Science of the Total Environment 960 (2025) 178312

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Machine learning-based identification of animal feeding operations in the United States on a parcel-scale

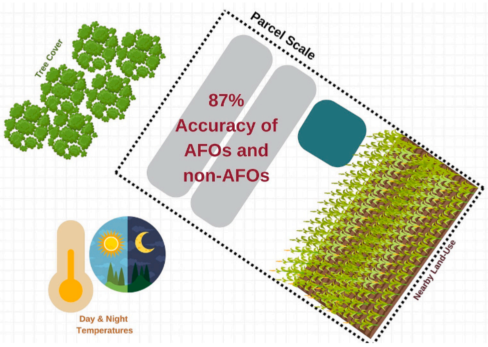
Arghajeet Saha [a], Barira Rashid [a], Ting Liu [a], Lorrayne Miralha [b], Rebecca L. Muenich [a,*]

[a] *Department of Biological and Agricultural Engineering, University of Arkansas, United States of America*
[b] *Department of Food, Agricultural and Biological Engineering, The Ohio State University, United States of America*

HIGHLIGHTS

- Random forest model to predict AFO locations without aerial images
- U.S.-wide model had accuracy of 87 %.
- Accuracy varied based on region and available validation data.
- Land use, land surface temperatures, and tree cover were key predictors.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Jurgen Mahlknecht

Keywords:
Livestock
Remote sensing
Manure
CAFO
SHAP

ABSTRACT

The increasing global demand for meat and dairy products, fueled by rapid industrialization, has led to the expansion of Animal Feeding Operations (AFOs) in the United States (US). These operations, often found in clusters, generate large amounts of manure, posing a considerable risk to water quality due to the concentrated waste streams they produce. Accurately mapping AFOs is essential for effective environmental and disease management, yet many facilities remain undocumented due to variations in federal and state regulations. Current techniques for mapping AFOs in the US rely on a mix of manual digitization, aerial imaging, and image processing. By applying a machine learning-based random forest (RF) classification method to a socio-environmental dataset that excluded aerial images in this work, we overcame some of the limitations associated with aerial image-based approaches, enhancing mapping accuracy to 87 %. We used publicly available environmental, nutrient-focused, and socioeconomic data downscaled to the parcel level, which more accurately reflects farm boundaries and operations than previous methods. Our study incorporates 58 variables, with canopy cover, surrounding vegetation, day and nighttime land surface temperatures, and phosphorus from animals identified as key predictors of AFO presence. The relevance of these variables varies across states, influenced by whether the dominant land covers are human-induced, like croplands, or natural, such as savannas and grasslands. Thus, our public-data based approach, easily replicable, not only improves the precision of AFO detection, but also facilitates the monitoring of nutrient flows at the parcel level—critical for nutrient budgeting and recovery, water quality management, and disease risk assessment and tracing.

[*] Corresponding author.
E-mail address: rlogsdo@uark.edu (R.L. Muenich).

https://doi.org/10.1016/j.scitotenv.2024.178312
Received 21 August 2024; Received in revised form 3 December 2024; Accepted 26 December 2024
Available online 6 January 2025

1. Introduction

Over the past fifty years, the United States (US) animal agriculture industry has witnessed a significant transformation with the rise of confined facilities largely replacing smaller, dispersed family-owned farms with open-air enclosures (MacDonald et al., 2018). Animal Feeding Operations (AFOs), as these facilities are defined in the Clean Water Act, feed animals for at least 45 days a year within the confined portion of the facility and avoid growing grass or forage in the confined area where animals are kept; essentially distinguishing them from grazing operations (United States Environmental Protection Agency [USEPA], 2009). While confinement areas do not sustain crops, it is common for AFOs to be part of larger agricultural operations that include crop production for feed (Centner, 2010). These crops are typically grown in fields near the confinement areas (Centner, 2010).

Since the 1960s, with improvements in agricultural technology and animal breeding programs, AFOs have evolved into larger, high-density, intensive operations driven by a focus on production efficiency to meet escalating consumer demand (Mallin and Cahoon, 2003; Key et al., 2017; Walljasper, 2018). Larger AFOs, as defined by total number of animals varying by animal type, called Concentrated Animal Feeding Operations (CAFOs), have helped to double milk production, triple meat production, and quadruple egg production (Pew Commission, 2009), but also have attracted stricter regulations and more scrutiny as they produce massive quantities of manure, with USDA-based census data and livestock reports estimating it between 1.2 and 1.37 billion tons (wet weight) annually—almost 3 to 20 times more than the human waste generated in US each year (USEPA, 2005). A typical small CAFO, housing no more than 300 cattle, was estimated to generate as much waste as produced by the urine and feces of 16,000 humans (Sierra Club, 2021).

While manure from animal facilities provides beneficial nutrients like phosphorus (P) and nitrogen (N), its overapplication can cause ecological harm. In the US, animal manure constitutes approximately 50 % of P applications to landscapes (Bouwman et al., 2017). As a result, excessive nutrients can lead to eutrophication, damaging water bodies (Devlin and Brodie, 2023), while leaks from manure lagoons may pollute groundwater (Rudko et al., 2023). Additionally, during overflow events, these leaks can affect surface water systems (Raff and Meyer, 2022; Aneja et al., 2003). Antibiotics and pharmaceuticals used in animal feed can also be delivered to surface waters from animal operations, raising concerns about microbial resistance (Lopatto et al., 2019; Kümmerer, 2004; West et al., 2011; Burkholder et al., 2007). The proximity to these operations also poses health risks to nearby residents, leading to diseases like methemoglobinemia and hyperthyroidism and causing adverse reproductive outcomes (Kronberg and Ryschawy, 2019; Ward et al., 2005; Seffner, 1995; Arbuckle et al., 1988). Furthermore, increased atmospheric particulate matter from CAFOs can degrade air quality, exacerbating respiratory and other health problems (Kümmerer, 2004; Cole et al., 2000) and even causing leukemia (Fisher et al., 2020) among local populations.

Any AFO meeting the EPA's CAFO definition based on the threshold of animal types (Hribar, 2010) must adhere to the National Pollutant Discharge Elimination System (NPDES) program under the Clean Water Act. Initially, the 2003 rule required all operations meeting criteria to be defined as a CAFO to obtain NPDES permits and collect essential data on their operations and manure management. However, changes in 2008 revised the rule to allow facilities to opt out of the permit requirement if they demonstrated no potential for waste discharge into the Waters of the United States (WOTUS). Thereby, the implementation of NPDES permits now varies across states—some mandate NPDES permits for all CAFOs as defined in the Clean Water Act, others only for those discharging into WOTUS, and a few states enforce stricter regulations than federal standards (Rosov et al., 2020). Even for states that require all federally defined CAFOs to have a permit, because the permits are inventory-based, clusters of facilities under inventory thresholds may go under-regulated and unmonitored across the US (Miralha et al., 2022). These inconsistencies make it challenging to compare data across states and to develop a comprehensive national understanding of the impacts of CAFO operations. Privacy laws further complicate this by allowing farmers to withhold location information (Steinzor and Huang, 2012). As a result, there is a pressing need to design a framework that can map all animal operations to better evaluate impacts and improve management.

Historically, efforts to map and monitor AFOs have relied on ground investigations and aerial surveys conducted by various non-profit and public interest groups. Notable among these were initiatives like the 'Exposing Fields of Filth' project, spearheaded by the Environmental Working Group (EWG) and the Water Keeper Alliance that mapped CAFOs during flooding seasons (EWG, 2016). These traditional methods, however, require extensive manual labor and are resource and time intensive. In recent years, with the development of machine learning tools and remote sensing data availability, approaches to locate animal operations have veered towards utilizing aerial imagery and deep learning algorithms. Using an image classification technique over high-resolution images from the US Department of Agriculture's National Agricultural Imagery Program (NAIP), Handan-Nader and Ho (2019) were able to train a convoluted neural network that identified an additional 589 poultry AFOs (or a 15 % increase) when compared with manual estimates across North Carolina. Robinson et al. (2022) used image labeling, segmentation, and object-based filtering techniques to train a convoluted neural network on 42-Terabytes (TB) of 1-m NAIP imagery and identified poultry AFOs with up to 83 % accuracy across ten counties in California. Around the same time, Zhu et al. (2022) used over 86,000 georeferenced aerial images from Sentinel-1 and Sentinel-2 to develop a multi-sensor database, Methane Tracking Emission Reference Database (Meter-ML), that identified methane emitting AFOs with up to 91.5 % precision (i.e., the proportion of positive identifications that are correct).

However, these advanced techniques also come with challenges, including significant computational demands for image processing and data development. Additionally, the performance of models based on image classification tends to decline when the images are not centered around the facility but are tiled over the entire study area. For example, in Handan-Nader and Ho's (2019) model, the recall (or the proportion of actual positives correctly identified) drops from 93.72 % to 54 % from image decentering. In a recent study (Saha, 2022), only 83 % of publicly available AFO location data points were correctly centered on the AFO structures. Moreover, these image processing methods also risk missing AFOs not adequately represented in training images and are constrained by changes in and around animal facilities over time or due to spatial variations by state. For example, swine AFOs in North Carolina primarily have waste lagoons located next to the barn structures whereas swine facilities in many portions of the Midwest have manure lagoons beneath the structure, making the development of an image-based swine model applicable to multiple regions challenging.

Given these issues, there is a critical need to develop a more robust and efficient method of mapping AFOs that utilizes easily sourced datasets to detect animal operations without taking significant time or data and human labor. To address this need, we present an innovative machine learning-based method that leverages publicly available datasets from various governmental agencies and public institutions, without the inclusion of computationally intensive aerial imagery, and enables the detection of AFOs with high accuracy by employing a parcel-scale data synthesis.

2. Materials and methods

For this work we developed a random forest (RF) classification model that uses environmental raster data (Section 2.1.2), atmospheric data (Section 2.1.3), census data (Section 2.1.4), county level manure nutrient data (Section 2.1.5), soil P data (Section 2.1.6) and locations of meat processing facilities (Section 2.1.6) to predict AFO locations (Section 2.1.1) at a parcel scale (Section 2.1.7). A tree-based model like RF has generally been observed to better represent complex relationships between datasets of different categories, so we therefore selected a RF model for our work (Ahmad et al., 2017; Chowdhury et al., 2020). The RF model was trained using the target variable and 58 predictor variables. About 80 % of the total input dataset (comprising predictor and target variables) was used for training, while the remaining 20 % was used for testing, following commonly applied splits for training machine learning models (Hino et al., 2018). The model also used 500 trees after a brief tuning process to detect complex patterns and interactions between datasets and a seed determined by the length of the training dataset minus one was used to make sure that every time the model is run it gives the same results.

2.1. *Description of input datasets*

2.1.1. *Target variable*

Since the goal of our study was to identify where AFOs are located, one part of our target variable comprised the latitude and longitude of these facilities. Despite challenges with available AFO data due to variations in state regulations, we successfully gathered, verified (ground-truthed), geocoded, and digitized AFO locations into point shapefiles using ArcGIS-based tools for eighteen US states: Alabama (AL), Arizona (AZ), Florida (FL), Indiana (IN), Iowa (IA), Michigan (MI), Minnesota (MN), Mississippi (MS), Missouri (MO), North Carolina (NC), Ohio (OH), Oregon (OR), Pennsylvania (PA), Tennessee (TN), Texas (TX), Wisconsin (WI), South Carolina (SC), and Louisiana (LA), as documented in Supplemental Table S1.

The other part of our target variable consisted of non-AFO locations. These points were located near AFOs but under separate ownership (e.g. parcels) to ensure they were not part of the same AFO facility. The goal of including points close to AFOs was to enable the model to distinguish environmental signatures not directly impacted by the AFOs but nearby. The points were distributed across diverse land covers, with the majority coming from cropland (37.3 %), followed by savanna (35.6 %), forest areas (12.6 %), grasslands (7.5 %), residential areas (6 %), wetlands (0.6 %), shrublands (0.3 %), barren lands (0.1 %), and water bodies (0.01 %).

Incorporating non-AFO locations was critical since our modeling task involved predicting whether a location was an AFO or not. By defining our target variable into two classes— 'non-AFO' and 'AFO'—the RF model developed in our study could effectively learn from the predictor variables and differentiate between these two categories. The target variable input data for our study was developed in a binary format, with '1' denoting an AFO location and '0' denoting a non-AFO location. In total, 9410 AFO locations and 2519 non-AFO locations were considered in our study, with the distribution of these locations detailed in Table 1.

Table 1
Distribution of AFO and non-AFO locations across eighteen US states in the complete model input, including both training and testing datasets.

Geographical region[a]	State abbreviation	Percentage (%) of total input data (Training and testing)	
		AFO	Non-AFO
Midwest	IA, IN, MI, MN, MO, OH, WI	39.66	10.13
Northeast	PA	2.46	0.26
West	AZ, CA, OR	3.65	0.47
Southeast	AL, FL, LA, MS, NC, SC	31.08	9.48
Southwest	TX	1.85	0.74

[a] Geographical regions are based on maps developed by CDC's National Centre for Health Statistics (more details can be found here: https://www.cdc.gov/nchs/hus/sources-definitions/geographic-region.htm).

2.1.2. *Terrestrial predictor variables ($n = 14$)*

In addition to the target variable, the training and testing dataset of our RF model included predictor variables. The first predictor variable set comprised of terrestrial environmental metrics derived from MODIS (Moderate Resolution Imaging Spectroradiometer) satellite imagery captured between 2017 and 2018. Available in raster format with spatial resolutions ranging from 250 to 1000 m per pixel grid, these metrics encompass Land Use Land Cover (LULC), Percent Tree Cover (PTC), Leaf Area Index (LAI), Normalized Difference Vegetation Index (NDVI), Land Surface Temperature during both day (LST Day) and night (LST Night), and Evapotranspiration (ET). More details can be found in Supplemental Table S2. Thereafter, we simplified the MODIS LULC data into eight key categories relevant to our target variable locations - cropland, grassland, savannas, forest, shrubland, wetland, urban, and barren. Savannas, as defined in our study, represent a mixed ecosystem intermingled with woodlands, grass, and other herbaceous vegetation (Fowler and Beckage, 2019). For each LULC category, we created raster maps for the US where pixels representing a specific land cover were assigned a value of 1, and all other pixels were assigned a value of 0. The method on how these land categories were determined can be found in Miralha et al. (2021).

Both PTC and NDVI are dimensionless indices derived from MODIS data; PTC is measured annually, and NDVI, which assesses plant health or greenness, is updated every 16 days based on optimal pixel values. The LAI, calculated by MODIS every 8 days, quantifies canopy density. It measures the total leaf area per unit ground area for broadleaf species and half the total needle surface area for conifers (Myeni et al., 1997). LST for day and night is recorded in Kelvin and is averaged over an 8-day period. ET, measured in millimeters per 8-day interval, captures the total water loss due to soil evaporation and plant transpiration per pixel area. More details on the various MODIS products can also be found in Supplemental Table S2.

Although the NLCD dataset offers land cover information at a higher spatial resolution (30 m), MODIS satellite imagery emerged as the preferred option for our study, particularly due to its proven effectiveness in CAFO detection. For example, Martin et al. (2018) identified frequent misclassifications of CAFOs as natural wetlands using land cover data from NLCD, a problem not observed with MODIS. Furthermore, MODIS has shown higher accuracy in representing forest cover, evidenced by a lower mean squared error (Song et al., 2014). Additional studies, such as those documented by Miralha et al. (2021), further underscored the superiority of MODIS in detecting AFOs compared to NLCD.

In total, our study considered 14 terrestrial environmental variables, 8 of which were land cover categories. For our analysis, we considered only the summer season (June 15th to September 15th) values for the vegetation-based variables since this period coincides with peak growing season, enabling us to capture the best overall health and density of vegetation. This approach also helped us avoid issues arising from cloud cover and snow effects.

The relationship between these land cover alterations and changes in terrestrial variables has been a focal point of ecological research for decades. As early as 1985, Tucker et al. established a direct correlation between deforestation and the decline in NDVI, highlighting how land use changes impact vegetation indices. Similarly, Wickham et al. (2012) observed that forests, which have lower albedo due to their ability to absorb and retain more solar radiation, exhibit lower daytime LST and higher nighttime LST compared to croplands and grasslands. Further studies, such as those by Ishtiaque et al. (2016), Yu et al. (2019), and Miralha et al. (2021), have linked shifts from forest to cropland or shrublands with increases in daytime LST and decreases in PTC, ET, and LAI.

However, this relationship is not always straightforward and can be influenced by a variety of competing environmental factors. For instance, Pettorelli et al. (2005) showed that NDVI can decline in natural land covers such as grasslands and savannas due to multiple stressors including drought, diseases, overgrazing, and fires. In agricultural settings, Wardlow and Egbert (2008) noted that NDVI typically rises from the planting phase to the senescence phase, reflecting the growth and maturation of crops, before declining post-harvest. Asner et al. (2003) found that while mature forest trees often have higher LAI values than most crops due to their perennial nature and layered foliage, well-managed crops with high planting densities can achieve similar high LAI values, but only during the peak growing season. Quintanar et al. (2009) observed that LST can also be significantly influenced by anthropogenic modifications, like manure lagoons. These lagoons, affected by strong thermal stratification from solar radiation, exhibit higher daytime latent heat flux and, due to heat redistribution, lower nighttime surface temperatures. For ET, Allen et al. (2011) further associated its increase with factors such as low albedo and extensive irrigation practices. Thereby, by integrating these terrestrial variables into our study, we can effectively highlight the dynamic nature and complexity of their interactions with land cover changes. Influenced by both natural events and human activities, these variables are essential for achieving our study's objective to accurately characterize the diverse environmental conditions surrounding AFOs across the eighteen states under examination.

2.1.3. *Atmospheric predictor variables (n = 3)*

The second set of environmental variables emphasizes the impact of atmospheric emissions on air quality, particularly focusing on nitrogenous pollutants such as ammonia and nitrogen oxides released from AFOs and agricultural areas.

Ammonia emissions, originating from crop fields, grazing fields, and various feedlot operations such as housing, storage, retention pools, and lagoons, significantly affect human health by contributing to the formation of aerosols (USEPA, 2004). Nitrogen oxides released from crop fields and agricultural machinery also form hazardous aerosols, severely impacting air quality and human health (USEPA, 2004). Therefore, aerosols, particularly fine particulate matter (PM2.5; particles with a diameter of $\leq$2.5 μm) near agricultural regions, predominantly consist of ammonium and nitrate particulates due to the conversion of ammonia and nitrogen oxides into these particulate forms (Hristov, 2011).

According to a study by Battye et al. (1994), about 43.4 % of the anthropogenic ammonia emissions within the US came from the cattle industry, followed by the swine and poultry industries, which contribute 26.7 % and 10.1 %, respectively. Fertilizer application was responsible for 9.5 % of these emissions (Battye et al., 1994). On the other hand, an air quality report by USEPA (2004) listed about 4 % of nitrogen oxide to be coming from on-field agricultural sources.

Currently, within the agricultural sector, CAFO-based air emissions are addressed by state regulations alongside federal laws such as the USEPA's Clean Air Act (CAA), the Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA), and the Emergency Planning and Community Right-to-Know Act (EPCRA) (Moses and Tomaselli, 2017). However, the USEPA does not have an accepted methodology to measure hazardous emissions from animal waste. Instead, the agency recommends that farm owners use past data, process-based models, engineering approximations, and best judgment to quantify the releases.

To address this challenge and ensure a consistent dataset across the US, we utilized air quality data developed through a satellite-derived method by Van Donkelaar et al. (2019). This method combines Aerosol Optical Depth (AOD) measurements from MODIS, Multi-angle Imaging SpectroRadiometer (MISR), and SeaWiFS instruments with the GEOS-Chem chemical transport model to generate aerosol projections. The results are then calibrated using regional ground-based observations of both total PM2.5 and its components. Finally, the calibrated data is analyzed using Geographically Weighted Regression (GWR) to derive the final compositional estimates.

For our study, we considered three variables — PM2.5 and its compositional estimates of nitrate (NO_3^-) and ammonium (NH_4^+). PM2.5 represents the total particulates in the air, while the free radicals, NO_3^- and NH_4^+, serve as proxy representations of nitrogenous emissions. The concentrations of these variables were in micrograms per cubic meter (μg m^{-3}) of air within a 1.1 km-pixel grid.

2.1.4. *Social and economic predictor variables (n = 35)*

The third predictor set is comprised of social and economic variables obtained from the US Census Bureau (2016) that offers a comprehensive demographic and financial overview at the county level. The main goal of including socioeconomic variables was to incorporate aspects of environmental justice and environmental racism since studies have found that in many areas across the US, AFOs cluster around low-income minority communities (Nicole, 2013). For example, a survey of CAFOs in North Carolina — one of the largest swine producers in the US according to the USDA in 2017 — found that these operations were disproportionately located in communities with low socioeconomic status and high minority populations (Son et al., 2021). Further analysis by Quist et al. (2022) quantified the disparity, revealing that the proportion of Black, Hispanic, and American Indian residents living within 3 miles of a CAFO was 1.42, 1.42, and 2.20 times higher, respectively, compared to White residents.

A study by Wilson et al. (2002) found that in Mississippi, hog-based livestock operations had high percentages of low-income, African American residents living nearby, while in Ohio, Hispanic communities were disproportionately affected by CAFOs (Lenhardt and Ogneva-Himmelberger, 2013). A study by Nicole (2013) noted that Black residents in CAFO-dominated regions in NC often face high rates of poverty. The 21 social variables include information related to population demographics, such as total population count, gender distribution, racial composition, and voter registration data by gender. The data also differentiated educational attainment into specific age groups and education levels, from those with less than a ninth-grade education to those holding graduate or professional degrees. Gender-based variables were considered to better understand whether differences in employment, education, and political engagement between sexes—particularly in communities with high voting participation—impact CAFO establishment. An anthropological study by Sterling (2015) highlighted that in La Salle County, Illinois, a community mobilization led to rezoning and the denial of permits for AFOs, with many the protesting citizens being community college graduates, faculty members, and scientists.

The remainder 14 variables were economic in nature and captured household income across a wide spectrum, from those earning less than US$10,000 annually to households with incomes over US$200,000. This income data also provided information about health insurance coverage, by sub-dividing the population into groups having private, public, and no health insurance. Details of all the socioeconomic variables are in Supplemental Table S3. Although US census data is available at the tract level, we chose to use county-level estimates because county-level aggregates provide us with broader, more consistent coverage and helps us avoid the noise associated with tract-level information, particularly in sparsely populated regions. Moreover, since many regulatory decisions regarding the establishment of animal operations are made at the county level, county-level census data align well with decision-making frameworks. Using county-level data also simplifies processing and reduces computational complexity, which was also one of the goals of this study.

2.1.5. *Manure nutrient-based predictor variables (n = 4)*

The nutrient variables were derived from the USGS calculations of manure per county by Falcone (2021). These variables quantify the nutrients generated in each US county from animals, where the animal numbers are producer figures from their inventory. The use of these variables in our study was also associated with the underlying assumption that counties with a greater number of animals in AFOs produce more manure, and the fact that these data are based on animal inventories from the National Agriculture Statistics Service's Agricultural Census. While quinquennial, county-level data cannot help identify specific operations, we hypothesized that it would be correlated with AFOs. The 4 variables used included common animal N (kilogram - kg) and common animal P (kg), as well as other animal N (kg) and other animal P (kg). In our study, we will assume common animal N and P as nutrients generated by livestock (beef, dairy, hog, swine) and poultry, while other animal N and P are nutrients produced by sheep and horses.

2.1.6. *Additional predictor variables ($n = 2$)*

We also incorporated soil P concentrations and the proximity of each target variable to the nearest meat processing plant (MPP) as two additional variables. The soil P concentration, measured in milligrams of phosphorus per kg of soil (mg kg^{-1}), estimates the phosphorus content in the top 5 centimeters (cm) of the soil profile and was sourced from a USGS database developed by Smith et al. (2013). The USGS database consisted of soil core samples collected from 4857 sites (1 site sampled per 1600 km^2) across conterminous US. Thereafter, we interpolated the sample soil *P* values from these sites using ArcGIS-based tools to generate a 10-km spatial map of the US. The soil P variable was included because findings by Long et al. (2018) and Waldrip et al. (2023) demonstrated that soils surrounding AFOs have substantial P build-up due to manure overapplication and nutrient-rich runoff from these operations. The MPP variable was included to measure how far AFOs are from animal processing plants, as slaughter plants generally tend to be near feedlots (MacDonald et al., 2000). The MPP locations were retrieved from the Meat, Poultry, and Egg Product Inspection directory maintained by the Food Safety and Inspection Service (FSIS) under the US Department of Agriculture (USDA) (USDA-FSIS, 2023). The distance between MPP and each target variable location was estimated using the Point Distance analysis tool in ArcGIS Pro version 3.2.0, measured in km.

2.1.7. *Parcel data*

Parcels are spatial geographic units that delineate the perimeter boundaries of land ownership units. A single parcel is a polygon shapefile characterized by curves and angles that represent the contours of a property. As a result, a parcel provides a more accurate spatial representation of a property unit, unlike pixel grids in raster data (see Fig. 1-a and b), which may cross parcel boundaries in a non-uniform manner. Previous studies have used parcel data to improve representation and model agricultural management units (Kalcic et al., 2015). We acquired nation-wide parcel data through a collaboration with Regrid (https://regrid.com/company) under their 'Data with Purpose' program, which provides academics access to such data at a flexible licensing fee.

2.2. *Data processing and simulation of the random forest classification model*

After the input variables were collected, we processed them using ArcGIS Pro version 3.2.0 to scale all variables to the parcel scale (Fig. 1). In cases where the data were larger than a particular parcel, that value was assigned to the parcel. In cases where there were multiple values for one data within a parcel, we took a weighted average value, or in cases of categorical data like land use, we selected the majority category.

For each state, the average time for processing each terrestrial variable ranged from 20 to 62 min, with those with lower spatial resolution taking less time. The average downscaling time per state for each air quality variable was about 28 min. Meanwhile, the downscaling of socio-economic variables averaged about 8 min per state.

The final downscaled file had 59 columns, where 58 columns were the predictor variables, and one column was the target variable. The number of rows represented parcels where either AFO or non-AFO locations were identified. Each row included a single observation of a parcel, with values for 58 predictor variables and 1 target variable class (AFO or non-AFO). This final tabular file was then normalized using a linear function to ensure that all the variables used in our input dataset were on the same scale (between 0 and 1), and then used as input for our RF model.

The RF model was initially analyzed at the national level, encompassing eighteen states. Its performance was assessed to determine how effectively it identified AFO parcels, the number of non-AFO parcels it misidentified as AFOs, and the accuracy with which it recognized non-AFOs. This assessment was conducted using a confusion matrix, one of the most common machine-learning evaluation techniques, focusing on its precision score. A confusion matrix associated with a classification model illustrates the relationship between the predicted and actual classifications made by the model (Visa et al., 2011). Subsequently, predictive variables influencing the model's predictions were analyzed to see how and to what extent they aided the model in classifying a parcel as an AFO.

Generally, identifying these variables—otherwise known in machine learning as feature selection—in high-dimensional models, such as ours, is both a boon and a curse. While more dimensions can help us explore patterns, they may also add complexity and result in data redundancy. To address this problem, we used Shapley Additive exPlanation (SHAP)-value-based feature selection (Lundberg and Lee, 2017).

The SHAP framework is model-agnostic and incorporates principles

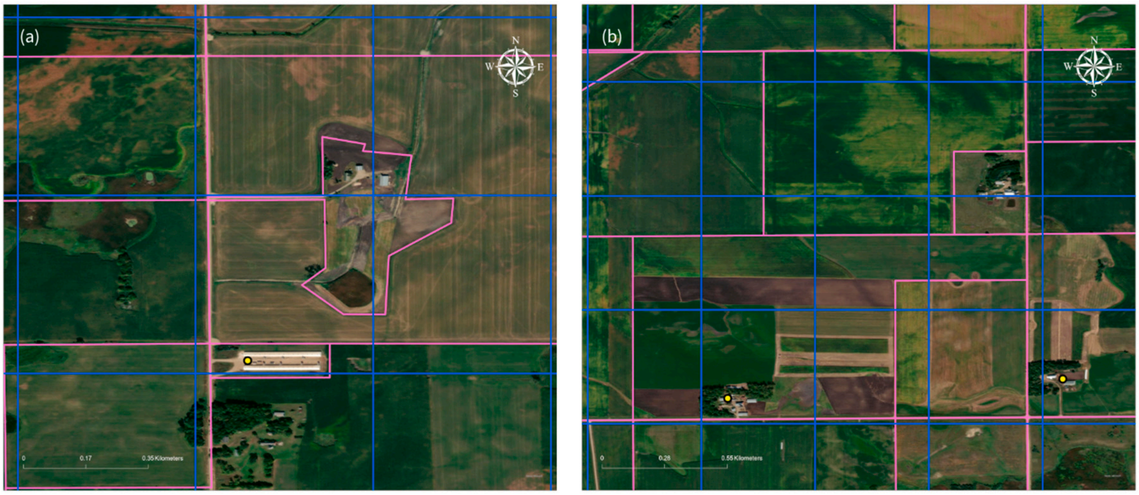


Fig. 1. Orthoimages (Fig. a and b) of Animal Feeding Operations (AFOs) in Wisconsin (highlighted by yellow points on each image) with pink lines representing parcel boundaries and blue lines representing a 500-meter land cover pixel grid. Figure (a) illustrates a small AFO facility with the edges of the parcel border representing a semi-cleared vegetation of a crop field around the AFO. In figure (b) it is clear that the AFO parcel has multiple grid cells representing it. These examples highlight the importance of processing-scale in machine learning models.

from game theory to assess the contribution of each feature towards the predicted outcome of a model for each individual instance. Unlike the SHAP framework, traditional methods like permutation importance provide a global perspective and assess feature importance across the entire dataset. While other impurity-based feature selection methods, using Gini coefficients or Mean Square Error, tend to strongly favor features having high cardinality. Therefore, the SHAP values generated using this framework quantified the contribution of each feature as the difference between the actual model output for that sample and the expected value (or baseline) of the model over the dataset. Following this methodology, the prediction of our machine learning model, f(x), can be expressed as:

$$f(x) = \text{base value} + \text{sum of SHAP values}$$

where *base value* is the average prediction across all data points in the absence of any features. (i.e., the model's output without any input features). It serves as the reference against which the contributions of individual features are measured.

For our study, we used a SHAP-based summary plot (Fig. 2-c) to determine the important features that affect our model prediction. The features impacting our model output are listed along the vertical axis, ranked from top to bottom based on their overall importance, while the horizontal axis represents the SHAP values. The line crossing through the SHAP value of zero is called the baseline. Features with positive SHAP values positively impact the target prediction, while those with negative values adversely affect it. A SHAP value of zero means the feature has no impact.

Each dot on the plot (Fig. 2-c) is a SHAP value for a feature and an individual prediction, where color indicates the actual value of the feature for that prediction – from low (purple) to high (yellow). SHAP does not classify features 'high' or 'low' in an absolute sense. It considers a feature's value relative to the distribution of that feature within the data. Dots far to the right of the baseline indicate that the feature has increased our model's capacity to identify the target, in this case, an AFO. Conversely, dots to the left of the baseline signify that the feature decreases the prediction output. The greater the spread of dots from the baseline, the stronger the impact (either positive or negative) on the model's output. Since land cover types are represented in binary format, their values are depicted in either yellow or purple. Other continuous features are represented by a range of shades within the color gradient illustrated in Fig. 2-c. For ease of explanation, the values of continuous features are divided into three parts — low (ranging from 0 to 0.33), moderate (ranging from 0.34 to 0.66), and high (ranging from 0.67 to 1). Additionally, the air emission-based features were further assessed using a scatter plot to better understand their source contributions. This analysis was particularly relevant as both NH_4^+ and NO_3^- particulates are emitted by both AFO and non-AFO sources.

Building on our national-scale results, we further evaluated the effectiveness of our model at the state level by focusing on the interaction between the ten most sensitive features, excluding land cover. Since land cover often dominates model analyses, exploring the other parameters allows us to better understand the strength of our model and the subtle interactions that support the prediction of AFO parcels. By

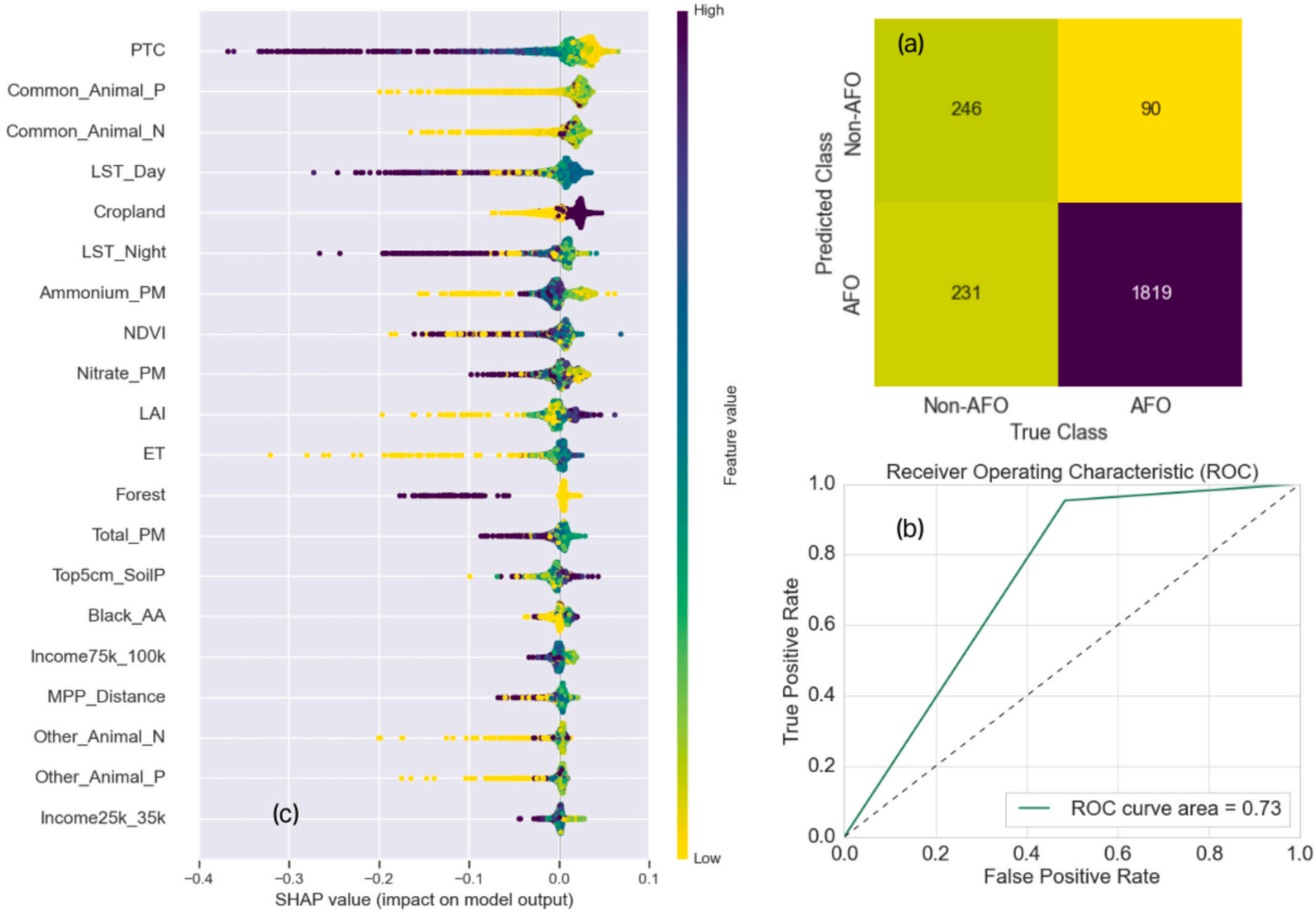


Fig. 2. National scale results of the random forest classifier showing (a) confusion matrix illustrating the difference in model prediction between true class (actual target variable label) and predicted class (predicted target variable label), (b) Receiver Operating Characteristic curve, and (c) Shapley Additive explanation (SHAP) summary plot.

isolating these features, we aimed to gain robust insights into regional terrestrial signatures, anthropogenic impacts, air and soil composition, and demographic patterns. This detailed analysis will help us develop future frameworks designed to enhance the usability and predictive accuracy of the model across diverse environments and geographic regions.

To facilitate these analyses, we first determined the dominant land cover type for each state by analyzing the most prevalent land covers in our test parcels. We classified states as having a dominant land cover if it accounted for >50 % of the test parcels within that state. Based on this criterion, the eighteen states were grouped into three categories - ten states where cropland was the dominant land cover, six states predominantly covered by savanna, and two states where no single land cover type was dominant. Thereafter, we segmented our test data according to these three categories and processed each segment through our RF classifier model to generate SHAP summaries. The summaries will identify the key features influencing model predictions for each category, specifically highlighting the drivers behind correct AFO predictions, false positives, and false negatives.

3. Results and discussion

The RF classification model developed for our study detected AFO parcels (true positives) and non-AFO parcels (true negatives) at an overall accuracy of 87 %. Accuracy is defined as the number of correct predictions from total number of predictions. We tested the model with 2386 parcels, of which 1909 were AFOs and 477 were non-AFOs. The model identified 2050 parcels as AFOs and 336 as non-AFOs.

The model demonstrated a high recall of 0.95, accurately identifying about 95 % of test AFO parcels. This assessment can be further elaborated from the confusion matrix (Fig. 2-a), which indicated that out of 1909 AFOs, the model could accurately identify 1819 of them. Since recall, or true positive rate, measures how much of a particular target the model correctly identifies, it is intrinsically dependent on the proportion of false negatives predicted by the model. Therefore, higher false negatives result in a lower recall value. In our study, only 5 % of AFO parcels were falsely identified as non-AFOs. The model also achieved a precision score of 0.89, indicating that 89 % of the predicted AFOs were true positives. Precision is influenced by the rate of false positives, which was about 11 % for AFO predictions.

In machine learning, precision and recall metrics focus exclusively on predicting true positives, like AFO parcels. However, our research extended this concept to evaluate model performance in identifying non-AFO parcels nationally to discern the underlying data-dependent factors that could enhance our model. Our findings reveal that the model successfully identified only 52 % of the actual non-AFO parcels, missing nearly half. This discrepancy is primarily due to the composition of our testing data, where only 19 % of samples were non-AFOs, leading to a significant class imbalance. This imbalance resulted in an elevated false positive rate for non-AFO predictions—27 %, which translates to approximately 231 out of 477 non-AFO parcels being incorrectly classified as AFOs by our RF model. The F1 score, which measures the harmonic balance between recall and precision, was estimated to have a weighted average value of 0.86, indicating a solid balance between precision and recall and demonstrating that our model is highly effective at distinguishing the dominant positive class (AFO parcels) with minimal errors.

The effectiveness of our RF model can be further understood with the help of the Receiver Operating Characteristic Curve (ROC; Fig. 2-b). The ROC curve diagnoses the classification ability of our model and has two parameters – True Positive Rate (TPR), signifying recall, and False Positive Rate (FPR). Initially, the ROC curve rises sharply, indicating a high TPR with only a minimal increase in FPR. This result demonstrates the ability of our classifier to effectively distinguish between classes with minimal false positives. However, the curve starts to flatten after reaching the center, suggesting that an increase in sensitivity diminishes classifier performance and leads to more false positives. Despite these challenges, an area under the curve (AUC) score of 0.73 indicates that the classifier performs reasonably well in differentiating between the two classes, albeit with some false positives.

3.1. *Key variables (features) influencing our model predictions on a national scale*

The model results on a national scale, comprising eighteen states, for the twenty most sensitive features, as evidenced by the SHAP plot (Fig. 2-c), revealed PTC as the most influential feature in our predictions. The longer spread of high PTC values on the left of the baseline demonstrated that tree cover is negatively associated with AFOs. This finding aligns with existing studies that suggested that areas adjacent to animal operations often undergo deforestation (Miralha et al., 2021). In addition to PTC, the nutrients (N and P) generated from livestock and poultry (i.e., Common_Animal_N and Common_Animal_P, respectively) also significantly impacted model predictions. However, manure P was identified to be more sensitive, as it was found to have more inherent variability towards the classes, AFO and non-AFO. This is reflected in the Spearman correlation coefficient for P (0.31), which, while not very high, is slightly higher than that for N (0.28), suggesting a stronger monotonic relationship with the class labels.

The lack of significant canopy cover around AFO parcels is also supported by thermal signatures, where moderate daytime LST and low-to-moderate nighttime LST—reflective of diverse crop management practices and grazing—were identified as significant features helping our model detect AFOs accurately. This observation is further validated by the spread of low SHAP cropland values to the right of the baseline, which indicatesthat most AFO parcels are found in croplands.

Both NH_4^+ and NO_3^- particulates (Ammonium_PM and Nitrate_PM, respectively, in Fig. 2-c) exhibited mixed impacts on AFO prediction, indicating that AFOs emit a range of these particulates. To better understand this phenomenon, we used a scatter plot to observe their distribution across both AFO and non-AFO parcels under different land covers, as shown in Fig. 3. From these observations, we identified two distinct clusters - one includes parcels that release significant amounts of both N-based compounds (emissions to the right of the short blue dash lines at $y = 0.67$), another consists of parcels emitting moderate levels of both NO_3^- and NH_4^+ (emissions between the short pink dash lines at $y = 0.34$ and short blue dash lines $y = 0.67$), and the third represents parcels with low NH_4^+ and NO_3^- emissions (emissions left of the short pink dash lines at $y = 0.34$).

The widespread overlap of atmospheric NH_4^+ and NO_3^- particulates in the first cluster suggests many non-AFO and AFO parcels have comparable emission rates. High emissions from non-AFO parcels in agricultural areas are often attributed to intensive fertilization practices and agricultural machinery, or due to intensive pasture. In some cases, high nitrate emission rates from non-AFO parcels may also result from industries (USEPA, 2004), particularly if these parcels are located near urban centers. High emissions from AFOs typically originate from parcels that either house numerous animals or feature intensively managed cropland or herbaceous vegetation, such as savannas and croplands, alongside an AFO.

Moderate to low NH_4^+ and NO_3^- emissions are often associated with parcels having semi-managed savannas or grasslands, smaller AFO facilities, or a combination of small facilities and semi-managed vegetation. The relationship between low or moderate NH_4^+ and NO_3^- emissions is non-linear, underscoring that in many such areas, localized factors such as animal types, barn design, feed, management practices, pasture, regulatory oversight, and climate conditions influence the predominance of either form of emission.

On a national scale, the study revealed a nuanced impact of NDVI and LAI, on predicting AFO presence across eighteen states. Generally, the model identifies AFOs in parcels exhibiting low-to-moderate NDVI and LAI values (Fig. 2-c), signifying various stages of crop cultivation

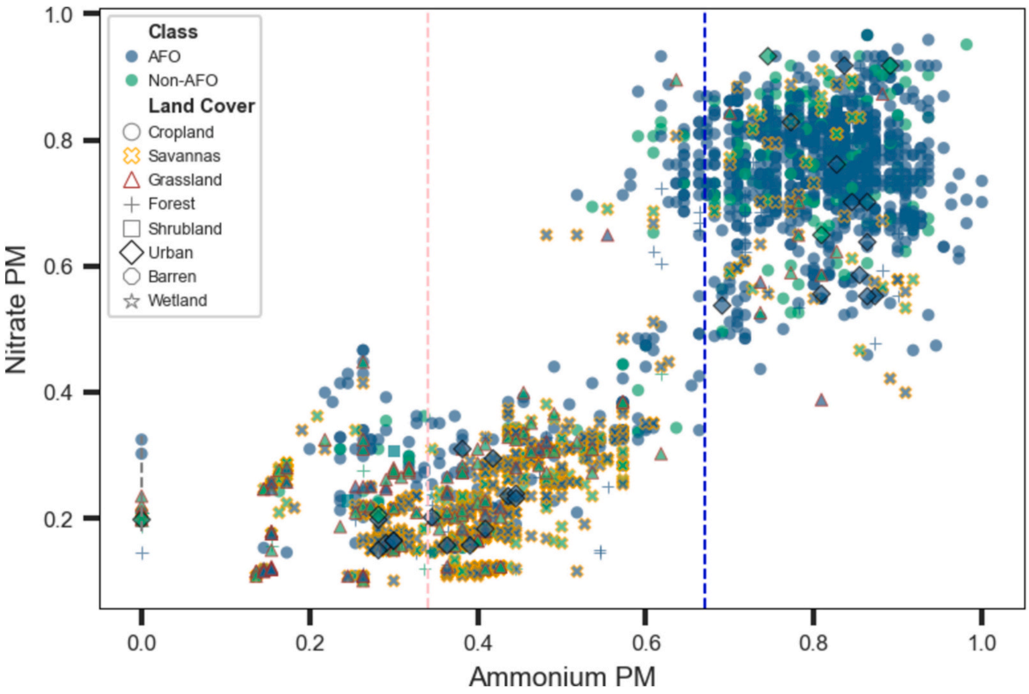


Fig. 3. Scatter plot showing the distribution Nitrate Particulate Matter (Nitrate PM) and Ammonium Particulate Matter (Ammonium PM) across various land uses present in our test data for both Animal Feeding Operation (AFO) and non-AFO based sources.

and grazing. The relevance of these observations is underscored by the timing of our data collection, which coincided with peak summer—a period that aligns with diverse crop growth and harvest schedules across the US.

However, we also note exceptions where high NDVI and LAI values strongly suggest the presence of AFOs. These anomalies can be representative of areas with healthy natural land cover or are the result of regulatory influences. For instance, a study by Miralha et al. (2021) observed that CAFOs in Michigan's Lower Peninsula exhibited higher NDVI and LAI, likely due to regulated agricultural practices. Similarly, Qi et al. (2017) reported improvements in NDVI around CAFOs following the implementation of environmental regulations, suggesting that such measures can significantly impact the surrounding vegetation's health and density.

The ET rate and forest cover had a lesser impact compared to other features but were still significant enough to affect model output (Fig. 2-c). Moderate ET values tended to favor AFO prediction signaling that most AFOs have some measure of biomass loss around them. A spread of low ET values to the left of the baseline are indicative of barren and urban land parcels, as they have strong negative impact on AFO predictions. The high forest cover values and their negative impact on AFO identification corroborated our earlier PTC-focused observation that a dense tree cover does not surround the majority of AFOs across the states considered in our study.

Although most values of these features centered around the baseline, the spread of high values to the right indicated that in several cases, AFO locations had higher P levels in the top 5 cm of soil (Top5Cm_SoilP) and were located near an MPP as shown in Fig. 2-c. High soil P levels can be indicative of soils near operations that house large numbers of animals in a livestock-dense county. Additionally, the elevated soil P can also be an artifact of the type of animal, such as poultry, which releases more P per ton of manure than swine or cattle (Lorimor et al., 2004).

Other features like N and P from sheep and horses (i.e., Other_Animal_N and Other_Animal_P, respectively) had more neutral to negative impact, suggesting that nutrients from sources like sheep and horses might be less indicative of AFO locations than those from other more intensive animal operations, which coincides with typical animal management practices for these animals.

Social features had a smaller impact on improving AFO location accuracy compared to environmental factors. However, a slight shift to the right of the baseline suggested that communities with a higher population of African Americans (i.e., Black_AA) were more prevalent in some AFO areas (Fig. 2-c). This observation validated exiting studies (Son et al., 2021) which identified significant black population around CAFOs across many US states.

Our SHAP-based economic assessment revealed that higher concentrations of AFOs tended to be in counties with fewer households earning within the US$75,000 to US$100,000 (i.e., Income 75k_100k; Fig. 2-c) range, alongside a low proportion of households with incomes between US$25,000 and US$35,000 ((Income25k_35k); Fig. 2-c). This pattern indicates a generally lower economic status in these counties, aligning with existing research that often links AFOs to economically disadvantaged areas (Son et al., 2021; Nicole, 2013). However, given that of all the income categories analyzed, which ranged from less than US$10,000 to over US$200,000, these two income brackets were specifically highlighted as significant points towards a need for further research using more detailed data, such as that from census tracts, to better understand the dynamics between race and income around AFOs.

3.2. State-wise analysis of model results

The state-wise assessment to identify the dominant land cover type across the eighteen states found that cropland was the most dominant land cover (i.e., land cover most commonly associated with test parcels) in all the Midwestern states considered for our study. Other states with cropland as the common land cover were PA in the Northeast, OR in the West, and TN in the Southeast. On the other hand, the Southeastern states had savanna as the dominant land cover (i.e., land cover commonly found around test parcels). Therefore, dominant land cover was used to present and group state-based results.

3.2.1. States with croplands as the dominant land cover

The results indicated that our model achieved an average accuracy of 0.89 and an average recall of 0.95 in states where cropland is the predominant land use (Table 2). Specifically, in MI, OH, PA, and TN, the model achieved a perfect recall score of 1.0, successfully identifying all AFOs within these states. In contrast, IA, MO, IN, and MN recorded recall scores above 0.93, whereas WI had a notably lower recall of 0.72. OR also performed well, with a recall score of 0.96. Despite TN's high recall rate, publicly available data on AFO locations were limited, accounting for only 0.24 % of the total training parcels and 0.13 % of the total testing parcels. More work is needed to combine georeferencing and new data to identify more AFOs in states with limited public data.

The relatively lower recall score for WI can be attributed to the model incorrectly classifying about 8.3 % of the AFO parcels as false negatives. Furthermore, in WI the model incorrectly identified 21.8 % of non-AFO parcels as false positives, which consequently reduced its precision score, as detailed in Table 2. Although MN, OH, IN, and IA had false positive rates of approximately 13.9 %, 13.6 %, 10.7 %, and 10.2 % respectively, they exhibited significantly low or negligible false negative rates, which contributed to their higher recall scores (Table 2).

Features Influencing AFO Parcel Predictions in States Dominated by Cropland — About 92 % of the AFOs in the ten cropland-dominated states had crop-specific land cover, 4 % had savannas, 3 % had grasslands, and 1 % had forests. As a result, across these states, the lack of tree cover (low PTC) emerged as the most influential factor for identifying AFO parcels, as evidenced in the SHAP summary plot (Fig. 4-a). In three OR parcels, however, moderate PTC values indicated the presence of AFOs, suggesting that these AFOs were adjacent to forests.

Although savannas also feature mixed vegetation, the absence of woody canopies near AFO parcels in these states suggests land cover alteration. A study by Fowler and Beckage (2019) validated this observation and indicated that savannas in crop growing regions across the US have been severely impacted by row-crop expansion, land management, and grazing.

Considering the land cover scenario around AFOs in these regions, changes in LST were found to be sensitive towards detecting AFO presence. Based on the model results on Fig. 4-a, moderate-to-high nighttime LST and high daytime LST had a positive influence on AFO prediction. Interestingly, low nighttime LST in these states was typically attributed to non-AFO urban parcels. This phenomenon can be influenced by the extensive use of heat-absorbing materials like asphalt and concrete in urban areas. These materials absorb heat during the day and release it at night, keeping the surface temperature cooler (Azevedo et al., 2016). Additionally, urban areas often incorporate high-albedo materials like white paint and light-colored rooftops, which contribute to the same cooling effect (Taha et al., 1992).

In addition to the top three features, the model effectively utilized information from livestock and poultry manure nutrients, particularly P, to enhance its predictions of AFO locations. Counties with higher manure P production yielded increased model confidence in identifying AFO parcels within those areas (Fig. 4-a). This trend underscores the usefulness of self-reported census data on animal numbers, suggesting that more animals are likely indicative of more AFOs or intensive operations. Furthermore, this county-level census data can serve as a "bounding" dataset for refining predictions and guiding future research efforts in AFO studies. The ability of the model to identify AFO parcels in cropland dominated states was also improved by observing low-to-moderate values of NDVI and LAI, along with moderate values of ET. These indicators are typical of biomass degradation from activities such as harvesting and cutting in croplands, as well as grazing in savannas and grasslands. Since ET can serve as a proxy for irrigation activity, the positive correlation of moderate ET values can also suggest that some AFOs in these states have irrigated fields within their property (parcel).

The educational level of residents in a county also influenced the model's predictions in these cropland-dominated states. The majority of AFO parcels identified by the model were located in counties where only a small percentage of residents over the age of 25 held an associate degree (denoted as E25_AD). While this observation was not influential at the national scale, it corresponds with findings from a Midwest-focused study by Carrel et al. (2016), which analyzed swine CAFOs in Iowa and noted generally lower levels of college education among residents in these areas.

Since most cropland-dominated states were within the 'Corn Belt,' which contributed 35 % of US agricultural ammonia emissions in 2014 (Hu et al., 2021), NO_3^- particulates positively influenced the identification of AFOs (Fig. 4-a). However, as illustrated in Fig. 3, parcels used for crop cultivation and those housing AFOs can demonstrate similar trends in air pollution levels, making it challenging to distinguish between these two land uses based on this metric alone. As a result, the

Table 2
State level description of RF classification results, test data and land cover.

State	Model Results						Test Data Statistics						
					Percentage (%) of Test Data		Percentage (%) of Test Data						
	Accuracy Score	Precision Score (AFO)	Recall Score (AFO)	Recall Score (non-AFO)	Predicted False Positives	Predicted False Negatives	Samples in State	Cropland	Shrubland	Forest	Savannas	Grasslands	Urban
Alabama	0.80	0.76	1.00	0.43	20.00	0.00	0.84	15.00	0.00	5.00	70.00	10.00	0.00
Arizona	0.33	1.00	0.33	0.00	0.00	66.67	0.13	33.00	67.00	0.00	0.00	0.00	0.00
Florida	0.71	0.33	0.05	0.96	2.78	26.39	3.02	6.00	0.00	0.00	57.00	37.00	0.00
Iowa	0.88	0.89	0.99	0.08	10.75	1.10	19.11	99.00	0.00	0.00	1.00	0.00	0.00
Indiana	0.85	0.89	0.94	0.27	10.26	5.13	13.08	95.00	0.00	1.00	3.00	1.00	0.00
Louisiana	0.96	0.95	1.00	0.90	3.51	0.00	2.39	4.00	0.00	0.00	70.00	13.00	13.00
Michigan	0.98	0.98	1.00	0.50	2.04	0.00	2.05	94.00	0.00	0.00	4.00	0.00	2.00
Minnesota	0.81	0.84	0.93	0.36	13.91	5.22	4.82	89.00	0.00	3.00	6.00	3.00	0.00
Missouri	0.93	0.94	0.98	0.54	5.26	1.75	4.78	73.00	0.00	3.00	18.00	6.00	0.00
Mississippi	0.88	0.88	0.98	0.44	10.62	1.47	11.44	4.00	0.00	5.00	89.00	2.00	0.00
North Carolina	0.83	0.85	0.94	0.42	12.68	4.51	14.88	30.00	0.00	4.00	63.00	3.00	0.00
Ohio	0.86	0.67	1.00	0.81	13.64	0.00	0.92	55.00	0.00	0.00	9.00	5.00	31.00
Oregon	0.93	0.96	0.96	0.73	3.26	3.26	3.86	51.00	0.00	10.00	9.00	29.00	1.00
Pennsylvania	0.98	0.98	1.00	0.75	1.56	0.00	2.68	86.00	0.00	5.00	9.00	0.00	0.00
South Carolina	0.94	0.96	0.97	0.50	3.67	2.75	9.14	8.00	0.00	2.00	77.00	10.00	3.00
Tennessee	1.00	1.00	1.00	1.00	0.00	0.00	0.13	67.00	0.00	0.00	33.00	0.00	0.00
Texas	0.82	0.84	0.94	0.44	13.85	4.62	2.72	47.00	0.00	2.00	5.00	45.00	1.00
Wisconsin	0.70	0.50	0.72	0.69	21.88	8.33	4.02	54.00	0.00	19.00	21.00	1.00	5.00

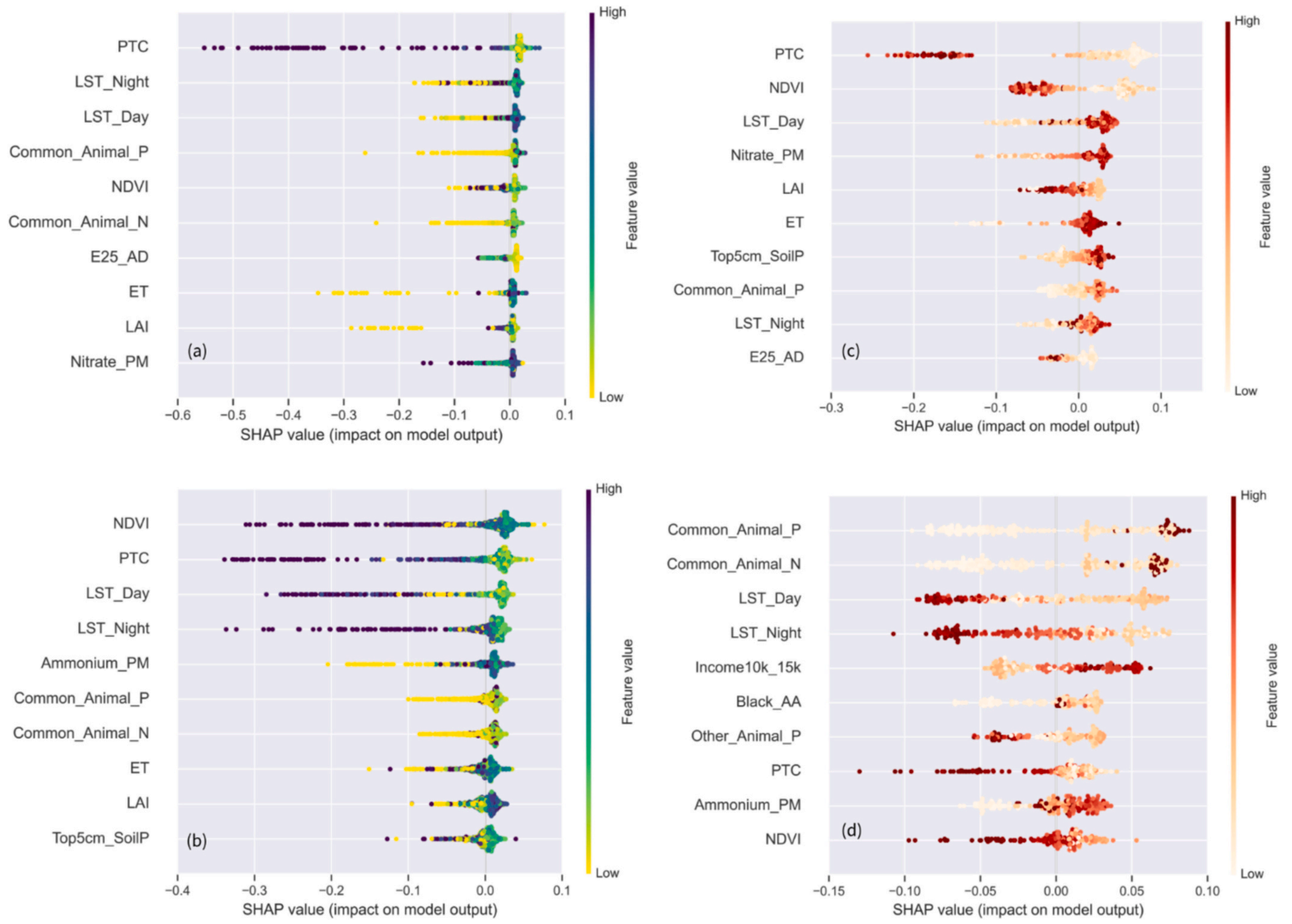


Fig. 4. State level results of the Random Forest classification model showing Shapley Additive exPlanation (SHAP) summary plot of features influencing (a) Animal Feeding Operation (AFO) parcel predictions in cropland-dominated states, (b) AFO parcel predictions in savanna-dominated states, (c) false positive prediction of AFO parcels in cropland-dominated states, and (d) false positive prediction of AFO parcels in savanna-dominated states.

model leveraged other features to distinguish AFO from non-AFO parcels more effectively. Thus, while NO_3^- emissions may be a significant factor in identifying AFO land parcels, their influence was less decisive than the other features.

The ability of our model to correctly identify AFO land parcels with high recall is crucial for nutrient sustainability and the protection of at-risk communities through environmental justice initiatives, particularly in IA and other contiguous Upper Midwestern states. The Upper Midwest hosts the majority of the US swine operations, with IA producing about 31.4 % of the total swine livestock in the US. MN and IN follow with 12 % and 5.5 % respectively (USDA, 2015). However, more than any other state, IA puts communities around AFOs at risk by maintaining one of the lowest rates of NPDES permitting, thereby making the state an attractive destination for major meat-based brands such as Hormel, Smithfield, and Cargill (Kolbe, 2013; Carrel et al., 2016). Based on the NPDES implementation records, of the 12,367 AFOs registered in the Iowa Department of Natural Resources database, only 4 % were under NPDES regulations.

Features Influnecing False Positive Predictions of AFO Parcels in States Dominated by Cropland — While 57 % of the total false positives in our study occurred in cropland-dominated states, approximately 91 % of these false positives originated from crop-growing parcels. A SHAP analysis (see Fig. 4-c) exploring the reasons behind these statistics revealed that, similar to actual AFO identification, the lack of tree cover was the most decisive factor leading our model to misidentify non-AFO parcels as AFOs, indicating that in these states AFO and their surroundings are more environmentally degraded than many cultivated areas. As a result, crop parcels characterized by comparatively lower NDVI and more extreme daytime LST were erroneously classified by the model as having an AFO facility.

Further analysis of our model results also corroborated that environmentally vulnerable crop parcels in agriculturally intensive states were more likely to be misidentified as AFOs by the model. For instance, parcels with high NO_3^- emissions and elevated topsoil P levels (Fig. 4-c) were also frequently misclassified by the model as AFOs. The likely cause of these misclassifications can be due to the spread of manure nutrients beyond the boundaries of AFO parcels, especially in counties with high livestock and poultry-based manure production. Among manure nutrients, P proved to be particularly influential, highlighting its sensitivity in detecting environmental patterns indicative of AFOs.

In addition to these dominant features, other factors, although limited in their influence, also contribute to the model's tendency to misclassify. These features also act as filters by mirroring common characteristics found in actual AFOs in these cropland-dominated states. In our study, ET, and education levels of residents above the age of 25 (E25_AD) were found to be two such features (Fig. 4-c).

The remaining 9 % of false positive predictions occurred in savanna and grassland parcels in the Upper Midwest. These areas exhibit terrestrial characteristics similar to those of AFO parcels, likely influenced by intense grazing, which contributed to their misclassification.

Elevated NO_3^- and soil P levels, likely from the accumulation of pasture manure and urine, along with their presence in counties known for high livestock and poultry-based manure P outputs, further corroborate this analysis. Additionally, the educational levels of individuals over the age of 25 (E25_AD) in these regions were comparable to those residing near AFO parcels, which also increased their likelihood of misclassification.

3.2.2. States with savannas as the dominant land cover

In addition to cropland, the RF model developed for this study also demonstrated considerable reliability in identifying AFOs in states where savannas were the predominant land cover. States with such characteristic account for about 42 % of the test parcels. According to our results (Table 2), AL and LA achieved the highest recall scores of 1.0, effectively identifying all AFO parcels within these states. They were followed by MS, SC, and NC, each with recall scores above 0.94. However, FL recorded a significantly lower recall score of 0.05, mainly because 26.4 % of its AFO parcels were misclassified as false negatives by the model. Despite this, the model achieved a non-AFO recall score of 0.96 in FL, indicating that nearly all actual non-AFO parcels were correctly identified.

The effectiveness of our model in identifying AFOs is particularly crucial in NC, a state recognized as a nutrient hotspot and responsible for contributing 13 % of total swine production in the US (USDA, 2015). Research by Yang et al. (2016) further highlighted this and noted that this contribution has been increasing ever since. Between 1930 and 2012, N and P levels generated from animal manure in NC increased by approximately 70 %, leading to a 40 % and 70 % rise in N and P loading to water bodies, respectively (Yang et al., 2016).

Features Influencing AFO Parcel Predictions in States Dominated by Savanna — In states dominated by savannas, our analysis found that 76 % of AFOs are located within this land cover. Additionally, 17 % are found in croplands, 6 % are in grasslands, and 1 % are situated in forests. NDVI emerged as the most effective indicator for identifying AFO parcels in this region. Higher NDVI values, indicative of healthier plants, led the model to classify parcels as non-AFO, while low-to-moderate NDVI values swayed the model towards AFO predictions. The pattern can be attributed to the greater prevalence of vegetation creating a more distinct NDVI spread, unlike in cropland-dominated states where a greater proportion of area is under anthropogenic impact resulting in less NDVI variation. This is evident from the narrower spread in NDVI values around the baseline, as shown in Fig. 4-a. Additionally, factors such as low to moderate tree cover, along with low-to-moderate daytime and nighttime LST values—indicators of reduced biomass conditions in croplands, savannas, and grasslands—also played a significant role in enhancing the accuracy of AFO location predictions, as detailed in Fig. 4-b.

Moderate-to-high levels of NH_4^+ particulates were also observed to positively influence AFO prediction in savanna-dominated states, supporting the notion that even in regulated surroundings, AFOs can cause some form of environmental deterioration. As a result, within the context of model training, air quality surrounding grassland and savanna parcels substantially influenced the model's tendency to predict the presence of AFO parcels. This effect was more pronounced than the air quality around AFOs and crop-growing parcels since they often exhibit similar patterns (Fig. 3).

In these savanna-dominated states, the model also demonstrated increased sensitivity to even slight changes in manure nutrient data. As a result, it identified more AFO parcels in counties with moderate-to-high levels of livestock and poultry-based manure nutrients, primarily P, viewing these areas as more animal-intensive. This observation aligns with findings from Bian et al. (2021), who reported that since 1980, manure N and P production have become increasingly concentrated, creating nutrient hotspots in the Southeastern US, especially in NC, while neighboring areas have experienced a decline in these nutrients.

Additionally, moderate ET values positively correlated with the prediction of AFO parcels, reflecting attributes of herbaceous vegetation, crop growth, and even irrigation. For the LAI, both moderate and high values shifted the model towards predicting AFOs. However, since low NDVI values sometimes correlate more strongly with AFOs, it can be inferred that even in some areas where the vegetation is dense, its overall health is compromised by factors such as nutrient deficiency, water stress, or diseases. Furthermore, although less influential than other features, higher soil P levels positively impacted the model's ability to identify AFOs, reflecting a pattern consistent with manure nutrient generation.

Features Influencing False Positive Predictions of AFO Parcels in States Dominated by Savanna — In these six states, AL recorded the highest false positive rate at 20 %, followed by NC at 13 %, and MS at 10.6 %, as detailed in Table 2. Similar to AFO prediction, the RF model was found to be sensitive to slight changes in livestock and poultry-based manure nutrients, especially P. A SHAP summary plot (Fig. 4-d) indicated that misidentification of non-AFO to AFO parcels primarily occurred in counties exhibiting moderate-to-high values for manure nutrients. Contributing features supporting the misclassification also included moderate daytime LST, low-to-moderate nighttime LST, and moderate-to-high NH_4^+ emissions, as illustrated in Fig. 4-d.

The misidentification of non-AFO parcels as AFOs in these six states can also be attributed to the prevalence of non-AFO parcels in areas with a significant presence of African American families earning between US $10,000 and US$15,000 (Income10k_15k; Fig. 4-d). This observation is consistent with the broader national trends identified in our study and corroborated by various studies mentioned previously, which noted that residents living near these facilities predominantly belong to communities of color and are financially disadvantaged.

3.2.3. States with mixed land cover as the dominant land cover

States such as TX and AZ, which feature mixed land cover, accounted for 3 % of the test parcels, as detailed in Table 2. In TX, approximately 47 % of the parcels were designated for agricultural use, and about 45 % were covered by grasslands. The remaining parcels were comprised of savannas (5 %), forests (2 %), and urban pockets (1 %). In contrast, AZ parcels were predominantly under shrubland (67 %), with crops making up the remaining 33 %.

For TX, the RF model achieved a high recall score of 0.94 and recorded a false positive rate of 13.8 %. In contrast, AZ saw a much lower recall score of 0.33, with a substantial 66.7 % of AFOs misclassified as false negatives (Table 2). The limited dataset from AZ, which constituted only 0.13 % of the testing data, contributed to these poor results. While the model successfully identified an AFO within an agricultural parcel in AZ, it failed to recognize AFOs in parcels characterized by shrubland. These findings underscore the need for more comprehensive data that encompasses the full range of geographical and climatic regions across the US. As result, given the superior data quality in TX among states with mixed land cover, we focused our detailed analysis exclusively on the factors influencing model performance within this state.

Features Influencing AFO Parcel Predictions in States Dominated by Mixed Land Use —Based on SHAP summary plot for TX (Supplemental Fig. S1-a) low nighttime LST, indicative of biomass loss from agricultural activity, along with high NO_3^- particulates and low-to-moderate daytime LST, were identified as strong predictors for AFO parcels near crop fields. On the other hand, AFO parcels near grasslands and savannas were associated with moderate-to-high nighttime LST, moderate daytime LST, and moderate NO_3^- particulates. Counties generating high livestock and poultry-based manure nutrients significantly influenced the identification of AFO parcels within them, while counties moderate nutrient values had a negative impact. Among nutrients, P demonstrated particular sensitivity, a trend that is consistent across various states. Additionally, the proximity of AFO facilities to meat processing plants also aided the model in detecting AFO parcels. Our results revealed that most AFOs in Texas are located a short distance from processing facilities, with a significant proportion of these being dairy operations (Rompala, 2023).

Features Influencing False Positives Predictions of AFO Parcels in States Dominated by Mixed Land Use — The analysis of false positives in TX using SHAP values (Supplemental Fig. S1-b) showed that parcels incorrectly identified as AFOs were often located in areas where households typically earn between US$25,000 and US$35,000 annually (Income25k_35k). Additionally, moderate nighttime LST and high NO_3^- emissions were also significant contributors to the model incorrectly classifying non-AFO parcels as AFOs. Furthermore, the model frequently misclassified non-AFO parcels in counties with elevated levels of N and P from livestock and poultry manure.

The impact of socioeconomic factors in TX reflects the environmental justice concerns commonly associated with AFOs across the US. Additional metrics indicative of poverty, such as low educational qualification (less than 9th grade; see E25_less_9 in Supplemental Fig. S1-b) and lack of health insurance coverage (i.e., with_public_HIC), also strongly influenced false positive predictions. In contrast, higher income levels (Income75k_100k) had the opposite effect. These findings are further supported by a study from Salzano (2023), which noted a positive association between poverty and high exposure to CAFOs in TX.

3.2.4. *Features influencing false negative predictions of AFO parcels across all states*

Although smaller in proportion, our model also misidentified 4.7 % of actual AFO parcels as non-AFOs (false negatives). Understanding these misclassifications is essential since they impact the overall accuracy of our model predictions. These false negative predictions can be considered as 'rare events', representing AFO parcels in conditions not commonly seen in most states. A SHAP analysis (Supplemental Fig. S2) conducted across eighteen states, with a particular focus on Florida, revealed that the RF model tends to misclassify AFOs as non-AFOs in parcels with high nighttime LST and substantial tree cover, suggesting that although the model can recognize some AFOs near low albedo areas such as forests, it may not do so consistently without additional features indicating otherwise. The capacity of the model for false positive predictions also increased in parcels that had high atmospheric NO_3^- but were located in counties with low N and P production from livestock and poultry operations, highlighting the limitations of our air quality data where atmospheric influences from non-agricultural sources can skew results. Additionally, the model's tendency to mistakenly categorize an AFO as a non-AFO was also positively influenced by counties having comparatively higher educational levels among residents, specifically those with associate or bachelor's degrees (as shown in Supplemental Fig. S2). This finding stands in contrast to the general trend observed in AFO locations, where the majority of surrounding populations tend to be less educated and poor (Son et al., 2021).

In our study covering eighteen states, the model successfully identified AFO and non-AFO parcels with an accuracy exceeding 80 % in fourteen of them, as illustrated in Supplemental Fig. S3. Notably, the model achieved an accuracy of over 85 % in most states across the Midwest, Northeast, West, and Southeast regions where cropland or savanna predominated as the land cover. FL, however, was an exception; as one of the largest states in the study, it recorded an accuracy of only 71 %. A map representing all the parcel locations, as well as those correctly identified as AFOs, non-AFOs, and those incorrectly identified as false positives and negatives, is illustrated in Fig. 5-a and b.

Predicting AFOs where there are none may help to improve larger-scale ecosystem management as these areas may be at risk from nearby AFO activities or future AFO expansion. This predictive capability facilitates early intervention and enhances monitoring of potential deteriorations in water, soil, and air quality, thus preempting environmental issues before they escalate. Additionally, by highlighting these ecologically vulnerable parcels, our findings can drive policy changes and direct resources towards improving the health and welfare of affected communities.

Considering both national and state-level results, PTC emerged as the most robust predictor in determining whether an area qualifies as an AFO, contributing to accurate true positive identifications. Across various states, the effectiveness of PTC is reinforced by the NDVI, daytime and nighttime LST, and the amount of P from manure produced by livestock and poultry. However, the extent to which these parameters influence AFO prediction varies by state, due to differing environmental and operational characteristics. As a result, instead of relying on a uniform national model, there is a crucial need for regional machine learning models that utilize only the features specific to each region to identify AFO locations. By adopting this approach, the models could provide more precise and locally relevant insights about AFO prevalence, taking into account the extent of anthropogenic activity as well as the effects of different animal types (e.g., cattle vs. poultry) on variables such as manure composition, atmospheric emissions, and soil nutrient accumulation.

3.3. *Study limitations*

Although our study focuses on a parcel scale, many of our predictive features operate at scales larger than individual parcels. While this approach reduces computational complexity in predicting AFO locations, it also leaves significant room for improvement. Environmental data with resolutions ranging between 250 and 1000 m can blend signals from multiple land-use types within a single pixel, thereby obscuring granular information and causing feature signals from smaller parcels to be lost entirely. Similarly, socioeconomic-based predictive features aggregated at the county level fail to capture diverse parcel-specific characteristics. Additionally, census data, the source of the socioeconomic features, typically are estimated at fixed intervals and can fail to capture dynamic changes around individual parcels.

Our RF modeling effort can also be affected by uncertainties arising from the soil P predictor. Since the soil P data is derived from an interpolated soil P map, it assumes spatial continuity. This assumption may not accurately reflect actual conditions, particularly in areas with sparse sampling points. Moreover, in areas where sample points are unevenly distributed, interpolation can be less reliable, leading to potential biases in predictions. Finally, we are limited by the training data available for our model. Future efforts should focus on expanding these data to additional land uses and regions, as well as incorporating data such as animal type and facility structure. We also did not have an equal amount of AFO and non-AFO data points for our training process. While additional testing showed little impact on our results when the ratio of AFO to non-AFOs was changed, future work could explore the expanse and variability of non-AFO data for its impact on model predictions.

4. Conclusion

We provided the first study identifying AFO locations on a parcel level, utilizing a unique approach that overcomes the limitations commonly faced by traditional mapping methods based on aerial images. By leveraging publicly available datasets that include geographical, environmental, and socio-economic variables, our study developed a RF-based machine learning model that effectively discerned patterns and identified properties with AFO facilities with high accuracy.

Although the development of new large language models significantly advances image classification through multimodal learning, feature extraction, and semantic understanding, the method developed for our study is not only easy to implement and replicable but also aims to democratize information access. Our approach uniquely addressed AFO identification by focusing on publicly available environmental data and socioeconomic data, many of which are also available globally. Additionally, our method provided a robust understanding of terrestrial environmental patterns in landscapes dominated by crop and animal agriculture across various US regions. It proved particularly effective in identifying AFOs and nutrient hotspots in the Midwestern and Southeastern states, underscoring its significance as an essential tool for

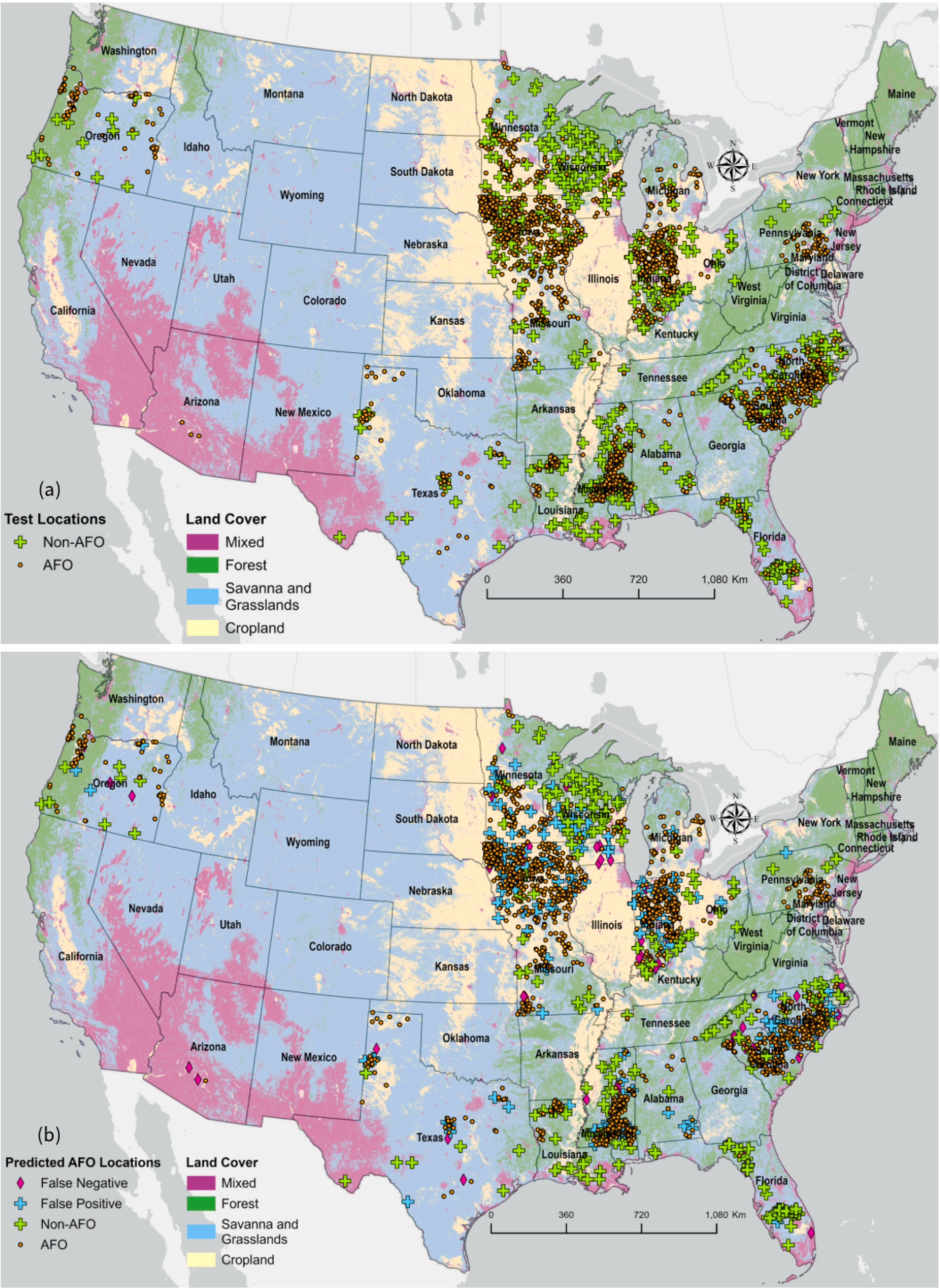


Fig. 5. Map showing (a) test Animal Feeding Operation (AFO) and non-AFO locations across the eighteen US states considered in our study. The background depicts four primary land cover groups across contiguous US, where Forest refers to forest land cover, and mixed refers to all land covers other than forests, savannas, grasslands, and croplands, and (b) predicted results from the random forest classification model used in our study showing false positives, false negatives, correctly predicted non-AFO (i.e., non-AFO) and AFO (i.e., AFO) locations, against the same land cover background.

environmental monitoring and nutrient management studies. By incorporating manure nutrients as features of animal agriculture, our model identified P as a more significant ecological marker for predicting AFO locations than N, further emphasizing its potential in nutrient management studies, especially in developing targeted source-sink-based Pmanagement strategies. Additionally, given the ease of replicability due to the use of public data sources, our approach could be used globally to identify AFOs in other countries, by focusing on extent of tree cover (PTC) and vegetation indices as metrics for identifying AFO locations.

Although our current approach focused on locating AFOs at a parcel level was successful, it was limited to only eighteen states. Efforts are ongoing to extend this strategy to cover the entire US. Future work will also involve updating our input datasets with high-resolution land cover and methane emissions data. Since manure lagoons are well-documented sources of methane emissions, incorporating this information could enhance the accuracy and utility of our model, improving its application in ongoing environmental monitoring and management efforts.

CRediT authorship contribution statement

Arghajeet Saha: Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis, Conceptualization. **Barira Rashid:** Data curation. **Ting Liu:** Data curation. **Lorrayne Miralha:** Data curation, Writing – review & editing. **Rebecca L. Muenich:** Writing – review & editing, Visualization, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Rebecca Muenich reports financial support was provided by National Science Foundation. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This project was supported by the Science and Technologies for Phosphorus Sustainability (STEPS) Center under National Science Foundation (NSF) award number CBET-2019435. All conclusions are those of the authors and do not necessarily reflect the views of the NSF. Regrid's Data With Purpose program (www.regrid.com/purpose) provided the parcel data used in this research. Regrid is a leading provider of nationwide land parcel data and matches a range of geospatial data to parcel boundaries nationally.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2024.178312.

Data availability

Data will be made available on request.

References

Ahmad, M.W., Mourshed, M., Rezgui, Y., 2017. Trees vs neurons: comparison between random forest and ANN for high-resolution prediction of building energy consumption. Energ. Buildings 147, 77–89. https://doi.org/10.1016/j.enbuild.2017.04.038.

Allen, R.G., Pereira, L.S., Howell, T.A., Jensen, M.E., 2011. Evapotranspiration information reporting: I. Factors governing measurement accuracy. Agric Water Manag 98 (6), 899–920. https://doi.org/10.1016/j.agwat.2010.12.015.

Aneja, V.P., Nelson, D.R., Roelle, P.A., Walker, J.T., Battye, W., 2003. Agricultural ammonia emissions and ammonium concentrations associated with aerosols and precipitation in the Southeast United States. J. Geophys. Res. Atmos. 108 (D4). https://doi.org/10.1029/2002jd002271.

Arbuckle, T.E., Sherman, G.J., Corey, P.N., Walters, D., Lo, B., 1988. Water nitrates and CNS birth defects: a population-based case-control study. Archives of Environmental Health: An International Journal 43 (2), 162–167. https://doi.org/10.1080/00039896.1988.9935846.

Asner, G.P., Scurlock, J.M., Hicke, A., J., 2003. Global synthesis of leaf area index observations: implications for ecological and remote sensing studies. Glob. Ecol. Biogeogr. 12 (3), 191–205. https://doi.org/10.1046/j.1466-822x.2003.00026.x.

Azevedo, J.A., Chapman, L., Muller, C.L., 2016. Quantifying the daytime and night-time urban heat island in Birmingham, UK: a comparison of satellite derived land surface temperature and high-resolution air temperature observations. Remote Sens. (Basel) 8 (2), 153. https://doi.org/10.3390/rs8020153.

Battye, R., Battye, W., Overcash, C., Fudge, S., 1994. Development and selection of ammonia emission factors. Retrieved from. https://www.osti.gov/biblio/6763800.

Bian, Z., Tian, H., Yang, Q., Xu, R., Pan, S., Zhang, B., 2021. Production and application of manure nitrogen and phosphorus in the United States since 1860. Earth Syst. Sci. Data 13, 515–527. https://doi.org/10.5194/essd-13-515-2021.

Bouwman, A.F., Beusen, A.H.W., Lassaletta, L., Van Apeldoorn, D.F., Van Grinsven, H.J. M., Zhang, J., Ittersum Van, M., K., 2017. Lessons from temporal and spatial patterns in global use of N and P fertilizer on cropland. Sci. Rep. 7 (1), 40366. https://doi.org/10.1038/srep40366.

Burkholder, J., Libra, B., Weyer, P., Heathcote, S., Kolpin, D., Thorne, P.S., Wichman, M., 2007. Impacts of waste from concentrated animal feeding operations on water quality. Environ. Health Perspect. 115 (2), 308–312. https://doi.org/10.1289/ehp.8839.

Carrel, M., Young, S.G., Tate, E., 2016. Pigs in space: determining the environmental justice landscape of swine concentrated animal feeding operations (CAFOs) in Iowa. Int. J. Environ. Res. Public Health 13 (9), 849. https://doi.org/10.3390/ijerph13090849.

Centner, T.J., 2010. Empty Pastures: Confined Animals and the Transformation of the Rural Landscape. University of Illinois Press. Retrieved from. https://www.google.com/books/edition/Empty_Pastures/Bom4ztdHxMkC?hl=en&gbpv=1&dq=Centner,+T.+J.+(2010).+Empty+pastures:+Confined+animals+and+the+transformation+of+the+rural+landscape.+University+of+Illinois+Press.&pg=PP2&printsec=frontcover.

Chowdhury, R., Heng, K., Shawon, M.S.R., Goh, G., Okonofua, D., Ochoa-Rosales, C., Gonzalez-Jaramillo, V., Bhuiya, A., Reidpath, D., Prathapan, S., Shahzad, S., 2020. Dynamic interventions to control COVID-19 pandemic: a multivariate prediction modelling study comparing 16 worldwide countries. Eur. J. Epidemiol. 35, 389–399. https://doi.org/10.1007/s10654-020-00649-w.

Cole, D., Todd, L., Wing, S., 2000. Concentrated swine feeding operations and public health: a review of occupational and community health effects. Environ. Health Perspect. 108 (8), 685–699. https://doi.org/10.2307/3434721.

Devlin, M., Brodie, J., 2023. Nutrients and Eutrophication. In: Reichelt-Brushett, A. (Ed.), Marine Pollution – Monitoring, Management and Mitigation. Springer Nature Switzerland, Cham, pp. 75–100. https://doi.org/10.1007/978-3-031-10127-4_4.

Environmental Working Group (EWG), 2016. Exposing Fields of Filth EWG Online. https://www.ewg.org/research/exposing-fields-filth.

Falcone, J.A., 2021. Estimates of county-level nitrogen and phosphorus from fertilizer and manure from 1950 through 2017 in the conterminous United States (No. 2020-1153). In: US Geological Survey. https://doi.org/10.3133/ofr20201153.

Fisher, J.A., Freeman, L.E.B., Hofmann, J.N., Blair, A., Parks, C.G., Thorne, P.S., Ward, M.H., Jones, R.R., 2020. Residential proximity to intensive animal agriculture and risk of lymphohematopoietic cancers in the Agricultural Health Study. Epidemiology 31 (4), 478–489. https://doi.org/10.1097/ede.0000000000001186.

Fowler, N.L., Beckage, B., 2019. Savannas of North America. In: Savanna Woody Plants and Large Herbivores, pp. 123–150. https://doi.org/10.1002/9781119081111.ch5.

Handan-Nader, C., Ho, D.E., 2019. Deep learning to map concentrated animal feeding operations. Nature Sustainability 2 (4), 298–306. https://doi.org/10.1038/s41893-019-0246-x.

Hino, M., Benami, E., Brooks, N., 2018. Machine learning for environmental monitoring. Nature Sustainability 1, 583–588. https://doi.org/10.1038/s41893-018-0142-9.

Hribar, C., 2010. Understanding concentrated animal feeding operations and their impact on communities. Retrieved from. https://stacks.cdc.gov/view/cdc/59792.

Hristov, A.N., 2011. Contribution of ammonia emitted from livestock to atmospheric fine particulate matter (PM2. 5) in the United States. J. Dairy Sci. 94 (6), 3130–3136. https://doi.org/10.3168/jds.2010-3681.

Hu, C., Griffis, T.J., Frie, A., Baker, J.M., Wood, J.D., Millet, D.B., Yu, Z., Yu, X., Czarnetzki, A.C., 2021. A multiyear constraint on ammonia emissions and deposition within the US corn belt. Geophys. Res. Lett. 48 (6), e2020GL090865. https://doi.org/10.1029/2020GL090865.

Ishtiaque, A., Myint, S.W., Wang, C., 2016. Examining the ecosystem health and sustainability of the world's largest mangrove forest using multi-temporal MODIS products. Sci. Total Environ. 569, 1241–1254. https://doi.org/10.1016/j.scitotenv.2016.06.200.

Kalcic, M.M., Frankenberger, J., Chaubey, I., Prokopy, L., Bowling, L., 2015. Adaptive targeting: engaging farmers to improve targeting and adoption of agricultural conservation practices. JAWRA Journal of the American Water Resources Association 51 (4), 973–991. https://doi.org/10.1111/1752-1688.12336.

Key, N., McBride, W.D., Ribaudo, M., Sneeringer, S., 2017. Trends and Developments in Hog Manure Management: 1998-2009 (SSRN Scholarly Paper No. 2981722). Social Science Research Network, Rochester, NY.

Kolbe, E.A., 2013. Won't you be my neighbor: living with concentrated animal feeding operations. Iowa L. Rev. 99, 415. https://doi.org/10.2139/ssrn.2387639.

Kronberg, S.L., Ryschawy, J., 2019. Negative impacts on the environment and people from simplification of crop and livestock production. In: Agroecosystem Diversity. Academic Press, pp. 75–90. https://doi.org/10.1016/B978-0-12-811050-8.00005-4.

Kümmerer, K., 2004. Resistance in the environment. J Antimicrobial Chemotherapy 54 (2), 311–320. https://doi.org/10.1093/jac/dkh325.

Lenhardt, J., Ogneva-Himmelberger, Y., 2013. Environmental injustice in the spatial distribution of concentrated animal feeding operations in Ohio. Environmental Justice 6 (4), 133–139. https://doi.org/10.1089/env.2013.0023.

Long, C.M., Muenich, R.L., Kalcic, M.M., Scavia, D., 2018. Use of manure nutrients from concentrated animal feeding operations. J. Great Lakes Res. 44 (2), 245–252. https://doi.org/10.1016/j.jglr.2018.01.006.

Lopatto, E., Choi, J., Colina, A., Ma, L., Howe, A., Hinsa-Leasure, S., 2019. Characterizing the soil microbiome and quantifying antibiotic resistance gene dynamics in agricultural soil following swine CAFO manure application. PloS One 14 (8), e0220770. https://doi.org/10.1371/journal.pone.0220770.

Lorimor, J., Powers, W., Sutton, A., 2004. Manure Characteristics, Second edition. MidWest Plan Service, Ames, IA. MWPS-18 Section 1. 24 p. Retrieved from. https://www.canr.msu.edu/uploads/files/ManureCharacteristicsMWPS-18_1.pdf.

Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, p. 30. Retrieved from. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

MacDonald, D.D., Ingersoll, C.G., Berger, T.A., 2000. Development and evaluation of consensus-based sediment quality guidelines for freshwater ecosystems. Arch. Environ. Contam. Toxicol. 39, 20–31. https://doi.org/10.1007/s002440010075.

MacDonald, J.M., Hoppe, R.A., Newton, D., 2018. Three Decades of Consolidation in U.S. Agriculture (No. EIB-189). USDA Economic Research Service. Retrieved from. https://ageconsearch.umn.edu/record/276247/?v=pdf.

Mallin, M.A., Cahoon, L.B., 2003. Industrialized animal production—a major source of nutrient and microbial pollution to aquatic ecosystems. Popul. Environ. 24, 369–385. https://doi.org/10.1023/a:1023690824045.

Martin, K.L., Emanuel, R.E., Vose, J.M., 2018. Terra incognita: the unknown risks to environmental quality posed by the spatial distribution and abundance of concentrated animal feeding operations. Sci. Total Environ. 642, 887–893. https://doi.org/10.1016/j.scitotenv.2018.06.072.

Miralha, L., Muenich, R.L., Schaffer-Smith, D., Myint, S.W., 2021. Spatiotemporal land use change and environmental degradation surrounding CAFOs in Michigan and North Carolina. Sci. Total Environ. 800, 149391. https://doi.org/10.1016/j.scitotenv.2021.149391.

Miralha, L., Sidique, S., Muenich, R.L., 2022. The spatial organization of CAFOs and its relationship to water quality in the United States. J. Hydrol. 613, 128301. https://doi.org/10.1016/j.jhydrol.2022.128301.

Moses, A., Tomaselli, P., 2017. Industrial animal agriculture in the United States: concentrated animal feeding operations (CAFOs). In: International Farm Animal, Wildlife and Food Safety Law, pp. 185–214. https://doi.org/10.1007/978-3-319-18002-1_6.

Myeni, R.B., Keeling, C.D., Tucker, C.J., Asrar, G., Nemani, R.R., 1997. Increased plant growth in the northern high latitudes from 1981 to 1991. Nature 89, 698–702. https://doi.org/10.1038/386698a0.

Nicole, W., 2013. CAFOs and environmental justice: the case of North Carolina. Environ. Health Perspect. 121, a182–a189. https://doi.org/10.1289/ehp.121-a182.

Pettorelli, N., Vik, J.O., Mysterud, A., Gaillard, J.M., Tucker, C.J., Stenseth, N.C., 2005. Using the satellite-derived NDVI to assess ecological responses to environmental change. Trends Ecol. Evol. 20 (9), 503–510. https://doi.org/10.1016/j.tree.2005.05.011.

Pew Commission on Industrial Animal Farm Production, 2009. Putting meat on the table: industrial farm animal production in America. Retrieved from. http://www.ncifap.org/_images/PCIFAPFin.pdf.

Qi, J., Xin, X., John, R., Groisman, P., Chen, J., 2017. Understanding livestock production and sustainability of grassland ecosystems in the Asian Dryland Belt. Ecol. Process. 6 (1), 22. https://doi.org/10.1186/s13717-017-0087-3.

Quintanar, A.I., Mahmood, R., Loughrin, J.H., Lovanh, N., Motley, M.V., 2009. A system for estimating Bowen ratio and evaporation from waste lagoons. Appl. Eng. Agric. 25 (6), 923–932. https://doi.org/10.13031/2013.29238.

Quist, A.J., Johnston, J.E., Fliss, M.D., 2022. Disparities of industrial animal operations in California, Iowa, and North Carolina. In: Earth Justice, pp. 1–30. Retrieved from. https://earthjustice.org/wp-content/uploads/quistreport_cafopetition_oct2022.pdf.

Raff, Z., Meyer, A., 2022. CAFOs and surface water quality: evidence from Wisconsin. Am. J. Agric. Econ. 104 (1), 161–189. https://doi.org/10.1111/ajae.12222.

Robinson, C., Chugg, B., Anderson, B., Ferres, J.M.L., Ho, D.E., 2022. Mapping industrial poultry operations at scale with deep learning and aerial imagery. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 15, 7458–7471. https://doi.org/10.1109/jstars.2022.3191544.

Rompala, A., 2023. A Spatial Analysis of Beef Production and Its Environmental and Health Impacts in Texas. Master's thesis. University of Southern California. Retrieved from. https://spatial.usc.edu/wp-content/uploads/formidable/12/Amanda-Rompala-thesis.pdf.

Rosov, K.A., Mallin, M.A., Cahoon, L.B., 2020. Waste nutrients from US animal feeding operations: regulations are inconsistent across states and inadequately assess nutrient export risk. J. Environ. Manage. 269, 110738. https://doi.org/10.1016/j.jenvman.2020.110738.

Rudko, N., Muenich, R.L., Garcia, M., Xu, T., 2023. Development of a point-source model to improve simulations of manure lagoon interactions with the environment. J. Environ. Manage. 325, 116332. https://doi.org/10.1016/j.jenvman.2022.116332.

Salzano, L., 2023. Characterizing Populations Living Near Concentrated Animal Feeding Operations: Implications for Health Equity and Environmental Justice. Doctoral dissertation. Yale University. Retrieved from. https://elischolar.library.yale.edu/cgi/viewcontent.cgi?article=2332&context=ysphtdl.

Seffner, W., 1995. Natural water contents and endemic goiter—a review. International Journal of Hygiene and Environmental Medicine 196 (5), 381–398. Retrieved from. https://pubmed.ncbi.nlm.nih.gov/7727020/.

Saha, A., 2022. Mapping Animal Feeding Operations in the United States: Improvements Seen with Parcel Data. https://steps-center.org/p-week/sps-abstracts/.

Sierra Club, 2021. Why are CAFOs bad?. In: Sierra Club Michigan Chapter Retrieved from. https://www.sierraclub.org/michigan/why-are-cafos-bad.

Smith, D.B., Cannon, W.F., Woodruff, L.G., Solano, F., Kilburn, J.E., Fey, D.L., 2013. Geochemical and mineralogical data for soils of the conterminous United States (No. 801). In: US Geological Survey. https://doi.org/10.3133/ds801.

Son, J.Y., Muenich, R.L., Schaffer-Smith, D., Miranda, M.L., Bell, M.L., 2021. Distribution of environmental justice metrics for exposure to CAFOs in North Carolina, USA. Environ. Res. 195, 110862. https://doi.org/10.1016/j.envres.2021.110862.

Song, X.P., Huang, C., Feng, M., Sexton, J.O., Channan, S., Townshend, J.R., 2014. Integrating global land cover products for improved forest cover characterization: an application in North America. International Journal of Digital Earth 7 (9), 709–724. https://doi.org/10.1080/17538947.2013.856959.

Steinzor, R.I., Huang, L.Y., 2012. Manure in the Bay: A Report on Industrial Animal Agriculture in Maryland and Pennsylvania, 1206. Center for Progressive Reform Briefing Paper. https://ssrn.com/abstract=2079716.

Sterling, E.A., 2015. Linkages Between Concentrated Animal Feeding Operation (CAFO) Expansion and County Board Politics in Rural Illinois. Northern Illinois University. Retrieved from. https://www.proquest.com/docview/1722054862?pqorigsite=gscholar&fromopenview=true.

Taha, H., Sailor, D.J., Akbari, H., 1992. High-albedo Materials for Reducing Building Cooling Energy Use. https://doi.org/10.2172/10178958.

US Census Bureau. (2016). US Department of Commerce. Retrieved from https://data.census.gov/.

USDA, 2015. Overview of the United States Hog Industry. National Agricultural Statistics Service, Washington, DC, USA. Retrieved from. https://downloads.usda.library.cornell.edu/usda-esmis/files/rr171x21v/cz30pw658/js956j577/hogview-10-29-2015.pdf.

USDA FISIS, 2023. Meat, poultry and egg product inspection directory. Retrieved from. https://www.fsis.usda.gov/inspection/establishments/meat-poultry-and-egg-product-inspection-directory.

USEPA, 2004. Air Quality Criteria for Particulate Matter, volume II. US Environmental Protection Agency, Research Triangle Park. Retrieved from. https://cfpub.epa.gov/ncea/risk/recordisplay.cfm?deid=87903.

USEPA, 2005. Detecting and mitigating the environmental impact of fecal pathogens originating from confined animal feeding operations: review. Retrieved from. https://nepis.epa.gov/Exe/ZyNET.exe/P10089B1.txt?ZyActionD=ZyDocument&Client=EPA&Index=2000%20Thru%202005&Docs=&Query=&Time=&EndTime=&SearchMethod=1&TocRestrict=n&Toc=&TocEntry=&QField=&QFieldYear=&QFieldMonth=&QFieldDay=&UseQField=&IntQFieldOp=0&ExtQFieldOp=0&XmlQuery=&File=D%3A%5CZYFILES%5CINDEX%20DATA%5C00THRU05%5CTXT%5C00000024%5CP10089B1.txt&User=ANONYMOUS&Password=anonymous&SortMethod=h%7C-&MaximumDocuments=1&FuzzyDegree=0&ImageQuality=r75g8/r75g8/x150y150g16/i425&Display=hpfr&DefSeekPage=x&SearchBack=ZyActionL&Back=ZyActionS&BackDesc=Results%20page&MaximumPages=1&ZyEntry=4#.

USEPA, 2009. Animal feeding operations. Retrieved from. http://cfpub.epa.gov/npdes/home.cfm?program_id=7.

Van Donkelaar, A., Martin, R.V., Li, C., Burnett, R.T., 2019. Regional estimates of chemical composition of fine particulate matter using a combined geoscience-statistical method with information from satellites, models, and monitors. Environ. Sci. Technol. 53 (5), 2595–2611. https://doi.org/10.1021/acs.est.8b06392.

Visa, S., Ramsay, B., Ralescu, A.L., Van Der Knaap, E., 2011. Confusion matrix-based feature selection. Maics 710 (1), 120–127. Retrieved from. https://www.researchgate.net/profile/Jennifer-Seitzer-2/publication/220833258_Using_a_Genetic_Algorithm_to_Evolve_a_D_Search_Heuristic/links/545a2bea0cf2bccc49132577/Using-a-Genetic-Algorithm-to-Evolve-a-D-Search-Heuristic.pdf#page=126.

Waldrip, H.M., Campbell, T.N., Koziel, J.A., Watts, D.B., Torbert, H.A., 2023. Legacy phosphorus in Alabama Hartsells soil after long-term amendment with broiler litter, 52 (4), 897–906. https://doi.org/10.1002/jeq2.20462.

Walljasper, C., 2018. Large Animal Feeding Operations on the Rise. Investigate Midwest. Retrieved from. https://investigatemidwest.org/2018/06/07/large-animal-feeding-operations-on-the-rise-2/.

Ward, M.H., DeKok, T.M., Levallois, P., Brender, J., Gulis, G., Nolan, B.T., VanDerslice, J., 2005. Workgroup report: drinking-water nitrate and health—recent findings and research needs. Environ. Health Perspect. 113 (11), 1607–1614. https://doi.org/10.1289/ehp.8043.

Wardlow, B.D., Egbert, S.L., 2008. Large-area crop mapping using time-series MODIS 250 m NDVI data: an assessment for the US Central Great Plains. Remote Sens. Environ. 112 (3), 1096–1116. https://doi.org/10.1016/j.rse.2007.07.019.

West, B.M., Liggit, P., Clemans, D.L., Francoeur, S.N., 2011. Antibiotic resistance, gene transfer, and water quality patterns observed in waterways near CAFO farms and wastewater treatment facilities. Water Air Soil Pollut. 217, 473–489. https://doi.org/10.1007/s11270-010-0602-y.

Wickham, J.D., Wade, T.G., Riitters, K.H., 2012. Comparison of cropland and forest surface temperatures across the conterminous United States. Agric. For. Meteorol. 166, 137–143. https://doi.org/10.1016/j.agrformet.2012.07.002.

Wilson, S.M., Howell, F., Wing, S., Sobsey, M., 2002. Environmental injustice and the Mississippi hog industry. Environ. Health Perspect. 110 (Suppl. 2), 195–201. https://doi.org/10.1289/ehp.02110s2195.

Yang, Q., Tian, H., Li, X., Ren, W., Zhang, B., Zhang, X., Wolf, J., 2016. Spatiotemporal patterns of livestock manure nutrient production in the conterminous United States from 1930 to 2012. Sci. Total Environ. 541, 1592–1602. https://doi.org/10.1016/j.scitotenv.2015.10.044.

Yu, B., Shang, S., Zhu, W., Gentine, P., Cheng, Y., 2019. Mapping daily evapotranspiration over a large irrigation district from MODIS data using a novel hybrid dual-source coupling model. Agric. For. Meteorol. 276, 107612. https://doi.org/10.1016/j.agrformet.2019.06.011.

Zhu, B., Lui, N., Irvin, J., Le, J., Tadwalkar, S., Wang, C., Ouyang, Z., Liu, F.Y., Ng, A.Y., Jackson, R.B., 2022. METER-ML: a multi-sensor earth observation benchmark for automated methane source mapping. https://arxiv.org/abs/2207.11166.