Spatio-temporal downscaling of gridded crop model yield estimates based on machine learning

C. Folberth[a], A. Baklanov[b,c], J. Balkovič[a,d], R. Skalský[a,e], N. Khabarov[a], M. Obersteiner[a]

[a] International Institute for Applied Systems Analysis, Ecosystem Services and Management Program, Schlossplatz 1, A-2361 Laxenburg, Austria, folberth@iiasa.ac.at, balkovic@iiasa.ac.at, skalsky@iiasa.ac.at, khabarov@iiasa.ac.at, oberstei@iiasa.ac.at

[b] International Institute for Applied Systems Analysis, Advanced Systems Analysis Program, Schlossplatz 1, A-2361 Laxenburg, Austria, baklanov@iiasa.ac.at

[c] National Research University Higher School of Economics, Soyuza Pechatnikov str., 16, St. Petersburg, Russian Federation

[d] Department of Soil Science, Faculty of Natural Sciences, Comenius University in Bratislava, Ilkovičova 6, 842 15 Bratislava, Slovak Republic, balkovic@fns.uniba.sk

[e] National Agricultural and Food Centre, Soil Science and Conservation Research Institute, Trencianska 55, 824 80 Bratislava, Slovak Republic, r.skalsky@vupop.sk

Corresponding author:

Christian Folberth

Schlossplatz 1

A-2361 Laxenburg, Austria

E-Mail address: folberth@iiasa.ac.at

Highlights:

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

Abstract

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

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

1 Introduction

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

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

While this spatial resolution provides sufficient detail for robust assessments at macro scales such as the country level, there is increasing concern that GGCM estimates and hence impact assessments at coarse resolutions often miss actual on-ground conditions. As only average or dominant characteristics present within each grid are considered for simulations, assumptions and data may not match actually farmed land (e.g. Folberth et al., 2016) and farming practices (e.g. Reidsma et al., 2009). In addition, they may omit farm-level heterogeneity present at the sub-grid level (Ewert et al., 2011), which is essential for local to regional decision-making and stakeholder information (Rosenzweig et al., 2018).

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

An approach lending itself to address these issues in an efficient and flexible way is the use of meta-models built from coarser GGCM simulations. This allows for deriving estimates of crop yields and associated agricultural externalities at high, virtually scale-free, spatial resolutions without requirements for setting up high-resolution crop model infrastructures including their comprehensive data requirements. There is no scientific literature on crop meta-model development for spatio-temporal predictions across scales known to the authors. The potentially most closely related field is the recently evolving crop model emulator development at the grid cell level. Examples are the development of regressions along climate change trajectories as such (e.g. Blanc and Sultan, 2015; Blanc, 2017) or the use of global crop model simulations with artificial alterations of climate variables to retrieve estimates of climate change impacts for assessment studies based on regressions along temperature, precipitation, and CO_2 concentrations (Ruane et al., 2017; Rosenzweig et al., 2018). The production of high-resolution crop yield surfaces in contrast is foremost accomplished using simplified crop model algorithms (e.g. IIASA/FAO, 2012) or purely statistical approaches (e.g. Mueller et al., 2012). Common to all referenced approaches is that they (a) are based on narrow sets of *a priori* selected covariates based on modelers' assumptions and (b) do not allow for or have not been tested for the joint evaluation of agricultural productivity and externalities. Crop model emulators are in addition typically parameterized at the grid level, which renders them spatially determined and scale-depended.

The presently most flexible approaches for data-driven development of models with high accuracy can be found in the field of machine learning. Machine learning is a collective term for a wide range of data analysis and data-driven forecasting techniques. The most advanced techniques are characterized by the ability to digest large amounts of covariates (herein *syn.* features, *syn.* predictors) to provide predictions for both numeric and categorical variables with algorithms of high complexity and flexibility, which determine the relevance of provided covariates themselves (e.g. Witten et al., 2016). Examples of methodologic approaches are neural networks, various forms and derivatives of regression trees, as well as clustering techniques. While simpler methods such as multiple linear or lasso regressions are typically computationally faster and straightforward to interpret, they show typically a substantially lower performance. Within agricultural sciences, applications are to date mostly limited to processing and analyzes of remote sensing data (e.g. Duro et al., 2012; Ali et al., 2015). Few exceptions are the development of crop nutrient response models for studying yield responses in sub-Saharan Africa based on field trial data (Hengl et al., 2017b) and the use of data mining tools for identifying crop growth limitations (Delerce et al., 2016).

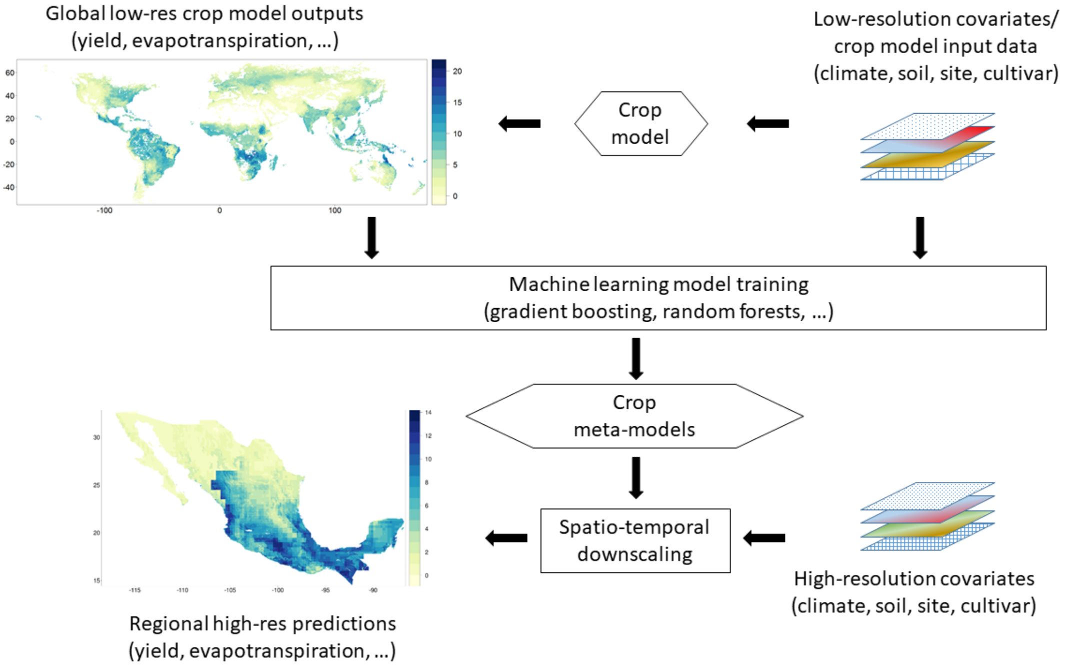
Global low-res crop model outputs
(yield, evapotranspiration, ...)
Crop
model
Low-resolution covariates/
crop model input data
(climate, soil, site, cultivar)
Machine learning model training
(gradient boosting, random forests, ...)
Crop
meta-models
Spatio-temporal
downscaling
High-resolution covariates
(climate, soil, site, cultivar)
Regional high-res predictions
(yield, evapotranspiration, ...)

Figure 1. Schematic representation of the downscaling approach presented in this study. Machine-learning derived meta-models trained on global crop model outputs and covariates at a comparably low spatial resolution are used for producing regional estimates of corresponding variables at a higher spatial resolution.

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

2 Methods and Data

2.1 Gridded crop model description

Crop simulations were carried out using a gridded version of the Environmental Policy Integrated Climate model (EPIC). EPIC was initially developed to assess the impacts of management on crop yields (Williams, 1995). It has constantly been updated to cover additional processes such as effects of elevated atmospheric CO_2 concentration on plant growth (Stockle et al., 1992), detailed soil organic matter cycling (Izaurralde et al., 2006, Izaurralde et al., 2012), and an extended number of crop types and cultivars (e.g. Kiniry et al., 1995; Gaiser et al., 2010) among others (see Gassman, 2004). More details of the crop growth model are provided in Supplementary Text S1.

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

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

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

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

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

2.3 Crop model input data

2.3.1 Climate data

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

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

2.3.2 Soil data

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

2.3.3 Topography

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

2.3.4 Land use

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

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

2.3.5 Crop management

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

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

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

2.4 Machine learning framework

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

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

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

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

2.4.1 Extreme gradient boosting

Similar to other boosting methods, extreme gradient boosting is an ensemble learning technique that sequentially builds the model: each tree is fit on a modified version of the original training data set. I.e., every new tree uses information from previously grown trees. This is the key difference to random forests (see below). Extreme gradient boosting generalizes boosting methods by allowing minimization of an arbitrary differentiable loss function. In this study, we employed the R package XGBoost for extreme gradient boosting, a highly efficient realization of the gradient boosting approach that showed the best performance in recent machine learning challenges (Chen and Guestrin, 2016). Being a learning algorithm with high flexibility, extreme gradient boosting is prone to overfitting, especially, if training data are scarce, which is not the case here. Typically, parameter tuning is done by performing an exhaustive grid search along parameter dimensions using the default parameters as the reference point. This was here not considered meaningful due to the vast amount of training data, rendering a full grid search computationally inefficient and unneeded, due to extremely low error obtained already in a limited grid search. I.e., we tuned only key parameters for shrinkage and learning (eta, max_depth, nrounds; Table S1). In our case, the default parameter values resulted in stable but improvable performance with R^2=0.94 for the test dataset. This suggested to increase the maximum tree depth and local variation of the learning rate (eta). The grid search resulted in R^2=0.99 for both training and test data with eta=0.15 or 0.30 and max_depth=15 or 20. The lowest RMSE in both training and test data was obtained with eta=0.15 and max_depth=20 in a five-fold cross validation (Table S2). Although this parameter set results in a marginal overfit, it also showed the best performance in regression metrics and mean absolute error (MAE; not shown), the main performance indicators used herein (see section 2.5.1). It was hence selected for performing the predictions. Extending the grid search to by increasing the rounds of tree building (nrounds) from 60 to 100 provided only a negligible increase in performance (Table S2). Resulting parameters were hence eta=0.15, max_depth=20, and – to ensure very high accuracy - nrounds=100.

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

2.4.2 Random forests

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

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

2.4.3 Feature importance

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

2.4.4 Machine learning features and feature engineering

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

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

Abbreviation	Variable description
Climate variables (VARs; transient)	
TMX	Maximum temperature [°C]
TMN	Minimum temperature [°C]
GDD	Growing degree days [°C]
RAD	Solar radiation [MJ m^{-2}]
PET	Potential evapotranspiration [mm]
PRCP	Total precipitation [mm]
WET	Wet day frequency [d]
CMD	Climatic moisture deficit (PRCP-PET) [mm]
Temporal aggregates and derivatives of climate variables (transient)	
*VAR*_X	Monthly value for month X {1:12} since planting (e.g. "TMX_1")
*VAR*sd_X	Standard deviation of mean value in month X {1:12} (e.g. "TMXsd_1")
*VAR*avYRcal	Average of climate variable in calendar year (January to December)
*VAR*sumYRcal	Sum of climate variable in calendar year (January to December)
*VAR*avYRgs	Average of climate variable in growing season year (12 months from planting)
*VAR*sumYRgs	Sum of climate variable in growing season year (12 months from planting)
*VAR*skYRgs	Skew of climate variable in growing season year (12 months from planting)
*VAR*avGS	Average of climate variable in growing season (planting month to harvest)
*VAR*sumGS	Sum of climate variable in growing season (planting month to harvest)
*VAR*skGS	Skew of climate variable in growing season (planting month to harvest)
Soil and site variables (static)	
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^{-1}]
CEC	Cation exchange capacity in topsoil [cmol kg^{-1}]
EC	Electric conductivity in topsoil [mmho cm^{-1}]
ROK	Coarse fragment (rock) content in topsoil [%]
BD	Bulk density in topsoil [g cm^{-3}]
CARB	Carbonate content in topsoil [%]
OC	Organic carbon content in topsoil [%]
PAW	Total plant available water capacity [m^3 m^{-3}]
HG	Soil hydrologic group (water infiltration potential) [-]
SLP	Hill slope [%]
Cultivar and growing season variables (static)	
PHU	Potential heat units/growing degree days from planting to maturity [°C]
LVP	Length of vegetation period. Average days from planting to maturity [d]
Target variables (transient)	
YLDG	Maize crop yield [t ha^{-1}]
CAW	Crop available water [mm]
GSET	Growing season evapotranspiration [mm]

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

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

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

To evaluate the importance of raw and engineered climate features, the machine learning models were trained with various feature subsets (Table 2). Soil and site data, PHU, and LVP were considered in all scenarios to evaluate the importance of climate variables only. Annual climate data can be considered the most general feature set. Growing season climate considers the mean or sum of climatic conditions experienced by the crop. Monthly data in turn account for intra-seasonal variability and climate effects in certain growth stages. The complete climate feature set takes all aspects into account and solely lets the algorithm select the most relevant features. Thereby, months beyond the sixth from planting were excluded to keep the number of features at a reasonable extent, considering that maize cultivars hardly require >180 days to reach maturity.

Table 2. Climate feature subsets used in the analyses. Besides indicated climate features (see 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	*VAR*avYRcal, *VAR*sumYRcal	23
growing season climate	*VAR*avGS, *VAR*sumGS	23
monthly climate	*VAR*_X (with X ≤ 6)	63
complete climate	all features except *VAR*_X with x ≥ 7	151

2.5 Performance metrics and model evaluation

2.5.1 Machine learning model performance compared to crop model simulations

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

The coefficient of determination R^2 was calculated according to

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

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

MAE was calculated as

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

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

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

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

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

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

2.5.2 Model performance compared to regional statistics

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

2.6 Computational framework

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

3 Results

3.1 Global scale model performance for crop yields

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

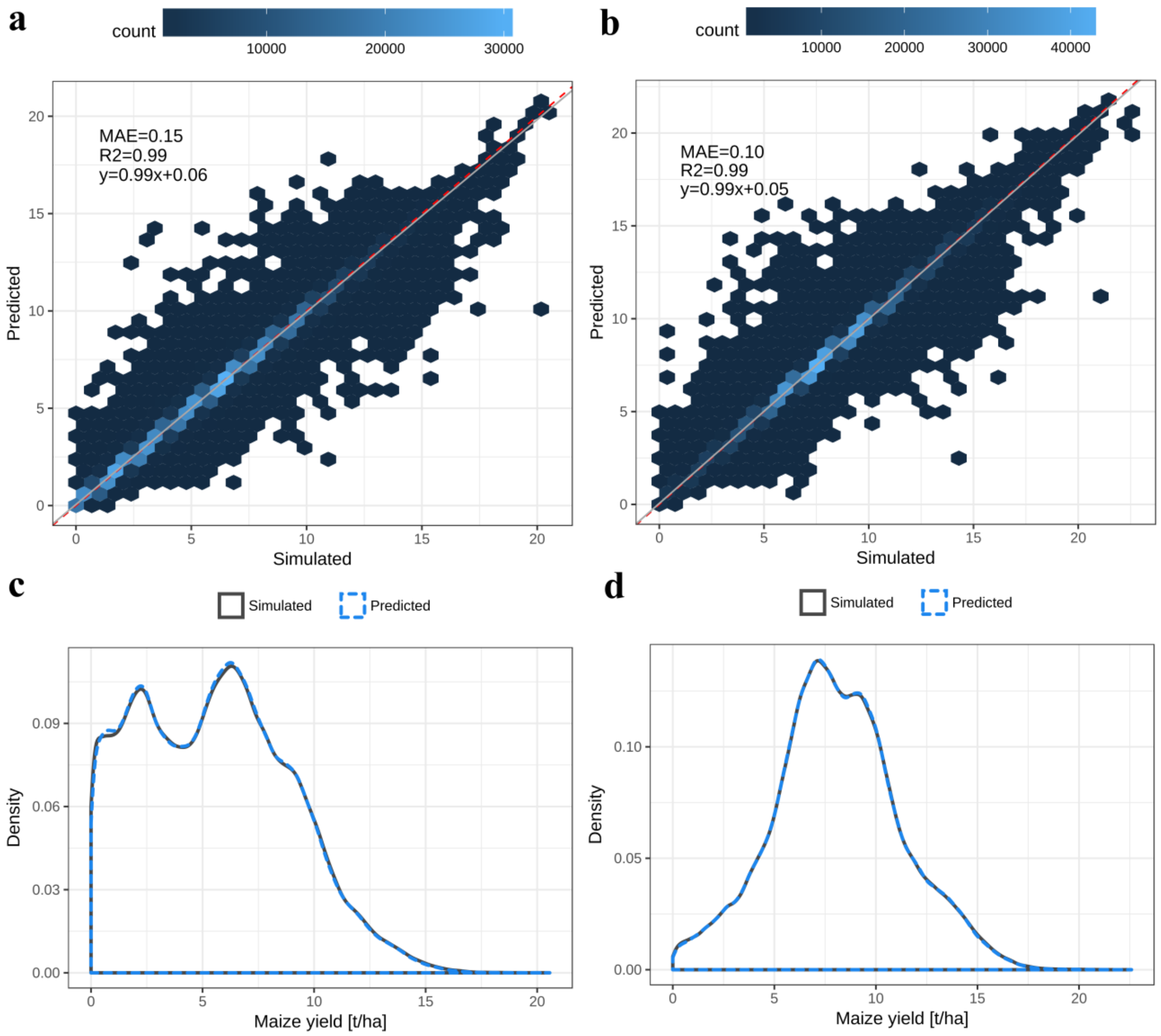

Figure 2. Hexbin and regression plots for EPIC simulated and extreme gradient boosting predicted crop yields in the validation dataset (25% of total samples) for (a) rainfed and (b) irrigated conditions and corresponding density distributions for (c) rainfed and (d) irrigated conditions. Red dashed and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

3.2 Performance of crop yield predictions for Mexico

3.2.1 General performance and patterns

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

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

Using the more parsimonious climate feature sets decreases model performance (Table S5) similar to the global scale validation data (Table S4). The largest decrease occurs for the most set of growing season climate data, while again hardly any difference is found when using the monthly climate features.

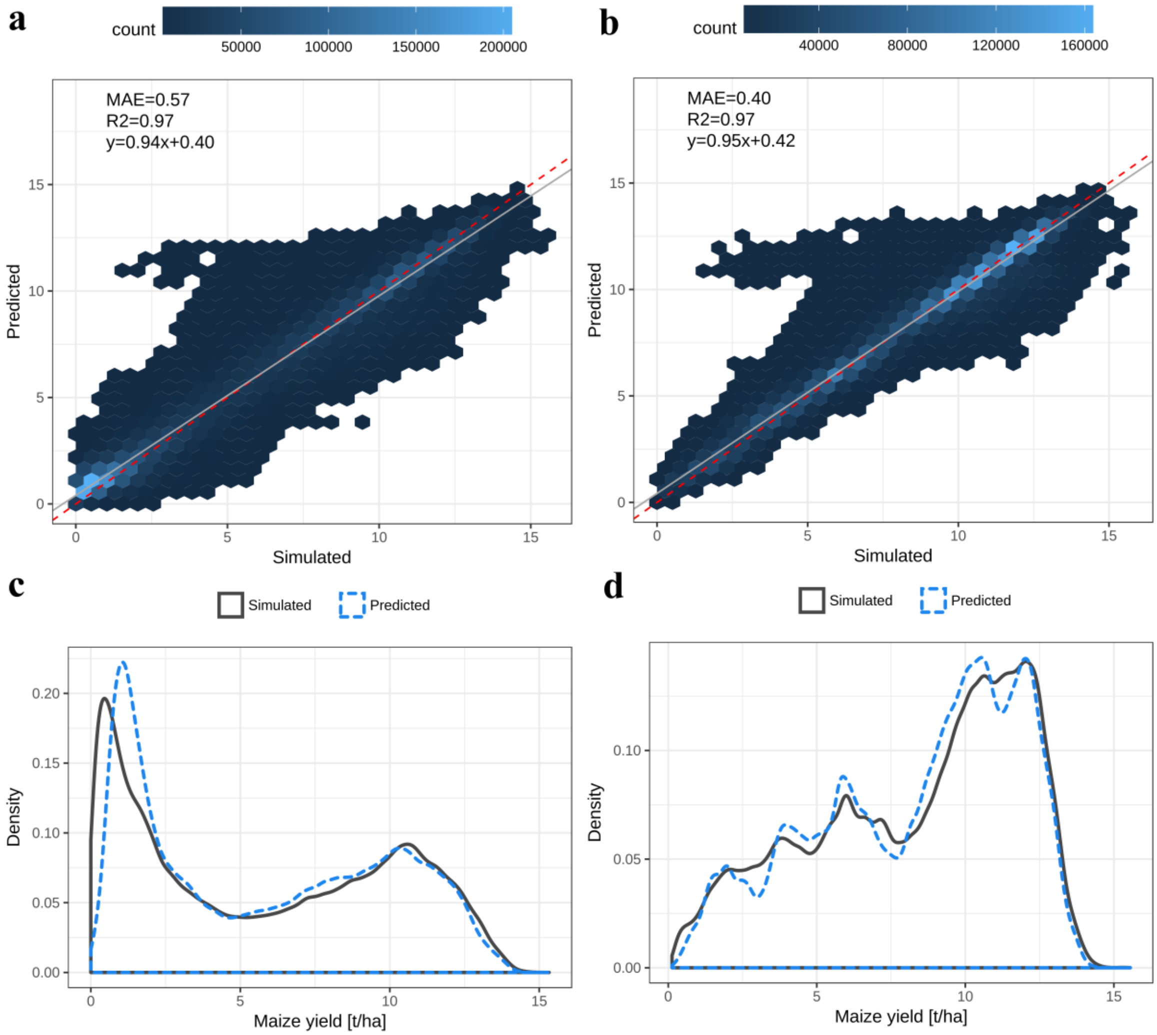

Figure 3. Same as Figure 2 but comparing the high-resolution downscaled predictions and benchmark EPIC simulations for Mexico.

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

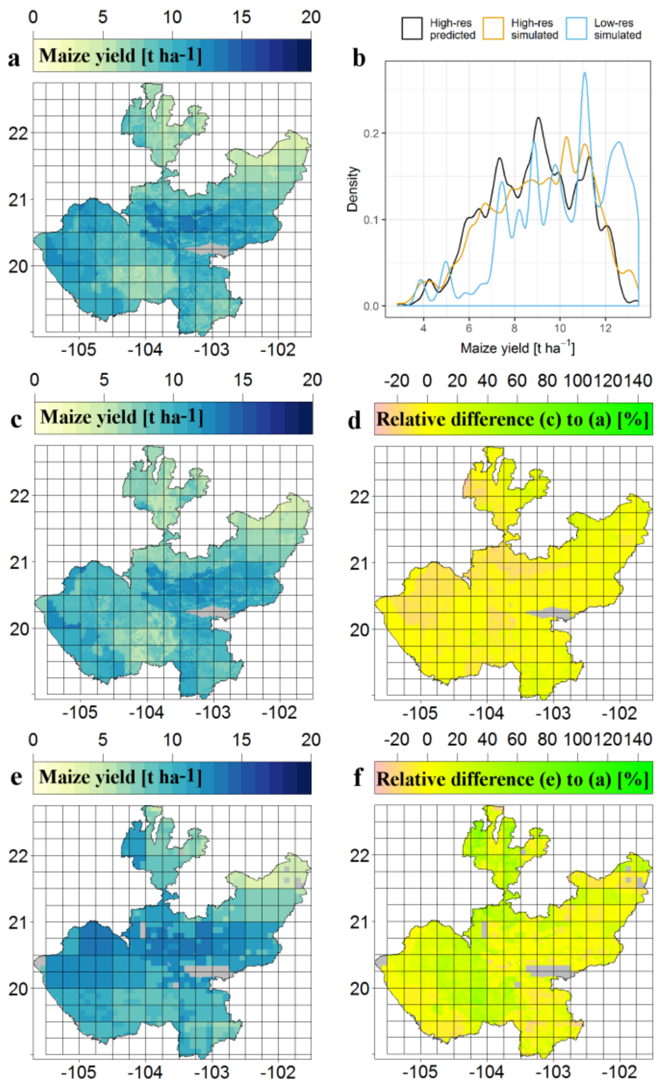

Figure 4. Examples of rainfed maize yields for the year 2000 in the state of Jalisco from (a) high-resolution EPIC simulation, (c) high-resolution machine learning prediction, and (e) global low-resolution simulation. (b) Shows the corresponding density distributions for which yield estimates from the low-resolution simulations have been resampled to the higher resolution to obtain at consistent sample sizes. (d) and (f) show the relative differences of (c) and (e) compared to (a), respectively. Regressions and statistics are presented in Figure S5a,b. The rectangular grid represents the 0.25° x 0.25° climate grid.

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

3.2.2 Reproduction of inter-annual crop yield variability

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

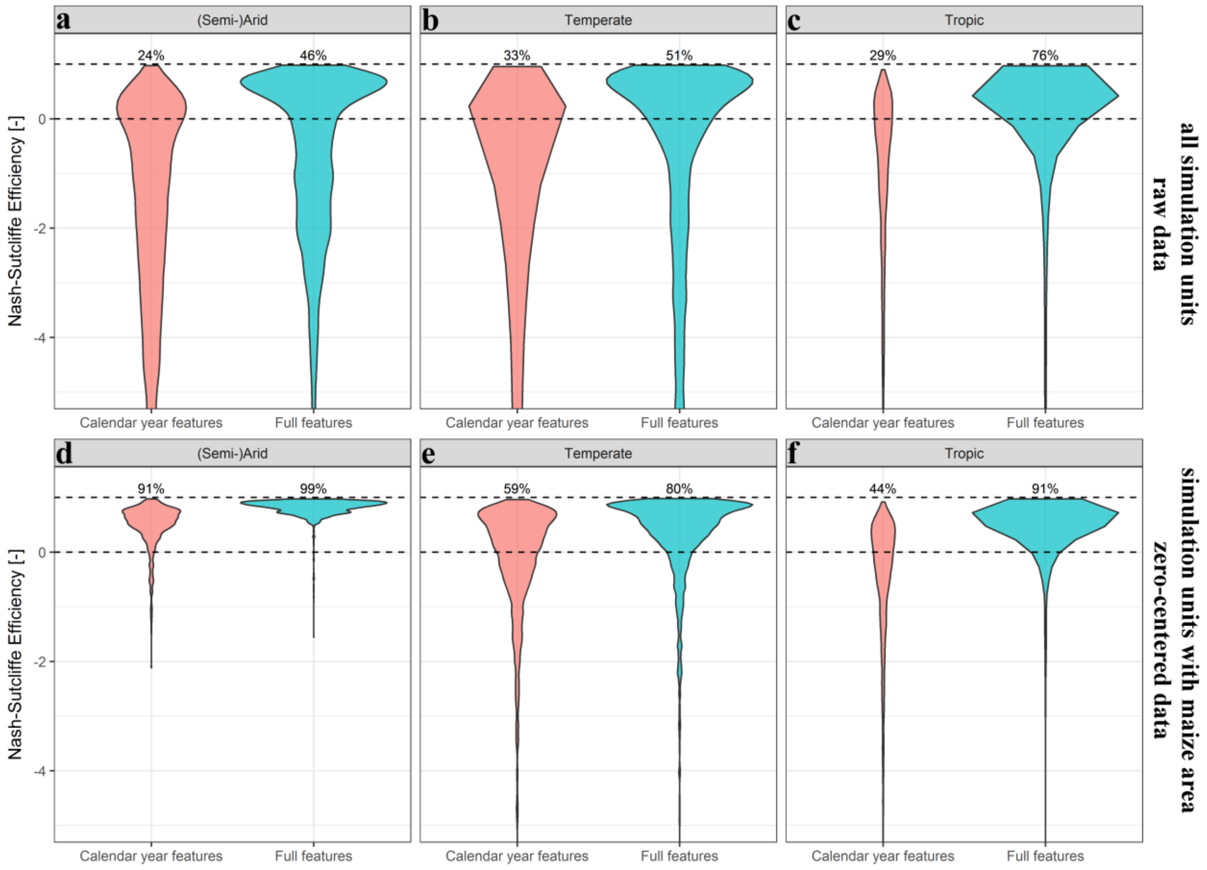

Figure 5. Violin plots of Nash-Sutcliffe Efficiency disaggregated by major Koeppen-Geiger climate regions (see section 2.5) for the feature subsets using calendar year climate variables only or all climate features. (a-c) All simulation units of Mexico with raw data or (d-f) only simulation units with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

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

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

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

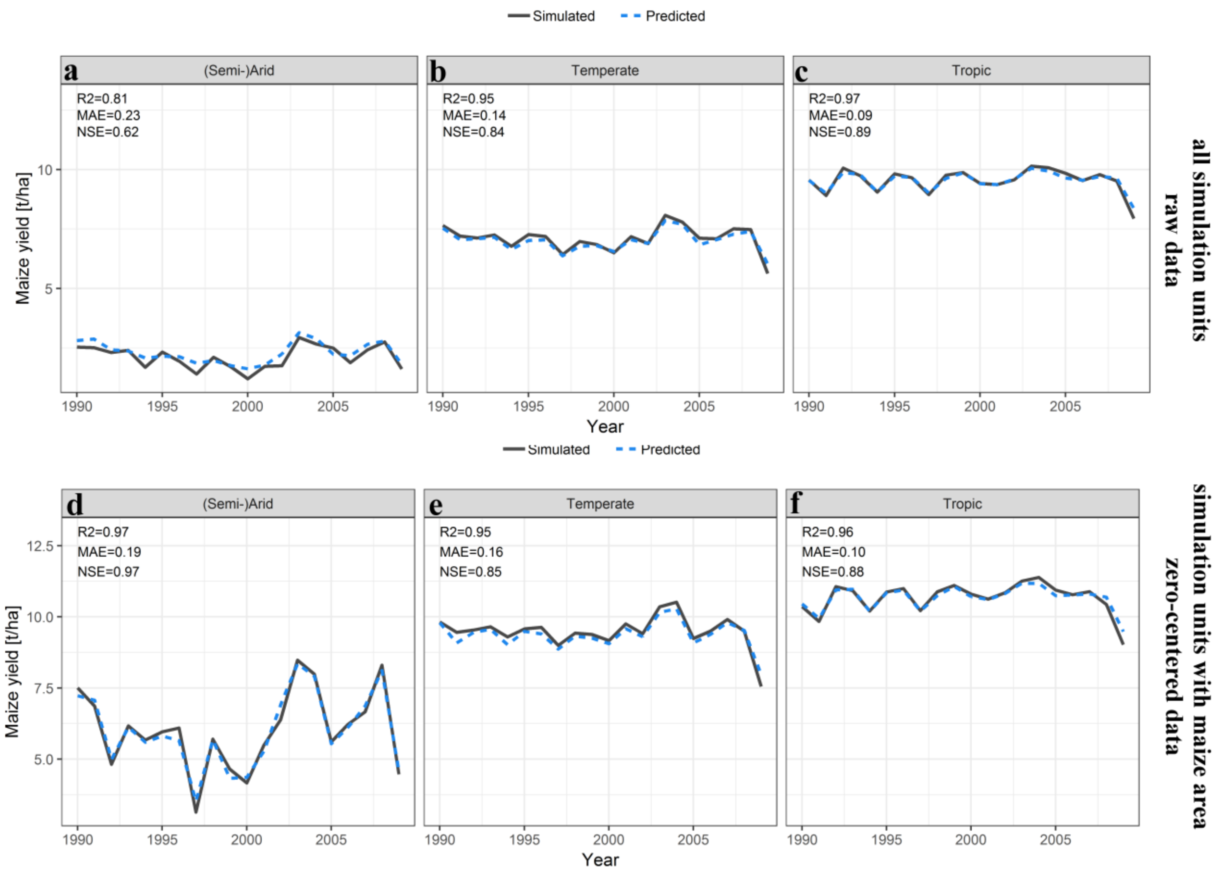

Figure 6. Inter-annual dynamics of mean rainfed yields for each Koeppen-Geiger climate region of Mexico (see section 2.5) considering (a-c) all simulation units or (d-f) only simulation units intersecting with substantial maize harvest areas (see section 2.3.4).

3.3 Feature importance and the role of feature engineering

With rainfed water supply only, the sum of precipitation during the growing season (PRCPsumGS) is the by far most important predictor (Figure 7a), followed by calendar year precipitation PRCPsumYRcal, PHU, and LVP. Temperature, radiation, and soil-related features are of moderate to minor importance. Soil variables matter only with respect to water availability, driven by depth and PAW, which is a composite of texture, SOC, and depth. Other soil variables, which are mostly related to nutrient availability, matter less due to the estimation of yield potentials. With sufficient irrigation, the temporally static cultivar and management characteristics PHU and LVP are the most important features, followed by the annual growing degree day sum GDDsumYRcal and a wider set of transient climate features, which are expectedly related to temperature and solar radiation (Figure 7b). Precipitation and ET-related features do not occur among the top ranking features except for CMD_4. Among the soil characteristics, again depth and PAW are the most relevant features.

Comparing the variable importance of different subsets of features for model training (Figure S8; see Table 2 for feature subsets) shows that for rainfed water supply, precipitation- and cultivar-related features are consistently the most important predictors (Figure S8a,c,e). Beyond, the ranking of features depends on the feature set with PET derivatives exhibiting rather low importance among climate features. Notably, soil characteristics beyond depth and PAW are typically lowest ranking if occurring at all. With sufficient irrigation water supply, PHU and LVP are consistently the most important features (Figure S8b,d,f) followed predominantly by temperature and radiation indices. As for rainfed yield estimates, depth and PAW occur in all feature set as moderately higher-ranking covariates. Precipitation- and PET-related features are only present in the parsimonious models with 23 features in total, except for CMD_4 in the model based on monthly features.

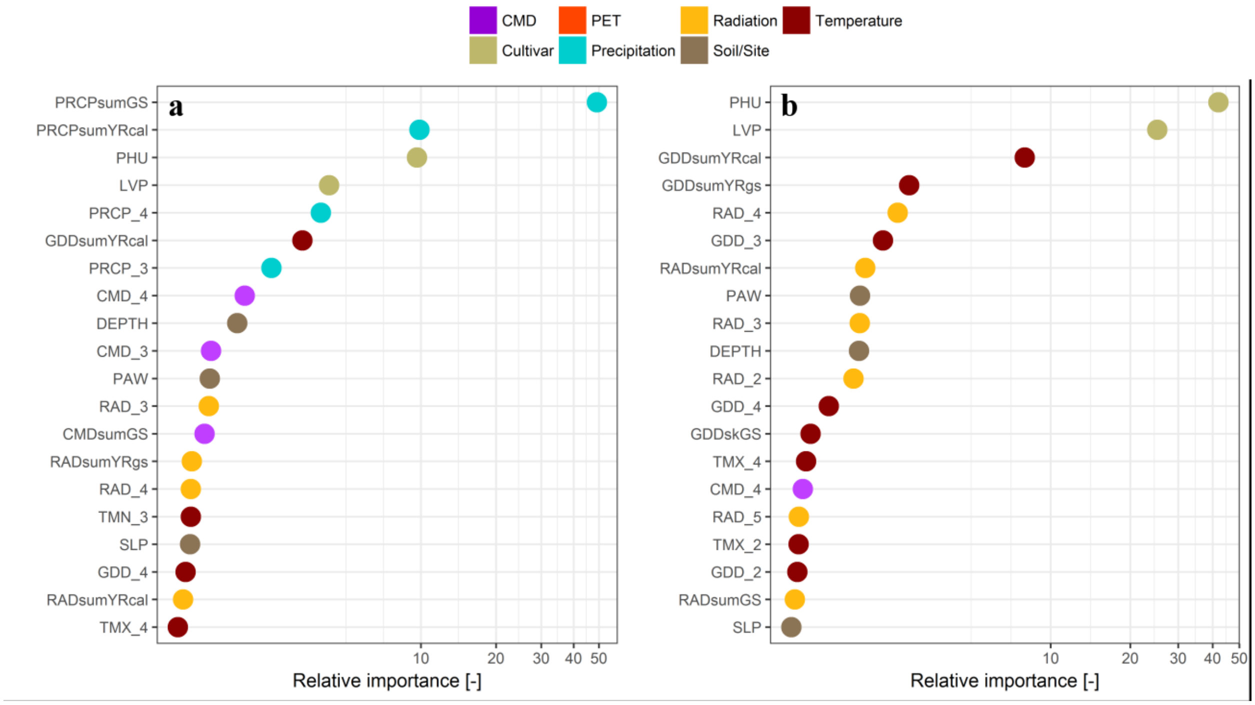

Figure 7. Feature importance for the extreme gradient boosting models for (a) rainfed and (b) irrigated conditions. Only top 20 features (see Table 1 for details) are shown. The x-axis is log(x+1) transformed for better readability.

3.4 Random forests models compared to extreme gradient boosting

Statistical coefficients for the random forests predictions in the global validation dataset are highly comparable to those from extreme gradient boosting (Table S4) with a marginal tendency towards lower slopes and higher intercept and slope under rainfed conditions. Predictions for Mexico in turn (Table S5) result typically in slightly higher intercepts and MAE as well, but higher R^2 especially for the parsimonious feature sets under irrigated conditions.

NSE statistics in contrast are almost consistently poorer. For the full set of climate covariates under rainfed water management, the numbers of simulation units with NSE>0 are in all cases lower or virtually equal (Table S8; c.f. Table S6). Most notable difference are apparent for the models trained on the full climate feature set in tropic regions. This is even more pronounced for irrigated conditions, where the number of simulation units with NSE<0 is up to 40% lower than the extreme gradient boosting predictions (Table S9; c.f. Table S7).

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

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

3.5 Reproduction of reported inter-annual yield variability

The evaluation of inter-annual yield variability for the top producing municipios in Jalisco (Figure S15) shows that NSE is positive in the majority of municipios and hence satisfactory in all crop yield predictions from both EPIC and the extreme gradient boosting models. Lowest median performance was found for the global simulations (EPIC global), followed by the high-resolution EPIC simulations at the scale of Mexico (EPIC high-res) with a slight tendency towards higher NSE. Interestingly, the median NSE for extreme gradient boosting predictions (Predicted high-res) is higher than for the EPIC simulations at the same resolution. This is mainly due to one municipio with rather poor performance in the simulations, while the predictions (Predicted high-res) do not achieve very high performance in other municipios where EPIC simulations result in up to NSE=0.8. The overall best rendition of inter-annual yield variability is produced by the machine learning predictions using 1k-resolution monthly climate surfaces (Predictions 1k) and more so if a national soil data product is used (Predictions 1k CRU x INEGI) with a median NSE of 0.42 as opposed to 0.20 in the high-resolution EPIC simulations (EPIC high-res). The CRU x HWSD combination in contrast results in a lower median but higher maximum NSE.

4 Discussion and Conclusions

4.1 Model performance for downscaling of yield estimates

Performance of the meta-models for spatio-temporal downscaling of crop yield estimates is exceptionally high in terms of linear regression statistics, and mean bias for both machine learning methods (Table S4; Table S5). While the results are highly comparable among the two methods, extreme gradient boosting shows moderately better results especially for inter-annual yield variability (cf. Tables S6-9), which is of ample importance for climate impact studies (e.g. Müller et al., 2017). In essence, substantial deviations of predictions from EPIC simulations occur only for very low yields. Even here, this applies foremost to their absolute magnitude while inter-annual yield variability is typically still very well reproduced although this is not an implicit goal of the machine learning model optimization. In addition, the high skill in reproducing irrigated yields stands out, as crop yield variability is known to be more strongly dominated by variability in precipitation than temperature in most regions (e.g. Frieler et al., 2017).

Our results can hardly be compared to existing literature, as the spatio-temporal downscaling of crop model outputs via meta-models has not yet been addressed to the authors' knowledge. Within the closely related, recently emerging field of crop model emulators, Blanc and Sultan (2015) and Blanc (2017) developed polynomial models to predict yields for various crops under climate change using unique parameterizations for the statistical models at the grid cell level. Besides weather and soil data, they include CO_2 as an additional dimension. These structural differences (a) grid-cell level in the references vs scale-free approach here and (b) no CO_2 dimension in the present study render the comparison of results difficult. The authors of the cited studies conclude that the statistical models provide reasonable results in the longer term. However, the visual comparison of inter-annual yield variability for the Corn Belt during the historic time period in Blanc and Sultan (2015) and the regional predictions presented in this study suggest that the polynomial models may be suitable at the global scale and for longer term assessments but not for regional impact studies. A similar statistical approach has been employed by Oyebamiji et al. (2015) for a single GGCM finding that 62-93% of crop yield variability produced by the GGCM can be explained by their multiple tier statistical model, which was as well parameterized at the grid cell level. This indicates that so far no other methodologic approaches can provide as accurate and flexible crop meta-models as the ones presented herein, which are also virtually scale-free, free from *a priori* assumptions on relevant features, and truly data-driven.

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

4.2 Feature engineering and feature importance

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

The feature importance of models for rainfed yield prediction is quite straightforward with precipitation and other water-related features strongly dominating (Figure 7a). Static variables PHU and LVP follow thereafter, rendering water availability the main driver for inter-annual yield variability, while especially PHU – a composite of growing season length and long-term temperatures – may rather serve as a proxy for the overall yield potential and thermal growth conditions. If monthly climate statistics are considered, the third and fourth months have the largest influence on rainfed yield predictions. This relates to the aforementioned non-linearity of crop growth requirements and the crop's higher sensitivity during the second half of the growing season.

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

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

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

4.3 Predictions of agricultural externalities

Agricultural externalities were assessed supplementary (Supplementary Text S2) to evaluate the potential of machine learning algorithms to predict these as well, which is an essential advantage of integrated crop growth models compared to purely statistical methods of crop yield estimation. The very good results for GSET show that this is in principle feasible. The slightly lower performance for CAW in turn indicates that there are limits under extreme conditions: The very high values that are underestimated here (Figure S9c,d) occur in simulation units with moderate to high precipitation, low slopes, and soils with high infiltration potential (not shown). Capturing also such combinations may require an extension of the training data set (see section 4.6). Overall, however, the results show that the computational framework used for yield predictions can flexibly be transferred to other crop model outputs. Limitations can still be expected for agro-environmental externalities that occur intermittently with daily peaks such as emissions of certain greenhouse gases.

4.4 Differences and advantages of employed machine learning approaches

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

Although the quantification of prediction uncertainty is beyond the scope of this study, it is worth mentioning that for random forests there are established methods to quantify prediction intervals and hence uncertainties associated with predictions (e.g. Meinshausen, 2006) for which no readily applicable methods have been developed for gradient boosting. Provided that the meta-model predictions show very high accuracy but outliers still occur, this may become of great importance for applications of downscaled yield estimates e.g. in land use change studies as well as in the quantification of trade-offs and benefits of (potential) meta-model error and improved coverage of landscape heterogeneity. We can hence conclude that within the scope of this study, the extreme gradient boosting approach appears most suitable, but still the selection of the most appropriate method needs to be made on a per case basis of a specific study.

4.5 Model performance benchmarked against reported local yields

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

4.6 Outlook

The meta-models presented herein can readily provide robust estimates within the domain of the training data, providing a solid proof of concept that machine learning bears great potential for building readily applicable crop meta-models for spatio-temporal downscaling applications. It is likely, however, that regional and specific local conditions are not represented within the global feature ranges and their combinations. In addition, crop cultivars are often adapted to regional conditions, e.g. in terms of temperature requirements and maturity classes. Here, we found that specific, extremely rare climate-soil combinations led to a systematic underestimation of the growing season soil water balance CAW. An option to train a meta-model for such conditions in a systematic way is to simulate artificial combinations of atmospheric, soil, cultivar, and management conditions that go beyond the combinations inherently occurring in the global database. This allows for covering an enhanced space of potentially prevailing plant growth conditions at finer resolutions. A similar approach has recently been undertaken within the GGCMI initiative (Elliott et al., 2015), altering atmospheric and management conditions in each simulation unit (resp. 0.5° x 0.5° grid cell) along the dimensions CO_2, temperature, precipitation, and N fertilizer (CTWN; Ruane et al., 2017) to develop crop model emulators for climate change impact studies among others. This can hence serve as a blueprint for extending as well the training data extent as well as its dimensionality for a wider range of applications and environments.

Acknowledgements

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

http://webarchive.iiasa.ac.at/~folberth/downscaling_paper/. The authors declare no conflicts of interest.

References

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

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

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

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

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

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

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* (pp. 785–794). New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785

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

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

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

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

Ewert, F., van Ittersum, M. K., Heckelei, T., Therond, O., Bezlepkina, I., & Andersen, E. (2011). Scale changes and model linking methods for integrated assessment of agrienvironmental systems. *Agriculture, Ecosystems & Environment*, *142*(1), 6–17. https://doi.org/10.1016/j.agee.2011.05.016

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

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

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

Fischer, G., F. Nachtergaele, O. Prieler, S. Teixeira, E. Tóth, G. van Velthuizen, H. Verelst, L. Wiberg, D. (2012). *Global AgroقEcological Zones. (GAEZ* International Institute for Applied Systems Analysis, Laxenburg, Austria and FAO, Rome, Italy. Retrieved from http://www.iiasa.ac.at/Research/LUC/GAEZv3.0/docs/GAEZ_Model_Documentation.pdf

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

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

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

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

Frieler, K., Schauberger, B., Arneth, A., Balkovič, J., Chryssanthacopoulos, J., Deryng, D., … Levermann, A. (2017). Understanding the weather signal in national crop-yield variability. *Earthôs Future*, *5*(6), 605–616. https://doi.org/10.1002/2016EF000525

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

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

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

Gerik, T., Williams, J., Francis, L., Greiner, J., Magre, M., Meinardus, A., Steglich, E., & Taylor, R. (2015). *Environmental Policy Integrated Climate Model - Userôs Manual Vers 0810*. Blackland Research and Extension Center, Texas A&M AgriLife, Temple, USA. Retrieved from http://agrilife.org/epicapex/files/2015/10/EPIC.0810-User-Manual-Sept-15.pdf

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

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

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

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

Hengl, T., Leenaars, J. G. B., Shepherd, K. D., Walsh, M. G., Heuvelink, G. B. M., Mamo, T., … Kwabena, N. A. (2017b). Soil nutrient maps of Sub-Saharan Africa: assessment of soil nutrient content at 250 m spatial resolution using machine learning. *Nutrient Cycling in Agroecosystems*, *109*(1), 77–102. https://doi.org/10.1007/s10705-017-9870-x

INEGI, 2004. Información Nacional sobre Perfiles de Suelo v1.2. Instituto Nacional de Estadística, Geografía e Informática. http://www.inegi.org.mx/geo/contenidos/recnat/edafologia/

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

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

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

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

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

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

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

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

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

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

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

Oyebamiji, O. K., Edwards, N. R., Holden, P. B., Garthwaite, P. H., Schaphoff, S., & Gerten, D. (2015). Emulating global climate change impacts on crop yields. *Statistical Modelling*, *15*(6), 499–525. https://doi.org/10.1177/1471082X14568248

Peel, M. C., Finlayson, B. L., & McMahon, T. A. (2007). Updated world map of the Köppen-Geiger climate classification. *Hydrol. Earth Syst. Sci.*, *11*(5), 1633–1644. https://doi.org/10.5194/hess-11-1633-2007

Portmann, F.T., Siebert, S., & Döll, P. (2010). MIRCA2000 - global monthly irrigated and rainfed crop areas around the year 2000: a new high-resolution dataset for agricultural and hydrological modeling. *Global Biogeochem. Cycles 24*, *GB1011*, https://doi.org/10.1029/2008GB003435

Pugh, T. A. M., Müller, C., Elliott, J., Deryng, D., Folberth, C., Olin, S., … Arneth, A. (2016). Climate analogues suggest limited potential for intensification of production on current croplands under climate change. *Nature Communications*, *7*, 12608. https://doi.org/10.1038/ncomms12608

R Development Core Team (2008). *R: A language and environment for statistical computing, R Foundation for Statistical Computing*. Vienna, Austria, Retrieved from https://www.R-project.org.

Reidsma, P., Ewert, F., Boogaard, H., & Diepen, K. van. (2009). Regional crop modelling in Europe: The impact of climatic conditions and farm characteristics on maize yields. *Agricultural Systems*, *100*(1), 51–60. https://doi.org/10.1016/j.agsy.2008.12.009

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

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

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

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

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

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

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

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

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

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

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

Stockle, C. O., Williams, J. R., Rosenberg, N. J., & Jones, C. A. (1992). A method for estimating the direct and climatic effects of rising atmospheric carbon dioxide on growth and yield of crops: Part I—Modification of the EPIC model for climate change analysis. *Agricultural Systems*, *38*(3), 225–238. https://doi.org/10.1016/0308-521X(92)90067-X

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

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

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

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

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

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

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

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