6.1

LINKED ENVIRONMENTS FOR ATMOSPHERIC DISCOVERY (LEAD): A CYBERINFRASTRUCTURE FOR MESOSCALE METEOROLOGY RESEARCH AND EDUCATION

[1,2]Kelvin K. Droegemeier, [3]V. Chandrasekar, [4]Richard Clark, [5]Dennis Gannon, [6]Sara Graves, [7]Everette Joseph, [8]Mohan Ramamurthy, [9]Robert Wilhelmson, [1]Keith Brewster, [8]Ben Domenico, [1]Theresa Leyton, [7]Vernon Morris, [8]Donald Murray, [5]Beth Plale, [6]Rahul Ramachandran, [9]Daniel Reed, [6]John Rushing, [1]Daniel Weber, [8]Anne Wilson, [1,2]Ming Xue, [4]Sepideh Yalda

[1]Center for Analysis and Prediction of Storms and [2]School of Meteorology
University of Oklahoma
Norman, Oklahoma

[3]Colorado State University
Fort Collins, Colorado

[4]Millersville University
Millersville, Pennsylvania

[5]Indiana University
Bloomington, Indiana

[6]University of Alabama in Huntsville
Huntsville, Alabama

[7]Howard University
Washington, DC

[8]University Corporation for Atmospheric Research
Boulder, Colorado

[9]National Computational Science Alliance
Urbana, Illinois

1. MOTIVATION AND OVERVIEW

Each year across the United States, floods, tornadoes, hail, strong winds, lightning, and winter storms – so-called mesoscale weather events -- cause hundreds of deaths, routinely disrupt transportation and commerce and result in annual economic losses greater than $13B (Pielke and Carbone 2002). Although mitigating the impacts of such events would yield enormous economic and societal benefits, research leading to that goal is hindered by rigid information technology (IT) frameworks that cannot accommodate the *real time, on-demand, and dynamically-adaptive* needs of mesoscale weather research; its disparate, *high volume* data sets and streams; and the tremendous *computational demands* of its numerical models and data assimilation systems.

Corresponding Author Address: Professor Kelvin K. Droegemeier, CAPS, Sarkeys Energy Center, Room 1110, 100 E. Boyd Street, Norman, OK, 73019-0628; kkd@ou.edu.

In response to this pressing need for a comprehensive national *cyberinfrastructure* in mesoscale meteorology, particularly one that can interoperate with those being developed in other relevant disciplines, we are addressing the fundamental IT and meteorology research challenges needed to create an integrated, scalable framework -- known as *Linked Environments for Atmospheric Discovery (LEAD)* -- for identifying, accessing, preparing, assimilating, predicting, managing, analyzing, mining, and visualizing a broad array of meteorological data and model output, *independent of format and physical location*.

A transforming element of LEAD, which is supported by a 5-year National Science Foundation Large Information Technology Research (ITR) grant, is *dynamic workflow orchestration and data management*. These capabilities provide for the use of analysis tools, forecast models, and data repositories *not in fixed*

configurations or as static recipients of data, as is now the case, but rather as dynamically adaptive, on-demand, Grid-enabled systems that can a) change configuration rapidly and automatically in response to weather; b) continually be steered by new data; c) respond to decision-driven inputs from users; d) initiate other processes automatically; and e) steer remote observing technologies to optimize data collection for the problem at hand.

LEAD is designed for use principally by the meteorological higher education and operations research communities and will be developed as a phased set of prototypes (§9) that embody an increasingly complete and sophisticated set of tools and capabilities. The starting point is the highly successful Internet data distribution (IDD) and thematic server (THREDDS) infrastructure of the University Corporation for Atmospheric Research (UCAR) Unidata Program. Atop this foundation the LEAD team and its collaborators are performing the basic research needed to support the unique requirements of mesoscale meteorology research and education.

2. ATMOSPHERIC SCIENCES DRIVERS

2.1. Directions in Mesoscale Research and Supporting Infrastructure

A significant component of today's knowledge base in mesoscale meteorology derives from numerical simulation, with computer models playing a central role in research, education and operational forecasting. Twenty-five years ago, only a few three-dimensional (3D) "hero simulations" of a severe thunderstorm in idealized settings, using a grid of order 10^5 points, could be performed at select high-performance computing sites. Each such simulation required laborious manual analysis. Today, sophisticated mesoscale forecast models, representing all relevant atmospheric processes and in some cases coupled with hydrologic and oceanographic models, are being operated locally, in real time[1], at dozens of universities, Federal research laboratories, and even private companies. This same technology, applied in a non-real time research framework, now can be used to generate hundreds of simulations for assessing complex atmospheric behavior in real or idealized settings. Such capability is particularly important for ensemble prediction, where multiple, concurrently valid forecasts are generated from slightly different initial conditions, from different models, or by using different options within the same model.

Unfortunately, these and other advances in mesoscale research have far outpaced the cyberinfrastructure needed to support them. The availability of increasingly sophisticated models, including the new Weather Research and Forecast system (hereafter WRF) being developed as a dual research-operational resource (Michalakes et al., 2000); much broader access to affordable and extremely powerful computers; the proliferation of high-performance networks; and rapidly growing private sector interest in customized numerical weather prediction (NWP) all have catalyzed demand for flexible software frameworks for weather research, especially that conducted in real time. Cyberinfrastructure, while advancing, has notable limitations that, if left unaddressed, will widen the gap between mesoscale tools and their execution environments, stifling advances in research and education.

2.2. Real Time Weather Research Challenges

Most experimental real time forecasts, especially those created at universities, are initialized with pre-processed analyses from the National Weather Service (NWS) – with no real time observations added – owing to the complexity of managing data flows, dealing with multiple and changing data formats, and synchronizing complex data ingest, processing, and forecasting software. Consequently, the primary research benefit of local forecasts – namely, the use of much finer model grid spacings, and more sophisticated dynamical and physical frameworks, than those employed operationally by the NWS – is offset by an inability to assimilate observations, especially those available locally (e.g., from local highway, agricultural, and electric utility mesonetworks) and collected on fine scales (e.g., from Doppler radar).

This limitation has important scientific implications: the spacing between model grid points and observations used in model initialization should be reasonably similar; otherwise, the use of fine grids is unjustified, except in regions of significant orographic or coastal forcing, or for features represented within coarse models (e.g., fronts) that collapse nonlinearly on finer grids. Further, this inability to use both local and non-local real time observations inhibits realization of the potential usefulness of the WRF because the research

[1] Real time is defined here as the transmission or receipt of information about an event nearly simultaneously with its occurrence, or the processing of data immediately upon receipt or request.

version contains none of the orchestration[2] components needed for real time data acquisition, assimilation and prediction. Because WRF will be used heavily within the research and education communities (Michalakes et al., 2000), where IT support to develop orchestration capabilities is scarce, the consequences are obvious and potentially severe. LEAD also is giving special attention to on-demand, user-driven capability, and to scalability, so that models and other mesoscale research and education tools can be run locally, if resources allow, or be run across the Grid with automatic resource scheduling – all in a manner transparent to the user.

2.3. Non Real Time Weather Research Challenges

Non-real time "research mode" simulations needed to understand increasingly complex behaviors at the mesoscale, including ultra fine-scale experiments (domains of 10^7 or more points) that resolve both a tornado and its parent thunderstorm, produce enormous volumes of output, the analysis of which today proceeds as it did a quarter century ago – through the laborious study of each run individually, even though parameter spaces may encompass hundreds to thousands of experiments. Tools are needed for end-to-end orchestration of workflows; to assimilate all available observations; to prepare initial and boundary conditions; and to mine massive amounts of resulting information to reveal complex physical relations, verify against observations, visualize four-dimensional behavior, and curate and catalog all results and their meta data for future use in digital libraries.

Not surprisingly, *the associated IT limitations are not confined to computer modeling, but extend to diagnostic case study analyses and the creation of climatological data bases*, both of which require identifying, gathering, synthesizing, and managing vast quantities of information across a broad array of observing technologies and formats.

3. COMPUTER SCIENCE (CS) AND INFORMATION TECHNOLOGY (IT) DRIVERS

The nimble weather research relevant to LEAD involves highly nonlinear dependencies among filtering, analysis, pattern recognition, mining, and simulations that are stream driven, have complex feedback loops, and have dynamic structure. Indeed, the exact choice and sequencing, as well as the scale of the tasks, may change in response to the evolving weather or to user commands. Achieving these capabilities requires significant, fundamental computer science/information technology advances in six key areas, described below. Together, they must guarantee on-demand response, dynamic flexibility and autonomic behavior across the entire system.

Workflow orchestration and coordination must enable construction and scheduling of parameterized execution task graphs, with data sources drawn from real-time sensor streams and outputs – themselves forming streams for later use by either other computations or data mining and visualization tools. In turn, *data streaming mechanisms* must support robust, high bandwidth transmission of multi-sensor data to and from potentially large numbers of geographically distributed sites. A *distributed monitoring and performance evaluation infrastructure* is necessary to enable soft real-time performance guarantees for ensemble executions and data streams by providing estimates of resource behavior and performance for the workflow orchestration. Hints from the performance monitoring and prediction infrastructure must guide initial ensemble scheduling and trigger re-evaluation when execution does not satisfy expectations. *Data management* involves not only the storage and cataloging of observational data, but also of model output and results from data mining. The latter requires correlation of ensemble outputs, perhaps at multiple geographically dispersed sites.

The dynamic processing environment of LEAD imposes additional requirements on *data mining* tools, including hazardous weather detection algorithms that identify and classify specific features being sensed in real-time. Because LEAD needs to adapt to changing conditions, the mining components must detect faults, allow incremental processing (interrupt / resume), and estimate run time and memory requirements based on properties of the data (e.g., number of samples, dimensionality). Finally, the resulting data must be *semantically rich*, enabling use by diverse tools and applications.

4. THE UNIFYING CONCEPT OF LINKED ENVIRONMENTS

The foundation of LEAD is a series of interconnected, heterogeneous virtual IT "Grid environments" that are linked at several levels to

[2] Orchestration refers to the software needed to control and coordinate processes, data and workflows, and software components such as models, assimilation systems, and data mining engines.

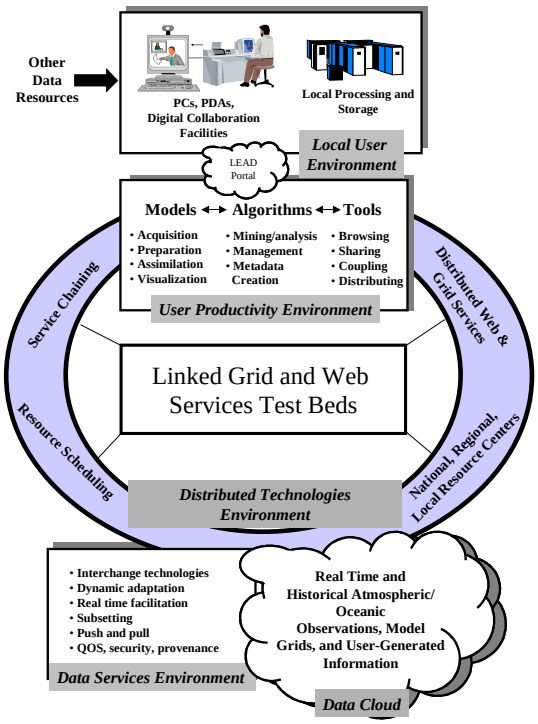

Figure 1. Functional concept of linked environments.

enable data transport, service chaining, interoperability, and distributed computation. Shown in Figure 1, these environments provide a complete, distributed framework within which users can identify, obtain, and work with observational and user-generated data where the problem being addressed, the relevant data, and the computational resources can change with time and be dependent upon or control one another.

From the user perspective, these environments include a deployable system consisting of the LEAD Toolkit and "MyLEAD," both within the *Local User Environment*. MyLEAD represents a virtual information space for controlling information flows, posting results for access by others, and managing interconnected processes. The *User Productivity Environment* contains models, tools and algorithms for operating on data and other information available within the LEAD *Data Cloud* – a virtual space of existing servers that provides access to all types of geophysical data (§6).

The *Data Services Environment* handles the complexities of data transport, formatting and interoperability using interchange technologies. Users download and process data locally using *Productivity Environment* tools, or if the data are too large, operate on them remotely, using the *Distributed Technologies Environment* to schedule resources and distribute work across the Grid. In this environment, the Grid workflow infrastructure autonomously computes scheduling constraints, dynamically acquires resources, recovers from component errors and adapts to changing plans. A set of Linked Grid and Web Services Testbeds (§6,10) maintain a rolling archive of several months of recent data (plus selected historical data); provide tools for operating on these data (e.g., mining engines, algorithms); and serve as a framework (i.e., a mini Grid) for developing distributed Web services capabilities.

The LEAD software component architecture leverages ongoing research in Grid infrastructure and middleware, augmented by advanced, distributed event monitoring and data stream management tools. Applications, models, and user tools are based upon the Open Grid Service Architecture (OGSA) now being formulated in the Global Grid Forum (Foster et al. 2003). As a component architecture extension of the Web Services standards now being adopted by industry, OGSA allows us to leverage ongoing academic and industry software development and standards. Applications required for each environment are layered upon this base, with the MyLEAD virtual environment providing the advanced information and services needed for dynamically orchestrating workflows of application services and data. Semantic descriptions of both data and software facilitate interoperability, allowing these disparate components to be used together effectively and dynamically to solve user problems.

5. EMPOWERING MESOSCALE RESEARCH AND EDUCATION: A REPRESENTATIVE SCENARIO

Given the multiple ways LEAD can be utilized, the capabilities and benefits to be enabled are best illustrated by a *use case scenario*. Although this particular example touches all aspects of LEAD to demonstrate its full potential, *LEAD is not geared solely to the high-end user community*, but rather brings to all users practical, sustainable integrated Grid and Web Services.

Suppose a graduate student wishes to understand why some severe thunderstorms produce a succession of mesocyclones and multiple tornadoes, while others do not. The first step, involving capabilities to be made available during Phase I of LEAD (§9), requires establishing a climatology of observed storm behavior. The Web-enabled LEAD portal, which is the access point to all of LEAD's capabilities, allows the

student to search and locate, access, and decode all required data – including ten years of WSR-88D (NEXRAD) Doppler radar data, along with upper air observations and NWS model forecasts, hourly surface observations, weekly land surface data, 6-minute precipitable water data from GPS satellites, and 15 minute satellite radiance data – all for the contiguous United States.

Because the study concerns only intense thunderstorms, a mining engine within the LEAD toolkit is applied to the NEXRAD data to identify only those dates and times when severe thunderstorms were present. Using the LEAD portal, the student then accesses the appropriate subset of data, which are too voluminous to be stored locally and must be stored on a LEAD Testbed site. There, the disparate, asynchronous and distributed observations are combined via a data assimilation system to yield a set of dynamically consistent, gridded, three-dimensional fields of all principal meteorological variables at five minute intervals. Using feature detection and pattern recognition techniques, the student applies a data mining engine to the assimilated data sets to catalog all cyclic versus non-cyclic storms, the existence of tornadoes, and the surrounding environmental conditions associated with each. The resulting metadata, along with the assimilated data sets, are then automatically available on the student's MyLEAD virtual information space for use by the broader community, even though the data physically reside at the LEAD Testbed.

Using capabilities to be made available during Phase II of LEAD (§9), the student then develops a parameter space of 500 idealized numerical simulations designed to provide a physical-dynamical understanding of the storm cycling process. The simulations produce hundreds of terabytes of output, and mining techniques are used to correlate cyclic storm behavior with environmental characteristics and internal storm dynamics. The simulation output and its metadata again are accessible on the student's MyLEAD server and are automatically published to digital library catalogs [e.g., National Science Digital Library (NSDL) and Digital Library for Earth System Education (DLESE)]. To examine the predictability of cyclic storm behavior in an operational environment, the student, in collaboration with operational researchers at the NOAA/FSL, uses LEAD orchestration tools to configure a set of 50 real-time, high-resolution WRF ensemble forecasts. Because she wants to run the ensembles only when thunderstorms are actually forming, the student applies data mining tools to streaming real time feeds from all NEXRAD radars in the U.S. to identify storm locations. With Phase III LEAD capabilities (§C9), mining tools trigger the WRF ensemble system over appropriate domains, which in turn automatically requests. Grid computing resources with sufficient priority to provide results significantly faster than the weather unfolds. As storms form in new regions and intensify in others, additional ensemble forecasts are spawned and finer grids are emplaced to capture their evolution. This on-demand requirement for additional resources is handled automatically by the Grid.

In this example, data mining, atmospheric modeling, and computing systems were interoperating and responding to the weather, whereas the observing systems were providing data independent of all other activities. In contrast, arrays of dynamically adaptive, collaborating remote sensors – which reconfigure in real time to sense multiple phenomena – now are being developed or refined to optimize the collection of atmospheric data. Examples include the CHILL research radar at Colorado State University (CSU) and the dynamically adaptive, collaborative phased array radars being developed by the new NSF Center for Collaborative Adaptive Sensing of the Atmosphere (CASA; http://casa.umass.edu).

LEAD will develop, in Phases II and III (§9), the capability for algorithms and models to guide the collection of data by dynamically responsive remote sensors. In the scenario above, once a data mining engine detected a severe thunderstorm, it could change the radar's mode of operation to scan only that storm at high time resolution, or with different polarization diversity, so as to optimally provide observations for hazardous weather detection and model initialization.

6. ENABLING TECHNOLOGIES AND THE LEAD DATA CLOUD

The IT and meteorology research in LEAD is not *ab initio*, but rather builds upon several enabling technologies pioneered by the LEAD team. The first is the UCAR Unidata Local Data Manager (LDM) (Rew et al., 1990; Fulker et al. 2000), a software package for event-driven data distribution within the Unidata Internet Data Distribution (IDD) network. The Collaborative Radar Acquisition Field Test (Project CRAFT) (Droegemeier et al. 2000), which compresses and transmits, via the Internet, real time WSR-88D (NEXRAD) Level II radar data, uses LDM. The Unidata Thematic Real Time Environmental Data Distributed Services (THREDDS) (Davis and Caron 2002; Caron, 2002) project is developing an extensible

framework for distributed thematic data for a scientific data web, and Unidata's Integrated Data Viewer (IDV) provides analysis and visualization capabilities for data residing on THREDDS servers[3]. The Advanced Regional Prediction System (ARPS) (Xue et al. 2000, 20001, 2003) and the new Weather Research and Forecast (WRF) model (Michalakes et al. 2000) provide advanced tools for assimilation and modeling. With NASA funding, Unidata and NCSA are merging NetCDF and HDF so the former can be used in high performance computing environments, be more efficient, and work with parallel I/O interfaces and large arrays. These enhancements are being made with an eye toward implementation in WRF.

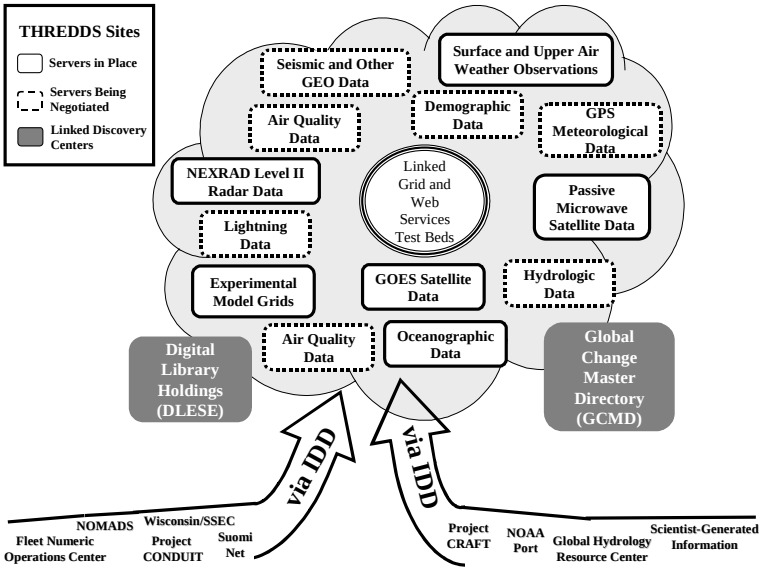

Figure 2. The LEAD Data Cloud.

The Algorithm Development and Mining (ADaM) System at the University of Alabama in Huntsville (UAH) provides distributed data mining components that may be accessed across the Web or Grid for federated data analysis solutions. These components provide data mining capabilities, such as phenomena detection and feature extraction, for large scientific data sets (Hinke et al. 1997; Ramachandran et al. 2000; Rushing et al. 2001; Hinke et al. 2001). OGSA work at Indiana (Foster et al. 2001, 2002) builds upon software component frameworks for Grid applications (Krishnan 2001; Govindaraju et al. 2002), and upon collaboration with IBM on its Business Process Execution Language for Web Services (BPEL4WS) workflow language (Cubera et al. 2003). The Earth Science Markup Language (ESML), developed at the UAH, encodes structural and semantic information required for data-to-application or model interchange (Ramachandran et al. 2001a,b; 2003). The dQUOB project (Plale 2002; Plale and Schwan 2003) at Indiana provides an advanced streaming technology for filtering data for applications requiring low latency. The LEAD monitoring architecture is based in part upon the Illinois Autopilot Grid Toolkit (Ribler et al. 2001), while the NOAA Operational Model Archive and Distribution System (NOMADS) (Rutledge et al. 2002) and Meteorological Assimilation Data Ingest System (MADIS) (Barth et al. 2002) represent options for accessing historical as well as real time gridded model output and observations.

Finally, the CHILL radar, operated by Colorado State University (CSU) and enabled as a remotely-controllable device, along with phased-array Doppler radars to be deployed in Oklahoma as part of the new NSF Engineering Research Center for Collaborative Adaptive Sensing of the Atmosphere (http://casa.umass.edu), will be used to develop and test observation system steering capabilities.

The enabling data technologies described above collectively create a "cloud" of distributed information resources we call the LEAD Data Cloud. Shown in Figure 2, it consists of the data, metadata, simulations, and other data-related services pertinent to mesoscale meteorology. Its foundation is a series of existing and future THREDDS servers, to which IDD provides a variety of real time data and metadata (over 35GB daily). IDD is an event-driven "push" system within which some 150 LDM sites relay data to one another using hierarchical distribution "trees" (Rew et al. 1990; Fulker et al. 2000). This non-centralized topology, coupled with multiple ingest sites, provides scalability and redundancy. LEAD leverages the significant investments made to date in THREDDS servers located at the: NCDC; National Geophysical Data Center; Space Science and Engineering Center; Lamont Doherty Earth Observatory, Pacific Marine Environment

[3] The phrase "THREDDS server" is used to describe a combination of OPeNDAP (Open-source Project for a Network Data Access Protocol) data services and THREDDS catalog services.

Laboratory; National Center for Atmospheric Research, Climate Diagnostics Center; Fleet Numerical Meteorological and Oceanographic Center; George Mason University/Center for Oceans Land Atmosphere; UAH, and OU.

The THREDDS infrastructure is supplemented by four Linked Grid and Web Services Testbeds (see also §9), which leverage their institution's particular expertise. For example, the data mining engine at one of the Testbed sites monitors incoming data at another to detect specific weather events. Upon such detection, a high-resolution local model run is initiated at another Testbed site, or across the Grid, with the output sent in real-time, via IDD, to another site for further mining and analysis. In this manner, the Testbeds serve both as development and production environments.

7. RESEARCH PROGRAM

The research components of LEAD consist of three distinct but highly synergistic elements: (1) basic IT and CS research driven by the unique needs of mesoscale meteorology to enable the linked environments system described above; (2) mesoscale meteorology research that utilizes these capabilities to address important scientific problems; and (3) the development, deployment, and refinement of tools and technologies by researchers, educators, and operational development technologists. We summarize below the principal components of the first two and refer the reader to the LEAD web site (http://lead.ou.edu) for further details. Information about deployment may be found in §9.

CS/IT Research Topics

- *Workflow orchestration* – the construction and scheduling of execution task graphs with data sources drawn from real-time sensor streams and outputs;
- *Data streaming* – to support robust, high bandwidth transmission of multi-sensor data;
- *Distributed monitoring and performance evaluation* -- to enable soft real-time performance guarantees by estimating resource behavior.
- *Data management* – for storage and cataloging of observational data, model output and results from data mining.
- *Data mining tools* – that detect faults, allow incremental processing (interrupt / resume), and estimate run time and memory requirements based on properties of the data (e.g., number of samples, dimensionality).
- *Semantic and data interchange technologies* – to enable use of heterogeneous data by diverse tools and applications.

Meteorology Research Topics

- *Data analysis system (ADAS) for the WRF model* – adaptation of the CAPS/ARPS Data Assimilation System (ADAS) to allow WRF users to assimilate a wide variety of observations in real time, especially those obtained locally;
- *Orchestration system for the WRF model* – to allow users to manage flows of data, model execution streams, creation and mining of output, linkages to other software;
- *Fault tolerance in the WRF model* – to accommodate interrupts in streaming data and user execution commands on the Grid;
- *Continuous model updating* -- to allow numerical models to be steered continually by observations;
- *Hazardous weather detection* – to identify hazardous features in forecasts, observations, and assimilated data sets using data mining technologies;
- *Storm-scale ensemble forecasting – to create multiple*, concurrently valid forecasts from slightly different initial conditions, from different models, or by using different options within the same or multiple models.

8. EDUCATION AND OUTREACH PROGRAM

LEAD contains a three-phase education and outreach program coordinated by Millersville University (MU) and Howard University (HU). It is designed to assess the effectiveness of LEAD technologies for education, provide critical input and feedback to IT developers, and facilitate knowledge transfer to a community of users (educators, researchers and students).

The first phase involves establishing education objectives to help shape the evolution and environment of LEAD, and to fuse the goals and enabling technologies into applications that are scalable and especially congruent with educational requirements, specifications, and standards.

During this phase, *LEAD Education Testbeds* (OU, Illinois, UAH, Millersville, and Howard) will engage successful national science and technology education initiatives (e.g., DLESE, NSDL, UCAR Windows to the Universe, DataStreme, and AMS Project ATMOSPHERE) to build on best practices that will help steer the development of an education-friendly LEAD User Productivity Environment. Education testbeds will initiate the development of tele-collaborative projects using LEAD technology with distinct goals tied to undergraduate and pre-college curricular improvements, including the NSF-funded Visual Geophysical Exploration Environment (VGEE) (Bramer et al. 2002; Pandya et al. 2002), which couples visualization (using Unidata's IDV) with data probes for inquiry-based exploration of the relationships between concepts and data.

The second phase commences with the flow of proto-tools and proto-technologies, including user documentation, for evaluation and refinement. Undergraduate and graduate students will join project participants to refine prototypes using assessment metrics that emphasize applicability, functionality, accessibility, scalability, and extensibility in the LEAD Educational Testbeds. In parallel, the MU Testbed will engage teacher-partners in evaluating prototypes for use in pre-college education, including a directed effort to align LEAD science and enabling technologies with National Science Education and National Educational Technology Standards, giving special attention to the need for the changing emphases on teaching, professional development, science content, and assessment. Outcomes from this collaborative assessment will provide critical feedback to the LEAD developers, resulting in progressive refinements to the tools and technologies as IT research proceeds toward deployable applications.

The third and culminating component focuses on deploying and integrating LEAD applications into higher education and pre-college learning environments to incite curricular changes as bold as the LEAD concept itself. Partner institutions have identified several curricular areas where LEAD capabilities will be integrated to drive innovation in meteorology and computer science courses (e.g., real time, dynamic experimentation in numerical weather prediction, computational fluid dynamics, algorithm design, networking, data management), and to other disciplines because of its inherent extensibility (e.g., oceanography, ecology). MU will assume a lead role for the integration of tools and services into undergraduate and pre-college curricula, while Howard University will coordinate graduate level WRF workshops (years 4 and 5) incorporating LEAD capabilities.

The dissemination of LEAD to the education community will occur via national and regional workshops and short courses at professional venues. This will be supplemented by the distribution of tele-collaborative tutorials and learning materials designed initially by the LEAD participants, and later, leveraging Unidata's success in building communities, by contributions from the community of users.

9. PROJECT ORGANIZATION AND TIMELINE

The interdisciplinary nature of LEAD requires a clear understanding of the roles to be played by the meteorology, IT, and education communities, as well as lines of communication, dependencies, and feedback loops. Consequently, the research in LEAD will be conducted in three phases, each striking a careful balance between basic inquiry and the development of practical tools and capabilities.

Figure 3 (inspired by the Grid Physics Network (GriPhy; http://www.griphyn.org/), presents this information and highlights knowledge transfer to the public policy and social sciences communities, as well as to other organizations that may use LEAD operationally or even commercialize some of its components. This diagram also highlights the fact that education is woven throughout the entire LEAD project and plays an especially critical role in assessing its effectiveness via the Education Test Beds.

The specific goals for LEAD, by phase, are:

Phase I (October, 2003 – April, 2005): ADAS front-end to WRF users; establish test beds; prototype LEAD Portal and mining and orchestration tools on Test Beds

Phase II (April, 2005 – October, 2006): Dynamically adaptive experiments on Test Beds; WRF and orchestration at partner sites including NCAR/NOAA Developmental Test Bed Center

Phase III (October, 2006 – October, 2008): Tools deployed via Unidata and components operational at the NCAR/NOAA Model Developmental Test Bed Center.

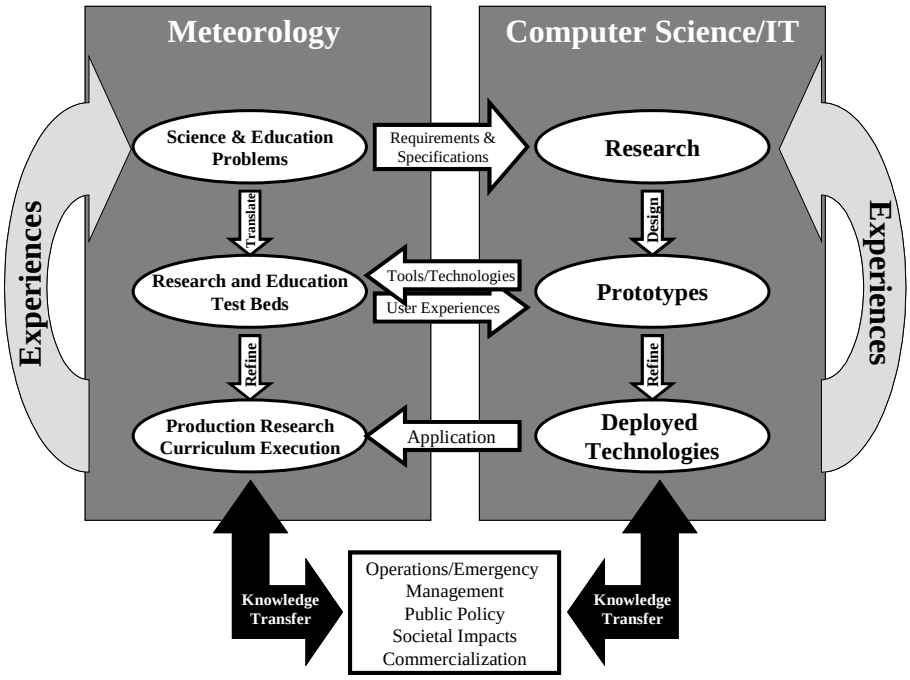

Figure 3. Lead development methodology (inspired by the GriPhyN Project).

10. INTEGRATION, DEPLOYMENT AND SUSTAINENANCE

The transformation by LEAD of basic research into useful, deployable and sustainable technologies centers around four Grid and Web Services Testbeds (Oklahoma, Illinois, UCAR, and Alabama in Huntsville) as the LEAD-specific development environment for CS and meteorological experimentation. Unidata and the University of Illinois have the primary responsibility for fusing research results into system components suitable for use by the environmental science communities that Unidata serves. As tools reach sufficient maturity, they will be rigorously tested in the crucible of Unidata's providers and academic users. Final products, improved by experimentation in the Testbed setting, will be subjected to Unidata's stringent release-engineering process that readies them for use in the large, diverse, demanding Unidata community.

LEAD capabilities also will be tested within the operations research environment of the NOAA Forecast Systems Laboratory. Education testbeds will play an integral role as well, and finally, WRF-related technologies to be developed by LEAD are envisioned to become part of the planned NOAA-NCAR Modeling Developmental Testbed Center.

11. SYNERGY WITH OTHER CYBER-INFRASTRUCTURE PROJECTS

A distinguishing characteristic of LEAD is its emphasis on dynamically adaptive, on demand, real time, fault tolerant Grid computing, and consequently it enjoys synergies with a number of other activities in the geosciences. The Earth System Modeling Framework is developing tools to enhance ease of use, performance, portability, interoperability, and reuse in climate, numerical weather prediction, and data assimilation applications. The Department of Energy's Earth System Grid (ESG) Project is focusing on access to climate model output, and both the NOAA Operational Model Archive and Distribution System (NOMADS) and the Meteorological Assimilation Data Ingest System (MADIS) represent synergistic options for accessing historical as well as real time gridded model output and observations. Considerable synergy also exists between LEAD and the two-year MEAD (Modeling Environment for Atmospheric Discovery) effort funded by the NCSA Alliance (see preprint in this volume by Wilhelmson et al.). MEAD (its name derived from LEAD) is an early limited prototype focused on providing a research environment for idealized ensemble simulations using a coupled atmosphere/ocean system (WRF/ROMS).

12. ACKNOWLEDGMENTS

LEAD is funded by the National Science Foundation under the following Cooperative Agreements: ATM-0331594 (Oklahoma), ATM-0331591 (Colorado State), ATM-0331574 (Millersville), ATM-0331480 (Indiana), ATM-0331579 (Alabama in Huntsville), ATM03-31586 (Howard), ATM-0331587 (UCAR), and ATM-0331578 (Illinois at Urbana-Champaign).

13. REFERENCES

Barth, M.F., P.A. Miller and A.E. MacDonald, 2002: MADIS: The meteorological data assimilation ingest system. Preprints, *Symposium on Observations, Data Assimilation, and Probabilistic Prediction*, Amer. Meteor. Soc.

Bramer, D. J., T. Scheitlin, R. Deardorff, D. Elliott, K. Hay, M. R. Marlino, D. Middleton, R. Pandya, M. K. Ramamurthy, M. Weingroff, and R. B. Wilhelmson, 2002: Using an Interactive Java-Based Environment to Facilitate Visualization Comprehension. *Preprints*, 18th Conference on IIPS. Amer. Meteor. Soc., Orlando, FL.

Caron, J., 2002: DODS aggregation and THREDDS catalog services. *Preprints*, *18th Int. Conf. on Interactive Information Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology*, 13-17 January, Amer. Meteor. Soc., Orlando, Florida, 54.

Curbera, F., Goland, Y., Klein, J., Leymann, F., Roller, D., Thatte, S., Weerawarana, S., *"Business Process Execution Language for Web Services, Version 1.0",* http://www-106.ibm.com/developerworks/webservices/klibaray/ws-bpel.

Davis, E.R. and J. Caron, 2002: THREDDS: A geophysical data/metadata framework. *Preprints*, *18th Int. Conf. on Interactive Information Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology*, 13-17 January, Amer. Meteor. Soc., Orlando, Florida, 52-53.

Droegemeier, K.K. and Co-Authors, 2002: Project CRAFT: A test bed for demonstrating the real time acquisition and archival of WSR-88D Level II data. Preprints, *18th Int. Conf. on Interactive Information Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology*, 13-17 January, Amer. Meteor. Soc., Orlando, Florida, 136-139.

Foster, I., Kesselman, C., Nick, J., and Tuecke, S., *"The Physiology of the Grid: An Open Grid Services Architecture of Distributed Systems Integration",* Technical Report, http://www.globus.org/research/papers.html, 2002.

Foster, I., Kesselman, C., and Tuecke, S., *"The Anatomy of the Grid: Enabling Scalable Virtual Organizations",* International Journal of Supercomputer Applications, 15(3), 2001.

Foster, I., Kesselman, K., Nick, J., Tuecke, S. "The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration", to appear in *Grid Computing: Making the Global Infrastructure a Reality*, Fox G., Ed. Wiley, 2003.

Fulker, D., R.K. Rew, and A.Wilson, 2000: A TCP/IP-Based System for Reliably Disseminating Near-Real-Time Meteorological Data. CBS Technical Conference on WMO Information Systems and Services,Geneva, Switzerland, 27-28 November.

Fulker, D., R.K. Rew, and A.Wilson, 2000: A TCP/IP-Based System for Reliably Disseminating Near-Real-Time Meteorological Data. CBS Technical Conference on WMO Information Systems and Services,Geneva, Switzerland, 27-28 November.

Govindaraju, M., Krishnan, S., Chiu, K., Slominski, A., Gannon, D., Bramley, R., *XCAT 2.0: A Component Based Programming Model for Grid Web Services*, Technical Report, Computer Science Department, Indiana University, 2002.

Hinke, T., J. Rushing, S. Heggere, S. Ranganath, and S.J. Graves, "Target-independent data mining for scientific data: Capturing transients and trends for phenomena mining", Proceedings of the 3rd International Conference On Data Mining (KDD-97), Newport Beach, CA, August 14-17, 1997.

Hinke, T. H., J. Rushing, H. Ranganath and S. J. Graves, "Techniques and Experience in Mining Remotely Sensed Satellite Data", *Artificial Intelligence Review* (AIRE, S4): Issues on the Application of Data Mining, pp 503-531, 2001.

Krishnan, K., *The XCAT Science Portal,* IEEE/ACM Supercomputing, November 2001.

Michalakes, J., S. Chen, J. Dudhia, L. Hart, J. Klemp, J. Middlecoff, and W. Skamarock: Development of a next-generation regional weather research and forecast model. Proceedings of the Ninth ECMWF Workshop on the use of Parallel Processors in Meteorology, Reading, U.K.,

November 13-16, 2000. Argonne National Laboratory preprint ANL/MCS-P868-0101.

Pandya, R., D. Bramer, D. Elliott, K. Hay, M. Marlino, D. Middleton, M. Ramamurthy, T. Scheitlin, M. Weingroff, and R. Wilhelmson, 2002: An inquiry-based learning strategy from the Visual Geophysical Exploration Environment (VGEE). *Preprints*, 11th Symposium on Education, Amer. Meteor. Soc., Orlando, FL.

Pielke, R.A. and R. Carbone, 2002: Weather impacts, forecasts, and policy. Bull. Amer. Meteor. Soc., **83**, 393-403.

Plale, B., "Leveraging Run-Time Knowledge about Event Rates to Improve Memory Utilization in Wide Area Data Stream Filtering," *Proceedings of the Eleventh International Symposium on High Performance Distributed Computing,* August 2002.

Plale, B. and Schwan, K., "Dynamic Querying of Streaming Data with the dQUOB System", *To appear in IEEE Transactions in Parallel and Distributed Systems*, 2003.

Ramachandran R., H. Conover, S. Graves and K. Keiser, "Algorithm Development and Mining (ADaM) System for Earth Science Applications", Second Conference on Artificial Intelligence, 80th AMS Annual Meeting, Long Beach, CA, January, 2000.

Ramachandran, R. M. Alshayeb, B. Beaumont, H. Conover, S. Graves, X. Li, S. Movva, A. McDowell and M. Smith, "Earth Science Markup Language: A Solution for Generic Access to Heterogeneous Data Sets." NASA Earth Science Technology Conference 2001, Greenbelt, MD, August 28-30, 2001a.

Ramachandran R., M. Alshayeb, B. Beaumont, H. Conover, S. Graves, N. Hanish, X. Li, S. Movva, A. McDowell and M. Smith, "Earth Science Markup Language", 17th Conference on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology, 81st AMS Annual Meeting, Albuquerque, NM, January, 2001b.

Ramachandran, R., S. Graves, H. Conover and K. Moe, "Earth Science Markup Language", Submitted to *Computers & Geosciences Journal*, Accepted with revisions, 2003.

Rew, R. K. and G. Davis, 1990: Distributed data capture and processing in a local area network. *Preprints*, 6th Int. Conf. on Interactive Information and Processing Systems for Meteorology, Oceanography and Hydrology, February, Anaheim, CA, Amer. Meteor. Soc., 69-72.

Ribler, R. L., Simitci, H. and Reed, D. A. "The Autopilot Performance-Directed Adaptive Control System," *Future Generation Computer Systems*, special issue (Performance Data Mining), 18 (1) pp. 175-187, September 2001.

Rushing, J., H. Ranganath, T. Hinke, and S. J. Graves, "Using Association Rules as Texture Features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(8) pp. 845-858, August 2001.

Rutledge, G.K. and Co-Authors, 2002: The NOAA operational model archive and distribution system (NOMADS). Preprints, *13th Symposium on Global Change and Climate Variations*, Amer. Meteor. Soc.

Xue, M., K. K. Droegemeier, and V. Wong, 2000: The Advanced Regional Prediction System (ARPS) - A multiscale nonhydrostatic atmospheric simulation and prediction model. Part I: Model dynamics and verification. *Meteor. and Atmos. Physics*., **75**, 161-193.

Xue, M., K. K. Droegemeier, V. Wong, A. Shapiro, K. Brewster, F. Carr, D. Weber, Y. Liu, and D.-H. Wang, 2001: The Advanced Regional Prediction System (ARPS) - A multiscale nonhydrostatic atmospheric simulation and prediction tool. Part II: Model physics and applications. *Meteor. and Atmos. Physics*, **76**, 134-165.

Xue, M., D.-H. Wang, J.-D. Gao, K. Brewster, and K. K. Droegemeier, 2003: The Advanced Regional Prediction System (ARPS), storm-scale numerical weather prediction and data assimilation. *Meteor. and Atmos. Physics*, **82**, 139-170.