Supplement of Earth Syst. Sci. Data, 18, 8801–8819, 2026
https://doi.org/10.5194/essd-18-8801-2026-supplement
© Author(s) 2026. CC BY 4.0 License.



Open Access Earth System Science Data

Supplement of

A spatially explicit dataset of agriculture liming across the contiguous United States

Samuel Shou-En Tsao et al.

Correspondence to: Samuel Shou-En Tsao (samuel.tsao@yale.edu)

The copyright of individual parts of the supplement might differ from the article licence.

Section S1

1. Liming data processing

The county level data for liming, agricultural and environmental variables was joined with the U.S. county-level shapefile in 2012

Notes for Joining the Datasets

Miami-Dade County, Florida: Previously known as Dade County, this county was renamed Miami-Dade County in 1997, and its FIPS code changed from 12025 to 12086. Thus, the county level census data with the original FIPS identifier 12025 was joined to the county level shapefile with GEOID identifier of 12086.

Broomfield County, Colorado: Created in 2001 from parts of Adams, Boulder, Jefferson, and Weld counties, Broomfield County (FIPS 8014) did not exist during the 1954-1987 period. Therefore, we assumed its liming application to be 0 for this time frame. In this case there is a GEOID identifier of 8014, but no original FIPS identifier in the county level census data.

Baltimore, MD; St. Louis, MO; and independent cities in Virginia: Independent cities, such as Baltimore, St. Louis, and several cities in Virginia, are administratively independent from any county and are not included in county-level agricultural data from the U.S. Bureau of the Census. Since agricultural data is typically reported at the county level and these independent cities do not have significant agricultural activity, lime application is assumed to be 0 for these regions. In these cases, in our published dataset, there are GEOID identifiers but no original FIPS identifiers in the county level census data.

1974 Limed Area: The 1974 data do not report limed area (acres). Therefore, for 1974, we interpolated the limed area (acres) and lime application rate (tons per acre) using data from the previous census years, 1969 and 1978.

Cropland Area: Cropland area values recorded as 4.000000e+10 are treated as errors and marked as missing values.

Table S1. Predictor Variables by Category, Temporal Range, and Source

Category	Variable	VIF> 10	Exclude in final analysis	Temporal Range	Source
Agriculture	Land in farms (acres)			Yearly varying	U.S. Census of Agriculture
	Harvested crop land (acres) *Transformed to Cropland harvested percentage (%)*				
	Irrigated land (acres)				
	Total Cropland (acres)				
	Cropland used only for pasture (acres) *Transformed to Cropland pastured percentage (%)*				
	Cropland not harvested/pastured (acres) *Transformed to Cropland not harvested/pastured percentage (%)*				
	Approximate land area (acres)				
	Corn for grain or seed (bu.)				
	Corn for silage (bu.)				
	green chop (tons, green)				
	Wheat for grain (bu.)				

Category	Variable	VIF> 10	Exclude in final analysis	Temporal Range	Source
	Sorghum for grain or seed (bu.)			Yearly varying	U.S. Census of Agriculture
	Cotton (bales)				
	Tobacco (pounds)				
	Soybeans, beans (bu.)				
	Alfalfa & alfalfa mixtures (tons)				
Nutrient Input	N fertilizer (kg)	True		Yearly varying	Falcone (2021)
	P fertilizer (kg)	True	Excluded		
	N manure (kg)	True			
	P manure (kg)	True	Excluded		
Climate	Mean Precipitation (in.)			Static (Decadal mean)	PRISM Climate Group
	Mean Temperature (°F)				
Soil Properties	Soil pH			Static	Walkinshaw et al. (2023), UC Davis Soil Properties
	pH recommendations				
	Clay (%)	True			
	Silt (%)	True	Excluded		
	Sand (%)	True			
	Cation Exchange Capacity (cmol(+)/kg)				

Category	Variable	VIF> 10	Exclude in final analysis	Temporal Range	Source
	Soil Organic Matter (kg/m²)				
	$CaCO_3$ stock (kg/m²)				
	Rock Fragments (0–25 cm) %				
	Bulk Density (g/cm³),				
	Soil Depth (cm),				
	Electrical Conductivity (dS/m)				
	Sodium Adsorption Ratio				
	Depth to Restrictive Layer (cm)				
	Available Water Storage (cm)	VIF~9.9	Exclude		
	Sat. Hyd. Conductivity (Ksat) – Mean				
Lithology	Carbonate Sedimentary Rocks (%)			Static	Hartmann & Moosdorf (2012); GLiM database
Other Physical Factors	Wind Erodibility Index (tons/ac/yr)			Static	Walkinshaw et al. (2023)

Table S2 Counties or independent cities that did not report lime application weight and limed area, and the associated missing land area and cropland area

Year	Variable	Missing counties or independent cities	Missing counties or independent cities (%)	Missing land area (%)	Missing cropland area (%)
1954	Lime (acres)	55	1.8	1.9	0.2
1954	Lime (tons)	55	1.8	1.9	0.2
1959	Lime (acres)	105	3.4	4.7	2.4
1959	Lime (tons)	105	3.4	4.7	2.4
1964	Lime (acres)	99	3.2	4.6	2.3
1964	Lime (tons)	99	3.2	4.6	2.3
1969	Lime (acres)	621	19.9	34.6	23.8
1969	Lime (tons)	542	17.4	32.1	18.8
1974	Lime (acres)	3,117	100.0	100.0	100.0
1974	Lime (tons)	542	17.4	32.1	18.8
1978	Lime (acres)	222	7.1	12.1	5.6
1978	Lime (tons)	145	4.7	5.2	4.1
1982	Lime (acres)	161	5.2	5.1	6.4
1982	Lime (tons)	174	5.6	6.0	6.7
1987	Lime (acres)	180	5.8	5.4	6.9
1987	Lime (tons)	186	6.0	5.7	7.1

Fig. S1

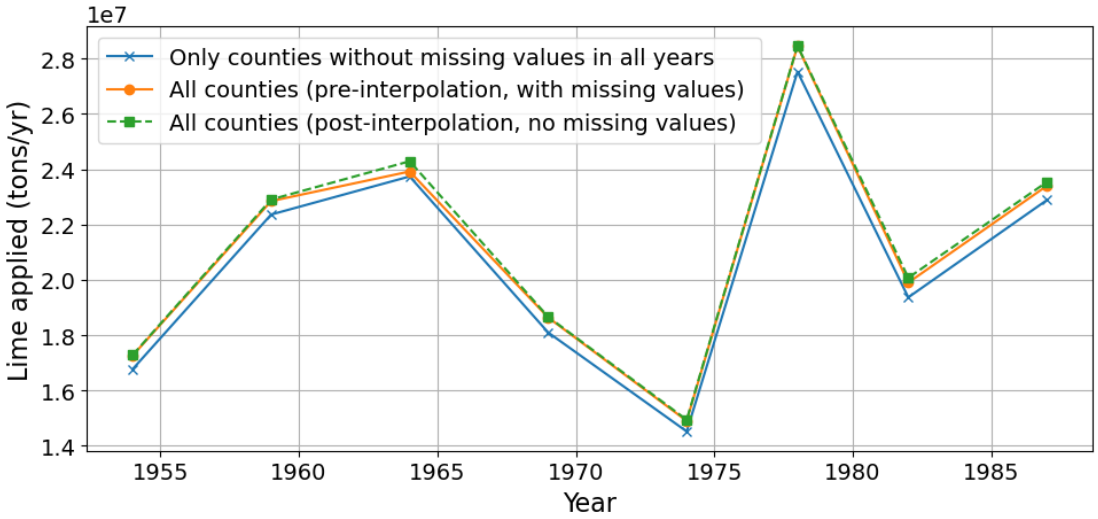

Fig. S1 Comparison of total agricultural lime applied in the U.S. before and after temporal interpolation and data filling. The blue line represents the sum of the counties that have valid data in all years, while the orange line shows the sum of all counties data, before interpolation and data filling. The green dashed line represents the sum of all counties data, after interpolation and data filling. All three lines demonstrate similar trends, with minor differences indicating the effect of missing data, data interpolation and filling.

Fig. S2

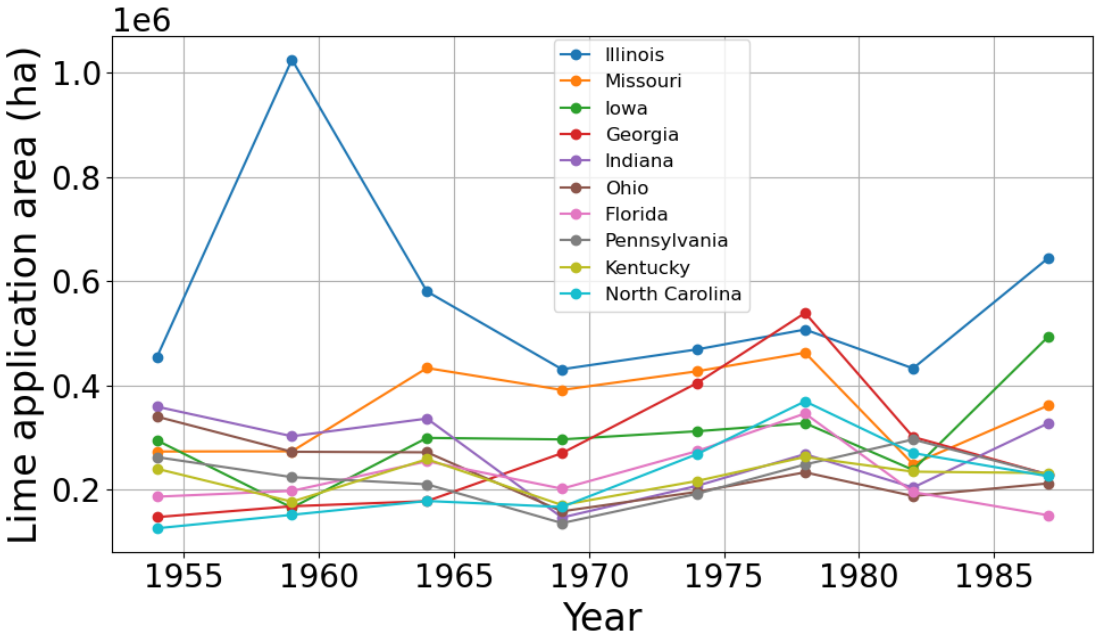

Fig. S2 Temporal variation in the total limed area (hectares, ha) for the top 10 states in the continental U.S. from 1954 to 1987. The states are ranked in descending order based on total limed area across the period, with Illinois having the largest area, followed by Missouri, Iowa, Georgia, Indiana, Ohio, Florida, Pennsylvania, Kentucky, and Mississippi.

Fig. S3

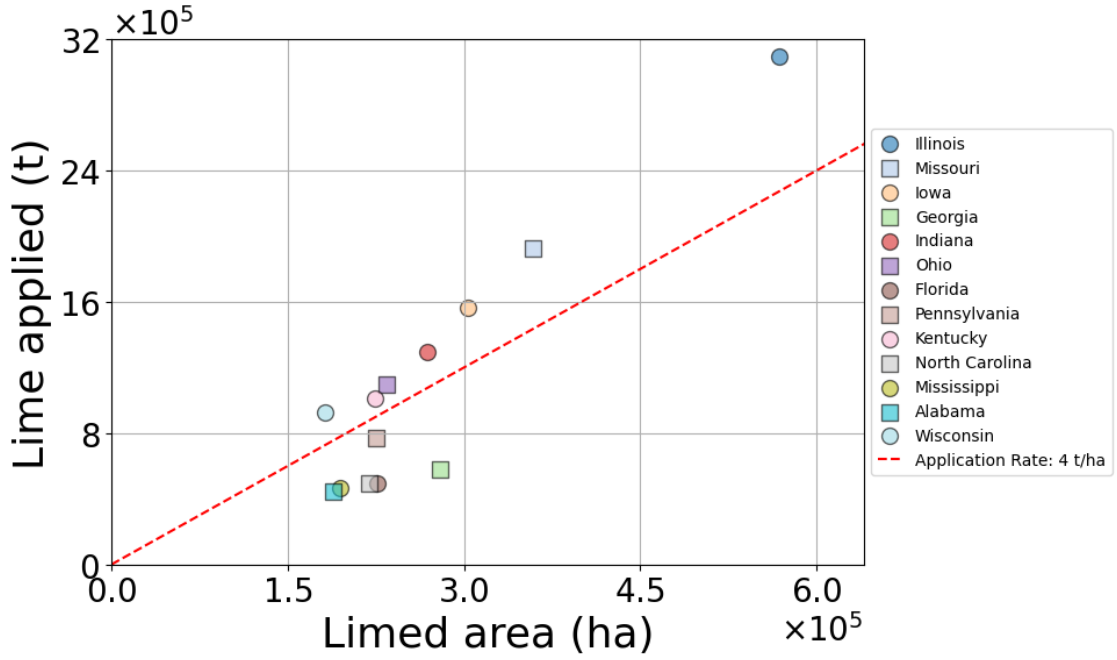

Fig. S3

Total limed weight (metric tons, t) versus total limed area (hectares, ha) for the 13 states with the highest average limed area. The values represent the averages from 1954 to 1987. The red dashed line indicates a lime application rate of 4 metric tons per hectare.

Fig. S4

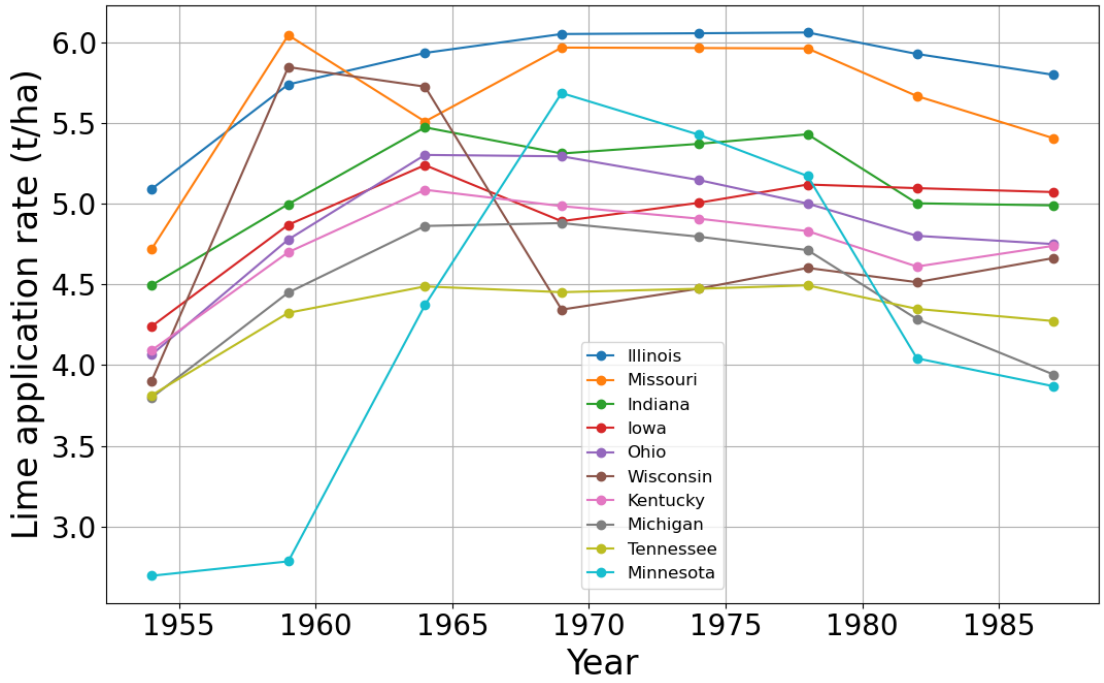

Fig. S4 Temporal variation in lime application rate (metric tons/hectare) for the top 10 states in the continental U.S. from 1954 to 1987. The states are ranked in descending order based on average lime application rate across the period, with Illinois having the largest area, followed by Missouri, Indiana, Iowa, Wisconsin, Kentucky, Michigan, Tennessee, Minnesota.

Random Forest Analysis

2.1 Feature selection

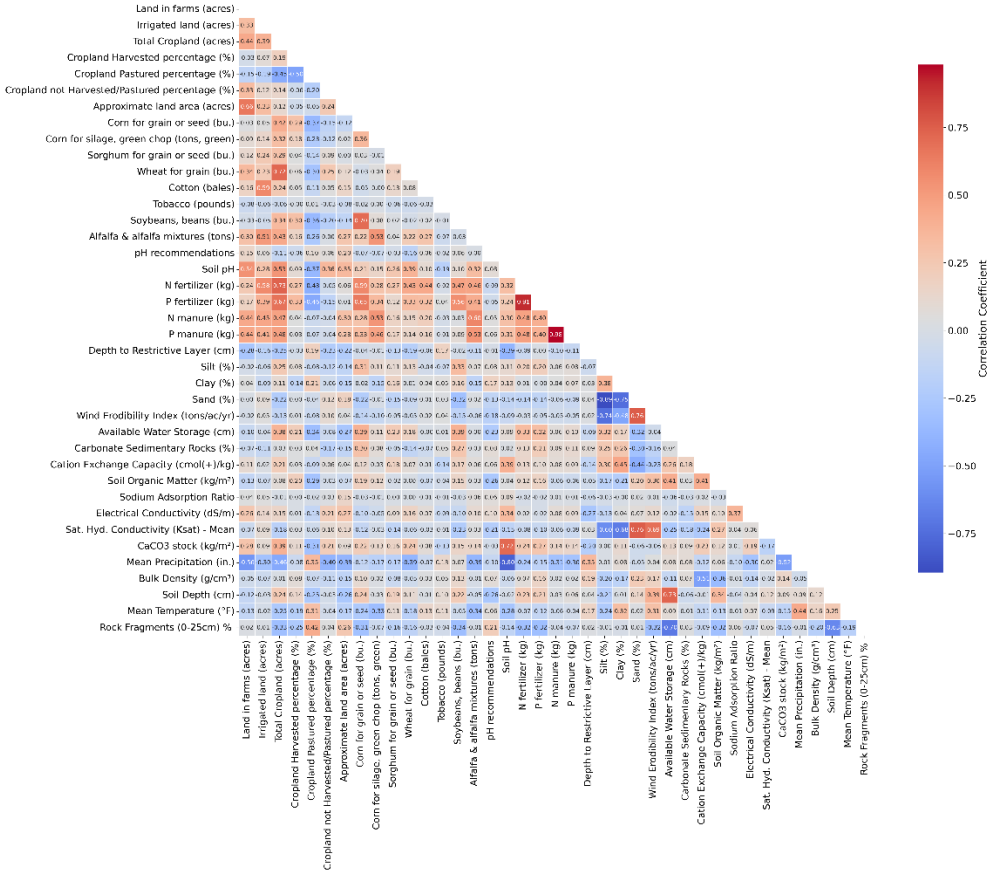

Fig. S5 Correlation matrix of predictors and Lime applied (tons), showing Pearson correlation coefficients. Each cell represents the correlation between two variables, with values closer to 1 (red) indicating a strong positive linear relationship, values closer to -1 (blue) indicating a strong negative linear relationship, and values near 0 (white) indicating no linear correlation. The heatmap highlights significant correlations, such as between N manure and P manure (0.96), and between N fertilizer and P fertilizer (0.91), which suggest potential collinearity among specific predictors.

Table S3 Variance Inflation Factor (VIF) Analysis of final list of predictor variables after removing collinear variables

Variable	VIF
Total Cropland (acres)	7.98
Soil pH	7.22
N fertilizer (kg)	7.21
Sand (%)	7.12
Wind Erodibility Index (tons/ac/yr)	5.59
Clay (%)	4.89
Corn for grain or seed (bu.)	4.85
Mean Precipitation (in.)	4.84
Wheat for grain (bu.)	4.69
Sat. Hyd. Conductivity (Ksat) - Mean	3.90
Mean Temperature (°F)	3.52
Soybeans, beans (bu.)	3.33
Cation Exchange Capacity (cmol(+)/kg)	3.23
Irrigated land (acres)	3.14
N manure (kg)	3.05
Soil Organic Matter (kg/m²)	3.03
Land in farms (acres)	3.01
Rock Fragments (0–25 cm) (%)	2.96

Variable	VIF
$CaCO_3$ stock (kg/m^2)	2.88
Approximate land area (acres)	2.56
Bulk Density (g/cm^3)	2.52
Alfalfa & alfalfa mixtures (tons)	2.51
Cropland Pastured percentage (%)	2.47
Soil Depth (cm)	2.40
Corn for silage, green chop (tons, green)	2.13
Cotton (bales)	2.07
Electrical Conductivity (dS/m)	1.58
Cropland not Harvested/Pastured percentage (%)	1.57
Carbonate Sedimentary Rocks (%)	1.46
Cropland Harvested percentage (%)	1.44
Sorghum for grain or seed (bu.)	1.39
pH recommendations	1.38
Depth to Restrictive Layer (cm)	1.37
Sodium Adsorption Ratio	1.21
Tobacco (pounds)	1.13

2.2

Random Forest modeling without pH recommendations, soil variables

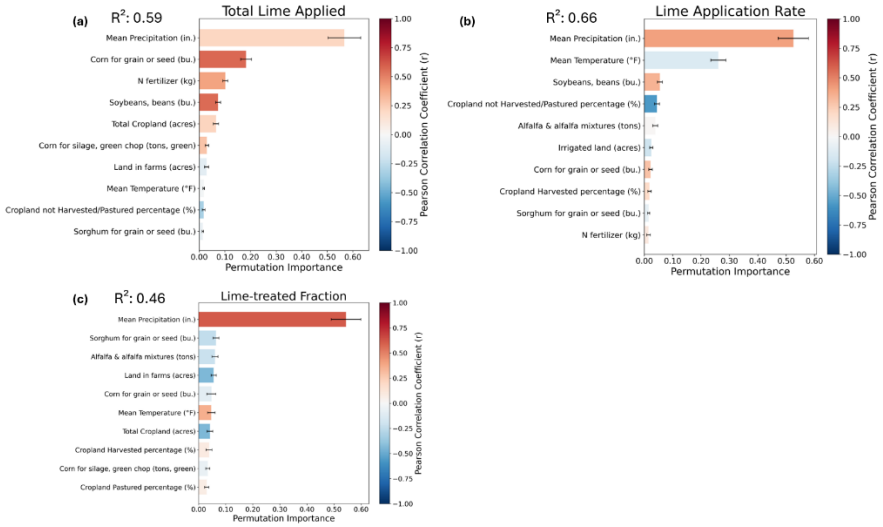

Fig. S6 Statistical and machine learning analysis of average annual (a) lime applied per county per year, normalized by county land area (t ha^{-1} yr^{-1}), (b) lime application rate (t ha^{-1} yr^{-1}), and (c) lime-treated fraction (yr^{-1}) per county across the three census years from 1978 to 1987. For (b) and (c) we ran the analysis only on the counties that have greater than 10% cropland by area, to minimize the influence of counties that have minimal agriculture. Here compared to Fig. 9, we removed in the analysis the soil variables and pH recommendation data that does not exactly match the temporal period of our study. The bar length represents the permutation importance of all variables in the random forest model, with the error bar showing the standard error for the permutation importance across the 5-folds of cross validation training. Only the top ten variables with the highest permutation importance are shown here. The colors represent the Pearson correlation coefficients with the target variable.

Random Forest modelling with pH recommendations, soil variables

- Lime applied per year

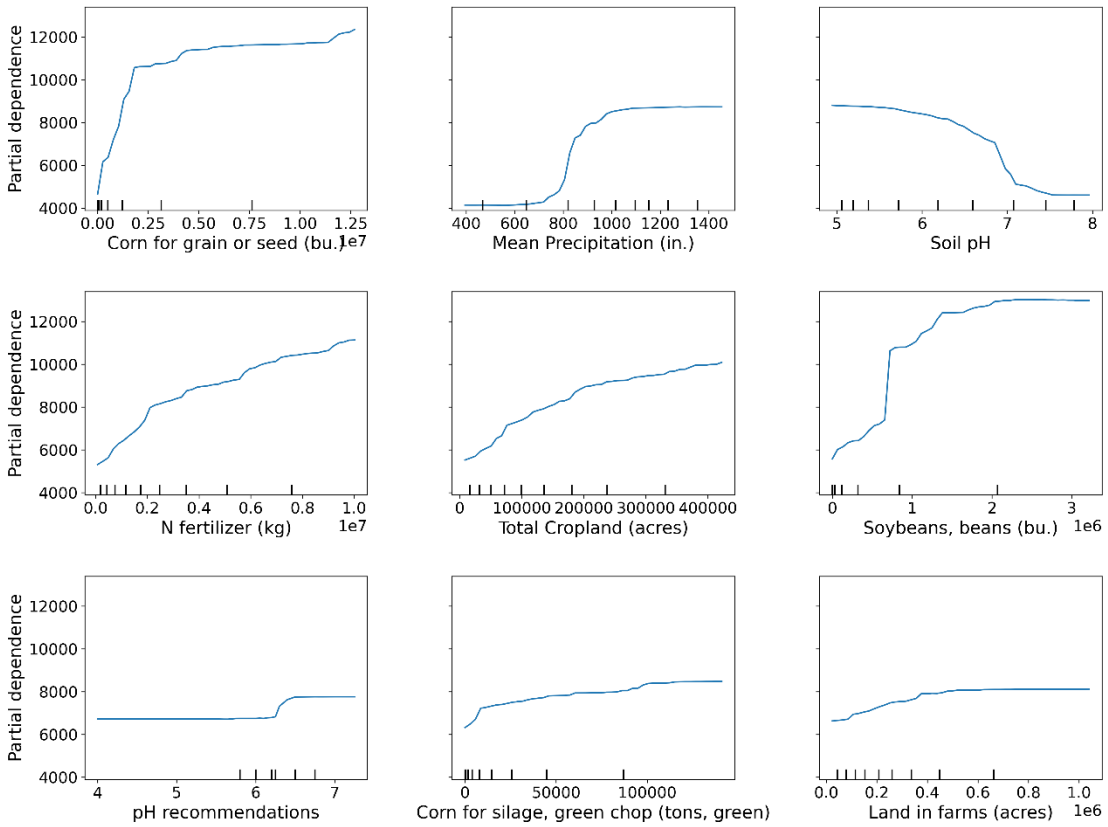

Fig. S7 The partial dependence plots of the variables that have the higher feature importances with annual lime application (metric tons per year) in each county, showing the effect of each variable on the prediction results.

-Lime application rate

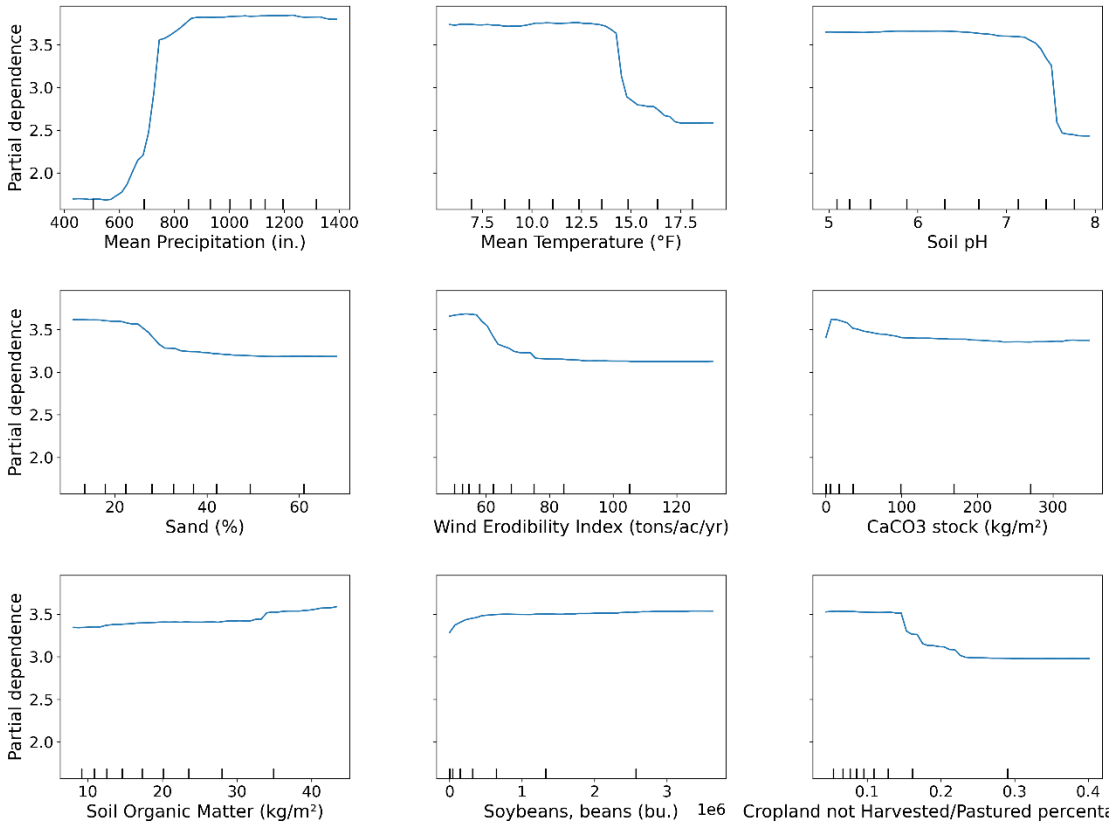

Fig. S8 The partial dependence plots of the variables that have the higher feature importances with lime application rate (limed applied per limed area; t/ha) in each county, showing the effect of each variable on the prediction results.

-Lime-treated fraction (limed area/cropland area)

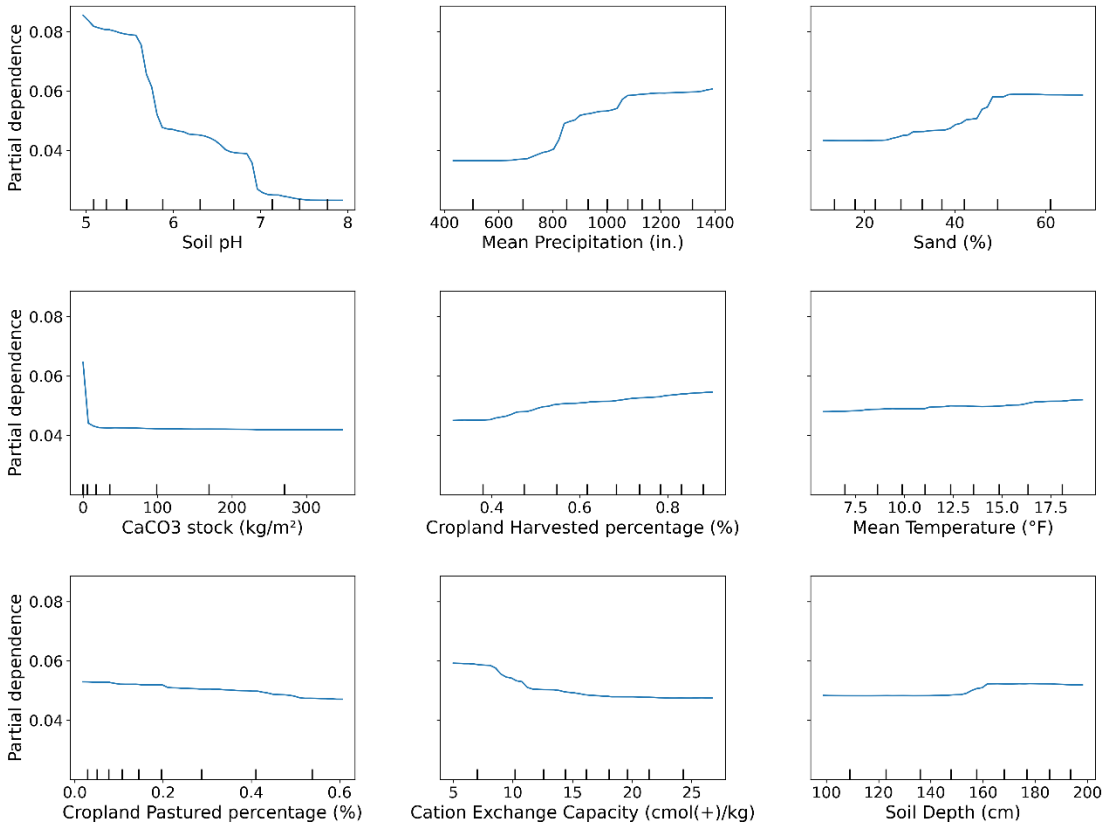

Fig. S9 The partial dependence plots of the variables that have the higher feature importances with lime-treated fraction (limed area/cropland area) in each county, showing the effect of each variable on the prediction results.

Sensitivity Analysis of different cropland thresholds.

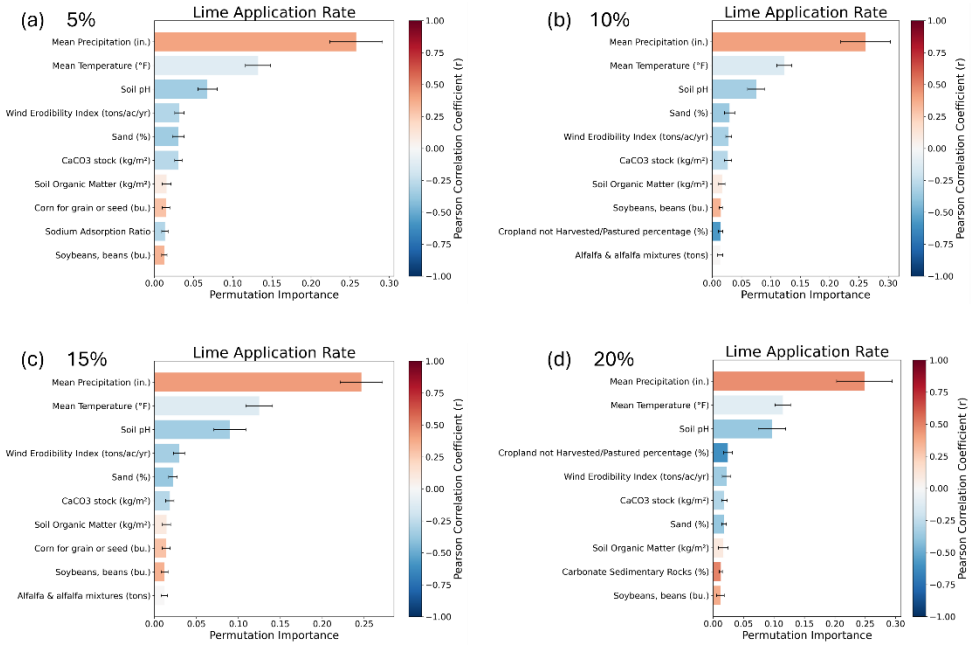

Fig S10 Statistical and machine learning analysis of average annual lime application rate (t ha^{-1} yr^{-1}) per county across the three census years from 1978 to 1987. We ran the analysis with counties that have at least (a) 5%, (b) 10%, (c) 15%, (d) 20% cropland by area, to minimize the influence of counties that have minimal agriculture. The bar length represents the permutation importance of all variables in the random forest model, with the error bar showing the standard error for the permutation importance across the 5-folds of cross validation training. Only the top ten variables with the highest permutation importance are shown here. The colors represent the Pearson correlation coefficients with the target variable.

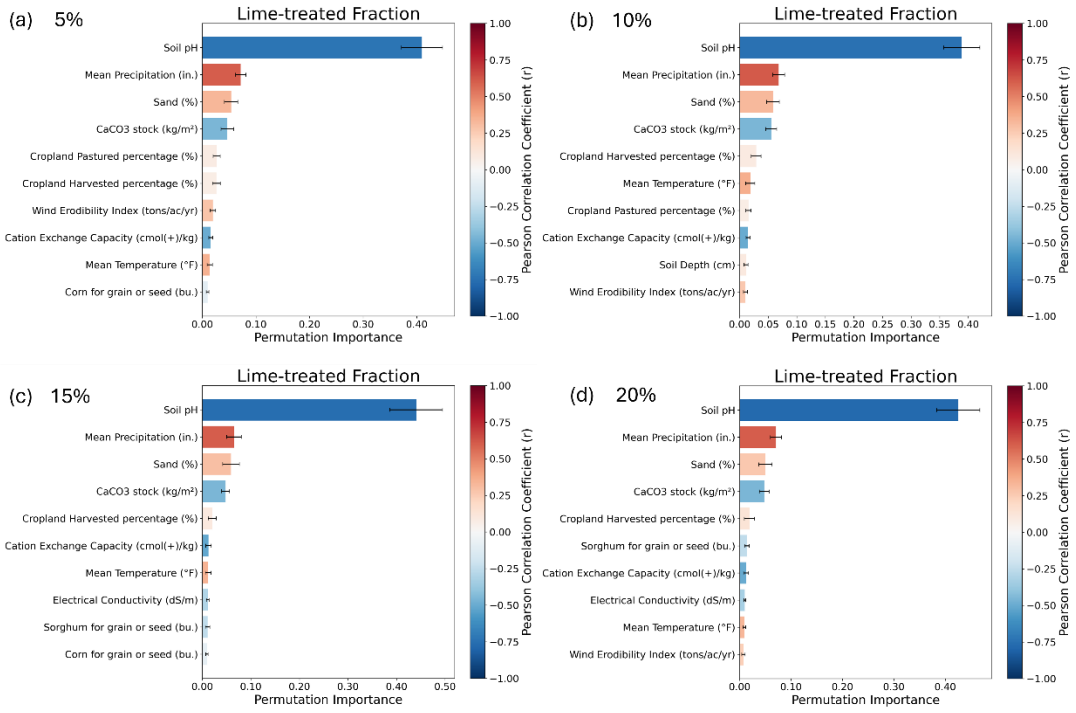

Fig S11 Statistical and machine learning analysis of average annual lime-treated fraction (yr^{-1}) per county across the three census years from 1978 to 1987. We ran the analysis with counties that have at least (a) 5%, (b) 10%, (c) 15%, (d) 20% cropland by area, to minimize the influence of counties that have minimal agriculture. The bar length represents the permutation importance of all variables in the random forest model, with the error bar showing the standard error for the permutation importance across the 5-folds of cross validation training. Only the top ten variables with the highest permutation importance are shown here. The colors represent the Pearson correlation coefficients with the target variable.

2. Spatial Pattern of agricultural and environmental parameters.

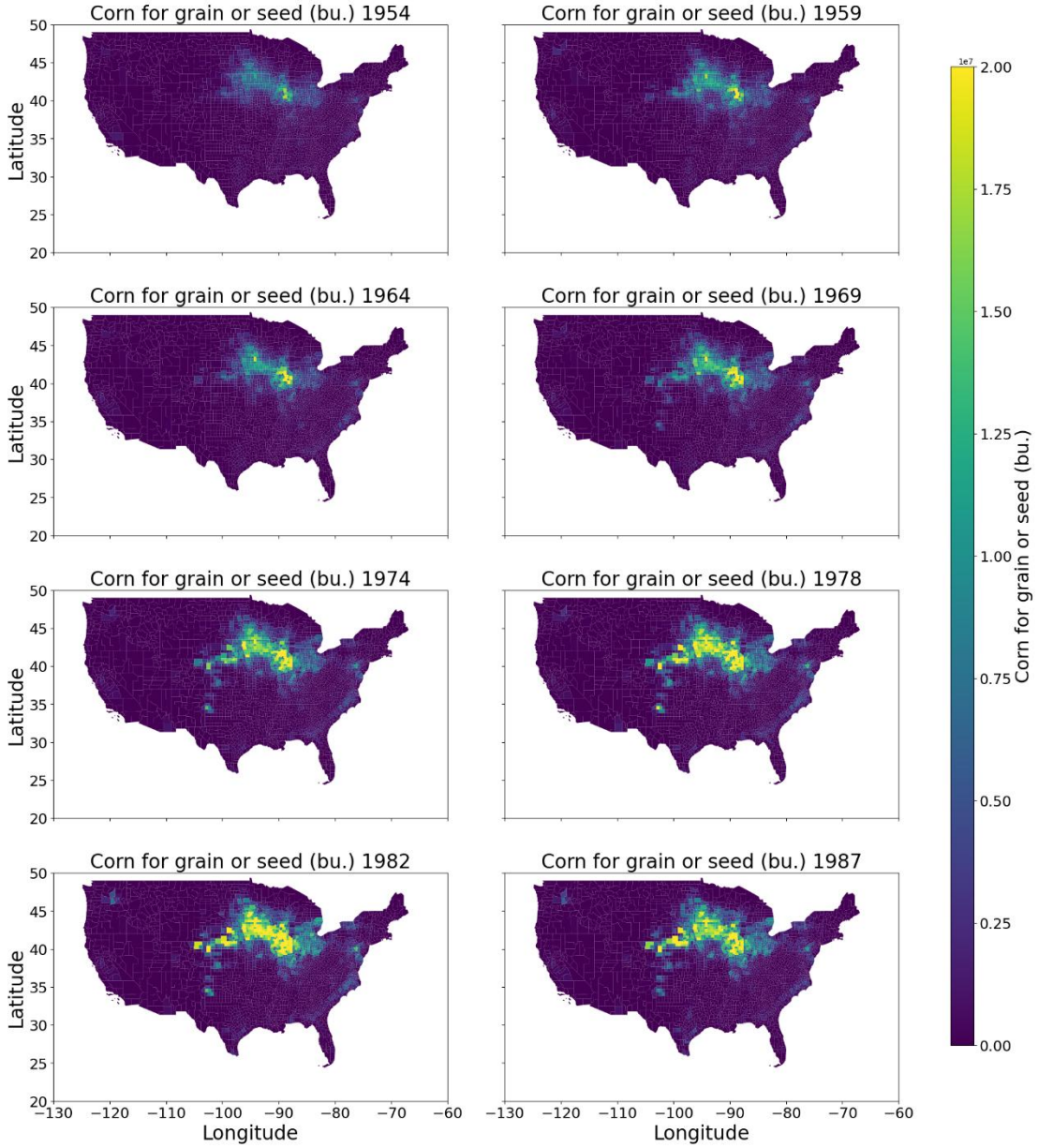

Fig. S12 Corn production for grain or seed (bu.) by county in the continental United States for eight census years from 1954 to 1987, as reported by the Census Bureau. The maps display the spatial distribution of corn production. Missing data is not shown and may be addressed through interpolation or spatial filling methods as described in the Methods section.

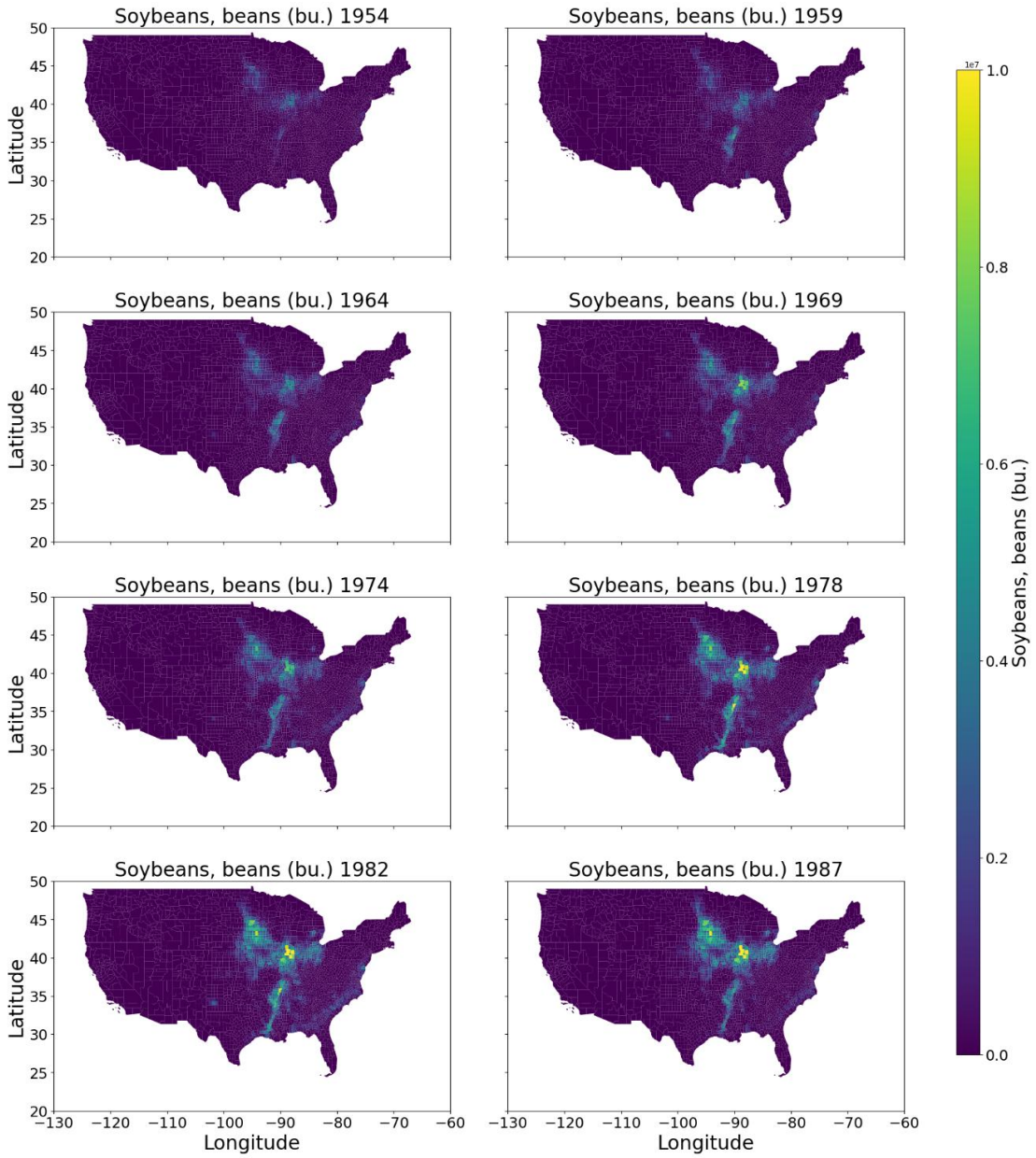

Fig. S13 Soybean production (bu.) by county in the continental United States for eight census years from 1954 to 1987, as reported by the Census Bureau. Missing data is not shown and may be addressed through interpolation or spatial filling methods, as described in the Methods section."

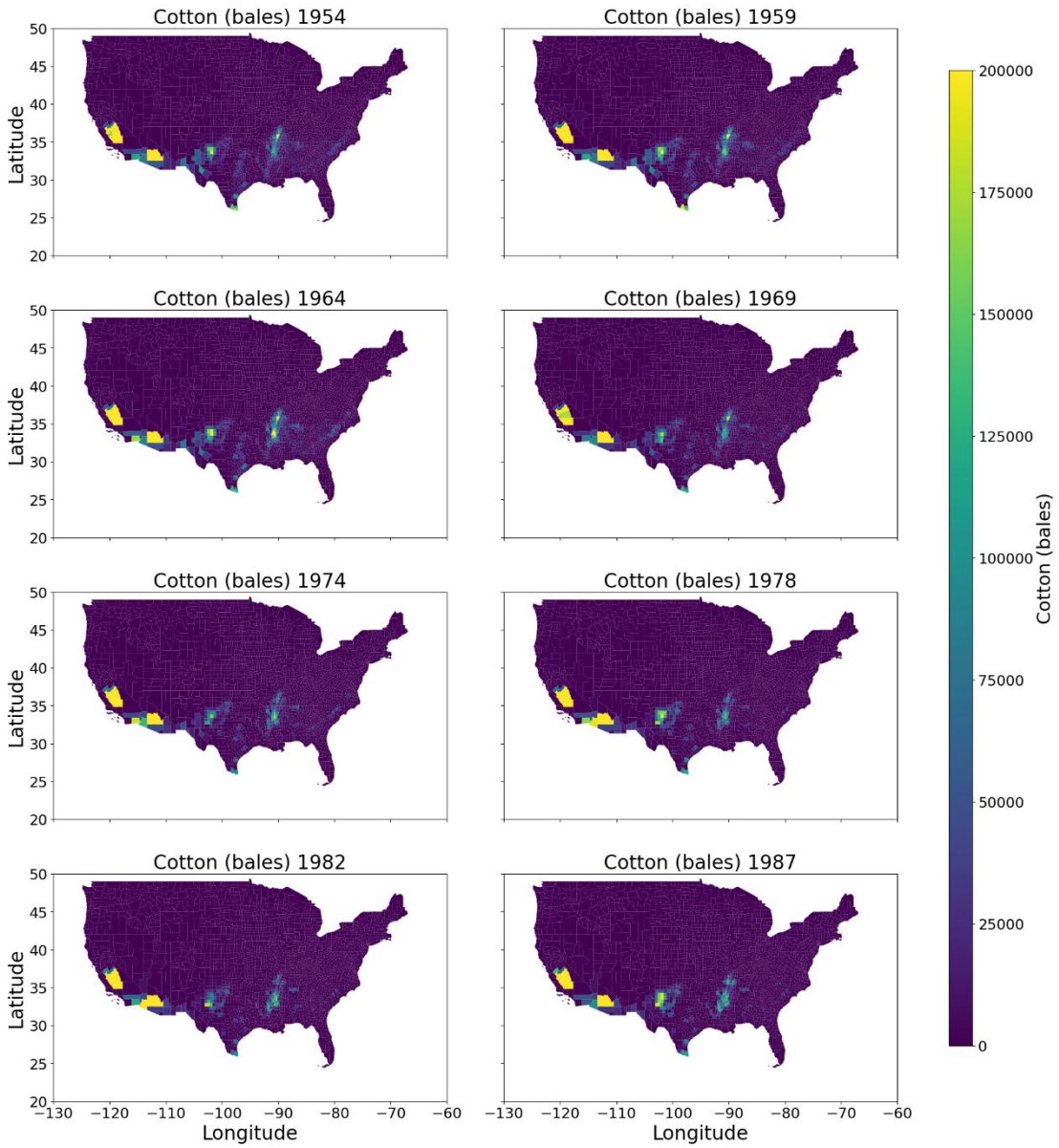

Fig. S14 Cotton production (bales) by county in the continental U.S. for eight census years from 1954 to 1987, as reported by the Census Bureau. The maps display the spatial variation of cotton production, with missing data temporarily interpolated and spatially filled, as described in the Methods section.

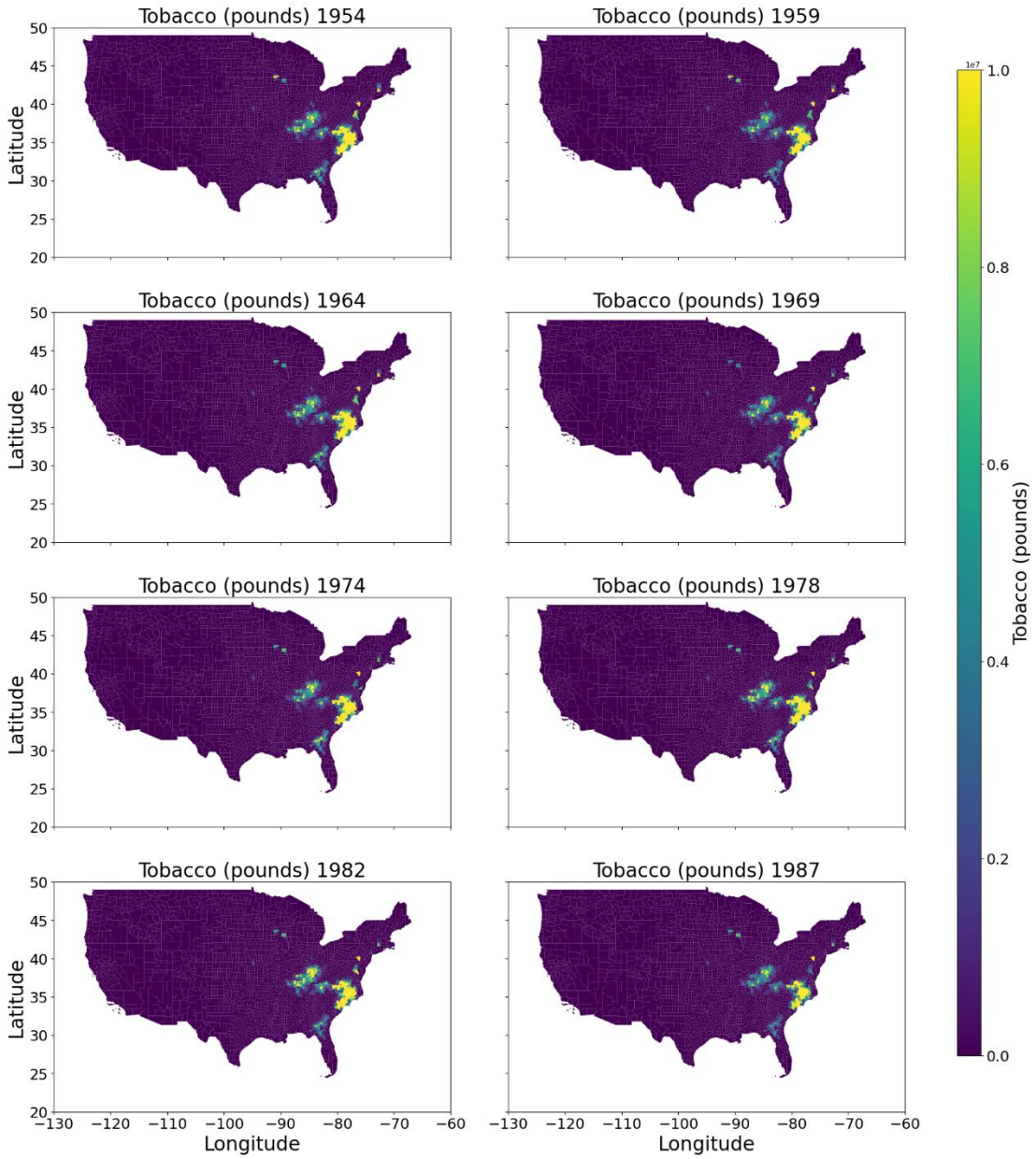

Fig. S15 Tobacco production (pounds) by county in the continental U.S. for eight census years from 1954 to 1987, as reported by the Census Bureau. The maps display the spatial variation of tobacco production across counties, with missing data temporarily interpolated and spatially filled, as described in the Methods section.

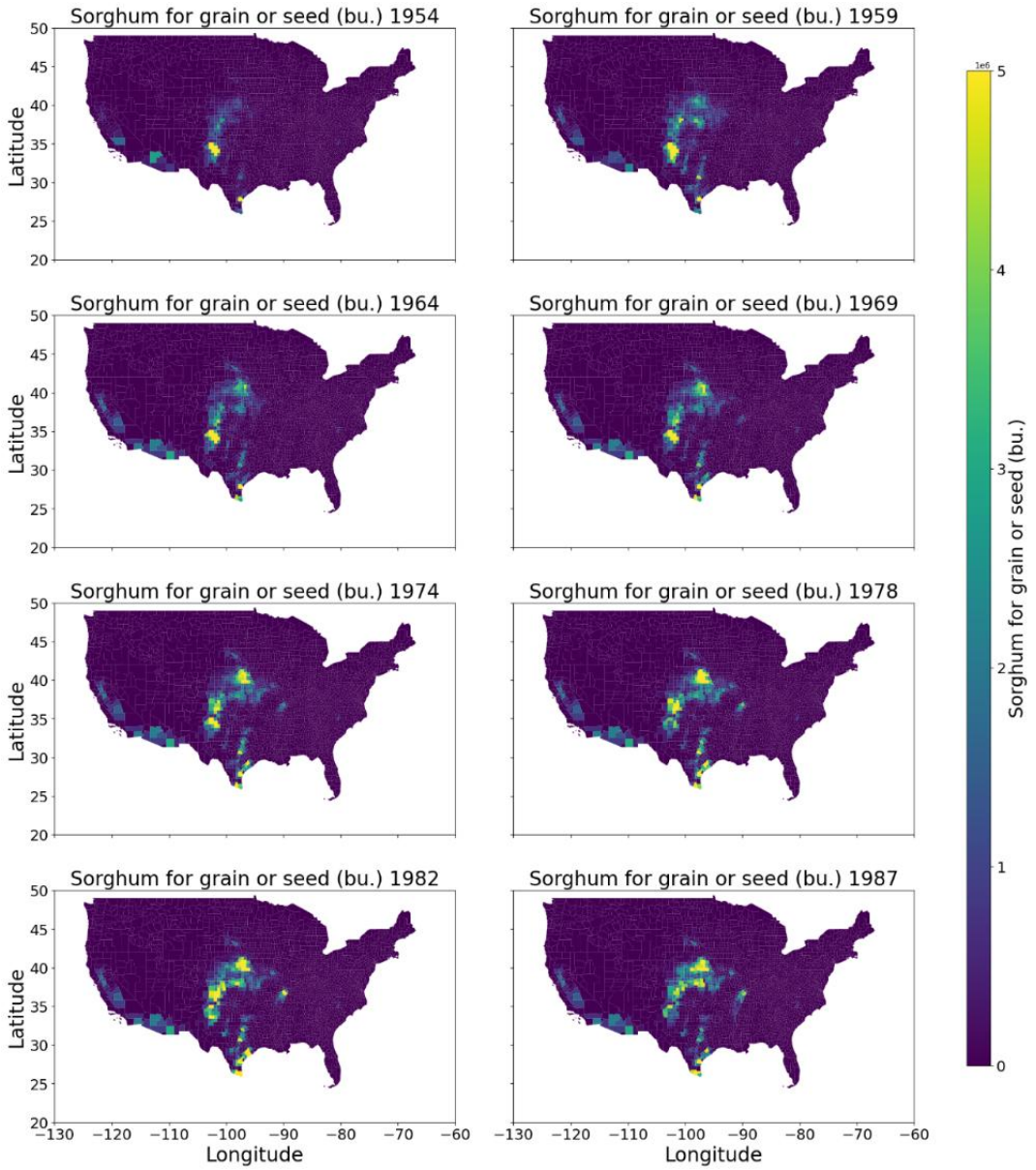

Fig. S16 Sorghum production for grain or seed (bu.) by county in the continental United States for eight census years from 1954 to 1987, as reported by the Census Bureau. Missing data is temporarily interpolated and spatially filled, as described in the Methods section.

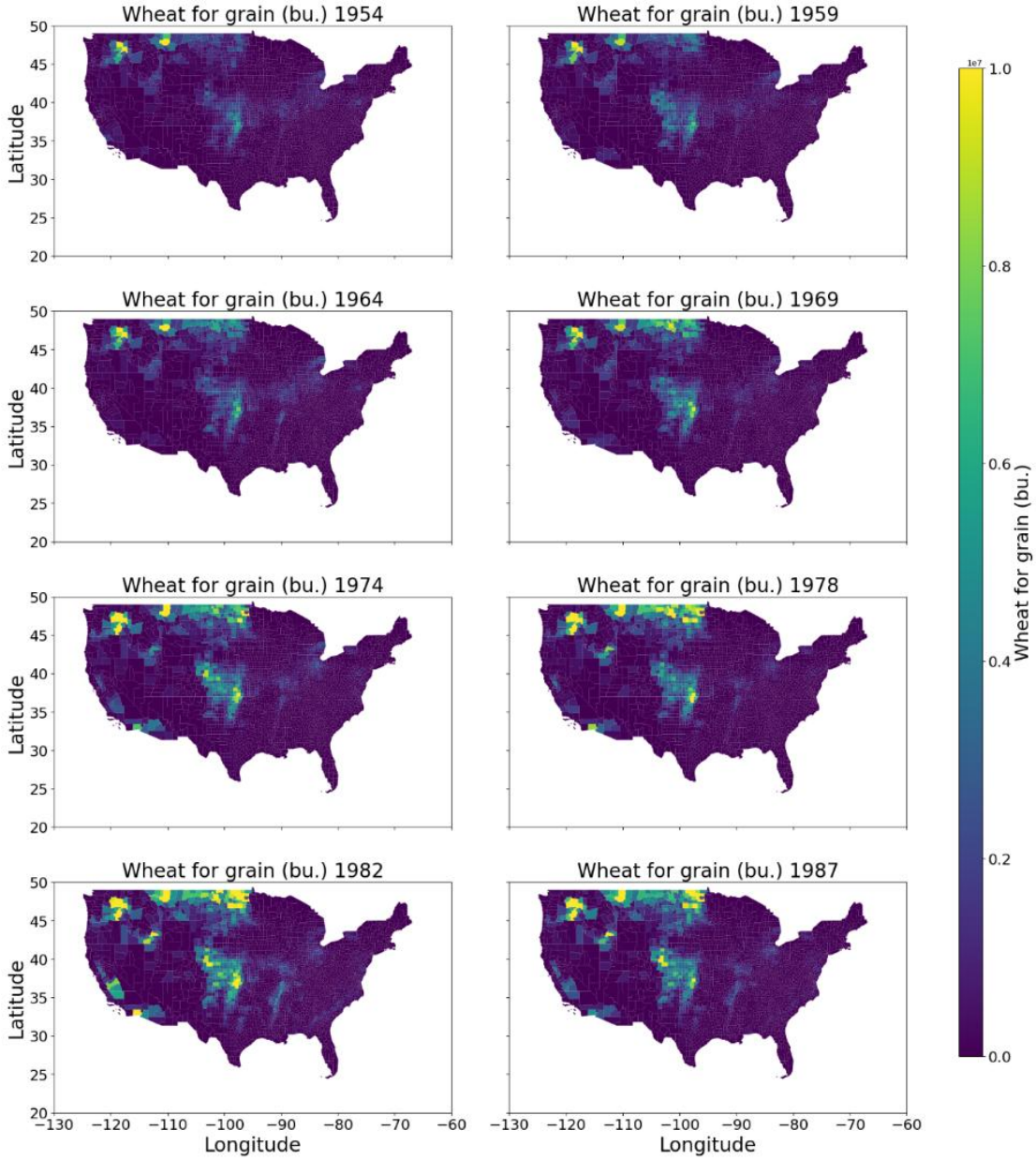

Fig. S17 Wheat production for grain (bu.) by county in the continental United States for eight census years from 1954 to 1987, as reported by the Census Bureau. The maps depict the spatial distribution of wheat production. Missing data is temporarily interpolated and spatially filled, as described in the Methods section.

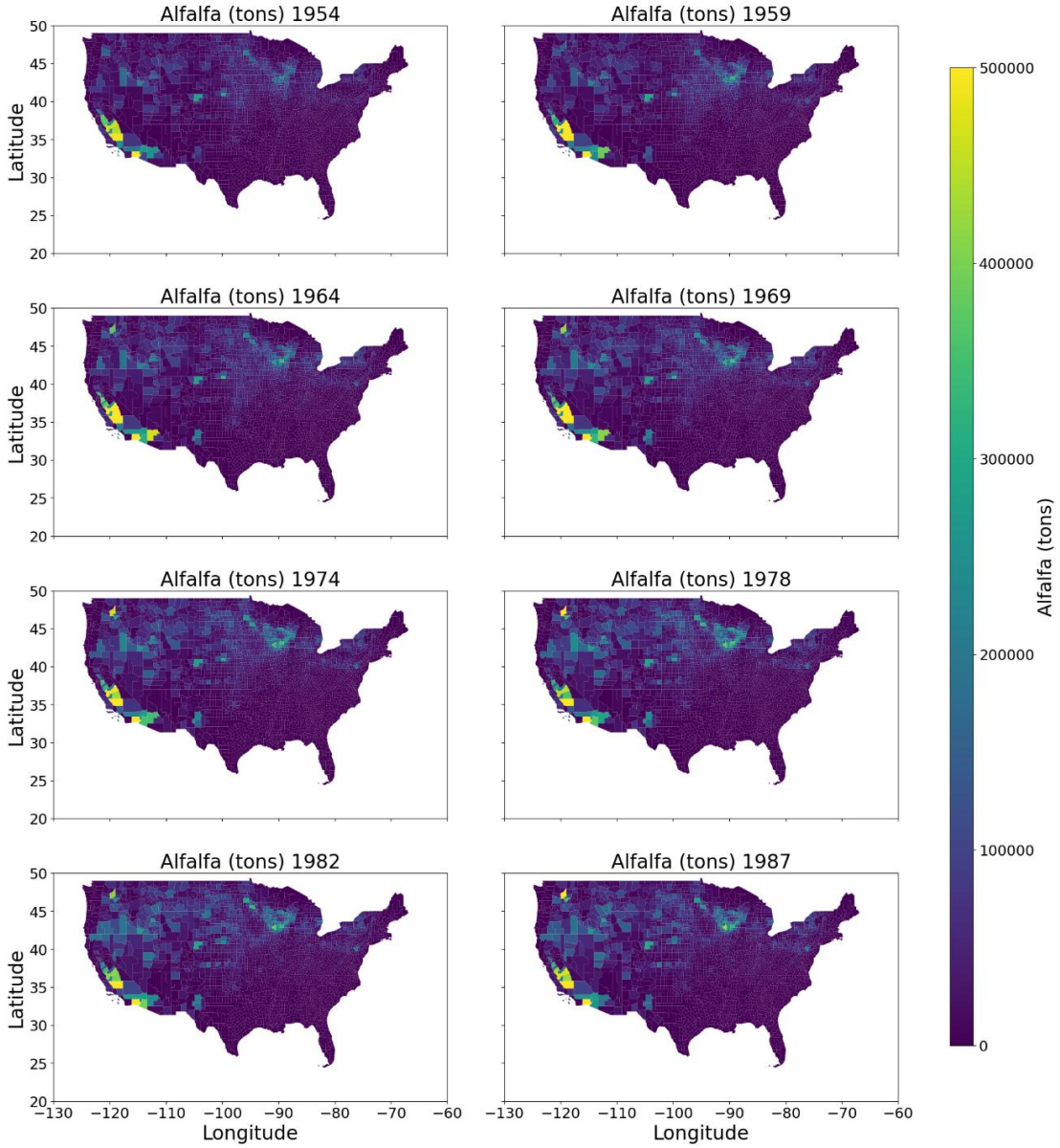

Fig. S18 Alfalfa production (tons) by county in the continental U.S. for eight census years from 1954 to 1987, as reported by the Census Bureau. The maps display the spatial variation of alfalfa production across counties, with missing data temporarily interpolated and spatially filled, as described in the Methods section.

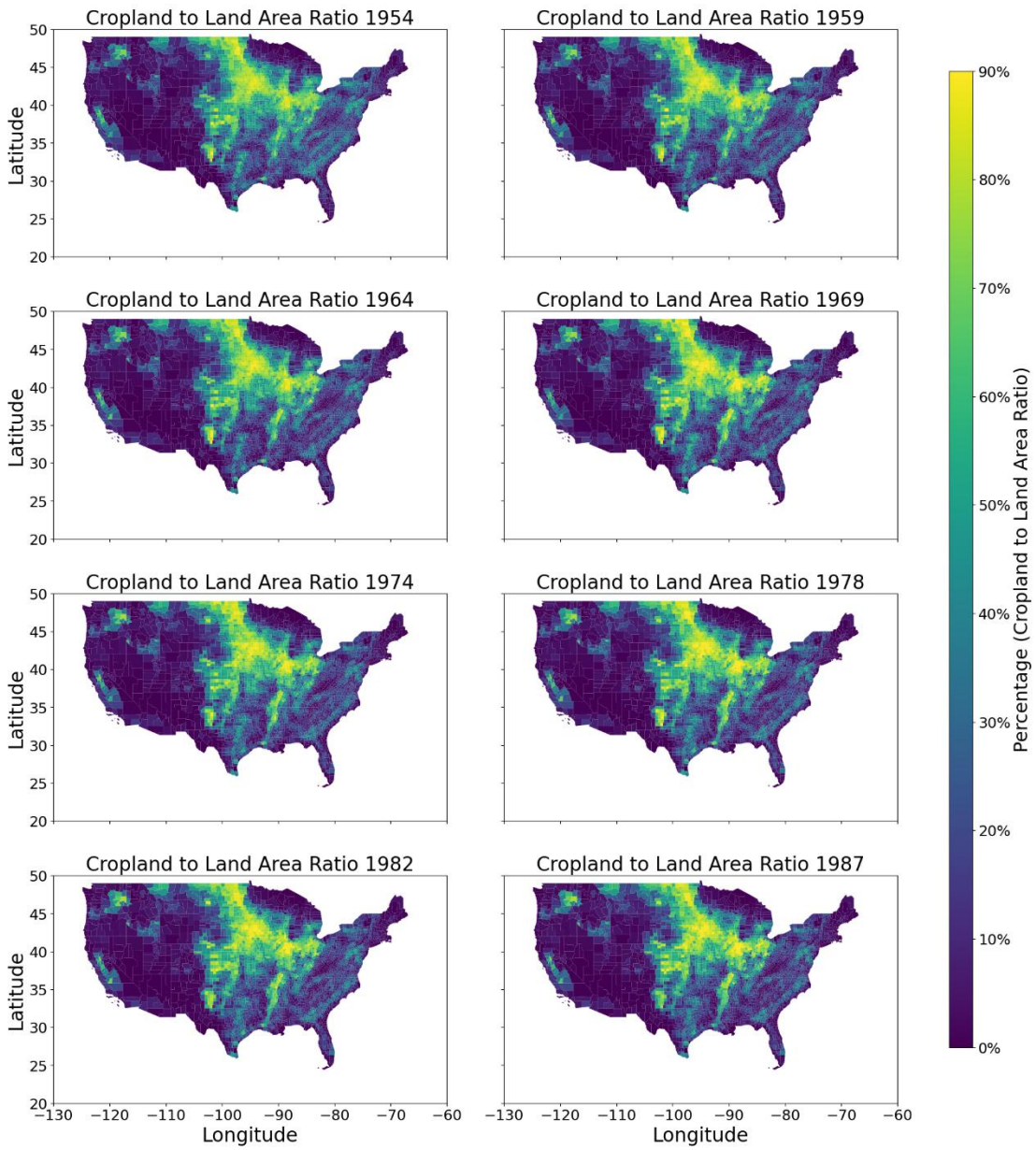

Fig. S19 Ratio of total cropland divided by the approximate land area per county for eight census years from 1954 to 1987, as reported by the Census Bureau. The approximate land area is calculated based on the shapefile of each county.

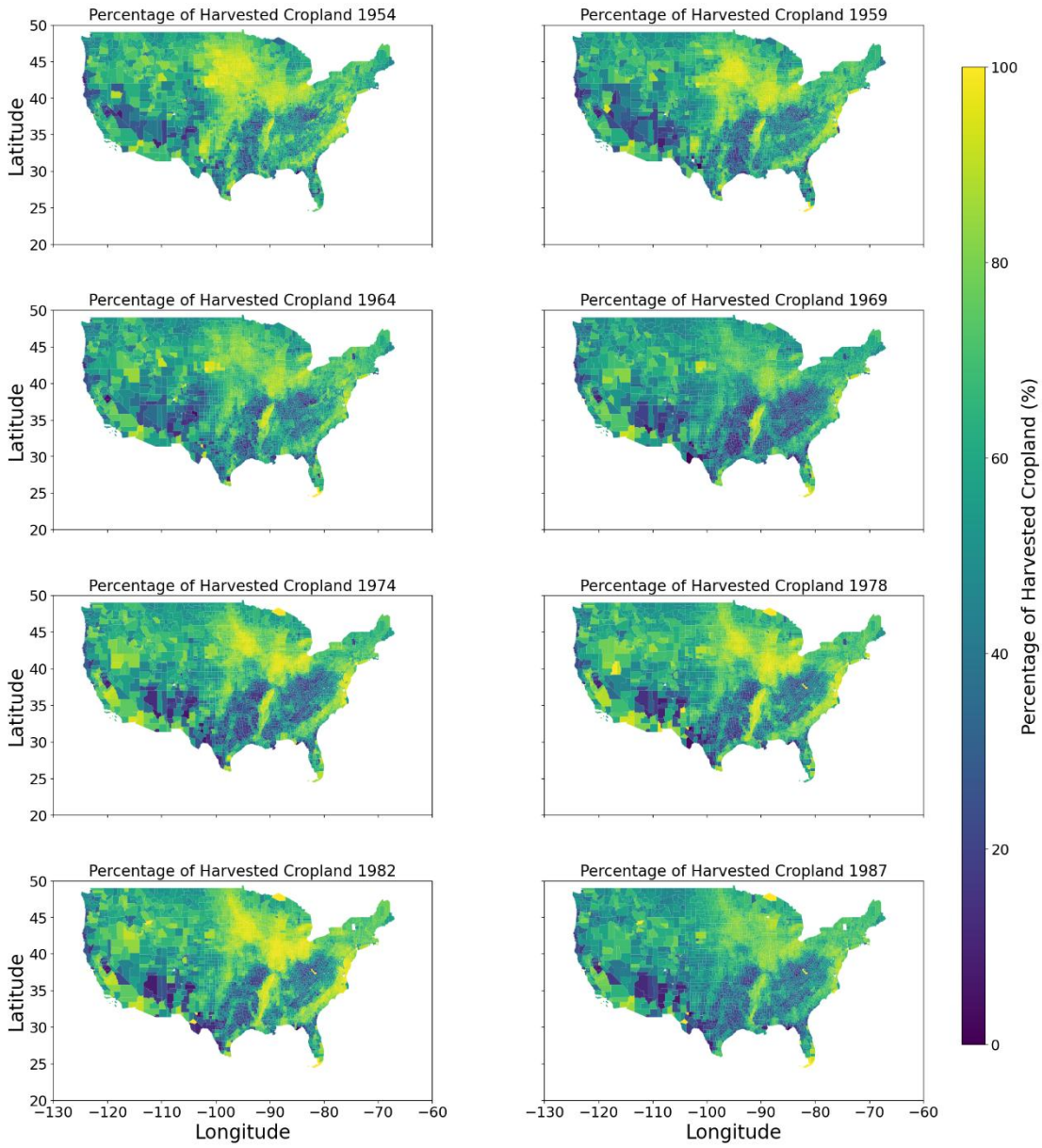

Fig. S20 Percentage of harvested cropland among the total cropland per county for eight census years from 1954 to 1987, as reported by the Census Bureau.

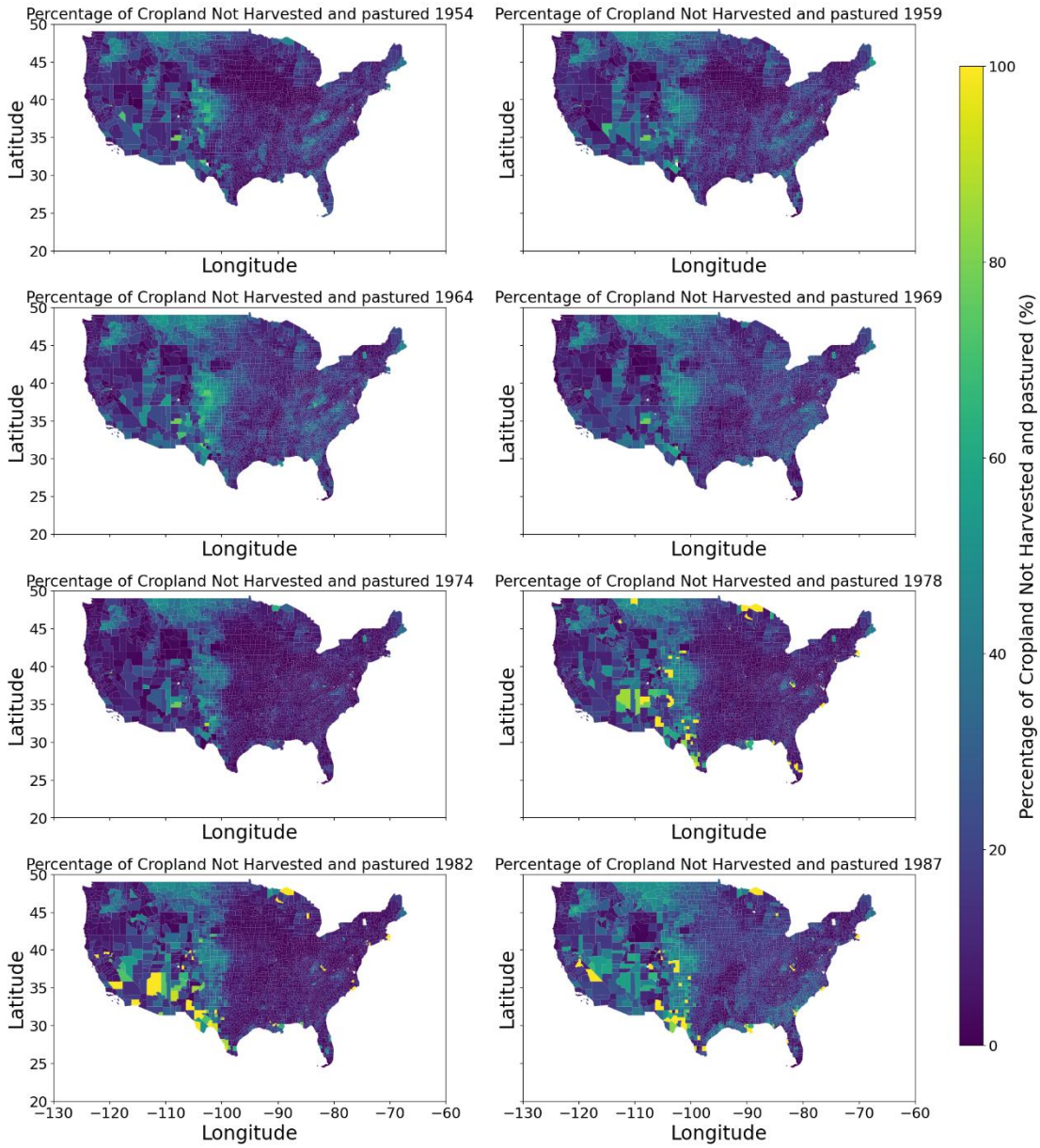

Fig. S21 Percentage of total cropland that is not harvested and not pastured for the eight census years from 1954 to 1987, as reported by the Census Bureau.

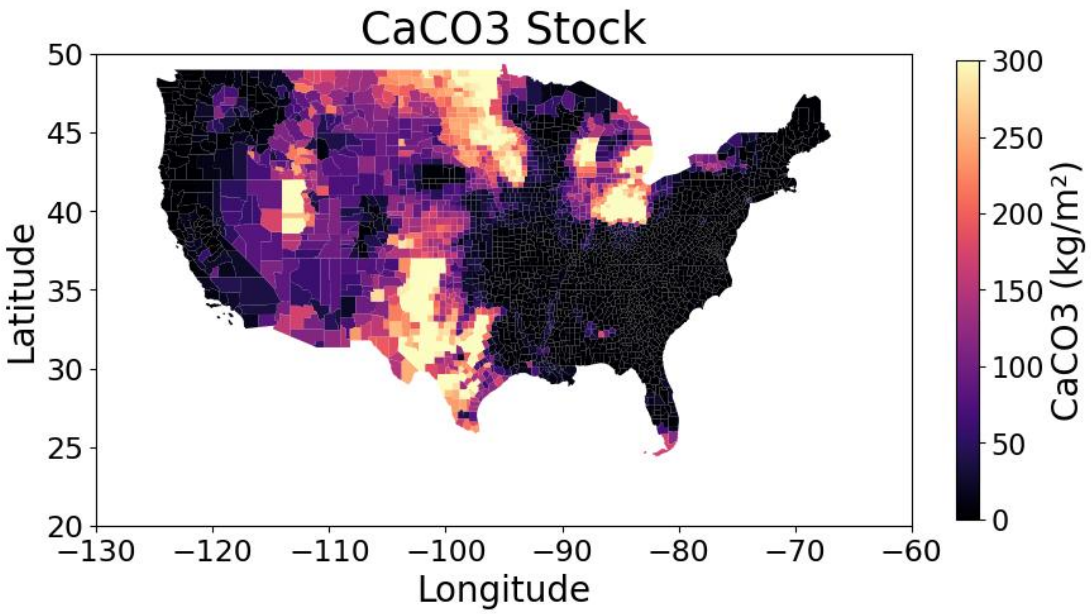

Fig. S22 Spatial distribution of $CaCO_3$ stock (kg/m^2) across counties in the continental United States. The map illustrates areas with varying levels of $CaCO_3$ stock, with higher concentrations predominantly located in the central and western regions.

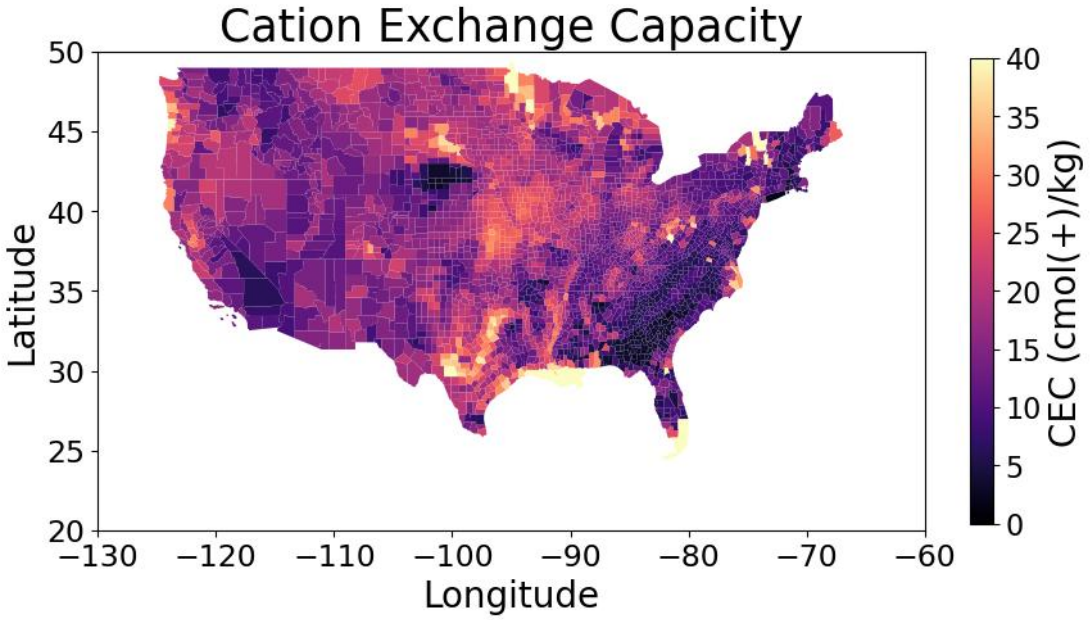

Fig. S23 Spatial distribution of Cation Exchange Capacity (CEC) (cmol(+)/kg) across counties in the continental United States. The map highlights regional variations in soil's ability to retain and exchange cations,

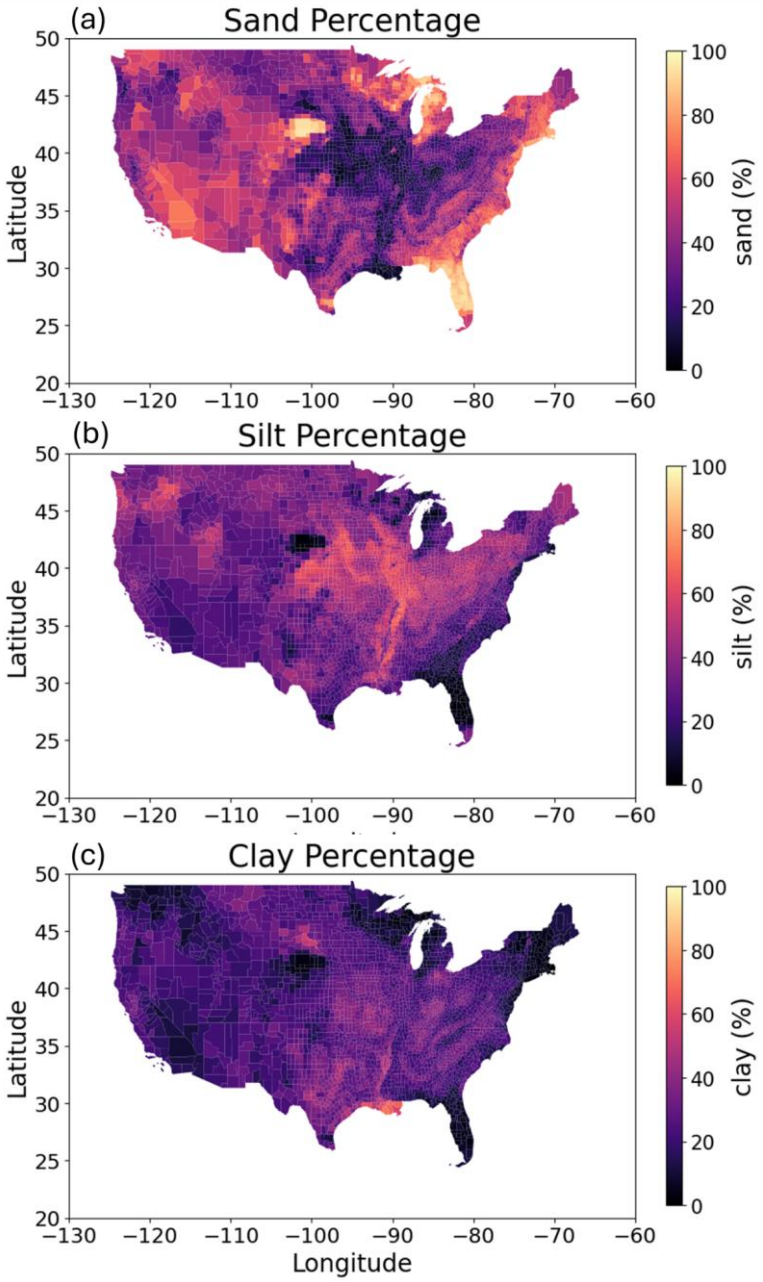

Fig. S24 Spatial distribution of percentage of (a) sand, (b) silt, (c) clay across counties in the continental United States. The map highlights regional variations in soil particle sizes.

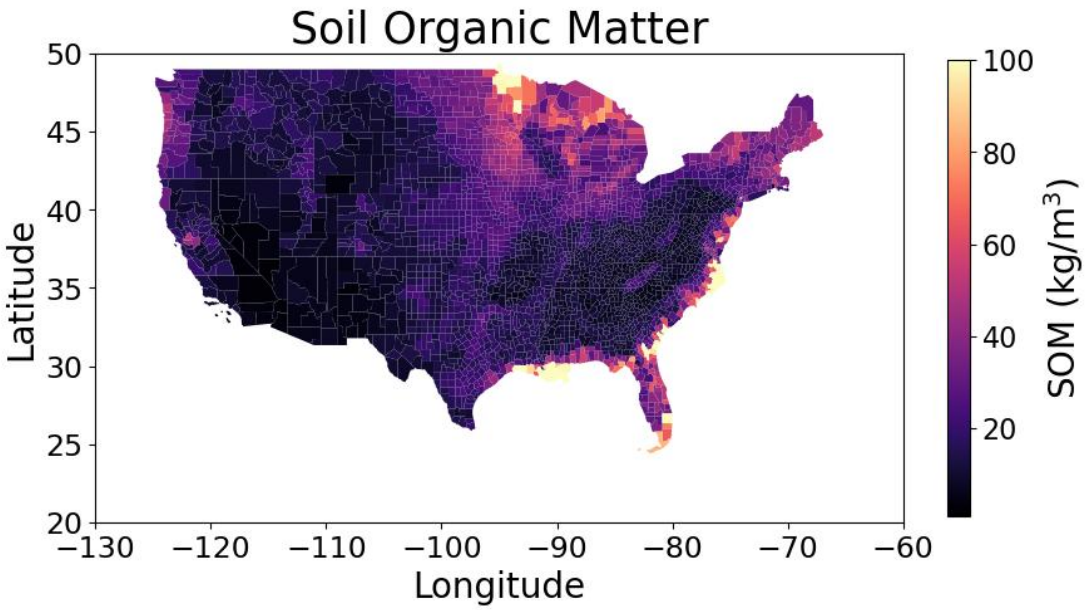

Fig. S25 Spatial distribution of soil organic matter (kg/m^3) across counties in the continental United States.

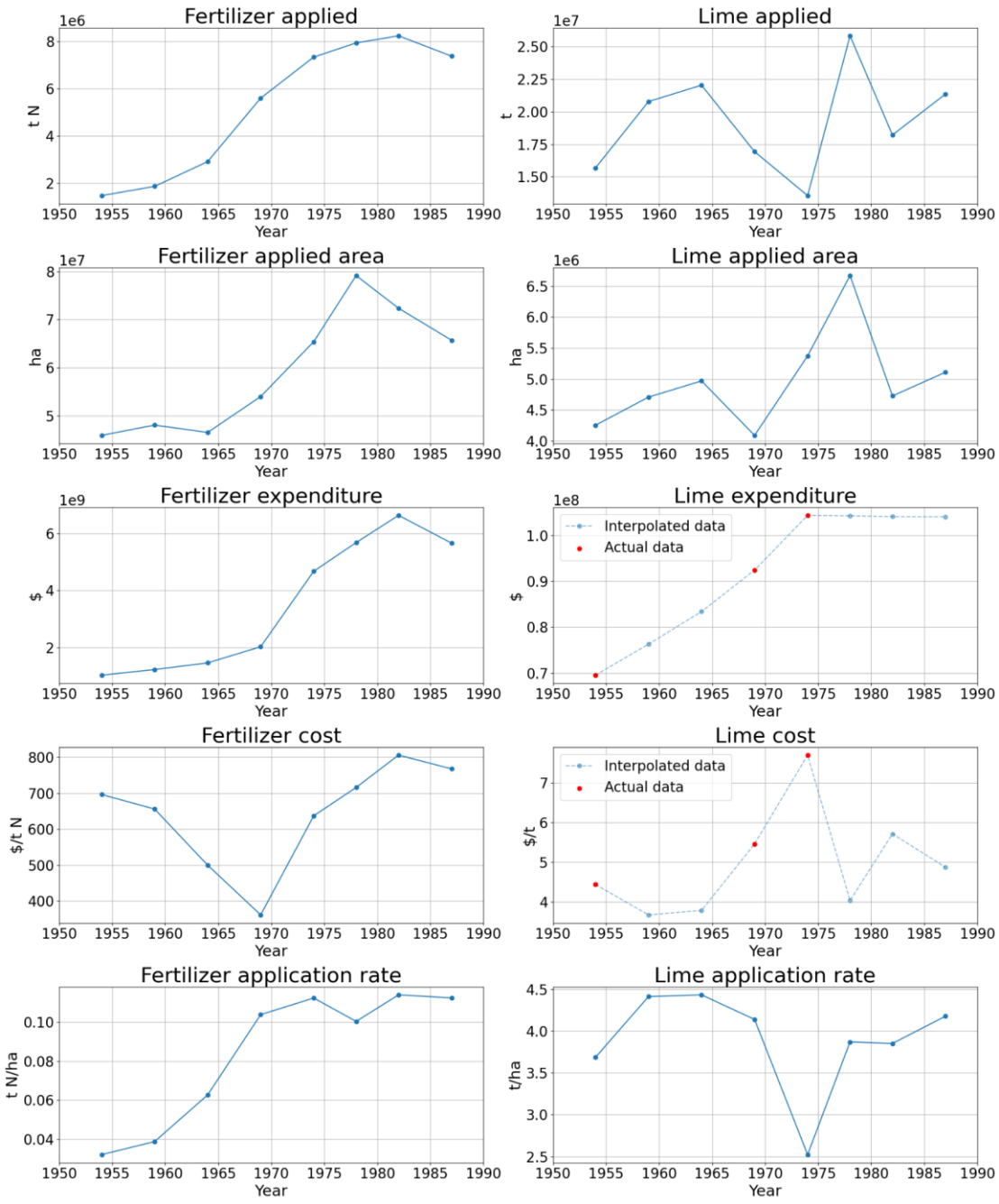

Fig. S26 Trends in total fertilizer and lime use, cost, area, and application rates in the United States from 1950 to 1990. Fertilizer applied (weight of N) is from the county-level data of Falcone (2021), while all the other variables are from the Census reports. Cost for fertilizer and lime is calculated by dividing the total expenditure to the amount of application. Application rate is calculated from dividing the total weight of application to the applied area. For lime expenditure and cost, red dots represent available actual data, and dotted blue lines represent interpolated estimates used to complete missing years.

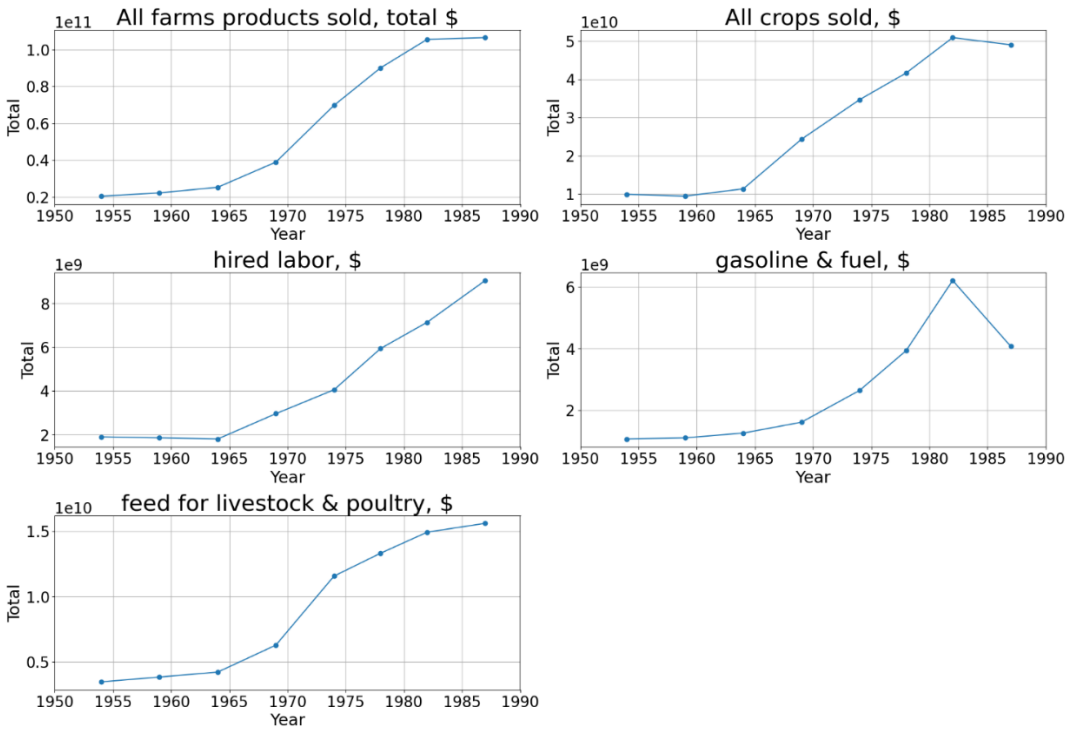

Fig. S27 Trends in total sales and input expenditures in U.S. agriculture from 1950 to 1990. This figure presents the inflation-agnostic trends (in nominal USD). While overall sales of agricultural products and crops steadily increased across the decades, input costs—especially labor, fuel, and feed—also rose, reflecting growing capital and operational intensities in U.S. farming.

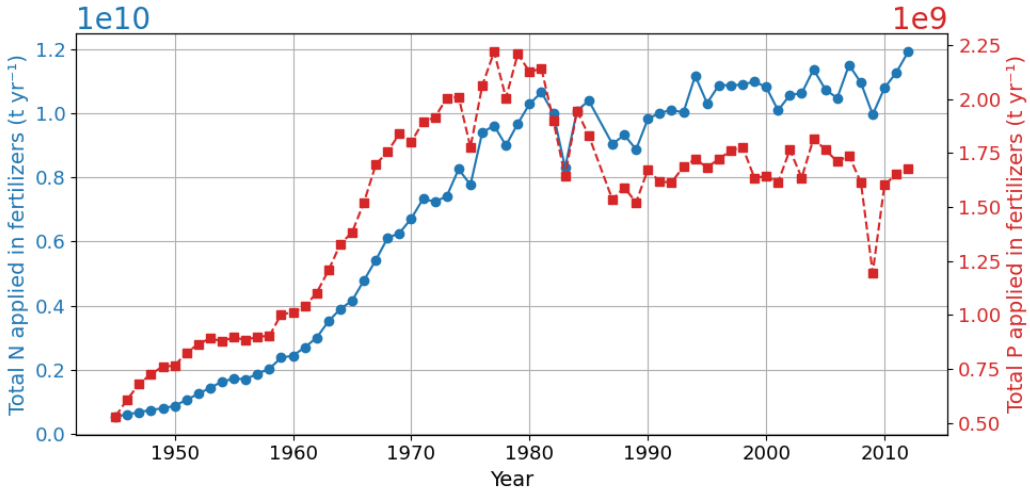

Fig. S28 Annual total N and P input via fertilizers (metric tons per year) within 1945-2012, based on the county-level total data from Falcone (2021).

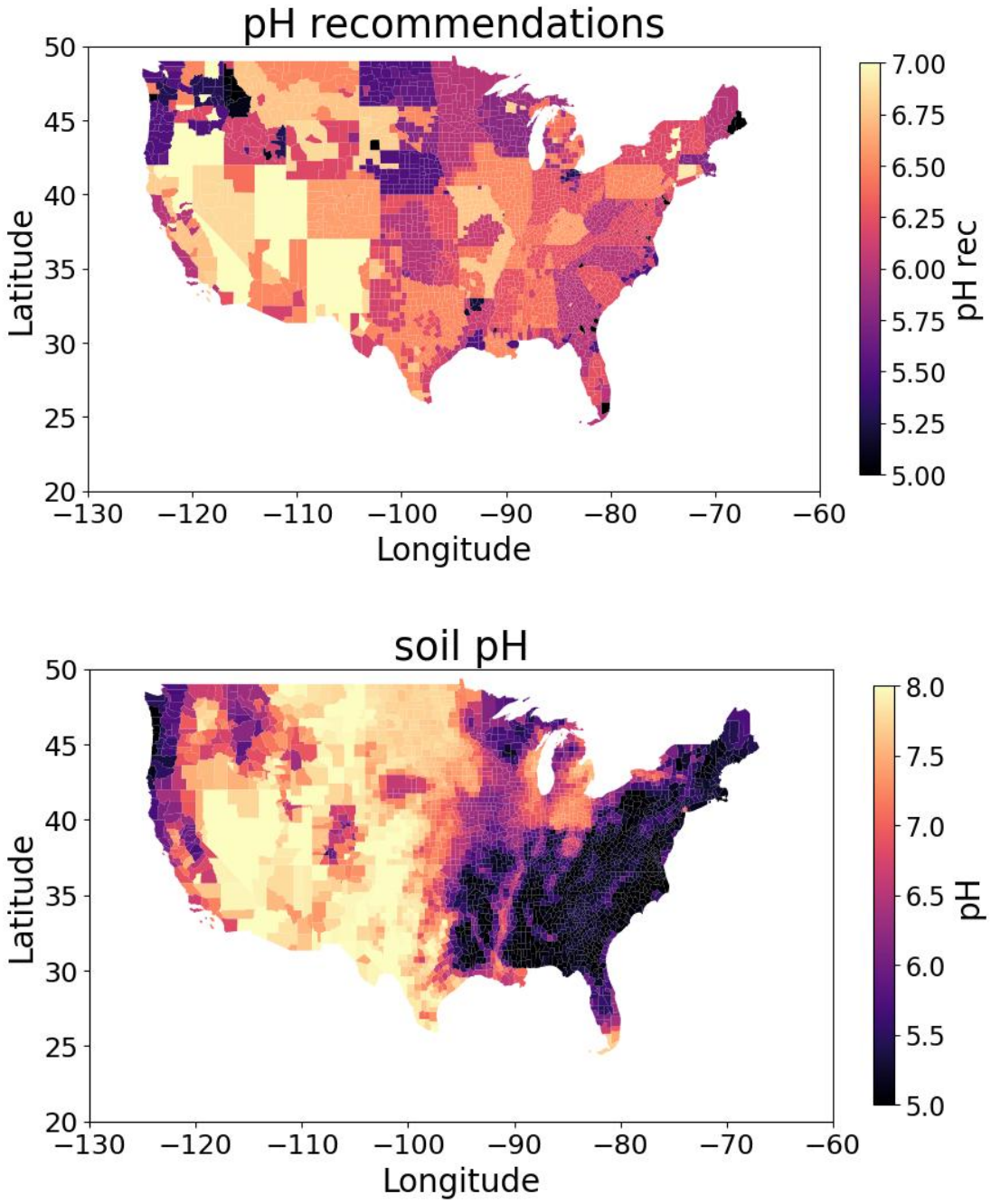

Fig. S29 Annual total N and P input via fertilizers (metric tons per year) within 1945-2012, based on the county-level total data from Falcone (2021).