INTRODUCTION

Cellulosic biofuel, obtained from the lignocellulose of dedicated bioenergy crop and agricultural residuals, possesses great potential for greenhouse gas reduction (Field et al., 2018; Khanna et al., 2020) and various local environmental benefits such as nutrient reduction (Housh et al., 2015; Zhong et al., 2018) and biodiversity improvement (Landis et al., 2018; Robertson et al., 2017). One of the key challenges for promoting cellulosic biofuel production and achieving a more sustainable future is land availability for bioenergy crops. Ideally, bioenergy crops should minimize...
diversion of agricultural lands to avoid competition with food and feed crops (Stoof et al., 2015); they should also not be grown on land under natural forests or grasslands, which are carbon sinks and sources of ecosystem services (Kim et al., 2009). A promising solution to this problem is to grow bioenergy crops on marginal lands (Cai et al., 2011; Gelfand et al., 2013), that is, the agricultural land that farmers might be willing to convert from crop production to bioenergy crops because of low productivity, low profitability, high environmental sensitivity, difficulty in land management, etc. (Jiang et al., 2019; Khanna et al., 2021).

Numerous studies have attempted to identify marginal land for bioenergy crops from different approaches. Some identify “biophysical marginal land” based on land productivity (Cai et al., 2011; Yang et al., 2020), climate and land characteristics (e.g., slope, soil, etc.; Kang et al., 2013), and/or environmental sensitivity (Gelfand et al., 2013). Others track the history of land-use change and identify marginal land as abandoned land (Campbell et al., 2008) or land undergoing intensive land-use changes (e.g., a switch between forest and agricultural land; Cai et al., 2011). Researchers also attempt to identify “economic marginal land” based on the profitability of food and bioenergy crops (Swinton et al., 2011).

These studies use secondary information to identify marginal land for bioenergy crop production. They assume that farmers and landowners (the ultimate decision-makers for marginal land availability) will make land-use choices fully based on those biophysical factors and/or economic theories. However, as demonstrated by many studies (Lee et al., 2018; Sullivan-Wiley & Teller, 2020; Walters, 2017), farmers’ decisions/perceptions about land use depend to a large extent on their preferences, experiences, and socioeconomic backgrounds (e.g., age, education, and farm size). To the best of the authors’ knowledge, only one study (Skevas et al., 2016) has attempted to understand the behavioral aspect of marginal land availability, which identified socioeconomic factors such as land rent and environmental concerns to be significant. More research is needed to improve the understanding of the impacts of various socioeconomic factors, along with land biophysical factors, on farmers’ perceptions of marginal land availability.

Another limitation in previous marginal land studies is the lack of understanding about the drivers of farmers’ heterogeneous perceptions of marginal land availability. Numerous studies have demonstrated the existence of heterogeneous behavior rules in farmers’ land-use decisions (Fewell et al., 2016; Garrod et al., 2012; Zhou et al., 2020). However, little is known about the heterogeneity of farmers’ perceptions on marginal land availability, especially how the dominant causes of land marginality vary among farmers. Answers to this question are of critical importance for the design of more targeted bioenergy crop promotion policies.

In this study, we address the above limitations by applying an explainable neural network (NN) model to examine the data obtained from a survey of Midwestern US farmers’ perceptions of marginal land. Farmers’ socioeconomic backgrounds, as well as the biophysical characteristics of their lands, are used to explain their responses of marginal land availability based on the NN model. The heterogeneity in farmers’ perceptions about classifying their land as marginal is addressed through a deep Taylor decomposition (DTD; Montavon et al., 2017) of the NN model. Through the DTD, we identify the distinctive responses of different farmers to changes in biophysical and socioeconomic factors, and the dominant reason why, at an individual level, part of their land is marginal. We further apply the model to Midwestern states to identify patterns of marginal land availability and associated dominant influencing factors. The analysis is expected to advance understandings about the heterogeneous drivers of marginal land availability and provide implications for more targeted outreach and promotion strategies for bioenergy crops.

### 2 METHODS

We hypothesize that farmers’ perceptions of marginal land availability are affected by both biophysical and socioeconomic factors, as demonstrated in other land-use decision studies (Eaton et al., 2019; Skevas et al., 2016). Figure 1
provides a general workflow of the steps to develop the machine learning model and to derive farmers’ perceptions of marginal land. The socioeconomic factors, as well as farmers’ responses of marginal land availability, are derived from a survey described in Section 2.1. Biophysical factors are derived from a set of remote sensing data and farm boundary data from Farm Market ID (2020), as described in Section 2.2. The procedures for data preprocessing are also described in Section 2.2. The NN model is trained and validated in Section 2.3 using the survey samples. The trained NN model is interpreted in Section 2.3 through the DTD. Finally, in Section 2.4, the trained NN model is applied to land pixels in the US Midwestern states to derive a map of marginal land likelihood (MLL) and associated dominating factors.

2.1 Farmers’ survey

A survey for growers in the US Midwestern states was conducted in late 2019 and early 2020. The survey starts by collecting farmers’ perceived availability of low-quality land, that is, if a farmer considers at least part of his/her farm as low-quality land. We further collect the top three reasons for the farmers who claim low-quality land and the current land uses on their low-quality lands. Other information regarding the farmers’ background characteristics (e.g., demographic conditions), economic conditions, land management objectives, and environmental attitudes is also collected. Refer to section I of the supporting information (SI) for the survey questions.

In the survey, the phrase “low-quality land” is used instead of “marginal land” to provide better clarity for farmers. Low-quality land is defined in the survey as “land which has relatively low crop yield due to high slope, sandy soil, waterlogging, prone to flooding, or other undesirable characteristics.” Other factors (e.g., low profitability and difficulty to access machinery) are also listed as possible reasons for farmers to justify marginal land from various perspectives (Table 2). Since no quantifiable measure of land quality is provided in the survey, farmers’ responses represent their own justification of relatively low-quality land.

The survey was delivered to 2500 farmers in the Midwest via both Internet (online) and mail. The 2500 farmer samples were selected from the Farm Market ID dataset based on farm size to ensure coverage of a wide range of farm sizes. We obtained 242 responses out of the 2500 farmers with an overall response rate of 9.7%. Among all the respondents, 204 reported the availability of low-quality land on their farms.

2.2 Data preprocessing

A large set of potential influencing factors (including both biophysical and socioeconomic factors, Table 1) are constructed as candidate input features for the machine learning model to explain low-quality land (marginal land) availability. Major data sources in Table 1 include the survey, remote sensing products, and the county-level or state-level census from the United States Department of Agriculture (USDA). Refer to section I in the SI for a full explanation of variables in Table 1. The biophysical factors for each respondent are extracted based on the farm boundary data derived from Farm Market ID (an example of the farm boundary is shown in Figure S1).

Features in Table 1 with missing data are filled using the multivariate iterative imputation method (Royston, 2005). For a particular multivariate dataset, the iterative imputation method estimates the missing value of one feature by deriving a regression function between this feature and the other features (i.e., imputation). Each feature is imputed sequentially, and the imputation is repeated 10 times for better convergence (Royston, 2005). Afterward, each feature in Table 1 is min-max normalized into the range of [0, 1].

We then use a backward feature selection (BFS) approach (Abe, 2005) to identify the machine learning model’s input features. Implemented in the Python package “Scikit-learn” (Pedregosa et al., 2011), the BFS identifies the optimal combination of a prespecified number of features for machine learning models through an iterative approach. A key step in BFS is to estimate each input variable’s feature relevance measure, calculated from a random forest classifier (Breiman, 2001). It should be noted that NN is not used for the BFS step because it does not provide a feature relevance measure. A trial-and-error analysis is performed to select the number of features associated with the largest binary classification accuracy.

2.3 Neural network modeling

We use the widely applied feedforward artificial NN to extract the relationships between selected input features and farmers’ binary perception of marginal land availability (PML). An advantage of the NN model is its strong flexibility to approximate a relationship with precision. Another benefit of NN is its ability to interpret farmers’ heterogeneous sensitivities to the various influencing factors, which is introduced in section III of the SI.

The flexibility of NN (and other machine learning models) requires relatively large training samples to constrain the model from learning the random errors in the training dataset (Raudys & Jain, 1991). Based on the results of empirical experiments, a binary classification machine learning model with a small number of input features (smaller than 10) requires at least 100–200 training samples (Vabalas et al., 2019; Zhang & Ling, 2018). This study has a sample size of 204 and the number of selected input features for the NN model is 6 (see section II in the SI), which makes the sample size appropriate for the machine learning model.
The selection of the NN hyper-parameter (including NN architecture, activation function, etc.), as well as the training of the NN model, is introduced in section II in the SI. We use fivefold cross-validation suggested by Yadav and Shukla (2016) to validate the NN model. The validation performance is measured by the area under the receiver operating characteristic curve statistic ($AUC$; Fielding & Bell, 1997). Ranging between 0 and 1, $AUC > 0.5$ indicates a random classifier better than a random guess, and $AUC = 1$ indicates a perfect classifier.

To interpret the model, we develop a partial dependence plot (PDP; Molnar, 2019) for the trained NN model. The PDP illustrates the marginal impact of different input features on the $MLL$ by accounting for the average effects of the other features in the model. The PDP is implemented using the Python package PDPbox (SauceCat, 2018). A positive value in the PDP represents a positive impact of an input feature, and vice versa.

To further interpret the NN model, we apply the DTD (Montavon et al., 2017) of the NN model with each modeling sample. For each sample (farmer), DTD calculates a unique combination of feature relevancies $R_p$, which represents the importance of each influencing factor $p$ in explaining the farmer’s perception about their land being marginal. The values of $R_p$ are non-negative, and their sum equals the NN prediction, that is, the $MLL$. In order to make the feature relevancies comparable for different farmers, the $R_p$ values of each prediction are normalized:

$$NFR_p = \frac{R_p}{\sum_p R_p},$$

where $NFR_p$ is the normalized feature relevance for the $p$th input feature. Full details about the DTD and feature relevance measures are introduced in section III of the SI.

### 2.4 Model prediction for Midwestern states

After the NN model for marginal land availability is trained and validated, we apply the model to land pixels in Midwestern states with an existing land use of cropland, pastureland, or grassland (Wickham et al., 2014). We also calculate the $NFR$ values for each input feature and denote the dominating factor as the feature with the largest $NFR$ value for each prediction.

The values of biophysical factors at each land pixel are collected from the sources shown in Table 1. County average values for the socioeconomic factors in Table 1 are obtained.
from the USDA 2017 agriculture census dataset (USDA, 2017). The county average farm size is calculated from county-specific farm data for 12 classes of farm size (Table S2). The remote sensing data for model prediction are projected to World Geodetic System (WGS) 84 coordinates and resampled to 250 m resolution using the “Bilinear” resampling methods in ArcGIS (ESRI, 2012). All other county average socioeconomic data are uniformly converted into raster data throughout the county with the above resolution and projection.

3 | RESULTS

3.1 | Survey results and model assessment

Among the 204 effective responses, 133 report at least part of their land as low-quality land (marginal land); 109 indicate the reasons why their land is marginal. The result (Table 2) suggests that less optimal biophysical conditions (“low soil quality,” “high slope,” and “prone to flooding”) account for most of the farmers’ choice of low-quality land availability, but there is also a considerable number of farmers who relate their decisions to economic factors (e.g., “low profitability”) and infrastructure conditions (e.g., “difficult to access machinery”; Table 2). It should be noted that the factors in Table 2 are only available for farmers with marginal land. Further explanation of farmers’ perceptions of marginal land availability (especially for the farmers with no marginal land) therefore requires use of the factors in Table 1.

The feature selection procedure also identifies both biophysical and socioeconomic factors as inputs for the NN model. Based on the recursive feature selection result shown in Figure S2, the following six factors are selected: farm size (Total_area), growing season precipitation (Prcp_grow), root zone soil water capacity (Rootaws), average slope (Mean_slope), growing season mean temperature (Tmean_grow), and growing season diurnal range of temperature (Trange_grow). Spatial distributions of the selected features are shown in Figure S3.

The NN model achieves an AUC value of 0.72 in the cross-validation dataset (Figure 2), comparable to the model performances of similar studies (Rizzo et al., 2014; Wimberly et al., 2017). The PDP in Figure 3 shows that the MLL generally increases with Total_area, Mean_slope, and Tmean_grow; and decreases with Prcp_grow, Rootaws, and Trange_grow. Some nonlinear relationships between the input features and MLL can be identified in Figure 3. For example, the MLL increases quickly with farm size when Total_area is smaller than 2000 acres but slowly when Total_area is larger than 2000 acres. For the land slope, the PDP curve increases rapidly when Mean_slope is less than 4% but is relatively flat when Mean_slope is greater than 4%. As can be seen from Figure 3, the partial dependence patterns only show the marginal impact of an input feature under the average condition of the other features in the NN (Molnar, 2019). The relationship between the MLL and an input feature might vary under different survey samples, as shown in Section 3.2.

To derive the uncertainties shown in Figure 3, a bootstrapping method (Freedman, 1981) is applied to generate 200 realizations of the NN model by resampling the training data with replacement. The shaded areas in Figure 3 show the standard deviation of the bootstrapping variance. The result suggests a notable but acceptable level of uncertainty caused by the relatively small sample size used in this study; the general relationships identified above are not affected by the uncertainties.

3.2 | Model interpretations

The scatterplots in Figure 4 show near-linear relationships between the six influencing factors and their NFR values. The signs of the correlation coefficients are opposite between the positive (the farmer has marginal land) and negative (the farmer has no marginal land) prediction samples. More specifically, we find that for the farmers with marginal land, the

| Table 2 | Farmers’ response for the reason of low-quality land |
|---------|-----------------------------------------------|
| Reason                          | Low profitability | Low soil quality | High slope | Prone to flooding | Difficult to access machinery | Other reason |
| Frequency                        | 51               | 73              | 44         | 40               | 23                           | 10           |
explaining powers (NFRs) of Total_area, Mean_slope, and Tmean_grow increase with their values, while the explaining powers of Prcp_grow, Rootaws, and Trange_grow decrease with their values. In contrast, for farmers without marginal land, the explaining powers of Total_area, Mean_slope, and Tmean_grow decrease with their values, while the explaining powers of Prcp_grow, Rootaws, and Trange_grow increase with their values.

Averaging over the surveyed farmers, among the six influencing factors, Trange_grow has the largest NFR value for positive predictions and Total_area for negative predictions (Table 3). The average NFR values for other influencing factors also vary between the positive and negative predictions but with less significant magnitudes (Table 3).

It should be noted that the collection of additional survey responses could allow for the construction of a more complex NN model with more input features and deeper hidden layers. The NFR–feature value relationships of a more complex model might be different from that identified from Figure 4. Nevertheless, the results presented here provide an important preliminary understanding of the socioeconomic and biophysical factors’ impacts on farmers’ heterogeneous perceptions of land being marginal. The results further show how the relative importance of these factors varies among farmers.
TABLE 3 Average normalized feature relevance values of the six influencing factors for the positive and negative prediction samples. Total_area: farm size; Prcp_grow: growing season precipitation; Rootaws: root zone soil water capacity; Mean_slope: average slope; Tmean_grow: growing season mean temperature; Trange_grow: growing season diurnal range of temperature

| Input feature     | Positive predictions | Negative predictions |
|-------------------|----------------------|----------------------|
| Total_area        | 0.088 (0.058)        | 0.249 (0.158)        |
| Prcp_grow         | 0.132 (0.065)        | 0.191 (0.091)        |
| Rootaws           | 0.249 (0.110)        | 0.212 (0.131)        |
| Mean_slope        | 0.118 (0.097)        | 0.143 (0.089)        |
| Tmean_grow        | 0.144 (0.104)        | 0.102 (0.091)        |
| Trange_grow       | 0.268 (0.151)        | 0.103 (0.071)        |

*Values in bracket are the bootstrapping standard deviation of the average normalized feature relevance values.

**Values in bold are the influencing factors with highest normalized relevancies for the positive and negative predictions, respectively.

3.3 | Marginal land availability and its dominant influencing factors in Midwestern states

We evaluate the spatial likelihood of marginal land availability in Midwestern states by applying our trained NN model to the cropland, pastureland, and grassland in this region. The summary statistics of our survey samples and the overall statistics of Midwestern farmers from the USDA NASS dataset are similar (Table S3), suggesting good representativeness of our survey.

The results suggest that farmers in the western parts of North Dakota, South Dakota, and Nebraska are generally more likely to claim marginal land than farmers elsewhere in Midwestern states (Figure 5a). In these

FIGURE 5 Pixel-level prediction of (a) marginal land likelihood and (b) the dominating factors. Total_area: farm size; Prcp_grow: growing season precipitation; Rootaws: root zone soil water capacity; Mean_slope: average slope; Tmean_grow: growing season mean temperature; Trange_grow: growing season diurnal range of temperature
western locations, the relatively low root zone soil water capacity (Rootaws, Figure S3) explains the marginal land existence for most land pixels (Figure 5b). An exception is a few counties where farms are large (Figure S3), and Total_area is identified as the dominating factor (Figure 5b). Farmers in the traditional Corn Belt region (Iowa, Illinois, Indiana, southern Michigan, western Ohio, eastern Nebraska, eastern Kansas, southern Minnesota, and parts of Missouri) claim relatively low MLLs (Figure 5a), but there are also local variations as a result of farmer heterogeneity (e.g., the mixture of yellow and green colors in Iowa in Figure 5a).

The impact of farmer heterogeneity can be better identified in the map of dominant influencing factors (Figure 5b) where local transitions of dominant influencing factors can be spotted throughout the figure. In addition to the spatial patterns identified above in the western parts of North Dakota, South Dakota, and Nebraska, and the traditional Corn Belt region, Mean_slope is found to be the most important factor in the high slope regions of Wisconsin and southern Nebraska (Figure 5b; Figure S3). In general, Total_area is identified as the major factor for most of the land pixels. Trange_grow, Rootaws, and Mean_slope explain the marginal land availabilities in most other land pixels.

We further classify the results from Figure 5 by the existing land use of each pixel: cropland, pastureland, and grassland (Figure 6). Figure 6a suggests that cropland has the lowest MLL among the three land-use types, followed by pastureland and grassland. Our results confirm findings from previous studies that marginal land is primarily located on non-croplands (Cai et al., 2011; Campbell et al., 2008; Yang et al., 2020). Nevertheless, there is still a significant number of cropland pixels with an MLL over 0.5, which suggests a considerable amount of marginal land is located in existing croplands. We also identify a change in the dominant factor between the different types of land uses (Figure 6b). The most important factor in Midwestern states is Total_area for cropland, Trange_grow for pastureland, and Total_area for grassland.

4 | DISCUSSION

4.1 | Farmers' perceptions of marginal land and its influencing factors

Our analysis of the survey data demonstrates that farmers’ perceptions of marginal land availability are affected by both biophysical and socioeconomic factors. Total_area appears to be a controlling factor, especially for farms growing food crops (Figure 6b). Farmers with large farm sizes are generally more likely to claim marginal land. The controlling role of farm size can be further identified in Figure 7, which shows an increase in MLL with farm size as well as the increasingly dominant role of Total_area in explaining MLL with an increase in farm size. This observation could be understood through several mechanisms. From a physical perspective, the trend might occur based on our binary definition of marginal land availability and due to land heterogeneity (Dalgaard et al., 2011). From a human perspective, farmers having different farm sizes might have different land management objectives and constraints. There is evidence that farmers with larger farm sizes tend to have larger emphasis on environmental consequences (Willock et al., 2008); meanwhile, larger farms are less likely to have enough labor to manage each piece of the land and therefore treat a greater portion of the land as marginal. Further separation of the impacts of farm sizes could be achieved by incorporating additional survey samples and input features in the NN model.

The other influencing factors used in the NN model are biophysical factors related to land productivity (potential crop yield). Rootaws controls soil efficiency of rainfall retention for crop evapotranspiration (He & Wang, 2019) and is found to positively affect crop yields; Mean_slope negatively affects crop yields through its impact on soil water and nutrient losses (Jiang & Thelen, 2004). The effects of three climate factors (Tmean_grow, Prcp_grow, and Trange_grow) on land productivity are complex. Each crop usually has an optimal temperature and rainfall, above which the crop production can be negatively affected by increased evapotranspiration loss (for temperature) and flooding loss (for rainfall). Below optimal temperature and
rainfall impair crop production because of low-temperature plant failure and lack of water to support crop growth (Leng & Hall, 2019; Porter & Gawith, 1999; Rosenzweig et al., 2002). \( \text{Trange}_{\text{grow}} \) represents the changes in the daytime and nighttime temperature. This factor might positively affect crop yield through reduced respiration loss during the night or negatively affect crop yield through increased evapotranspiration loss during the daytime (Lobell, 2007). In the US Midwestern states, \( \text{Prcp}_{\text{grow}} \) and \( \text{Trange}_{\text{grow}} \) are found to be positively correlated with crop yields (Goldblum, 2009; Leng & Huang, 2017) and \( \text{Tmean}_{\text{grow}} \) negatively correlated with crop yields (Leng, 2017). As expected, we find the impacts of the investigated biophysical factors on farmers’ perceptions of marginal land (Figures 3 and 4) confirm their relationships with crop yields.

Interestingly, in Figure 4 and Table 3, we identify two distinct behavior rules between farmers with and without marginal land. The perceptions of farmers without marginal land are mainly explained by their (usually) small farm size, and the perceptions of farmers with marginal land are mostly explained by other (usually unfavorable) biophysical factors. There are also opposite signs of the NFR–feature relationships between farmers with and without marginal land. These distinct differences imply a potential divergence of land management objectives between groups of farmers (Willock et al., 2008), which shows the presence of heterogeneity among farmers.

4.2 | Limitations

Because of the limited sample size, our NN model includes only six influencing factors for predicting \( \text{MLL} \): five relate to biophysical factors and one (\( \text{Total}_{\text{area}} \)) represents compounding impacts of both biophysical and socioeconomic factors (see discussion in Section 4.1). In reality, farmers’ socioeconomic characteristics, for example, land management objectives and education level, could affect their perceptions of marginal land availability (Skevas et al., 2014, 2016). Incorporation of these factors can potentially clarify the compounding impacts represented by \( \text{Total}_{\text{area}} \) in this study (see Section 4.1 for details).

Another limitation due to a relatively small sample size (204) is the relatively high cross-validation variance, as shown in Figure 2. However, as shown in Figure 3 and Table 3, the relatively high cross-validation variance does not affect the general relationships identified in this study. A relatively low sample size (100–300) is also considered acceptable in many machine learning analyses of survey results (Khondoker
et al., 2016; Oskar et al., 2020). In general, Vabalas et al. (2019) and Zhang and Ling (2018) recommended that the minimum sample size for a binary classification machine learning model should be 100–200. Nevertheless, a larger number of survey samples could help improve the validity and predictive power of the NN model.

Our results of marginal land availability might contain uncertainties and biases if the respondents and nonrespondents have systematic differences, leading to nonresponse bias (Peytchev, 2013). In this study, one source of nonresponse bias is the potential for respondents and nonrespondents to differ in the explanatory variables (e.g., farm size). Such differences can lead to biases in the NN model parameters (Krawczyk, 2016) and hence the interpretation about the dominating factors of farmers’ marginal land perception in Figure 5b. They might also lead to biases in the estimate of the overall marginal land availability in the Midwest. For example, if farmers with larger farms are more likely to respond to the survey, the result will overestimate the marginal land availability in the Midwest due to the positive impact from the farm size, as identified in this study. Another source of nonresponse bias relates to the potential for respondents to have systematically higher or lower marginal land availability than nonrespondents because of some factors not investigated in this study. Such differences might lead to biases in the estimate of overall marginal land availability in the Midwest (Figure 5a); however, these differences should not affect the interpretations of the model results (Figure 5b) if the underlying mechanism of nonresponse does not correlate to the explanatory variables used in this study. Nevertheless, nonresponse bias is a common problem in survey studies, and a low response rate (our response rate is 9.7%) might not necessarily be its cause (Groves & Peytcheva, 2008). In this study, we do not anticipate the first type of nonresponse bias (related to explanatory variables) because the respondents have similar distributions to the Midwestern farmers in the six influencing factors (Table S3). To avoid the second type of nonresponse bias (caused by undetected reasons), further study with improved survey design is needed, for example, adjust survey design based on the results from preliminary survey responses (Groves & Heeringa, 2006).

### 4.3 Implications and conclusions

In the survey, 133 farmers report their current land use on marginal lands. Over half of them allocate at least part of their marginal land for food crop production, about one third grow hay or pasture, and only five farmers utilize their marginal land for bioenergy crops (Table 4). Such a result confirms our findings in Figure 6a that a considerable part of marginal land might come from existing croplands. Thus, our result implies that at least part of bioenergy cropland will replace existing croplands in the future. In the debate over the “food versus fuel dilemma,” researchers have suggested growing bioenergy crops on non-cropland to avoid competition between fuel and food production (Campbell et al., 2008; Graham-Rowe, 2011). Our findings agree with the others (Cai et al., 2011; Emery et al., 2017; Skevas et al., 2016; Yang et al., 2020) that some potential marginal land for planting bioenergy crops is already used for food crops. However, converting marginal land with food crops to energy crops would not cause major competition between food and fuel because the land is already low quality. In addition, researchers suggest that converting part of existing cropland into perennial bioenergy crops could help reduce nutrient loss, diversify the near monoculture landscape in the US Midwestern states, and improve farmland biodiversity. The above mechanisms can subsequently improve crop production for the unconverted lands (Landis et al., 2018; Robertson et al., 2017; Werling et al., 2014).

The goal of this study is to identify the marginal land (Figure 5a) that could potentially be used for bioenergy crop production (especially when some policy incentives are available, e.g., cost-sharing, additional financial incentives, and government conservation programs; Bergtold et al., 2014). However, given farmers’ land management decisions are still mostly driven by economic profitability (Augustenborg et al., 2012), their adoption of bioenergy crops will not be limited to the low-quality land we have identified. Higher quality lands can also be used for bioenergy crops when favorable market conditions are available (Brown et al., 2016). For example, approximately 40% of US corn croplands were used for ethanol production in 2019 (USDA, 2020).

That said, the map of influencing factors in Figure 5b could help policymakers design more targeted policies to promote bioenergy crop production on marginal land. One such policy we suggest, given the controlling role of farm size on farmers’ perceptions of marginal land availability, is to target large farm owners and operators to encourage bioenergy crop adoption. Such a recommendation is consistent with previous theoretical and empirical analysis (Khanna et al., 2017; Lynes et al., 2016). Larger farms are also expected to be more willing to adopt bioenergy crops because of the economy of scale (larger farm size compensates the usually high fixed cost of bioenergy crop production; Khanna et al., 2017).

### TABLE 4 Existing land use on marginal land in the surveyed farms

| Land use        | Food crop | Hay/pasture | Bioenergy crop | Conservation | Other |
|-----------------|-----------|-------------|----------------|--------------|-------|
| Frequency       | 95        | 39          | 5              | 17           | 10    |
Overall, our study highlights the importance of socio-economic factors in explaining farmers’ heterogeneous perceptions about their land being marginal. Though a full explanation of the impacts of various factors would require additional survey data and probably a more complex machine learning model structure, the results of this study provide valuable information for researchers to improve their understanding and modeling of farmers’ land-use choices, particularly for bioenergy crop adoption.

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. Carrie Leibensperger and Dr. Brenna Ellison led the survey design and implementation; Xiaomin Piao processed part of the survey data; Michael Bianchini proofread the manuscript.

DATA AVAILABILITY STATEMENT
Data used in this study are available at the Illinois Data Bank https://doi.org/10.13012/B2IDB-3074705_V1.

ORCID
Pan Yang https://orcid.org/0000-0001-7609-7230
Madhu Khanna https://orcid.org/0000-0003-4994-4451

REFERENCES
Abe, S. (2005). Modified backward feature selection by cross validation. In European Symposium on Artificial Neural Networks, Bruges, Belgium, 163–168.
Augustenborg, C. A., Finn, J., Mcbennett, L., Connolly, V., Priegnitz, U., & Müller, C. (2012). Farmers’ perspectives for the development of a bioenergy industry in Ireland. Global Change Biology Bioenergy, 4(5), 597–610. https://doi.org/10.1111/j.1757-1707.2011.01151.x
Bergtold, J. S., Fewell, J., & Williams, J. (2014). Farmers’ willingness to produce alternative cellulosic biofuel feedstocks under contract in Kansas using stated choice experiments. Bioenergy Research, 7, 876–884. https://doi.org/10.1007/s12155-014-9425-9
Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
Brown, C., Bakam, I., Smith, P., & Matthews, R. (2016). An agent-based modelling approach to evaluate factors influencing bioenergy crop adoption in north-east Scotland. GCB Bioenergy, 8(1), 226–244. https://doi.org/10.1111/gcbb.12261
Brown, J. F., & Perez, M. S. (2014). Merging remote sensing data and national agricultural statistics to model change in irrigated agriculture. Agricultural Systems, 127, 28–40. https://doi.org/10.1016/j.agsy.2014.01.004
Cai, X., Zhang, X., & Wang, D. (2011). Land availability for biofuel production. Environmental Science and Technology, 45(1), 334–339. https://doi.org/10.1021/es103338e
Campbell, J. E., Lobell, D. B., Genova, R. C., & Field, C. B. (2008). The global potential of bioenergy on abandoned agriculture lands. Environmental Science and Technology, 42(15), 5791–5794. https://doi.org/10.1021/es800052w
Dalgaard, T., Hutchings, N., Dragosits, U., Olesen, J. E., Kjeldsen, C., Drouet, J. L., & Cellier, P. (2011). Effects of farm heterogeneity and methods for upsampling on modelled nitrogen losses in agricultural landscapes. Environmental Pollution, 159(11), 3183–3192. https://doi.org/10.1016/j.envpol.2011.02.043
Eaton, W. M., Burnham, M., Running, K., Hinrichs, C. C., & Sella, T. (2019). Symbolic meanings, landowner support, and dedicated bioenergy crops in the rural northeastern United States. Energy Research & Social Science, 52, 247–257. https://doi.org/10.1016/j.erss.2019.02.005
Emery, I., Mueller, S., Qin, Z., & Dunn, J. B. (2017). Evaluating the potential of marginal land for cellulosic feedstock production and carbon sequestration in the United States. Environmental Science and Technology, 51(1), 733–741. https://doi.org/10.1021/acs.est.6b04189
ESRI. (2012). ArcGIS Desktop: Release 10.1. Environmental Systems Research Institute.
Farm Market ID. (2020). Retrieved from https://www.farmmarketid.com/
Fewell, J. E., Bergtold, J. S., & Williams, J. R. (2016). Farmers’ willingness to contract switchgrass as a cellulosic bioenergy crop in Kansas. Energy Economics, 55, 292–302. https://doi.org/10.1016/j.eneco.2016.01.015
Field, J. L., Evans, S. G., Marx, E., Easter, M., Adler, P. R., Dinh, T., Willson, B., & Paustian, K. (2018). High-resolution technoeconomic modelling of a bioenergy landscape to identify climate mitigation opportunities in cellulosic ethanol production. Nature Energy, 3(3), 211–219. https://doi.org/10.1038/s41560-018-0088-1
Fielding, A. H., & Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. Environmental Conservation, 24(1), 38–49. https://doi.org/10.1017/S0376892997000088
Freedman, D. A. (1981). Bootstrap regression models. The Annals of Statistics, 9(6), 1218–1228. https://doi.org/10.1214/aos/1176345638
Garrod, G., Ruto, E., Willis, K., & Powe, N. (2012). Heterogeneity of preferences for the benefits of environmental stewardship: A latent-class approach. Ecological Economics, 76, 104–111. https://doi.org/10.1016/j.ecolecon.2012.02.011
Gelfand, I., Sahajpal, R., Zhang, X., Izaurralde, R. C., Gross, K. L., & Müller, C. (2012). Farmers’ perspectives for the development of a bioenergy industry in Ireland. Global Change Biology Bioenergy, 5(1), 334–353. https://doi.org/10.1111/j.1757-1707.2011.01151.x
Goldblum, D. (2009). Sensitivity of corn and soybean yield in Illinois to air temperature and precipitation: The potential impact of future climate change. Physical Geography, 30(1), 27–42. https://doi.org/10.2747/0272-3646.30.1.27
Graham-Rowe, D. (2011). Agriculture: Beyond food versus fuel. Nature, 493(7433), 514–517. https://doi.org/10.1038/nature11811
Groves, R. M., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: A meta-analysis. Public Opinion Quarterly, 72(2), 167–189. https://doi.org/10.1093/poq/nfn011
Skevas, T., Swinton, S. M., & Hayden, N. J. (2014). What type of landowner would supply marginal land for energy crops? *Biomass and Bioenergy, 67*, 252–259. https://doi.org/10.1016/j.biombioe.2014.05.011

Stoof, C. R., Richards, B. K., Woodbury, P. B., Fabio, E. S., Brumbach, A. R., Cherney, J., Das, S., Geohring, L., Hansen, J., Hornesky, J., Mayton, H., Mason, C., Ruestow, G., Smart, L. B., Volk, T. A., & Steenhuis, T. S. (2015). Untapped potential: Opportunities and challenges for sustainable bioenergy production from marginal lands in the Northeast USA. *BioEnergy Research, 8*(2), 482–501. https://doi.org/10.1007/s12155-014-9515-8

Sullivan-Wiley, K., & Teller, A. (2020). The integrated socio-perceptual approach: Using ecological mental maps and future imaginaries to understand land use decisions. *Global Environmental Change, 64*, 102151. https://doi.org/10.1016/j.gloenvcha.2020.102151

Swinton, S. M., Babcock, B. A., James, L. K., & Bandaru, V. (2011). Higher US crop prices trigger little area expansion so marginal land for biofuel crops is limited. *Energy Policy, 39*(9), 5254–5258.

Thornton, P. E., Thornton, M. M., Mayer, B. W., Wilhelmi, N., Wei, Y., Devarakonda, R., & Cook, R. B. (2014). Daymet: Daily surface weather data on a 1-km grid for North America, version 2. : Oak Ridge National Lab (ORNL).

U.S. Geological Survey. (2013). LANDFIRE US elevation. Retrieved from https://www.landfire.gov/slope.php

USDA Economic Research Service. (2020). *Feed grains: Yearbook tables.* https://www.ers.usda.gov/data-products/feed-grains-database/feed-grains-yearbook-tables/

USDA National Agricultural Statistics Service. (2014). *Gridded Soil Survey Geographic (gSSURGO) Database for the United States of America and the Territories, Commonwealths, and Island Nations served by the USDA-NRCS.*

USDA National Agricultural Statistics Service. (2017). 2017 census of agriculture. Retrieved from www.nass.usda.gov/AgCensus

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLoS ONE, 14*(11), e0224365. https://doi.org/10.1371/journal.pone.0224365

Walters, B. B. (2017). Explaining rural land use change and reforestation: A causal-historical approach. *Land Use Policy, 67*, 608–624. https://doi.org/10.1016/j.landusepol.2017.07.008

Werling, B. P., Dickson, T. L., Isaacs, R., Gaines, H., Gratton, C., Gross, K. L., Liere, H., Malmstrom, C. M., Meehan, T. D., Ruan, L., Robertson, B. A., Robertson, G. P., Schmidt, T. M., Schrotenboer, A. C., Teal, T. K., Wilson, J. K., & Landis, D. A. (2014). Perennial grasslands enhance biodiversity and multiple ecosystem services in bioenergy landscapes. *Proceedings of the National Academy of Sciences of the United States of America, 111*(4), 1652–1657. https://doi.org/10.1073/pnas.1309492111

Wickham, J., Homer, C., Vogelmann, J., McKerrow, A., Mueller, R., Herold, N., & Coulston, J. (2014). The multi-resolution land characteristics (MRLC) consortium – 20 years of development and integration of USA national land cover data. *Remote Sensing, 6*(8), 7424–7441. https://doi.org/10.3390/rs6087424

Willcock, J., Deary, I. J., Edwards-Jones, G., Gibson, G. J., McGregor, M. J., Sutherland, A., Dent, J. B., Morgan, O., & Griewe, R. (2008). The role of attitudes and objectives in farmer decision making: Business and environmentally-oriented behaviour in Scotland. *Journal of Agricultural Economics, 59*(2), 286–303. https://doi.org/10.1111/j.1477-9552.1999.tb00814.x

Wimberly, M. C., Janssen, L. L., Hennessy, D. A., Luri, M., Chowdhury, N. M., & Feng, H. (2017). Cropland expansion and grassland loss in the eastern Dakotas: New insights from a farm-level survey. *Land Use Policy, 63*, 160–173. https://doi.org/10.1016/j.landusepol.2017.01.026

Yadav, S., & Shukla, S. (2016). Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. *2016 IEEE 6th International Conference on Advanced Computing (IACC) (pp. 78–83).* IEEE.

Yang, P., Zhao, Q., & Cai, X. (2020). Machine learning based estimation of land productivity in the Contiguous US using biophysical predictors. *Environmental Research Letters, 15*, 074013. https://doi.org/10.1088/1748-9326/ab865f

Zhang, Y., & Ling, C. (2018). A strategy to apply machine learning to small datasets in materials science. *NPJ Computational Materials, 4*(1), 1–8. https://doi.org/10.1038/s41524-018-0081-z

Zhong, J., Yu, T. E., Clark, C. D., English, B. C., Larson, J. A., & Cheng, C. L. (2018). Effect of land use change for bioenergy production on feedstock cost and water quality. *Applied Energy, 210*, 580–590. https://doi.org/10.1016/j.apenergy.2017.09.070

Zhou, X., Ma, W., Li, G., & Qiu, H. (2020). Farm machinery use and maize yields in China: An analysis accounting for selection bias and heterogeneity. *Australian Journal of Agricultural and Resource Economics, 64*(4), 1282–1307. https://doi.org/10.1111/1467-8489.12395

**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section.

**How to cite this article:** Yang P, Cai X, Khanna M. Farmers’ heterogeneous perceptions of marginal land for biofuel crops in US Midwestern states considering biophysical and socioeconomic factors. *GCB Bioenergy*. 2021;13:849–861. https://doi.org/10.1111/gcbb.12821