Agricultural and Forest Meteorology 264 (2019) 1–15

Contents lists available at ScienceDirect

Agricultural and Forest Meteorology

journal homepage: www.elsevier.com/locate/agrformet

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
[b] International Institute for Applied Systems Analysis, Advanced Systems Analysis Program, Schlossplatz 1, A-2361 Laxenburg, Austria
[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
[e] National Agricultural and Food Centre, Soil Science and Conservation Research Institute, Trencianska 55, 824 80 Bratislava, Slovak Republic

ARTICLE INFO

Keywords:
Meta-model
Extreme gradient boosting
Random forests
Maize yield
Agricultural externalities
Climate features

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.

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., 2015), GGCMs provide base layers of input data for agro-economic or integrated assessment models (IAMs; Müller and Robertson (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

* Corresponding author.
E-mail addresses: folberth@iiasa.ac.at (C. Folberth), baklanov@iiasa.ac.at (A. Baklanov), balkovic@iiasa.ac.at, balkovic@fns.uniba.sk (J. Balkovič), skalsky@iiasa.ac.at, r.skalsky@vupop.sk (R. Skalský), khabarov@iiasa.ac.at (N. Khabarov), oberstei@iiasa.ac.at (M. Obersteiner).

https://doi.org/10.1016/j.agrformet.2018.09.021
Received 30 May 2018; Received in revised form 23 September 2018; Accepted 29 September 2018

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).

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 Fig. 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, 2012), and an extended number of crop types and cultivars (e.g. Kiniry et al., 1995; Gaiser et al., 2010) among others (see Gassman et al. (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 1.3×10^5 globally and 2.3×10^5 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