The LEAD Roadmap: an Integrative Framework for Utilizing and Sharing Community Resources to Enhance Understanding and Impact of Weather[1]

Public Version

February 21, 2009

Table of Contents

Introduction

Linked Environments for Atmospheric Discovery (LEAD; Plale et al. 2004, 2006; Droegemeier et al. 2005, 2007; Droegemeier 2008) is a National Science Foundation (NSF) Large Information Technology Research (ITR) grant that has pioneered new approaches for integrating complex weather data, assimilation, modeling, mining, and cyberinfrastructure systems. LEAD empowers researchers and students with capabilities heretofore available at only a few major universities and research or operational centers around the world, and it does so using a service-oriented architecture similar in many respects to the familiar Amazon.com where storage and compute resources and services are "in the cloud". By managing the complexity of inter-operative cyber tools and providing flexibility and ease in how they can be linked, LEAD allows users to focus their time on solving the science and engineering problems at hand, providing a means for more deeply understanding the tools and techniques being applied rather than the nuances of data formats, communication protocols, and job execution environments.

Foundational to LEAD is the notion that today's static environments for observing, predicting, and educating about mesoscale weather are fundamentally inconsistent with the manner in which such weather actually occurs, namely, with often unpredictably rapid onset and evolution, heterogeneity, and spatial and temporal intermittency. To mitigate this inconsistency, LEAD has created an integrated, scalable framework in which meteorological analysis tools, forecast models, and data repositories can operate as ***dynamically adaptive, on-demand, grid-enabled systems***, thus allowing them to a) change configuration rapidly and automatically in response to weather; b) respond to decision-driven inputs from users; c) initiate other processes automatically; and with the potential to d) steer remote observing technologies to optimize data collection for the problem at hand.

The execution of analysis tools and forecast models requires powerful compute resources on which to run. While some researchers have access to ample compute resources, resources in the departments of academic institutions are often limited and university computational resources are often batch scheduled so they lack responsiveness to the dynamic nature of weather. LEAD has overcome these problems by working closely with the TeraGrid project, an NSF funded nationally distributed high performance computational and network resource for scientific

---

[1] Primary contact is Beth Plale, Indiana University, plale@cs.indiana.edu , 812 855 4373.

investigation. LEAD pioneered a simple means by which researchers can use the TeraGrid without having to solve the complex setup required, particularly with respect to security mechanisms. Further, LEAD has eliminated the need for the individual scientist to write a proposal for compute cycles (an allocation proposal) to justify their work, making rapid start-up and smaller scale use possible. LEAD has also engaged in collaboration with other computer science projects to solve the problem of allowing high priority jobs to execute immediately on the TeraGrid, eliminating delays in job scheduling queues where an hour or longer wait is not uncommon. For initiatives like "warn on forecast" where time is of an essence, reduced job queue delay is critical. Capability such as is described above falls under the banner of virtual organizations. ***LEAD has pioneered virtual organizations for the weather community*** through a suite of tools that can be reused for the creation and maintenance of other virtual organizations as well. Furthermore, providing such powerful tools in an easy-to-learn environment to smaller universities advances the democratization of national computing resources.

LEAD has made advancements in reproducibility of atmospheric science research outcomes through provenance collection. The provenance of a data product is a record of its lineage, or trace of the execution history that resulted in the product. The provenance of a forecast model result, for instance, captures information about the executable version of the model, the configuration parameters, the input data products, the execution environment, and the owner. Provenance enables data to be properly attributed and captures critical parameters about the model run so the quality of the result can be ascertained. Proper provenance is essential to providing *reproducible* scientific computing results.

In the last 6 years, LEAD has successfully introduced and demonstrated a new paradigm for weather research. In the process it has built a research framework and undertook substantial engagement of the weather research and education community given the balanced research/engineering nature of the effort. The feedback from user engagement has been encouraging, as users talk of significantly reduced setup time, and easy access to data and numerical weather prediction tools not available elsewhere. This has enabled students, such as Howard University masters student Hector Lima to engage in much richer and deeper scientific discovery in meteorological course work and research. Mr. Lima used the LEAD Portal's to combine data and results from WRF simulations, which enabled him to study the development of gravity waves events within the vicinity of the Appalachian Mountains, and was able to investigate limitations of the model to forecast these events.

The LEAD PIs over the last 6 months have undertaken an extensive examination of fundamental contributions of the project, and more importantly, an effort to identify the unique value it provides to the atmospheric science community, and to the broader scientific community as the nation grapples with economic uncertainty and a recognition of the need for more creative solutions to energy sources and the environment. As a result of several meetings including two face-to-face meetings in Washington, DC and Indianapolis Fall 2008, we have developed this roadmap document as a way of establishing an intended future path forward for LEAD and to help engage the community in a discussion of LEAD's strengths and weaknesses. ***The scenario we envision for the future is LEAD's deployment as a long-term community resource.*** This is an ambitious goal. However, the resource would be funded by multiple organizations; would be made openly available; and would support research, education, and operational experimentation including, adaptive observations and the emerging concept of "warn on forecast" (e.g., Burgess et al. 2005; Stumpf et al. 2007; Kuhlman et al. 2008).

Related Computational Frameworks

The weather and climate research and education communities have several computational research frameworks that either exist or are under consideration. A discussion of them here is a means to clarify LEAD's contribution with respect to other projects that might be more familiar to different audiences. The NCAR Data Assimilation Research Section (DAReS), part of the Institute for Mathematics Applied to Geosciences (IMAGe), provides centralized data assimilation expertise which can be coordinated with existing observational and modeling expertise. Plans include hosting a research facility to run assimilations, and providing cloud-computing access to the facility and resources. LEAD could contribute expertise and tools to this effort as it has already built such a facility, one that utilizes and hides the details of running in a cloud environment (in LEAD's case it is the Teragrid.) Unlike DAReS, LEAD provides support for end-to-end modeling, analysis and visualization workflows. However, LEAD is an integrative framework, and it is entirely feasible that an assimilation step carried out in a LEAD workflow could actually be a call out to execute an assimilation located on the DAReS facility.

The NCAR Earth System Modeling Framework (ESMF) is a software framework for building and coupling weather, climate, and related models. Researchers who use ESMF must write couplers that code how data is to be transferred between components. ESMF couplers can be quite complex data structures that for instance, map the data from *n* parallel tasks in model *P1* to *m* parallel tasks in model *P2*. As LEAD has recognized, ESMF likewise recognizes the importance of provenance collection and is building support for collection and representation into the framework. LEAD and ESMF are complementary in focus. The Developmental Testbed Center (DTC) has brought together computational resources, a software framework, and organizational structure for purposes of speeding the transition of research developments into the operational setting. It accomplishes this through experiments involving operational forecasters and academics in the investigation of forecast model research. Bernardet et al. (Bernardet et al. 2008) describe a recent comparison of the WRF model, ARW core to WRF/ NMM core in an end-to-end forecast system. LEAD could bring to DTC a workflow execution capability that may have precluded the need in Bernardet et al. to write workflow execution scripts from scratch. In order for LEAD to host a study as was carried out by DTC, LEAD as a community resource would have to examine its policies and assess the broader interest in the end-to-end pieces of a DTC study. Too, the 2008 DTC study required significant dedicated computational resources that would likely have had to have been negotiated with Teragrid outside the shared LEAD compute cycle allocation.

The WRF Portal (Govett and Smith 2008), an effort of the Joint NWP Test Bed (JNT), of the Developmental Test Bed Center at NCAR began not long after LEAD. From its inception, LEAD has maintained close communication with JNT. In the early stages of LEAD, it became clear that the DTC was taking a different approach to modeling with a much shorter time horizon than LEAD, focusing on developing a desktop-client WRF configuration tool that executes WRF on a dedicated cluster. As time progressed and both the LEAD and JNT efforts matured, conversations were restarted regarding possible strategies for integrating key elements from these two important projects. A more thorough technical analysis recently was conducted regarding integrating the JNT WRF configuration client into the LEAD infrastructure, and NSF has provided supplemental funding to support it. This fusion of portal technologies will allow JNT users to gain full benefit from the many aspects of LEAD that are not included in the WRF Portal (Droegemeier et al. 2009), such as access to shared national infrastructure computing resources and multiply nested grids, and vice versa.

Roadmap for Future

The LEAD roadmap builds from a record of strong success in both research and outreach. Advancements in weather research have appeared in published literature and been discussed at community conferences and meetings. Advancements have additionally been made available to the community as advanced algorithms available for use through LEAD. Substantial computer science research has been carried out in building the LEAD cyberinfrastructure which has contributed fundamental ideas to the sub-disciplines of e-Science and cyberinfrastructure through publications. More importantly, the LEAD CI as a working platform in use by domain science researchers has itself been platform in which extensive computer science experimental evaluations have and continue to be carried out. Outreach, which is described in more detail below, has not only put LEAD in front of researchers, such as at WRF workshops, but has resulted in pedagogic advances to teaching atmospheric science and utilizing real time data in the classroom.

The roadmap for LEAD focuses the system as an ***integrative framework that enables components owned, managed, and possibly even executed outside the framework, to be easily included in the framework.*** An extremely high level view of LEAD is one having two layers: a science layer and a cyberinfrastructure layer. As shown in Figure 1, the science layer includes the observational and forecast model data collections; analysis, data mining, and modeling and verification tools; and the visualization tools that enable researchers and educators to carry out their work. The cyberinfrastructure layer at the bottom of the figure is a middleware layer between the user tools and the remote computational and storage resources (i.e., TeraGrid); that further provides value add on top of base capability for constructing workflows and collecting and managing provenance and data. The vision of LEAD as an integrative framework has impact on both the science and the cyberinfrastructure layers, where, as depicted by the arrows, contributions from the outside community can be made at all layers and functionalities, from analysis codes to new data collections to new cyberinfrastructure pieces. Early evidence of the benefit of such a modular approach is in the uptake and reuse of components in the cyberinfrastructure layer. Many are, for instance, available through the open source Open Grid Computing Environments (OGCE) (Alameda et al. 2007) repository. Components are additionally being deployed to extend the GridChem Science Gateway. Specifically, the LEAD workflow suite is being used to enable support for coupled simulations. LEAD cyberinfrastructure tools are also being used in an Indiana University chemical informatics drug discovery application used to study tumors (Dong et al. 2007).

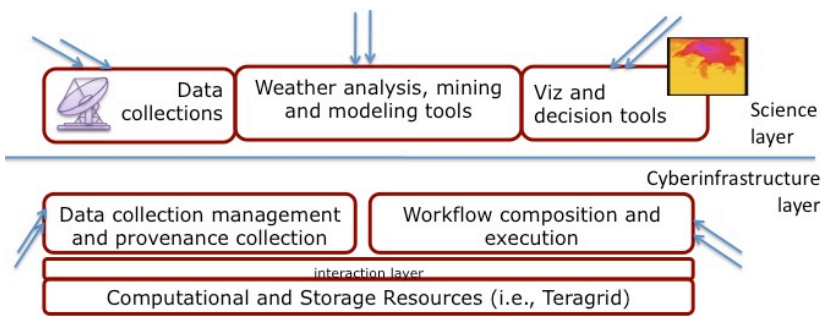

Figure 1. High level view of LEAD as a science layer toolbox for atmospheric scientists, and a cyberinfrastructure layer underlying and facilitating the research work. The notion of LEAD as an integrative framework impacts all aspects of the architecture.

Through roadmap process, LEAD has identified several target audiences. We envision building from a base of successes we have seen so far in reaching out to ***WRF users***, as evidenced in part by recent outreach activities listed in Table 1. The goals of LEAD bear a strong similarity to the emerging concept of “warn on forecast”, and we expect to have inroads to ***the warn on forecast community*** as evidenced by our participation in the annual Storm Prediction Center/NSSL Spring Program in the Hazardous Weather Testbed (Brewster et al., 2008). In the classroom the framework could be used in a range of courses from introductory to upper level undergraduate and graduate content courses, such as numerical modeling, dynamics, and mesoscale meteorology courses. LEAD would have particular value for ***undergraduate and graduate independent research projects***. Students should gain early research experiences by using LEAD support for Campus Weather Service as early as Feb 2009.

Other potential roles for LEAD include on-demand, tailored numerical forecasting for guiding crew and instrument deployment in ***field programs*** (discussions now are underway to demonstrate this during VORTEX2 in the Spring of 2009), and we see science value in LEAD continuing to reach out for use ***where weather forecasts and data contribute to non-meteorological science***, such as crop modeling research by Western Michigan University Scientists as part of a project funded by USDA. Several months ago, LEAD investigators were contacted by FAA about their interest in the LEAD cyberinfrastructure for use in the next generation air traffic control 4D

Data Cube. The FAA has said the LEAD service oriented architecture and data system were exactly what they envision for the 4D Data Cube, that no other system today, to their knowledge, is similar to LEAD. Work plan development is in its early stages, but is underway.

Community Engagement
Conducted LEAD workshop at *2008 AMS Annual Meeting*
Hazardous Weather Testbed
Demonstrated evidence of value of LEAD in real graduate and undergraduate research projects (see Year 5 Annual report)
Engaged in *Spring 2007 Weather Challenge*
Completed LEAD Education Usability Study in collaboration with UCAR (Dr. Lynne Davis)
NOAA Spring Experiment
90-min tutorial at annual *WRF Workshop*, June 27, 2008
VOTEX-2 (planned)

Table 1. Recent LEAD community engagement and outreach

Short-Term Efforts

The ***short-tem efforts along the roadmap*** are to redouble outreach efforts through the remaining period of the grant; build in support for the WRF 3.1 tool suite including WPS and WRF3D VAR and other tools; increase support for user-deployed data sets; and harden the software infrastructure. Current and planned tool and data set support is shown in Tables 2-5. By way of background, however, we give a brief introduction to user interaction with LEAD through the portal.

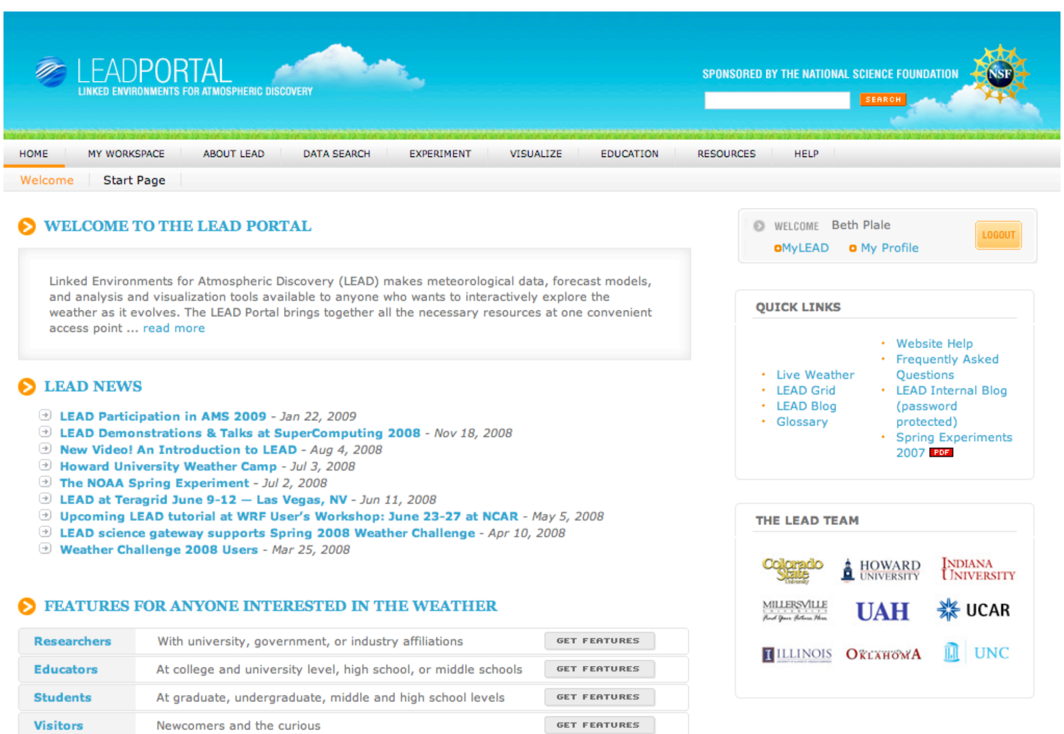

Figure 2. Meteorology researchers/educators interact with the LEAD framework by logging in to the LEAD portal, and selecting from amongst the options in the main menu bar across the top of the page.

A researcher/educator interacts with the LEAD framework through the LEAD portal, as shown in Figure 2. Upon logging in, the scientist/educator can navigate though the portal by clicking on the tabs on the main menu. Key capabilities are shown across the main menu including ‘My Workspace’ which provides remote persistent storage for user’s modeling and analysis activities; ‘Data Search’ which lets users search over a dozen or so collections of meteorologically relevant surface and upper air observational data and model data. The ‘Experiment’ tab takes the user to a wizard that walks the user through setting up analysis and model runs. Users can use canned runs (i.e., execute a WRF forecast run using University of Oklahoma ADAS 10km Analysis Data for Initial Conditions and NCEP NAM 40Km for setting Lateral Boundary interpolated by ARPS Preprocessing v5.2.9 and Forecast with WRF (ARW Core) v2.2.) The ‘Visualize’ tab describes the visualization support available in LEAD, which is the Unidata Integrated Data Viewer (IDV). The ‘Education’ tab takes the user through guided modules targeting exploratory education in upper high school and undergraduate education. The modules include live interaction with the full portal functionality in a controlled setting by allowing students to query for real time data, and pull this data into IDV for instant viewing of current weather conditions.

The value of the framework to the community is directly proportional to the tools and data it provides. Tables 2 and 3 list the tools that are currently supported for assimilation, modeling, analysis, mining, and visualization. Table 4 lists the data collections that can be accessed and downloaded. The tools can be linked in different ways, and can use the most recent observational or model data products available. As such, LEAD provides a means for more deeply understanding the tools and techniques being applied rather than the nuances of data formats, communication protocols, and job execution environments.

Currently Supported Assimilation and Modeling Tools	**Description**
WRF model v2.2	ARW core
ARPS2WRF v5.2.9	ARPS to WRF format conversion
ADAS Analysis v5.2.9	Interpolate ADAS data to user selected domain of the 3 dimensional ARPS Forecast Grid
WRFstatic v5.2.9	Generate static surface data needed by WRF
Arpstrn v5.2.9	Extract subset of terrain data from static terrain files
ext2arps v5.2.9	NAM initialization file format conversion to ARPS
Arpsintrp v5.2.9	ADAS initialization file format conversion to ARPS
Ext2arps v5.2.9	Interpolate external model data and extract lateral boundary files
arps2wrf v5.2.9	Format conversation ARPS to WRF
Unidata IDV	Latest version

Table 2. Currently supported suite of forecasting and analysis tools.

Currently Supported Data Mining Tools	**Description**
ITSC Storm Detection Algorithm (SDA)	Detect storm pattern from WSR88D-II reflectivity field.
ITSC Spatial Clustering Algorithm	Uses output from SDA to cluster storms into region.
ADaM 4.0	Suite of 140+ data mining algorithms (pattern recognition, image processing, clustering and shape detection) available as LEAD services.

Table 3. Currently supported suite of data mining tools

Dataset	# days
NEXRAD Level II	183 days
UPC Case Studies – Hurricane Katrina	--
Upper Air Observations (Rawinsondes)	600 days
UPC Steered Regional WRF model grids	130 days
Station data (METAR, LAPS)	600 days
NOAA Profiler data	600 days
OU ADAS 10 km CONUS	45 days
NAM 40 km CONUS	300 days
NAM 80 km CONUS	300 days

Table 4. Data sets currently indexed for use analysis, modeling, data mining, and assimilation tasks.

The ***roadmap targets early alignment of LEAD services to the broader WRF audience.*** These are shown in Table 5 with expected availability date of late Spring 2009. LEAD additionally targets integration of the LEAD/NCSA Parametric Workflow Engine as a service call from the LEAD workflow engine; this should bring capability being used in research and experimental forecasting (see http://rt.atmos.uiuc.edu/trigger/ for current forecasts being triggered in response to changing weather conditions) to a broader community.

To be added components:	
WRF model v3.1	ARW core. Upgrade with March NCAR release.
WRF WPS v3.1	March NCAR release.
WRF 3D VAR v3.1	With modifications to include MADIS_WRF-Var interface
MADIS	NOAA/ESRL Meteorological Assimilation Data Ingest System
WRF Portal interoperability	Interoperability with JNT WRF Portal
LEAD/NCSA Parametric Workflow Engine (PWE)	Incorporate as a service call from the LEAD workflow engine; investigate web-start of Rich Client GUI
ARPS Verif	Verification of models compared to observations
User-imported data sets	Researcher adds their data catalog into system

Table 5. Additional tools scheduled for addition to LEAD with an expected availability of Summer 2009.

Long term Efforts

The outcomes and impacts realized to date by LEAD provide a strong foundation and high level of confidence upon which to establish a ***persistent LEAD framework for integrating weather/climate research models and data into end-use knowledge***. We seek partnerships to continue the provisioning of the cyberinfrastructure for use in the atmospheric sciences, particularly targeting the emerging "warn on forecast" concept that is being pioneered by NOAA groups in Norman, Oklahoma, and that has involved LEAD as part of the yearly Hazardous Weather Test Bed experiments (Brewster et al. 2008). LEAD's extensibility makes it ideally suited to respond to the emerging need for mesoscale information in important economic sectors including energy security, public health and safety, transportation, water resources, and food production (NRC 2008). We have an existing partnership with Western Michigan University in food production, and are interested in partnering with others in this exciting space.

LEAD anticipates continued use of TeraGrid for its large-scale computational resource needs and the Indiana University Data Capacitor and mass store system for large-scale data storage through the end of the TeraGrid effort. As of this document, TeraGrid is undergoing a reframing as the XD program. We will work with the primary parties during the XD planning phase to convey LEAD's unique requirements and benefits.

References

Alameda, J. et al., 2007: The Open Grid Computing Environments collaboration: portlets and services for science gateways, *Concurrency and Computation Practice and Experience* 19 (6), 921.

Bernardet et al., 2008: The Developmental Testbed Center and its Winter Forecasting Experiment, *Bulletin of the American Meteorological Society*, pp. 611-627.

Brewster, K.A., D.B. Weber, S. Marru, K.W. Thomas, D. Gannon, K. Droegemeier, J. Alameda, S. J. Weiss, 2008: On-Demand Severe Weather Forecasts Using TeraGrid via the LEAD Portal. *Proceedings TeraGrid08*, Las Vegas, NV.

Burgess, D. W., D. C. Dowell, L. J. Wicker, A. Witt, 2005: Detailed comparison of observed and modeled tornadogenesis. Preprints, *32nd Conf. on Radar Meteor.,* Albuquerque, NM, Amer. Meteor. Soc

Dong, X., K.E. Gilbert, R. Guha, R. Heiland, J. Kim, M.E. Pierce, G.C. Fox, and D.J. Wild. 2007: Web Service Infrastructure for Chemoinformatics. *J. Chem. Inf. Model*, 47(4):1303–1307

Droegemeier, K.K. and Co-Authors, 2005: Service-oriented environments in research and education for dynamically interacting with mesoscale weather. *Computing in Science and Engineering*, **7**, 12-29.

Droegemeier, K.K., and Co-Authors, 2007: A new paradigm for mesoscale meteorology: Grid and web service-oriented research and education in LEAD. Preprints, *23rd Conf. On Integrated Information and Processing*, 15-18 January, San Antonio, TX, Amer. Meteor. Soc.

Droegemeier, K.K., 2008: Transforming the Sensing and Numerical Prediction of High Impact Local Weather Through Dynamic Adaptation. *Phil. Trans. of the Royal Soc*. A. 1-20.

Droegemeier, K.K., B. Plale, M. Ramamurthy, and C. Mattocks, 2009: A new approach for using web services, grids, and virtual organizations in mesoscale meteorological research, Preprints, *25th Conf. on Integrated Information and Processing*, Phoenix, AZ, Amer. Meteor. Soc.

Govett, M.W. and J.S. Smith, 2008: A public release of the WRF Portal: A graphical user interface for WRF. Preprints, *24th Conf. On Integrated Information and Processing*, 21-24 January, New Orleans, LA, Amer. Meteor. Soc.

Kuhlman, K.M., T.M. Smith, G.J. Stumpf, K.L. Ortega, and K.L. Manross, 2008: Experimental probabilistic hazard information in practice: Results from the 2008 EWP spring program. Preprints, *24th Conf. on Severe Local Storms*, Savannah. GA, Amer. Meteor. Soc., Paper 12.2.

National Research Council (NRC) 2008: Observing Weather and Climate from the Ground Up: A Nationwide Network of Networks, National Academies Press, Washington DC.

Plale, B., J. Alameda, R. Wilhelmson, D. Gannon, S. Hampton, A. Rossi, and K.K. Droegemeier, 2004: User-oriented active management of scientific data with my LEAD. *IEEE Internet Computing*, **9**, 27-34.

Plale, B., D. Gannon, J. Brotzge, K.K. Droegemeier and Co-Authors, 2006: CASA and LEAD: Adaptive cyberinfrastructure for real-time multiscale weather forecasting. *IEEE Computer*, **39**, 66-74.

Stumpf, G.J., T.M. Smith and K.A. Scharfenberg, 2007: The future of US severe weather warning operations. Preprints, *4th European Conf. on Severe Storms*, 10-14 September, Trieste, Italy, 2 pp. (Available at http://www.essl.org/ECSS/2007/abs/06-Forecasts/1177965431.stumpf-1-sec06.oral.pdf.)

Appendix A. Roadmap Participants

Jay Alameda, NCSA, University of Illinois Urbana Champaign

Keith Brewster, University of Oklahoma

Rich Clark, Millersville University

Kelvin Droegemeier, University of Oklahoma

Sara Graves, ITSC, University of Alabama Huntsville

Everette Joseph, Howard University

Gopi Kandaswamy, RENCI, University of North Carolina

Suresh Marru, Indiana University

Beth Plale, Indiana University

Rahul Ramachandran, ITSC, University of Alabama Huntsville

Mohan Ramamurthy, Unidata

Bob Wilhelmson, NCSA, University of Illinois Urbana Champaign

Sepi Yalda, Millersville University