Using D2K Data Mining Platform for Understanding the Dynamic Evolution of Land-Surface Variables

Praveen Kumar[1], Peter Bajcsy[2], David Tcheng[2], David Clutter[2], Vikas Mehra[1], Wei-Wen Feng[2], Pratyush Sinha[1], and Amanda B. White[1]
[1]Department of Civil and Environmental Engineering, [2]National Center of Supercomputing Applications, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801 [e-mail:kumar1@uiuc.edu]

***Abstract*- The objective of our project is to develop data mining and knowledge discovery in databases (KDD) techniques, using the "Data to Knowledge" (D2K) platform developed by National Center for Supercomputing Application (NCSA), to facilitate analysis, visualization and modeling of land-surface variables obtained from the TERRA and AQUA platforms in support of climate and weather applications. The project is developing capacity to access very large multivariate datasets; represent heterogeneous data types; integrate multiple GIS data sets stored in many GIS file formats; analyze variable relationships and model their dependencies using cluster and grid computing; and visualize input data, as well, as extracted features, integrated data sets and data mining results.**

I. INTRODUCTION

The objective of our research project is to develop data mining and knowledge discovery in databases (KDD) techniques, using the "Data to Knowledge" (D2K) platform developed by National Center for Supercomputing Application (NCSA), to facilitate analysis, visualization and modeling of land-surface variables obtained from the TERRA and AQUA platforms in support of climate and weather applications.

The project targets to address the science question: "How is the global Earth system changing?" In particular it focuses on the theme: What factors influence/modulate the changes in global ecosystem? The specific science questions that this project is focused on are:

1) How are evolving surface variables such as vegetation indices, temperature, and emissivity, as obtained from the TERRA and AQUA platforms, dynamically linked?
2) How do they evolve in response to climate variability such as ENSO (El Niño Southern Oscillation)? and
3) How are they dependent on temporally invariant factors such as topography (and derived variables such as slope, aspect, nearness to streams), soil characteristics, land cover classification, etc?

Answers to these questions, at the continental to global scales will enable us to develop better parameterization of the relevant processes in forecast models for weather, and inter-seasonal to inter-annual climate prediction. However, answering these questions at the continental to global scale requires the ability to perform analysis of a multitude of variables using very large datasets. The proposed data mining system is targeted to build this capability for Earth science datasets being collected by NASA.

The salient features of the project are:

Science: The project will explore scientific questions that have not been addressed before and utilize new spatio-temporal data mining algorithms on remote-sensing and related data. The scientific novelty of our work consists of (1) data driven exploration of relationships between ecosystem and topography related variables, and (2) enabling improved parameterization for climate and weather models. The computer science novelty of the project includes (1) large data size processing using cluster and grid computing, and (2) the application of data mining techniques to land-surface variables and scientific modeling.

Technology: The project will develop capacity to access very large multivariate datasets; represent heterogeneous data types; integrate multiple GIS data sets stored in many GIS file formats; analyze variable relationships and model their dependencies using cluster and grid computing; and visualize input data, as well, as extracted features, integrated data sets and data mining results.

The strengths of our project lie in (1) NCSA supercomputing power and storage available for processing and large data storage, (2) the well-established D2K environment as a platform for software development, (3) the GIS library of software tools called I2K (Image to Knowledge) that enables us to perform GIS data ingestion, data integration and data visualization, (4) HDF file format development at NCSA and supported in I2K, and (4) implementations of several data mining techniques in D2K.

II. BACKGROUND

Very large volumes of data have been collected for hydroclimatological studies through satellites such as TERRA and AQUA and numerous other sources. However, the volume of available data has far out stretched our ability to effectively use them for hypothesis testing, modeling and prediction. This is primarily because we continue to use traditional tools of scientific inquiry, such as statistical analysis or data assimilation, over these large datasets. There are several limitations of these methods. These techniques do not work very well for incomplete, inconclusive, or heterogeneous datasets. The use of these datasets are limited to small fragments of the entire volume because the data reside on multiple nodes of various organizations and are available in different formats that are not easy to integrate. This limits our

ability for formulating and testing hypothesis. In addition, our scientific vision is stymied due to the use of fragmented and limited datasets, and our ability to handle only "few variables" at a time. This limits the nature of hypothesis that are proposed and tested. The value of data is typically predicated on the ability to extract higher level information: information useful for decision support, for exploration, and for better understanding of the phenomena generating the data. Our traditional physics based and data driven approaches of scientific inquiry breakdown as the volume and dimensionality of data increases, thereby reducing the value of observed data.

The premise of the project is: techniques for exploring large datasets are now becoming available but have not been extensively applied for the exploration of scientific data, and in particular for hydroclimatological studies; scientific inquiry methods developed for small datasets or “few variable" problems may not be effective for large datasets or “many variable" problems; and there are pressing scientific questions that need answers and can be answered by effectively exploring the available observational data. During the last several years, data mining or automatic knowledge discovery in databases (KDD) tools capable of identifying implicit knowledge in databases have become available and these tools address some of the limitations identified above. Their use in commercial settings has lead to very successful applications. However, their specialized use for various scientific problems is limited, but initial work is underway [1, 2, 3, 4]. Data mining application to scientific data will enable us to develop hypothesis about relationships of variables from observed data. These new hypothesis combined with the existing understanding of the physical processes we already have can result in an improved understanding and novel formulations of physical laws and an improved predictive capability, for both climate and weather.

The research project will use the D2K (http://alg.ncsa.uiuc.edu/do/tools/d2k) application environment for data mining. D2K is a rapid, flexible data mining and machine learning system that integrates analytical data mining methods for prediction, discovery, and deviation detection, with data and information visualization tools. It offers a visual programming environment that allows users to connect programming modules together to build data mining applications and supplies a core set of modules, application templates, and a standard API for software component development. All D2K components are written in Java for maximum flexibility and portability. Major features that D2K provides to an application developer include:

1) Visual Programming System Employing a Scalable Framework
2) Robust Computational Infrastructure
   a. Enables processor intensive applications
   b. Supports distributed computing
   c. Enables data intensive applications
   d. Provides low overhead for module execution
3) Flexible and Extensible Architecture
   a. Provides plug and play subsystem architectures and standard APIs
   b. Promotes code reuse and sharing
   c. Expedites custom software developments
   d. Relieves distributed computing burden
4) Rapid Application Development (RAD) Environment
5) Integrated Environment for Models and Visualizatio
6) *D2K Module Development*: NCSA’s Automated Learning Group (ALG) has developed hundreds of modules that address every part of the KDD process. Some data mining algorithms implemented include Naive Bayesian, Decision Trees, and apriori, as well as visualizations for the results of each of these approaches. In addition, ALG has developed modules for cleaning and transforming data sets and a number of visualization modules for deviation detection problems. Modules have also been created for specific projects and collaborations.

   ALG NCSA is continuing development of modules with the short-term goal of enhancing the cleaning and transformation modules, improving the data mining algorithms and continuing development of feature subset selection modules. Long-term, ALG plans to continue development of modules for predictive modeling, image analysis and textual analysis, particularly toward enabling them for distributed and parallel computing. This type of work expedites the process of applying the latest research developments to be used on real-world applications.
7) *D2K-driven Applications*: D2K can be used as a stand-alone application for developing data mining applications or developers can take advantage of the D2K infrastructure and D2K modules to build D2K-driven applications such as the ALG application I2K-Image to Knowledge. These applications employ D2K functionality in the background, using modules dynamically to construct applications. They present their own specialized user interfaces specific to the tasks being performed. Advantages of coupling with D2K to build highly functional data mining applications such as these include reduced development time through module reuse and sharing, and access to D2K distributed computing and parallel processing capabilities.

III. METHODOLOGY AND RESULTS

To support various data formats, a common interface is designed to visualize, preprocess and analyze the data. Some of the supported data formats include hierarchical data formats (HDF), digital elevation model (DEM), and geographical information system (GIS) supported vector files. The overall system architecture has been divided into four parts (Fig. 1). These components are explained below:

1) *Read raster data using I2K:* I2K is an image analysis tool, designed to automate processing of huge datasets and is capable of analyzing multi-dimensional and multivariate image data. When analyzing multiple geographic datasets over the same geographic area, it is necessary to preprocess and integrate heterogeneous datasets. I2K is a key component for preprocessing, visualizing and integrating the diverse datasets. I2K uses HDF libraries to load HDF data, and links to ArcGIS Engine functionalities to operate on GIS data formats. Fig. 1 shows the visualization of different scientific data sets: Snow cover, Albedo, LST (Land Surface Temperature), FPAR (fraction of Photosynthetic active radiation) and DEM.

Fig. 2 shows Graphical User Interface (GUI) associated with the visualization of HDF data in I2K. An HDF file may contain more than one scientific data set. User can select the scientific dataset (SDS) for display. Once the image is loaded, user can zoom, crop and play all spectral bands. Geographical information and image related information associated with the data sets can also be viewed by selecting GeoInfo and ImageInfo options respectively in the menu bar.

2) *ArcGIS Engine*: It is a complete library of GIS components which can be embedded into custom applications. I2K links to these libraries for features extractions e.g. calculation of slope, aspect, and flow accumulation grid from DEM. These derived variables are used for analysis along with the remote sensing data sets.

3) *Create Relational Database:* Creating a user Database (Fig. 3) is a data preprocessing and integration step. Different scientific datasets such as Enhanced Vegetation Index (EVI), Albedo, Leaf Area Index (LAI), Emissivity and Sea Surface Temperature (SST) are at different spatial and temporal resolution. Also there is quality assurance and quality control (QA/QC) data associated with each scientific variable. QA/QC data provide information about the quality of data for each pixel inside a scientific dataset.

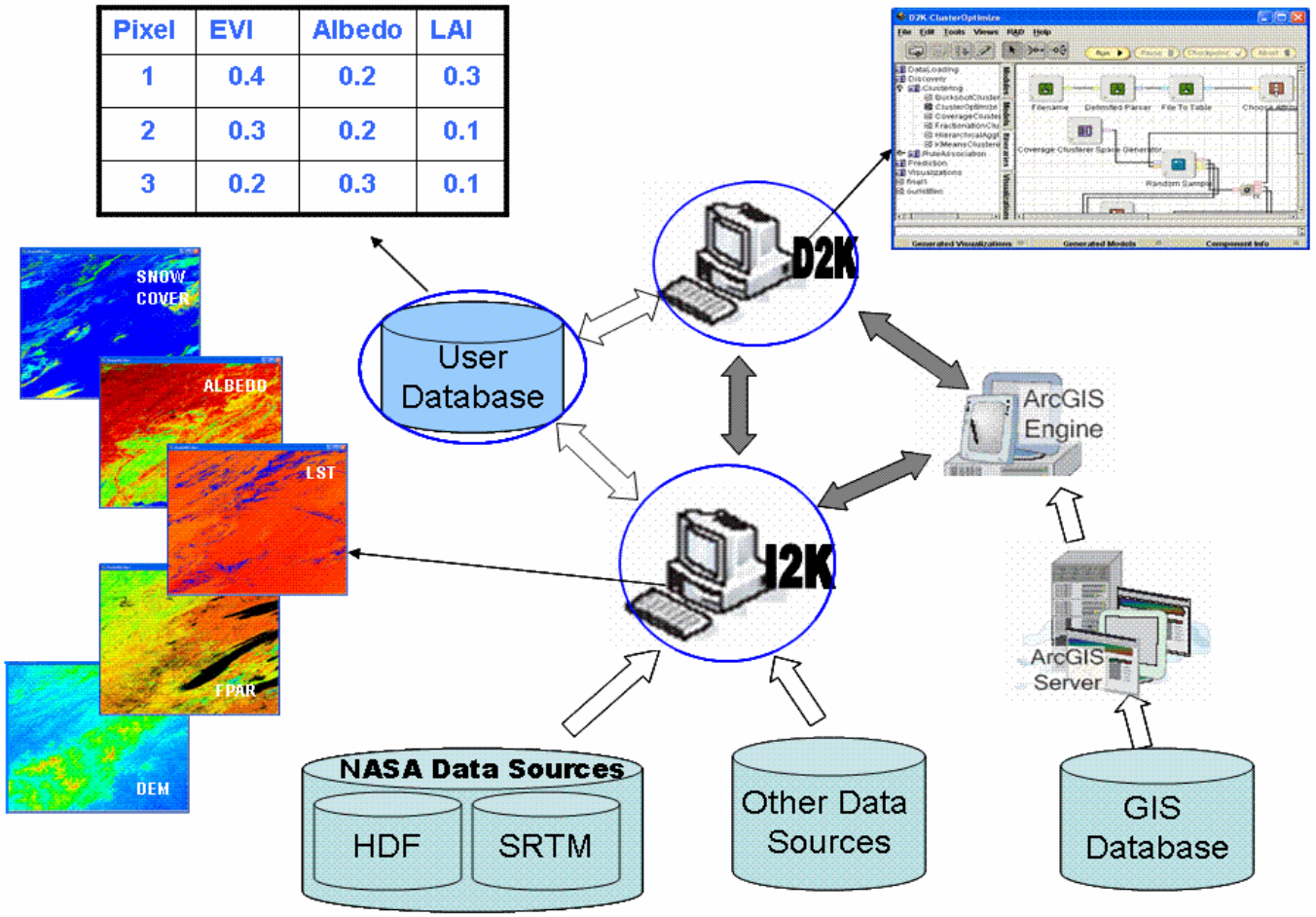


Fig. 1. Illustration of overall system architecture for data ingestion, preprocessing, integration, visualization and data analysis using various data mining algorithms. I2K reads all data sets from different data sources and visualize (snow cover, Albedo). It calls GIS functions using ArcGIS engine interface to perform feature extraction tasks (slope, aspects). All measured and derived variable are ingested in to database after preprocessing (spatial and temporal adjustment, removing bad pixels using QA/QC). D2K is used to analyze this database and results are visualized in I2K.

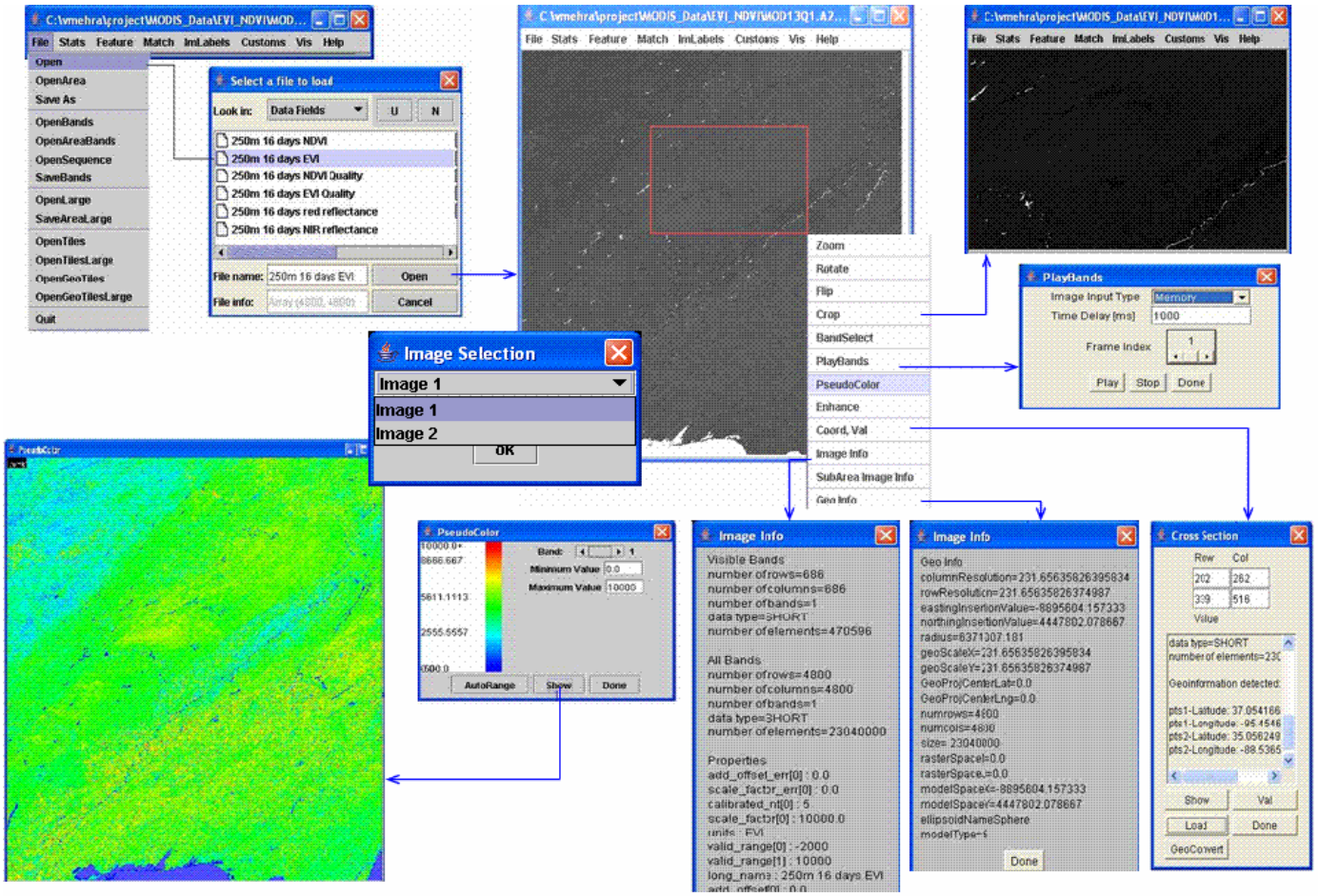


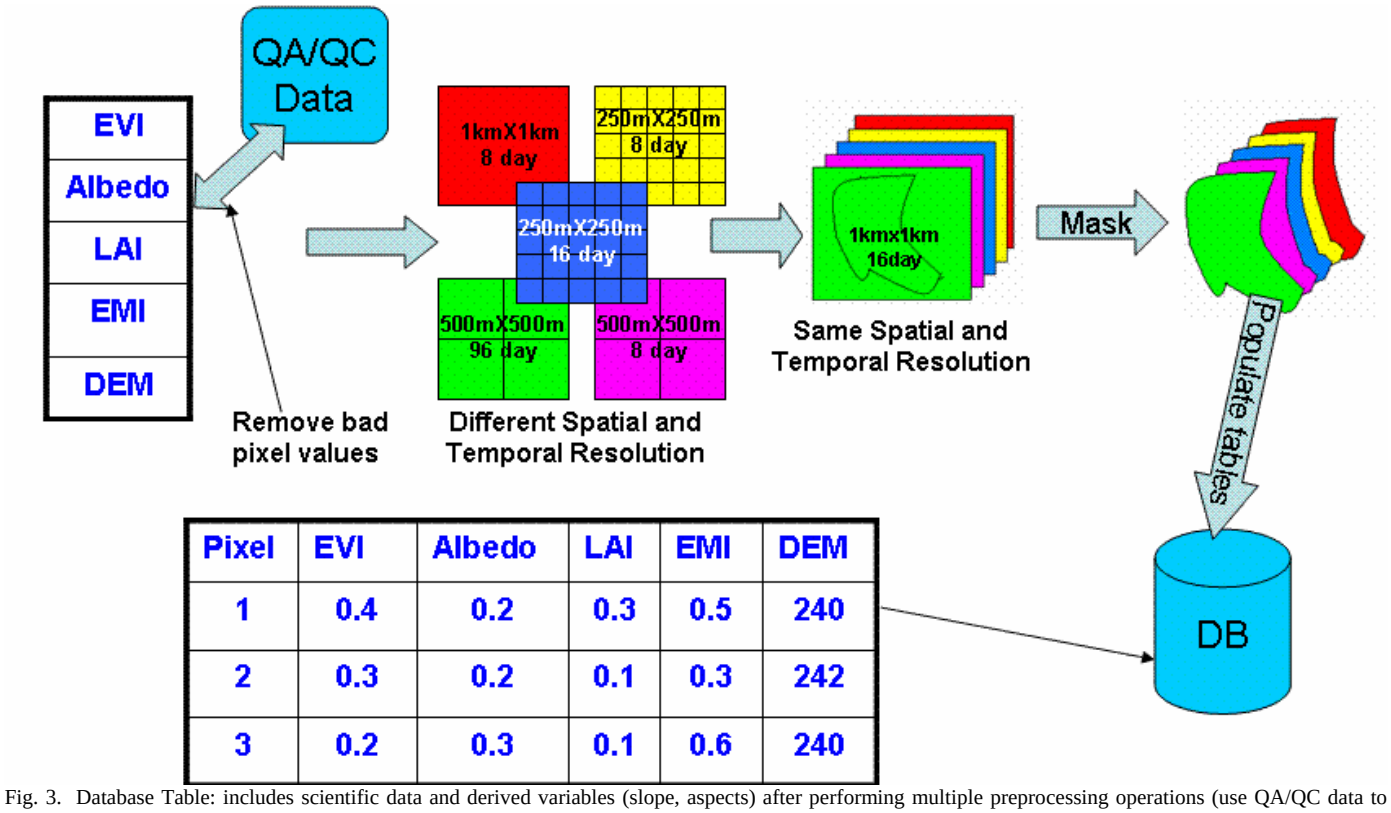
Fig. 2. Generic tool for loading different datasets. Interactive visualization environment (zoom, crop, geographical information, play all spectral bands of data) for integrating data mining and visualization processes.

To create an analysis database, we need to choose a unique spatial and temporal resolution. This is done by upscaling or downscaling the data. The unique spatial and temporal resolution is supplied by user as an input before creating the database. User may be interested in analyzing the data for a particular region only (Fig. 3). In that case (s)he can create a mask by selecting the area that (s)he wants to analyze. QA/QC data is used to remove bad pixel values e.g. no data values or bad pixel data received by satellite due to clouds. This option is again provided by the user. After all the above processing is done, integrated scientific and derived data sets are written into a database (Fig. 3).

4) Use D2K for data mining: This task plays the central role to enable automatic knowledge discovery through data mining. D2K uses database created in the above step as an input. It has modules for a variety of algorithms like multiple regression, Naïve Bayes, Decision Tree, and Neural Network to find various characteristic of data sets. Scientific question which we aim to answer are: (1) identify the dependence of the dynamically evolving variables on each other and their temporal scales of variability; and identify the roles of climate variability as a determinant of the variability in the dynamically observed quantities (2) identify how land-surface characteristics (elevation, slope, aspects, soil properties etc) further modulate the dynamical evolution of vegetation.

Overall procedure can be summarized as follows (Fig. 1):

STEP 1. Ingest all data sets using I2K
STEP 2. Visualize each scientific data set as necessary
STEP 3. Use native I2K functions along with ArcGIS Engine links to perform various feature extraction tasks.
STEP 4. Use QA/QC to remove bad quality pixel
STEP 5. Perform upscaling or downscaling of SDS to get unique spatial and temporal resolution
STEP 6. Mask the data set
STEP 7. Integrate data by writing all SDS and derived variables from SDS into database
STEP 8. Use D2K to run data mining algorithms on database created in above step.
STEP 9. Visualize results in I2K or GIS.



Pixel	EVI	Albedo	LAI	EMI	DEM
1	0.4	0.2	0.3	0.5	240
2	0.3	0.2	0.1	0.3	242
3	0.2	0.3	0.1	0.6	240

Fig. 3. Database Table: includes scientific data and derived variables (slope, aspects) after performing multiple preprocessing operations (use QA/QC data to remove bad pixel or no data values, spatial and temporal sampling adjustments, masking data sets, and error checking) and data integration.

IV. SUMMARY AND CONCLUSION

With the recent emergence of field of data mining, there is a need for a system that can handle large data sets and data assembly, preprocessing, and integration tasks. We are developing I2K as a common interface which can load HDF data from NASA data sources, supports visualization as well as data preprocessing and integration tasks. Further, it can use and extend functionalities present in ArcGIS using ArcEngine. Data mining algorithms present in D2K are applied on integrated data sets to find various patterns and relations between different variables. The understanding developed through our analyses will enable us to better parameterize the various natural processes for weather and climate models and thereby improving their predictability.

ACKNOWLEDGMENT

The support of this work is provided by the National Aeronautics and Space Administration (NASA) (Grant numbers NNG04GP78G and ESSF/O2-0000-0216) and National Science Foundation (NSF) (Grant number EAR 04-12859). Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflects the views of NASA or NSF.

REFERENCES

[1] Bajcy, P., P. Groves, S. Saha, T.J. Alumbaugh, and D. Tcheng, " A system for territorial partitioning based on GIS raster and vector data," Technical Report NCSA-ALG-03-0002, February 2003.

[2] Koperski, K., "A progressive refinement approach to spatial data mining,", Ph.D thesis , Simon Fraser University, pp. 175, April 1999.

[3] Mesrobian et al., "Mining geophysical data for knowledge," IEEE EXPERT, pp. 34-44, 1996.

[4] White, A. B., P. Kumar, and D. Tcheng, "A data mining approach for understanding topographic control on climate-induced inter-annual vegetation variability over the United States," to appear in Remote Sensing of Environment, 2005.[i]

To appear in the Proceedings of the 2005 NASA Earth-Sun System Technology Conference, Maryland.