of the predictions for each k in VS1 between the model estimated from a certain *TS* and the same model estimated from the previous *TS*. As our dependent variable is a composition, we used Aitchison's *total variance* (Pawlowsky-Glahn & Egozcue, 2001) to measure the variation in prediction between models estimated from the adjacent *TS* at each k. We then plotted the distribution of total variances over the *K* sample points in VS1. (A more detailed explanation including formula is provided in the Appendix Sect. B.) We expect the differences in prediction caused by the information added through increasing a sample size (and hence also the variance between adjacent sample sizes) to decline with increasing sample size as confounding between input factor effects reduces with increasing length of the Sobol' sequence.

⁷ Note: The measure corresponds to the traditional classification accuracy (proportion of correctly classified observations), e.g. used in logistic regression, here calculated from category totals rather than individual classification results. The division by two is necessary to avoid double-counting of misclassified areas, since given the fixed total arable area any overestimation of area in one crop category will necessarily also appear as an underestimation in other categories.

To assess stability of input factor influence in a comparable way across methods, we followed the permutation feature importance approach of Breiman (2001): to calculate the importance of input factor l in estimated model m, we randomly permuted the values of this input factor across simulation runs to break the relationship of this factor with the outputs, and re-estimated m on this permutation. The difference between the resulting model accuracy and the original model accuracy is regarded as a measure of input factor importance. Input factors were then ranked by importance and the ranking subjected to factorwise comparison with the ranking obtained at the previous, smaller *TS*.

2.3.4 Meta-Evaluation of Sequential Approach

While the previous steps would be performed in any practical application of the approach, for this article we did an additional step to evaluate the performance of the whole workflow: We evaluated performance of the surrogate models that were selected based on TS and VS1 in an additional 5000-element validation sample (VS2) corresponding to design points 5001-10000 of the Sobol' sequence. In practice, we do not expect modelers to generate a second VS with the original FLM. Rather, it here represents input factor combinations for which modelers expect to receive a good prediction by the surrogate model without having simulated them with the FLM.⁸ With this meta-validation on VS2, we evaluated in how far our iterative estimation and validation strategy (employing TS and VS1) has generated a generalizable surrogate model that provides robust interpolation for a wider sample.

2.4 Surrogate Modeling Methods

We explored the capacity of four different modeling methods to capture the compositional output of the farm-level model.

2.4.1 Multinomial-Logistic Regression

As a classical regression method, we employed *multinomial-logistic regression* (*MNL*), more specifically a baseline-category logistic model with observationspecific regressors and grouped data (Agresti, 2013)—making use of the fact that under certain simplifying assumptions, compositional output data such as simulated crop areas can be understood as multinomial data.

Estimation of a classical MNL model on our dataset was complicated by strong interaction and segmentation effects in the input-output mapping between prices and crop areas simulated by the FLM. Mutual price relations, e.g. the price ratio between product sales prices and production input prices and the sales price ratios between different products have a strong influence on crop choice. Effects of prices are often nonlinear and can be subject to breaks or shifts, i.e. show a segmented response: For

⁸ The typical purpose of a surrogate model is to predict outcomes not yet simulated with the original model. Subsequent extensions of the Sobol' sequence add a sample of the points between and around the space-filling sample of points already simulated, hence a sample of the points one might possibly like the surrogate model to predict.

example, there may be a specific ratio of their respective price expectations where one crop becomes more profitable than the other for a significant group of farm agents. Crop areas respond very differently to price variation above or below such a threshold ratio. The location of such segmentation points is not known *a priori* but must be determined during the analysis. Automatic backward and forward selection of regressors and interaction terms can hardly remedy these challenges as the possible locations of segmentation points would still have to be manually prespecified.

In our MNL application, we dealt with these challenges by manually and iteratively examining residuals and expanding functional forms. To avoid overfitting, we used Akaike's Information Criterion (AIC) to decide between different proposed model structures. (See Sect. A.1 of the Appendix for the details on MNL model selection and estimation in this study.)

2.4.2 Non-parametric, Machine Learning Methods

As an alternative, we tested three non-parametric, machine-learning methods, which are designed to largely avoid the pre-specification of functional forms and detect input factor interactions and segmentations automatically: *Multivariate Adaptive Regression Splines (MARS)*(Friedman, 1991), *Random forest regression (RF)* (Breiman, 2001; Hastie et al., 2009 and *Extreme gradient boosting (XGB)* (Chen & Guestrin, 2016).

Fitting non-parametric methods always has to strike a balance between a better fit and the danger of overfitting to the sample, which in all three methods is controlled through internal cross-validation and hyperparameters that define limits on model complexity, learning rates, or minimum improvement thresholds. These hyperparameters can be tuned for a specific application. Details on the methods and associated hyperparameter search and cross-validation procedures can be found in the Appendix (A.3, A.4, A.5).

While we expected the non-parametric methods to detect the effects of input factor interactions such as price ratios automatically, we tested this capacity by estimating them with two different input factor sets, from which the algorithms could freely form predictor terms: In the first set, *without ratios*, we only included the 22 price coefficients and uncertain input factors. In the second set, *with ratios*, we explicitly also included the ratios of all other price factors to the wheat price and to the fodder barley price leading to a set of 22 + 19 = 41 input variables. It is important to note that the second set did in principle not provide more information to the algorithm than is implicitly contained in the price coefficients of the first set already.

All three methods were originally designed for continuous, unconstrained data and are not per se suitable for use with compositional data. We therefore applied the *isometric-log-ratio transformation (ilr)* (Egozcue et al., 2003) to transform our *J*-dimensional compositional simulation output into an unconstrained, continuous J - 1 Euclidean vector space, on which statistical approaches designed for continuous data can be used. (Outcomes predicted by the surrogate models were then backtransformed into the compositional simplex space by inverse *ilr*).

Compared with other transformation methods such as the softmax function, the additive log-ratio transformation, or the centered log ratio transformation, the *ilr* transformation is, at the same time, symmetric, isometric (preserves distances) and subcompositionally consistent (although not unique), which makes it the theoretically best choice for input data estimation (Egozcue et al., 2003). (For more details see the Sect. A.2 in the Appendix).

As all logarithmic transformations, the *ilr* only works with strictly positive data, whereas our dataset contained a non-negligible amount of zeros. We worked around this problem by adding a small quantity (= one hectare) to all categories in all runs. With area totals around 25,000 hectares, the associated distortion was expected to be minimal.

3 Results

3.1 Predictive Performance and Model Selection

Figure 2 provides an overview of the predictive performance of the best-performing surrogate model candidate generated by each surrogate modeling method for each training sample size. As a benchmark, we indicate the predictive accuracy of using the mean crop area shares of the respective training sample as a predictor (null model), which achieved a share of correctly classified plots of 0.69 (without any major changes over the training sample sizes). All surrogate modeling methods achieved considerably higher predictive accuracy surpassing 0.9 for a sufficiently large TS size. There is a clear ranking of methods with XGB achieving up to 0.96, followed by RF and MNL with 0.94 and 0.93, respectively. The best MARS model reached a somewhat lower Scc of 0.91.

Even at the lowest TS size of 500 the predictive accuracy was above 0.9, except for MARS which started with 0.875. Performance response to training sample size increases was strongest below 1000 sample points and then flattened out, though smaller improvements are still visible up to 5000 (except for MNL). The ranking of methods remained stable over all TS sizes. The average predictive performance on VS2 (the meta-validation sample) hardly differed from performance in VS1. Worst-case performance was expectedly worse in VS2 compared with VS1. This is not surprising as VS2 was much larger and hence more likely to contain an extreme combination of input values. Notably, however, the degradation of worst-case performance from VS1 to VS2 was smaller for XGB and RF than for MARS, which shows the most inferior worst-case performance, even worse than the benchmark of using the sample average crop areas for prediction.

3.2 Functional Forms and Hyperparameter Choices

Detailed outcomes of the model (respectively hyperparameter) selection process for each modeling method can be found in the appendix. Three observations common to all methods are worth highlighting:



Fig. 2 Predictive accuracy of the best performing hyperparameter setting, resp. functional form, by surrogate modeling method and training sample size in validation samples VS1 and VS2. (All explicitly including price ratios in the set of independent variables)

(i) Complexity: Matching the FLM input-output mapping required complex surrogate models. The optimal number of trees converged around 512 trees for RF and between 256 and 512 trees for XGB. The chosen MNL model comprises 720 coefficients. For MARS allowing interactions of degree 2 increased predictive



Fig. 3 Comparison of predictive accuracy of the best performing hyperparameter setting, resp. functional form, WITHOUT and WITH explicit inclusion of price ratios as input factors by surrogate modeling method and training sample size

accuracy by 0.05 compared with no interactions. (Allowing higher degrees of interaction did not noticeably improve performance in the case of MARS.)

(ii) Explicit inclusion of price ratios: The explicit inclusion of price ratios as input factors for estimation considerably improved predictive performance. The left pane in Fig.3 shows the lower predictive accuracy of the best models without explicitly included price ratios. Predictive accuracy decreased by more than 0.05 in the case of MARS and MNL. For RF and XGB the effect decreased with TS size, but is still noticeable at TS=5000.

(iii) Robustness of selection: The ranking of best-performing choices for the key hyperparameters was mostly stable over increasing TS sizes. The best-performing degree of interaction (MARS), number of trees (RF) and functional form (MNL) performed best already at TS size 500 and ranking between choices hardly fluctuated further on. For XGB, at lower TS sizes a more restricted number of trees performed somewhat better, but the difference to the performance of the number of trees found optimal for larger TS sizes was rather small.

In all cases, predictive performance on VS2 was similar to VS1.

3.3 Stability of Predictions

Figure 4 illustrates the stability of predictions over increasing training sample sizes. Each violin graph shows how the total variance in prediction between surrogate models estimated from two adjacent TS sizes (indicated above the distribution) was distributed over the sample points in the VS1 associated with the larger of the two

TS. The sequence of violin graphs over increasing TS sizes depicts how the variance in prediction developed with increasing TS size. We can observe that the predictions of the best-fitting MNL model were extremely varying between training sample sizes, while the models found by the other methods showed much lower variance in prediction between estimation from different sample sizes. (Variation was slightly lower if using a simpler functional form for MNL, but still considerably higher than for the other methods.) Total variance decreased with increasing size of the training sample. This decrease is much stronger in MNL than in the other methods.

3.4 Stability of Input Factor Importance

Figure 5 shows differences in feature importance ranking between models estimated from adjacent TS sizes. Each violin graph shows the distribution of rank difference over the 41 input factors (22 original factors and explicit price ratios). A positive frequency at 10 means that at least one input factor moved in importance by ten ranks between the two TS sizes.

Similar to the variance in prediction also the variance in input factor importance was high for the estimated MNL functional form. In this case, however, also the MARS model showed considerable variation in ranking. XGB showed some variation in the lower TS sizes, where also the optimal number of trees varied. RF showed no variation at all.

4 Discussion

Our experiment was motivated by the question: "Can we estimate an accurate surrogate model to predict compositional simulation outcomes of a farm-level model from a moderately sized sample?" Overall, the results look very encouraging. All tested methods achieved a share of correctly classified area beyond 0.9, considerably better than the null benchmark of using the respective training sample average (0.69). There is a clear ranking of methods, with XGB performing best, followed by RF. XGB is able to achieve a very high fit of 0.96 at larger sample sizes and 0.93 already at lower sample sizes. Certainly, whether this level of expected predictive accuracy is sufficient and the differences between methods are practically relevant cannot be answered in general. It very much depends on the purpose for which the surrogate model is intended to be used in practice. It will also have to be seen in context with the uncertainty and inaccuracy of the original FLM itself.

4.1 Robustness

The reported performance of MNL has to be interpreted with care as it showed considerable instability in prediction and input factor effect estimation across samples. This underlines the importance of cross-validation and robustness diagnostics during training. While the other methods employed some form of cross-validation already in the training and selection process (GCV in MARS, inbuilt ensemble methods and explicit 3-fold cross-validation in RF and XGB), control of



Fig. 4 Total variance of prediction of the best selected model per method between adjacent training sample sizes. Each violin graph shows the distribution of Aitchison's total variance between models estimated from TS_s and TS_{s-1} over the runs of $VS1_s$. (Note: The vertical axis has a logarithmic scale)

overfitting relied only on comparing AICs for functional form selection in MNL. Importantly, this type of instability did not become apparent through a decay of performance from training to validation sample - an alternative check typically used to detect potential overfitting. (The difference between training sample and validation sample performance is stronger in RF than MNL, for example.)

Our experiment also highlights that already a small additional validation sample (here VS1) can provide a good estimate of performance further out-of-sample (here represented by VS2). We mainly attribute this to the representative sample structure ensured through the space-filling, increasingly dense representation of the input space achieved through the Sobol' sequence. (We emphasize that this validation sample does not make proper cross-validation during training obsolete, but serves as an additional control.)

4.2 The Importance of Domain Knowledge

The considerable performance gain achieved by explicitly including price ratios as predictors alongside the individual price coefficients was somewhat surprising for the automatic, non-parametric methods MARS, XGB and RF from which we would have expected automatic discovery of these interactions (unlike for manually-specified MNL). Although the effect may reduce for larger TS sample sizes, this strongly hints at the usefulness and importance of including domain knowledge, i.e. knowledge about the underlying processes, also in sample interpolation tasks, at least for restricted sample sizes.

4.3 Computational Demand and Minimal Sample Size

Runtime for a single prediction (t_{pred}) was about 0.1 s with RF and 0.003-0.005 s with the other methods, and in all cases much lower than the time for an original simulation run ($t_{simrun} = 3$ min).

Time required to estimate a surrogate model for each TS ($t_{estim,SM}(N_{estim})$) generally depends on the size of the hyperparameter search space considered (which was likely more extensive than necessary for RF and XGB in our case), (ii) the number of processors that can be used in parallel, which not only depends on the method but also on the available resources, and (iii) the specific software implementation of the algorithm.

Pure total computation time for training and validation over all candidate functional forms/hyperparameter settings was fastest for MNL (less than 1 h) and MARS (about 3 h) employing a single processor. (Model selection in MNL regression, however, was not automatized and involved a considerable, unmeasured amount of human work time and brain power that is difficult to quantify and stretched over weeks.) The RF tuning process required about 14 h employing 16 processors of a High-Performance Compute (HPC) node. Our very comprehensive



Fig. 5 Stability of feature importance of independent variables. Each violin graph shows the difference between the feature importance rank of independent variables estimated from TS_s vs. TS_{s-1} , as distributed over the 41 independent variables

search process for XGB required about 94 h using a HPC node with 40 cores.⁹ More information on time demands can be found in the respective appendices.

However, we did not systematically benchmark timings under comparable and reproducible conditions and did not optimize code structure, parallelization and size of hyperparameter search space towards time efficiency in training. (Moreover, in this particular study, MARS and MNL were run in R, while XGB and RF were run in Python.) All timing information should therefore be understood as first rough indications. Especially, the results of our comprehensive random searches for RF and XGB showed large shares of hyperparameter combinations that hardly differed in performance, so that in practice search spaces and hence computation times can most likely be reduced by a factor of five without relevant loss of fit, especially if systematic searches instead of random searches are used.

We tested a sequential approach to determine a sufficient number of evaluations with the original model. While here we directly simulated an extended range of sample sizes, in practice, a modeler will first simulate only the smallest TS size, estimate a surrogate model, test its performance with a small validation sample, then simulate the additional runs for the second TS size, estimate and test, simulate the additional runs for the second TS size, estimate and test, simulate the additional runs for the third TS size, and so on. The modeler will stop increasing TS size when computational resources are exhausted or when the fit in the validation sample remains stable and without fluctuations. What constitutes a negligible increase will, again, be subject to the intended purpose of the surrogate model and trade-off considerations between additional effort and fit. In our example, convergence could arguably be attested at a TS size of 3000 based on graphical inspection for all approaches, but even at 1500 runs this could be defensible in practice. Nevertheless, even if a modeler had been able to simulate only 500 runs, RF and XGB would already have given decent approximations in our case.

One can argue that the use of a surrogate model becomes efficient only when the number of predictions for which it is eventually used surpasses the number of original model runs necessary to train it (plus the number of runs possible during the time needed for the training). Nevertheless, in applications such as integration into large-scale models and interactive result exploration with stakeholders the model response is time-critical and the input factor combination to be evaluated cannot be anticipated and simulated in advance. Moreover, if the surrogate model is used to address data-privacy constraints or harmonize heterogeneous model

⁹ For XGB, separate training processes over all TS and samples of 300 hyperparameter settings with 3fold cross-validation were run for each candidate level for the hyperparameter n_{tree} ("number of trees"), each on a separate node with 40 processors. Smaller n_{tree} required much shorter training times than longer ones, as a doubling of n_{tree} roughly corresponded to a doubling of computation time. The 94 h result from adding up the time over nodes. As nodes were run in parallel, too, total waiting time was actually shorter and determined by the highest $n_{tree} = 1024$ for the model with price ratios, which required about 46 h for training. Further parallelization over TS and hyperparameter samples is possible and can decrease real waiting times using more processors more efficiently.

implementations, time spent on surrogate model estimation is usually well-invested and not prohibitive.

4.4 Scope for Performance Improvement

Since the MPMAS-CSJ model is nominally deterministic,¹⁰ we can expect a fully adaptable surrogate modeling method such as XGB to eventually achieve near-perfect fit if the training sample size is sufficiently increased.

For the given sample sizes, the performance and efficiency might potentially be improved by an adaptive sampling scheme that starts with a small initial space-filling design and subsequently prioritizes regions of high variance in prediction for additional farm-level model runs instead of using the Sobol' sequence throughout (Gramacy and Lee, 2009). Regions of poor fit would receive higher weight similar to the resampling logic used in Gradient boosting. However, while within Gradient boosting existing information would receive adapted resampling weights, adaptive sampling would generate new, additional information for these regions using the simulation model.

4.5 Generalizability

The sequential sampling approach to provide training data combined with performance and stability assessments in validation samples is generally applicable to support adequate surrogate model development for agroeconomic models. It builds on the theoretical properties of representative sampling by low discrepancy sequences and control of sampling error through cross-validation and is model-free (nonparametric).

The exact relationship between achievable goodness-of-fit and necessary original model evaluations observed in our experiment is most likely not generalizable. From a theoretical point of view, this relationship depends very much on the complexity of the simulated input-output relationships. The greater the complexity the more repetitions will be needed. Complexity may differ considerably between farm-level model applications. Nevertheless, we believe that our test case provides a sufficiently complex benchmark that represents well the typical input-output relationships in simulation data generated by farm-level models. We observed a high level of input factor interactions and segmented response functions (illustrated in the description of the manual MNL estimation in the appendix) that required complex surrogate models. Ratios between price expectations for different items are important to understand area response. Response to input factors and input factor ratios is not constant, but differs between segments along the input ranges and as a function of other input factors. Transition between segments is not always smooth, but sometimes abrupt at breakpoints.

¹⁰ Unsystematic, random effects in the simulation data cannot fully be ruled out: Indifferences with respect to optimality in combination with numerical complexities and tolerances in the mathematical optimization process may sometimes lead to different, unpredictable outcomes for similar inputs that may go unnoticed and are difficult to detect.

The good performance of XGB is in line with findings in machine learning contests for a large variety of complex prediction tasks (Bentéjac et al., 2021). There is hence reason to expect this to hold also for many agricultural economic applications. Apart from the baseline scenario (B), we also tested the procedure with climate scenario C2 of Troost & Berger (2015). As Troost and Berger discussed, this scenario represents a major structural break compared with the baseline as it i.a. removes an important crop rotation constraint reducing complementarity between crops. For the somewhat less complex price-crop area interrelationships in C2, the results we obtained were comparable except that RF performed very similar to XGB —even slightly better at lowest sample sizes (see Appendix C).

While we focused on aggregated land use shares as potential link to large-scale integrated assessment models, the principles are applicable to continuous output (omitting the *ilr* transformation) and should also be transferable for surrogate modeling of disaggregate farm-level land use shares. Two additional challenges will have to be addressed for the latter case: (i) Compared with regional aggregates, the share of crop area categories with zero area is typically considerably higher for individual farm agents, so the replacement of zeroes by small positive amounts will require additional scrutiny. (ii) Structured sampling should be extended to the farm agent population to select samples of farm agents for training, respectively validation, whereas in our aggregate version the full farm agent population was used always.

The estimated surrogate models themselves are not generalizable. They represent price-land use relationships only for the structural scenario, time frame, and study area simulated with the FLM. Relying on fundamental economic principles and a generic disaggregated formulation, the FLM itself can simulate farmer reactions under structurally very different conditions, such as climate change scenarios (Antle, 2019; Troost & Berger, 2015). Surrogate models estimated from one structural scenario can, however, not simply be used to extrapolate to other structural scenarios. After simulation of several structural scenarios with the FLM, either a separate surrogate model would have to be estimated for each structural scenario or an encompassing surrogate model would have to be estimated over a sufficiently large number of structural variations.

5 Conclusions

We conducted a systematic experiment to analyze in how far regional crop area share responses to price changes simulated by a detailed farm-level model can be efficiently and comprehensively summarized in the form of surrogate models using econometric and machine-learning methods.

We found that combining extreme gradient boosting with the isometric log-ratio transformation (XGB + ilr) provided a straightforward surrogate modeling approach that is able to achieve a high and robust fit of the complex compositional response surface, albeit connected with higher computational cost. Random forest regression (RF + ilr) appeared as a viable compromise between computational cost and predictive performance, especially for less complex input-output relationships. In our

experiments, both outperformed multivariate adaptive regression splines (MARS + ilr) as well as a manually-specified multinomial logistic regression model.

In addition, our experiments highlighted important lessons for the practical use of surrogate modeling with micro-scale agricultural economic models: (i) Even when using a highly representative experimental design, explicit cross-validation and diagnostics for robustness must be an essential component in surrogate modeling in order to prevent overfitting in model selection and avoid unstable and biased predictions. (ii) Explicitly including ratio terms between input variables based on economic domain knowledge can strongly improve efficiency of surrogate model estimation even for highly automatic non-parametric methods and increase performance especially at lower sample sizes, when combined with proper crossvalidation. (iii) Using an adequate sampling design in combination with appropriate cross-validation can help to keep the necessary number of original model runs low. The Sobol' sequence provides a straightforward space-filling sequential design which avoids unnecessary re-evaluations of the original model when increasing the sample size. In our case, it allowed robust estimation from a moderately sized sample and we recommend it as a default choice. Adaptive sequential designs could potentially further improve efficiency.

Our results show that, if these lessons are heeded, reliable surrogate models for aggregate compositional outcomes of micro-scale agricultural economic models can be estimated. This greatly facilitates the integration of micro-scale models into large-scale integrated assessment models - not only by potential gains in computational efficiency but also by providing an option to harmonize integration of models of heterogeneous provenance and allowing for model sharing despite privacy-restricted simulation model input data. In this way, they can make a significant contribution in collaboration towards the structure-rich computational agricultural economics as envisaged by Antle (2019). While our experiments focused on aggregate regional outcomes, the methods presented here could also substantially support calibration, sensitivity analysis, and interactive result exploration with stakeholders (Mössinger et al., 2022) if future research can confirm their successful applicability at a more disaggregate level.

Appendix A: Detailed Methodology and Results for Each Surrogate Modeling Method

A.1 Multinomial Logistic Regression (MNL)

As a classical regression method, we tested multinomial-logistic regression (MNL), more specifically a baseline-category logit model with observation-specific regressors¹¹ and grouped¹² data (Agresti, 2013). The categories correspond to the crops

¹¹ This is different from MNL models commonly used in discrete choice analysis which typically contain category-specific regressors or mix category- and observation-specific regressors.

¹² Ungrouped data has one row per observation with one multinomial dependent variable indicating the category assigned to the observation. Grouped data aggregates observations with the same values for regressors and uses a vector of category counts as a dependent variable. Statistical properties differ. The

chosen and the basic observational units categorized are the plots of the farm agents. While real-world plot sizes vary between and within farm holdings and also the MPMAS-CSJ model allows arbitrary small plot sizes, we used an assumed plot size of one hectare for estimating the multinomial-logistic model, rounding all land use category areas to full hectares. (Many spatially explicit farm-level or agent-based models, including other MPMAS applications, work with fixed-size raster cells making this a common simplification.)

In the present analysis, we did not distinguish plots by individual plot characteristics such as soil types or size of the farm they belong to. We only used 'plot' as an auxiliary observational unit that allowed us to treat the areas as count data and apply multinomial logistic regression. The observation-specific regressors in the MNL regression were the input factor values used in the simulation run. All plots within one simulation run shared the same values for the regressors and observations were grouped by simulation run.

The baseline-category logit model consisted of J - 1 = 8 equations, each modeling the log odds of a plot being categorized into category *j* instead of the baseline category *J* as a function of the input factors.

$$\ln \frac{\pi_j(\mathbf{x})}{\pi_J(\mathbf{x})} = \alpha_j + \beta_j \mathbf{x}, \quad j = 1, \dots, J - 1$$
(5)

The probability for each category then results as

$$\pi_j(\mathbf{x}) = \frac{\exp(\alpha_j + \beta_j \mathbf{x})}{1 + \sum_{h=1}^{J-1} \exp(\alpha_h + \beta_h \mathbf{x})}, \quad j = 1, \dots, J-1$$
(6)

For our purposes, we treated the estimated probability of an individual plot to fall into a specific category as equivalent to the maximum likelihood estimate for the share of the category in the total simulated area.

For estimation, we used the function multinom from the R package nnet (Ripley and Venables, 2021), one of the few that allowed entering the data in grouped format. Grouping is essential for our application as transforming the data into rows for thousands of individual plot observations would lead to prohibitive memory requirements.

To cope with interactions, segmentation and nonlinearities in the land use response to prices, we iteratively analyzed different functional forms for the regressors. In each iteration, additional interaction and segmentation terms where manually selected based on an in-depth analysis of model residuals. Akaike's Information Criterion (AIC) was used to compare extended functional forms. Overall, in this iterative model selection process, we tested 30 different functional forms for multinomial-logistic regression (mlr1-mlr30).

We started with a purely additive model of the price factors (mlr1) and then added the model uncertainty factors in a second model (mlr2). Given the theoretical

Footnote 12 continued

deviance of MNL models for grouped data is approximately chi-square distributed, which is not the case for ungrouped data (Agresti, 2013).

importance of price relationships for crop profitability, we added first order price interactions (mlr3), respectively in an alternative version the ratios of all other price factors to the wheat price as predictors (mlr4). Noticing more pronounced predictive errors at high wheat price levels, we added a square term for the wheat price factor (mlr5) and additionally interaction terms of the wheat price factor with selected model uncertainty parameters (mlr6). Observing that the residuals showed a strong pattern with respect to the malting-barley-to-fodder-barley-price ratio $\frac{pc_{mb}}{pc_{fb}}$ and that this response seemed to be segmented (see left panel in Fig. 7), we added $\frac{pc_{mb}}{pc_{m}}$ (as well as the ratios of fertilizer, rapeseed and ready-mix fooder prices to fodder barley) to the model as well as segmentation dummies interacting with $\frac{pc_{mb}}{pc_{th}}$. We varied the number and locations of the segmentation (mlr9-mlr13, mlr15) and let the segmentation dummies also interact with the malting barley, fodder barley and wheat price coefficients (mlr14, mlr16-mlr20) and finally segmented also the malting barley price to allow different coefficients for its interactions, testing different segmentation points (mlr20-mlr30). The R script with all 30 functional forms took about one hour to run on a normal desktop computer (no explicit parallelization). However, the analysis and design of functional forms was done iteratively and manually and required considerable working time.

Finally, we selected the model with the highest AIC as the best model obtained. In addition, we applied a forward selection and backward elimination procedure over the full set of terms included in our analysis using the R package stepAIC (which internally uses AIC for model selection), but could not achieve improvements in AIC over the manually selected model.

Figure 6 gives an overview over the development of AICs over the step-wise increases of the training sample size for a selection of functional forms. As the AIC depends on the sample and provides only relative information, the graph depicts the relative difference to the best performing model, i.e. the one with the lowest AIC, for each sample size. As can be observed, the best performing functional form mlr29



Fig. 6 Relative AIC of selected MNL models

showed the lowest AIC at all sample sizes. Although it is a manually specified model, this best estimated model is highly complex comprising 720 coefficients.

The ranking of the major groups of functional forms remained mostly stable (left pane), while ranking within the last group of functional forms that mainly differ in the location of segmentation points is only slightly more variable (right pane). The simpler functional forms achieved average shares of correctly classified plots of 0.85, while the better models achieve slightly over 0.92, with very minor differences. These values were virtually stable over all training sample sizes and, more importantly, a ranking between different functional forms also remained stable.

A.2 Isometric Log-Ratio Transformation (ilr)

A number of approaches are available which transform compositional data into an unconstrained continuous Euclidean vector space and enable the use of statistical approaches designed for continuous data with the transformed data.

On the one hand, the softmax function, the multivariate generalization of the logistic function, is typically being used in machine learning and is equivalent to a multinomial logistic formulation of the model.

On the other hand, Aitchison (1986) identified the simplex as the natural sampling space of compositional data and suggested two transformations: the additive log-ratio transformation (*alr*) and the centered log ratio transformation (*clr*). The *alr* projects a *K*-dimensional compositional vector on a K - 1-dimensional unconstrained vector and has been widely used in statistical analysis. It is, however, not symmetrical in components and not isometric (preserving distances). For contrast, the *clr*, which applies the logarithm of the share of a component normalized by the geometric mean of all components, is symmetric and easy to interpret, but it maintains a K-dimensional vector whose sum is constrained to zero (Egozcue et al., 2003).

Egozcue et al. (2003) complemented Aitchisons concept with the isometric logratio transformation (*ilr*) which starts from the *clr* and projects it into an orthonormal space of K - 1 components. An *ilr* is unconstrained, isometric and subcompositionally consistent, but not unique: Many orthonormal bases are possible based on different sequential binary partitions of the categories. These binary partitions may be chosen such as to ease interpretation. In the resulting Euclidean K - 1 Euclidean vector space, standard statistical regression techniques can be applied. For the isometric log ratio transformation, we use the corresponding function from the R package *compositions* (van den Boogaart et al., 2021).

All logarithmic transformations only work with strictly positive data, whereas our dataset contains a non-negligible amount of zeros. We worked around this problem by adding a small quantity (= one hectare) to all categories in all instances. With area totals around 25,000 hectares the distortion was expected to be minimal.

A.3 Multivariate Adaptive Regression Splines (MARS)

Multivariate Adaptive Regression Splines (MARS) (Friedman, 1991) is a nonparametric method that we chose for being particularly suited to tasks involving



Fig. 7 Residuals of selected MNL models estimated from training sample of size 1500: Residuals plotted in relation to the malting barley-to-fodder barley price ratio

segmented regression, nonlinearity and strong interactions. It estimates regression functions of the form

$$f(x) = c_1 + \sum_{i=2}^{k} c_i B_i$$
(7)

where B_i stands for a hinge function of the form $\max(0, x - d)$ or $\max(0, d - x)$, or a product of several such hinge functions (interactions between variables). The hinge functions allow for a continuous recursive partitioning of the input factor space that gives the technique its flexibility. That means it can automatically include segments and estimate their boundaries. A MARS model is estimated in two passes: In a forward pass, the algorithm subsequently adds pairs of complementary hinge functions for one factor or factor interaction. It always chooses a partitioning for the factor or factor interaction that generates the strongest reduction in prediction error. This continues until an overall limit of terms is reached or a threshold of minimum improvement is not surpassed anymore. In the following backward pass, the algorithm prunes terms from the equation in order to improve generalizability and reduce over-fitting. It prunes always the term that shows the least effectiveness with respect to a Generalized Cross Validation criterion.

Apart from the limit on the number of terms, the minimum improvement threshold and the penalty used in the calculation of the GCV, the most important hyperparameter is the degree of interactions between factors that is allowed. We used the R package "mda" (Leisch et al., 2020) to apply MARS to the ilr-transformed simulation results. The R script with 25 combinations of hyperparameter settings took about 3.3 hours to run on a normal desktop computer (no explicit parallelization).

Hyperparameter and feature set variation in the MARS approach showed that the strongest effect on performance were caused by changing the allowed degree of interactions and by the explicit inclusion of price coefficient ratios into the feature set (ratios of other prices to wheat, resp. fodder barley prices). The two hyperparameters 'thresholds for minimum improvement' and 'GCV penalty' did not have a noticeable effect on goodness-of-fit.

As Fig. 8 shows, including price coefficient ratios increased the share of correctly classified plots by about 0.05 and allowing interactions increased it by another 0.05. Allowing degrees of interaction higher than two did not noticeably affect goodness-of-fit and limiting to two-way-interactions shows the best worst-case performance among comparable models. The selected two degree model with ratios included contains between 40 (at sample size 500) and 60-65 terms (at sample sizes beyond 2000) and reaches an average share of correctly classified plots of about 0.91. As with MNL, the ranking of different MARS models was stable over increasing sizes of the training sample, performance did only marginally improve beyond 2,000 runs in the training sample and the small validation sample VS1 was a good predictor of average performance in the larger validation sample VS2, while worst-case performance was better predicted by a larger training sample than by VS1.

A.4 Random Forest Regression (RF)

Non-parametric ensemble-based methods achieve flexibility of fitting and control against overfitting by aggregating over ensembles of simple, smaller regression models estimated on subsets of the data. Random forest regression (Breiman, 2001) estimates an ensemble (a 'forest') of regression trees, each trained on a bootstrapped subset B consisting of both randomly drawn samples and features (independent variables) of the data set (Breiman, 2001; Hastie et al., 2009). This concept of training a model on B random bootstrap samples is also referred to as *bagging* (short for bootstrap aggregation). Each regression tree within the random forest is grown by step-wise partitioning the data. For this purpose, on each node those variables and thresholds with the strongest effect on a pre-defined loss function are identified and used to split the data into new leaves, or daughter nodes. This process of data partitioning is repeated until either the loss function equals zero or a pre-defined stopping criterion, such as a minimum sample size inside each leaf or the maximum number of partitions (maximum depth of tree), is reached. The eventual predictions of random forest regressions correspond to the average (or 'bagged') results of the individual regression trees, making them rather resistant against overfitting (Breiman, 2001; Hastie et al., 2009).

We used the RandomForestRegressor methods of the scikit-learn Python package (Pedregosa et al., 2011) to perform the estimations on the ilr-transformed MPMAS-CSJ simulation results varying the following hyperparameters: The number of trees *nTrees*; the maximum number of the independent variables included in an individual tree ('max. feature': $\in \{n, \sqrt{n}, 0.3n\}$, with *n* being the number of eligible independent features); the maximum depth of an individual tree ('max. depth' $\in [5, 500]$). As a loss functions we used the mean squared error (MSE) as well as the mean absolute error. In addition, we used two different sets of eligible features: One (WithoutRatios) comprising all L=22 input factors as they are, and one (WithRatios)



Fig. 8 Average and minimum of Share of correctly classified plots of MARS models in the validation and training samples



Fig. 9 Out-of-bag error by RF number of trees at each TS size. Rows distinguish models explicitly including price ratios from models not including them

in which we additionally included price ratios (i.e. ratios between some of the 22 input factors) as independent variables for RF.

Bergstra and Bengio (2012) argue that in most models very few hyperparameters are actually affecting the performance and randomized grid search leads to the same or even better results compared to a full grid search, while requiring significantly less computational power. Following the approach of Probst et al. (2019) and Oshiro et al. (2012), we combined grid and random search. In a first step, we identified the amount of trees at which the out-of-bag score converged, using default hyperparameter values. For this purpose and for each training sample size, we trained and predicted a RF model on 12 levels for *nTrees*, ranging between 1 and 2048 trees, duplicating the value at each step, which required about 9 minutes employing 16 processors in parallel. This resulted in an optimal *nTrees* of 512 for all training sample sizes as well as models both with and without ratios (Fig. 9).

In the neighborhood of this optimal *nTrees* at default values (number of trees 256, 512, 1024) we then conducted a random grid search (with 60 random draws) to test different combinations of 'max. feature', 'max.depth' and loss function. We used the *RandomizedGridCrossValidation* package from Scikit-learn in Python over 60 random draws¹³ out of a pre-defined hyperparameter grid, evaluating each draw with 3-fold cross-validation (Pedregosa et al., 2011). This required about 13 hours on 16 processors for all three *nTrees* levels. Individual TS tuning required between 2 min at TS = 500 and 84 min at TS = 5000 for *nTrees* = 512.

¹³ Reproducibility was ensured by setting a global seed at the beginning of each script.

Finally, we used the best set of hyperparameters at each level of training size and number of trees to predict VS1 and VS2. (Predictions on the second to fifth best sets were tested as well.)

Figure 10 shows the average and minimum share of correctly classified plots in VS1 for the four best-ranked hyperparameter sets identified in the random search at each tree size, training sample size and with and without explicit inclusion of price ratios. Explicitly including price ratios in the feature set has the strongest effect on goodness-of-fit, although the effect is lower than for the other methods, with an increase in Scc about 0.04 at TS = 500 and reducing to 0.015 at TS = 5000. Varying tree size around the optimum identified around convergence or varying within the best ranking hyperparameter set does not noticeably change fit.

Similar to the other methods, the predictive performance did not differ between VS1 and VS2. A larger discrepancy is observable between training and validation samples. Predictive performance in the training sample is about 0.025 higher, pointing to some degree of overfitting (Fig. 11).

A.5 Extreme Gradient Boosting (XGB)

Similar to random forest regression, gradient boosting regression also generates an ensemble of simple regression trees ('weak learners') that minimizes an arbitrary differentiable loss function (Friedman, 2002). Contrary to random forest regression, however, the bootstrapped samples to estimate an individual tree are not all created at the start, but trees are estimated sequentially. At each iteration, the predictive performance of the whole ensemble generated so far is assessed and the next bootstrapped sample is created assigning a higher weight to observations not yet well matched by the ensemble (Hastie et al., 2009). Compared with RF, a more accurate prediction is typically possible, but at the same time the risk for overfitting is higher and hyperparameters have to be chosen carefully to avoid this (Storm et al., 2020).

For our ilr-transformed MPMAS-CSJ simulation data, we used Extreme gradient boosting (XGB, Chen and Guestrin, 2016) as implemented in the XGBoost Python package (xgboost developers, 2021), a specific GB implementation using a more regularized model formulation to control over-fitting and efficient computation including partial parallelization. Contrary to RF, where each tree is independent from the others, it is not possible to parallelize the ensemble itself in GB, because each tree is dependent on the previous ones. XGB parallelizes the construction of nodes within each tree.

For tuning XGB hyperparameters, at each training sample size we exponentially increased the allowed number of trees *nTrees* starting with eight and doubling until no further improvement was visible. For each *nTrees* tested, we conducted a random search (300 iterations) with 3-fold cross-validation for five critical XGB hyperparameters suggested by Bentéjac et al. (2021): learning rate, minimum loss reduction (gamma), the maximum depth of the tree, the fraction of features to be evaluated at each split, and the subsampling rate. The best hyperparameter combination for each TS size and number of trees was recorded. The random search tuning process over all TS sizes was conducted using 40 processors in parallel and required between 8 min at *nTrees* = 8 and without price ratios and 14 h at *nTrees* = 1024 without price ratios.



Fig. 10 Average and minimum share of correctly classified plots of the best random forest models in VS1 and VS2. Grayscale distinguishes with vs. without explicitly included price ratios. Linestyle distinguishes maximum number of trees. Facets distinguish the four best-ranked hyperparameter combinations in the random search

With price ratios explicitly included, times tripled to 25 min at nTrees = 8 and 46 hours for nTrees = 1024.

For each TS, Fig. 12 shows the development of average Scc over an increase in the number of trees (in each case showing the best ranked combination of the other



Fig. 11 Average and minimum share of correctly classified plots of the best random forest models in training sample. Grayscale distinguishes with vs. without explicitly included price ratios. Linestyle distinguishes maximum tree size. Facets distinguish the four best-ranked hyperparameter combinations in the random search

five hyperparameters). While the optimal number of trees varied among the lower TS sizes, it converged to 512 for the larger TS sizes. While 512 was not necessarily the best choice at lower TS sizes, it still provided very good results.

Figure 13 shows the average and minimum share of correctly classified plots in VS1. Similar to the other methods, explicitly including price ratios in the feature set had the strongest effect on goodness-of-fit: it increased the average Scc by 0.04-0.05 at lower and about 0.02 at higher TS sizes. Note: Without explicit price ratios and at lower training sample sizes, a large number of trees was detrimental as it lead to clear overfitting (better performance in the training sample, but worse performance in the validation samples; not shown in graphs here).

Appendix B: Using Aitchison's Total Variance to Measure Stability Of Predictions

Aitchison's total variance is a measure of dispersion for compositional data (simplex space). Its properties correspond to the properties of the variance in real vector space (Pawlowsky-Glahn and Egozcue, 2001; Pawlowsky-Glahn and Buccianti, 2011).

The first step in the calculation of the total variance is the construction of the variation matrix \mathbf{M} .

The $J \times J$ elements of the variation matrix each represent the variance in the logratio between the area in one of the J crop categories and the area in one other



Fig. 12 Average share of correctly predicted plots in VS1 depending on max. number of trees in XGBoost for each TS size



Fig. 13 Average and minimum share of correctly classified plots in VS1 of the best XGB models, with and without explicit inclusion of price ratios



Fig. 14 Results for climate scenario C2: Predictive accuracy of the best performing hyperparameter setting, resp. functional form by surrogate modeling method and training sample size in validation samples VS1 and VS2. (All explicitly including price ratios in the set of independent variables)



Fig. 15 Results for climate scenario C2: Comparison of predictive accuracy of the best performing hyperparameter setting, resp. functional form WITHOUT and WITH explicit inclusion of price ratios as input factors by surrogate modeling method and training sample size

category at row k. In our case, a separate total variance and hence a separate variation matrix is calculated for each row k in VS1.

$$\mathbf{M}_{k} = \begin{bmatrix} m_{k,1,1} & m_{k,1,2} & \dots & m_{k,1,J} \\ \vdots & \vdots & \ddots & \vdots \\ m_{k,J,1} & m_{k,J,2} & \dots & m_{k,J,J} \end{bmatrix}$$
(8)

with $m_{k,j1,j2} = \operatorname{var}\left(\ln\frac{AreaSM_{k,j1}}{AreaSM_{k,j2}}\right)$ and $j1, j2 = 1, \ldots, J$.

The variance in each cell (indicated by var in the formula) is here the traditional variance of the log ratio between the two categories over the predictions for row k by the different surrogates included in the comparison. In Fig. 4, we presented the variance calculated over predictions from two adjacent TS, but this can be extended to three or more adjacent sample sizes to include more than only the last TS increase in the assessment of convergence. (We calculated also for three adjacent variables, but did not see a relevant difference and decided to include the simpler version with two.) We used the R package 'compositions' (van den Boogaart et al. (2021) to calculate the variation matrix.

The total variance for row k is then calculated by summing over all elements of the variation matrix for k and dividing by J (Pawlowsky-Glahn and Egozcue, 2001).

$$\text{totvar}_{k} = \frac{1}{J} \sum_{jl=1}^{J} \sum_{j2>jl}^{J} m_{k,j1,j2} = \frac{1}{2J} \sum_{jl=1}^{J} \sum_{j2>jl}^{J} \text{var}\left(\ln\frac{AreaSM_{k,j1}}{AreaSM_{k,j2}}\right).$$
(9)

For the graph in Fig. 4, we then plotted the distribution of $totvar_k$ over all k as a violin graph.

Appendix C: Results for Scenario C2

While the main part of the article used the the baseline scenario (B) of Troost and Berger (2015) as a test case for surrogate modeling, we also tested the procedure with their climate scenario C2, which represents a major structural break compared with the baseline. As Troost and Berger discussed, this scenario i.a. removes an important crop rotation constraint that leads to somewhat less complex price-crop area interrelationships. The results obtained were similar, except that XGB and RF show very similar performances in this case and at lowest sample sizes RF performed slightly better (Figs. 14, 15).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10614-022-10276-0.

Acknowledgements We acknowledge funding by the Federal Ministry of Education and Research of Germany (BMBF) for the project SimLearn (011S19073C) and by Deutsche Forschungsgemeinschaft (DFG) for the project FOR-1695. The authors acknowledge support by the state of Baden-Württemberg through bwHPC and by DFG through grant INST 35/1134-1 FUGG. Adam Törös developed initial multinomial-logistic regression models in his M.Sc. thesis.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declaration

Conflicts of interest The authors declare that there is no conflict of interest due to material or financial interests in the subject of the research.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Agresti, A. (2013). Categorical data analysis (3rd ed.). Hoboken, New Jersey: John Wiley & Sons.

Aitchison, J. (1986). The Statistical Analysis of Compositional Data. *Monographs on Statistics and Applied Probability*. London [u.a.]: Chapman and Hall, 1st ed.

Antle, J. M. (2019). Data, economics and computational agricultural science. American Journal of Agricultural Economics, 101, 365–382. https://doi.org/10.1093/ajae/aay103.

- Asher, M. J., Croke, B. F. W., Jakeman, A. J., & Peeters, L. J. M. (2015). A review of surrogate models and their application to groundwater modeling. *Water Resources Research*, 51, 5957–5973. https://doi. org/10.1002/2015WR016967
- Baustert, P., & Benetto, E. (2017). Uncertainty analysis in agent-based modelling and consequential life cycle assessment coupled models: A critical review. *Journal of Cleaner Production*, 156, 378–394. https://doi.org/10.1016/j.jclepro.2017.03.193
- Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54, 1937–1967. https://doi.org/10.1007/s10462-020-09896-5
- Berger, T., & Troost, C. (2014). Agent-based modelling of climate adaptation and mitigation options in agriculture. *Journal of Agricultural Economics*, 65, 323–348. https://doi.org/10.1111/1477-9552. 12045
- Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13, 281–305.
- Boogaart, K. G. van den, Tolosana-Delgado, R., & Bren, M. (2021). R package compositions: Compositional data analysis. https://CRAN.R-project.org/package=compositions.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. https://doi.org/10.1023/A: 1010933404324.
- Brown, C., Holman, I., & Rounsevell, M. (2021). How modelling paradigms affect simulated future land use change. *Earth System Dynamics*, 12, 211–231. https://doi.org/10.5194/esd-12-211-2021
- Buysse, J., Huylenbroeck, G. V., & Lauwers, L. (2007). Normative, positive and econometric mathematical programming as tools for incorporation of multifunctionality in agricultural policy modelling. *Agriculture, Ecosystems and Environment, 120,* 70–81. https://doi.org/10.1016/j.agee.2006.03.035.
- Campolongo, F., Cariboni, J., & Saltelli, A. (2007). An effective screening design for sensitivity analysis of large models. *Environmental Modelling and Software*, 22, 1509–1518. https://doi.org/10.1016/j. envsoft.2006.10.004.
- Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, KDD '16. New York, NY, USA: Association for Computing Machinery, 785–794, https://doi.org/10.1145/2939672. 2939785.
- Domínguez, I. P., Bezlepkina, I., Heckelei, T., Romstad, E., Lansink, A. O., & Kanellopoulos, A. (2009). Capturing market impacts of farm level policies: a statistical extrapolation approach using biophysical characteristics and farm resources. *Environmental Science and Policy*, 12, 588–600. https://doi.org/ 10.1016/j.envsci.2009.02.006.
- Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., & Barceló-Vidal, C. (2003). Isometric Logratio Transformations for compositional data analysis. *Mathematical Geology*, 35, 279–300. https://doi. org/10.1023/A:1023818214614.
- Friedman, J. H. (1991). Multivariate adaptive regression splines. The Annals of Statistics, 19, 1-67.
- Friedman, J. H. (2002). Stochastic gradient boosting. Computational Statistics and Data Analysis, 38, 367– 378. https://doi.org/10.1016/S0167-9473(01)00065-2.
- Gramacy, R. B., & Lee, H. K. H. (2009). Adaptive design and analysis of supercomputer experiments. *Technometrics*, 51, 130–145. https://doi.org/10.1198/TECH.2009.0015
- Happe, K., Kellermann, K., & Balmann, A. (2006). Agent-based analysis of agricultural policies: An illustration of the agricultural policy simulator AgriPoliS, its adaptation, and behavior. *Ecology and Society*, 11, 49.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.
- Jakeman, A., Letcher, R., & Norton, J. (2006). Ten iterative steps in development and evaluation of environmental models. *Environmental Modelling and Software*, 21, 602–614.
- Kleijnen, J. P. C. (2017). Regression and Kriging metamodels with their experimental designs in simulation: A review. European Journal of Operational Research, 256, 1–16. https://doi.org/10.1016/ j.ejor.2016.06.041
- Kremmydas, D., Athanasiadis, I., & Rozakis, S. (2018). A review of agent based modeling for agricultural policy evaluation. Agricultural Systems, 164, 95–106.
- Lamperti, F., Roventini, A., & Sani, A. (2018). Agent-based model calibration using machine learning surrogates. *Journal of Economic Dynamics and Control*, 90, 366–389. https://doi.org/10.1016/j.jedc. 2018.03.011

- Lee, J. S., Filatova, T., Ligmann-Zielinska, A., Hassani-Mahmooei, B., Stonedahl, F., Lorscheid, I., Voinov, A., Polhill, G., Sun, Z., & Parker, D. C. (2015). The complexities of agent-based modeling output analysis. *The Journal of Artificial Societies and Social Simulation*, 18, https://doi.org/10. 18564/jasss.2897
- Leisch, F., Hornik, K., Ripley, B. D., Narasimhan, B., Hastie, T., & Tibshirani, R. (2020). R package Mda: Mixture and Flexible Discriminant Analysis. https://CRAN.R-project.org/package=mda.
- Lengers, B., Britz, W., & Holm-Müller, K. (2014). What drives marginal abatement costs of greenhouse gases on dairy farms? A meta-modelling approach. *Journal of Agricultural Economics*, 65, 579–599. https://doi.org/10.1111/1477-9552.12057.
- Lippe, M., Bithell, M., Gotts, N., Natalini, D., Barbrook-Johnson, P., Giupponi, C., Hallier, M., Hofstede, G. J., Le Page, C., Matthews, R. B., Schlüter, M., Smith, P., Teglio, A., & Thellmann, K. (2019). Using agent-based modelling to simulate social-ecological systems across scales. *GeoInformatica*, 23, 269–298. https://doi.org/10.1007/s10707-018-00337-8
- Mössinger, J., Troost, C., & Berger, T. (2022). Bridging the gap between models and users: A lightweight mobile interface for optimized farming decisions in interactive modeling sessions. *Agricultural Systems*, 195, 103315. https://doi.org/10.1016/j.agsy.2021.103315
- Müller, B., Hoffmann, F., Heckelei, T., Müller, C., Hertel, T. W., Polhill, J. G., van Wijk, M., Achterbosch, T., Alexander, P., Brown, C., Kreuer, D., Ewert, F., Ge, J., Millington, J. D. A., Seppelt, R., Verburg, P. H., & Webber, H. (2020). Modelling food security: Bridging the gap between the micro and the macro scale. *Global Environmental Change*, 63, 102085. https://doi.org/10.1016/j.gloenvcha.2020. 102085
- Müller-Hansen, F., Schlüter, M., Mäs, M., Donges, J. F., Kolb, J. J., Thonicke, K., & Heitzig, J. (2017). Towards representing human behavior and decision making in Earth system models—An overview of techniques and approaches. *Earth System Dynamics*, 8, 977–1007. https://doi.org/10.5194/esd-8-977-2017.
- Oshiro, T. M., Perez, P. S. and Baranauskas, J. A. (2012). How Many Trees in a Random Forest? In Perner, P. (ed.), *Machine Learning and Data Mining in Pattern Recognition*, Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 154–168, doi:10.1007/978-3-642-31537-4 13.
- Pawlowsky-Glahn, V., & Buccianti, A. (2011). Compositional data analysis: theory and applications. London: John Wiley & Sons.
- Pawlowsky-Glahn, V., & Egozcue, J. J. (2001). Geometric approach to statistical analysis on the simplex. Stochastic Environmental Research and Risk Assessment, 15, 384–398. https://doi.org/10.1007/ s004770100077
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12, 2825–2830.
- Probst, P., Wright, M. N., & Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. WIREs Data Mining and Knowledge Discovery, 9, e1301. https://doi.org/10.1002/widm.1301
- Reidsma, P., Janssen, S., Jansen, J., & van Ittersum, M. K. (2018). On the development and use of farm models for policy impact assessment in the European Union—A review. *Agricultural Systems*, 159, 111–125.
- Ripley, B., & Venables, W. (2021). R package nnet: Feed-Forward Neural Networks and Multinomial Log-Linear Models. https://CRAN.R-project.org/package=nnet.
- Salle, I., & Yıldızoğlu, M. (2014). Efficient sampling and meta-modeling for computational economic models. *Computational Economics*, 44, 507–536. https://doi.org/10.1007/s10614-013-9406-7.
- Saltelli, A., Tarantola, S., Campolongo, F., & Ratto, M. (2004). Sensitivity analysis in practice—A guide to assessing scientific models. Chichester: Wiley.
- Schreinemachers, P., & Berger, T. (2011). MP-MAS: An agent-based simulation model of humanenvironment interaction in agricultural systems. *Environmental Modelling and Software*, 26, 845– 859.
- Seidel, C., & Britz, W. (2019). Estimating a dual value function as a meta-model of a detailed dynamic mathematical programming model. *Bio-Based and Applied Economics Journal*, 8, 75–99. https://doi. org/10.13128/bae-8147.
- Storm, H., Baylis, K., & Heckelei, T. (2020). Machine learning in agricultural and applied economics. European Review of Agricultural Economics, 47, 849–892. https://doi.org/10.1093/erae/jbz033
- Tarantola, S., Becker, W., & Zeitz, D. (2012). A comparison of two sampling methods for global sensitivity analysis. *Computer Physics Communications*, 183, 1061–1072. https://doi.org/10.1016/j.cpc.2011. 12.015

- Troost, C., & Berger, T. (2015). Dealing with uncertainty in agent-based simulation: Farm-level modeling of adaptation to climate change in Southwest Germany. *American Journal of Agricultural Economics*, 97, 833–854. https://doi.org/10.1093/ajae/aau076
- Troost, C., Walter, T., & Berger, T. (2015). Climate, energy and environmental policies in agriculture: Simulating likely farmer responses in Southwest Germany. *Land Use Policy*, 46, 50–64. https://doi. org/10.1016/j.landusepol.2015.01.028
- van der Hoog, S. (2019). Surrogate modelling in (and of) agent-based models: A prospectus. Computational Economics, 53, 1245–1263. https://doi.org/10.1007/s10614-018-9802-0.
- van Wijk, M., Rufino, M., Enahoro, D., Parsons, D., Silvestri, S., Valdivia, R., & Herrero, M. (2014). Farm household models to analyse food security in a changing climate: A review. *Global Food Security*, 3, 77–84. https://doi.org/10.1016/j.gfs.2014.05.001
- van Wijk, M. T. (2014). From global economic modelling to household level analyses of food security and sustainability: How big is the gap and can we bridge it? Food Policy 49. Part, 2, 378–388. https://doi. org/10.1016/j.foodpol.2014.10.003
- Wuertz, D., Setz, T., & Chalabi, Y. (2021). R Package fOptions: Rmetrics—Pricing and Evaluating Basic Options. https://CRAN.R-project.org/package=fOptions.
- xgboost developers (2021). XGBoost documentation Python Package Introduction (1.4.0). https:// xgboost.readthedocs.io/en/latest/python/python_intro.html.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.