10.12 EARLY LEAD: A WRF ENSEMBLE DEMONSTRATING A DATA MINING CAPABILITY

Richard D. Clark* and David Fitzgerald
Millersville University of Pennsylvania, Millersville, Pennsylvania

Thomas Baltzer
Unidata Program Center, UCAR, Boulder, CO

Rahul Ramachandran
University of Alabama-Huntsville, Huntsville, AL

Everette Joseph and Sen Chiao
Howard University, Washington, DC

1. INTRODUCTION

In response to a pressing need for a comprehensive national cyberinfrastructure in mesoscale meteorology, particularly one that can interoperate with those being developed in other relevant disciplines, the National Science Foundation in 2003 funded a Large Information Technology Research (ITR) grant known as Linked Environments for Atmospheric Discovery (LEAD). This multi-disciplinary effort involving nine institutions and some 80 scientists and students is addressing the fundamental IT and meteorology research challenges needed to create an integrated, scalable framework for identifying, accessing, decoding, assimilating, predicting, managing, analyzing, mining, and visualizing a broad array of meteorological data and model output, independent of format and physical location (Droegemeier et al, 2005).

One of the major goals of LEAD is to develop and deploy technologies that will enable mesoscale ensemble forecasting using a dynamically adaptive, on-demand, grid-enabled system. Simply put, LEAD is creating the IT needed to allow people (students, faculty, research scientists, operational practitioners) and atmospheric tools (radars, numerical models, data assimilation systems, data mining engines, hazardous weather decision support systems) to interact with weather. And while mesoscale meteorology is the particular problem to which LEAD concepts are being applied, the methodologies and infrastructure being developed are extensible to other domains such as biology, oceanography, and geology (Droegemeier et al, 2005).

EarlyLEAD is a test-of-concept effort within the LEAD project spearheaded by the education testbeds to autonomically identify significant mesoscale features using a data mining tool and invoke the WRF model to follow their development. The purpose of EarlyLEAD is to demonstrate and evaluate instantiations of LEAD technologies and to bring a subset of LEAD capabilities into computing environments that are not likely to have authorization on the TeraGrid, even when LEAD is fully functional. EarlyLEAD demonstrates a small subset of the tools and applications that are being developed for LEAD, and consists of 1) a data mining tool, 2) the WRF modeling system run at three testbed sites, 3) an open-source protocol to access stored data, 4) a data services catalog, and 5) a visualization tool. By adapting tools from different sources and combining them to create a new set of products, EarlyLEAD scales the LEAD technologies to faculty and students for use now, while LEAD researchers continue to develop the more sophisticated applications and tools for dynamic workflow orchestrations for meteorology research across the TeraGrid.

2. PHENOMENON EXTRACTION

EarlyLEAD begins with the most recent NAM or WRF forecast. The output from the model is mined by a Phenomena Extraction Algorithm (PEA) developed at the University of Alabama at Huntsville (UAH). PEA is a specialization of the Intelligent Data Thinning Algorithm, which is part of a more comprehensive Algorithm Development and Mining (ADaM) toolset, one of several

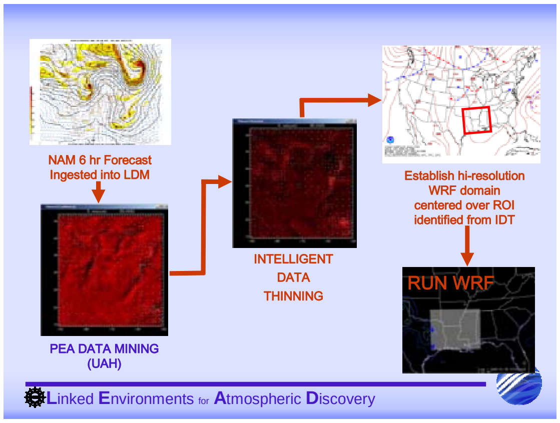
*Corresponding author address: Richard D. Clark, Dept. of Earth Sciences, P.O. Box 1002, Millersville University, Millersville, PA 17551; Richard.clark@millersville.edu

applications that comprise the Service-oriented architecture being developed by LEAD researchers.

The PEA is a feature extraction methodology designed to identify and extract regions representing phenomena in a geospatial data set. A phenomenon is defined as any state or process known through the senses rather than by intuition or reasoning, and thus is an observable event, especially something special or unusual. A phenomenon in a geospatial data set can be identified as a region significantly different from the rest of the scene. Statistically, therefore, this region of geophysical phenomenon can be characterized as having higher or lower than the average background intensity value and/or having higher intensity variations or gradients as compared to the remaining data points.

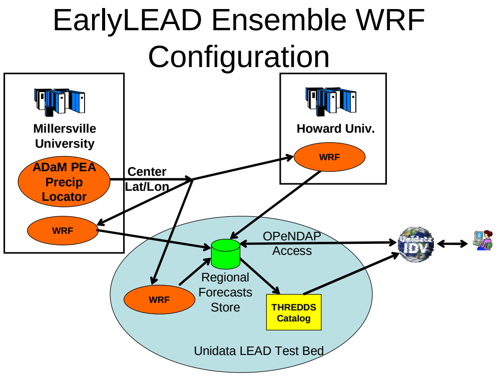
The PEA algorithm has two components. The first component is a recursive decomposition algorithm to divide the data into smaller regions, and statistical tests to evaluate these regions forms the second component. The KD-tree data decomposition strategy is used in the PEA algorithm to partition the original image into sub-regions hierarchically. The KD-tree stands for the k-dimensional binary tree. In each partitioning, a region is recursively split into two sub-regions, thus an image can be effectively represented as a number of sub-images hierarchically. At each level of decomposition, statistical F hypothesis test and student's T-test are applied to the data in the sub-region. These statistical tests determine if the region contains data points with abnormal intensities or large variations of intensities compared to the data statistics of global image. If these statistical hypotheses are validated, the sub-region is selected as a candidate region representing the phenomena of interest and followed by further decomposition on that region. If the statistical hypotheses fail, the region does not contain the phenomena of interest and further decomposition on this region is terminated. The recursive splitting of a region of interest proceeds until the size of sub-region reaches six (6) pixels or less. These remnant regions are the phenomena of interest. More detail on the PEA algorithm can be found in Ramachandran et al., (2006); this session.

Fig. 1 is an illustration of the Early LEAD methodology using the PEA as a data mining tool. User-defined forecast quantities, such as the u and v components of wind velocity or ω (vertical velocity) and precipitation fields, are autonomically selected from the initial model output and subjected to the PEA. The algorithm identifies and outputs the center latitude and longitude of regions of interest (ROI), and prioritizes according to their significance. The PEA can be adjusted by the user to extract the ROI according to geographical location, pre-selected meteorological phenomena, and other user preferences. Once the locations of ROIs have been prioritized, the center latitude and longitude of the ROI with the highest significance is used as the basis for establishing a high resolution WRF model domain over that region. The WRF is initialized with the model (NAM or WRF) output, then run to produce a short-term forecast. The output from the WRF or NAM forecast is again subjected to the PEA to identify where to move the WRF domain. This sequential process of output-to-PEA-to-WRF-to-output steers the WRF domain to follow the interesting weather phenomenon identified through the PEA.



3. EarlyLEAD ENSEMBLE

Fig. 2 is a schematic of the EarlyLEAD configuration enabling the Steered Weather Research & Forecasting (WRF) Model ensemble effort. This configuration consists of operational and compute infrastructure at three of the LEAD institutions; Millersville University (MU), Howard University (HU), and the UCAR Unidata Program Center (UPC). The infrastructure relies upon software (the ADaM Phenomena Extraction Algorithm or PEA developed at a fourth LEAD institution, the University of Alabama at Huntsville (UAH).



Operationally, EarlyLEAD works as follows: the PEA is run at MU against the North American Model (NAM) precipitation fields to determine the center latitude and longitude of the highest concentration of forecast precipitation. Once this location has been determined, it is distributed to HU and UPC for use in configuring the center latitude and longitude for their Regional Scale WRF runs. It is also used to configure the WRF run at MU. The WRF is being run with differing setup configurations to allow for an ensemble comparison. These configurations can include but are not limited to differences in cumulus, radiation, boundary layer, and microphysical parameterizations, or nudging of the initial conditions.

After the three WRF runs have completed, they are delivered to the UPC LEAD testbed system where they are placed in an OPeNDAP (Open-source Project for a Network Data Access Protocol) accessible location. THREDDS (Thematic Real-time Environmental Distributed Data Services) catalogs of these files are then generated (Caron et al., 2005). In this way, end users can make use of the Integrated Data Viewer (IDV) to browse the catalogs to view all fields available and perform intercomparisons of the model output from the three institutions (Murray et al., 2005). IDV bundles to automate intercomparison of the latest precipitation fields will be made available to the community.

4. SUMMARY

EarlyLEAD should not be viewed as a surrogate for the meteorology and IT research efforts that are central to the LEAD mission. Instead, EarlyLEAD focuses on the testing of a simple workflow that uses a data mining tool to select regions of interest in a prescribed way to establish a WRF domain at three institutions, concluding with the creation of an ensemble product that can be cataloged, accessed, and visualized. By updating the WRF model with new NAM or WRF output, the WRF domain can be steered to follow an interesting weather phenomenon. However, in contrast to LEAD, which will demonstrate the potential benefits of dynamic adaptability and a cyberinfrastructure that supports complex forecast systems that change configuration on demand in response to the weather, EarlyLEAD is static during the simulation. In this regard, EarlyLEAD resembles an autonomous floater, where the PEA takes the place of the observer in locating regions of interest. Nonetheless, EarlyLEAD has already proved useful as an effort where LEAD components such as PEA can be demonstrated, evaluated and refined.

5. ACKNOWLEDGEMENTS

LEAD is a Large Information Technology Research (ITR) Grant funded by the National Science Foundation under the following Cooperative Agreements: ATM-0331594 (University of Oklahoma), ATM-0331591 (Colorado State University), ATM-0331574 (Millersville University), ATM-0331480 (Indiana University), ATM-0331579 (University of Alabama-Huntsville), ATM0331586 (Howard University), ATM-0331587 (University Corporation for Atmospheric Research), and ATM-0331578 (University of Illinois at Urbana-Champaign, with a sub-contract to the University of North Carolina).

6. REFERENCES

Caron et al., 2005: THREDDS Catalog Specification, Web page found at: http://www.unidata.ucar.edu/projects/THREDDS/tech/catalog/InvCatalogSpec.html

Murray et al., 2005: The Integrated Data Viewer (IDV), Web page found at: http://www.unidata.ucar.edu/software/idv/

Ramachandran, R., X. Li, S. Graves, R. D. Clark, and D. Fitzgerald, 2006: PEA: Phenomena Extraction Algorithm. 22nd International Conference on Interactive Information Processing Systems (IIPS), 86th AMS Annual Meeting, Atlanta, Georgia.