Geographical Research Letter

Supporting Information for

On the use of SMAP Soil Moisture for Forecasting NDVI Over CONUS Cropland Regions

Manh-Hung Le[1,2], John D. Bolten[1,*], Kristen M. Whitney[1], David M. Johnson[3], Rick Mueller[3], and Iliana E. Mladenova[4]

[1] Hydrological Sciences Laboratory, NASA Goddard Space Flight Center, Greenbelt, MD, USA

[2] Science Applications International Corporation, Greenbelt, MD, USA

[3] U.S. Department of Agriculture, National Agricultural Statistics Service, Washington, DC, USA

[4] U.S. Department of Agriculture, Foreign Agricultural Service, Washington, DC, USA

Contents of this file

Text S1. Crop dataset processing

For crop dataset processing, we initially obtained data from the USDA NASS (National Agricultural Statistics Service). These datasets include county-level crop yield data and a 30-m Crop Data Layer (CDL) for four crop types: corn, cotton, soybeans, and wheat. To minimize noise from different land covers, we selected grid cells identified as consistently cultivating a single crop over time. We used CDL data from 2008–2022, choosing grid cells that appeared in at least three different years. Additionally, for each crop type, based on the Topologically Integrated Geographic Encoding and Referencing (TIGER) administrative boundaries, we focused on the most productive states, which collectively contribute to over 90% of overall crop productivity (see Supplementary Table S2). To explore the impact of water resource management (i.e., irrigation levels), we employed the Moderate Resolution Imaging Spectroradiometer (MODIS) Irrigated Agriculture Dataset for the United States (MIrAD-US) 250m product. MIrAD-US provides a detailed geospatial dataset of irrigated agriculture across the CONUS, enabling a binary classification of each pixel as rainfed or irrigated (where total irrigation exceeds zero). This classification allowed us to assess the influence of irrigation on our NDVI forecasting model systematically.

For integrating information from grid cells regarding *RZSM*, NDVI, irrigation, and aridity index for each crop type, we followed this procedure: First, we screened NDVI grid cells. VIIRS NDVI values often have many missing values during the non-growing season, especially at high latitudes. To retain as many relevant grid cells as possible, especially for crop types located in high latitudes, we limited our study period from Day of Year (DOY) 49 (DOY049) to DOY289, considering 16 Days of Year annually, with a requirement of at least two values from DOY049 to DOY081 and at least 10 values from DOY097 to DOY289 for the period of 2012 to 2021. Next, we checked *RZSM* grid cells for any missing values at the same locations, alongside irrigation information and the aridity index. Subsequently, based on the crop data and county boundaries, we retained grid cells within the targeted states as outlined in Supplementary Table S2. Detailed data processing flowchart can be found at Supplementary Figure S1 as well.

Figure S1 Data processing flowchart. Blue text indicates spatial resolution of the datasets in each box.

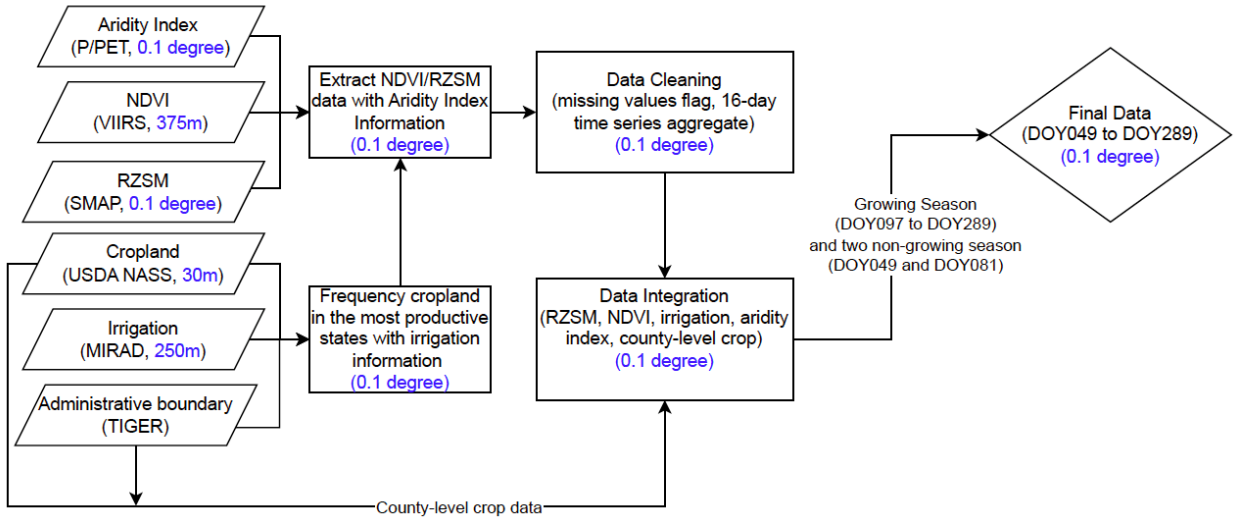
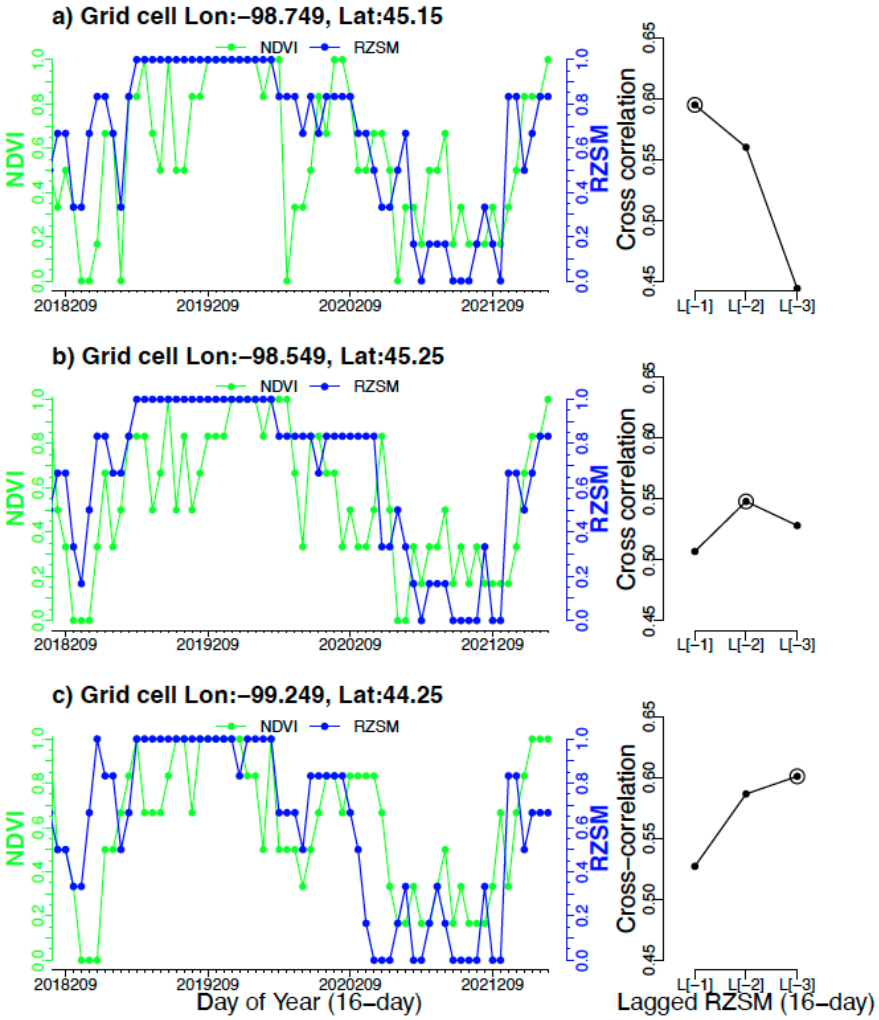


Figure S2 Illustration of Cross-Correlation between Standardized Rank NDVI and Standardized Rank RZSM



Standardized rank NDVI and standardized rank RZSM profiles at three corn grid cells, along with cross-correlation up to a lagged time of three (in 16-day intervals), were analyzed between standardized rank NDVI and standardized rank RZSM. For instance, the top-left panel shows a low rank in NDVI for the year 2021 and a similar rank for RZSM. The top-right panel displays the cross-correlation between lagged RZSM (L-1, L-2, and L-3) and NDVI(L), with the highest cross-correlation observed between RZSM (L-1) and NDVI(L), indicated by a large circle. Regarding the cross-correlation between lagged RZSM and NDVI at the middle right and bottom right, the highest correlations were observed between RZSM (L-2) and NDVI(L), and RZSM (L-3) and NDVI(L), respectively.

Figure S3 Spatial *NRMSE* assessment for *CLIM* and three forecasted DAPI lead times across all studied grid cells. The figure displays the difference in *NRMSE* between *DAPI* lead times and *CLIM*. Red color coding denotes better performance of forecasted *DAPI*.

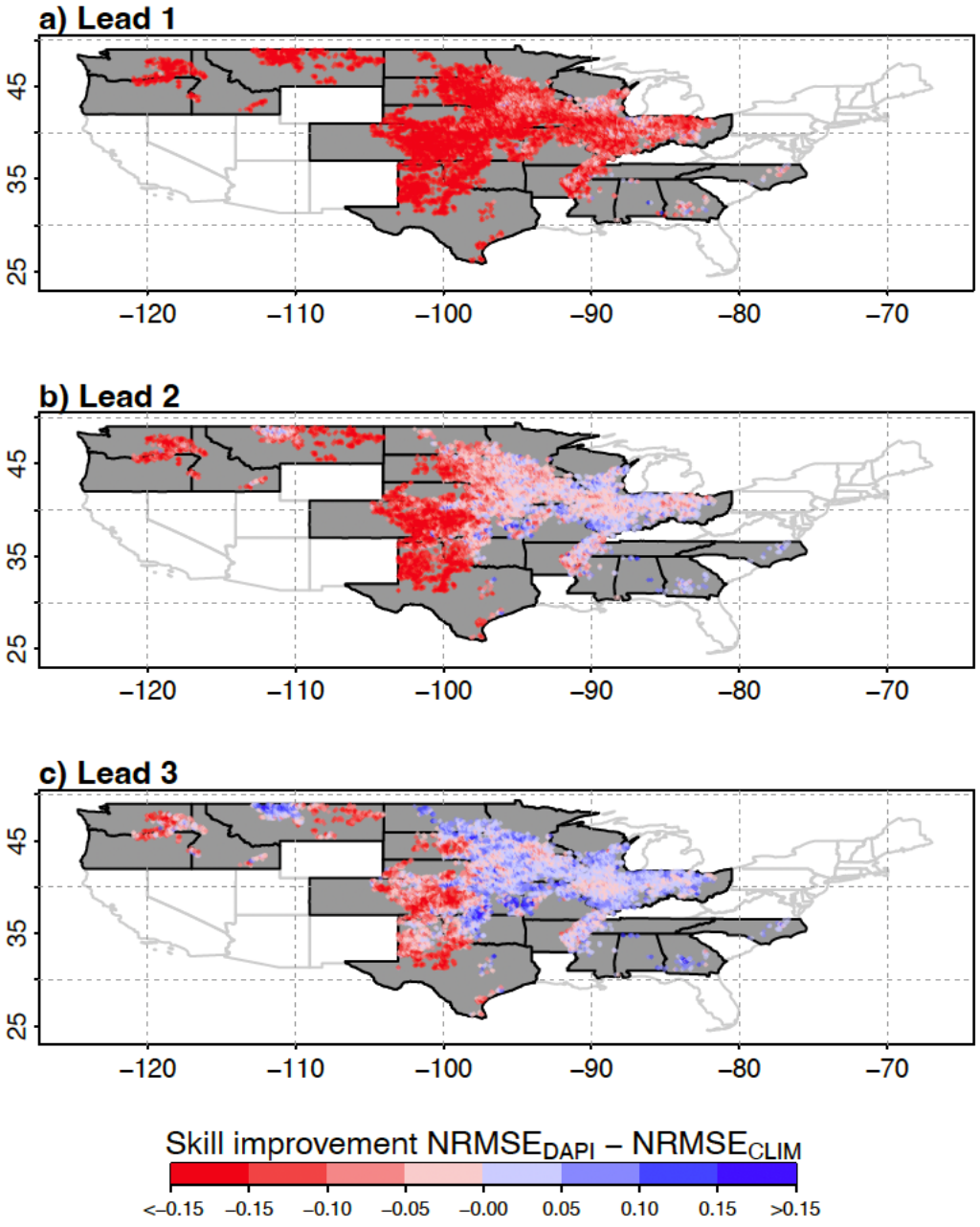


Figure S4 Split plot of maximum auto correlation of standardized rank NDVI for each crop type based on water-limited and energy-limited environments.

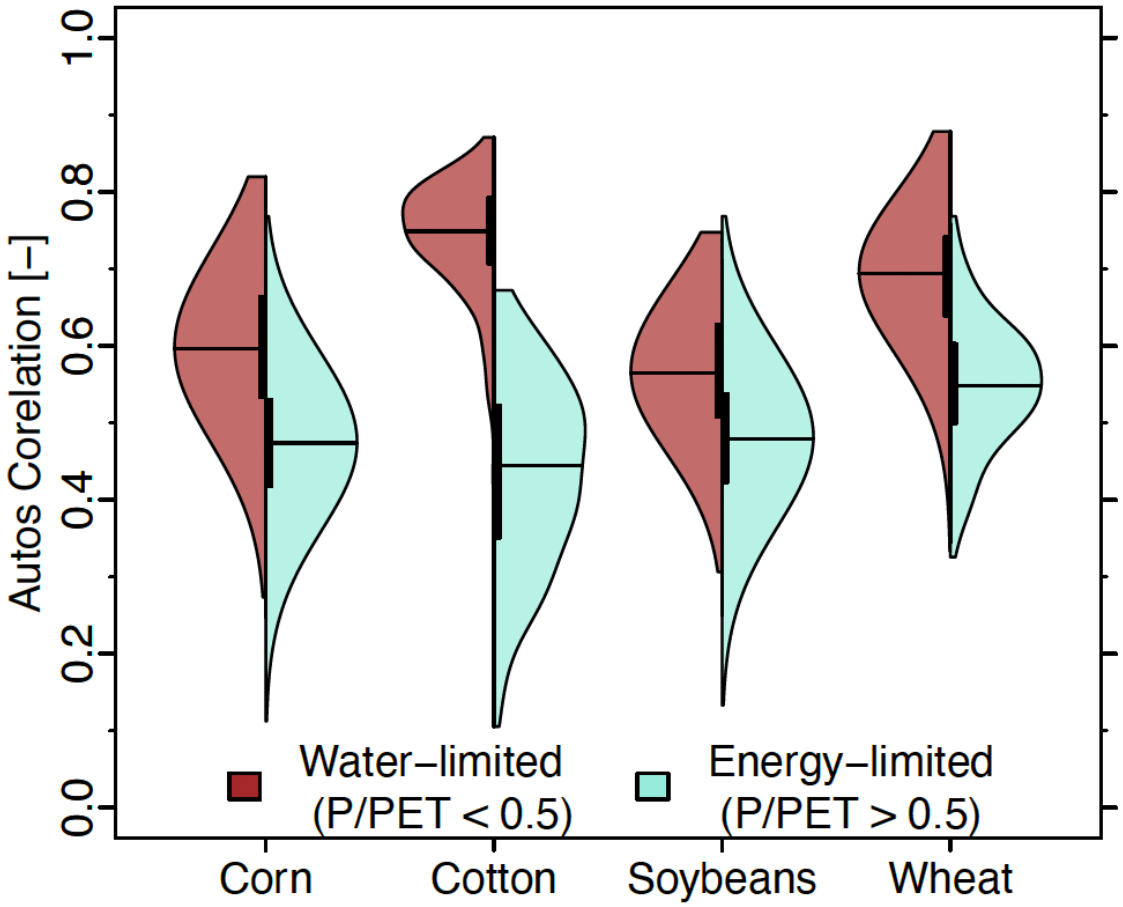


Figure S5 Intercomparison between *DAPI* with and without *SM* for three different lead times. Spatial distribution of grid cells classified as improvement (red color code, where the *NRMSE* of the DAPI forecasting model with *SM* is lower than the *NRMSE* of the *DAPI* forecasting model without *SM*). Non-improvement grid cells represent the opposite condition (blue color code). Value in parentheses indicates the percentage of grid cells over the total number of grid cells.

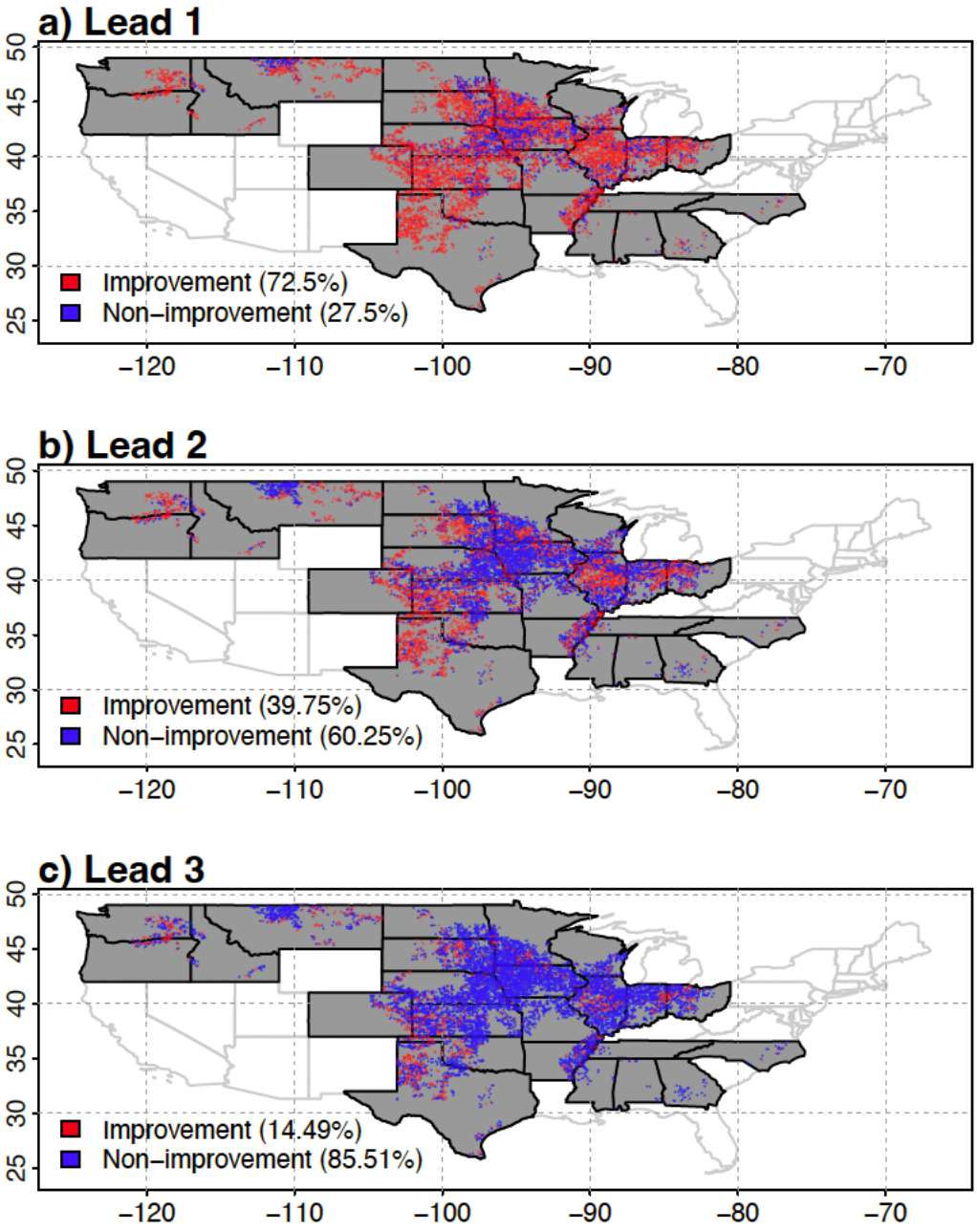


Figure S6 a) Distribution of single-crop grid cells (blue color) and multiple-crops grid cells (red color) and b) boxplot of the difference between *NRMSE* for multiple crops and *NMRSE* for single crop across different lead times.

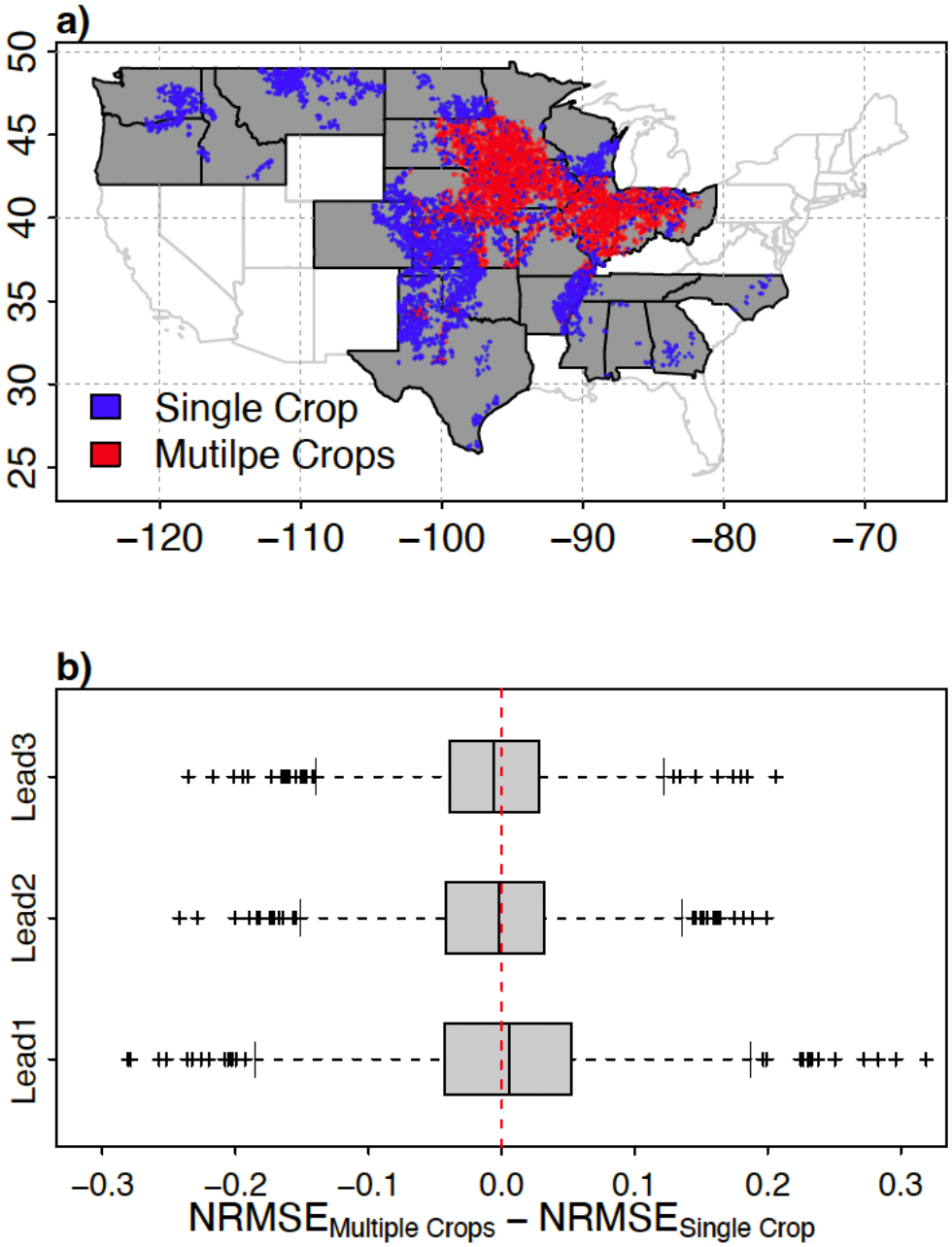


Figure S7 a) Distribution of rainfed grid cells (blue color) and irrigated grid cells (red color) and b) boxplot of the difference between *NRMSE* for multiple crops and *NMRSE* for single crop across different lead times.

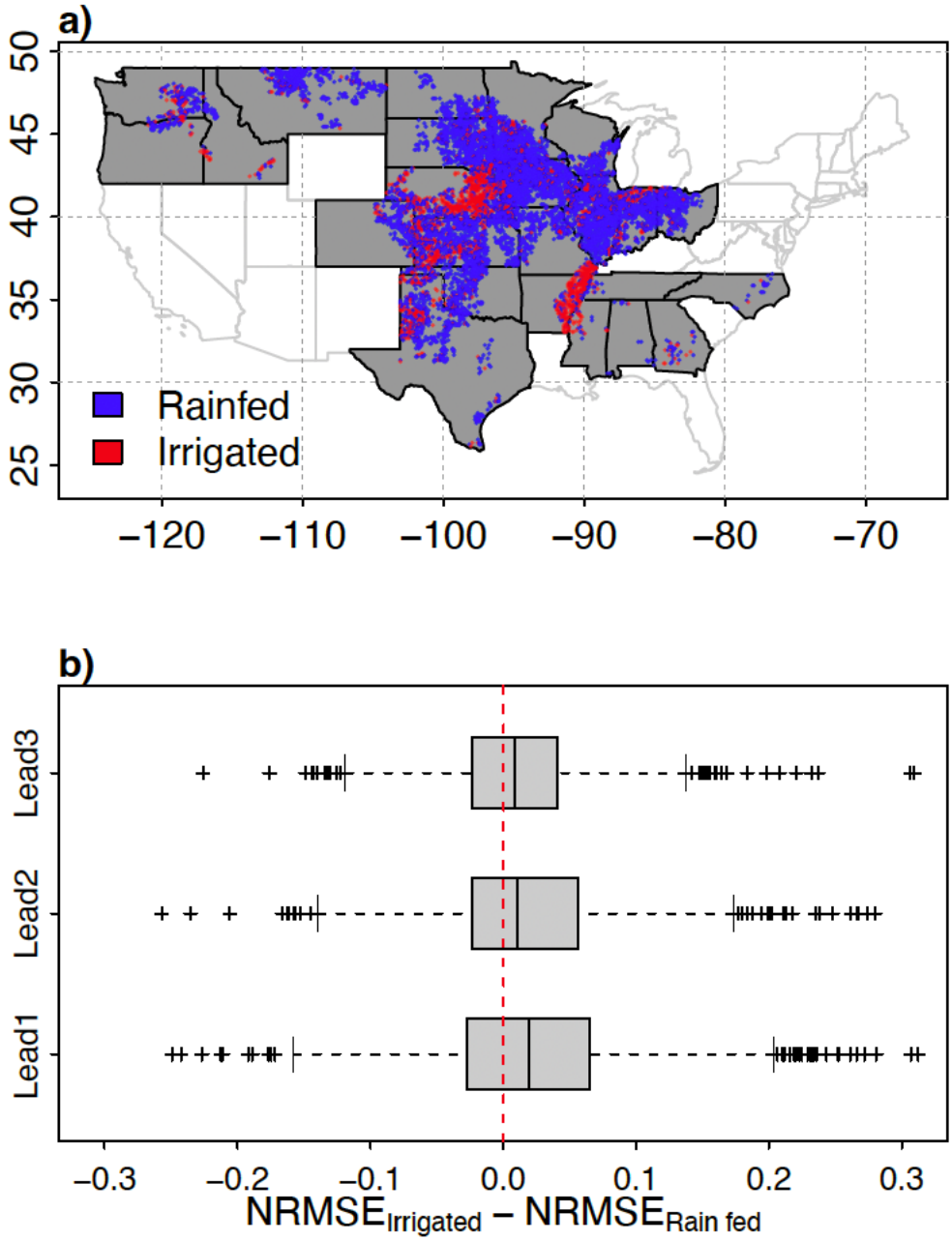


Table S1. Summary of datasets used in this study

Data Type	Data Description	Native Spatio/Temporal Resolution	Processed Spatio/ Temporal Resolution	Available Period	Used Period
Root Zone Soil Moisture	Palmer Water Model + SMAP DA	10km/ 1 day	10km/ 16 day	Apr 2015 –Aug 2022	Apr 2015-Dec 2021
NDVI	Visible Infrared Imaging Radiometer Suite (VIIRS)	375 m/ 8 day	10 km/ 16 day	Feb 2012-present	Feb 2012-Dec 2021
Cropland	USDA National Agricultural Statistics Service (NASS) Crop Data Layer (CDL)	30m/ Annual	10 km/ -	2008-2022	2008-2022
Crop yield	USDA NASS	-/ Annual		2003-2022	2016-2021
Administration	TIGER states and counties boundaries	-	-	-	-
Irrigation	Moderate Resolution Imaging Spectroradiometer (MODIS) Irrigated Agriculture Dataset for the United States (MIrAD-US)	250 m/ Annual	-	2007, 2012, 2017	2017
Aridity Index	Ratio between long term precipitation (P) and long-term potential evapotranspiration (PET)	10 km/ Long-term	Annual	Monthly long term	Average P and PET for growing season (April to November)

Table S2. Summary of total studied grid cells for each crop types

Crop Type	Total States	Total Counties	Total Grid cells	Grid cells per County	State Names
Corn	10	522	5026	9.6	Iowa, Illinois, Nebraska, Minnesota, Indiana, South Dakota, Kansas, Ohio, Wisconsin, Missouri
Cotton	8	136	551	4.1	Texas, Georgia, Mississippi, Arkansas, Oklahoma, North Carolina, Alabama, Tennessee
Soybeans	11	540	5007	9.3	Iowa, Illinois, Nebraska, Minnesota, Indiana, South Dakota, Kansas, Ohio, Wisconsin, Missouri
Wheat	10	212	2471	11.7	Kansas, Washington, Oklahoma, Montana, Colorado, Texas, Idaho, Oregon, South Dakota, Nebraska

Table S3: Descriptive statistics of VIF for regression models for each crop type and forecasting lead time. Our regression models have two predictors—lagged NDVI anomalies and lagged RZSM anomalies, which are represented by ndviano and rzsmano, respectively. N represents the total sample (grid cells). p95, median, mean, and p05 refer to the 95th percentile, median value, mean value, and 5th percentile of VIF across grid cells.

Crop Types	Lead Time	Predictors	n	p95	median	mean	p05
corn	1	ndviano	5026	1.19	1.02	1.05	1.00
corn	1	rzsmano	5026	1.19	1.02	1.05	1.00
corn	2	ndviano	5026	1.19	1.02	1.05	1.00
corn	2	rzsmano	5026	1.19	1.02	1.05	1.00
corn	3	ndviano	5026	1.19	1.02	1.05	1.00
corn	3	rzsmano	5026	1.19	1.02	1.05	1.00
cotton	1	ndviano	551	1.48	1.18	1.20	1.01
cotton	1	rzsmano	551	1.48	1.18	1.20	1.01
cotton	2	ndviano	551	1.39	1.14	1.16	1.01
cotton	2	rzsmano	551	1.39	1.14	1.16	1.01
cotton	3	ndviano	551	1.39	1.13	1.16	1.01
cotton	3	rzsmano	551	1.39	1.13	1.16	1.01
soybeans	1	ndviano	5007	1.19	1.02	1.05	1.00
soybeans	1	rzsmano	5007	1.19	1.02	1.05	1.00
soybeans	2	ndviano	5007	1.19	1.02	1.05	1.00
soybeans	2	rzsmano	5007	1.19	1.02	1.05	1.00
soybeans	3	ndviano	5007	1.19	1.02	1.05	1.00
soybeans	3	rzsmano	5007	1.19	1.02	1.05	1.00
wheat	1	ndviano	2471	1.40	1.13	1.15	1.01
wheat	1	rzsmano	2471	1.40	1.13	1.15	1.01
wheat	2	ndviano	2471	1.33	1.11	1.13	1.00
wheat	2	rzsmano	2471	1.33	1.11	1.13	1.00
wheat	3	ndviano	2471	1.32	1.11	1.13	1.00
wheat	3	rzsmano	2471	1.32	1.11	1.13	1.00

Table S4 Summaries of t-tests for the difference between $NRMSE_{DAPI}$ and $NRMSE_{CLIM}$. A significantly negative difference in NRMSE was observed for the first two lead times of cotton and wheat, as well as for lead time 1 of corn and soybeans. For lead time 3 of cotton and wheat, we also observed a slightly negative difference (small magnitude but statistically significant) NRMSE. On the other hand, we observed a slightly positive difference in NRMSE (but statistically significant) for corn and soybeans.

Lead times	Statistics	Corn	Cotton	Soybeans	Wheat
DAPI - Lead 1	t-stats	-110.38	-79.458	-105.38	-170.69
	95% confident interval	[-0.183, -0.177]	[-0.333, -0.317]	[-0.174, -0.168]	[-0.374, -0.366]
	p-value	2.20E-16	2.20E-16	2.20E-16	2.20E-16
DAPI - Lead 2	t-stats	-27.78	-0.17254902	-23.749	-81.406
	95% confident interval	[-0.042, -0.037]	[-0.168, -0.153]	[-0.036, -0.032]	[-0.176, -0.168]
	p-value	2.20E-16	2.20E-16	2.20E-16	2.20E-16
DAPI - Lead 3	t-stats	3.327	-18.649	9.3243	-38.29
	95% confident interval	[0.017, 0.066]	[-0.064, -0.052]	[0.009, 0.0014]	[-0.079, -0.071]
	p-value	0.0008	2.20E-16	2.20E-16	2.20E-16