Integrating Unsupervised Classification and Expert Knowledge to Develop Phenoregion Maps Using Remotely Sensed Imagery

Forrest M. Hoffman[†‡], Jitendra Kumar[‡], William W. Hargrove*

[†]University of California - Irvine, [‡]Oak Ridge National Laboratory, and *USDA Forest Service

November 17, 2013

4th SC Workshop on Petascale (Big) Data Analytics: Challenges and Opportunities

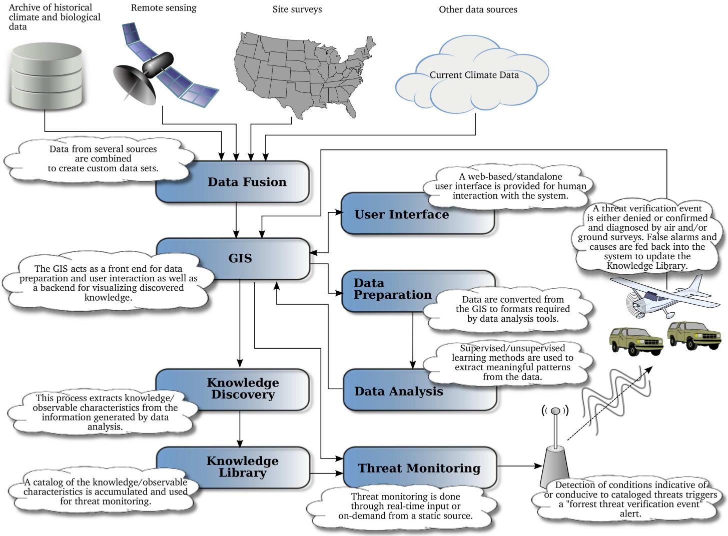
Denver, Colorado, USA

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.

Design Plan for the *ForWarn* Early Warning 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.

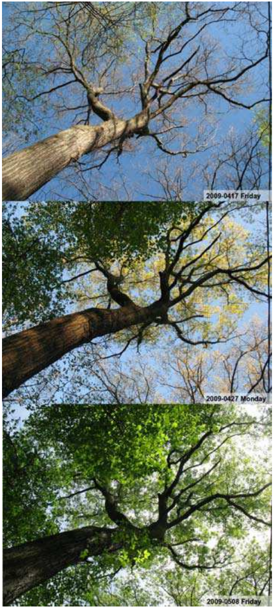
$$\mathrm{NDVI} = \frac{(\sigma_{\mathrm{nir}} - \sigma_{\mathrm{red}})}{(\sigma_{\mathrm{nir}} + \sigma_{\mathrm{red}})} \tag{1}$$

- 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

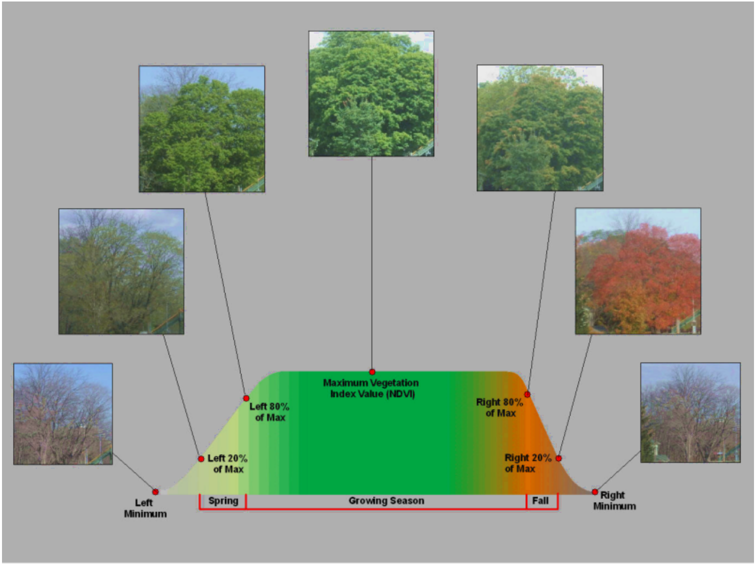
- The Moderate Resolution Imaging Spectroradiometer (MODIS) is a key instrument aboard the Terra (EOS AM, N$\rightarrow$S) and Aqua (EOS PM, S$\rightarrow$N) satellites.
- Both view the entire surface of Earth every 1 to 2 days, acquiring data in 36 spectral bands.
- The MOD 13 product provides Gridded Vegetation Indices (NDVI and EVI) to characterize vegetated surfaces.
- Available are 6 products at varying spatial (250 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 250 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.
- *ForWarn* 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.



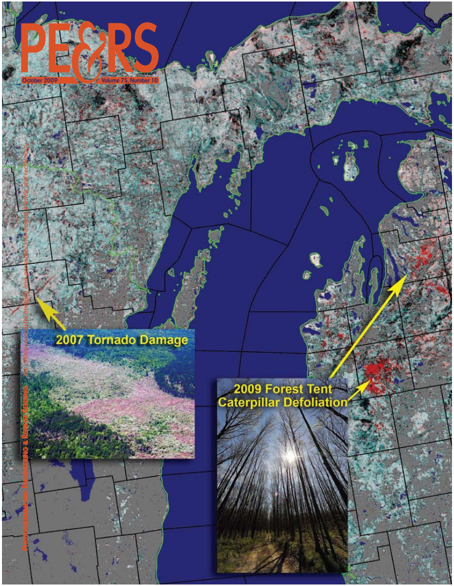
Up-looking photos of a scarlet oak showing the timing of leaf emergence in the spring (Hargrove et al., 2009).

Annual Greenness Profile Through Time

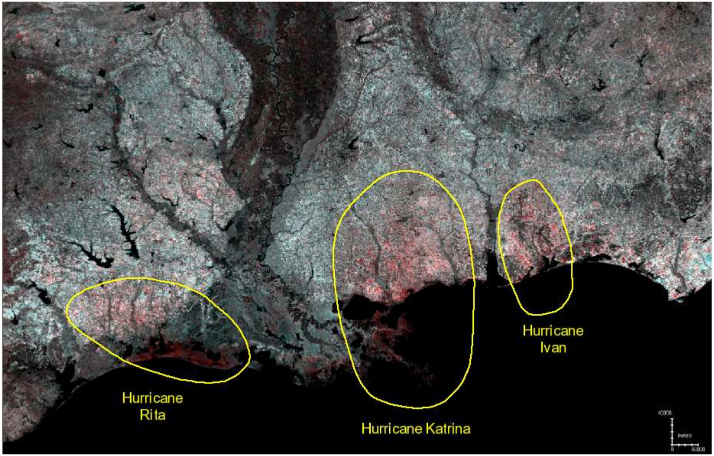


- To detect vegetation disturbances, the current NDVI measurement is compared with the normal, expected baseline for the same location.
- Substantial decreases from the baseline represent potential disturbances.
- Any increases over the baseline may represent vegetation recovery.
- Maximum, mean, or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

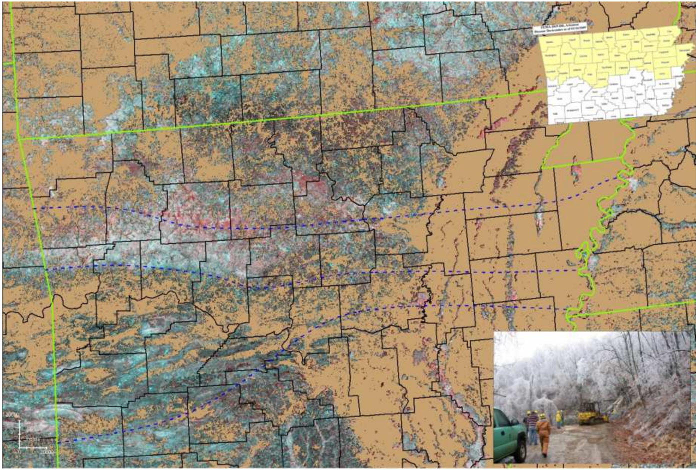


Three Hurricanes

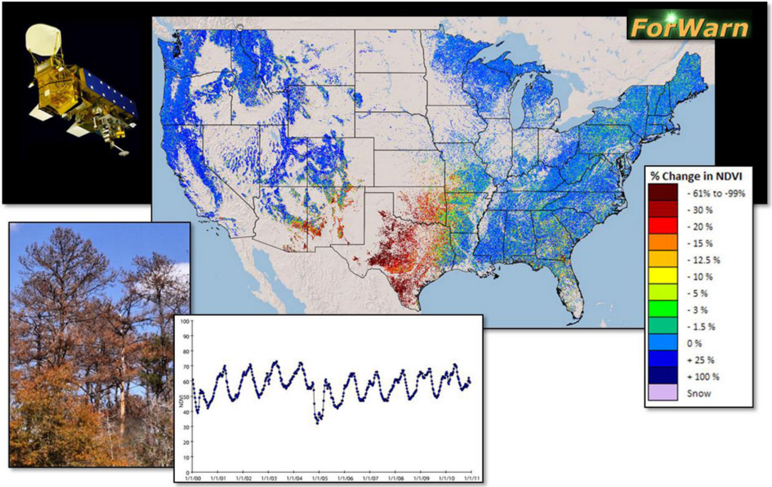


Computed by assigning 2006 20% left value to green & blue, and 20% left from 2004 to red (Hargrove et al., 2009). Red depicts areas of reduced greenness, primarily east of storm tracks and in marshes.

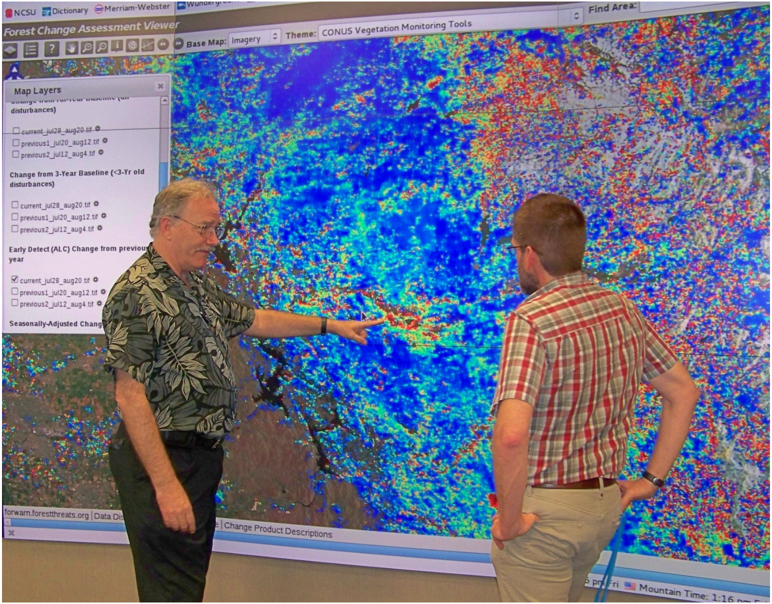
Arkansas Ozarks Ice Storm, Jan. 26–29, 2009



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.

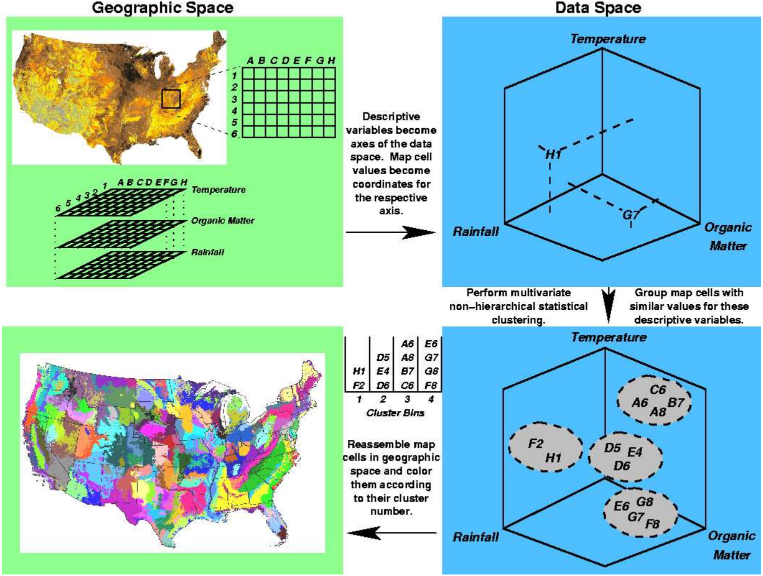


ForWarn is a forest change recognition and tracking system that uses high-frequency, moderate resolution satellite data to provide near real-time forest change maps for the continental United States that are updated every eight days. Maps and data products are available in the **Forest Change Assessment Viewer** at `http://forwarn.forestthreats.org/fcav/`



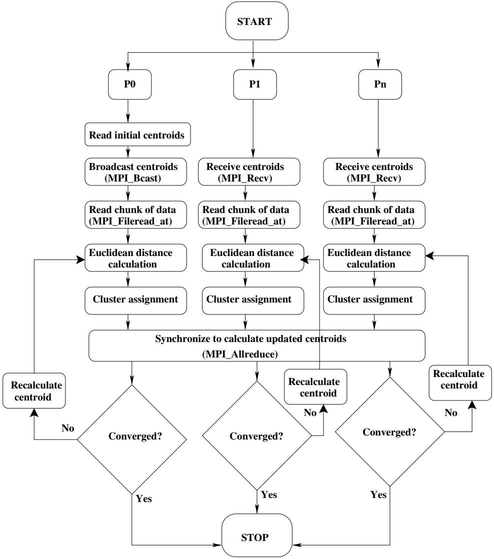
ForWarn researchers get EVEREST-sized look at woodland disturbances

Geospatiotemporal Data Mining



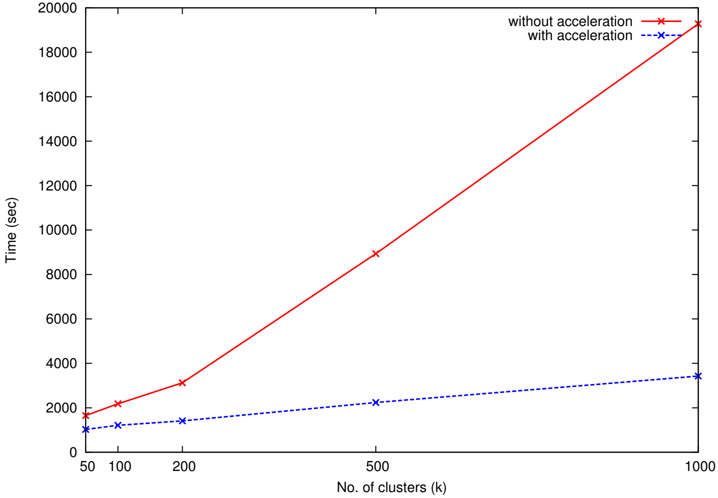
Parallel Cluster Analysis

- We developed a parallel "*masterless*" *k*-means cluster analysis algorithm.
- Acceleration technique exploits the triangular inequality to dramatically reduce the number of distance calculations
- Optimized parallel I/O:
  - Lustre tuning and optimization for Spider filesystem at the Oak Ridge Leadership Computing Facility (OLCF)
  - Two-stage parallel I/O scheme



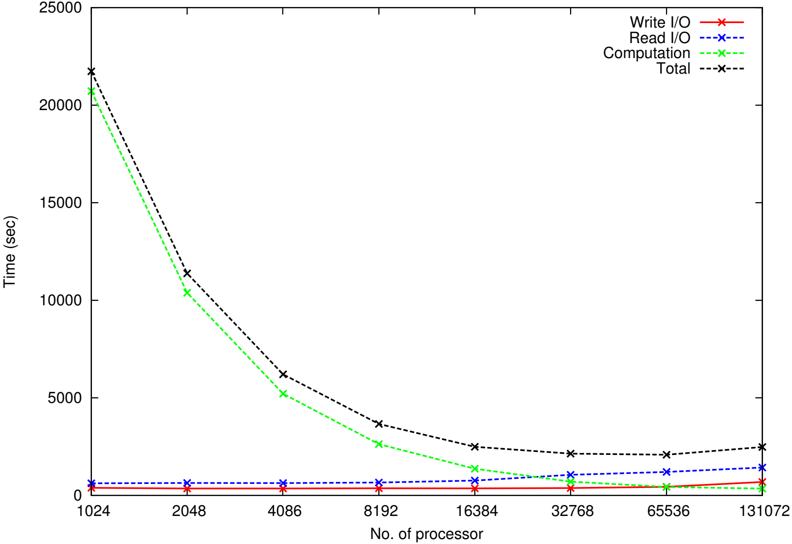
Parallel k-means algorithm performance

Computation time for increasing problem size (k) using 1024 processors on the OLCF Titan supercomputer



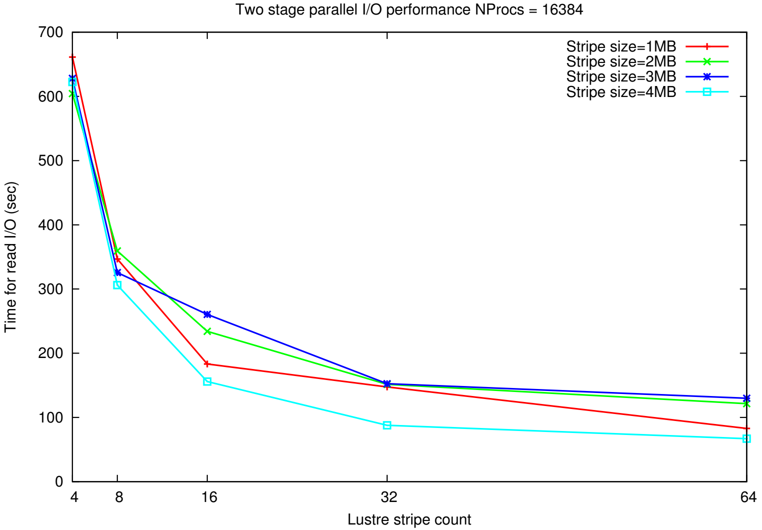
Parallel k-means computational performance

Weak scaling test for k=1000 on the OLCF Titan supercomputer



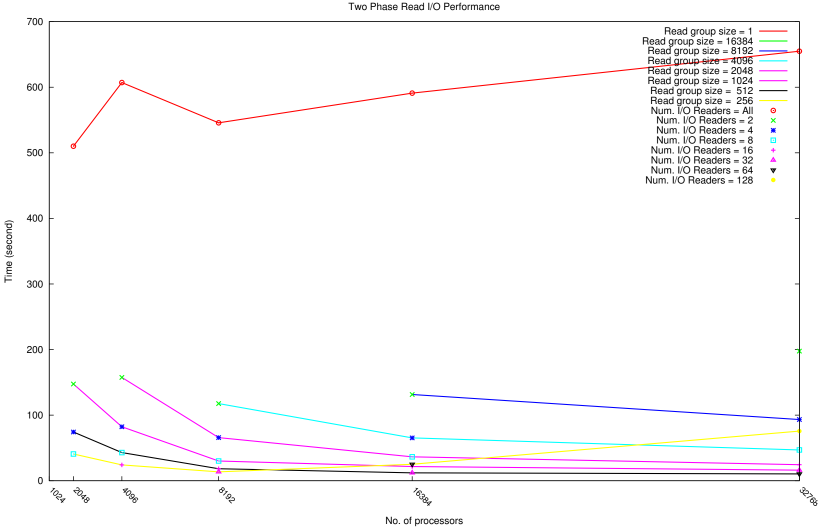
Lustre filesystem tuning for optimized I/O performance

Tuning Lustre stripe size/count on the OLCF Titan supercomputer



Two stage parallel I/O performance

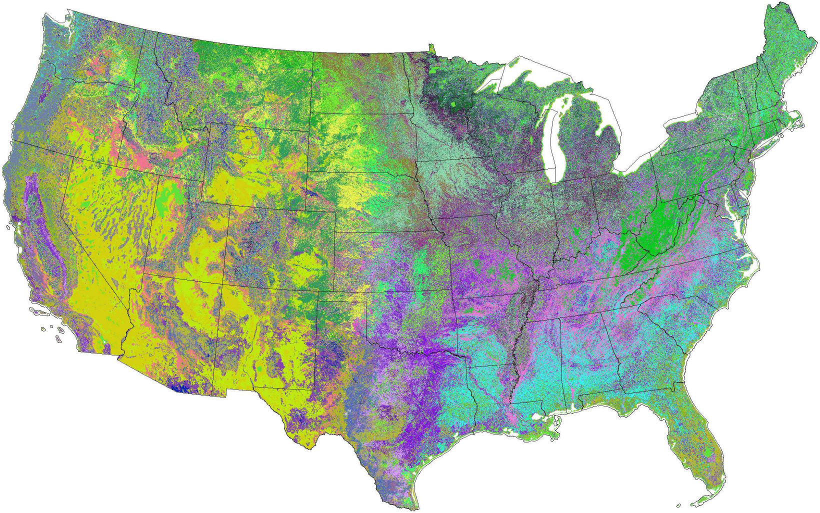
Two stage I/O for 16384 cores on the OLCF Titan supercomputer



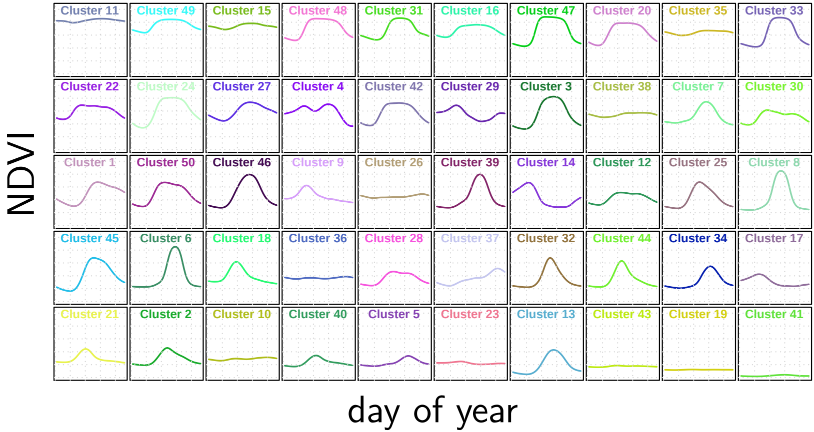
Clustering MODIS NDVI into Phenoregions

- Hoffman and Hargrove previously used *k*-means clustering to detect brine scars from hyperspectral data (Hoffman, 2004) and to classify phenologies from monthly climatology and 17 years of 8 km NDVI from AVHRR (White et al., 2005).
- This data mining approach requires high performance computing to analyze the entire body of the high resolution MODIS NDVI record for the continental U.S.
- $>$87B NDVI values, consisting of $\sim$146.4M cells for the CONUS at 250 m resolution with 46 maps per year for 13 years (2000–2012), analyzed using *k*-means clustering.
- The annual traces of NDVI for every year and map cell are combined into one 327 GB single-precision binary data set of 46-dimensional observation vectors.
- Clustering yields 13 phenoregion maps in which each cell is classified into one of *k* phenoclasses that represent prototype annual NDVI traces.

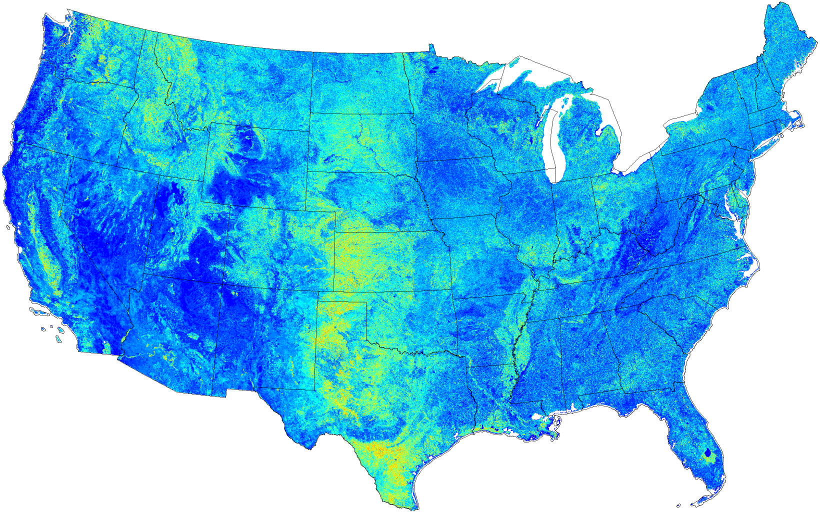
50 Phenoregions for year 2012 (Random Colors)



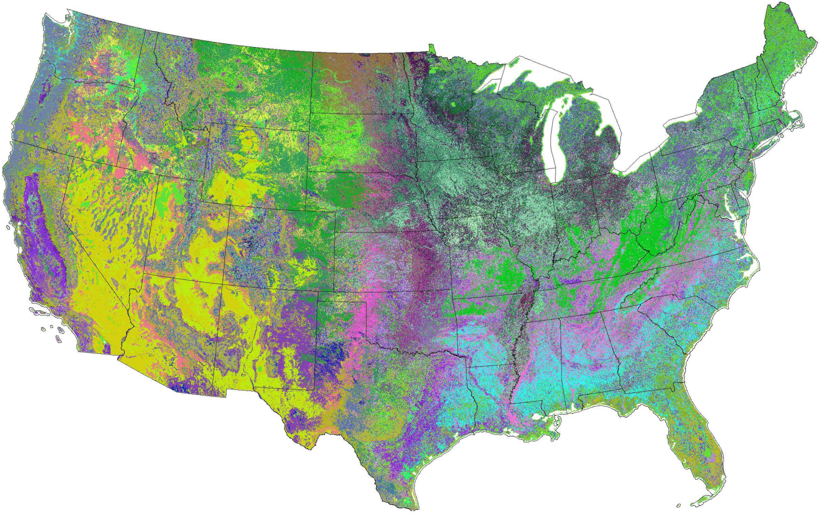
50 Phenoregion Prototypes (Random Colors)



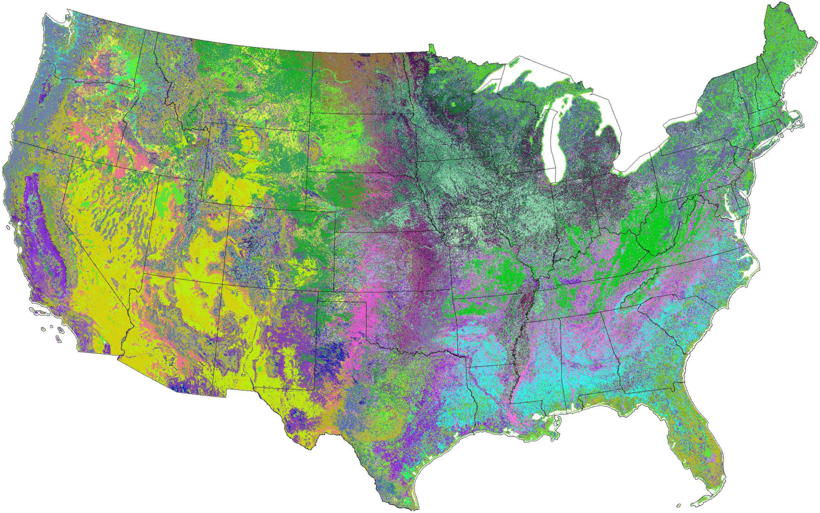
50 Phenoregions Persistence



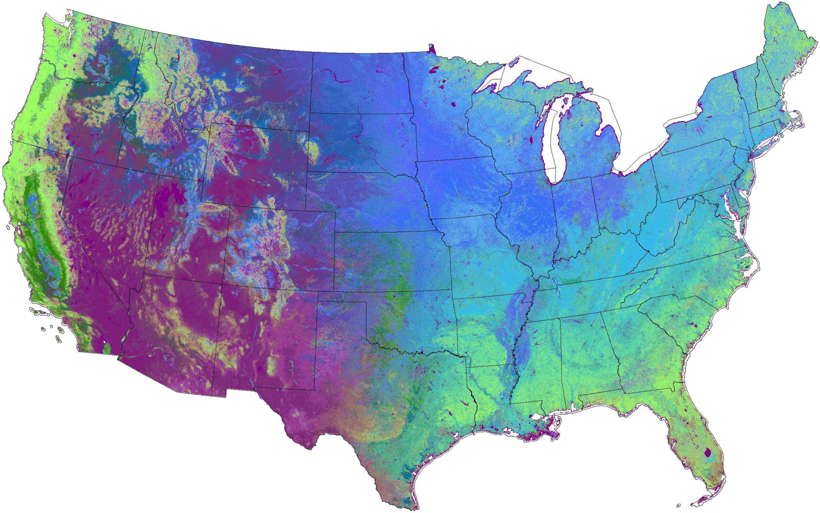
50 Phenoregions Mode (Random Colors)



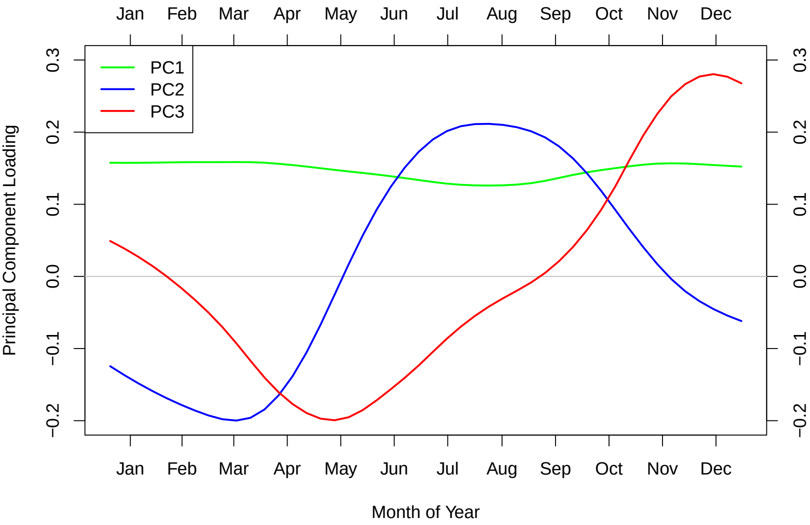
50 Phenoregions Max Mode (Random Colors)



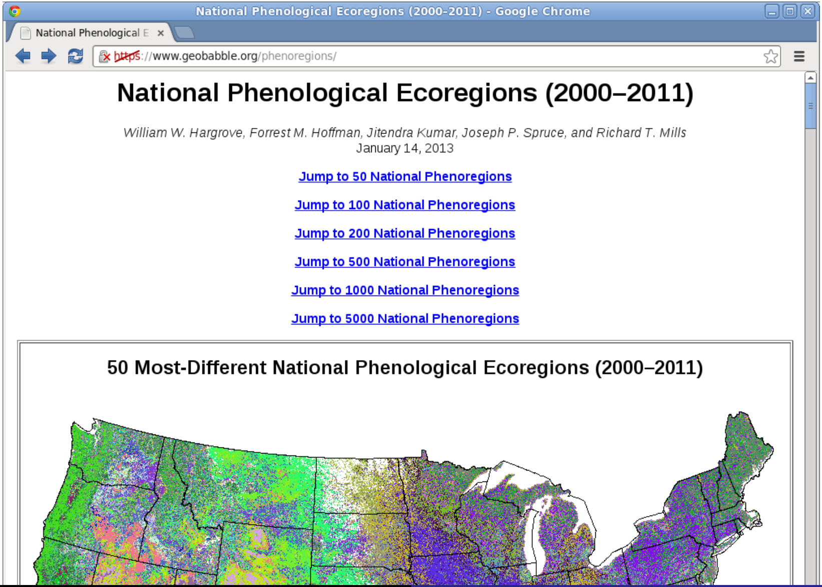
50 Phenoregions Max Mode (Similarity Colors)



50 Phenoregions Max Mode (Similarity Colors Legend)

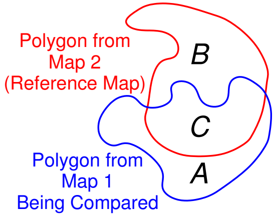


Phenoregions Clearinghouse



Mapcurves: A Method for Comparing Categorical Maps

- Hargrove et al. (2006) developed a method for quantitatively comparing categorical maps that is
  - independent of differences in resolution,
  - independent of the number of categories in maps, and
  - independent of the directionality of comparison.



Goodness of Fit (GOF) is a unitless measure of spatial overlap between map categories:

$$\text{GOF} = \sum_{\text{polygons}} \frac{C}{B+C} \times \frac{C}{A+C}$$

- GOF provides "credit" for the area of overlap, but also "debit" for the area of non-overlap.
- Mapcurves comparisons allow us to reclassify any map in terms of any other map (*i.e.*, color Map 2 like Map 1).
- A greyscale GOF map shows the degree of correspondence between two maps based on the highest GOF score.

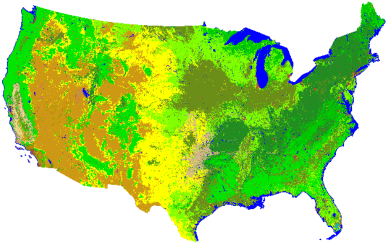
Two 2-Way Comparisons with Land Cover Maps

Cluster	IGBP Land Cover	Olson's Global Ecoregions
1	Grasslands	cool grasses and shrubs
2	Evergreen Needleleaf Forest	cool conifer forest
3	Croplands	corn and beans cropland
4	Cropland/Natural Vegetation Mosaic	cool forest and field
5	Open Shrublands	semi desert sage
6	Grasslands	cool conifer forest
7	Grasslands	hot and mild grasses and shrubs
8	Cropland/Natural Vegetation Mosaic	cool forest and field
9	Grasslands	hot and mild grasses and shrubs
10	Open Shrublands	semi desert shrubs
11	Croplands	corn and beans cropland
12	Evergreen Needleleaf Forest	conifer forest
13	Open Shrublands	semi desert shrubs
14	Savannas	savanna (woods)
15	Grasslands	hot and mild grasses and shrubs
16	Evergreen Needleleaf Forest	cool conifer forest
17	Evergreen Needleleaf Forest	cool conifer forest
18	Evergreen Needleleaf Forest	cool conifer forest
19	Deciduous Broadleaf Forest	deciduous broadleaf forest
20	Deciduous Broadleaf Forest	deciduous broadleaf forest
21	Deciduous Broadleaf Forest	cool broadleaf forest
22	Open Shrublands	semi desert sage
23	Grasslands	cool grasses and shrubs
24	Grasslands	semi desert sage
25	Croplands	woody savanna

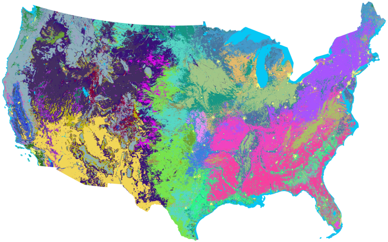
Two 2-Way Comparisons with Land Cover Maps

Cluster	IGBP Land Cover	Olson's Global Ecoregions
26	Evergreen Needleleaf Forest	conifer forest
27	Evergreen Needleleaf Forest	cool conifer forest
28	Water	inland water
29	Croplands	woody savanna
30	Grasslands	cool grasses and shrubs
31	Croplands	cool crops and towns
32	Water	inland water
33	Grasslands	cool grasses and shrubs
34	Open Shrublands	semi desert shrubs
35	Grasslands	hot and mild grasses and shrubs
36	Deciduous Broadleaf Forest	cool broadleaf forest
37	Evergreen Needleleaf Forest	deciduous broadleaf forest
38	Evergreen Needleleaf Forest	cool conifer forest
39	Grasslands	hot and mild grasses and shrubs
40	Croplands	broadleaf crops
41	Cropland/Natural Vegetation Mosaic	cool fields and woods
42	Croplands	corn and beans cropland
43	Mixed Forests	cool broadleaf forest
44	Croplands	deciduous broadleaf forest
45	Cropland/Natural Vegetation Mosaic	cool forest and field
46	Cropland/Natural Vegetation Mosaic	crops, grass, shrubs
47	Evergreen Needleleaf Forest	crops, grass, shrubs
48	Croplands	corn and beans cropland
49	Deciduous Broadleaf Forest	cool broadleaf forest
50	Grasslands	cool grasses and shrubs

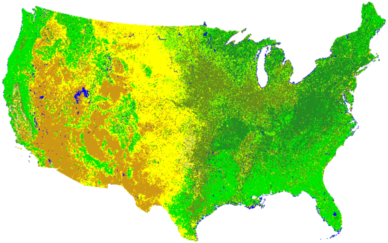
Phenoregions Reclassed Using Land Cover Types



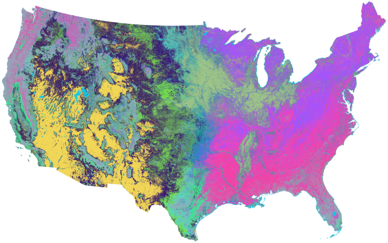
(a) IGBP Land Cover



(c) Olson's Global Ecoregions

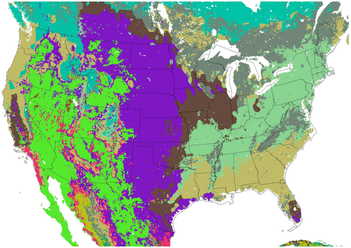


(b) 50 Phenoregions Reclassed

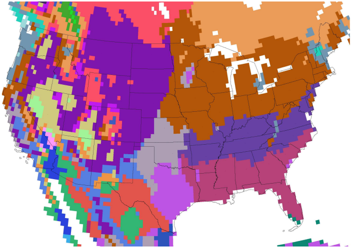


(d) 50 Phenoregions Reclassed

Expert-Derived Land Cover/Vegetation Type Maps



Foley Land Cover



Holdridge Life Zones

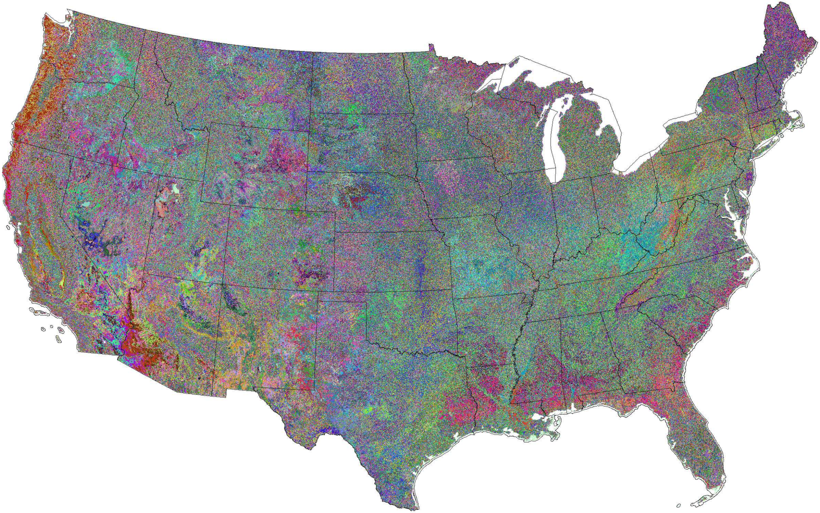
	Expert Map	# Cats
1.	DeFries UMd Vegetation	12
2.	Foley Land Cover	14
3.	Fedorova, Volkova, and Varlyguin World Vegetation Cover	31
4.	GAP National Land Cover	578
5.	Holdridge Life Zones	25
6.	Küchler Types	117
7.	BATS Land Cover	17
8.	IGBP Land Cover	16
9.	Olson Global Ecoregions	49
10.	Seasonal Land Cover Regions	194
11.	USGS Land Cover	24
12.	Leemans-Holdridge Life Zones	26
13.	Matthews Vegetation Types	19
14.	Major Land Resource Areas	197
15.	National Land Cover Database 2006	16
16.	Wilson, Henderson, & Sellers Primary Vegetation Types	23
17.	Landfire Vegetation Types	443

Label Stealing: Having your cake and eating it too!

- Clustering is an unsupervised classification technique, so phenoregions have no descriptive labels like **Eastern Deciduous Forest Biome**.
- **Label stealing** allows us to perform automated "supervision" to "steal" the best human-created descriptive labels to assign to phenoregions.
- We employ the **Mapcurves GOF** to select the best ecoregion labels from ecoregionalizations drawn by human experts.
- We consider an entire library of ecoregion and land cover maps, and choose the label with the highest GOF score for every phenoregion polygon.

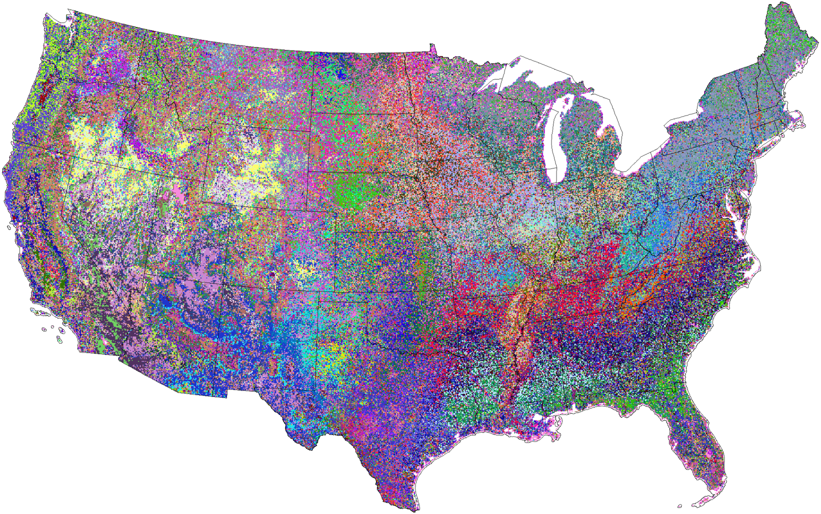
Patchwork Crazy Quilt of Multiple Land Cover Types

1000 Phenoregions Max Under (Random Colors)

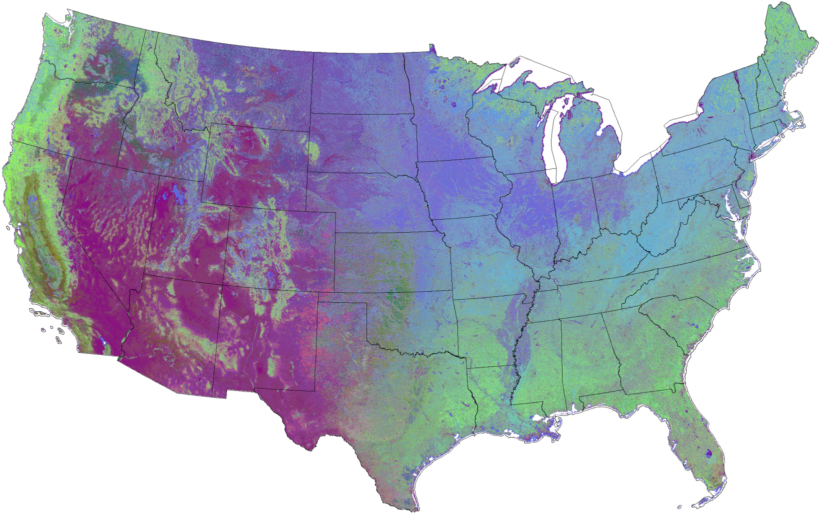


Category	Land Cover Label	Land Cover Map
1	Acadian Low-Elevation Spruce-Fir-Hardwood Forest	landfire vegetation type
2	Agriculture-Pasture and Hay	landfire vegetation type
3	Alpine meadows & barren	ktlamb
4	Barren	landcover.slcr
5	Barren or Sparsely Vegetated	landcover.usgs
6	Bluestem/Grama	ktlamb
7	Bluestem Hills, MLRA 76	mlra
8	Boreal Evergreen Forest/Woodland	foleylandcover
9	Boreal	fvvcode
10	Boreal moist forest	holdridgezonesnormal
11	Broadleaf Deciduous Forest	landcover.usgs
12	Brown Glaciated Plain, MLRA 52	mlra
13	California Central Valley and Southern Coastal Grassland	GAP 240m laea
14	California Central Valley Mixed Oak Savanna	GAP 240m laea
15	California oakwoods	ktlamb
16	California steppe	ktlamb
.	.	.
.	.	.
.	.	.
222	Warm temperate moist forest	holdridgezonesnormal
223	Warm Temperate Moist Forest	leemansholdridgezones
224	[water]	ktlamb
225	Water	landcover.slcr
226	Western Great Plains Mesquite Woodland and Shrubland	GAP 240m laea
227	Western Great Plains Shortgrass Prairie	landfire vegetation type
228	Western ponderosa	ktlamb
229	Western Rio Grande Plain, MLRA 83B	mlra
230	Western spruce/Fir	ktlamb
231	Wheatgrass/Bluegrass	ktlamb
232	Wheatgrass/Needlegrass	ktlamb
233	Willamette and Puget Sound Valleys, MLRA 2	mlra
234	Woodland/Cropland Mosaic	landcover.usgs
235	Woody wetlands	NLCD2006 240m laea

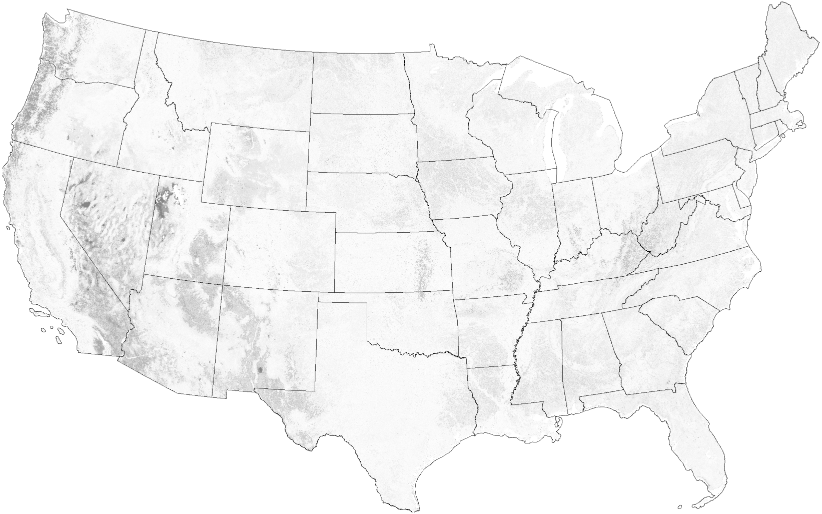
1000 Phenoregions Reclassed into 235 Land Cover Types



1000 Phenoregions Reclassed into 235 Land Cover Types



1000 Phenoregions Reclassed Goodness of Fit

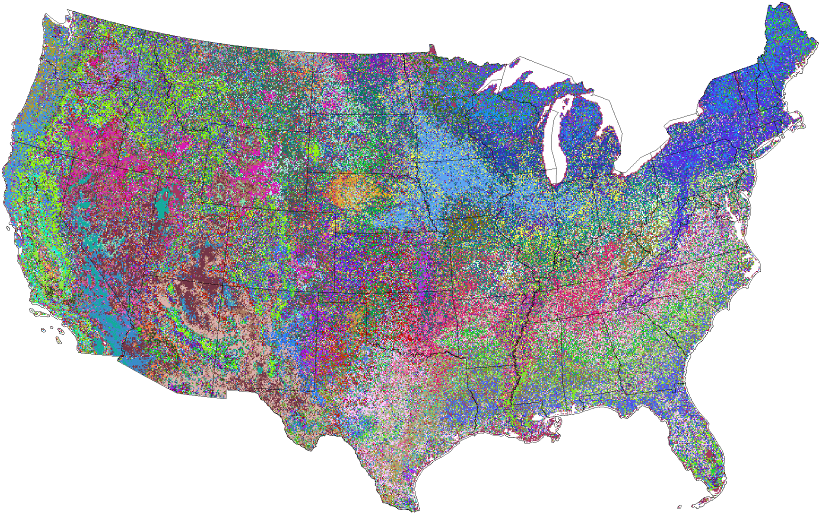


Composition of the 235 Land Cover Types Map

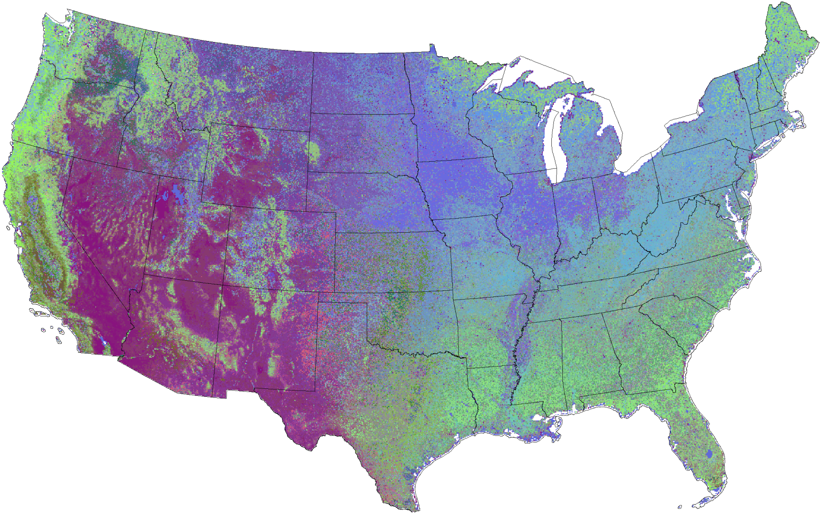
	Map	Cats	WCats	WClusts	%Area
10.	Seasonal Land Cover Regions	194	43	160	19.45
9.	Olson Global Ecoregions	49	12	96	12.36
3.	Fedorova, Volkova, and Varlyguin World Vegetation Cover	31	4	93	10.69
17.	Landfire Vegetation Types	443	27	85	9.09
6.	Küchler Types	117	34	81	7.87
14.	Major Land Resource Areas	197	42	107	7.18
12.	Leemans-Holdridge Life Zones	26	8	54	5.27
11.	USGS Land Cover	24	7	21	4.85
4.	GAP National Land Cover	578	19	124	4.48
5.	Holdridge Life Zones	25	9	38	4.15
2.	Foley Land Cover	14	7	48	3.86
15.	National Land Cover Database 2006	16	8	47	3.24
13.	Matthews Vegetation Types	19	5	18	2.49
16.	Wilson, Henderson, & Sellers Primary Vegetation Types	23	2	9	1.46
7.	BATS Land Cover	17	4	10	1.23
8.	IGBP Land Cover	16	3	4	0.80
1.	DeFries UMd Vegetation	12	2	5	0.25
	TOTAL		**235**	**1000**	**100%**

#	Category	Land Cover Label	Land Cover Map	Percent Area
1	176	Subboreal	fvvcode	5.28%
2	179	Subtropical	fvvcode	4.25%
3	73	Evergreen Coniferous Forest	landcover.usgs	3.87%
4	67	Open Shrubland	foleylandcover	3.74%
5	35	corn and beans cropland	landcover.oge	3.48%
6	29	cool conifer forest	landcover.oge	2.93%
7	32	Cool temperate moist forest	holdridgezonesnormal	2.55%
8	64	Desert Shrubland/Grassland (Creosote, Saltbush, Mesquite, Sand Sage)	landcover.slcr	2.27%
9	55	Deciduous Forest (Oak, Hickory, Sweet Gum, Southern Pines) with Cropland and Pasture	landcover.slcr	2.25%
10	28	cool broadleaf forest	landcover.oge	2.23%
11	66	Sparsely Vegetated Desert Shrublands	landcover.slcr	2.14%
12	188	Warm temperate moist forest	holdridgezonesnormal	2.06%
13	180	Subtropical moist forest	holdridgezonesnormal	2.05%
14	160	semi desert sage	landcover.oge	1.87%
.	.	.	.	
.	.	.	.	
.	.	.	.	
187	120	Northern hardwoods/Spruce	ktlamb	0.01%
188	102	Laurentian-Acadian Alkaline Conifer-Hardwood Swamp	landfire vegetation type	0.01%
189	51	NASS-Vineyard	landfire vegetation type	0.01%
190	2	Alpine meadows & barren	ktlamb	0.01%
191	143	Pseudotsuga menziesii Forest Alliance	landfire vegetation type	0.01%
192	134	Olympic and Cascade Mountains, MLRA 3	mlra	0.01%
193	79	Evergreen Needleleaf Forest (Lodgepole Pine and Douglas Fir)	landcover.slcr	0.01%
194	125	North Pacific Maritime Mesic Subalpine Parkland	GAP 240m laea	0.00%
195	80	Evergreen Needleleaf Forest (Lodgepole Pine, Englemann Spruce, Ponderosa Pine)	landcover.slcr	0.00%
196	157	Saltbrush/Greasewood	ktlamb	0.00%
197	106	Mediterranean California Red Fir Forest	GAP 240m laea	0.00%

1000 Phenoregions Reclassed into 197 Land Cover Types



1000 Phenoregions Reclassed into 197 Land Cover Types



Uses for Label Stealing

- Borrowing ecoregion, land cover, or vegetation type labels for unsupervised classifications.
- Automated attribution of disturbance agents through comparison of a *ForWarn* disturbance map with ADS aerial sketchmaps, wildfire perimeters, tornado track maps, and fuel treatment maps through time.
- Determination of the most important driving variable for phenoregions maps through comparison with separate maps of slope, aspect, solar input, elevation, soil types, etc.
- Automated recognition of species composition of forest vegetation through comparison of a phenoregions map with individual tree species range maps.

AGU Fall Meeting Session

IN006. Big Data in the Geosciences: New Analytics Methods and Parallel Algorithms

Co-conveners: Jitendra Kumar (ORNL), Robert Jacob (ANL), Don Middleton (NCAR), and Forrest Hoffman (ORNL)

Confirmed Invited Speakers:

- Gary Geernaert (U.S. Dept. of Energy)
- Matt Hancher (Google Earth Engine)
- Jeff Daily (Pacific Northwest National Laboratory)
- William Hargrove (USDA Forest Service)

Earth and space science data are increasingly large and complex, often representing long time series or high resolution remote sensing, making such data difficult to analyze, visualize, interpret, and understand. The proliferation of heterogeneous, multi-disciplinary observational and model data have rendered traditional means of analysis and integration ineffective. This session focuses on development and applications of data analytics (statistical, data mining, machine learning, etc.) approaches and software for the analysis, assimilation, and synthesis of large or long time series Earth science data that support integration and discovery in climatology, hydrology, geology, ecology, seismology, and related disciplines.

Fifth Workshop on Data Mining in Earth System Science

ICCS 2014: "Big Data Meets Computational Science"

Fifth Workshop on Data Mining in Earth System Science (DMESS 2014)

Co-conveners: Forrest Hoffman, Jitendra Kumar (ORNL), J. Walter Larson (Australian National University), Miguel D. Mahecha (Max Planck Institute for Biogeochemistry)

The "explosion" of heterogeneous, multi-disciplinary Earth science data has rendered traditional means of integration and analysis ineffective, necessitating the application of new analysis methods and the development of highly scalable software tools for synthesis, assimilation, comparison, and visualization. This workshop explores various data mining approaches to understanding Earth science processes, emphasizing the unique technological challenges associated with utilizing very large and long time series geospatial data sets. Especially encouraged are original research papers describing applications of statistical and data mining methods—including cluster analysis, empirical orthogonal functions (EOFs), genetic algorithms, neural networks, automated data assimilation, and other machine learning techniques—that support analysis and discovery in climate, water resources, geology, ecology, and environmental sciences research.

Full paper submissions are due December 15.

Acknowledgments

This research was sponsored by the U.S. Department of Agriculture Forest Service, Eastern Forest Environmental Threat Assessment Center (EFETAC) and the U.S. Department of Energy Biological and Environmental Research (BER) program. This research used resources of the Oak Ridge Leadership Computing Facility 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, University of Tennessee, Department of Physics and Astronomy, Knoxville, Tennessee, USA, November 2004.

Michael A. White, Forrest 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.

William W. Hargrove, Forrest M. Hoffman, and Paul F. Hessburg. Mapcurves: A quantitative method for comparing categorical maps. *J. Geograph. Syst.*, 8(2):187–208, July 2006. doi: 10.1007/s10109-006-0025-x.