Abstract

From field-scale measurements to global climate simulations and remote sensing, the growing body of very large and long time series Earth science data are increasingly difficult to analyze, visualize, and interpret. Data mining, information theoretic, and machine learning techniques—such as cluster analysis, singular value decomposition, block entropy, Fourier and wavelet analysis, phase-space reconstruction, and artificial neural networks—are being applied to problems of segmentation, feature extraction, change detection, model-data comparison, and model validation. The size and complexity of Earth science data exceed the limits of most analysis tools and the capacities of desktop computers. New scalable analysis and visualization tools, running on parallel cluster computers and supercomputers, are required to analyze data of this magnitude. This workshop will demonstrate how data mining techniques are applied in the Earth sciences and describe innovative computer science methods that support analysis and discovery in the Earth sciences.

Keywords:
Data mining, remote sensing, high performance computing, segmentation, change detection, synthesis, visualization

1. Introduction

The Workshop on Data Mining in Earth System Science (DMESS 2011) continues and expands upon a primary theme of the GeoComputation 2009 Workshop [1], which was held in conjunction with the International Conference on Computational Science (ICCS 2009) in Baton Rouge, Louisiana, USA. As the name states, DMESS 2011 is focused on applications of data mining techniques to studies in the Earth sciences. Spanning many orders of magnitude...
in time and space scales, Earth science data are increasingly large and complex and often represent very long time series, making such data difficult to analyze, visualize, interpret, and understand. Moreover, advanced electronic data storage technologies have enabled the creation of large repositories of observational data, while modern high performance computing capacity has enabled the creation of detailed empirical and process-based models that produce copious output across all these time and space scales. The resulting “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, comparison, and visualization. This workshop explores various data mining approaches to understanding Earth science data, emphasizing the technological challenges associated with utilizing very large and long time series geospatial data sets.

2. Earth Science Data

Observational and modeled data acquired or generated by the various disciplines within the realm of the Earth sciences encompass temporal scales of seconds (1 s) to millions of years (10^{13} s) and spatial scales of microns (10^{-6} m) to tens of thousands of kilometers (10^7 m). Because of rapid technological advances in sensor development, computational capacity, and data storage density, the volume, complexity, and resolution of Earth science data are increasing equally rapidly. Moreover, combining, integrating, and synthesizing data across Earth science disciplines offers new opportunities for scientific discovery that are only beginning to be realized. In fact, the rise of data-intensive scientific pursuits, in Earth sciences and other disciplines, has led some visionaries to proclaim it the fourth paradigm of discovery alongside the traditional experimental, theoretical, and computational archetypes [2]. Data-centric science, however, also poses unique technological and social challenges, many of which are exacerbated by the sheer size of the data sets involved.

The promise of scientific advances from data mining and synthesis has stimulated an increase in the number of users of Earth science data within the research community. Worldwide interest in sustainability and environmental policy, as well as mounting political pressure from climate change skeptics, has added decision-makers and the general public to the growing list of data users. Open and user-friendly access to Earth science data is required, particularly for climate change data, and decision-makers often need distilled data products for assessing impacts and planning and implementing climate adaptation and mitigation strategies. Organized global climate modeling activities, like the Coupled Model Intercomparison Project (CMIP) that coordinates simulations in support of the United Nations’ Intergovernmental Panel on Climate Change (IPCC) assessment reports, can generate tens of terabytes to several petabytes of simulation results in raw form [3], and are now made available to the research community and the general public through a series of distributed, interconnected servers called the Earth System Grid (ESG) [4]. For the IPCC Fourth Assessment Report (AR4), ESG distribution of CMIP Phase 3 model output resulted in hundreds of new papers analyzing various aspects and implications of the simulated climate change scenarios. Additionally, efforts are underway to develop composited, summary data from future simulation output collections that will be more directly useful for decision-makers and public users.

Observational data pose their own challenges. Satellite remote sensing data tend to be very large and their size has grown as spatial and temporal resolutions have increased. Meanwhile, small ecological data sets, often the most useful for synthesis, may be the most difficult to preserve, distribute, and use [5]. Such data must be well curated and their provenance must be formally documented. Data format standards and metadata conventions are needed for both observational and model data. The continually evolving Climate and Forecast (CF) Metadata Convention [6] is a framework that has successfully served the climate modeling community. Heterogeneous data resulting from multi-disciplinary fields in Earth sciences may require entirely new metadata languages [5]. The workflow associated with processing, quality control, gap-filling, analyzing, and synthesizing data should also be documented so that all those steps can be reproduced by other researchers. Scientific workflow systems are being developed to provide such capabilities. Pioneering efforts to automate and document the entire process—from data acquisition and generation to synthesis and publication—are being undertaken by the DataONE project (http://www.dataone.org/) in the context of establishing federated data systems [5].

Climate modeling activities like CMIP place new demands on the measurements community to provide observations and measurement uncertainties useful for assessing model fidelity and for validation during model development [7]. The international research community needs agreed-upon standards for model evaluation [8] and benchmarks for scientific performance of simulation models. Hence, the International Land Model Benchmarking project
(ILAMB: http://www.ilamb.org/) was recently established to develop benchmarks for terrestrial biogeochemistry models that run within Earth System Models (ESMs), which are presently being used to carry out CMIP Phase 5 experiments in support of the IPCC Fifth Assessment Report (AR5), expected to appear in 2013. By adopting the CF Metadata Convention and advocating for similar standards from the measurements community, model developers and climate scientists can more easily, quickly, and frequently perform detailed model evaluations and data-model intercomparisons. Another goal of ILAMB is to create a reusable, open source framework for evaluating cost functions and generating diagnostics for data-model intercomparison projects, eliminating the need for project organizers to reinvent the architecture each time and allowing such model assessments to be made on a regular basis. By utilizing only freely available observational data and openly distributing the code for its model evaluation tools, ILAMB seeks to improve the scientific process by achieving a new standard for scientific openness and transparency [9].

In addition to the data management issues of provenance, curation, metadata creation, and public distribution, today’s large and complex Earth science data often cannot be synthesized and analyzed using traditional methods or on single-processor desktop computers. Instead, new methods of analysis must be brought to bear on the problem and, in many cases, this requires development of new, highly scalable software tools that take advantage of large distributed-memory parallel computational resources. Data mining, machine learning, and high performance visualization approaches are increasingly filling this void and can often be deployed on parallel cluster computers or supercomputers. Examples of these approaches and their implementations are outlined here and described in the accompanying workshop papers.

3. Data Mining Approaches

A wide variety of data mining, machine learning, and information theoretic techniques are now being applied to the growing body of Earth science data. Cluster analysis has proven to be useful for segmentation, feature extraction, network analysis, change detection, model intercomparison, and model-data comparison in a number of Earth science applications [10]. Block entropy can be used as a classifier for dynamical systems. Spectral methods are frequently employed for decomposing periodic phenomena. Artificial neural networks and model tree ensembles have been used to refine models and to empirically up-scale and extrapolate point measurements.

Ecoregions—land areas that are relatively homogeneous with respect to a collection of observable environmental characteristics—have traditionally been developed by humans using expert judgment. However, stratification of geographic regions based on high resolution synoptic bioclimatic observational data using cluster analysis has now become an accepted method for delineating ecoregions [11]. The same method may be used to stratify climate observational or model data, not only across space, but also through time, resulting in time-evolving ecoregions or climate regimes [12] that can be used to better inform terrestrial biosphere models. Such clustered climate regimes can be tracked from present locations into alternative forecasted futures, allowing exploration of changes in size and location of home ranges for particular species [13], as well as new ecoregion compositions for particular locations [14]. In the paper included in this DMESS section, Sisneros et al., have integrated a flexible stratification method, similar to that used in cluster analysis, into a high performance visualization system to show how life zone boundaries are likely to change in the next century according to a projection from a global climate model. This suggests that scientific visualization, too often relegated to the final step of data analysis, can be one of the methods for exploration of very large data, especially in large-scale immersive visualization environments [15]. Querying for specific features in large geospatial data, like that generated from satellite remote sensing, requires new tools or languages that can exploit high performance computing resources. One example, a new high throughput data query language, was presented at the previous workshop [16].

Connecting the representativeness of locations where informative geophysical and ecological measurements are made (e.g., eddy flux towers) to other locations of corresponding ecoregion types where measurements are difficult, cost-prohibitive, or impossible to make is one area where multivariate geospatial cluster analysis has been used to improve analysis and quantify the representation of continental-scale sampling networks. A comprehensive analysis of network representativeness, indicating which ecoregions are well-represented by sampling, was performed for the AmeriFlux network of eddy covariance CO$_2$ flux tower sites in the conterminous United States [17, 18]. While such large-scale sampling networks are rarely designed and built based on such a data-intensive analysis, the same technique was recently used to establish the 20 sampling domains within the U.S. for the National Ecological Observatory.
analyses could be used to diagnose changes in periodic cycles (the model can handle. At the GeoComputation 2009 workshop, Brooks showed how wavelet and Fourier transform which can be important when assimilating measurements that need to be filtered to be representative on scales that automated and robust classifiers could be particularly useful for benchmarking the variability within model output, the system’s information content are viable classifiers for precipitation measured at Australian weather stations. Such covariance CO tree ensemble (MTE) method to perform global empirical up-scaling of observational data from the FLUXNET eddy assimilation images of the Earth’s subsurface. Other notable and recent applications have combined an ANN with a model of data wave seismograms reduces the time- and labor-intensive processing steps involved in creating high resolution images of the Earth’s subsurface. Other notable and recent applications have combined an ANN with a model tree ensemble (MTE) method to perform global empirical up-scaling of observational data from the FLUXNET eddy covariance CO2 measurement sites located throughout the world. This effort has produced global spatial estimates of gross primary production (GPP) [28], a global estimate of the temperature sensitivity of heterotrophic respiration (Q10) [29], and the global trend in land evapotranspiration [30].

4. High Performance Computing

To realize the promise of new scientific discovery from very large and long time series Earth science data, increasing capacity from high performance computing resources is required. Traditional analysis methods and algorithms are
insufficient for analyzing and synthesizing such large data sets, and those algorithms rarely scale out onto distributed-memory parallel computer systems. Therefore, new analysis techniques and scalable algorithms and software tools must be developed to enable analysis, exploration, and visualization of today’s Earth science data. Fortunately, the rapidly increasing computational power of state-of-the-art supercomputers provides opportunities for development of such tools. Often these same supercomputing resources are used to run simulation model experiments, so analysis and visualization may simply be another step in the scientific workflow process in the same computing environment.

Because of a scientific interest in developing empirical ecoregions based on observed data, Hoffman and Hargrove developed a parallel k-means clustering algorithm [31] that they implemented on an early Beowulf-style parallel cluster computer they constructed from surplus personal computers [32]. That code has been continually used and improved for environmental data analysis on machines ranging from laptops to the largest supercomputers in the world. Recent improvements to that code, including adoption of a triangle-inequality-based acceleration technique and “warping” of unassigned/empty cluster centroids, have significantly reduced the time to solution [10], and a new technique for initial centroid determination has improved the statistical performance of the clustering result. These enhancements have enabled the cluster analysis of large satellite data sets for identification of forest disturbances. In a paper included here, Kumar et al., present a fully distributed version of the k-means algorithm that includes several of the modifications developed by Hoffman et al. [10, 23]. This implementation avoids master-slave parallelism and is designed and tested for analysis of large data sets using state-of-the-art supercomputers. This implementation scales to tens of thousands of processors and has been tested on seven years of MODIS NDVI data at a resolution of 250 m² over the conterminous U.S.

5. Acknowledgments

The DMESS 2011 co-conveners—FMH, JWL, and RTM—wish to thank the Workshop Program Committee for their assistance in reviewing submitted papers. The Program Committee consisted of Michael W. Berry, Bjorn-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. The submitted manuscript has been authored by a contractor of the U.S. Government; accordingly, the U.S. Government retains a non-exclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes.

6. References


URL: http://www.geobabble.org/flux-ecoregions/


F. M. Hoffman, Analysis of reflected spectral signatures and detection of geophysical disturbance using hyperspectral imagery, Master’s thesis, Department of Physics and Astronomy, University of Tennessee, Knoxville, Tennessee, USA (Nov. 2004).


