Data Mining in Earth System Science (DMESS 2011)

Forrest M. Hoffman[a,b], J. Walter Larson[c,d,e], Richard Tran Mills[a,f], Bjørn-Gustaf J. Brooks[g], Auroop R. Ganguly[a], William W. Hargrove[h], Jian Huang[f], Jitendra Kumar[a], and Ranga R. Vatsavai[a]

[a]Oak Ridge National Laboratory, [b]University of California-Irvine, [c]Argonne National Laboratory, [d]University of Chicago, [e]Australian National University, [f]University of Tennessee, [g]University of Wisconsin, [h]U.S. Dept. of Agriculture-Forest Service

June 2, 2011

International Conference on Computational Science (ICCS 2011)

Nanyang Technological University, Singapore

Introduction

- Earth science data span many orders of magnitude in space and time scales.
- These data are increasingly large and complex, often representing long time series, making them difficult to analyze, visualize, interpret, and understand.
- Electronic data storage and high performance computing capacity enable creation of large data repositories and detailed empirical and process-based models.
- The resulting "explosion" of heterogeneous, multi-disciplinary Earth science data requires use of new analysis methods and development of highly scalable software tools.

Earth Science Data

- Observational and modeled data encompass temporal scales of seconds to millions of years (10^0–10^{13} s) and spatial scales of microns to tens of thousands of kilometers (10^{-6}–10^7 m).
- Integrating and synthesizing data across Earth science disciplines offers new opportunities for scientific discovery.
- The rise of data-centric science is becoming recognized as the *fourth paradigm of discovery* alongside the experimental, theoretical, and computational archetypes (Hey et al., 2009).
- However, the promise of data-intensive Earth science has yet to be realized because of the unique technological and social challenges it poses.

Model Results

- Open and user-friendly access to Earth science data is required—particularly for climate science—as interest in sustainability and environmental policy has added decision-makers and the public to the list of data users.
- Organized global climate modeling activities, like the Coupled Model Intercomparison Project (CMIP), can generate tens of terabytes to several petabytes of simulation results (Overpeck et al., 2011).
- CMIP results are now made available to the research community and the public through distributed, interconnected servers called the Earth System Grid (ESG; Williams et al., 2009).
- Composited, summary data from collections of simulation output are being developed to make model results more directly useful outside of the climate science community.

Observational Data

- Satellite remote sensing data tend to be very large and grow quickly as spatial and temporal resolutions increase.
- Meanwhile, small ecological data sets are often the most valuable for synthesis, but may be the hardest to preserve, distribute, and use (Reichman et al., 2011).
- Data curation and provenance must be formally documented; data format standards and metadata conventions are needed.
- Scientific workflow systems are being developed to document and automate data processing, quality control, gap-filling, analysis, and synthesis.
- The DataONE project (`http://www.dataone.org/`) is pioneering technologies to automate and document every step, from data acquisition and generation to synthesis and publication.

Model Validation Using Measurements

- Model evaluation places new demands on the measurements community to provide observations and uncertainties useful for assessing model fidelity (Randerson et al., 2009).
- Researchers need agreed-upon standards for benchmarks of scientific model performance.
- The International Land Model Benchmarking (ILAMB) project (`http://www.ilamb.org/`) was recently established to develop benchmarks for terrestrial biogeochemistry models.
- ILAMB will create a reusable and extensible, open source framework for evaluating metrics and generating diagnostics.
- By using freely available observational data and distributing its evaluation tools, ILAMB seeks to achieve a new standard for scientific openness and transparency (Kleiner, 2011).

Data Mining Approaches

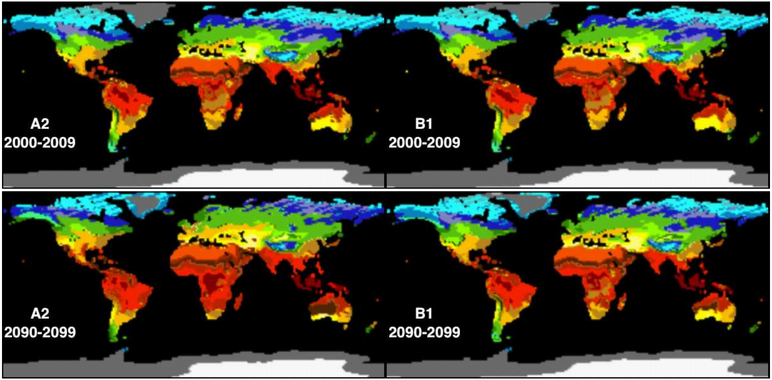
- Much of today's large and complex Earth science data cannot be synthesized and analyzed using traditional methods on small desktop computers.
- Data mining algorithms and tools can be used to extract knowledge and information from observations and model data.
- Data mining, machine learning, and high performance visualization approaches that exploit distributed-memory parallel computational resources offer promising alternatives.
- Techniques include:
  - cluster analysis,
  - block entropy,
  - spectral and wavelet methods,
  - artificial neural networks, and
  - regression tree and model tree ensembles.

Cluster Analysis

- Clustering approaches have become an accepted method for stratifying environments and delineating ecoregions (Hargrove and Hoffman, 2004).
- The same method has proven useful for stratifying climate observational or model data, not only across space, but also through time (Hoffman et al., 2005).
- Further extension of clustering to comprehensive analysis of sampling network representativeness has been performed for existing measurement sites (Hargrove et al., 2003) and for the design of the new National Ecological Observatory Network (NEON) domains (Keller et al., 2008).

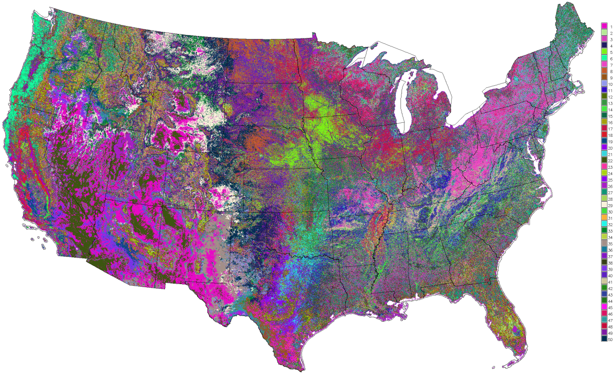
Cluster Analysis

Presented today: Sisneros *et al.* have integrated a similar, flexible stratification method into a high performance visualization system to demonstrate how life zone boundaries might change under scenarios of climate change.



Cluster Analysis

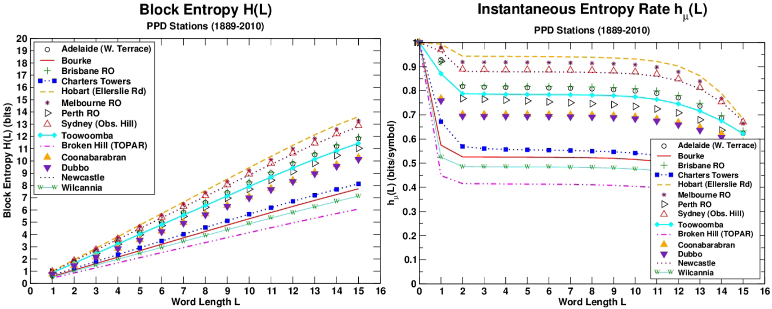
Presented today: Mills *et al.* present an updated analysis from seven years of satellite data, demonstrating the utility of cluster analysis in stratifying phenological behavior and in detecting forest disturbances from mountain pine beetle, wildfire, etc.



Block Entropy Methods

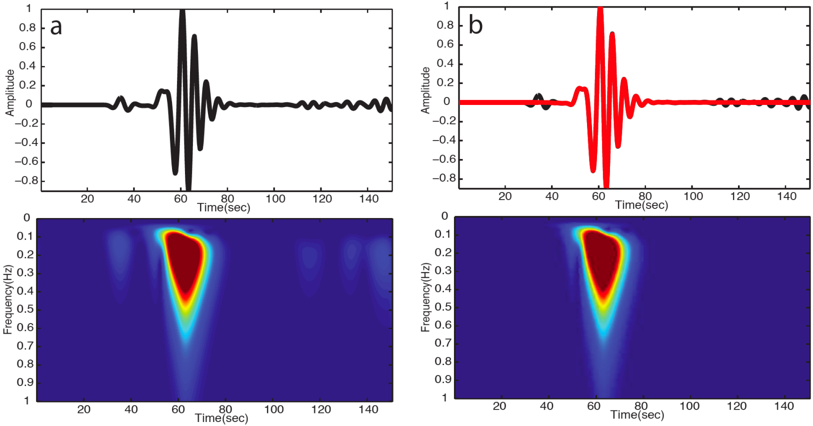
Information theoretic techniques, including the block entropy method, are useful for detecting and classifying state changes in environmental variables.

Presented today: Larson *et al.* present results from use of block entropy as a classifier for dynamical behavior in observed meteorological time series data from Australian weather stations.



Artificial Neural Networks

Presented today: Diersen *et al.* describe the use of ANN and an Importance-Aided Neural Network (IANN) to the refinement of structural models used to create full-wave tomography images.

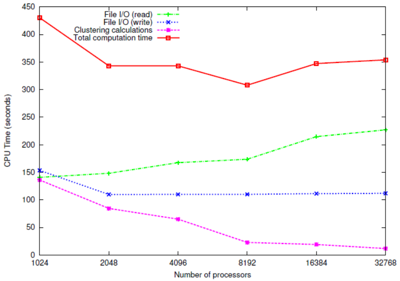


High Performance Computing

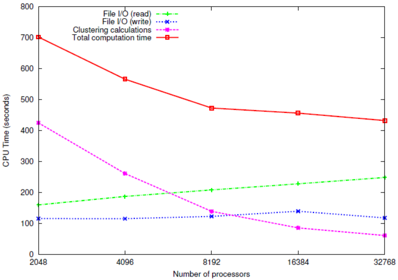
- Increasing computational capacity is required to realize the promise of new scientific discovery from Earth science data.
- New analysis techniques and highly scalable algorithms and software tools must be developed to enable analysis, exploration, and visualization of these data.
- Fortunately, the rapidly increasing computational power of supercomputers provides opportunities for development of such tools.
- Analysis and visualization could be another step in the scientific workflow process in the same computing environment as used for model experiments.

High Performance Computing

Presented today: Kumar *et al.* present a fully distributed version of a k-means clustering algorithm that includes several performance enhancement modifications and was designed and tested specifically for analysis of very large Earth science data sets using state-of-the-art supercomputers.



$k = 50$



$k = 1000$

Acknowledgments

Co-conveners:

- Forrest M. Hoffman (forrest@climatemodeling.org)
- J. Walter Larson (larson@mcs.anl.gov)
- Richard Tran Mills (rtm@utk.edu)

The DMESS 2011 co-conveners wish to thank the Workshop Program Committee for their assistance in reviewing submitted papers. The Program Committee consisted of **Michael W. Berry, Bjørn-Gustaf J. Brooks, Rebecca A. Efroymson, Sara J. Graves, William W. Hargrove, Jian Huang, Robert L. Jacob, Jitendra Kumar, Vipin Kumar, and Ranga R. Vatsavai**.

Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. Argonne National Laboratory is managed by UChicago Argonne, LLC, for the U.S. Department of Energy under Contract No. DE-AC02-06CH11357.

References

W. W. Hargrove and F. M. Hoffman. Potential of multivariate quantitative methods for delineation and visualization of ecoregions. *Environ. Manage.*, 34(Supplement 1):S39–S60, Apr. 2004. doi:10.1007/s00267-003-1084-0.

W. W. Hargrove, F. M. Hoffman, and B. E. Law. New analysis reveals representativeness of the AmeriFlux Network. *Eos Trans. AGU*, 84(48):529, 535, Dec. 2003. doi:10.1029/2003EO480001.

T. Hey, S. Tansley, and K. Tolle, editors. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Corporation, Redmond, Washington, USA, Oct. 2009. ISBN 978-0-9825442-0-4.

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.

M. Keller, D. Schimel, W. Hargrove, and F. Hoffman. A continental strategy for the National Ecological Observatory Network. *Front. Ecol. Environ.*, 6(5):282–284, June 2008. doi:10.1890/1540-9295(2008)6[282:ACSFTN]2.0.CO;2. Special Issue on Continental-Scale Ecology.

K. Kleiner. Data on demand. *Nature Clim. Change*, 1(1):10–12, Apr. 2011. doi:10.1038/nclimate1057.

J. T. Overpeck, G. A. Meehl, S. Bony, and D. R. Easterling. Climate data challenges in the 21st century. *Science*, 331(6018):700–702, Feb. 2011. doi:10.1126/science.1197869.

J. T. Randerson, F. M. Hoffman, P. E. Thornton, N. M. Mahowald, K. Lindsay, Y.-H. Lee, C. D. Nevison, S. C. Doney, G. Bonan, R. Stöckli, C. Covey, S. W. Running, and I. Y. Fung. Systematic assessment of terrestrial biogeochemistry in coupled climate-carbon models. *Global Change Biol.*, 15(10):2462–2484, Oct. 2009. ISSN 1365-2486. doi:10.1111/j.1365-2486.2009.01912.x.

O. J. Reichman, M. B. Jones, and M. P. Schildhauer. Challenges and opportunities of open data in ecology. *Science*, 331(6018):703–705, Feb. 2011. doi:10.1126/science.1197962.

D. N. Williams, R. Drach, R. Ananthakrishnan, I. T. Foster, D. Fraser, F. Siebenlist, D. E. Bernholdt, M. Chen, J. Schwidder, S. Bharathi, A. L. Chervenak, R. Schuler, M. Su, D. Brown, L. Cinquini, P. Fox, J. Garcia, D. E. Middleton, W. G. Strand, N. Wilhelmi, S. Hankin, R. Schweitzer, P. Jones, A. Shoshani, and A. Sim. The Earth System Grid: Enabling access to multimodel climate simulation data. *Bull. Am. Meteorol. Soc.*, 90(2): 195–205, Feb. 2009. doi:10.1175/2008BAMS2459.1.