**MAIN MANUSCRIPT**

**Title:** Machine Learning to Predict Biomass Sorghum Yields under Future Climate Scenarios

**Authors:**
Tyler Huntington[1,2], Xinguang Cui[1,2,3], Umakant Mishra[1,2,4]*, Corinne D. Scown[1,2,5,6]*

[1]Biosciences Area, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA, 94720, USA
[2]Joint BioEnergy Institute, 5885 Hollis Street, Emeryville, CA, 94608, USA
[3]Department of Aerospace Engineering, Huazhong University of Science and Technology, 1037 Luoyu Road, Wuhan, Hubei, 430074, China
[4]Environmental Science Division, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL, 60439, USA
[5]Energy Technologies Area, 1 Cyclotron Road, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA
[6]Energy & Biosciences Institute, University of California, 2151 Berkeley Way, Berkeley, CA, 94704, USA

*Co-corresponding authors
Scown: Email - cdscown@lbl.gov, Phone – 510-486-4507
Mishra: Email - umishra@anl.gov, Phone – 630-252-1108

**Keywords:** Sorghum, Machine Learning, Yield Modeling, Bioenergy, Climate Change, Biofuels

**Abstract**

Crop yield modeling is critical in the design of national strategies for agricultural production, particularly in the context of a changing climate. Forecasting yields of bioenergy crops at fine spatial resolutions can help to evaluate near- and long-term pathways to scaling up bio-based fuel and chemical production, and to understand the impacts of abiotic stressors such as severe droughts and temperature extremes on potential biomass supply. We used a large dataset of 28,364 *Sorghum bicolor* yield samples (uniquely identified by county and year of observation), environmental variables, and multiple approaches to analyze historical trends in sorghum productivity across the U.S. We selected the most accurate machine learning approach (a variation of Random Forest) to predict future trends in sorghum yields under four greenhouse gas emission scenarios and two irrigation regimes. We identified irrigation practices, vapor pressure deficit, and time (a proxy for technological improvement) as the most important predictors of sorghum productivity. Our results showed a decreasing trend of sorghum yields over future years (on average 2.7% from 2018 to 2099), with greater decline under a high greenhouse gas emissions scenario (3.8%) and in the absence of irrigation (4.6%). Geographically, we observed the steepest predicted declines in the Great Lakes (8.2%), Upper Midwest (7.5%), and Heartland (6.7%) regions. Our study demonstrates the use of machine learning to identify environmental controllers of sorghum biomass yield and predict yields with reasonable accuracy. These results can inform the development of more realistic biomass supply projections for bioenergy if sorghum production is scaled up.

## Introduction

Crop yield predictions are crucial to the development of agricultural strategies that maximize output from a limited supply of arable land. While such modeling techniques have historically focused on food crops,[1–3] their applicability to dedicated bioenergy crops is increasingly relevant in view of ambitious renewable energy targets that require substantial increases in domestic biofuel production.[4] Strategies for scaling up cellulosic biofuel production rely on accurate predictions of energy crop yields to forecast long-term feedstock availability near potential biorefineries.[5] These projections must account for future climate scenarios in which severe droughts and extreme temperatures are increasingly common.[6] The ability to model and compare yield variability of different candidate bioenergy crops in response to environmental factors can also inform trait engineering efforts aimed at optimizing crop performance at the field level.[7] Sorghum is particularly interesting to explore in the bioenergy context because of its high biomass yields, drought tolerance, and the availability of historical yield data in the U.S. (while many other potential bioenergy crops are not commercially produced).[8] Previous modeling work has identified drivers of spatio-temporal variability in crop yields, including climatological factors such as temperature and precipitation,[9,10] and soil characteristics including water-holding capacity[11] and clay content.[12]

There are two broad approaches to crop yield simulation: process-based modeling and statistical modeling. Process-based models rely on fundamental mechanisms of plant physiology to simulate crop growth under a given set of resource conditions and management practices.[13] They tend to perform well at the scale of experimental field trial plots but are difficult to validate at larger spatial scales where input parameters are less easily controlled and measured.[14] For

example, agronomic practices may differ across geographic regions and there is rarely sufficient information on farmer behavior to integrate these differences into a process-based yield model.

In contrast to process-based models, data driven statistical models characterize relationships between crop yield and environmental parameters—often climate, soil, and crop management variables—using historical datasets. Multiple linear regression has been the predominant statistical technique of choice, primarily due to its ease of use and interpretability. However, numerous studies have shown this assumption of linear relationships between predictors and dependent variable may not hold when multiple climate and soil variables are used to model crop yields.[15,16]

Nonlinear machine learning (ML) algorithms offer an alternative to conventional statistical modeling methods. The ML approach includes a family of algorithms that do not presume any mechanistic relationships and instead seek to learn a function that best maps input parameters to an output. One subclass of ML techniques is a family of algorithms that are built upon decision trees which can be used for both classification and regression tasks.[17] Breiman et al. (1984) first demonstrated the use of Classification and Regression Tree analysis that now encompasses a broad suite of derived methods, all of which share the decision tree as their fundamental building block. Among these are a variety of ensemble learning techniques, such as Random Forest, which employs bootstrap aggregation to build a strong predictive model by averaging predictions over a collection of individually weak learners.[17] Several studies have explored the predictive capabilities of Random Forest approaches for crop yield modeling[18] demonstrated the use of Random Forest regression for estimating vegetation biomass in wetland systems from WorldView-2 multispectral satellite data, which achieved a 19% reduction in prediction error relative to a baseline stepwise linear regression model. Similarly, it has been

shown that Random Forest approach substantially outperforms multiple linear regression for predicting crop yields of wheat, maize, and potato at both regional and global spatial scales.[12]

Building on these previous studies, we evaluated a number of approaches to predict yields of biomass sorghum as a dedicated bioenergy crop. We then used the best-performing ML approach to predict region-specific bioenergy sorghum yields through 2099 in the United States under four different greenhouse gas emission scenarios, as defined by the Intergovernmental Panel on Climate Change, and two water-supply regimes: rainfed-only and irrigated. The objective of this study is to better understand how different linear and tree-based modeling techniques compare in their ability to predict county-level biomass sorghum yields in the U.S., establish an approach for simulating long-term trends and variability in sorghum yields under different climate scenarios, and use that approach to assess potential future biomass sorghum yields across the U.S.

## Materials and Methods

This study relies entirely on publicly-available historical datasets and future projections, all of which are outlined in Table S1 in the Supplementary Information (SI), along with details about temporal and geographic aggregation methods; a list is also shown in Figure 1. We sought to incorporate as many relevant datasets and predictor variables as possible in the initial regression analyses, including several predictors that may not have direct mechanistic links with yield (e.g. year, latitude, longitude) to evaluate whether they can serve as proxies for parameters that are not otherwise well quantified. However, because the ultimate goal is to develop a predictive model,

we did not incorporate data that could not reasonably be projected through 2099 (e.g. market prices).

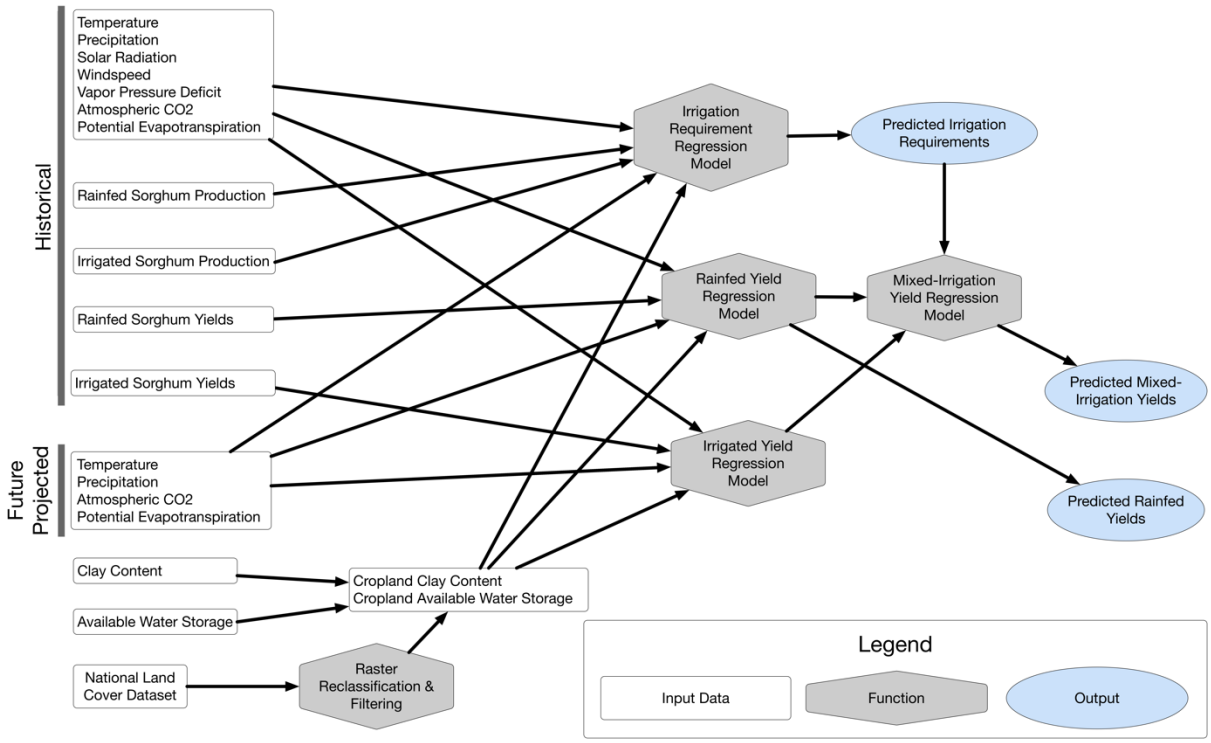


Figure 1. Schematic methodology used in this study for sorghum yield prediction over the continental United States. Greater detail on input variables and data sources can be found in the Supplementary Information.

*Modeling Irrigation Practices*

Particularly important for building bioenergy feedstock production scenarios is the choice of whether to allow for irrigation. Most recent studies assume that all dedicated bioenergy crops will be exclusively rainfed.[5] However, unless regulatory policy will prohibit or disincentivize

irrigation of bioenergy crops, an arguably more realistic approach would be to attempt to predict whether farmers will choose to irrigate, given local conditions. This mixed-irrigation scenario, requires both knowledge of historical irrigation practices across growing regions and an ability to predict when long-term climatic trends may cause farmers to begin or cease irrigating. Historical irrigation practices for sorghum vary over time and space as a function of climate, local market prices for biomass, farmer preferences, and a host of other factors. Of the counties and years for which sorghum yield was reported, 39.9% of county-year data points were accompanied by information as to whether the crop was irrigated. Because the use of irrigation is a crucial input for the predictive model, we chose to systematically populate these missing historical values with estimated breakdowns of irrigated vs. rainfed production.

We began by computing percent production irrigated (PPI) values for every county-year combination for which disaggregated values of irrigated and non-irrigated sorghum production data were available. Because yield is reported at the county level, PPI is equivalent to the fraction of land irrigated in each county. A county's PPI in a given year was defined as the proportion of total sorghum production reported to come from irrigated fields to that county's total sorghum production.[19] Since the irrigation status of reported production values were only specified in less than half (39.9%) of the NASS data, we performed a two-step method for imputing PPI values that could not be computed directly. First, if a county-year specific PPI could not be determined because irrigated and rainfed sorghum production were not reported separately, we used the mean PPI for that county averaged over the values from all years for which PPI could be computed. In water-constrained regions such as the Northern High Plains, evidence suggests that irrigated area varies annually as a function of both precipitation (included in model) and commodity prices (excluded from model), so these average values are an

approximation. If no such mean could be determined for a given county because irrigation practices were not specified in any years, then we computed the inter-annual mean PPI of the state in which the county is located. If the PPI for a county in a specific year could not be determined by these methods due to insufficient data, then no PPI record was included for that county-year combination. The resulting dataset was used to train an Extremely Randomized Trees (ERT) random forest model for predicting PPI as a function of the same of input parameters used in the yield regression analyses. The purpose of building this model was to forecast future changes in irrigation practices in response to the changing climate, which then informed our mixed-irrigation predictive yield model. Additional details about the PPI predictor model can be found in the SI.

*Predicting Future Agronomic Practices and Sorghum Yields*

To predict future sorghum yields for the years 2018-2099, we explored a set of possible climate scenarios corresponding to the Intergovernmental Panel on Climate Change (IPCC) Representative Concentration Pathways (RCPs) 2.6, 4.5 and 6.5 and 8.5. Future temperature and precipitation data at 30 km resolution were obtained from the NASA Earth Exchange downscaled climate projections (NEX-DCP30) which provides estimates from an ensemble of Global Circulation Models (GCMs) conducted under the Coupled Model Intercomparison Project Phase 5 (CMIP5). We obtained future predictions of atmospheric $CO_2$ concentrations two different Carbon Cycle Models, BERN and ISAM, at ten year intervals from 2020 to 2100.[20] Historical and future Potential Evapotranspiration (PET) values were obtained from the Environmental Protection Agency's EnviroAtlas database.[21]

To explore yield variations related to irrigation practices, we used two separate ERT models to simulate separate irrigation regimes: a rainfed-only model in which we assumed that no irrigation will be used for biomass sorghum production, and a mixed-irrigation model in which farmers' choices to irrigate were predicted using the ERT model described above. Our decision to use ERT random forest model was based on a comparative analysis of several linear and tree-based modeling approaches which revealed ERT to be the highest performing. Details of this supporting analysis can be found in the Supplementary Information section. The fundamental difference between our two ERT yield models lies in the historical data used to train each one. For the rainfed model, we restrict historical training data to yield samples explicitly reported as coming from rainfed fields. The mixed-irrigation model makes use of all available data, but also accounts for potential shifts in farmers' behavior in response to climate change (reflected in predicted shifts in PPI). For example, farmers in a county where the growing-season precipitation falls below the minimum requirement for rainfed sorghum (11-14 inches[22]) by 2099 may decide to irrigate, even if sorghum in that county has historically been all or mostly rainfed.

For the mixed-irrigation model, we first estimated the proportion of total sorghum tonnage produced on irrigated fields for each county in each year using the PPI predictor model described above. We then trained an ERT model on all historical yield data and used it to predict irrigated yield value for every future year. Additionally, we used the rainfed model to predict a rainfed yield value. Finally, to determine a mixed-irrigation yield value for each prediction sample, we take the mean of the rainfed and irrigated predictions, weighted by the PPI to approximate a mixed-irrigation yield value; that is, a value for counties in which only a portion of total sorghum production is likely to be irrigated. One challenge for rainfed-only predictions is that, in some arid regions of the U.S., little or no yield data exists for rainfed crops because

irrigation is a necessity. To avoid unnecessary extrapolation, we constrained yield predictions in the rainfed scenario to counties for which historical rainfed sorghum yield values have been reported, or the county's total precipitation during the growing season in the prediction year is greater than 10 inches. This cutoff was set based on literature values for the minimum rainfall requirement for successful sorghum cultivation.[22]

## Results

### *Evaluating Predictive Performance and Relative Variable Importance*

Before applying the ERT models to generate future sorghum yield predictions, we tested and applied that model in a predictive capacity. To test the ERT model's extrapolation capabilities over time (for future yield predictions), we used a temporal cutoff to split our data into calibration and validation sets. Specifically, data from 1958 through 1987 was used for model training, and data from 1988 through 2016 was set aside as the test set for validation. We adopted this approach to evaluate model performance in a manner that best resembled the ultimate use of our model, which is to predict sorghum yields over a contiguous span of future years through 2099, using all available historical data from the preceding years for model training. The final ERT model is, of course, trained on data sampled from the full 1958-2016 timeframe. Figure 2 shows the results of temporally-binned calibration and validation. Validation predictions closely tracked the average trend of observed yields from 1987 to 1995, while generally tending to underestimate empirical values from 1995 onwards.

The under-prediction after 1995 may be partially related to fluctuations in market prices and the impacts on farmer behavior; for example, the market price of grain sorghum nearly doubled between 1994 and 1996 and spiked again in 2007-2008 and 2010-2013.[23] The cause-and-effect in the relationship between yields and market price can vary; farmers may respond to high prices by applying additional fertilizer and irrigation or high prices may be a consequence of a drop in yield due to drought or disease. If market prices increase because of speculation about increasing demand, for example, this could incentivize farmers to irrigate and fertilize more liberally, thus resulting in a yield increase. The latter is likely what occurred in 2012 when market prices were at a near-historic high and yields dropped (see Figure 2).[24] Because market prices cannot be predicted through 2099 with any reasonable confidence, we decided to exclude market prices from the final model, but did explore its impact in a sensitivity analysis as discussed further below. Another factor that our model does not currently capture is changes in quantities of water and fertilizer applied to sorghum because of limited historical data availability. The quantity of nitrogen applied to corn, assuming it is an appropriate proxy, has increased by approximately 10% since 1995 after remaining steady for the previous two decades.[25]

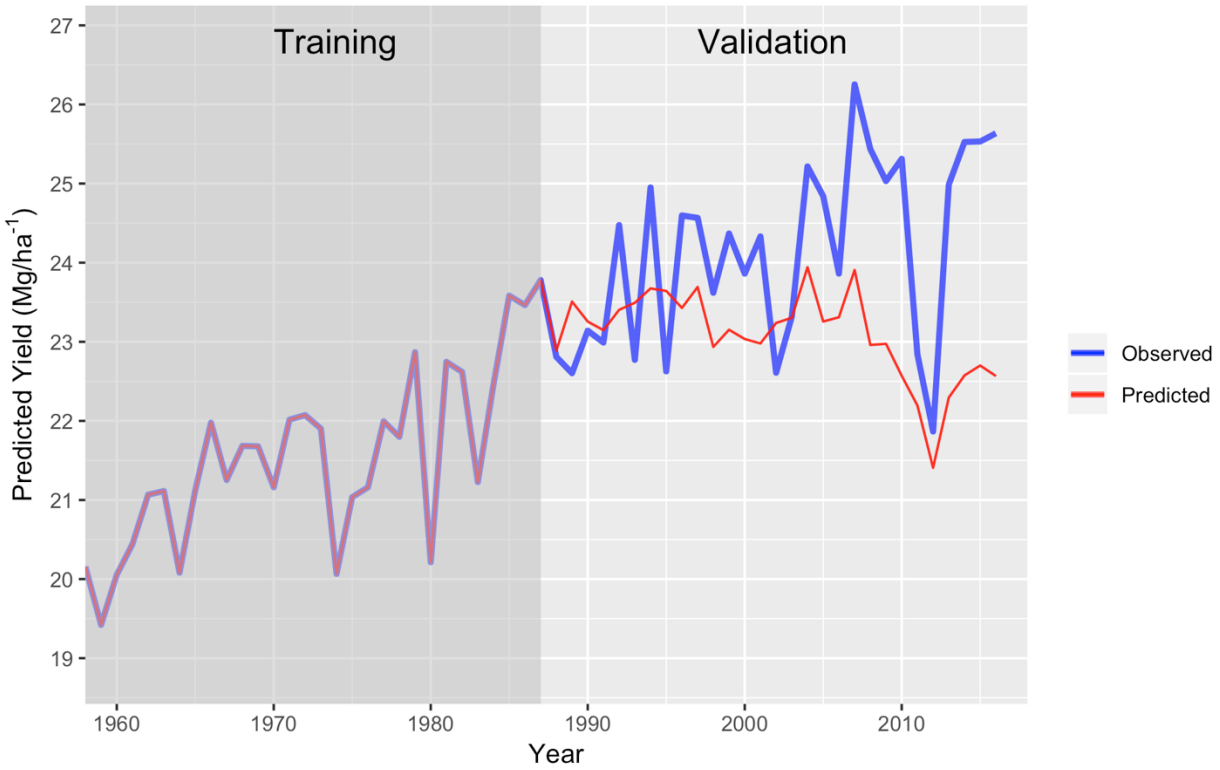


Figure 2. Time series of sorghum yields by year, showing both observed and predicted values using the Extremely Randomized Trees model. Values along the Y-axis represent averages of all observations/predictions in the corresponding year.

Without more granular data on agronomic practices, possible phenotypic improvements from breeding efforts, or even changes in how farmers report yield data, it is difficult to explain differences between model predictions and actual yields with certainty. Extreme events that are not well captured in our modeling approach can impact yields. Tornadoes and flooding can either destroy crops or shorten the growing season by delaying fall planting dates. Even if we tuned the historical model to identify incidences of flooding, incorporating these effects into the future projections based on CMIP5 climate modeling results would be near-impossible. Conversely, the impacts of drought are fairly well captured in our approach; it is notable that the model successfully predicted the dip in yield due to the U.S. drought of 2012.[24]

Variable importance values help to reveal which parameters are driving the model's predictions. These values were calculated from Gini impurities for the top fifteen ranked predictors (Figure 3). Unsurprisingly, whether the crop is irrigated or rainfed was ranked as the most important factor in predicting sorghum yield, followed by vapor pressure deficit (difference between vapor pressure and saturation vapor pressure in the air) during the third month of the growing season, year, and longitude. Year, latitude, and longitude are, of course, not directly linked to yield in any sort of mechanistic way, but the ERT model has identified them as useful proxies for variables not otherwise captured in the input data. For example, the high importance of the y ear variable likely reflects the impact of technological advances in agronomic practices, intensification (including increases in fertilizer application), and crop breeding efforts which have improved yields over the long-term.[26–28] The Supplementary Information includes additional discussion of alternative strategies to capturing these long-term increases, such as detrending. The top ten ranked predictors accounted for a bulk of the sum total of all importance values, and lower-ranked variables all shared similar scores of around 0.012.

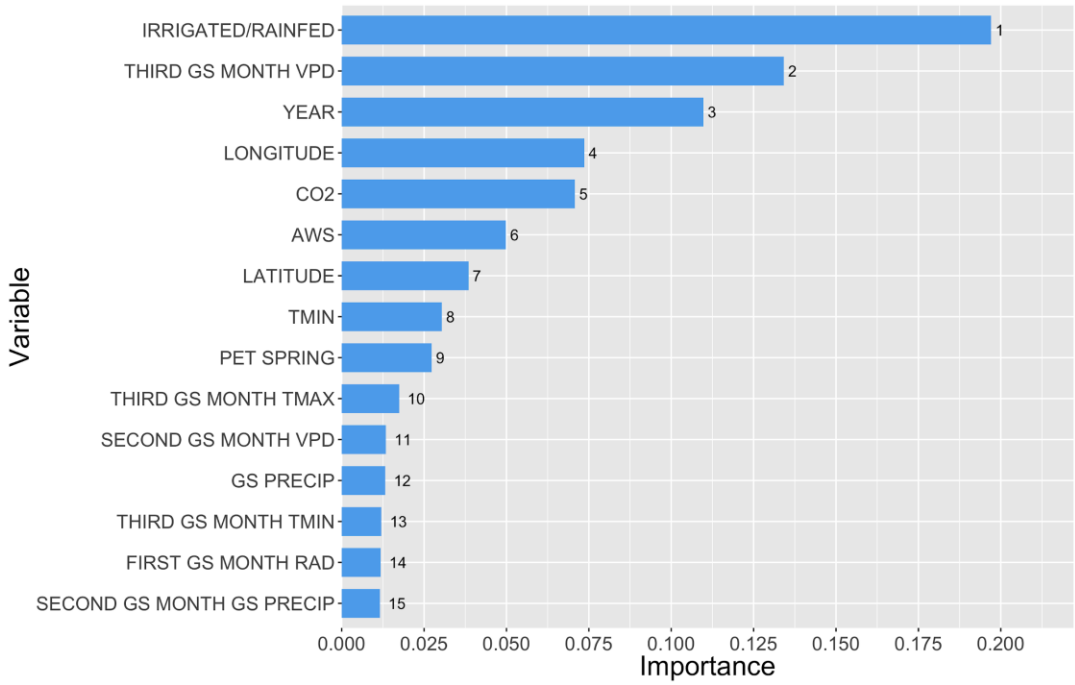

Figure 3. Variable importance values (calculated by Gini impurities) of top fifteen predictors in the Extremely Randomized Trees model. Importance values of all variables sum to one, and variables not shown all have similar importance values of around 0.012. Numbers to the right of each bar denote the importance ranking of each variable. Variable name abbreviations used along the Y-axis are provided in Table S1.

We note that there are many environmental, agronomic and economic variables may have not been considered in our model which likely play a role in determining annual sorghum yields. Our selection of input parameters was largely based upon extensive review of the literature and data availability at the spatial (county) and temporal (annual) scale necessary for our analysis. We conducted a sensitivity analysis which considers the market price of grain sorghum as an additional input parameter to ERT model. The inclusion of price as a predictor did not improve predictive accuracy and when variable importance values were computed for this new model (Figure S6), price was ranked fairly low compared to pre-existing input parameters, suggesting a limited influence of market forces on yields as the other variables in our model. Interestingly, we observe some substantial changes in the relative importance values of other variables in the model with price included. Notably, the importance of Year is significantly lower (possibly explained by collinearity with sorghum price) and no longer appears in the top 15 highest ranked predictors. Additionally, we see a reduction in the importance of Vapor Pressure Deficit in the third month of the growing season a relative increase in the importance of latitude and longitude. These changes likely reflect a combination of multicollinearity between input parameters and the non-deterministic nature of the ERT algorithm. Consequently, we are wary of drawing mechanistic inferences about input parameters from these importance values alone.

*Generating Future Sorghum Yield Projections*

ERT model predictions for the years 2018 to 2100 indicate a gradual decreasing trend in sorghum yields across the U.S. As expected, the rainfed-only scenario results in Figure 4 indicate that yields will be comparatively low in the western U.S. if farmers do not irrigate, regardless of year or climate scenario. The rainfed predictions are only provided for counties where historical USDA yield data exists that is explicitly labeled as rainfed (many yield data points are not accompanied by information on irrigation practices). These same regions where rainfed yields are low, or insufficient rainfed yield data exists, are the highest-yielding when irrigation practices are unconstrained. The decreases in yield through 2099 become more severe in higher-GHG-emissions scenarios (Figure S7). This pattern holds under both rainfed-only and mixed-irrigation regimes, although there is considerable regional variation in the magnitude of the yield declines (Figure 5). In particular, we see the most pronounced yield declines projected in the Great Lakes (8.2%), Upper Midwest (7.5%, Heartland (6.7%), Northeastern (5.6%), Northern Plains and Eastern Mountain (5.5%) with moderate to stable trends in the Southern (4.1%), Delta (4%), Southern Plains (2.3%) and Pacific (0.7%) regions from 2018 to 2099 and a slight upward trend in the Mountain region (+2.3%) (Figure 5).

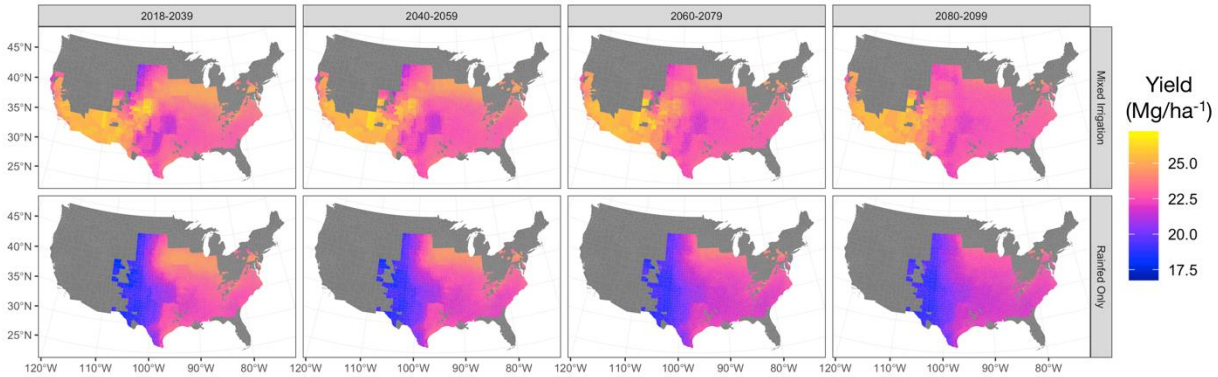


Figure 4. Geographic distributions of ERT predicted sorghum yields under mixed-irrigation and rainfed scenarios, assuming the highest emissions pathway (RCP 8.5). Input values for precipitation, minimum and maximum temperature are CMIP5 ensemble means.

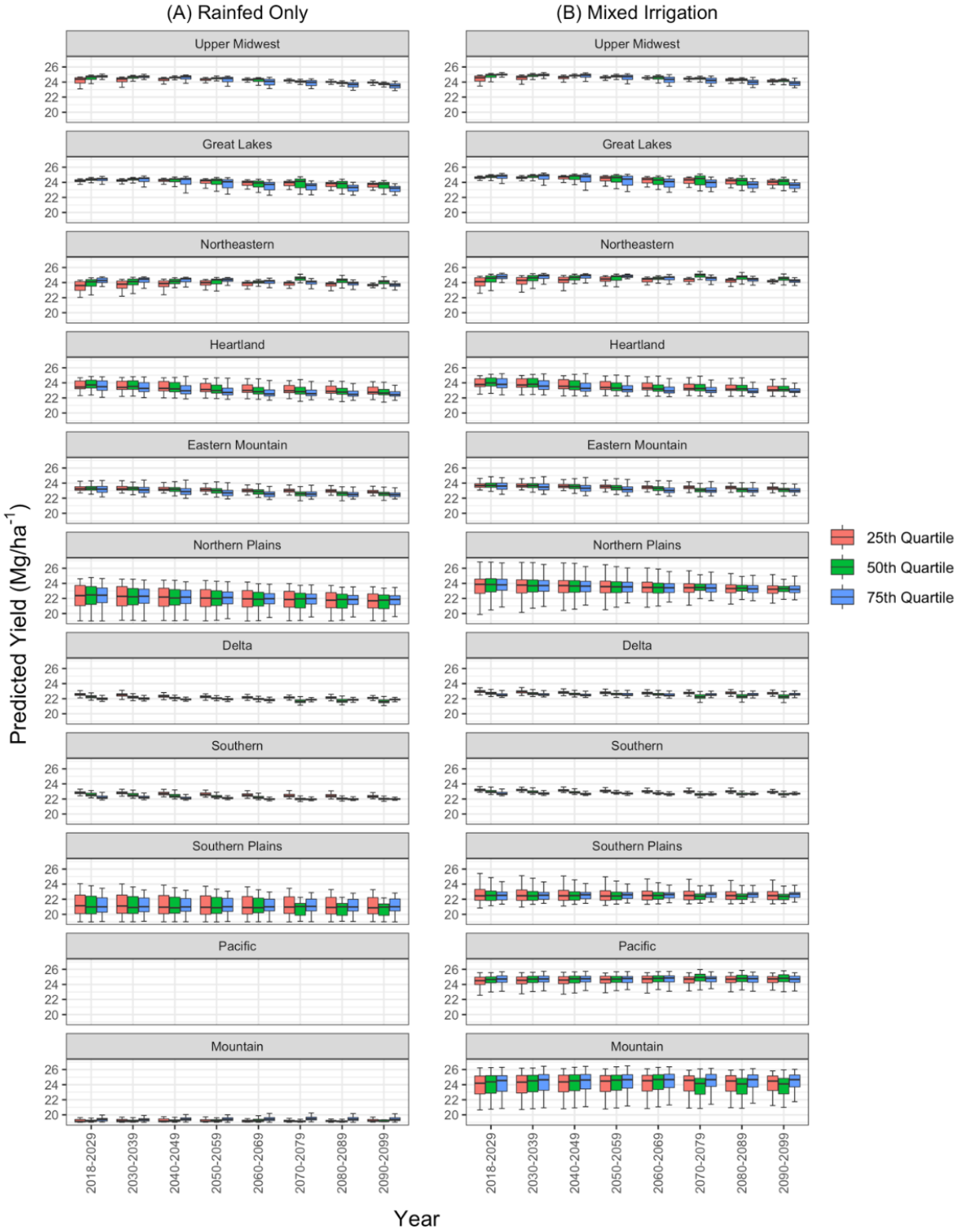
(A) Rainfed Only
(B) Mixed Irrigation
Upper Midwest
Great Lakes
Northeastern
Heartland
Eastern Mountain
Northern Plains
Delta
Southern
Southern Plains
Pacific
Mountain
Predicted Yield (Mg/ha$^{-1}$)
Year
2018-2029
2030-2039
2040-2049
2050-2059
2060-2069
2070-2079
2080-2089
2090-2099
25th Quartile
50th Quartile
75th Quartile

Figure 5. Distributions of predicted sorghum yields over future years, using the 25th, 50th and 75th quartiles of CMIP5 ensemble values for input climate variables under (A) rainfed and (B) mixed irrigation scenarios. An emissions scenario of RCP 8.5 is assumed. Values in each boxplot are ten-year averages taken over the years corresponding to its position along the X-axis.

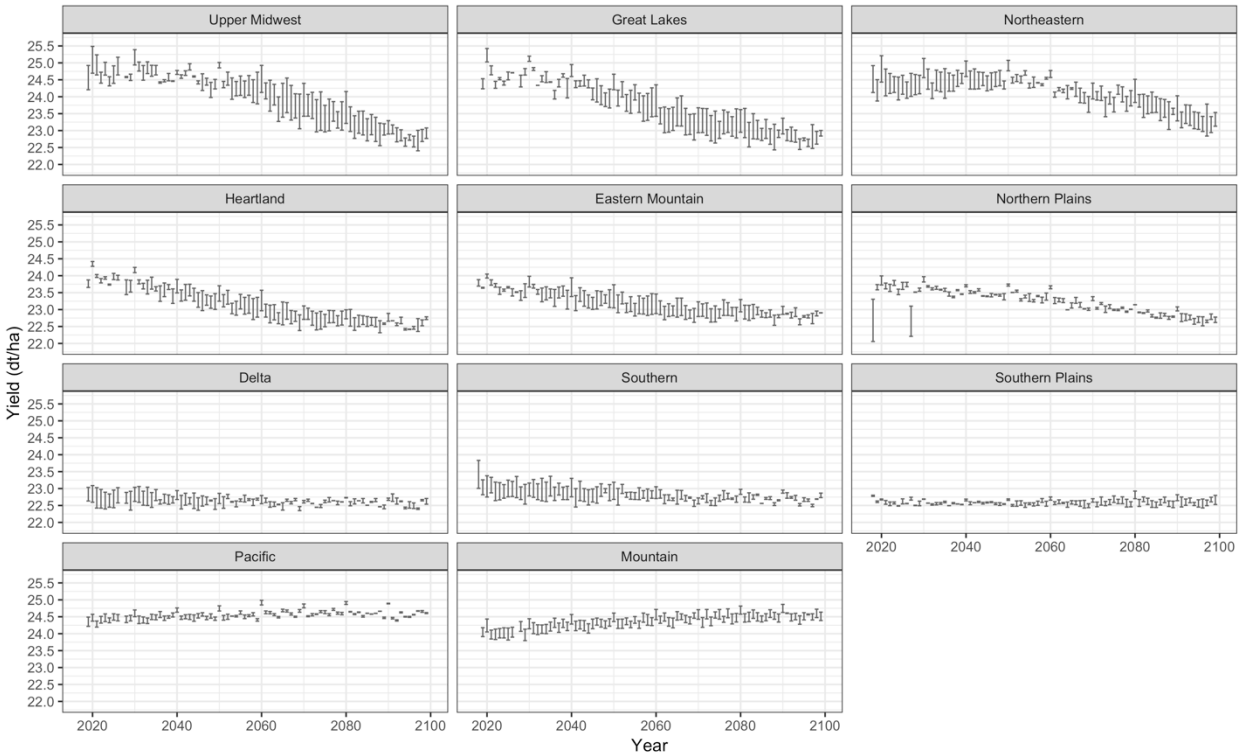


Figure 6. Sensitivity of predicted sorghum yields (assuming mixed irrigation and the RCP 8.5 scenario) to climatological input parameters. The lower and upper bounds on each bar correspond to yield predictions made using the 25th and 75th quartile values respectively of precipitation, minimum and maximum temperature values from the CMIP5 ensemble of global circulation models for the corresponding year along the X-axis.

The differences in yield trends across regions are best understood in the context of the CMIP5 ensemble values for input climate values, and the sensitivity of yield to those values. The Pacific and Mountain regions have the least-negative yield trends (a near-negligible decrease in Pacific and a 2.3% increase in Mountain). These two regions are unique in that the climate projections show a curious and abrupt increase in TMIN values during the growing season (see Figure S4). Most regions exhibit a slight downward shift in the range ($25^{th}$-$75^{th}$ percentile) of

growing season precipitation values (Figure S5). We also noted substantial regional variation in the sensitivity of yields to region-specific input climate parameters. Yield predictions made using the 25$^{th}$, 50$^{th}$, and 75$^{th}$ quartiles of CMIP5 ensemble values for input climate variables revealed that western regions (Mountain and Pacific) were least sensitive to the quartile of climate data used, while all other geographic regions exhibited greater variation in quartile-specific distributions (Figure 5; Figure 6). The greatest ranges in quartile-specific yield predictions corresponded to the Great lakes, Heartland and Upper Midwest, and Northeastern regions (Figure 5; Figure 6).

The narrowing range of projected yield distributions for each 10-year period, especially in the Midwest and Northern and Southern Plains (Figure 5), is likely attributed to the convergence of our ERT model as values of input climate parameters become geographically homogenous at the county-level. That is, as regional climatic values became less dissimilar under future emissions scenarios, we observed diminishing regional variability in corresponding yields. In reality, extreme events that are not captured well in long-term climate projections, such as droughts, flooding, and storms, are likely to drive year-to-year variations in yield. In particular strong storms and high winds that result in lodging are not well captured in our model, although modern farming equipment can be used to harvest lodged biomass, so the impact of lodging on yields is unclear.

## Conclusion

Our results indicate that tree-based machine learning algorithms, specifically Random Forest and Extremely Randomized Trees regression, can offer robust predictive performance for

bioenergy crop modeling. The RF and ERT models evaluated in our study outperformed linear prediction techniques on the basis of validation prediction errors using a temporal split of historical yield records for training and validation. The superior performance of the RF and ERT models can likely be attributed to the non-linear relationships between our predictor variables and yield, and the likelihood of complex interactions between predictors. Specifically, the tree-building subroutine, which recursively selects the single best predictor at each tree node, forces a given decision tree to be affected by only one variable at a time, circumventing issues of multicollinearity and interactions. Although this results in individual decision trees that perform weakly on their own, the aggregate performance as a forest is typically far stronger. Thus, the ensembling component of these algorithms enables them to account for nuanced effects of different predictors captured in different trees and merge them together into robust consensus estimates.

The spatial variation and magnitude of sorghum yields predicted by our ERT model generally align with those of other regional-scale studies, although the regional variability in yields predicted by mechanistic models is substantially greater. For example, Lee[29] generated yield potential maps for rainfed biomass sorghum across the continental U.S. by combining field trial data from six states over five years and a statistical-mechanistic framework called the Parameter-elevation Regressions on Independent Slopes Model (PRISM). PRISM uses temperature (maximum, mean, and minimum), total precipitation, evapotranspiration, soil evaporation (averaged over over the years 1981 to 2010) and soil properties (salinity, pH, and drainage) to simulate spatially heterogeneous biomass yields. The model also incorporates crop coefficient, average rooting depth, and water stress response factor. Based on these inputs, the model predicts yields for the years 2009 to 2015. Our ML approach uses a number of similar

soil and climate datasets that affect crop growth, but does not include all the same heterogenous soil conditions (our ML approach uses only available water storage and clay content). Salinity, pH, and drainage were not included due to limited data availability at the necessary spatial scale and coverage for our analysis. Our review of the literature found that these soil properties were not typically used in other statistical and machine learning approaches to yield prediction.[9,12,16] Ideally, all available soil data should be incorporated in future models. Nevertheless, the near-term predictions of our rainfed ERT model spatially align those of PRISM, similarly predicting the highest rainfed sorghum yields in the Upper Midwestern Region, with values ranging from 18 – 22 Mg $ha^{-1}$. However, it is worth noting that, while our ERT yield predictions outside of the Upper Midwest region do not drop below 17 Mg $ha^{-1}$, the PRISM-ELM model predicts much lower value ranges for all other regions. This is most pronounced in the Pacific, Mountain, Northern and Southern Plains regions where PRISM-ELM estimates yields as low as 1 to 10 Mg $ha^{-1}$.

Another comprehensive study prior to ours that offers predictions of biomass sorghum yield is the U.S. Department of Energy Billion Ton Report[5] which utilized a version of PRISM developed expressly for use in that study: PRISM Environmental-Model (PRISM-EM). PRISM-EM was used to estimate biomass sorghum yield within a complex agro-economic and environmental simulation framework named the Policy Analysis System (POLYSYS) for a variety of scenarios.[5] PRISM-EM combines biweekly precipitation and temperature values from PRISM, soil pH, drainage, and salinity data from the Soil Survey Geographic (SSURGO) Database's soil pH, drainage, and crop-specific water-use and temperature-tolerance relationships to estimate yields.[5]

Despite differences in input parameters, our ERT model and the Billion Ton Study version of the POLSYS/PRISM-EM model (Figure S8) output similar in yield predictions across the Eastern, Upper Midwest and Great Lakes regions. However, where POLYSYS/PRISM-EM predicts higher and lower extremes in other regions, our ERT model maintains fairly consistent mid-range values. This regional homogeneity of mid to high yield values over large swaths of the U.S. is projected into future years by the ERT model as seen in Figure 5. A model based on historical data such as ours will not have trained on records of rainfed yields as low as those predicted by PRISM-ELM or POLYSYS/PRISM-EM; any land yielding as low as 1-10 Mg $ha^{-1}$ of sorghum biomass under rainfed conditions will either not be planted at all or will be irrigated to improve yields. It is possible that lower-than-expected crop yields in the ranges predicted by PRISM-ELM and POLYSYS/PRISM-EM for the Western U.S. do occur, and in that case (for grain types only), farmers may choose to collect crop insurance payments. In this case, it is likely that yields would not be reported to USDA, as the crop must then be destroyed or put to an alternative use.

If the higher rainfed yields predicted in our model for the Southern Plains, Northern Plains, and Mountain regions can be achieved, this could have beneficial land-use change implications. For example, relatively high-yielding biomass sorghum across a wide geographic swath of the U.S., particularly in regions that are considered marginal based on their climate, allows for higher-value agricultural land to be cultivated with more profitable crops. In other words, if sorghum will grow relatively well across a larger area than has been predicted by previously-developed models, a target amount of gross biomass tonnage can be achieved using a more diverse range of land conversion strategies. However, if mid-to-low-yielding land is prioritized for growing sorghum, an obvious outcome will be a larger total land area requirement.

Depending on the geographic distribution and value of such lands—both economic and agricultural—this may or may not noticeably increase logistics costs for delivering biomass to biorefineries.

Our analysis contributes to a growing field of biomass yield prediction informed by historical data and real-world agronomic practices and lays the foundation for follow-up work on scaling up bioenergy crop production. Further analysis is needed to compare and understand the performance differences between tree-based ML algorithms and process-based models for predicting sorghum yields. The interplay between commodity prices and agricultural yields is also largely ignored in bioenergy crop yield projections. Higher market prices will incentivize farmers will apply additional fertilizer and irrigation water. Ultimately, the most robust predictive yield models will likely emerge from a hybrid of statistical and process-based methods. Linking these two modeling paradigms will be a critical next step to improve the accuracy, precision, and interpretability of bioenergy crop models. Additionally, incorporating mechanistic aspects of process-based modeling will better enable for targeting specific traits and underlying genotypes that promote better crop performance under future climate scenarios.

## Acknowledgements

This work was part of the DOE Joint BioEnergy Institute (http://www.jbei.org) supported by the US Department of Energy, Office of Science, Office of Biological and Environmental Research, through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the US Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a

non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. We would like to thank Daniel Putnam and Jeff Dahlberg for their valuable feedback on our analysis.

## References

1. Cerrato, M. E. & Blackmer, A. M. Comparison of models for describing; corn yield response to nitrogen fertilizer. *Agron J* **82,** 138 (1990).
2. Parry, M. L., Rosenzweig, C., Iglesias, A., Livermore, M. & Fischer, G. Effects of climate change on global food production under SRES emissions and socio-economic scenarios. *Glob. Environ. Change* **14,** 53–67 (2004).
3. Lobell, D. B. & Field, C. B. Global scale climate-crop yield relationships and the impacts of recent warming. *Environ. Res. Lett.* **2,** 014002 (2007).
4. Dale, B. E. *et al.* Take a closer look: biofuels can support environmental, economic and social goals. *Environ. Sci. Technol.* **48,** 7200–7203 (2014).
5. Langholtz, M. H., Stokes, B. J. & Eaton, L. M. *2016 Billion-Ton Report: Advancing Domestic Resources for a Thriving Bioeconomy*. (EERE Publication and Product Library, 2016). doi:10.2172/1271651
6. Sillmann, J., Kharin, V. V., Zwiers, F. W., Zhang, X. & Bronaugh, D. Climate extremes indices in the CMIP5 multimodel ensemble: Part 2. Future climate projections. *J. Geophys. Res. Atmos.* **118,** 2473–2493 (2013).
7. Heslot, N., Akdemir, D., Sorrells, M. E. & Jannink, J.-L. Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. *Theor. Appl. Genet.* **127,** 463–480 (2014).
8. Cui, X., Kavvada, O., Huntington, T. & Scown, C. D. Strategies for near-term scale-up of cellulosic biofuel production using sorghum and crop residues in the U.S. *Environmental Research Letters* (2018). doi:10.1088/1748-9326/aae6e3
9. Ray, D. K., Gerber, J. S., MacDonald, G. K. & West, P. C. Climate variation explains a third of global crop yield variability. *Nat. Commun.* **6,** 5989 (2015).
10. Deryng, D., Conway, D., Ramankutty, N., Price, J. & Warren, R. Global crop yield response to extreme heat stress under multiple climate change futures. *Environmental Research Letters* **9,** 034011 (2014).
11. Li, R. *et al.* Effects on soil temperature, moisture, and maize yield of cultivation with ridge and furrow mulching in the rainfed area of the Loess Plateau, China. *Agricultural Water Management* **116,** 101–109 (2013).
12. Jeong, J. H. *et al.* Random forests for global and regional crop yield predictions. *PLoS One* **11,** e0156571 (2016).
13. Raes, D., Steduto, P., Hsiao, T. C. & Fereres, E. AquaCrop—the FAO crop model to simulate yield response to water: II. Main algorithms and software description. *Agronomy Journal* **101,** 438–447 (2009).
14. Lobell, D. B. & Asseng, S. Comparing estimates of climate change impacts from process-based and

statistical crop models. *Environmental Research Letters* **12,** 015001 (2017).
15. Lobell, D. B., Bänziger, M., Magorokosho, C. & Vivek, B. Nonlinear heat effects on African maize as evidenced by historical yield trials. *Nat. Clim. Chang.* **1,** 42–45 (2011).
16. Schlenker, W. & Roberts, M. J. Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. *Proc. Natl. Acad. Sci. USA* **106,** 15594–15598 (2009).
17. Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. Classification and Regression Trees, The Wadsworth Statistics and Probability Series, Wadsworth International Group, Belmont California (pp. 356). (1984).
18. Mutanga, O., Adam, E. & Cho, M. A. High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. *International Journal of Applied Earth Observation and Geoinformation* **18,** 399–406 (2012).
19. Deines, J. M., Kendall, A. D. & Hyndman, D. W. Annual Irrigation Dynamics in the U.S. Northern High Plains Derived from Landsat Satellite Data. *Geophys. Res. Lett.* **44,** 9350–9360 (2017).
20. Solomon, S., Qin, D., Manning, M., Averyt, K. & Marquis, M. *Climate change 2007-the physical science basis: Working group I contribution to the fourth assessment report of the IPCC*. **4,** (Cambridge university press, 2007).
21. Pickard, B. R., Daniel, J., Mehaffey, M., Jackson, L. E. & Neale, A. EnviroAtlas: A new geospatial tool to foster ecosystem services science and resource management. *Ecosystem Services* **14,** 45–55 (2015).
22. Martin, J. H. Climate and sorghum. *Yearbook of agriculture. USDA, Washington, DC* 343–347 (1941).
23. The World Bank. World Bank GEM Commodities. (2019). at <https://datacatalog.worldbank.org/dataset/gem-commodities>
24. Rippey, B. R. The U.S. drought of 2012. *Weather and Climate Extremes* **10,** 57–64 (2015).
25. United States Department of Agriculture, E. R. S. Fertilizer Use and Price. (2019). at <https://www.ers.usda.gov/data-products/fertilizer-use-and-price.aspx>
26. Evans, L. T. The natural history of crop yield: A combination of improved varieties of crop plants and technological innovations continues to increase productivity, but the highest yields are approaching limits set by biological constraints. *Am. Sci.* **68,** 388–397 (1980).
27. Tester, M. & Langridge, P. Breeding technologies to increase crop production in a changing world. *Science* **327,** 818–822 (2010).
28. Lobell, D. B. & Asner, G. P. Climate and management contributions to recent trends in U.S. agricultural yields. *Science* **299,** 1032 (2003).
29. Lee, D. K. *et al.* Biomass production of herbaceous energy crops in the United States: field trial results and yield potential maps from the multiyear regional feedstock partnership. *Glob. Change Biol. Bioenergy* **10,** 698–716 (2018).
30. Daly, C., Halbleib, M. D., Hannaway, D. B. & Eaton, L. M. Environmental limitation mapping of potential biomass resources across the conterminous United States. *Glob. Change Biol. Bioenergy* **10,** 717–734 (2018).