

# Big Data in the Geosciences: Data Mining Methods for Characterizing Ecoregions, Designing Sampling Networks, Detecting Forest Threats, and Understanding Climate Change Predictions

Forrest M. Hoffman<sup>†</sup>,  
Jitendra Kumar<sup>†</sup>, and  
William W. Hargrove<sup>‡</sup>

<sup>†</sup>Oak Ridge National Laboratory and  
<sup>‡</sup>USDA Forest Service, Eastern Forest  
Environmental Threat Assessment  
Center (EFETAC)



**Smoky Mountains  
Computational Sciences  
and Engineering Conference**

**September 5, 2013**



U.S. DEPARTMENT OF  
**ENERGY**



**Climate Change  
Science Institute**

AT OAK RIDGE NATIONAL LABORATORY



**OAK RIDGE NATIONAL LABORATORY**

MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

Data Mining for Climate Change Model Intercomparison

Sampling Domain Representativeness

Developing Phenoregion Maps Using Remotely Sensed Imagery

# Data Mining for Climate Change Model Intercomparison

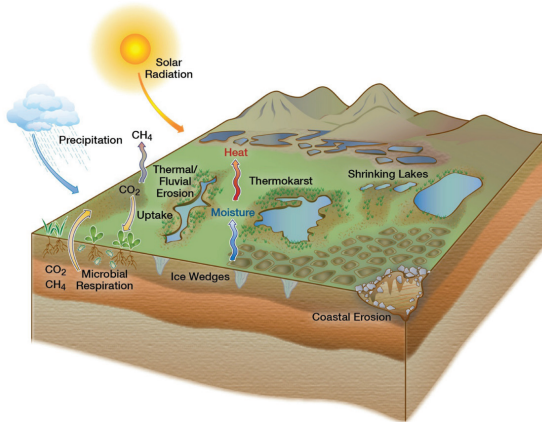
Hoffman et al. (2005)

Data Mining for Climate Change Model Intercomparison

Sampling Domain Representativeness

Developing Phenoregion Maps Using Remotely Sensed Imagery

# Next-Generation Ecosystem Experiments (NGEE Arctic) Representativeness and Scaling



*The Next-Generation Ecosystem Experiments (NGEE Arctic) project is supported by the Office of Biological and Environmental Research in the DOE Office of Science.*



# Quantitative Sampling Network Design

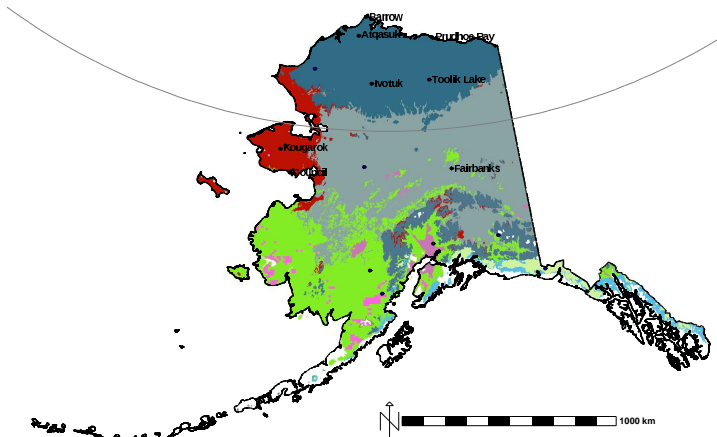
- ▶ Resource and logistical constraints limit the frequency and extent of observations, necessitating the development of a systematic sampling strategy that objectively represents environmental variability at the desired spatial scale.
- ▶ Required is a methodology that provides a quantitative framework for informing site selection and determining the representativeness of measurements.
- ▶ Multivariate spatiotemporal clustering (MSTC) was applied at the landscape scale ( $4 \text{ km}^2$ ) 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.

# 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

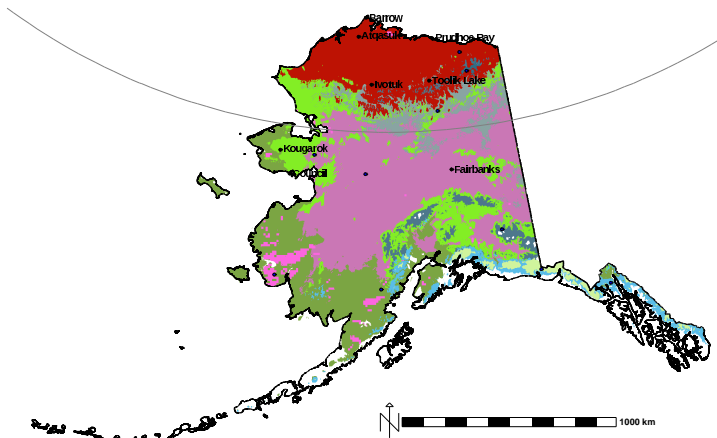
# 10 Alaska Ecoregions (2000–2009)



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

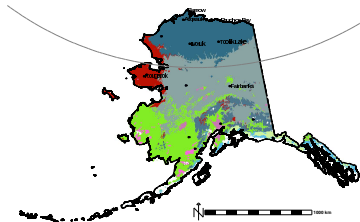


# 10 Alaska Ecoregions (2090–2099)

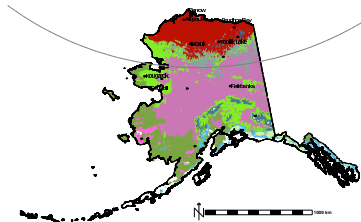


Each ecoregion is a different random color. Blue filled circles mark locations most representative of mean conditions of each region.

# 10 Alaska Ecoregions, Present and Future



2000–2009

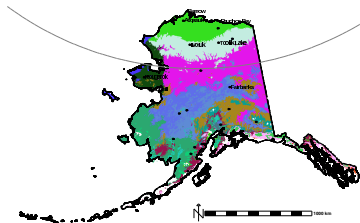


2090–2099

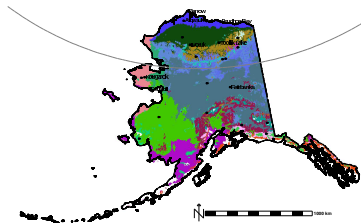
*Since the random colors are the same in both maps, a change in color represents an environmental change between the present and the future.*

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

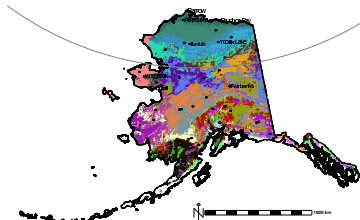


2090–2099

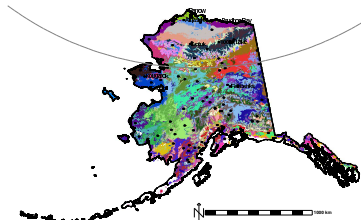
*Since the random colors are the same in both maps, a change in color represents an environmental change between the present and the future.*

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

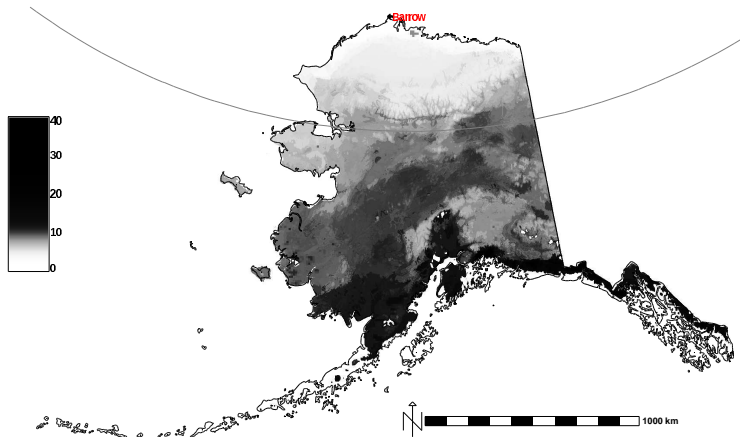
*Since the random colors are the same in both maps, a change in color represents an environmental change between the present and the future.*

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.

# NGEE Arctic Site Representativeness

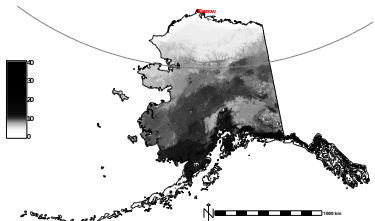
- ▶ This representativeness analysis uses the standardized  $n$ -dimensional data space formed from all input data layers.
- ▶ In this data space, the Euclidean distance between a sampling location (like Barrow) and every other point is calculated.
- ▶ These data space distances are then used to generate 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”

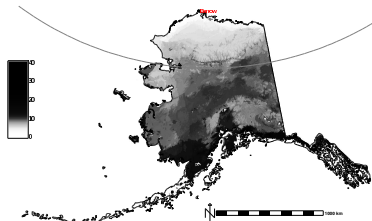


Light-colored regions are well represented and dark-colored regions are poorly represented by the sampling location listed in **red**.

# Present vs. Future Barrow-ness



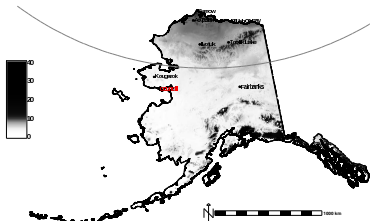
2000–2009



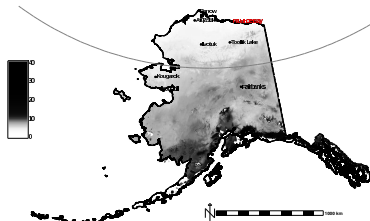
2090–2099

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

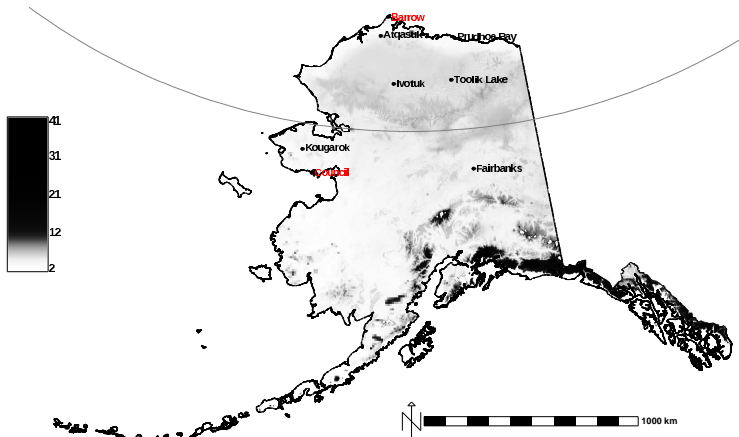


Prudhoe Bay

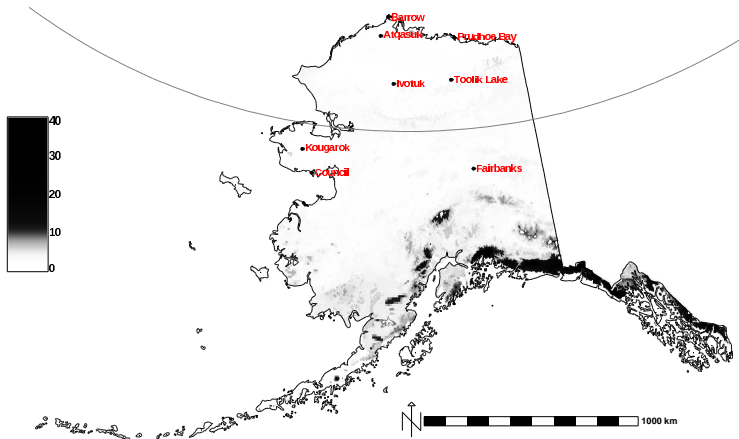
Representativeness analysis was performed for sites at Barrow, Council, Atkasuk, Ivotuk, Kougarok, Prudhoe Bay, Toolik Lake, and Fairbanks.



# 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

<b>Sites</b>				Toolik		Prudhoe	
	Council	Atqasuk	Ivotuk	Lake	Kougarok	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				Toolik		Prudhoe	
	Council	Atqasuk	Ivotuk	Lake	Kougarok	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

		<i>Future (2090–2099)</i>								
		Barrow	Council	Atqasuk	Ivotuk	Toolik Lake	Kougarok	Prudhoe Bay	Fairbanks	
<i>Present (2000–2009)</i>	Sites	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 is in press:

Hoffman, F. M., J. Kumar, R. T. Mills, and W. W. Hargrove (2013) "Representativeness-Based Sampling Network Design for the State of Alaska." *Landscape Ecol.*, in press.  
doi:10.1007/s10980-013-9902-0.

Data Mining for Climate Change Model Intercomparison

Sampling Domain Representativeness

Developing Phenoregion Maps Using Remotely Sensed Imagery



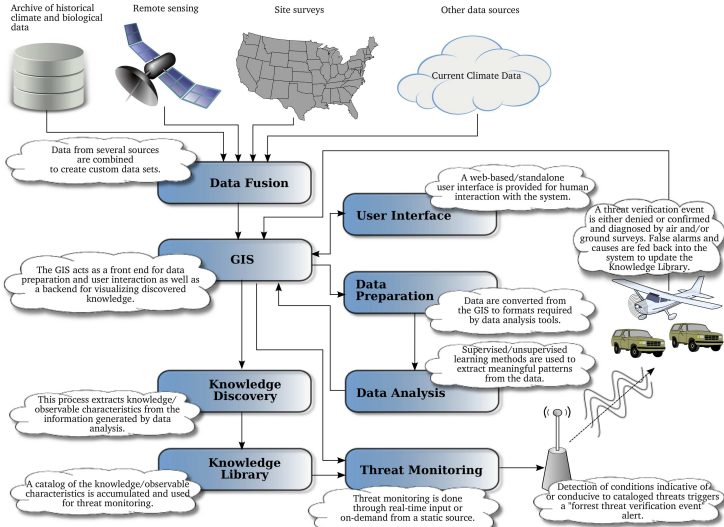
The USDA Forest Service, NASA Stennis Space Center, and DOE Oak Ridge National Laboratory are creating a system to monitor threats to U.S. forests and wildlands at two different scales:

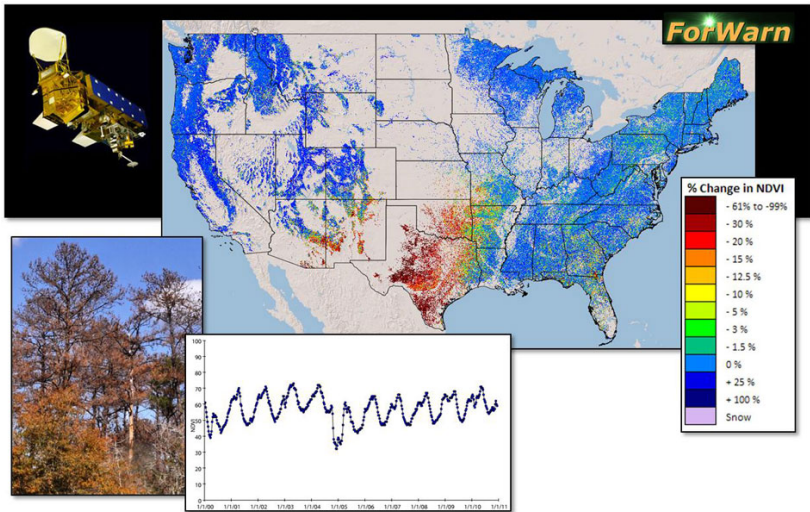
- ▶ **Tier 1: Strategic** — The *ForWarn System* that routinely monitors wide areas at coarser resolution, repeated frequently — a *change detection system* to produce alerts or warnings for particular locations may be of interest
- ▶ **Tier 2: Tactical** — Finer resolution airborne overflights and ground 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



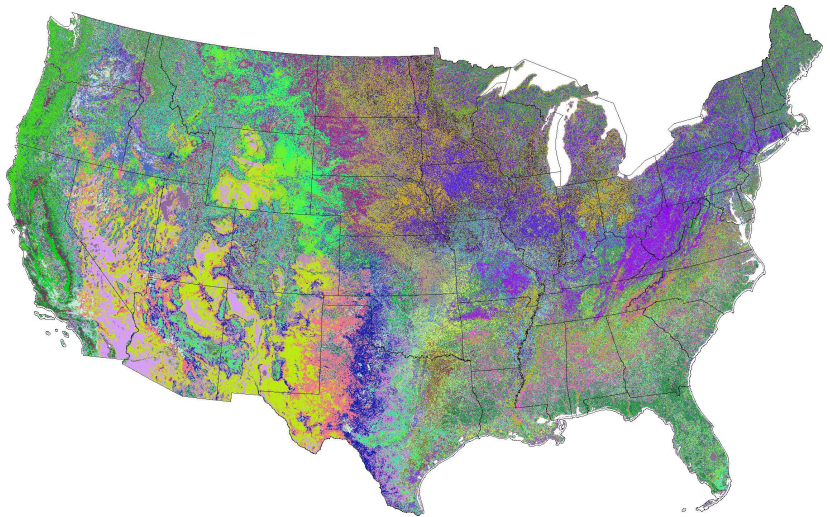


*ForWarn* is forest change recognition and tracking system that uses high-frequency, moderate resolution satellite data to provide near real-time forest change maps for the continental United States that are updated every eight days. Maps and data products are available in the **Forest Change Assessment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

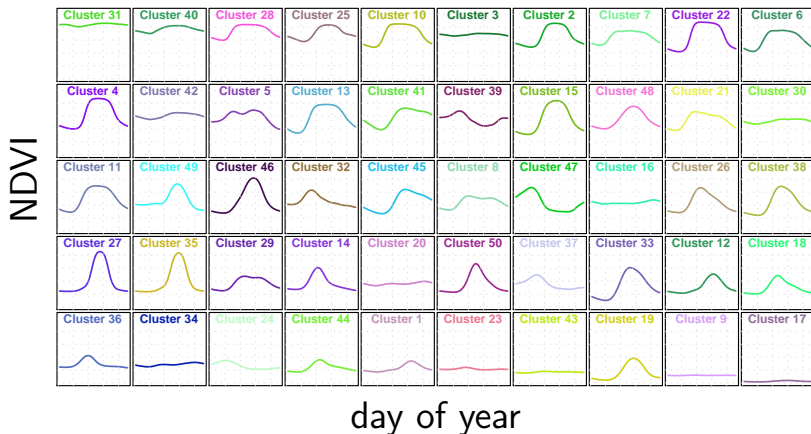
## Clustering MODIS NDVI into Phenoregions

- ▶ 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, using high performance computing, was applied to the entire body of the high resolution MODIS NDVI record for the continental U.S.
- ▶ >80B NDVI values, consisting of  $\sim 146.4\text{M}$  cells for the CONUS at 250 m resolution with 46 maps per year for 12 years (2000–2011), analyzed using  $k$ -means clustering.
- ▶ The annual traces of NDVI for every year and map cell are combined into one 323 GB single-precision binary data set of 46-dimensional observation vectors.
- ▶ Clustering yields 12 maps in which each cell is classified into one of  $k$  phenoclasses, and phenoregions form representative prototype annual NDVI traces.

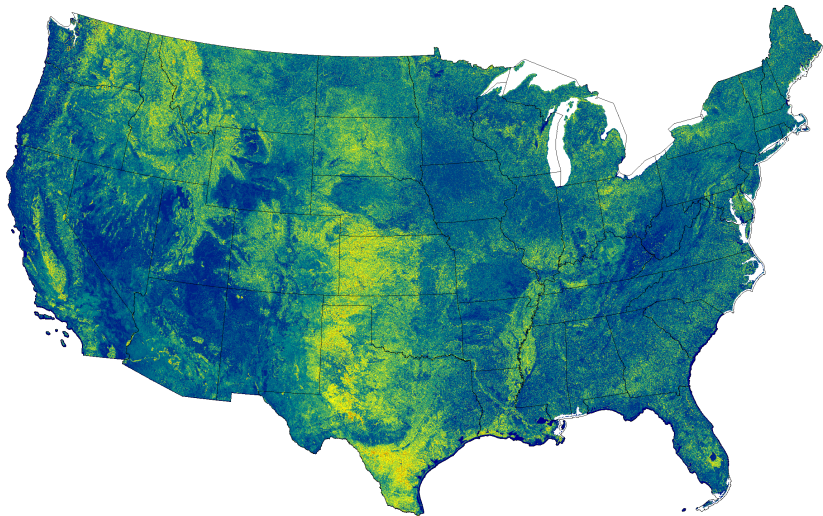
# 50 Phenoregions for year 2011 (Random Colors)



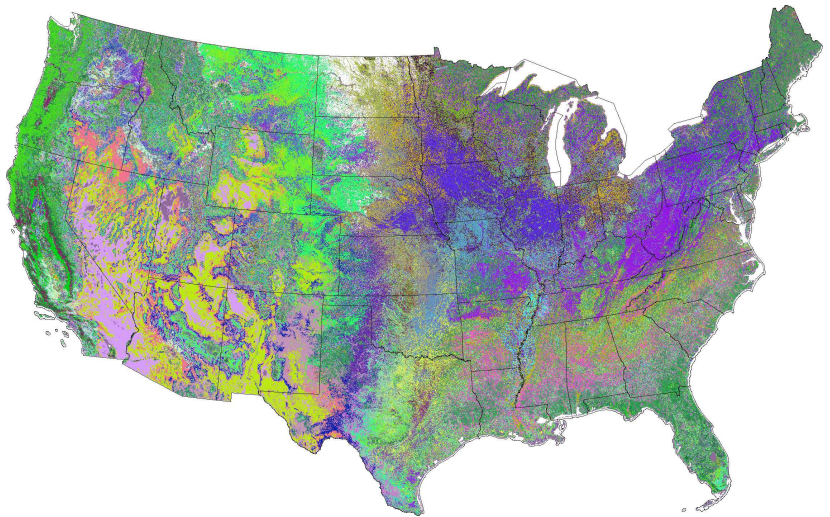
# 50 Phenoregion Prototypes (Random Colors)



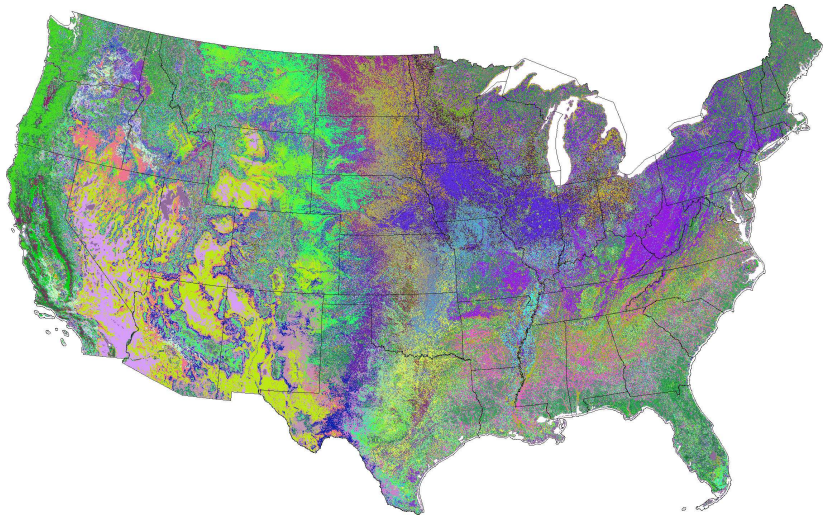
# 50 Phenoregions Persistence (Random Colors)



# 50 Phenoregions Mode (Random Colors)

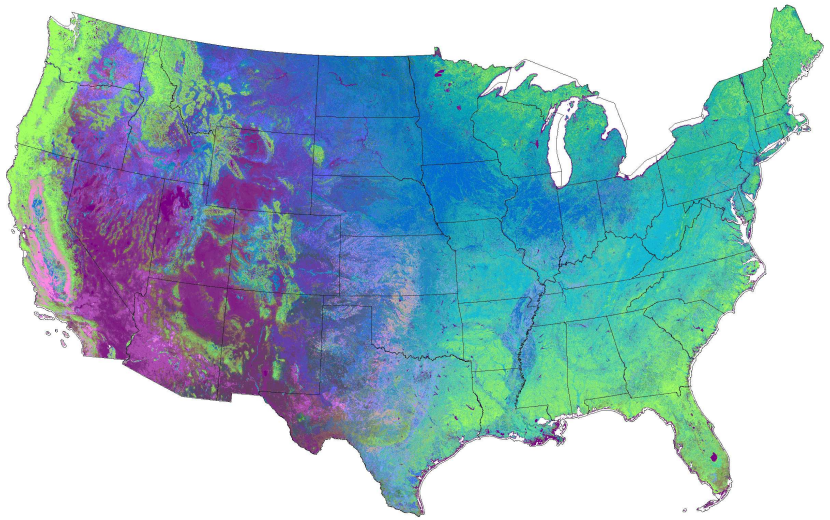


# 50 Phenoregions Max Mode (Random Colors)

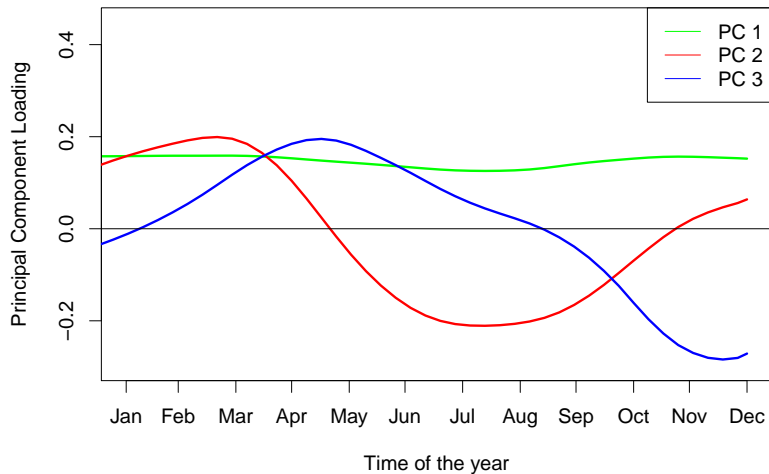




# 50 Phenoregions Max Mode (Similarity Colors)



# 50 Phenoregions Max Mode (Similarity Colors Legend)



# Phenoregions Clearinghouse

National Phenological Ecoregions (2000-2011) - Google Chrome

National Phenological E x

<https://www.geobabble.org/phenoregions/>

## National Phenological Ecoregions (2000–2011)

*William W. Hargrove, Forrest M. Hoffman, Jitendra Kumar, Joseph P. Spruce, and Richard T. Mills*  
January 14, 2013

[Jump to 50 National Phenoregions](#)

[Jump to 100 National Phenoregions](#)

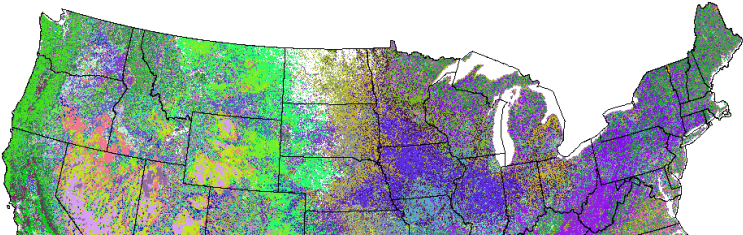
[Jump to 200 National Phenoregions](#)

[Jump to 500 National Phenoregions](#)

[Jump to 1000 National Phenoregions](#)

[Jump to 5000 National Phenoregions](#)

### 50 Most-Different National Phenological Ecoregions (2000–2011)



## **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 Presenters:**

- ▶ 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.

# Fourth Workshop on Data Mining in Earth System Science



## ICCS 2014: “Big Data Meets Computational Science”

### Fourth 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.**

See  
Jitendra  
Kumar's  
poster at  
5:30 p.m.

## Classification and Delineation of Large Earth Science Data

Jitendra Kumar<sup>1</sup>, Forrest M. Hoffman<sup>1</sup>, William W. Hargrove<sup>1</sup>

<sup>1</sup>Oak Ridge National Laboratory, <sup>2</sup>USDA Forest Service

### Climatic Data Analytics

- Identification of ecotones or climate zones is important for defining and studying climatic regimes, predicting suitable species ranges, and delineating environmental and ecological sampling domains
- Model diagnostics and intercomparison
- Knowledge discovery from model and observation data
- Increasing volumes of climate data calls for improved data analytics algorithms and computational tools

### Parallel k-means Clustering

- We have developed a highly scalable parallel k-means clustering algorithm tool (Figure 1)
- New acceleration schemes improve the computational efficiency of the clustering algorithm (Figure 2)

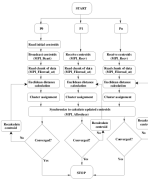


Figure 1: The parallel k-means algorithm

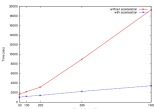


Figure 2: Accelerating k-means

### Scaling and Optimization

- We have optimized the Multivariate Spatio-Temporal Clustering (MSTC) tool for excellent parallel performance on T3an Cray X96 at ORNL (Figure 3)
- Two phase (read + scatter) parallel IO was implemented using MPI IO and optimized for performance on Lustre filesystem on OLCF machines (Figure 4)
- The tool has been applied for a wide range of data sets on a University of ORNL facility

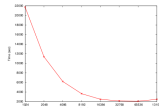


Figure 3: Parallel scaling (k=1000, NDVI 2000-2011)

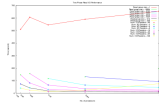


Figure 4: Parallel I/O performance and optimization

### Geo-spatial Analysis and Visualization

- We have built an Open Source tool chain for analysis and visualization (Figure 5, 6)
- This framework was designed and optimized to utilize high performance computing resources for analysis of large Earth Science data sets



Figure 5: Open source tools for analysis and data sharing

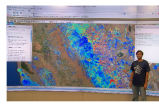


Figure 6: The EVEREST Visualization Facility provides a unique opportunity for analysis of very large simulation output and high resolution data products

### Forest Threat Detection

- USDA Forest Service, NASA, DOE ORNL, and USGS developed an early warning system for forest threats
- The ForRisk system uses phenology derived from NDVI observations from MODIS every 8 days (Figure 7)

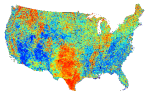


Figure 7: Integrated NDVI disturbance map

### Next Generation Ecosystem Experiments (NGEE) - Arctic

- NGEE is a model inspired field measurement program focused on the Arctic and other critical regions (Figure 8)
- Quantitative methodology developed for stratifying domain and determining representativeness of sites (Figure 9)

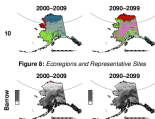


Figure 9: Site and Network Representativeness

### Climate Model Diagnostics and Intercomparison

- Cluster analysis makes large, multivariate time-series projections from Earth System Models understandable
- Results from CMIP5 historical and future climate under the RCP 8.5 scenario were analyzed (Figure 10, 11)
- Temperature, precipitation, and soil moisture were used in unsupervised classification

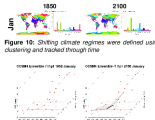


Figure 11: Centroids form a skeleton in state space

# Acknowledgments



U.S. DEPARTMENT OF  
**ENERGY**

---

Office of Science



This research was sponsored by the U.S. Department of Energy's Biological and Environmental Research (BER) program and 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

- F. 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, Nov. 2004.
- F. M. Hoffman, W. W. Hargrove, D. J. Erickson, and R. J. Oglesby. Using clustered climate regimes to analyze and compare predictions from fully coupled general circulation models. *Earth Interact.*, 9(10):1–27, Aug. 2005. doi:10.1175/EI110.1.
- F. M. Hoffman, J. Kumar, R. T. Mills, and W. W. Hargrove. Representativeness-based sampling network design for the State of Alaska. *Landscape Ecol.*, 2013. doi:10.1007/s10980-013-9902-0. In press.
- M. A. White, F. Hoffman, W. W. Hargrove, and R. R. Nemani. A global framework for monitoring phenological responses to climate change. *Geophys. Res. Lett.*, 32(4):L04705, Feb. 2005. doi:10.1029/2004GL021961.