Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 4 (2011) 1450–1455

Procedia Computer Science

International Conference on Computational Science, ICCS 2011

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[h], William W. Hargrove[i], Jian Huang[f], Jitendra Kumar[a], Ranga R. Vatsavai[h]

[a]*Computational Earth Sciences Group, Oak Ridge National Laboratory, Oak Ridge, TN, USA*
[b]*Department of Earth System Science, University of California, Irvine, CA, USA*
[c]*Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, USA*
[d]*Computation Institute, University of Chicago/Argonne National Laboratory, Chicago, IL, USA*
[e]*School of Computer Science, The Australian National University, Canberra, ACT, AUSTRALIA*
[f]*Department of Electrical Engineering & Computer Science, University of Tennessee, Knoxville, TN, USA*
[g]*Center for Climatic Research, University of Wisconsin, Madison, WI, USA*
[h]*Geographic Information Science and Technology Group, Oak Ridge National Laboratory, Oak Ridge, TN, USA*
[i]*Eastern Forest Environmental Threat Assessment Center (EFETAC), USDA Forest Service, Asheville, NC, USA*

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

Email addresses: forrest@climatemodeling.org (Forrest M. Hoffman), larson@mcs.anl.gov (J. Walter Larson), rmills@ornl.gov (Richard Tran Mills), bjorn@climatemodeling.org (Bjørn-Gustaf J. Brooks), gangulyar@ornl.gov (Auroop R. Ganguly), hnw@geobabble.org (William W. Hargrove), huangj@eecs.utk.edu (Jian Huang), jkumar@climatemodeling.org (Jitendra Kumar), vatsavairr@ornl.gov (Ranga R. Vatsavai)

doi:10.1016/j.procs.2011.04.157

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

Network (NEON), a 30-year nationwide study of climate and ecology [19, 20]. Additionally this geospatial data mining method was previously applied to remotely sensed hyperspectral imagery for detection of brine scar disturbances across a regional landscape [21]. When further applied across a time series of geospatial data, multivariate geospatiotemporal clustering was useful for characterizing the time-evolving dynamical behavior of Earth system processes. For example, combining monthly climatology with normalized difference vegetation index (NDVI) data from 17 years of 8 km^2 Advanced Very High Resolution Radiometer (AVHRR) images produced regions of similar phenological behavior, called phenoregions. Similar analysis using NDVI from the Moderate Resolution Imaging Spectroradiometer (MODIS) suggests that spatiotemporal clustering is useful for change detection in satellite imagery and may serve as a key component in a national-scale early warning system for detection of threats to forest health [22, 23]. This is highlighted by a paper included here in which Mills *et al.*, present an updated analysis from seven years of MODIS NDVI, demonstrating the utility of cluster analysis in detecting forest disturbances from mountain pine beetle, wildfire, hurricane landfall, and accompanying coastal salt-water intrusion.

With recent emphasis on biofuel development for reducing dependency on oil and reducing carbon emissions from energy production and consumption, the landscape of many countries is going to change dramatically in the coming years. In the United States, continuous corn production is becoming a dominant cropping pattern as the practice of soybean and wheat rotations is reduced to maximize total corn production. It is also expected that more pasture lands will be converted to switchgrass (*Panicum virgatum*) in the coming years, which may positively impact climate change because of its superior carbon balance properties with proper management practice. However, monitoring natural resources, especially crop biomass over large geographic regions using remote sensing poses several challenges and opportunities. Existing change detection techniques are not adequate or scalable for continuous monitoring. On the other hand, characterizing changes requires accurate classification of remote sensing images. Spatiotemporal data mining, especially the techniques that exploit the subtle multidimensional signals through the joint use of high temporal resolution (*e.g.*, MODIS) data and moderate- and fine-spatial resolution (*e.g.*, Advanced Wide Field Sensor or AWiFS) satellite images, has proven to be highly useful for extracting multi-temporal biomass change information [24], including crop types [25]. Spatial and spatiotemporal data mining may also be applied to the extraction of recurrence patterns of climate extremes from model results [26].

Other important Earth Science data mining techniques include those capable of detecting and classifying state changes in environmental variables, especially when it is necessary to compare time series across many locations, for example using information theoretic and spectral methods. Also in this DMESS section, Larson *et al.*, present a paper exploring the use of block entropy as a classifier for dynamical behavior in observed meteorological time series data. Their symbolic dynamics analysis system shows that the randomness (h_μ) and "memory" (E) components of the system's information content are viable classifiers for precipitation measured at Australian weather stations. Such automated and robust classifiers could be particularly useful for benchmarking the variability within model output, which can be important when assimilating measurements that need to be filtered to be representative on scales that the model can handle. At the GeoComputation 2009 workshop, Brooks showed how wavelet and Fourier transform analyses could be used to diagnose changes in periodic cycles (*e.g.*, annual precipitation intensity and frequency) between multiple model simulations as a way of understanding how changes in regular cycles may feed back onto other natural systems in global climate models [27].

Artificial neural networks (ANNs) are increasingly applied to model or classify Earth science data. In a paper included here, 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. Employing ANN and IANN for classification 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 CO_2 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 (Q_{10}) [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 m^2 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, 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. 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

[1] Y. Xue, F. M. Hoffman, D. Liu, GeoComputation 2009, in: G. Allen, J. Nabrzyski, E. Seidel, G. D. van Albada, J. Dongarra, P. M. Sloot (Eds.), Proceedings of the 9th International Conference on Computational Science (ICCS 2009), Vol. 5545 of Lecture Notes in Computer Science (LNCS), Springer-Verlag, Heidelberg, 2009, pp. 345–348. doi:10.1007/978-3-642-01973-9_38.

[2] T. Hey, S. Tansley, K. Tolle (Eds.), The Fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Corporation, Redmond, Washington, USA, 2009.

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

[4] 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, A. Sim, The Earth System Grid: Enabling access to multimodel climate simulation data, Bull. Am. Meteorol. Soc. 90 (2) (2009) 195–205. doi:10.1175/2008BAMS2459.1.

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

[6] B. Eaton, J. Gregory, B. Drach, K. Taylor, S. Hankin, J. Caron, R. Signell, P. Bentley, G. Rappa, H. Höck, A. Pamment, M. Juckes, NetCDF Climate and Forecast (CF) metadata conventions, version 1.5, Tech. rep. (Oct. 2010).

[7] 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, I. Y. Fung, Systematic assessment of terrestrial biogeochemistry in coupled climate-carbon models, Global Change Biol. 15 (10) (2009) 2462–2484. doi:10.1111/j.1365-2486.2009.01912.x.

[8] G. A. Alexandrov, D. Ames, G. Bellocchi, M. Bruen, N. Crout, M. Erechtchoukova, A. Hildebrandt, F. Hoffman, C. Jackisch, P. Khaiter, G. Mannina, T. Matsunaga, S. T. Purucker, M. Rivington, L. Samaniego, Technical assessment and evaluation of environmental models and software: Letter to the editor, Environ. Modell. Softw. 26 (3) (2011) 328–336, Thematic issue on the assessment and evaluation of environmental models and software. doi:10.1016/j.envsoft.2010.08.004.

[9] K. Kleiner, Data on demand, Nature Clim. Change 1 (1) (2011) 10–12. doi:10.1038/nclimate1057.
[10] F. M. Hoffman, W. W. Hargrove, R. T. Mills, S. Mahajan, D. J. Erickson, R. J. Oglesby, Multivariate Spatio-Temporal Clustering (MSTC) as a data mining tool for environmental applications, in: M. Sànchez-Marrè, J. Béjar, J. Comas, A. E. Rizzoli, G. Guariso (Eds.), Proceedings of the iEMSs Fourth Biennial Meeting: International Congress on Environmental Modelling and Software Society (iEMSs 2008), 2008, pp. 1774–1781.
[11] W. W. Hargrove, F. M. Hoffman, Potential of multivariate quantitative methods for delineation and visualization of ecoregions, Environ. Manage. 34 (Supplement 1) (2004) S39–S60. doi:10.1007/s00267-003-1084-0.
[12] F. M. Hoffman, W. W. Hargrove, D. J. Erickson, R. J. Oglesby, Using clustered climate regimes to analyze and compare predictions from fully coupled general circulation models, Earth Interact. 9 (10) (2005) 1–27. doi:10.1175/EI110.1.
[13] K. M. Potter, W. W. Hargrove, F. H. Koch, Predicting climate change extirpation risk for central and southern Appalachian forest tree species, in: J. S. Rentch, T. M. Schuler (Eds.), Proceedings from the Conference on Ecology and Management of High-Elevation Forests of the Central and Southern Appalachian Mountains, General Technical Report NRS-P-64, Newton Square, Pennsylvania, 2010, pp. 179–189.
[14] J. Westervelt, W. Hargrove, Forecasting climate-induced ecosystem changes on army installations, Technical Report ERDC/CERL, in preparation (2011).
[15] P. Fox, J. Hendler, Changing the equation on scientific data visualization, Science 331 (6018) (2011) 705–708. doi:10.1126/science.1197654.
[16] C. R. Johnson, M. Glatter, W. Kendall, J. Huang, F. M. Hoffman, Querying for feature extraction and visualization in climate modeling, in: G. Allen, J. Nabrzyski, E. Seidel, G. D. van Albada, J. Dongarra, P. M. Sloot (Eds.), Proceedings of the 9th International Conference on Computational Science (ICCS 2009), Vol. 5545 of Lecture Notes in Computer Science (LNCS), Springer-Verlag, Heidelberg, 2009, pp. 416–425. doi:10.1007/978-3-642-01973-9_46.
[17] W. W. Hargrove, F. M. Hoffman, B. E. Law, New analysis reveals representativeness of the AmeriFlux Network, Eos Trans. AGU 84 (48) (2003) 529, 535. doi:10.1029/2003EO480001.
[18] W. W. Hargrove, F. M. Hoffman, A flux atlas for representativeness and statistical extrapolation of the AmeriFlux network, Technical Memorandum ORNL/TM-2004/112, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA (Apr. 2004).
URL http://www.geobabble.org/flux-ecoregions/
[19] M. Keller, D. Schimel, W. Hargrove, F. Hoffman, A continental strategy for the National Ecological Observatory Network, Front. Ecol. Environ. 6 (5) (2008) 282–284, Special Issue on Continental-Scale Ecology. doi:10.1890/1540-9295(2008)6[282:ACSFTN]2.0.CO;2.
[20] D. Schimel, W. Hargrove, F. Hoffman, J. McMahon, NEON: A hierarchically designed national ecological network, Front. Ecol. Environ. 5 (2) (2007) 59. doi:10.1890/1540-9295(2007)5[59:NAHDNE]2.0.CO;2.
[21] 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).
[22] W. W. Hargrove, J. P. Spruce, G. E. Gasser, F. M. Hoffman, Toward a national early warning system for forest disturbances using remotely sensed phenology, Photogramm. Eng. Rem. Sens. 75 (10) (2009) 1150–1156.
[23] F. M. Hoffman, R. T. Mills, J. Kumar, S. S. Vulli, W. W. Hargrove, Geospatiotemporal data mining in an early warning system for forest threats in the United States, in: Proceedings of the 2010 IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2010), 2010, pp. 170–173, invited. doi:10.1109/IGARSS.2010.5653935.
[24] V. Chandola, R. R. Vatsavai, Scalable time series change detection for biomass monitoring using Gaussian process, in: NASA Conference on Intelligent Data Understanding (CIDU), 2010, pp. 69–82.
[25] G. Jun, R. R. Vatsavai, J. Ghosh, Spatially adaptive classification and active learning of multispectral data with Gaussian processes, in: ICDM Workshop on Spatial and Spatiotemporal Data Mining (SSTDM), 2009, pp. 597–603.
[26] A. R. Ganguly, K. Steinhaeuser, Data mining for climate change and impacts, in: Proceedings of the 2008 IEEE International Conference on Data Mining: Workshop on Climate Data Mining, 2008, pp. 385–394.
[27] B.-G. J. Brooks, Applying wavelet and fourier transform analysis to large geophysical datasets, in: G. Allen, J. Nabrzyski, E. Seidel, G. D. van Albada, J. Dongarra, P. M. Sloot (Eds.), Proceedings of the 9th International Conference on Computational Science (ICCS 2009), Vol. 5545 of Lecture Notes in Computer Science (LNCS), Springer-Verlag, Heidelberg, 2009, pp. 426–434. doi:10.1007/978-3-642-01973-9_47.
[28] C. Beer, M. Reichstein, E. Tomelleri, P. Ciais, M. Jung, N. Carvalhais, C. Rödenbeck, M. A. Arain, D. Baldocchi, G. B. Bonan, A. Bondeau, A. Cescatti, G. Lasslop, A. Lindroth, M. Lomas, S. Luyssaert, H. Margolis, K. W. Oleson, O. Roupsard, E. Veenendaal, N. Viovy, C. Williams, F. I. Woodward, D. Papale, Terrestrial gross carbon dioxide uptake: Global distribution and covariation with climate, Science 329 (5993) (2010) 834–838. doi:10.1126/science.1184984.
[29] M. D. Mahecha, M. Reichstein, N. Carvalhais, G. Lasslop, H. Lange, S. I. Seneviratne, R. Vargas, C. Ammann, M. A. Arain, A. Cescatti, I. A. Janssens, M. Migliavacca, L. Montagnani, A. D. Richardson, Global convergence in the temperature sensitivity of respiration at ecosystem level, Science 329 (5993) (2010) 838–840. doi:10.1126/science.1189587.
[30] M. Jung, M. Reichstein, P. Ciais, S. I. Seneviratne, J. Sheffield, M. L. Goulden, G. Bonan, A. Cescatti, J. Chen, R. de Jeu, A. J. Dolman, W. Eugster, D. Gerten, D. Gianelle, N. Gobron, J. Heinke, J. Kimball, B. E. Law, L. Montagnani, Q. Mu, B. Mueller, K. Oleson, D. Papale, A. D. Richardson, O. Roupsard, S. Running, E. Tomelleri, N. Viovy, U. Weber, C. Williams, E. Wood, S. Zaehle, K. Zhang, Recent decline in the global land evapotranspiration trend due to limited moisture supply, Nature 467 (7318) (2010) 951–954. doi:10.1038/nature09396.
[31] F. M. Hoffman, W. W. Hargrove, Multivariate geographic clustering using a Beowulf-style parallel computer, in: H. R. Arabnia (Ed.), Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA '99), Vol. III, CSREA Press, 1999, pp. 1292–1298.
[32] W. W. Hargrove, F. M. Hoffman, T. Sterling, The do-it-yourself supercomputer, Sci. Am. 265 (2) (2001) 72–79.
URL http://www.sciam.com/article.cfm?articleID=000E238B-33EC-1C6F-84A9809EC588EF21