ENVIRONMENTAL RESEARCH
LETTERS

LETTER • OPEN ACCESS

Machine learning based estimation of land productivity in the contiguous US using biophysical predictors

To cite this article: Pan Yang *et al* 2020 *Environ. Res. Lett.* **15** 074013

Environmental Research Letters

LETTER

Machine learning based estimation of land productivity in the contiguous US using biophysical predictors

Pan Yang[1,2], Qiankun Zhao[1,2] and Ximing Cai[1,2]

[1] Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, United States of America
[2] DOE Center for Advanced Bioenergy and Bioproducts Innovation, University of Illinois at Urbana-Champaign,Urbana, IL, United States of America

E-mail: xmcai@illinois.edu

Keywords: land productivity, marginal land, land use, machine learning

Supplementary material for this article is available online

OPEN ACCESS

RECEIVED
17 November 2019

REVISED
2 April 2020

ACCEPTED FOR PUBLICATION
3 April 2020

PUBLISHED
22 June 2020

Abstract

Estimation of land productivity and availability is necessary to predict land production potential, especially for the emerging bioenergy crop production, which may compete land with food crop production. This study provides land productivity estimates in the contiguous United States (CONUS) through a machine learning approach. Land productivity is defined as the potential in producing agricultural outputs given biophysical properties including climate, soil, and land slope. The land productivity is approximated by the potential yields of six major crops in the CONUS, i.e. corn, soybean, winter wheat, spring wheat, cotton, and alfalfa. This quantitative relationship is then applied to estimating the availability of marginal land for bioenergy crop production in the CONUS. Furthermore, the levels of uncertainties associated with land productivity and marginal land estimates are quantified and discussed. Based on the modeling results, the total marginal land of the CONUS ranges 55.0–172.8 mha, but the 95% inter-percentile distance of the estimated productivity index reaches up to 60% of its expected value in data-scarce regions. Finally, in a cross-check analysis, marginal lands estimated based on biophysical criteria are found to be comparable to those based on an economic criterion.

1. Introduction

The high precision of agricultural management nowadays calls for accurate estimation of land productivity, i.e. the capability of a piece of land in supporting agricultural production based on its biophysical conditions (e.g. climate, land slope, and soil condition) [1, 2]. A pressing need of land productivity assessment is the identification of marginal land that is not highly productive (that should be used to produce food crops for the humanity) but can be used for growing dedicated bioenergy crops (e.g. miscanthus, switchgrass, and sweet sorghum). Existing studies that estimate land productivity and land availability for biomass production are mostly subject to significant uncertainties due to incomplete data [3, 4], insufficient resolution [5], and even some unknown factors [5, 6]. This study provides land productivity estimates in the contiguous United States (CONUS) through a machine learning approach, based on which marginal lands available for bioenergy production are estimated using both biophysical and economic criteria.

A simple method to estimate land productivity is to classify the land through simple soil taxonomy rules (i.e. empirical knowledge about land productivity and soil taxonomy features). Although the resulting land productivity index from even such a simple method is proved to correlate well with the observed county level crop yield [3], uncertainty treatment in the estimation requires advanced methods [6–8]. For example, national commodity crop productivity index (NCCPI) is another land productivity index that is derived through fuzzy logic models [4], which estimates the land productivity through a set of fuzzy if-then rules based on empirical knowledge and has been widely used for deriving land use decisions [9, 10]. Limitations on uncertainty treatment exist in previous studies, including subjective components (e.g. the subjective choice of membership function and its parameter values in a fuzzy logic model [7]), unrecognized uncertainties in the input data for

Table 1. Datasets used in this study.

Data	Databases	Spatial resolution	Time period	Sources
Soil	Gridded Soil Survey Geographic (gSSURGO)	10 m	2017	USDA [15]
Slope	Global Terrain Slope (GTS)	30 arc-second	2006	Fischer *et al* [16]
Temperature and precipitation	Daymet data set	1 km	2008–2017, monthly	Oak Ridge National Laboratory (ORNL) Distributed Active Archive Center (DAAC) [17]
Evapotranspiration	NCEP North American Regional Reanalysis	32 km	2008–2017, yearly	NOAA/OAR/ESRL [18]
Land cover	Cropland Data Layer (CDL)	30 m	2008–2017, yearly	USDA [14]
Land cover	National Land Cover Database (NLCD)	30 m	2016	Multi-Resolution Land Cover Characteristics (MRLC) Consortium [19]
Irrigation	MIrAD-US	250 m	2012	USGS [20]
Gross Primary Productivity (GPP)	MODIS 250 productivity	250 m	2008–2017, yearly	Robinson *et al* [21]
Crop price, crop yield[a], land rent	NASS Surveys	County/state	2008–2017, yearly	NASS [22]
Crop specific production cost	USDA Economic Research Service (ERS)	US/major production region	2008–2017, yearly	USDA [23]

[a]In this paper, for grain crops, yield refers to the yield of grain; for alfalfa, yield refers to above-ground biomass.

generating those indices [11–13], and complex relations between uncertainty sources in the land observation data (e.g. correlation between soil and land slope [6]).

To address these limitations, this study updates our previous research on fuzzy logic models based land productivity [6] through a set of specially designed ML models, which (i) avoid relying on empirical knowledge and subjective judgements that can occur with the use of a completely data-driven approach; (ii) provide an estimate of the uncertainty associated with the ML model prediction; and (iii) automatically extract the information embedded in data by assuming no pre-defined relationship between the predictors and the predictand. In addition, this study is benefited from updated remote sensing data with improved spatial resolutions (up to ~250 m resolution, see table 1). We use long-term yields of six major crops (corn, soybean, winter wheat, spring wheat, cotton, the five most grown crops in CONUS [14]; and alfalfa, a major crop in high slope regions [14]) in the contiguous United States (CONUS) as proxies of land productivity, and use machine learning (ML) algorithms to estimate the potential yields of these six major crops and their associated uncertainties at ~250 m resolution. A productivity index is then derived based on the estimated long-term average potential yields of the six major crops during 2008–2017. To the best of our knowledge, this is the first productivity index that links directly to the actual crop yields through advanced ML techniques. This quantified relationship together with a set of biophysical and economic criteria are applied to estimating potential land available for bioenergy crop production (here referred as 'marginal land') at ~250 m resolution in the CONUS. Overall, this study is expected to contribute to the existing literature of agricultural land management and biofuel development in the following aspects: (i) accurate estimation of land productivity through a machine learning approach, (ii) improved identification of marginal lands for producing bioenergy crops, and iii) reliable quantification of uncertainty involved in land productivity and marginal land estimations.

2. Methods

2.1. Model overview

Figure 1 provides a general flow chart for deriving the land productivity index and potential marginal lands. The first step is to collect and preprocess spatial data from table 1. Detailed characteristics of data from table 1 and their preprocessing procedures are introduced in Section I in the supporting information (SI), while their usage in this study is introduced in the subsequent sections. All the spatial data are projected to the World Geodetic System (WGS) 84 coordinate and resampled to 250 m resolution using the 'Bilinear' (for continuous data) and 'Majority' (for categoric data) resampling methods in ArcGIS [24]. In addition, all time-series of the data in table 1 are collected during 2008–2017, and their average values over these years are calculated and used in the ML models.

We then adopt a two-step ML approach for each of the six major crops: we first downscale the

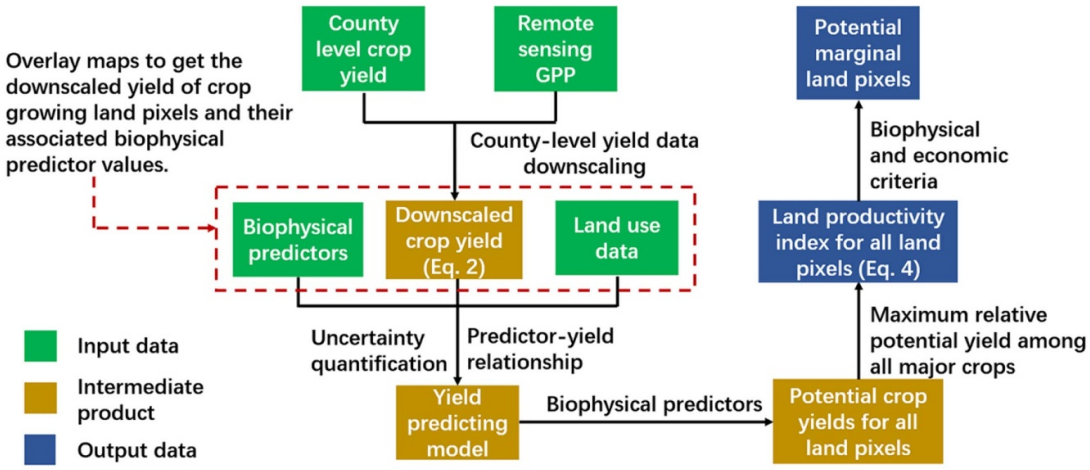

Figure 1. Flow chart for deriving land productivity index and identifying potential marginal lands.

county-level crop yield data (table 1) to a ~250 m resolution based on the remote sensing gross primary productivity (GPP) data (equation 2). Second, we estimate the pixel-level potential yields and yield uncertainties across the CONUS based on a model trained with the downscaled crop yield data from equation 1 as the target and climate, land slope, and soil properties as inputs. Finally, we estimate the land productivity index and marginal land through a set of rules.

2.2. Crop yield downscaling and modeling

Figure 2 provides more details about the estimation of crop yields and their associated uncertainties through the two-step ML approach. The first step crop yield downscaling is implemented through a Gaussian process model (GPM), and the second step yield estimation through a random forest model (RFM). GPM is a Bayesian inference method that has been proved to be able to provide excellent accuracy estimations as well as error bars (i.e. uncertainties) for geophysical studies [25–27]. The solid statistical foundation and non-parametric structure of GPM make it appealing especially for applications with small datasets (e.g. our case for downscaling county-level crop yield data), where the information-to-noise ratio is low [25, 28]. For larger datasets (e.g. the downscaled crop yield at ~250 m resolution), GPM is no longer suitable because of its high computational burden [29]; while neural network (especially deep neural network) and RFM are among the most widely used ML models and those with the best performing [25, 30–33]. Our preliminary test shows neural network and RFM have a similar performance. RFM is selected because of its simpler structure and lower computation burden. The tree-based RFM also holds an advantage of 'built-in' resistance to overfitting because of its 'bagging' process [34], while other popular ML models require major modifications to achieve such performance [35, 36].

The first step GPM is used to estimate the following variable α at pixel-level (~250 m), which is further used to estimate downscaled crop yield:

$$\alpha_i = YD_i/GPP_i \quad (1)$$

where α_i is the ratio between crop yield YD_i and the growing season GPP value of a specific crop i (GPP_i). The GPMs are trained with long-term (2008–2017) average county level crop yield data as the target and county average temperature and water availability data as inputs, as suggested by other studies [37–39]. The water availability information is represented by two variables: the ratio of evapotranspiration (ET) over precipitation and the percentage of land being irrigated. GPM assumes α_i to follow a normal distribution as shown in figure 2, a common assumption in crop yield estimation models [40]. Section II in the SI provides details about the development of GPM for estimating α_i, and table 1 shows the sources of input data. A 5-fold cross-validation as suggested by Yadav and Shukla [41] is used to validate the GPMs trained in this study. It should be noted that, when estimating pixel-level α_i, the ratio of irrigated land (one GPM input to represent the water availability information) is set to zero, which will decrease the land productivity estimate for irrigated land. By this treatment, the land productivity index would only represent the 'natural' productivity of land based on the biophysical properties, but not the 'artificial' increase of yield from irrigation. The GPMs are trained at a county level, where all the inputs and targets reflect the county-averaged values. Then the GPMS are reapplied at a pixel level to estimate the pixel-level crop yields. Since the temperature and water infrastructure conditions (the GPM inputs) are usually homogeneous within a county (especially with the large spatial scope of the

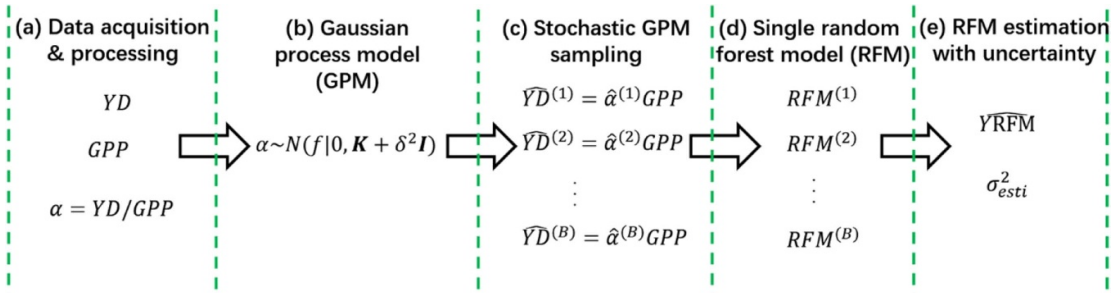

Figure 2. Schematic diagram showing the processes of the two-step ML approach applied to each of the six major crops considered in this study, refer to section 2.2 and sections II and III in the SI for the definition of variables.

Table 2. Inputs for the random forest regression model.

Category	Variable description
Average temperature	Monthly average air temperatures from January to December
Diurnal temperature range	Monthly average diurnal air temperature ranges from January to December
Average precipitation	Monthly average precipitations from January to December
Slope	Percentages of the area falling into eight levels of slope classes: 0%–0.5%, 0.5%–2%, 2%–5%, 5%–10%, 10%–15%, 15%–30%, 30%–45%, and >45%
Soil	Soil depth, soil available water storage and soil organic carbon in 6 levels of soil: 0–5 cm, 5–20 cm, 20–50 cm, 50–100 cm, 100–150 cm, and >150 cm

GPMs), the additional uncertainty caused by the difference between the resolutions in the GPM training and application phases would be limited.

As illustrated in figure 1, the second step RFM in our two-step ML approach is then trained with downscaled crop yield data from the first step GPM. To further capture the uncertainty associated with the GPM, for each crop, an ensemble of random forests are developed, each trained with a stochastic realization of normally distributed downscaled crop yields in GPM (figure 2); the variances among the ensemble of random forests are included as one source of uncertainty in the final RFM crop yield estimation. We sample an ensemble of downscaled crop yields from the GPM to prepare the training targets for the subsequent RFM in figure 2 by equation (2):

$$\widehat{YD}_i^{(j)} = \hat{\alpha}_i^{(j)} GPP_i \tag{2}$$

where variables with hat '^' refer to their pixel-level estimated values from a trained GPM, and the superscript in parentheses '(j)' refers the ensemble number. Inputs to the RFM include the climate variables, slope information, and soil information (table 2). Table 1 provides references for the data sources of the variables shown in table 2.

To account for the uncertainty associated in RFM, we adopt a modified bootstrapping approach (MBA) [42] (figure 2). We train a separate random forest with each of the stochastic realizations of the downscaled crop yields (equation 2), where each random forest consists of a small number of regression trees. To balance the estimation accuracy and computation burden, the number of separate random forest and the number of regression trees in each random forest are selected as 50 and 10 respectively according a previous study [42], which results in 500 separate regression trees for each RFM and a 20% Monte Carlo error for variance estimation according to the MBA [42]. The expected yield for each crop i $\widehat{YRFM}_i$ t ha^{-1} is calculated as the average of all the 500 regression tree outputs; the variance of $\widehat{YRFM}_i$ calculated based on the within/between random forest variances.

A separate RFM is developed for each of the major crops. Data in table 2 are split into training (50%) and testing (50%) sets, and around 36.8% of the training data are unused in the training process and are retrieved as an out-of-bag (oob) validation set for hyperparameter selection [43]. Details about the development of RFM and hyperparameter selection are provided in section III in the SI.

2.3. Estimation of productivity

We then approximate the land productivity using the maximum value of the normalized potential yield across all six major crops, as estimated in section 2.2. To make the potential yields of different crops comparable, we normalize the deterministic RFM estimated yields for each crop with its 99th percentile:

$$NY_i = \widehat{YRFM}_i / YRFM_i^{99} \tag{3}$$

where NY_i is the normalized potential yield for crop i, and $YRFM_i^{99}$ the 99th percentile of $\widehat{YRFM}_i$ t ha^{-1}. The

land productivity index, defined as maximum productivity (*MP*), is then calculated as the maximum value of NY_i over all major crops for each pixel:

$$MP = \max_{i \in \boldsymbol{VS}} (NY_i) \tag{4}$$

where ***VS*** denotes a viable set of crops for a particular pixel. A specific crop *i* is said to be 'viable' for land pixels in counties that have a record of growing crop *i* (identified based on the CDL data). Since the performance of data-driven models like RFM tend to deteriorate for input combinations not seen during the training phase, the introduction of a viable set concept helps avoid such a problem. If the viable set ***VS*** is empty for a particular pixel, *MP* is assigned the NY_i value of winter wheat at that pixel because of its wide spread over the CONUS (figure S2 in the SI). Such an operation might result in overestimation/underestimation of the MP value of that particular pixel, but we do not anticipant a significant effect on the overall pattern of the productivity in the CONUS, given that it is uncommon that a pixel would have empty VS [14]. Also, as all the row crops considered in this study are mostly grown under low slope conditions (more than 99.9% are grown at lands with average slope classes smaller than 4, while the slope class ranges from 1 (most flat) to 8 (most steep), as shown in table 2), *MP* of grid cells with average slope class larger than 4 are assigned the NY_i values of a pasture crop alfalfa (which is more likely to be grown in slope lands) to avoid overestimating productivity at those regions [44].

We quantify the uncertainty of *MP* through a Monte Carlo approach, and calculate the *MP* value of one Monte Carlo realization ($\widehat{NY_i}^{(j)}$) as:

$$\widehat{MP}^{(j)} = \max_{i \in \boldsymbol{VS}} \left(\widehat{NY_i}^{(j)} \right) \tag{5}$$

where the superscript *j* is the realization number, $\widehat{NY_i}^{(j)}$ is sampled based on the RFM estimated variance of NY_i (equation S8, section III in the SI). We randomly generate 1000 realizations of $\widehat{MP}^{(j)}$ and calculate its 95% inter-percentile distance (*IPD*), the distance between 2.5th and 97.5th percentile of the aforementioned 1000 realizations of $\widehat{MP}^{(j)}$) to represent the uncertainty associated with *MP*. A larger value of *IPD* represents a higher level of uncertainty.

2.4. Identification criteria for potential land available for energy crop production

One potential application of the derived *MP* index is to identify marginal land available for bioenergy crop production. We develop a series of scenarios for marginal land identification based on biophysical criteria, and cross-check and justify the results using an economic criterion.

Two aspects of land properties are considered to develop biophysical criteria: the productivity and current land use [6]. For land productivity, the MP values are classified into three categories: high, medium, and low, based on the distribution of MP values of currently cultivated land. For current land use, similar to our previous study [6], two land use scenarios are used in combination with the productivity categories to identify marginal land: one scenario constrains marginal land to only current cultivated land, including crop land and pasture land, with low to medium productivity; the other scenario is bolder and assumes current grassland and shrubland with medium productivity could also be identified as marginal land for growing energy crops. However, the current study does not consider existing forest lands with improved land use classification that differentiates forest land from a coarse classification of mixed cropland, forest, grassland, and shrubland [19] as potential land for bioenergy production due to a growing concern on producing energy crops at the expense of deforestation.

The thresholds for breaking the productivity categories could be identified based on fuzzy logic rules [6], or by constraining the productivity estimate with other variables, e.g. land cover [45] or environmental vulnerabilities [46]. Since such constraining variables are not directly available, we test a series of thresholds via a trial and error method, and identify two sets of thresholds (which will be further validated with profit estimate) on productivity classification: criterion I assigns P25 and P50 (i.e. 25th and 50th percentiles) of the current crop land MP values over the CONUS as the break points from low to median and from medium to high productivity, respectively; criterion II assigns P10 and P25 as those break points. It should be noted that many other combinations of thresholds could be used, but the above two combinations are shown here to represent a realistic range for marginal land acreage. Criterion I represents an aggressive scenario that the maximum MP value of marginal land is higher than the MP values of the vast majority (i.e. 90th percentile) of existing grassland and shrubland (table S1 stacks.iop.org/ERL/15/074013/mmedia); criterion II represents a conservative scenario that the maximum MP value of marginal land reaches just the medium MP values (i.e. 50th percentile) of existing grassland and shrubland. According to the productivity classification criteria and land use criteria as specified above, four scenarios of marginal land estimation are formulated, as displayed in table 3 in the result section.

We use the commonly used economic criterion of positive profit to cross-check the marginal land estimation based on the biophysical criteria [47, 48], i.e. the land that does not pay off the investments and costs when growing regular crops is identified as marginal land. The maximum potential profit (*MPF*) is calculated as:

Table 3. Definitions of scenarios of biophysical and economic criteria for marginal land and their associated total acreages in CONUS.

Scenario/criteria	Description	Acreage (mha)
S1	Current crop and pasture land with $MP \leqslant$ P50[a]	109.2
S2	Current crop and pasture land with $MP \leqslant$ P25[a]	55.1
S3	S1 + current grass and shrub land with $MP \in$ [P25, P50]	175.6
S4	S2 + current grass and shrub land with $MP \in$ [P10, P25]	152.7
Economic	Current crop and pasture land with MPF < 0	73.8

[a]P10, P25 and P50 are the 10th, 25th and 50th percentile of crop *MP* values

$$F_i = YRFM_i \times CP_i - R - OC_i - MC_i, \quad \text{for } i \in [\text{corn, soy, Wwheat, Swheat, cotton}] \tag{6}$$

$$MPF = \max_{i \in \boldsymbol{VS}} (F_i) \tag{7}$$

where F_i is the potential profit of growing crop i ($\$\ ha^{-1}$), CP_i the price of crop i ($\$\ t^{-1}$), R the rent ($\$\ ha^{-1}$), OC_i the operation cost of growing crop i ($\$\ ha^{-1}$), and MC_i the capitalized machinery and equipment costs for crop i ($\$\ ha^{-1}$). Data sources for CP_i and R are shown in table 1, and the costs OC_i and MC_i are calculated with data from USDA Economic Research Service [23] (ERS) (see section I in the SI for the details). The viable set ***VS*** in equation (7) is identical to that in equations (4) and (5). If the *MPF* of a land pixel is negative, the land might be marginal from an economic perspective. The marginal land estimate based on the economic criterion is used as a cross-check with those based on the biophysical criteria. Theoretically, the estimate of marginal land conforms to the economic theory on farmers' land use decision in real life. However, the national level economic data available (crop prices and land rent rates) only has a state-level resolution and may involve large unidentified uncertainty. Also, crop prices and land rents are dynamic and change with the evolution of market conditions, and it is difficult to predict their change beforehand. Therefore, the economically identified marginal land cannot be used directly to identify candidate locations for energy crop production; instead, the marginal land acreage and its spatial distribution are only used to cross-check if any major inconsistencies exist between the marginal land estimated from the biophysical and economic criteria.

After the marginal lands are identified, we apply an irrigation filter to exclude any land pixels with existing irrigation infrastructures (refer to figure S3 for the regions with existing irrigation infrastructures). This follows our assumption that marginal land for bioenergy crops will not include irrigated land, i.e. the current land with irrigation facilities is used for food and fiber crops, especially high-valued crops such as vegetable and fruit in arid and semi-arid regions; future land development for bioenergy production usually does not consider irrigation due to existing water stress around the world [49].

3. Results and discussion

3.1. Machine learning performances

The GPM for crop yield downscaling has good performance in the 5-fold cross-validation results shown in figure S3 in the SI. In terms of the coefficients of determination (R^2) and root mean square error (RMSE), the GPM has at least the same level of performance as the work by Marshall *et al* [37], who developed a crop production efficiency model to adjust yield estimates of corn, soybean, and wheat using GPP. The GPM performances for cotton and alfalfa are also comparable with other studies [50, 51]. The RFM oob validation results are shown in figure S4, and the R^2 values for corn, soybean, winter wheat, spring wheat, and cotton all reach to a high value of ~0.9, and 0.83 for alfalfa. The RFM with the testing dataset shows similar performances (figure S5) to that of the oob validation set. The RFM also adequately estimates the uncertainty associated with the crop yield estimations. As is shown in figure S6, most of the downscaled crop yield data are within the standard deviation of the RFM estimations. The coverage ratios (the percentage of downscaled crop yields falling in their associated RFM confidence intervals, see figure S6) of 95% confidence intervals of the crop yield estimations are 0.95, 0.90, 0.92, 0.93, 0.86, and 0.83 for corn, soybean, winter wheat, spring wheat, cotton, and alfalfa, respectively, which are considered acceptable given the 20% Monte Carlo error allowed for our study (see section 2.2).

3.2. Spatial distribution of productivity

The derived *MP* value shows similar general patterns as other productivity indices (e.g. NCCPI [4] and the soil productivity index [3]). The value appears to be high in the Midwest corn-belt region (figure 3(a)) that is usually considered as the most productive region in the US. Major attributes of the corn belt region include flat topography, medium temperature, sufficient amount of rainfall (the new figures 3(c)–(e)), and deep and fertile soil, representing high land productivity given appropriate drainage facilities to reduce extra soil moisture before the crop growing season. The western US along and west to the Rocky Mountains has low *MP* values, mostly due to the high slope (figure 3(e)) and low precipitation (figure 3(c)).

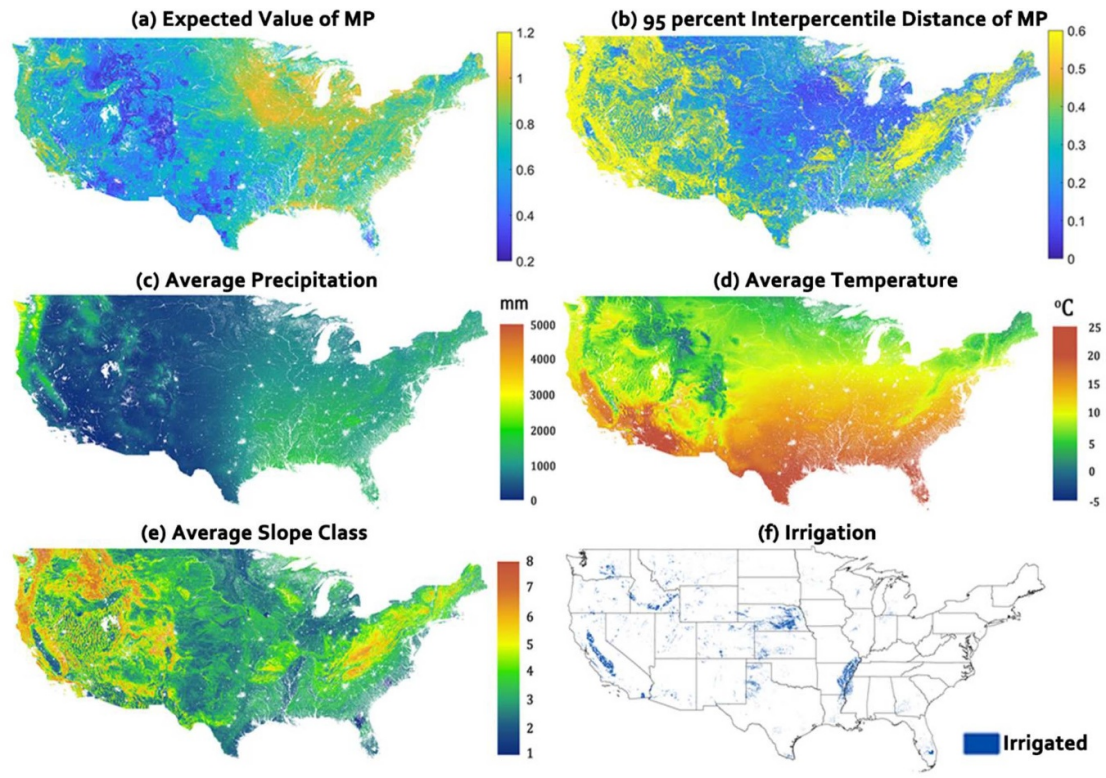

Figure 3. Spatial distribution of (a) maximum productivity (*MP*) expected value, (b) *MP* 95% inter-percentile distance, (c) average annual cumulative precipitation, (d) annual average temperature, (e) average slope class, and (f) irrigated agricultural land in the CONUS.

Between the corn belt and the western US is a corridor with moderate *MP* values. The lower Mississippi region and California Central Valley are major agricultural regions in the US but show moderate *MP* values as compared to that in the corn belt. The major reason for relatively low *MP* values in these two regions is the water deficit from precipitation and evapotranspiration, and irrigation is necessary (figure 3(f)). Note that the irrigation impact is removed when calculating the *MP* value (section 2.1).

The productive corn-belt region shows relatively small *IPD* values (which accounts for 10%–20% of the expected *MP* values in these pixels, figure 2(b)). In comparison, this ratio between *IPD* and *MP* in the western US and high slope regions reach to over 60% (figure S7). Such high *IPD* values could mainly be attributed to the smaller number of observed crop cultivation in those regions (and thus more unseen input combinations in the RFM model) as is shown in figure S2, and the complex impact of slope on the potential crop yields [52, 53].

The *MP* values of current cultivated land are higher than other land covers (as can be seen in figure S8, showing the histograms of *MP* values of current land uses). The ranking of *MP* values for each current land use category from high to low (based on their percentiles shown in table S3) is: cultivated, developed (urban), forest, herbaceous, shrubland, and barren land.

3.3. Spatial distribution of marginal land available for bioenergy crop production

The marginal lands identified according to different biophysical criteria S1–S4 show a considerable level of variability in both their spatial distributions (figure 4(a)) and total area (table 3). The total area of marginal land ranges from 55.1 (from scenario S2) to 175.6 mha (from scenario S3), which fall in the range of 43–179 mha reported by other studies [6, 47, 54, 55]. Several marginal land 'hot spots' that appear in all scenarios S1–S4 are identified, including the corridor region between the Midwest corn belt and the western US, part of the lower Mississippi region and California central valley (that is cultivated but not irrigated according to our land use and irrigation data), and southeastern states.

As can be seen in figure 4(b), the marginal lands identified in S1 and S2 mostly associate with high probability, suggesting high confidence in identifying these lands as marginal. Among all four scenarios, the 'hot spots' of marginal lands identified in figure 4(a) are associated with high probabilities, which suggests an agreement of the marginal land 'hot spots.' Less confidence is found with the marginal lands identified in the western US from S3 and S4. This is mainly because of the lack of crop data and the impact of complex landscapes (i.e. frequently varying slopes) in these regions.

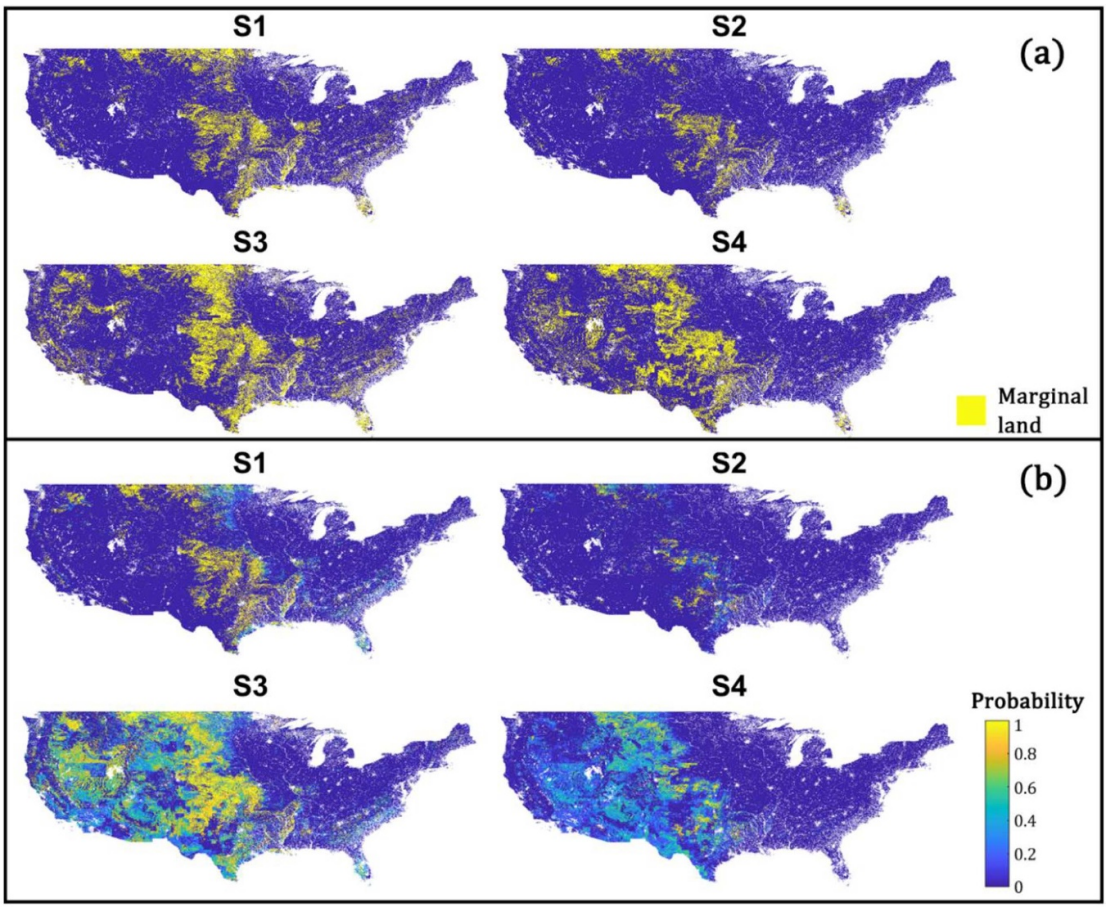

Figure 4. (a) Marginal lands identified deterministically according to biophysical criteria S1–S4 and (b) probability of land being marginal given biophysical criteria S1–S4; refer to table 3 for the definition of marginal land criteria for S1–S4.

We then conduct a cross-check by comparing the marginal land results (S1 and S2) based on the biophysical criteria to those derived from the economic criterion (MLEC). Figure 5 shows a similar spatial pattern of marginal land identified from the three criteria; the total area of marginal land of MLEC is 73.8 mha, which falls in the range between those of S1 and S2 (table 3). This supports the productivity thresholds as tested in section 2.4. Nevertheless, we acknowledge that land productivity is one among many factors (e.g. the economic profitability, environmental vulnerability, and accessibility of land) for identifying actual marginal land potentially available for bioenergy crop production.

4. Implications and limitations

This study contributes to the existing literature of land productivity estimation by adopting a data-driven approach that incorporates the relationships embedded in the data in estimating land productivity. This study is also the first to provide an estimate of uncertainty associated with the land productivity and marginal land, which is proved to be non-trivial. Though our scope of land productivity estimation and marginal land identification is limited within the CONUS, the ML model we have developed is general and could be applied to other regions and for solving other issues (e.g. hydroclimate modeling [56] and geochemistry analysis [57]) in the world (see a recent example of ML based estimation of land suitable of growing cassava [58]).

The high IPD values in some regions (e.g. those with a high slope) can result in a chance to significantly overestimate or underestimate the land productivity in these areas. Therefore, we anticipate substantial uncertainty with the marginal land estimates in those regions (figure 4(b)). The uncertainty must be assessed with caution before the land estimate is used for land use decisions, especially for energy crops, considering the crop market risk that can be caused by the land availability uncertainty. Usually the model performance could be improved when the model scope moves from a national level to a local level (state or county) because of the inclusion of locally relevant data. The model performance could also be improved by incorporating the technique of transfer learning, i.e. inserting the knowledge learned from a large dataset to a model trained with a smaller dataset. If the processes represented by the two datasets are similar, transfer learning could significantly lower

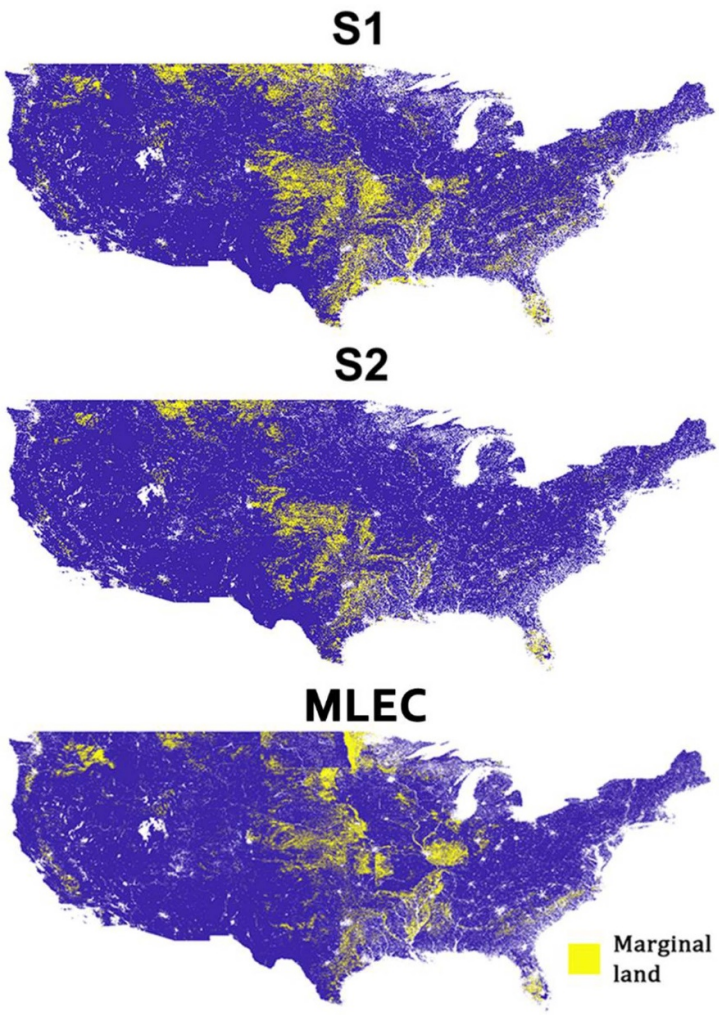

Figure 5. Marginal lands estimated derived from the economic criterion (MLEC) and biophysical criteria (S1 and S2) in the CONUS.

down the uncertainty of the model trained with the small dataset [59].

Our quantification of uncertainty in land productivity and marginal land estimation might be affected by several assumptions made with the two-step ML approach and crop yield post-processing procedures. First, errors in uncertainty estimation might propagate through the two-step ML approach, and one potential source of this error could be the Monte Carlo sampling error (MC error) for the RFM. In general, the Monte Carlo error is smaller if more individual regression trees are included in an RFM, and our choice (500 trees) results from a balance of accuracy (about 20% MC error) and the computation burden. Future studies could consider to include more regression trees or adopt advanced variance reduction methods (e.g. Wager *et al* [60]) to potentially reduce the MC error. Second, our calculation of *MP* value is based on the assumption that all the six major crops are equally good indicators of land productivity. Such an assumption is made to allow practically meaningful calculation of *MP* value for agricultural decisions, but it might underestimate some uncertainties associated with the potentially different yield-environment responses for different crops. Incorporating crop specific processes and criteria for more accurate land productivity uncertainty estimation goes beyond the scope of this study and will be one of the future research issues following this study.

This study trains the model for calculating *MP* values with 10-year average crop yields for the major crops in the CONUS, and the potential impact of yield increase as a result of technology improvement [61, 62] is not considered as a factor. Since the *MP* is calculated in a comparative manner, we expect our estimate of *MP* to be stable from time to time. However, the *MP* value could potentially be changed as a result of climate change or new crop development. For example, the northwestern US might be more productive in the future as a result of more humid climate [8]. Also, the southwestern US might be more productive if new, reliable drought-tolerant crops [63] are developed and popularized.

The marginal lands in this study are identified assuming no irrigation and no deforestation for bioenergy crop. The land acreage would be increased

if irrigation is allowed for growing bioenergy crop, especially for the intensively irrigated areas in the lower Mississippi and California central valley. Also, this study identifies marginal lands with biophysical properties potentially suitable for solving the competition between food and bioenergy production in a long-term manner, but the actual lands available for growing bioenergy crop in a short-term could largely be affected by a series of factors: economic viability [64], farmer's willingness [65], environmental vulnerability [46], infrastructure availability [66], agricultural policies [67], etc. It could be further complicated by some potential environmental consequences of large-scale energy crop plantation, e.g. loss of biodiversity [68], degradation of soil quality [69], and additional use of water [70]. Local criteria for marginal land identification could be developed to reflect the region-specific biophysical and socio-economic conditions. The marginal land identified in this study could be used as a base to consider these conditions for further exploration of land availability for bioenergy crop production.

Acknowledgments

This work was funded by the DOE Center for Advanced Bioenergy and Bioproducts Innovation (U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under Award Number DE-SC0018420). Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the U.S. Department of Energy. We acknowledge the suggestions on the presentation of methods and results from three anonymous reviewers. We thank Avin Arefzadeh for collecting the crop yield and economic data, and Professor Madhu Khanna, Professor Kaiyu Guan, Dr. Chongya Jiang, and Dr. Deepayan Debnath for commenting on the methods and results.

Data availability

The data that support the findings of this study are openly available. The soil property data is stored at www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/home/?cid=nrcs142p2_053628. The slope data is available at https://webarchive.iiasa.acat/Research/LUC/Products-Datasets/global-terrain-slope.html. Monthly temperature and precipitation data are available at https://daymet.ornl.gov/. Annual evapotranspiration data is available at www.esrl.noaa.gov/psd/data/gridded/data.narr.html. Cropland Data Layer land use data is available at www.nass.usda.gov/Research_and_Science/Cropland/SARS1a.php, and National Land Cover Database land use data is available at www.mrlc.gov/. The irrigation map can be found at https://earlywarning.usgs.gov/USirrigation. The gross primary productivity data is available at www.ntsg.umt.edu/project/landsat/landsat-productivity.php. Crop price, crop yield, land rent data are available at https://quickstats.nass.usda.gov. Crop specific production cost data can be found at www.ers.usda.gov/. The maximum productivity index and the identified marginal lands for the biophysical and economic criteria, as well as their associated uncertainty are freely available at the Illinois Data Bank https://databank.illinois.edu/datasets/IDB-4584681.

ORCID iDs

Pan Yang https://orcid.org/0000-0001-7609-7230
Ximing Cai https://orcid.org/0000-0002-7342-4512

References

[1] Lubowski R N, Plantinga A J and Stavins R N 2008 What drives land-use change in the United States? A national analysis of landowner decisions *Land Econ.* **84** 529–50
[2] Irwin E G and Geoghegan J 2001 Theory, data, methods: developing spatially explicit economic models of land use change *Agric. Ecosyst. Environ.* **85** 7–24
[3] Schaetzl R J, Krist Jr F J and Miller B A 2012 A taxonomically based ordinal estimate of soil productivity for landscape-scale analyses *Soil Sci.* **177** 288–99
[4] Dobos R, Sinclair H and Hipple K 2012 National commodity crop productivity index (NCCPI) user guide v2.0 (Lincoln, NE: USDA NRCS National Soil Survey Center)
[5] Nalepa R A and Bauer D M 2012 Marginal lands: the role of remote sensing in constructing landscapes for agrofuel development *J. Peasant Stud.* **39** 403–22
[6] Cai X, Zhang X, and Wang D 2010 Land availability for biofuel production *Environ. Sci. Technol.* **45** 334–9
[7] Feng Q, Chaubey I, Engel B, Cibin R, Sudheer K and Volenec J 2017 Marginal land suitability for switchgrass, Miscanthus and hybrid poplar in the Upper Mississippi River Basin (UMRB) *Environ. Modell. Softw.* **93** 356–65
[8] Zhang X and Cai X 2011 Climate change impacts on global agricultural land availability *Environ. Res. Lett.* **6** 014014
[9] Wightman J L, Ahmed Z U, Volk T A, Castellano P J, Peters C J, DeGloria S D, Duxbury J M and Woodbury P B 2015 Assessing sustainable bioenergy feedstock production potential by integrated geospatial analysis of land use and land quality *Bioenergy Res.* **8** 1671–80
[10] Hellwinckel C, Clark C, Langholtz M and Eaton L 2016 Simulated impact of the renewable fuels standard on US Conservation Reserve Program enrollment and conversion *GCB Bioenergy* **8** 245–56
[11] Marotzke J and Forster P M 2015 Forcing, feedback and internal variability in global temperature trends *Nature* **517** 565
[12] Malone B P, Styc Q, Minasny B and McBratney A B 2017 Digital soil mapping of soil carbon at the farm scale: a spatial downscaling approach in consideration of measured and uncertain data *Geoderma* **290** 91–99
[13] Goulden T, Hopkinson C, Jamieson R and Sterling S 2016 Sensitivity of DEM, slope, aspect and watershed attributes to LiDAR measurement uncertainty *Remote Sens. Environ.* **179** 23–35
[14] USDA NASS Research and Science Cropland Data Layer Releases (available at: www.nass.usda.gov/Research-and-Science/Cropland/Release/) (Accessed: 10 June 2019)

[15] USDA, Gridded Soil Survey Geographic (gSSURGO) Database for the United States of America and the Territories, Commonwealths, and Island Nations served by the USDA-NRCS (available at: https://gdg.sc.egov.usda.gov/) (Accessed: 20 January)
[16] Fischer G, Nachtergaele F, Prieler S, Van Velthuizen H, Verelst L and Wiberg D 2008 *Global Agro-ecological Zones Assessment for Agriculture (GAEZ 2008)* vol 10 (Rome: IIASA, Laxenburg, Austria and FAO)
[17] Thornton P E, Thornton M M, Mayer B W, Wei Y, Devarakonda R, Vose R S and Cook R B 2016 *Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 3* (Oak Ridge, TN: ORNL DAAC)
[18] Mesinger F, DiMego G, Kalnay E, Mitchell K, Shafran P C, Ebisuzaki W, Jović D, Woollen J, Rogers E and Berbery E H 2006 North American regional reanalysis *Bull. Am. Meteorol. Soc.* **87** 343–60
[19] Wickham J, Homer C, Vogelmann J, McKerrow A, Mueller R, Herold N and Coulston J 2014 The multi-resolution land characteristics (MRLC) consortium—20 years of development and integration of USA national land cover data *Remote Sens.* **6** 7424–41
[20] Brown J F and Pervez M S 2014 Merging remote sensing data and national agricultural statistics to model change in irrigated agriculture *Agric. Syst.* **127** 28–40
[21] Robinson N P, Allred B W, Smith W K, Jones M O, Moreno A, Erickson T A, Naugle D E and Running S W 2018 Terrestrial primary production for the conterminous United States derived from Landsat 30 m and MODIS 250 m *Remote Sens. Ecol. Conserv.* **4** 264–80
[22] USDA NASS (USDA National Agricultural Statistics Service) Quick Stats Database (available at: www.nass.usda.gov/Quick_Stats/) (Accessed: 20 January 2019)
[23] USDA Economic Research Service, Commodity Costs and Returns (available at: www.ers.usda.gov/data-products/commodity-costs-and-returns/commodity-costs-and-returns/#Recent%20Cost%20and%20Returns) (Accessed: 20 January)
[24] ESRI 2011 *ArcGIS Desktop: Release 10* (Redlands, CA: Environmental Systems Research Institute)
[25] Camps-Valls G, Verrelst J, Munoz-Mari J, Laparra V, Mateo-Jimenez F and Gomez-Dans J 2016 A survey on Gaussian processes for earth-observation data analysis: a comprehensive investigation *IEEE Geosci. Remote Sens. Mag.* **4** 58–78
[26] Han J, Zhang Z, Cao J, Luo Y, Zhang L, Li Z and Zhang J 2020 Prediction of winter wheat yield based on multi-source data and machine learning in China *Remote Sens.* **12** 236
[27] Campos-Taberner M, García-Haro F J, Camps-Valls G, Grau-Muedra G, Nutini F, Crema A and Boschetti M 2016 Multitemporal and multiresolution leaf area index retrieval for operational local rice crop monitoring *Remote Sens. Environ.* **187** 102–18
[28] Fang D, Zhang X, Yu Q, Jin T C and Tian L 2018 A novel method for carbon dioxide emission forecasting based on improved Gaussian processes regression *J. Cleaner Prod.* **173** 143–50
[29] Williams C K and Rasmussen C E 2016 *Gaussian Processes for Machine Learning* vol 2 (Cambridge, MA: MIT Press) p 4
[30] Sadler J M, Goodall J L, Morsy M M and Spencer K 2018 Modeling urban coastal flood severity from crowd-sourced flood reports using Poisson regression and Random Forest *J. Hydrol.* **559** 43–55
[31] Wang H, Magagi R, Goïta K, Trudel M, McNairn H and Powers J 2019 Crop phenology retrieval via polarimetric SAR decomposition and Random Forest algorithm *Remote Sens. Environ.* **231** 111234
[32] Crane-Droesch A 2018 Machine learning methods for crop yield prediction and climate change impact assessment in agriculture *Environ. Res. Lett.* **13** 114003
[33] Ma L, Liu Y, Zhang X, Ye Y, Yin G and Johnson B A 2019 Deep learning in remote sensing applications: a meta-analysis and review *ISPRS J. Photogramm. Remote Sens.* **152** 166–77
[34] Breiman L 2001 Random Forests *Mach. Learn.* **45** 5–32
[35] Sun X, Ren X, Ma S and Wang H 2017 meprop: sparsified back propagation for accelerated deep learning with reduced overfitting *Proc. 34th Int. Conf. on Machine Learning* vol 70 pp 3299–308
[36] Han H and Jiang X 2014 Overcome support vector machine diagnosis overfitting *Cancer Inform.* **13** CIN-S13875
[37] Marshall M, Tu K and Brown J 2018 Optimizing a remote sensing production efficiency model for macro-scale GPP and yield estimation in agroecosystems *Remote Sens. Environ.* **217** 258–71
[38] Alganci U, Ozdogan M, Sertel E and Ormeci C 2014 Estimating maize and cotton yield in southeastern Turkey with integrated use of satellite images, meteorological data and digital photographs *Field Crops Res.* **157** 8–19
[39] Ryan E M, Ogle K, Peltier D, Walker A P, De Kauwe M G, Medlyn B E, Williams D G, Parton W, Asao S and Guenet B 2017 Gross primary production responses to warming, elevated CO_2, and irrigation: quantifying the drivers of ecosystem physiology in a semiarid grassland *Glob. Change Biol.* **23** 3092–106
[40] Newlands N K, Zamar D S, Kouadio L A, Zhang Y, Chipanshi A, Potgieter A, Toure S and Hill H S 2014 An integrated, probabilistic model for improved seasonal forecasting of agricultural crop yield under environmental uncertainty *Front. Environ. Sci.* **2** 17
[41] Yadav S and Shukla S 2016 Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification *2016 IEEE 6th Int. Conf. Advanced Computing (IACC)* (Piscataway, NJ: IEEE) pp 78–83
[42] Sexton J and Laake P 2009 Standard errors for bagged and random forest estimators *Comput. Stat. Data Anal.* **53** 801–11
[43] Cao D S, Yang Y N, Zhao J C, Yan J, Liu S, Hu Q N and Liang Y Z 2012 Computer-aided prediction of toxicity with substructure pattern and random forest *J. Chemom.* **26** 7–15
[44] 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
[45] Fritz S, See L, Van Der Velde M, Nalepa R A, Perger C, Schill C, McCallum I, Schepaschenko D, Kraxner F and Cai X 2013 Downgrading recent estimates of land available for biofuel production *Environ. Sci. Technol.* **47** 1688–94
[46] Gopalakrishnan G, Cristina Negri M and Snyder S W 2011 A novel framework to classify marginal land for sustainable biomass feedstock production *J. Environ. Qual.* **40** 1593–600
[47] Emery I, Mueller S, Qin Z and Dunn J B 2016 Evaluating the potential of marginal land for cellulosic feedstock production and carbon sequestration in the United States *Environ. Sci. Technol.* **51** 733–41
[48] Lewis S M and Kelly M 2014 Mapping the potential for biofuel production on marginal lands: differences in definitions, data and models across scales *ISPRS Int. J. Geo-Inf.* **3** 430–59
[49] Mekonnen M M and Hoekstra A Y 2016 Four billion people facing severe water scarcity *Sci. Adv.* **2** e1500323
[50] Johnson D M 2016 A comprehensive assessment of the correlations between field crop yields and commonly used MODIS products *Int. J. Appl. Earth Obs. Geoinf.* **52** 65–81
[51] He M, Kimball J, Maneta M, Maxwell B, Moreno A, Beguería S and Wu X 2018 Regional crop gross primary productivity and yield estimation using fused landsat-MODIS data *Remote Sens.* **10** 372
[52] Green T R and Erskine R H 2004 Measurement, scaling, and topographic analyses of spatial crop yield and soil water content *Hydrol. Process* **18** 1447–65
[53] Jiang P and Thelen K 2004 Effect of soil and topographic properties on crop yield in a north-central corn–soybean cropping system *Agron. J.* **96** 252–8

[54] Campbell J E, Lobell D B, Genova R C and Field C B 2008 The global potential of bioenergy on abandoned agriculture lands *Environ. Sci. Technol.* **42** 5791–4
[55] Nijsen M, Smeets E, Stehfest E and van Vuuren D P 2012 An evaluation of the global potential of bioenergy production on degraded lands *GCB Bioenergy* **4** 130–47
[56] Gentine P, Pritchard M, Rasp S, Reinaudi G and Yacalis G 2018 Could machine learning break the convection parameterization deadlock? *Geophys. Res. Lett.* **45** 5742–51
[57] Fang K, Shen C, Kifer D and Yang X 2017 Prolongation of SMAP to spatiotemporally seamless coverage of continental US using a deep learning neural network *Geophys. Res. Lett.* **44** 11,030–11,039
[58] Jiang D, Wang Q, Ding F, Fu J and Hao M 2019 Potential marginal land resources of cassava worldwide: a data-driven analysis *Renewable Sustainable Energy Rev.* **104** 167–73
[59] Weiss K, Khoshgoftaar T M and Wang D 2016 A survey of transfer learning *J. Big Data* **3** 9
[60] Wager S, Hastie T and Efron B 2014 Confidence intervals for random forests: the jackknife and the infinitesimal jackknife *J. Mach. Learn. Res.* **15** 1625–51
[61] Long S P, Marshall-Colon A and Zhu X-G 2015 Meeting the global food demand of the future by engineering crop photosynthesis and yield potential *Cell* **161** 56–66
[62] DeLucia E H, Chen S, Guan K, Peng B, Li Y, Gomez-Casanovas N, Kantola I B, Bernacchi C J, Huang Y and Long S P 2019 Are we approaching a water ceiling to maize yields in the United States? *Ecosphere* **10** e02773
[63] Nuccio M L, Paul M, Bate N J, Cohn J and Cutler S R 2018 Where are the drought tolerant crops? An assessment of more than two decades of plant biotechnology effort in crop improvement *Plant Sci.* **273** 110–19
[64] Anand M, Miao R and Khanna M 2019 Adopting bioenergy crops: does farmers' attitude toward loss matter? *Agric. Econ.* **50** 435–450
[65] Rizzo D, Martin L and Wohlfahrt J 2014 Miscanthus spatial location as seen by farmers: a machine learning approach to model real criteria *Biomass Bioenergy* **66** 348–63
[66] Ng T L, Cai X and Ouyang Y 2011 Some implications of biofuel development for engineering infrastructures in the United States *Biofuels, Bioprod. Biorefin.* **5** 581–92
[67] Renwick A, Jansson T, Verburg P H, Revoredo-Giha C, Britz W, Gocht A and McCracken D 2013 Policy reform and agricultural land abandonment in the EU *Land Use Policy* **30** 446–57
[68] Sauerbrei R, Aue B, Krippes C, Diehl E and Wolters V 2017 Bioenergy and biodiversity: intensified biomass extraction from hedges impairs habitat conditions for birds *J. Environ. Manage.* **187** 311–19
[69] Fernando A L, Duarte M P, Almeida J, Boléo S and Mendes B 2010 Environmental impact assessment of energy crops cultivation in Europe *Biofuels Bioprod. Biorefin.* **4** 594–604
[70] McCalmont J P, Hastings A, McNamara N P, Richter G M, Robson P, Donnison I S and Clifton-Brown J 2017 Environmental costs and benefits of growing Miscanthus for bioenergy in the UK *GCB Bioenergy* **9** 489–507