Argonne NATIONAL LABORATORY

Scalable Unsupervised Learning Approaches for Analysis of Large Geospatiotemporal Datasets

Richard Tran Mills, Argonne National Laboratory

LANS informal seminar, Argonne, IL

April 25, 2018

Credits

R. Mills is the speaker, but this talk summarizes collaborative efforts of many people, and contains contributed material or results from

- William Hargrove, USDA Forest Service, Southern Research Station
- Forrest Hoffman, ORNL
- Jitendra Kumar, ORNL
- Salil Mahajan, ORNL
- Steve Norman, USDA Forest Service, Southern Research Station
- Sarat Sreepathi, ORNL
- Vamsi Sripathi, Intel Corporation
- Joseph Spruce, NASA Stennis Space Center

Who is this guy?

Biosketch:

- May 2017, joined staff in MCS/LANS. "Remote" employee based in Portland, OR
- January 2014—May 2017, HPC Earth System Models Architect at Intel
  - Part of the Many Integrated Core (MIC) program developing Xeon Phi
  - Co-design in weather, climate, and Earth system models
- August 2004—January 2014, Staff scientist at ORNL
  - Started in NCCS/OLCF, moved to CSM, then Environmental Sciences Division
  - Joint faculty appointment (departments of Earth and Planetary Sciences; Computer Science) University of Tennessee, 2010—2014
- 2001—2004, DOE Computational Science Graduate Fellow, Dept. of Computer Science at William and Mary. Practicum in Earth and Environmental Sciences Division, LANL

What am I known for?

- One of the original developers of PFLOTRAN (`www.pflotran.org`), massively parallel hydrologic flow and reactive transport code.
- Long-time occasional (now full-time) contributor to PETSc development.
- Will largely avoid talking about either of these things today!

Introduction to what I **will** talk about today

- Increasing availability of high-resolution geospatiotemporal data sets from varied sources:
  - Observatory networks
  - Remote sensing platforms
  - Computational Earth system models
- New possibilities for knowledge discovery and mining of ecological data sets fused from disparate sources.
- Traditional tools impractical for analysis/synthesis of data sets this large: Need new approaches to utilize complex memory hierarchies and high levels of available parallelism in state-of-the-art high-performance computing platforms.
- We have adapted pKluster—an open-source tool for accelerated k-means clustering we use for many geospatiotemporal applications—to effectively utilize state-of-the art multi- and manycore processors, such as the second-generation Intel Xeon Phi ("Knights Landing") processor, as well as GPGPUs.
- Have also developed a parallel PCA/SVD tool that complements some of our clustering-based approaches.

Talk Outline

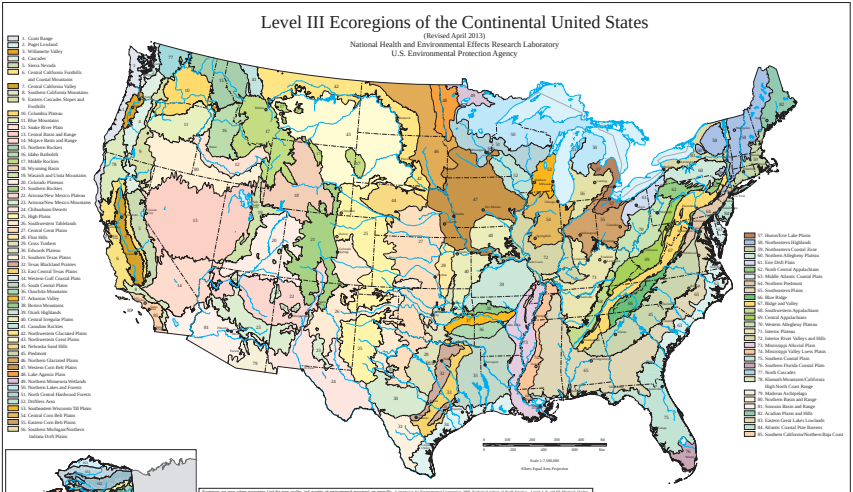
1. Some history: The "Stone Soupercomputer" and quantitative ecoregion delineation
2. Optimizations to the pKluster parallel *k*-means code
   - 2.1 "Accelerated" *k*-means using the triangle inequality
   - 2.2 Optimizations for AVX2 and AVX-512 multi- and many-core CPUs
3. A geo-spatio-temporal application: Early warning system for threats to forest ecosystems
   - 3.1 Cluster-based approaches
   - 3.2 Principal components analysis (PCA)-based approaches
4. Extra credit (time and audience interest permitting):
   Speculative application: Using machine-learning for "scale-bridging" in land surface hydrology

Scalable k-means Clustering with pKluster

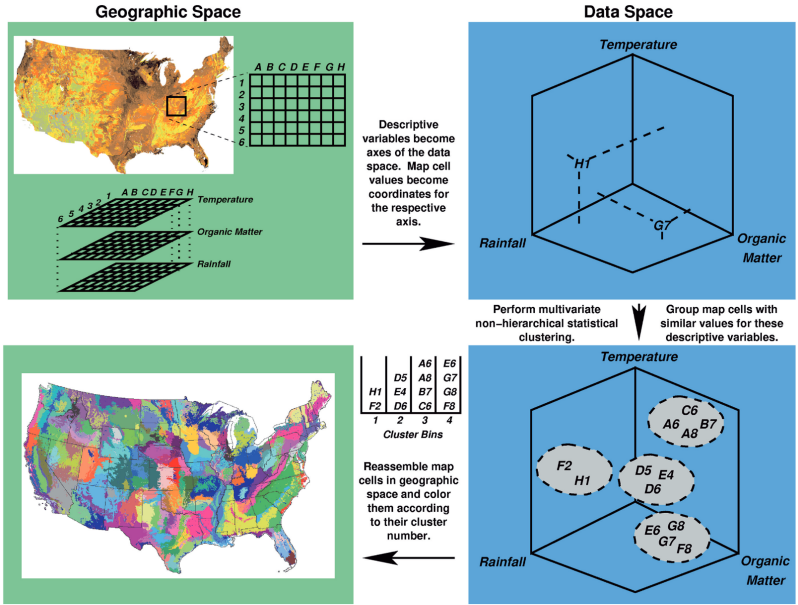
Our distributed-memory clustering code has a long history...

Figure: Originally developed in 1996–1997 for use on the Stone Soupercomputer, a very early Beowulf-style cluster constructed entirely out of surplus parts (see "The Do-It-Yourself Supercomputer", *Scientific American*, 265 (2), pp. 72-79, 2001.)

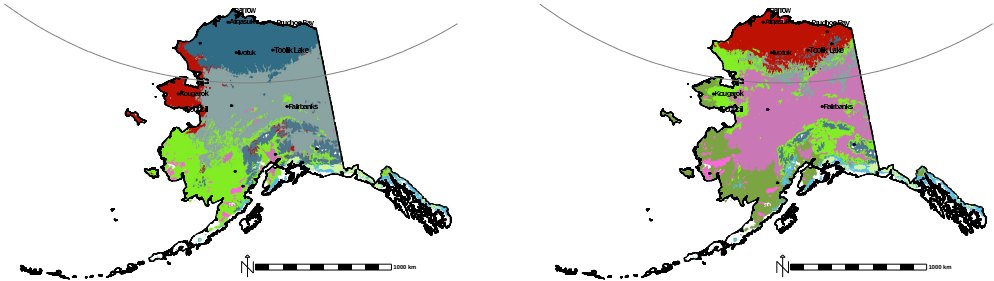
Original motivation: Replacing hand-drawn ecoregionalizations



Quantitative Ecoregionalization through Multivariate Spatio(-Temporal) Clustering



Quantitative Ecoregionalization through Time: Sampling Network Design

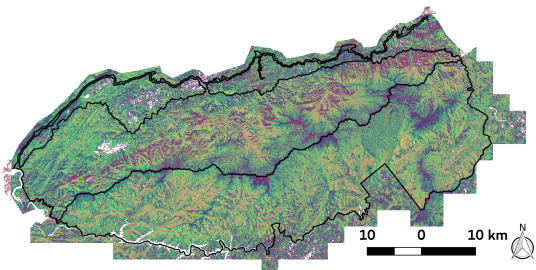


(a) 10 ecoregions, present (2000-2009)

(b) 10 ecoregions, future (2090-2099)

Figure: Geospatiotemporal clustering of a combination of observational data and downscaled general circulation model results projects dramatic shifts in location of Alaska ecoregions using downscaled 4 km GCM results. Arctic tundra projected to be at 0.78% of current extent by 2099. DOI: 10.1007/s10980-013-9902-0. **2014 US-IALE Outstanding Paper in Landscape Ecology.**

GSMNP LiDAR-derived canopy structure classification



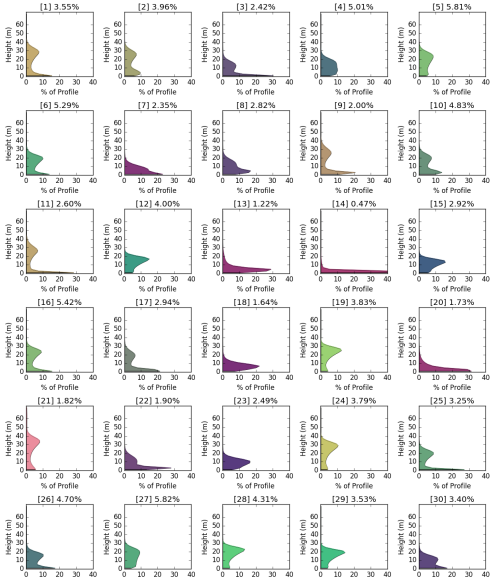


Figure: Map (above) showing the 30 most-different classes of vegetation canopy structure, as identified by k-means clustering (right) for the Great Smoky Mountains National Park.

https://www.climatemodeling.org/~jkumar/pubs/Kumar_ICDM_20151117.pdf

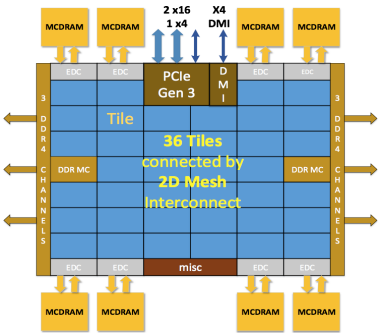
- When pKluster was initially written, on-node parallelism was virtually nonexistent on commodity PCs; the focus was purely on distributed-memory parallelism.
- Because of extreme heterogeneity of the cluster, a master-slave parallel programming paradigm was used (provides dynamic load-balancing).
  - On modern systems, a fully-distributed, masterless approach may be more efficient.
  - We work with the master-slave version here, because some techniques used here introduce load imbalance even on homogeneous machines.

Features:

- Runs on any machine (or cluster) with C89 (or higher) C compiler and an MPI implementation.
- Option to 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.
- Support for clustering observation vectors with many zero entries (e.g., species occurrence data).
- **Fast!** Suitable for clustering multi-terabyte data sets.
  - Implements "accelerated" *k*-means algorithm.
  - Optimizations for manycore CPU and GPGPU systems.

Manycore Computing Architectures

- In recent years, the number of compute cores and hardware threads has been dramatically increasing.
- Seen in GPGPUS, "manycore" processors such as the Intel Xeon Phi, and even on standard server processors (e.g., Intel Xeon Skylake).
- There is also increasing reliance on data parallelism/fine-grained parallelism.
  - Current Intel Xeon processors have 256-bit vector registers and support AVX2 instructions.
  - Second-generation Intel Xeon Phi processors and Intel Skylake Server processors have 512-bit vectors/AVX512 instructions.



At left, "Knights Landing" (KNL) Xeon Phi processor:

- Up to 36 tiles interconnected via 2D mesh
- Tile: 2 cores + 2 VPU/core + 1 MB L2 cache
- Core: Silvermont-based, 4 threads per core, out-of-order execution
- Dual issue; can saturate both VPUs from a single thread
- 512 bit (16 `floats` wide) SIMD lanes, AVX512 vector instructions
- High bandwidth memory (MCDRAM) on package: 490+ GB/s bandwidth on STREAM triad[2]

Benchmarking Platforms and Problem

	Intel(R) Xeon(R) CPU E5-2697 v4	Intel(R) Xeon(R) Gold 6148	Intel(R) Xeon Phi(TM) CPU 7250
Code Name	Broadwell (BDW)	Skylake (SKX)	Knights Landing (KNL)
Sockets	2	2	1
Cores	36	40	68
Threads (HT enabled)	72	80	272
CPU Clock (GHz)	2.3	2.4	1.4
HBM	-	-	16 GB
Memory	128 GB @ 2400 MHz	192 GB @ 2666 MHz	98 GB @ 2400 MHz
ISA	AVX2	AVX512{F, DQ, CD, BW, VL}	AVX512{F,PF, ER, CD}

Benchmark problem: GSMNP LiDAR clustering

- 1.5 million observations
- 74 coordinates
- $k = 2000$ clusters

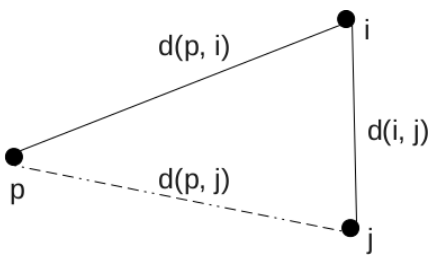
Parallel k-means clustering algorithm

- Centralized master-worker paradigm
- Start from some initial centroids (chosen offline)
- Master:
  - Broadcasts centroids and aliquot assignment to workers
  - Collects new cluster assignments from workers
  - Recomputes centroids
- Workers, for an assigned aliquot:
  - Compute observation-to-centroid distances
  - Assign each observation to closest centroid

Figure: Illustration of k-means iteration for $k = 3$. `https://commons.wikimedia.org/wiki/File:K-means_convergence.gif`

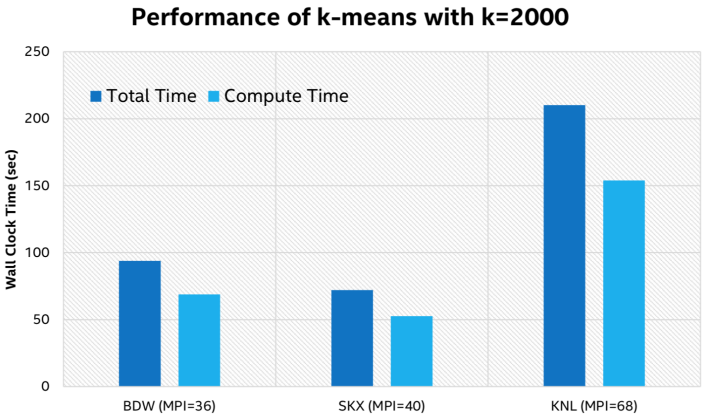
Accelerated k-means clustering

- Classical k-means actually performs far more distance calculations than required!
- 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.



$d(i,j) \leq d(p,i) + d(p,j)$
$d(i,j) - d(p,i) \leq d(p,j)$
if $d(i,j) \geq 2d(p,i)$:
$d(p,j) \geq d(p,i)$
without calculating the distance $d(p,j)$

Baseline (accelerated k-means) Performance

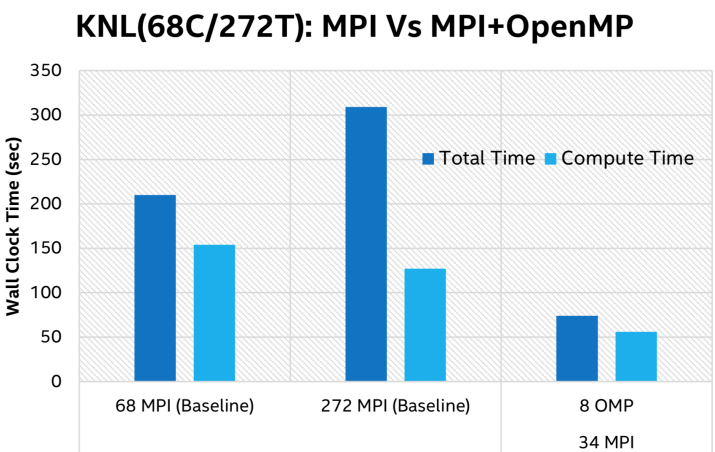


- 1.3X speedup on SKX vs. BDW
- Significant slowdown (2.2X) on KNL vs. BDW

Effective Use of Hyperthreads

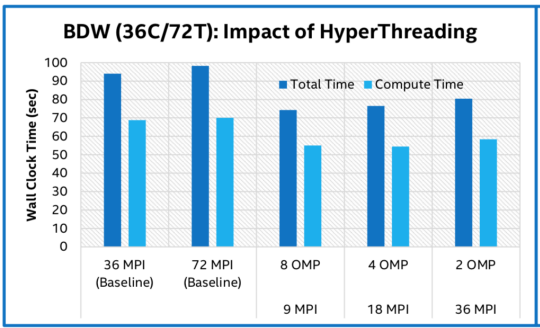
- Using a pure MPI approach (one MPI rank per core), performance of the accelerated k-means clustering approach is surprisingly poor on the "Knights Landing" (KNL) processor.
- Using two MPI ranks per core slightly decreases time in the actual clustering calculation, but slightly increases total time due to greater overhead in master-worker coordination.
- This suggests that using more available hardware threads can improve performance on KNL, if we can avoid increasing master-worker overhead.

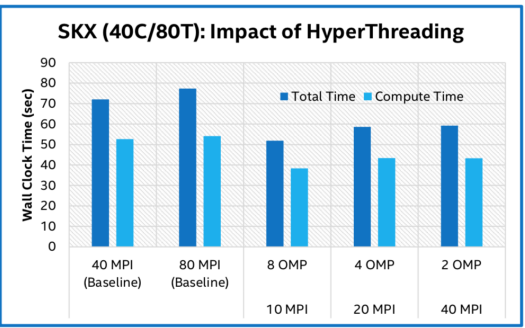
Performance Optimizations: OpenMP Parallelism on KNL



- Hybrid MPI-OpenMP version of distance calculation function effectively utilizes FMA units and reduces the bottleneck on rank 0.
- Use dynamic loop scheduling to smooth load imbalance due to triangle inequality (many observations in an aliquot might skip point-to-centroid distance calculation).
- Pin each MPI to a KNL "tile" and spawn 8 threads (4 threads per core).
- 2.8X improvement.

Performance Optimizations: OpenMP Parallelism on BDW and SKX



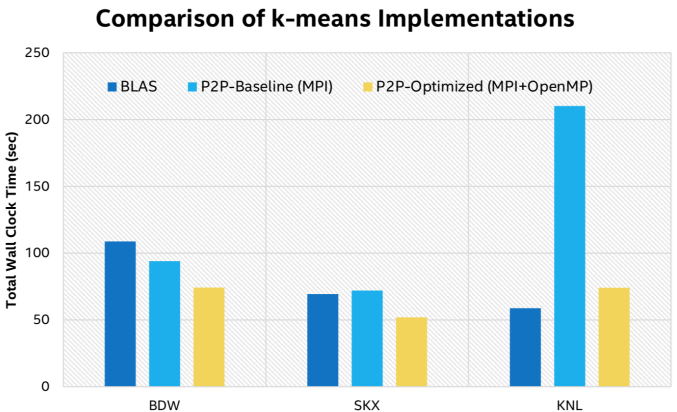


- Hybrid MPI-OpenMP implementation enables to effectively use hyper threads/logical threads
- BDW: 26% improvement with 9 MPI and 8 OMP
- SKX: 38% improvement with 10 MPI and 8 OMP

Improving computational intensity

- Can achieve greater computational intensity of the observation–centroid distance calculations by expressing the calculation in matrix form:
  - For observation vector x_i and centroid vector z_j, the squared distance between them is $D_{ij} = \left\|x_i - z_j\right\|^2$.
  - Via binomial expansion, $D_{ij} = \|x_i\|^2 + \left\|z_j\right\|^2 - 2x_i \cdot z_j$
  - The matrix of squared distances can thus be expressed as $D = \overline{x}\mathbf{1}^\intercal + \mathbf{1}\overline{z}^\intercal - 2X^\intercal Z$, where X and Z are matrices of observations and centroids, respectively, stored in columns, $\overline{x}$ and $\overline{z}$ are vectors of the sum of squares of the columns of X and Z, and $\mathbf{1}$ is a vector of all 1s.
- Above expression can be calculated in terms of a level-3 BLAS operation (xGEMM), followed by two rank-one updates (xGER, a level-2 operation).
- We use highly optimized BLAS implementations from Intel's MKL and NVIDIA cuBLAS to speed up distance calculations on Xeon Phi and GPGPUs, respectively.
- Distance calculations using above formulation can be dramatically faster than the straightforward loop over vector distance calculations when many distance comparisons must be made.
- Using the matrix formulation for distance comparisons in early k-means iterations is straightforward; a more complicated approach we hope to explore is using the matrix formulation in combination with the acceleration techniques described above, in which only a subset of observation–centroid distances are calculated.

Performance Summary



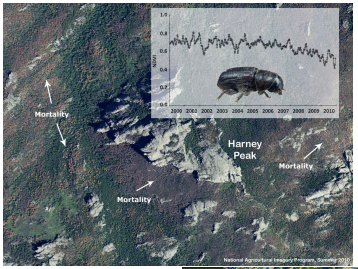
- BLAS formulation provides the best performance on KNL (despite doing many more distance calculations than P2P calculations using triangle-inequality "acceleration"), slightly slower then P2P distance calculation on SKX.
- Overall performance improvements:
  - KNL: 3.5X
  - BDW: 1.3X
  - SKX: 1.4X

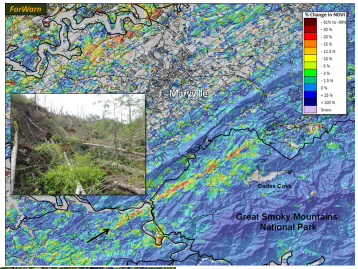
Future Directions: pKluster Software Development

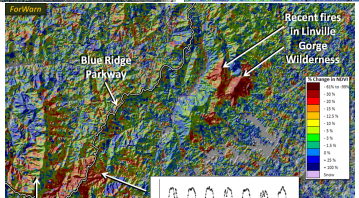
- Investigate hybrid approach combining accelerated k-means method and matrix formulation within the same iteration.
- Re-implement a fully distributed, masterless approach in the current version of the code to handle cases in which master-slave overhead is high (e.g., many cases on `KNL`).
- Add support for emerging high-capacity, non-volatile memory technologies.
- Supported open-source release under Apache License 2.0.

Application: Early detection of forest threats via remote sensing

- Early identification of forested areas threatened by insects, disease, drought, or other agents can be critical to preventing long-term or irreversible damage to forest ecosystems.
- $>$ 600 million acres of forest/wildlands in the United States, so regularly monitoring any significant fraction of these lands through aerial surveys and ground-based inspections is infeasible.
- Thus, many threats go unnoticed until it is too late to easily mitigate or correct them.







Phenology

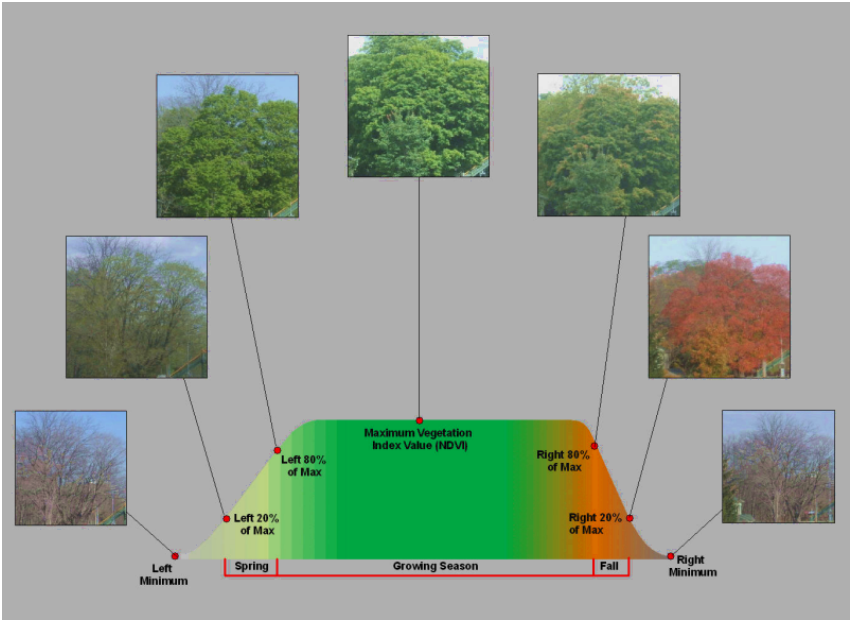
- Phenology is the study of periodic plant and animal life cycle events and how these are influenced by seasonal and interannual variations in climate.
- *ForWarn* is interested in deviations from the "normal" seasonal cycle of vegetation growth and senescence.
- NASA Stennis Space Center has developed a set of National Phenology Datasets based on MODIS NVDI.
- NDVI exploits the strong differences in plant reflectance between red and near-infrared wavelengths to provide a measure of photosynthetic capacity or "greenness" from remote sensing measurements.

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

- NVDI ranges from -1 to 1; 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.

Up-looking photos of a scarlet oak showing the timing of leaf emergence in the spring.

Prototypical NDVI ("Greenness") Profile Over One Year



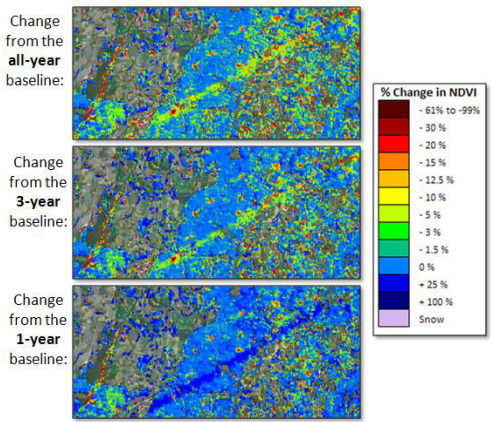
MODIS MOD13 NDVI Product

- 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 products at varying spatial (231 m, 1 km, 0.05°) and temporal (8-day, 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 Aqua and Terra MODIS products at 231 m resolution, processed by NASA Stennis Space Center.

The Forest Change Assessment Viewer

ForWarn ~~is currently~~ will soon resume providing interactive forest disturbance detection maps though the U.S. Forest Change Assessment Viewer: http://forwarn.forestthreats.org/fcav

- Maps computed through raster map arithmetic approaches: current NDVI compared with values from some historical baseline.
- E.g., current maximum NDVI observed over a 24-day window at a given location may be compared with the maximum NDVI for the same location/window observed over a set of previous years.

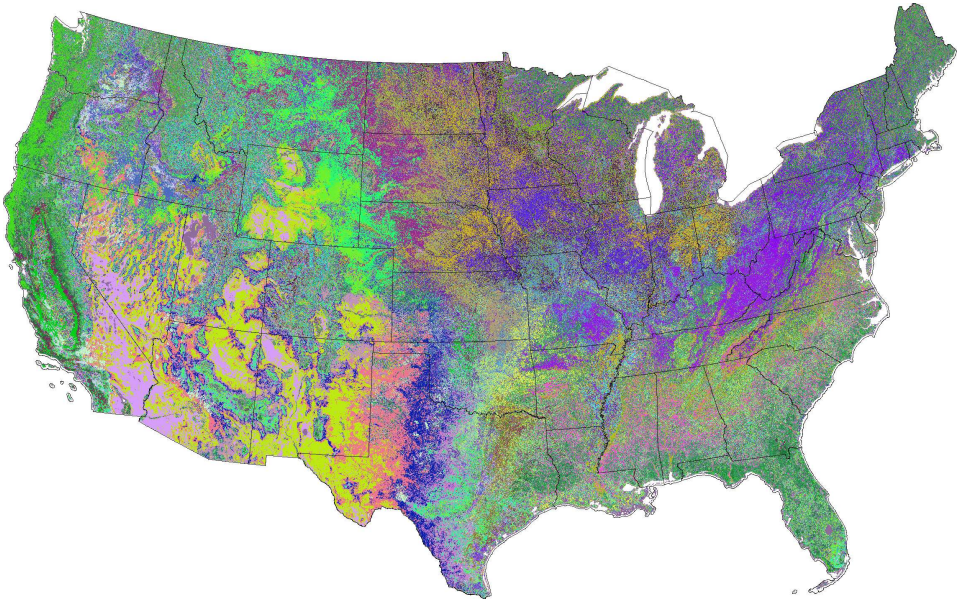


- A difficulty with map-arithmetic approaches:
  identification of appropriate parameters (maximum NDVI, 20% "spring" NDVI, etc.) to use, since the appropriate choice of parameters may vary by region and/or type of disturbance.
- To complement such approaches, we desire automated, unsupervised approaches to determine "normal" seasonal/inter-seasonal variation at each geographic location, using the full volume of NDVI data (almost 400 GB in single precision).
- Some approaches we have had success with are based on k-means clustering (see ICCS/DMESS 2011 papers by Mills et al., Kumar et al.).

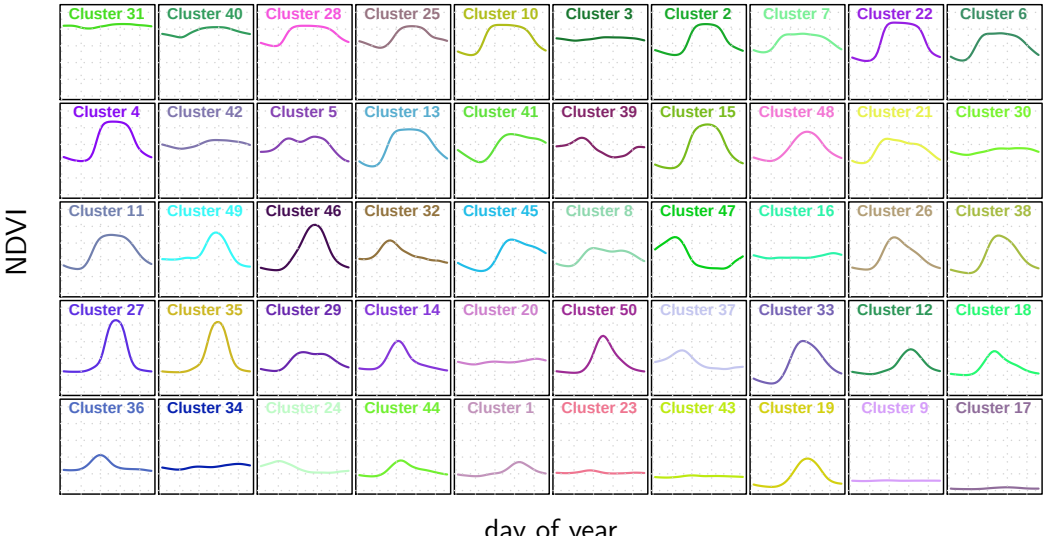
Clustering the NDVI data set

- For each year and each grid cell in the CONUS, construct an observation vector of 46 NDVI values representing the seasonal NDVI trace for that year/location.
- All observation vectors are combined into a data matrix with 46 columns and hundreds of millions of rows (each year corresponds to 146.4 million rows; 25 GB of single-precision data per year).
- Data are standardized and then clustered using a highly-parallel k-means clustering code.
- Cluster assignments are then mapped back to each map cell and year from which each observation came, yielding one map per year in which each cell is classified into one of k clusters or "phenoclasses".
- These can be viewed as forming a dictionary of prototypical annual NDVI traces.

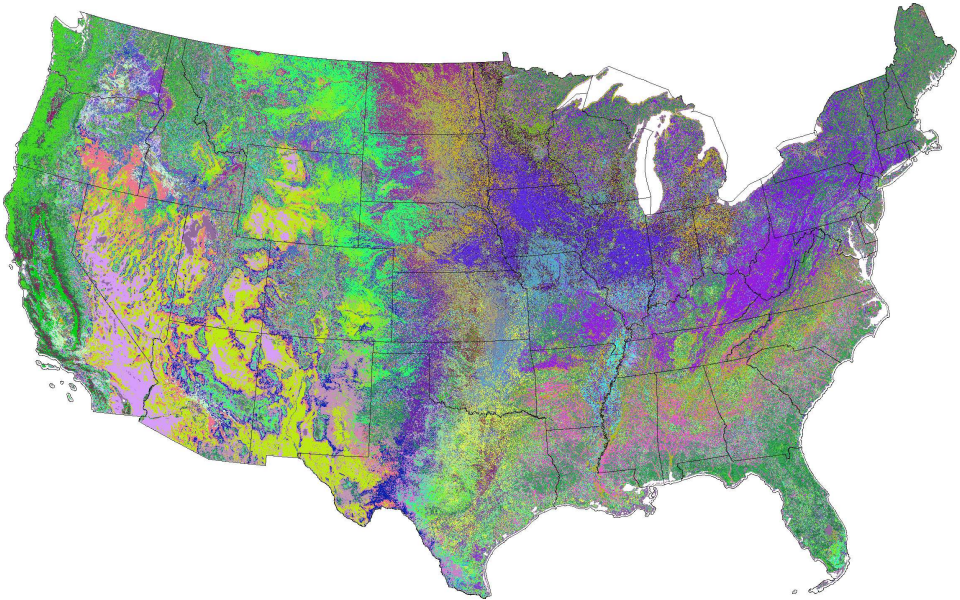
50 Phenoregions for year 2011 (Random Colors)



50 Phenoregion Prototypes (Random Colors)



50 Phenoregions Max Mode (Random Colors)

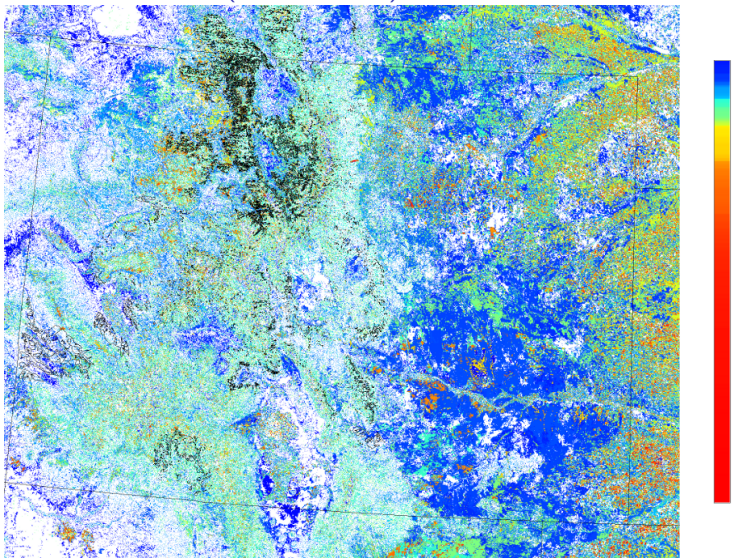


Clustering-based disturbance detection

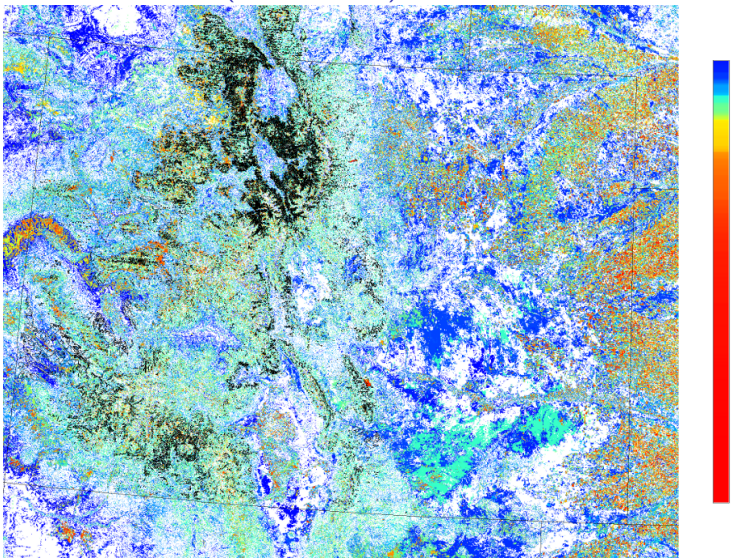
Disturbance or recovery can be detected by analyzing the history of phenoclass assignment. E.g.:

- Look for significant deviation from the statistical mode of cluster assignments for that location.
- Look for a large Euclidean "transition" distance between the currently assigned cluster centroid and those from a prior year or years.

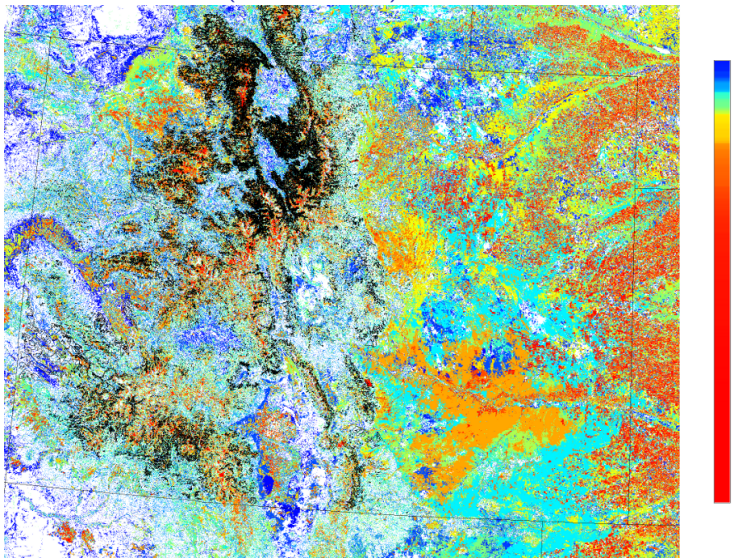
Mountain Pine Beetle in Colorado for (2004 – 2003)



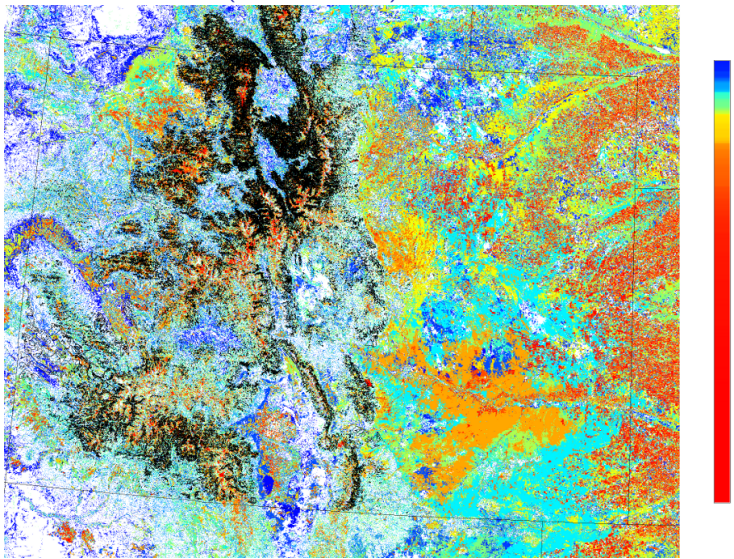
Mountain Pine Beetle in Colorado for (2005 – 2003)



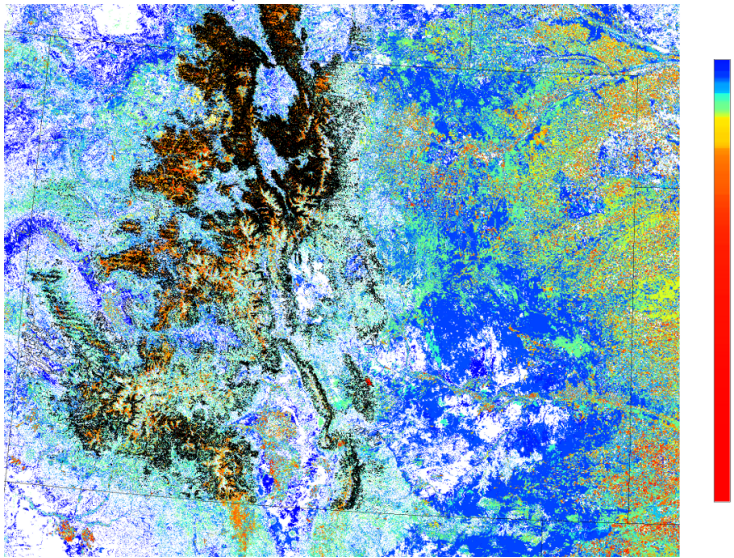
Mountain Pine Beetle in Colorado for (2006 – 2003)



Mountain Pine Beetle in Colorado for (2007 – 2003)



Mountain Pine Beetle in Colorado for (2008 – 2003)

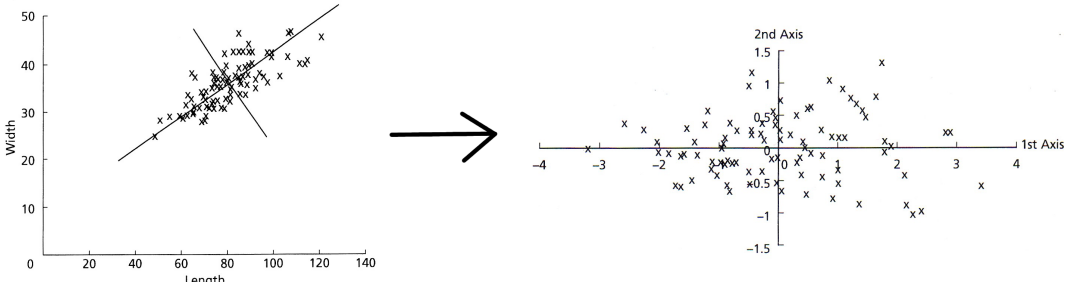


PCA/SVD approaches for threat detection

- Our clustering-based approaches can flag a wide range of disturbances, particularly those involving high mortality events such as fire, storms, or mountain pine beetle outbreaks.
- Slower-acting agents, such as hemlock woolly adelgid, that cause a gradual decline in forest health are more difficult.
- Also, the annual phenology of some areas is highly influenced by interannual climate variability: grasslands, for instance, experience rapid greenup after precipitation and do not have smooth annual cycles.
- These areas tend to display a large transition distance from year to year even when there is essentially no real change in the vegetation health.
- To remedy these shortcomings, we have been exploring the use of principal components analysis (PCA) (or the related SVD) as a complementary approach.

A complementary approach: Principal component 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

Varimax-rotated loadings for top 3 components

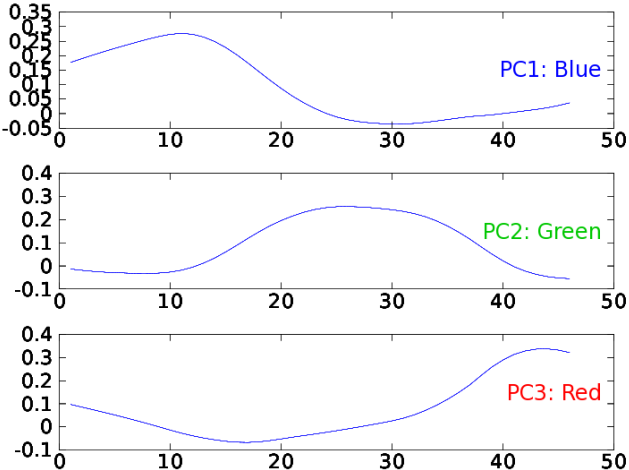
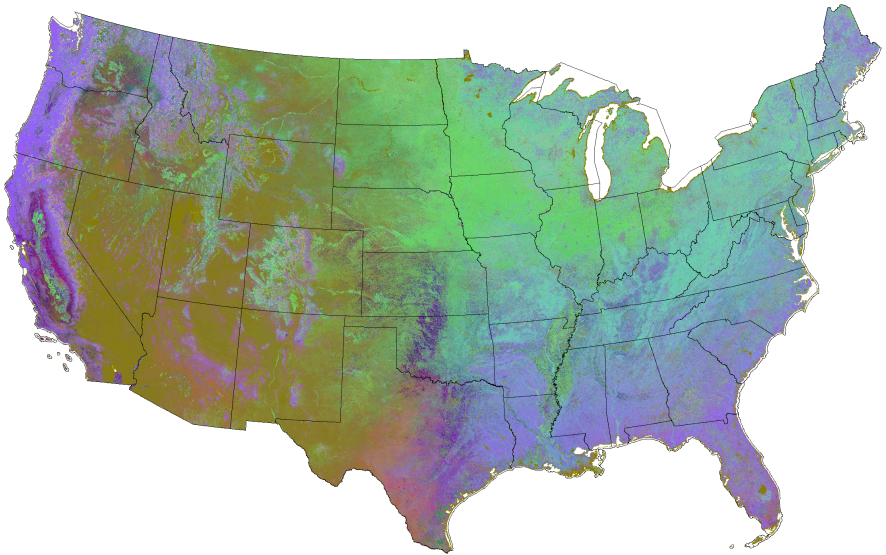


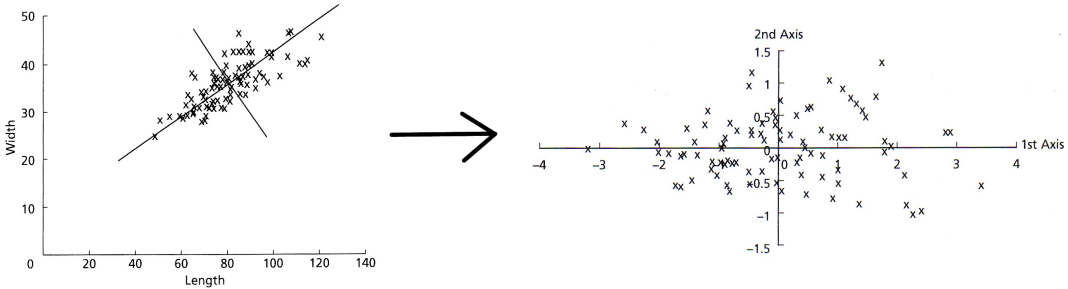
Figure: The loadings (coefficients in the linear combination of the 46 original variables) along the three varimax-rotated principal axes. The x-axis corresponds to the eight-day NDVI-acquisition windows and loadings are shown on the y-axis.

$k = 1000$ map for year 2000, similarity colored



A complementary approach: Principal component 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.

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" ($m >> n$) nature of our matrices: ($m >> n$)
  - Form reduced factorization $\mathbf{A} = \mathbf{QR}$ (via parallel PLAPACK routine)
  - Gather the matrix $\mathbf{R}$ to process 0.
  - Process 0 calls LAPACK DGESVD to compute the SVD $\mathbf{R} = \mathbf{USV}^T$.
  - Optionally, back transform $\mathbf{Q}$ to get $\mathbf{Q} \leftarrow \mathbf{QU}$.
  - Final SVD is: $\mathbf{A} = \mathbf{QSV}^T$
- A serial bottleneck exists where the SVD of $\mathbf{R}$ is computed, but this matrix is so small (only 46×46 for our NDVI data set) that this serial portion is essentially negligible.

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: $x = \hat{x} + \tilde{x}$
  - $P = \begin{bmatrix} v_1 & v_2 & \dots & v_q \end{bmatrix}$
  - $\hat{x} = PP^T x = Cx$ and $\tilde{x} = (I - PP^T)x = \tilde{C}x$
  - Abnormal if SPE $= \|\tilde{x}\|^2 = \left\|\tilde{C}x\right\|^2$ 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.

Detecting anomalies within single year, single 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.

NEON Domains

Detecting anomalies within single year, single 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.

Colorado and Southern Wyoming, 2008

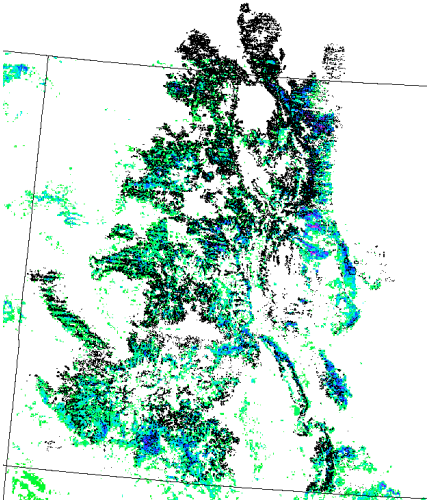


Figure: Portion of the Southern Rockies–Colorado Plateau NEON Domain for year 2008, showing map cells scoring in the 85th percentile. Black polygons show damaged areas noted in aerial detection surveys; extensive damage due to mountain pine beetle and sudden aspen decline are evident.

Vicinity of Louisiana Coast: Hurricane-induced disturbance

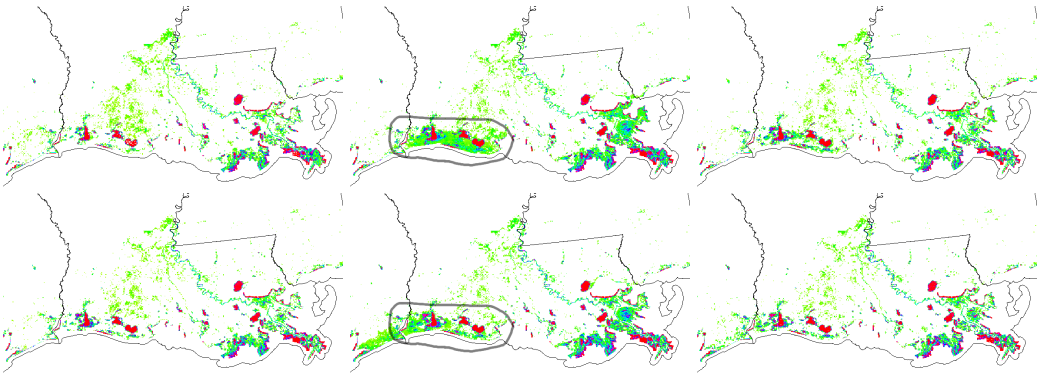


Figure: Portions of the PCA-based anomaly maps (map cells scoring in the 90th percentile are shown) for the Southeast NEON Domain for years 2004–2009, showing the area in the vicinity of the Louisiana coast. From left to right, the top row shows years 2004, 2005, and 2006, respectively, and the bottom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

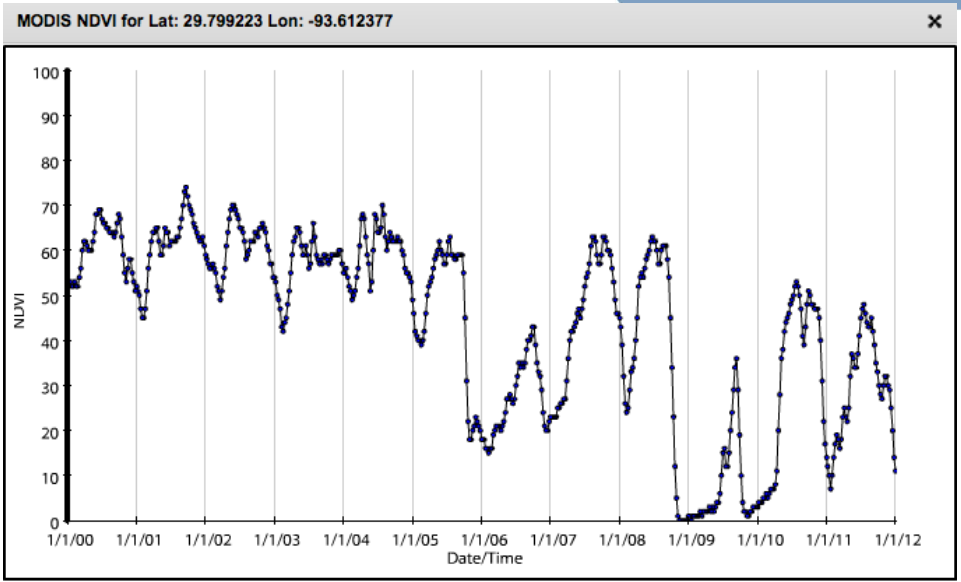


Figure: NDVI trajectory as viewed via the Forest Change Assessment Viewer for a location (close to the center of the circled region in the previous figure) near the coast in southwestern Louisiana showing apparent hurricane-induced mortality from events in 2005 and 2008.

Southern Appalachians: Hemlock decline

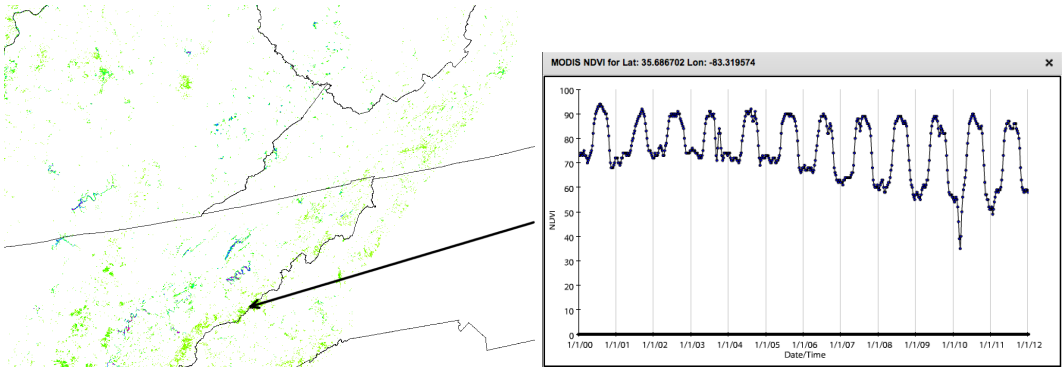


Figure: At left, a portion of the PCA-based anomaly map (map cells scoring in the 90th percentile are shown) for the Southern Appalachians/Cumberland Plateau NEON Domain for year 2010. The arrow indicates a location thought to be affected by hemlock woolly adelgid, and the corresponding NDVI trajectory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

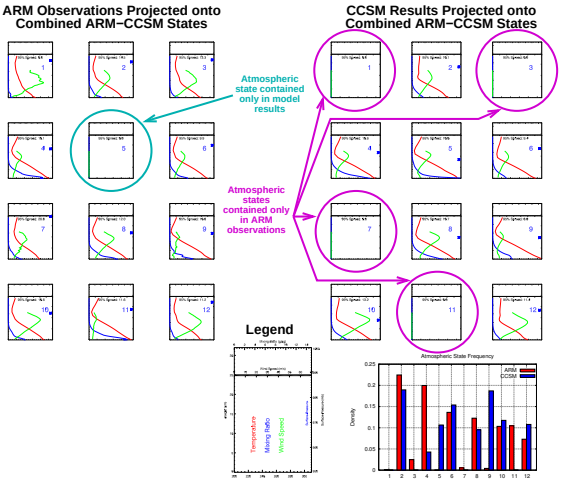
Future Directions: Possible Science Goals

We have a few scalable tools suitable for analyzing large (multi-TB) geo-spatio-temporal data sets. What other interesting things could we do with them?

- Potential questions of interest:
  - How are global plant distributions affect by climate change?
  - What are the implications for global carbon budgets and feedbacks to climate?
  - What changes do we expect to key events like onset of growing season?
  - What changes do we expect to suitable growing ranges for crops?
  - Are there policy implications for agriculture and ensuring the food supply?
- Could combine analysis to all of the MODIS vegetative phenology record with global fine-scale meteorological reanalysis and possibly other ancillary data layers.
  - Enables attribution of vegetation changes to climate or other events.
  - Study directly observed vegetation responses to extreme events.
- Could analyze high-resolution and/or multi-model ensemble Earth system model simulations:
  - Project changes to distribution of eco-phenoregions (identified by the historical analysis) for different climate change scenarios.
  - Combine with crop physiology models to project changes in yields.
  - Combine with urban growth models or population models to assess resource planning, policy scenarios, and crop futures.
- Another item of interest: model-data and model-model comparison

Cluster analysis employed to compare ARM observational data at the Southern Great Plains (SGP) site with corresponding 6-hourly output from an integration of the Community Climate System Model (CCSM) run under the IPCC SRES A2 scenario for the current decade.

- State 5 (very high humidity and temperature at the surface) has no analog in the observational data.
- States 1, 3, and 7 have very low frequency in the observations (see frequency plot), so their absence from model predictions does not suggest a problem.
- State 11 (high humidity and temperature with very low wind shear), is never predicted by CCSM.
- CCSM predicts over-abundance of state 9 (low humidity and high temperature conditions) while under-representing state 4 (moderate humidity, temperature, and shear conditions).
- Misrepresentation of atmospheric states in CCSM over the SGP site could have impacts on predictions of cloud formation and hence the local radiation budget.



https://www.climatemodeling.org/arm/

Bonus: Machine-learning based scale bridging in hyperresolution simulations

A daunting grand challenge problem for an exascale-class machine—hyperresolution land surface modeling—and a possible (possibly not possible) approach (or two).

Motivation for hyperresolution land surface models

- Land surface models (LSMs) usually run at spatial resolutions at which it is computationally feasible to run a coupled climate model (about $O(10)$ km for a current very high-resolution global model).
- There are compelling arguments for running LSMs (with more realistic hydrologic process models) at dramatically higher resolutions.
- "Hyperresolution" models would enable representation of several important processes related to C and N cycling, e.g.,
  - Accurate prediction of denitrification rates; requires fine resolution of topographic heterogeneities and regions of high soil moisture where redox conditions are favorable to denitrification.
  - CO_2 outgassing from river channel systems too fine to represent in current resolution models; these fluxes can be very significant.

Amazon Basin fine-scale river channel systems

An estimated 0.5 Gt/year of carbon is outgassed in the Amazon Basin, much of it from small streams.

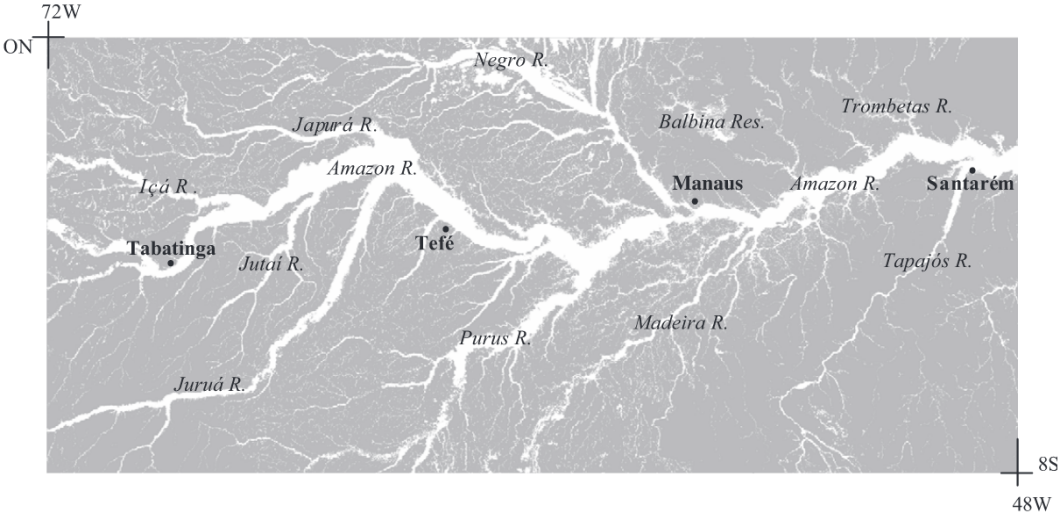
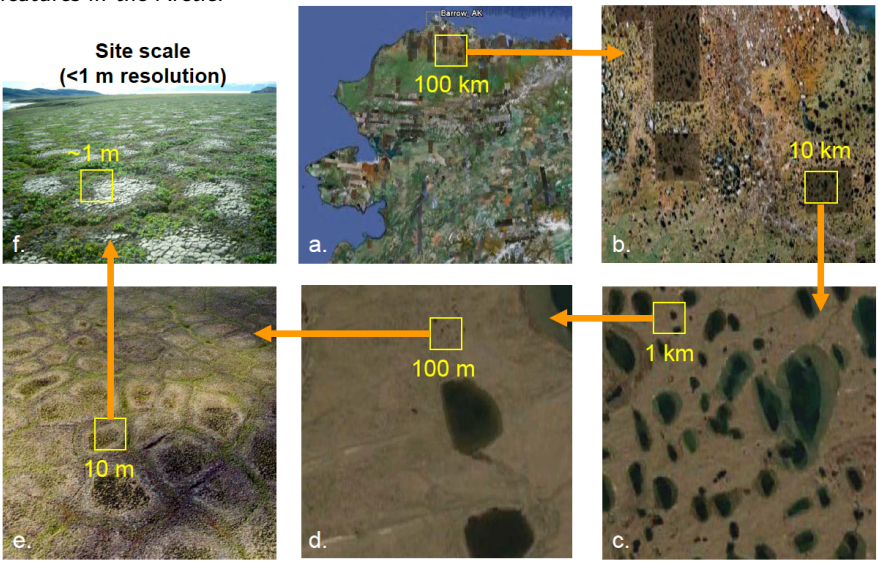


Figure: The central Amazon basin; wetlands (white pixels) occupy 17% of the total area. Figure from Hess et al. 2003, *Remote Sensing of Environment*, v. 87.

- Sophisticated surface–subsurface hydrology and reactive transport models—needed for accurate representation of surface/subsurface water dynamics and biogeochemical processes at high resolutions—have recently been coupled to LSMs.
- Some of these models (e.g., PFLOTRAN, ParFlow) can fully utilize leadership-class supercomputers, enabling dramatically higher-resolution simulations of global hydrology and biogeochemistry in LSMs as supercomputer power grows.
- But *even with all foreseeable advances in computing power and solver algorithms, global LSM simulations will not resolve some important processes that are difficult to represent via sub-grid parameterizations.*

Scaling challenge: Permafrost-affected Arctic regions

Linear scaling assumptions are a poor fit for the complex organization of fine- and intermediate-scale features in the Arctic.



NGEE–Arctic upscaling/downscaling approach

- Construct mechanistic, process-resolving models accurate at small scale.
- Conduct series of simulations over sequence of nested computational domains ranging from fine to global climate-grid resolutions.
- Ensemble of simulations at finer scale analyzed to produce parameter values for a coarser scale;
- Coarser scale simulations analyzed to generate parameters for yet coarser simulations, or for improved boundary conditions for repeating finer-scale runs.
- This is an off-line approach.

Representativeness-based sampling network design

- The scaling approach cannot rely solely on simulations; it must be constrained by observations.
- Geospatiotemporal clustering (GSTC) can be used to stratify sampling domains, inform site selection, and provide a basis for up-scaling and extrapolating measurements to land areas within and beyond the sampling domains.

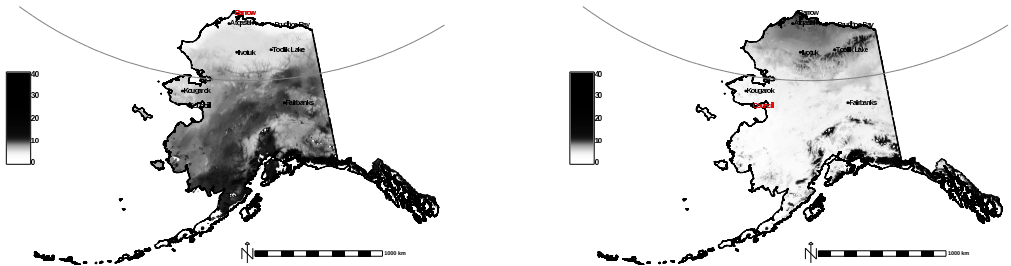
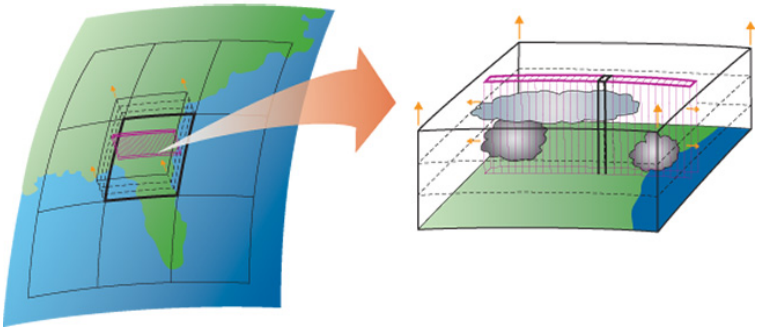


Figure: Representativeness (closeness in terms of Euclidean distance in the data-space of eco-climatic variables used in the cluster analysis) for present-day conditions for two potential NGEE–Arctic sites considered in [?] (Barrow on the left, Council on the right).

Superparameterization in atmospheric models



From http://www.ucar.edu/communications/quarterly/summer06/cloudcenter.jsp.

- Embed small-scale 2D or quasi-3D model inside each "global scale" cell.
- Embedded models employ periodic lateral boundary conditions, don't interact with each other (except indirectly through fluxes on the global-scale grid).
- Couple by enforcing the property that the horizontal average of a small-scale variable is exactly equal to the value of the corresponding large-scale value.

Superparameterization-inspired online upscaling in hydrologic models

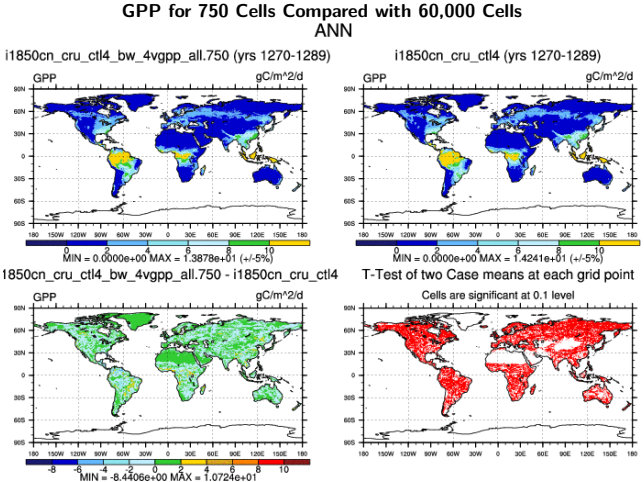
- Many hydrologic processes work in an intrinsically 3D space, but embedding 3D small-scale grids in every global grid cell is computationally infeasible.
- *Could the same GSTC-based techniques used for choosing ecological sampling sites be used to choose a sparse set of "measurement" sites where fine-scale, 3D models are to be embedded*?
  - Periodically run an on-line clustering using the parameters and state variables from all cells in the global grid.
  - Group cells with similar states, properties, and forcings together into clusters.
  - In representative subset of cells from each cluster, run an embedded fine-scale model.
  - Map quantities of interest, e.g., biogeochemical reaction rates, from the sparse collection of fine-scale models back to other global cells that are members of the same (or similar) clusters.

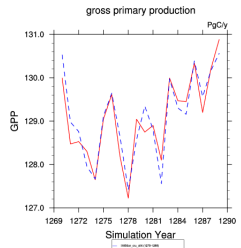
Cluster-based upscaling from sparse fine-scale models to global

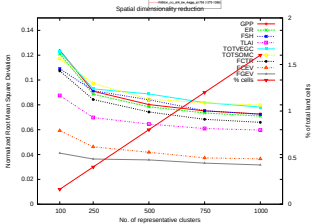
- Simplest approach is "paint by numbers": if only one member of a cluster has an embedded model, assign the same upscaled value from that member to all others of the cluster; if there are multiple members of the cluster that run embedded fine-scale models, assign other members the value from the closest (in data space) member or some average.
- Alternatively, construct model response surfaces from the ensemble of fine-scale simulations within a cluster.
  - Such an approach was used by Luxmoore et al. in [**?**] and [**?**] to scale the results of small-scale plant physiological responses to various environmental factors up through a hierarchy of models to the regional ecosystem scale.

Cluster-based spatial sparsification for global CLM simulations

- 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 currently computationally untenable.
- Cluster analysis uses the CRU-NCEP climate data, plant functional type (PFT) characteristics, and steady-state modeled quantities.







Open questions

I believe that the geospatiotemporal clustering-guided approach outlined presents a viable way to adaptively and parsimoniously select locations in which embedded ultra high-resolution model patches should be run, but there are many open questions:

- What level of division (i.e., the number of clusters) should be employed?
- How many different levels of scale must be used?
- What is the best (or at least adequate) approach for upscaling?
- What is a good approach to prescribing boundary conditions for the embedded fine-scale models?

I hope to explore these and other questions with the climate science and computational mathematics communities.

Data assimilation/Bayesian approaches?

Alternative idea: Treat results derived from fine-scale models as "measurements" in a data assimilation (DA) framework.

- DA systems estimate true state by blending data from mathematical models with available observations, according to estimates of the expected errors.
- In atmospheric prediction, major source of errors is initial conditions, and DA system adjusts the model by determining corrections to initial state.
- We need to perturb evolution of the global model state; done naively, will violate conservation and positivity.
- However, Jacobs and Ngodock [**?**] have constructed schemes honoring physical conservation laws within 4DVAR by applying corrections to flux terms instead of conserved quantities themselves.
- Is such an approach feasible here?

References cited


F. M. Hoffman, J. Kumar, R. T. Mills, and W. W. Hargrove, "Representativeness-based sampling network design for the State of Alaska," *Landscape Ecology*, vol. 28, no. 8, pp. 1567–1586, Oct. 2013.


R. J. Luxmoore, W. W. Hargrove, M. L. Tharp, W. M. Post, M. W. Berry, K. S. Minser, W. P. Cropper Jr, D. W. Johnson, B. Zeide, R. L. Amateis *et al.*, "Signal-transfer modeling for regional assessment of forest responses to environmental changes in the southeastern united states," *Environmental Modeling & Assessment*, vol. 5, no. 2, pp. 125–137, 2000.


R. J. Luxmoore, W. W. Hargrove, M. Lynn Tharp, W. Mac Post, M. W. Berry, K. S. Minser, W. P. Cropper, D. W. Johnson, B. Zeide, R. L. Amateis *et al.*, "Addressing multi-use issues in sustainable forest management with signal-transfer modeling," *Forest Ecology and Management*, vol. 165, no. 1, pp. 295–304, 2002.


G. Jacobs and H. Ngodock, "The maintenance of conservative physical laws within data assimilation systems," *Monthly weather review*, vol. 131, no. 11, pp. 2595–2607, 2003.