agriculture

MDPI

Supplementary Information: Leveraging important covariate groups for corn yield prediction

Data Repository & Reproducibility Statement

We have made all raw data collected for this project from the USDA-NASS, DayMet group, MirRAD project, USDA CDL, and HWSD available at our GitHub page, https://github.com/blschum/corn-yield-infill/data/data-raw. In addition, all scripts used to select variables (variable-selection.Rmd/.html), impute data (linear-interpolation.Rmd/.html), tune RFs (tune-ranger-models.Rmd), build RFs (cornyield-manuscript-analysis.Rmd), and visualize our results (cornyield-manuscript-visualizations.Rmd, cornyield-SI-visualizations.Rmd), can be found at the bean-collaboration branch of our GitHub repository (https://github.com/blschum/corn-yield-infill/). Finally, we have made our infilled corn yield data product from 2008-2018, available at https://github.com/blschum/corn-yield-infill/results/ranger-results.

We believe in open data science and in making research more reproducible; if you run into issues in any of our code, please reach out to blschum on GitHub to address them.

We hope these clean datasets will be used in future research to further elucidate links between crop yields and future cropscapes.

Table S1. Full list of available historical biophysical predictors. From this full list, we select only the nine climatic features most predictive of corn yield and intrinsic soil properties likely to remain stable through time, as described in the manuscript.

Variable Name	Description
Climate	
Growing degree days	GDD: an indicator of cumulative temperature exposure; the sum of maximum daily temperatures within a crop-specific tolerance range (10°C to 30°C for corn) over the growing season
Total precipitation	TP: sum of precipitation (in millimeters) throughout the growing season
Mean annual temp.	bio1 = Mean annual temperature
Mean diurnal range	bio2 = Mean diurnal range (mean of max temp - min temp)
Isothermality	bio3 = Isothermality (bio2/bio7) (* 100)
Temp. seasonality	bio4 = Temperature seasonality (standard deviation *100)
Max warm temp.	bio5 = Max temperature of warmest month
Min cold temp.	bio6 = Min temperature of coldest month
Temp. range	bio7 = Temperature annual range (bio5-bio6)
Mean wet temp.	bio8 = Mean temperature of the wettest quarter
Mean dry temp.	bio9 = Mean temperature of driest quarter
Mean warm temp.	bio10 = Mean temperature of warmest quarter
Mean cold temp.	bio11 = Mean temperature of coldest quarter
Total precip.	bio12 = Total (annual) precipitation
Precip. wet month	bio13 = Precipitation of wettest month
Precip. dry month	bio14 = Precipitation of driest month
Precip. seasonality	bio15 = Precipitation seasonality (coefficient of variation)
Precip. wet quarter	bio16 = Precipitation of wettest quarter
Precip. dry quarter	bio17 = Precipitation of driest quarter
Precip. warm quarter	bio18 = Precipitation of warmest quarter
Precip. cold quarter	bio19 = Precipitation of coldest quarter
Soil characteristics	
Subsoil pH	S_PH_H2O: (-log(H^+)) soil reaction; a measure of the acidity alkalinity of the subsoil (5 classes with specific agronomic significance)
Base saturation	T_BS: (%) measure the sum of exchangeable cations (nutrients) Na, Ca, Mg, and K as a percentage of the overall exchange capacity of the topsoil
Calcium carbonate	T_CACO3: (% weight) total topsoil lime content; calcium carbonate is the active ingredient in agricultural lime. Low levels enhance soil structure and are generally beneficial for crop production while higher concentrations may induce iron deficiency and limit the water storage capacity of soils
CEC clay	T_CEC_CLAY: (Cmol/kg) topsoil cation exchange capacity of the clay fraction (classes 1-4)
CEC soil	T_CEC_SOIL: (Cmol/kg) Cation exchange capacity (total nutrient fixing capacity of a soil; topsoil with low CEC have little resilience and cannot build up stores of nutrients); the clay content, OM content, and clay type determine the total nutrient storage capacity; values > 10 cmol/kg are considered satisfactory for most crops (class 1-5)
Clay	T_CLAY: (% weight) percentage topsoil clay (diameter less than 0.002 mm; composed of fine-grained materials that are plastic when wet and hardens when heated; hydrated silicates or aluminum)
Sodium perc.	T_ESP: (%) exchangeable topsoil sodium percentage; indicates levels of sodium hazards in crops
Gravel	T_GRAVEL: (% vol.) volume percentage topsoil gravel (materials larger than 2 mm)
Organic carbon	T_OC: (% weight) percentage of topsoil organic carbon; OC with pH is the best simple indicator of the health status of soils (moderate to high amounts of organic carbon are associated with fertile soils with good structure (codes 1-5, where 1 = very poor in organic carbon)
Subsoil pH	S_PH_H2O: subsoil pH (H2O) (-log(H^+)), soil reaction; a measure of the acidity alkalinity of the soil (5 classes with specific agronomic significance)
Bulk density	T_REF_BULK_DENSITY: (kg/dm3) property of topsoil particulate materials; the mass of many particles of the material / volume (space between particles and the space inside of pores of individual particles) they occupy
Silt	T_SILT: (% weight) percentage topsoil silt (produced by mechanical weathering of rock as opposed to chemical weathering which produces clay; ranges in size from 0.002 to 0.050/0.0625 mm)
Exchangeable bases	T_TEB: (% weight) total exchangeable bases in topsoil; the sum of exchangeable cations in a soil: sodium (Na), calcium (Ca), magnesium (Mg), and potassium (K)
Electrical conductivity	T_ECE: (dS/m) topsoil electrical conductivity; crops vary significantly in their resistance and response to salt in soils (levels indicate agronomic relevant limits)
Sand	T_SAND: (% weight) Percentage sand (particles ranging in diameter from 0.0625 to 2 mm)
Topography	
Slope	SLOPE: Slope (degrees)
Elevation	ELEVATION: Elevation (meters)
Irrigation	

Irrigation	PERC_IRR: Percentage of agricultural land in every county utilizing irrigation (includes all land irrigated by artificial/controlled means, including lagoon wastewater distributed by sprinkler or flood system); measured as the number of agricultural acres irrigated and standardized by the total number of agricultural acres operated, per county. When NASS data was unavailable, % ag acres irrigated backfilled using linear interpolation of MIrAD data.
Diversity	
Shannon's Diversity Index	SDI_CDL_AG: ($\geq$ 0, without limit) A measure of landscape *diversity*; measured as the proportional abundance of each land use category in a county and used as a relative index to compare across landscapes or the same landscape at different times. SDI increases as richness and evenness increase.
Farm inputs/management	
Fertilizer	fert: Total expense of fertilizers, including lime and soil conditioners, rock phosphate and gypsum, and the cost of custom application, per agricultural acre; measured as total expense in USD $ and standardized by the total number of agricultural acres operated, per county.
Chemicals	chem: Total expense of chemicals, including insecticides, herbicides, fungicides, and other pesticides, and the cost of custom application (excludes commercial fertilizer purchased), per agricultural acre; measured as total expense in USD $ and standardized by the total number of agricultural acres operated, per county.
Labor	labor_expense: Total expense of all laborers, per agricultural acre; measured as the total expense of laborers (hired, contract, and migrant) in USD $ and standardized by the total number of agricultural acres operated, per county.
Machinery	machinery: Total asset value of agricultural machinery, per agricultural acre; measured as total machinery assets in USD $ and standardized by the total number of agricultural acres operated, per county.
Corn acreage	perc_corn: Percentage of agricultural land in every county cultivated in corn.
Farm assistance	
Government receipts	gvt_prog: Total cash receipts of government programs, per agricultural acre; measured in USD $ per operation.[1]
Insurance	insur_acres: Percentage of agricultural acres with crop insurance; measured as the number of crop acres with insurance and standardized by the total number of acres, per county.[2]
Farm(er) characteristics	
Years farming	exp: Average number of years experience on present operation.
% farming as primary occupation	occup: Percentage of operators in a county whose primary occupation is farming, standardized by the total number of operators in the county.
% tenants	tenant: Percentage of agricultural acres operated by tenants (producers who operate land they rent from others *and/or* land they worked on shares for others); measured as the number of agricultural acres operated by tenants and standardized by the total number of agricultural acres operated, per county.
Median farm size	acres_per_op: Median farm size (in acres per operation) in a county.

[1] This category consists of direct payments from the government and includes: payments from Conservation Reserve Program, Wetlands Reserve Program, Farmable Wetlands Program, and Conservation Reserve Enhancement Program; loan deficiency payments; disaster payments; other conservation programs; and all other federal farm programs under which payments were made directly to farm operators. Commodity Credit Corporation (CCC) proceeds, local and state government agricultural program payments, and federal crop insurance payments are not tabulated in this category ([25], p. 759).

[2] Agricultural land enrolled in any Federal, private, or other crop insurance programs, measured as the total number of acres with insurance and standardized by the total number of agricultural acres in a county ([25], p. 761).

Table S2. Average model performance. Comparing results using the full set of historical (a) available climate, soil, and topography predictors listed in SI Table 1 and (b) the subset included in the final manuscript for biophysical and farm(er) RF ensembles with default[Δ] and tuned hyperparameters using a 75/25 train test model (see github.com/XXXX/corn-yield-infill/tune-RF.Rmd).

	RMSE	R^2	MAE	MAPE
Biophysical models – default parameters*				
a) all available	16.74	0.817	12.62	0.113
b) subset	16.79	0.816	12.61	0.114
Biophysical models - tuned†				
a) all available	16.74	0.817	12.62	0.113
b) subset	16.75	0.817	12.63	0.113
Farm(er) models – default parameters				
a) all available	16.76	0.817	12.58	0.114
b) subset	17.70	0.796	13.35	0.124
Farm(er) models – tuned◦				
a) all available	16.75	0.817	12.56	0.113
b) subset	17.39	0.803	13.06	0.118

[Δ] Default hyperparameters for ranger: ntree = 500, mtry = sqrt(ncol), nodesize = 5, sample fraction = 0.632

† Hyperparameter tuning for biophysical models: ntree = 2000, mtry = 7, node size = 4, sample fraction = 0.800

◦ Hyperparameter tuning for farm(er) models: ntree = 2000, mtry = 19, node size = 4, sample fraction = 0.800

Table S3. RF variable importance rankings and accuracy metrics for biophysical models. Comparing results with mtry = 1 – 15 (by 2), min.node.size = 4, sample.fraction = 0.8, and ntree = 2,000 on biophysical data.

	mtry =1	**mtry = 3**	**mtry = 5**	**mtry = 7**	**mtry = 9**	**mtry = 11**	**mtry = 13**	**mtry = 15**
Performance statistics on *n* = 4,246 county-years								
RMSE	19.70	17.61	17.31	17.22	17.23	17.23	17.24	17.25
% variance explained	0.747	0.798	0.804	0.807	0.806	0.806	0.806	0.806
Permutation variable importance								
1 (most)	GDD	GDD	PERC_IRR	PERC_IRR	PERC_IRR	PERC_IRR	PERC_IRR	PERC_IRR
2	FRR	PERC_IRR	GDD	GDD	GDD	YEAR	FRR	YEAR
3	PERC_IRR	YEAR	YEAR	YEAR	YEAR	GDD	YEAR	FRR
4	YEAR	FRR	FRR	FRR	FRR	FRR	GDD	GDD
5	S_PH_H20	BV2	BV2	BV2	BV2	BV18	BV18	BV18
6	BV2	S_PH_H2O	BV18	BV18	BV18	BV2	BV2	BV2
7	ELEVATION	BV18	S_PH_H2O	SDI_CDL_AG	SDI_CDL_AG	SDI_CDL_AG	SDI_CDL_AG	SDI_CDL_AG
8	BV4	BV9	SDI_CDL_AG	S_PH_H2O	S_PH_H2O	S_PH_H2O	S_PH_H2O	S_PH_H2O
9	BV19	SDI_CDL_AG	BV9	BV9	BV9	BV9	BV9	BV9
10	BV9	BV4	ELEVATION	ELEVATION	ELEVATION	BV4	BV4	BV4
11	T_CEC_SOIL	ELEVATION	BV4	BV4	BV4	ELEVATION	ELEVATION	ELEVATION
12	BV18	T_CEC_SOIL	T_CEC_SOIL	T_CEC_SOIL	BV19	BV19	BV19	BV19
13	SDI_CDL_AG	BV19	BV19	BV19	T_CEC_SOIL	T_CEC_SOIL	TP	SLOPE
14	T_REF_BULK_ DENSITY	TP	TP	TP	TP	TP	SLOPE	TP
15	SLOPE	SLOPE	SLOPE	SLOPE	SLOPE	SLOPE	T_CEC_SOIL	T_CEC_SOIL
16	TP	T_REF_BULK_ DENSITY	T_OC	T_OC	T_OC	T_OC	T_OC	T_OC
17	T_OC	T_OC	BV15	BV15	BV15	BV15	BV15	BV15
18	BV15	BV15	T_REF_BULK_ DENSITY	T_REF_BULK_ DENSITY	T_REF_BULK_ DENSITY	T_REF_BULK_ DENSITY	BV8	BV8
19 (least)	BV8	BV8	BV8	BV8	BV8	BV8	T_REF_BULK_ DENSITY	T_REF_BULK_ DENSITY

Table S4. RF variable importance rankings and accuracy metrics for farm(er) models. Comparing results with mtry = 5 – 19 (by 2), min.node.size = 4, sample.fraction = 0.8, and ntree = 2,000 on farm(er) data (farm(er) covariates **bolded**).

	mtry =5	**mtry = 7**	**mtry = 9**	**mtry = 11**	**mtry = 13**	**mtry = 15**	**mtry = 17**	**mtry = 19**
Performance statistics on *n* = 4,246 county-years								
RMSE	17.82	17.57	17.47	17.38	17.34	17.31	17.30	17.30
% variance explained	0.793	0.797	0.801	0.803	0.804	0.805	0.805	0.805
Permutation variable importance								
1 (most)	YEAR	PERC_IRR	PERC_IRR	PERC_IRR	PERC_IRR	PERC_IRR	PERC_IRR	PERC_IRR
2	PERC_IRR	YEAR	YEAR	YEAR	YEAR	YEAR	YEAR	YEAR
3	GDD	GDD	GDD	GDD	GDD	GDD	GDD	GDD
4	**chem**	**fert**	**fert**	**fert**	**fert**	**fert**	**fert**	**fert**
5	**fert**	**chem**	BV18	BV18	BV18	BV18	BV18	BV18
6	BV18	BV18	**chem**	**chem**	**chem**	**chem**	**chem**	**chem**
7	BV2	BV2	BV2	perc_corn	BV2	BV2	BV2	BV2
8	perc_corn	perc_corn	perc_corn	BV2	perc_corn	perc_corn	perc_corn	SDI_CDL_AG
9	**insur_acres**	**insur_acres**	**insur_acres**	S_PH_H2O	S_PH_H2O	S_PH_H2O	S_PH_H2O	S_PH_H2O
10	S_PH_H2O	S_PH_H2O	S_PH_H2O	**insur_acres**	**insur_acres**	SDI_CDL_AG	SDI_CDL_AG	perc_corn
11	**gvt_prog**	**gvt_prog**	**gvt_prog**	BV9	BV9	BV9	BV9	BV9
12	FRR	BV9	BV9	gvt_prog	govt_prog	**insur_acres**	**insur_acres**	**insur_acres**
13	BV9	ELEVATION	ELEVATION	SDI_CDL_AG	SDI_CDL_AG	BV4	BV4	BV4
14	**machinery**	FRR	BV4	ELEVATION	**gvt_prog**	**gvt_prog**	ELEVATION	ELEVATION
15	ELEVATION	**machinery**	FRR	BV4	ELEVATION	ELEVATION	**gvt_prog**	**gvt_prog**
16	BV4	BV4	**machinery**	FRR	BV4	FRR	**machinery**	**machinery**
17	T_CEC_SOIL	T_CEC_SOIL	T_CEC_SOIL	**machinery**	**machinery**	T_CEC_SOIL	T_CEC_SOIL	T_CEC_SOIL
18	SDI_CDL_AG	SDI_CDL_AG	SDI_CDL_AG	T_CEC_SOIL	FRR	**machinery**	FRR	FRR
19	TP	TP	TP	TP	T_CEC_SOIL	TP	TP	TP
20	BV19	BV19	BV19	BV19	TP	BV19	BV19	BV19
21	**labor_expense**	**labor_expense**	**labor_expense**	**labor_expense**	BV19	**labor_expense**	**labor_expense**	**labor_expense**
22	**acres_per_op**	**acres_per_op**	SLOPE	SLOPE	**labor_expense**	SLOPE	SLOPE	SLOPE
23	T_REF_BULK_ DENSITY	T_REF_BULK_ DENSITY	**acres_per_op**	**acres_per_op**	SLOPE	**acres_per_op**	BV15	BV15
24	SLOPE	SLOPE	T_REF_BULK_ DENSITY	T_REF_BULK_ DENSITY	**acres_per_op**	BV15	**acres_per_op**	TOC
25	T_OC	T_OC	T_OC	T_OC	T_OC	T_OC	T_OC	**acres_per_op**
26	BV15	BV15	BV15	BV15	BV15	BV8	BV8	BV8
27	**tenant**	BV8	BV8	BV8	T_REF_BULK_ DENSITY	T_REF_BULK_ DENSITY	T_REF_BULK_ DENSITY	T_REF_BULK_ DENSITY
28	BV8	**tenant**	**tenant**	**tenant**	**tenant**	**tenant**	**tenant**	**tenant**
29	**occup**	**occup**	**occup**	**occup**	**occup**	**occup**	**occup**	**occup**
30 (least)	**exp**	**exp**	**exp**	**exp**	**exp**	**exp**	**exp**	**exp**

Table S5. List of counties excluded due to missing data across census years, with Census FIPS codes. Note: See SI Figure 3 for map of counties excluded.

GEOID	County	State
04001	Apache	Arizona
04005	Coconino	Arizona
04007	Gila	Arizona
04012	La Paz	Arizona
04019	Pima	Arizona
05013	Calhoun	Arkansas
05025	Cleveland	Arkansas
05039	Dallas	Arkansas
06003	Alpine	California
06015	Del Norte	California
06051	Mono	California
06057	Nevada	California
06075	San Francisco	California
06105	Trinity	California
08014	Broomfield	Colorado
08019	Clear Creek	Colorado
08031	Denver	Colorado
08047	Gilpin	Colorado
08053	Hinsdale	Colorado
08065	Lake	Colorado
08079	Mineral	Colorado
08097	Pitkin	Colorado
08111	San Juan	Colorado
08117	Summit	Colorado
11001	District of Columbia	
12003	Baker	Florida
12005	Bay	Florida
12037	Franklin	Florida
12045	Gulf	Florida
12073	Leon	Florida
12077	Liberty	Florida
12087	Monroe	Florida
12089	Nassau	Florida
12103	Pinellas	Florida
12123	Taylor	Florida
12129	Wakulla	Florida
13029	Bryan	Georgia
13039	Camden	Georgia
13053	Chattahoochee	Georgia
13059	Clarke	Georgia
13063	Clayton	Georgia
13065	Clinch	Georgia
13089	DeKalb	Georgia
13097	Douglas	Georgia
13101	Echols	Georgia
13127	Glynn	Georgia
13179	Liberty	Georgia
13183	Long	Georgia
13191	McIntosh	Georgia
13215	Muscogee	Georgia
13225	Peach	Georgia
13239	Quitman	Georgia
13247	Rockdale	Georgia
13263	Talbot	Georgia
13265	Taliaferro	Georgia
13281	Towns	Georgia
13289	Twiggs	Georgia

16003	Adams	Idaho
16015	Boise	Idaho
16017	Bonner	Idaho
16079	Shoshone	Idaho
16085	Valley	Idaho
17031	Cook	Illinois
17043	DuPage	Illinois
18013	Brown	Indiana
21013	Bell	Kentucky
21071	Floyd	Kentucky
21095	Harlan	Kentucky
21115	Johnson	Kentucky
21119	Knott	Kentucky
21121	Knox	Kentucky
21129	Lee	Kentucky
21131	Leslie	Kentucky
21133	Letcher	Kentucky
21159	Martin	Kentucky
21193	Perry	Kentucky
21195	Pike	Kentucky
22033	East Baton Rouge	Louisiana
22051	Jefferson	Louisiana
22059	LaSalle	Louisiana
22071	Orleans	Louisiana
22087	St. Bernard	Louisiana
22089	St. Charles	Louisiana
23023	Sagadahoc	Maine
24510	Baltimore	Maryland
25007	Dukes	Massachusetts
25019	Nantucket	Massachusetts
25025	Suffolk	Massachusetts
26013	Baraga	Michigan
26039	Crawford	Michigan
26053	Gogebic	Michigan
26083	Keweenaw	Michigan
26095	Luce	Michigan
26119	Montmorency	Michigan
26135	Oscoda	Michigan
26143	Roscommon	Michigan
26153	Schoolcraft	Michigan
27031	Cook	Minnesota
27075	Lake	Minnesota
27123	Ramsey	Minnesota
28129	Smith	Mississippi
29510	St. Louis	Missouri
30053	Lincoln	Montana
30061	Mineral	Montana
30093	Silver Bow	Montana
31075	Grant	Nebraska
32009	Esmeralda	Nevada
32011	Eureka	Nevada
32017	Lincoln	Nevada
32021	Mineral	Nevada
32029	Storey	Nevada
32510	Carson City	Nevada
33003	Carroll	New Hampshire
33019	Sullivan	New Hampshire
34003	Bergen	New Jersey
34009	Cape May	New Jersey
34013	Essex	New Jersey
34017	Hudson	New Jersey

34029	Ocean	New Jersey
34031	Passaic	New Jersey
34039	Union	New Jersey
35021	Harding	New Mexico
35028	Los Alamos	New Mexico
35049	Santa Fe	New Mexico
36005	Bronx	New York
36041	Hamilton	New York
36047	Kings	New York
36059	Nassau	New York
36061	New York	New York
36079	Putnam	New York
36081	Queens	New York
36085	Richmond	New York
36087	Rockland	New York
36113	Warren	New York
36119	Westchester	New York
37055	Dare	North Carolina
37129	New Hanover	North Carolina
41007	Clatsop	Oregon
41041	Lincoln	Oregon
41057	Tillamook	Oregon
42023	Cameron	Pennsylvania
42045	Delaware	Pennsylvania
42053	Forest	Pennsylvania
42101	Philadelphia	Pennsylvania
42103	Pike	Pennsylvania
44001	Bristol	South Dakota
44003	Kent	South Dakota
47171	Unicoi	Tennessee
48007	Aransas	Texas
48103	Crane	Texas
48135	Ector	Texas
48183	Gregg	Texas
48261	Kenedy	Texas
48269	King	Texas
48301	Loving	Texas
48377	Presidio	Texas
48475	Ward	Texas
48495	Winkler	Texas
49009	Daggett	Utah
49019	Grand	Utah
49047	Uintah	Utah
51013	Arlington	Virginia
51027	Buchanan	Virginia
51051	Dickenson	Virginia
51057	Essex	Virginia
51510	Alexandria	Virginia
51520	Bristol	Virginia
51530	Buena Vista	Virginia
51540	Charlottesville	Virginia
51570	Colonial Heights	Virginia
51580	Covington	Virginia
51590	Danville	Virginia
51595	Emporia	Virginia
51600	Fairfax	Virginia
51610	Falls Church	Virginia
51620	Franklin	Virginia
51630	Fredericksburg	Virginia
51640	Galax	Virginia
51650	Hampton	Virginia

51660	Harrisonburg	Virginia
51670	Hopewell	Virginia
51678	Lexington	Virginia
51680	Lynchburg	Virginia
51683	Manassas	Virginia
51685	Manassas Park	Virginia
51690	Martinsville	Virginia
51700	Newport News	Virginia
51710	Norfolk	Virginia
51720	Norton	Virginia
51730	Petersburg	Virginia
51735	Poquoson	Virginia
51740	Portsmouth	Virginia
51750	Radford	Virginia
51760	Richmond	Virginia
51770	Roanoke	Virginia
51775	Salem	Virginia
51790	Staunton	Virginia
51820	Waynesboro	Virginia
51830	Williamsburg	Virginia
51840	Winchester	Virginia
53009	Clallam	West Virginia
53031	Jefferson	West Virginia
53045	Mason	West Virginia
53051	Pend Oreille	West Virginia
53055	San Juan	West Virginia
53059	Skamania	West Virginia
53069	Wahkiakum	West Virginia
54005	Boone	Wisconsin
54007	Braxton	Wisconsin
54009	Brooke	Wisconsin
54013	Calhoun	Wisconsin
54015	Clay	Wisconsin
54017	Doddridge	Wisconsin
54019	Fayette	Wisconsin
54021	Gilmer	Wisconsin
54039	Kanawha	Wisconsin
54041	Lewis	Wisconsin
54045	Logan	Wisconsin
54047	McDowell	Wisconsin
54051	Marshall	Wisconsin
54055	Mercer	Wisconsin
54059	Mingo	Wisconsin
54073	Pleasants	Wisconsin
54081	Raleigh	Wisconsin
54085	Ritchie	Wisconsin
54089	Summers	Wisconsin
54091	Taylor	Wisconsin
54095	Tyler	Wisconsin
54101	Webster	Wisconsin
54103	Wetzel	Wisconsin
54109	Wyoming	Wisconsin
55037	Florence	Wyoming
55078	Menominee	Wyoming
55125	Vilas	Wyoming

Table S6. The RMSE accuracy results for one iteration of PyCaret package using all possible explanatory variables. Results are provided using a 75/25 training/test approach, along with the results for 5-fold cross validation.

Model Name	RMSE ~ Training/Test Results	RMSE ~ 5-fold Cross Validation Results
Extra Trees Regressor	16.82	15.69
Light Gradient Boosting Machine	18.04	17.15
Random Forest Regressor	18.10	16.82
Gradient Boosting Regressor	21.11	21.02
K Neighbors Regressor	23.41	22.28
Decision Tree Regressor	25.62	25.26
AdaBoost Regressor	25.62	26.60
Ridge Regression	25.90	26.48
Linear Regression	25.90	26.48
Bayesian Ridge	25.95	26.50
Lasso Regression	26.41	26.93
Elastic Net	26.58	27.07
Huber Regressor	29.42	29.71
Least Angle Regression	26.46	26.48
Orthogonal Matching Pursuit	32.28	32.82
Dummy Regressor	38.98	39.38
Lasso Least Angle Regression	38.98	39.38
Passive Aggressive Regressor	34.40	43.21

Figure S1. Count of missingness across Survey years. Note: Where category is 0, a county reported yield data across all 11 years (e.g., Corn Belt), where the category is 10, a county reported only 1 year of yield data (e.g., southern Arizona). The "missing" category refers to all counties where NASS reported no corn yields from 2008-2018; these counties may be "missing" because they produced no corn, or, more likely, because there were not enough reporting growers to meet NASS statistical disclosure requirements.

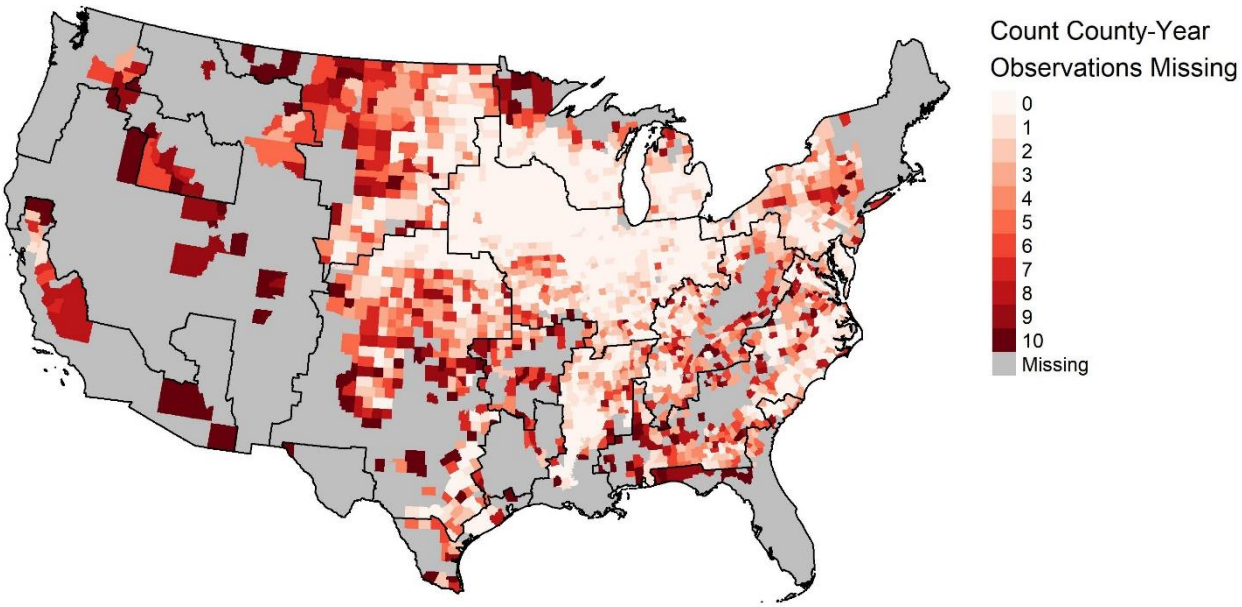


Figure S2. Correlation matrix for continuous predictors and corn yield (see Table S1, above for variables, units, and descriptions associated with variable names listed below).

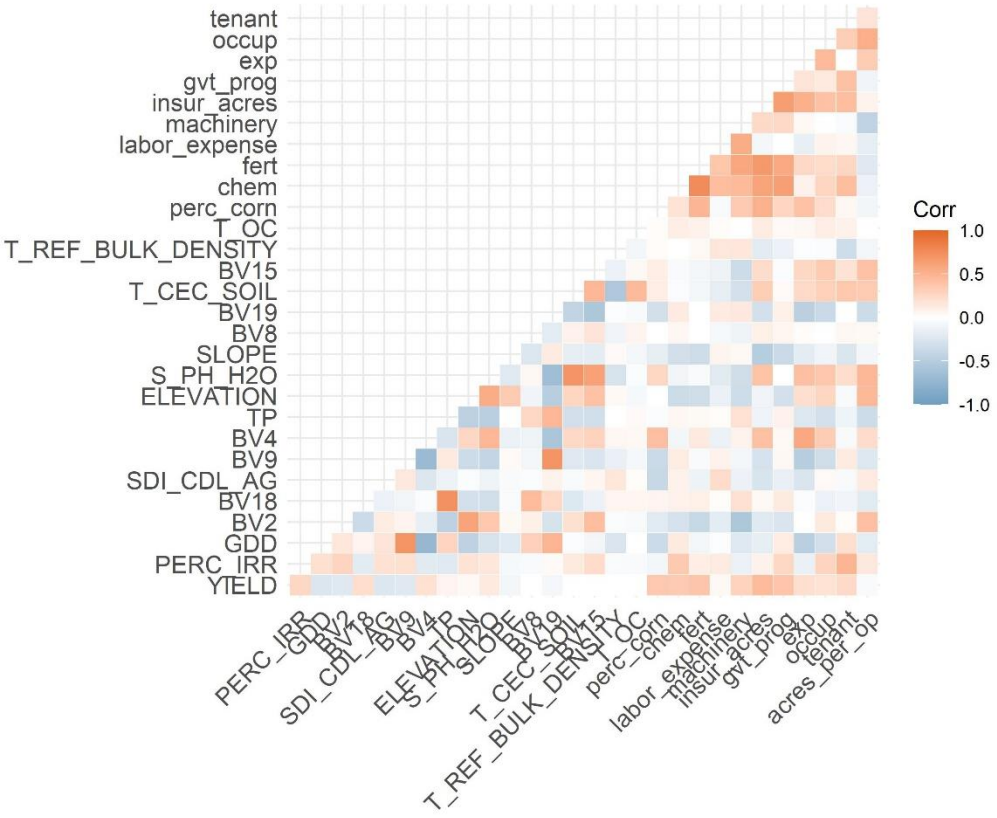


Figure S3. Map of counties excluded due to missing data across census years. Note: "Missing" here refers to all other US counties not removed due to CoA imputation.

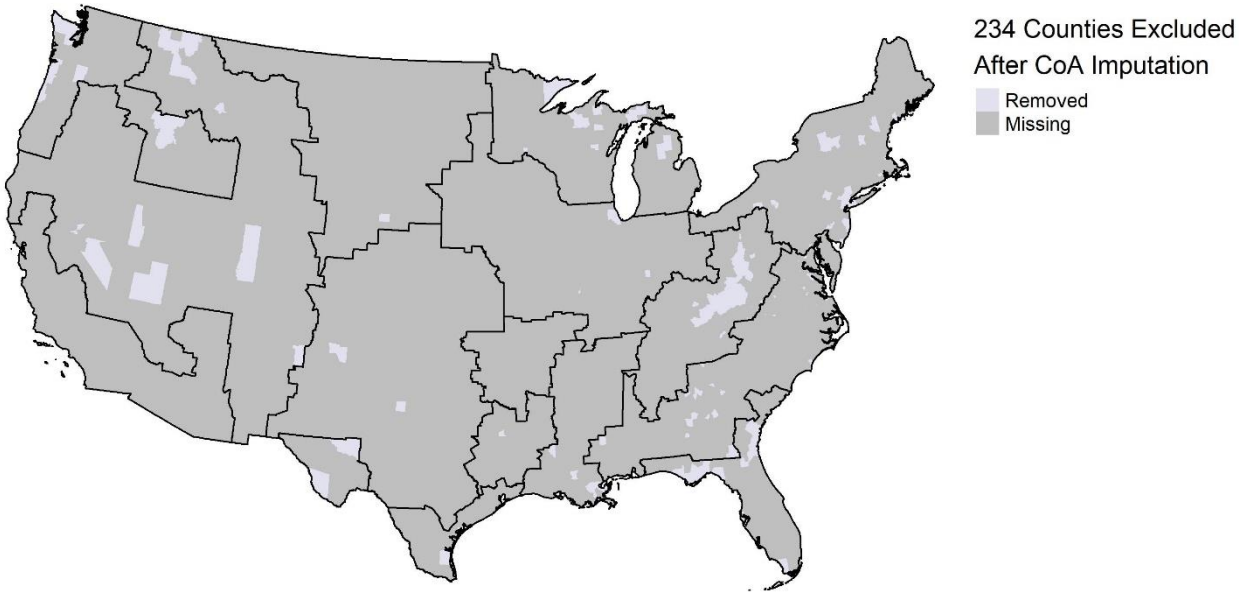


Figure S4. Biophysical RF ensemble model performance based on five-fold cross validation. A) Average error in model predictions in bushels/acre observed across ensemble models (n = 50 replications) and years (2008-2018). Counties in which predicted yields were less (greater) than observed yields appear in shades of red (blue). B) Observed v. predicted plot for the full panel dataset. The dashed line indicates a 1:1 relationship, the solid blue line shows a linear regression between the observed and predicted yields.

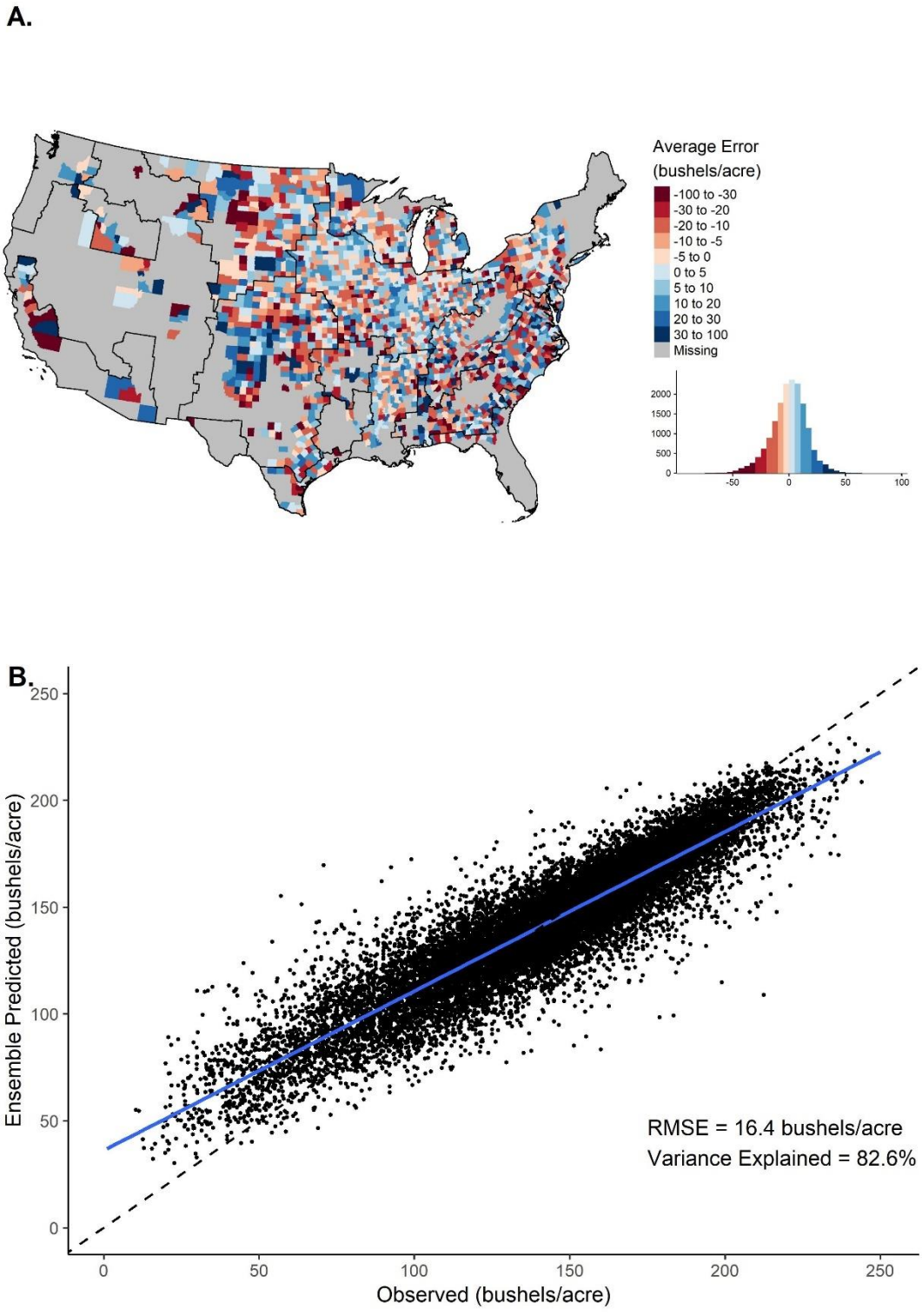


Figure S5. Farm(er) RF ensemble model performance based on five-fold cross validation. A) Average error in model predictions in bushels/acre observed across ensemble models ($n = 50$ replications) and years (2008-2018). Counties in which predicted yields were less (greater) than observed yields appear in shades of red (blue). B) Observed v. predicted plot for the full panel dataset. The dashed line indicates a 1:1 relationship, the solid blue line shows a linear regression between the observed and predicted yields.

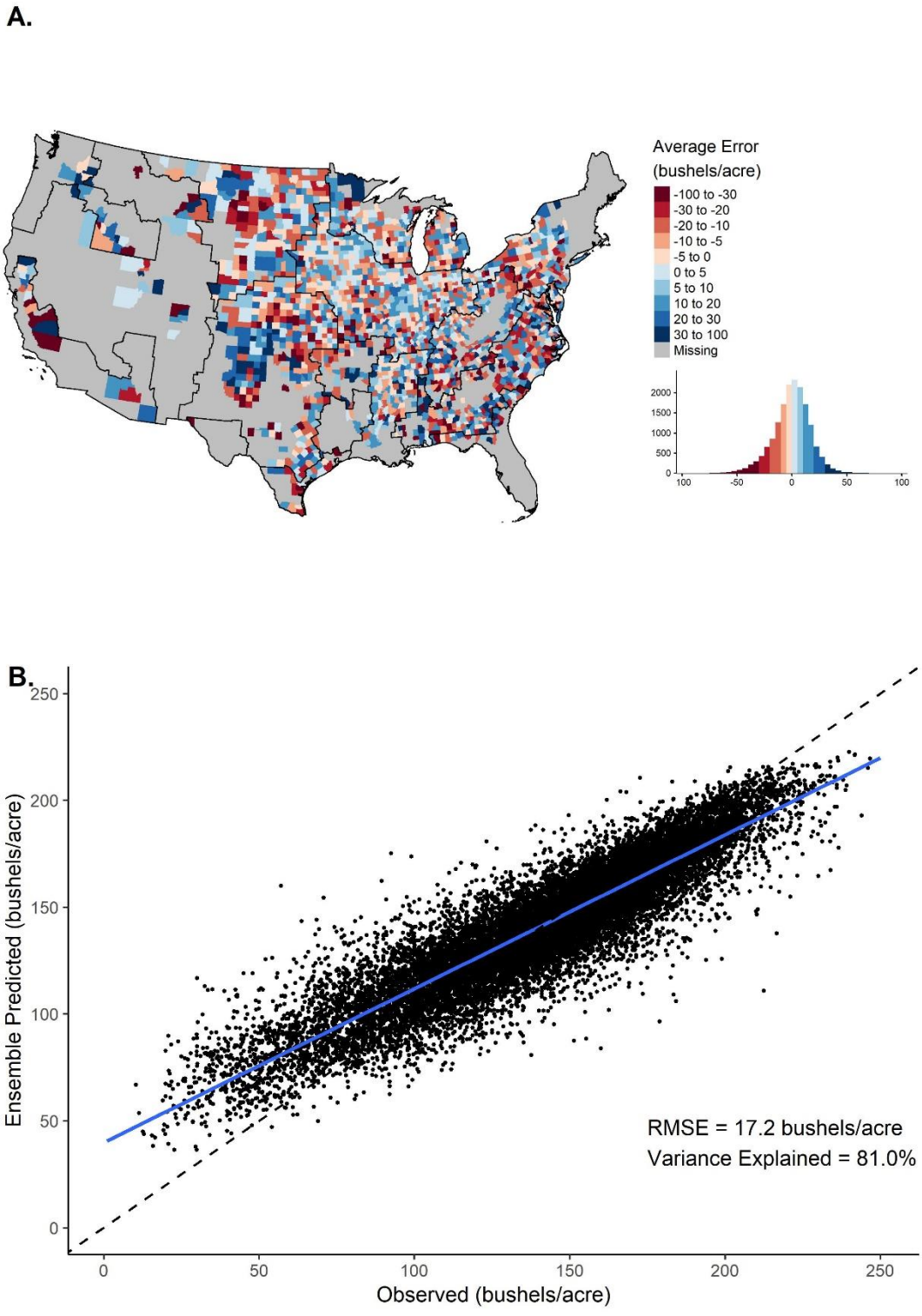


Figure S6. Group exclusion (climate + space + time) RF ensemble model performance based on five-fold cross validation. A) Average error in model predictions in bushels/acre observed across ensemble models (n = 50 replications) and years (2008-2018). Counties in which predicted yields were less (greater) than observed yields appear in shades of red (blue). B) Observed v. predicted plot for the full panel dataset. The dashed line indicates a 1:1 relationship, the solid blue line shows a linear regression between the observed and predicted yields.

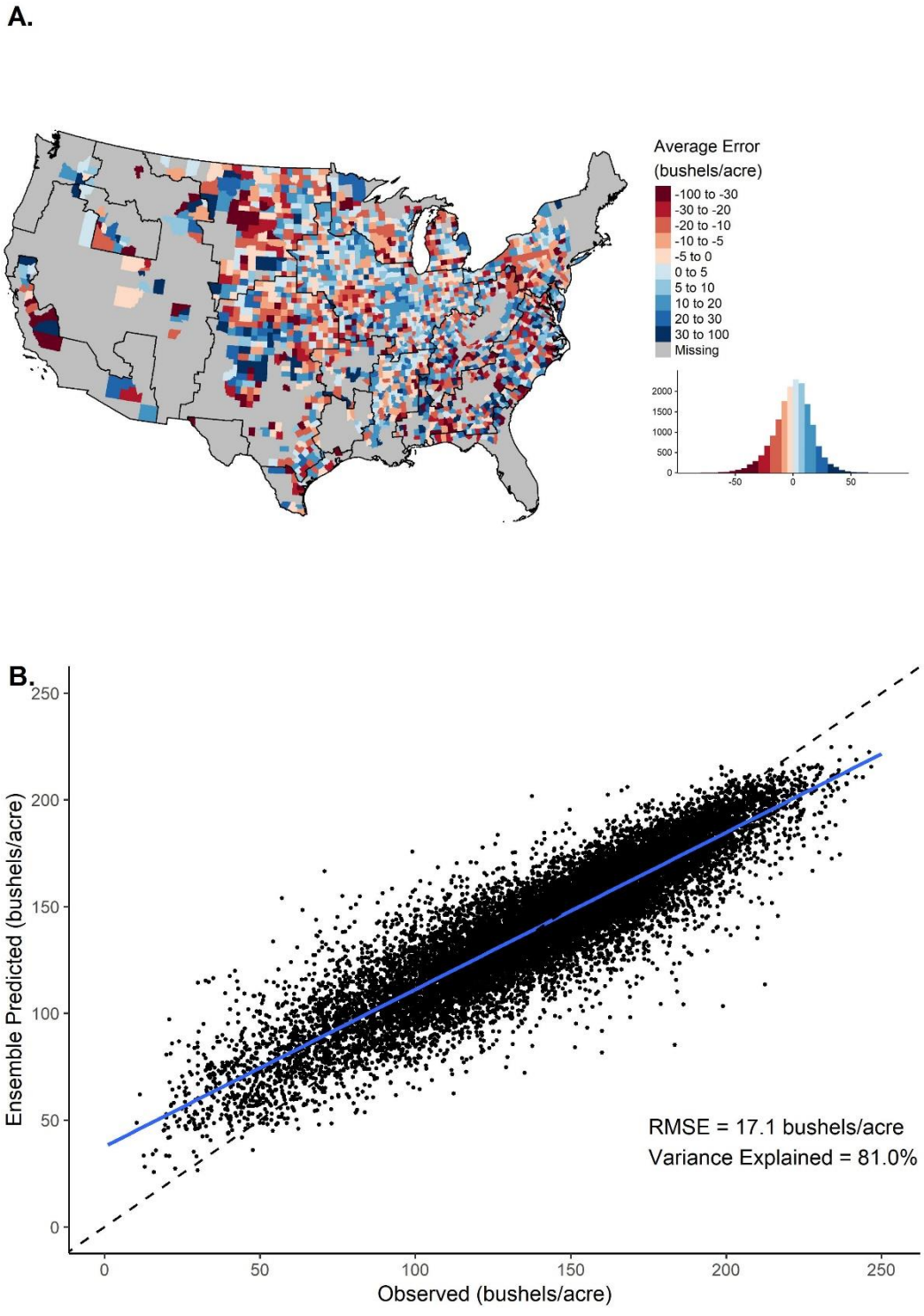


Figure S7. Boxplots of absolute percentage errors (APE) via 5-fold cross validation (one iteration) for all counties as organized by number of missing years. This figure shows that counties with less years of missing yield measurements (1-4 missing years) tend to have slightly smaller relative errors than counties with 5 or more missing years of yield measurement. It also shows that counties with missing values for almost all years tend to have the most extreme relative errors.

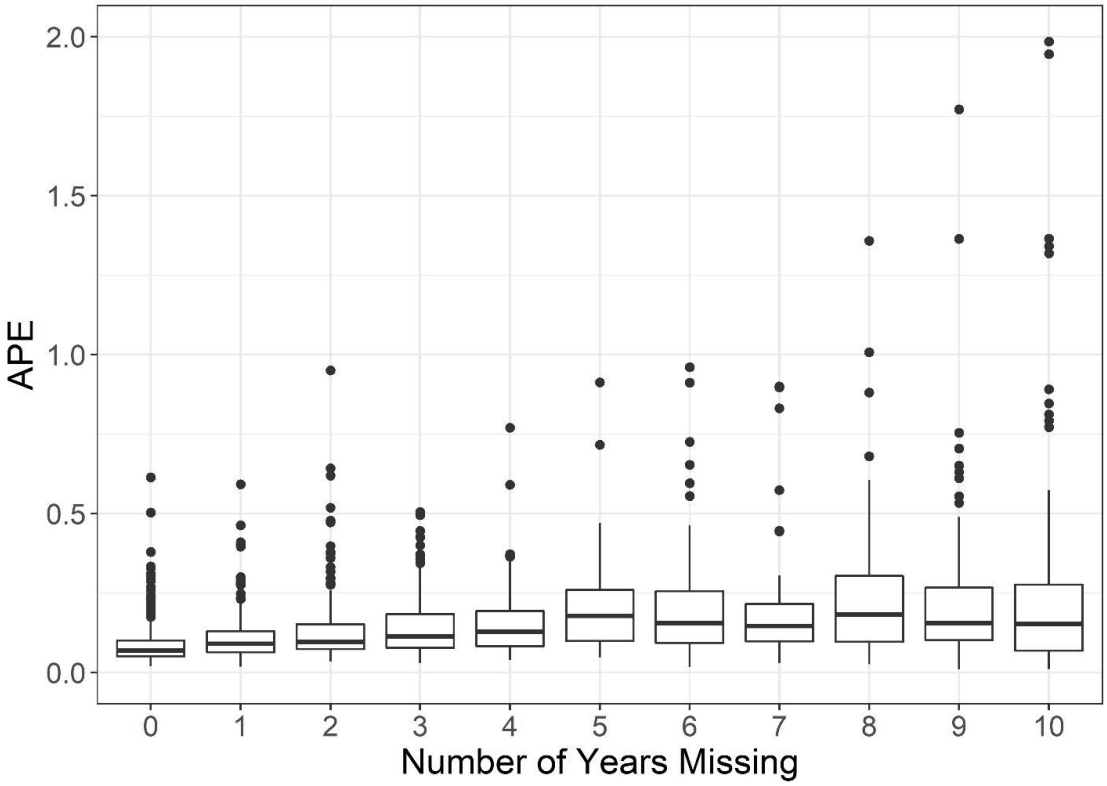
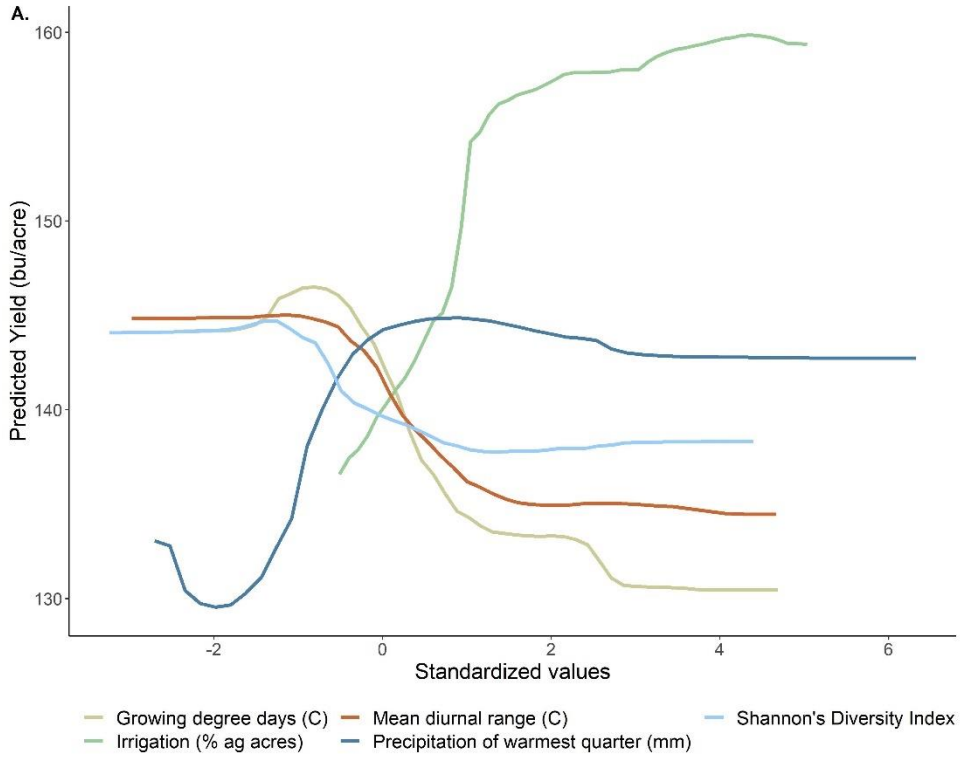


Figure S8. Partial dependence plots for consistently important variables in the a) biophysical ensemble and b) farm(er) ensemble. Partial dependence is the dependence of the outcome on one predictor after averaging out the effects of all other predictors in the model [37]. Partial dependence plots graphically characterize the relationship between an individual predictor (standardized, here) and the predicted values of yield (see Figure S9a-j for unstandardized relationships).



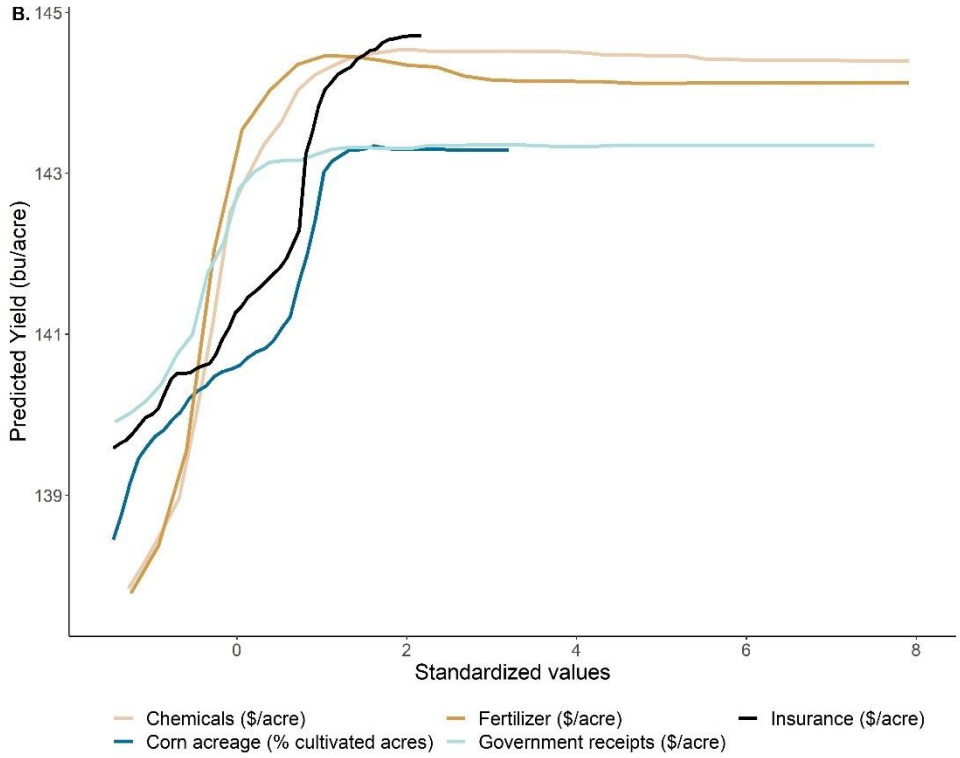


Figure S9. Partial dependence plots of important variables from biophysical and farm(er) ensembles (raw data, not standardized). Partial dependence is the dependence of the outcome on one predictor after averaging out the effects of all other predictors in the model [37]. Partial dependence plots graphically characterize the relationship between an individual predictor and the predicted values of yield. Partial dependence on: A) chemicals ($/acre); B) % cultivated area in corn (perc_corn); C) fertilizers ($/acre); D) government receipts ($/acre); E) growing degree days (GDD); F) insurance ($/acre); G) irrigation (PERC_IRR); H) mean diurnal range (BV2); I) precipitation of the warmest quarter (BV18); J) Shannon's Diversity Index (SDI_CDL_AG).

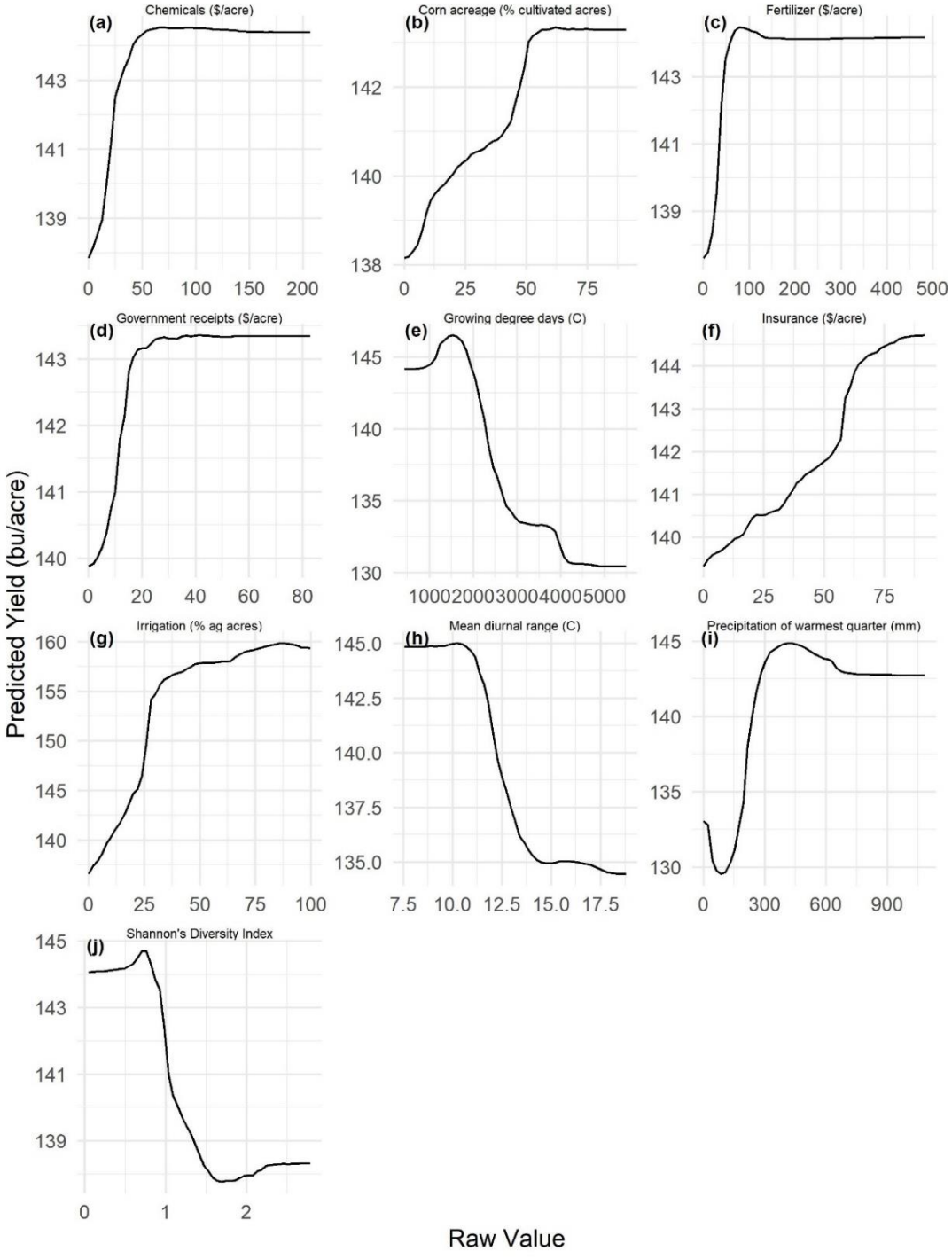


Figure S10. Partial dependence plots of important variables from the reduced ensemble (raw data, not standardized). Partial dependence is the dependence of the outcome on one predictor after averaging out the effects of all other predictors in the model [37]. Partial dependence plots graphically characterize the relationship between an individual predictor and the predicted values of yield. Partial dependence on: A) growing degree days (GDD); B) Irrigation (PERC_IRR); C) Latitude (lat); D) Longitude (lon); E) mean diurnal range (BV2); F) Mean temperature of the driest quarter (BV9); G) mean temperature of the wettest quarter (BV8); H) Precipitation of the coldest quarter (BV19); I) Precipitation of the warmest quarter (BV18); J) Precipitation seasonality (BV15); K) Temperature seasonality (BV4); L) Total precipitation (TP); and M) Year.

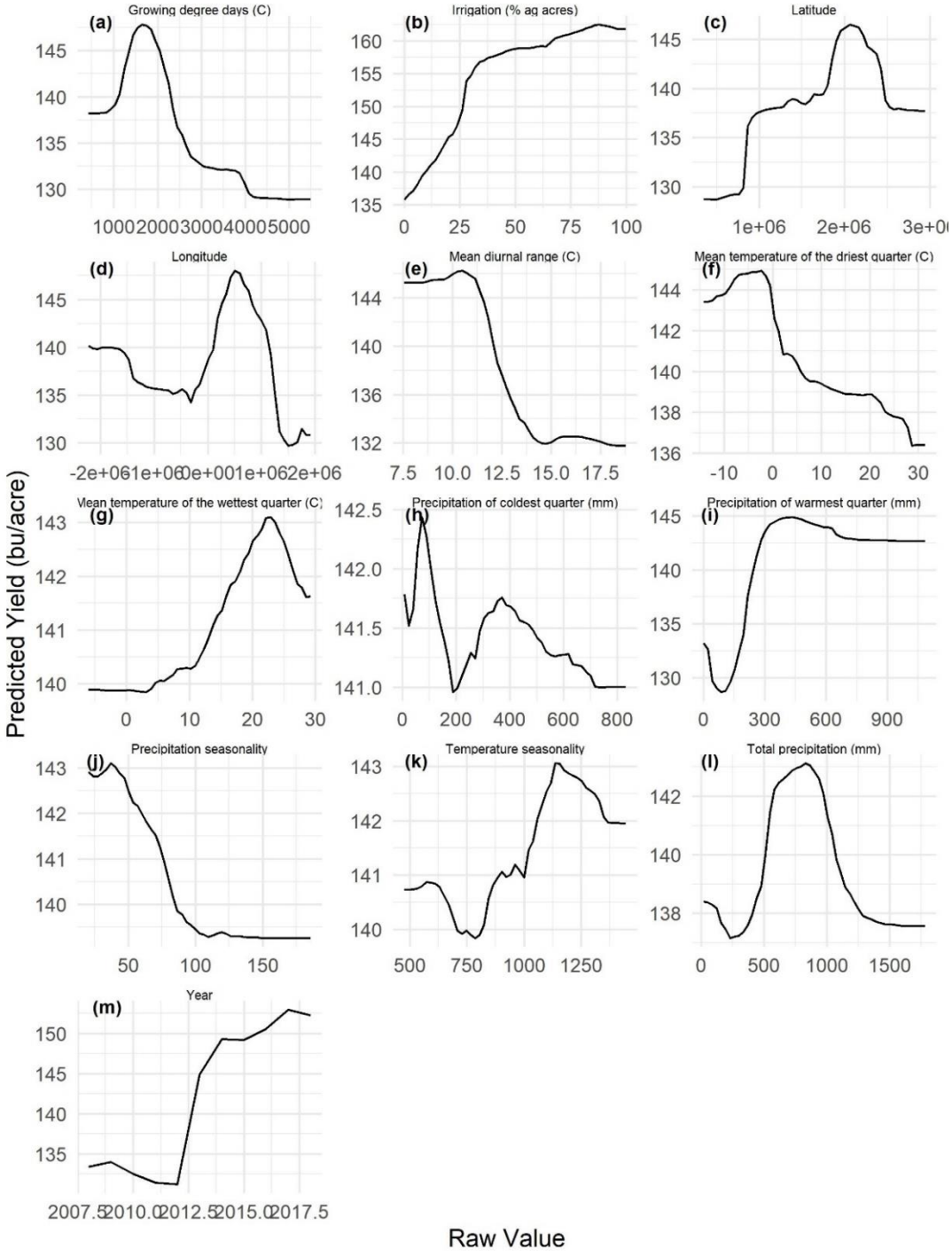


Figure S11. Bivariate choropleth constructed by binning county-level average corn yield (bushels/acre) and percent acres cultivated in corn on agricultural lands into thirds; each tercile is then paired and binned into distinct categories. Yellow indicates counties with high average corn yields and an agricultural landscape dominated by corn production, while purple indicates counties with low average yields and a low percentage of agricultural acres in corn. Gray counties indicate missing data.

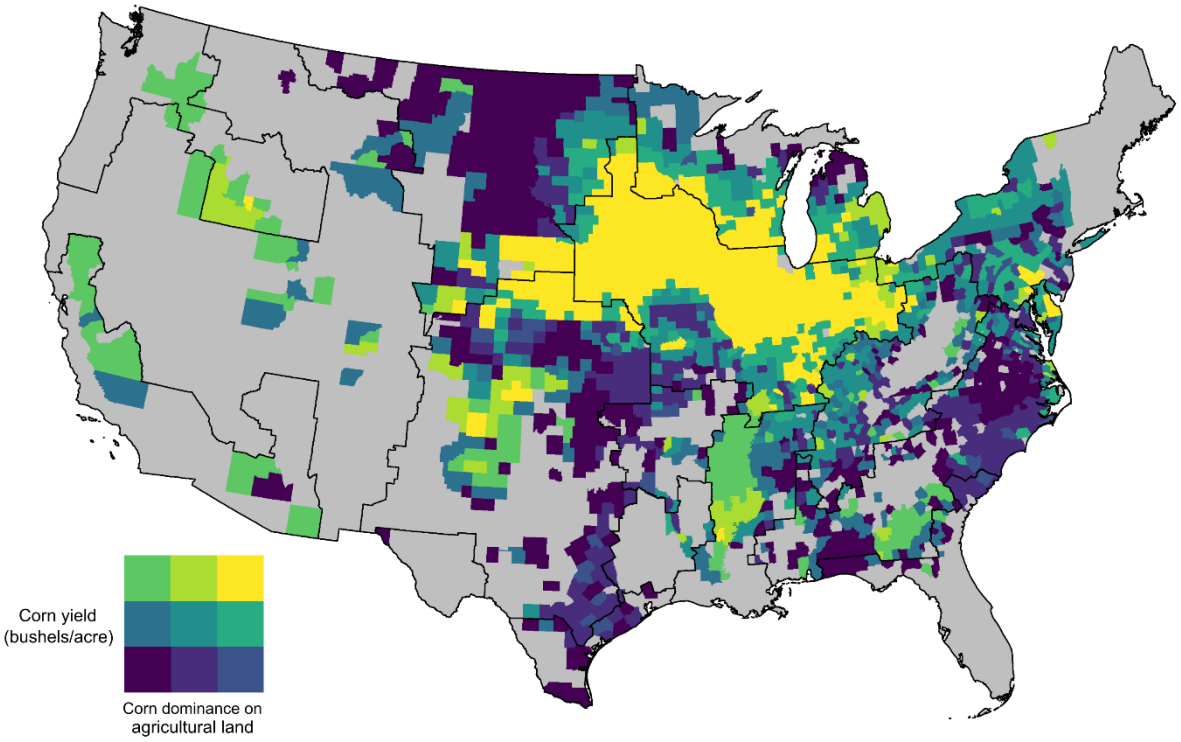


Figure S12. Percent irrigated acreage on agricultural lands across study years (2008-2018) faceted by prediction class.