

LETTER • **OPEN ACCESS**

Predicting spatial and temporal variability in crop yields: an inter-comparison of machine learning, regression and process-based models

To cite this article: Guoyong Leng and Jim W Hall 2020 *Environ. Res. Lett.* **15** 044027

View the [article online](#) for updates and enhancements.

Environmental Research Letters



LETTER

OPEN ACCESS

RECEIVED
15 December 2019

REVISED
5 February 2020

ACCEPTED FOR PUBLICATION
28 February 2020

PUBLISHED
20 April 2020

Original content from
this work may be used
under the terms of the
[Creative Commons
Attribution 4.0 licence](#).

Any further distribution
of this work must
maintain attribution to
the author(s) and the title
of the work, journal
citation and DOI.



Predicting spatial and temporal variability in crop yields: an inter-comparison of machine learning, regression and process-based models

Guoyong Leng^{1,2} and Jim W Hall²

¹ Key Laboratory of Water Cycle and Related Land Surface Processes, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, People's Republic of China

² Environmental Change Institute, University of Oxford, Oxford OX1 3QY, United Kingdom

E-mail: lenggy@igsnr.ac.cn

Keywords: climate change, crop yield, machine learning, statistical model, crop model

Supplementary material for this article is available [online](#)

Abstract

Pervious assessments of crop yield response to climate change are mainly aided with either process-based models or statistical models, with a focus on predicting the changes in average yields, whilst there is growing interest in yield variability and extremes. In this study, we simulate US maize yield using process-based models, traditional regression model and a machine-learning algorithm, and importantly, identify the weakness and strength of each method in simulating the average, variability and extremes of maize yield across the country. We show that both regression and machine learning models can well reproduce the observed pattern of yield averages, while large bias is found for process-based crop models even fed with harmonized parameters. As for the probability distribution of yields, machine learning shows the best skill, followed by regression model and process-based models. For the country as a whole, machine learning can explain 93% of observed yield variability, followed by regression model (51%) and process-based models (42%). Based on the improved capability of the machine learning algorithm, we estimate that US maize yield is projected to decrease by 13.5% under the 2 °C global warming scenario (by ~2050 s). Yields less than or equal to the 10th percentile in the yield distribution for the baseline period are predicted to occur in 19% and 25% of years in 1.5 °C (by ~2040 s) and 2 °C global warming scenarios, with potentially significant implications for food supply, prices and trade. The machine learning and regression methods are computationally much more efficient than process-based models, making it feasible to do probabilistic risk analysis of climate impacts on crop production for a wide range of future scenarios.

1. Introduction

Understanding climate impacts on crop yields is fundamental for enhancing agricultural resilience in a changing climate. Numerous studies have been conducted for assessing the impacts of climate change on crop yield at the regional and global scales (Olesen and Bindi 2002, Jones and Thornton 2003, Parry *et al* 2004, Tao *et al* 2006, Yao *et al* 2007, Schlenker and Roberts 2009, Reidsma *et al* 2010, Schlenker and Lobell 2010, Bindi and Olesen 2011, Müller *et al* 2011, Urban *et al* 2012, Rosenzweig *et al* 2014). In general, high temperature exerts negative impact on yields directly through heat stress and indirectly via

soil moisture deficits, while low precipitation tends to induce closure of stoma, reduction of carbon uptake and decrease of yields (Peng *et al* 2004, Schlenker and Roberts 2009, Butler and Huybers 2013, Asseng *et al* 2015; Ray *et al* 2015, Liu *et al* 2016, Zhao *et al* 2016, Schauburger *et al* 2017, Wang *et al* 2017).

Despite tremendous efforts to quantify yield response to climate variation, substantial uncertainties still exist in the estimation of climate impacts on crop yield (Asseng *et al* 2013, Wheeler and von Braun 2013, Challinor *et al* 2014b, Wang *et al* 2017), which can be attributed to the complexity of the relevant processes and the scarcity of relevant datasets. In general, two types of models have been widely

used in the literature: statistical models (Lobell and Field 2007, Tebaldi and Lobell 2008, Zampieri *et al* 2017, Gaupp *et al* 2019, Leng and Hall 2019) and process-based models (Rosenzweig *et al* 2014, Waha *et al* 2015, Deryng *et al* 2016, Müller *et al* 2017). Whilst each approach has its own strengths and weakness, only a few studies have compared the estimates between statistical models and crop models (Leng 2017a, 2017b; Lobell and Asseng 2017, Roberts *et al* 2017).

For the most part, the focus of these studies, and inter-comparison exercises such as the Agricultural Modelling Inter-comparison and Improvement Project (AgMIP); (Rosenzweig *et al* 2013) has been on understanding the potential future changes in average agricultural yields given future climate scenarios. There has been much less attention to the impacts of climate variation and extremes, though these effects can be particularly significant for global food systems if they result in large reductions in yield, interruptions to supply and price increases (Gaupp *et al* 2019). Empirically characterizing the variability in crop yields is more challenging than characterizing general trends, in particular because extreme events are by definition rare. Furthermore, using crop models to characterize variability in time involves many computationally expensive simulations, whilst simulating variation in space requires high resolution datasets that may not be available.

Some more studies have recently examined the effects of droughts on agricultural production (Leng and Hall 2019) and the impacts of concurrent droughts in different locations (Gaupp *et al* 2019). Although these studies have addressed large food producing regions, they have not sought to characterize both the spatial and temporal response of yields to climatic variability. Better resolution of yield variations in space and time is needed in order to understand the risks of climatic variability for agricultural productions. Here, our hypothesis is that neither established crop models or conventional regression models are well-suited to this task. Crop models have been mainly used for understanding and simulating trends in crop yields and agricultural production (Elliott *et al* 2014, Rosenzweig *et al* 2014, Zhao *et al* 2017), in order to inform policy decisions, whilst regression models are traditionally linear, which may not be appropriate given the complex response of crop yields to external perturbations.

Recently, machine-learning (ML) has emerged as a powerful tool for environmental analysis (Chlingaryan *et al* 2018), as well as climate impact assessment on crop yield (Jeong *et al* 2016, Johnson *et al* 2016, Feng *et al* 2018, 2019, Cai *et al* 2019, Vogel *et al* 2019). ML often shows better performance compared to conventional linear regression models (Feng *et al* 2018), as it can capture non-linear relationships, handle the interactions among predictors and do not assume a certain shape of response function

(e.g. linear or polynomial) (Breiman 2001, Shalev-Shwartz and Ben-David 2014). We hypothesize that these characteristics of ML will enable better prediction of variability of crop yields in space and time. However, a comprehensive inter-comparison of process-based crop models, regression and ML is lacking, especially for prediction of variation and extremes in crop yields in relation to spatially and temporally varying climatic effects.

In this study, we choose the US maize yield as an example to demonstrate the potential utilization of ML for assessing climate impacts on crop yields, through comparing the simulations by process-based models, regression and ML in a consistent manner. Specifically, the following scientific questions are addressed in this study: (1) How ML performs in assessing climate impacts on maize yields in the United States? (2) How large are the uncertainties arising from climate impact assessment models? (3) How the average, variability and extremes of US maize yield would change in the future under global warming scenarios? Through a county-level analysis, we aim to identify where maize yield estimation is most uncertain and where assessment of yield response to climate change is relatively more promising, thus providing critical information for climate risk assessment and effective adaptations.

2. Materials and methods

2.1. Crop yields and climate data

Maize yield census data is obtained from the US Department of Agriculture (USDA)'s National Agriculture Statistics Survey's (NASS) Quick Stats database (www.nass.usda.gov/Quick_Stats). Simulated maize yields by 11 process-based models (supplementary table S1 (stacks.iop.org/ERL/15/044027/mmedia)) are obtained from the Agricultural Modelling Inter-comparison and Improvement Project (AgMIP) (Rosenzweig *et al* 2013) and the Inter-Sectoral Impact Model Inter-comparison Project (ISI-MIP) (Warszawski *et al* 2014). Some of the 11 process models were calibrated with regard to certain parameters (supplementary table S1). For example, the parameter maximum LAI, harvest index and the factor for scaling leaf-level photosynthesis to stand level are calibrated in the LPJmL model, while the parameter fertilizer application rate and potential harvest index are calibrated in the GEPIC model. More details on the process models can refer to (Müller *et al* 2017). The crop models have been run under three different configurations (Elliott *et al* 2015). The default setup represents each modelling group's own assumptions, while the key input data, parameters and definitions provided by AgMIP are adopted by the models under the setting of 'fullharm'. Models that consider nitrogen are also run without nitrogen stress under the 'harmnon' setting. Five climate model simulations (supplementary table S2)

from the Coupled Model Intercomparison Project Phase 5 (CMIP5) (Taylor *et al* 2012) under Representative Concentration Pathways RCP8.5 (Moss *et al* 2010) are used to drive crop models for simulating future yield changes. These climate model projections are statistically downscaled to 1/8 degree and bias-corrected against observations using bias-correction and spatial-downscaling approach (BCSD) (Wood *et al* 2004, Leng *et al* 2016). The historical period 1980–2010 is selected because it overlaps with both census yield data and process-based model simulations.

2.2. Machine learning and regression model

Besides process-based crop models, we use ML and a traditional multiple linear regression model to simulate maize yield. Here, the Random Forest algorithm (Breiman 2001) which has been successfully used in previous studies (Hoffman *et al* 2018, Feng *et al* 2019, Vogel *et al* 2019) is adopted. The Random Forest algorithm is a non-parametric ML method and relies on an ensemble of decision trees through two randomization steps: (1) each decision tree is constructed based on a bootstrapped sub-sample dataset, with the decision rule depending on a random sub-set of candidate predictor variables; (2) These processes are repeated at every decision split to overcome the limitations of single decision tree, thus avoiding the potential overfitting issue (Breiman 2001). Compared to traditional statistical approaches (e.g. multiple linear regression), Random Forests are expected to have a number of advantages, as they do not assume a certain shape of response function (e.g. linear or polynomial). That is, the complex interactions and nonlinear relations among the variables could be handled.

Following previous studies (Lobell and Asner 2003, Lobell *et al* 2011, Leng 2017a, 2017b), two climate variables and the year term are selected in this study as inputs for both the ML and regression models. That is, the growing season (June–July–August) temperature and precipitation are selected to represent the first-order climate impacts on yields, while the change trend of yield is accounted for by the year term in the model. Climatic extremes are not explicitly considered in this study, since it is more difficult for CMIP5 climate models to predict future changes in climate extremes than climatology. Specifically, the multiple linear regression model is constructed with temperature (T), precipitation (P) and year (Y) as the predictant and maize yield (MY) as the dependent variable:

$$MY_{c,y} = \beta_0 + \beta_1 T_{c,y} + \beta_2 P_{c,y} + Y_y + \alpha_s + \varepsilon_{c,y} \quad (1)$$

where β_{0-2} are the intercept and coefficients, c is for the county and y is the year, α_s is the intercept representing constant regional effects, $\varepsilon_{c,y}$ denotes the error term. The same time series of three predictors are also used as inputs into the ML model, which is run under

the default settings with $ntree$ (the number of trees to grow in the forest) set to be 500 and $mtry$ (the number of randomly selected predictor variables at each node) being 2. Here, a ‘complete’ model refers to the one that includes both climatic factors governing yield variability and the year term representing the change trend of yields.

2.3. Simulation and analysis

In this study, 80% of the dataset was randomly selected to calibrate both the ML and regression model, with the remaining data used for validations. This procedure was executed 100 times with different randomly selected calibration and validation datasets to evaluate model stability. The ‘complete’ models which include both climatic factors and the year term (see equation (1)) are developed for simulating historical maize yields, based on which inter-comparisons of the long-term mean and year-to-year variation of maize yield are conducted among process-based models, regression model and ML. To assess the portion of yield variability that is contributed by climate variations, a ‘reduced’ model which includes climate variables only is constructed. Specifically, both climate variables and yields are detrended using the least squares linear regression method, before they are used for fitting the ‘reduced’ models. The simulated yield anomaly is then compared with observed yield anomaly to evaluate the variance of yield anomalies that can be explained by climate variations for each maize growing county. Country-level values are obtained based on aggregated yield anomalies through weighted arithmetic mean with harvested area as weighting factors. This calculation ensures that grid cells with large harvesting areas have greater impacts on regional yield anomalies.

2.4. Projection of yield change under the 1.5 °C and 2 °C warming scenarios

Most previous studies have focused on how crop yields will change on average in a changing climate, with less emphasis on predicting potential changes in low yields as a consequence of climate variations. To illustrate the usefulness of ML method for this purpose, we examine how climate change could impact upon the likelihood of low yields defined as the historical 10th percentile yield. Instead of looking at the far future periods such as the end of 21st century, we limit our predictions to the 1.5 °C and 2 °C warming scenarios (relative to pre-industrial levels) (Rogelj *et al* 2015, UNFCCC 2015). It is noted that the reference period 1980–2010 is 0.6 °C warmer than pre-industrial levels, implying that the 1.5 °C and 2 °C warming targets are actually corresponding to global warming of 0.9 °C and 1.4 °C above the reference period, respectively (Lissner and Fischer 2016). By this, most of the historical training climate lies inside

future temperature and precipitation ranges (supplementary figure S1).

Specifically, the gridded temperature and precipitation corresponding to the 1.5 °C and 2 °C warming world relative to pre-industrial era are first extracted from climate model simulations using the time-slice method (Schewe *et al* 2014, Leng *et al* 2015, Gosling *et al* 2016, Lissner and Fischer 2016), and aggregated to the county scale with weights based on the harvest area. Changes in temperature and precipitation relative to the reference period are then calculated for each county and used for perturbing the observed climate time series to construct future climate scenarios. Future maize yields simulated by ML driven with the climate scenarios are compared with yields in the reference period for evaluating the changes in yield average and low yields. The above processes are repeated for each climate model, based on which the multi-model ensemble mean is used for analysis with the inter-model spread indicating the uncertainty from climate models.

3. Results

Figure 1 shows the long-term mean county-level maize yield simulated by the regression model, ML algorithm (i.e. Random Forests) and AgMIP models under different settings against observations for the period 1980–2010. It is evident that process-based models under the default settings show large bias in yield simulations, not only in the spatial distribution pattern but also in the magnitude of long-term mean yields. Such bias remains substantial when models are run even under the fullharm settings (i.e. assuming no limits of water and nutrient applications). When fed with observed management parameters (e.g. growing season, fertilizer input) in the harmonized settings (Elliott *et al* 2015), simulated yields by process-based models tend to match closer with observations in most of the country, though overestimation still exists in southern parts of the Mid-west corn belt and much of southeastern US. The varying performance under different management settings suggest that representation of agricultural management is an important uncertainty source for crop yield simulations.

Compared to process-based crop models, both regression and ML models show better skills in reproducing the observed spatial distribution patterns of maize yield across the country due to the trend (supplementary figure S2), though they are driven by only three predictors (i.e. temperature, precipitation and time). Similar findings are obtained when looking at the extremes of maize yield, with regression and ML models better capturing the observed spatial patterns of 10th percentile yield (supplementary figure S3). However, caution should be exercised when interpreting the results from the above inter-comparisons, because both regression model and ML have been

trained against observations, while limited calibration is done in the simulations of process-based crop models (Elliott *et al* 2015). Nevertheless, it is still important to recognize the promising skills of the regression model and ML algorithm for yield simulations, especially given that substantial inputs and computer costs are required for driving process-based models.

As for the year-to-year variation of maize yield, all models show similar spatial distribution patterns with larger values of coefficient of variance (CV) in the northern and southeastern parts of the country (figure 2). Compared to observations, both regression model and ML tend to underestimate the magnitude of historical yield variability, which may be induced to some extent by the assumption of fixed growing season, though the bias is relatively small (supplementary figure S4). Process-based AgMIP crop models show mixed performances and the ensemble mean gives a overestimation in much of western country, which is mainly due to the underestimation of yield averages (figures 1(a)–(c)). Similar to the simulation of yield averages, the performance of process-based models depends on the settings of agricultural management, and the sign of bias even differs among the different settings in the Midwest corn belt (supplementary figures S4(a)–(c)). Overall, process-based crop models tend to perform better in simulating the year-to-year variability of maize yield than its long-term averages. The results have great implications for enhancing the agricultural resilience to environmental risks, since it is the year-to-year variability that dominates the stability of food supply and food security.

Figure 3 shows the probability density distribution (PDF) of maize yield from observations and simulations by regression model, ML and process-based crop models under the harmnon setting. For the country as a whole, process-based models show large bias in simulating historical maize yield distributions, not only in the peak but also in the PDF tails. Instead, both regression model and ML well capture the observed shape of the probability density curve, but show slight bias in the peak of PDF. Specifically, the regression model tends to shift the peak of yield PDF to the right of observations, while ML gives an estimate shifting to the left. As for yield anomaly (after excluding the change trends), process-based models have largely underestimated the peak value of yield PDF. This suggests that the impacts of climate variations on yields may have been underestimated in the process-based models, since yield anomaly is mainly driven by climate variations. The regression model captures well the observed peak value of yield PDF, but showing substantial bias in the negative tail of yield distribution. ML outperforms both regression model and process-based models not only in the peak but also in the tails of yield PDF, demonstrating the value of using ML for analysis of the yield

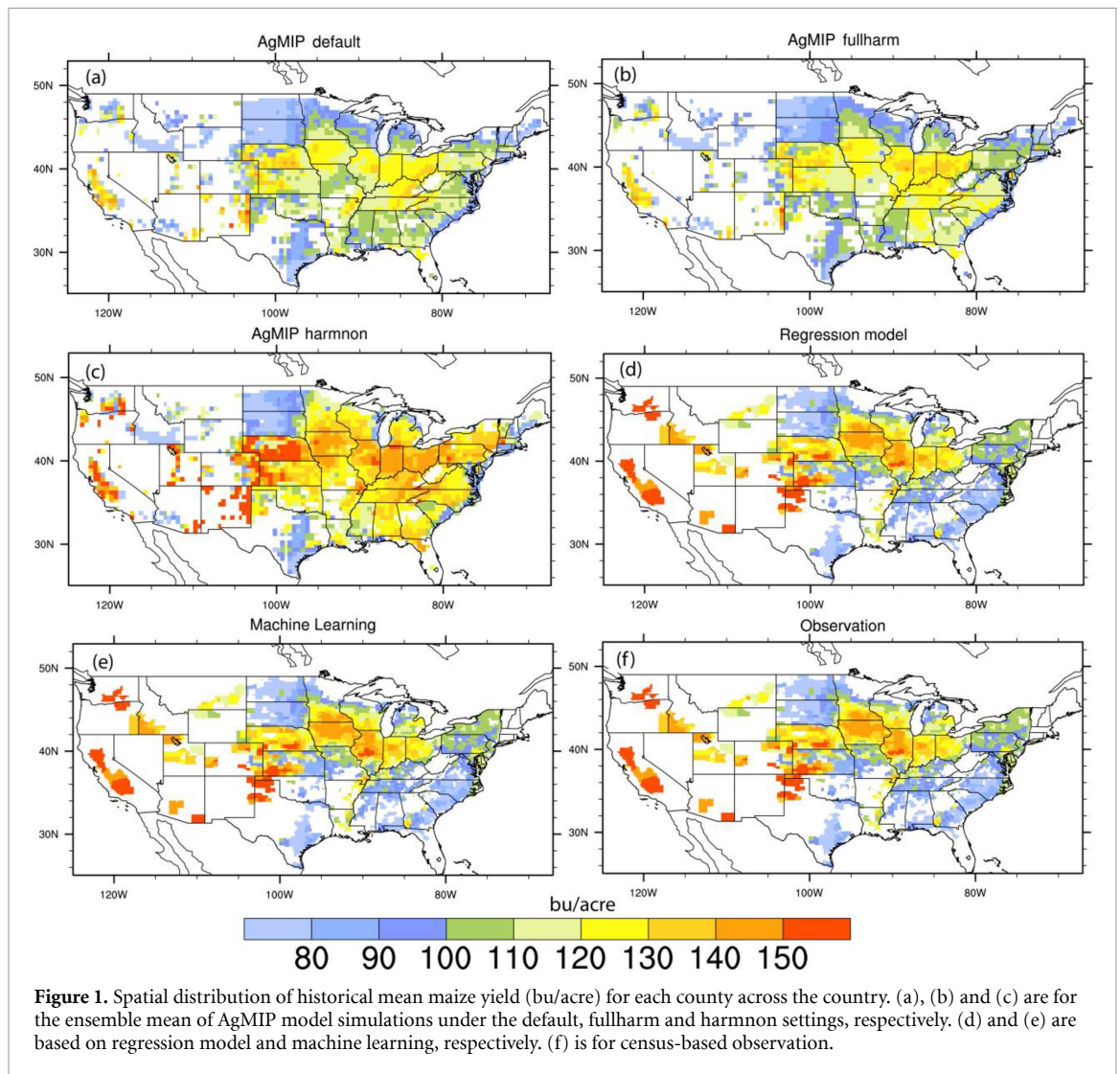


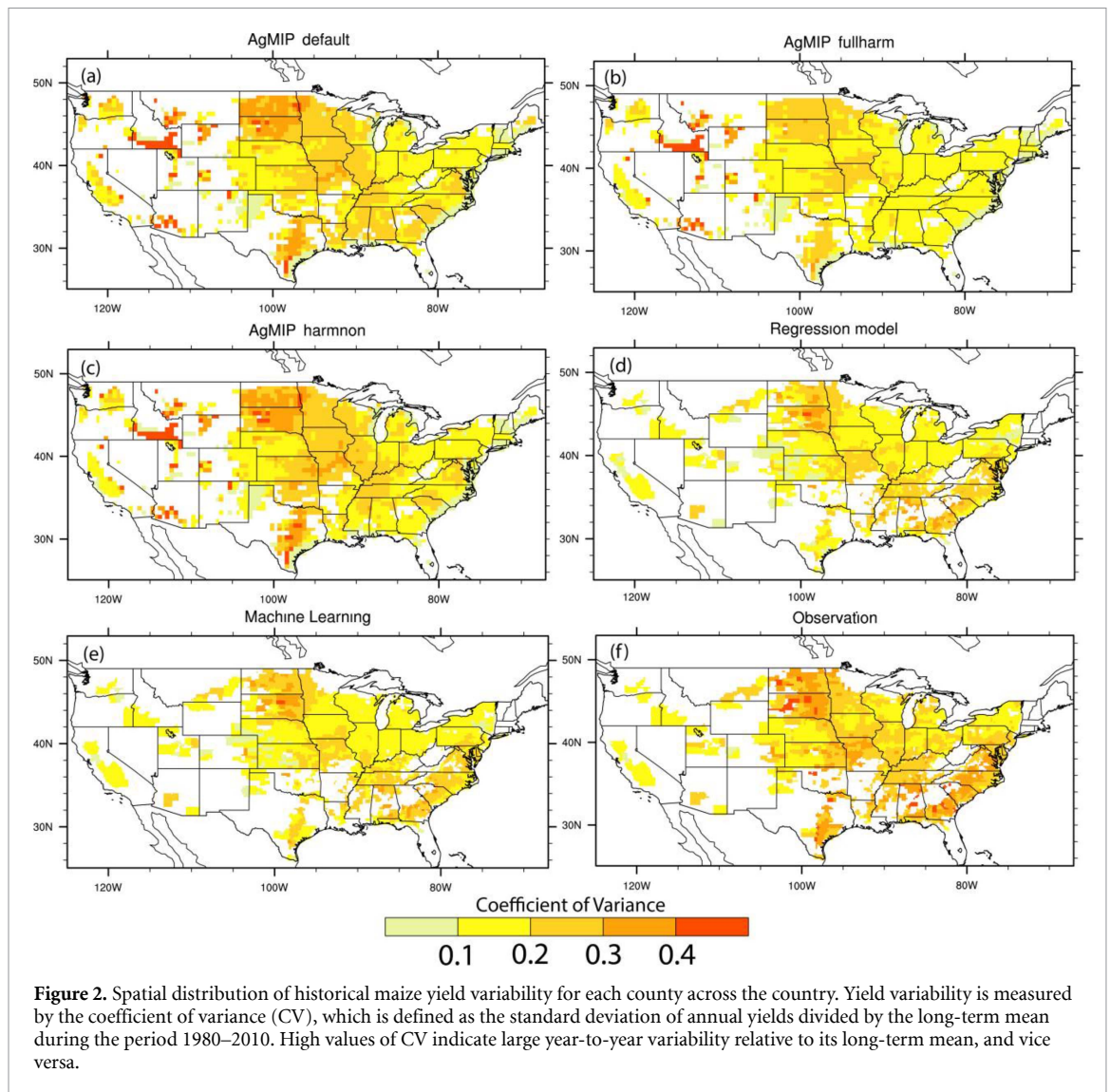
Figure 1. Spatial distribution of historical mean maize yield (bu/acre) for each county across the country. (a), (b) and (c) are for the ensemble mean of AgMIP model simulations under the default, fullharm and harmnon settings, respectively. (d) and (e) are based on regression model and machine learning, respectively. (f) is for census-based observation.

variability and the risks of extreme low yields. Similar findings are observed for the top 6 maize producing states, but with regression model showing worse skill in reproducing the PDF of yield anomaly in Nebraska, Indiana and South Dakota. Overall, ML exhibits the best skill in simulating the PDF of maize yield and its anomaly, followed by the regression model and process-based models.

To assess the portion of yield variability that is contributed by climate variations, we perform ML and regression simulations driven by climate variables only (see method section 2.3). Both climate variables and yields are detrended, before they are used for fitting the models. Overall, our analysis indicates that climate variations can explain more than 40% of year-to-year variability of maize yield for the country as a whole (figure 4). However, such estimates depend on the models that link yield variability to climate variations. Specifically, process-based models give the smallest estimates, with only 42% of yield variability explained by the inter-annual climate variations. Regression models which empirically fit the linear climate-yield relation indicates that 51% of

US maize yield variability is induced by climate variations. Based on the ML algorithm, climate variation can explain 93% of historical yield variability, suggesting that previous estimates of climate impacts on yield anomaly may have been largely underestimated.

Given the promising performance of ML, we simulate US maize yield changes under the 1.5 °C and 2 °C warming scenarios by driving the ML model with climate model projections (see method section 2.4). As shown in figure 5, a shift of yield probability distribution curve to the left is observed, which is more pronounced under the 2 °C warming (by ~2050 s) than the 1.5 °C warming (by ~2040 s). This suggests that US maize yield is likely to decrease in the future, especially under the 2 °C warming world. On average, yield is projected to decrease by 9.0% and 13.5% under the 1.5 °C and 2 °C warming scenarios, respectively, which are broadly consistent with previous estimations (Schlenker and Roberts 2009, Butler and Huybers 2013). Spatially, a decrease in maize yield is predicted for most US counties under the 1.5 °C and 2 °C global warming scenarios (figures 6(a) and (b)). The most severe yield loss is



projected in the hottest states in the central and southeastern US, while the northern US where temperature is lower than the optimal level for crop growth is likely to benefit from future warming climate. The magnitude of maize yield loss tends to grow progressively in much of the country with increase in temperature, especially in the Southeastern growing areas. Using the ML model, we predict that maize yield could decrease by more than 15% in most of central and southeastern part of the country under the 2 °C global warming. Constraining global temperature rise to 1.5 °C would lead to smaller yield decrease in these hot-spot regions. The results have great implications for informing targeted adaptation measures, through identifying the hotspot regions where future climate change impacts are most severe.

Figure 6 further shows that extreme low yield defined as the historical 10th percentile yield will occur more frequently in the future. This change in the frequency of extremely low yields potentially has great implications for food security, given that an extreme yield anomaly would typically cause larger

impacts on the price and markets. Spatially, most of the maize growing areas will experience a substantial increase in the probability of extreme low yields, while a slight decrease is projected in the northern parts of the country due to the benefits of climate warming (figures 6(c) and (d)). In Texas, Kansas, Missouri and Southeastern parts of the country, the ML model projects that the probability of extreme yield anomalies would even increase from 10% up to 30% under 2 °C warming.

4. Discussions

Understanding climate impacts on crop yields is critical for adaptation and mitigations. Previous assessment relies on regression models or process-based models. This study demonstrates the potential utilization of ML for assessing climate impacts on crop yields, and a comprehensive comparison among process-based models, regression and ML is conducted, which is extremely valuable in improving

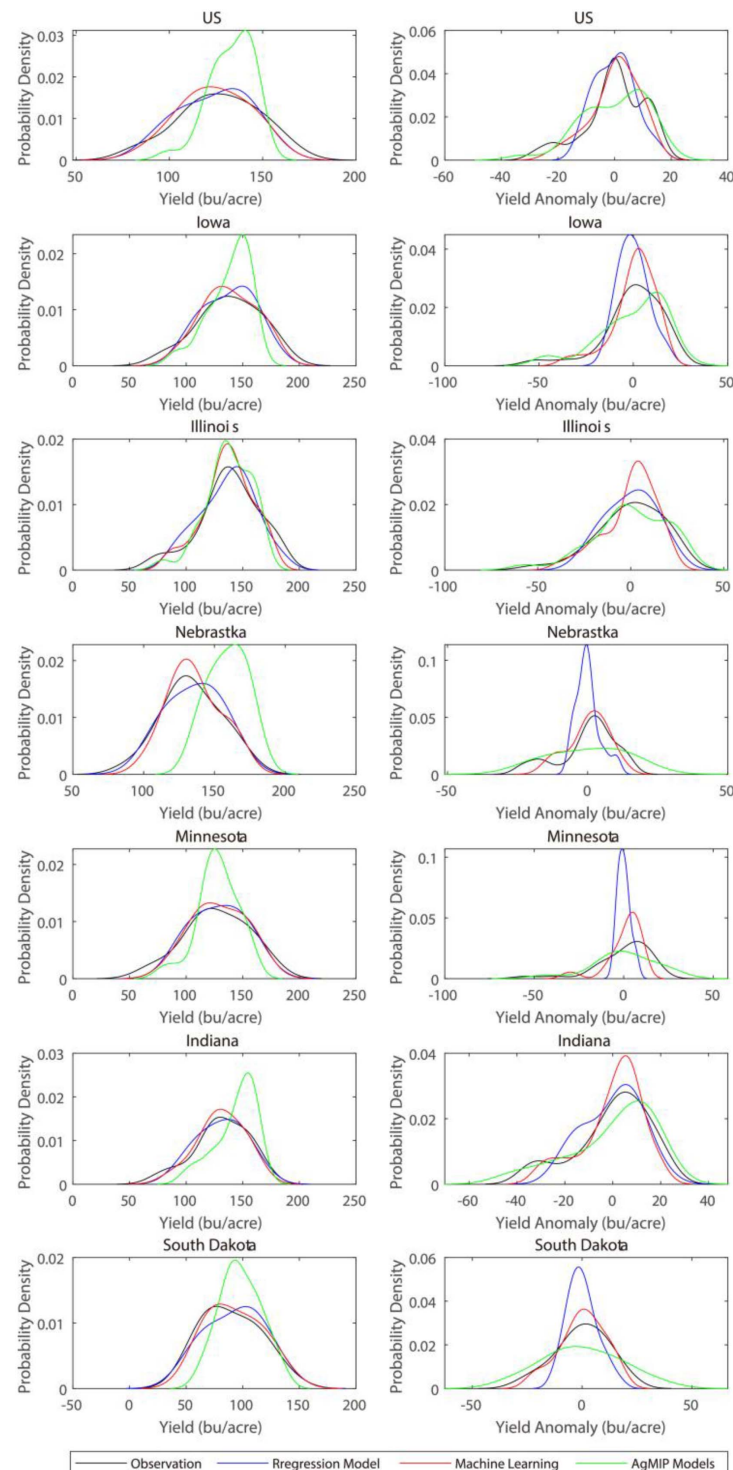
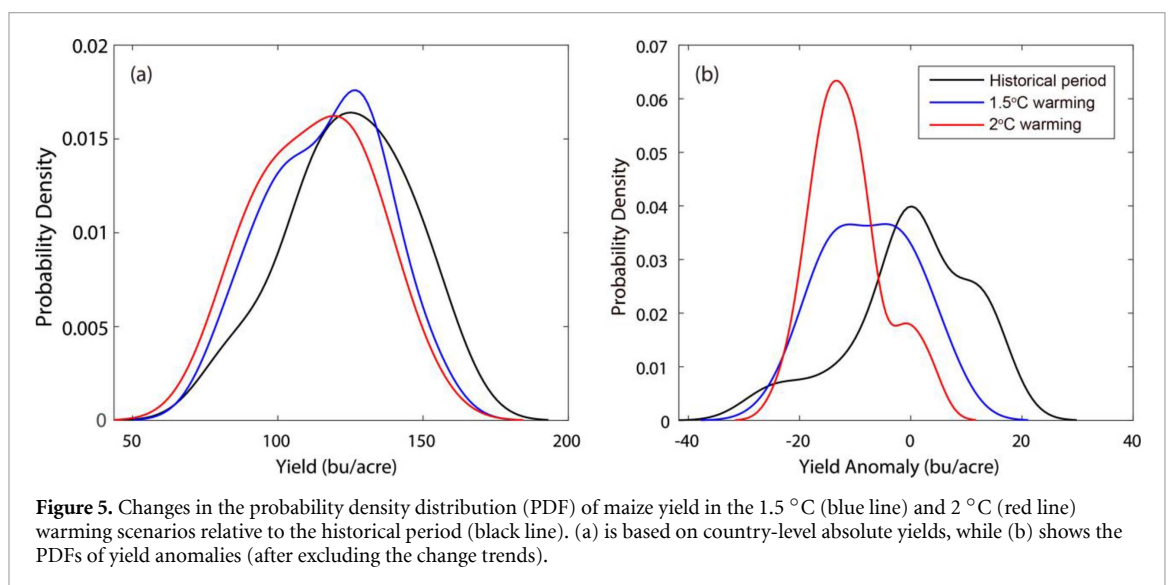
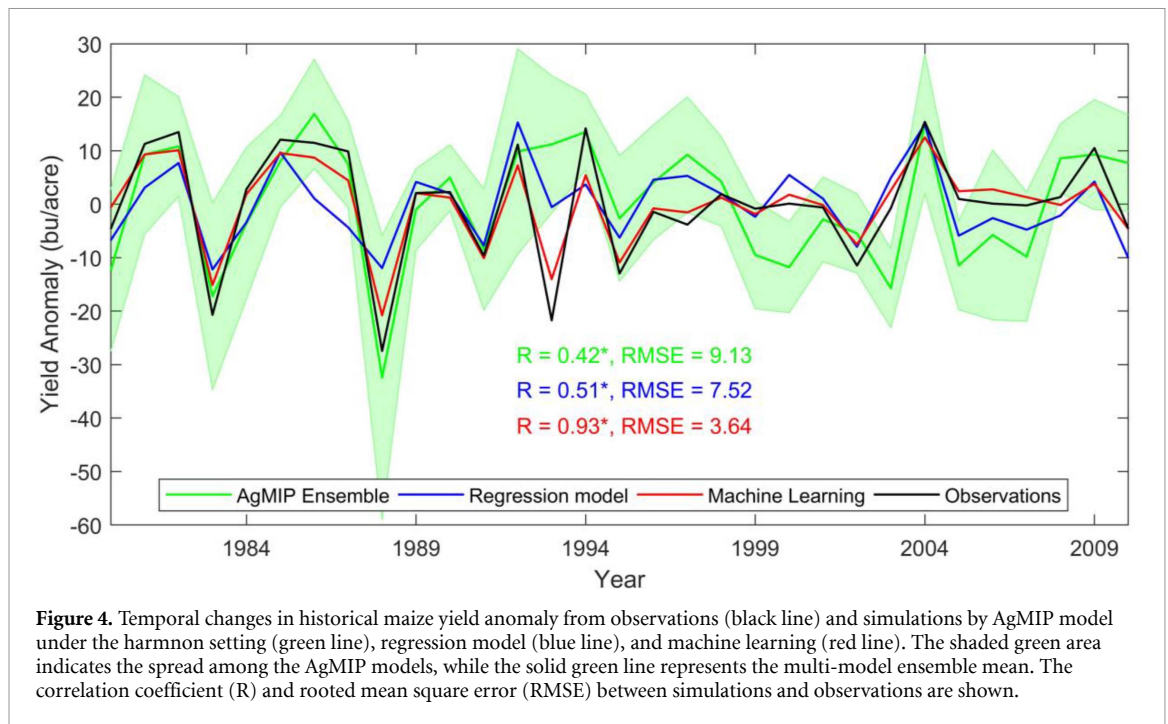


Figure 3. Probability density distribution (PDF) of maize yield from observations and simulations by regression model, machine learning and AgMIP models under the harmnon setting for the country as a whole and top six maize producing states. The left column shows the PDF of the absolute maize yield, while the right column is based on yield anomaly after excluding its change trend. Annual gridded yields are aggregated to the country and state scales with weights determined by the harvest area, based on which the PDFs are fitted.

our understanding of the implications of alternative methodological approaches. However, there are several limitations that should be acknowledged when interpreting the results of this study.

First, we choose June-July-August (JJA) as the major growing season for US maize, following the literature (Lobell *et al* 2013). Sacks and Kucharik

(2011) reported that US maize growing period has increased by 12 d for the period 1981–2005. Since our study is based on the seasonal mean climate (i.e. average of growing season climate), a shift of few days would not affect the overall conclusions. Indeed, Lobell and Field (2007) showed that the particular choice of growing season did not greatly affect

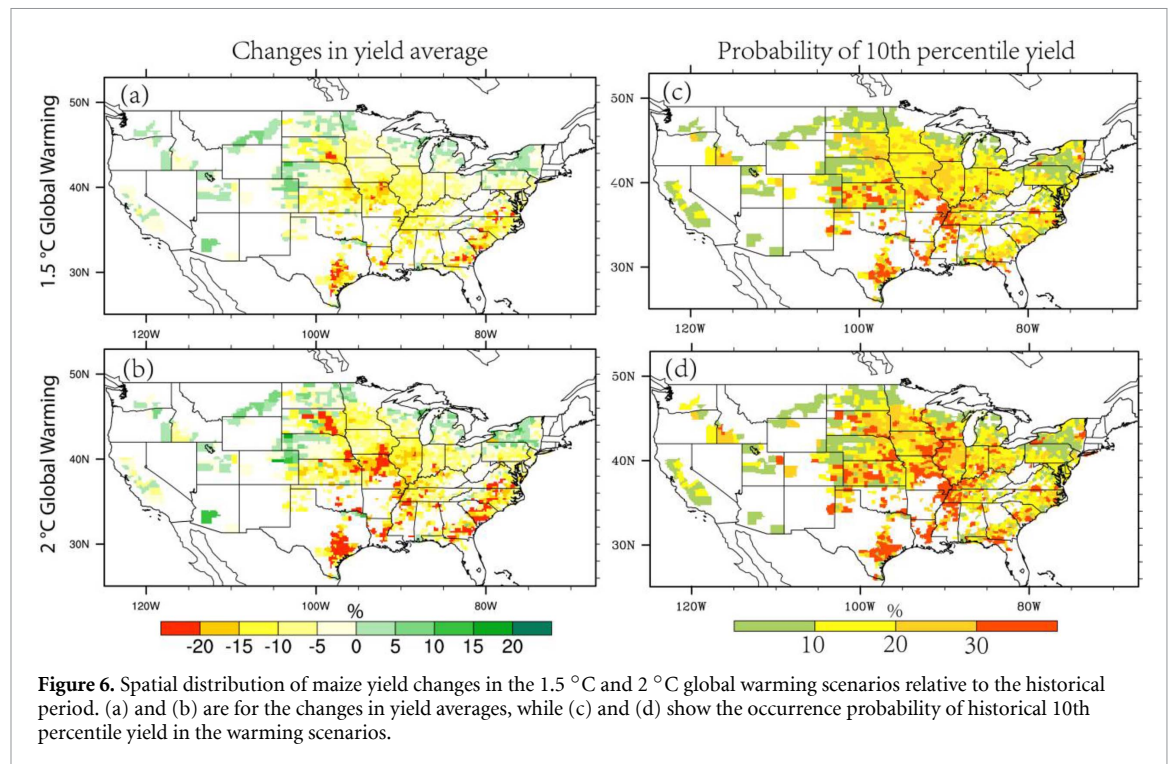


their estimation of climate impacts on crop yields. Schlenker and Roberts (2009) confirmed this by conducting a series of sensitivity analysis and found qualitatively similar results for eight alternative specifications of US maize growing season. Therefore, we choose the JJA period as the major growing season for US maize, and a fixed growing season has been widely adopted in many previous works (Lobell and Asner 2003).

Second, there are several other climatic factors (e.g. vapor pressure deficit) that are neglected in this study. Given limitations in both statistical and process-based models, the goal of this paper is not to present an accurate yield prediction, but rather to compare and understand more fully their different level of performances. Therefore, the commonly adopted temperature and precipitation are selected,

since they are the major climatic predictors of yields and their effects are represented in both statistical and process-based models. We show that even conditioned on temperature and precipitation only, ML and regression model tend to show promising performances. Further sensitivity analysis could be conducted to assess the benefits of including other climatic variables, though we suspect that this would be relatively small compared to precipitation and temperature.

Third, process-based models are not fully calibrated, while regression model and ML are trained against observations. It is well recognized that fully calibrating the process-based models is non-trivial, given the large numbers of uncertain parameters and lack of site-specific data. Especially, calibration of 11 process models would require tremendous efforts and



coordination of different crop model development groups. To demonstrate the importance of model calibrations, we use and compare the simulations under three different settings (i.e. the default, fullharm and harmnon), which vary in the treatment of certain parameters. It is noted that some of the parameters related to managements are harmonized under the fullharm and harmnon settings, though they are not fully calibrated. Such a treatment does lead to better performance in yield simulations, especially under the harmnon settings (see the results section).

Fourth, our estimates of future yield changes are conducted without considering adaptations (Butler and Huybers 2013, Challinor *et al* 2014a), CO₂ fertilization effects (Mcgrath and Lobell 2013, Deryng *et al* 2016), technology change (Najafi *et al* 2018) and other important factors such as pests, weeds and nutrients (Tubiello *et al* 2007). For example, the negative temperature impacts on crop yields could be reduced through soil mulching (Qin *et al* 2015), conservation tillage (Karlen *et al* 2013) and multiple cropping (Seifert and Lobell 2015). Recent observation-based studies also showed that irrigation can well buffer and mitigate the negative climate impacts on US maize yield (Troy *et al* 2015, Leng 2017a), though irrigation may be constrained by water availability, in particular in more severe drought scenarios.

5. Conclusions

While predicting crop yield response to climate change and variability has received a lot of attentions, it is not well explored in terms of yield variability and

extremes. Here, we use the process-based models and traditional regression model as well an ML algorithm to investigate climate impacts on the average, variability and extremes of crop yields, using US maize yield as an example. Compared to process-based models, both regression model and ML show better skills in simulating yield averages, but still tend to slightly underestimate the year-to-year variability of yields in major production regions. Whilst the ensemble mean of process-based models can well reproduce the observed patterns of maize yield variability, there are substantial biases in predictions of yield averages. Notably, observed yield anomalies (after excluding the trend term) are almost completely reproduced by ML ($R^2 = 0.93$), pointing to the promising skill of ML in assessing yield anomalies induced by climate variations. This also suggests that climate variation may have played a more dominant role in yield anomaly than previous estimations based on regression and process-based crop models, which reproduce less than half of observed yield anomalies.

We also found that the performance of process-based models largely depends on the parameterizations of agricultural managements (e.g. planting date and irrigation). When fed with harmonized management parameters, current state-of-art process-based models show improved skills in major production regions, though substantial bias remains in southeastern US. This suggests that representation of agricultural managements is an importance uncertainty source in process-based crop models. Nonetheless, the empirical (regression and ML) methods do not explicitly use parameterization of agricultural management, but are still able to better

represent yield variations in many cases, conditioned only on climate variables.

Based on the historically trained ML model, US maize yield is projected to decrease by 9.0% and 13.5% under the 1.5 °C (by ~2040 s) and 2 °C (by ~2050 s) global warming scenarios, respectively. However, such estimates cannot be interpreted as predictions, as many factors influencing yields and the potential adaptation effects are not included. Given the ML algorithm's improved capability for predicting yield variations, we used it to examine the future change in the frequency of extreme low yields. Using the yield that is currently at the 10th percentile in the yield distribution, we predicted how frequently (i.e. at what percentile of the yield distribution) this would occur in future. Nationally we found that yields would be less than or equal to this benchmark low yield in 19% and 25% of years in 1.5 °C and 2 °C global warming scenarios, with potentially significant implications for food supply, prices and trade.

Understanding crop yield response to climate change and variability is fundamental for mitigating agricultural risks within the context of global warming. Previous assessments of climate impacts are mainly aided with either process-based models or regression models. Here, we show that machine-learning algorithm is powerful for predicting crop yields, even though the model is conditioned on only two climate variables (i.e. temperature and precipitation). Process-based models will continue to be invaluable tools for projecting long term trends and the impacts of future changes, but could be supplemented by machine learning to downscale temporal and spatial patterns of yield average, variability and extremes. This will be particularly important for analysis of the risk of low yields in climatic extremes.

Acknowledgments

We thank the editor and two anonymous reviewers for their constructive comments and suggestions that led to substantial improvements of the manuscript. This research was funded by the Wellcome Trust, Our Planet Our Health (Livestock, Environment and People—LEAP), award number 205212/Z/16/Z.

Data availability statement

The data that support the findings of this study are openly available. Maize yield census data is available at the US Department of Agriculture (USDA)'s National Agriculture Statistics Survey's (NASS) Quick Stats database (www.nass.usda.gov/Quick_Stats/). Process-based crop model simulations are available at www.isimip.org/gettingstarted/data-access/. Observed climate data sets are downloaded from [www.engr.scu.edu/~emaurer/gridded_obs/index_](http://www.engr.scu.edu/~emaurer/gridded_obs/index_gridded_obs.html)

[gridded_obs.html](http://www.engr.scu.edu/~emaurer/gridded_obs/index_gridded_obs.html). The bias-corrected CMIP5 climate scenarios are obtained from [ftp://gdo-dcp.ucllnl.org/pub/dcp/archive/cmip5/hydro/BCSD_mon_forc_nc/](http://gdo-dcp.ucllnl.org/pub/dcp/archive/cmip5/hydro/BCSD_mon_forc_nc/).

References

- Asseng S, Ewert F, Martre P, Rötter R, Lobell D, Cammarano D, Kimball B, Ottman M J, Wall G and White J 2015 Rising temperatures reduce global wheat production *Nat. Clim. Change* **5** 143–7
- Asseng S, Ewert F, Rosenzweig C, Jones J, Hatfield J, Ruane A, Boote K, Thorburn P, Rötter R and Cammarano D 2013 Uncertainty in simulating wheat yields under climate change *Nat. Clim. Change* **3** 827–32
- Bindi M and Olesen J E 2011 The responses of agriculture in Europe to climate change *Reg. Environ. Change* **11** 151–8
- Breiman L 2001 Random forests *Mach. Learn.* **45** 5–32
- Butler E E and Huybers P 2013 Adaptation of US maize to temperature variations *Nat. Clim. Change* **3** 68–72
- Cai Y, Guan K, Lobell D, Potgieter A B, Wang S, Peng J, Xu T, Asseng S, Zhang Y and You L 2019 Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches *Agric. Forest Meteorol.* **274** 144–59
- Challinor A J, Watson J, Lobell D, Howden S, Smith D and Chhetri N 2014a A meta-analysis of crop yield under climate change and adaptation *Nat. Clim. Change* **4** 287
- Challinor A J, Watson J, Lobell D, Howden S, Smith D and Chhetri N 2014b A meta-analysis of crop yield under climate change and adaptation *Nat. Clim. Change* **4** 287–91
- Chlingaryan A, Sukkarieh S and Whelan B 2018 Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: a review *Comput. Electron. Agric.* **151** 61–69
- Deryng D, Elliott J, Folberth C, Müller C, Pugh T A, Boote K J, Conway D, Ruane A C, Gerten D and Jones J W 2016 Regional disparities in the beneficial effects of rising CO₂ concentrations on crop water productivity *Nat. Clim. Change* **6** 786–90
- Elliott J, Deryng D, Müller C, Frieler K, Konzmann M, Gerten D, Glotter M, Flörke M, Wada Y and Best N 2014 Constraints and potentials of future irrigation water availability on agricultural production under climate change *Proc. Natl. Acad. Sci.* **111** 3239–44
- Elliott J, Müller C, Deryng D, Chrysanthacopoulos J, Boote K, Büchner M, Foster I, Glotter M, Heinke J and Iizumi T 2015 The global gridded crop model intercomparison: data and modeling protocols for Phase 1 (v1.0) *Geosci. Model Dev.* **8** 261–77
- Feng P, Wang B, Li Liu D, Waters C and Yu Q 2019 Incorporating machine learning with biophysical model can improve the evaluation of climate extremes impacts on wheat yield in south-eastern Australia *Agric. Forest Meteorol.* **275** 100–13
- Feng P, Wang B, Li Liu D, Xing H, Ji F, Macadam I, Ruan H and Yu Q 2018 Impacts of rainfall extremes on wheat yield in semi-arid cropping systems in eastern Australia *Clim. Change* **147** 555–69
- Gaupp F, Hall J, Mitchell D and Dadson S 2019 Increasing risks of multiple breadbasket failure under 1.5 and 2° C global warming *Agric. Syst.* **175** 34–45
- Gosling S N, Zaherpour J, Mount N J, Hattermann F F, Dankers R, Arheimer B, Breuer L, Ding J, Haddeland I and Kumar R 2016 A comparison of changes in river runoff from multiple global and catchment-scale hydrological models under global warming scenarios of 1° C, 2° C and 3° C *Clim. Change* **141** 577–95
- Hoffman A L, Kemanian A R and Forest C E 2018 Analysis of climate signals in the crop yield record of sub-Saharan Africa *Glob. Change Biol.* **24** 143–57
- Jeong J H, Resop J P, Mueller N D, Fleisher D H, Yun K, Butler E E, Timlin D J, Shim K-M, Gerber J S and Reddy V R

- 2016 Random forests for global and regional crop yield predictions *PLoS One* **11** e0156571
- Johnson M D, Hsieh W W, Cannon A J, Davidson A and Bédard F 2016 Crop yield forecasting on the Canadian Prairies by remotely sensed vegetation indices and machine learning methods *Agric. Forest Meteorol.* **218** 74–84
- Jones P G and Thornton P K 2003 The potential impacts of climate change on maize production in Africa and Latin America in 2055 *Global Environ. Change* **13** 51–59
- Karlen D L, Kovar J L, Cambardella C A and Colvin T S 2013 Thirty-year tillage effects on crop yield and soil fertility indicators *Soil Tillage Res.* **130** 24–41
- Leng G 2017a Evidence for a weakening strength of temperature-corn yield relation in the United States during 1980–2010 *Sci. Total Environ.* **605** 551–8
- Leng G 2017b Recent changes in county-level corn yield variability in the United States from observations and crop models *Sci. Total Environ.* **607** 683–90
- Leng G and Hall J 2019 Crop yield sensitivity of global major agricultural countries to droughts and the projected changes in the future *Sci. Total Environ.* **654** 811–21
- Leng G, Huang M, Voisin N, Zhang X, Asrar G R and Leung L R 2016 Emergence of new hydrologic regimes of surface water resources in the conterminous United States under future warming *Environ. Res. Lett.* **11** 114003
- Leng G, Tang Q, Huang S and Zhang X 2015 Extreme hot summers in China in the CMIP5 climate models *Clim. Change* **135** 669–81
- Lissner T K and Fischer E M 2016 Differential climate impacts for policy-relevant limits to global warming: the case of 1.5 °C and 2 °C *Earth Syst. Dyn.* **7** 327
- Liu B, Asseng S, Müller C, Ewert F, Elliott J, Lobell D B, Martre P, Ruane A C, Wallach D and Jones J W 2016 Similar estimates of temperature impacts on global wheat yield by three independent methods *Nat. Clim. Change* **6** 1130–6
- Lobell D B and Asner G P 2003 Climate and management contributions to recent trends in US agricultural yields *Science* **299** 1032–32
- Lobell D B and Asseng S 2017 Comparing estimates of climate change impacts from process-based and statistical crop models *Environ. Res. Lett.* **12** 015001
- Lobell D B and Field C B 2007 Global scale climate–crop yield relationships and the impacts of recent warming *Environ. Res. Lett.* **2** 014002
- Lobell D B, Hammer G L, Mclean G, Messina C, Roberts M J and Schlenker W 2013 The critical role of extreme heat for maize production in the United States *Nat. Clim. Change* **3** 497–501
- Lobell D B, Schlenker W and Costa-Roberts J 2011 Climate trends and global crop production since 1980 *Science* **333** 616–20
- Mcgrath J M and Lobell D B 2013 Regional disparities in the CO₂ fertilization effect and implications for crop yields *Environ. Res. Lett.* **8** 014054
- Moss R H, Edmonds J A, Hibbard K A, Manning M R, Rose S K, Van Vuuren D P, Carter T R, Emori S, Kainuma M and Kram T 2010 The next generation of scenarios for climate change research and assessment *Nature* **463** 747–56
- Müller C, Cramer W, Hare W L and Lotze-Campen H 2011 Climate change risks for African agriculture *Proc. Natl. Acad. Sci.* **108** 4313–5
- Müller C, Elliott J, Chrysanthacopoulos J, Arneth A, Balkovic J, Ciais P, Deryng D, Folberth C, Glotter M and Hoek S 2017 Global gridded crop model evaluation: benchmarking, skills, deficiencies and implications *Geosci. Model Dev.* **10** 1403
- Najafi E, Devineni N, Khanbilvardi R M and Kogan F 2018 Understanding the changes in global crop yields through changes in climate and technology *Earth Future* **6** 410–27
- Olesen J E and Bindi M 2002 Consequences of climate change for European agricultural productivity, land use and policy *Eur. J. Agron.* **16** 239–62
- Parry M L, Rosenzweig C, Iglesias A, Livermore M and Fischer G 2004 Effects of climate change on global food production under SRES emissions and socio-economic scenarios *Global Environ. Change* **14** 53–67
- Peng S, Huang J, Sheehy J E, Laza R C, Visperas R M, Zhong X, Centeno G S, Khush G S and Cassman K G 2004 Rice yields decline with higher night temperature from global warming *Proc. Natl. Acad. Sci. USA* **101** 9971–5
- Qin W, Hu C and Oenema O 2015 Soil mulching significantly enhances yields and water and nitrogen use efficiencies of maize and wheat: a meta-analysis *Sci. Rep.* **5** 16210
- Ray D K, Gerber J S, Macdonald G K and West P C 2015 Climate variation explains a third of global crop yield variability *Nat. Commun.* **6** 5989
- Reidsma P, Ewert F, Lansink A O and Leemans R 2010 Adaptation to climate change and climate variability in European agriculture: the importance of farm level responses *Eur. J. Agron.* **32** 91–102
- Roberts M J, Braun N O, Sinclair T R, Lobell D B and Schlenker W 2017 Comparing and combining process-based crop models and statistical models with some implications for climate change *Environ. Res. Lett.* **12** 095010
- Rogelj J, Luderer G, Pietzcker R C, Kriegler E, Schaeffer M, Krey V and Riahi K 2015 Energy system transformations for limiting end-of-century warming to below 1.5 [deg] C *Nat. Clim. Change* **5** 519–27
- Rosenzweig C, Elliott J, Deryng D, Ruane A C, Müller C, Arneth A, Boote K J, Folberth C, Glotter M and Khabarov N 2014 Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison *Proc. Natl. Acad. Sci.* **111** 3268–73
- Rosenzweig C, Jones J, Hatfield J, Ruane A, Boote K, Thorburn P, Antle J, Nelson G, Porter C and Janssen S 2013 The agricultural model intercomparison and improvement project (AgMIP): protocols and pilot studies *Agric. Forest Meteorol.* **170** 166–82
- Sacks W J and Kucharik C J 2011 Crop management and phenology trends in the US Corn Belt: Impacts on yields, evapotranspiration and energy balance *Agric. Forest Meteorol.* **151** 882–94
- Schauberger B, Archontoulis S, Arneth A, Balkovic J, Ciais P, Deryng D, Elliott J, Folberth C, Khabarov N and Müller C 2017 Consistent negative response of US crops to high temperatures in observations and crop models *Nat. Commun.* **8** 13931
- Schewe J, Heinke J, Gerten D, Haddeland I, Arnell N W, Clark D B, Dankers R, Eisner S, Fekete B M and Colón-González F J 2014 Multimodel assessment of water scarcity under climate change *Proc. Natl. Acad. Sci.* **111** 3245–50
- Schlenker W and Lobell D B 2010 Robust negative impacts of climate change on African agriculture *Environ. Res. Lett.* **5** 014010
- Schlenker W and Roberts M J 2009 Nonlinear temperature effects indicate severe damages to US crop yields under climate change *Proc. Natl. Acad. Sci.* **106** 15594–8
- Seifert C A and Lobell D B 2015 Response of double cropping suitability to climate change in the United States *Environ. Res. Lett.* **10** 024002
- Shalev-Shwartz S and Ben-David S 2014 *Understanding Machine Learning: From Theory to Algorithms* (Cambridge: Cambridge university press)
- Tao F, Yokozawa M, Xu Y, Hayashi Y and Zhang Z 2006 Climate changes and trends in phenology and yields of field crops in China, 1981–2000 *Agric. Forest Meteorol.* **138** 82–92
- Taylor K E, Stouffer R J and Meehl G A 2012 An overview of CMIP5 and the experiment design *Bull. Am. Meteorol. Soc.* **93** 485–98
- Tebaldi C and Lobell D 2008 Towards probabilistic projections of climate change impacts on global crop yields *Geophys. Res. Lett.* **35** L08705
- Troy T, Kipgen C and Pal I 2015 The impact of climate extremes and irrigation on US crop yields *Environ. Res. Lett.* **10** 054013

- Tubiello F N, Soussana J-F and Howden S M 2007 Crop and pasture response to climate change *Proc. Natl. Acad. Sci.* **104** 19686–90
- UNFCCC 2015 Adoption of the Paris Agreement *Proposal by the President (Draft Decision)* (Geneva (Switzerland): United Nations Office) p 32
- Urban D, Roberts M J, Schlenker W and Lobell D B 2012 Projected temperature changes indicate significant increase in interannual variability of US maize yields *Clim. Change* **112** 525–33
- Vogel E, Donat M G, Alexander L V, Meinshausen M, Ray D K, Karoly D, Meinshausen N and Frieler K 2019 The effects of climate extremes on global agricultural yields *Environ. Res. Lett.* **14** 054010
- Waha K, Huth N, Carberry P and Wang E 2015 How model and input uncertainty impact maize yield simulations in West Africa *Environ. Res. Lett.* **10** 024017
- Wang E, Martre P, Zhao Z, Ewert F, Maïorano A, Rötter R P, Kimball B A, Ottman M J, Wall G W and White J W 2017 The uncertainty of crop yield projections is reduced by improved temperature response functions *Nat. Plants* **3** 17102
- Warszawski L, Frieler K, Huber V, Piontek F, Serdeczny O and Schewe J 2014 The inter-sectoral impact model intercomparison project (ISI-MIP): project framework *Proc. Natl. Acad. Sci.* **111** 3228–32
- Wheeler T and von Braun J 2013 Climate change impacts on global food security *Science* **341** 508–13
- Wood A W, Leung L R, Sridhar V and Lettenmaier D 2004 Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs *Clim. Change* **62** 189–216
- Yao F, Xu Y, Lin E, Yokozawa M and Zhang J 2007 Assessing the impacts of climate change on rice yields in the main rice areas of China *Clim. Change* **80** 395–409
- Zampieri M, Ceglar A, Dentener F and Toreti A 2017 Wheat yield loss attributable to heat waves, drought and water excess at the global, national and subnational scales *Environ. Res. Lett.* **12** 064008
- Zhao C, Liu B, Piao S, Wang X, Lobell D B, Huang Y, Huang M, Yao Y, Bassu S and Ciais P 2017 Temperature increase reduces global yields of major crops in four independent estimates *Proc. Natl. Acad. Sci.* **114** 9326–31
- Zhao C, Piao S, Wang X, Huang Y, Ciais P, Elliott J, Huang M, Janssens I A, Li T and Lian X 2016 Plausible rice yield losses under future climate warming *Nat. Plants* **3** 16202