



22 **Highlights:**

- 23 • Machine learning allows for highly accurate downscaling of GGCM outputs
- 24 • Increasing detail of climate features improves prediction accuracy
- 25 • Feature importance ranks in the order climate  $\geq$  cultivar  $>$  soil and topography
- 26 • Approach is scale-free and does not require prior assumptions on feature importance
- 27 • It enables the development of robust downscaling tools with low user bias

28

29 **Abstract**

30 Global gridded crop models (GGCMs) are essential tools for estimating agricultural crop yields  
31 and externalities at large scales, typically at coarse spatial resolutions. Higher resolution  
32 estimates are required for robust agricultural assessments at regional and local scales, where the  
33 applicability of GGCMs is often limited by low data availability and high computational  
34 demand. An approach to bridge this gap is the application of meta-models trained on GGCM  
35 output data to covariates of high spatial resolution. In this study, we explore two machine  
36 learning approaches – extreme gradient boosting and random forests - to develop meta-models  
37 for the prediction of crop model outputs at fine spatial resolutions. Machine learning algorithms  
38 are trained on global scale maize simulations of a GGCM and exemplarily applied to the extent of  
39 Mexico at a finer spatial resolution. Results show very high accuracy with  $R^2 > 0.96$  for  
40 predictions of maize yields as well as the hydrologic externalities evapotranspiration and crop  
41 available water with also low mean bias in all cases. While limited sets of covariates such as  
42 annual climate data alone provide satisfactory results already, a comprehensive set of predictors  
43 covering annual, growing season, and monthly climate data is required to obtain high  
44 performance in reproducing climate-driven inter-annual crop yield variability. The findings  
45 presented herein provide a first proof of concept that machine learning methods are highly  
46 suitable for building crop meta-models for spatio-temporal downscaling and indicate potential  
47 for further developments towards scalable crop model emulators.

48 **Keywords:** meta-model, extreme gradient boosting, random forests, maize yield, agricultural  
49 externalities, climate features

## 50 **1 Introduction**

51 In recent years, global gridded crop models (GGCMs) - combinations of a crop model  
52 and global sets of gridded data - have become essential tools for estimating crop yields and  
53 agricultural externalities under a wide range of environmental and management conditions (e.g.  
54 Müller et al., 2017). Besides the direct provision and interpretation of model outputs for crop  
55 yields alone (e.g. Rosenzweig et al., 2014) or their joint evaluation with externalities such as  
56 crop water use (Liu et al., 2013; Elliott et al., 2014), GGCMs provide base layers of input data  
57 for agro-economic or integrated assessment models (IAMs; Müller and Nelson, 2014) e.g. for  
58 land use change analyses and optimization (e.g. Havlík et al., 2011).

59 The present global standard resolution of input data is  $0.5^\circ \times 0.5^\circ$  corresponding to  
60 approx. 50 km x 50 km near the equator. This is foremost determined by climate data, which are  
61 rarely available at higher resolutions at a global scale. Further common input data are  
62 management information and in most cases soil data and topography (Müller et al., 2017). The  
63 latter two are available at increasingly fine resolutions well below 1 km (Hengl et al., 2017a,  
64 Jarvis et al., 2008), while management is typically reported at national or subnational  
65 administrative levels (e.g. Sacks et al., 2010; Mueller et al., 2012). In few cases, simulations are  
66 run at the sub-grid level accounting for some heterogeneity in soil and topography (Skalský et  
67 al., 2008; Balkovič et al., 2014). Regardless of the spatial resolution, each simulation unit is  
68 treated as a homogenous field in the crop model.

69 While this spatial resolution provides sufficient detail for robust assessments at macro  
70 scales such as the country level, there is increasing concern that GGCM estimates and hence  
71 impact assessments at coarse resolutions often miss actual on-ground conditions. As only

72 average or dominant characteristics present within each grid are considered for simulations,  
73 assumptions and data may not match actually farmed land (e.g. Folberth et al., 2016) and  
74 farming practices (e.g. Reidsma et al., 2009). In addition, they may omit farm-level  
75 heterogeneity present at the sub-grid level (Ewert et al., 2011), which is essential for local to  
76 regional decision-making and stakeholder information (Rosenzweig et al., 2018).

77         Applying gridded crop models at very high spatial resolutions on the other hand increases  
78 computational demand substantially and is often limited by data availability as outlined above.  
79 Foremost climate data at suitable temporal resolutions for crop models - which is typically a  
80 daily time step (Müller et al., 2017) - are hardly available at fine spatial resolutions. The  
81 presently highest resolving global daily dataset known to the authors has  $0.25^\circ \times 0.25^\circ$  (Ruane et  
82 al., 2015), while regional products may have resolutions of up to  $0.11^\circ \times 0.11^\circ$  (Haylock et al.,  
83 2008). Temporally coarser data e.g. with a monthly time step, however, are available at very fine  
84 resolutions up to  $<1$  km (e.g. Wang et al., 2016; Fick and Hijmans, 2017).

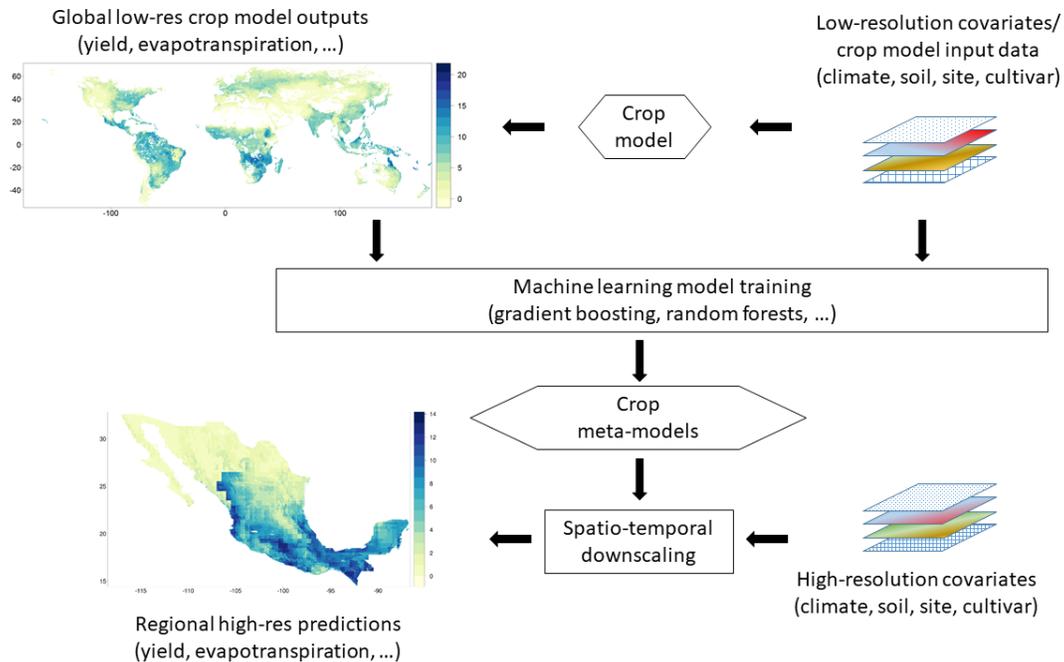
85         An approach lending itself to address these issues in an efficient and flexible way is the  
86 use of meta-models built from coarser GGCM simulations. This allows for deriving estimates of  
87 crop yields and associated agricultural externalities at high, virtually scale-free, spatial  
88 resolutions without requirements for setting up high-resolution crop model infrastructures  
89 including their comprehensive data requirements. There is no scientific literature on crop meta-  
90 model development for spatio-temporal predictions across scales known to the authors. The  
91 potentially most closely related field is the recently evolving crop model emulator development  
92 at the grid cell level. Examples are the development of regressions along climate change  
93 trajectories as such (e.g. Blanc and Sultan, 2015; Blanc, 2017) or the use of global crop model

94 simulations with artificial alterations of climate variables to retrieve estimates of climate change  
95 impacts for assessment studies based on regressions along temperature, precipitation, and CO<sub>2</sub>  
96 concentrations (Ruane et al., 2017; Rosenzweig et al., 2018). The production of high-resolution  
97 crop yield surfaces in contrast is foremost accomplished using simplified crop model algorithms  
98 (e.g. IIASA/FAO, 2012) or purely statistical approaches (e.g. Mueller et al., 2012). Common to  
99 all referenced approaches is that they (a) are based on narrow sets of *a priori* selected covariates  
100 based on modelers' assumptions and (b) do not allow for or have not been tested for the joint  
101 evaluation of agricultural productivity and externalities. Crop model emulators are in addition  
102 typically parameterized at the grid level, which renders them spatially determined and scale-  
103 depended.

104         The presently most flexible approaches for data-driven development of models with high  
105 accuracy can be found in the field of machine learning. Machine learning is a collective term for  
106 a wide range of data analysis and data-driven forecasting techniques. The most advanced  
107 techniques are characterized by the ability to digest large amounts of covariates (herein *syn.*  
108 features, *syn.* predictors) to provide predictions for both numeric and categorical variables with  
109 algorithms of high complexity and flexibility, which determine the relevance of provided  
110 covariates themselves (e.g. Witten et al., 2016). Examples of methodologic approaches are  
111 neural networks, various forms and derivatives of regression trees, as well as clustering  
112 techniques. While simpler methods such as multiple linear or lasso regressions are typically  
113 computationally faster and straightforward to interpret, they show typically a substantially lower  
114 performance. Within agricultural sciences, applications are to date mostly limited to processing  
115 and analyzes of remote sensing data (e.g. Duro et al., 2012; Ali et al., 2015). Few exceptions are

116 the development of crop nutrient response models for studying yield responses in sub-Saharan  
117 Africa based on field trial data (Hengl et al., 2017b) and the use of data mining tools for  
118 identifying crop growth limitations (Delerce et al., 2016).

119



120

121 **Figure 1.** Schematic representation of the downscaling approach presented in this study.

122 Machine-learning derived meta-models trained on global crop model outputs and  
123 covariates at a comparably low spatial resolution are used for producing regional  
124 estimates of corresponding variables at a higher spatial resolution.

125 In this study, we evaluate machine learning as an approach for building crop meta-  
126 models. The focus is on the feasibility to use low-resolution global crop simulations of maize  
127 yield potential for predictions at a high resolution, here exemplary the extent of Mexico, as

128 depicted schematically in Figure 1. Non-nutrient and pest limited yield potentials (Lobell et al.,  
129 2009) with and without sufficient water supply were selected as a target variable as they allow  
130 for a thorough evaluation of climate-related covariates without inference from soil nutrient  
131 trajectories. Two of the presently most flexible and in recent competitions best performing  
132 (Fernández-Delgado, 2014; Chen and Guestrin, 2016) machine learning approaches for numeric  
133 predictions, extreme gradient boosting and random forests, are tested and compared against crop  
134 model simulations carried out at the finer resolution. Objectives of the study are to (a) evaluate  
135 the meta-model performance in downscaling the low-resolution global yield simulation to high-  
136 resolution predictions in the study region of Mexico, (b) identify most important covariates  
137 required by the meta-model, and (c) test the approach for predictions of selected agricultural  
138 externalities across scales. To provide an exemplary application case, machine learning model  
139 predictions are performed at a very high spatial resolution (1 km x 1 km) in major producing  
140 areas and benchmarked against reported inter-annual yield variability, a key performance  
141 indicator for climate change impact assessments (Müller et al., 2017). Finally, an outlook  
142 provides suggestions for further steps to extend the models' capabilities.

## 143 **2 Methods and Data**

### 144 2.1 Gridded crop model description

145 Crop simulations were carried out using a gridded version of the Environmental Policy  
146 Integrated Climate model (EPIC). EPIC was initially developed to assess the impacts of  
147 management on crop yields (Williams, 1995). It has constantly been updated to cover additional  
148 processes such as effects of elevated atmospheric CO<sub>2</sub> concentration on plant growth (Stockle et

149 al., 1992), detailed soil organic matter cycling (Izaurrealde et al., 2006, Izaurrealde et al., 2012),  
150 and an extended number of crop types and cultivars (e.g. Kiniry et al., 1995; Gaiser et al., 2010)  
151 among others (see Gassman, 2004). More details of the crop growth model are provided in  
152 Supplementary Text S1.

153 The gridded version of EPIC used here, EPIC-IIASA (Balkovič et al., 2014), runs the  
154 EPIC model for a given set of simulation units derived from intersecting homogenous response  
155 units (soil and topography), administrative borders, and climate grids (Skalský et al., 2008).  
156 Thereby, each simulation unit is treated as a representative, homogenous field.

## 157 2.2 Study regions, delineation of simulation units, and simulation period

158 Simulations and meta-model predictions were performed (a) at the global scale at a  
159 coarse spatial resolution and (b) for Mexico at a finer resolution. The latter was selected as an  
160 exemplary study region as it encompasses the three major climates tropic, temperate, and (semi-  
161 )arid and has a large coverage of maize harvest areas. The basic spatial resolutions at the two  
162 scales were grids of 5' (global) and 0.5' (Mexico), respectively, serving also as basic references  
163 for spatial harmonization of all underlying input data (topography, soil, and land cover).  
164 Individual pixels were aggregated to homogeneous response units (HRUs) based on slope,  
165 altitude and soil classes. HRU provide aggregated spatial units which are expected to be  
166 homogenous in their bio-physical response and relatively stable over time. The basic bio-  
167 physical drivers assumed for an HRU are hardly adjustable by farmers, which allows for  
168 analyzing impacts of the same management practices employed across a variety of natural  
169 conditions. Intersecting HRUs with administrative units (countries globally and states for

170 Mexico) and the climate grids of  $0.5^\circ \times 0.5^\circ$  and  $0.25^\circ \times 0.25^\circ$  resolution at the global and  
171 Mexican scale, respectively, resulted in final simulation units with a total number of  $1.3 \times 10^5$   
172 globally and  $2.3 \times 10^5$  for Mexico. Spatially explicit inputs for EPIC on topography and soil were  
173 then calculated as mean (altitude) or majority (slope, soil) values across all pixels within the  
174 simulation unit. Additional evaluations were carried out for the Mexican state of Jalisco, which is  
175 the top rainfed maize producing state in the country according to Servicio the Información  
176 Agroalimentaria y Pesquera (SIAP, 2018b).

177 Simulations were performed for the years 1980-2010 based on climate data coverage  
178 (Section 2.3.1) and evaluated for the period 1990-2009 as the crop model equilibrates during the  
179 first simulation years and the global simulations used for training machine learning models did  
180 not provide outputs for the year 2010 in regions with growing seasons crossing years.

## 181 2.3 Crop model input data

### 182 2.3.1 Climate data

183 Gridded climate data were obtained from the publicly available AgMERRA climate  
184 dataset (Ruane et al., 2015) at spatial resolutions of  $0.5^\circ \times 0.5^\circ$  for global simulations and  
185 predictions and  $0.25^\circ \times 0.25^\circ$  for the study region of Mexico. AgMERRA covers the period  
186 1980-2010 and combines data from the Modern-Era Retrospective Analysis for Research and  
187 Applications (MERRA; Rienecker et al., 2011), station data, and remotely sensed datasets and  
188 has been bias corrected using stations from agricultural land only. The high-resolution version  
189 was obtained from the providers' website directly, the coarser resolution was provided through  
190 the Global Gridded Crop Model Intercomparison (GGCMI) project (Elliott et al., 2015).

191 Although higher resolution monthly climate data would be available for the study region (e.g.  
192 Wang et al., 2016) allowing for higher resolution meta-model predictions, these would not allow  
193 for benchmarking against EPIC simulations requiring daily climate data.

#### 194 2.3.2 Soil data

195 Soil data were retrieved from the Harmonized World Soil Database v1.2 (HWSD;  
196 FAO/IIASA/ISRIC/ISS-CAS/JRC, 2012) at both spatial scales. For each grid cell at 5' (global)  
197 or 0.5' (Mexico) resolution, the dominant soil type of the largest soil mapping unit was selected  
198 as the representative soil type. Soil characteristics considered in EPIC and the machine learning  
199 approaches are depth, texture, coarse fragment content, bulk density, soil organic carbon content,  
200 pH, electric conductivity, cation exchange capacity, base saturation, and carbonate content  
201 (Table 1).

#### 202 2.3.3 Topography

203 For the global setup, elevation data were adopted from GTOPO30 (USGS, 2002)  
204 calculating the mean elevation in each simulation unit. Slope classes were obtained from the  
205 Global Agro-ecological Zones Assessment for Agriculture (GAEZ; Fischer et al., 2012). For the  
206 high-resolution setup constructed for Mexico, both elevation and slopes were derived from the  
207 SRTM 4.1 database provided by CIAT-CSI (Jarvis et al., 2008).

#### 208 2.3.4 Land use

209 Global low-resolution simulations were carried out for all simulation units presently  
210 containing cropland according to at least one of the datasets Global Land Cover 2000 database

211 (Global Land Cover 2000 database, 2003) or SPAM (You et al., 2017). For Mexico, simulations  
212 were done for all simulation units and MIRCA2000 was used for identifying simulation units  
213 containing relevant maize harvest area, here defined as >5% of total area. Selected analyses were  
214 restricted to these in order to evaluate model performance for the whole land and relevant  
215 cropland only.

### 216 2.3.5 Crop management

217 Maize was used as a model crop due to its extensive cultivation globally and in Mexico.  
218 Default crop parameters from the EPIC model were used, which reflect a high-yielding variety  
219 adapted to warm climate (Kiniry et al., 1995). Crop growing seasons were adopted at both scales  
220 from Sacks et al. (2010) as provided by Elliott et al. (2015). PHU were calculated from planting  
221 to harvest using long-term monthly climate data for the whole time-period covered by the  
222 AgMERRA climate dataset (1980-2010) at each spatial resolution separately.

223 To obtain non-nutrient limited maize yield potentials (Lobell et al., 2009), mineral N  
224 fertilizer was applied automatically by the EPIC model based on plant stress to avoid plant  
225 growth limitations due to nutrient deficits, which may cause trends in yields over time due to  
226 nutrient mining. The maximum applied amount of fertilizer was set to 500 kg N ha<sup>-1</sup> yr<sup>-1</sup>, which  
227 is commonly more than sufficient for maximizing maize yields (e.g. Folberth et al., 2013).  
228 Simulations were carried out with water supply either from precipitation only (rainfed) or with  
229 sufficient supplementary irrigation water supply (fully irrigated). Irrigation water was applied  
230 based on plant stress analogously to fertilizer with an annual maximum volume of 2000 mm.

231 Other management practices were kept at a basic level with four operations in each season: field  
232 cultivation, planting, harvest, and stover removal.

## 233 2.4 Machine learning framework

234 We test two state-of-the-art tree-based ensemble methods, extreme gradient boosting and  
235 random forests. Ensemble methods employ a collection of learning algorithms to achieve better  
236 predictive power than could be gained from any of these algorithms alone. For ensembles such as  
237 extreme gradient boosting and random forests, it is typical to use trees as building blocks to  
238 allow for invariance to scaling of inputs and complex interactions between features. Since  
239 ensembles have additional parameters responsible for aggregation of learning algorithms, they  
240 have more flexibility in fitting training data than single-algorithm approaches do. Thus,  
241 ensembles are more prone to overfitting. Overfitting is prevented through out-of-bag error  
242 monitoring, n-fold cross-validation, correction of the ensemble by regularization that makes the  
243 training procedure more conservative, and testing on the holdout dataset covering 25% of  
244 observations (see below). Both extreme gradient boosting and random forests are insensitive to  
245 multiple correlation of covariates with respect to prediction accuracy and overfitting. The  
246 quantification of variable importance, however, may be affected if covariates are strongly  
247 correlated (see Section 2.4.3).

248 Crop model simulation data (serving here as observations) for building machine learning  
249 models was randomly split into training and validation sets containing 75% and 25% of samples,  
250 respectively, which is a common split ratio in machine learning. About  $19.5 \times 10^5$  samples  
251 (simulation units x simulation years) were used for model training and  $6.5 \times 10^5$  for validation.

252 Machine learning models were built separately for the two water management scenarios, rainfed  
253 or sufficiently irrigated, within the statistical computing software R (R Development Core Team,  
254 2008) using the packages specified in the following sections.

255 To streamline the presentation of results, the main body of the paper focuses on results  
256 from extreme gradient boosting. The evaluation of the random forests models is presented in the  
257 SI and discussed within the main body where relevant.

#### 258 2.4.1 Extreme gradient boosting

259 Similar to other boosting methods, extreme gradient boosting is an ensemble learning  
260 technique that sequentially builds the model: each tree is fit on a modified version of the original  
261 training data set. I.e., every new tree uses information from previously grown trees. This is the  
262 key difference to random forests (see below). Extreme gradient boosting generalizes boosting  
263 methods by allowing minimization of an arbitrary differentiable loss function. In this study, we  
264 employed the R package XGBoost for extreme gradient boosting, a highly efficient realization of  
265 the gradient boosting approach that showed the best performance in recent machine learning  
266 challenges (Chen and Guestrin, 2016). Being a learning algorithm with high flexibility, extreme  
267 gradient boosting is prone to overfitting, especially, if training data are scarce, which is not the  
268 case here. Typically, parameter tuning is done by performing an exhaustive grid search along  
269 parameter dimensions using the default parameters as the reference point. This was here not  
270 considered meaningful due to the vast amount of training data, rendering a full grid search  
271 computationally inefficient and unneeded, due to extremely low error obtained already in a  
272 limited grid search. I.e., we tuned only key parameters for shrinkage and learning ( $\eta$ ),

273 max\_depth, nrounds; Table S1). In our case, the default parameter values resulted in stable but  
274 improvable performance with  $R^2=0.94$  for the test dataset. This suggested to increase the  
275 maximum tree depth and local variation of the learning rate (eta). The grid search resulted in  
276  $R^2=0.99$  for both training and test data with eta=0.15 or 0.30 and max\_depth=15 or 20. The  
277 lowest RMSE in both training and test data was obtained with eta=0.15 and max\_depth=20 in a  
278 five-fold cross validation (Table S2). Although this parameter set results in a marginal overfit, it  
279 also showed the best performance in regression metrics and mean absolute error (MAE; not  
280 shown), the main performance indicators used herein (see section 2.5.1). It was hence selected  
281 for performing the predictions. Extending the grid search to by increasing the rounds of tree  
282 building (nrounds) from 60 to 100 provided only a negligible increase in performance (Table  
283 S2). Resulting parameters were hence eta=0.15, max\_depth=20, and – to ensure very high  
284 accuracy - nrounds=100.

285         Since extreme gradient boosting may produce negative predictions even if the training  
286 data does not have them, the lower boundary was set to zero and all predictions below corrected  
287 to this value. This was the case for rainfed crop yields in 0.1% of samples with predictions of up  
288 to  $-0.19 \text{ t ha}^{-1}$  in the validation set and 0.02% of the predictions for Mexico with up to  $-0.08 \text{ t ha}^{-1}$   
289 <sup>1</sup>. Irrigated crop yield predictions were affected in the validation set only with up to  $-0.09 \text{ t ha}^{-1}$  in  
290  $<0.01\%$  of samples.

#### 291 2.4.2 Random forests

292         In contrast to boosting methods, tree ensembles build a number of models in parallel  
293 from which average predictions are derived. Bagging is a basic approach to introduce an

294 ensemble that consists of a number of decision trees trained on random subsets of data  
295 (bootstrapped training samples). Random forests (Breiman, 2001) employ not only bagging (row  
296 sub-sampling) but also column sub-sampling, i.e., every time a split in a tree is examined for a  
297 random subset of candidate features drawn from the full set of features. This effectively de-  
298 correlates the trees. As reported in a recent meta-study of machine learning algorithms  
299 (Fernández-Delgado, 2014), random forests was identified as the best family of classifiers. In  
300 this study, random forests models were constructed using the R package h2o, which serves as a  
301 link to the H2O.ai machine learning cluster environment (The H2O.ai team, 2017).

302         As random forests are less prone to overfitting, global parameters were tuned to achieve a  
303 reasonable balance between performance and computational demand, which increases linearly  
304 with number of trees and tree depth. Major parameters to adjust in random forest are number of  
305 trees (ntrees), maximum tree depth (max\_depth), and a number of features considered for each  
306 split decision (mtries). The latter is per default one third of total features for numeric predictions.  
307 Starting from the default values ntree=50, max\_depth=20, and mtries=[number of features]\*0.3,  
308 we found an increase in performance in terms of regression coefficients and MAE of the test  
309 dataset up to max\_depth=30 with negligible improvements if ntree was increased from 50 to 80  
310 (Figure S1). Further increasing the parameter values provides a marginal increase, but would not  
311 justify the increase in computational demand, which is already at any point substantially higher  
312 than for extreme gradient boosting (see also section 4.4). Increasing or decreasing the parameter  
313 mtries from about 33% of feature number as a default to 20% or 50% affected model  
314 performance only marginally as well with no changes in  $R^2$  or slope and changes by  $\pm 0.01 \text{ t ha}^{-1}$   
315 in intercept and MAE.

### 316 2.4.3 Feature importance

317 Both methods determine feature importance internally. To obtain an overall summary of  
318 the importance of predictors, the residual sum of squares (for regression) or the Gini index (for  
319 classification; Breiman et al. 1984) are used. For ensembles of regression trees, the total amount  
320 by which the residual sum of squares is decreased by splits over a fixed feature is calculated and  
321 then average over all trees. Larger values point to predictors that are more important. Likewise,  
322 in the case of ensembles of classification trees, the total amount that the Gini index is reduced  
323 due to splits is cumulated over a given feature and averaged over all trees. For both machine  
324 learning methods, we present the relative importance of each feature as percentage. Due to  
325 differences in the estimation of feature importance, it is not feasible to compare importance  
326 across different algorithms quantitatively. In addition, multiple correlated features, which can be  
327 expected here at least among soil characteristics or (monthly) climate variables, are known to  
328 bias the quantification of feature importance (Tološi and Lengauer, 2011). E.g., if two features  
329 included in an extreme gradient boosting model are perfectly correlated, each of them will  
330 receive 50% of the actual importance. For these reasons, we focus in the evaluation of feature  
331 importance foremost on the ranking of features rather than their quantitative contributions.

### 332 2.4.4 Machine learning features and feature engineering

333 **Table 1.** Features and target variables used in machine learning experiments. Several statistics  
334 were calculated for each climate variable VAR in the first section of the table as listed in  
335 the second section. Averages were calculated for the temperature indices TMX and TMN,  
336 sums for all others. Total number of features is 247, the maximum number used in model

337 training is 151 (Table 2). The attributes transient and static in the section headings refer  
 338 to the temporal dimension.

Abbreviation	Variable description
<b>Climate variables (VARs; transient)</b>	
TMX	Maximum temperature [°C]
TMN	Minimum temperature [°C]
GDD	Growing degree days [°C]
RAD	Solar radiation [MJ m <sup>-2</sup> ]
PET	Potential evapotranspiration [mm]
PRCP	Total precipitation [mm]
WET	Wet day frequency [d]
CMD	Climatic moisture deficit (PRCP-PET) [mm]
<b>Temporal aggregates and derivatives of climate variables (transient)</b>	
VAR_X	Monthly value for month X {1:12} since planting (e.g. "TMX_1")
VARsd_X	Standard deviation of mean value in month X {1:12} (e.g. "TMXsd_1")
VARavYRcal	Average of climate variable in calendar year (January to December)
VARsumYRcal	Sum of climate variable in calendar year (January to December)
VARavYRgs	Average of climate variable in growing season year (12 months from planting)
VARsumYRgs	Sum of climate variable in growing season year (12 months from planting)
VARskYRgs	Skew of climate variable in growing season year (12 months from planting)
VARavGS	Average of climate variable in growing season (planting month to harvest)
VARsumGS	Sum of climate variable in growing season (planting month to harvest)
VARskGS	Skew of climate variable in growing season (planting month to harvest)
<b>Soil and site variables (static)</b>	
DEPTH	Total soil depth [m]
SAND	Sand content in topsoil [%]
CLAY	Clay content in topsoil [%]
PH	pH in topsoil [-]
SB	Sum of bases in topsoil [cmol kg <sup>-1</sup> ]
CEC	Cation exchange capacity in topsoil [cmol kg <sup>-1</sup> ]
EC	Electric conductivity in topsoil [mmho cm <sup>-1</sup> ]
ROK	Coarse fragment (rock) content in topsoil [%]
BD	Bulk density in topsoil [g cm <sup>-3</sup> ]
CARB	Carbonate content in topsoil [%]
OC	Organic carbon content in topsoil [%]
PAW	Total plant available water capacity [m <sup>3</sup> m <sup>-3</sup> ]
HG	Soil hydrologic group (water infiltration potential) [-]
SLP	Hill slope [%]
<b>Cultivar and growing season variables (static)</b>	
PHU	Potential heat units/growing degree days from planting to maturity [°C]
LVP	Length of vegetation period. Average days from planting to maturity [d]
<b>Target variables (transient)</b>	
YLDG	Maize crop yield [t ha <sup>-1</sup> ]
CAW	Crop available water [mm]
GSET	Growing season evapotranspiration [mm]

339

340 Features are based on crop model input data, i.e. soil, climate and management  
341 specifications as described in Section 2.3. Daily climate data were in a first step aggregated to  
342 monthly sums or averages depending on the variable. For each simulation unit, the month of  
343 planting was designated as month 1 to harmonize the order of months from planting globally.  
344 Subsequently, annual and growing season values were calculated for (a) the growing season  
345 months (based on the static length of reported vegetation period (LVP)), (b) the calendar year,  
346 and (c) a year starting from the planting month (Table 1). This process is referred to as feature  
347 engineering, i.e. the specification of model features beyond raw data based on expert knowledge.

348 Soil variables were foremost adopted for the topsoil, which has the largest impact on crop  
349 growth. Only variables with high importance for water availability, depth, plant available water  
350 capacity (PAW; difference of water contents at field capacity and wilting point), and hydrologic  
351 soil group (HG) refer to the whole soil profile. Additional characteristics considered potentially  
352 relevant for the meta-models were hill slope as a site characteristic and PHU and LVP as cultivar  
353 characteristics.

354 Models were built for three target variables: maize crop yield (yield hereafter), growing  
355 season ET (GSET), and crop available water (CAW). The latter is a balance of initial soil  
356 humidity at the beginning of the growing season, growing season precipitation and irrigation  
357 water if provided, surface runoff, and percolate.

358 To evaluate the importance of raw and engineered climate features, the machine learning  
359 models were trained with various feature subsets (Table 2). Soil and site data, PHU, and LVP  
360 were considered in all scenarios to evaluate the importance of climate variables only. Annual  
361 climate data can be considered the most general feature set. Growing season climate considers

362 the mean or sum of climatic conditions experienced by the crop. Monthly data in turn account for  
 363 intra-seasonal variability and climate effects in certain growth stages. The complete climate  
 364 feature set takes all aspects into account and solely lets the algorithm select the most relevant  
 365 features. Thereby, months beyond the sixth from planting were excluded to keep the number of  
 366 features at a reasonable extent, considering that maize cultivars hardly require >180 days to  
 367 reach maturity.

368 **Table 2.** Climate feature subsets used in the analyses. Besides indicated climate features (see  
 369 Table 1 for details), soil and site data, PHU, and LVP were considered in all training sets.

Feature subset	Climate features considered	Number of features
annual climate	<i>VARavYRcal</i> , <i>VARsumYRcal</i>	23
growing season climate	<i>VARavGS</i> , <i>VARsumGS</i>	23
monthly climate	<i>VAR_X</i> (with $X \leq 6$ )	63
complete climate	all features except <i>VAR_X</i> with $x \geq 7$	151

370

## 371 2.5 Performance metrics and model evaluation

### 372 2.5.1 Machine learning model performance compared to crop model simulations

373 Model performance was assessed using linear regression of (a) meta-model predictions  
 374 against the validation subset of global EPIC simulations and (b) downscaling predictions against  
 375 the high-resolution benchmark simulations for Mexico. Mean absolute error (MAE) was used as  
 376 a metric for mean model bias. Nash-Sutcliffe efficiency (NSE) was used as an indicator for the  
 377 accuracy of inter-annual yield variability.

378 The coefficient of determination  $R^2$  was calculated according to

379 
$$R^2 = 1 - \frac{\sum_{i=1}^n (Y_{ref,i} - \hat{Y}_{pred})^2}{\sum_{i=1}^n (Y_{ref,i} - \bar{Y}_{ref})^2} \quad (1)$$

380 where  $i$  is the number of the sample point (one simulation year-location) considered,  $n$  is  
 381 the total number of sample points across simulation units and years,  $Y_{ref}$  is the reference crop  
 382 yield,  $\hat{Y}_{pred}$  is the fitted yield, and  $\bar{Y}_{ref}$  is the arithmetic mean of reference samples.

383 MAE was calculated as

384 
$$MAE = \frac{\sum_{i=1}^n |Y_{pred,i} - Y_{ref,i}|}{n} \quad (2)$$

385 where  $Y_{pred,i}$  is the machine learning model predicted value for data point  $i$  and  $Y_{ref,i}$  is the  
 386 corresponding EPIC simulated reference value.

387 NSE is a common metric for model performance over time, used especially in hydrology  
 388 (Nash and Sutcliffe, 1970). It is calculated using the same variables as the prior metrics but  
 389 separately for each simulation unit over time according to

390 
$$NSE = 1 - \frac{\sum_{t=1}^m (Y_{ref,t} - Y_{pred,t})^2}{\sum_{t=1}^m (Y_{ref,t} - \bar{Y}_{ref})^2} \quad (3)$$

391 where  $Y_{pred,t}$  is the yield estimated by the meta model for year  $t$  and  $Y_{ref,t}$  the  
 392 corresponding reference. NSE can range from  $-\infty$  to  $+1$  with  $NSE > 0$  indicating that model  
 393 predictions are more useful than the mean of reference data. As NSE is sensitive to both absolute  
 394 values and their temporal dynamics, it was in addition calculated for zero-centered yield values  
 395 (sample mean removed) in order to assess inter-annual yield variability alone, which is  
 396 considered a vital GCM evaluation characteristic for climate (change) impact assessments (e.g.  
 397 Müller et al., 2017).

398 Evaluations were partly carried out at the level of major Koeppen-Geiger climate regions  
399 (Figure S2) following the rules of Peel et al. (2007). Koeppen-Geiger regions were identified for  
400 each 0.25° x 0.25° climate grid for the 31-year climatology of the AgMERRA dataset 1980-  
401 2010.

#### 402 2.5.2 Model performance compared to regional statistics

403 The EPIC model itself and the global gridded EPIC-IIASA framework have been  
404 evaluated and validated thoroughly at various scales from the agricultural plot (Kiniry et al.,  
405 1995; Gassmann et al., 2004; Izaurrealde et al., 2006) to regional (Gaiser et al., 2010; Folberth et  
406 al., 2012) and global assessments (Balkovič et al., 2014; Müller et al., 2017) finding good  
407 agreement with reported yields. Here we provide a brief evaluation of model performance in  
408 terms of inter-annual yield variability expressed as NSE (eq. (3)) for the top ten maize producing  
409 municipios (second-level administrative units) of the major maize producing state Jalisco, where  
410 crop management can be considered fairly stable and data quality reasonable. This also illustrates  
411 an exemplary application of the machine learning framework. Reported maize yields were  
412 obtained from SIAP (2018a). Crop yields are reported since the year 2003 at the second  
413 administrative level, resulting in an evaluation period from 2003-2009 considering the time  
414 period for crop model simulations (see Section 2.2). Besides the machine learning predictions  
415 corresponding to the high-resolution input data for the crop simulations at the scale of Mexico  
416 (see Section 2.3.1), predictions were also produced using monthly climate surfaces from  
417 ClimateNA 5.60 (Wang et al., 2016) at a spatial resolution of 1 km x 1 km and a national soil  
418 dataset (INEGI, 2004) besides HWSD to assess the impact of higher resolution climate data and  
419 regional soil data products, a major application opportunity for the methodology presented

420 herein. Maize planting dates recorded in the year 2017 were obtained from SIAP (2018b). All  
421 yields were de-trended linearly to correct for changes in management intensity.

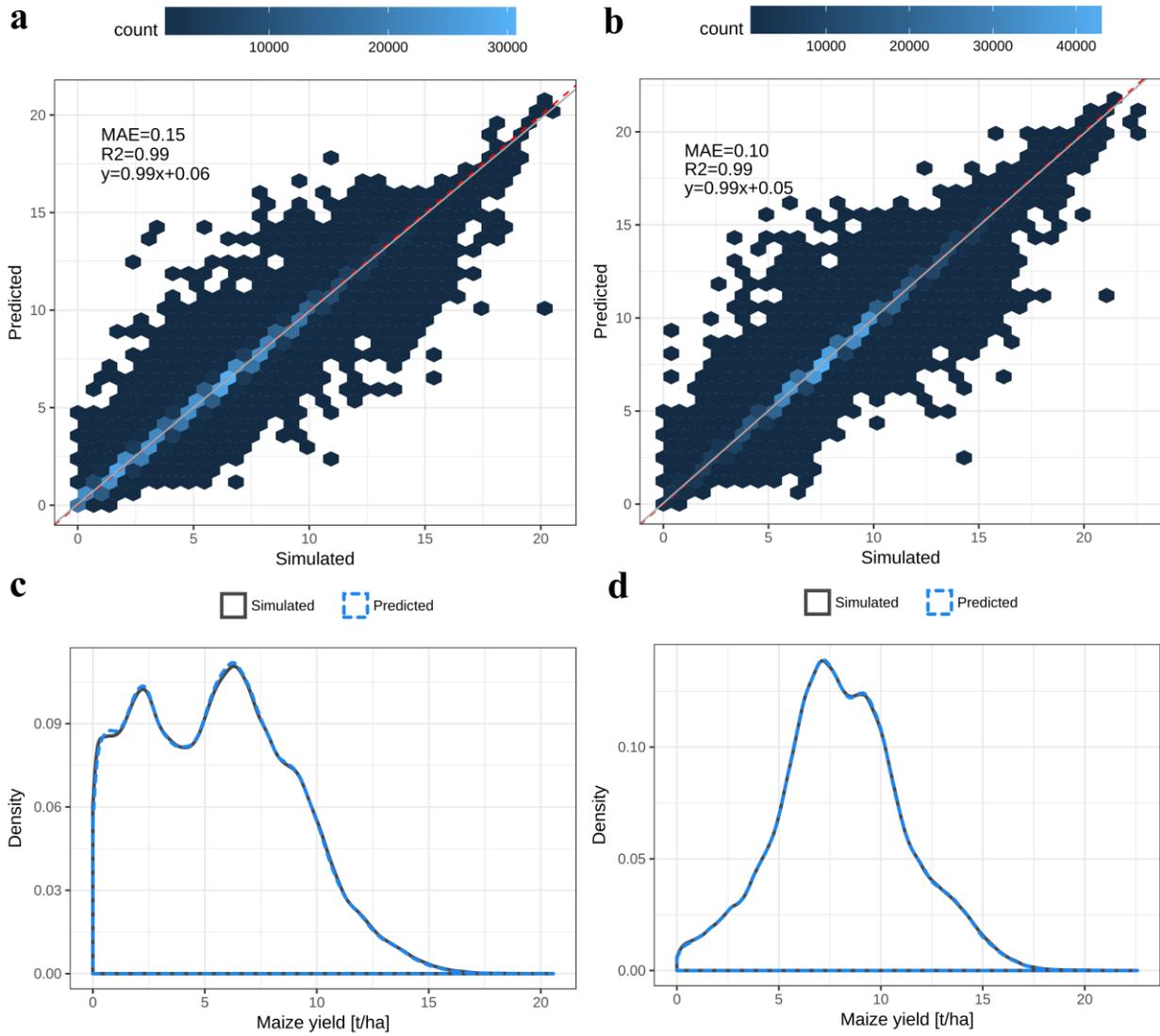
## 422 2.6 Computational framework

423 All computations, evaluations and plotting were done within the R software environment  
424 (R Development Core Team, 2008). Machine learning models were built using the packages  
425 specified in sections 2.4.1 and 2.4.2. Figures were produced using ggplot2 (Wickham, 2009).  
426 Statistical analyses beyond linear regression were carried out with hydroGOF (Zambrano-  
427 Bigiarini, 2017).

## 428 **3 Results**

### 429 3.1 Global scale model performance for crop yields

430 The global extreme gradient boosting meta-models for irrigated and rainfed maize yields  
431 based on the full climate features show a near perfect fit and low mean bias in both cases (Figure  
432 2a,b). Large over- and underestimations in predictions are rare. The first occur foremost at low  
433 simulated yields, the latter at high ones with a negative trend beyond 12 t ha<sup>-1</sup> (see Figure S3a,b  
434 for residual plots). For rainfed yields, noticeable deviations in density distributions of EPIC  
435 simulated and extreme gradient boosting predicted yields occur below 2 t ha<sup>-1</sup> and around 6-7 t  
436 ha<sup>-1</sup> (Figure 2c). The density distributions are nearly identical for irrigated yields (Figure 2d).



437

438 **Figure 2.** Hexbin and regression plots for EPIC simulated and extreme gradient boosting  
 439 predicted crop yields in the validation dataset (25% of total samples) for (a) rainfed and  
 440 (b) irrigated conditions and corresponding density distributions for (c) rainfed and (d)  
 441 irrigated conditions. Red dashed and grey solid lines in (a) and (b) show 1:1 line and  
 442 regression, respectively. See section 3.4 and SI for random forest models.

## 443 3.2 Performance of crop yield predictions for Mexico

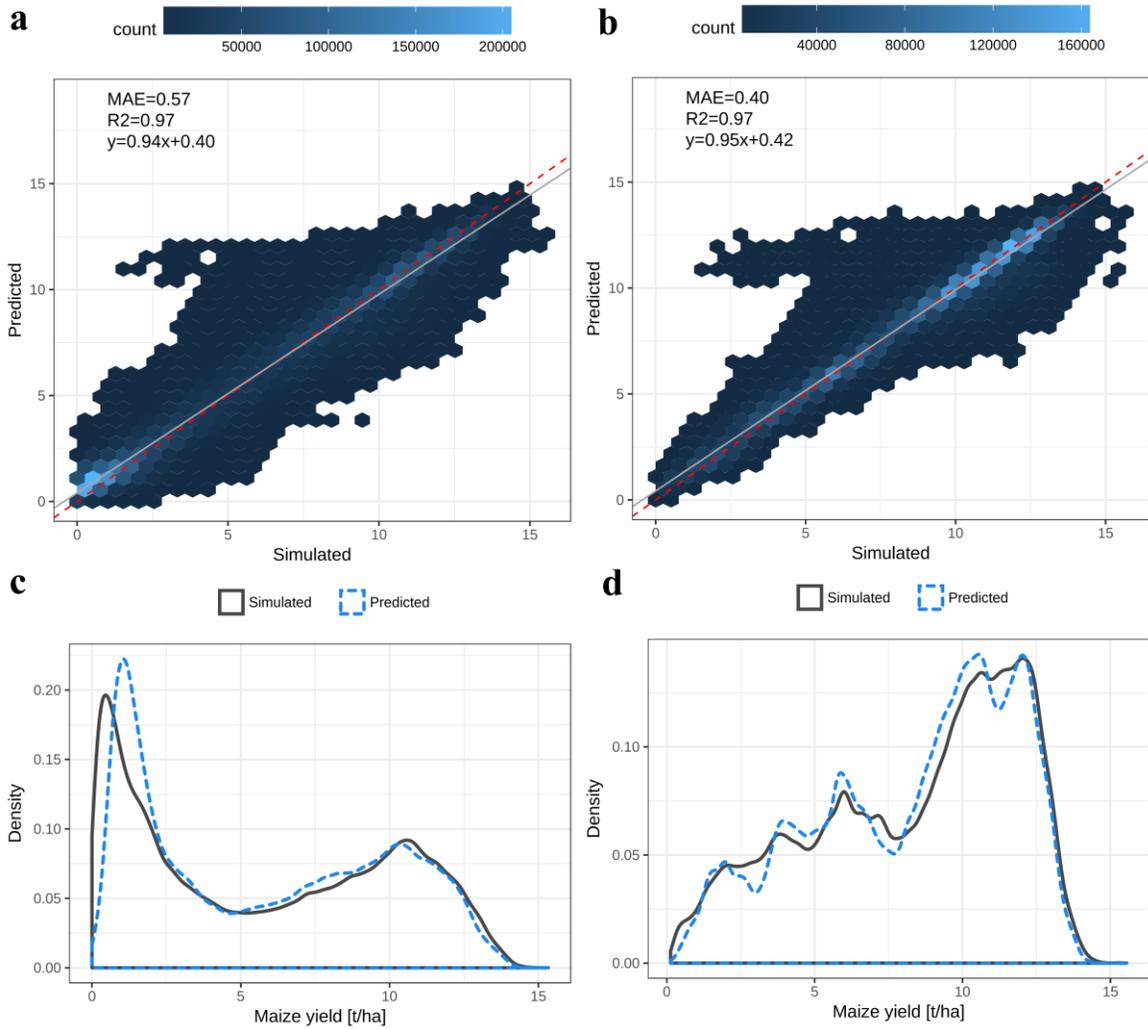
### 444 3.2.1 General performance and patterns

445 The accuracy of rainfed and irrigated yield predictions for Mexico at a high spatial  
446 resolution (Figure 3a,b) is nearly up to that of the global validation data with 97% of variance of  
447 EPIC simulated yields explained by the extreme gradient boosting models in both cases. Slopes  
448 of the linear regressions are lower and the intercepts are higher than at the global scale indicating  
449 biases at the lower and upper bounds of simulated yields. MAE increases by up to 0.5 t ha<sup>-1</sup> but  
450 is still considerably low concerning the mean of crop yield estimates. Overestimations by >100%  
451 occur in both water management scenarios with a cluster of data points around 3.5 t ha<sup>-1</sup> of EPIC  
452 simulated yields. These are related to remaining nitrogen stress in few simulations (0.5% of  
453 samples) due to extreme soil-climate combinations on which the automatic fertilizer application  
454 of up to 500 kg N yr<sup>-1</sup> does not suffice to fulfill plant requirements caused by vast losses of N in  
455 runoff. Removing these simulations has no discernible effect on model performance (Figure S4).

456 The distributions of rainfed yield estimates and predictions exhibit a bimodal pattern with  
457 over- and underestimation especially at the lower bound where the peak is shifted by about 1 t  
458 ha<sup>-1</sup> (Figure 3c). This is to a lesser extent also the case for the distributions of irrigated yield  
459 estimates and predictions (Figure 3d). In addition, irrigated yields predicted by the extreme  
460 gradient boosting model exhibit clustering, i.e. with overestimation peaks around 4, 5.5, and 10 t  
461 ha<sup>-1</sup> and valleys at 3 and 12 t ha<sup>-1</sup>, while EPIC simulated yields show a smoother distribution.

462 Using the more parsimonious climate feature sets decreases model performance (Table  
463 S5) similar to the global scale validation data (Table S4). The largest decrease occurs for the

464 most set of growing season climate data, while again hardly any difference is found when using  
465 the monthly climate features.

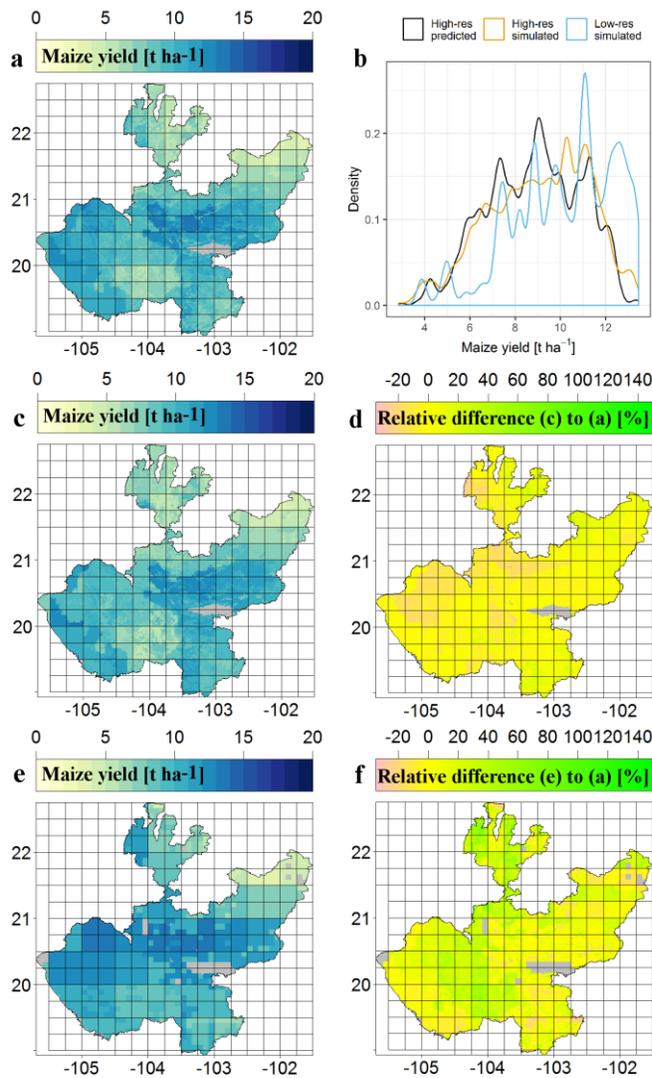


466

467 **Figure 3.** Same as Figure 2 but comparing the high-resolution downscaled predictions and  
468 benchmark EPIC simulations for Mexico.

469 Comparing low-resolution simulations, high-resolution simulations, and high-resolution  
470 machine learning predictions at the scale of a single state of Jalisco for rainfed maize yields in

471 the year 2000 shows that the machine learning predictions can fairly well reproduce the  
472 heterogeneity seen in the high-resolution simulations (Figure 4a,c). Notable differences are  
473 apparent in the region west of  $-104.5^{\circ}$  and north of  $20^{\circ}$ , where the predictions are about 20%  
474 lower than the simulation results and parts of the southern and northern state where predictions  
475 are up to 40% higher (Figure 4d). Overall, the distributions of yields agree fairly well (Figure  
476 4b), but the predictions omit moderate and very high yields, indicating peaks around 7.5 and 9 t  
477  $\text{ha}^{-1}$  and a valley at  $10.5 \text{ t ha}^{-1}$ , which are not present in the simulations. Still, yield predictions  
478 and simulations are correlated with  $R^2=0.87$  (Figure S5a).



479

480

**Figure 4.** Examples of rainfed maize yields for the year 2000 in the state of Jalisco from (a)

481

high-resolution EPIC simulation, (c) high-resolution machine learning prediction, and (e)

482

global low-resolution simulation. (b) Shows the corresponding density distributions for

483

which yield estimates from the low-resolution simulations have been resampled to the

484

higher resolution to obtain at consistent sample sizes. (d) and (f) show the relative

485

differences of (c) and (e) compared to (a), respectively. Regressions and statistics are

486

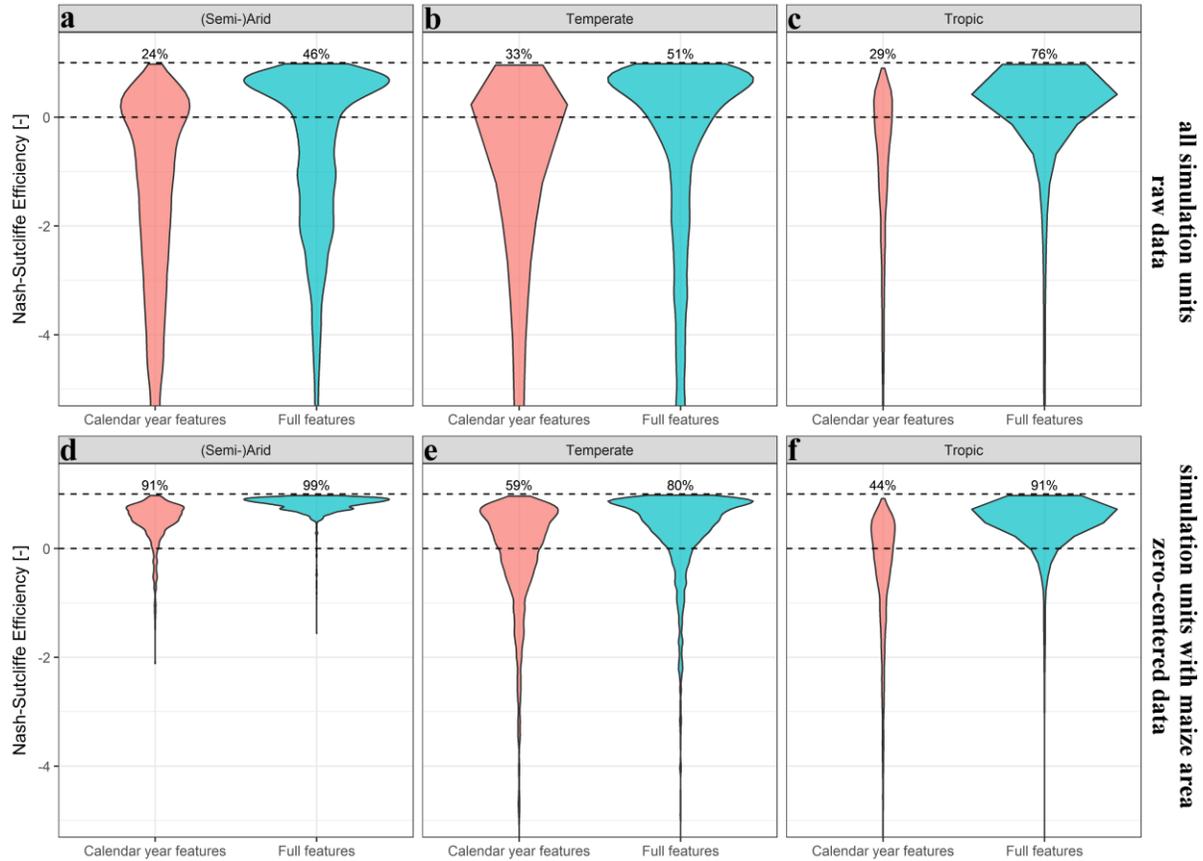
presented in Figure S5a,b. The rectangular grid represents the  $0.25^\circ \times 0.25^\circ$  climate grid.

487           Expectedly, low-resolution EPIC estimates (Figure 4e) agree only with respect to large-  
488 scale patterns. Substantial overestimation by up to 60% occur in the central parts and  
489 underestimation by up to 30% foremost in the west but also scattered at the subgrid level (Figure  
490 4f). The yield distribution is biased towards higher yield estimates (Figure 4b) and the coefficient  
491 of determination is  $R^2=0.64$  (Figure S5b). The arithmetic means at the state level are  $9.06 \text{ t ha}^{-1}$   
492 for the high-resolution simulations,  $8.85 \text{ t ha}^{-1}$  for the predictions, and  $10.15 \text{ t ha}^{-1}$  for the low-  
493 resolution simulations, corresponding to an overestimation by 11.98% for the low-resolution  
494 simulations and an underestimation by 2.31% for the extreme gradient boosting predictions.  
495 Hence, despite remaining differences, the high-resolution predictions reproduce the  
496 corresponding simulations quite robustly compared to the EPIC outputs derived from more  
497 granular input data.

### 498 3.2.2   Reproduction of inter-annual crop yield variability

499           NSE is greater than zero in around 20-30% of all simulation units for predictions of  
500 rainfed yields by the model based on calendar year climate features alone (Figure 5a-c). The  
501 model trained with the full set of climate features in contrast shows a substantially better  
502 performance, especially in tropic climates. If simulated and predicted yields are zero-centered  
503 and only present cropland is considered, NSE performance turns out substantially better for both  
504 feature sets (Figure 5d-f) and again to a very high degree for the extreme gradient boosting  
505 model trained on the full climate feature set.

506



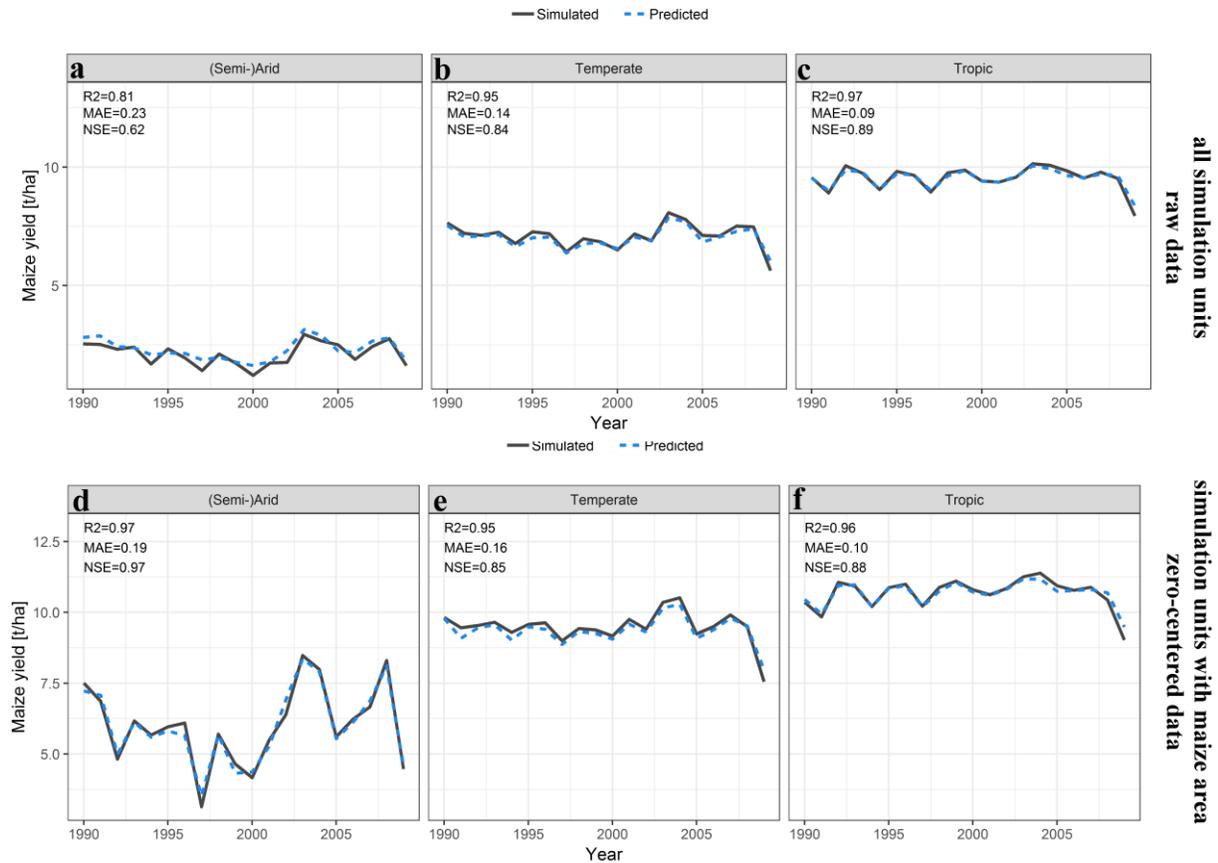
507

508 **Figure 5.** Violin plots of Nash-Sutcliffe Efficiency disaggregated by major Koeppen-Geiger  
 509 climate regions (see section 2.5) for the feature subsets using calendar year climate  
 510 variables only or all climate features. (a-c) All simulation units of Mexico with raw data  
 511 or (d-f) only simulation units with >5% maize harvest area and zero-centered yield  
 512 variability. Percentages indicate the fraction of simulation units with  $NSE > 0$ .  
 513 Complementary statistics are provided in Table S6. The extent of the y-axis was limited  
 514 to -5 for better readability.

515           With sufficient irrigation water supply, NSE performance is overall lower while the  
516 patterns remain quite similar, resulting in only few simulation units with  $NSE > 0$  for the model  
517 based on annual climate data (Figure S6a-f). A key difference to rainfed yield estimates is the  
518 lower performance in (semi-)arid regions, where inter-annual yield variability decreases  
519 substantially if sufficient water is supplied.

520           At a higher level of spatial aggregation – here the arithmetic mean for major Koeppen-  
521 Geiger climate regions –, inter-annual dynamics are well represented when considering all  
522 simulation units (Figure 6a-c). Similar to the distributions presented above (Figure 5),  
523 performance is best in tropic climates and poorest in (semi-)arid regions, but NSE is in all cases  
524 well above zero and  $MAE < 0.25 \text{ t ha}^{-1}$ . If only present cropland is considered (Figure 6d-f),  
525 performance decreases marginally in tropic and temperate climates, while it improves  
526 substantially in (semi-)arid climate where mostly highly arid simulation units are now neglected  
527 and predominantly simulation units with erratic rainfall remain (not shown). Foremost the latter  
528 climate region shows that the yield predictions can quite well reflect both yield peaks and  
529 valleys.

530           If sufficient irrigation water is supplied, the agreement with EPIC simulations in terms of  
531 NSE decreases substantially in temperate climate if all simulation units are considered but  
532 remains very similar in tropics and (semi-)arid climate (Figure S7a-c). For present cropland  
533 alone, the agreement in terms of NSE decreases most in (semi-)arid climate compared to rainfed  
534 yield estimates, followed by temperate regions. Predictions for the tropics still show very good  
535 agreement.



536

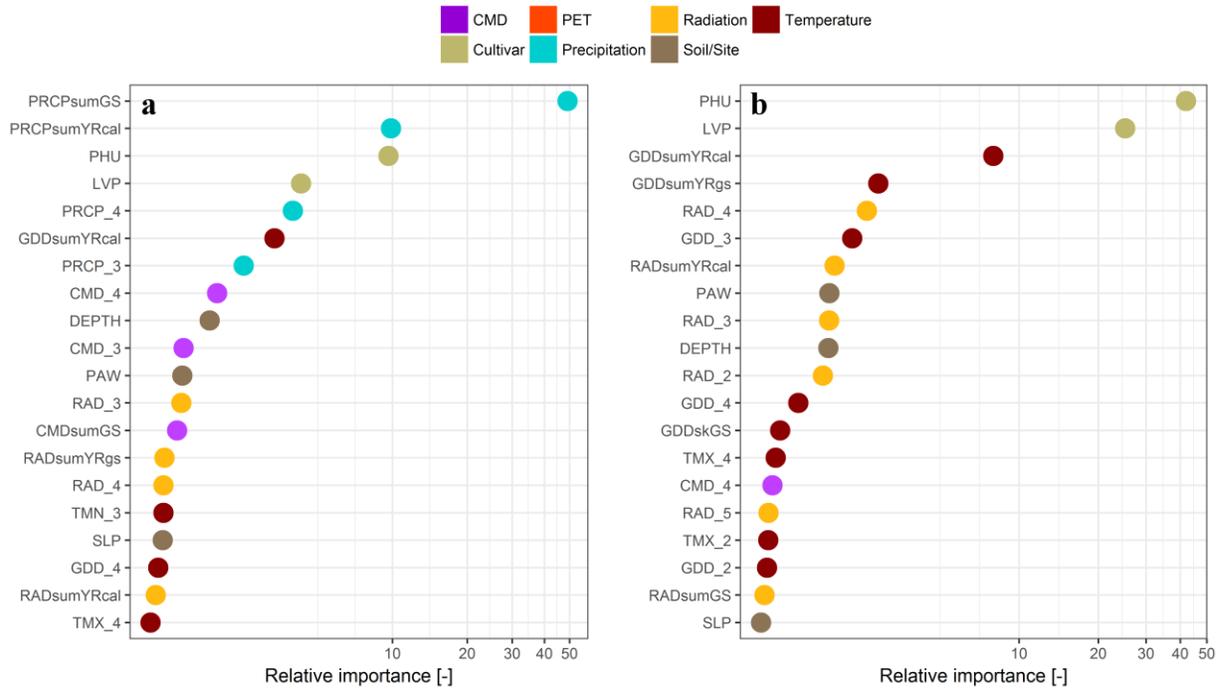
537 **Figure 6.** Inter-annual dynamics of mean rainfed yields for each Koeppen-Geiger climate region  
 538 of Mexico (see section 2.5) considering (a-c) all simulation units or (d-f) only simulation  
 539 units intersecting with substantial maize harvest areas (see section 2.3.4).

540 3.3 Feature importance and the role of feature engineering

541 With rainfed water supply only, the sum of precipitation during the growing season  
 542 (PRCPsumGS) is the by far most important predictor (Figure 7a), followed by calendar year  
 543 precipitation PRCPsumYRcal, PHU, and LVP. Temperature, radiation, and soil-related features  
 544 are of moderate to minor importance. Soil variables matter only with respect to water

545 availability, driven by depth and PAW, which is a composite of texture, SOC, and depth. Other  
546 soil variables, which are mostly related to nutrient availability, matter less due to the estimation  
547 of yield potentials. With sufficient irrigation, the temporally static cultivar and management  
548 characteristics PHU and LVP are the most important features, followed by the annual growing  
549 degree day sum GDDsumYRcal and a wider set of transient climate features, which are  
550 expectedly related to temperature and solar radiation (Figure 7b). Precipitation and ET-related  
551 features do not occur among the top ranking features except for CMD\_4. Among the soil  
552 characteristics, again depth and PAW are the most relevant features.

553         Comparing the variable importance of different subsets of features for model training  
554 (Figure S8; see Table 2 for feature subsets) shows that for rainfed water supply, precipitation-  
555 and cultivar-related features are consistently the most important predictors (Figure S8a,c,e).  
556 Beyond, the ranking of features depends on the feature set with PET derivatives exhibiting rather  
557 low importance among climate features. Notably, soil characteristics beyond depth and PAW are  
558 typically lowest ranking if occurring at all. With sufficient irrigation water supply, PHU and  
559 LVP are consistently the most important features (Figure S8b,d,f) followed predominantly by  
560 temperature and radiation indices. As for rainfed yield estimates, depth and PAW occur in all  
561 feature set as moderately higher-ranking covariates. Precipitation- and PET-related features are  
562 only present in the parsimonious models with 23 features in total, except for CMD\_4 in the  
563 model based on monthly features.



564

565 **Figure 7.** Feature importance for the extreme gradient boosting models for (a) rainfed and (b)  
 566 irrigated conditions. Only top 20 features (see Table 1 for details) are shown. The x-axis  
 567 is  $\log(x+1)$  transformed for better readability.

568 3.4 Random forests models compared to extreme gradient boosting

569 Statistical coefficients for the random forests predictions in the global validation dataset  
 570 are highly comparable to those from extreme gradient boosting (Table S4) with a marginal  
 571 tendency towards lower slopes and higher intercept and slope under rainfed conditions.

572 Predictions for Mexico in turn (Table S5) result typically in slightly higher intercepts and MAE  
 573 as well, but higher  $R^2$  especially for the parsimonious feature sets under irrigated conditions.

574 NSE statistics in contrast are almost consistently poorer. For the full set of climate  
 575 covariates under rainfed water management, the numbers of simulation units with  $NSE > 0$  are in

576 all cases lower or virtually equal (Table S8; c.f. Table S6). Most notable difference are apparent  
577 for the models trained on the full climate feature set in tropic regions. This is even more  
578 pronounced for irrigated conditions, where the number of simulation units with  $NSE < 0$  is up to  
579 40% lower than the extreme gradient boosting predictions (Table S9; c.f. Table S7).

580 Accordingly, predictions aggregated to Koeppen-Geiger regions show also a poorer fit,  
581 but differences are here less pronounced and apparent foremost in NSE statistics (Figure S12 and  
582 Figure S13). This is most evident under rainfed conditions in (semi-)arid regions if all simulation  
583 units are considered (Figure S12a-c). Under irrigated conditions, NSE is even negative in (semi-  
584 )arid climates, no matter whether all simulation units are considered or present cropland only,  
585 (Figure S13a,d) and in temperate climate if all simulation units are considered (Figure S13b).

586 Variable importance remains structurally similar among feature subsets and water supply  
587 regimes (Figure S14) compared to extreme gradient boosting (Figure 7; Figure S8) concerning  
588 the overall ranking of features with some predictors moving up or down a few positions. A  
589 striking differences, however, is that random forests rank also variables indicating distributions,  
590 i.e. standard deviation, among the more important features, while extreme gradient boosting  
591 predictions are foremost relying on sums and averages.

### 592 3.5 Reproduction of reported inter-annual yield variability

593 The evaluation of inter-annual yield variability for the top producing municipios in  
594 Jalisco (Figure S15) shows that NSE is positive in the majority of municipios and hence  
595 satisfactory in all crop yield predictions from both EPIC and the extreme gradient boosting  
596 models. Lowest median performance was found for the global simulations (EPIC global),

597 followed by the high-resolution EPIC simulations at the scale of Mexico (EPIC high-res) with a  
598 slight tendency towards higher NSE. Interestingly, the median NSE for extreme gradient  
599 boosting predictions (Predicted high-res) is higher than for the EPIC simulations at the same  
600 resolution. This is mainly due to one municipio with rather poor performance in the simulations,  
601 while the predictions (Predicted high-res) do not achieve very high performance in other  
602 municipios where EPIC simulations result in up to NSE=0.8. The overall best rendition of inter-  
603 annual yield variability is produced by the machine learning predictions using 1k-resolution  
604 monthly climate surfaces (Predictions 1k) and more so if a national soil data product is used  
605 (Predictions 1k CRU x INEGI) with a median NSE of 0.42 as opposed to 0.20 in the high-  
606 resolution EPIC simulations (EPIC high-res). The CRU x HWSD combination in contrast results  
607 in a lower median but higher maximum NSE.

## 608 **4 Discussion and Conclusions**

### 609 4.1 Model performance for downscaling of yield estimates

610 Performance of the meta-models for spatio-temporal downscaling of crop yield estimates  
611 is exceptionally high in terms of linear regression statistics, and mean bias for both machine  
612 learning methods (Table S4; Table S5). While the results are highly comparable among the two  
613 methods, extreme gradient boosting shows moderately better results especially for inter-annual  
614 yield variability (cf. Tables S6-9), which is of ample importance for climate impact studies (e.g.  
615 Müller et al., 2017). In essence, substantial deviations of predictions from EPIC simulations  
616 occur only for very low yields. Even here, this applies foremost to their absolute magnitude  
617 while inter-annual yield variability is typically still very well reproduced although this is not an

618 implicit goal of the machine learning model optimization. In addition, the high skill in  
619 reproducing irrigated yields stands out, as crop yield variability is known to be more strongly  
620 dominated by variability in precipitation than temperature in most regions (e.g. Frieler et al.,  
621 2017).

622 Our results can hardly be compared to existing literature, as the spatio-temporal  
623 downscaling of crop model outputs via meta-models has not yet been addressed to the authors'  
624 knowledge. Within the closely related, recently emerging field of crop model emulators, Blanc  
625 and Sultan (2015) and Blanc (2017) developed polynomial models to predict yields for various  
626 crops under climate change using unique parameterizations for the statistical models at the grid  
627 cell level. Besides weather and soil data, they include CO<sub>2</sub> as an additional dimension. These  
628 structural differences (a) grid-cell level in the references vs scale-free approach here and (b) no  
629 CO<sub>2</sub> dimension in the present study render the comparison of results difficult. The authors of the  
630 cited studies conclude that the statistical models provide reasonable results in the longer term.  
631 However, the visual comparison of inter-annual yield variability for the Corn Belt during the  
632 historic time period in Blanc and Sultan (2015) and the regional predictions presented in this  
633 study suggest that the polynomial models may be suitable at the global scale and for longer term  
634 assessments but not for regional impact studies. A similar statistical approach has been employed  
635 by Oyebamiji et al. (2015) for a single GGCM finding that 62-93% of crop yield variability  
636 produced by the GGCM can be explained by their multiple tier statistical model, which was as  
637 well parameterized at the grid cell level. This indicates that so far no other methodologic  
638 approaches can provide as accurate and flexible crop meta-models as the ones presented herein,

639 which are also virtually scale-free, free from *a priori* assumptions on relevant features, and truly  
640 data-driven.

641 The very high accuracy of the machine learning models also allowed for detection of an  
642 anomaly in the high-resolution EPIC simulations for Mexico, in which the automatic fertilizer  
643 application failed due to extreme combinations of climate and soil (see Figure 3a,b and  
644 associated text). This indicates that the method should also be tested for quality control of crop  
645 model simulations.

#### 646 4.2 Feature engineering and feature importance

647 The evaluation of different feature subsets shows that even very basic features from  
648 annual climate provide robust results when it comes to general regression metrics. This  
649 highlights that these features should contain sufficient information for providing at least long-  
650 term mean crop yield and agricultural externalities surfaces. Monthly climate data are essential,  
651 in contrast, to provide predictions of very high accuracy (Table S4, Table S5) and to capture  
652 inter-annual crop-climate response accurately as reflected in the EPIC model (Figure 6). This can  
653 be expected as crop growth processes are typically non-linear (Bonhomme, 2000) and crops'  
654 sensitivity to temperature and water supply can shift throughout the growing season. That is, for  
655 instance, the case for drought stress susceptibility of maize yield formation, which is largest  
656 during the second half of the growth cycle for maize (e.g. Gaiser et al., 2010) and is reflected in  
657 the EPIC model within the calculation of an actual HI based on water stress (see section 2.1).

658 The feature importance of models for rainfed yield prediction is quite straightforward  
659 with precipitation and other water-related features strongly dominating (Figure 7a). Static

660 variables PHU and LVP follow thereafter, rendering water availability the main driver for inter-  
661 annual yield variability, while especially PHU – a composite of growing season length and long-  
662 term temperatures – may rather serve as a proxy for the overall yield potential and thermal  
663 growth conditions. If monthly climate statistics are considered, the third and fourth months have  
664 the largest influence on rainfed yield predictions. This relates to the aforementioned non-linearity  
665 of crop growth requirements and the crop’s higher sensitivity during the second half of the  
666 growing season.

667         If sufficient water is supplied (Figure 7b), temperature- and solar radiation-related  
668 features come to the fore. In the first case, these are not minimum or maximum temperatures  
669 indices as such, but again growth effective temperature sums (here GDD). This corresponds  
670 directly to the estimation of phenologic development in the EPIC model (see section 2.1), which  
671 is driven by HU accumulation, while very high and very low temperatures cause stresses to the  
672 crop, which is over large areas typically of minor importance compared to water deficits (e.g.  
673 Schauburger et al., 2017). It is striking, however, that among the transient climate features, not  
674 the growing season sum of GDD (GDDsumGS) is the most important feature, but annual GDD  
675 (GDDsumYRcal). An explanation is that growing season features were calculated for the months  
676 of the average length of vegetation period (feature LVP). Hence, GDDsumGS may in some years  
677 exceed or fall below the actual PHU requirement, while GDDsumYRcal is a more robust annual  
678 temperature index.

679         The low importance of soil covariates can be expected due to the simulation of yield  
680 potentials. As shown in an earlier study (Folberth et al., 2016), the EPIC model itself is rather  
681 insensitive to soil data if yield potentials are simulated, even more so with sufficient irrigation.

682 Hence, the only soil covariates of relevance here relate to water availability, i.e. soil depth and  
683 PAW. Nutrient-related soil covariates in turn may even outweigh the importance of climate  
684 features if no or little nutrients are supplied exogenously as nutrient supply can affect crop yields  
685 by more than an order of magnitude (e.g. Folberth et al., 2013). Still, the spatial detail in Figure  
686 4a,b shows that despite the low importance of soil and site covariates, yield patterns are very  
687 well reproduced at the sub-climate grid ( $0.25^\circ \times 0.25^\circ$ ) level. This indicates that the soil and site  
688 signal is sufficiently represented in the crop yield meta-model despite the comparably low  
689 ranking of soil and site features (Figure 7). An increase in the importance of soil and site features  
690 was found for the meta-model to predict crop available water (Supplementary Text S2), where  
691 various hydrologically relevant covariates such as slope and soil hydrologic group rank higher  
692 than for crop yield predictions or GSET (Figure S11). This emphasizes that approaches free from  
693 assumptions on feature importance are required at least when moving away from crop yield  
694 predictions towards agricultural externalities.

#### 695 4.3 Predictions of agricultural externalities

696 Agricultural externalities were assessed supplementary (Supplementary Text S2) to  
697 evaluate the potential of machine learning algorithms to predict these as well, which is an  
698 essential advantage of integrated crop growth models compared to purely statistical methods of  
699 crop yield estimation. The very good results for GSET show that this is in principle feasible. The  
700 slightly lower performance for CAW in turn indicates that there are limits under extreme  
701 conditions: The very high values that are underestimated here (Figure S9c,d) occur in simulation  
702 units with moderate to high precipitation, low slopes, and soils with high infiltration potential  
703 (not shown). Capturing also such combinations may require an extension of the training data set

704 (see section 4.6). Overall, however, the results show that the computational framework used for  
705 yield predictions can flexibly be transferred to other crop model outputs. Limitations can still be  
706 expected for agro-environmental externalities that occur intermittently with daily peaks such as  
707 emissions of certain greenhouse gases.

#### 708 4.4 Differences and advantages of employed machine learning approaches

709 Differences between the applied machine learning algorithms have been touched upon  
710 above and are here summarized and complemented. In this study, random forests were found to  
711 have lower performance in predictions with respect to inter-annual yield variability but showed  
712 overall similar predictive accuracy, while also the importance of features for crop yield  
713 predictions remained comparable (see section 3.4). From a practical point, however, the  
714 computational cost of random forests is far higher than that of extreme gradient boosting. In the  
715 case of the full climate feature set, it was here about nine hours versus one on the same 32 core  
716 cluster (Figure S16). Even if the number of trees was reduced, which may not cause substantial  
717 trade-offs in accuracy (Figure S1), the time requirement can be assumed at least four times  
718 higher. While common gradient boosting methods may show low computational performance  
719 due to sequential tree building, the extreme gradient boosting approach has *markedly* high  
720 efficiency due to parallelization as already evaluated in its original publication (Chen and  
721 Guestrin, 2016).

722 Although the quantification of prediction uncertainty is beyond the scope of this study, it  
723 is worth mentioning that for random forests there are established methods to quantify prediction  
724 intervals and hence uncertainties associated with predictions (e.g. Meinshausen, 2006) for which

725 no readily applicable methods have been developed for gradient boosting. Provided that the  
726 meta-model predictions show very high accuracy but outliers still occur, this may become of  
727 great importance for applications of downscaled yield estimates e.g. in land use change studies as  
728 well as in the quantification of trade-offs and benefits of (potential) meta-model error and  
729 improved coverage of landscape heterogeneity. We can hence conclude that within the scope of  
730 this study, the extreme gradient boosting approach appears most suitable, but still the selection of  
731 the most appropriate method needs to be made on a per case basis of a specific study.

#### 732 4.5 Model performance benchmarked against reported local yields

733 The performance evaluation against reported yields for ten major producing municipios  
734 (Section 3.5) shows that both EPIC and the extreme gradient boosting models perform  
735 satisfactorily for major producing regions. Thereby, the use of high-resolution monthly climate  
736 surfaces substantially improves the quality of yield predictions. Further targeted evaluations  
737 beyond the scope of this paper will be required to assess under which circumstance the crop  
738 model itself or the meta-model may perform better or poorer and what the impact of  
739 uncertainties and spatial resolutions in climate, soil, management, and land use data as well as  
740 crop model parameterization or meta-model error is as has been done before for single crop  
741 models (Folberth et al., 2012a) and crop model ensembles (e.g. Angulo et al., 2014).

#### 742 4.6 Outlook

743 The meta-models presented herein can readily provide robust estimates within the  
744 domain of the training data, providing a solid proof of concept that machine learning bears great  
745 potential for building readily applicable crop meta-models for spatio-temporal downscaling

746 applications. It is likely, however, that regional and specific local conditions are not represented  
747 within the global feature ranges and their combinations. In addition, crop cultivars are often  
748 adapted to regional conditions, e.g. in terms of temperature requirements and maturity classes.  
749 Here, we found that specific, extremely rare climate-soil combinations led to a systematic  
750 underestimation of the growing season soil water balance CAW. An option to train a meta-model  
751 for such conditions in a systematic way is to simulate artificial combinations of atmospheric,  
752 soil, cultivar, and management conditions that go beyond the combinations inherently occurring  
753 in the global database. This allows for covering an enhanced space of potentially prevailing plant  
754 growth conditions at finer resolutions. A similar approach has recently been undertaken within  
755 the GGCMi initiative (Elliott et al., 2015), altering atmospheric and management conditions in  
756 each simulation unit (resp.  $0.5^\circ \times 0.5^\circ$  grid cell) along the dimensions  $\text{CO}_2$ , temperature,  
757 precipitation, and N fertilizer (CTWN; Ruane et al., 2017) to develop crop model emulators for  
758 climate change impact studies among others. This can hence serve as a blueprint for extending as  
759 well the training data extent as well as its dimensionality for a wider range of applications and  
760 environments.

761

## 762 **Acknowledgements**

763 C.F., J.B., R.S., and M.O. were supported by European Research Council Synergy grant  
764 ERC-2013-SynG-610028 Imbalance-P. Support from the Basic Research Program of the  
765 National Research University Higher School of Economics for A.B. is gratefully acknowledged.  
766 All data and scripts required for training machine learning algorithms and evaluating against  
767 EPIC benchmark simulations are accessible online at

768 [http://webarchive.iiasa.ac.at/~folberth/downscaling\\_paper/](http://webarchive.iiasa.ac.at/~folberth/downscaling_paper/). The authors declare no conflicts of  
769 interest.

## 770 **References**

771 Ali, I., Greifeneder, F., Stamenkovic, J., Neumann, M., & Notarnicola, C. (2015). Review of  
772 Machine Learning Approaches for Biomass and Soil Moisture Retrievals from Remote  
773 Sensing Data. *Remote Sensing*, 7(12), 16398–16421. <https://doi.org/10.3390/rs71215841>

774 Balkovič, J., van der Velde, M., Skalský, R., Xiong, W., Folberth, C., Khabarov, N., ...  
775 Obersteiner, M. (2014). Global wheat production potentials and management flexibility  
776 under the representative concentration pathways. *Global and Planetary Change*, 122,  
777 107–121. <https://doi.org/10.1016/j.gloplacha.2014.08.010>

778 Blanc, E., & Sultan, B. (2015). Emulating maize yields from global gridded crop models using  
779 statistical estimates. *Agricultural and Forest Meteorology*, 214–215, 134–147.  
780 <https://doi.org/10.1016/j.agrformet.2015.08.256>

781 Blanc, É. (2017). Statistical emulators of maize, rice, soybean and wheat yields from global  
782 gridded crop models. *Agricultural and Forest Meteorology*, 236, 145–161.  
783 <https://doi.org/10.1016/j.agrformet.2016.12.022>

784 Bonhomme, R. (2000). Bases and limits to using ‘degree.day’ units. *European Journal of*  
785 *Agronomy*, 13(1), 1–10. [https://doi.org/10.1016/S1161-0301\(00\)00058-7](https://doi.org/10.1016/S1161-0301(00)00058-7)

786 Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.  
787 <https://doi.org/10.1023/A:1010933404324>

788 Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of*  
789 *the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data*  
790 *Mining* (pp. 785–794). New York, NY, USA: ACM.  
791 <https://doi.org/10.1145/2939672.2939785>

792 Delerce, S., Dorado, H., Grillon, A., Rebolledo, M. C., Prager, S. D., Patiño, V. H., ... Jiménez,  
793 D. (2016). Assessing Weather-Yield Relationships in Rice at Local Scale Using Data  
794 Mining Approaches. *PLOS ONE*, *11*(8), e0161620.  
795 <https://doi.org/10.1371/journal.pone.0161620>

796 Duro, D. C., Franklin, S. E., & Dubé, M. G. (2012). A comparison of pixel-based and object-  
797 based image analysis with selected machine learning algorithms for the classification of  
798 agricultural landscapes using SPOT-5 HRG imagery. *Remote Sensing of Environment*,  
799 *118*, 259–272. <https://doi.org/10.1016/j.rse.2011.11.020>

800 Elliott, J., Müller, C., Deryng, D., Chryssanthacopoulos, J., Boote, K. J., Büchner, M., ...  
801 Sheffield, J. (2015). The Global Gridded Crop Model Intercomparison: data and  
802 modeling protocols for Phase 1 (v1.0). *Geosci. Model Dev.*, *8*(2), 261–277.  
803 <https://doi.org/10.5194/gmd-8-261-2015>

804 EROS Data Center (2000), Global Land Cover Characteristics Database v2.0.  
805 [https://lta.cr.usgs.gov/glcc/globdoc2\\_0](https://lta.cr.usgs.gov/glcc/globdoc2_0), USGS Long Term Archive.

806 Ewert, F., van Ittersum, M. K., Heckelei, T., Therond, O., Bezlepkina, I., & Andersen, E. (2011).  
807 Scale changes and model linking methods for integrated assessment of agri-

808 environmental systems. *Agriculture, Ecosystems & Environment*, 142(1), 6–17.  
809 <https://doi.org/10.1016/j.agee.2011.05.016>

810 FAO/IIASA/ISRIC/ISS-CAS/JRC (2012). Harmonized World Soil Database v1.21.  
811 <http://webarchive.iiasa.ac.at/Research/LUC/External-World-soil-database/HTML/>,  
812 International Institute for Applied Systems Analysis, Laxenburg, Austria.

813 Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we Need Hundreds  
814 of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning*  
815 *Research*, 15, 3133–3181.

816 Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces  
817 for global land areas. *International Journal of Climatology*, 37(12), 4302–4315.  
818 <https://doi.org/10.1002/joc.5086>

819 Fischer, G., F. Nachtergaele, O. Prieler, S. Teixeira, E. Tóth, G. van Velthuizen, H. Verelst, L.  
820 Wiberg, D. (2012). *Global Agro-Ecological Zones (GAEZ v3.0): Model Documentation*.  
821 International Institute for Applied Systems Analysis, Laxenburg, Austria and FAO,  
822 Rome, Italy. Retrieved from  
823 [http://www.iiasa.ac.at/Research/LUC/GAEZv3.0/docs/GAEZ\\_Model\\_Documentation.pdf](http://www.iiasa.ac.at/Research/LUC/GAEZv3.0/docs/GAEZ_Model_Documentation.pdf)  
824 [f](#)

825 Folberth, C., Yang, H., Wang, X., & Abbaspour, K. C. (2012a). Impact of input data resolution  
826 and extent of harvested areas on crop yield estimates in large-scale agricultural modeling  
827 for maize in the USA. *Ecological Modelling*, 235–236, 8–18.  
828 <https://doi.org/10.1016/j.ecolmodel.2012.03.035>

829 Folberth, C., Gaiser, T., Abbaspour, K. C., Schulin, R., & Yang, H. (2012b). Regionalization of a  
830 large-scale crop growth model for sub-Saharan Africa: Model setup, evaluation, and  
831 estimation of maize yields. *Agriculture, Ecosystems & Environment*, *151*, 21–33.  
832 <https://doi.org/10.1016/j.agee.2012.01.026>

833 Folberth, C., Yang, H., Gaiser, T., Abbaspour, K. C., & Schulin, R. (2013). Modeling maize  
834 yield responses to improvement in nutrient, water and cultivar inputs in sub-Saharan  
835 Africa. *Agricultural Systems*, *119*, 22–34. <https://doi.org/10.1016/j.agsy.2013.04.002>

836 Folberth, C., Skalský, R., Moltchanova, E., Balkovič, J., Azevedo, L. B., Obersteiner, M., &  
837 Velde, M. van der. (2016). Uncertainty in soil data can outweigh climate impact signals  
838 in global crop yield simulations. *Nature Communications*, *7*, 11872.  
839 <https://doi.org/10.1038/ncomms11872>

840 Frieler, K., Schauburger, B., Arneth, A., Balkovič, J., Chryssanthacopoulos, J., Deryng, D., ...  
841 Levermann, A. (2017). Understanding the weather signal in national crop-yield  
842 variability. *Earth's Future*, *5*(6), 605–616. <https://doi.org/10.1002/2016EF000525>

843 Gaiser, T., de Barros, I., Sereke, F., & Lange, F.-M. (2010). Validation and reliability of the  
844 EPIC model to simulate maize production in small-holder farming systems in tropical  
845 sub-humid West Africa and semi-arid Brazil. *Agriculture, Ecosystems & Environment*,  
846 *135*(4), 318–327. <https://doi.org/10.1016/j.agee.2009.10.014>

847 Gassman, P.W., Williams, J.R., Benson, V.W., Izaurrealde, R.C., Hauck, L.M., Jones, C.A.,  
848 Atwood, J.D., Kiniry, J.R., & Flowers, J.D. (2004). *Historical Development and*

849           *Applications of the EPIC and APEX models*. ASAE/CSAE Meeting Paper No. 042097.  
850           Retrieved from <https://www.card.iastate.edu/products/publications/pdf/05wp397.pdf>

851   Gerik, T., Williams, J., Francis, L., Greiner, J., Magre, M., Meinardus, A., Steglich, E., & Taylor,  
852           R. (2015). *Environmental Policy Integrated Climate Model - User's Manual Version*  
853           0810. Blackland Research and Extension Center, Texas A&M AgriLife, Temple, USA.  
854           Retrieved from [http://agrillife.org/epicapex/files/2015/10/EPIC.0810-User-Manual-Sept-](http://agrillife.org/epicapex/files/2015/10/EPIC.0810-User-Manual-Sept-15.pdf)  
855           [15.pdf](http://agrillife.org/epicapex/files/2015/10/EPIC.0810-User-Manual-Sept-15.pdf)

856   Global Land Cover 2000 database (2003). European Commission, Joint Research Centre, Ispra,  
857           Italy. Retrieved from <https://forobs.jrc.ec.europa.eu/products/glc2000/glc2000.php>

858   Havlík, P., Schneider, U. A., Schmid, E., Böttcher, H., Fritz, S., Skalský, R., ... Obersteiner, M.  
859           (2011). Global land-use implications of first and second generation biofuel targets.  
860           *Energy Policy*, 39(10), 5690–5702. <https://doi.org/10.1016/j.enpol.2010.03.030>

861   Haylock M. R., Hofstra N., Klein Tank A. M. G., Klok E. J., Jones P. D., & New M. (2008). A  
862           European daily high-resolution gridded data set of surface temperature and precipitation  
863           for 1950–2006. *Journal of Geophysical Research: Atmospheres*, 113(D20).  
864           <https://doi.org/10.1029/2008JD010201>

865   Hengl, T., Jesus, J. M. de, Heuvelink, G. B. M., Gonzalez, M. R., Kilibarda, M., Blagotić, A., ...  
866           Kempen, B. (2017a). SoilGrids250m: Global gridded soil information based on machine  
867           learning. *PLOS ONE*, 12(2), e0169748. <https://doi.org/10.1371/journal.pone.0169748>

868   Hengl, T., Leenaars, J. G. B., Shepherd, K. D., Walsh, M. G., Heuvelink, G. B. M., Mamo, T., ...  
869           Kwabena, N. A. (2017b). Soil nutrient maps of Sub-Saharan Africa: assessment of soil

870 nutrient content at 250 m spatial resolution using machine learning. *Nutrient Cycling in*  
871 *Agroecosystems*, 109(1), 77–102. <https://doi.org/10.1007/s10705-017-9870-x>

872 INEGI, 2004. Información Nacional sobre Perfiles de Suelo v1.2. Instituto Nacional de  
873 Estadística, Geografía e Informática.  
874 <http://www.inegi.org.mx/geo/contenidos/reclat/edafologia/>

875 Izaurrealde, R. C., Williams, J. R., McGill, W. B., Rosenberg, N. J., & Jakas, M. C. Q. (2006).  
876 Simulating soil C dynamics with EPIC: Model description and testing against long-term  
877 data. *Ecological Modelling*, 192(3), 362–384.  
878 <https://doi.org/10.1016/j.ecolmodel.2005.07.010>

879 Izaurrealde, R.C., McGill, W.B., & Williams, J.R. (2012). *Development and application of the*  
880 *EPIC model for carbon cycle, greenhouse gas mitigation, and biofuel studies*. In: Liebig,  
881 M.A., Franzluebbers, A.J., & Follet, R.F. (eds.). *Managing Agricultural Greenhouse*  
882 *Gases*, San Diego, USA: Academic Press.

883 Jarvis, A., Reuter, H.I., Nelson, A., & Guevara, E. (2008). Hole-filled SRTM for the globe  
884 Version 4, available from the CGIAR-CSI SRTM 90m Database. Retrieved from  
885 <http://srtm.csi.cgiar.org>

886 Kiniry, J. R., Williams, J. R., Major, D. J., Izaurrealde, R. C., Gassman, P. W., Morrison, M., ...  
887 Zentner, R. P. (1995). EPIC model parameters for cereal, oilseed, and forage crops in the  
888 northern Great Plains region. *Canadian Journal of Plant Science*, 75(3), 679–688.  
889 <https://doi.org/10.4141/cjps95-114>

890 Liu, J., Folberth, C., Yang, H., Röckström, J., Abbaspour, K., & Zehnder, A. J. B. (2013). A  
891 Global and Spatially Explicit Assessment of Climate Change Impacts on Crop Production  
892 and Consumptive Water Use. *PLOS ONE*, 8(2), e57750.  
893 <https://doi.org/10.1371/journal.pone.0057750>

894 Lobell, D. B., Cassman, K. G., & Field, C. B. (2009). Crop Yield Gaps: Their Importance,  
895 Magnitudes, and Causes. *Annual Review of Environment and Resources*, 34(1), 179–204.  
896 <https://doi.org/10.1146/annurev.envIRON.041008.093740>

897 Meinshausen, N. (2006). Quantile Regression Forests. *J. Mach. Learn. Res.*, 7, 983–999.

898 Mueller, N. D., Gerber, J. S., Johnston, M., Ray, D. K., Ramankutty, N., & Foley, J. A. (2012).  
899 Closing yield gaps through nutrient and water management. *Nature*, 490(7419), 254–257.  
900 <https://doi.org/10.1038/nature11420>

901 Müller, C., & Robertson, R. D. (2014). Projecting future crop productivity for global economic  
902 modeling. *Agricultural Economics*, 45(1), 37–50. <https://doi.org/10.1111/agec.12088>

903 Müller, C., Elliott, J., Chryssanthacopoulos, J., Arneth, A., Balkovic, J., Ciais, P., ... Yang, H.  
904 (2017). Global gridded crop model evaluation: benchmarking, skills, deficiencies and  
905 implications. *Geosci. Model Dev.*, 10(4), 1403–1422. [https://doi.org/10.5194/gmd-10-](https://doi.org/10.5194/gmd-10-1403-2017)  
906 [1403-2017](https://doi.org/10.5194/gmd-10-1403-2017)

907 Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I —  
908 A discussion of principles. *Journal of Hydrology*, 10(3), 282–290.  
909 [https://doi.org/10.1016/0022-1694\(70\)90255-6](https://doi.org/10.1016/0022-1694(70)90255-6)

- 910 Oyebamiji, O. K., Edwards, N. R., Holden, P. B., Garthwaite, P. H., Schaphoff, S., & Gerten, D.  
911 (2015). Emulating global climate change impacts on crop yields. *Statistical Modelling*,  
912 *15*(6), 499–525. <https://doi.org/10.1177/1471082X14568248>
- 913 Peel, M. C., Finlayson, B. L., & McMahon, T. A. (2007). Updated world map of the Köppen-  
914 Geiger climate classification. *Hydrol. Earth Syst. Sci.*, *11*(5), 1633–1644.  
915 <https://doi.org/10.5194/hess-11-1633-2007>
- 916 Portmann, F.T., Siebert, S., & Döll, P. (2010). MIRCA2000 - global monthly irrigated and  
917 rainfed crop areas around the year 2000: a new high-resolution dataset for agricultural  
918 and hydrological modeling. *Global Biogeochem. Cycles* *24*, GB1011,  
919 <https://doi.org/10.1029/2008GB003435>
- 920 Pugh, T. A. M., Müller, C., Elliott, J., Deryng, D., Folberth, C., Olin, S., ... Arneth, A. (2016).  
921 Climate analogues suggest limited potential for intensification of production on current  
922 croplands under climate change. *Nature Communications*, *7*, 12608.  
923 <https://doi.org/10.1038/ncomms12608>
- 924 R Development Core Team (2008). *R: A language and environment for statistical computing*, R  
925 *Foundation for Statistical Computing*. Vienna, Austria, Retrieved from [https://www.R-](https://www.R-project.org)  
926 [project.org](https://www.R-project.org).
- 927 Reidsma, P., Ewert, F., Boogaard, H., & Diepen, K. van. (2009). Regional crop modelling in  
928 Europe: The impact of climatic conditions and farm characteristics on maize yields.  
929 *Agricultural Systems*, *100*(1), 51–60. <https://doi.org/10.1016/j.agsy.2008.12.009>

930 Rienecker, M. M., Suarez, M. J., Gelaro, R., Todling, R., Bacmeister, J., Liu, E., ... Woollen, J.  
931 (2011). MERRA: NASA's Modern-Era Retrospective Analysis for Research and  
932 Applications. *Journal of Climate*, 24(14), 3624–3648. [https://doi.org/10.1175/JCLI-D-11-](https://doi.org/10.1175/JCLI-D-11-00015.1)  
933 [00015.1](https://doi.org/10.1175/JCLI-D-11-00015.1)

934 Rosenzweig, C., Elliott, J., Deryng, D., Ruane, A. C., Müller, C., Arneth, A., ... Jones, J. W.  
935 (2014). Assessing agricultural risks of climate change in the 21st century in a global  
936 gridded crop model intercomparison. *Proceedings of the National Academy of Sciences*,  
937 111(9), 3268–3273. <https://doi.org/10.1073/pnas.1222463110>

938 Rosenzweig, C., Ruane, A. C., Antle, J., Elliott, J., Ashfaq, M., Chatta, A. A., ... Wiebe, K.  
939 (2018). Coordinating AgMIP data and models across global and regional scales for 1.5°C  
940 and 2.0°C assessments. *Phil. Trans. R. Soc. A*, 376(2119), 20160455.  
941 <https://doi.org/10.1098/rsta.2016.0455>

942 Ruane, A. C., Goldberg, R., & Chryssanthacopoulos, J. (2015). Climate forcing datasets for  
943 agricultural modeling: Merged products for gap-filling and historical climate series  
944 estimation. *Agricultural and Forest Meteorology*, 200, 233–248.  
945 <https://doi.org/10.1016/j.agrformet.2014.09.016>

946 Ruane, A. C., Rosenzweig, C., Asseng, S., Boote, K. J., Elliott, J., Ewert, F., ... Thorburn, P. J.  
947 (2017). An AgMIP framework for improved agricultural representation in integrated  
948 assessment models. *Environmental Research Letters*, 12(12), 125003.  
949 <https://doi.org/10.1088/1748-9326/aa8da6>

950 Sacks, W. J., Deryng D., Foley J. A., & Ramankutty N. (2010). Crop planting dates: an analysis  
951 of global patterns. *Global Ecology and Biogeography*, 19(5), 607–620.  
952 <https://doi.org/10.1111/j.1466-8238.2010.00551.x>

953 Schauburger, B., Archontoulis, S., Arneth, A., Balkovic, J., Ciais, P., Deryng, D., ... Frieler, K.  
954 (2017). Consistent negative response of US crops to high temperatures in observations  
955 and crop models. *Nature Communications*, 8, 13931.  
956 <https://doi.org/10.1038/ncomms13931>

957 Sharpley, A.N., & Williams, J.R. (1990). *EPIC – Erosion/Productivity Impact Calculator: 1.*  
958 *Model Documentation* (US Department of Agriculture Technical Bulletin 1768).  
959 Retrieved from <http://agrilife.org/epicapex/files/2015/05/EpicModelDocumentation.pdf>

960 SIAP (2018a). Avance de Siembras y Cosechas - Resumen nacional por estado. Servicio de  
961 Información Agroalimentaria y Pesquera (SIAP). Retrieved from  
962 <https://www.gob.mx/siap>

963 SIAP (2018b). Estadística de Producción Agrícola. Servicio de Información Agroalimentaria y  
964 Pesquera (SIAP). Retrieved from <http://infosiap.siap.gob.mx/gobmx/datosAbiertos.php>

965 Skalský, R., Tarasovičová, Z., Balkovič, J., Schmid, E., Fuchs, M., Moltchanova, E.,  
966 Kindermann, G., & Scholtz, P. (2008). Geo-bene global database for biophysical  
967 modelling v. 1.0. Concepts, methodologies and data: Retrieved from [http://www.geo-](http://www.geo-bene.eu/files/Deliverables/Geo-BeneGlbDb10%28DataDescription%29.pdf)  
968 [bene.eu/files/Deliverables/Geo-BeneGlbDb10%28DataDescription%29.pdf](http://www.geo-bene.eu/files/Deliverables/Geo-BeneGlbDb10%28DataDescription%29.pdf)

969 Stockle, C. O., Williams, J. R., Rosenberg, N. J., & Jones, C. A. (1992). A method for estimating  
970 the direct and climatic effects of rising atmospheric carbon dioxide on growth and yield

971 of crops: Part I—Modification of the EPIC model for climate change analysis.  
972 *Agricultural Systems*, 38(3), 225–238. [https://doi.org/10.1016/0308-521X\(92\)90067-X](https://doi.org/10.1016/0308-521X(92)90067-X)

973 The H2O.ai team (2017). h2o: R Interface for H2O. R package version 3.14.0.3. Retrieved from  
974 <https://CRAN.R-project.org/package=h2o>

975 Toloşi, L., & Lengauer, T. (2011). Classification with correlated features: unreliability of feature  
976 ranking and solutions. *Bioinformatics*, 27(14), 1986–1994.  
977 <https://doi.org/10.1093/bioinformatics/btr300>

978 Wang, T., Hamann, A., Spittlehouse, D., & Carroll, C. (2016). Locally Downscaled and Spatially  
979 Customizable Climate Data for Historical and Future Periods for North America. *PLOS*  
980 *ONE*, 11(6), e0156720. <https://doi.org/10.1371/journal.pone.0156720>

981 Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York, USA: Springer.

982 Williams, J.R. (1995). *The EPIC Model*. In: Singh, V. P. (ed.). *Computer Models of Watershed*  
983 *Hydrology*, Water Resources Publications.

984 Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine*  
985 *learning tools and techniques*. Burlington, USA: Morgan Kaufmann.

986 You, L., Wood-Sichra, U., Fritz, S., Guo, Z., See, L., & Koo, J. (2017). Spatial Production  
987 Allocation Model (SPAM) 2005 v3.2. March 6, 2018. Retrieved from  
988 <http://mapspam.info>.

989 Zambrano-Bigiarini, M. (2014). *hydroGOF: Goodness-of-fit functions for comparison of*  
990 *simulated and observed hydrological time series*. R package version 0.3-8. Retrieved  
991 from <https://CRAN.R-project.org/package=hydroGOF>