# A. PROJECT SUMMARY

During the past several years, a number of key information technology (IT) and scientific advancements have positioned the meteorology and computer science communities to join forces with a view toward creating practical, sustainable technologies that promise to revolutionize the understanding and numerical prediction of nature's most deadly and societally-disruptive natural phenomena – mesoscale weather. Such weather includes individual thunderstorms and their larger aggregates; hurricanes; banded frontal precipitation systems; local winter storms; coastal circulations; orographic flows; and pollution episodes. In aggregate, these and related events in the United States each year claim dozens of lives and have an economic impact of several billion dollars.

The scientific advances and societal benefits associated with the accurate prediction of mesoscale weather are enormous; however, presently available computational and data management frameworks are unable to accommodate the real time, on-demand, and dynamically-adaptive nature of mesoscale problems; the complexities associated with its vastly disparate, high volume and bandwidth data; and the tremendous computational demands of its models and data assimilation systems. ***To overcome these limitations, we propose to conduct IT research that will enable an integrated, scalable framework -- known as Linked Environments for Atmospheric Discovery (LEAD) -- for use in accessing, preparing, assimilating, predicting, managing, mining/analyzing, and displaying a broad array of meteorological and related information, independent of format and physical location***.

The LEAD concept involves a series of interconnected IT "environments" that provide a complete framework within which users can identify, obtain, and work with observational, computer model, and user-generated information – and do so in a distributed setting where real time data streams and decision making are important, and where both the problem being addressed and the computational resources can change dynamically with time. The IT research will establish the *LEAD Portal*, operated within the *Local User Environment*, which will serve as the user's primary window to the LEAD world. The Portal will provide all linkages to the *User Productivity Environment*, which contains a broad array of tools, models, and algorithms for operating on and visualizing data and other information made available within the *Data Cloud*. Users will be able to download data to their local site and operate on them using local copies of *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. The complexities of multiple data types, formats, and projections will be handled by the *Data Services Environment* using data interchange technologies. This environment also will facilitate the real time distribution of time-critical data, provide security, and make available within the Data Cloud information generated by users.

The LEAD environments are supplemented by *Linked Grid and Web Services Test Beds*, which will provide significant online holdings of particular types of data, as well as related tools, and will allow a user to request data and services without technical knowledge of the underlying infrastructure. Because LEAD will utilize the power of desktop workstations yet have the capability to access distributed resources across the Grid, *its full functionality will be available to all institutions regardless of size or type.*

LEAD will deploy practicable, sustainable technologies for meteorological research, education, and operations, with specific emphasis on facilitating societal impacts research and policymaking. Related research topics deal with adaptive numerical weather analysis and simulation; deterministic and ensemble forecasting with assessment of use, value, and risk; scale interaction and predictability; and storm climatologies/threats with related implications to societal vulnerability and public policy. Education will be facilitated through the development of inquiry-based learning capabilities underpinned by LEAD technologies, with full scalability from the pre-college through graduate levels. Educational Test Beds will help shape LEAD research into applications that are congruent with educational requirements, national standards, and evaluation metrics.

By leveraging the many existing linkages between LEAD investigators and Minority Serving Institutions, traditionally underrepresented groups will be included in the research and educational elements of LEAD at all levels. Technology dissemination will be coordinated with the UCAR Unidata Program, which impacts tens of thousands of university students, educators, operational practitioners, and researchers each year at more than 130 participating organizations.

# C. PROJECT DESCRIPTION

## C.1. The Challenge: IT Infrastructures for Mesoscale Meteorology

During the past several years, a number of key information technology (IT) and scientific advancements (e.g., Abilene, terascale computers and storage devices, high-resolution observing systems, vastly improved data assimilation and modeling systems) have positioned the meteorology and computer science communities to join forces with a view toward *revolutionizing the understanding and prediction of nature's most deadly and societally-disruptive natural phenomena – mesoscale weather*. Such weather includes individual thunderstorms and their larger aggregates; hurricanes; banded frontal precipitation systems; local winter storms; coastal circulations; orographic flows; and pollution episodes. Although the term "mesoscale" suggests a particular scale, it actually encompasses spatial scales ranging from one to hundreds of kilometers and temporal scales ranging from tens of minutes to more than a day. Further, mesoscale events often have considerable impact on larger scale circulations and cycles as well as highly local bio- and eco-system processes.

The economic and societal benefits associated with understanding and accurately predicting mesoscale weather are enormous; however, presently available computational and data management frameworks are unable to accommodate the real time, on-demand, and dynamically-adaptive nature of mesoscale problems; the complexities associated with its vastly disparate, high volume and bandwidth data; and the tremendous computational demands of its models and data assimilation systems. ***To overcome these limitations, we propose to conduct IT research that will enable an integrated, scalable framework -- known as Linked Environments for Atmospheric Discovery (LEAD; http://lead.ou.edu) -- for use in accessing, preparing, assimilating, predicting, managing, mining/analyzing, and displaying a broad array of meteorological and related information, independent of format and physical location***. A *key feature of LEAD* will be the ability for users to work in a seamless manner on resources ranging from local desktop workstations or wireless devices to multiple, distributed high-performance systems, the latter of which will, via Grid and distributed Web services, accommodate on-demand allocation and rapid adaptation to changing problem configuration. A rich set of tools will empower users to formulate and seek answers to critical questions – *to discover new knowledge* -- by interacting with information in radically new ways. An *essential deliverable of LEAD* will be the deployment of practicable, sustainable technologies for research, education, and operations, with specific emphasis on their use in societal impacts research and policymaking. Thus, LEAD will serve as a common framework for linking students with their mentors, atmospheric researchers with operational practitioners and policymakers, and science educators with disciplinary experts.

### C.1.a. Meteorology Research Challenges

Weather analysis (more recently referred to as data mining) and prediction have a rich history that began over a century ago with the largest scales of motion, notably planetary waves and synoptic-scale cyclones and fronts. In the middle of the last century, the growing volume of observational data was used not only to analyze weather events, but also to initialize computer forecast models. Model grid spacings at that time were ~300 km; today's operational global and hemispheric models utilize equivalent grid spacings of ~50 km, while limited-area regional/synoptic models operate on grids of 15-20 km spacing. Although such representations of the atmosphere are vastly better than those used even a decade ago, they remain inadequate for capturing explicitly the most intense and locally disruptive weather. Indeed, between approximately 15 and 5 km grid spacing, no clear scale separation exists for convective clouds, i.e., they cannot be resolved explicitly, and closure assumptions regarding their representation as sub-grid scale phenomena are not applicable [1]. Research conducted during the past several years at the Center for Analysis and Prediction of Storms (CAPS), University of Oklahoma (OU) [2-7]; the National Center for Atmospheric Research (NCAR) [8,9]; and elsewhere [10,11] clearly demonstrates that grid spacings of 1-3 km are required to adequately capture the evolution of convective storms and related phenomena. Further, many model runs (an ensemble) are needed to quantify forecast uncertainty. *For these and other reasons, it is clear that the next major advancement in atmospheric prediction will come with the application of storm-resolving models (1-3 km grid spacings) that are initialized with observations of comparable resolution, and that are operated not on fixed schedules or in fixed configurations, as is now the case for coarser-grid operational systems, but in a manner that responds rapidly to the weather itself and to decision-driven inputs from the user.*

The scientific challenges needed to realize this capability, and to mine/analyze the huge volumes of observational data and model output -- significantly faster than the weather evolves -- are enormous. They include developing a better understanding of land-atmosphere interactions; of the mutual interactions and the communication of errors among scales; of complex cloud microphysical processes and their interaction with dynamics; and of the conversion of model output into useful knowledge and decision making information. The data volumes associated with these challenges place mesoscale meteorology among the most data intensive areas of science. Atmospheric observing systems range from ground-based radars to satellites to sensors on commercial aircraft, the data volumes from which, in aggregate, exceed 500 terabytes per year, with data intervals ranging from seconds to hours. In many cases, the data volume is expected to increase by an order of magnitude during the next 5 years. Particularly challenging are the *numerous sources and formats involved, and especially the requirements of real time delivery and processing owing to the time-critical nature of weather forecasting.* Rather than simply combining such data graphically or digitally, modeling systems employ variational control theory to optimally *assimilate* observations and model background fields, subject to constraints, so as to yield a *physically complete and dynamically consistent set of gridded state variables and parameters* [12]. This computationally intensive process also *retrieves quantities that cannot be observed directly* via fitting observations to a model [13-15], making assimilated data sets extremely valuable for analysis/mining.

### C.1.b. IT Research Challenges

Future mesoscale weather analysis/mining and numerical prediction will require far more than incremental extensions of today's capabilities. Consequently, they give rise to a number of unique IT challenges in computation, data, middleware, and physical infrastructure. To illustrate, Figure 1 presents a concept of operational mesoscale prediction in which successively finer resolution sub-grids are nested within a coarser outer grid forecast as a means of capturing the details of local, high-impact weather -- while at the same time representing the larger-scale environment. *This computation is highly dynamic* because a) it covers a broad spectrum of interacting time and space scales, ranging from the mesoscale environment (size of states and evolving over hours) to individual storms (the size of small towns and lasting less than an hour); b) new grids continually are initiated over regions of developing weather and removed over regions where active weather is dissipating; c) as storms develop, move across the country and decay, the national data delivery infrastructure (right panel) must respond rapidly by shifting data, networking, and computational resources to areas of active weather ; d) the end user may wish to initiate new forecasts on demand, or increase their frequency, due to weather developments near a major asset; and e) different parts of the computation (e.g., individual grids) can be run at different locations in a manner that maximizes throughput and load balances the entire problem. *This notion of dynamic networking linked to dynamic computation, data acquisition, and real time decision making is the essence of the National Computation and Information Grid [16] and distributed Web services [17-20] concepts.*

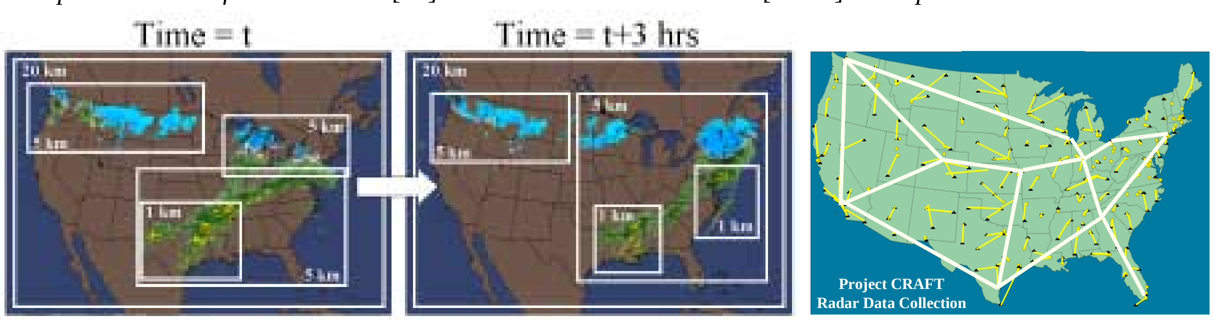


*Figure 1. Dynamic nature of mesoscale prediction: the nested domains in left two panels change with time as the weather evolves. Abilene/I2-based system that feeds NEXRAD Doppler radar and other data to the forecast grids must adapt to dynamically evolving bandwidth requirements to ensure quality of service.*

Realizing this scenario is difficult because the resources of large national service providers frequently are tightly coupled to that provider's data holdings, physical location and hardware. *In LEAD, no central facility will exist;* rather, data and computational resources (tools and services) will be distributed. Conducting research and educating students in such an environment requires close attention to performance issues owing to the need for processing and transferring enormous amounts of data in ***real time*** *(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).* Further, LEAD must be multilingual in its data formats, structures and scales in order to realize success in linking all the necessary components into viable solutions. The current IT environment is not suited for these tasks. Emerging distributed computing technologies address many of the system-level obstacles, and researchers within the LEAD team and elsewhere currently are investigating new methods of sorting through the myriad data formats and structures to provide the necessary interoperability (C.3.e). However, in order to use this information seamlessly for research, education and public dissemination, new paradigms for organization, searching and analysis/mining must be researched and developed. LEAD will provide the focus necessary to direct these efforts toward a practicable implementation, with emphasis on four specific IT areas, described below.

(1) Parallel and Distributed Computation. Numerical models and their related data assimilation systems – which represent key research tools in mesoscale meteorology -- are among today's most computationally-intensive parallel applications. The infrastructure necessary for their distributed execution by researchers and students – utilizing the richness of facilities available on the Grid – is absent. (2) Data. The enormous diversity of spatial and temporal scales in mesoscale meteorology is a key factor in the preparation of data for use in a wide variety of applications. Furthermore, *real-time data streams are an essential part of Earth science education,* especially for analysis and nowcasting. Current format variations and geophysical parameter inconsistencies make such multi-source use extremely difficult. (3) Middleware. Historically, advanced applications requiring access to network resources and services have had to provide such access themselves. Middleware that provides standardized, interoperable methods for accessing network resources and services is essential for mesoscale meteorology applications. Services such as collaboration tools, Quality of Service (QoS), and security mechanisms likewise are vital for allowing users to work seamlessly across a broad array of resources. (4) Physical Infrastructure. A complex system of distributed hardware, middleware, and data requires a robust infrastructure in order to function effectively. The framework must easily accommodate on-demand requests, rapidly changing problem configurations both locally -- and at multiple sites -- simultaneously, and must facilitate real time needs. These requirements dovetail well with both Grid and Web services paradigms.

## C.2. The Vision for LEAD

### C.2.a. Concept, Audience, Capabilities to be Enabled

The LEAD framework, shown in Figure 2, is founded upon a series of *virtual environments* that conceptually are familiar to all users of meteorological information (data, tools, connectivity, services), and which are *linked at several levels* to enable data transport, service chaining, interoperability, and distributed computation. Although the foundations for these linkages currently are evolving within the Internet2, Grid, and Web communities, the challenges involved in chaining together diverse services to work on multiple, disparate data and hardware in a truly distributed environment are daunting. Consequently, much of the research in LEAD will be directed toward developing the IT capabilities necessary to effectuate linkages -- both within and among the virtual LEAD environments -- so as to accommodate the unique aspects of mesoscale meteorology research and education (real time data and computations; resources available on demand; dynamic adaptability to the weather; and data- and compute-intensive applications).

LEAD will be designed for use principally by the research and higher education communities, which via Unidata encompasses *tens of thousands of students* (C.7). Some capabilities will be of significant value to educators in grades 6-12 (C.5). The *LEAD Portal*, operated within the *Local User Environment* (i.e., the totality of resources available at the user's location, which could be mobile), will be the primary window to the LEAD world, providing linkages to the *User Productivity Environment*, which contains a broad array of tools, models, and algorithms for analyzing/mining and visualizing information made available within the *Data Cloud* and from other sources. Users will be able to download data to their local site and operate on them using local copies of Productivity Environment tools or, if the data are too large, the user can operate on them remotely, using the *Distributed Technologies Environment* to schedule resources and distribute work across the LEAD infrastructure.

The complexities of multiple data types, formats, projections, and related issues will be handled by the *Data Services Environment* using data interchange technologies. This environment also will facilitate the real time distribution of time-critical data, provide security, and make available within the Data Cloud information generated by users. Four *Linked Grid and Web Services Test Beds* will maintain a rolling archive of several months of recent data, as well as selected historical data collections, and will serve as a framework for developing linked chains of Grid and Web services capabilities. Because LEAD will utilize the power of desktop workstations yet have the capability to access distributed resources, *its full functionality will be available to all institutions regardless of size or type. Thus, LEAD will contribute to bridging the "digital divide" that today tends to stratify institutions.*

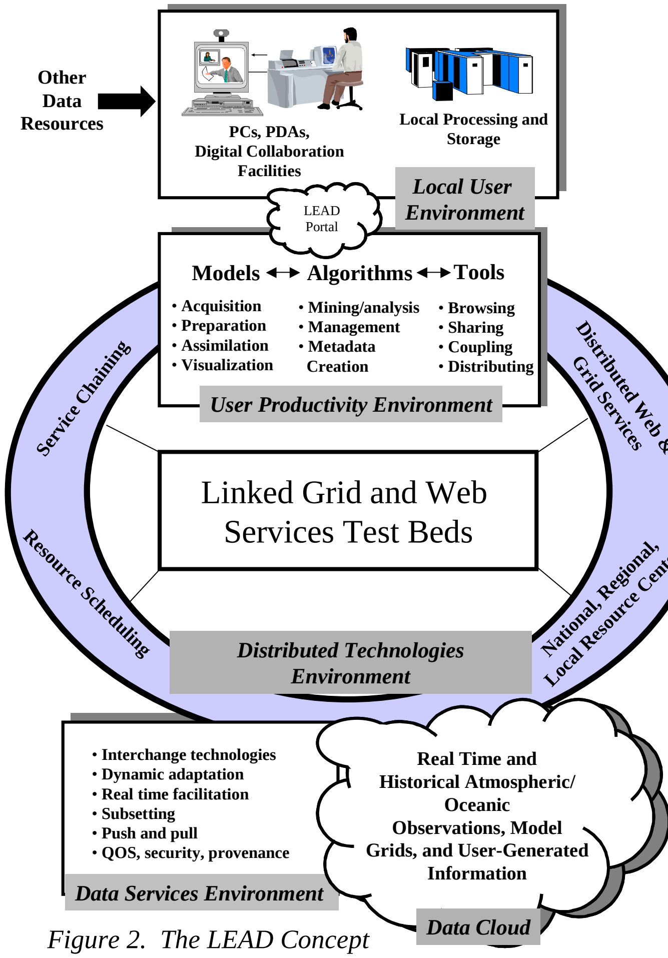


*Figure 2. The LEAD Concept*

### C.2.b. Guiding Principles

LEAD is guided by an overall philosophy that stresses research leading to the *deployment of practicable, sustainable technologies for research and education*. Within this philosophy, LEAD has chosen to focus principally on mesoscale meteorology because of its unique research and IT challenges, relevance to society, importance in public policy, and because mesoscale meteorology encompasses many of the needs relevant to other areas, such as climate. To broaden the effort in a meaningful way would require resources far beyond those available in the present solicitation. However, the infrastructure created within LEAD will be extensible to other disciplines, especially within geosciences, *and will have initial linkages to the ocean sciences, solid Earth sciences, and social sciences communities* (C.3.b).

The interdisciplinary LEAD research effort will require a clear understanding of the roles to be played by the meteorology, IT, policy/social sciences, and education communities, as well as lines of communication, dependencies, and feedback loops. Figure 3 (inspired by the Grid Physics Network, or GriPhyN project; 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, as described in C.5, that education is woven throughout the entire LEAD project and plays an especially critical role in assessing its effectiveness via Education Test Beds.

The LEAD technologies will be portable, utilize as little proprietary software as possible, and be *extensively documented both in the code and via user guides and tutorials*. Further, we will leverage existing efforts, including those of the Grid. For example, GriPhyN is pioneering technology for virtual data management, and the Global Grid Forum (GGF), in which LEAD is well represented, has over a dozen working groups actively defining standards for security, resource co-scheduling, performance analysis, data archive management, event notification and user services. Most recently, the GGF has been considering a new standard for managing Grid services, called the Open Grid Services Architecture (OGSA), which draws upon emerging Web services protocols that will be heavily used by LEAD. *Note that LEAD is not dependent upon the Grid, but rather will focus on broad Web services, including those of the Grid.*

## C.3. <u>Proposed IT Research</u>

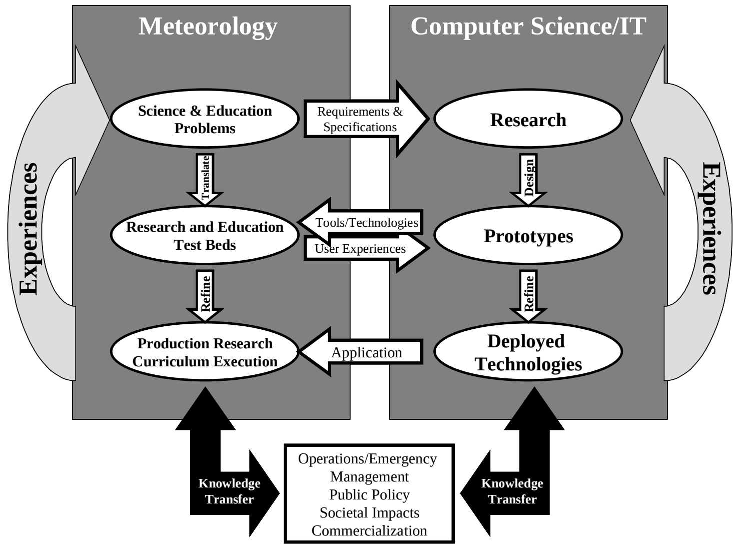


*Figure 3. Lead development methodology (inspired by the GriPhyN Project).*

The proposed IT research is built around nearly a dozen enabling activities (C.3.a), all of which have been independent in their evolution and have involved members of the LEAD team. It is important to note, however, that these efforts encompass important but relatively focused objectives that merely have set the stage for the current proposal. They cannot, on their present, mostly independent courses, achieve the bold goals outlined herein. Indeed, the "linked environment" concept at the heart of LEAD extends well beyond simply making the enabling activities work side-by-side. Rather, our goal is to create true interoperability in a dynamic, wide-area context. We envision users composing, in real time, data streams and numerical experiments into dynamic systems that solve real problems, with the output of one or more systems directing the activities of others. The details of our research are described below, and a *table summarizing milestones for the entire effort is presented in the Management Plan (C.11).*

### C.3.a. Enabling Technologies

The first enabling technology is the University Corporation for Atmospheric Research (UCAR) Unidata **Local Data Manager (LDM)** [21,22], a software package for event-driven data distribution. LDM drives the **Unidata Internet Data Distribution (IDD)** network, which is a community effort involving over 130 institutions that relay atmospheric data in near real time. One IDD data stream is that of the **Cooperative Opportunity for NCEP Data Using IDD Technology (CONDUIT) Project** [23]. It consists of high-resolution gridded fields generated by US operational forecast models. Another is from the **Collaborative Radar Acquisition Field Test (Project CRAFT)** [24], which compresses and transmits, via the Internet in real time, WSR-88D (NEXRAD) Level II radar data.

The **Thematic Realtime Environmental Data Distributed Services (THREDDS)** [25,26] project is developing an extensible framework for distributed thematic data for a scientific data web. THREDDS links closely with the **Digital Library for Earth System Education (DLESE; http://dlese.org)** and other projects to ensure that data collections can be referenced and searched using digital library systems. The **Unidata MetApps Project** [27] is a collaborative effort to develop software tools for learning and research in the atmospheric sciences. Using Java to provide platform independence, MetApps can be used interactively on local computers, via the Web, or on remote computers. The data to be analyzed may exist locally or be distributed on THREDDS servers.

The CAPS **Advanced Regional Prediction System (ARPS),** and the new **Weather Research and Forecast (WRF)** model, represent advanced tools for assimilation and modeling. The ARPS [4,28,29] is a fully automated, scalable-parallel, non-hydrostatic forecast *system* designed principally for high-impact local weather. Similar to the ARPS but being developed by several organizations including CAPS, the WRF [30] is a community-wide, dual-purpose research and operational system. In the areas of data mining/analysis and knowledge discovery, the **Warning Decision Support System (WDSS)** [31], developed by the National Severe Storms Laboratory (NSSL), applies sophisticated algorithms [32-36] to detect hazardous weather from NEXRAD radar observations. The **Algorithm Development and Mining (ADaM) System** was created by the Information Technology and Systems

Center at the University of Alabama in Huntsville (UAH) to mine large scientific data sets for geophysical phenomena detection and feature extraction [37-40]. ADaM was the first data mining application employed on the NASA Information Power Grid (IPG) [41]. **D2K (Data to Knowledge)**, developed at the National Computational Science Alliance (NCSA), is a rapid application development environment for creating and managing sophisticated, distributed data mining processes [42,43]. Finally, the fabric within which the above elements will be modified, extended and *linked* is the World Wide Web and the more specialized **National Computation/Information Grid** [16], the latter of which represents a broad array of computing and storage devices, federated data repositories, remotely controlled instruments, and sensors.

### C.3.b. Data Cloud with Linked Grid and Web Services Test Beds

The enabling technologies described above are components which, when taken together, create a "cloud" of distributed information resources we call the LEAD *Data Cloud*. Shown in Figure 4, it consists of the data, metadata, simulations, and other data-related services pertinent to mesoscale meteorology – including information generated by individual scientists (simulation results, analyses, field program data). Its foundation is a series of existing and future THREDDS servers, shown in Figure 4 by the solid and dashed rounded ovals, respectively, to which IDD provides a variety of real time data and metadata (approximately 10 gigabytes per day and, at peak times, over a gigabyte per hour). IDD is an event-driven "push" system within which some 130 LDM sites relay data to one another using hierarchical distribution "trees" [21,22]. This non-centralized topology, coupled with multiple ingest sites, provides scalability and redundancy. LEAD will leverage the significant investments made to date in THREDDS servers located at: National Climatic Data Center; National Geophysical Data Center; Space Science and Engineering Center; Lamont Doherty Earth Observatory, Pacific Marine Environment 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.

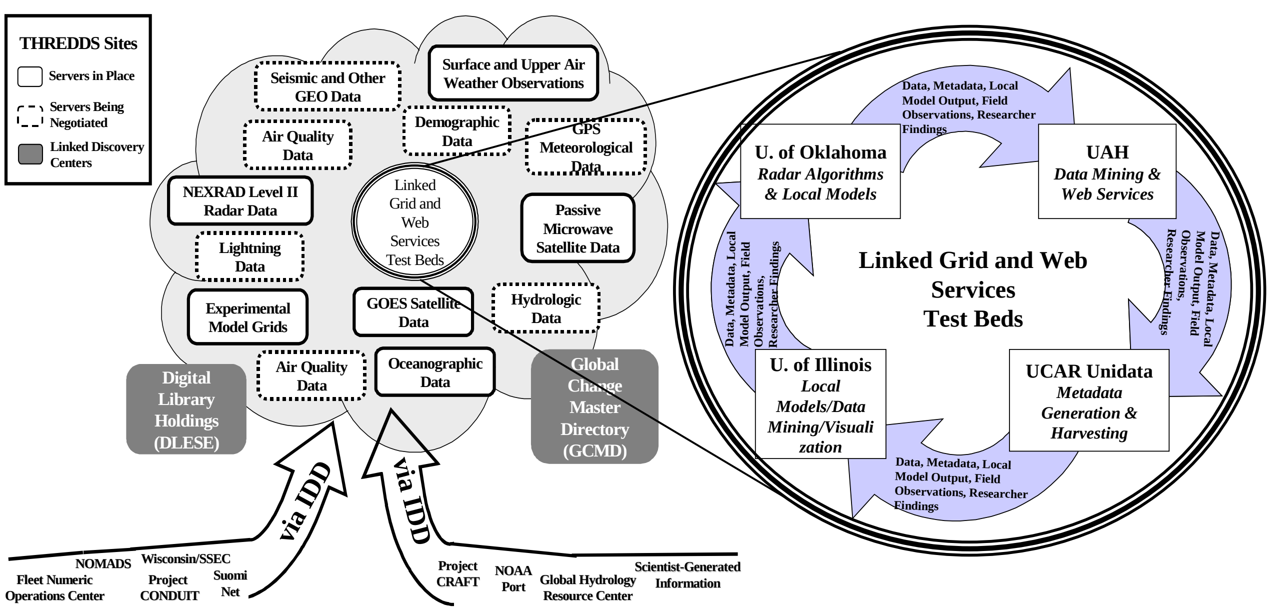


*Figure 4. The LEAD Data Cloud with Linked Grid and Web Services Test Beds.*

The THREDDS infrastructure will be supplemented by four *Linked Grid and Web Services Test Beds* that will 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 (C.3.e). These sites will leverage their institution's particular expertise but will be linked by chaining (C.3.e). For example, the data mining engine at one of the Test Bed sites will be able to monitor incoming data at another to detect specific weather events. Upon such detection, a high-resolution local model run can be initiated at another Test Bed 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 Test Beds serve both as development and production environments, and leverage the already significant investments made by the participating institutions in data ingest and analysis hardware. To accommodate the capabilities required by the Test Beds, we will use a system like the Storage Resource Broker (SRB) [44], which is a data-handling environment that supports support information and resource discovery in distributed, heterogeneous systems. We also will evaluate peer-to-peer storage and access solutions, networked storage, and other technologies. The final component of the Data Cloud is linked discovery centers (gray boxes in Figure 4), by which THREDDS incorporates metadata generated at server sites, and by users in the community, into central discovery systems at DLESE and the Global Change Master Directory (GCMD). Thus, users searching the discovery sites for environmental information will be led to LEAD data and tools where appropriate.

In order for the Data Cloud to meet the demands of LEAD, LDM must be upgraded to accommodate the rapidly increasing volumes of meteorological data. One strategy involves sub-setting, for which LDM must provide a much finer classification of products than currently is supported, and which includes a more dynamic subscription mechanism. In this context, Unidata is studying the use of Network News Transport Protocol (NNTP) and the Netnews network. NNTP supports several important capabilities such as a virtually unlimited number of hierarchical name spaces for groups that allow classification of data having much finer granularity. Data products can be cross-posted to multiple newsgroups, thus providing multiple views. Automated routing will be required so that LDM can respond quickly to dynamical subscription changes related to rapidly developing weather. For sites having marginal computing facilities, a simple "receive only," platform-independent version of the LDM is needed. *This latter capability will help ensure that LEAD is available to all institutions regardless of their size or resources*. Note that NNTP uses a method of automated routing called the "flooding" algorithm, in which messages are redundantly relayed to multiple downstream sites that intentionally are highly interconnected. As a result, a data product reaches its destination by the fastest route possible. NNTP also supports both push and pull transmission, allowing for more dynamic data requests. Our research will seek to identify mechanisms for implementing NNTP in ways that ensure a high quality of service without compromising security. For example, recipient sites must ensure that a product actually came from the intended originating site and was not modified in transmission. Further, sites must be protected from denial-of-service attacks. Meeting these conditions will require additional security mechanisms and the digital signing of products, coupled with news-specific utilities such as moderated newsgroups and control messages.

### C.3.c. Local User Environment and LEAD Portal

The LEAD Data Cloud and its component technologies discussed above represent the raw materials for meteorological research and education. The *Local User Environment* represents the totality of resources available to an individual user, a classroom of students, or to a scientist working in the field. This includes but is not limited to local workstations, storage devices, networks, and software packages. The *LEAD Portal* represents the user's window to the LEAD world and will combine existing tools, such as the platform-independent Unidata MetApps, with other features specific to the LEAD environment. Through the Portal, users will be able to drive all tools in the *User Productivity Environment*, as well as access tutorials and other learning materials.

<u>Data Discovery and Access</u>. Users will be able to search through the LEAD distributed information space to locate and access data objects using such established Earth science catalogs as the Global Change Master Directory [http://gcmd.nasa.gov] and/or Federal Geographic Data Committee Clearinghouse [www.fgdc.gov/clearinghouse/clearinghouse.html], as well as DLESE (note that three of LEAD's senior personnel have significant involvement and responsibility within DLESE). Because client-server technology, such as DODS (Distributed Oceanographic Data Systems), is being built into the THREDDS servers and client applications, end users will be able to integrate different datasets from the distributed servers from within data analysis and display software running on their own local computers. The options include "thin" client, browser-based access via the LEAD portal as well as fully interactive, "thick" clients such as MetApps.

Data Visualization. The platform-independent MetApps tools will enable users to analyze, visualize and interact directly with data integrated from diverse sources. Specifically, MetApps is an environment for producing three-dimensional views or 2D slices; interactively probing data to retrieve values or create profiles and time-series plots; animating visualizations both temporally and spatially; and applying functions and/or user defined equations to compute derived parameters. LEAD information also will be available through OGC-compliant Web Map Servers for image rendering, through Web Coverage Servers for distributed data access/retrieval, and through Web Feature Servers for access to meteorological features and events [www.opengis.org]. Additionally, software for linking visualization generation with feature detection will be made available, in addition to the visualization capabilities in AdaM and D2K (which include scatter plot matrices, parallel coordinates, and decision trees).

Collaboration Tools. Access Grid technologies [www.fp-mcs.anl.gov/fl/accessgrid/default.htm] and off-the-shelf desktop collaboration tools (e.g., inexpensive Web-based video conferencing systems) will be used to facilitate interaction among users of the LEAD environment. In addition, Electronic Notebooks, which can be thought of as histories containing executable scripts and links to data used and generated, can be shared with collaborators who can use them to call-up any data object or redo computations. The processing steps used to produce a set of results can be edited, modified and re-executed to carry out "what if" experiments.

### C.3.d. User Productivity Environment

The *User Productivity Environment* represents the set of tools available to the researcher, student, or operational practitioner to carry out all desired tasks in LEAD. Data Preparation Tools include services such as preprocessing, sub-setting, data format translation, gridding, fusion, and automated metadata generation. Both ARPS and WRF include a broad array of tools for data description as well as assimilation, interrogation, and weather prediction. Their variational data assimilation components produce from a set of limited observations a complete and physically consistent set of gridded fields, to which a variety of interrogation tools can be applied. Data assimilation also can be facilitated by the interchange technology to be described in section C.3.e, and the ready availability of proposed data translation services. The dynamic application of distributed data preparation services throughout the LEAD environment can provide “data-in-a-box” to the various models and analysis applications.

Model Tools represent a broad array of resources for data description as well as assimilation, interrogation, and prediction. The centerpieces will be the ARPS and WRF, which are advanced meteorological data assimilation and prediction systems. LEAD will enable these systems to be run in a variety of modes (C.4): a single forecast could run on a local or remote machine; many forecasts could be run in parallel on different computers within the Grid to create an ensemble; multiple fine-resolution grids, nested within a larger coarse-grid domain, could each be computed at a different facility; and so on. In all cases, data mining/analysis and visualization could be performed locally or distributed across the Grid. *It is important to recognize, however, that the term “model” has a much broader representation within LEAD, extending to models of data, models of networks (especially critical for the dynamically adaptive nature of mesoscale research), and learning models of models.* Such models also will be accommodated within LEAD.

Data Mining/Analysis Algorithms will be utilized to detect features and anomalies and to discover knowledge from various online data stores and assimilated analyses. The data mining/analysis and knowledge discovery component of LEAD will be built around ADaM and D2K (see section C.3.a). D2K provides the means to develop information visualization techniques and services that will be integrated with other LEAD components. Both ADaM and D2K are built on scalable, modular architectures to facilitate the integration of new mining or analysis components. Further, each system will be enhanced to leverage current and emerging distributed service protocols, Grid services, and/or lightweight peer-to-peer connections. These features will enable loose coupling or linked chaining of algorithms across the two systems in order to adapt to the intense data processing requirements of the LEAD environment, and to address a wider variety of problems than either system alone could solve. The modular architecture of ADaM [45] will allow for integration with spatial visualization packages such as the UAH-developed Space Time Toolkit (STT) [46], MetApps, and OpenGIS compatible map servers.

Data Mining/Analysis Research will address incoming data streams, model results, and information abstracted from them. One type of mining/analysis is feature extraction and phenomena detection using observations and model results. Such mining/analysis can be performed either in real time, when an anomalous or targeted feature is detected in a data stream, or in a later analysis of the data. Descriptions of all such mining/analysis results (features and phenomena detected) will be abstracted into a concept hierarchy [47] and maintained in a spatial data store within the LEAD distributed information space (C.3.e). This abstracted feature information, typically not available today, will be useful for performing trend analysis and knowledge discovery as well as further mining for associations/causal relationships between various phenomena.

### C.3.e. Data Services Environment

Data and information management are key to LEAD. Together with the Distributed Technologies Environment, they form part of the LEAD middleware foundation and support the Local User and Productivity Environments discussed above. Not only must LEAD keep pace with large volumes and multiple types of data, but it also must track data stored in the Grid and Web Services Test Beds; user-generated analyses published to the Data Cloud; and the wealth of information collected about each data item. The distributed information space therefore incorporates the following: real time, historical, and user generated science data (C.3.b); basic content information to catalog data files in the Test Bed for later search and retrieval (discovery metadata; C.3.c); structural and semantic information to enable data usage and exchange (ESML; described at the end of this section); data processing lineage, or historical record, of each data type/file; and additional semantic information about particular meteorological events or significant features of a data file and dynamic personalized indices.

Our approach to the complex problem of data and information management is multifaceted. For data discovery, LEAD will leverage existing and emerging metadata standards where appropriate. For example, THREDDS is implementing the Dublin Core to insure compatibility with STEME (Scientific, Technology, Engineering and Mathematics Education) digital library standards. LEAD will track the DLESE-adopted IMS+, FGDC, ESML, OpenGIS, and ISO TC19005 standards to determine whether to migrate to one of them or provide crosswalks from Dublin Core. In addition, LEAD research will extend these standards or develop new conventions for exchanging new types of information. For example, we will examine conventions used in the GIS communities for describing geographic features (streets, rivers, etc) in order to define descriptions of meteorological features and events. Harvesting systems will be implemented to insure that the resulting search metadata are incorporated into the GCMD and the NSDL collections via DLESE. We also will leverage existing tools generally agreed upon as best practice in the Grid and Web communities, where applicable. The Globus Replica Catalog [48] will be employed to maintain replication consistency among the Test Bed sites.

The distributed information space will extend the research space to *three efforts targeting next generation data and information management*. The first is the Virtual Organization Index (VOIndex), which addresses the problem of efficient querying and timely decision making for unstructured data. (A 'virtual organization' is a group of people related by common interest and not bound by time or organizational boundaries.) Because of timeliness issues in mesoscale meteorology, we propose to use VOIndex as a personalized index associated with a user or a virtual organization. VOIndex can be thought of as a lightweight database with index support for heterogeneous data types. It has an active representation and a compact persistent representation. When called into use, as from the LEAD Portal when an experiment begins, VOIndex instantiates itself as a personal runtime service. When the experiment has concluded, VOIndex collapses into its persistent storage representation. VOIndex achieves heterogeneity through a set of indexes specialized for XML documents (ESML, web pages), temporal data, and geospatial data. This structure is well suited to versioning time sequenced data [49], a key requirement of LEAD. The data themselves are not resident in the VOIndex; instead, the index contains pointers in the form of URLs to files or objects. We additionally will explore efficient index representation for representing the local analysis data and policies for publishing this new information into the LEAD data cloud.

The second effort, dQUOB [50], is a lightweight middleware system for managing temporal data in streams. It has two relevant strengths that derive from its relational database view of real-time data [51,52]. First, data streams can be joined by a simple SQL join query. Applied recursively, the operation aggregates any number of streams of moving data while simultaneously filtering out uninteresting data. Second, dQUOB's ability for rapid decision making over continuous streams of data will be extended and applied to the problem of versioning time sequences and assignment of provenance values. Results can be utilized by VOIndex and by a traditional database management system, the latter for subsequent data mining/analysis.

The final effort involves data interchange technologies using the UAH Earth Science Markup Language (ESML). Interchange technologies are important to LEAD because of the many variations in scale, projection, structure, and format in the data to be used. ESML provides a *standard method for describing* the wide variety of data formats so that applications can use this standard description in reading data. ESML and associated software tools will provide the interchange technology to facilitate search, visualization, analysis and modeling independent of data type or format [53]. *Note that ESML is not a new data format*; rather, it is an XML-based description file generated by data providers or data users, external to the actual data. These data descriptions go far beyond traditional metadata to include both structural and semantic information needed to effect a practical runtime interpretation of a data set [54-55]. Tools such as automated metadata generators and ESML editors will make ESML definition tasks easier and more seamless, greatly facilitating data use. ESML software libraries will be incorporated into LEAD tools, and can be used to build data translation services to format data as needed.

### C.3.f. Distributed Technologies Environment

LEAD will be based upon two levels of distributed systems technology. At the core is *Abilene*, an Internet2 national high-performance research backbone. All LEAD partner sites are connected to Abilene at OC12 (622 Mbps) speed or greater. The NCSA and San Diego Supercomputing Center (SDSC) TeraGrid systems are linked via a 40 gigabit/second network and also connected to Abilene. The Abilene Network Operations Center (NOC) [56], located at Indiana University (a LEAD PI institution), will be working to monitor, measure and optimize LEAD applications. This will be of particular importance in studying weather-related network bottlenecks.

The second level of technology is the backbone of fixed resources within the Linked Grid and Web Services Test Beds (section C.3.b), which will be built using our experience from the first generation of large Grid projects and massive storage devices at the NCSA and SDSC (C.8). We will leverage virtual data management technology being pioneered by the GriPhyN project and, as with GriPhyN, all of our technology will scale to handle large volumes of data and potentially very large numbers of concurrent users. This will require the use of smart virtual data caches as a "built out" of our Distributed Grid and Web Services Test Beds. The GGF, in which LEAD is well represented, has over a dozen working groups actively defining Grid standards. *However, LEAD requires us to go beyond this static Grid infrastructure and confront a second generation of Grid problems that currently are beyond the scope of GGF standards*. The "Linked Environments" in LEAD are much more heterogeneous and dynamic than currently deployed Grids. In many scenarios, LEAD users may be at remote or mobile locations and may be establishing on-the-fly collaborations to respond to a time-critical situation. Such collaborations may rely on non-traditional trust relationships in order to marshal the needed resources, which may reside in vastly different administrative domains. These users will apply the LEAD User Productivity Environment, via the LEAD Portal (C.3.c-d), to build distributed, reliable, high-throughput, real-time applications by composing remote services and remote data sources from a laptop or wireless handheld device. These issues will be explored using the Linked Grid and Web Services Test Beds, in collaboration with researchers at NCSA and SDSC.

Distributed Services and Component Technologies. Addressing the problems described above will require extending current Grid research in several ways. *First*, we will incorporate advances in Web services and Grid technology [57-58,16] that allow data and computational services to be accessed /controlled by simple XML based messaging protocols, and that can adapt the data streams to different, often very limited bandwidths. These technologies are the foundation for the planned OGSA [59,60]. *Second*, we will draw upon peer-to-peer application frameworks and Grid Portals [61] to allow dynamic and mobile collaborations to be constructed rapidly. *Finally*, we will deploy a software component technology that allows users to rapidly compose and control remote LEAD services that use Data Cloud (C.3.b) resources and our Enabling Technologies (C.3.a) [62]. For example, a user could employ the LEAD Portal to access a graphical composition tool to link data mining/analysis components at one site to WRF simulations running on the TeraGrid and live data from another site.

Parallel and Distributed Algorithms. The significant processing requirements for LEAD tools, especially those relying upon real-time data streams, require the use of parallel processing techniques. Thus, effort will be directed toward parallelizing those tools that can be executed using distributed computational resources, ranging from Linux clusters to the TeraGrid. Scaling of these tools for use on the latest high performance computing platforms is a high priority in LEAD. Further, the decomposition of LEAD computing services, science models, and data sets will be studied for compatibility with parallel processing practices and adapted to this paradigm where possible.

## C.4. Proposed Meteorology and Social Sciences/Public Policy Research

We are taking a somewhat novel approach to the non-IT research in LEAD by attacking key problems in mesoscale meteorology that a) have strong societal and/or public policy dimensions; b) depend upon and feed into LEAD technologies; and c) are fully integrated into our education and research plan. Our strategy will be to conduct the physical sciences research needed to develop a deeper understanding of the meteorological aspects of the four areas below, while simultaneously and synergistically studying social and public policy elements.

Adaptive Numerical Weather Analysis and Simulating Modeling: The Evolving Roles of Industry and Academia. Most current operational weather prediction systems throughout the world run on fixed time schedules, and in fixed configurations, regardless of user need or the weather being depicted (exceptions include models of specific threats, such as hurricanes). Using the WRF and ARPS, LEAD will develop the middleware needed to adapt data analysis and simulative modeling computations to the evolving weather. Such capability has profound implications for operational weather forecasting in the United States, where today all forecasts are produced centrally and the public policy infrastructure is built upon an assumption of centralization. As mesoscale models resolve finer-scale features, a distributed prediction scenario may be most effective for the future, where individual weather forecast offices or private entities operate a local model, or control locally nested sub-domains of a model run centrally, in an effort to deal most effectively with weather that is driven by local physical influences (e.g., terrain) and has local impact. Further, because private companies now are beginning to operate their own prediction models, especially for mesoscale weather, the appropriate roles of government and industry are straining the existing legal and political mechanisms of governance. LEAD research in this area will focus on reconciling scientific and technological advances with mechanisms of governance and accountability, with specific emphasis on institutional, legal, administrative, and partnership structures for the provision of weather services.

Deterministic and Ensemble Forecasting: Assessing Use, Value and Risk Using Calibrated Probabilities. The uncertainty associated with specifying forecast model initial conditions is especially acute at the mesoscale, and thus ensemble forecasting (i.e., the generation of several forecasts, all valid for the same time, that are started from slightly different initial conditions, are produced by different models, or utilize different options within the same model) is viewed as essential [63]. However, the generation of initial conditions, and the traditional averaging of ensemble forecast members to generate a mean forecast, present particular challenges at the scale of individual storms. Another especially notable challenge is the verification of fine-scale forecasts and an assessment of their value, and we will investigate the above issues with the ARPS and WRF models, including multi-model ensembles [63] and super-ensembles [64]. Regardless of the outcome of this work, however, even the most accurate forecast is of little value if it is not used properly. To this end, an understanding of information required by decision makers, and effective ways of communicating it, are essential. We will conduct research in this area via our education test beds, feedback from the Unidata user community, and interaction with operational forecasters.

Scale Interaction and Predictability: Assessment of Targeted Investments in the National Technology Infrastructure. Decision makers lack the knowledge (as opposed to information) necessary to prioritize observational programs and plans according to their contributions to science and society [65]. Absent such information, decisions regarding observational infrastructure often are made on an ad hoc or even political basis. One result is that unhealthy competition for scarce resources develops: scientists compete with other scientists (e.g., weather versus climate), research vies with operations (e.g., NASA versus NOAA), and various platform advocates coalesce into warring "tribes" (e.g., satellite versus in situ). Such prioritization is important because the observational system is the backbone of the weather forecasting enterprise. By studying the predictability of mesoscale weather and especially the impact of observing systems via data denial experiments with real and simulated data, we will develop an understanding of which data are important, and why, and link this work to policy frameworks for observing strategies [66].

Storm Feature Climatologies: Storm Threats, Societal Vulnerability and Public Policy Implications. Considerable effort has been invested during the past 10 years to create algorithms that identify weather hazards (e.g., hail, mesocyclones, icing, tornadoes, microburst), in real time, using Doppler weather radar [28-36] and other sensors. Each time a new observing tool is deployed, however, the algorithms (e.g., from WDSS) must be adjusted or new ones created. In contrast, such algorithms applied to assimilated data sets can take advantage of all atmospheric variables and do not need to be changed when new data sources are added. The ubiquity of new fine-scale atmospheric observations (e.g., from Doppler radars) also provides tremendous opportunity to understand mesoscale meteorology through the development of highly detailed storm climatologies, based in part on the identification algorithms. Such information will be of vital importance to urban planners, the insurance industry, weather-sensitive industries, and policy makers. Using the data mining/analysis tools in LEAD, we will compare hazard identification based on raw sensor output with assimilated data sets, and also expand our present work in detailed climatologies (http://mesocyclone.ou.edu), which are based on single algorithms and not data mining.

## C.5. Education Research / Implementation Plan and Collaborative University Test Beds

Partnered as a collaborative education test bed, Millersville University (MU) will coordinate and manage a *three-phase plan* designed to assess the effectiveness of LEAD technologies for education, provide critical input and feedback, and facilitate knowledge transfer to a community of users in concert with the LEAD development methodology (Fig. 3). LEAD will entrain students and their mentors to create the foundation for the development of authentic inquiry-based learning environments that are scalable from the graduate down to pre-college levels.

***The first phase of the LEAD education plan is directed toward establishing education objectives*** that will help shape the evolution and environment of LEAD, and 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 Test Beds (OU, Illinois, UAH, Millersville, and Howard) will engage successful national science and technology education initiatives, such as DLESE, UCAR Windows to the Universe, and AMS Project ATMOSPHERE and DataStreme, to build on best practices that will help steer the development of an education-friendly User Productivity Environment. MU already has strong ties to several of these initiatives. A collaborative network of teacher-partners from pre-college schools (likely to be "thin" clients), each located in proximity to one of the LEAD partner institutions for person-to-person interaction, will be invited and trained to assist in the evaluation of LEAD prototypes. Education Test Beds will initiate the development of tele-collaborative projects using LEAD technology with distinct goals tied to undergraduate and pre-college curricular improvements. In addition, the education plan at Illinois includes enhancements to the NSF-funded Visual Geophysical Exploration Environment (VGEE) [67,68] for use in the LEAD Educational Test Beds, which couple visualization (using MetApps) 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 Test Beds. In parallel, the MU Test Bed 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 in the LEAD education plan will focus 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***. Participating institutions have identified several courses where LEAD capabilities will be integrated to drive substantive change in meteorology and computer science, and to other disciplines because of its inherent extensibility. MU will assume a lead role for the integration of tools and services into undergraduate and pre-college curricula. In select courses, students will work in small groups in authentic inquiry-based learning environments, experiencing first-hand evolving weather events and their societal impacts in real-time (e.g., via web stories for major news organizations). 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. MU is experienced in developing and implementing workshops for in-service teachers, and will follow this path for bringing LEAD capabilities to the greater pre-college community.

## C.6. Strategies for Increasing the Involvement of Traditionally Underrepresented Groups

Of the 16 PIs in LEAD, 5 are women, 11 are men, and 1 of the men is African American. Traditionally underrepresented groups will be included in the research and educational elements via several mechanisms. First, the 3 WRF workshops to be coordinated by Howard University in years 3-5 will involve 10-15 students, half from Howard and the other half drawn from minority serving institutions (MSIs). Second, graduate and undergraduate student recruiting will leverage existing linkages to the LEAD participating institutions, including Clark-Atlanta (3+2 program with OU), REUs (which themselves emphasize the issue), SOARS (operated by NCAR), and Howard's programs with Jackson State. The two doctoral students at Howard most likely will be from underrepresented groups. Third, the Educational Test Beds coordinated by Millersville University will recruit heavily from MSIs, including Hispanic and Tribal Colleges, and select pre-college teacher partners from a diverse demographic to participate in the development of the LEAD Portal and assessment of LEAD technologies. Finally, the Unidata community of 93 educational institutions, which represents the core user base for LEAD, includes 14 MSIs (3 of which are HBCUs) and several accredited minority postsecondary institutions, all of which will be targeted for early involvement in LEAD curriculum development.

## C.7. User Integration, Operational Support, and Expected Impacts: The Unidata Community and Technology Dissemination for National Impact

The individuals associated with LEAD are involved in numerous major initiatives (C.11) that have a direct pipeline to or impact on operations. Consequently, LEAD will have substantial national visibility and input from a variety of communities, the largest and most important of which is the Unidata academic community, which directly impacts *tens of thousands of students each year*. The success of Unidata IDD clearly demonstrates the meteorological community's willingness and ability to take software developed centrally, and independently implement and support its use in a broad range of departments. To facilitate technology dissemination, LEAD will rely on Unidata's extensive mechanisms for training and user support. As additional LEAD technologies are implemented, the Unidata community will recommend which to incorporate into the supported Unidata collection. However, the impact of LEAD is likely to extend *well beyond the academic community*, just as the current IDD/LDM has been adopted by private companies and operational government centers. This extended community of users almost certainly will be interested in adopting LEAD technology as it evolves, and in supporting its continued operation – particularly in the context of homeland security, which for tracking biological and chemical materials requires fine-scale atmospheric information *immediately*. Finally, LEAD will play a vital role in the development of the community-wide WRF model by creating capabilities that are critical to its use as a research and educational tool, and that ultimately will be important for its operational implementation. Several LEAD senior personnel are involved in WRF or WRF-related activities (Droegemeier, Xue, Wilhelmson, Joseph, Gannon).

## C.8. Differentiation From and Synergy with Other Projects

Although several initiatives in Earth sciences data now are underway, all differ significantly from LEAD. For example, the *Earth System Modeling Framework* (http://www.esmf.ucar.edu/) is developing tools to enhance ease of use, performance, portability, interoperability, and reuse in climate, numerical weather prediction, and data assimilation applications. It emphasizes code development and maintenance tools and not data transport, mining/analysis, or distributed computing. The *National Climatic Data Center* is the major long-term archive for meteorological data. Although it does not provide the types of tools, data interchange technologies, distributed services, and accessibility proposed by LEAD, it is highly complementary to LEAD and will link with it via CONDUIT and CRAFT. The *GriPhyN Project* has many objectives similar to those of LEAD, though is focused on the physics community and does not address many of the issues critical to meteorology (e.g., real time, on-demand). In addition, GriPhyN projects, such as Atlas, are less heterogeneous than the potential LEAD user base, which numbers in the tens of thousands and spans more 130 institutions including community colleges and a spectrum of university departments (e.g., oceanography, geosciences, hydrology, meteorology and physics). Considerable synergy exists between LEAD and the *TeraGrid* effort at NCSA and SDSC, the latter of which will provide high performance facilities and expertise in intelligent data management. NCSA also will be the focal point for integrating the LEAD Environments and Grid-enabled LEAD software for execution on both the TeraGrid and other emerging Grids. Letters from NCSA and SDSC committing to these collaborations are contained in the Supplementary Materials.

## C.9. <u>Division of Funding, Responsibilities, and Personnel</u>

*Table 1. Division of funding and work among all participating institutions as well as the total number of personnel to be involved over the 5 year period (e.g., if 2 graduate student lines are budgeted over 5 years, 4 to 5 students likely will be funded owning to graduation and are so counted in the table).*

| Institution | Principal Responsibilities | Budget | % of Budget | #Faculty or Senior Invest. | #Post Docs or Research Staff | #Graduate Students | #Under Grad Students | #Technical Staff |
|---|---|---|---|---|---|---|---|---|
| **University of Oklahoma** | Lead Institution. Project leadership and management; meteorology research; Grid and Web Services Test Bed; Educational Test Bed | **$2,830,853** | 19.4% | 3 | 4 | 7 | 0 | 1 |
| **Howard University** | Meteorology research; Educational Test Bed | **$697,789** | 4.8% | 2 | 1 | 2 (+10 workshop travel) | 1 | 1 |
| **University of Alabama in Huntsville** | Data mining, interchange technology, distributed web services; Grid and Web Services Test Bed site; Educational Test Bed | **$2,774,082** | 19.0% | 3 | 3 | 7 | 1 | 1 |
| **Millersville University** | Education Coordinator; Education Test Bed | **$669,554** | 4.6% | 2 | 0 | 0 | 8 | 2 |
| **Indiana University** | Web and Open Grid services, QOS, security, and data storage research | **$2,061,735** | 14.1% | 2 | 1 | 8 | 0 | 1 |
| **University of Illinois** | User interface, data mining, data storage, and meteorology research; Grid and Web Services Test Bed; Educational Test Bed | **$2,450,834** | 16.8% | 2 | 3 | 5 | 0 | 4 |
| **UCAR Unidata** | Data distribution and user interface research; Grid and Web Services Test Bed | **$2,172,146** | 14.8% | 2 | 2 | 0 | 3 | 3 |
| **University of Colorado** | Public policy and societal impacts research | **$956,001** | 6.5% | 1 | 4 | 3 | 5 | 1 |
| TOTALS | | **$14,612,994** | **100.0%** | **17** | **18** | **32** | **18** | **14** |

## C.10. <u>Results from Prior NSF Support</u>

**Kelvin K. Droegemeier** (ATM-9220009) "Science and Technology Center for the Analysis and Prediction of Storms" developed the Advanced Regional Prediction System (ARPS), which was the world's first mesoscale model designed for highly parallel computers. It also developed techniques for assimilating Doppler radar and other fine-scale data into storm-resolving models, demonstrated the practical predictability of thunderstorms and related mesoscale phenomena, and studied numerous aspects of convective storm dynamics.

**Ming Xue** (ATM-9909077) “A New Joint Weather Research and Prediction (WRF) Model” includes developing 3D variational techniques for small-scale (unbalanced) flows with emphasis on the use of Doppler radar data, evaluating case studies for model verification, exploring optimal assimilation/forecast strategies, calibration and initialization of soil models, and developing and testing the dynamic and computational frameworks of WRF.

**Sara Graves** (ANI-9729627) "A Testbed for a Multi-network Infrastructure for Research", in collaboration with Alabama A&M University (an HBCU) and the Alabama Research and Education Network (AREN), focused on the development of a state-wide high-performance networking infrastructure for research and education and a high-performance connection to Abilene, an Internet2 network.

**Sepideh Yalda and Richard Clark** (ATM-9909711) “Workstations to Support Interactive Learning and Research in Atmospheric Science Using Unidata Applications” enabled the acquisition of workstations to complete the hardware configuration of the Synoptic/Dynamics Lab to facilitate full utilization of Unidata applications in several upper level meteorology and computer applications (Fortran, IDL) courses.

**Everett Joseph** (ATM-9909190) "Upgrading a Training and Research Lab for CSTEA and HUPAS". This support enabled the acquisition of a UNIX workstation to serve as the local host for the UNIDATA LDM. The workstation was added to a laboratory being developed as facility for student research and instructional applications in several areas of atmospheric sciences.

**Dennis Gannon** (EIA-9975020) “GrADS: Grid Advanced Development Systems” and related projects have focused on parallel computation, object oriented methods for distributed scientific applications, and problem solving environments for scientific computation. Work in Grid applications resulted in an award for heterogeneous distributed computation as part of the I-Way project at SC98 which was the first real Grid applications test bed.

**Beth Plale** (EIA-9973834) "Applying Database Techniques to Management of Large Data Flows in Scientific Applications" explored the abstraction of data streams as database in the context of scientific, wide area parallel and distributed applications. The major outcome was successful performance of the prototype and discovery that certain well-established query optimization heuristics yield less than optimal queries in a data stream environment.

**Mohan Ramamurthy** (ATM-9730985) "Mesoscale Ensemble Forecasting and Predictability Studies.” Investigated the use of ensemble forecasting for mesoscale phenomena and developed a comprehensive multi-model, multi-analysis ensemble prediction system which includes (MM5, WRF, RSM and Eta) and analyses from multiple data assimilation systems (NCEP/Eta, NCEP/AVN, ECMWF, NOGAPS and GEMS).

**Robert Wilhelmson and Mohan Ramamurthy** (DUE-9972491). Along with other investigators developed a Visual Geophysical Exploration Environment (VGEE) that offers a framework for technology-mediated, inquiry-based approach to help students connect theoretical and abstract understanding to real world meteorological phenomena using state-of-the-art simulations, visualizations, interactive java tools, and tutorials.

**Ben Domenico, Don Murray, Anne Wilson** (ATM-0130792) “Unidata 2003” Implemented significant improvements to and expanded LDM, incorporated new data streams into IDD, and developed a new set of platform-independent data analysis and display tools (MetApps) that access data from a wide variety of distributed data sources. This work established the foundation for the NSDL THREDDS project.

**Roger Pielke, Jr**. (SBR-9708983) "Use and Misuse of Prediction in the Earth Sciences” investigated the role of prediction in formulating environmental policies including planning for and the response to natural and anthropogenic hazards; managing natural resources; and regulating environmental impacts. It convened several workshops to assemble a diverse group of people involved in prediction, and one outcome of the grant was the publication of a book titled “Prediction: Science, Decision-Making and the Future of Nature (Island Press 2000).”

## C.11. Project Management

The governance of a project as large and physically distributed as LEAD requires a balance between the bureaucratic structure needed to manage a complex effort, and the flexibility that is essential in an academic setting for accommodating new ideas or technological developments. As the director of a former NSF Science and Technology Center (which remains in operation), the PI of the lead institution (K. Droegemeier), who will serve as the overall Project Director for LEAD, has extensive experience formulating and leading multi-institutional, multi-disciplinary collaborative projects. The management plan for LEAD reflects this experience and has been formulated to ensure that: a) the tactical implementation plan developed from the funded proposal is based upon the very latest information, represents a consensus among all senior investigators, and can be modified with time as the project proceeds; b) project goals and strategies are clearly articulated, understood and most importantly accepted by all participants, particularly students, so they have a clear understanding of how their work fits into the overall LEAD concept; c) progress throughout the entire project is monitored on a continuous basis so that problems can be identified and addressed quickly, especially those regarding inadequate performance or the need to change directions or re-allocate resources; d) input from end users, especially those associated with the education components, feeds back into the research process; and e) directions for and results from LEAD are scrutinized by experts, with their input an integral part of the management process.

It is important in any large team project to address the extent to which investigators are working together or have done so in the past. This issue bears on the management plan for obvious reasons, and the figure below shows that the LEAD investigators are involved with numerous and significant mutual collaborations, in most cases on large projects that underpin LEAD (e.g., THREDDS, CRAFT, WRF, MetApps, CONDUIT, VGEE, IDD). Consequently, these collaborations not only will help ensure effective project management, but also clearly demonstrate the considerable leveraging that is associated with LEAD.

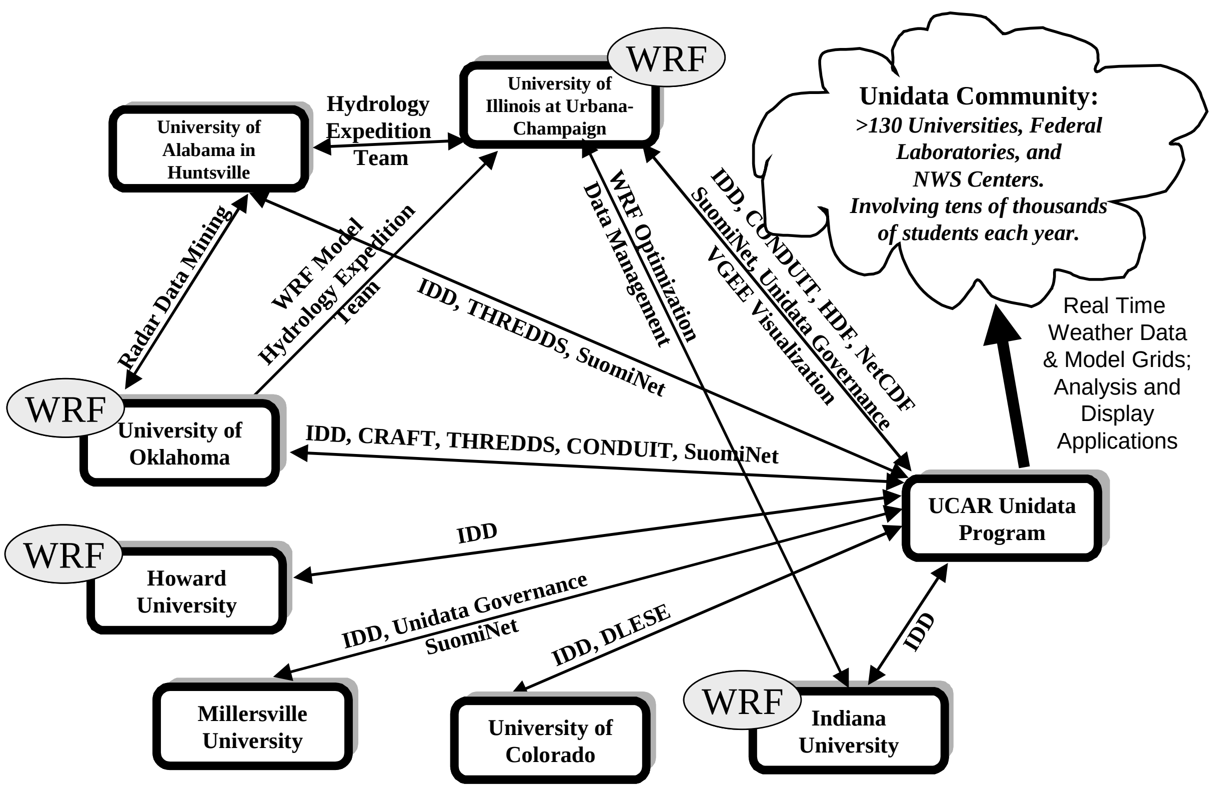


*Figure 5. Existing collaborations among LEAD participants.*

*Project Director* K. Droegemeier (Figure 6) will have overall responsibility for ensuring the success of LEAD and will serve as its principal point of contact with the NSF. He also will orchestrate LEAD's overall scientific direction, enforce the highest standards in ethics and research quality, monitor overall progress relative to established milestones (Table 2), promote the dissemination of results, and work to ensure that information regarding LEAD is being communicated as broadly as possible. He will be assisted by a full-time *Project Coordinator*, also located at OU, who will manage the day-to-day operations of the entire LEAD collaboration. The associated responsibilities include but are not be limited to ensuring effective communication and sharing of results among all participants; coordinating meetings; collecting and assembling information for project reports; managing the LEAD web site; and most importantly, monitoring progress compared to milestones and identifying problems. This individual *will be a scientist* who also has strong skills in organization management. Droegemeier hired such a person several years ago as the assistant director of CAPS and has a long history of working with and mentoring such support staff. Every effort will be made to draw this hire from traditionally underrepresented groups.

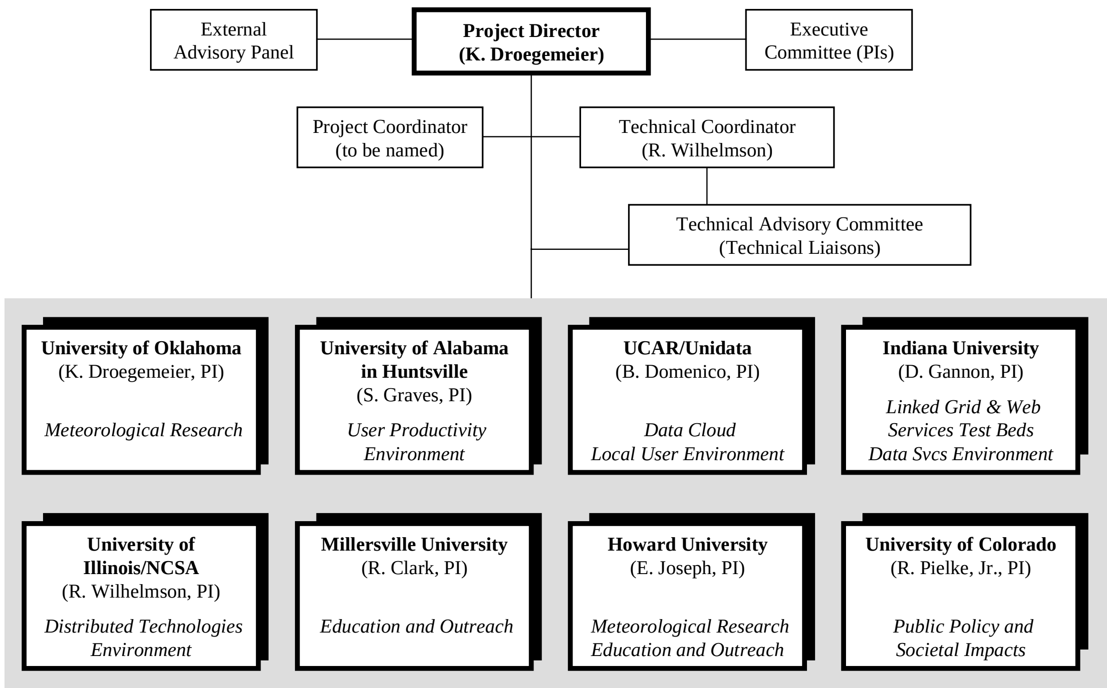


*Figure 6. Proposed LEAD management structure. The scientific thrust areas for which each institution has scientific oversight are shown. Note, however, that each institution will work on multiple areas, as described throughout the narrative.*

Each of the principal scientific thrust areas in LEAD will be overseen by one of the local (i.e., institutional) PIs (Figure 6): Data Cloud and Local User Environment (Domenico); Linked Grid and Web Services Test Beds as well as Data Services Environment (Gannon); User Productivity Environment (Graves); Distributed Technologies Environment (Wilhelmson); Meteorological Research (Droegemeier); Education and Outreach (Clark and Joseph); and Public Policy and Societal Impacts (Pielke). In this capacity they will have responsibility for helping set research directions, ensure timely progress, and identify problems and new opportunities. A *Technical Liaison*, drawn from among each institution's scientific participants, will assist with these duties. Collectively, the Liaisons will compose the *Technical Advisory Committee*, which will be chaired by R. Wilhelmson (NCSA) as the project *Technical Coordinator*. In this role he will oversee technical development, testing, and deployment. This position is particularly important in light of NCSA's commitment to various LEAD activities, as noted in the letter from NCSA Director D. Reed, and of its involvement in overall Grid integration.

*Table 2. Selected Yearly Project Milestones.*

| Research Component | Year 1 | Year 2 | Year 3 | Year 4 | Year 5 |
|---|---|---|---|---|---|
| **Local User Environment and LEAD Portal** | Initial portal and collaboration tools | Data search and access tools | Query capability for concept hierarchy | User access to service chaining | Full access to LEAD environment and tools |
| **User Productivity Environment** | Initial version of mining algorithms; data preparation design | Grid and Web services for mining; initial data tools | Mining concept hierarchy; framework for customization | Parallel distributed mining; MetApps linked to Web services | Mining for causal relationships and decision support |
| **Data Cloud Grid and Web Services Test Beds** | Test Beds; NNTP testing; product groups; metadata generation | Implement NNTP; metadata delivery to DLESE and GCMD | Integrate IDD/ THREDDS & Web services | Integrate IDD/ THREDDS, & Grid services | Integrate Test Beds into other sites; refine |
| **Data Services Environment** | ESML for initial data, integration; deploy Open Grid Services | Define concept hierarchy, mods to ESML, integrate OGSI | Metadata generator, data interchange; Web services testing | Full ESML integration with applications | Complete ESML integration; deploy web services |
| **Distributed Technologies Environment** | Test Beds; job launch, high performance file transfer; Web services | Metadata and archive tools, real time peer collaboration protocols | Grid services, advanced flow management and monitoring | Service chaining; rapid flow management control; Web services | Deploy Web services with LEAD Portal |
| **Education and Outreach** | Objectives and Test Beds; pre-college teacher-partner sites | Develop tele-collaborative projects & learning materials | Deploy proto-technologies; incorporate metrics | Deploy applications; integrate tools, organize short courses/wkshops | Disseminate LEAD to education and Unidata communities |
| **Meteorological Research** | Globus in WRF/ARPS; ensembles; scale interaction; mining | Mining of assimilated data sets; prediction testing/ensembles | Real time distributed prediction; ensemble verification | Continue real time prediction testing and assimilating mining | Full dynamic operation of WRF and relational mining |
| **Public Policy and Societal Impacts Research** | Web impacts and policy outreach backbone; Technology assessmt | Technology assessmt in Test Beds; design transfer methods | Apply prescriptive and descriptive methods for tool evaluation | Reconcile supply and demand for LEAD products and services | Full implementation of information value/use assessment |

The 8 PIs in LEAD will serve on an *Executive Committee (EC)* to advise the Director on all facets of the project (Figure 6). Chaired by the Director, it will meet monthly, via the Access Grid and desktop video conferencing, and will hold two face-to-face meetings per year. Every effort will be made to leverage PI attendance at national meetings (e.g., Supercomputing) so that most of the LEAD travel funding can be used for research interactions. Individual project researchers will collaborate regularly via the Access Grid and desktop video conferencing, and sufficient travel support has been requested to accommodate several face-to-face collaborations per year as well as attendance at conferences. The Technical Coordinator will facilitate monthly virtual meetings of the Technical Liaisons, and once each year, an “all hands” meeting will be held at one of the participating sites. Finally, an *External Advisory Panel (EAP)*, consisting of 5 to 6 experts drawn from areas relevant to LEAD, will meet yearly to advise the Director and EC on all aspects of the project (travel for these meetings has been requested in the form of participant support costs).

During the past 2 decades, Unidata has been the focal point for delivering meteorological and related information to the academic community. To effectively integrate LEAD into this community and obtain feedback from it, the LEAD senior leadership will continue its long-standing close interaction with the national Unidata governance structure. At the present time, LEAD PIs play key roles in all 3 Unidata governance bodies: Policy Committee (M. Ramamurthy, Member), User Committee (R. Clark, Chair), and THREDDS Technical Task Force (B. Domenico, Chair). In addition, close ties will be maintained between LEAD and NSDL (headquartered at UCAR and led by former Unidata Director D. Fulker), and DLESE (Program Center headquartered at UCAR, with LEAD Co-PI R. Wilhelmson serving on its Steering Committee).

**C.12. References Cited**


[1] Molinari, J., 1993: An overview of cumulus parameterizations in mesoscale models. In The Representation of Cumulus Convection in Numerical Models, *Meteor. Monograph*, **46**, Amer. Meteor. Soc., 155-158.

[2] Droegemeier, K.K., 1997: The numerical prediction of thunderstorms: Challenges, potential benefits, and results from realtime operational tests. *WMO Bulletin*, **46**, 324-336.

[3] Weygandt, S.S., A. Shapiro and K.K. Droegemeier, 2002: Retrieval of initial forecast fields from single-Doppler observations of a supercell thunderstorm. Part II: Thermodynamic retrieval and numerical prediction. *Mon. Wea. Rev*, **130**, 454-476.

[4] Xue, M., D.-H. Wang, J.-D. Gao, K. Brewster, and K. K. Droegemeier, 2002: The Advanced Regional Prediction System (ARPS), storm-scale numerical weather prediction and data assimilation. In press for *Meteor. and Atmos. Physics*.

[5] Carpenter, R.L. Jr., K.K. Droegemeier, G.M. Bassett, S.S. Weygandt, D.E. Jahn, S. Stevenson, W. Qualley, and R. Strasser, 1999: Storm-scale numerical weather prediction for commercial and military aviation, Part 1: Results from operational tests in 1998. *Preprints*, 8th Conf. on Aviation, Range, and Aerospace Meteorology, 10-15 January, Amer. Meteor. Soc., Dallas, TX, 209-211.

[6] Wang, D., K.K. Droegemeier, D. Jahn, K.-M. Xu, M. Xue, and J. Zhang, 2001: NIDS-based intermittent diabatic assimilation and application to storm-scale numerical weather prediction. *Preprints*, 14th Conf. on Num. Wea. Pred., 30 July - 2 August, Amer. Meteor. Soc., Fort Lauderdale, Florida, J125-J128.

[7] Yoo, H.-D., K.K. Droegemeier, K. Brewster, S.-Y. Lee, and C.-H. Cho, 2002: Impact of radar data assimilation on the Chorwon-Yonchon 1996 heavy rainfall event: Preliminary results. *Preprints*, 3rd Joint Korea-US Workshop on Storm- and Meso-Scale Weather Analysis and Prediction, 21-22 February, Boulder, CO, 157-163.

[8] Weisman, M.L., W.C. Skamarock, and J.B. Klemp, 1997: The resolution dependence of explicitly modeled convective storms. *Mon. Wea. Rev*., **125**, 527-458.

[9] Sun, J. and N.A. Crook, 1998: Dynamical and microphysical retrieval from Doppler radar observations using a cloud model and its adjoint. Part II: Retrieval experiments of an observed Florida convective storm. *J. Atmos. Sci*., **55**, 835–852.

[10] Belair, S. and J. Mailhot, 2001: Impact of horizontal resolution on the numerical simulation of a midlatitude squall line: Implicit versus explicit condensation. *Mon. Wea. Rev*., **129**, 2362-2376.

[11] Mass, Clifford F., David Ovens, Ken Westrick, Brian A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? *Bull. Amer. Meteor. Soc*., **83**, 407–430.

[12] LeDimet, F.X. and O. Talagrand, 1986: Variational algorithms for analysis and assimilation of meteorological observations: Theoretical aspects. *Tellus*, **38A**, 97-110.

[13] Sun, J. and N.A. Crook, 1997: Dynamical and microphysical retrieval from Doppler radar observations using a cloud model and its adjoint. Part I: Model development and simulated data experiments. *J. Atmos. Sci*., **54**, 1642-1661.

[14] Gao, J., M. Xue, K.K. Droegemeier, and A. Shapiro, 2001: A 3-D variational method for single-Doppler velocity retrieval applied to a supercell storm case. *Preprints*, 30th Conf. on Radar Meteorology, 19-25 July, Amer. Meteor. Soc., Munich, Germany, 456-458.

[15] Gao, J., M. Xue, K. Brewster, F. Carr, and K.K. Droegemeier, 2001: A three-dimensional variational data assimilation scheme for a storm-scale model. *Preprints*, 14th Conf. on Num. Wea. Pred., 30 July - 2 August, Amer. Meteor. Soc., Fort Lauderdale, Florida, J72-J74.

[16] Foster, I. and C. Kesselman, 1999: *The Grid: Blueprint for a new computing infrastructure*. Morgan Kauffman, San Francisco, 677pp.

[17] Vaughan-Nichols S. J., 2002: Web Services: Beyond the Hype. *IEEE Computer*, February.

[18] Roy J., and A. Ramanujan, 2002: Understanding Web Services. *IEEE IT Proceedings*, November/December.
[19] McIlraith S. A., T. Cao Son and H. Zeng, 2001: Semantic Web Services. *IEEE Intelligent Systems*, March/April.
[20] Predonzani P., A. Sillitti and T. Vernazza, 2001: Components and Data Flow Applied to the Integration of Web Services. IECON 01, The 27th Annual Conference of the IEEE Industrial Electronics Society.
[21] 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.
[22] 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.
[23] Chiswell, S.R., L. Miller, and P. Hirschburg, 1999: Toward expanding model data access through Unidata's Internet Data Distribution System. Paper 13.10. *Preprints*, 15th Int. Conf. on Interactive Information and Processing Systems for Meteorology, Dallas, TX, Amer. Meteor. Soc.
[24] 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.
[25] 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.
[26] 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.
[27] Murray, D. and C. Murphy, 2001: Community software design: The Unidata Java-based Metapps project. Proc., *17th International Conference on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology*, Amer. Meteor. Soc., January, Albuquerque, New Mexico.
[29] 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.
[29] 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.
[30] 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.
[31] Eilts, M.D. and Co-Authors, 1996: Severe weather warning decision support system. *Preprints*, 18$^{th}$ Conf. on Severe Local Storms, San Francisco, CA, Amer. Meteor. Soc., 536-540.
[32] Johnson, J. T. and Co-Authors, 1998: The storm cell identification and tracking algorithm: An enhanced WSR-88D algorithm. *Wea. and Forecasting*, **13**, 263-276.

[33] Witt, A. and Co-Authors, 1998: An enhanced hail detection algorithm for the WSR-88D. *Wea. and Forecasting*, **13**, 286-303.

[34] Stumpf, G. and Co-Authors, 1998: The National Severe Storms Laboratory mesocyclone detection algorithm for the WSR-88D. *Wea. and Forecasting*, **13**, 304-326.

[35] Mitchell, E.D. and Co-Authors, 1998: The National Severe Storms Laboratory tornado detection algorithm. *Wea. and Forecasting*, **13**, 352-366.

[36] Fulton, R.A. and Co-Authors, 1998: The WSR-88D rainfall algorithm. *Wea. and Forecasting*, **13**, 377-395.

[37] Hinke, T., J. Rushing, S. Heggere, S. Ranganath, and S.J. Graves, 1997: Target-independent data mining for scientific data: Capturing transients and trends for phenomena mining. Proc, 3rd Int. Conf. On Data Mining (KDD-97), Newport Beach, CA, August 14-17.

[38] Rushing, J., H. Ranganath, T. Hinke, S. J. Graves, "Using Association Rules as Texture Features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. **23**, No. 8, pp. 845-858, August 2001.

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

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

[41] Hinke, Thomas, J. Novotny, 2000: Data Mining on NASA's Information Power Grid. *Proceedings*, Ninth IEEE International Symposium on High Performance Distributed Computing, Pittsburgh, Pennsylvania, August 1-4.

[42] Shaw, M., C. Subramanian, G.W. Tan, M. Welge. Knowledge Management and Data Mining for Marketing. *Decision Support Science*, April, 1998.

[43] Welge, M., W. H. Hsu, L. S. Auvil, C. Bushell, J. Martirano, T. M. Redman, and D. Tcheng. Data to Knowledge (D2K): A Rapid Application Development Environment for Knowledge Discovery in Databases. Technical Report, National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, 1999.

[44] Baru, C., R. Moore, a. Rajasekar, and M. Wan, 1998: The SDSC Storage Resource Broker. *Proc*. CASCON'98 Conference , Toronto, Canada, Nov.30-Dec.3,,

[45] Ramachandran, R., H. Conover, S. J. Graves, K Keiser, S. Movva, and S. Tanner, 2001: Flexible Earth Science Data Mining System Architecture. Fourth Workshop on Mining Scientific Datasets, Seventh ACM SigKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, August 26-29.

[46] Botts, M., K. Keiser, H. Conover, S. J. Graves, 2000: Visualization of On-demand Virtual Data Products in a Distributed Environment. North American Cartographic Information Society Maps and the Internet Symposium, Knoxville, TN, October 11.

[47] He, Yubin, R. Ramachandran, U. S. Nair, K. Keiser, H. Conover, S. Graves, 2002: Earth Science Data Mining and Knowledge Discovery Framework. SIAM International Conference on Data Mining, April 11-13, Arlington, VA.

[48] Allcock, B., J. Bester, J. Bresnahan, A. Chervenak, I. Foster, C. Kesselman, S. Meder, V. Nefedova, D. Quesnel and S. Tuecke, 2001: Secure, Efficient Data Transport and Replica Management for High-Performance Data-Intensive Computing. IEEE Mass Storage Conference.

[49] Plale, B., 2001: Performance Impact of Streaming Doppler Radar Data on a Geospatial Visualization System, College of Computing, Georgia Institute of Technology, Atlanta, Georgia, GIT-CC-01-07.

[50] Plale, B, and K. Schwan, 2000: dQUOB: Managing Large Data Flows Using Dynamic Embedded Queries. *Proc*. 9th IEEE Intl. High Performance Distributed Computing (HPDC).

[51] Babu, S. and J. Widom, 2001: Continuous Queries over Data Streams. International Conference on Management of Data (SIGMOD).

[52] Terry, D., D. Goldberg, D. Nichols and B. Oki, 1992: Continuous Queries over Append-Only Databases. International Conference on Management of Data (SIGMOD).
[53] Ramachandran. R., H. Conover, S. Graves, and K. Keiser, 2002: Interchange Technology for applications to facilitate generic access to heterogenous data formats. IGARSS, Toronto, Canada, June 24-28.
[54] Ramachandran, R. M. Alshayeb, B. Beaumont, H. Conover, S. Graves, X. Li, S. Movva, A. McDowell and M. Smith, 2001: Earth Science Markup Language: A Solution for Generic Access to Heterogeneous Data Sets. NASA Earth Science Technology Conference 2001, August 28-30.
[55] Ramachandran R., M. Alshayeb, B. Beaumont, H. Conover, S. Graves, N. Hanish, X. Li, S. Movva, A. McDowell and M. Smith, 2001: Earth Science Markup Language. 17th Conference on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology, 81st AMS Annual Meeting, Albuquerque, NM, January.
[56] Indiana University, Global Network Operations Center (Global NOC), http://globalnoc.iu.edu.
[57] Johnston, W., D. Gannon, B. Nitzberg, A. Woo, B. Thigpen, L. Tanner, 2000: Computing and Data Grids for Science and Engineering." Proceedings of SC2000 IEEE, Dallas, November
[58] Foster, I., Kesselman, C., Tuecke S., 2001: The Anatomy of the Grid: Enabling Scalable Virtual Organizations. *International J. Supercomputer Applications*, **15**(3).
[59] Foster, I, Kesselman, C, Nick, J. Tuecke, S, 2002: The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration. Globus Research Papers, January.
[60] Tuecke, S. Czajkowski, K., Foster, I, Frey, J., Graham, C, Kesselman, C., 2002: Grid Service Specification. Globus Research Papers, February.
[61] Krishnan, S., Bramley, R., Govindaraju, M., Indurkar, R., Slominski, A., Gannon, D., Alameda, J., Alkaire, D, 2001: The XCAT Science Portal. Proceedings SC2001, Denver, November.
[62] Gannon, D., Bramley, R., Fox, G., Smallen S., Rossi, A., Ananthakrishnan R., Bertrand, F., Chiu, K., Farrellee, M., Govindaraju, M., Krishnan S., Ramakrishnan, L., Simmhan, Y., Slominski, A., Ma, Y., Olariu, C., Rey-Cenvaz N., 2001: Programming the Grid: Distributed Software Components, P2P and Grid Web Services for Scientific Applications, To appear *Journal of Cluster Computing*.
[63] Hou, D., E. Kalnay, and K.K. Droegemeier, 2001: Objective verification of the SAMEX '98 ensemble forecasts. *Mon. Wea. Rev*., **129**, 73-91.
[64] Krishnamurty, T.N. and Co-Authors, 1999: Improved weather and seasonal climate forecasts from multimodel superensemble. Science, 285, 1548-1550.
[65] Pielke, Jr., R. A., and R. Carbone, 2002: Weather Forecasts, Impacts and Policy: An Integrated Perspective. *Bull. of the Amer. Met. Soc*, **83**, 393-403.
[66] Emanuel, K. and Co-Authors, 1996: Observations in aid of numerical weather prediction for North America. Report of Prospectus Development Team #7, US Weather Research Program.
[67] 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.
[68] 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.