Uncertainties in a global crop model ensemble: effects of parameterization and setup harmonization on maize yield estimates and model performance

Christian Folberth[1], Joshua Elliott[2,3], Christoph Müller[4], Juraj Balkovič[1,5], James Chryssanthacopoulos[3], Roberto C. Izaurralde[6,7], Curtis D. Jones[6], Nikolay Khabarov[1], Wenfeng Liu[8], Ashwan Reddy[6], Erwin Schmid[9], Rastislav Skalský[1,10], Hong Yang[8,11], Almut Arneth[12], Philippe Ciais[13], Delphine Deryng[3], Peter J. Lawrence[14], Stefan Olin[15], Thomas A.M. Pugh[12,16], Alex C. Ruane[3,17], Xuhui Wang[13,18]

[1]International Institute for Applied Systems Analysis, Ecosystem Services and Management Program, 2361 Laxenburg, Austria

[2]University of Chicago and ANL Computation Institute, Chicago, IL 60637, USA

[3]Columbia University Center for Climate Systems Research and NASA Goddard Institute for Space Studies, New York, NY 10025, USA

[4]Potsdam Institute for Climate Impact Research, 14473 Potsdam, Germany

[5]Comenius University in Bratislava, Department of Soil Science, 842 15 Bratislava, Slovak Republic

[6]University of Maryland, Department of Geographical Sciences, College Park, MD 20742, USA

[7]Texas A&M University, Texas AgriLife Research and Extension, Temple, TX 76502, USA

[8]Eawag, Swiss Federal Institute of Aquatic Science and Technology, CH-8600 Duebendorf, Switzerland

[9]University of Natural Resources and Life Sciences, Institute for Sustainable Economic Development, 1180 Vienna, Austria

[10]Soil Science and Conservation Research Institute, National Agricultural and Food Centre, 82713 Bratislava, Slovak Republic

[11]Department of Environmental Sciences, University of Basel, Petersplatz 1, CH-4003 Basel, Switzerland

[12]Karlsruhe Institute of Technology, IMK-IFU, 82467 Garmisch-Partenkirchen, Germany

[13]Laboratoire des Sciences du Climat et de l'Environnement. CEA CNRS UVSQ Orme des Merisiers, F-91191 Gif-sur-Yvette, France

[14]National Center for Atmospheric Research, Earth System Laboratory, Boulder, CO 80307, USA

[15]Department of Physical Geography and Ecosystem Science, Lund University, 223 62 Lund, Sweden

[16]School of Geography, Earth & Environmental Science and Birmingham Institute of Forest Research, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom

[17]National Aeronautics and Space Administration Goddard Institute for Space Studies, New York, NY 10025, USA

[18]Peking University, Sino-French Institute of Earth System Sciences, 100871 Beijing, China

Correspondence to: Christian Folberth (folberth@iiasa.ac.at)

Abstract. Global gridded crop models (GGCMs) combine field-scale agronomic models or sets of plant growth algorithms with gridded spatial input data to estimate spatially explicit crop yields and agricultural externalities at the global scale. Differences in GGCM outputs arise from the use of different bio-physical models, setups, and input data. While GGCM ensembles have been employed in earlier studies foremost to bracket uncertainties in climate change impacts, we disentangle here differences in maize yield estimates from five GGCMs based on the public domain field-scale model Environmental Policy Integrated Climate (EPIC) that participate in the AgMIP Global Gridded Crop Model Intercomparison (GGCMI) initiative. Albeit using the same crop model, the GGCMs differ in input data, management assumptions, parameterization, geographic distribution of cultivars, and selection of released or experimental model subroutines. The analyses show that both absolute yield levels and inter-annual yield variability in the EPIC-based GGCMs are highly sensitive to soil parameterization and handling, crop management, and cultivar distributions. Harmonizing key input data substantially reduces mean bias and improves time-series correlation of yield estimates among models. All GGCMs show a moderate performance in reproducing reported absolute yield levels or inter-annual yield dynamics often exhibiting increased skill if nutrient deficits are eliminated. Our findings suggest that studies focusing on the evaluation of differences in bio-physical routines may require further harmonization of input data and management assumptions in order to eliminate background noise resulting from differences in model setups. Future sensitivity analyses will need to be extended from presently dominating cultivar parameterizations to agro-environmental processes. For agricultural climate impact assessments, employing a GGCM ensemble with its widely varying assumptions in setups appears the best solution for bracketing such uncertainties as long as comprehensive global data allowing for a thorough spatial parameterization are lacking. Finally, we recommend improvements in the documentation of setups and input data of GGCMs in order to allow for sound interpretability, comparability and reproducibility of published results.

Keywords: Environmental Policy Integrated Climate, EPIC, agricultural management; agro-ecologic systems; crop cultivars; soil data; global agriculture

1 Introduction

1.1 Global gridded crop models and their applications

The assessment of agricultural production systems at the global scale is not only relevant for studying the supply of food under global change processes such as climate change (e.g. Rosenzweig 2014, Wheeler and von Braun, 2013) but also with respect to its impact on terrestrial carbon cycling, (e.g. Pugh et al. 2015, Müller et al. 2007), water (e.g. Elliott et al. 2014, Jägermeyr et al. 2015) and nutrient budgets (Bodirsky et al. 2014, Bouwman et al. 2009). Global gridded crop models (GGCMs) are common tools to study global patterns of agricultural production and are typically combinations of (a) a core model that estimates crop yields and externalities of crop production for a given set of input data and (b) a model framework that processes specified input data and runs the model over large regions or the globe. The first can in general be grouped into field-scale crop models, ecosystem models, or empirical models (Müller et al., 2017). Model frameworks are commonly based on computational interfaces and georeferenced data from earth observations, statistical databases, or modelers' assumptions depending on data availability. Over the past decade, GGCMs evolved to become major tools for agricultural climate change impact

assessments (e.g. Tan and Shibasaki, 2003; Liu et al., 2013; Balkovič et al., 2014; Elliott et al., 2014; Folberth et al., 2014; Rosenzweig et al., 2014; Müller et al., 2015; Deryng et al., 2016) or studies on agricultural externalities (e.g. Bondeau et al., 2007; Liu et al., 2007; Fader et al., 2010; Stehfest et al., 2007), besides providing key data for land use change and agro-economic models (e.g. Havlík et al., 2011; Schneider et al., 2011; Müller and Robertson, 2014; Nelson et al., 2014).

Despite their wide use and substantial deviations among studies based on single GGCMs (e.g. Fader et al., 2010; Liu et al., 2013), there has been little systematic investigation of actual drivers behind these uncertainties, which can be grouped into simulated processes, input data, parameterization, and management assumptions. While these have been addressed in single GGCM studies before - typically for one singled out model component - we present here a first evaluation of drivers in differences among yield estimates produced by a partly harmonized GGCM ensemble. The focus is thereby on five GGCMs based on the field-scale model Environmental Productivity Integrated Climate (EPIC) to reduce differences in the representation of plant growth processes as such, which are often subject of model intercomparison studies. Instead, we focus here on the representation of environmental processes and management assumptions, which we hypothesize to also greatly affect model performance and comparability. The results are related to a wider ensemble of GGCMs based on different core models to identify setup components for potential further harmonization and implications for future uncertainty analyses.

The following section provides a brief overview of recent uncertainty assessments in field- and large-scale crop models. Thereafter, the Global Gridded Crop Model Intercomparison (GGCMI; Elliot et al., 2015) initiative, which the data used in the present study are based on, is introduced and the objectives of the present study are laid out.

1.2 Sources of uncertainty in crop models and their assessment

1.2.1 Field-scale models

Field-scale models, which GGCMs or their crop growth components are typically based on (Rosenzweig et al., 2014; Müller et al., 2017), have been subject to extensive uncertainty assessments over the past decades (e.g. Vanclooster et al., 1995; Müller et al., 2017), which has been extended from single model to ensemble studies within the Agricultural Model Intercomparison and Improvement Project (AgMIP; Rosenzweig et al., 2013). Due to the commonly good availability of environmental, management, and benchmark data, models can be well calibrated and validated at this scale (Cavero et al., 1998; Gaiser et al., 2010; Srivastava and Gaiser, 2010; Wang et al., 2013b) and be tested against measured crop yields and biomass (Asseng et al., 2013; Bassu et al., 2014), including in-season dynamics (Martre et al., 2015) and externalities (Sándor et al., 2016). Thereby, model performance and ensemble agreement were found to increase substantially if models are calibrated to field conditions and trial management (Bassu et al., 2014; Sándor et al., 2016). Besides the trend towards crop model ensemble studies, advancements in computational capacities have allowed for the development of auto-calibration and sensitivity analysis frameworks covering various thousand parameter sets and allowing for highly detailed studies of parameter sensitivity and calibration in field-scale and regional models (e.g. Wang et al., 2013a).

1.2.2 Global gridded crop models

GGCMs in contrast rely foremost on global gridded input data, aggregated or downscaled to the model resolution (typically 0.5°x0.5°). Average crop yields reported at administrative levels are typically the only available

benchmark (e.g. Liu, 2009; Fader et al., 2010; Folberth et al., 2013; Balkovič et al., 2014; Gbegbelegbe et al., 2016; Liu et al., 2016a), which limits the feasibility of thorough calibration. Uncertainty analyses for GGCMs or large-scale crop models have hence mostly addressed model sensitivity to climate (e.g. Fader et al., 2010; Osborne et al., 2013; Liu et al., 2013) and soil (e.g. Zhang et al., 2014; Waha et al., 2015; Folberth et al., 2016) input data from different sources, spatial resolution of input data (e.g. Folberth et al., 2012; Angulo et al., 2014), and deviations in outputs from various GGCMs (Rosenzweig et al., 2014; Elliott et al., 2014) or different crop models within the same framework (Angulo et al., 2014; Waha et al., 2015). While the relative impact of either component differs spatially, depending on climate and management among others, it is commonly found to be crop model > climate > soil averaged over large scales.

Fewer studies have addressed parameter sensitivity and calibration, mostly focusing on the parameterization of crop cultivars, which can differ substantially by region due to adaptation to regional environmental conditions (i.e. climate), management preferences (e.g. rotations), and qualitative characteristics (e.g. Zhao et al., 2014). For example, Liu (2009) performed a sensitivity analysis for five selected parameters in an early version of the GEPIC model and found yield estimates to be most sensitive to the four crop growth-related parameters whereas a runoff-related parameter was regionally of moderate importance. Xiong et al. (2016) calibrated crop parameters and plant density for rice in a global EPIC model, finding that phenology information is of great relevance for the magnitude of estimated climate change impacts. At the regional level, Iizumi et al. (2009) identified rice growth parameters to induce larger uncertainty under unprecedented warming than under past climate conditions. Similarly, Angulo et al. (2012) identified crop parameter calibration as an important driver in climate change impact magnitudes for a range of crops at the European scale.

In contrast to input data and crop parameters, uncertainties in crop and soil management (e.g. tillage regimes), parameterization of soil organic matter (SOM) cycling, and hydrologic model components are hardly considered although they have been shown to introduce substantial uncertainties at the field-scale scale (e.g. Basso et al., 2016; Sándor et al., 2016). Only two recent studies addressed sensitivity to environmental processes routines and parameterization, i.e. the selection of potential evapotranspiration (PET) routines at the global scale (Liu et al., 2016a) and the parameterization of processes relating to soil degradation at the European level (Balkovič et al., in preparation).

Common to all cited studies is that they either treat ensemble uncertainty as a given or cover *a priori* selected parameters and subroutines considered relevant by the respective researchers. While the latter is essential for singling out model sensitivity in certain processes, it does not allow for assessing the overall uncertainty in differing concepts behind GGCMs, which can deviate substantially due to intended use of crop model outputs, focus regions, and modelers' assumptions on representative parameterizations. For ensemble impact studies, the wide ranging differences among and often limited meta-data provided for the models do in turn not allow for an in-depth interpretation of deviations.

1.3 The Global Gridded Crop Model Intercomparison initiative and model outputs used in this study

A first appraisal of uncertainties in climate change impact estimates from an ensemble of GGCMs in their default setups coordinated by the AgMIP (Rosenzweig et al., 2013) and ISI-MIP (Warszawski et al., 2014) projects revealed that GGCMs – even if based on the same core model and forced with identical climate data - provide substantially differing yield estimates (Elliott et al., 2014; Rosenzweig et al., 2014). To study the underlying

drivers more systematically, the GGCMI initiative (Elliott et al., 2015) provided 14 participating research groups with input data for historic climate, fertilizer application rates, and growing seasons to assess the impact of these key input data harmonization on crop yield estimates. Thereby, GGCMs were forced (a) in their default setups with the same climate data, (b) in addition with harmonized growing seasons and fertilizer rates, and (c) in addition with sufficient nutrient supply to allow for a comparison of model behavior and remaining differences after step-wise harmonization and elimination of plant growth constraints. Any other model parameters and setup components that had been selected by the participating groups to their best knowledge were kept constant across these scenarios.

Among the 14 GGCMs participating in GGCMI phase 1, five are based on the field-scale model EPIC. Besides being the most frequent core model, EPIC has within the ensemble the most detailed representation of soil processes with options for various submodules and parameterizations, which allows for assessing the impacts of such routines that are included in an ever increasing number of field-scale and global crop models (e.g. Jones et al., 2016). Outputs from another seven GGCMs that contributed simulations for at least one of the harmonized scenarios are included to relate the EPIC-based ensemble to a wider range of models.

1.4 Objectives of the study

The objectives of the present study are to identify drivers for differences in maize yield estimates among the five EPIC-based GGCMs caused by model setup and parameterization in order to (a) identify key assumptions that drive differences in yield estimates in the EPIC-based GGCM ensemble, (b) derive priorities for further improvements in model input data (e.g. management aspects) or – in the absence of suitable data sets – harmonization in the wider GGCMI ensemble to address central uncertainties as a set of scenarios rather than random choices made by modelers (Confalonieri et al. 2016), and (c) provide a thorough documentation of GGCM setups and their implications as an aid in the potential further use of the publically available GGCMI phase 1 outputs and in the interpretation of earlier studies the participating models were used in (e.g. Elliott et al., 2014; Rosenzweig et al., 2014; Deryng et al., 2016; Müller et al., 2017; Schauberger et al., 2017; Frieler et al., forthcoming). A detailed evaluation is conducted for two EPIC-based GGCMs by step-wise introducing aggregated setup domains from one into the other in order to provide an exemplary quantification of the importance of cultivar setups, organic matter turn-over and nutrient cycling, hydrologic parameterization, soil data and handling, and management.

Evaluations address bias among models in absolute yield estimates, time-series correlation of yield estimates, and benchmarking against global and national yield statistics from FAO (FAO, 2014). The latter is constrained to countries for which reported yields can be considered reliable, i.e. the major maize producing countries and selected further countries for which production data have not been estimated by FAO. Rather than aiming at the identification of a single best model, which cannot be expected to exist due to spatially static parameterizations within each GGCM, the benchmarking serves foremost for evaluating the role of setup components in model performance.

2 Methods and Data

2.1 Global gridded crop models in this study

A total of 14 GGCMs contributed model outputs to GGCMI phase 1 (see Sect. 1.3), 12 of which provided outputs with harmonized input data (Section 2.2). These were selected for the present study to allow for comparison of model outputs produced with partial harmonization. The ensemble consists of five ecosystem model-based GGCMs (CLM-Crop, LPJ-GUESS, LPJmL, ORCHIDEE-Crop, PEGASUS) and seven crop model-based GGCMs, five of which use EPIC (EPIC-BOKU, EPIC-IIASA, EPIC-TAMU, GEPIC, PEPIC), one APSIM (pAPSIM), and one DSSAT (pDSSAT). Except for the EPIC-based GGCMs, all models differ in their combinations of phenology, photosynthesis, plant stress, soil, and management modules. Table S1-1 provides an overview of key GGCM characteristics. More detailed information are provided on the website of ISI-MIP (http://www.isimip.org) and in Müller et al. (2017). In the further text, the terms model and GGCM are used synonymously unless otherwise specified.

2.2 Crop management scenarios

Six crop management scenarios (Table 1) were simulated to quantify differences among GGCMs based on three steps of growing season and nutrient supply harmonization. Each of these was combined with two water management scenarios, rainfed only or sufficiently irrigated, resulting in a total of six management setup combinations per GGCM.

The default scenario represents each research group's assumptions on annual fertilizer application rates and growing seasons (see SI 1.3 for EPIC-based GGCMs). It serves for evaluating differences among GGCMs if only climate data are harmonized. The fully harmonized (fullharm) setup allows for identifying remaining differences if annual nutrient application rates and growing seasons are harmonized using the input data described in Sect. 2.6. The fully harmonized setup with sufficient nutrient application (harm-suffN, referred to as harmnon in the simulation protocol (Elliott et al. 2015)) aims to virtually eliminate plant nutrient deficits and consequently impacts of soil nutrient dynamics. This is expected to minimize differences among GGCMs resulting from the setup of fertilizer application and soil nutrient cycling. If a model supplied only the default (LPJ-GUESS, LPJmL) or fullharm setup (EPIC-TAMU) besides harm-suffN, the submitted data were used for both scenarios in evaluations to keep the number of models across scenarios constant.

Irrigation water is applied in all GGCMs up to a sufficient amount automatically based on plant water requirement in the irrigated management scenarios. The application takes place based on varying thresholds in each model (see Table 2 for EPIC-based GGCMs).

2.3 The Environmental Policy Integrated Climate (EPIC) field-scale model

The EPIC model was first developed in the 1980s to assess the impacts of soil management on crop yields (Williams et al., 1989). It has been updated frequently to cover e.g. 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 (Gassman, 2004). The presently publically available version is EPIC v.0810.

EPIC estimates potential biomass increase on a daily time-step based on light interception and conversion of CO_2 to biomass. Plant growth and phenology are calculated based on the daily accumulation of heat units. Potential biomass accumulation is constrained by water and nutrient (nitrogen (N) and phosphorus (P)) deficits, adverse temperature, and aeration stress. On each day of the crop growth period, the potential biomass gain is adjusted by the major plant growth-regulating factor to obtain the actual biomass increment. Hence, only one stress factor limits biomass accumulation on a given day. Root growth can be limited by soil strength, adverse soil temperature, and aluminum toxicity. At maturity, the model calculates crop yield based on above ground biomass and an actual harvest index HIa, which is estimated within a range given by potential HI (HI_{max}) and minimum HI under water stress (HI_{min}).

Besides plant growth and yield formation, EPIC estimates a wide range of environmental externalities, for example wind and water erosion rates, turnover and partitioning of organic matter (OM) based on the CENTURY model (Parton et al., 1994; Izaurralde et al., 2006), mineral N and P cycling, evapotranspiration (ET), fluxes of selected gases, and soil hydrologic processes. All of these have feedbacks on plant growth, mainly through nutrient and water availability. EPIC has one central plant growth module, but provides various subroutines for calculating several of the externalities, e.g. six methods for water erosion estimation, eleven methods for estimating field capacity (FC) and wilting point (WP) including static input of own estimates or data, and five options for potential evapotranspiration (PET) among others. Due to its modularity, various configurations even of the same EPIC version may be considered different agro-environmental models. While this allows for adjusting the model to site conditions for which one method may be more appropriate than the other, it introduces another dimension of uncertainty besides the numeric parameterization of processes itself. Further information on relevant subroutines are provided in SI 1.1.

2.4 Setups and parameterizations of EPIC-based global gridded crop models

All EPIC-based GGCMs except for EPIC-TAMU use EPIC v.0810 as the core model. The EPIC-TAMU version uses the experimental version v.1102 (Izaurralde et al. in prep.), which has additional routines mainly for OM and nutrient cycling, but the same plant growth module. As shown in Sect. SI 1.2, EPIC v.1102 produces virtually identical outputs in high-input regions but shows differences in low-input agriculture where nutrient cycling has larger impacts on plant growth. It may hence be considered another configuration of EPIC in this context.

All models have been applied in prior studies except EPIC-TAMU which has first been set up in the course of this project. Based on prior applications or modelers' parameter estimates suitable for global model runs, the EPIC-GGCMs differ substantially in their parameterization and selection of subroutines. E.g. the GEPIC model has been set up for reproducing small-holder agriculture in sub-Saharan Africa (Folberth et al., 2012), relying partly on parameters calibrated in West Africa (Gaiser et al., 2010) whereas EPIC-IIASA has frequently been applied in high-input regions such as the EU (Balkovič et al., 2013) or China (Xiong et al., 2014b). Outlines of model setups, purposes, and prior applications are provided in SI 1.3. Table 2 gives an overview of the setups and parameterizations grouped by hydrology, soil degradation, OM and nutrient cycling, crop management, and crop growth apart from cultivar definitions, which are described in Sect. 2.5. Numbers in braces below refer to column "No" in the table. As there are interactions among model processes such as hydrology and OM cycling, this grouping is tentative and partly owed to the model structure. Concerning the choice of major subroutines, three GGCMs use Penman-Monteith (PM; Monteith, 1965) for PET estimation (1) and two Hargreaves (HG;

Hargreaves and Samani, 1985) in different parameterizations. Only EPIC-IIASA uses prior estimated FC and WP parameters (9) while all other EPIC-based GGCMs estimate these parameters using Rawls method online (Rawls and Brakensiek, 1985). Water erosion (14) is considered in two of the models (and wind erosion in an additional one) with deviations in estimation method (16) and scaling of sediment yield (15). Three EPIC-GGCMs have a dynamic soil profile (19) with transient updating of profile depth, texture, OM and nutrient pools, and hydrology. In the two models with static soil profiles, soil texture and OM are re-initialized at the beginning of each year, but not mineral nutrient pools and soil humidity. All models are run transiently (20), except for GEPIC, which is run for each decade separately with a spin-up of 30 years (see SI 1.3.4). All three methods available for estimating denitrification (21) are used in the EPIC-GGCM ensemble. Numeric parameters agree in some cases among models, especially if default parameter values have been selected (e.g. microbial decay rate (22)) but differ in several cases among four to five EPIC-GGCMs as in the case for the N volatilization coefficient (26). Different values have also been selected for defining irrigation water and fertilizer application strategies (27-29), and EPIC-TAMU differs in addition from the other models in the parameterization of selected growth process (30-33).

2.5 Geographic distributions and parameterization of maize cultivars

Crop cultivars are here considered to be defined by HI_{min}, HI_{max}, and optimal temperature ranges only, whereas the heat unit requirement is prescribed by growing season input data. Between one and four different maize cultivars were planted within each EPIC-GGCM (Figure 1; Table S1-4). EPIC-IIASA uses four cultivars in its default setup (Figure 1a) that are attributed to major world regions based on climatic and economic characteristics. The same geographic distribution of cultivars was used for EPIC-IIASA in the harmonized setup scenarios except that the early and late maturing high-yielding cultivars 1 and 3 were merged as growing season length was defined according to common input data sets (see Sect. 2.6). EPIC-TAMU (Figure 1b) plants high- and low-yielding varieties. The latter is assigned to countries in which maize yields have stagnated or decreased within the past decades according to Ray et al. (2012). The high-yielding variety is assigned to all other regions. The same two maize cultivars were distributed in GEPIC and PEPIC (Figure 1c) based on the HDI. The high-yielding variety is planted in all countries with HDI≥80, which corresponds to high development. EPIC-BOKU used the high-yielding variety in all grid cells (Figure 1d).

2.6 Common input data

Climate forcing data based on the WFDEI GPCC dataset (Weedon et al., 2014) at a spatial resolution of 0.5°x0.5° were provided by the ISI-MIP and GGCMI projects. The climate data are based on temperature and solar radiation from ERA-interim (Dee et al., 2011) and precipitation from GPCC (Schneider et al., 2013). All EPIC-based GGCMs used soil data from the ISRIC-WISE database (Batjes, 2006) mapped to the Digital Soil Map of the World (FAO, 1995). For EPIC-BOKU and EPIC-IIASA, the 5000 soil profiles had been reduced to the original 120 soil typologic units WISE is based on (Skalský et al., 2008). Soil hydraulic parameters not provided in the WISE database (FC, WP and saturated conductivity) were estimated for EPIC-IIASA according to Schaap and Bouten (1996) and Wösten and Van Genuchten (1988) using the ROSETTA model. A comparison of key variables in both datasets in provided in Figure S1-3.

For the harmonized runs, nutrient application rates for N and P were based on crop-specific data from Mueller et al. (2012) to which manure application rates had been added proportionally. Separate planting dates and growing

season lengths for rainfed and irrigated management were based on Sacks et al. (2010), complemented by gap filling with data from the MIRCA2000 dataset (Portmann et al., 2010) and the LPJmL model (Waha et al., 2012). Both datasets were provided by the GGCMI project (Elliott et al., 2015). Default runs were carried out using individual fertilizer and growing season data within each GGCM.

2.7 Permutation of setup domains for GEPIC and EPIC-IIASA

To identify the importance of single data and parameterization domains within the EPIC-GGCMs, aggregated setup domains of EPIC-IIASA were step-wise introduced into the GEPIC model. Parameters and routines were grouped into the 6 domains (Table 3) of cultivar distribution (Cult), soil data (SoilD), soil handling (SoilP), nutrient-related coefficients (CoeffN), hydrologic coefficients (CoeffW), and crop management (Manage). The two GGCMs were selected based on providing stable fluctuating yields over time (Figure 2) and substantial differences in their setups (Table 2).

The GEPIC model was run with all 64 (2^6) resulting setup combinations using the land mask of EPIC-IIASA to ensure consistency. The evaluation focuses on rainfed yield estimates as these cover the whole range of uncertainty impacts. This approach serves for evaluating differences between GGCMs designed by different research groups and is hence fundamentally different from common parameter sensitivity analysis typically carried out with single GGCMs for parameters selected *a priori* based on a specific research subject.

Magnitudes of plant growth stresses are included to analyze drivers behind different yield estimates. Benchmarking against reported yields (see Sect. 2.8) at the country-level serves for quantifying the contribution of single setup domains to model performance besides model sensitivity for a given setup domain in contrasting countries.

2.8 Model evaluation and reference data

2.8.1 Yield aggregation and metrics for model agreement

Crop yields are compared and evaluated at the global, national, and grid level as well as clustered by Koeppen-Geiger regions (Figure S2-1). Model agreement is compared in relation to fertilizer application rates, mean annual precipitation (MAP), and cultivar distributions to identify drivers of deviations in yield estimates.

Global and national average yields (YD_{av}) were calculated from simulated rainfed and irrigated yields in each grid cell and the respective rainfed and irrigated harvested areas (Portmann et al., 2010) according to

$$YD_{av,c} = \frac{\sum_{g=1}^{m}\left[YD_{i,g} \times HA_{i,g} + YD_{r,g} \times HA_{r,g}\right]}{\sum_{g=1}^{m}\left[HA_{i,g} + HA_{r,g}\right]} \quad (1)$$

where $YD_{av,c}$ is the national average yield in country c, $YD_{i,g}$ is yield under irrigated conditions in grid cell g, $YD_{r,g}$ is yield under rainfed conditions in grid cell g, $HA_{i,g}$ is irrigated area in grid cell g, and $HA_{r,g}$ is rain fed area in each grid cell g, and m is the number of grid cells in country c. We acknowledge the uncertainty introduced from spatial aggregation (Porwollik et al., 2016) but as the focus is on a comparison among GGCMs, we consider this to be of minor importance here.

The coefficient of variation (CV) [%] was used as a metric for absolute bias among yields averaged throughout the study period (CV_{av}; e.g. Figure 3) as well as changes in inter-annual yield dynamics if model setup components are introduced from GEPIC into EPIC-IIASA (CV_t; e.g. Figure 9). The coefficient of variation is expressed as

$$CV = \frac{S}{\bar{X}} \times 100\% \quad (2)$$

where S is the standard deviation and $\bar{X}$ is the mean of yields throughout the evaluation period in each grid cell or globally aggregated. CV_{av} was calculated for the period 1980-2009 as the first simulation year 1979 did not have a complete growing season globally. CV_t was calculated the same way but after scaling the time series average to 1 in order to avoid bias caused by changes in the magnitude of yields.

The mean error (ME) was used in the same evaluations as a metric for absolute model bias including the sign of change:

$$ME = \frac{\sum_{i=1}^{n} YD_{est} - YD_{ref}}{n} \quad (3)$$

where YD_{est} is the yield estimate, YD_{ref} is the reference yield, and n is the number of years considered. In the permutation of model setup domains (see Sect. 2.7) difference are evaluated in relative terms compared to the original EPIC-IIASA setup. ME hence corresponds there (Figure 9) to the fraction of relative change [-].

To test the agreement in inter-annual yield variability, Pearson's correlation coefficient r was calculated for yield estimates in each grid cell. As one of the EPIC-based GGCMs exhibited a substantial decline in yields after the first simulation years, the evaluation period was here limited to 1980-1990 in order to avoid bias from unexpected model behavior later in the simulation period.

All evaluations were carried out with the statistics software R (R Development Core Team, 2008) using the packages ggplot2 (Wickham, 2009), corrplot (Wei and Simko, 2016), and the heatmap.2 function of gplots (Warnes et al., 2016) in a modified version from Müller et al. (2017) for visualization.

2.8.2 Benchmark metrics and data processing

Annual national average crop yields from the FAOSTAT database (FAO, 2014) were used for assessing model performance. Reported yields were de-trended by subtracting the 5-year moving mean in order to remove trends in yields due to changes in technology and management (Elliott et al., 2015; Müller et al., 2017). The mean bias in absolute yield estimates from reported yields was measured as mean error (ME, see Section 2.8.1). As a reference, the de-trended yields from FAO were multiplied by their mean of the period 1997-2003 for which fertilizer inputs are representative.

Model performance of the ensemble has been assessed in detail in Müller et al. (2017). Here, it serves predominantly for comparing differences in model skills in relation to differences in setups focusing on the two harmonized management scenarios to ensure comparability in key input data. Following the methodology of Müller et al., we used the time series correlation (tscorr) according to Pearson's correlation coefficient r between national average simulated and reported yields as the main metric. For the calculation of tscorr, also GGCM outputs were de-trended as described above for the reference data. A significance threshold of $p<0.1$ (at approx. $r>0.35$) was selected for defining good model performance. The statistical significance of differences in model performance with different management setups was measured using the analysis of variance (ANOVA) combined with Tukey's honest significant difference (HSD) test at $p=0.05$. Tukey's HSD indicates whether the means of two samples are significantly different, which is indicated by different letters (e.g. Figure 8).

The evaluation of model performance itself is often limited by the quality of benchmark (Müller et al., 2017) and landuse data (Porwollik et al., 2016), characteristics of climate data (Ruane et al., in preparation) and

representativeness of management data for a given region. Benchmarking itself is hence subject to substantial uncertainties and was hence here limited to major producers and other countries for which available benchmark and management data can be considered representative. These were selected based on whether (a) production and harvest area data had not been estimated by FAO and (b) harvested area did not fluctuate by >100% throughout the study period to account for the static cropland mask used in the aggregations.

3 Results

3.1 Effects of setup harmonization on global average maize yield estimates

If the EPIC-based GGCMs are run in their default setups, global average simulated maize yields differ by up to 124% annually (mean 95%) using the lowest estimate as a reference (Figure 2a; Table S3-1). This is mainly due to very high yield estimates from EPIC-BOKU of around 8 t ha^{-1}, while the other EPIC-based GGCMs have yield estimates of around 4-6 t ha^{-1} . The ranges decrease to 55% if harmonized planting dates and fertilizer application rates are used (Figure 2b) and further to 26% with sufficient nutrient supply (Figure 2c). The bias from reported yields varies greatly with the highest deviation for EPIC-BOKU (ME=3.7 t ha^{-1}) and the lowest for GEPIC (-0.25 t ha^{-1}) and PEPIC (0.44 t ha^{-1}) in the default setups (Table S3-2). The bias decreases for EPIC-BOKU and remains fairly constant for the other models in in the fullharm setup. With sufficient nutrient supply, the mean bias of the whole EPIC ensemble increases expectedly to 1.5-2.5 t ha^{-1}.The mean bias is not constant over time, however, with significant negative trends in yield estimates for PEPIC in all setup scenarios, and for EPIC-BOKU in the fullharm and harm-suffN scenarios. EPIC-IIASA in contrast shows a slight positive trend in its default setup. The inter-annual yield variability still exhibits similar patterns and the whole EPIC-based ensemble indicates a peak in global average yield in 2004, which is also apparent in the reported data.

3.2 Spatial differences in mean and inter-annual maize yield estimates

3.2.1 Variability in long-term mean yields

Spatially, the deviation of maize yield estimates among the EPIC-GGCMs is largest with the default setups in tropical and arid regions (defined in Figure S2-1) with CV_{av} of up to 224% and CV ≥50% in >44% of all grid cells (Figure 3a,b; Table S3-3). The most distinct differences were found in sub-Saharan Africa, particularly the countries of Angola and Mozambique. Large differences with CV_{av} around 100% occur also in South America, India, and Southeast Asia. The lowest differences with CV_{av} <100% and often <50% are found in mid and high latitudes of both hemispheres, where (a) fertilizer inputs are at moderate or high levels (Figure S2-2), (b) most models plant the same high-yielding cultivar (Figure 1) and (c) the climatology usually defines a narrow growing season window limiting differences among models in planting date assumptions. Rainfed cultivation results in larger differences among models in (semi-)arid regions of Central and West Asia, the Western USA and North-eastern Brazil. If irrigation water is applied, differences increase in most parts of sub-Saharan Africa and Central India, but decrease in most of North and South America, Central Asia, and Europe.

Harmonizing fertilizer and growing season data reduces the CV_{av} to ≤64% under rainfed and ≤54% under irrigated conditions in 75% of all grid cells (Figure 3c,d; Table S3-3). Spatial patterns remain largely similar to those found for the default managements. However, the CV_{av} increases in some regions after harmonization. E.g., in the low fertilizer input region Western Russia where EPIC-IIASA plants a cultivar adapted to colder climates and provides

high yields in comparison to the other EPIC-based GGCMs (Figure S3-2), CV_{av} increases from around 30-40% in the default setups to often 50-60% in the fullharm ensemble. Also in the semi-arid region of North-eastern Brazil, EPIC-IIASA provides very high yields, while EPIC-BOKU and EPIC-TAMU do so in Southern China. In both cases, GEPIC and PEPIC plant a low-yielding cultivar (Figure 1; Figure S3-2). However, cultivar distributions alone cannot explain these differences as national average yields are in various countries at comparable magnitudes for models with different cultivar distributions as is the case for South Africa and Kenya (Figure S3-2). This indicates that other processes can superimpose the impact of cultivar distributions.

The application of sufficient nutrients further reduces differences among EPIC-GGCMs. Under irrigated conditions, the CV_{av} is now in 50% of the grid cells ≤25% and under rainfed conditions ≤29% (Figure 3e,f; Table S3-3). Differences among model frameworks decrease especially in regions with low or moderate reported fertilizer application rates (see Figure S2-2) such as India, sub-Saharan Africa, and South America. Large CV_{av} remains in the same regions as in the fullharm setup, foremost in Southern China. An important remaining driver is the cultivar distribution: if all models plant one of the high-yielding Cultivars 1 or 2, CV_{av} is typically lowest (Figure S3-3). The deviation increases for regions in which all models that use Cultivar 4 plant this variety and is often 50-100% higher than the first option if none of the cultivars dominates, which mostly refers to regions in which GEPIC and PEPIC but not EPIC-IIASA and EPIC-TAMU plant cultivar 4. This effect is stronger under rainfed than under irrigated conditions.

3.2.2 Agreement in inter-annual yield dynamics

In the default setup, the median time-series correlation coefficient r is often around 0 (Figure 4a,b; see also Figure S3-5), except for temperate and cold regions in case of sufficient irrigation and additionally arid regions under rainfed conditions. In the latter case, still >40% of all grid cells have a correlation with at least $p<0.1$ (Table S3-4). Harmonization provides a slight improvement, foremost with a higher correlation in regions that already had a moderate agreement in the default setups (Figure 4c,d). Low agreement prevails especially in the tropics and along the Eurasian border. With sufficient nutrient supply (Figure 4e,f), there is a significant correlation in >69% under both irrigated and rainfed water supply, and a very high agreement at $p<0.01$ in 53% or 45%, respectively (Table S3-4). The largest deviations remain in the humid tropics of sub-Saharan Africa and South America, South(-East) Asia, and along the Eurasian border.

The cultivar distribution shows here a more diverse impact (Figure S3-4) than for the CV_{av}. Regardless of the water supply regime and cultivar definitions, the agreement is high in arid regions with lowest agreement if all models plant a high-yielding cultivar with irrigation. At overall lower agreement, the picture is similar in the tropics and also applies to the mixed cultivar definitions. Cold and foremost temperate regions in contrast show a gradient in decreasing model agreement from uniform planting of the high-yielding cultivar towards dominant low-yielding cultivar or mixed cultivar definitions. Due to substantial deviations in sample sizes of grid cells per cultivar distributions and climate region, however, the statistical significance could not be determined here.

3.3 Impact of fertilizer supply and precipitation on deviations in maize yield estimates

The model bias in terms of CV_{av} is in most climate regions highly correlated with the level of N fertilizer supply with decreasing differences in mean yield estimates among EPIC-GGCMs with increasing N application rates (Figure 5a-h; see Figure S2-2 for spatial fertilizer application rates). Although linear regressions are highly

significant in all climate region/water supply combinations, the explained variance is substantially higher under irrigated (Figure 5a-d) than under rainfed conditions (Figure 5e-h). CV_{av} is on average at about 60-85% in all climate regions at very low N application levels and highest in arid regions under irrigated conditions. It decreases on average to about 21-23% in arid and temperate regions and 16-17% in tropical and cold regions at applications rates above 200 kg N ha^{-1} yr^{-1}. Rainfed cultivation substantially dampens the effect of nutrient application rates as a driver for differences among models in (semi-)arid climates and leaves larger deviations among models at moderate to high application rates also in other climate regions.

In addition, the correlation among models increases with increasing fertilizer application rates (Figure 6a-h) and as for CV_{av} especially under irrigated conditions (Figure 6a-d), but with comparably little explained variance. The highest impact of fertilizer application can be found in cold and temperate regions under irrigated conditions (Figure 6b,c). In arid regions under rainfed conditions, the correlation is often already high at low fertilizer application rates (Figure 6e) indicating that that climatic drivers dominate here the model inter-correlation. In the tropics in contrast, where fertilizer application rates are commonly moderate to low, the correlation is at all application rates lower than in other regions (Figure 6d,h). Binning the fertilizer application rates (Figure S3-6) shows that there may rather be thresholds of application rates allowing for high model correlation at least in arid, cold and temperate regions where a substantial increase can be found at >150 kg N ha^{-1} compared to low fertilizer levels, again foremost with irrigation water supply (Figure S3-6a-d).

Comparing both statistical indicators CV_{av} and time-series correlation to mean annual precipitation (MAP) (Figure S3-7; Figure S3-8) in contrast shows little explained variance except for CV_{av} in arid regions (Figure S3-11e), which decreases with increasing MAP and explains the low correlation with fertilizer (Figure 5e). The density distributions indicate trends towards lower CV_{av} and (in non-arid regions) lower model agreement with increasing precipitation but no linear relationship.

3.4 Differences in model performance among EPIC-based GGCMs

The time series correlation coefficient r between annual national average yields and detrended reported yields shows very mixed performance of the EPIC-based GGCMs in the fullharm setup. EPIC-BOKU and EPIC-IIASA have a significant correlation in the largest number of countries (56) followed by EPIC-TAMU and finally GEPIC and PEPIC (Table 4). For the number of countries with highest performance (column “best”), however, EPIC-TAMU ranks second after EPIC-IIASA. GEPIC and PEPIC, notably the two GGCMs considering a dynamic soil profile and erosion (Table 2), exhibit here the poorest performance. In the harm-suffN scenario, the numbers for best performance converge among GGCMs with largest increases for GEPIC and PEPIC. The number of countries with significant correlation narrows down to a range of ±2 among all models except for PEPIC. The largest improvements are provided for EPIC-TAMU and GEPIC.

For the ten major maize producing countries, substantial variability in model performance can be observed as well for the fullharm (Figure 7a) as for the harm-suffN (Figure 7b) scenarios. With the fully harmonized setup, EPIC-IIASA shows in most cases the best performance, followed by PEPIC, and finally EPIC-BOKU and GEPIC. All models show high performance in the USA and France and low performance for Indonesia or Mexico (see also Table S3-5). For the harm-suffN scenario, the best performing models change in various countries, primarily those with overall low to moderate time-series correlation, such as Brazil, Indonesia, and Mexico with decreases in ensemble performance in the latter two (Figure 7b).

In various countries in which at least one GGCM has a high performance in the fullharm scenario, various EPIC-GGCMs achieve better results with sufficient nutrient supply, which can be observed for Argentina and to a lesser extent India (Table S3-5). While no direct relationship between the level of fertilizer supply and model performance could be detected, Figure 8a,d and Table S3-5 show that there is a trend towards better performance with high nutrient supply in the EPIC ensemble. This is foremost the case for GEPIC, for which the median correlation coefficient increases by >30% from fullharm to harm-suffN whereas most improvement can be found in countries with low to moderate fertilizer application rates, most notably Argentina and India. For Indonesia, the performance remains insignificant regardless of the fertilizer supply scenario and GGCM and even decreases slightly which is as well observed for Mexico and Ukraine. Extending the analysis to all countries is limited by benchmark data quality and changes in agronomic systems throughout the study period in parts of the world (Müller et al., 2017). For a wider range of countries for which neither yields have been estimated by FAO nor maize harvest area changed 100% throughout the study period, the observed trend appears robust but as well not significant (Figure S3-9).

3.5 Impact of single setup domains on maize yield estimates, plant stress and model performance

3.5.1 Differences in yield estimates and plant stress

The evaluation of differences in model setup domains between EPIC-IIASA and GEPIC (see Sect. 2.7) focuses on the relative difference from the complete EPIC-IIASA setup (Figure 9a-p). Absolute yields are provided in Figure S3-10. Magnitudes of plant stresses (Figure S3-11) cannot be related to differences in yield estimates directly as their impact depends on estimates of potential biomass growth in the model (driven e.g. by growing season length, climate and management; see also SI 1.1) besides the timing of the stress occurrence throughout the growing season. They are hence only addressed per panel but not among different managements. To allow for a more thorough understanding of the drivers of yield estimates and associated stresses, selected examples for single grid cells are provided in Figures S3-12 and Figure S3-13, and dominant stresses for contrasting setups in Figure S3-14.

If only the management is replaced in the EPIC-IIASA setup, yields increase slightly and show a change in inter-annual yield variability in terms of CV_t (Figure 9a; see Section 2.8.1) despite an increase in phosphorus (P) and water (W) deficits (Figure S3-11a). This is caused by the narrower row spacing in GEPIC (Table S1-3), which increases the estimate of potential biomass (see SI 1.1) often resulting in higher actual biomass estimates despite higher stress occurrence (see Figure S3-12 for grid level example). Replacing also the cultivars scales yields down and increases variability as GEPIC plants the low-yielding cultivar in a larger number of countries (Figure 1c,a), which outweighs the higher HI_{max} specified for the high-yielding cultivar (Table S1-4). Introducing the gCoeffN parameters into the setup (Figure 9b) increases yields in all cultivar x management combinations and affects inter-annual yield dynamics whereas nutrient-related stresses decrease (Figure S3-11b) due to more rapid turnover of organic matter (Table 2; see Figure S3-13 for point level example of eCoeffN vs gCoeffN). The slight increase in temperature (T) stress is hence a secondary effect due to the stress handling in the model selecting only the major limiting factor for biomass production on a given day (see Sect. 2.3). The gCoeffW parameters in turn import little change on yield variability but slightly scale yields up if combined with eCoeffN (Figure 9c).

Replacing the static soil handling of EPIC-IIASA by the dynamic decadal runs of GEPIC (Figure 9e-h) alters yield levels and inter-annual dynamics substantially with about 15% lower yields than in the corresponding eSoilP

scenarios. Nutrient deficits become the dominant growth constraint, especially in combination with eCoeffN (Figure S3-11e,g), which causes a slower release of nutrients from OM and higher volatilization of N. The higher P stress with gCoeffN (Figure S3-11f,h) is often a secondary effect of high N availability early in the simulation that cause more rapid P mining form the soil in low-P input regions and a concomitant increase in P stress (see Figure S3-13i for grid example).

The soil data of GEPIC (Figure 9i-p) result in an increase in yield estimates and changes in inter-annual yield variability in all scenarios (Figure 9e,g vs Figure 9m,p). This is driven by decreases in N and W stress and increases in P stress if a static soil profile eSoilP is employed or if the dynamic soil handling gSoilP is combined with eCoeffN (Figure S3-11i-p). The most significant difference between the soil datasets is in the estimation of hydraulic parameters field capacity (FC) and wilting point (WP) where EPIC-IIASA has typically higher values for the first and lower for the latter (Figure S1-3). Both parameters affect a wide range of processes in the model, among them the threshold for percolation of water and the optimal soil humidity for microbial processes (see SI 1.1). The gSoilD component hence allows for providing larger amounts of nutrients from OM as required soil humidity is reached earlier, but causes higher water stress as an effect of (a) lower water storage capacity and (b) higher model sensitivity to climate stresses caused by higher nutrient supply. In combination with the static soil profile gCoeffN (Figure 9j,l), nutrient stresses are virtually eliminated and yield estimates are foremost driven by climate (see also Figure S4-4d), potential biomass accumulation, and cultivar specification, which leaves the highest yields estimated with this setup and the cultivar distributions eCult and management gManage (same panels and Figure S3-7j,l).

Analyzing the correlation of the global weighted yields among all scenarios shows that the combination of eSoilD, gSoilP, and eCoeffN (Figure 9e,g) has the lowest agreement with the remainder of setups (Figure S3-15). In turn, the nutrient and OM turnover parameterizations and soil data of GEPIC (gCoeffN and gSoilD) as well as the soil handling of EPIC-IIASA (eSoilP) render the remaining setup resilient to other changes in setup domains (Figure S3-16). The other setup domains in contrast show bimodal distributions and hence depend more strongly on interactions with other setup components.

3.5.2 Impact of setup domains on model performance

The permutation of setup domains greatly affects the model performance in reproducing inter-annual yield variability (Figure 10). The maximum correlation coefficient increases in all countries compared to the basic fullharm setups of the two models (Figure 7a), except for Argentina where the full EPIC-IIASA setup still provides the highest skill (with slightly lower r due to different digit precisions in output files used in GEPIC). In various countries, the setup with the highest correlation coefficient also exceeds the highest value of the EPIC ensemble (Figure 7a), apart from China and France. This also applies to the performance of the EPIC ensemble in the harm-suffN scenario (Figure 7b) for countries with low to moderate performance such as Brazil and Mexico.

The countries show inherently different sensitivities to the exchanged setup domains with distinct patterns of positive or adverse impacts of setup components (Figure S3-17). In the US (Figure S3-17a), the correlation is considerably high with any setup whereas the OM and nutrient cycling parameterization eCoeffN typically provides a higher correlation. The opposite is the case for the corresponding GEPIC component gCoeffN, which has been derived from a calibration in semi-arid low-input regions. For China in contrast, the performance is overall moderate and model performance is most sensitive to soil data but also cultivar definitions (Figure S3-

17b). In Argentina and India (Figure S3-17c,d), two countries with low reported fertilizer rates (Table S3-5), performance shows a very high sensitivity to model setups ranging from r<0 to r>0.8. The most sensitive setup domain is here the soil handling with lowest results for gSoilP that introduces artefacts in low-input regions (see SI 1.3.4).

Most setups components however show bimodal distributions indicating their relative performance to depend on other parameters they are combined with. In the US for example, only the hydrologic parameters of GEPIC (gCoeffW) allow for very high correlation, although the corresponding parameters form EPIC-IIASA show on average a higher correlation. Yield reproducibility for Argentina shows in addition a considerable sensitivity to management, while soil data affect the performance more strongly for India. To further evaluate and disentangle such interactions, however, is beyond the scope of this study and will need to be subject to more detailed regional sensitivity assessments.

3.6 Relations of the EPIC-based GGCMs to the wider ensemble

The members of the wider GGCM ensemble differ substantially in the representation and detail of plant growth and agro-environmental processes, stress handling, consideration of soils, and fertilizer effects (Table S1-1). The continuous decrease in the spread among EPIC-based GGCMs with increasing level of harmonization and elimination of nutrient limitations is contrasted by an increasing spread for the non-EPIC-based GGCMs most notably in the harm-suffN scenario (Figure S3-18b,d,f). This is driven by very high yield estimates by two GGCMs and very low estimates by one, while three simulate yield potentials at a similar level as the EPIC ensemble. Besides this increase in spread among the GGCMs, the range of yields is in the fullharm scenario among the non-EPIC-based GGCMs about three times as large as among the EPIC-based.

The correlation of yield estimates from the whole GGCM ensemble at the grid cell level is expectedly far lower. While the EPIC-based GGCMs show a poor median correlation in the default setup and a significant increase with harmonization and elimination of nutrient deficits (Figure 4a-f), this is less so for the wider ensemble (Figure S3-19; Table S3-7). Various models show a fairly even distribution of correlation coefficients across management scenarios with the mode around r=0, most notably two ecosystem model based GGCMs CLM-crop and PEGASUS, although the peak of the latter moves to a high correlation in the harm-suffN scenario with sufficient irrigation for all EPIC-based GGCMs except EPIC-IIASA. The ecosystem model-based ORCHIDEE-crop shows a good agreement with various GGCMs, foremost EPIC-IIASA, EPIC-TAMU, and PEPIC in the fullharm simulations (no harm-suffN simulations were provided by ORCHIDEE-crop for maize), although this model differs substantially in the representation of plant growth, associated stresses, and representation of soil processes (Table S1-1). When considering irrigation, the agreement of the whole ensemble is comparably high in cold to temperate climates of the northern hemisphere and extends to parts of the tropics with increasing harmonization (Figure S3-20a,c,e). Under rainfed conditions, the agreement is high in most of the US, Europe and various arid regions and improves with increasing harmonization in the same areas (Figure S3-20b,d,f). The spatial patterns remain largely constant albeit with overall lower agreement if the EPIC-based models are removed from the ensemble (Figure S3-21a-f; Table S3-8).

The performance of the wider ensemble shows for the two scenarios with harmonization better skills if sufficient nutrients are supplied in terms of the best performance (Figure S3-22). However, in most cases, the fullharm and harm-suffN results are not significantly different for the best performing model or only harm-suffN simulations

are considered as is the case for LPJ-GUESS and LPJmL. While the EPIC-based and the two other crop model-based GGCMs (pDSSAT, pAPSIM) show a trend towards better performance with sufficient nutrient supply – mainly due to higher tscorr in few countries with a combination of moderate performance under fullharm and low to moderate reported fertilizer supply -, the only model in the ensemble that has a significant increase in model skill is PEGASUS (Figure 8a-f). This applies, however, only to the major producers and not the wider range of countries (Figure S3-9f) although a positive trend can be observed also there.

4 Discussion

4.1 Representations of agricultural systems and environmental conditions in the EPIC-based GGCM ensemble

4.1.1 Effects of model setups on plant growth limitations

The EPIC-based GGCMs investigated herein represent inherently different agricultural systems resulting in varying crop growth conditions and limitations. For EPIC-BOKU the average global yields in the default setup indicate the virtual elimination of nutrient stresses (Figure 2a) but in the harmonized scenarios the model setup causes a decrease in yields likely attributed to nutrient depletion (see Sect. 4.3). Similarly, EPIC-IIASA exhibits in its fullharm setup dominant climate stresses (Figure S3-11a; Figure S3-14a), which can be assumed to be highly similar in the default setup that is based on nearly identical input data (SI 1.3.2) and shows only slightly lower global average yields (Figure 2a). For GEPIC in contrast, productivity on large parts of the global cropland is limited by nutrient deficits (Figure S3-14b), which can also be expected for PEPIC based on the identical cultivar distribution, similar global average yield levels (Figure 2a), and the consideration of dynamic soil handling (Table 2). EPIC-TAMU presents a compromise including soil OM dynamics over time but no water erosion. The EPIC ensemble can hence be expected to bracket a range of agricultural systems especially in regions with mixed high-input and smallholder agricultural, which prevails across most of the tropics (e.g. Fritz et al., 2015). However, all GGCMs assume high-input agriculture in regions with reported high average fertilizer application rates, which neglects potential imbalance and heterogeneity in input systems in these regions.

4.1.2 Differences in model setups and their implications

The cultivar distributions in terms of harvest index and optimal temperature ranges reflect substantial diversity and rationales as well resulting in contrasting cultivar representations in most parts of the world except highly developed countries (Figure 1). The highest detail is provided here in the setup of EPIC-IIASA albeit none of the GGCMs can be considered representative for the range of global maize types (e.g. Hartkamp, 2000).

The parameterizations of soil OM and nutrient turnover encompass in various cases recommended parameter ranges (Gerik et al., 2004) and hence a wide range in assumptions on microbial process dynamics. For example, coefficients for slow to passive humus partitioning (Table 2, parameter 23) range from 0.003-0.05, resulting in higher or lower availability if nutrients from OC. The N volatilization coefficient (parameter 26) ranges from 0.005-0.7, which can greatly affect the estimation of gaseous N losses. Values for both parameters are at the edges of recommended ranges (Gerik et al., 2004) and hence bracket fairly extreme cases. While the microbial decay rate (parameter 22) is in a fairly narrow range, it affects OM turnover substantially as a direct multiplier in the microbial turnover term (see SI 1.1). Various field and regional studies showed that OM partitioning and turnover

parameters for CENTURY are subject to substantial uncertainty as C pools are conceptual representation and hence not possible to measure (e.g. Matthews et al., 2005). Hence, these parameters typically require to be calibrated to field conditions.

Hydrologic components in contrast reflect rather methodologic uncertainties in selecting different methods for estimations of PET, FC and WP, and parameterizing run-off processes. While a recent study based on PEPIC (Liu et al., 2016a) found that the PM method provided the best fitting average national yield estimates within the otherwise constant model framework other studies suggest that PET estimation requires calibration to local conditions in general (e.g. Gavilán et al., 2006). Also for the estimation of FC and WP, which have not been subject to sensitivity analyses in crop models to the authors' knowledge, hydrologic studies arrive at different conclusions depending on the field conditions and soils (Givi et al., 2004; Baroni et al., 2010). The substantial increase in model agreement with elimination of nutrient deficits indicates that hydrologic routines affect model agreement and performance at least within this ensemble to a lesser extent than soil handling and nutrient cycling, which is also supported by the comparison of EPIC-IIASA and GEPIC setup domains (e.g. Figure S3-17).

General management coefficients (Table 2, parameter 27-29) follow pragmatic assumptions rather than the representation of actual farming systems. Low triggers (i.e. high values for parameters 27 and 29) allow for a more rapid plant stress reduction but cause in the case of fertilizer an earlier consumption of the annul maximum rate, especially if also a low trigger is selected for irrigation water application, which may result in stronger nutrient leaching. Low fertilizer application triggers in contrast do not allow for full plant stress reduction in the harm-suffN scenario. The actual operations compared for GEPIC and EPIC-IIASA (Table S1-3) differ most substantially in the removal of plant residue and row spacing affecting long-term nutrient availability and potential biomass estimation both of which depend on socio-economic decisions and prevailing practices on-farm or locally.

In summary, the EPIC-based GGCMs can be considered contrasting representations of agricultural systems and agro-ecologic conditions globally. Due to lack of spatial parameterizations except for a limited number of cultivars, however, none of them can be expected to represent an optimal setup. Instead, they allow for bracketing uncertainties that may exist at small scales (Ewert et al., 2011), e.g. through the representation of low-input or high-input agricultural systems and associated plant growth limitations within a pixel, which GGCMs cannot consider in their presently too coarse spatial resolution. For a more targeted representation of agricultural systems, however, setups should be compiled in a more consistent way. E.g., GEPIC is rather representative for smallholder systems but the narrow planting densities and ploughing operations are more common to high-input agriculture.

4.2 Interpretability and reproducibility of GGCM experiments

The broad differences in setups also has implications for the interpretability of past studies. The characterization of GGCMs in ensemble studies has been fairly limited in the past (e.g. Rosenzweig et al., 2014) and became more comprehensive recently (Müller et al., 2017). The details of the EPIC-based GGCMs laid out herein in contrast allow for a well-grounded interpretation of the different behaviours of EPIC-BOKU and GEPIC under climate change in Rosenzweig et al. (2014): As EPIC-BOKU largely neglects nutrient limitations in its default setup, yield estimates show often a higher sensitivity to adverse climate change impacts, which is limited in GEPIC by superimposed nutrient deficits at least for nutrient-demanding crops such as maize. Also studies based on single

GGCMs are typically limited in their descriptions of parameter selection and management setups, which are not static but often adjusted on a per case basis.

The reproducibility of GGCM simulations, even if core models themselves are publically available, may remain limited even if in-depth characterizations as presented herein as various input data such as soil characteristics and management schedules may be are stored on a grid basis as is the case for EPIC. However, a full characterization as presented herein would allow at least for a first step in making GGCM experiments replicable.

4.3 Effects of harmonization on ensemble agreement

Harmonization expectedly decreases mean bias and increases correlation among the EPIC-based GGCMs in most parts of the world (Figure 2, Figure 3, Figure 4), foremost driven by nutrient supply and climatic conditions with remaining differences partly caused by cultivar distributions. This highlights that management assumptions alone greatly affect yield estimates at least among same family models (Figure 2a,b). The better agreement in terms of correlation among models among the wider ensemble (Figure S3-20; Figure S3-21) predominantly in regions with suboptimal water supply from precipitation indicates that the majority of models shows similar reactions to water stress as has been shown in an earlier study for the US including some of the models (Schauberger et al., 2017). The substantial divergence in the tropics in the harm-suffN scenario, where under non-nutrient limited conditions plant phenology and photosynthesis dominate biomass accumulation and yield formation, indicates that these processes exhibit differences that exceed the impact of nutrient supply in the EPIC ensemble (Figure 4c compared to Figure S3-21e,f). A wider range of diagnostic variables that were not collected in this experiment will be required to further pin down these differences.

The negative yield trend observed for EPIC-BOKU after harmonization indicates an incompatibility of the GGCM's parameterization with the fullharm and harm-suffN management scenarios apparently causing nutrient depletion despite assumed sufficient supply, which could not be further assessed here. Also one of the non-EPIC-based GGCMs, LPJ-GUESS shows very low yield estimates after setup harmonization, which is due to incompatibility between the default model parameterization and harmonized input data. These arbitrary behaviors show that harmonization of GGCMs even at the level of basic input data is not a trivial process but may rather need to be carried out iteratively to derive parameterizations that are robust across management scenarios.

4.4 The role of nutrient dynamics and soil handling in reproducing inter-annual yield variability

Nutrient limitations and soil handling were found to hamper substantially the reproduction of inter-annual yield estimates in various models and countries (e.g. Figure 7a,b; Figure 8). Especially the performance of GEPIC with its decadal run setup (see SI 1.3.4) massively improves with sufficient nutrient supply due to strongly dominating nutrient stresses in all regions with limited fertilizer application, which limits its sensitivity for climate stress analysis and potentially introduces artefacts from the decadal runs in the form of yield peaks at the beginning of each simulation decade (Figure S1-2). While yields were found to be driven by climate in about 30% of the global maize areas in a recent study (Ray et al., 2012) the superimposed nutrient stresses and inter-annual nutrient dynamics do not allow the model to pick-up this signal. Although stress handling can differ substantially among GGCMs based on different core models, this appears to be also the case for other models with apparently high sensitivity for nutrient deficits such as PEGASUS, which includes a multiplicative stress function for photosynthesis (Deryng et al., 2011). While these results appear contradictory to those of Müller et al. (2017) who

found no effect of eliminating nutrient stress on model performance within the ensemble, we selected here only countries for which data may be considered to have a certain level of reliability compared to all countries in the global dataset in the cited study. It also needs to be noted that various models exhibit the best performance in their default setups for which they have prior been calibrated and/or validated (Müller et al., 2017). These were not taken into account here for reasons of comparability in input data (see Section 2.8). While further evaluations of input data used in the default setups of each model would provide valuable information on whether these can provide in general better performance than global datasets often derived at administrative levels, going into these details will have to be subject of follow-up studies.

4.5 Implications for global model evaluation

Global and large-scale crop models are typically evaluated against long- or mid-term global or national average yields (e.g. Stehfest et al., 2007; Fader et al., 2012; Folberth et al., 2012; Gbegbelegbe et al., 2016) accounting for temporal dynamics only to a limited extent (e.g. Liu et al., 2007; Osborne et al., 2013; Balkovič et al., 2014). Inter-annual yield dynamics have come under scrutiny only recently (Müller et al., 2017).

While a good score in both mean bias and time-series correlation is desirable, the interaction of nutrient stresses and climate sensitivity poses a dilemma. Due to the static management setup in the GGCMs, good skill in reproducing reported time-series yield variability can be expected only for countries, in which inter-annual yield dynamics are dominated by climate (Müller et al., 2017). For countries with low to moderate fertilizer inputs, a model well capable of reproducing absolute yield levels may, however, not do so for inter-annual yield variability if inter-annual nutrient dynamics and nutrient-climate-stress interactions in the model control yield dynamics in the simulations. This is apparently the case here for EPIC-TAMU, GEPIC, PEPIC, and PEGASUS which show large improvements in reproducing inter-annual yield variability if sufficient nutrients are supplied (Figure 8; Figure S6-1; Figure S3-9; Table S3-5), at least for countries in which other members of the GGCM ensemble show good skills, indicating that yield dynamics can be picked up by GGCMs in general. This suggests that GGCMs set up for reproducing yield levels under low-input conditions should also be tested for reproducing inter-annual yield variability under non-nutrient limited conditions to assess both performance indicators. However, it needs to be noted that yield dynamics in many low-input regions are driven by technologic changes such as recurring development programs (e.g. Denning et al., 2009), which may hamper reproducibility in static management setups in general. Hence, a classification of drivers affecting inter-annual yield dynamics per country or region would be required to derive a general recommendation under what assumptions GGCMs are best to be evaluated.

4.6 Implications for global scale sensitivity analyses and model calibration

While recent years have seen a vast growth in sensitivity analyses and calibration efforts of crop and plant models at the field, local, and regional scales (e.g. Iizumi et al., 2009; Angulo et al., 2012; Wang et al., 2013a; Valade et al., 2014; Zhao et al., 2014) with increasing methodologic sophistication (e.g. Wang et al., 2013a), global crop models have been subject to such studies only to a very limited extent (e.g. Liu, 2009; Xiong et al., 2016). In both cases a focus is typically on directly plant growth related parameters (e.g. photosynthesis, leaf development, or temperature response) or these are identified as the most sensitive variables.

Acknowledging the narrow parameterization of cultivars herein, the detailed comparison of EPIC-IIASA and GEPIC shows that other model components such as soil microbial processes and hydrologic parameters may import substantial impact on yield estimates depending on the representation of soil processes (Figure 9; Figure S3-16; Figure S3-17). Such parameters should hence be included in future sensitivity analyses at the global scale in order to derive a full picture of model sensitivity globally. This has briefly been addressed for nitrogen turnover and leaching processes by Liu et al. (2016b) but will require further attention.

The large-scale calibration of crop models with a focus on crop parameters hence bears the potential pitfall of projecting uncertainty in environmental processes unto representation of plant genetic coefficients. A shortcoming in global calibration efforts in general is also the limited availability of benchmark data, which is presently restricted to reported crop yields of often limited quality (Section 2.8.1; Müller et al., 2017). This is contrasted by the availability and quality of environmental parameters and wider ranges of diagnostic variables at the field scale (e.g. Bassu et al., 2014; Basso et al., 2016; Sándor et al., 2016). A thorough calibration of both environmental processes and crop parameters based on field experiments and their upscaling based on zoning approaches as presented by Xiong et al. (2008) for rice yields China does hence rather lend itself for deriving global spatially explicit model parameterizations. Global crop model development may here profit from the growing number of trial sites and parameterizations covered by the AgMIP project.

5 Conclusions

The results presented herein highlight the importance of parameter choices in global-scale crop modeling studies that have not received much attention so far and are often not explicitly described. The parameter choices by individual modeling groups do not reflect poor vs. careful parameterization but the lack of reference and input data on many aspects of agricultural production systems, such as soil and management aspects. This leads to limited comparability of outputs from various GGCMs based on the same, or very similar, sets of biophysical algorithms, as demonstrated here for a set of five EPIC-based GGCMs. A more thorough documentation of model setups may hence be required in the future to allow for sound interpretations of model behavior in as well single model as ensemble studies. The differences in model setups translating into different outputs indicate that further steps of harmonization among GGCMs should be taken if plant growth algorithms or other submodules are to be compared globally. Soil data may lend themselves to be harmonized for avoiding differences in nutrient supply in low-input regions from SOM mineralization and especially differences in soil hydrology. While a compilation of global crop management practices cannot be expected in the short run, management practices regarding the timing of fertilizer application and thresholds for automatic fertilizer and water irrigation need to be harmonized to avoid deviations among GGCMs impairing the interpretability of plant water stress. To address this issue, model runs in GGCMI phase 2 will be performed with prescribed timing and rates of fertilizer application. However, the representation of contrasting agricultural systems and methodologic approaches presents an asset for impact studies as it allows for bracketing uncertainties in actual conditions at the sub grid level. This does not only affect the simulation of agricultural yields, but also has implications for simulations of soil carbon and water consumption in biogeochemical studies. Improvements may be required, though, in the representativeness of GGCMs for defined agricultural production systems. Instead of leaving the representation of different agricultural production systems to individual user choices, these could also be described in counterfactual scenarios in

intercomparison and other model application studies so that implications of these uncertainties could be explicitly discussed and interpreted.

Data availability

Model outputs used in this study are available through the GGCMI data archive at http://www.rdcep.org/research-projects/ggcmi.

Acknowledgements

We acknowledge the support and data provision by the Agricultural Intercomparison and Improvement Project (AgMIP). AA and TAMP were funded by the European Commission's 7th Framework Programme, under Grant Agreement number 603542 (LUC4C). CF, JB, NK, and RS were supported by the European Research Council Synergy Grant number ERC-2013-SynG-610028 (IMBALANCE-P). C.M. acknowledges financial support from the MACMIT project (01LN1317A) funded through the German Federal Ministry of Education and Research (BMBF).

References

Angulo, C., Gaiser, T., Rötter, R. P., Børgesen, C. D., Hlavinka, P., Trnka, M., & Ewert, F.: 'Fingerprints' of four crop models as affected by soil input data aggregation. Eur. J. Agron., 61, 35-48, 2014.

Asseng, S., Ewert, F., Rosenzweig, C., Jones, J. W., Hatfield, J. L., Ruane, A. C., Boote, K. J., Thorburn, P. J., Rotter, R. P., Cammarano, D., Brisson, N., Basso, B., Martre, P., Aggarwal, P. K., Angulo, C., Bertuzzi, P., Biernath, C., Challinor, A. J., Doltra, J., Gayler, S., Goldberg, R., Grant, R., Heng, L., Hooker, J., Hunt, L. A., Ingwersen, J., Izaurralde, R. C., Kersebaum, K. C., Muller, C., Naresh Kumar, S., Nendel, C., O'Leary, G., Olesen, J. E., Osborne, T. M., Palosuo, T., Priesack, E., Ripoche, D., Semenov, M. A., Shcherbak, I., Steduto, P., Stockle, C., Stratonovitch, P., Streck, T., Supit, I., Tao, F., Travasso, M., Waha, K., Wallach, D., White, J. W., Williams, J. R., and Wolf, J.: Uncertainty in simulating wheat yields under climate change, Nat. Clim. Change, 3, 827-832, 2013.

Balkovič, J., van der Velde, M., Schmid, E., Skalský, R., Khabarov, N., Obersteiner, M., Stürmer, B., and Xiong, W.: Pan-European crop modelling with EPIC: implementation, up-scaling and regional crop yield validation. Agric. Syst., 120, 61-75, 2013.

Balkovič, J., van der Velde, M., Skalský, R., Xiong, W., Folberth, C., Khabarov, N., Smirnov, A., Mueller, N. D., and Obersteiner, M.: Global wheat production potentials and management flexibility under the representative concentration pathways, Global Planet. Change, 122, 107-121, 2014.

Baroni, G., Facchi, A., Gandolfi, C., Ortuani, B., Horeschi, D., and Van Dam, J. C.: Uncertainty in the determination of soil hydraulic parameters and its influence on the performance of two hydrological models of different complexity. Hydrol. Earth Syst. Sci., 14, 251-270, 2010.

Basso, B., Dumont, B., Shcherbak, I., Paustian, K., Nendel, C., Grace, P., Thorburn, P., Cammarano, D., Asseng, S., Boote, K.J., Biernath, C., de Sanctis, G., Durand, J.-L., Grant, R., Gayler, S., Kent, J., Priesack, E., Ripoche, D., Ruane, A. C., Hatfield, J. L., Jones, J. W., Rosenzweig, C.: Accounting for the legacy of soil and crop management when assessing climate change impact on crop production. AgMIP 6th Global Workshop, Montpellier, France.

Bassu, S., Brisson, N., Durand, J. L., Boote, K., Lizaso, J., Jones, J. W., et al.: How do various maize crop models vary in their responses to climate change factors?. Global Change Biol., 20, 2301-2320, 2014.

Batjes, N.H.: ISRIC-WISE derived soil properties on a 5 by 5 arc-minutes global grid. ISRIC - World Soil Information, Wageningen, Netherlands, 2006.

Bodirsky BL, Popp A, Lotze-Campen H, Dietrich JP, Rolinski S, Weindl I, Schmitz C, Müller C, Bonsch M, Humpenöder F, Biewald A, and Stevanovic M.: Reactive nitrogen requirements to feed the world in 2050 and potential to mitigate nitrogen pollution, Nat. Commun., 5, 3858, 2014.

Bondeau, A., Smith, P. C., Zaehle, S., Schaphoff, S., Lucht, W., Cramer, W., Gerten, D., Lotze-Campen, H., Müller, C., Reichstein, M., and Smith, B.: Modelling the role of agriculture for the 20th century global terrestrial carbon balance. Global Change Biol., 13, 679-706, 2007.

Bouwman AF, Beusen AHW, and Billen G.: Human alteration of the global nitrogen and phosphorus soil balances for the period 1970-2050, Global Biogeochem. Cycles, 23, Gb0a04, 2009.

Cavero, J., Plant, R. E., Shennan, C., Williams, J. R., Kiniry, J. R., and Benson, V. W.: Application of EPIC model to nitrogen cycling in irrigated processing tomatoes under different management systems. Agric. Syst., 56, 391-414, 1998.

Confalonieri R, Orlando F, Paleari L, Stella T, Gilardelli C, Movedi E, Pagani V, Cappelli G, Vertemara A, Alberti L, Alberti P, Atanassiu S, Bonaiti M, Cappelletti G, Ceruti M, Confalonieri A, Corgatelli G, Corti P, Dell'Oro M, Ghidoni A, Lamarta A, Maghini A, Mambretti M, Manchia A, Massoni G, Mutti P, Pariani S, Pasini D, Pesenti A, Pizzamiglio G, Ravasio A, Rea A, Santorsola D, Serafini G, Slavazza M, and Acutis M.: Uncertainty in crop model predictions: What is the role of users?, Environ. Modell. Software, 81, 165-173, 2016.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., and Bechtold, P.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system:. Q. J. R. Meteorol. Soc., 137, 553-597, 2011.

Denning, G., Kabambe, P., Sanchez, P., Malik, A., Flor, R., Harawa, R., Nkhoma, P., Zamba, C., Banda, C., Magombo, C., Keating, M., Wangila, J., and Sachs, J.: Input subsidies to improve smallholder maize productivity in Malawi: toward an African Green Revolution. PLoS Biology 7, e1000023, 2009.

Deryng, D., Sacks, W. J., Barford, C. C., and Ramankutty, N.: Simulating the effects of climate and land management practices on global crop yield, Global Biogeochem. Cycles, 25, doi:10.1029/2009GB003765, 2011

Deryng, D., Elliott, J., Folberth, C., Müller, C., Pugh, T. A., Boote, K. J., Conway, D., Ruane, A. C., Gerten, D., Jones, J. W., Khabarov, N., , Olin, S., Schaphoff, S., Schmid, E., Yang, H., and Rosenzweig, C.: Regional disparities in the beneficial effects of rising CO_2 emissions on crop water productivity, Nat. Clim. Change, 6, 786-790, 2016.

Elliott, J., Deryng, D., Müller, C., Frieler, K., Konzmann, M., Gerten, D., et al.: Constraints and potentials of future irrigation water availability on agricultural production under climate change. Proc. Natl. Acad. Sci. U.S.A., 111, 3239-3244, 2014.

Elliott, J., Müller, C., Deryng, D., Chryssanthacopoulos, J., Boote, K. J., Büchner, M., Foster, I., Glotter, M., Heinke, J., Iizumi, T., Izaurralde, R. C., Mueller, N. D., Ray, D. K., Rosenzweig, C., Ruane, A. C., and

Sheffield, J.: The Global Gridded Crop Model Intercomparison: data and modeling protocols for Phase 1 (v1.0), Geosci. Model Dev., 8, 261-277, 2015.

Elshout, P. M. F., Van Zelm, R., Balkovič, J., Obersteiner, M., Schmid, E., Skalský, R., Van Der Velde, M., and Huijbregts, M. A. J.: Greenhouse-gas payback times for crop-based biofuels, Nat. Clim. Change, 5, 604-610, 2015.

Ewert, F., Van Ittersum, M.K., Heckelei, T., Therond, O., Bezlepkina, I., and Andersen, E.: Scale changes and model linking methods for integrated assessment of agri-environmental systems. Agric. Ecosyst. Environ., 142, 6–17, 2011.

Fader, M., Rost, S., Müller, C., Bondeau, A., and Gerten, D.: Virtual water content of temperate cereals and maize: present and potential future patterns, J. Hydrol. 384, 218–231, 2010.

FAO: FAO Digital Soil Map of the World, FAO, Rome, 1995.

FAO: FertiSTAT - Fertilizer Use Statistics, Food and Agricultural Organization of the UN, Rome, 2007.

FAO: FAOSTAT statistical database, Food and Agricultural Organization of the UN, Rome, 2014.

Folberth, C., Gaiser, T., Abbaspour, K. C., Schulin, R., and Yang, H.: Regionalization of a large-scale crop growth model for sub-Saharan Africa: model setup, evaluation, and estimation of maize yields, Agr. Ecosyst. Environ., 151, 21-33, 2012.

Folberth, C., Yang, H., Gaiser, T., Abbaspour, K. C., and Schulin, R.: Modeling maize yield responses to improvement in nutrient, water and cultivar inputs in sub-Saharan Africa, Agric. Syst., 119, 22-34, 2013.

Folberth, C., Yang, H., Gaiser, T., Liu, J., Wang, X., Williams, J., and Schulin, R.: Effects of ecological and conventional agricultural intensification practices on maize yields in sub-Saharan Africa under potential climate change, Environ. Res. Lett., 9, 044004, 2014.

Folberth, C., Skalský, R., Moltchanova, E., Balkovič, J., Azevedo, L. B., Obersteiner, M., and van der Velde, M.: Uncertainty in soil data can outweigh climate impact signals in global crop yield simulations, Nat. Commun., 7, 11872, 2016.

Frank, S., Schmid, E., Havlík, P., Schneider, U.A., Balkovič, J., Obersteiner, M., and Böttcher, H.: The dynamic soil carbon mitigation potential of European cropland, Global Environ. Change, 35, 269-278, 2015.

Frieler, K., Arneth, A., Balkovic, J., Chryssanthacopoulos, J., Deryng, J., Elliott, J., Folberth, C., Khabarov, N., Müller, C., Olin, S., Pugh, T. A. M., Schaphoff, S., Schewe, J., Schmid, E., Schauberger, B., Warszawski, L., and Levermann, A.: Understanding the weather-signal in national crop-yield variability, Earth's Future, accepted.

Fritz, S., See, L., McCallum, I., You, L., Bun, A., Moltchanova, E., et al.: Mapping global cropland and field size. Global Change Biol., 21, 1980-1992, 2015.

Havlík, P., Schneider, U. A., Schmid, E., Böttcher, H., Fritz, S., Skalský, R., Aoki, K., Cara, S. D., Kindermann, G., Kraxner, F., Leduc, S., McCallum, I., Mosnier, A., Sauer, T., and Obersteiner, M.: Global land-use implications of first and second generation biofuel targets, Energy Policy, 39, 5690–5702, 2011.

Gaiser, T., de Barros, I., Sereke, F., and Lange, F.-M.: 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, Agr. Ecosyst. Environ., 135, 318-327, 2010.

Gassman, P. W., Williams, J. R., Benson, V. W., Izaurralde, R. C., Hauck, L. M., Jones, C. A., Atwood, J. D., Kiniry, J. R., and Flowers, J. D.: Historical Development and Applications of the EPIC and APEX models, ASAE/CSAE Meeting Paper No. 042097, 2004.

Gavilán, P., Lorite, I. J., Tornero, S., and Berengena, J.: Regional calibration of Hargreaves equation for estimating reference ET in a semiarid environment, Agr. Water Manage., 81, 257-281, 2006.

Gbegbelegbe, S., Cammarano, D., Asseng, S., Robertson, Richard D., Chung, U., Adam, M., Abdalla, O., Payne, T., Reynolds, M., Sonder, K., Shiferaw, B., and Nelson, G. C.: Baseline simulation for global wheat production with CIMMYT mega-environment specific cultivars, Field Crops Res., in press. http://dx.doi.org/10.1016/j.fcr.2016.06.010, 2016

Gelfand, I., Sahajpal, R., Zhang, X., Izaurralde, R. C., Gross, K. L., and Robertson, G. P.: Sustainable bioenergy production from marginal lands in the US Midwest, Nature, 493, 514-517, 2013.

Gerik, T., Williams, J., Francis, L., Greiner, J., Magre, M., Meinardus, A., Steglich, E., and Taylor, R.: Environmental Policy Integrated Climate Model - User's Manual Version 0810, Blackland Research and Extension Center, Texas A&M AgriLife, Temple, USA, 2014.

Givi, J. A. V. A. D., Prasher, S. O., and Patel, R. M.: Evaluation of pedotransfer functions in predicting the soil water contents at field capacity and wilting point. Agr. Water Manage., 70, 83-96, 2004.

Hargreaves, G. H., and Samani, Z. A.: Reference crop evapotranspiration from temperature, Applied Engr. Agric., 1, 96-99, 1985.

Hartkamp, A. D. Maize production environments revisited: a GIS-based approach. CIMMYT, 2001.

Iizumi, T., Yokozawa, M., and Nishimori, M.: Parameter estimation and uncertainty analysis of a large-scale crop model for paddy rice: Application of a Bayesian approach. Agr. Forest. Meteorol., 149, 333-348, 2009.

Izaurralde, R. C., Williams, J. R., McGill, W. B., Rosenberg, N. J., and Quiroga Jakas, M. C.: Simulating soil C dynamics with EPIC: model description and testing against long-term data, Ecol. Model., 192, 362-384, 2006.

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

Jägermeyr, J., Gerten, D., Heinke, J., Schaphoff, S., Kummu, M., and Lucht, W.:, Water savings potentials of irrigation systems: global simulation of processes and linkages, Hydrol. Earth Syst. Sci., 19, 3073-3091, 2015.

Jones, J. W., Antle, J. M., Basso, B., Boote, K. J., Conant, R. T., Foster, I., Godfray, H.C.J., Herrero, M., Howitt, R.E. Janssen, S., Keating, B.A., Munoz-Carpena, R., Porter, C.H., Rosenzweig, C., Wheeler, T.R.: Brief history of agricultural systems modeling. Agric. Syst., https://doi.org/10.1016/j.agsy.2016.05.014, 2016.

Kiniry, J. R., Major, D.J., Izaurralde, R.C., Williams, J.R., Gassman, P.W., Morrison, M., Bergentine, R., and Zentner, R.P.: EPIC model parameters for cereal, oilseed, and forage crops in the northern Great Plains region, Can. J. Plant Sci., 75, 679-688, 1995.

Liu, J., Williams, J. R., Zehnder, A. J. B., and Yang, H.: GEPIC - modelling wheat yield and crop water productivity with high resolution on a global scale, Agric. Syst., 94, 478-493, 2007.

Liu, J.: A GIS-based tool for modelling large-scale crop-water relations, Environ. Modell. Softw., 24, 411-422, 2009.

Liu, J., Folberth, C., Yang, H., Rockström, J., Abbaspour, K., and Zehnder, A. J. B.: A global and spatially explicit assessment of climate change impacts on crop production and consumptive water use, PLoS One, 8, e57750, 2013.

Liu, W., Yang, H., Folberth, C., Wang, X., Luo, Q., and Schulin, R.: Global investigation of impacts of PET methods on simulating crop-water relations for maize, Agr. Forest. Meteorol., 221, 164-175, 2016a.

Liu, W., Yang, H., Liu, J., Azevedo, L. B., Wang, X., Xu, Z., Abbaspour, K. C., and Schulin, R.: Global assessment of nitrogen losses and trade-offs with yields from major crop cultivations, Sci. Total Environ., 572, 526-537, 2016b.

Lychuk, T. E., Izaurralde, R. C., Hill, R. L., McGill, W. B., and Williams, J. R.: Biochar as a global change adaptation: predicting biochar impacts on crop productivity and soil quality with the EPIC model, Mitig. Adapt. Strateg. Global Change, 20, 1437-1458, 2015.

Martre, P., Wallach, D., Asseng, S., Ewert, F., Jones, J. W., Rötter, R. P., et al.: Multimodel ensembles of wheat growth: many models are better than one. Global Change Biol., 21, 911-925, 2015.

Matthews, R.B. and Pilbeam, C.: Modelling the long-term productivity and soil fertility of maize/millet cropping systems in the mid-hills of Nepal. Agr. Ecosyst. Environ., 111, 119–139, 2005.

McGill, W. B., Hunt, H. W., Woodmansee, R. G., and Reuss, J. O.: Phoenix, a model of the dynamics of carbon and nitrogen in grassland soils, in: Clark, F. E. and Rosswall, T.: Terrestrial Nitrogen Cycles. Processes, Ecosystem Strategies and Management Impacts, Ecological Bullettins, 49-115, Stockholm, 1981.

Monteith, J. L.: Evaporation and environment. Symp. Soc. Exp. Biol., 19, 1965.

Mueller, N. D., Gerber, J. S., Johnston, M., Ray, D. K., Ramankutty, N., and Foley, J. A.: Closing yield gaps through nutrient and water management, Nature, 490, 254-257, 2012.

Müller, C.: Agriculture: harvesting from uncertainties, Nat. Clim. Change, 1, 253-254, 2011.

Müller, C., Eickhout, B., Zaehle, S., Bondeau, A., Cramer, W., and Lucht, W.: Effects of changes in CO2, climate, and land use on the carbon balance of the land biosphere during the 21st century, J. Geophys. Res. Biogeosci., 112, G02032, 2007.

Müller, C., and Robertson, R.: Projecting future crop productivity for global economic modeling. Agr. Econ., 45, 37-50, 2014.

Müller, C., Elliott, J., Chryssanthacopoulos, J., Deryng, D., Folberth, C., Pugh, T. A., and Schmid, E.: Implications of climate mitigation for future agricultural production, Environ. Res. Lett., 10, 125004, 2015.

Müller, C., Elliott, J., Chryssanthacopoulos, J., Arneth, A., Balkovič, J., Ciais, P., Deryng, D., Folberth, C., Glotter, M., Hoek, S., Iizumi, T., Izaurralde, R. C., Jones, C., Khabarov, N., Lawrence, P., Liu, W., Olin, S., Pugh, T. A. M., Ray, D., Reddy, A., Rosenzweig, C., Ruane, A. C., Sakurai, G., Schmid, E., Skalský, R., Song, C. X., Wang, X., de Wit, A., and Yang, H.: Global Gridded Crop Model evaluation: benchmarking, skills, deficiencies and implications, Geosci. Model Dev., doi:10.5194/gmd-10-1403-2017, 2017.

Nelson, G. C., Valin, H., Sands, R. D., Havlík, P., Ahammad, H., Deryng, D., Elliott, J., Fujimori, S., Hasegawa, T., Heyhoe, E., Kyle, P., Von Lampe, M., Lotze-Campen, H., Mason d'Croz, D., van Meijl, H., van der Mensbrugghe, D., Müller, C., Popp, A., Robertson, R., Robinson, S., Schmid, E., Schmitz, C., Tabeau, A., and Willenbockel, D.: Climate change effects on agriculture: Economic responses to biophysical shocks, Proc. Natl. Acad. Sci. U.S.A., 111, 3274-3279, 2014.

Osborne, T., Rose, G., and Wheeler, T.: Variation in the global-scale impacts of climate change on crop productivity due to climate model uncertainty and adaptation. Agr. Forest Meteorol., 170, 183-194, 2013.

Parton, W.J., Ojima, D.S., Cole, C.V., Schimel, D.S.: A general model for soil organic matter dynamics: sensitivity to litter chemistry, texture and management, in: Quantitative Modeling of Soil Forming Processes, SSSA Special Publication 39, Soil Science Society of America, Madison, USA, 1994.

Peel, M.C., Finlayson, B.L., McMahon, T.A.: Updated world map of the Koeppen-Geiger climate classification. Hydrol. Earth Syst. Sci., 11, 1633-1644, 2007.

Porwollik, V., Müller, C., Elliott, J., Chryssanthacopoulos, J., Iizumi, T., Ray, D. K., Ruane, A. C., Arneth, A., Balkovič, J., Ciais, P., Deryng, D., Folberth, C., Izaurralde, R. C., Jones, C. D., Khabarov, N., Lawrence, P. J., Liu, W., Pugh, T. A.M., Reddy, A., Sakurai, G., Schmid, E., Wang, X., de Wit, A., and Wu, X.: Spatial and temporal uncertainty of crop yield aggregations, Eur. J. Agron., doi: 10.1016/j.eja.2016.08.006, 2016.

Portmann, F. T., Siebert, S., and Döll, P.: MIRCA2000 - Global monthly irrigated and rainfed crop areas around the year 2000: A new high-resolution dataset for agricultural and hydrological modeling, Global Biogeochem. Cy., 24, GB 1011, doi:10.1029/2008GB003435, 2010.

Pugh, T.A.M., Arneth, A., Olin, S., Ahlström, A., Bayer, A.D., Goldewijk, K.K., Lindeskog, M., and Schurgers, G.: Simulated carbon emissions from land-use change are substantially enhanced by accounting for agricultural management, Environ. Res. Lett., 10, 124008, 2015.

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

Rawls, W. J., and Brakensiek, D. L.: Prediction of soil water properties for hydrologic modeling. Watershed Management in the Eighties, ASCE, 1985.

Ray, D. K., Ramankutty, N., Mueller, N. D., West, P. C., and Foley, J. A.: Recent patterns of crop yield growth and stagnation, Nat. Comm., 3, 1293, 2012.

Rosenzweig, C., Jones, J. W., Hatfield, J. L., Ruane, A. C., Boote, K. J., Thorburn, P., Antle, J. M., Nelson, G. C., Porter, C., Janssen, S., Asseng, S., Basso, B., Ewert, F., Wallach, D., Baigorria, G., and Winter, J. M.: The Agricultural Model Intercomparison and Improvement Project (AgMIP): Protocols and pilot studies, Agr. Forest Meteorol., 170, 166-182, 2013.

Rosenzweig, C., Elliott, J., Deryng, D., Ruane, A. C., Arneth, A., Boote, K. J., Folberth, C., Glotter, M., Khabarov, N., Müller, C., Neumann, K., Piontek, F., Pugh, T., Schmid, E., Stehfest, E., and Jones, J. W.: Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison, Proc. Natl. Acad. Sci. U.S.A., 111, 3268-3273, 2014.

Sacks, W.J., Deryng, D., Foley, J.A., and Ramankutty, N.: Crop planting dates: An analysis of global patterns, Global Ecol. Biogeogr., 19, 607-620, 2010.

Sándor, R., Ehrhardt, F., Basso, B., Bellocchi, G., Bhatia, A., Brilli, L., et al.: C and N models Intercomparison–benchmark and ensemble model estimates for grassland production. Adv. Anim. Biosci., 7, 245-247, 2016.

Schaap, M. G. and Bouten, W.: Modelling water retention curves of sandy soils using neural networks, Water Resour. Res., 32, 3033-3040, 1996.

Schauberger, B., Archontoulis, S., Arneth, A., Balkovic, J., Ciais, P., Deryng, D., et al.: Consistent negative response of US crops to high temperatures in observations and crop models. Nature Comm., 8, 13931, 2017.

Schneider, U. A., Havlík, P., Schmid, E., Valin, H., Mosnier, A., Obersteiner, M., Böttcher, H., Skalský, R., Balkovič, J., Sauer, T., and Fritz, S.: Impacts of population growth, economic development, and technical change on global food production and consumption, Agric. Syst., 104, 204–215, 2011.

Schneider, U., Becker, A., Finger, P., Meyer-Christoffer, A., Zuise, M., and Rudolf, B.: GPCC's new land surface precipitation climatology based on quality - controlled in situ data and its role in quantifying the global water cycle, Theor. Appl. Climatol., 115, 15-40, 2013.

Skalský, R., Tarasovičová, Z., Balkovič, J., Schmid, E., Fuchs, M., Moltchanova, E., Kindermann, G., and Scholtz, P.: Geo-bene global database for biophysical modelling v. 1.0. Concepts, methodologies and data, 2008.

Srivastava, A. K., and Gaiser, T.: Simulating biomass accumulation and yield of yam (Dioscorea alata) in the Upper Ouémé Basin (Benin Republic)-I. Compilation of physiological parameters and calibration at the field scale. Field Crops Res., 116, 23-29, 2010.

Stehfest, E., Heistermann, M., Priess, J. A., Ojima, D. S., and Alcamo, J.: Simulation of global crop production with the ecosystem model DayCent, Ecol. Model., 209, 203-219, 2007.

Stockle, C. O., Williams, J. R., Jones, C. A., and Rosenberg, N. J.: A method for estimating the direct and climatic effects of rising atmospheric carbon dioxide on growth and yield of crops. I. Modification of the EPIC model for climate change analysis, Agric. Syst., 38, 225-238, 1992.

Stolbovoy, V., Montanarella, L., and Panagos, P. (eds.): Carbon Sink Enhancement in Soils of Europe: Data, Modeling, Verification, JRC technical and scientific reports, Office for Official Publications of the European Communities, Luxembourg, 2007.

Tan, G., and Shibasaki, R.: Global estimation of crop productivity and the impacts of global warming by GIS and EPIC integration. Ecol. Modell., 168, 357-370, 2003.

Vanclooster, M., Viaene, P., Diels, J., and Feyen, J.: A deterministic evaluation analysis applied to an integrated soil-crop model. Ecol. Modell., 81, 183-195, 1995.

Valade, A., Ciais, P., Vuichard, N., Viovy, N., Caubel, A., Huth, N., Marin, F. and Martiné, J. F.: Modeling sugarcane yield with a process-based model from site to continental scale: uncertainties arising from model structure and parameter values. Geosci. Model Dev., 7, 1225-1245, 2014.

Waha, K., van Bussel, L. G. J., Müller, C., and Bondeau, A.: Climate-driven simulation of global crop sowing dates, Global Ecol. Biogeogr., 21, 247-259, 2012.

Wang, J., Li, X., Lu, L., and Fang, F.: Parameter sensitivity analysis of crop growth models based on the extended Fourier Amplitude Sensitivity Test method, Environ. Modell. Softw, 48, 171-182, 2013a.

Wang, J., Li, X., Lu, L., and Fang, F.: Estimating near future regional corn yields by integrating multi-source observations into a crop growth model. Eur. J. Agron., 49, 126-140, 2013b.

Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M., and Venables, B.: gplots: Various R Programming Tools for Plotting Data. R package version 3.0.1, 2016.

Warszawski, L., Frieler, K., Huber, V., Piontek, F., Serdeczny, O., and Schewe, J.: The Inter-Sectoral Impact Model Intercomparison Project (ISI-MIP): Project framework, Proc. Natl. Acad. Sci. U.S.A., 111, 3228-3232, 2013.

Weedon, G. P., Balsamo, G., Bellouin, N., Gomes, S., Best, M. J., and Viterbo, P.: The WFDEI meteorological forcing data set: WATCH Forcing Data methodology applied to ERA-Interim reanalysis data, Water Resour. Res., 50, 7505-7514, 2014.

Wei, T. Simko, V.: corrplot: Visualization of a Correlation Matrix. R package version 0.77, 2016.

Wheeler, T, and von Braun, J.: Climate change impacts on global food security, Science, 341, 508-513, 2013.

Wickham, H.: ggplot2: Elegant Graphics for Data Analysis, Springer, New York, 2009.

Williams, J. R., Jones, C. A., Kiniry, J. R., and Spanel, D. A.: The EPIC crop growth model, Trans. ASAE, 32, 497-511, 1989.

Wösten, J. H. M. and Van Genuchten, M. T.: Using texture and other soil properties to predict the unsaturated soil hydraulic functions, Soil. Sci. Soc. Am. J., 52, 1762-1770, 1988.

Xiong, W., Holman, I., Conway, D., Lin, E., and Li, Y.: A crop model cross calibration for use in regional climate impacts studies. Ecol. Modell., 213, 365-380, 2008.

Xiong, W., Balkovič, J., van der Velde, M., Zhang, X., Izaurralde, R. C., Skalský, R., Lin, E., Mueller, N., and Obersteiner, M.: A calibration procedure to improve global rice yield simulations with EPIC, Ecol. Modell., 273, 128-139, 2014a.

Xiong, W., van der Velde, M., Holman, I. P., Balkovič, J., Lin, E., Skalský, R., Porter, C., Jones, J., Khabarov, N., and Obersteiner, M.: Can climate-smart agriculture reverse the recent slowing of rice yield growth in China?, Agric. Ecosyst. Environ., 196, 125-136, 2014b.

Xiong, W., Skalský, R., Porter, C. H., Balkovič, J., Jones, J. W., and Yang, D.: Calibration-induced uncertainty of the EPIC model to estimate climate change impact on global maize yield, J. Adv. Model. Earth Syst., 8, 1358–1375, 2016.

Zhang, X., Izaurralde, R. C., Manowitz, D. H., Sahajpal, R., West, T. O., Thomson, A. M., Xu, M., Zhao, K., LeDuc, S. D., and Williams, J. R.: Regional scale cropland carbon budgets: evaluating a geospatial agricultural modeling system using inventory data, Environ. Modell. Softw., 63, 199-216, 2015.

Zhao, G., Bryan, B. A., and Song, X.: Sensitivity and uncertainty analysis of the APSIM-wheat model: Interactions between cultivar, environmental, and management parameters. Ecol. Modell., 279, 1-11, 2014.

Figures and tables

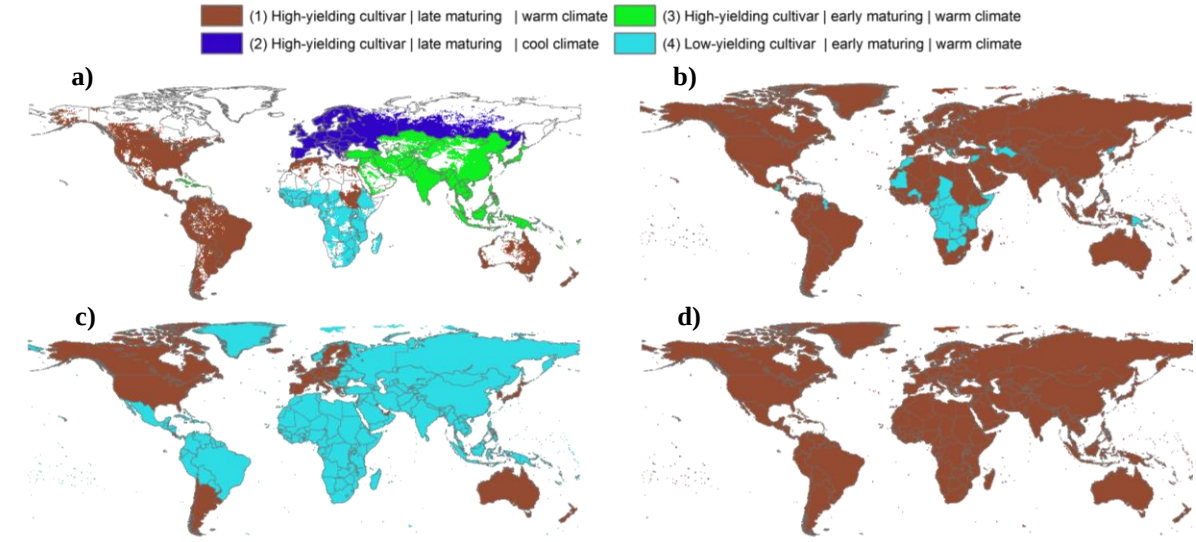


Figure 1: Distributions of maize cultivars in the model frameworks for (a) EPIC-IIASA, (b) EPIC-TAMU, (c) GEPIC and PEPIC, and (d) EPIC-BOKU. Differences in the parameterization of each cultivar are provided in Table S1-4. Numbers in braces (1-4) are used in the text to refer to the cultivars.

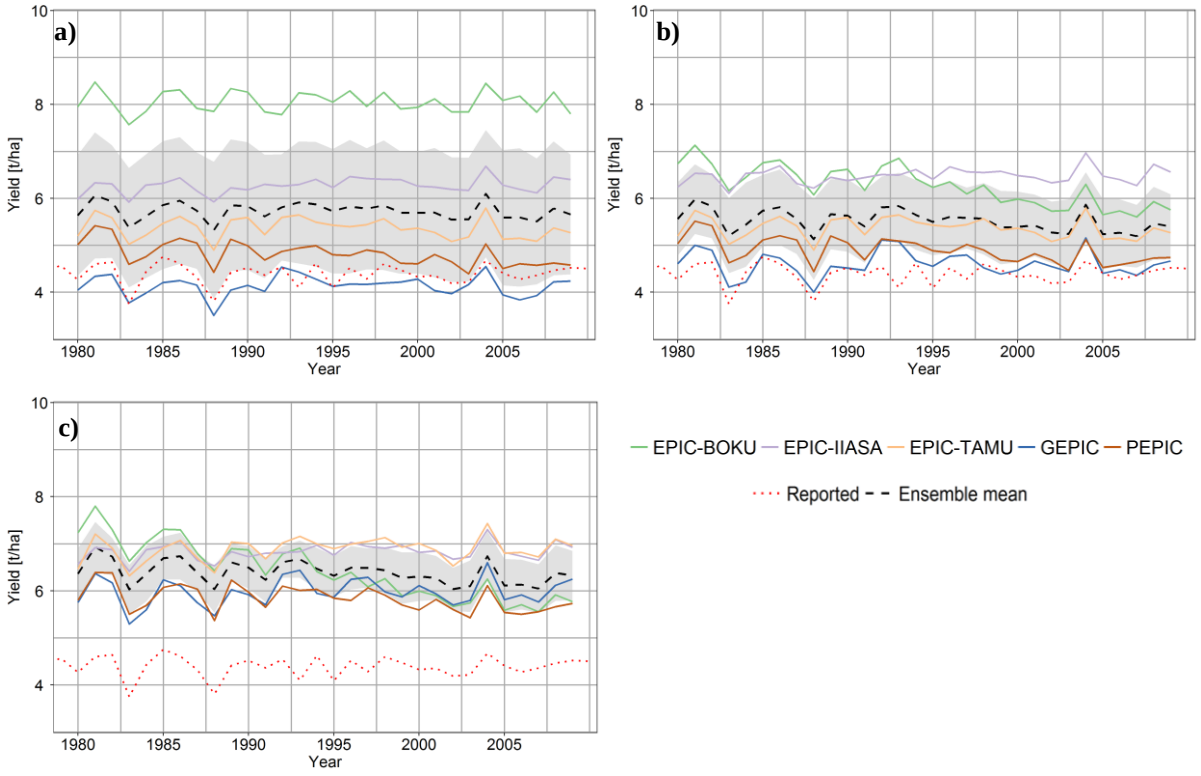


Figure 2: Global average area-weighted maize yield estimates of five EPIC-based GGCMs for the (a) default, (b) fully harmonized (fullharm), and (c) fully harmonized scenario with sufficient nutrient supply (harm-suffN) management scenario (Table 1). Reported yields are based on FAOSTAT (FAO, 2014) and have been detrended (see Sect. 2.8). The black dashed line represents the ensemble mean. The grey ribbon shows the 95% confidence interval of the ensemble mean. For EPIC-TAMU, outputs from the fully harmonized (fullharm) simulations were used as a substitute for missing default outputs to keep the number of EPIC-based GGCMs across management scenarios constant. Table S3-2 shows statistical coefficients of yield trends over time and ME relative to FAO reported yields. Corresponding linear regressions are displayed in Figure S3-1.

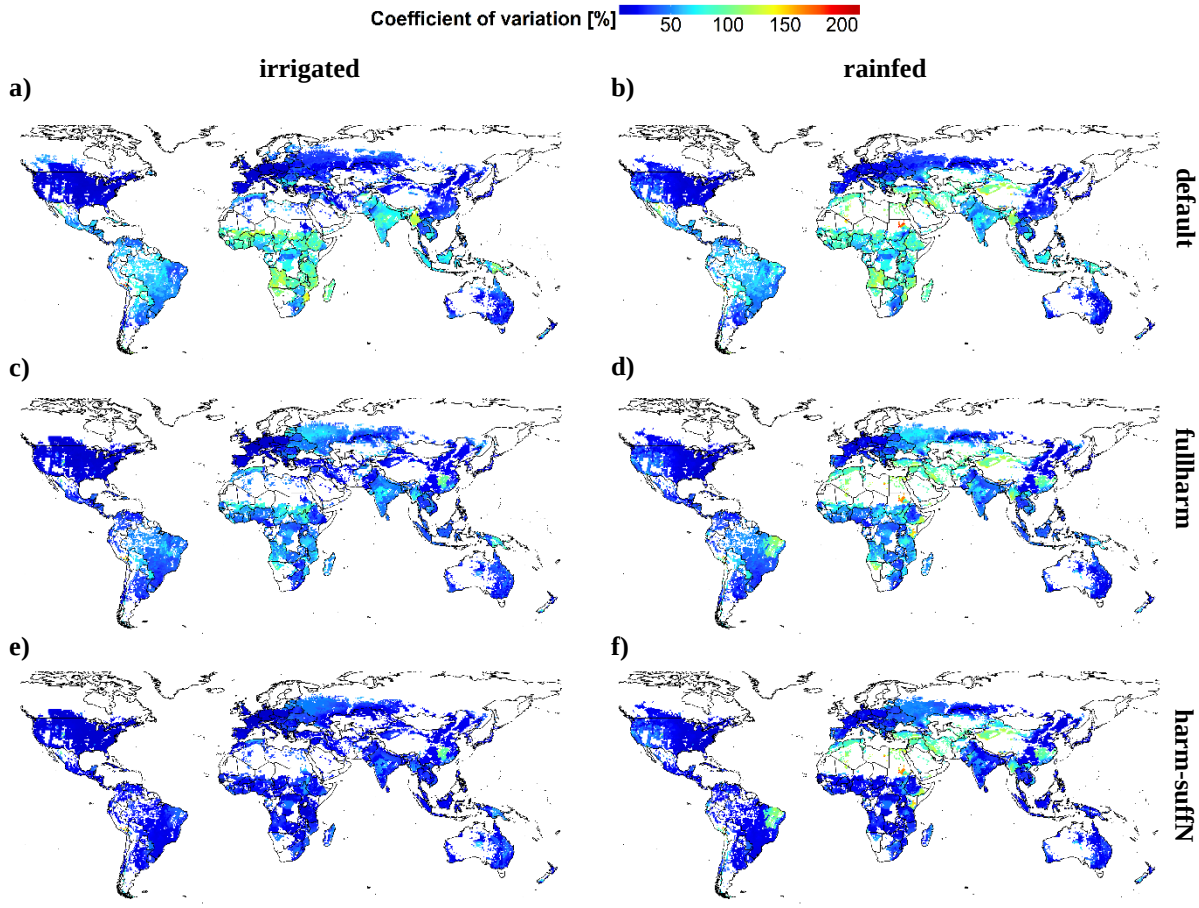


Figure 3: Coefficient of variation for long-term average maize yield estimates (CV_{av}) among EPIC-based GGCMs for each of the six crop management scenarios defined in Table 1. For EPIC-TAMU, outputs from the fully harmonized (fullharm) simulations were used as a substitute for missing default outputs to keep the number of models constant across management scenarios.

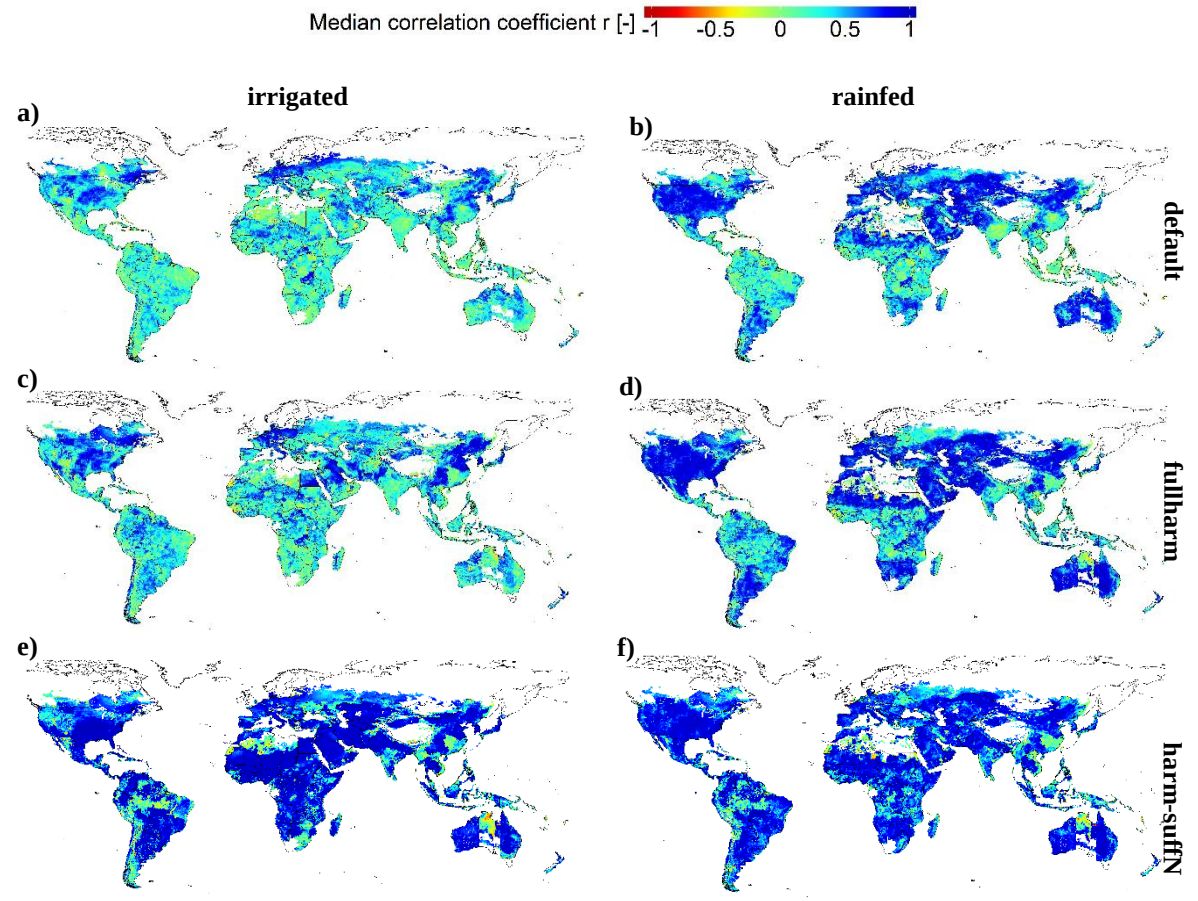


Figure 4: Median of time-series correlation coefficient r for maize yield estimates among EPIC-based GGCMs for each of the six crop management scenarios defined in Table 1. For EPIC-TAMU, outputs from the fully harmonized (fullharm) simulations were used also as default to keep the number of models constant across management scenarios. Fractions of grid cells per significance level are provided in Table S3-4.

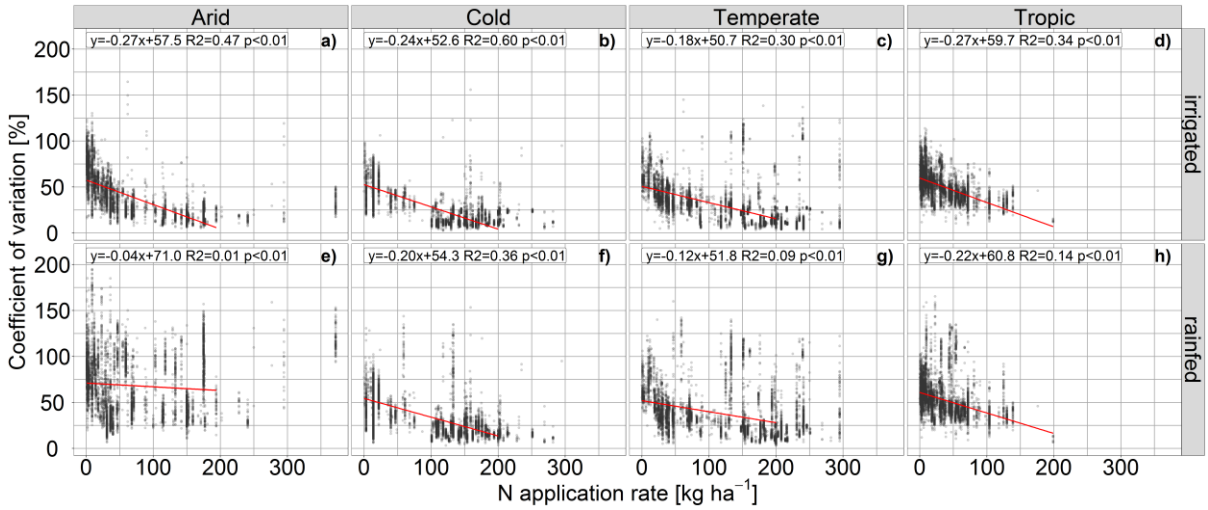


Figure 5: Coefficient of variation for maize yields among EPIC-based GGCMs compared to fertilizer application rates in the fully harmonized management scenario (fullharm) with sufficiently irrigated (a-d) or rainfed (e-h) water supply in each grid cell of four major climate regions. Linear regressions are limited to ≤200 kg N ha^{-1} which commonly corresponds to sufficient N supply (e.g. Folberth et al., 2013).

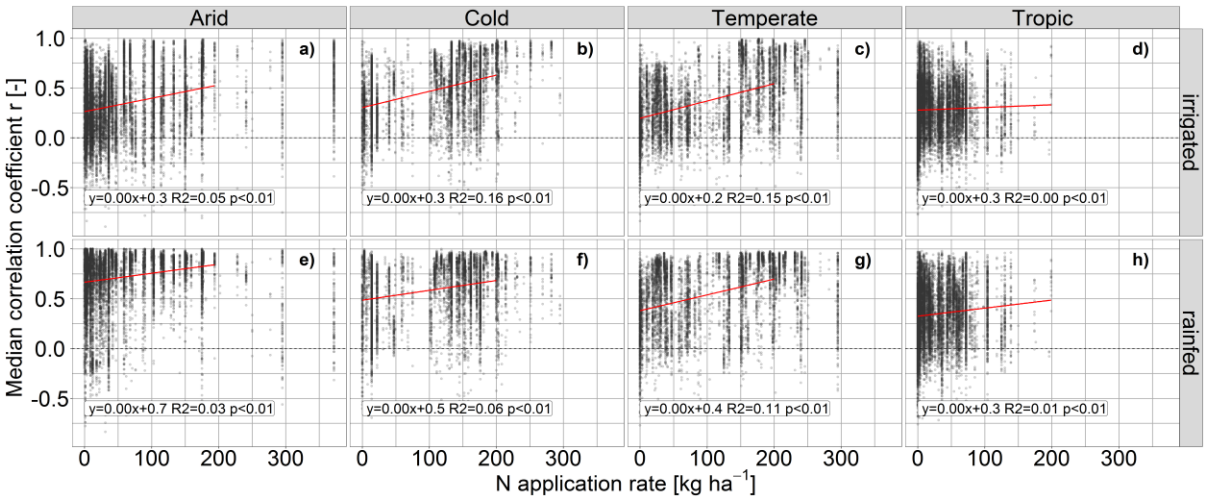


Figure 6: Median time-series correlation coefficient r for maize yields among EPIC-based GGCMs compared to fertilizer application rates in the fully harmonized management scenario (fullharm) with sufficiently irrigated (a-d) or rainfed (e-h) water supply in each grid cell of four major climate regions. Linear regressions are limited to ≤200 kg N ha^{-1} which commonly corresponds to sufficient N supply (e.g. Folberth et al., 2013).

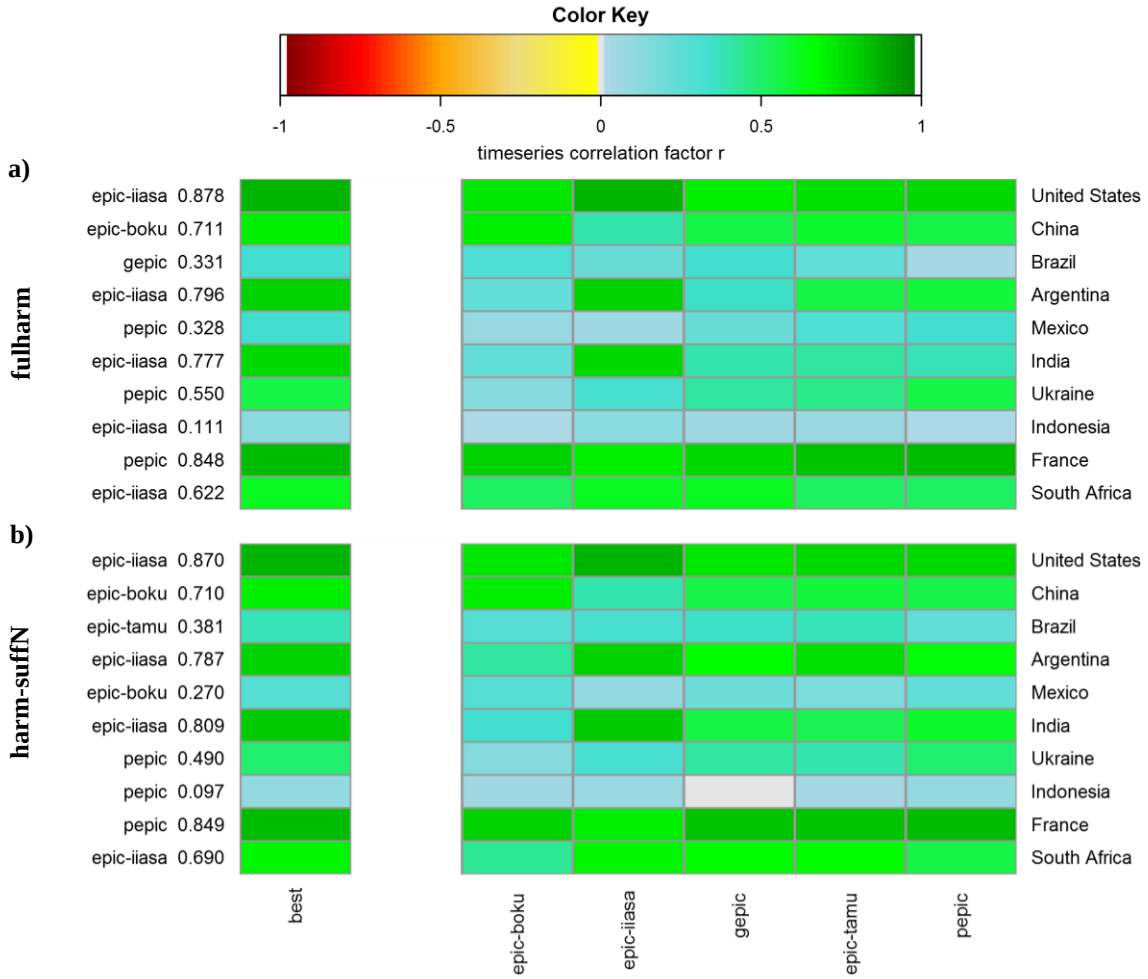
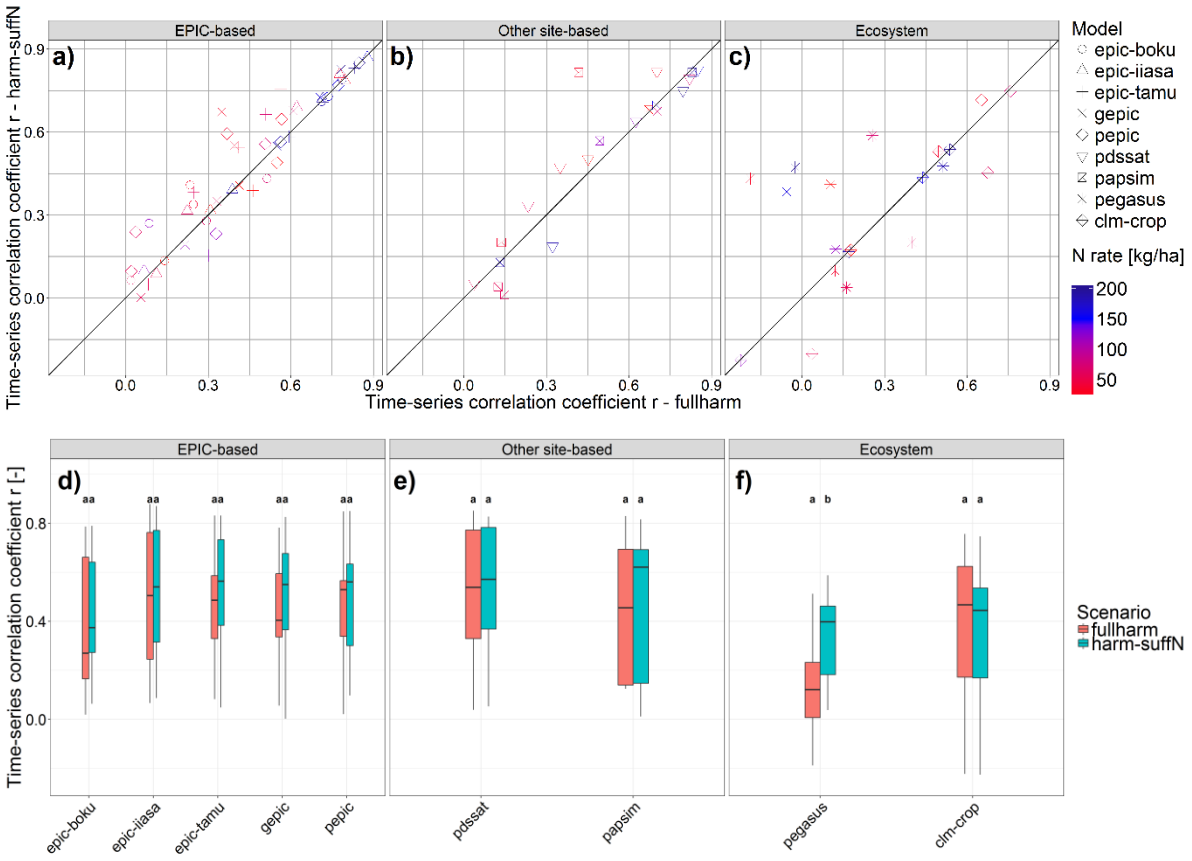


Figure 7: Time-series correlation coefficients against reported detrended yields for EPIC-based GGCMs (x-axis) in the top ten maize producing countries (right y-axis) for (a) the fullharm and (b) the harm-suffN simulations. The best performing GGCM including r value is displayed on the left y-axis. Correlation coefficients for each GGCM and country are provided in Table S3-5.



1180 **Figure 8: Differences in model performance between the fullharm and harm-suffN scenarios per GGCM and GGCM type (Table S1-1) for the ten major maize producing countries (see Figure 7). (a)-(c): Direct comparison of time-series correlation coefficients with different shapes for each model and color scale representing national N application rates. (d)-(f): Boxplots for each GGCM and setup scenario spanning the ten correlation coefficients in each scenario. Letters a/b above boxplots indicate a significant difference between scenarios based on ANOVA/Tukey's HSD test.**

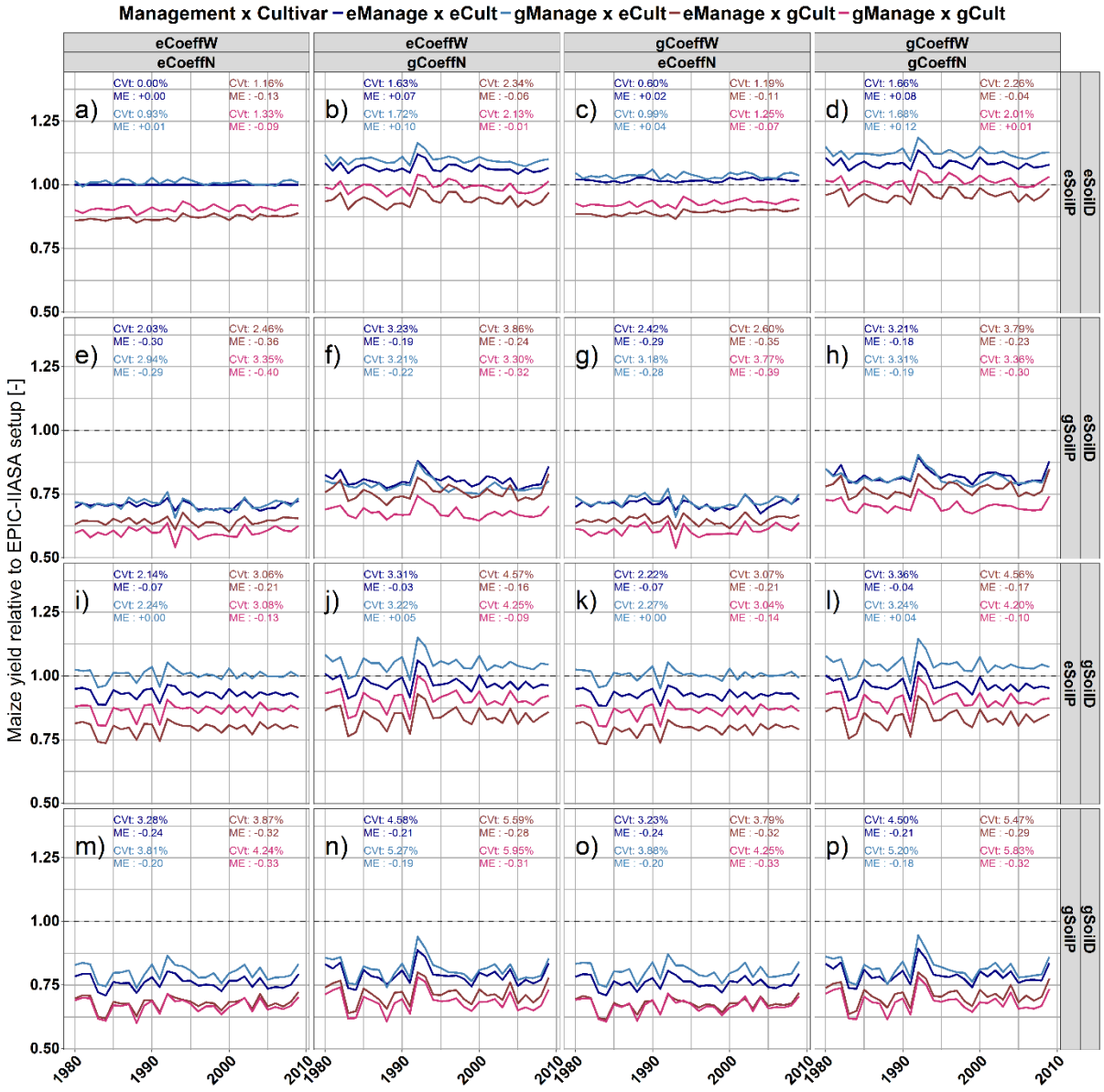


Figure 9: Relative difference in global average rainfed maize yields over a 29 year time period for 64 setup combinations based on the EPIC-IIASA and GEPIC setups (Table 3) compared to the complete EPIC-IIASA setup. e=EPIC-IIASA, g=GEPIC, Cult=cultivar definition and distribution, SoilD=soil data, SoilP=spin-up and soil handling, CoeffN=organic matter and nutrient cycling coefficients, CoeffW=hydrologic coefficients, Manage=crop management. CV_t=coefficient of variation over time normalized to mean=1. ME=mean error compared to the full EPIC-IIASA setup.

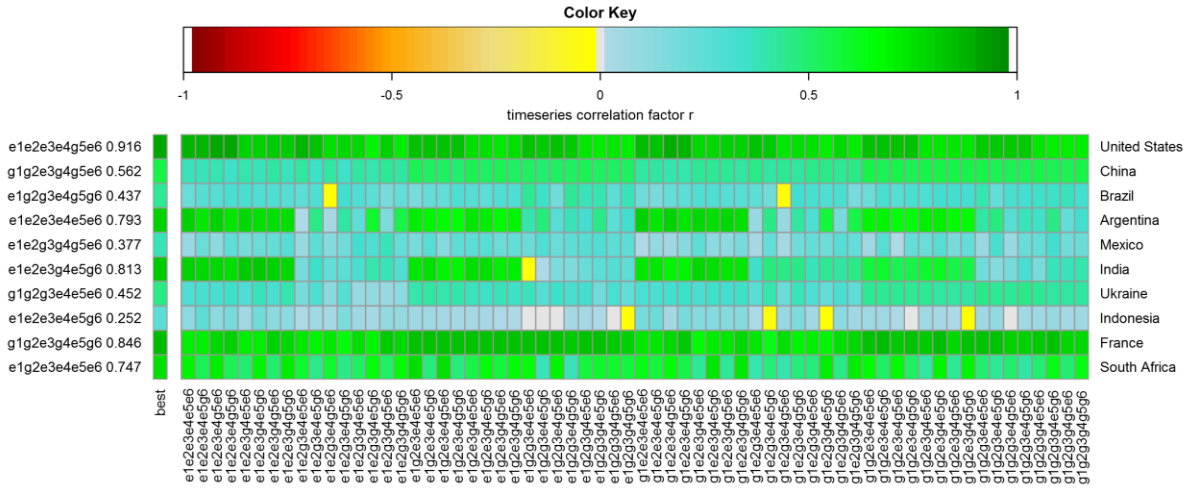


Figure 10: Time-series correlation coefficients for all EPIC-IIASA / GEPIC setup combinations with the fullharm scenario (x-axis) in the top ten maize producing countries (right y-axis) and the best performing setup including r value (left y-axis).

Table 1: Crop management scenarios based on Elliott et al. (2015). The default setup represents each modelling group's own assumptions and input data. The harmonized scenarios use the same growing season data and the same annual application rates for N and P (fullharm) or sufficient nutrient supply (harm-suffN) to avoid nutrient-related plant growth limitations. See Figure S2-2a,b for maps of harmonized N and P application rates.

Name	Abbreviation	Irrigation vol. [mm]	N [kg ha^{-1}]	P [kg ha^{-1}]	Growing season dates
Default, irrigated	default	sufficient	individual[1)]	individual[1)]	individual[1)]
Default, rainfed		-	individual[1)]	individual[1)]	individual[1)]
Fully harmonized, irrigated	fullharm	sufficient	harmon.[2)]	harmon.[2)]	harmon.[3)]
Fully harmonized, rainfed		-	harmon.[2)]	harmon.[2)]	harmon.[3)]
Harmonized & suff. nutrients, irrig.	harm-suffN	sufficient	sufficient	sufficient	harmon.[3)]
Harmonized & suff. nutrients, rainfed		-	sufficient	sufficient	harmon.[3)]

[1)] Based on each research group's assumptions and data

[2)] Harmonized fertilizer application rates based on Mueller et al. (2010) processed as described in Elliott et al. (2015)

[3)] Harmonized growing season data based on Sacks et al. (2010) with gap filling as described in Elliott et al. (2015)

Table 2: Differences in parameters and choice of subroutines for the participating EPIC-based GGCMs. A dash indicates that the parameter is not relevant for the respective model due to selection of subroutines. A brief explanation of parameters is provided in Table S1-2.

	No	Parameter	EPIC-BOKU	EPIC-IIASA	EPIC-TAMU	GEPIC	PEPIC
Hydrology	1	PET estimation method[1)]	PM	HG	PM	HG	PM
	2	Hargreaves exp. coefficient	-	0.6	-	0.5	-
	3	Hargreaves linear coefficient	-	0.0023	-	0.0032	-
	4	Soil evaporation-cover coefficient	0	0	0.15	0	0
	5	Soil cover-temperature function[2)]	1,30 8,95	1,30 8,95	1,05 3,95	1,30 8,95	1,30 8,95
	6	Soil evaporation coefficient	2.5	1.5	2.5	2.5	1.5
	7	Soil evaporation-depth function[2)]	10,50 100,95	10,50 100,95	10,70 100,95	10,50 100,95	10,50 100,95
	8	Plant water use-soil water tension function[2)]	100,01 1000,90	100,01 1000,90	500,01 1500,50	100,01 1000,90	100,01 1000,90
	9	FC, WP, and K_{sat} estimation[3)]	Rawls	static	Rawls	Rawls	Rawls
	10	Soil variable dependence of CN[4)]	SMI	depth	depth	SMI	SMI
	11	CN number index coefficient	1.5	1.2	1	0.5	1
	12	CN coefficient for standing dead residue	0.0	0.0	0.3	0.2	0.0
Soil degradation	13	Wind erosion considered[5)]	no	no	yes	yes	yes
	14	Water erosion considered [5)]	no	no	no	yes	yes
	15	Water erosion conservation practice[6)]	-	-	-	0.5	1.0
	16	Water erosion estimation method[7)]	-	-	-	MUSS	RUSL2
	17	Field length for wind erosion	-	2.00	1.00	1.24	2.00
	18	Field width for wind erosion	-	2.00	1.00	0.62	2.00
	19	Soil profile handling (static/dynamic)[8)]	stat.	stat.	dyn.	dyn.	dyn.
	20	Simulation continuity (transient/decadal)[9)]	trans.	trans.	trans.	dec.	trans.
OM / nutrient cycling	21	Denitrification method[10)]	EPIC	CI	AK	AK	AK
	22	Microbial decay rate	1.0	0.8	1.0	1.0	1.0
	23	Slow to passive humus coefficient	0.05	0.05	0.003	0.05	0.05
	24	Oxygen content-soil depth function[2)]	200,05 500,90	400,05 600,90	200,05 500,90	200,05 500,90	200,05 500,90
	25	Oxygen coefficient for microbial activity	0.90	0.99	0.80	0.90	0.90
	26	N volatilization coefficient	0.005	0.700	0.030	0.005	0.300
Manage-ment	27	Automatic irrigation trigger	0.90	0.80	0.99	0.90	0.90
	28	Maximum single water application [mm]	50	500	100	1000	500
	29	Automatic fertilizer application trigger[11)]	0.90	0.80	0.99	0.90	-
Growth	30	Coefficient allocating root growth	0.5	0.5	0.7	0.5	0.5
	31	Coefficient for root growth dist. by depth	10	10	7	10	10
	32	Root growth stress considered	no	no	yes	no	no
	33	Fraction of growing season from which HI_{min} affects yield formation	0.50	0.50	0.45	0.50	0.50

[1)] PM: Penman-Monteith; HG: Hargreaves

[2)] Parameters 5,7,8, and 24 are X and Y values (separated by commas) for two points (upper and lower pairs) defining the shape of sigmoid functions

[3)] Field capacity (FC) and wilting point (WP) can be estimated by 11 different methods or be an input in soil files. Saturated hydraulic conductivity (K_{sat}) can be estimated according to Rawls method or be input. For EPIC-IIASA these parameters were estimated based on the ROSETTA model as described in S1.2.2.

[4)] Describes the dependence of curve number (CN) estimation on soil moisture, which can be based on five methods, among them soil moisture gradient with profile depth or calculation of a daily soil moisture index (SMI)

[5)] Water and wind erosion can be turned on or off and water erosion is estimated by different methods (see below)

[6)] Water erosion rates are lowered by the given fraction (0 corresponds to virtually eliminated water erosion, 1 to no erosion control)

[7)] MUSS: Modified Universal Soil Loss Equation for Small Watersheds; RUSL2: Modified Revised Universal Soil Loss Equation

[8)] Static: annual re-initialization of soil profile, except water content and mineral nutrients; dynamic: transient updating of soil parameters throughout simulation

[9)] The GEPIC model is run separately for each decade as described in SI 1.3.4

[10)] EPIC: original EPIC method (Williams et al., 1989); CI: Cesar Izaurralde method (Izaurralde et al., 2012); AK; Armen Kemanian method (no reference)

[11)] The auto-fertilizer and irrigation triggers define at which stress level fertilizer or water are being applied. E.g., a value of 0.8 for the auto-fertilizer trigger implies that fertilizer is applied on a given day if potential biomass production would be limited by >20%.

Table 3: Composition of aggregated setup domains the comparison of GEPIC in EPIC-IIASA in the fully harmonized (fullharm) scenario (Table 1). Numbers in the first column are used in selected figures to keep annotation short, otherwise the abbreviation is used. Numbers in column "Parameters considered" refer to those in Table 2. When referencing the setup domain parameterizations from each model, e=EPIC-IIASA and g=GEPIC (e.g. eCult refers to cultivar setup of EPIC-IIASA)

No	Setup domain and abbreviation	Parameters considered	Effect in model
1	**Cultivars (Cult)**	• see Figure 1 for distribution of cultivars and Table S1-4 for differences in parameterization	• scaling of yields based on potential HI_{max} • higher sensitivity to water stress with lower HI_{min} • temperature ranges for optimal crop growth
2	**Soil data (SoilD)**	• Table 2: 9 • differences in hydrologic soil group definitions • sum of bases and saturated conductivity only in EPIC-IIASA soil files • ten soil layers in EPIC-IIASA • five soil layers in GEPIC	• soil hydrology • nutrient cycling • little difference in basic soil properties (see Figure S1-3)
3	**Soil handling (SoilP)**	• Table 2: 13-20 • decadal runs with dynamic soil handling in gSoilP • fully transient runs with static soil profile in eSoilP	• carry-over effects in transient runs but re-initialization of soil texture, depth and OM for EPIC-IIASA setup • carry-over effects for all soil variables including losses from erosion transient for each decade with 30yr spin-up (see SI 1.3.4)
4	**Parameterization of organic matter and nutrient cycling (CoeffN)**	• Table 2: 21-26	• nutrient fate and availability • e.g. denitrification, microbial mineralization, partitioning to OM pools
5	**Parameterization of hydrologic processes (CoeffW)**	• Table 2: 2, 3, 6, 10, 11, 12	• PET estimation • runoff and percolation • plant water deficit • OM and nutrient cycling
6	**Crop management (Manage)**	• Table 2: 27-29 • a list of crop management operations in both models is provided in Table S1-3	• short- and long-term nutrient availability • surface roughness and soil erodibility • potential biomass estimation

Table 4: Numbers of countries (out of 99 for which benchmark data and model outputs are available) in each harmonized setup scenario, in which each EPIC-based GGCMs has the highest (column "best") performance compared against reported yields within the EPIC ensemble and all countries (column "all") in which the correlation coefficient is significant at p<0.1 and positive.

Scenario	fullharm		harm-suffN	
Model	best	all	best	all
EPIC-BOKU	20	56	18	59
EPIC-IIASA	26	56	23	60
GEPIC	15	50	19	58
EPIC-TAMU	23	48	20	61
PEPIC	15	48	19	52