Classification and Delineation of Large Earth Science Data

Jitendra Kumar[α], Forrest M. Hoffman[α], William W. Hargrove[β]

[α]Oak Ridge National Laboratory, [β]USDA Forest Service

Climate Data Analytics

- Identification of ecoregions 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)

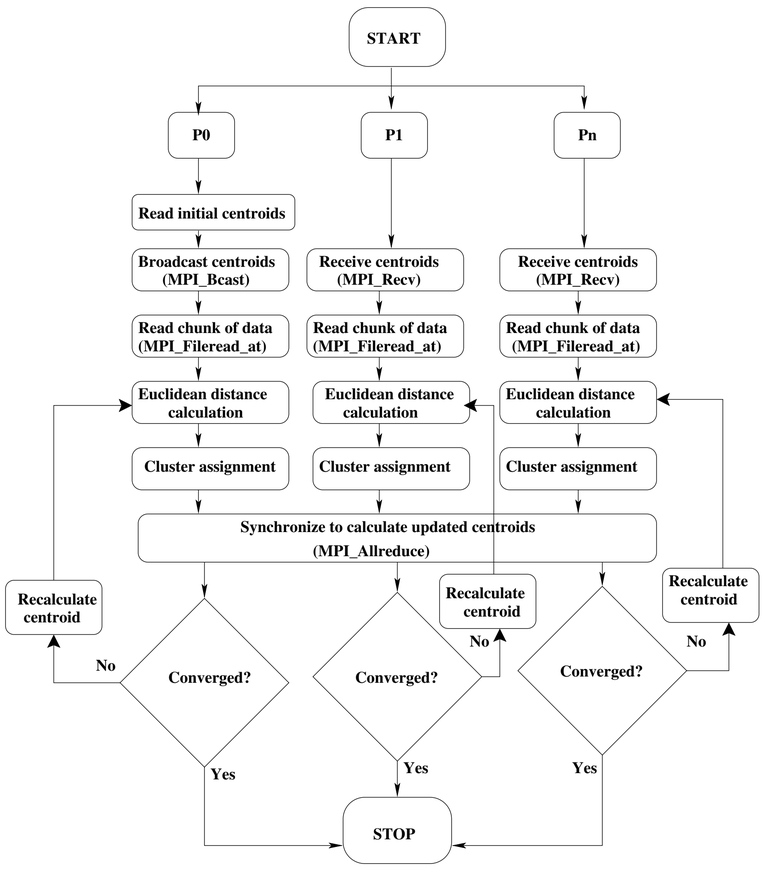

Figure 1: *The parallel k-means algorithm*

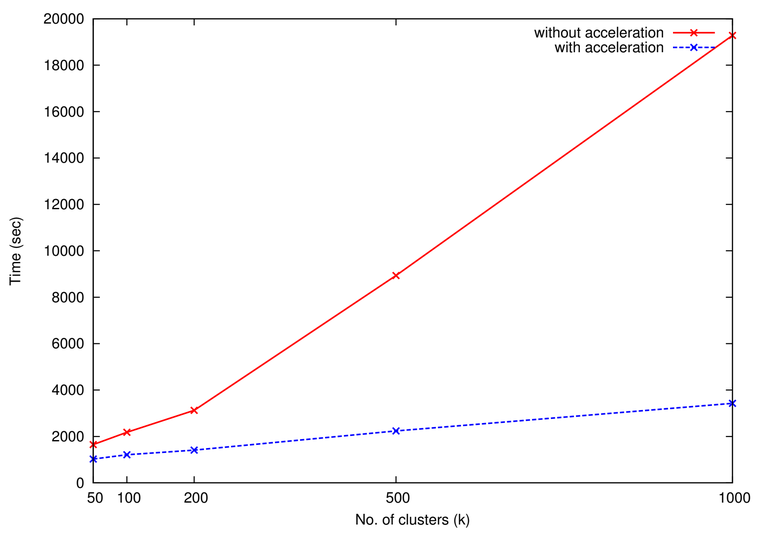

Figure 2: *Accelerating k-means*

Scaling and Optimization

- We have optimized the Multivariate Spatio-Temporal Clustering (MSTC) tool for excellent parallel performance on Titan Cray XK6 at ORNL (Figure 3)
- Two phase (read + scatter) parallel I/O was implemented using MPI I/O and optimized for performance on Lustre filesystem on OLCF machines (Figure 4)
- The tool has been applied for a wide range of data sets up to hundreds of GBs in size

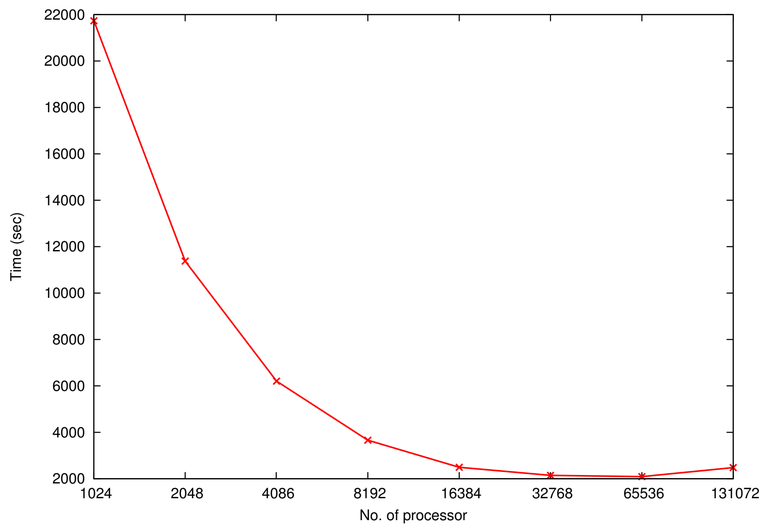

Figure 3: *Parallel scaling (k=1000, NDVI 2000–2011)*

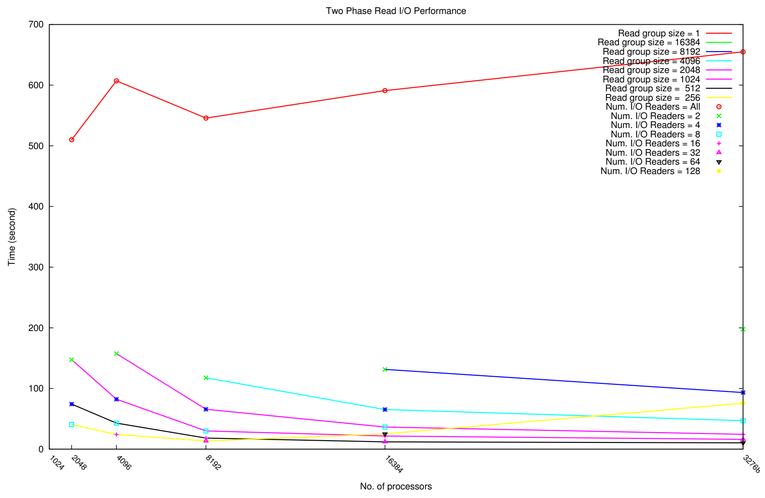

Figure 4: *Parallel I/O performance and optimization*

Geo-spatial Analysis and Visualization

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

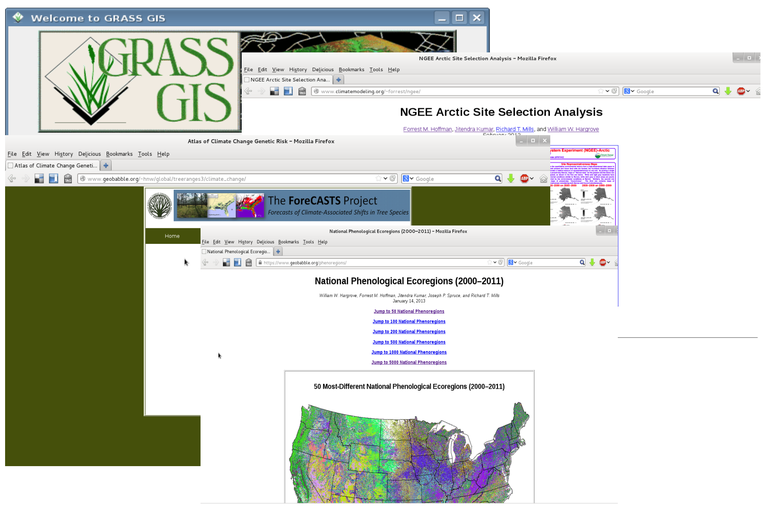

Figure 5: *Open source tools for analysis and data sharing*

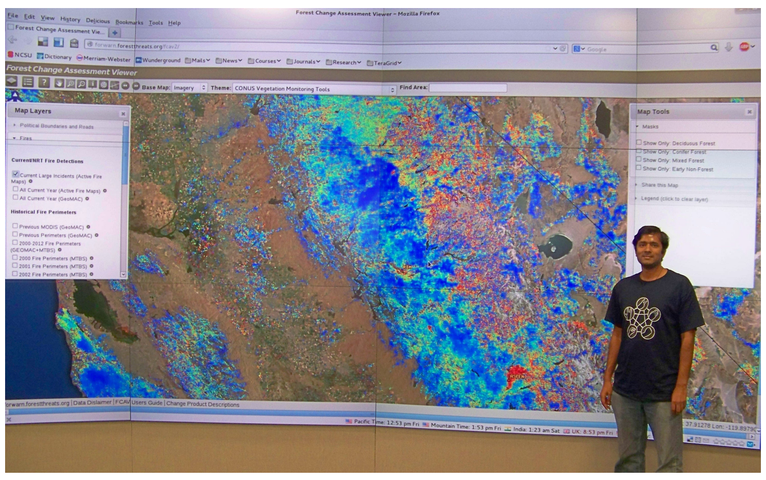

Figure 6: *The EVEREST Visualization Facility provides a unique opportunity for analysis of very large simulation output and high resolution data products*

Forest Threat Detection

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

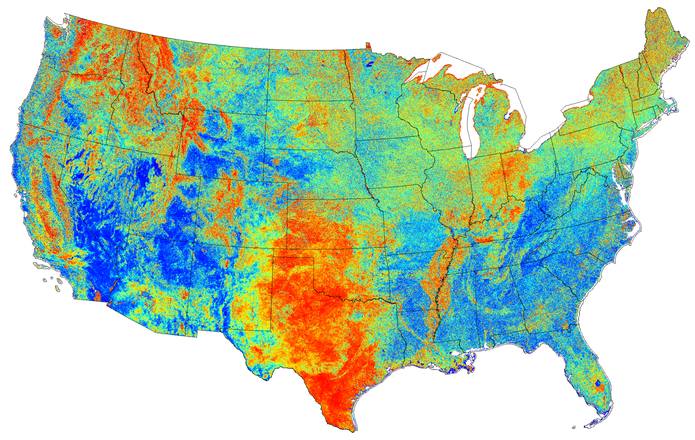

Figure 7: △ *Integreated NDVI disturbance map*

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 domains and determining representativeness of sites (Figure 9)

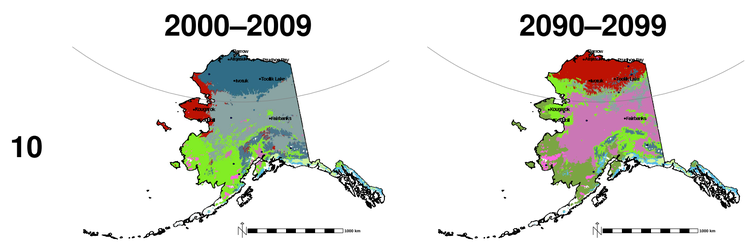

Figure 8: *Ecoregions and Representative Sites*

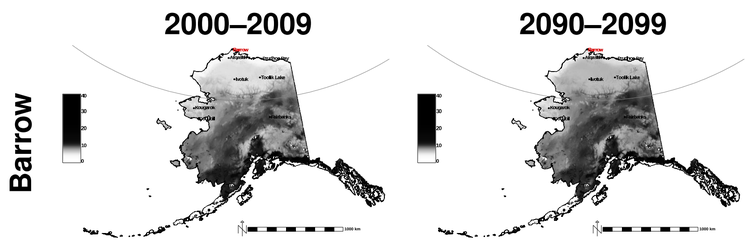

Figure 9: *Site and Network Representativeness*

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

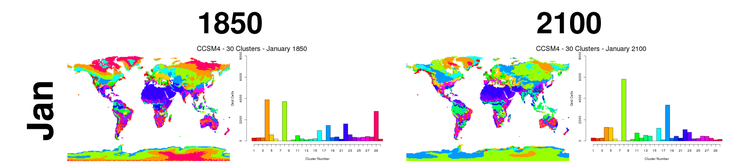

Figure 10: *Shifting climate regimes were defined using clustering and tracked through time*

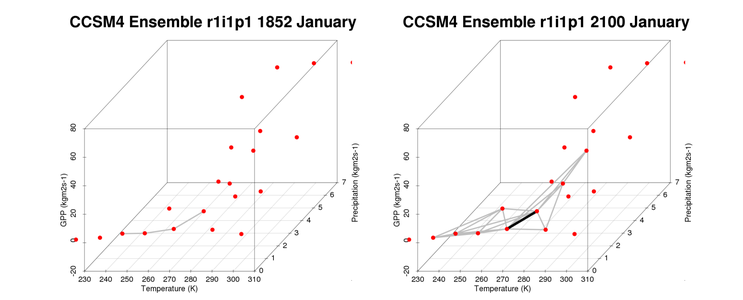

Figure 11: *Centroids form a skeleton in state space*

Climate Change Science Institute
AT OAK RIDGE NATIONAL LABORATORY

Jitendra Kumar: jkumar@climatemodeling.org
Forrest Hoffman: forrest@climatemodeling.org
William Hargrove: hnw@geobabble.org

OAK RIDGE National Laboratory