A Data Mining Methodology for Detecting Forest Threats and Mapping Representativeness

Forrest M. Hoffman[†], Jitendra Kumar[†], Richard T. Mills[†], William W. Hargrove[‡], and Joseph P. Spruce*

[†]Oak Ridge National Laboratory; [‡]USDA Forest Service, Eastern Forest Environmental Threat Assessment Center; and *NASA Stennis Space Center

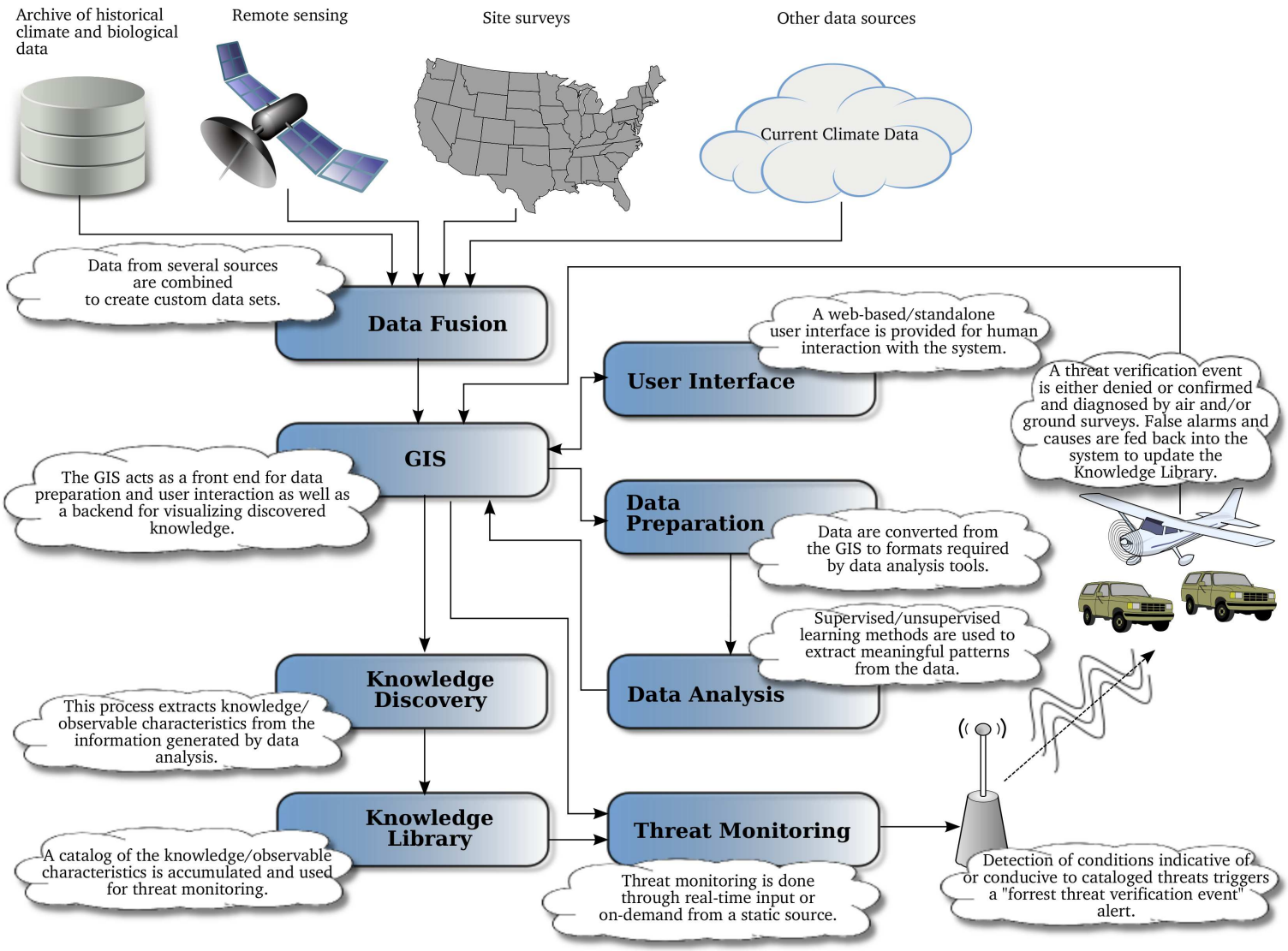
SAMSI/NCAR Workshop on Massive Datasets in Environment and Climate
National Center for Atmospheric Research (NCAR), Boulder, Colorado, USA
February 13–15, 2013

The USDA Forest Service, NASA Stennis Space Center, and DOE Oak Ridge National Laboratory are creating a system to monitor threats to U.S. forests and wildlands at two different scales:

- Tier 1: Strategic — The *ForWarn System* that routinely monitors wide areas at coarser resolution, repeated frequently — a *change detection system* to produce alerts or warnings for particular locations may be of interest
- Tier 2: Tactical — Finer resolution airborne overflights and ground inspections of areas of potential interest — *Aerial Detection Survey (ADS)* monitoring to determine if such warnings become alarms

Tier 2 is largely in place, but Tier 1 is needed to optimally direct its labor-intensive efforts and discover new threats sooner.

Overview of the Forest Incidence Recognition and State Tracking (FIRST) System



Normalized Difference Vegetation Index (NDVI)

- NDVI exploits the strong differences in plant reflectance between red and near-infrared wavelengths to provide a measure of "greenness" from remote sensing measurements.

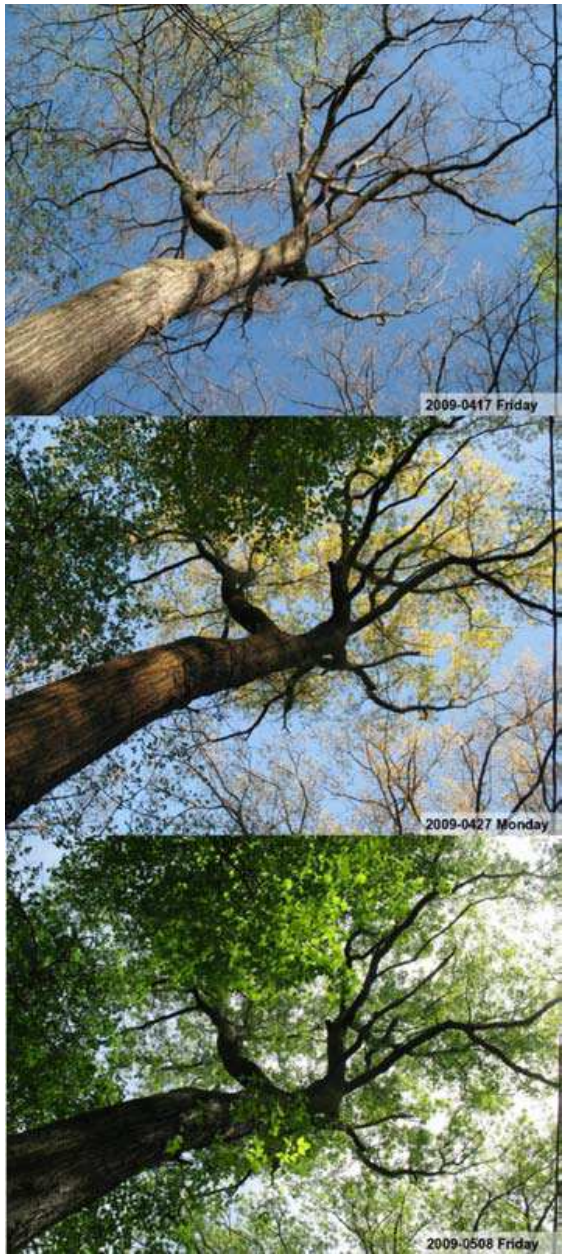
$$\text{NDVI} = \frac{(\sigma_{\text{nir}} - \sigma_{\text{red}})}{(\sigma_{\text{nir}} + \sigma_{\text{red}})} \tag{1}$$

- These spectral reflectances are ratios of reflected over incoming radiation, $\sigma = I_r/I_i$, hence they take on values between 0.0 and 1.0. As a result, NDVI varies between -1.0 and $+1.0$.
- Dense vegetation cover is 0.3–0.8, soils are about 0.1–0.2, surface water is near 0.0, and clouds and snow are negative.

MODIS MOD13 NDVI Product

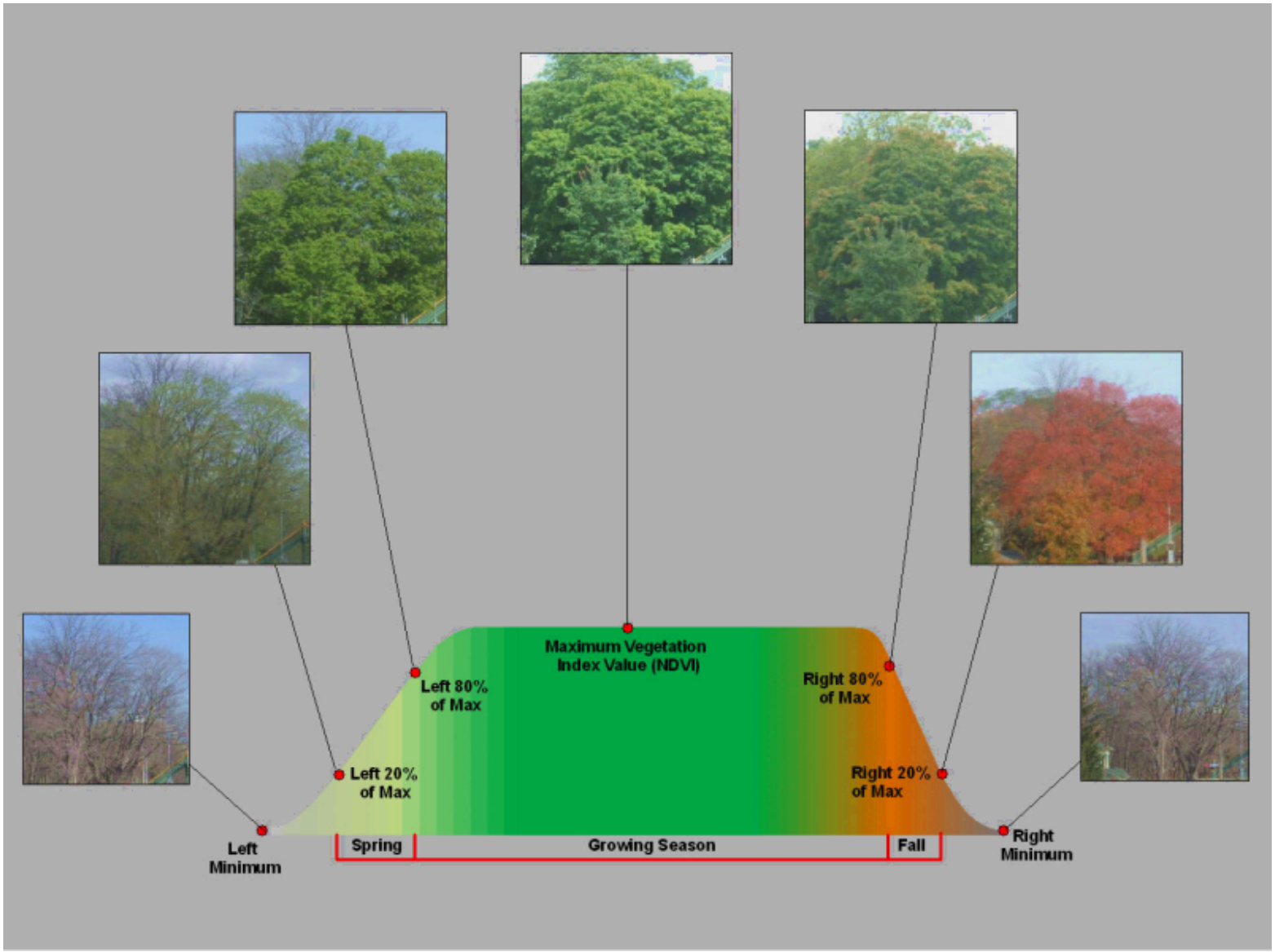
- The Moderate Resolution Imaging Spectroradiometer (MODIS) is a key instrument aboard the Terra (EOS AM, N→S) and Aqua (EOS PM, S→N) satellites.
- Both view the entire surface of Earth every 1 to 2 days, acquiring data in 36 spectral bands.
- The MOD 13 product provides Gridded Vegetation Indices (NDVI and EVI) to characterize vegetated surfaces.
- Available are 6 produces at varying spatial (231 m, 1 km, 0.05°) and temporal (16-day, monthly) resolutions.
- The Terra and Aqua products are staggered in time so that a new product is available every 8 days.
- Results shown here are derived from the 8-day Terra+Aqua MODIS product at 231 m resolution, processed by NASA Stennis Space Center.

- Phenology is the study of periodic plant and animal life cycle events and how these are influenced by seasonal and interannual variations in climate.
- FIRST is interested in deviations from the "normal" seasonal cycle of vegetation growth and senescence.
- NASA Stennis Space Center has developed a new set of National Phenology Datasets based on MODIS.
- Outlier/noise removal and temporal smoothing are performed, followed by curve-fitting and estimation of descriptive curve parameters.



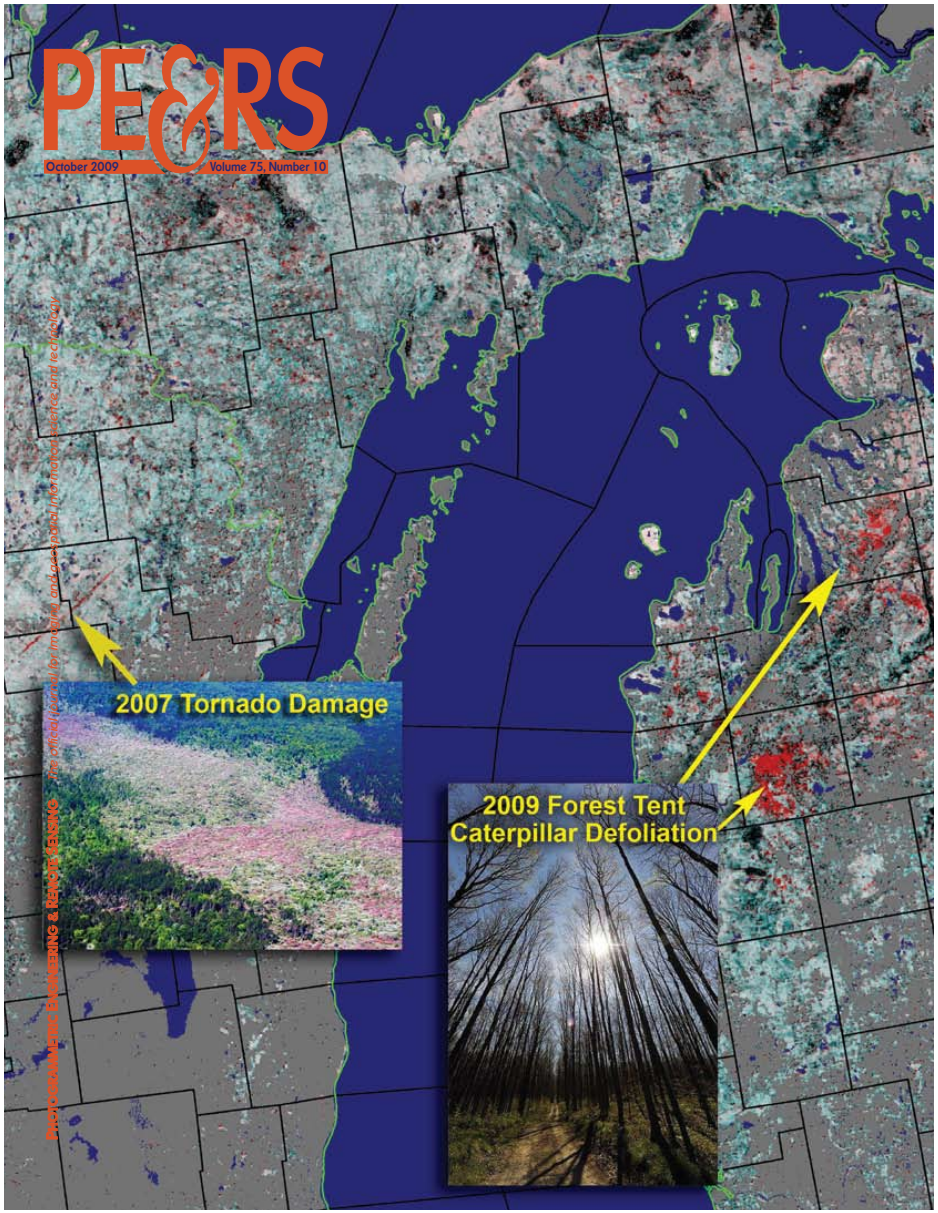
Up-looking photos of a scarlet oak showing the timing of leaf emergence in the spring (Hargrove et al., 2009).

Annual Greenness Profile Through Time

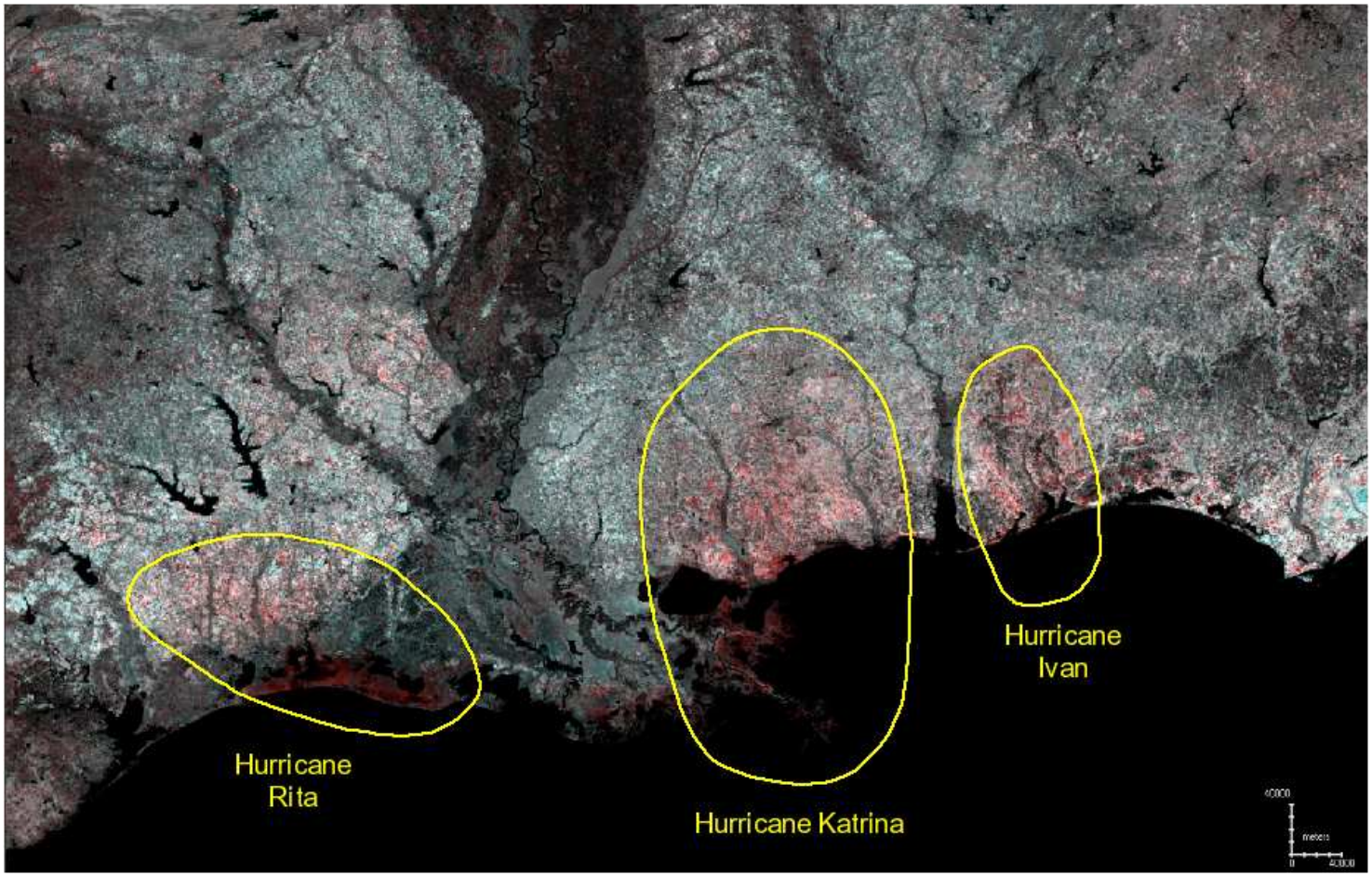


- To detect vegetation disturbances, the current NDVI measurement is compared with the normal, expected baseline for the same location.
- Substantial decreases from the baseline represent potential disturbances.
- Any increases over the baseline may represent vegetation recovery.
- Maximum, mean, or median NDVI may provide a suitable baseline value.

June 10–23, 2009, NDVI is loaded into blue and green; maximum NDVI from 2001–2006 is loaded into red (Hargrove et al., 2009).

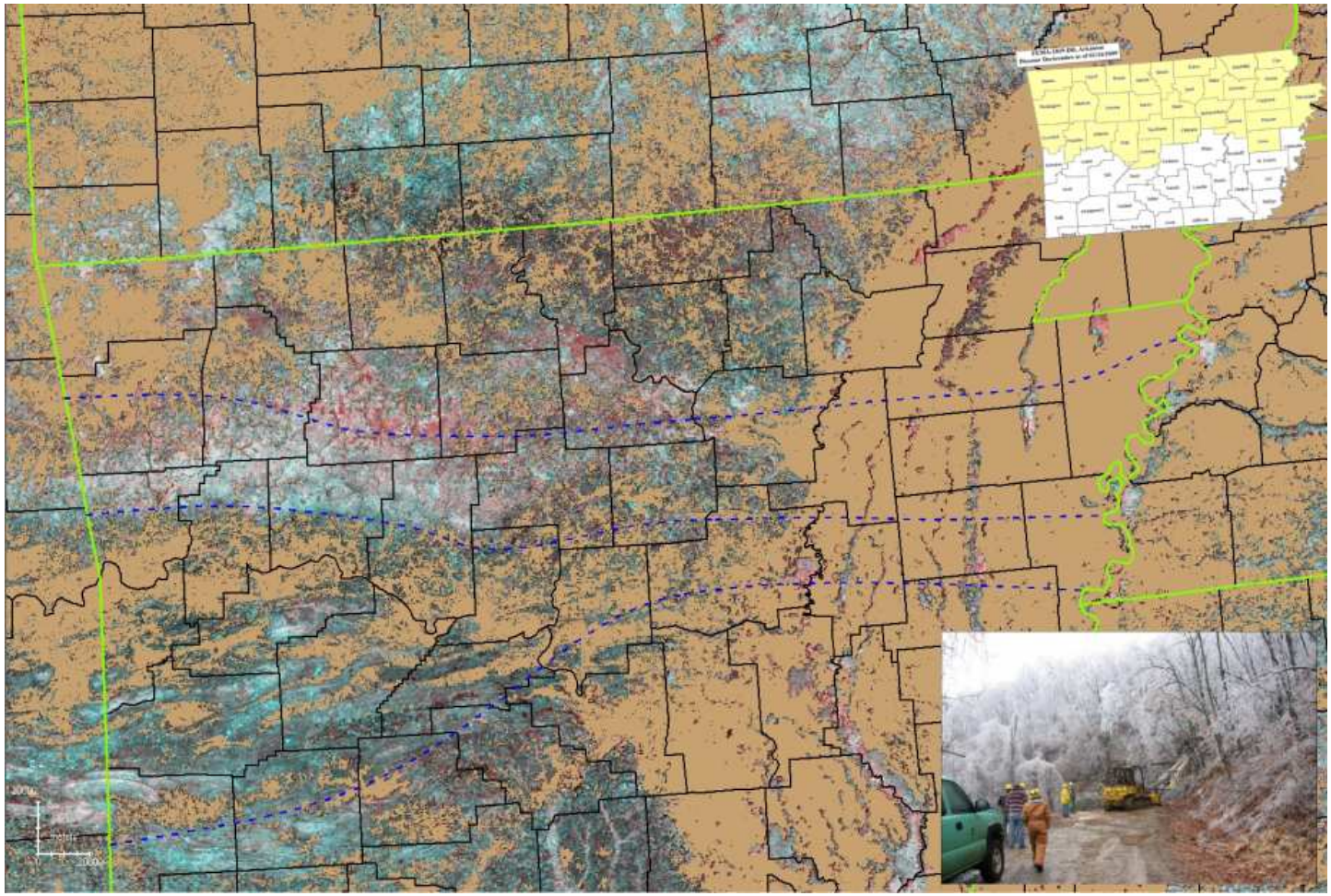


Three Hurricanes



Computed by assigning 2006 20% left value to green & blue, and 20% left from 2004 to red (Hargrove et al., 2009). Red depicts areas of reduced greenness, primarily east of storm tracks and in marshes.

Arkansas Ozarks Ice Storm, Jan. 26–29, 2009

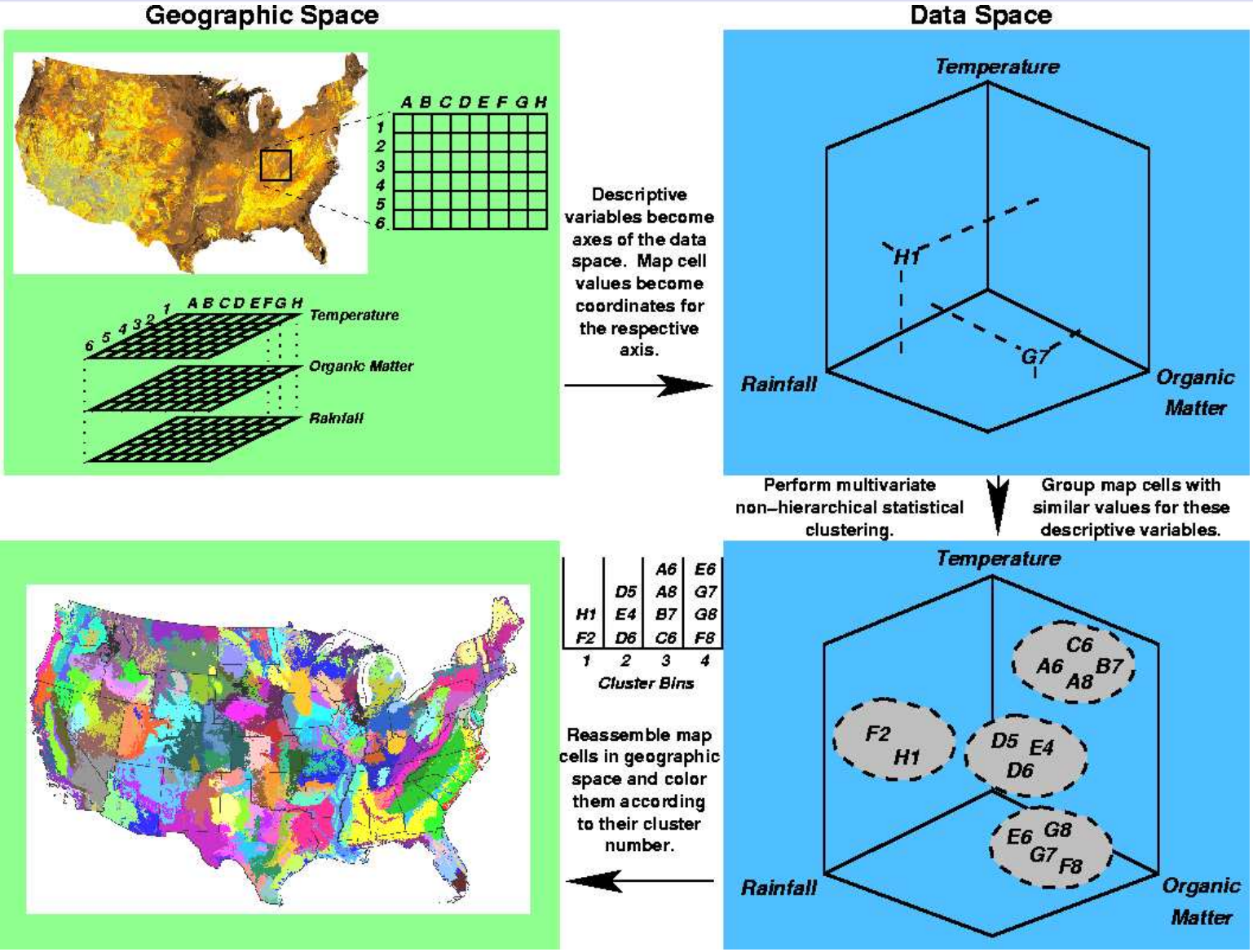


Computed by assigning 2009 max NDVI for June 10–July 15 into blue & green, and 2001–2006 max NDVI for June 10–July 27 into red. Storm resulted in 35,000 without power and 18 fatalities.

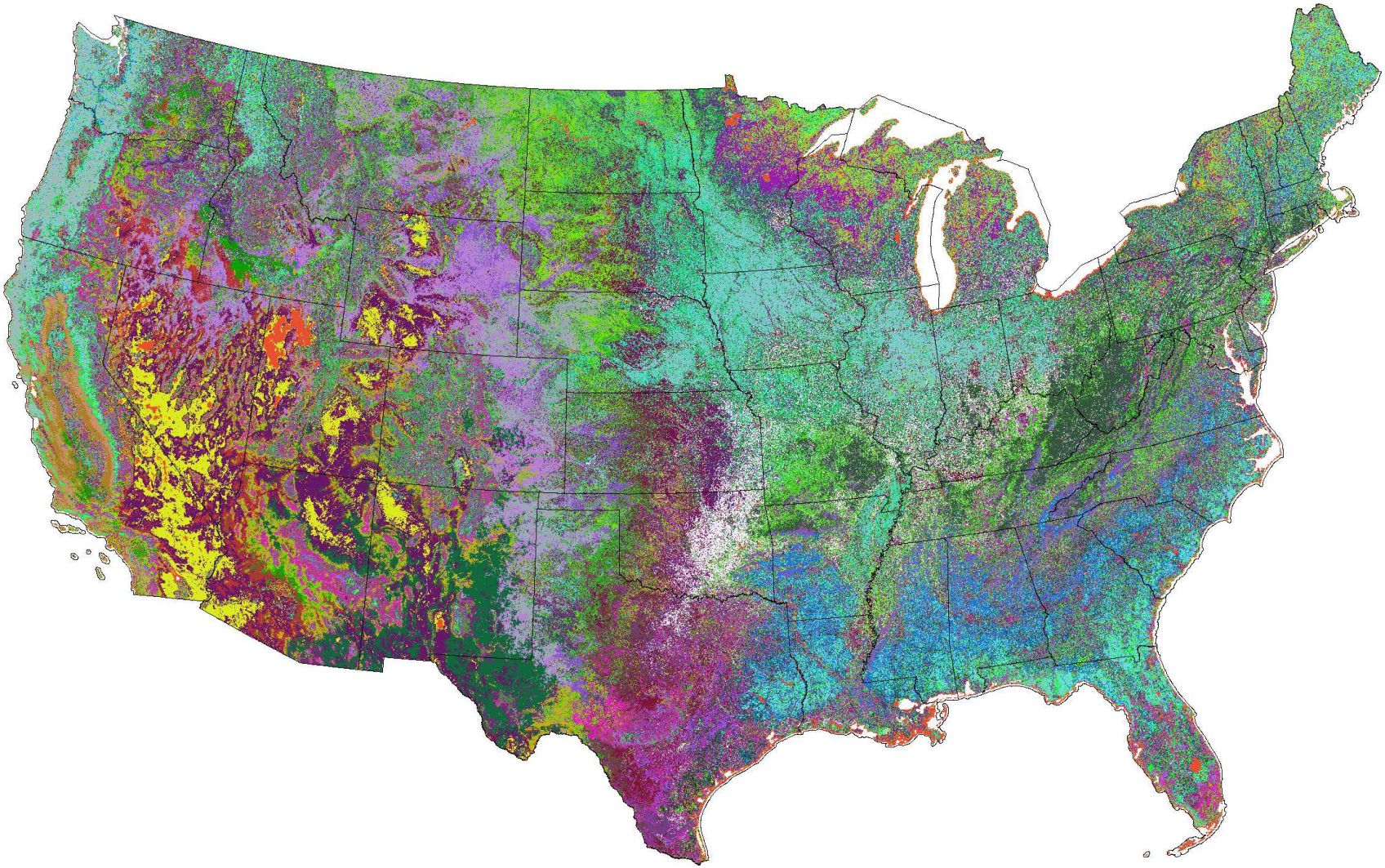
Data Mining for Change Detection

- Map arithmetic on selected parameters is good for studying the impact of known disturbances, but what is desired is an automated, unsupervised change detection system.
- A data mining approach, utilizing high performance computing (HPC) for the entire body of the very large, high resolution NDVI data history, appears to be the best approach.
- Hoffman and Hargrove previously employed a highly scalable k-means algorithm to automatically detect brine scars from hyperspectral remote sensing data (Hoffman, 2004) and for land surface phenology from monthly climatology and 17 years of 8 km NDVI from AVHRR (White et al., 2005).
- For only the current MODIS NDVI data for 11 years (2000–2012), 46 maps per year, at 231 m over the CONUS, single-precision data exceed 325 GB, requiring HPC resources.

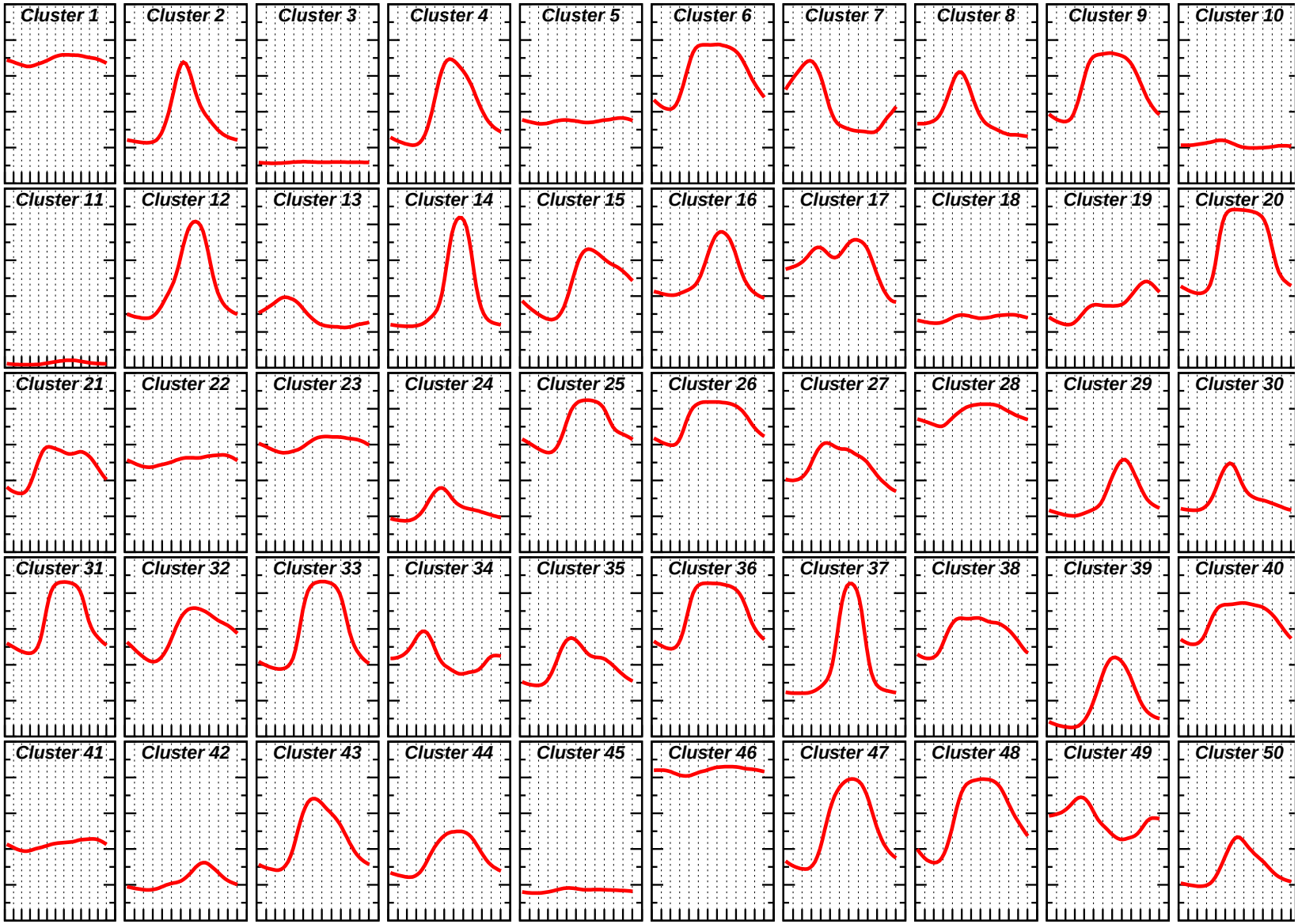
Geospatiotemporal Data Mining



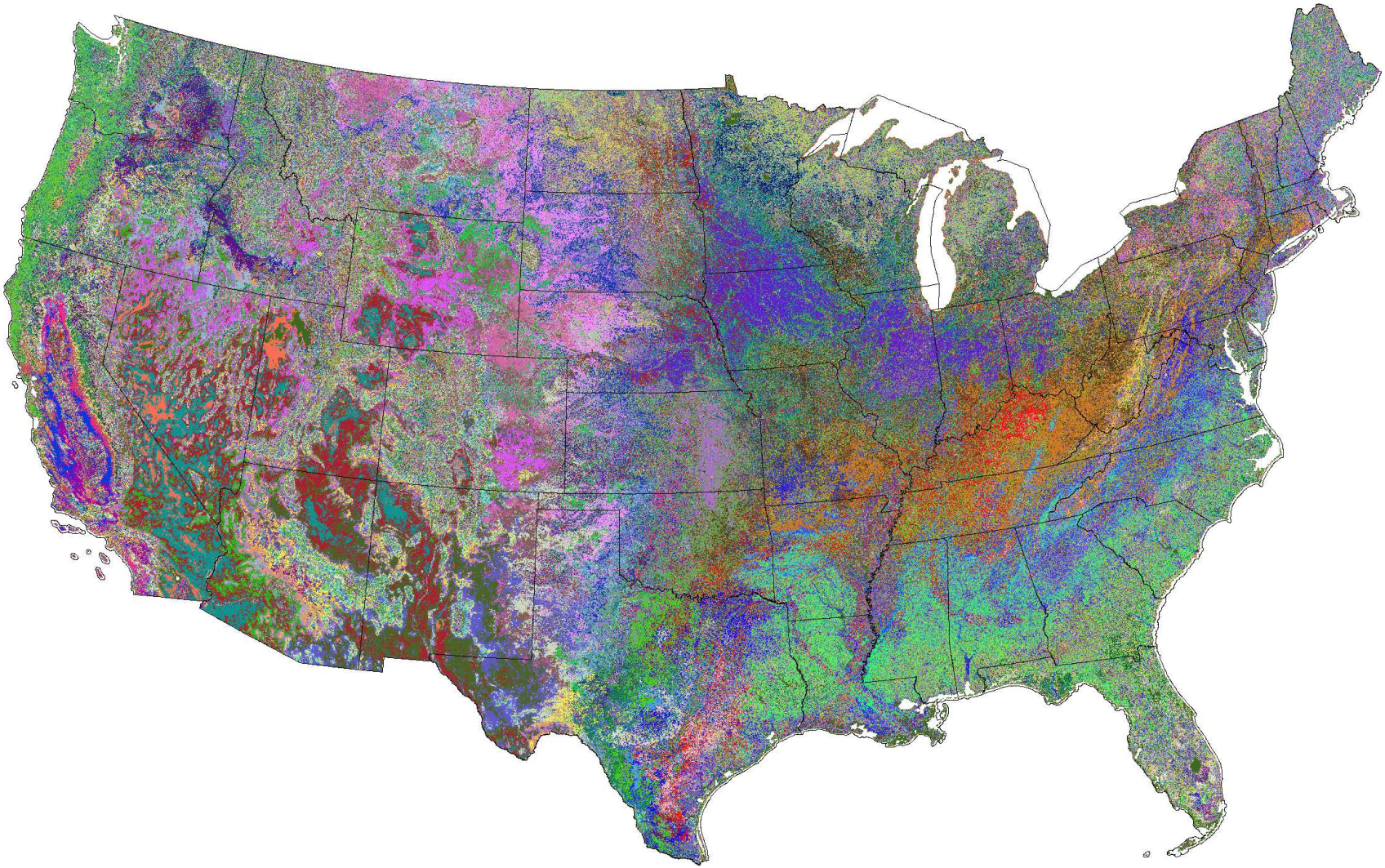
50 Phenoregions for Year 2010 (Random Colors)



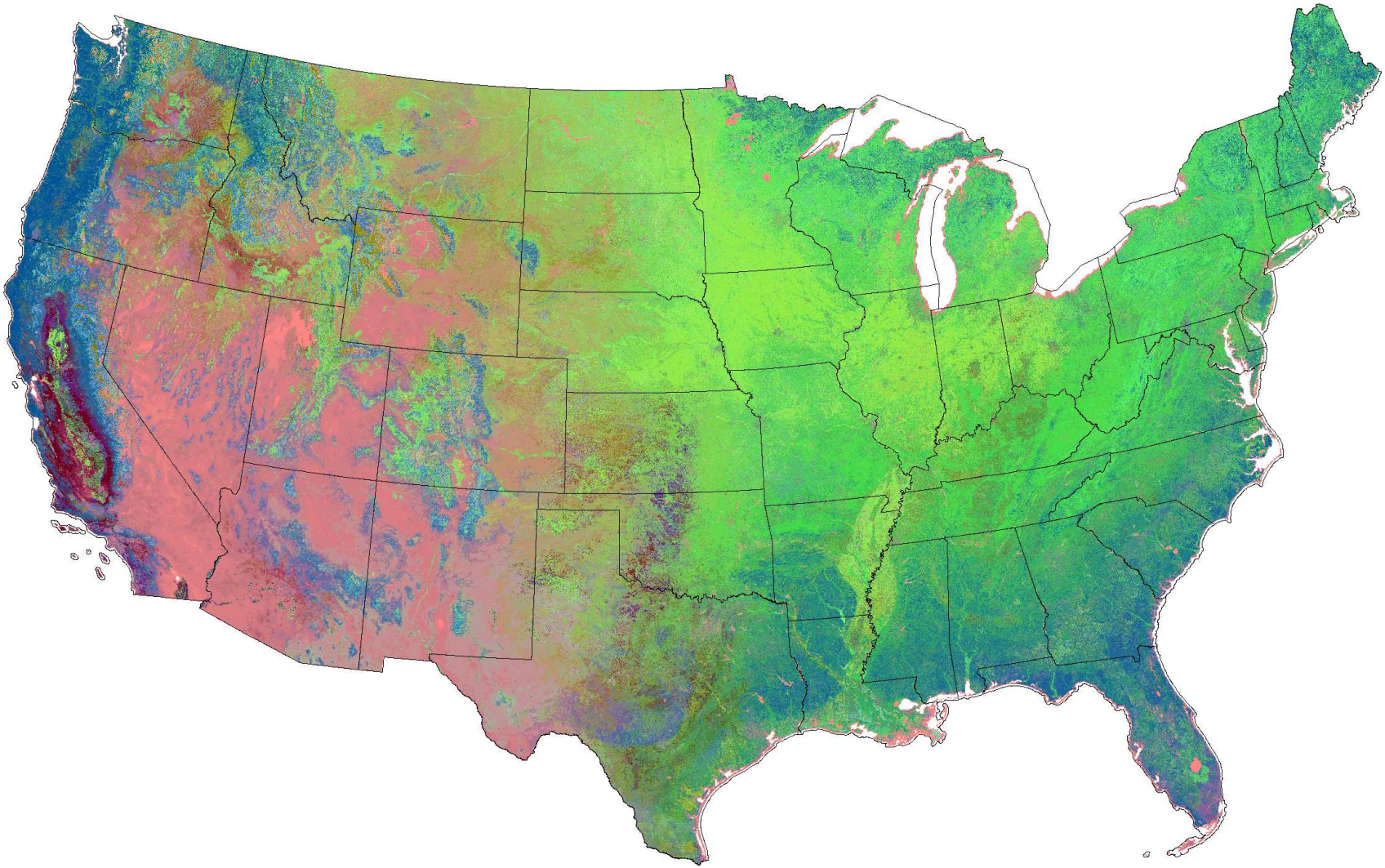
50 Phenoregion Prototypes



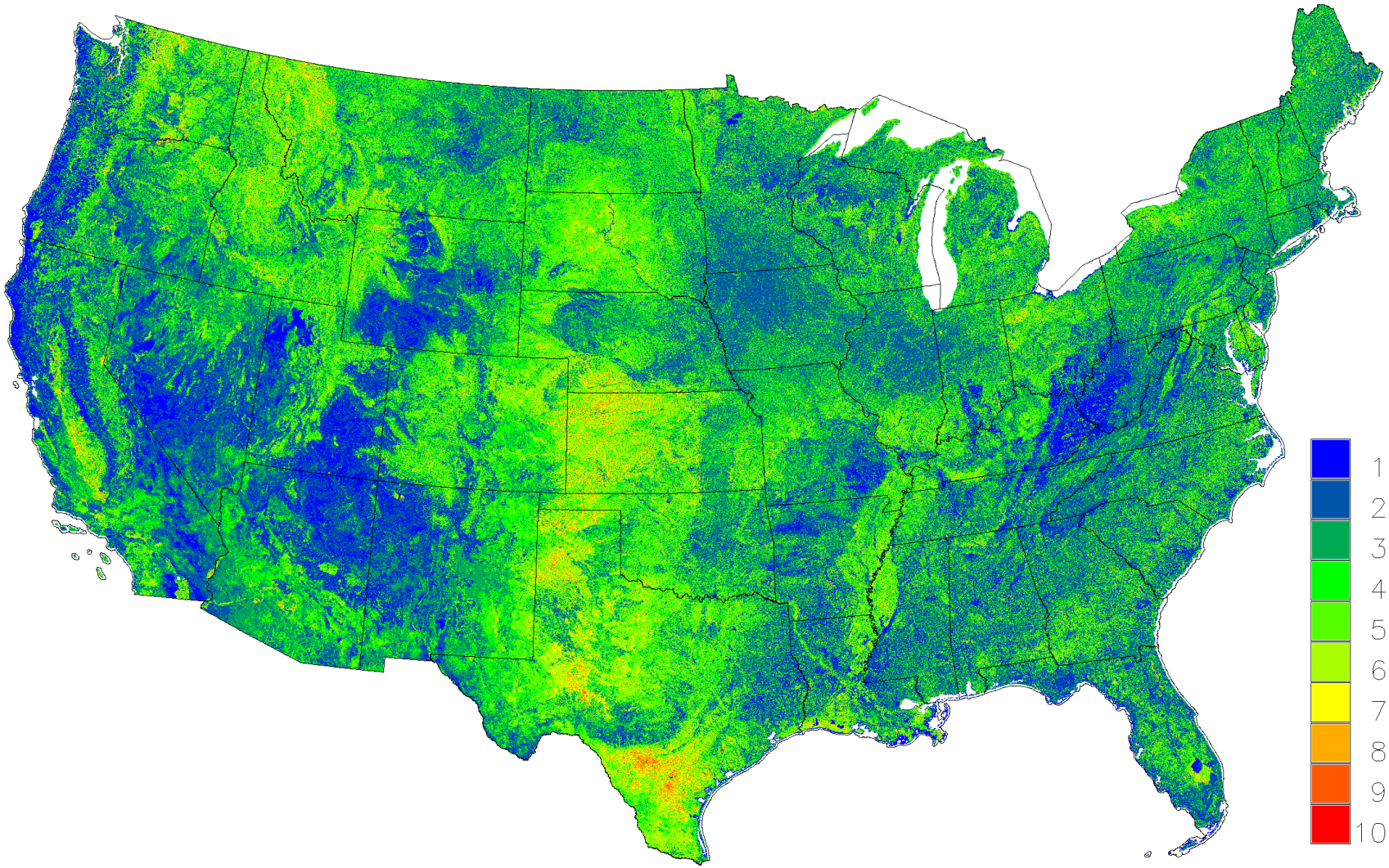
100 Phenoregions for Year 2010 (Random Colors)



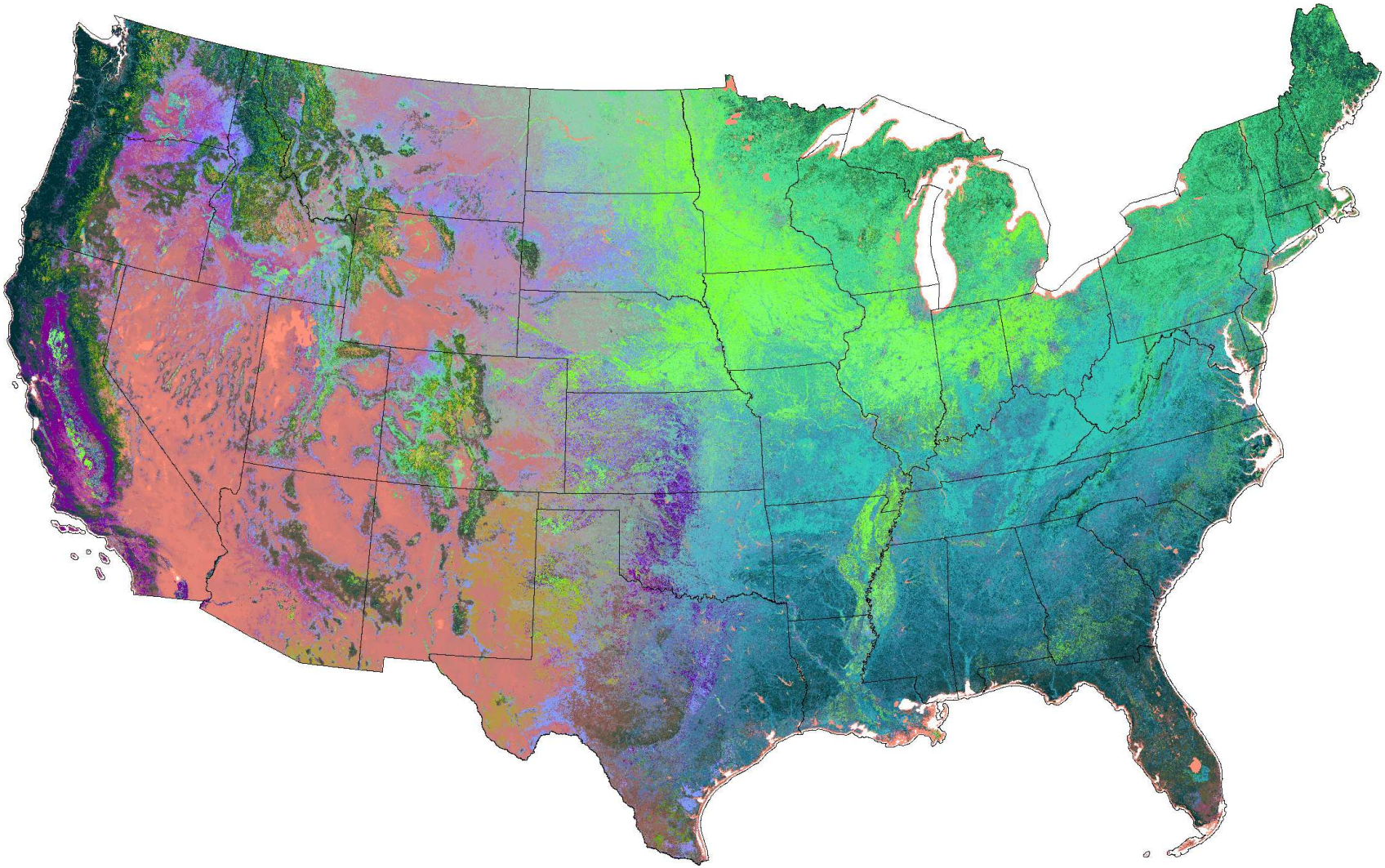
100 Phenoregions for Year 2010 (Similarity Colors)



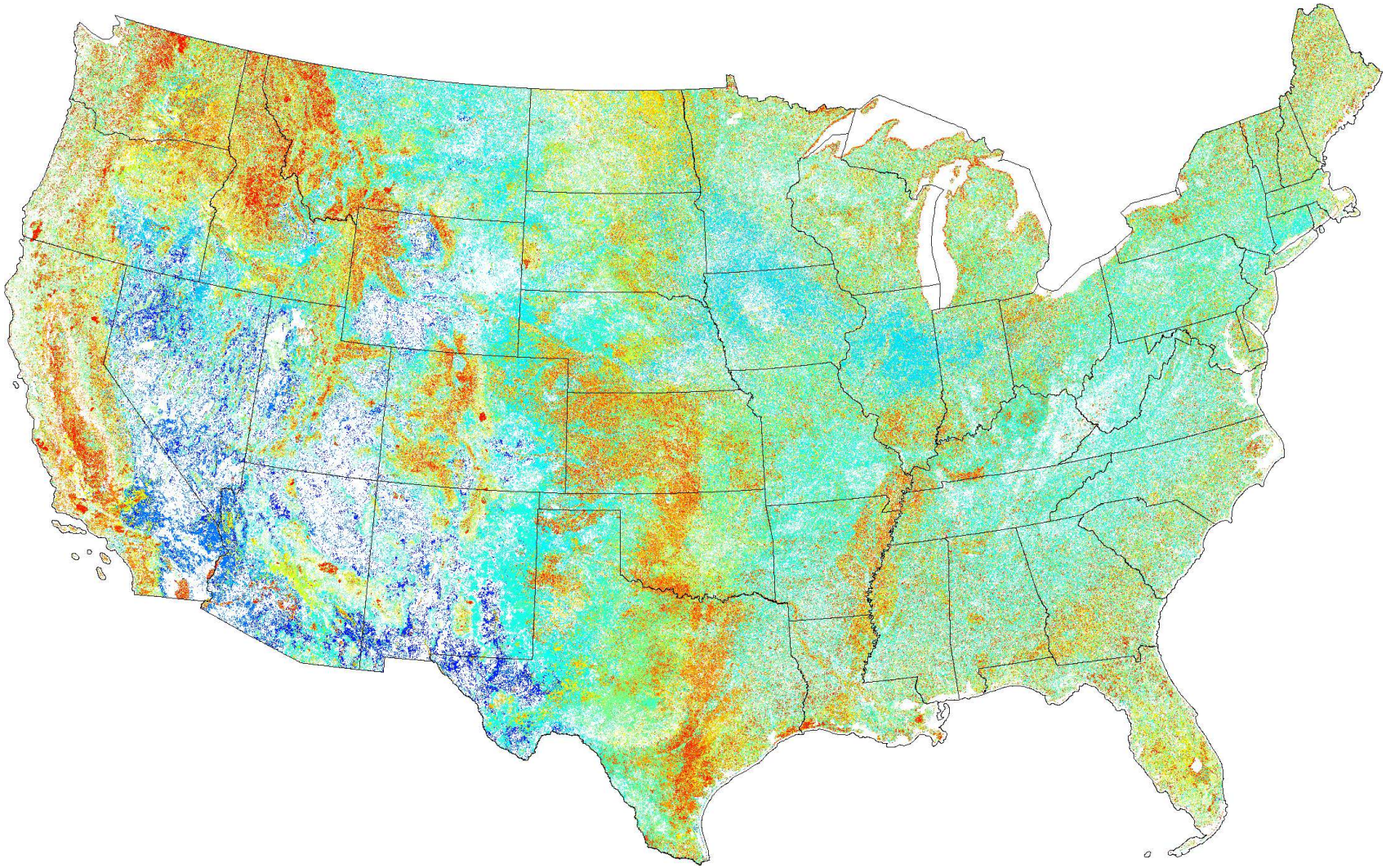
Cluster Persistence Map (2000–2009)



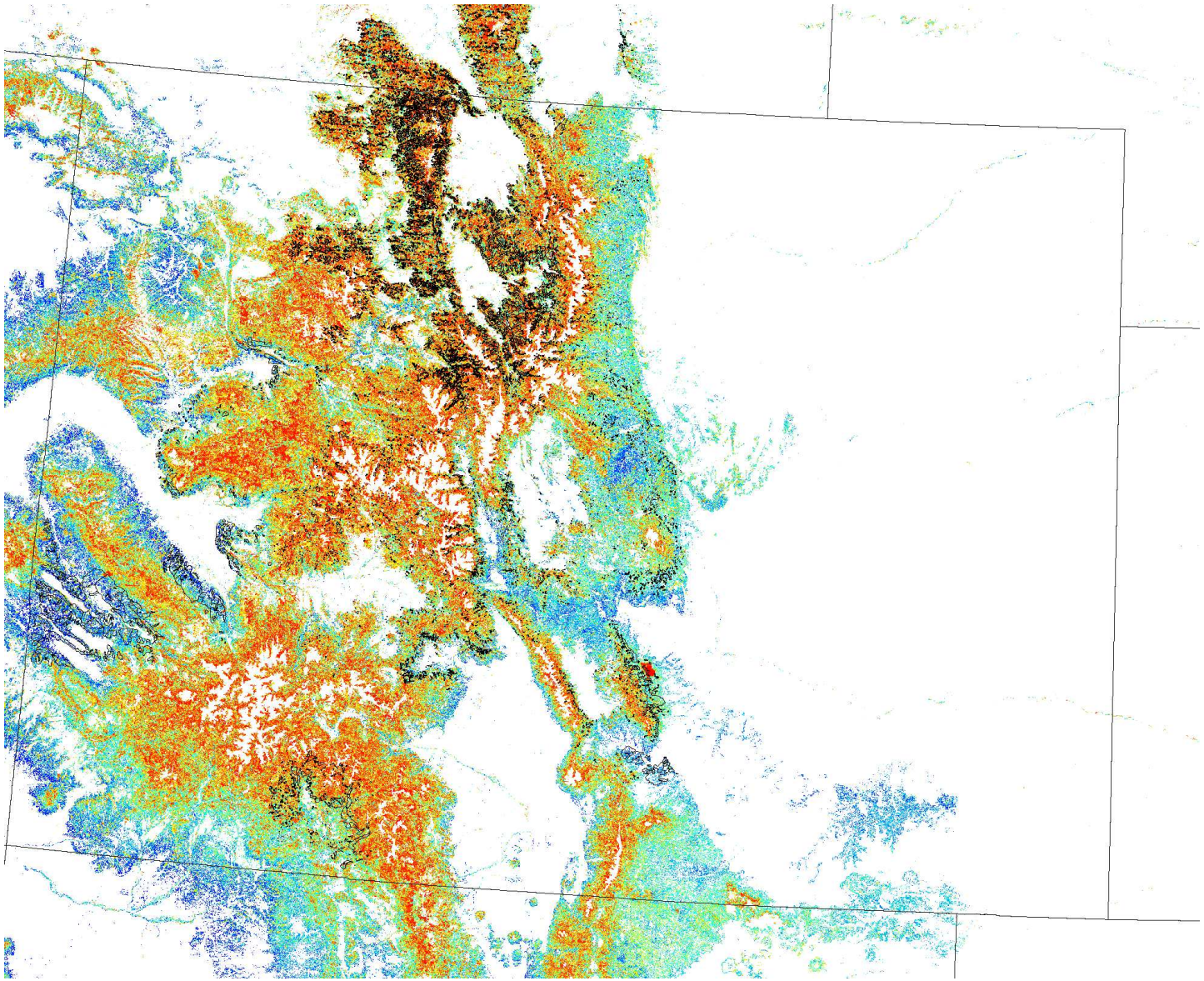
Cluster Mode Map (2000–2009)



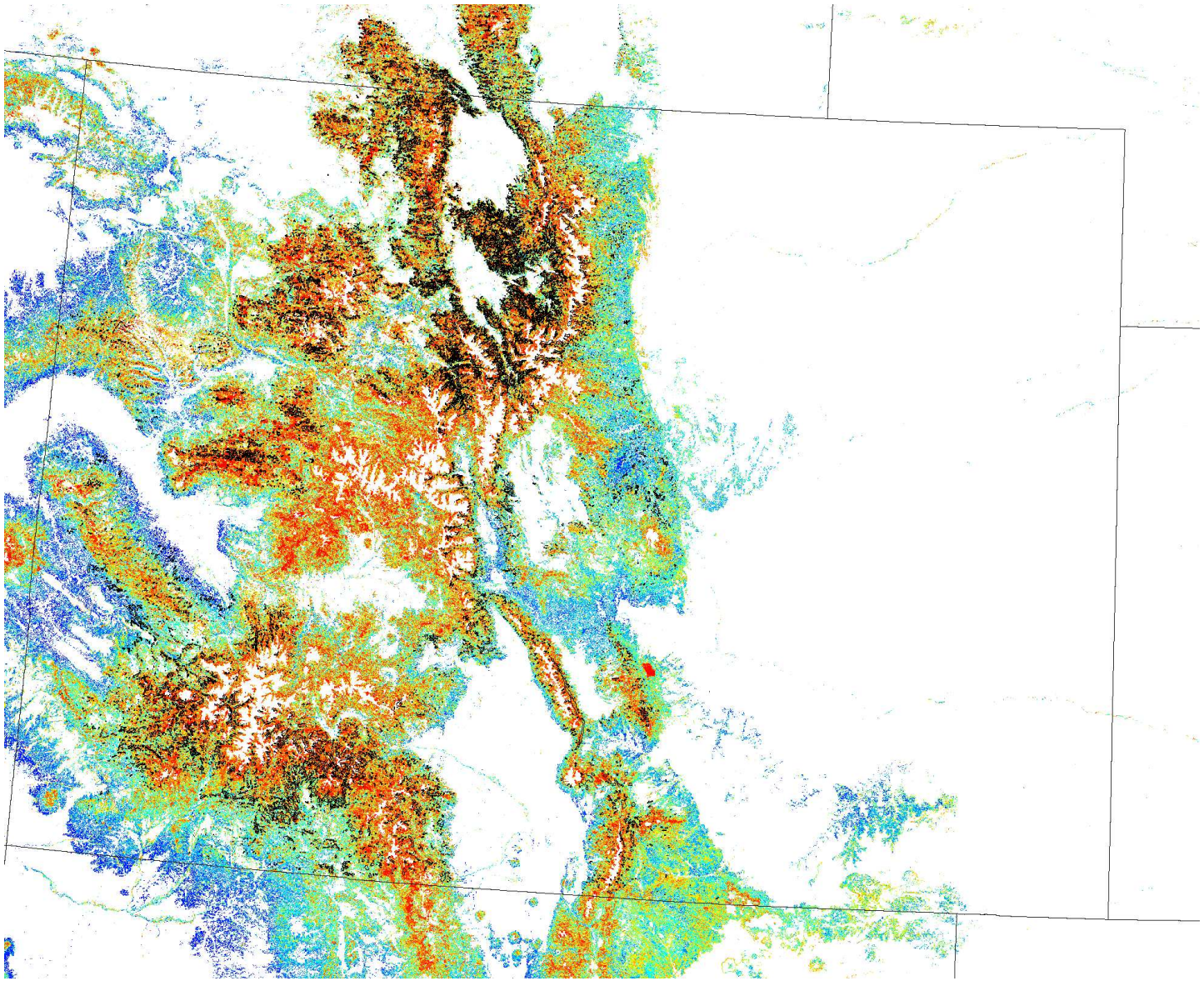
Cluster Transition Distances for 2009 – 2000 (2000–2009)



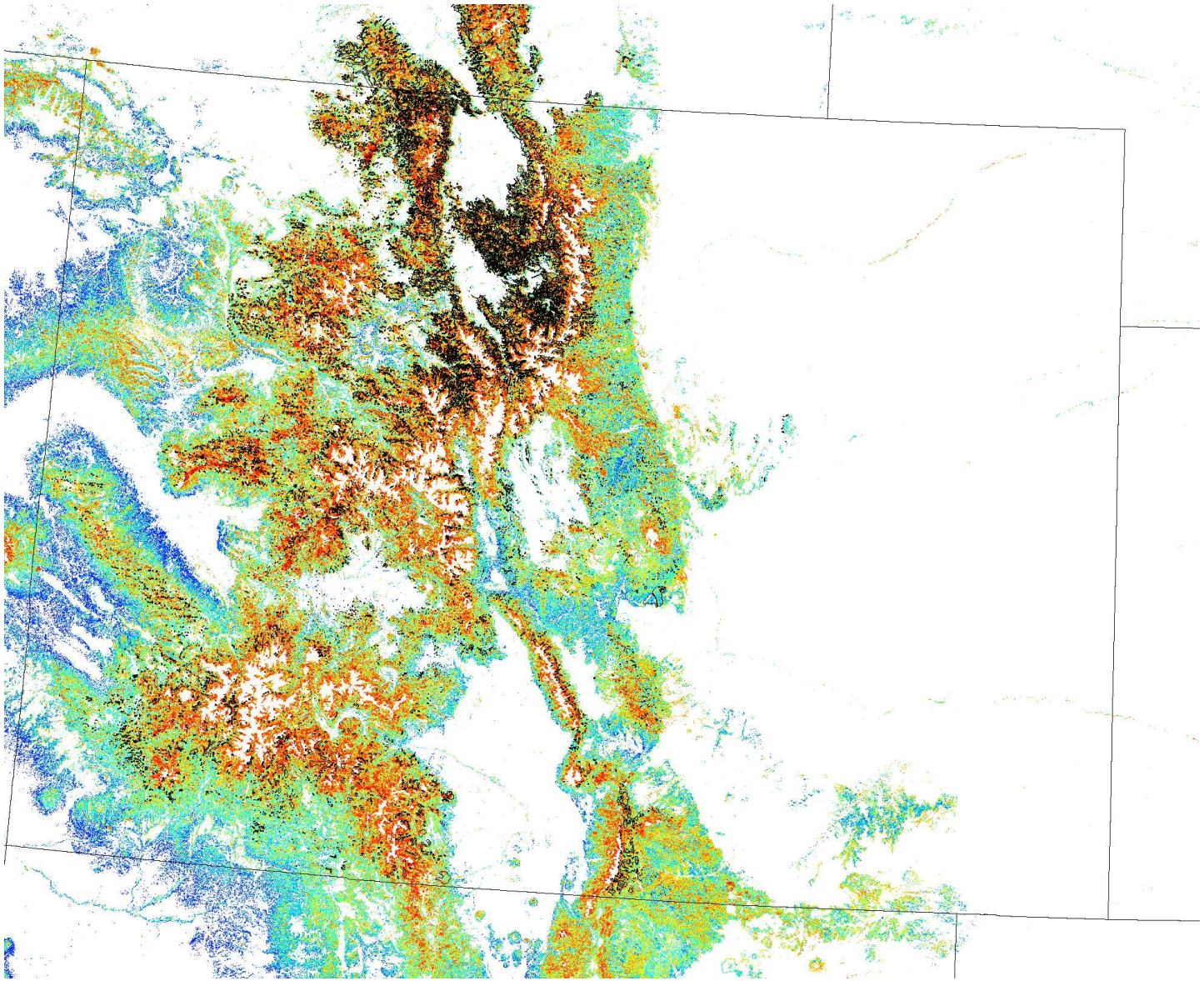
Mountain Pine Beetle in Colorado (2004)



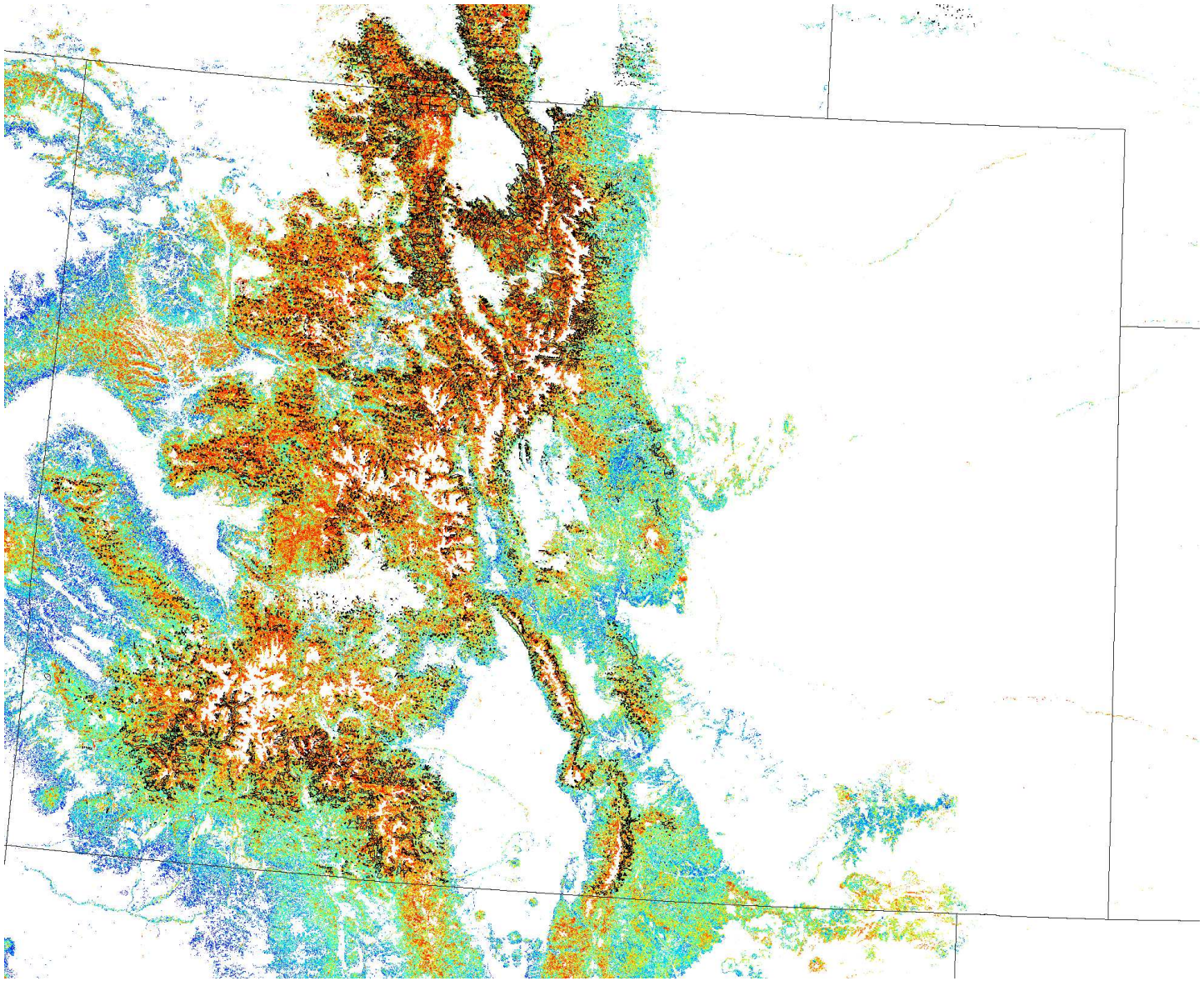
Mountain Pine Beetle in Colorado (2005)



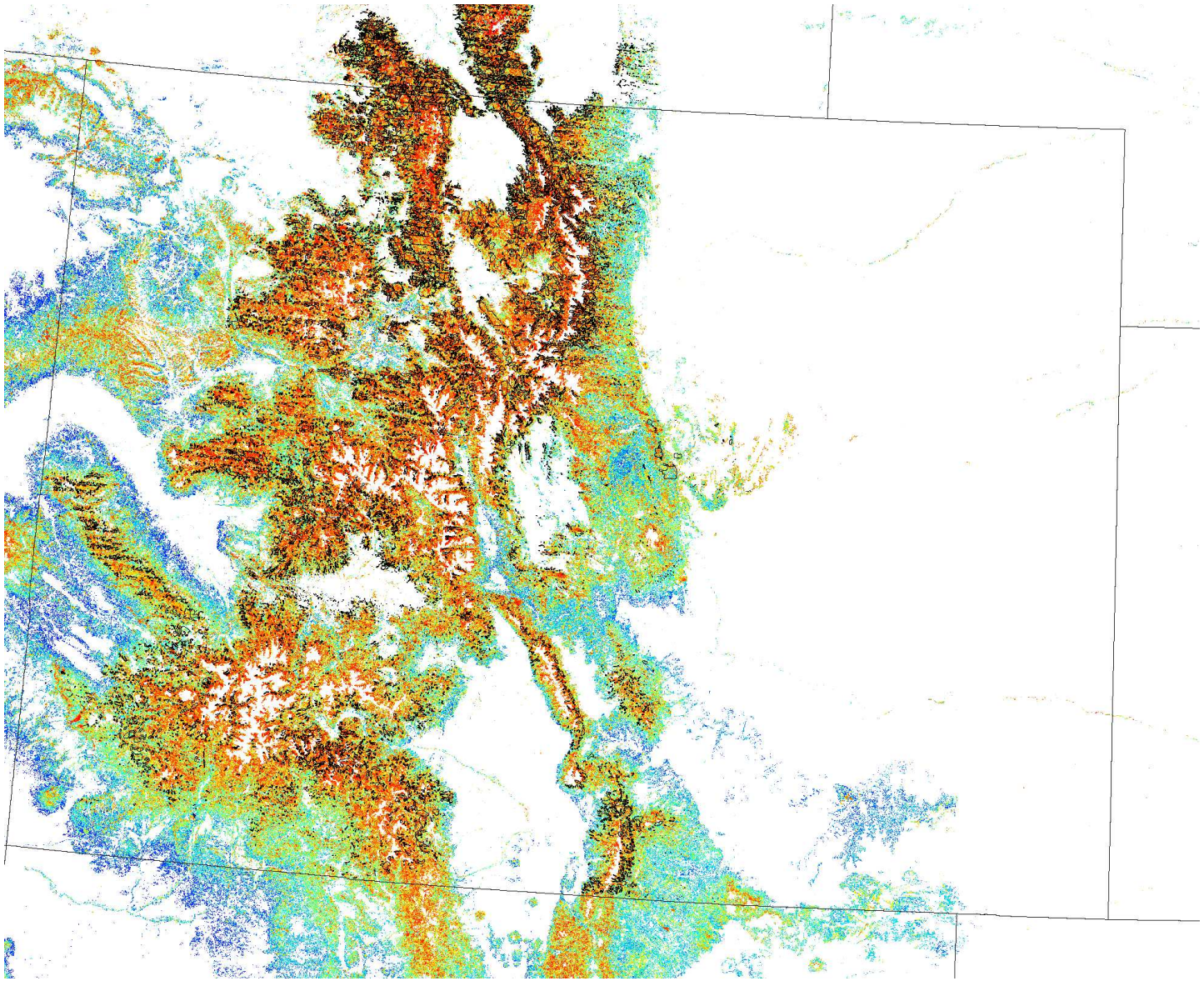
Mountain Pine Beetle in Colorado (2006)



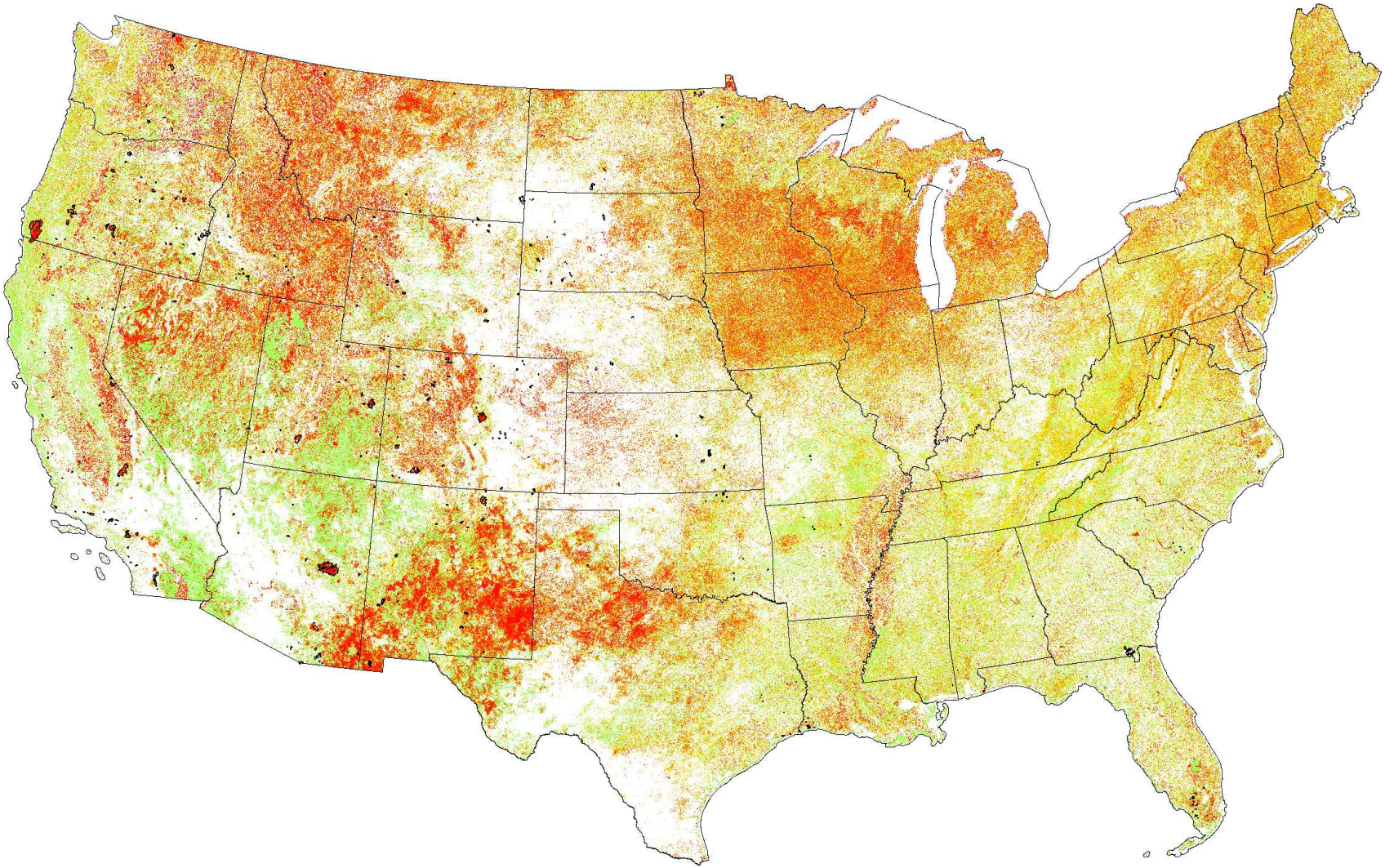
Mountain Pine Beetle in Colorado (2007)



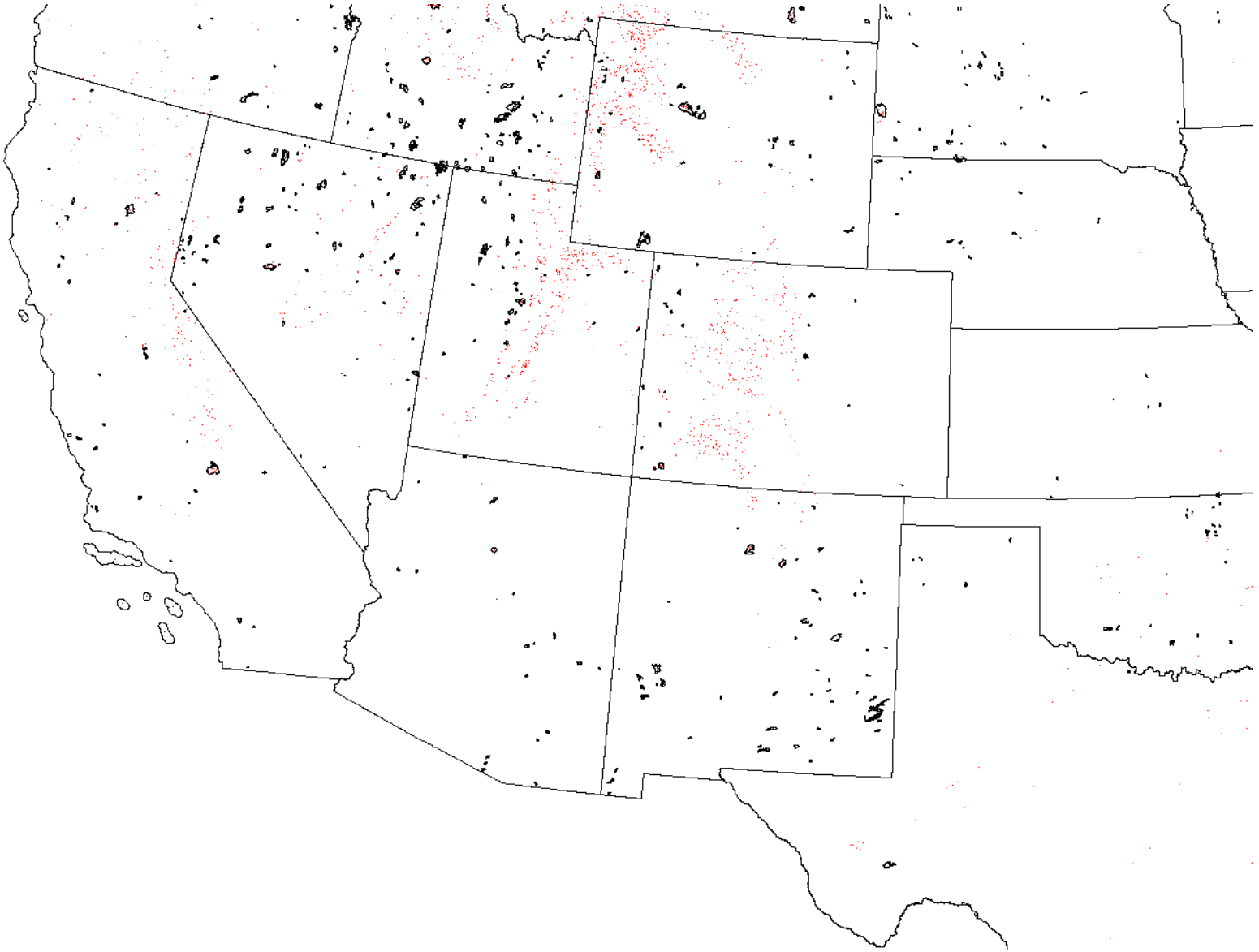
Mountain Pine Beetle in Colorado (2008)



ΔIntegrated NDVI for 2003 − 2002 (2000–2010, $k = 1000$)

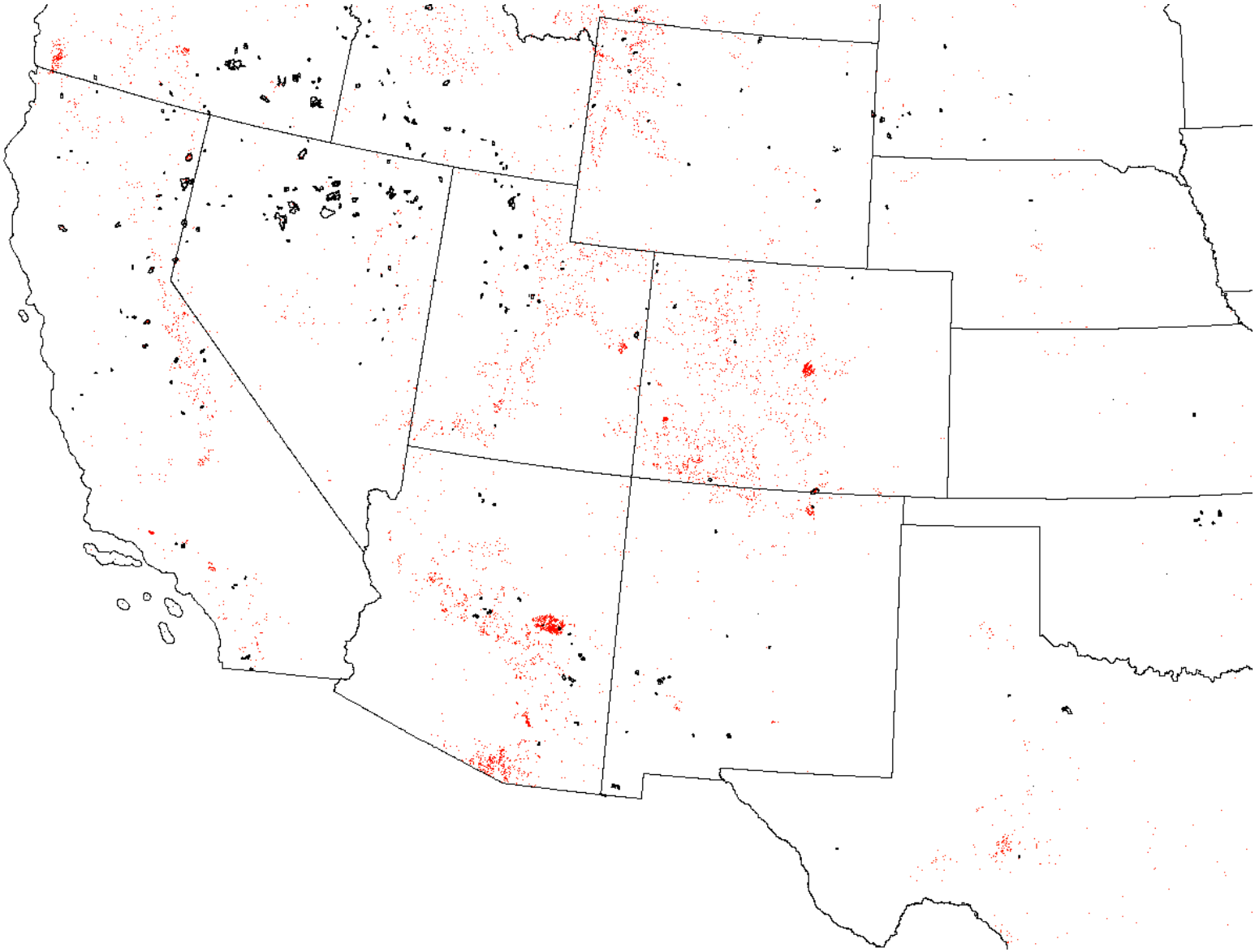


ΔIntegrated NDVI with Threshold for 2001 – 2000



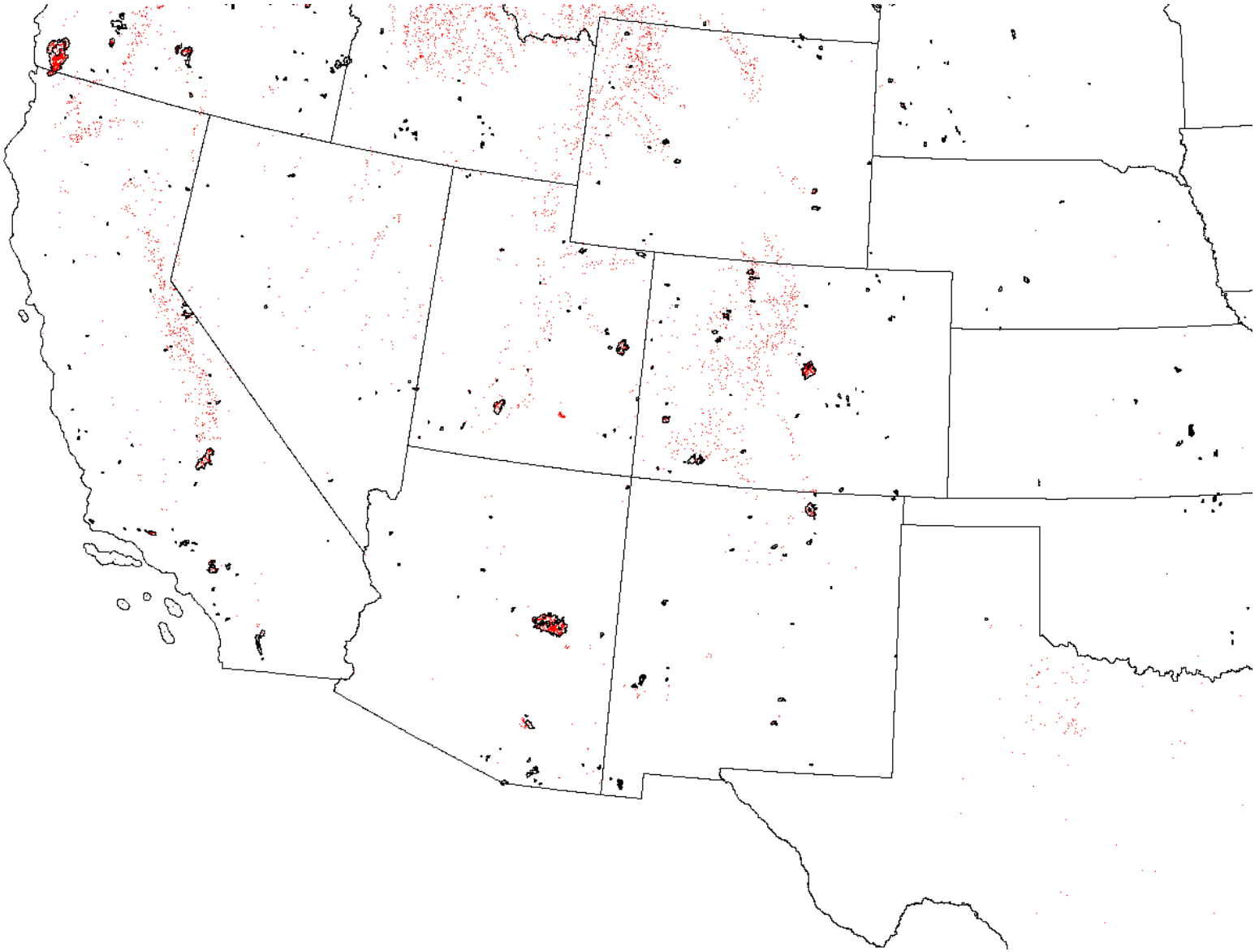
Kumar, et al., in prep.

ΔIntegrated NDVI with Threshold for 2002 – 2001



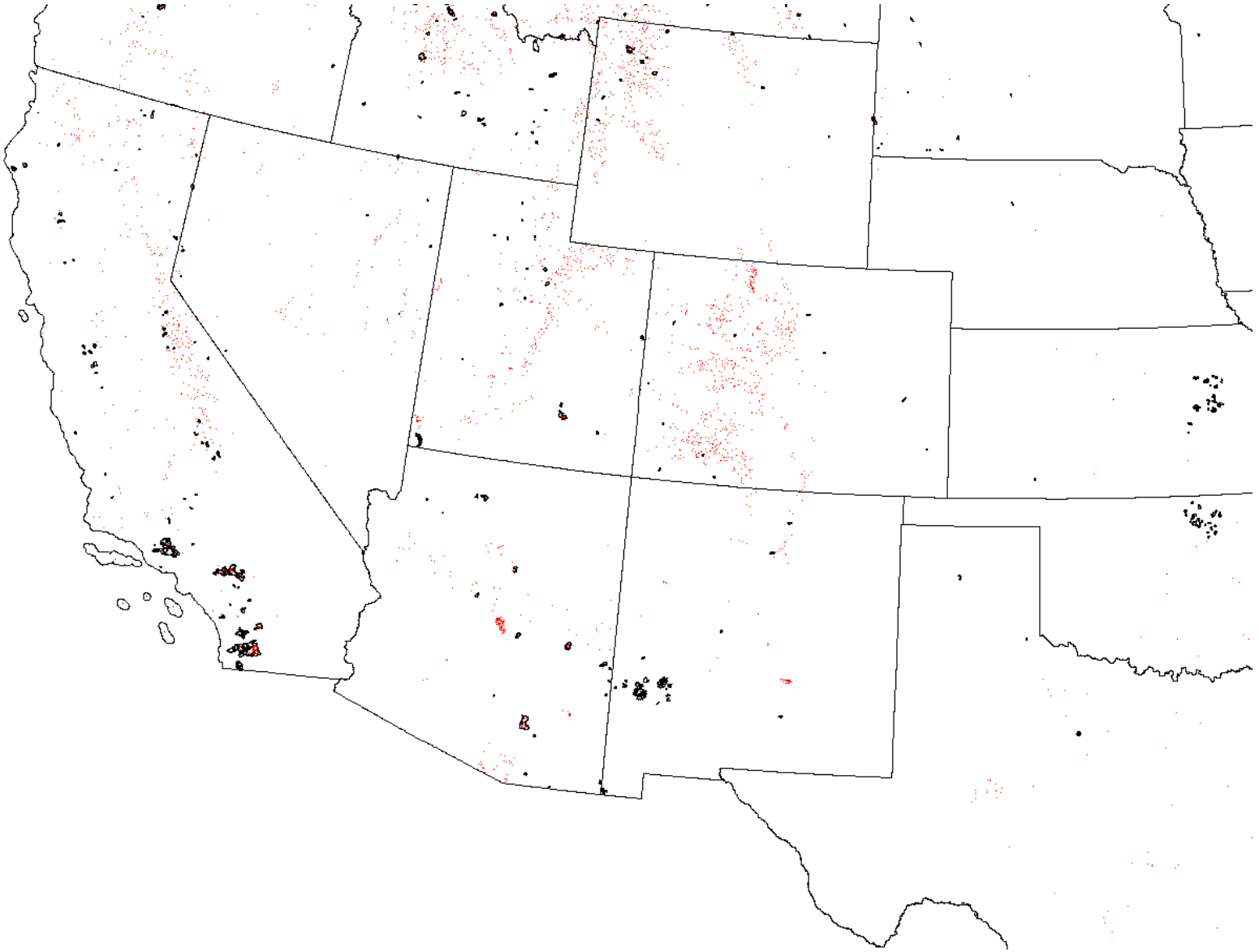
Kumar, et al., in prep.

ΔIntegrated NDVI with Threshold for 2003 – 2002



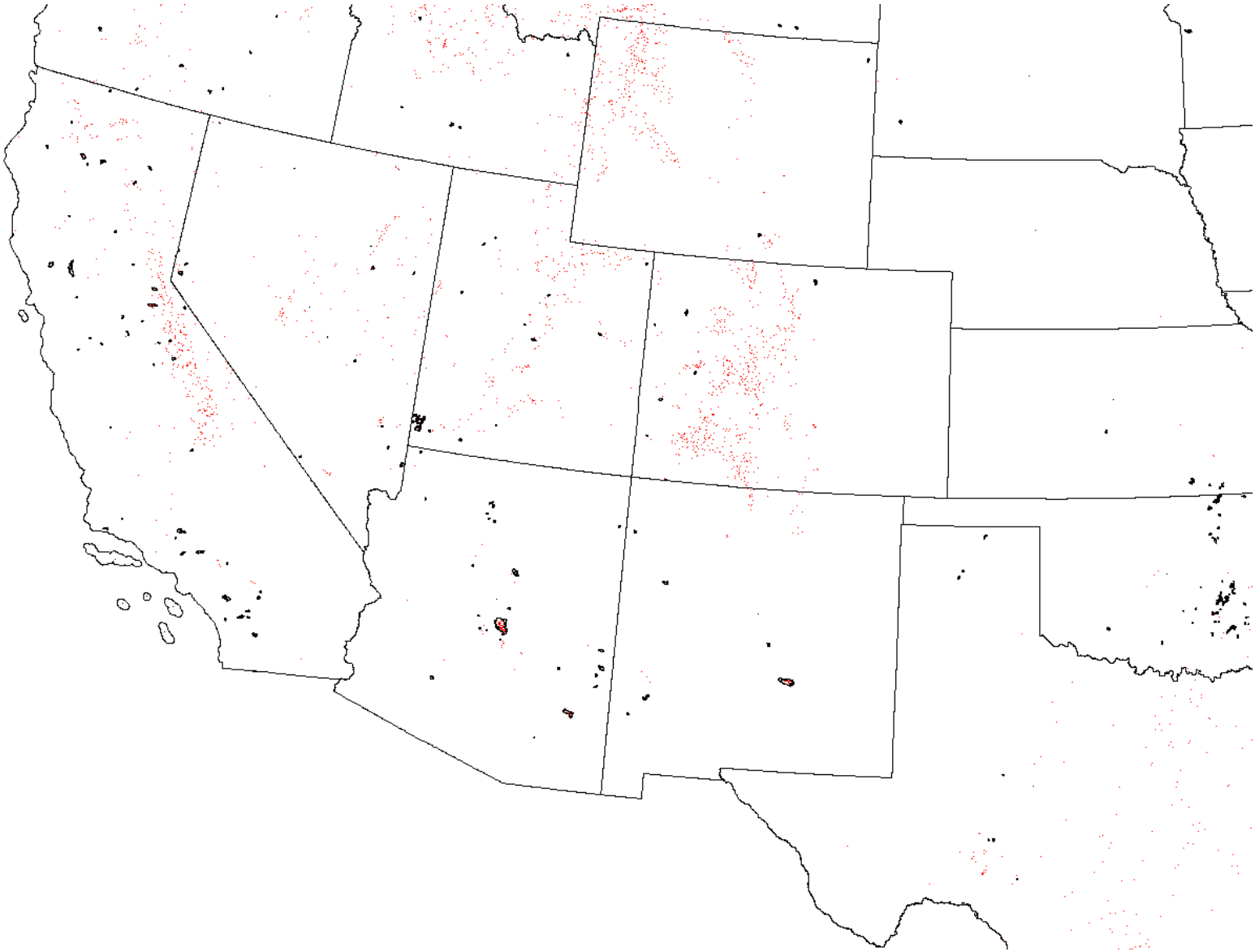
Kumar, et al., in prep.

ΔIntegrated NDVI with Threshold for 2004 – 2003



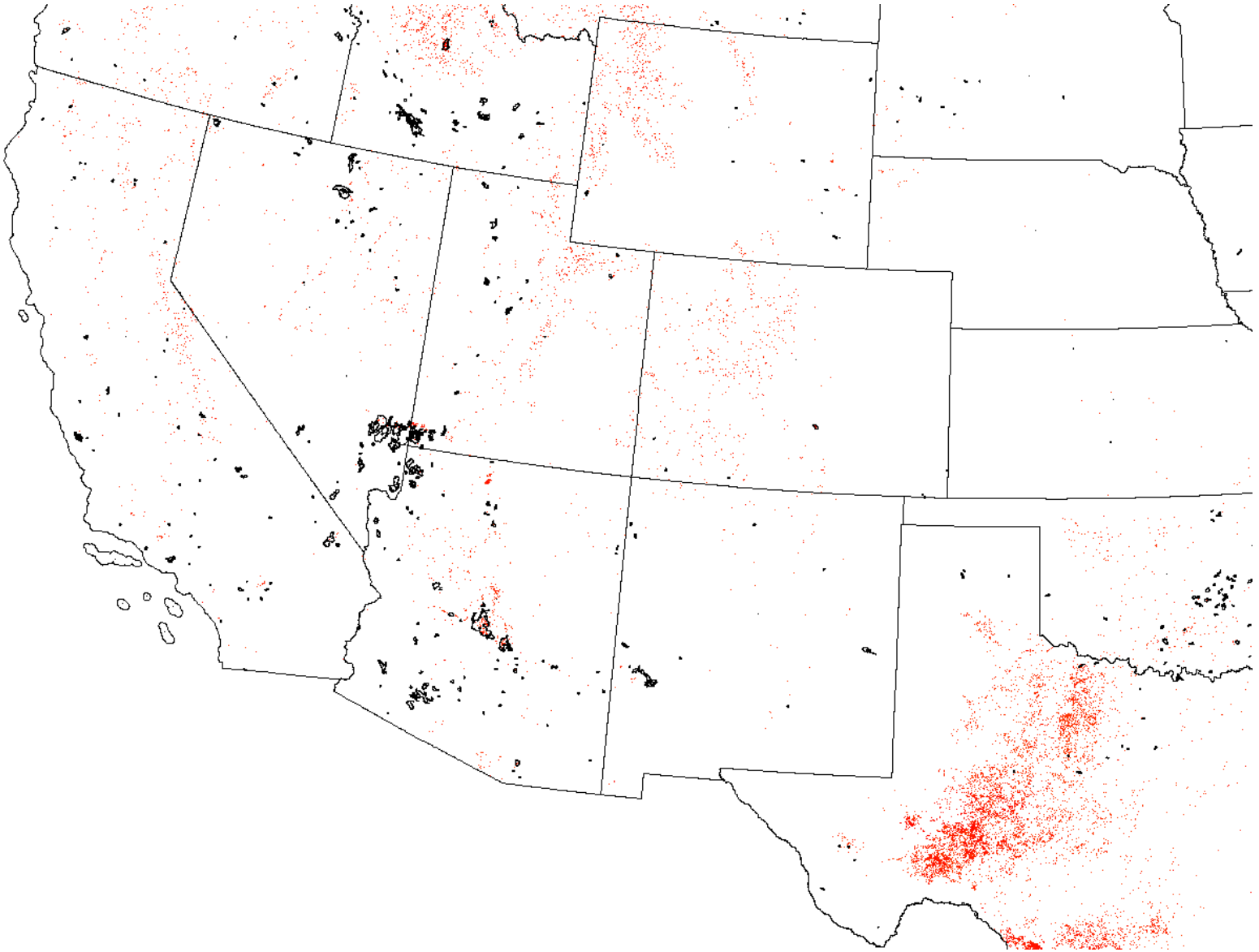
Kumar, et al., in prep.

ΔIntegrated NDVI with Threshold for 2005 – 2004



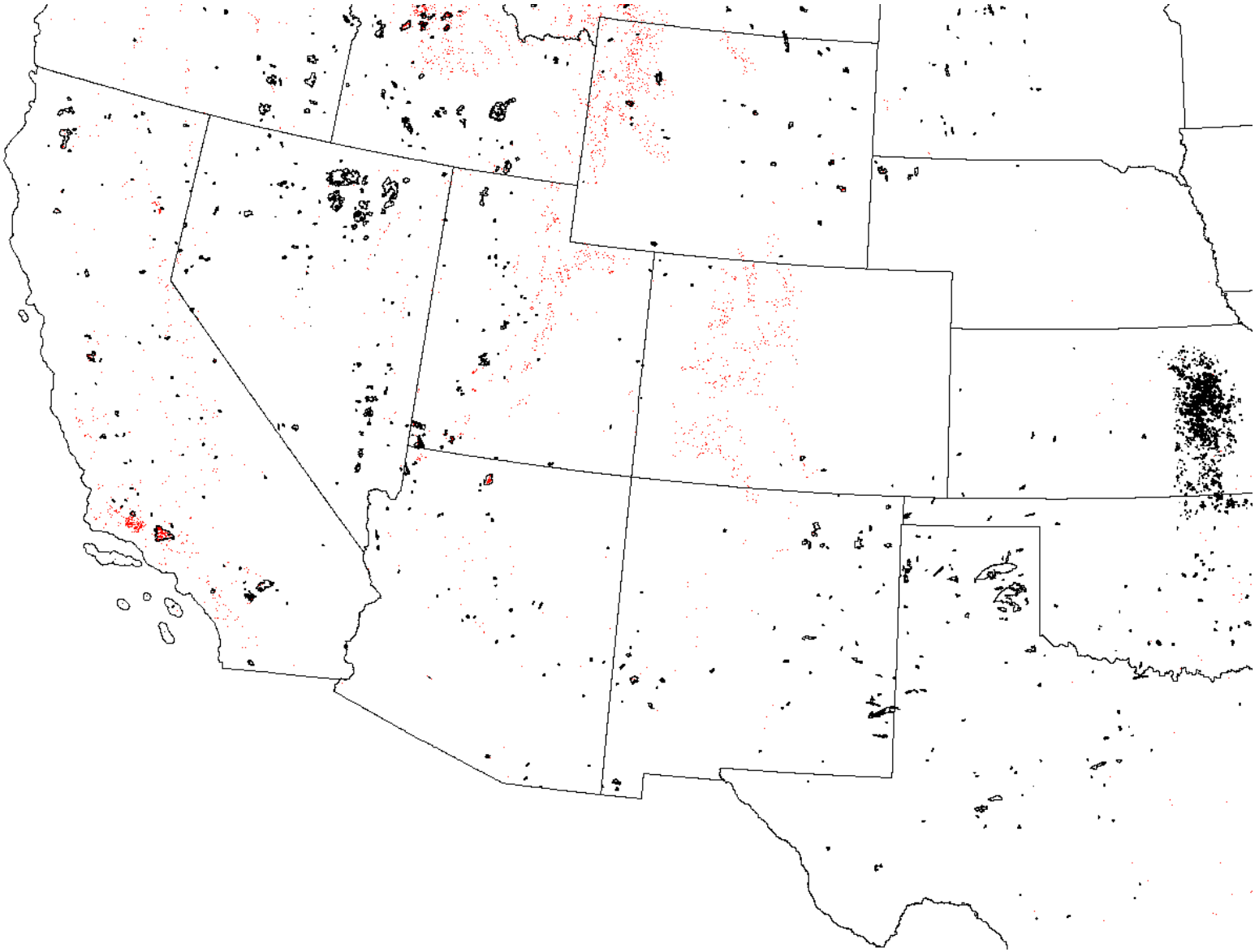
Kumar, et al., in prep.

ΔIntegrated NDVI with Threshold for 2006 – 2005



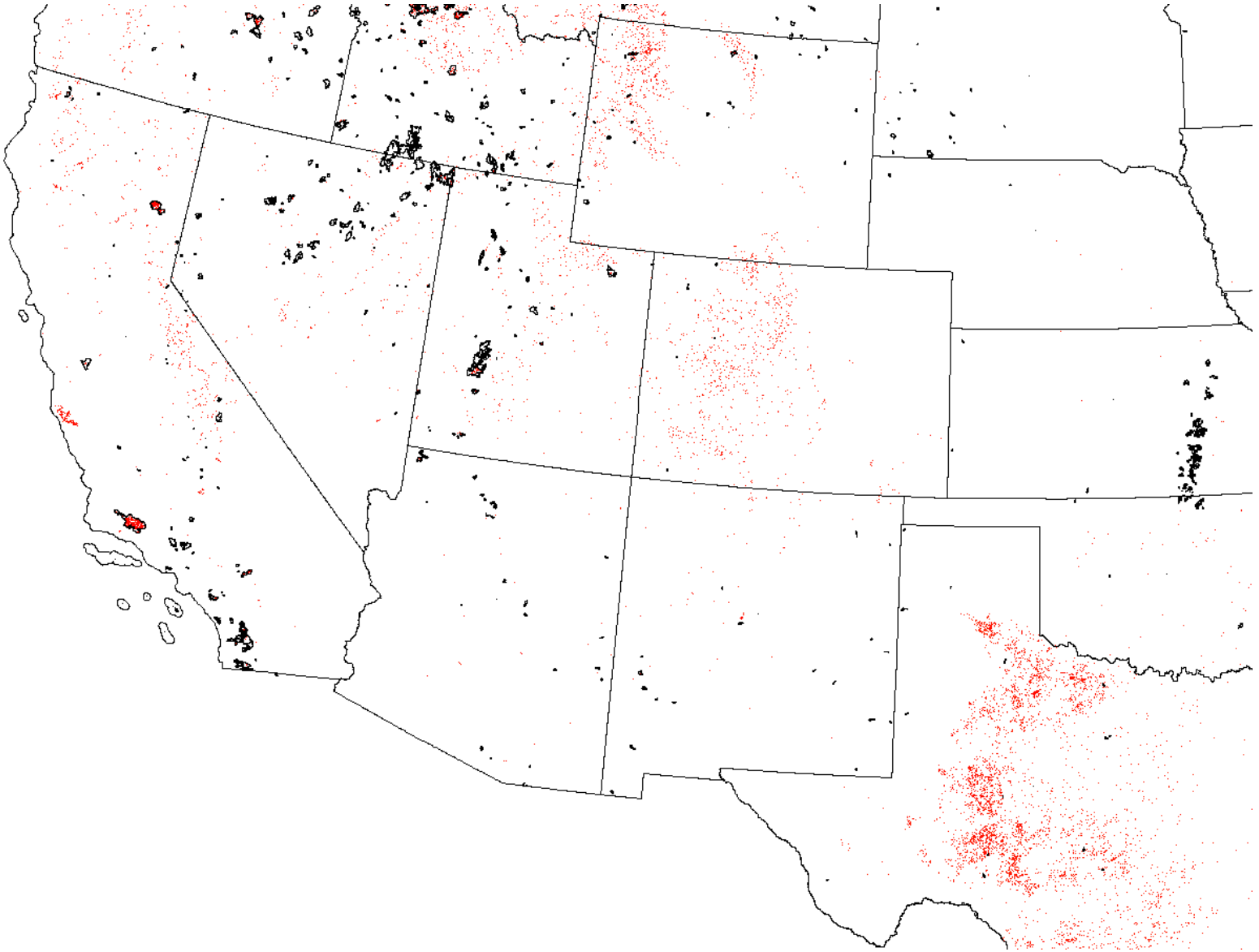
Kumar, et al., in prep.

ΔIntegrated NDVI with Threshold for 2007 – 2006



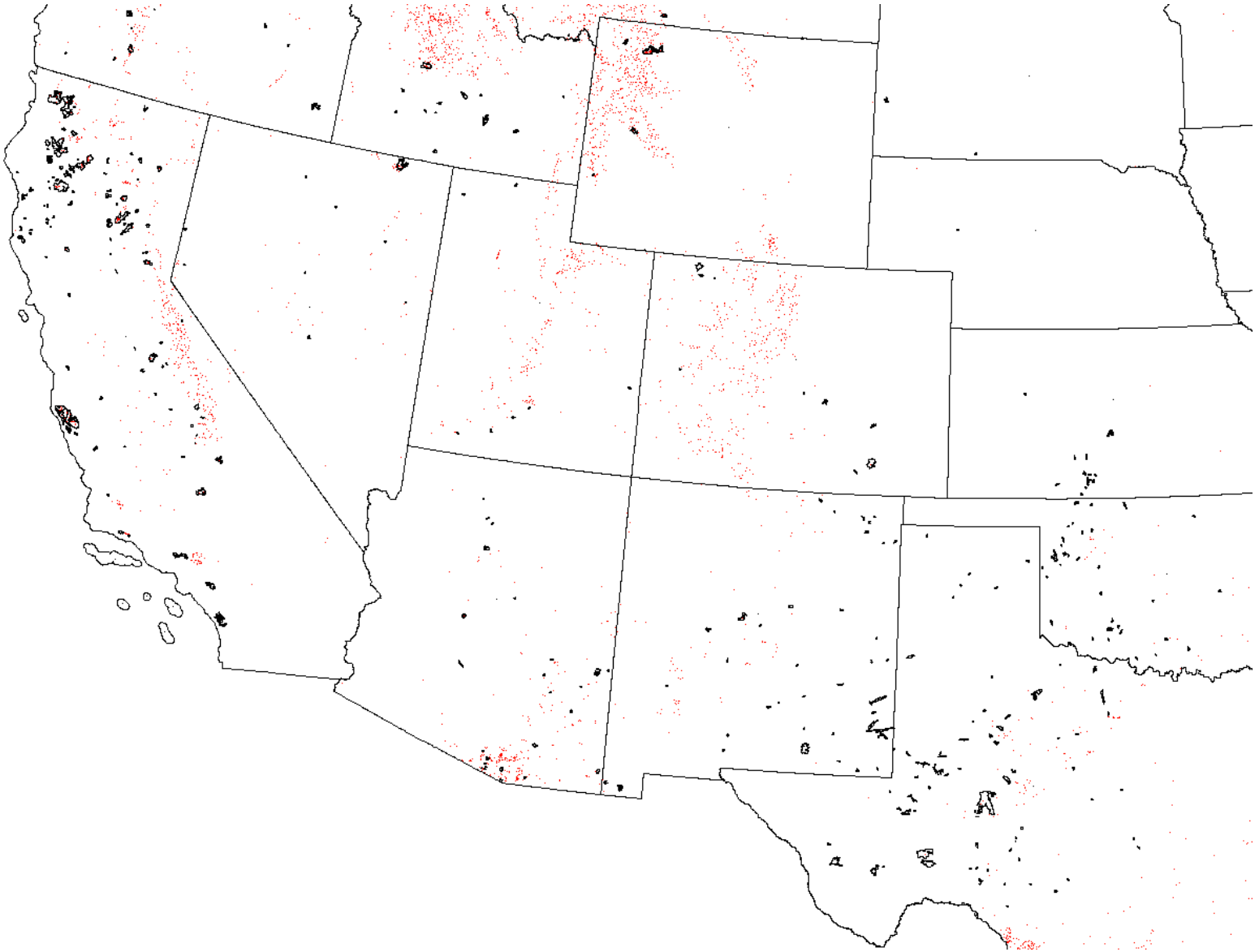
Kumar, et al., in prep.

ΔIntegrated NDVI with Threshold for 2008 – 2007



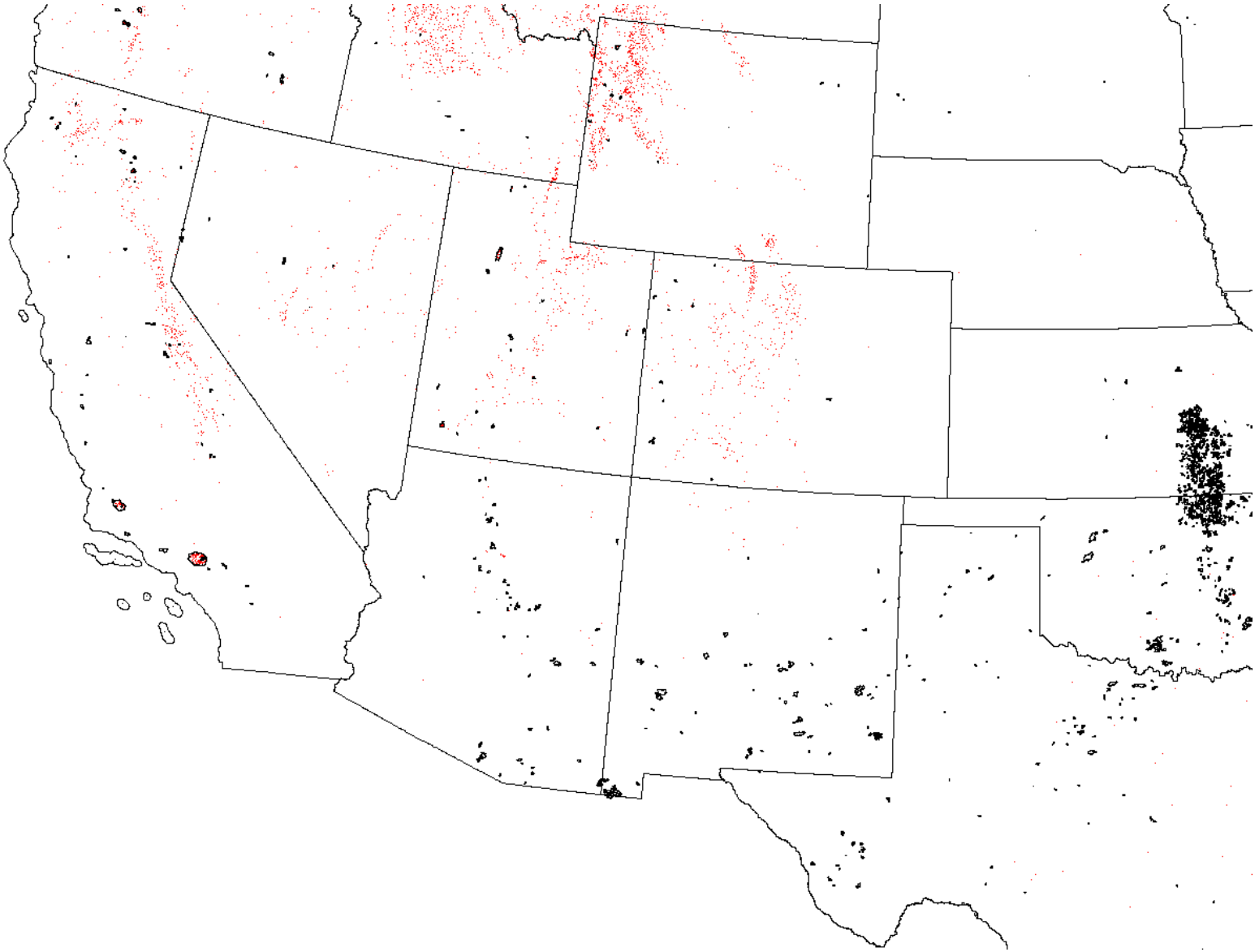
Kumar, et al., in prep.

ΔIntegrated NDVI with Threshold for 2009 – 2008



Kumar, et al., in prep.

ΔIntegrated NDVI with Threshold for 2010 – 2009



Kumar, et al., in prep.

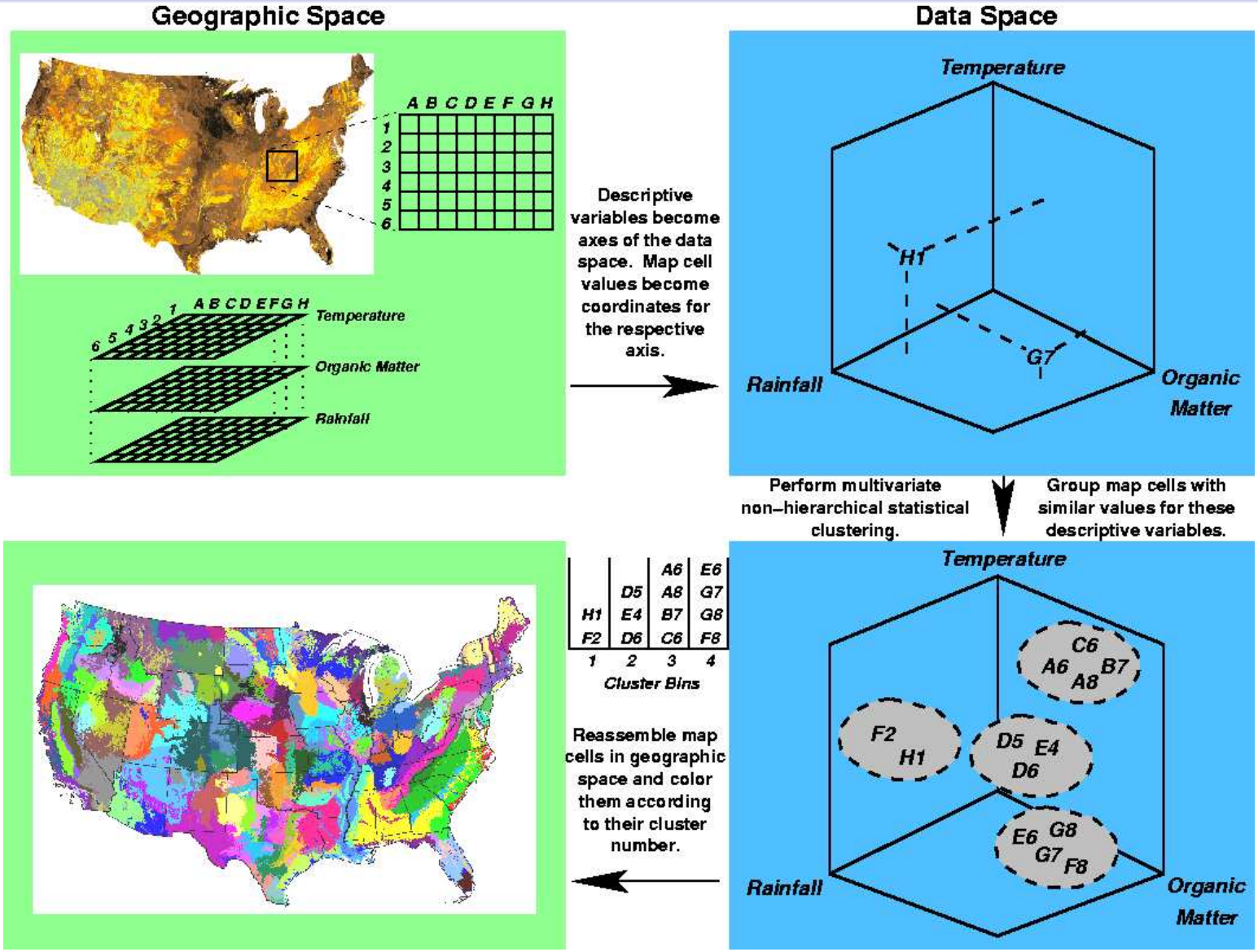
Summary and Future Work

- Initial results of geospatiotemporal cluster analysis of phenology from MODIS NDVI are promising, suggesting such analysis will be a key component in the ForWarn early warning system.
- The enhanced, accelerated k-means clustering algorithm enables the analysis of very large, high resolution remote sensing data.
- Determining "normal" phenological patterns is difficult due to interannual climate variability, spatially variable climate change trend, and relatively short satellite record.
- However, mortality events, like progressive Mountain Pine Beetle damage and wildfire, are easily detected.
- The next step is to establish generalized or biome-specific or event-specific thresholds based on interannual variability, continue to obtain validation from ADS and ground surveys, and track and accumulate both loss and new growth for carbon accounting.
- Future work will build a library of phenostate transitions attributed to pests or pathogens for individual biomes, allowing the system to hypothesize about causes of future disturbances detected.

Quantitative Sampling Network Design

- Resource and logistical constraints limit the frequency and extent of observations, necessitating the development of a systematic sampling strategy that objectively represents environmental variability at the desired spatial scale.
- Required is a methodology that provides a quantitative framework for informing site selection and determining the representativeness of measurements.
- Multivariate spatiotemporal clustering (MSTC) was applied at the landscape scale (4 km^2) for the State of Alaska to demonstrate its utility for representativeness and scaling.
- An extension of the method applied by Hargrove and Hoffman for design of National Science Foundation's (NSF's) National Ecological Observatory Network (NEON) domains.

Multivariate Spatiotemporal Clustering (MSTC)

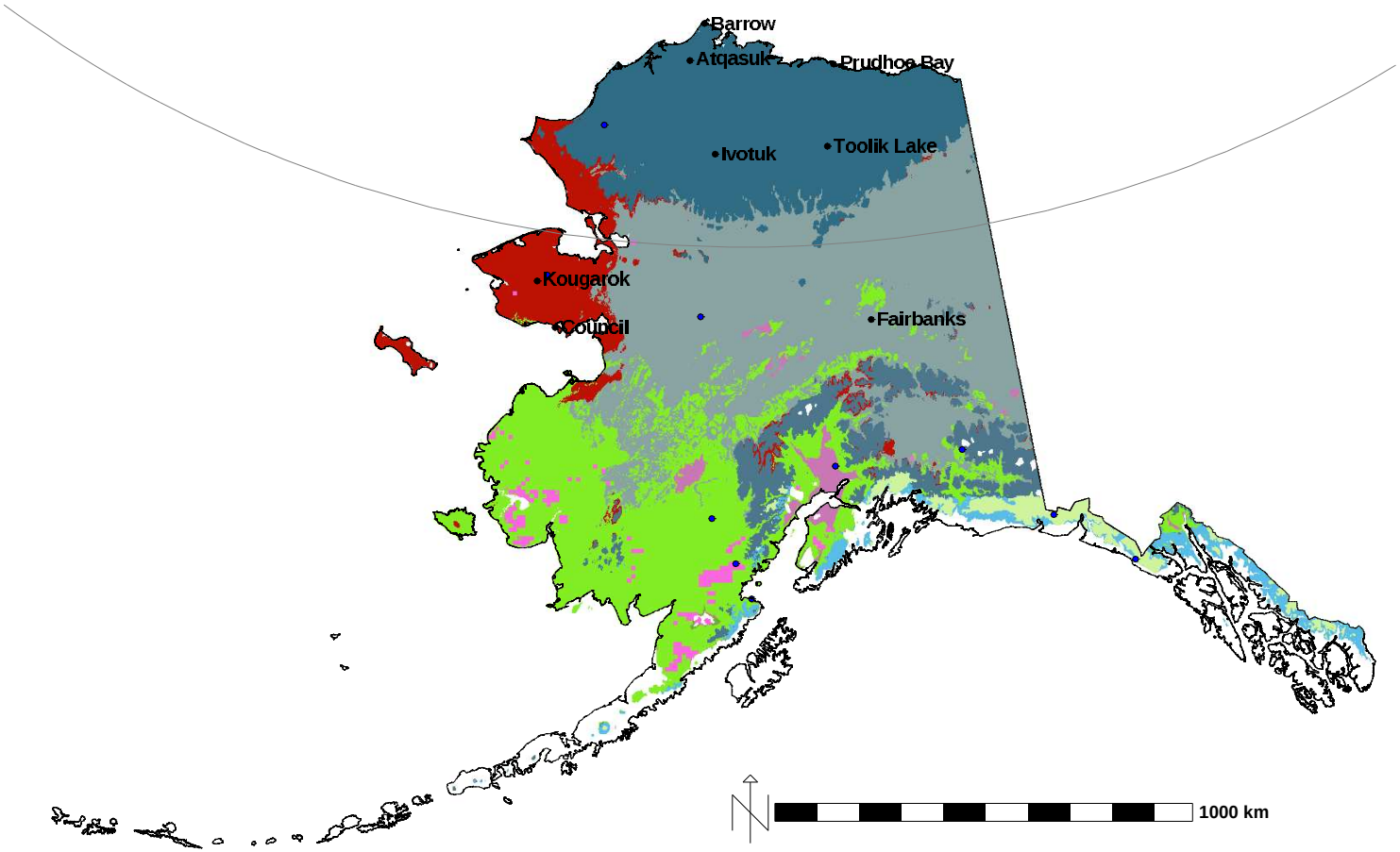


Data Layers

Table: 37 variables averaged for 2000–2009 and 2090–2099

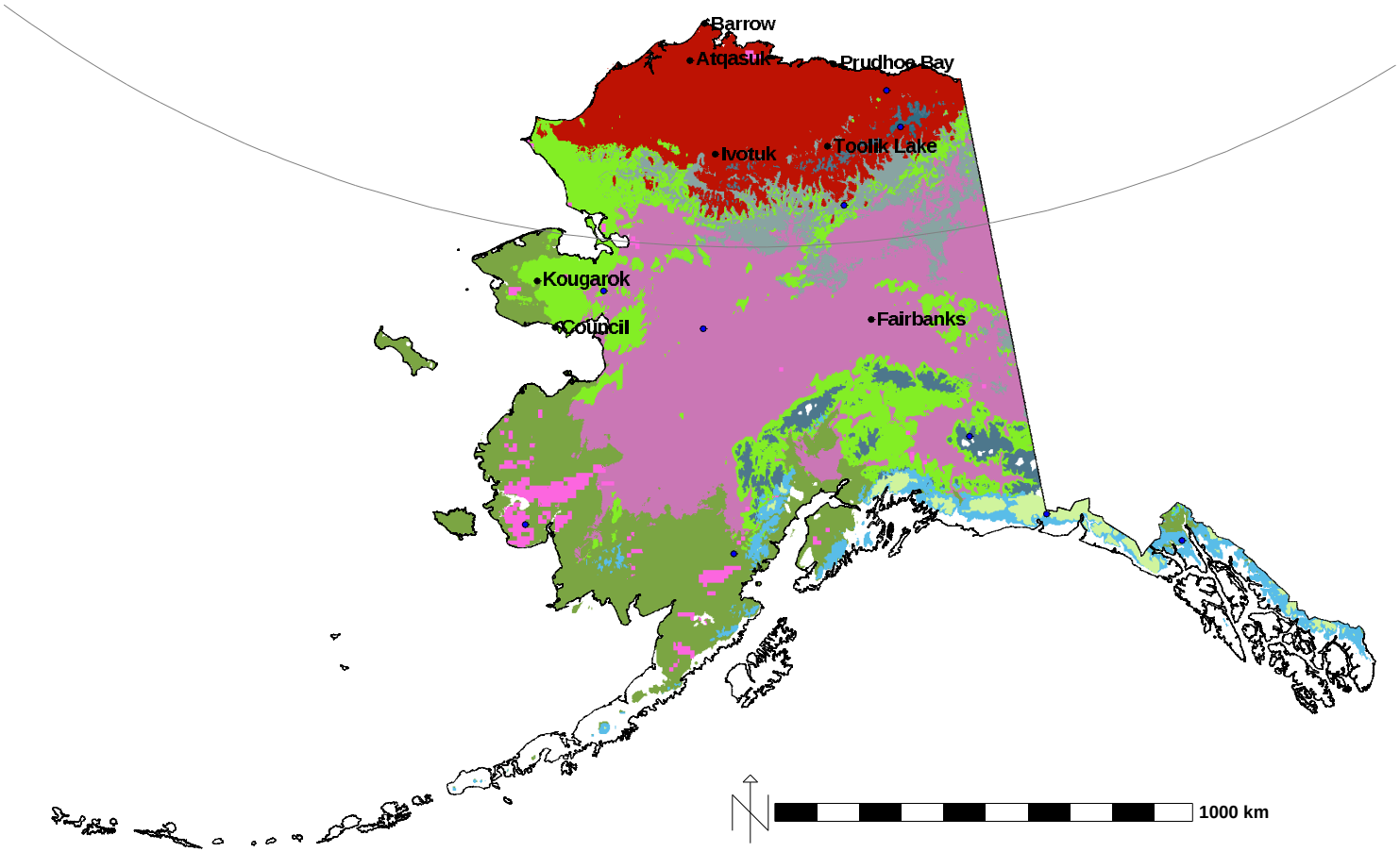
Description	Number/Name	Units	Source
Monthly mean air temperature	12	°C	GCM
Monthly mean precipitation	12	mm	GCM
Day of freeze	mean	day of year	GCM
	standard deviation	days	
Day of thaw	mean	day of year	GCM
	standard deviation	days	
Length of growing season	mean	days	GCM
	standard deviation	days	
Maximum active layer thickness	1	m	GIPL
Warming effect of snow	1	°C	GIPL
Mean annual ground temperature at bottom of active layer	1	°C	GIPL
Mean annual ground surface temperature	1	°C	GIPL
Thermal offset	1	°C	GIPL
Limnicity	1	%	NHD
Elevation	1	m	SRTM

10 Alaska Ecoregions (2000–2009)



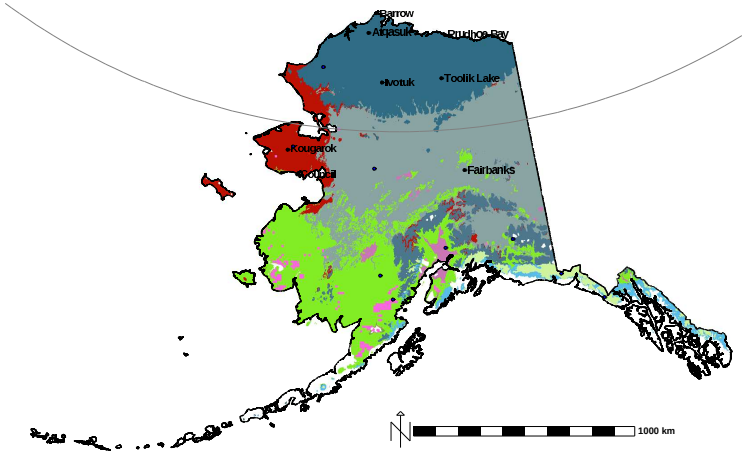
Each ecoregion is a different random color. Blue filled circles mark locations most representative of mean conditions of each region.

10 Alaska Ecoregions (2090–2099)

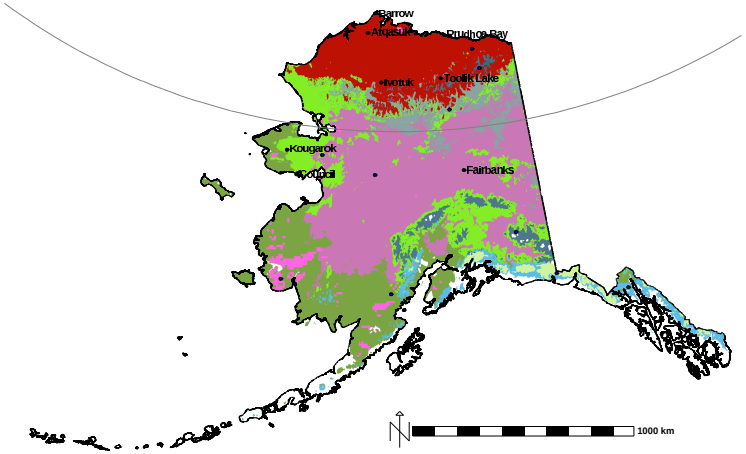


Each ecoregion is a different random color. Blue filled circles mark locations most representative of mean conditions of each region.

10 Alaska Ecoregions, Present and Future



2000–2009

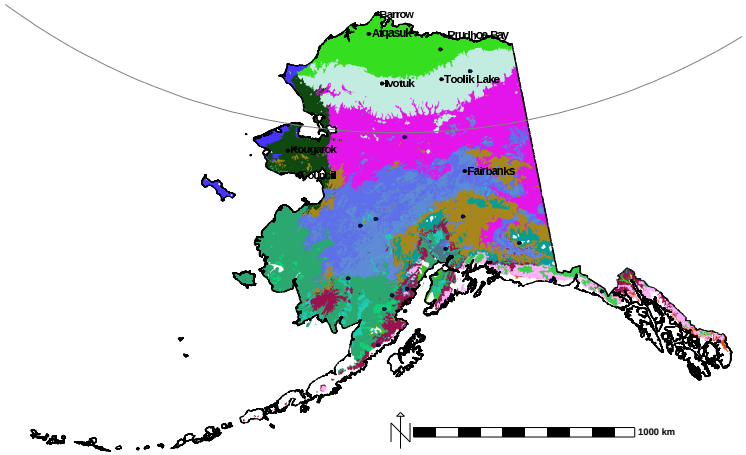


2090–2099

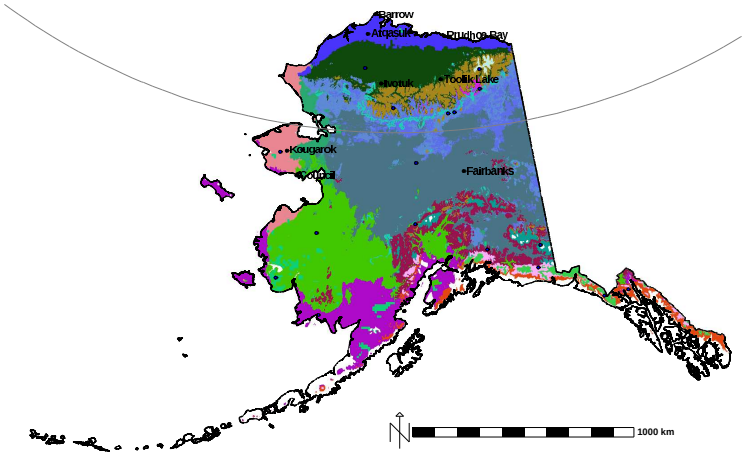
Since the random colors are the same in both maps, a change in color represents an environmental change between the present and the future.

At this level of division, the conditions in the large boreal forest become compressed onto the Brooks Range and the conditions on the Seward Peninsula migrate to the North Slope.

20 Alaska Ecoregions, Present and Future



2000–2009

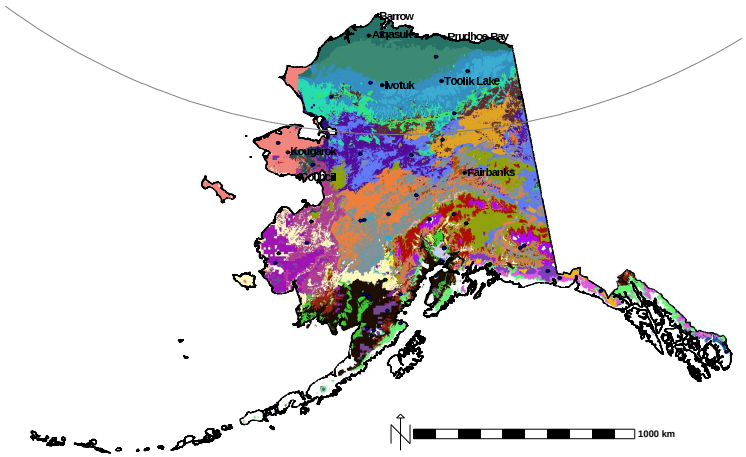


2090–2099

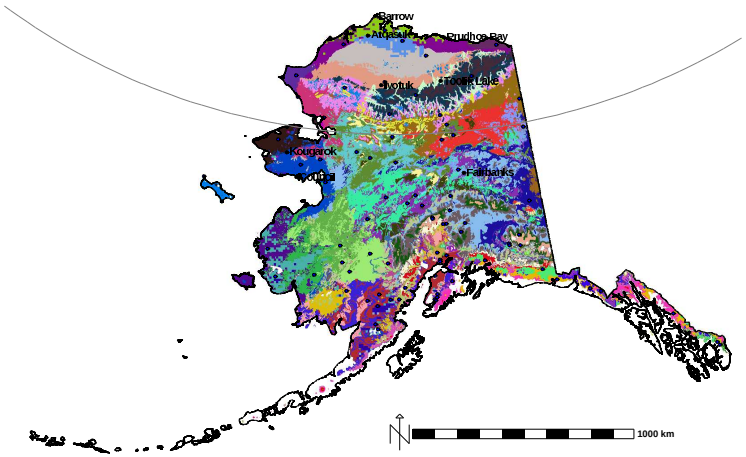
Since the random colors are the same in both maps, a change in color represents an environmental change between the present and the future.

At this level of division, the two primary regions of the Seward Peninsula and that of the northern boreal forest replace the two regions on the North Slope almost entirely.

50 and 100 Alaska Ecoregions, Present



$k = 50$, 2000–2009

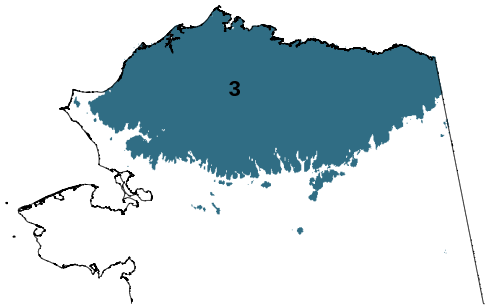


$k = 100$, 2000–2009

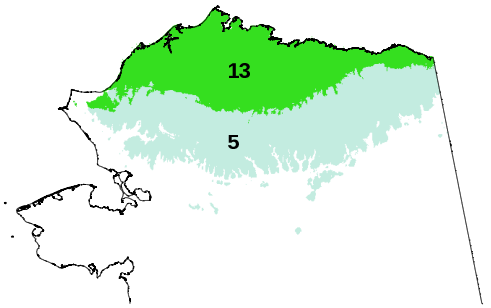
Since the random colors are the same in both maps, a change in color represents an environmental change between the present and the future.

At high levels of division, some regions vanish between the present and future while other region representing new combinations of environmental conditions come into existence.

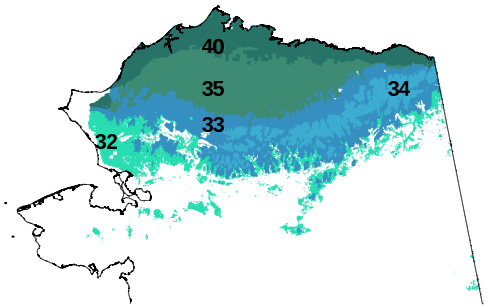
A Hierarchy of Ecoregions



(a) At $k = 10$, the North Slope is occupied by Ecoregion #3, which corresponds to the Arctic Tundra Level 2 ecological group.



(b) At $k = 20$, the North Slope is occupied by Ecoregion #5, corresponding to the Brooks Range ecoregion; and Ecoregion #13, corresponding to the Beaufort Coastal Plains ecoregion.

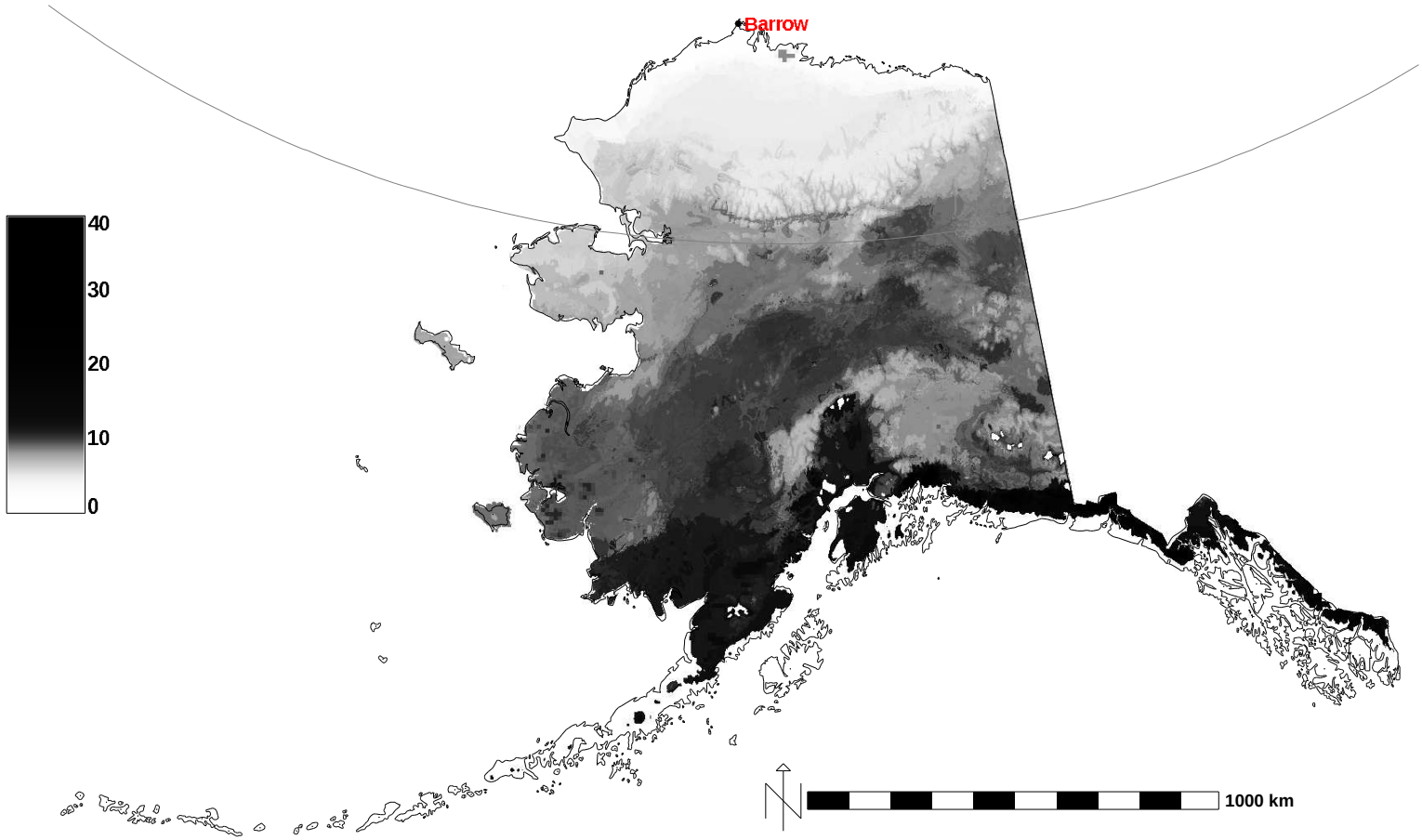


(c) At $k = 50$, the North Slope is occupied by Ecoregion #32, corresponding to the Intermontane Boreal ecological group; Ecoregions #33 and #34, corresponding to mid- and high-elevation of the Brooks Range ecoregion; Ecoregion #35, corresponding to the Brooks Foothills ecoregion; and Ecoregion #40, corresponding to the Beaufort Coastal Plains ecoregion.

NGEE Arctic Site Representativeness

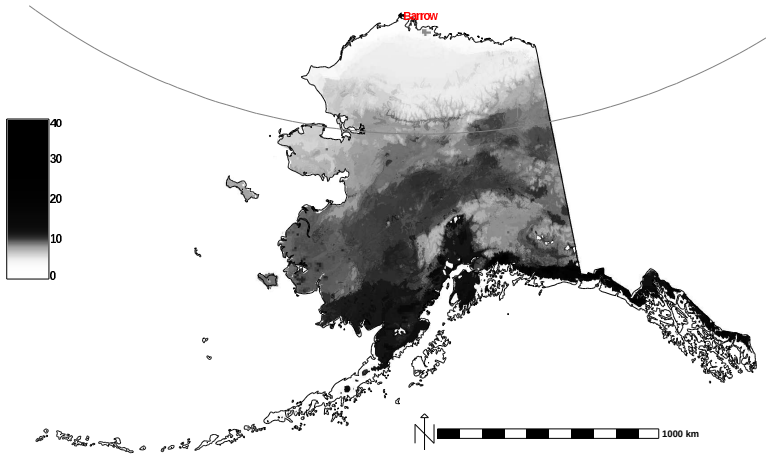
- This representativeness analysis uses the standardized n-dimensional data space formed from all input data layers.
- In this data space, the Euclidean distance between a sampling location (like Barrow) and every other point is calculated.
- These data space distances are then used to generate grayscale maps showing the similarity, or lack thereof, of every location to the sampling location.
- In the subsequent maps, white areas are well represented by the sampling location or network, while dark and black areas as poorly represented by the sampling location or network.
- This analysis assumes that the climate surrogates maintain their predictive power and that no significant biological adaptation occurs in the future.

Present Representativeness of Barrow or "Barrow-ness"

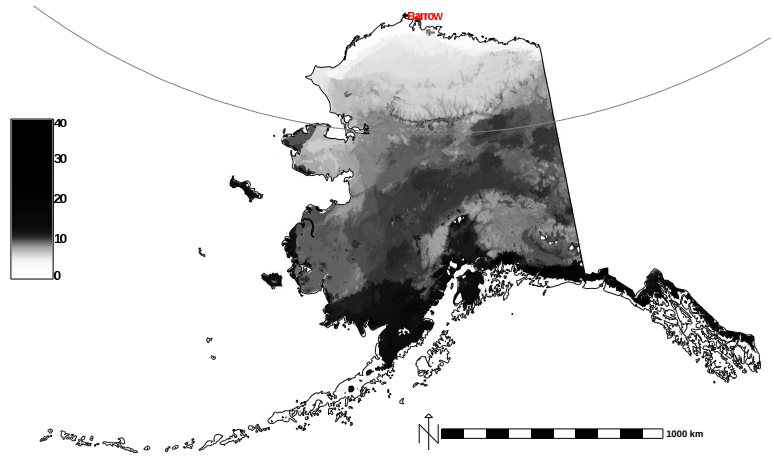


Light-colored regions are well represented and dark-colored regions are poorly represented by the sampling location listed in **red**.

Present vs. Future Barrow-ness



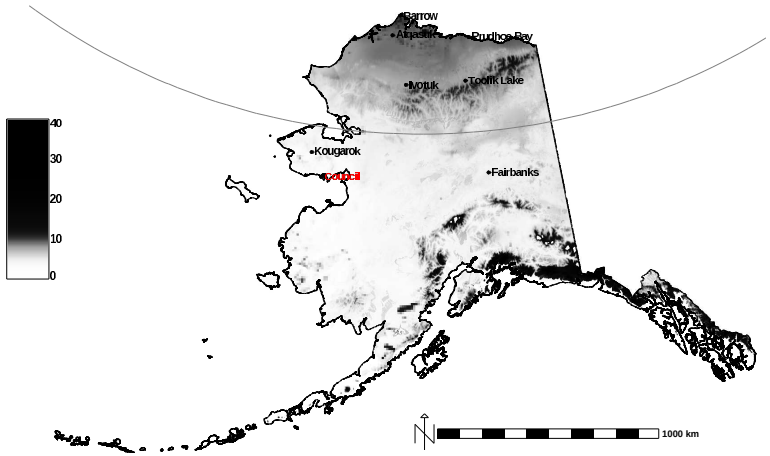
2000–2009



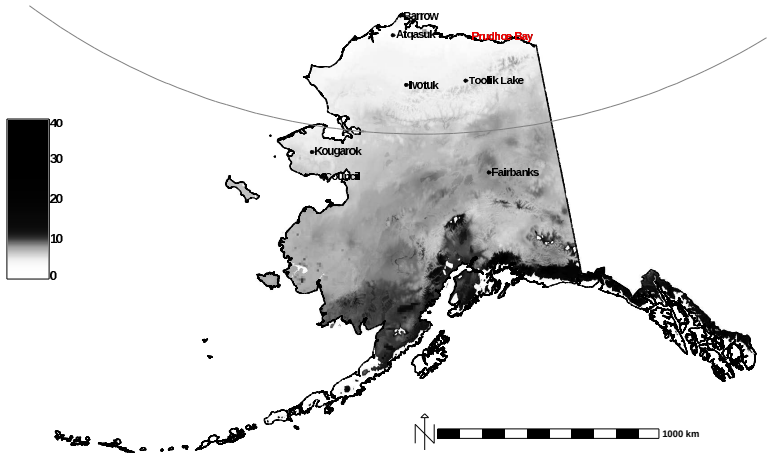
2090–2099

As environmental conditions change, due primarily to increasing temperatures, climate gradients increase and the representativeness of Barrow will be diminished in the future.

Council and Prudhoe Bay Representativeness



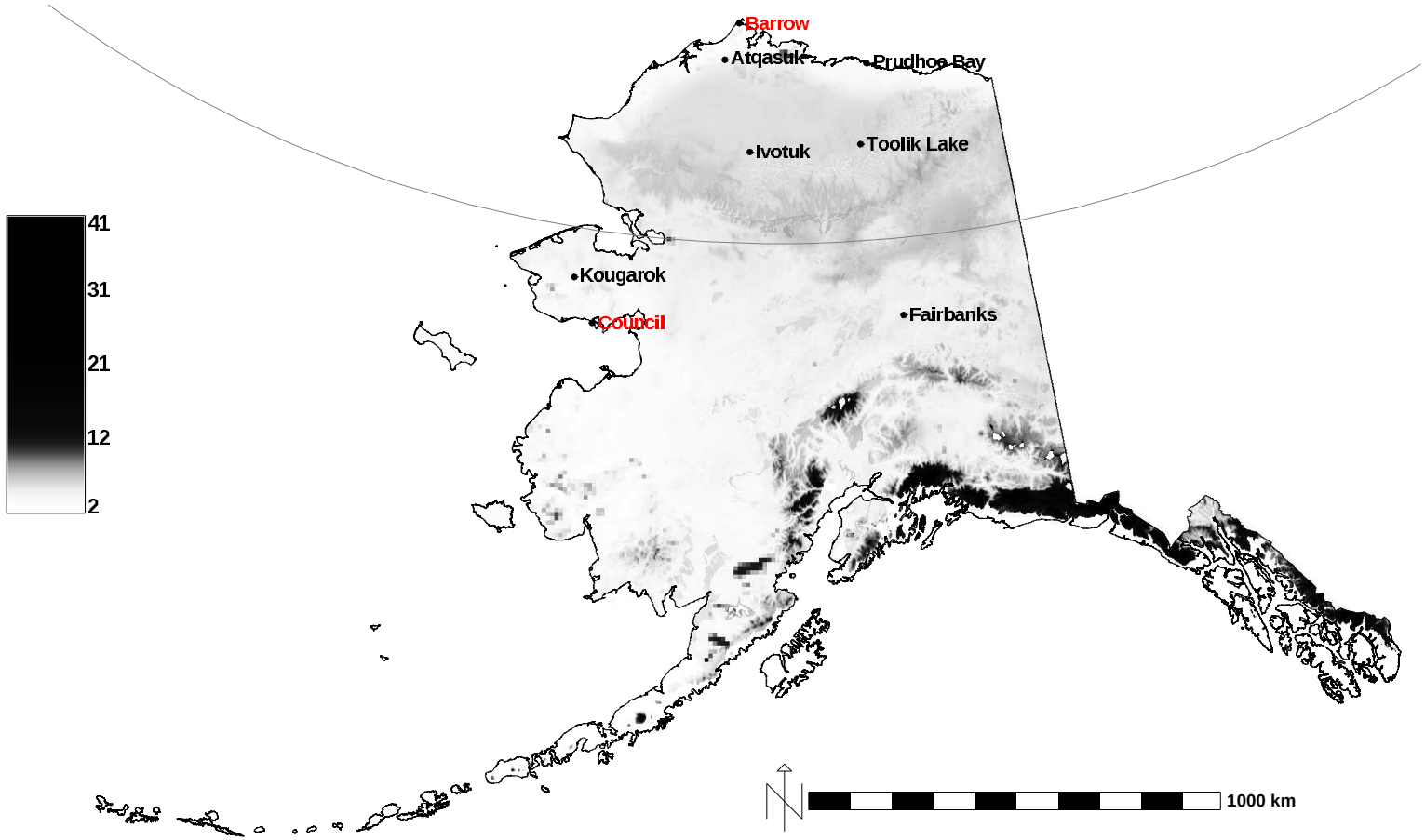
Council



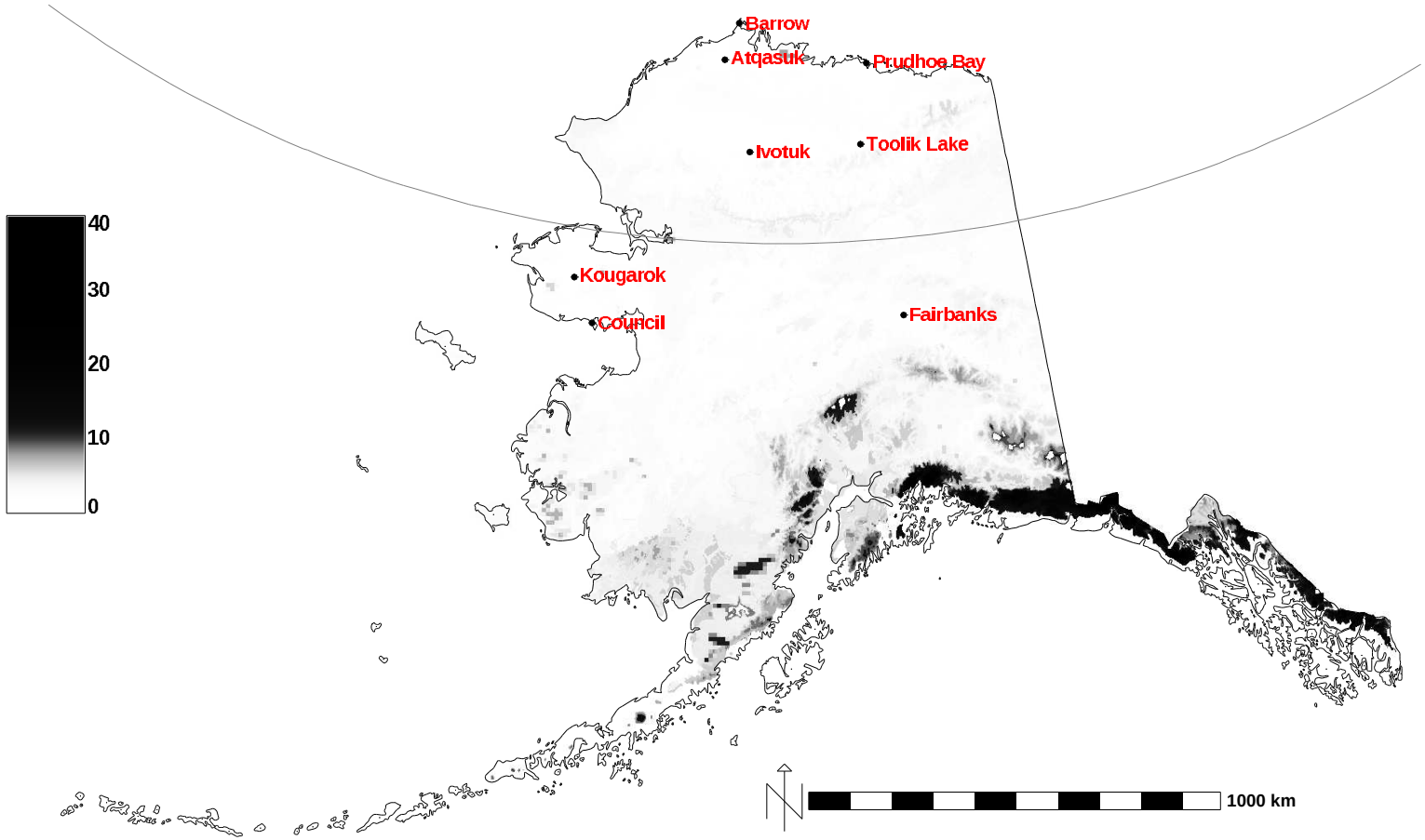
Prudhoe Bay

Representativeness analysis was performed for sites at Barrow, Council, Atqasuk, Ivotuk, Kougarok, Prudhoe Bay, Toolik Lake, and Fairbanks.

Network Representativeness: Barrow + Council



Network Representativeness: All 8 Sites



State Space Dissimilarity: 8 Sites, Present (2000–2009)

Table: Site state space distances for the present (2000–2009) with DEM

Sites	Council	Atqasuk	Ivotuk	Toolik Lake	Kougarok	Prudhoe Bay	Fairbanks
Barrow	9.13	4.53	5.90	5.87	7.98	3.57	12.16
Council		8.69	6.37	7.00	2.28	8.15	5.05
Atqasuk			5.18	5.23	7.79	1.74	10.66
Ivotuk				1.81	5.83	4.48	7.90
Toolik Lake					6.47	4.65	8.70
Kougarok						7.25	5.57
Prudhoe Bay							10.38

State Space Dissimilarity: 8 Sites, Future (2090–2099)

Table: Site state space distances for the future (2090–2099) with DEM

Sites	Council	Atqasuk	Ivotuk	Toolik Lake	Kougarok	Prudhoe Bay	Fairbanks
Barrow	8.87	4.89	6.88	6.94	8.04	4.18	11.95
Council		8.82	6.93	7.74	2.43	8.24	5.66
Atqasuk			5.86	5.84	8.15	2.30	10.16
Ivotuk				2.01	7.27	4.75	7.51
Toolik Lake					7.81	5.00	8.33
Kougarok						7.89	6.42
Prudhoe Bay							9.81

State Space Dissimilarity: 8 Sites, Present and Future

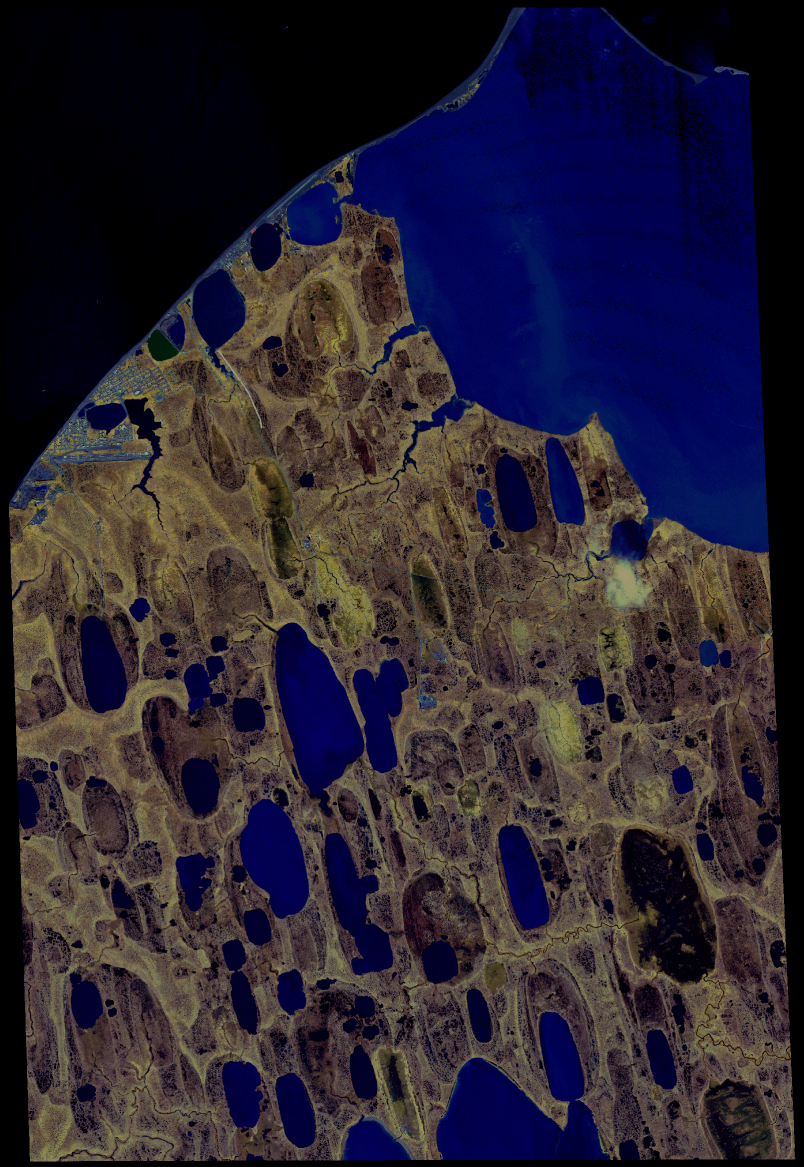
Table: Site state space distances between the present (2000–2009) and the future (2090–2099) with DEM

		Future (2090–2099)							
	Sites	Barrow	Council	Atqasuk	Ivotuk	Toolik Lake	Kougarok	Prudhoe Bay	Fairbanks
Present (2000–2009)	Barrow	3.31	9.67	4.63	6.05	5.75	9.02	3.69	11.67
	Council	8.38	1.65	8.10	5.91	6.87	3.10	7.45	5.38
	Atqasuk	6.01	9.33	2.42	5.46	5.26	8.97	2.63	10.13
	Ivotuk	7.06	7.17	5.83	1.53	2.05	7.25	4.87	7.40
	Toolik Lake	7.19	7.67	6.07	2.48	1.25	7.70	5.23	8.16
	Kougarok	7.29	3.05	6.92	5.57	6.31	2.51	6.54	5.75
	Prudhoe Bay	5.29	8.80	3.07	4.75	4.69	8.48	1.94	9.81
	Fairbanks	12.02	5.49	10.36	7.83	8.74	6.24	10.10	1.96

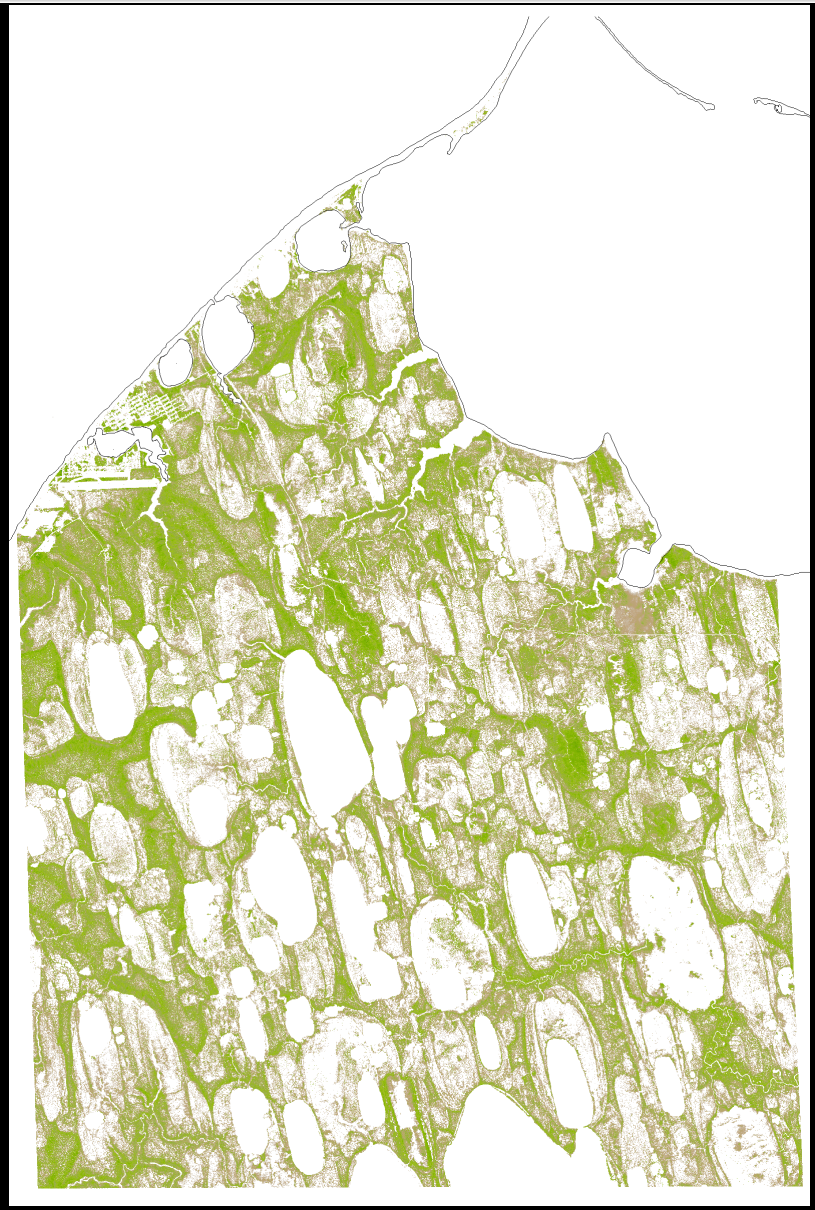
Representativeness: A Quantitative Approach for Scaling

- MSTC provides a quantitative framework for stratifying sampling domains, informing site selection, and determining representativeness of measurements.
- Representativeness analysis provides a systematic approach for up-scaling point measurements to larger domains.
- Methodology is independent of resolution, thus can be applied from site/plot scale to landscape/climate scale.
- It can be extended to include finer spatiotemporal scales, more geophysical characteristics, and remote sensing data.
- Paper describing the methodology has been submitted:

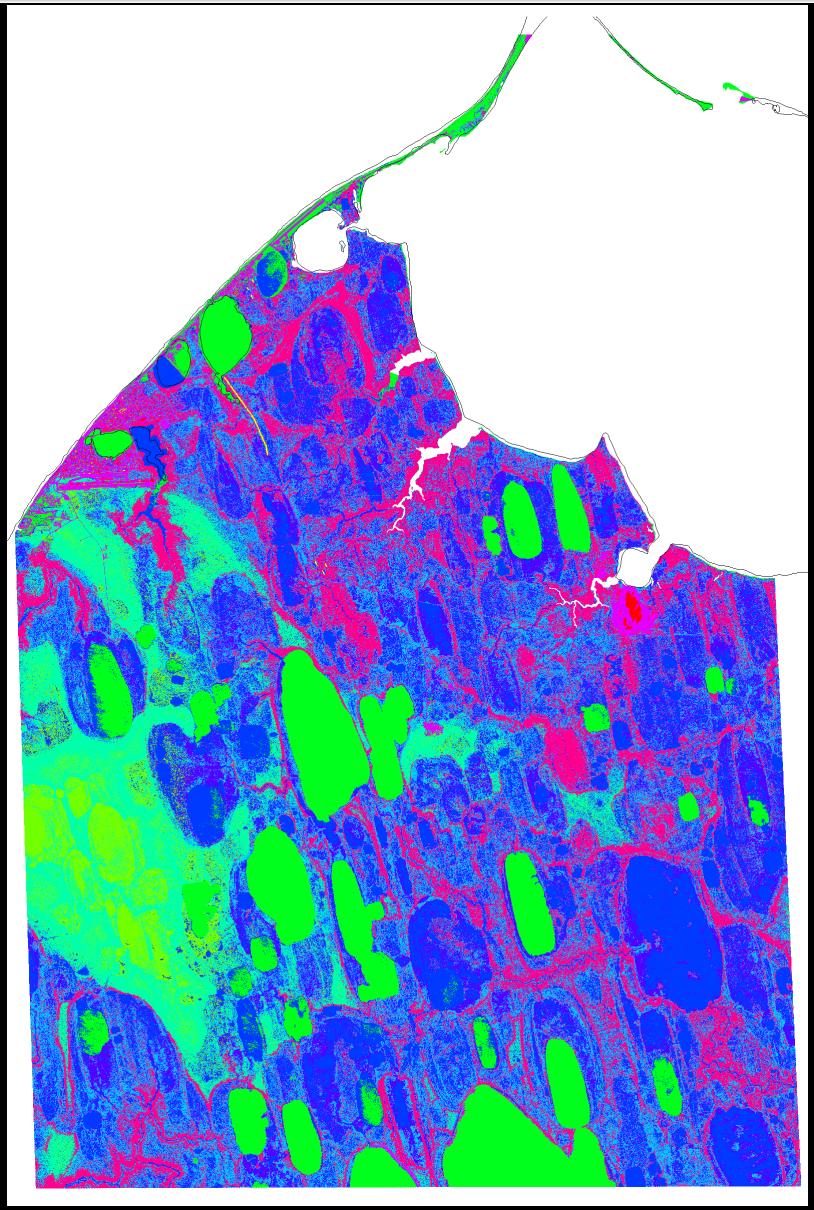
Hoffman, F. M., J. Kumar, R. T. Mills, and W. W. Hargrove (2012) "Representativeness-Based Sampling Network Design for the Arctic." *Landscape Ecol.*, submitted.



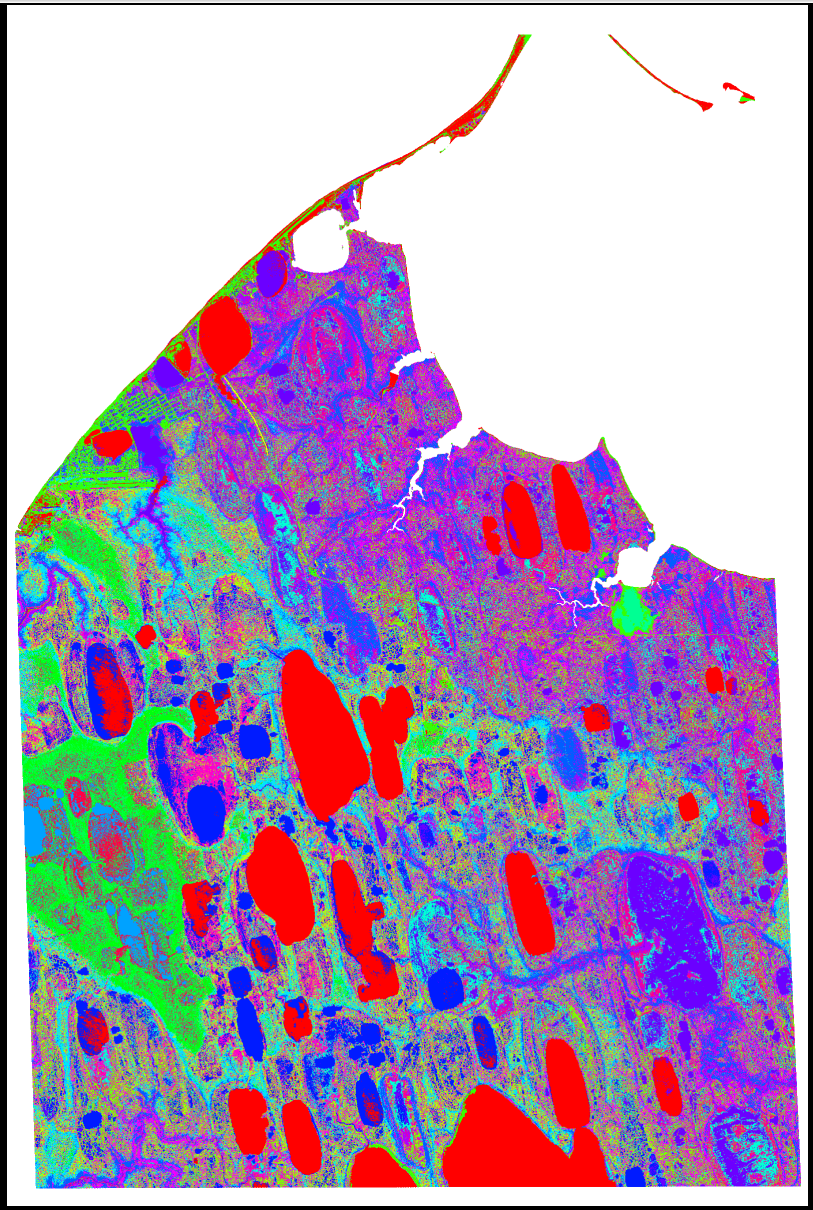
WorldView-2 RGB Composite



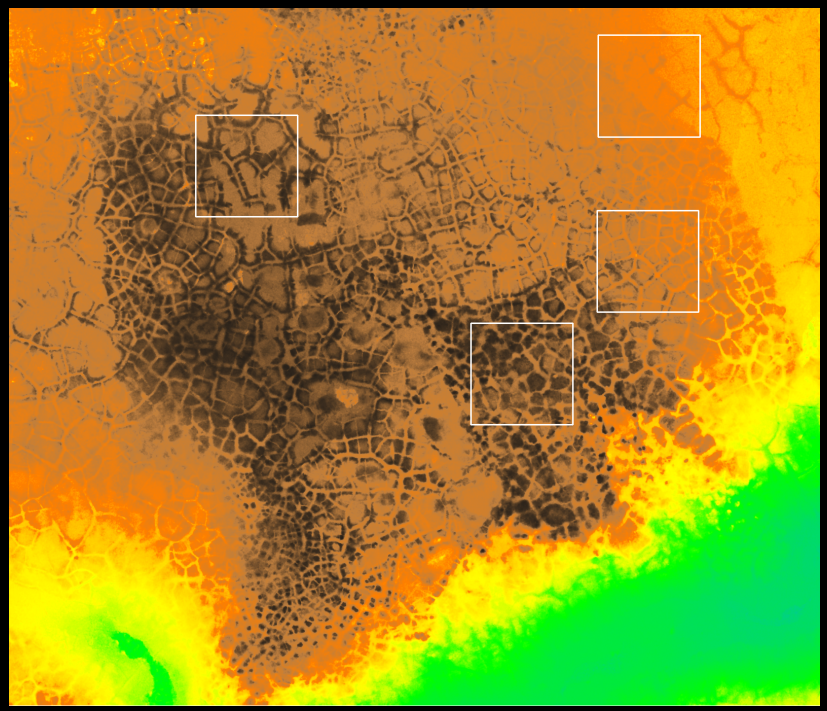
Calculated NDVI



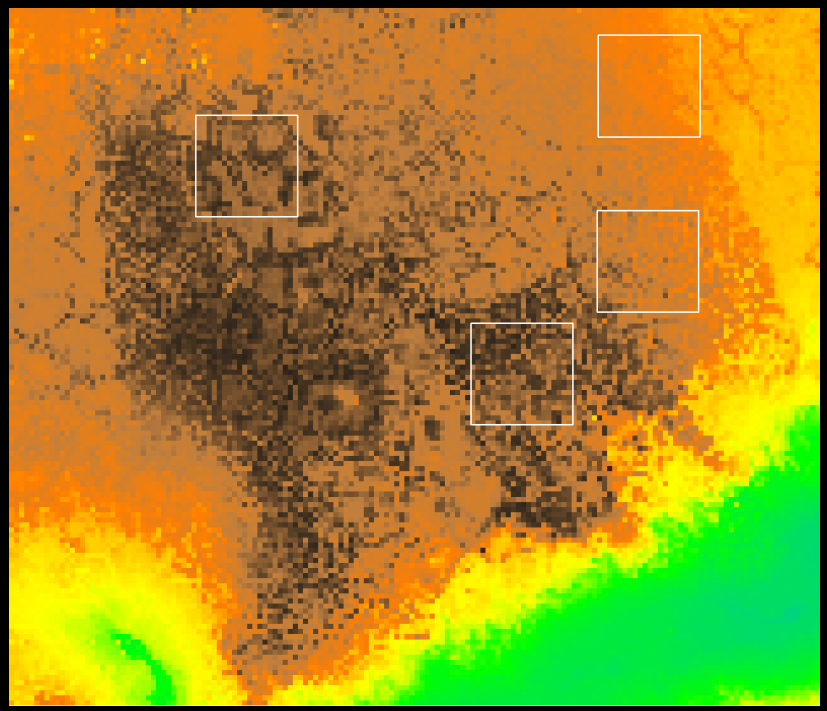
10 Clusters



20 Clusters

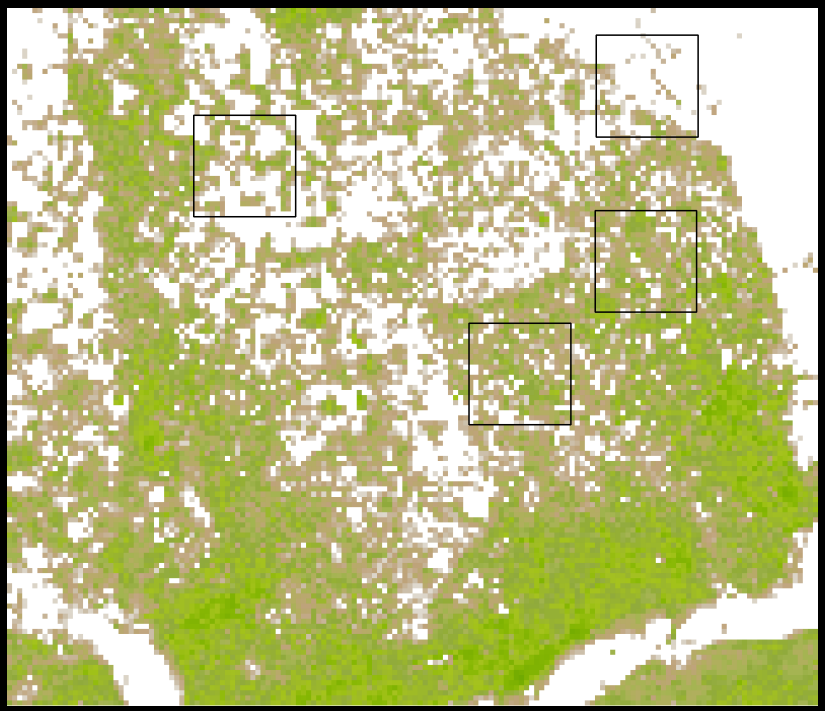


Original LiDAR (0.25 m)

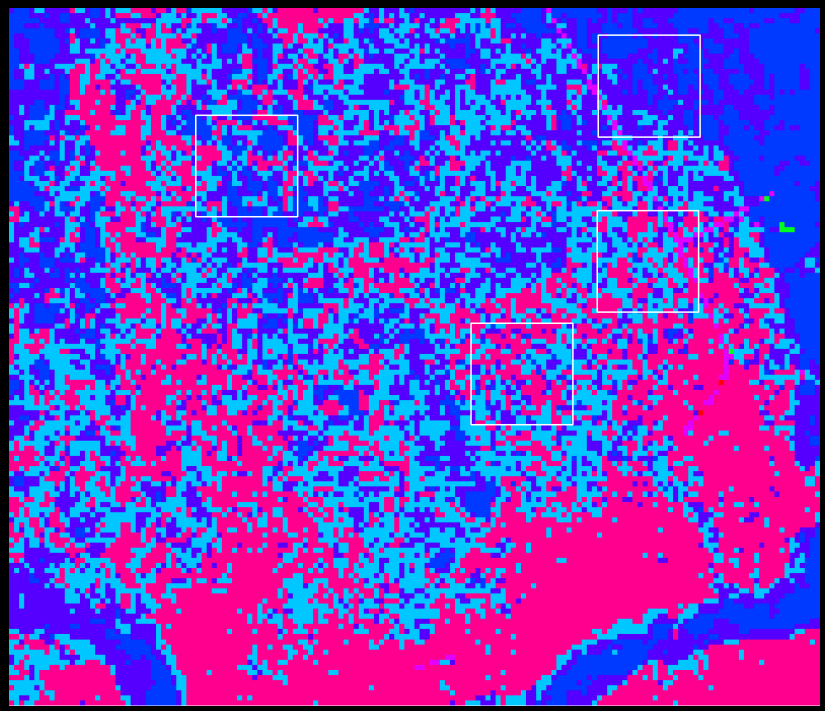


Degraded LiDAR (5.00 m)

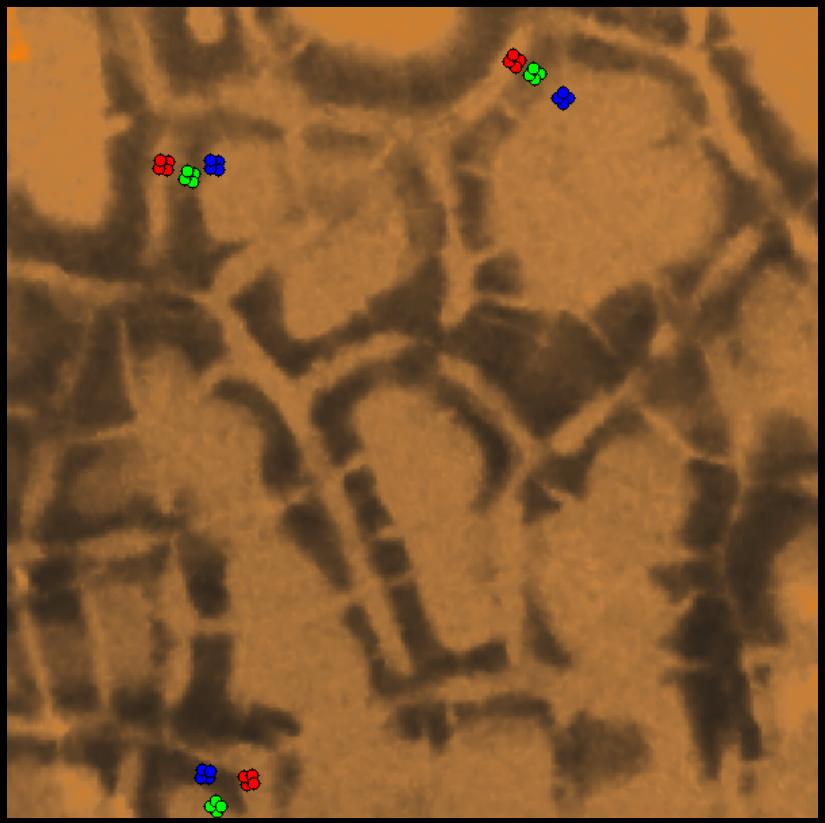
LiDAR from Craig Tweedie



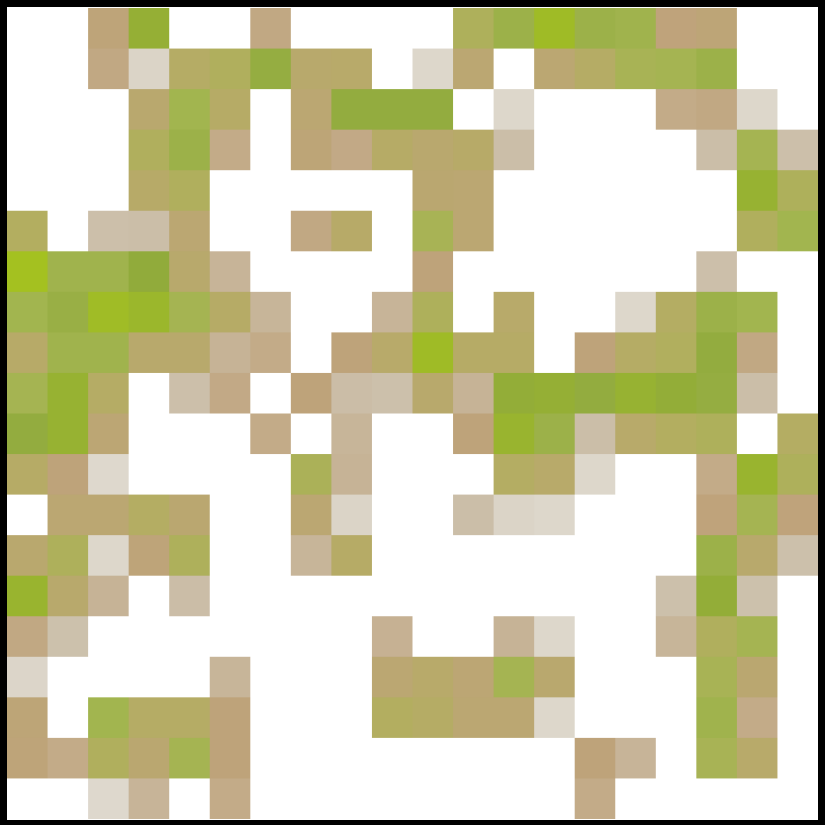
NDVI for Intensive Site



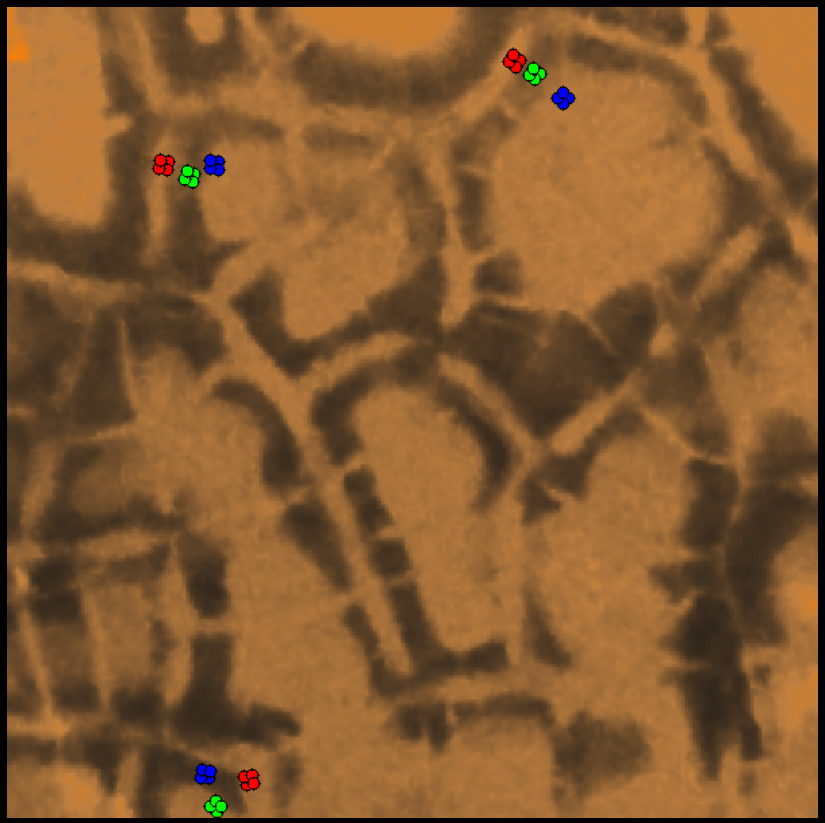
10 Clusters for Intensive Site



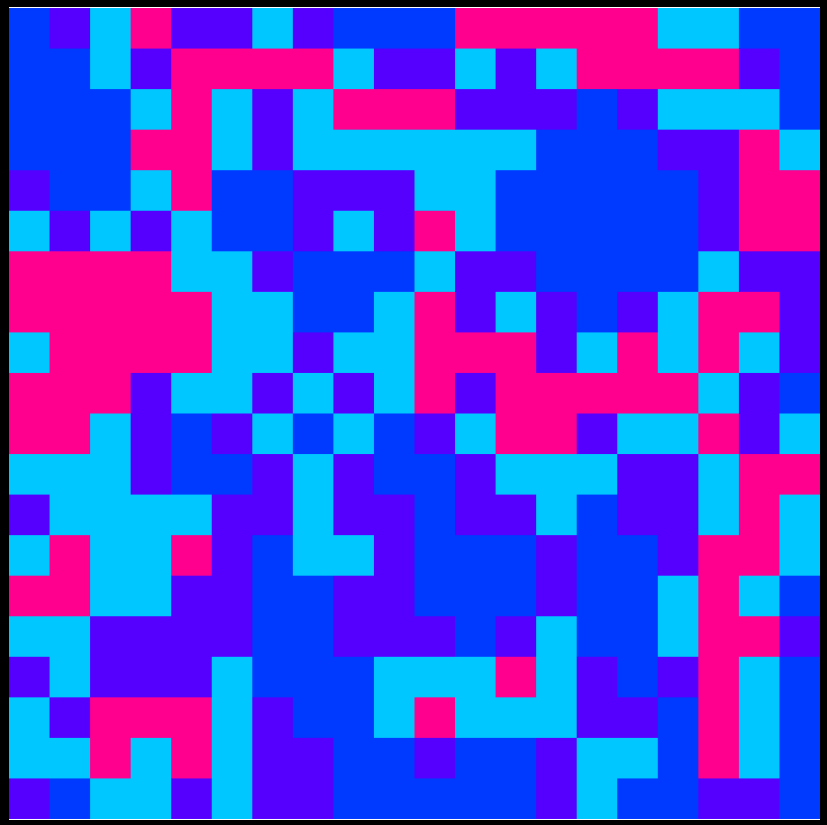
LiDAR for Site A



NDVI for Site A

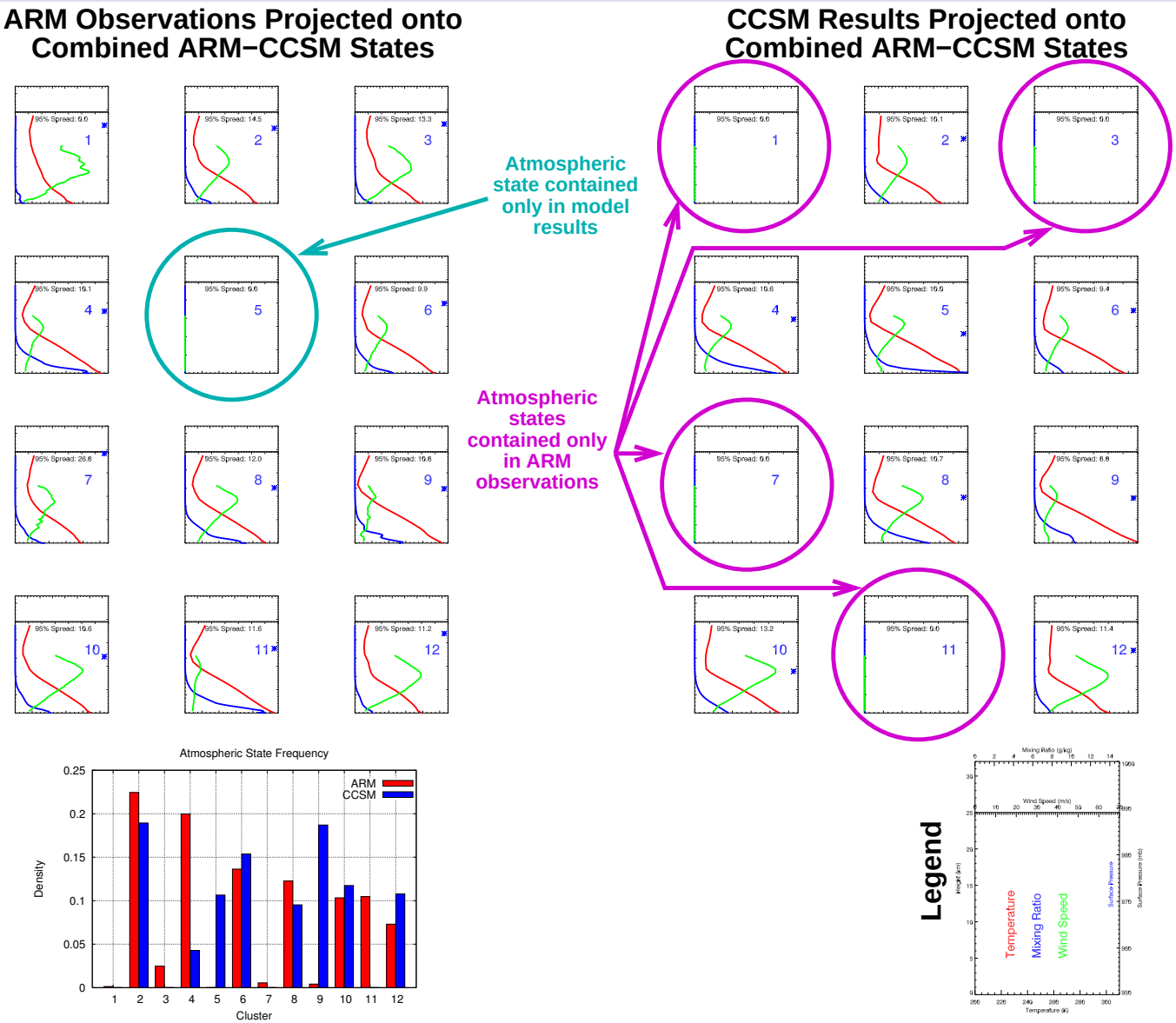


LiDAR for Site A



10 Clusters for Site A

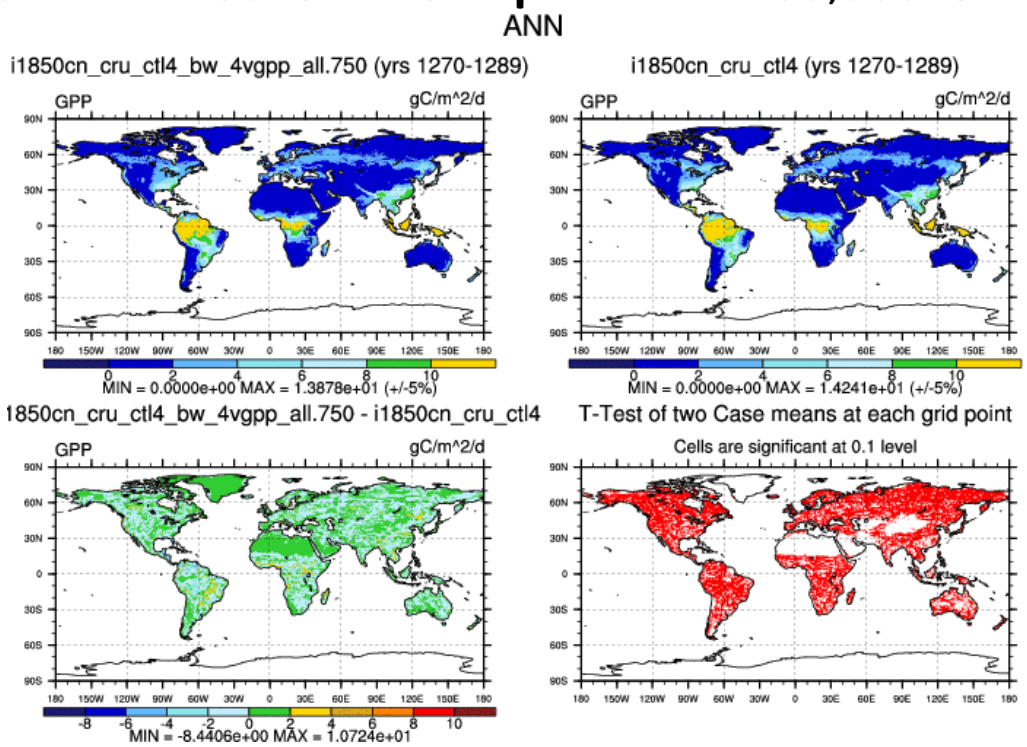
Atmospheric States from Observations and Models

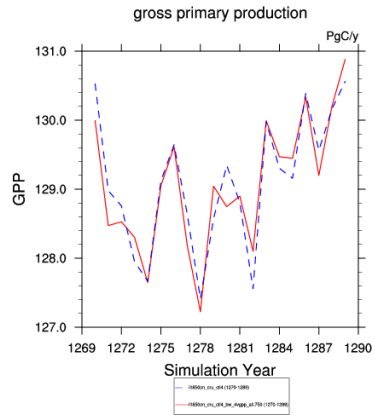


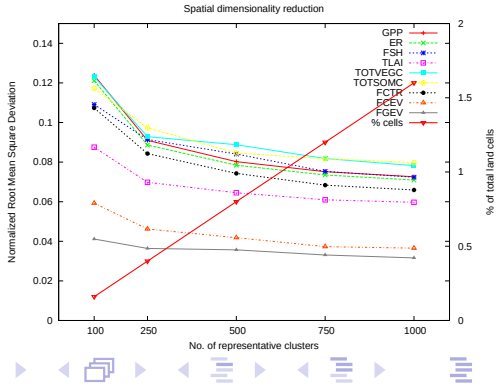
Spatial Dimensionality Reduction for Sensitivity Analysis

- Global CLM simulations at $0.5^\circ \times 0.5^\circ$ have ~60,000 grid cells that must be modeled in hundreds of 100–1000 y simulations, which is computationally untenable.
- Cluster analysis uses the CRU-NCEP climate data, plant functional type (PFT) characteristics, and steady-state modeled quantities.

GPP for 750 Cells Compared with 60,000 Cells







Acknowledgments

This research was sponsored by the U.S. Department of Agriculture Forest Service, Eastern Forest Environmental Threat Assessment Center (EFETAC). This research used resources of the National Center for Computational Science at Oak Ridge National Laboratory, which is managed by UT-Battelle, LLC, for the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

References

William W. Hargrove, Joseph P. Spruce, Gerald E. Gasser, and Forrest M. Hoffman. Toward a national early warning system for forest disturbances using remotely sensed phenology. *Photogramm. Eng. Rem. Sens.*, 75(10):1150–1156, October 2009.

Forrest M. Hoffman. Analysis of reflected spectral signatures and detection of geophysical disturbance using hyperspectral imagery. Master's thesis, Department of Physics and Astronomy, University of Tennessee, Knoxville, November 2004.

Michael A. White, Forrest M. Hoffman, William W. Hargrove, and Ramakrishna R. Nemani. A global framework for monitoring phenological responses to climate change. *Geophys. Res. Lett.*, 32(4):L04705, February 2005. doi:10.1029/2004GL021961.