Discovery of Changes from the Global Carbon Cycle and Climate System Using Data Mining

Vipin Kumar, Michael Steinbach, Pusheng Zhang, and Shashi Shekhar
Dept. Comp. Science and Eng.
University of Minnesota
Minneapolis, MN 55455
{kumar,steinbac,pusheng,shekhar}@cs.umn.edu

Pang-Ning Tan
Dept. Comp. Science and Eng.
Michigan State University
East Lansing, MI 48824
ptan@cse.msu.edu

Christopher Potter and Steven Klooster
NASA Ames Research Center
Moffett Field, CA 94035
cpotter@mail.arc.nasa.gov
sklooster@gaia.arc.nasa.gov

***Abstract*— The goal of our NASA sponsored project, "Discovery of Changes from the Global Carbon Cycle and Climate System Using Data Mining," is to better understand global scale patterns in biosphere processes, especially relationships between the global carbon cycle and the climate system. To that end, we have developed data mining techniques to efficiently find spatio-temporal patterns in large Earth Science data sets. One such technique finds ecosystem disturbances in long-term, non-stationary fractional photosynthetically active radiation (FPAR) time series data. A key contribution of this work, which was the subject of a recent NASA press release, is the development of an automated technique for detecting abrupt changes in FPAR that take into account the timing, location, and magnitude of such changes. Using this technique, scientists were able to estimate that nearly 9 Pg (peta-grams) of carbon have been lost from the terrestrial biosphere to the atmosphere as a result of large-scale ecosystem disturbance events over an 18-year time period. Also, a novel clustering technique was developed to identify regions of uniform behavior in spatio-temporal data. The clusters produced by these methods are useful in discovering climate indices because they identify significant regions of the ocean or atmosphere where the behavior is relatively uniform over the entire area. Some of the discovered clusters correspond to known climate indices while other clusters are variants of known indices that appear to provide better predictive power for some land areas and still other clusters may represent potentially new Earth science phenomena. Other contributions include visualization techniques for finding interesting associations using land cover information and a filter-and-refine approach for efficiently processing correlation-based queries. Such innovative data analysis tools and techniques can aid NASA scientists in analyzing the growing datasets generated by NASA's global observing satellites and offer an unprecedented opportunity for predicting and preventing future ecological problems by managing the ecology and health of our planet.**

I. Introduction

In this paper we present a brief overview of some results from our currently funded NASA IS-IDU project, *Discovery of Changes from the Global Carbon Cycle and Climate System Using Data Mining*, which is part of the Intelligent Systems (NRA2-37143) program. During this project we developed new data analysis and knowledge discovery techniques to investigate changes in the global carbon cycle and climate system. The detailed results of this research are available in the following papers: [6], [8], [9], [10], [11], [12], [13], [14], [16], [17], [18], [19], [22], [26], [27], [28]. After a quick description of the data that we used in our investigations, we describe some of our work.

II. Data

The types of data shown in Figure 1 are representative of the data considered in this project, i.e., the basic data elements are individual co-registered cells in grids which cover the entire surface of the earth with resolutions between 0.25 km and 50 km. (Land variables derived from EOS satellite data are available at resolutions as high as 0.25 km, while surface climatology data, e.g., temperature and precipitation, are only available for grids of resolution 50 km or greater.) At a particular moment in time, each grid point can be described by the values of different variables, e.g., Net Primary Production (NPP), temperature, precipitation, etc. The variable values for each grid point are available for periodic, discrete points in time with resolutions from 8 days to 1 month. These variable values can either be the result of observations (from satellites or other sources) or the result of model predictions.

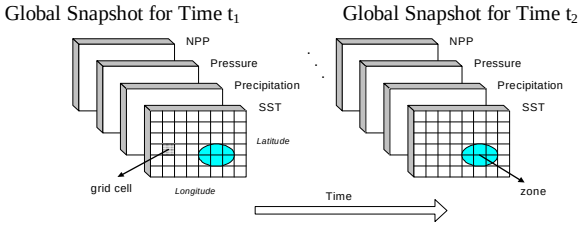


Fig. 1. Representation of Earth science data.

III. Detecting Ecosystem Disturbances

An ecosystem disturbance is an event that results in a sustained disruption of ecosystem structure and function, generally with effects that last for time periods longer than a single seasonal growing cycle for native vegetation. Physical disturbance categories include fires, hurricanes, floods, droughts, lava flows, and ice storms. Biogenic disturbance categories include the impacts of herbivorous insects, mammals, and pathogens. Anthropogenic disturbance categories include logging, deforestation, drainage of wetlands, clearing for cultivation, chemical pollution, and the introduction of

alien species to an ecosystem. Many of these events alter ecosystem productivity and resource availability (light and nutrient availability) for organisms on large spatial and temporal scales.

Ecosystem disturbances can also contribute to the current rise of carbon dioxide (CO2) levels in the atmosphere [15]. Because major 'pulses' of CO2 from terrestrial biomass loss can be emitted to the atmosphere during large disturbance events, the timing, location, and magnitude of vegetation disturbance is presently a major uncertainty in understanding global carbon cycles [3]. Elevated biogenic sources of CO2 have global implications for climatic change, which can in turn affect a vast number of species on Earth and the functioning of virtually all ecosystems.

We are developing a proven methodology to monitor and understand most ecosystem disturbance events and their historical regimes at a global scale. As a step in this direction, we have conducted studies to evaluate patterns in an 18-year record of global satellite observations of vegetation phenology from the Advanced Very High Resolution Radiometer (AVHRR) as a means to characterize major ecosystem disturbance events and regimes [13]. The FPAR absorbed by vegetation canopies worldwide has been computed at a monthly time interval from 1982 to 1999 and gridded at a spatial resolution of 0.5° latitude/longitude. Potential disturbance events of large extent (> 0.5 Mha) were identified in the FPAR time series by locating anomalously low values (FPAR-LO) that lasted longer than 12 consecutive months at any 0.5° pixel. An example of such a time series is shown in Figure 2.

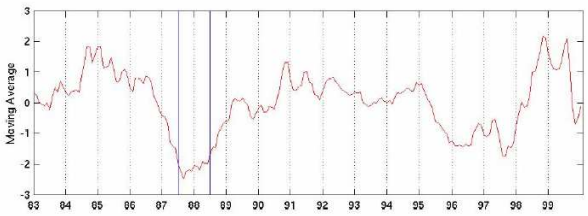


Fig. 2. A dip in FPAR in Mongoloia during 1987 was verified as an FPAR disturbance event due to wildfires.

Our study showed that nearly 400 Mha of the global land surface could be identified with at least one FPAR-LO event over the 18-year time series (Figure 3). The majority of these potential disturbance events occurred in tropical savanna and shrublands or in boreal forest ecosystem classes. Verification of potential disturbance events from our FPAR-LO analysis was carried out using documented records of the timing of large-scale wildfires at locations throughout the world. Disturbance regimes were further characterized by association analysis (Section 1.2.5) with historical climate anomalies. Assuming accuracy of the FPAR satellite record to characterize major ecosystem disturbance events, we estimate that nearly 9 Pg of carbon could have been lost from the terrestrial biosphere to the atmosphere as a result of large-scale ecosystem disturbance over this 18-year time series. (This finding was the subject of a NASA press release 03-51AR.) The use of high resolution MODIS products will improve these estimates.

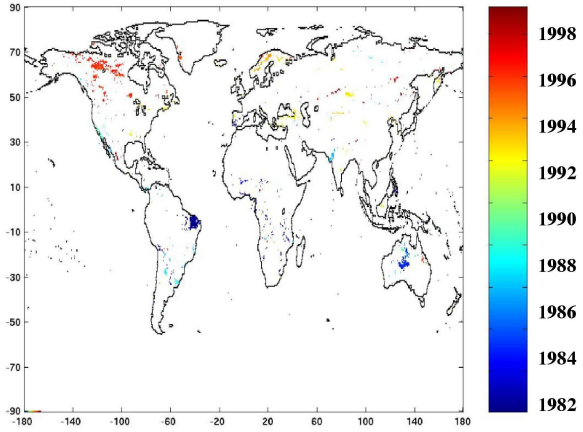


Fig. 3. First month and year for FPAR-LO lasting > 12 consecutive months over the period from 1982 to 1999.

IV. Exploring the Relationships Between Climate Indices and Net Ecosystem Production

Global teleconnnections [24], such as the El Nino Southern Oscillation (ENSO) [25] can be used to understand simultaneous variation in climate and related processes over widely separated points on the Earth. However, large-scale teleconnections between ocean and atmospheric processes and global NPP have yet to be demonstrated, and may escape ready detection without the aid of automated spatial-temporal analysis tools.

We applied correlation analysis to 17 years of ocean climate observations and model-estimated NEP on land to infer short-term (monthly to yearly) teleconnections between sea surface temperature and terrestrial carbon cycles [11]. The analysis suggests that, on a global level, combined climate indices can be used to predict net ecosystem carbon fluxes over more than 58 percent of the non-desert/ice covered land surface with a lead period of 2-6 months. A climate index is a time series that summarizes the behavior of the Earth's oceans or atmosphere and captures the relationships between the land and the oceans or the atmosphere. Teleconnections detected between ocean surface climate and seasonal carbon gain in terrestrial vegetation offer important capabilities for making inferences about the variability in the terrestrial carbon cycle of natural and agricultural ecosystems worldwide.

We have also investigated global teleconnections of climate to regional satellite-driven observations for predicted Amazon ecosystem production, mainly in the form of monthly estimates of net carbon exchange over the period 1982-1998 from the NASA-CASA model [10]. Results of our analysis suggest that anomalies of NPP and NEP predicted from the NASA-CASA model over large areas of the Amazon region east of 60° W longitude are strongly correlated with the Southern Oscillation Index (SOI). Extensive areas of the south-central Amazon also

show strong linkages of the FPAR and the NASA-CASA NPP anomaly record to the Arctic Oscillation (AO) index, which confirm a strong relation to southern Atlantic climate anomalies, with concurrent impacts on Amazon rainfall patterns. We further investigated processes for these teleconnections of global climate to Amazon ecosystem carbon fluxes and regional land surface climate.

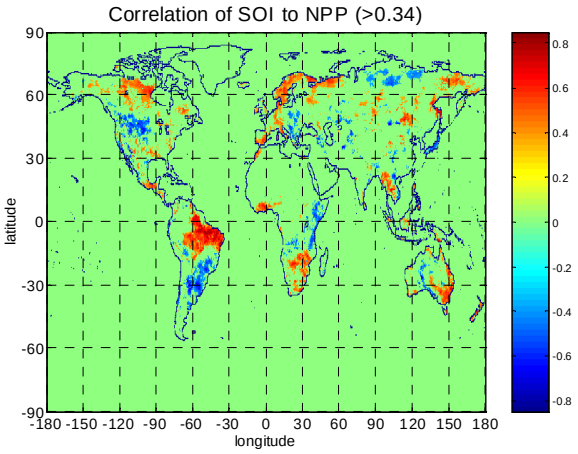


Fig. 4. Correlation of the Southern Oscillation Index (SOI) to NPP on land.

An additional type of correlation analysis that we have investigated involves the long-term (20 year) river discharge records from 30 of the world's largest river basins [14]. These records were an attempt to characterize surface hydrologic flows in relation to net precipitation inputs, ocean climate teleconnections, and human land/water use patterns. Comparisons of paired station records at upstream and downstream discharge locations within each major river basin suggest that the relatively 'natural' discharge signals represented in upstream discharge records are sustained in the downstream station records for nearly two-thirds of the undammed drainage basins selected. River basins that show the strongest localized climate control over historical discharge records, in terms of correlations with net basin-wide precipitation rates, are located mainly in the seasonally warm temperate and tropical latitude zones, as opposed to river basins located mainly in the higher latitude zones (above 45° N). Ocean climate indices, such as NINO1+2 and NINO3+4, correlate highly with historical interannual patterns in monthly river discharge for only four of the selected discharge station records, namely on the Amazon, Congo (Zaire), Columbia and Colorado (Arizona) rivers. Historical patterns of cropland development and irrigated areas may explain the weak climate correlations with interannual patterns in monthly river discharge rates for at least one-third of the major river drainages selected from the historical discharge data set.

V. Discovery of Climate Indices

Our interest in climate indices [7] arises from a desire to improve our understanding of teleconnections involving ocean temperature/pressure and terrestrial carbon flux. In the past,

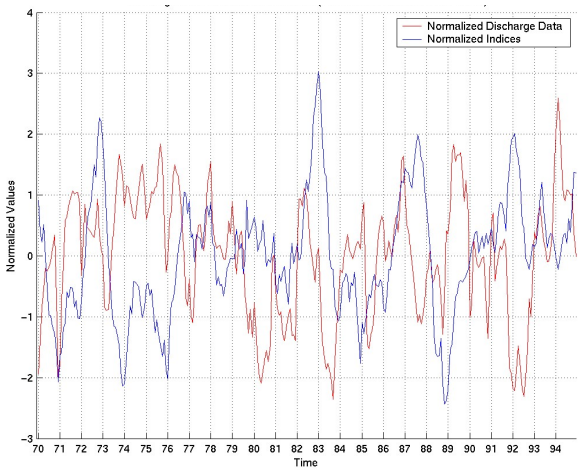


Fig. 5. Correlation between monthly river discharge for Amazon and NINO3+4 index.

Earth scientists have used observation and, more recently, eigenvalue analysis techniques, such as principal components analysis (PCA) and singular value decomposition (SVD), to discover climate indices [20]. These techniques are only useful for finding a few of the strongest signals and impose a condition that all discovered signals must be orthogonal to each other. We have developed an alternative methodology [17], [18], [16] for the discovery of climate indices that overcomes these limitations and is based on clusters that represent geographic regions with relatively homogeneous behavior. The centroids of these clusters are time series that summarize the behavior of these geographical areas.

Figure 6 shows the clusters produced by shared nearest neighbor (SNN) clustering [5] of sea level pressure data for the period 1958-1998 [17], [18], [16]. Many pairs of clusters in this clustering are highly correlated with the known climate indices. For example, clusters 13 and 20 are highly correlated with the Southern Oscillation Index (SOI), clusters 10 and 18 are correlated with the Arctic Oscillation index (AO), and clusters 7 and 10 are correlated with the North Atlantic Oscillation index (NAO).

We have also investigated clusters of SST. Four of these clusters are very highly correlated (correlation $>$ 0.9) with well-known climate indices, e.g., NINO 1+2, NINO 3, NINO 3.4, and NINO 4, and were located in approximately the same location as where these indices are defined [17], [18], [16]. The SST clusters that are less well correlated with known indices may represent new Earth science phenomena or weaker versions or variations of known phenomena. Indeed, some of these cluster centroids provide better 'coverage,' i.e., higher correlation to land temperature, for some areas of the land. This is illustrated in Figure 7, which compares the El Nino indices to that of clusters 62 (close to Brazil). Areas of yellow indicate where cluster 62 has higher correlation, while areas of blue indicate where the El Nino indices have higher correlation. Observe that cluster 62 'outperforms' the known indices for some areas of the land. The overall coverage of

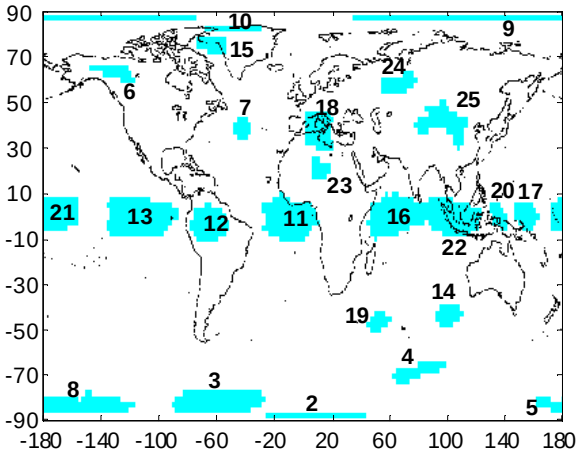


Fig. 6. 25 ocean clusters produced by SNN clustering of sea level pressure data for 1958-1998.

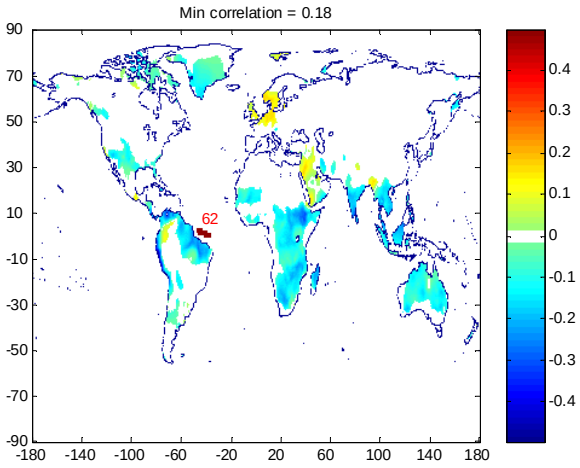


Fig. 7. Comparison of correlation of Cluster 62 vs. El Nino Indices to land Temp.

cluster 62 (measured in area weighted correlation) is similar to that of an El Nino based index, such as NINO 1+2, NINO 3, etc.)

VI. Discovering Associations among Anomalous NPP and Climate Events

Discovery of spatio-temporal relationships between global climate changes and land surface processes is crucial for understanding how the different elements of the ecosystem interact with each other. Such relationships are often captured using correlation analysis between time series or using a knowledge discovery technique known as association analysis for event sequences [2]. The goal of association analysis is to extract patterns in the form of rules or sets of events that will predict the occurrence of an event based on how frequent it co-occurs with other events.

We have previously applied association analysis to discover interesting relationships between anomalous NPP and climate events. Anomalous HI and LO events are first identified as values in the time series that deviate significantly above or below their monthly mean. For example, PREC-LO may suggest a well-below average precipitation or drought-related event. If events such as NPP-LO and PREC-LO are independent, then it can be shown that the probability they co-occur in the same month at the same location will be small. If these events co-occur more frequently than expected, then it is an indication of a non-random association [22], [21]. The patterns extracted using association analysis techniques are verified using the statistical chi-squared test (Lindgren, 1998).

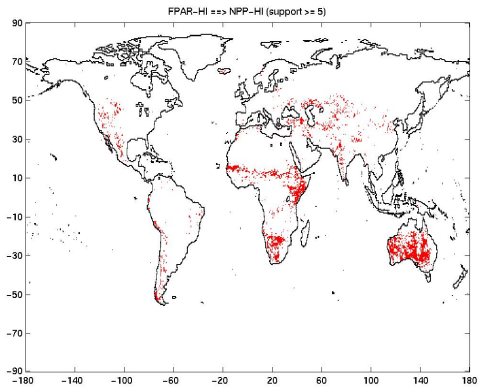
The patterns discovered using association analyses are also evaluated on the basis of their frequency of occurrence within major global vegetation type. For example, Figure 9a shows the regions where the events FPAR-HI and NPP-HI are observed together frequently. This pattern suggests that anomalously high values of FPAR, which means that the vegetation has generated more ”light-harvesting” photosynthetic capability than average, is often associated with abnormally high NPP values. Though this result is not surprising, further analyses reveal that such patterns are observed more frequently in regions that correspond to semi-arid annual grasslands, as shown in Figure 9b [22]. One possible explanation for such observation is that grasslands are vegetation that is able to more quickly take advantage of periodically high precipitation (and possibly solar radiation) than forests.

VII. Processing Correlation Queries

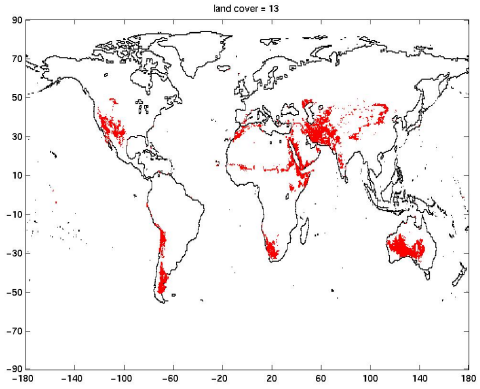
Massive amounts of data offer an unprecedented opportunity for researchers to discover potential nuggets of valuable information. The typical data in this project are *spatial time series data*, where each time series references a location on the Earth. Finding location pairs with highly correlated time series in spatial time series data is important for many application domains such as Earth science, epidemiology, ecology, climatology, and census statistics. For example, such queries were used to identify the land locations on the Earth where the climate was often affected by El Nino [26].

However, processing correlation queries in spatial time series data is computationally expensive because of the massive numbers of locations and time snapshots. Previous work [1], [4] did not incorporate spatial and temporal properties (e.g., neighborhoods) and thus these methods of correlation query processing suffer from inefficiency, i.e., the processing of correlation queries is substantially time consuming. Spatial and temporal properties are important information in spatial time series data, and should be considered in the design of efficient query processing methods to facilitate processing correlation queries. The design of such efficient query processing methods is crucial to organizations which make decisions based on large spatial time series data.

We have proposed filter-and-refine query processing algorithms [26], [27] to exploit spatial properties for facilitating correlation-based similarity queries on spatial time series data. Spatial time series data comply with Tobler’s first law of geography [23], which says that everything is related to everything else but nearby things are more related than distant things. Therefore, the attribute values of objects located in spatial



(a) Locations that support the association pattern {FPAR-HI, NPP-HI}



(b) Land locations that correspond to grassland and shrubland regions.

Fig. 8. Example of a non-random association pattern between FPAR-Hi and NPP-Hi events and the land locations where such pattern is observed frequently.

vicinity tend to be similar. The proposed algorithms first divide the data into a collection of disjoint groups based on spatial proximity. Each group might contain multiple time series of nearby locations together. The queries are then processed at the group level, instead of at the individual time series level to achieve the performance gains. Algebraic analyses using cost models and experimental evaluations using the real data were carried out to show that the proposed algorithms saved a large portion of computational cost, ranging from 40% to 98%. Further details are provided in [26], [27], [28].

VIII. Conclusion

In this paper, we provided an overview of our preliminary efforts to apply data mining techniques to the analysis of Earth Science data. We believe that our initial results are promising. For instance, we have been able to use our techniques to find patterns that represent well-known patterns, e.g., well-known climate indices. More importantly, we have also found new patterns that are not known to Earth Scientists, e.g., candidate climate indices and association patterns that relate land covers to rules that connect climate variables and NPP. While more evaluation is necessary to assess the Earth Science significance of these results, one of the major goals of our work is to produce new patterns and hypotheses for Earth Scientists to investigate, and we feel that we have made progress towards that goal.

Nonetheless, there are many tasks remaining both with respect to data mining and the application of data mining results. The main focus of the data mining work has been clustering and association analysis, and we have only lightly explored outlier detection, co-location mining, predictive modeling, and other data mining approaches. Also, while the Earth Science members of our team have evaluated and interpreted the data mining results that we have produced so far, which has led to publications in Earth Science journals [9], [10], [11], [12], [13], [14], there is much more to do within the scope of our current project. In the long term, we are hopeful that data mining can play an important role in helping Earth Scientists understand both global scale changes in biosphere processes and patterns, and the effects of widespread human activities. More broadly, improvements in data mining techniques made during the investigation of Earth Science data may have potential applications in other domains, such as transportation, business logistics, public health, and public safety.

Acknowledgment

This work was partially supported by NASA grant # NCC 2 1231 and by the Army High Performance Computing Research Center under the auspices of the Department of the Army, Army Research Laboratory cooperative agreement number DAAD19-01-2-0014. The content of this work does not necessarily reflect the position or policy of the government and no official endorsement should be inferred. Access to computing facilities was provided by the AHPCRC and the Minnesota Supercomputing Institute.

References

[1] R. Agrawal, C. Faloutsos, and A. Swami. Efficient Similarity Search In Sequence Databases. In *Proc. of the 4th Int'l Conference of Foundations of Data Organization and Algorithms*, 1993.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. of the 20th VLDB Conference*, 1994.

[3] J. G. Canadell, H. A. Mooney, D. D. Baldocchi, J. A. Berry, J. R. Ehleringer, C. B. Field, S. T. Gower, D. Y. Hollinger, J. E. Hunt, R. B. Jackson, S. W. Running, G. R. Shaver, W. Steffen, S. E. Trumbore, R. Valentini, and B. Y. Bond. Carbon metabolism of the terrestrial biosphere: a multi-technique approach for improving understanding. *Ecosystems*, 3:115–130, 2000.

[4] K. Chan and A. W. Fu. Efficient Time Series Matching by Wavelets. In *Proc. of the 15th Int'l Conference on Data Engineering*, 1999.

[5] L. Ertöz, M. Steinbach, and V. Kumar. Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data. In *Proc. of 3rd SIAM International Conference on Data Mining, San Francisco, CA, USA*, May 2003.

[6] J. Han, R. B. Altman, V. Kumar, H. Mannila, and D. Pregibon. Emerging scientific applications in data mining. *Communications of the ACM*, 45(8), august 2002.

[7] T. Karl, R. Knight, D. Easterling, , and R. Quayle. Indices of climate change for the united states. *Bulletin of the American Meteorological Society*, 77(2):279–292, 1996.

[8] V. Kumar, M. Steinbach, P. N. Tan, C. Potter, S. Klooster, and A. Torregrosa. Mining Scientific Data: Discovery of Patterns in the Global Climate System. In *Joint Statistical Meeting*, 2001.

[9] C. Potter, S. Klooster, R. Myneni, V. Genovese, P. Tan, and V. Kumar. Continental Scale Comparisons of Terrestrial Carbon Sinks. *Global and Planetary Change*, 39:201–213, 2003.

[10] C. Potter, S. Klooster, M. Steinbach, P. Tan, V. Kumar, S. Shekhar, and C. Carvalho. Understanding global teleconnections of climate to regional model estimates of amazon ecosystem carbon flux. *Global Change Biology, accepted for publication*, 2003.

[11] C. Potter, S. Klooster, M. Steinbach, P. Tan, V. Kumar, S. Shekhar, R. Nemani, and R. Myneni. Global Teleconnections of Ocean Climate to Terrestrial Carbon Flux. *Journal of Geophysical Research*, 108, 2003.

[12] C. Potter, S. Klooster, P. Tan, M. Steinbach, V. Kumar, and V. Genovese. Variability in Terrestrial Carbon Sinks over Two Decades. Part I: North America. *Earth Interactions*, 7, 2003.

[13] C. Potter, P. Tan, M. Steinbach, S. Klooster, V. Kumar, R. Myneni, and V. Genovese. Major Disturbance Events in Terrestrial Ecosystems Detected using Global Satellite Data Sets. *Global Change Biology*, 9(7):1005–1021, July 2003.

[14] C. Potter, P. Zhang, S. Klooster, V. Genovese, S. Shekhar, and V. Kumar. Land Use - Understanding Controls on Historical River Discharge in the World's Largest Drainage Basins. *Earth Interactions, accepted for publication*, 2003.

[15] C. S. Potter. Terrestrial biomass and the effects of deforestation on the global carbon cycle. *BioScience*, 49:769–778, 1999.

[16] M. Steinbach, P. Tan, V. Kumar, S. Klooster, and C. Potter. Discovery of Climate Indices Using Clustering. In *Proc. of the 9th ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining*, pages 446–455, August 2003.

[17] M. Steinbach, P. Tan, V. Kumar, C. Potter, and S. Klooster. Data Mining for the Discovery of Ocean Climate Indices. In *Proc of the Fifth Workshop on Scientific Data Mining*, 2002.

[18] M. Steinbach, P. Tan, V. Kumar, C. Potter, and S. Klooster. Temporal Data Mining for the Discovery and Analysis of Ocean Climate Indices. In *Proc of the KDD Workshop on Temporal Data Mining*, 2002.

[19] M. Steinbach, P. Tan, V. Kumar, C. Potter, S. Klooster, and A. Torregrosa. Clustering Earth Science Data: Goals, Issues and Results. In *Proc. of the Fourth KDD Workshop on Mining Scientific Datasets*, 2001.

[20] H. V. Storch and F. W. Zwiers. *Statistical Analysis in Climate Research*. Cambridge University Press, July 1999.

[21] P. Tan, J. Srivastava, V. Kumar, C. Potter, and S. Klooster. Selecting the right interestingness measure for association patterns. In *KDD02*, 2002.

[22] P. Tan, M. Steinbach, V. Kumar, C. Potter, S. Klooster, and A. Torregrosa. Finding Spatio-Temporal Patterns in Earth Science Data. In *Proc. of KDD Workshop on Temporal Data Mining*, 2001.

[23] W. R. Tobler. *Cellular Geography, Philosophy in Geography*. Gale and Olsson, Eds., Dordrecht, Reidel, 1979.

[24] Trenberth and Hurrell. Decadal atmosphere-ocean variations in the pacific. *Climate Dynamics*, 9:303–319, 1994.

[25] K. E. Trenberth, J. M. Caron, Stepaniak, D. P., and S. Worley. Evolution of el nino southern oscillation and global atmospheric surface temperatures. *J. Geophys. Research*, 107, 2002.

[26] P. Zhang, Y. Huang, S. Shekhar, and V. Kumar. Correlation Analysis of Spatial Time Series Datasets: A Filter-and-Refine Approach. In *the Proc. of the 7th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2003.

[27] P. Zhang, Y. Huang, S. Shekhar, and V. Kumar. Exploiting Spatial Autocorrelation to Efficiently Process Correlation-Based Similarity Queries. In *the Proc. of the 8th Int'l. Symp. on Spatial and Temporal Databases*, 2003.

[28] P. Zhang, S. Shekhar, Y. Huang, and V. Kumar. Spatial Cone Tree: An Index Structure for Correlation-based Similarity Queries on Spatial Time Series Data. In *the Proc. of the Int'l Workshop on Next Generation Geospatial Information*, 2003.