

Scalable Algorithms for Analysis of Large Geospatiotemporal Data Sets and Some Applications to Landscape Ecology

Richard Tran Mills, Intel Corporation, Hillsboro, OR; Forrest M. Hoffman and Jitendra Kumar, Oak Ridge National Laboratory, Oak Ridge, TN

1. Introduction

- The increasing availability of high-resolution geospatiotemporal data sets from sources such as observatory networks, remote sensing platforms, and computational Earth system models has opened new possibilities for knowledge discovery and mining of ecological data sets fused from disparate sources.
- Traditional algorithms and computing platforms are impractical for the analysis and synthesis of data sets of this size; however, new algorithmic approaches that can effectively utilize the complex memory hierarchies and the extremely high levels of available parallelism in state-of-the-art high-performance computing platforms can enable such analysis.
- We examine some of these approaches and a few practical applications to the analysis of climatic and remotely-sensed vegetation phenology data sets and speculate on some of the other applications that such scalable analysis methods may enable.

2. Accelerated k-means Clustering

- We have two implementations of accelerated k-means clustering, following two parallel programming models
 - A master-worker (MW) model: Central master assigns “aliquots” of work to workers. Facilitates dynamic load balancing but has memory and performance scalability limits due to single, central process.
 - Fully distributed (FD): All processes use static distribution of work. Very scalable, but no dynamic load balancing.
- We “accelerate” the k-means process using two techniques described by Phillips (doi:10.1109/IGARSS.2002.1026202):
 - Use the triangle inequality to eliminate unnecessary point-to-centroid distance computations based on the previous cluster assignments and the new inter-centroid distances.
 - Reduce evaluation overhead by sorting inter-centroid distances so that new candidate centroids c_j are evaluated in order of their distance from the former centroid c_i . Once the critical distance $2d(p, c_i)$ is surpassed, no additional evaluations are needed, as the nearest centroid is known from a previous evaluation.

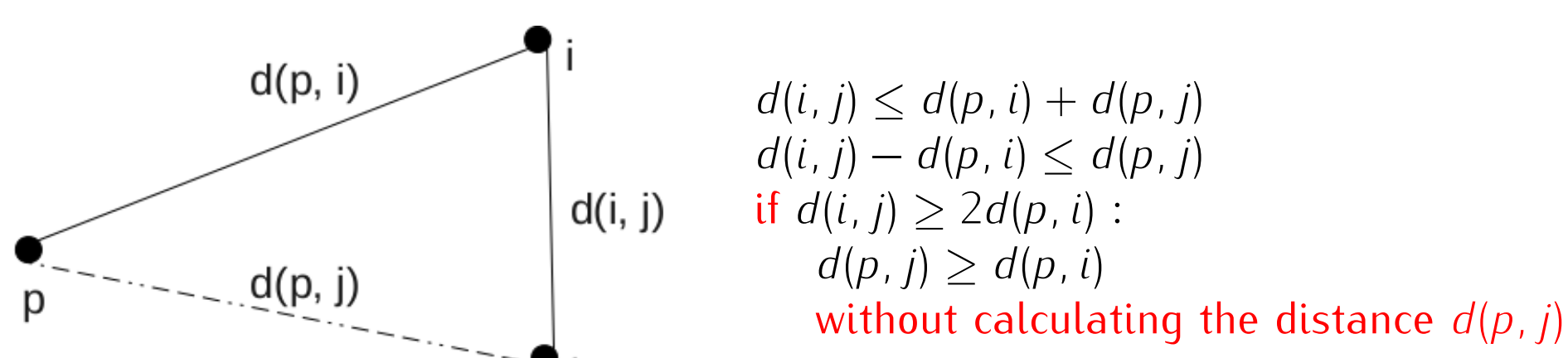


Figure 1: The triangle inequality is used to eliminate unnecessary distance calculations.

- We also improve cluster quality by moving or “warping” clusters that become empty to locations in data space where points that are farthest from their current cluster centroids reside.

2.1 Parallel Performance

- In 2011, we would use ~ 1024 AMD Opteron cores on a machine like Jaguar, the Cray XT5 at ORNL, for our analyses.
- In 2015, we can do larger analyses on a single compute node of Intel’s Endeavor cluster with Intel® Xeon® E7-8890 v3 (“Haswell-EX”) processors.
 - AVX2 instruction set: 256-bit (8 single precision floats) vector registers with dual-issue fused multiply-add
 - Four 18 core (36 thread) CPUs; over 500 GB DRAM

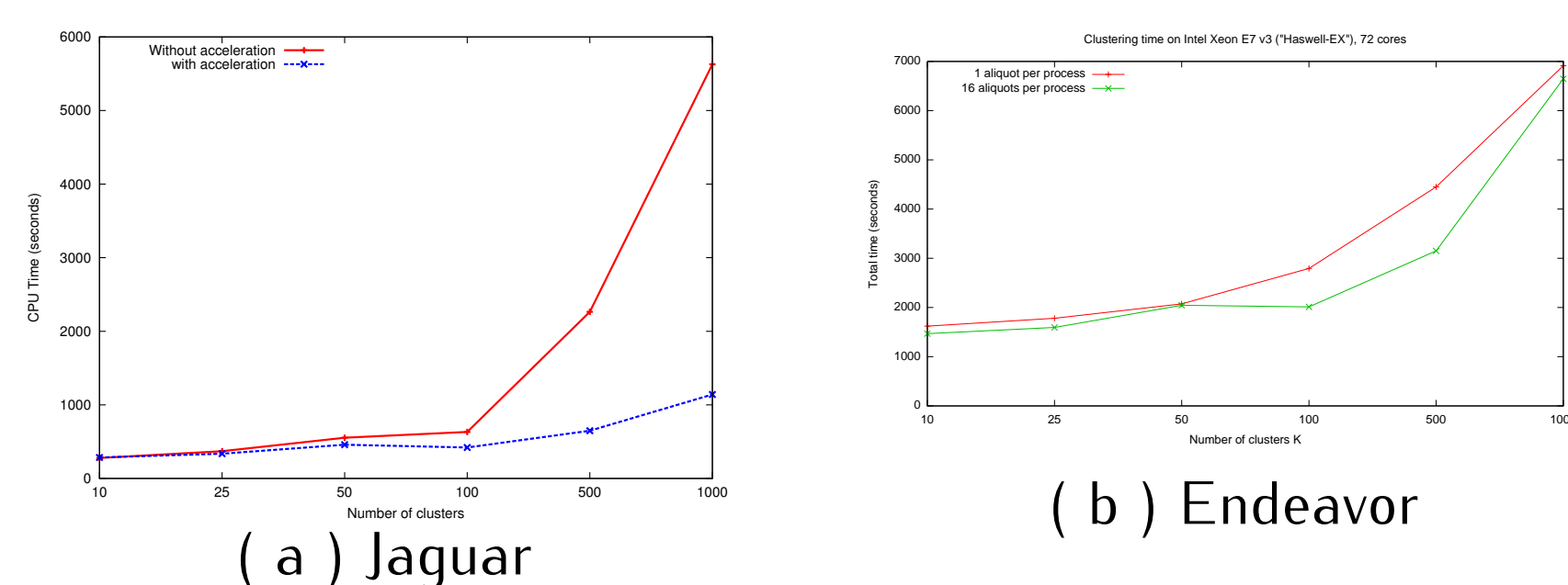


Figure 2: Times to cluster different versions of the 2000–2009 ForWarn phenology data set on (a) 1024 cores of the Jaguar Cray XT5, ca. 2011 at ORNL and (b) a single 72-core “Haswell-EX” node on Intel’s Endeavor cluster. The data set used on Jaguar is the 16 day product, while the one on Endeavor is the 8 day product and is therefore twice as large (251 GB in single precision).

- With acceleration, an equal distribution of observation vectors among processes does not guarantee load balance. Figure 2b illustrates the benefit of using smaller aliquots to enable dynamic load balancing in the MW clustering code.
- It may be possible to improve the percentage of peak FLOPS:
 - The distance calculations vectorize, but are expressed as level-1 BLAS operations and do not get good cache reuse.
 - It is possible to calculate the matrix of squared distances via level-2 (rank-1 updates) and level-3 (GEMM) BLAS operations, but calculates all distances.
 - Perhaps use a hybrid approach that uses level-2/3 BLAS part of the time, or to calculate only a portion of the distance matrix?

2.2 Applications: Quantitative Ecoregionalization and Change Detection

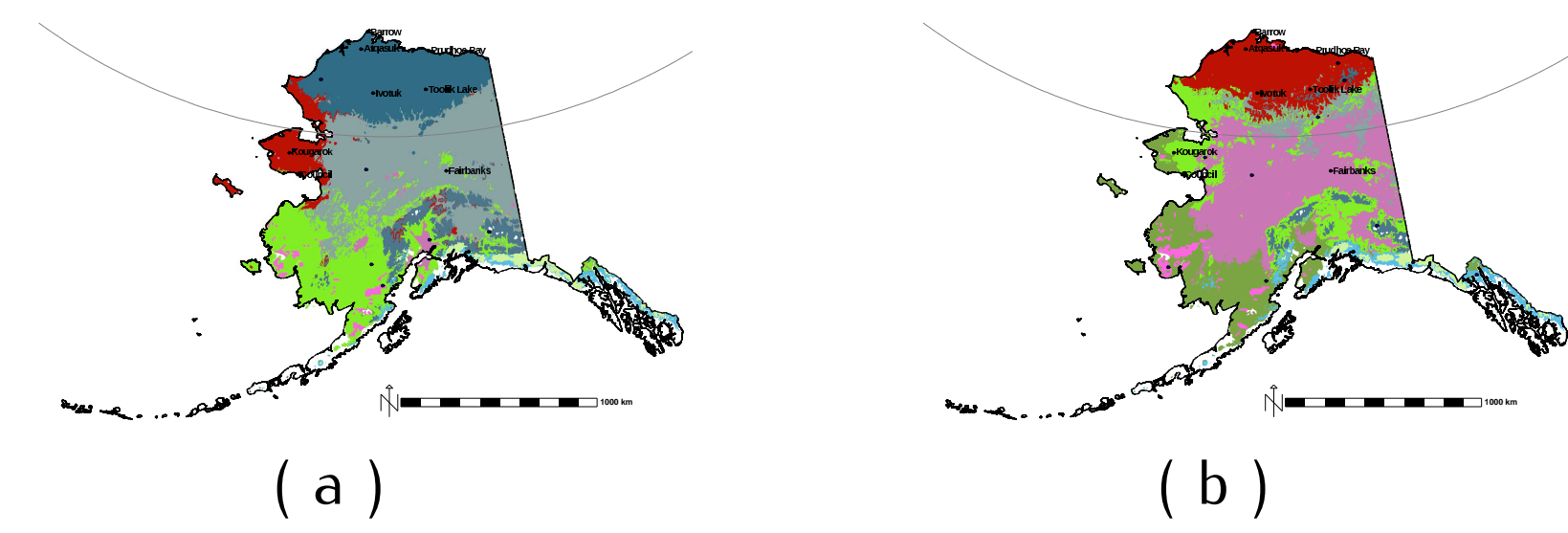


Figure 3: Geospatiotemporal clustering of a combination of observational data and downscaled general circulation model results projects dramatic shifts in location of Alaska ecoregions using downscaled 4km GCM results. Arctic tundra projected to be at 0.78% of current extent by 2099. DOI: 10.1007/s10980-013-9902-0.

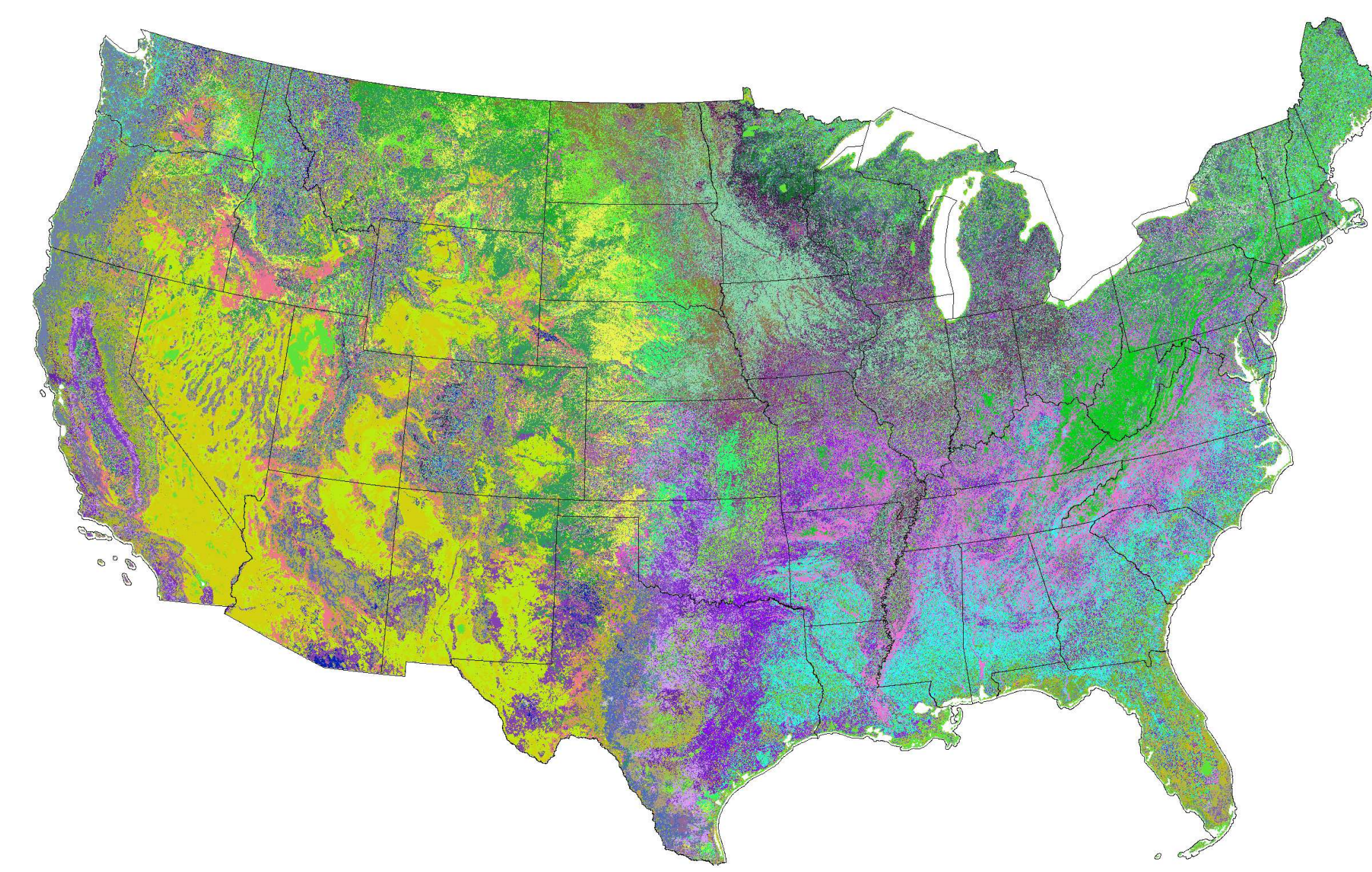


Figure 4: A map of “phenoregion” assignments for the year 2012, based on k -means analysis with $k = 50$ of the entire MODIS-derived ForWarn NDVI product for years 2000–2012. The body of observation vectors being clustered consists of the year-long NDVI time series for every map pixel, for each year. The map indicates cluster membership (in random colors) for the phenology observed in 2012 at each map pixel.

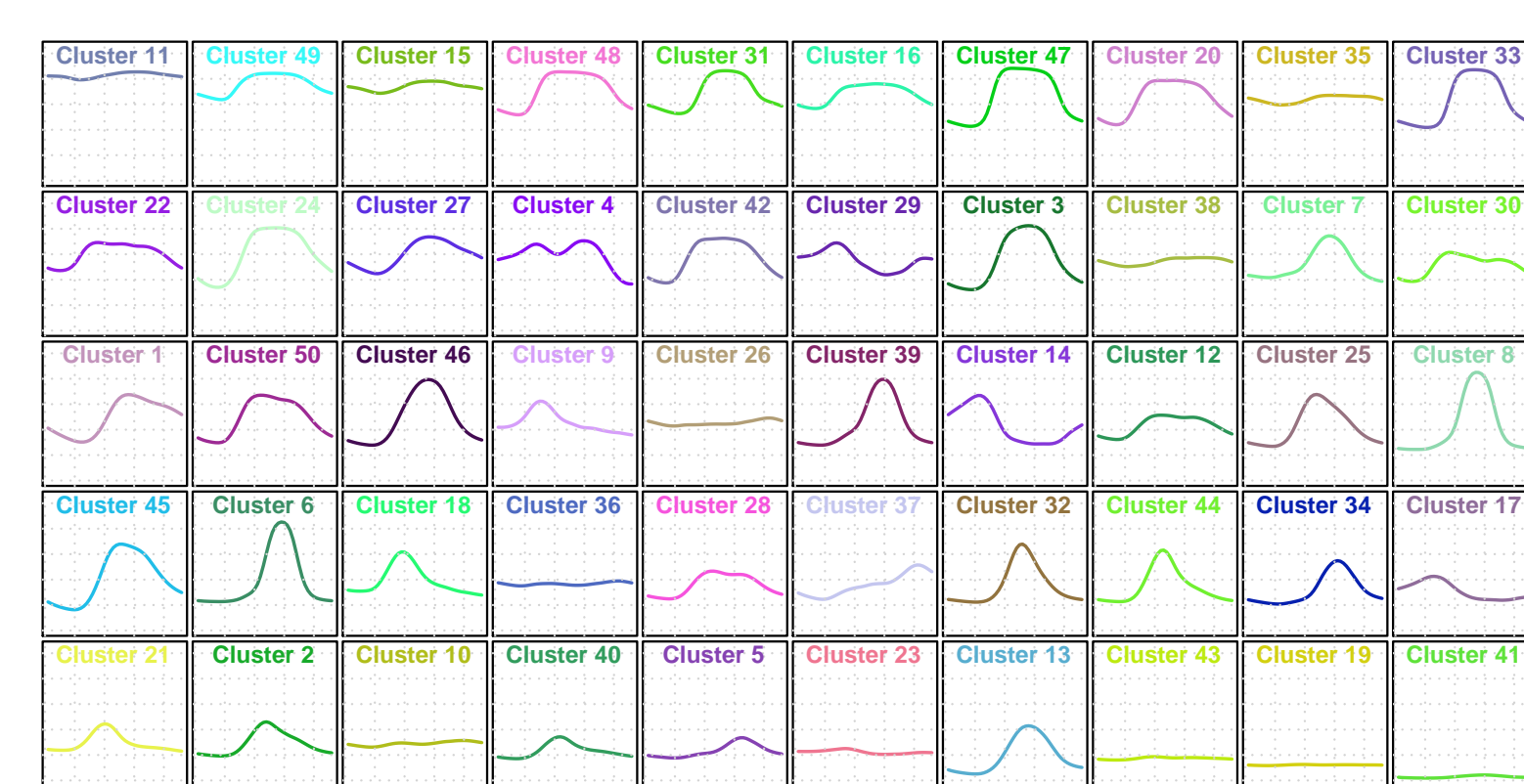


Figure 5: The fifty centroids (corresponding to “phenoregion” prototypes) used for the membership assignments in the map in Figure 4. The colors the centroid plot correspond to the map colors.

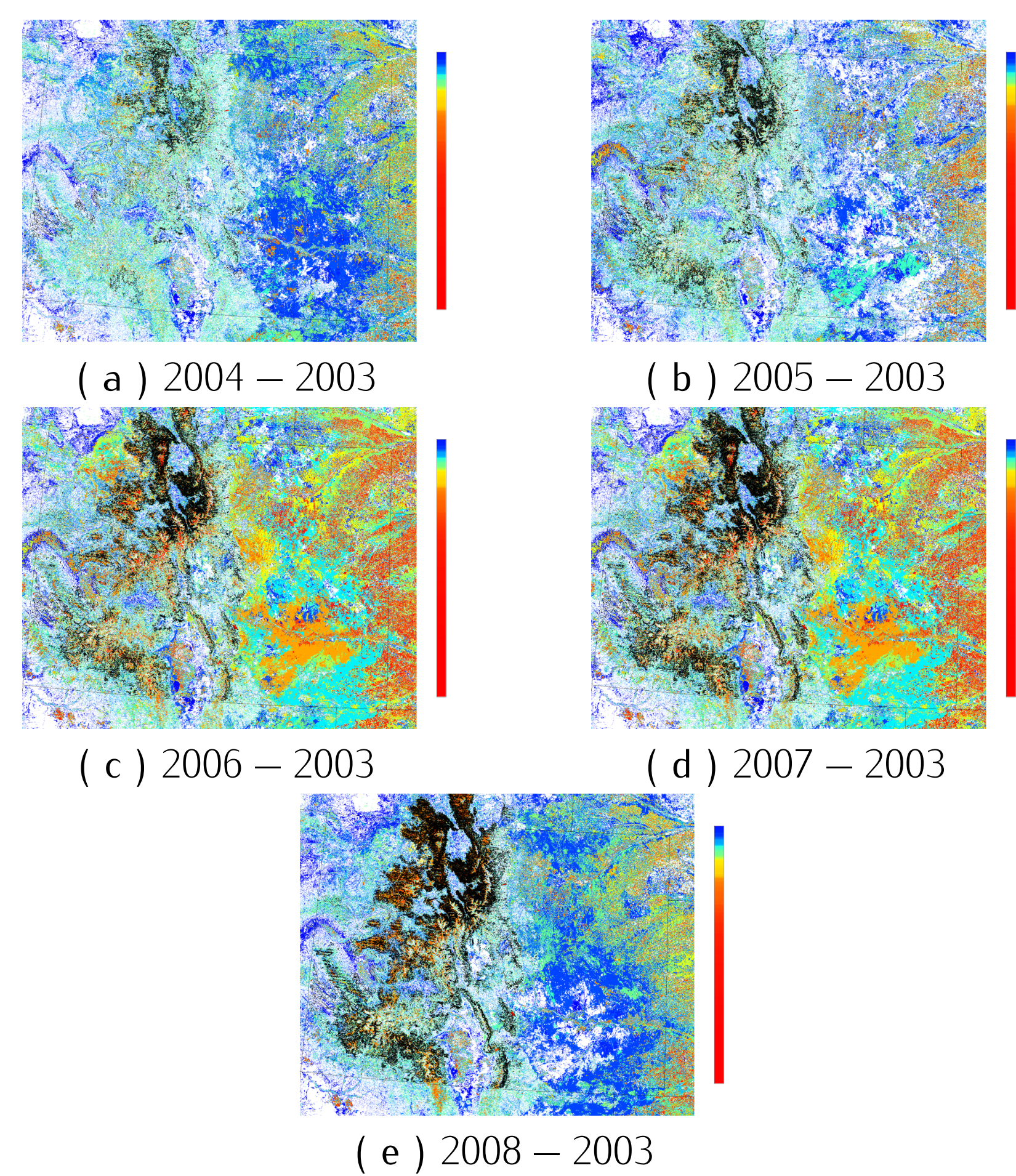
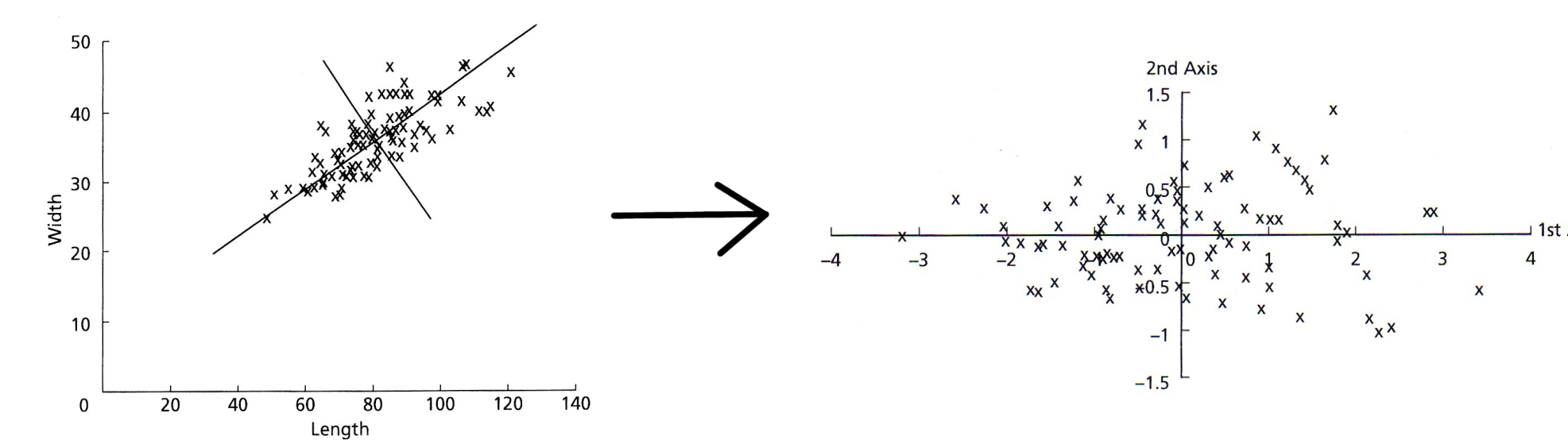


Figure 6: Maps showing the relative state space transition distances (how different phenoregion assignments are for given years) between years in Colorado and southern Wyoming. Pine beetle mortality correlates strongly with high transition distances. Black-outlined polygons are disturbed areas indicated on aerial sketch maps.

3. Principal Components Analysis

Principal Components Analysis (PCA) determines, for a p -dimensional data set, an orthogonal set of p new axes (linear combinations of the original p variables) such that the first axis explains the greatest variance, the second explains the next most variance, and so on.



- Commonly used to determine dominant patterns in data
- But can also be used to determine the anomalous patterns: Observations that score strongly on low order components do not follow the correlation structure of the data.

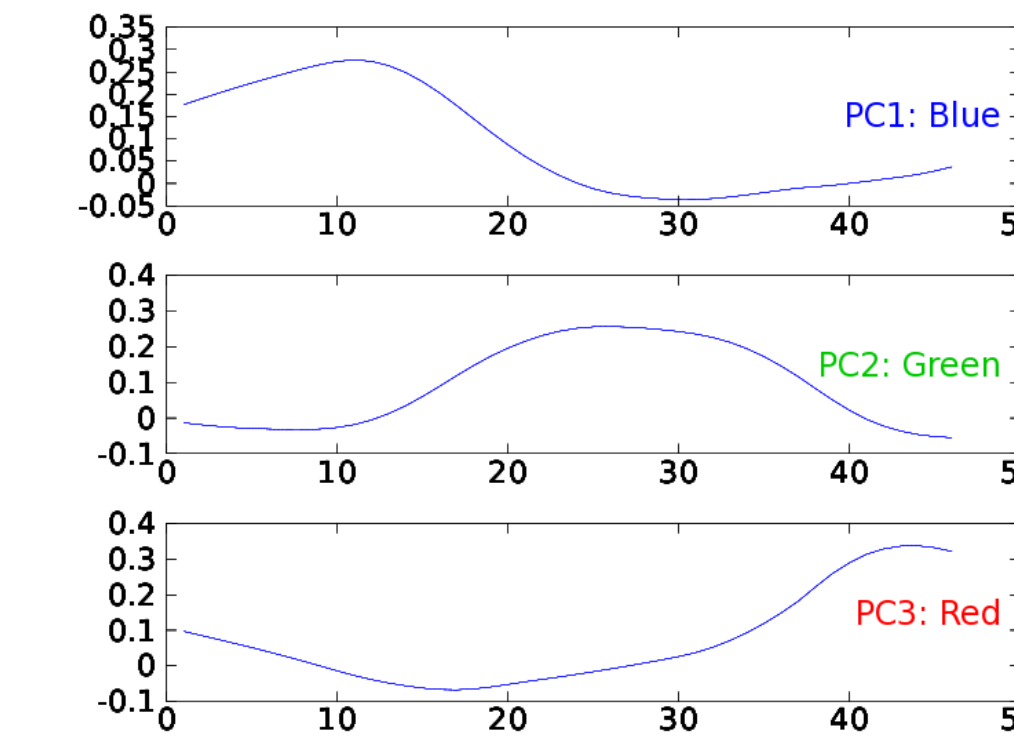


Figure 7: The loadings (coefficients in the linear combination of the 46 original variables) along the three varimax-rotated principal axes. The x -axis corresponds to the eight-day NDVI-acquisition windows and loadings are shown on the y -axis.

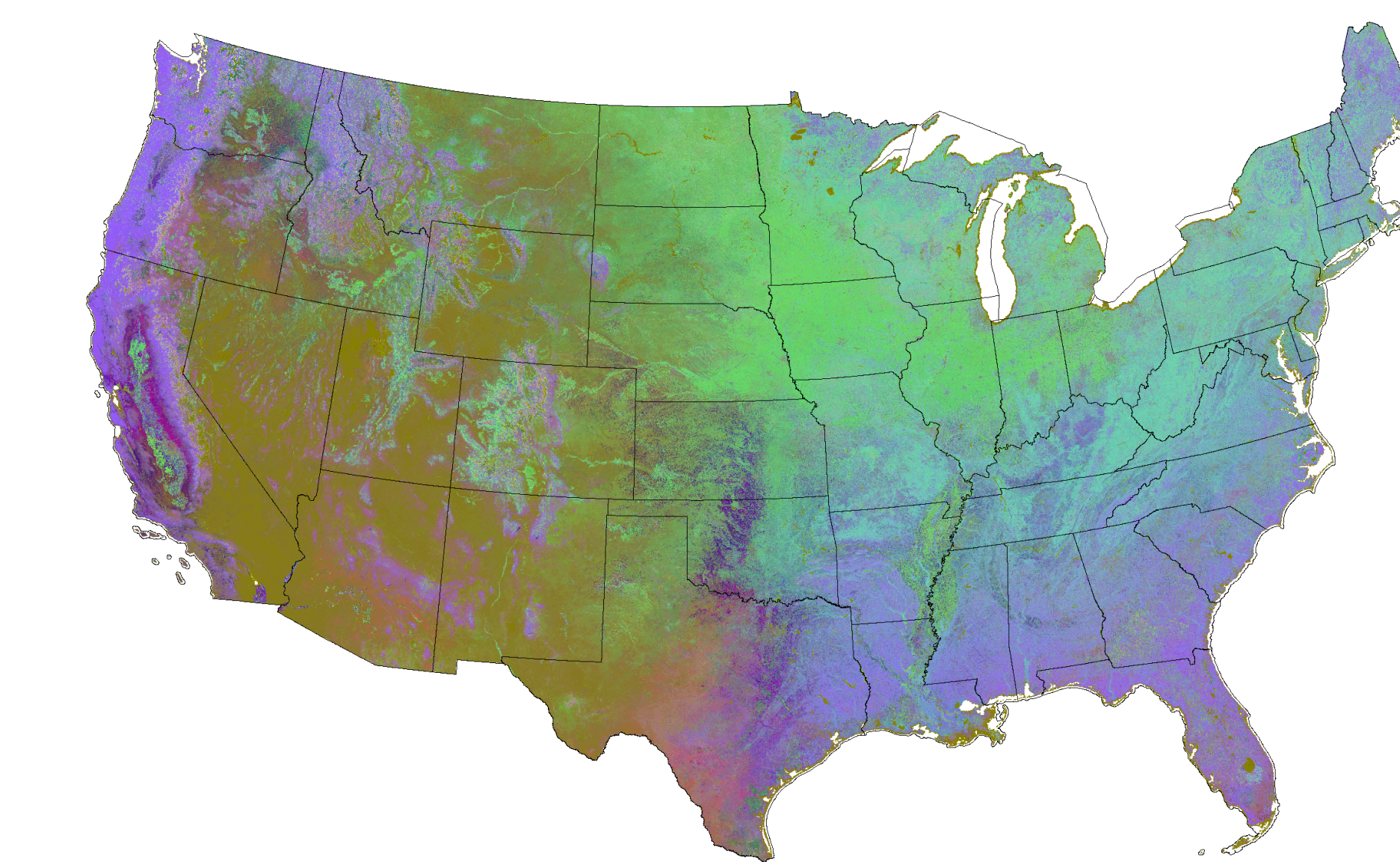


Figure 8: Phenoregion assignment map for year 2000 with $k = 1000$. Similarity colors are used to indicate cluster membership.

3.1 Parallel Principal Components Analysis Tool

- We have developed a prototype parallel tool to perform PCA.
- Rather than explicitly forming the covariance matrix, computes thin SVD of the adjusted data matrix.
- Uses the Lawson–Hanson–Chan factorization to exploit the “tall and skinny” ($m \gg n$) nature of our matrices: ($m \gg n$)
 - Form reduced factorization $A = QR$ (via parallel LAPACK routine)
 - Gather the matrix R to process 0.
 - Process 0 calls LAPACK DGEVD to compute the SVD $R = USV^T$.
 - Optionally, back transform Q to get $Q \leftarrow QU$.
 - Final SVD is: $A = QSV^T$
- A serial bottleneck exists where the SVD of R is computed, but this matrix is so small (only 46×46 for our NDVI data set) that this serial portion is essentially negligible.

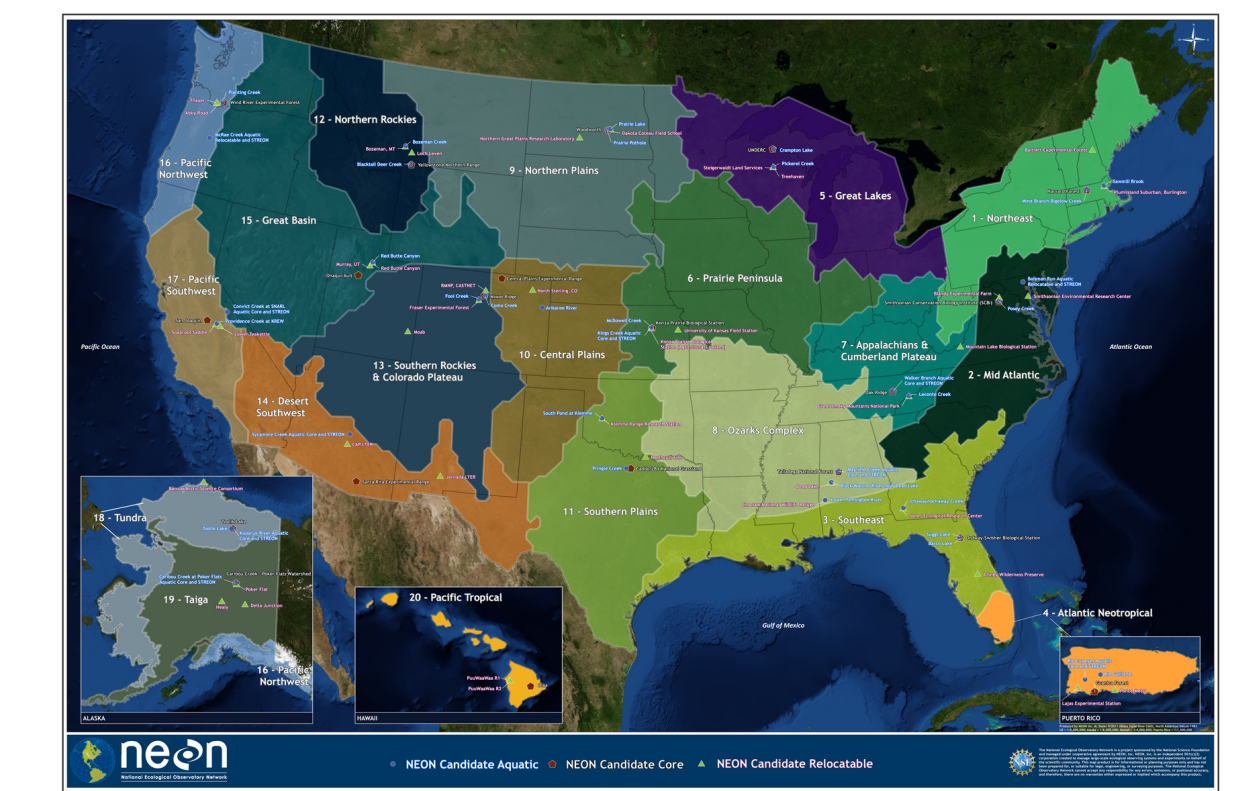
3.2 Detecting anomalous observations with PCA

- Can identify anomalies two complementary ways:
 - Look at sum of scores onto r lowest-order components: $\sum_{i=p-r+1}^p \frac{y_i^2}{\lambda_i}$ greater than some outlier threshold
 - Look at squared prediction error: How well an observation can be represented in subspace of q highest order components?
 - Idea: decompose into modeled and residual parts: $x = \hat{x} + \tilde{x}$
 - $P = [v_1 \ v_2 \ \dots \ v_q]$
 - $\hat{x} = PP^T x = Cx$ and $\tilde{x} = (I - PP^T)x = \tilde{C}x$
 - Abnormal if $SPE = \|\tilde{x}\|^2 = \|\tilde{C}x\|^2$ exceeds threshold
- Can also do cross-comparison: Construct subspace from one data set, then see how well observations from another can be represented in that space.

3.3 Detecting anomalies within a single year, single NEON domain

- These approaches will flag any observations that are somehow “unusual” for the collection of data from which the principal components have been calculated.
- Some judgement required: choice of NDVI data subset used in the PCA calculation will affect what constitutes a “normal” or “abnormal” observation.
- E.g., Extremely low NDVI may appear normal when using PCA based on national dataset due to presence of areas like the Mohave; appears anomalous when using PCA based only on humid Southeast.

- Here we use PCAs computed over single years and within a spatial domain conforming to the eco-climatic domains established by the National Ecological Observatory Network.



- In all examples, PC vectors 10–46 are used as the basis for the “abnormal” space, which explains 5–10% of the variance.
- In all of examples, certain features that are not disturbances but possess very anomalous NDVI traces (e.g., bodies of water) show up very strongly.

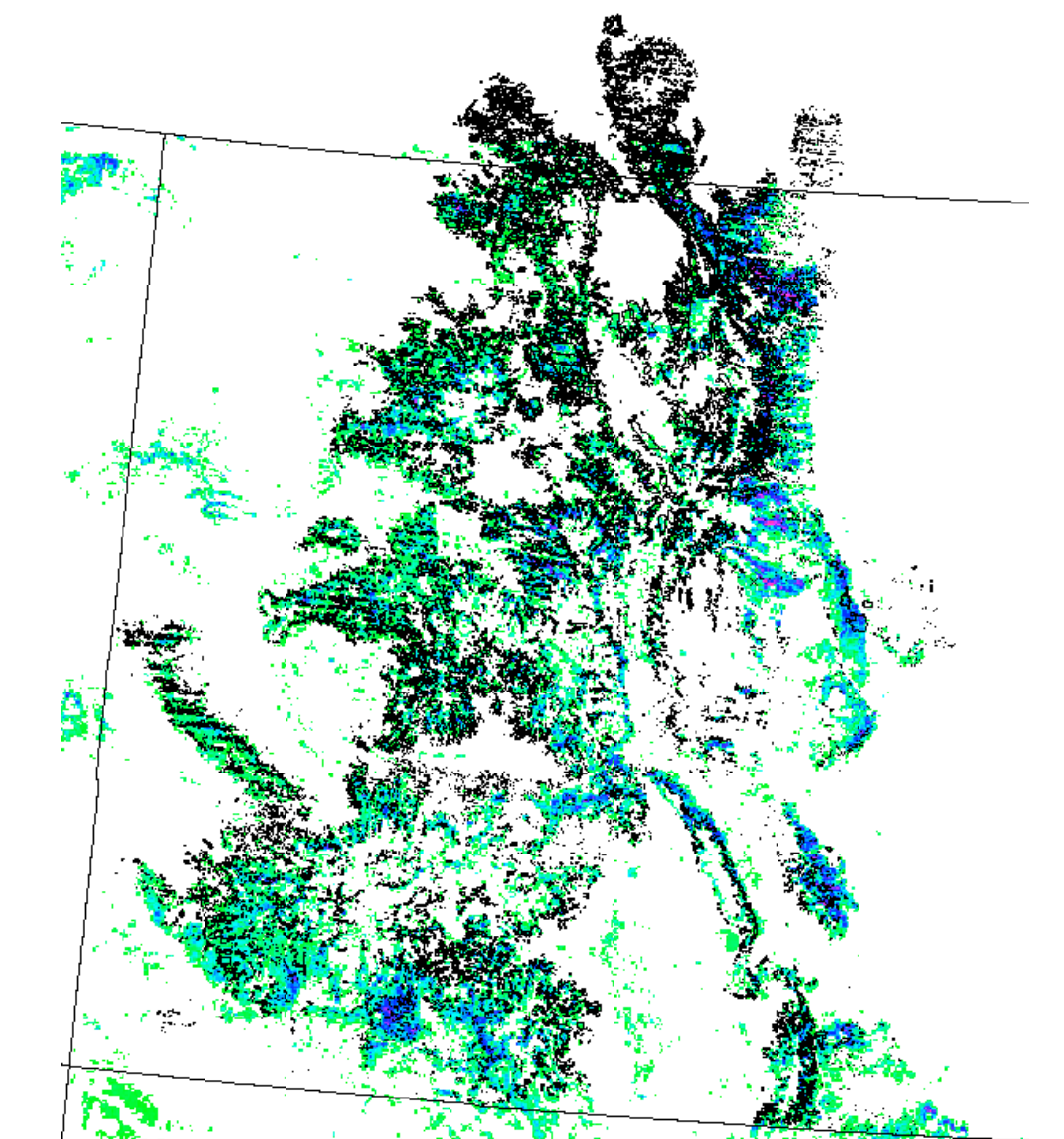


Figure 10: Portion of the Southern Rockies–Colorado Plateau NEON Domain for year 2008, showing map cells scoring in the 85th percentile. Black polygons show damaged areas noted in aerial detection surveys; extensive damage due to mountain pine beetle and sudden aspen decline are evident.

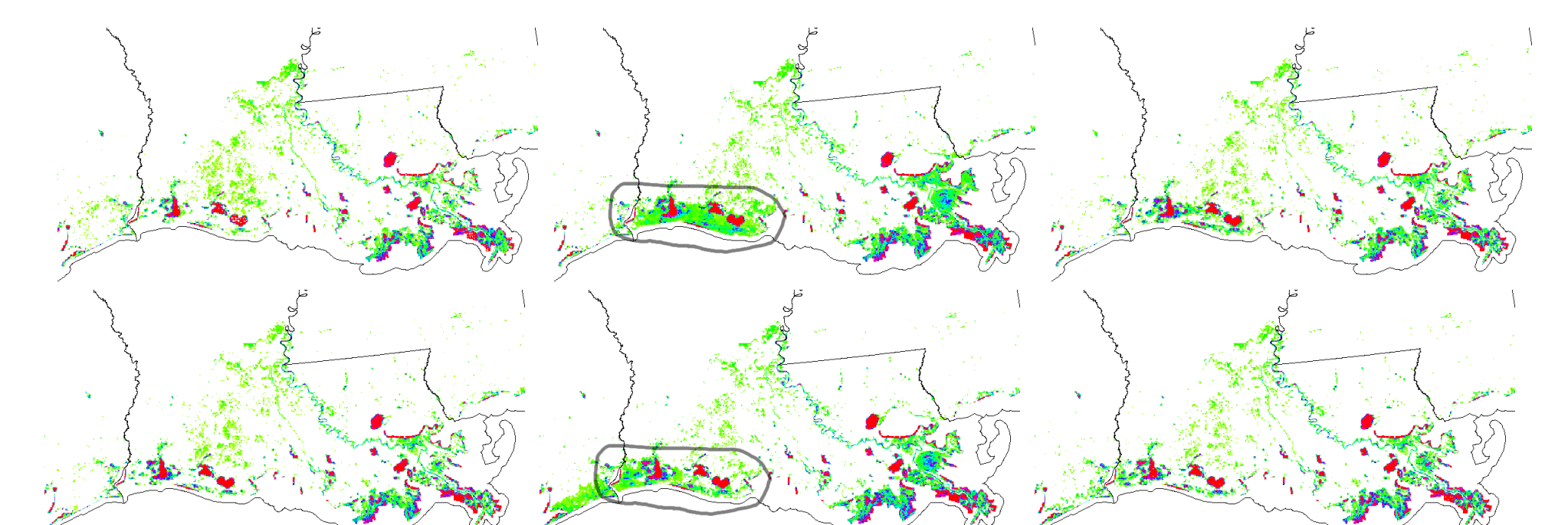


Figure 11: Portions of the PCA-based anomaly maps (map cells scoring in the 90th percentile) for the Southeast NEON Domain for years 2004–2009, showing the area in the vicinity of the Louisiana coast. From top row shows years 2004, 2005, and 2006, respectively, and the bottom row years 2007, 2008, and 2009. The affected regions are circled in the 2005 and 2008 maps. The prominent red features are water bodies.

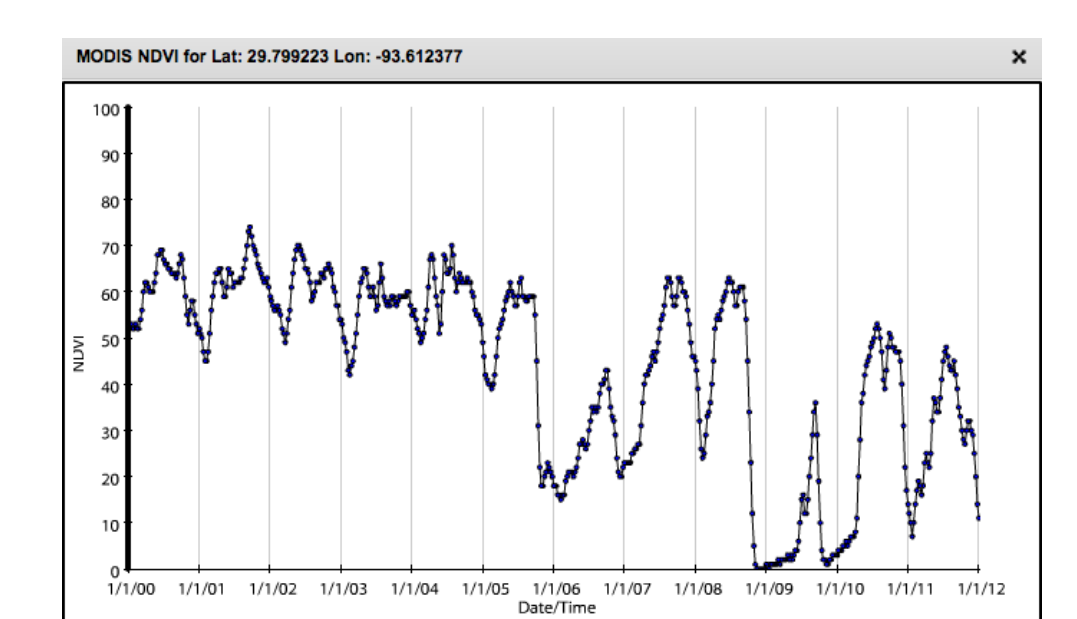


Figure 12: NDVI trajectory as viewed via the Forest Change Assessment Viewer for a location (close to the center of the circled region in Figure 11) near the coast in southwestern Louisiana showing apparent hurricane-induced mortality from events in 2005 and 2008.

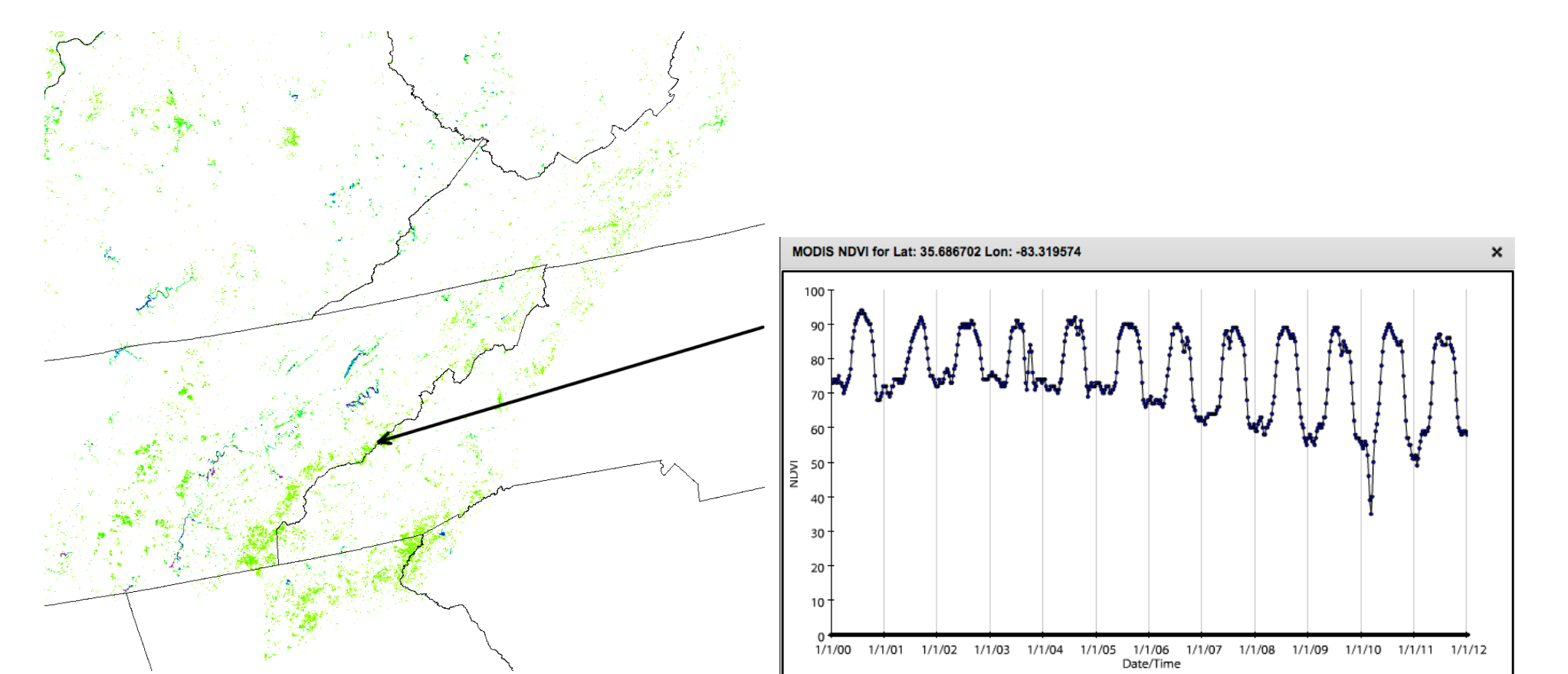


Figure 13: At left, a portion of the PCA-based anomaly map (map cells scoring in the 90th percentile) for the Southern Appalachians/Cumberland Plateau NEON Domain for year 2010. The arrow indicates a location thought to be affected by hemlock woolly adelgid, and the corresponding NDVI trajectory is shown at right.

4. Notices and Disclaimers

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document. Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade. This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps. The products and services described may contain defects or errors known as errata which may cause deviations from published specifications. Current characterized errata are available on request. Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-4725, or go to: <http://www.intel.com/design/literature.htm> Intel, Intel Xeon, Intel Xeon Phi™ are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States or other countries. Other brands and names may be claimed as the property of others. Copyright ©2014 Intel Corporation. All rights reserved. Optimization Notice Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSE4.2 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.