Vision of Cyberinfrastructure for End-to-End Environmental Explorations (C4E4)

R. S. Govindaraju[1]; B. Engel[2]; D. Ebert[3]; B. Fossum[4]; M. Huber[5]; C. Jafvert[6]; S. Kumar[7]; V. Merwade[8]; D. Niyogi[9]; L. Oliver[10]; S. Prabhakar[11]; G. Rochon[12]; C. Song[13]; and L. Zhao[14]

Abstract: Holistic approaches are needed for understanding and addressing a wide range of environmental issues that require multidisciplinary studies of complex and interlocking systems. The writers' vision of a cyberinfrastructure for end-to-end environmental exploration (C4E4) that combines data and modeling tools in an integrated environment across different spatial and temporal scales is presented. The overall goal behind C4E4 is to enable a broad environmental research and remediation community to address the challenges of environmental data management and integration in real-world settings. The St. Joseph Watershed in northern Indiana is chosen as a test bed in this effort. The C4E4 framework will allow researchers to combine heterogeneous data resources with state-of-the-art modeling and visualization tools through a user-friendly web portal. By engaging TeraGrid resources, C4E4 will have the computational resources to store, manipulate, and query large data sets, thereby facilitating new science. C4E4 will serve as a prototype, and provide valuable experience for scaling up to larger observatories at the national level. This paper presents the writers' vision and goals, initial efforts, and briefly describes how C4E4 can benefit the environmental community.

DOI: 10.1061/(ASCE)1084-0699(2009)14:1(53)

CE Database subject headings: Internet; Hydrology; Environmental engineering; Monitoring; Databases; Information management.

Introduction

The quality of our land, air, and water resources is under unprecedented pressures as a result of human activity. Many current vital questions in environmental sciences cannot be answered without conducting comprehensive studies based on data from various sources in hydrologic, atmospheric, agricultural sciences, and other related disciplines. As a result, an urgent need exists for the design and development of an enabling data infrastructure that helps integrate various data sources and tools, and provides easy access to researchers from multiple research communities. According to the National Science Foundation-(NSF) sponsored report on cyberinfrastructure (CI):

"Environmental research and education are characterized by a number of attributes that make cyberinfrastructure especially important for this field of scientific endeavor. Many environmental research activities are observationally oriented, rely on the integration and analysis of many kinds of data, and are highly collaborative and interdisciplinary. Much of the relevant data needs to be geospatially indexed and referenced, and there is a host of currently noninteroperable data formats and data manipulation approaches. Spatial scales vary from microns to thousands of kilometers; time scales range from microseconds (for some fast photochemical reactions) to centuries or millennia (for paleoclimate and Earth evolution studies); and data types range from written records and physical samples to long-term instrumental data or simulation model outputs." (NCAR 2003)

This paper presents an approach adopted by a group of inves-

[1]Professor, School of Civil Engineering, Purdue Univ., West Lafayette, IN 47907 (corresponding author). E-mail govind@ecn.purdue.edu

[2]Professor and Head, Dept. of Agricultural and Biological Engineering, Purdue Univ., West Lafayette, IN 47907.

[3]Professor, School of Electrical and Computer Engineering, Purdue Univ., West Lafayette, IN 47907.

[4]Managing Director, Discovery Park Cyber Center, Purdue Univ., West Lafayette, IN 47907.

[5]Associate Professor, Dept. of Earth and Atmospheric Sciences, Purdue Univ., West Lafayette, IN 47907.

[6]Professor, School of Civil Engineering, Purdue Univ., West Lafayette, IN 47907.

[7]Graduate Student, School of Civil Engineering, Purdue Univ., West Lafayette, IN 47907.

[8]Assistant Professor, School of Civil Engineering, Purdue Univ., West Lafayette, IN 47907.

[9]Assistant Professor of Regional Climatology, Indiana State Climatologist, Depts. of Agronomy and Earth and Atmospheric Sciences, Purdue Univ., West Lafayette, IN 47907.

[10]Managing Director, Discovery Park Center for the Environment, Purdue Univ., West Lafayette, IN 47907.

[11]Associate Professor, Dept. of Computer Science, Purdue Univ., West Lafayette, IN 47907.

[12]Associate Vice-President, Collaborative Research and Engagement; Director, Purdue Terrestrial Observatory; Chief Scientist, Rosen Center for Advanced Computing, Purdue Univ., West Lafayette, IN 47907.

[13]Senior Research Scientist, Rosen Center for Advanced Computing, Purdue Univ., West Lafayette, IN 47907.

[14]Research Scientist, Rosen Center for Advanced Computing, Purdue Univ., West Lafayette, IN 47907.

Note. Discussion open until June 1, 2009. Separate discussions must be submitted for individual papers. The manuscript for this paper was submitted for review and possible publication on June 13, 2007; approved on April 9, 2008. This paper is part of the ***Journal of Hydrologic Engineering***, Vol. 14, No. 1, January 1, 2009. ©ASCE, ISSN 1084-0699/2009/1-53–64/$25.00.

tigators, mostly at Purdue University, who are developing a prototype for a generalizable CI for environmental research and teaching purposes. This approach is called C4E4, which stands for *Cyberinfrastructure for End-to-End Environmental Explorations*. The writers' group includes a diverse mix of computer scientists, environmental engineers, hydrologists, atmospheric scientists, and education specialists. The plan brings together resources and expertise from different disciplines, and is aimed to engage the participation of representatives of many more areas and specialties than those in the group alone. The overall objective of C4E4 is to garner capabilities and tools already built and tested as part of other community efforts. These ongoing efforts include: (1) the successful NanoHUB (www.nanohub.org), which enables the nanoscience and nanotechnology communities to direct novel research and disseminates audiovisual lectures, demonstrations, and interactive teaching modules (see the Appendix); and (2) current developments in data engineering such as distributed storage, cataloging, metadata management, data transfer, data mining, and data fusion.

Background and Motivation for Building C4E4

In December 1999, a chemical spill of dimethyl dithiocarbonate caused a wastewater treatment plant in Anderson, Ind. to malfunction. The spill reached the White River, where toxic by-products (thiram, carbon disulphide, dimethylamine) formed, and over a period of a few weeks killed 80,000 fish in an 80 km stretch of river from Anderson to Indianapolis. The lingering effects of these toxins on populations of mussels, invertebrates, birds, mammals, and other wildlife in the area have yet to be assessed.

Events like these are intermittent and unpredictable, but they reoccur often enough to be a continual source of environmental degradation. As a result, streams in the Midwest, including Indiana, regularly fail to meet water quality standards with respect to nutrients, pesticides, suspended solids, pathogens (*E. coli*), PCBs, mercury, cyanide, dissolved oxygen, pH, and ammonia (USEPA 2000). Agriculture is the primary land use in most of the midwestern river basins, and artificial (tiled) drainage systems alter the patterns and mixing of runoff and groundwater. In the upper Midwest, pesticide contamination from the St. Joseph River (Clendon and Beaty 1987; Holtschlag and Nicholas 1998) and similar basins in Indiana also reach the Great Lakes and the Mississippi River (Goolsby et al. 1999, 2001; USGS 2000).

The degradation of air quality also ranks high in the catalogue of environmental impacts, and atmospheric deposition plays a pivotal role in determining the water and watershed quality of the Great Lakes region. Chemicals and particles deposited over land or water may have short-term effects on regional ecosystems and long-term effects on regional climate. In a world undergoing global climate change, the analysis of integrated complex interactions between the general circulation of the atmosphere, the biosphere, and downstream atmospheric chemistry and its related aerosol physics is difficult. As midlatitude temperatures rise, and droughts become more extreme, natural biomass burning will have an even more profound effect on regional air quality and climate in the heavily populated industrial regions of North America. From the southern Plains states to the Atlantic coast, increased black carbon, ozone, and other gaseous constituents will interact with the industrially generated sulfate aerosols to modify cloud albedos and modulate regional climate and air quality.

To address such a broad range of environmental issues, the construction of complex multimodel systems based on advanced software design principles is viewed as a first step. Examples of such systems include the Department of Energy's Dynamic Information Architecture System (http://www.dis.anl.gov/DIAS/) and the large family of coupled ocean-atmosphere models (http://www.pcmdi.llnl.gov/projects/cmip/index.php). The meteorological community continues to develop a suite of climate models with flux couplers that manage synchronous model execution and carry out realistic exchanges among atmosphere, sea-ice, land, and oceans. For example, the Rio Grande Coupled Model Project at Los Alamos National Laboratory included a regional atmospheric, land-surface hydrology model, an operational dynamic river wave/channel model, and a subsurface finite-element groundwater model. The Army Corps of Engineers uses similar suites of models at the Waterways Experimental Station. Community modeling efforts such as the Regional Atmospheric Modeling System, the Weather Research Forecasting (WRF) system, the Community Climate System Model, and the Community Multiscale Air Quality have been under development as large multiscale, multimedia efforts.

However, despite these activities, few systems (if any) have been systematically constructed to simulate multiscale coupled geospatial-ecological systems. Some of the components of real environmental observational frameworks lag behind in scope and scale (e.g., weather observation, water monitoring), and often lack such elements as georeferencing, metadata conventions, semantic sophistication, and even types of metadata that normally accompany sets of observations in most experimental environments. Consequently, a host of environmental problems have defied holistic solutions. The overall goal of building the C4E4 is to create a system that will enable a broad environmental research and remediation community to address the challenges of environmental data management and integration of existing and newly recovered research data into real-world applications. As a prototype, we expect to demonstrate fusion of data and models over the St. Joseph Watershed in northern Indiana.

C4E4 will allow researchers to combine heterogeneous data resources with state-of-the-art modeling and visualization tools. It will offer opportunities to address land, air, and water quality problems that are regionally important and that are significant to society. Environmental events occur and interact at numerous spatial and temporal scales. The monitoring, prediction, and regulation of adverse effects require the intelligent combination of existing data with proven and novel analysis, visualization, and experimental design methods.

Cyberinfrastructure Attributes

In order to foster broad participation, the CI will have several desirable attributes. For easy data discovery, C4E4 plans to support access to a suite of physical, hydrological, and ecophysiological observational data, as well as tools from heterogeneous environment domains. Moreover, it will support the scalable integration of interdisciplinary data to drive agricultural, pollution, health, economic, and political models. These models will strive to quantify the effects of human-induced changes in land use and urbanization, economic growth and consumption, and interactions of ecosystems and human health. Model outputs will be piped into other models or interpreted with decision-making and geospatially referenced database tools. Currently, several models (see Table 1) are already made available through individual web sites. Through a single portal, C4E4 will allow for easy prepara-

Table 1. A List of Candidate Models That Would Be Made Available through the C4E4 Portal

Model name	Brief description and URL
NAPRA WWW	Estimates impacts of agricultural management systems on surface and subsurface hydrology and water quality, and to identify location-specific environmentally friendly agricultural management practices (http://danpatch.ecn.purdue.edu/~napra)
L-THIA WWW	Estimates impacts of land use changes on hydrology and water quality (http://www.ecn.purdue.edu/runoff/lthianew)
ROMIN WWW	Provides assistance with environment-friendly land use planning (http://danpatch.ecn.purdue.edu/~romin)
SEDSPEC	Assists in analyzing runoff and erosion problem by determining the peak rate of runoff from the area and providing information about different types of runoff and erosion control structures (http://danpatch.ecn.purdue.edu/~sedspec)
WATERGEN	Provides online watershed delineation tool and serves as an interface for various spatial decision support systems (http://danpatch.ecn.purdue.edu/~watergen)
WHAT	Provides automated baseflow separation and a hydrograph analysis tool to complement the USGS daily streamflow web site with a Web GIS interface (http://danpatch.ecn.purdue.edu/~what)

Note: Adapted from Engel et al. (2007).

tion of input data files and launching of one or several of these models. It is expected that new models, specifically the Soil Water Assessment Tool (SWAT), will be included in this list (more in the section entitled "Initial Efforts").

C4E4 will enable a variety of users to set up scientific workflows combining the full suite of process and impact models with an array of static or streaming data sets. Users will be able to focus on scientific problems of interest free of data set heterogeneity and data access complexity problems. C4E4's links, portals, and underlying access systems will be designed to serve a large community of users over a long period of time with a minimal need for back-end support.

C4E4 will demonstrate its end-to-end capability via its simulations of the physical aspects of regional ecosystems. It will serve as an advanced form of "science gateway," connecting researchers worldwide via resources such as the TeraGrid and the Open Science Grid to data and computational resources worldwide. As a gateway to advanced CI, C4E4 will deepen and broaden scientific understanding across many disciplines. As a scalable, generic CI solution, it will serve as a template for future learning communities and online research. A recent NSF-sponsored workshop articulated a clear and urgent need for such a facility (NSF 2006).

The C4E4 framework will draw on the existing strengths of researchers, practitioners, and stakeholders in environmental science and engineering. These diverse sources of expertise will contribute to the development, mentoring, applications, and learning opportunities for numerous disciplines. C4E4 will ideally begin with local and regional data in need of study at multiple levels. The goal, however, will be to demonstrate, via these studies, the tools that can accept data from and guide experimental design within several of the national environmental observatories that are expected to be designed and deployed in the future.

Existing Data Infrastructure

C4E4 will build upon the achievements and developments of other studies that have yielded major data sets, and also draw from previous and current cyberinfrastructure projects already available within the community. For example, Purdue University is designated by the National Weather Service to receive real-time, nationwide WSR-88D (Weather Surveillance Radar 88 Doppler) data CERN (European Organization for Nuclear Research). The University is also a U.S.-Climate Monitoring System Tier 2 site in collaboration with CERN and Fermilab. These data resources are augmented by real-time multisensor satellite data (i.e., MODIS Terra & Aqua, AVHRR, GOES, and Feng Yun) and a wide array of near-real-time data products generated on a Linux Cluster, provided by the Purdue Terrestrial Observatory (PTO-http://www.itap.purdue.edu/pto/), as well as archival data amassed by Purdue's Laboratory for Applications of Remote Sensing (http://www.lars.purdue.edu/index.html), over the past 40 years, the Indiana statewide 0.6–1.0 m spatial resolution airborne LIDAR topographical reconnaissance missions, the USGS sponsored AmericaView and IndianaView Consortium spatial data holdings, the 0.3 m and 0.15 m leaf-on and leaf-off state orthophoto overflight data and the NASA Socio-Economic Data Applications Center archives.

The Office of the Indiana State Chemist (OISC) is responsible for pesticides and nutrient regulation within Indiana, including the monitoring of these substances in the state's waters. Numerous data have been collected by the OISC in support of their activities. Currently these data are not well organized; rather, they are in various spreadsheets and reports. All of these data will be made available to C4E4. They will be combined with databases on emerging contaminants, including some recently released data on veterinary and human antibiotics, prescription and nonprescription drugs, polycyclic aromatic hydrocarbons, hormones, and gasoline additives.

Ongoing and legacy studies have resulted in extensive data sets for the St. Joseph Watershed. These include static data in the form of Geographical Information Systems (GIS) data layers, including soil characteristics and topography. Similarly, soils, environmental, water quality, and hydrological data for this watershed have been compiled in previous studies. Twelve automated ISCO water quality sampling stations on seven drainage channels, two field-scale watersheds, two surface drainage systems in upper Cedar Creek subwatershed are currently collecting real-time meteorological, soil moisture, and water quality data. Five real-time, web-accessible weather, soil moisture/temperature, and streamflow stations are providing real-time information over the St. Joseph River Watershed (SJRW) study region. Rainfall is measured at all water quality stations. Further details are available at http://www.ars.usda.gov/research/projects/projects.htm?accn_no=411515.

These data, along with the socioeconomic, and other state and local legacy data will be available as temporal and geospatial layers to facilitate vulnerability assessment, hindcasting, nowcasting, data mining, data fusion, and generation of alternative future scenarios for decision support.

C4E4 Structure and Description

The following text discusses the infrastructure challenges for building the proposed C4E4 system and outline the writers' approach to these questions. Target user communities include

academic researchers investigating environmental scientific questions, farmers and agricultural groups, state and federal environmental data and quality monitoring agencies, and emergency responders.

C4E4 Framework Overview

Although a number of approaches are possible, stakeholder requirements focus the writers' perspective on developing the CI. In keeping with this priority, a regional ecosystem or ecosystems with a history of past and ongoing data collection efforts is first targeted. The next task will be to bring the system or systems "online," that is, C4E4 will begin with data streams from a space chosen as the best-available, already operational observing system. An example is the USGS's National Water Information System. Ongoing projects such as the Consortium of Universities for Advancement of Hydrologic Sciences Inc. (CUAHSI) Hydrologic Information System (HIS) have already created a framework for data extraction modules for the USGS sites, remote data proxy modules to support data access, and a common user interface component that can also enable viewing of other data resources (CUAHSI 2005). Besides bringing the system or systems online, the CI framework will develop data transformation modules to standardize and convert data to different formats required by various common data viewing and visualization tools.

The next item in the CI framework is the development of a workflow engine to permit the identification, extraction, assembly, and input of data into basic time-dependent, georeferenced modeling systems. The domain, species, and scale will vary according to the design, however, a watershed appears to be an optimum natural unit. In addition, the CI should also have a system to help users discover the relevant data sources available from the region. These should not only include data from the USGS sites, but also available weather and climate data (including radar data) and data from the global satellite downlink products. A data analysis toolset necessary to support the C4E4 scientific drivers will also be included, with a particular focus on turning outputs from the analysis process into inputs for existing hydrological, ecological, and other models that provide economic and health impact analyses. The final component in the CI will be a visualization module to enable two- and three-dimensional visualizations of model results.

C4E4 Cyberinfrastructure

The overall goal for C4E4 is to design and develop a distributed infrastructure that enables the environmental research and remediation community to combine heterogeneous data resources with modeling and visualization tools, in order to perform end-to-end scientific investigation. The computational integument of C4E4 is an important part of its anatomy. C4E4 will take advantage of existing resources such as the information framework developed by the CUAHSI HIS, CLEANER CyberCollaboratory (a web portal that facilitates joint working of a community available at http://cleaner.ncsa.uiuc.edu) and TeraGrid resources by integrating and customizing the modules for the study region. The integration of the heterogeneous data resources and end-to-end scientific computation will be achieved through a distributed data infrastructure and a highly data-driven workflow management environment. These and other aspects of the CI are described in the following.

Multidisciplinary Data Management System

C4E4 will leverage the Purdue TeraGrid multidisciplinary data management framework to manage data from different sources and provide multiple access points for users from communities with different levels of information technology expertise. The framework architecture consists of four layers: data capture layer, Storage Resource Broker (SRB) layer, application layer, and presentation layer (Zhao 2006). The base component is SRB, a client-server middleware developed at SDSC that provides a uniform interface to heterogeneous resources (Baru et al. 1998). It also allows users to discover data through logical attributes instead of physical file names and path names.

To further facilitate data discovery and processing that are often domain specific, a number of applications exist such as Open-source Project for a Network Data Access Protocol, (OPeNDAP), Thematic Realtime Environmental Distributed Data Services (THREDDS), and Hydrologic Data Access System (CUAHSI 2005). These servers operate with SRB and allow researchers to transform, combine, or subset data sets directly with existing OPeNDAP/THREDDS-enabled tools such as Integrated Data Viewer and MatLab (Sgouros 2004; Domenico et al. 2002). In addition, a Gridsphere-based data portal has been developed from customized JSR-168-compliant portlets, enabling easy data discovery, access, and sharing (Novotny et al. 2004).

As shown in Fig. 1, users may access data through various interfaces, including a user-friendly Gridsphere-based data portal; a set of SRB client tools including command line utilities and web/desktop interfaces; and application-specific tools, including clients enabled by OPeNDAP/THREDDS. The data management framework can have immediate impact on research communities by enabling the further development of powerful data-driven applications.

The success of C4E4 will largely depend on the identification and integration of additional data sources and in many cases data extraction methods that will need to be developed for them. As the C4E4 data sources grow, so will the need for new data transformation and access modules. A summary of possible data modules and other workflow components using sample end-to-end scenarios is discussed in the following. Another critical item in C4E4 will be to adapt a set of appropriate tools and applications into existing cyberinfrastructure frameworks such as the Rapid Application Infrastructure (Rappture) developed by the Network for Computational Nanotechnology's NanoHUB. This adaptation will enable C4E4 users to develop graphical user interfaces for their own applications that can then be shared with other C4E4 users, and will facilitate the exchange of knowledge and experience within the community.

Data-Driven Scientific Workflow Environment

Facing the challenges of heterogeneity and distribution of data sources, lack of existing metadata and metadata standards, diversity of data types, formats, and scales, and available interfaces to the data, the goal here is to develop a next-generation workflow-based system that will allow a variety of users to directly access and manipulate data relevant to the task of interest without first dealing with the details of identifying, extracting, and transforming the data. More specifically, the C4E4 workflow environment will provide integrated support for the following four components: (1) identification—the discovery of relevant data sources; (2) extraction—the retrieval of data and metadata;

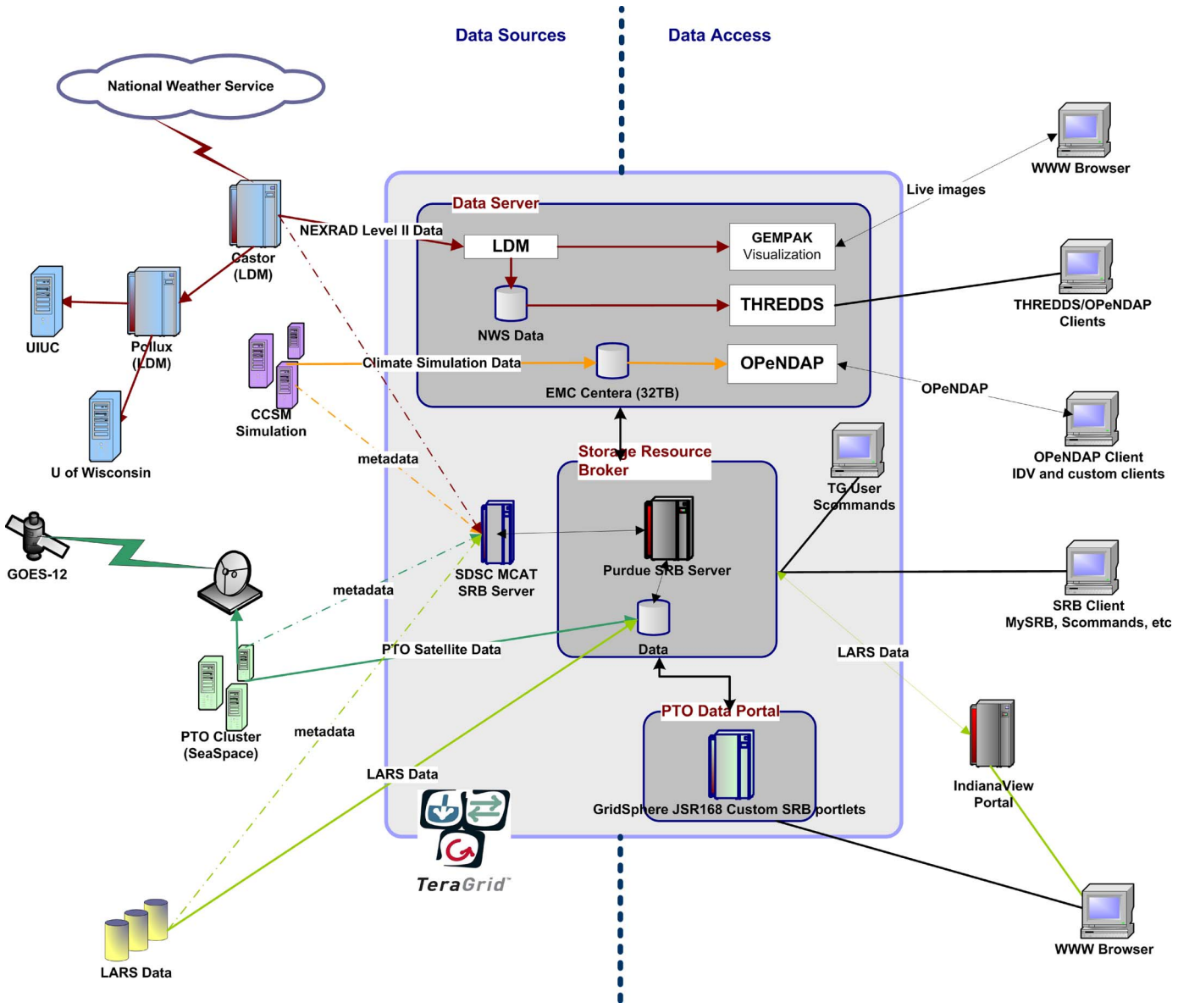

Fig. 1. Current status of Purdue multidisciplinary data management framework

(3) transformation—any reformatting required for further manipulation; and (4) output and knowledge creation, including the visualization of results and the archiving of results into the community's knowledge base. The following discusses the way in which each component will be tackled.

1. *Identification*. Access to desired environmental data might be hampered by the difficulty in identifying what pieces of data from a vast storehouse are relevant to the problem at hand. Added complexity arises from the large numbers of data sources and the wide variety of data. All types of users, ranging from beginners to advanced researchers, face this challenge. This problem can be addressed by developing a tool to help users identify available, relevant data. This identification tool could leverage existing tools (e.g., CLIPS, http://www.ghg.net/clips/CLIPS.html) with input from domain experts familiar with available data sources and from ontologies for environmental data [e.g., Semantic Web for Earth and Environmental Terminology (SWEET) from Jet Propulsion Laboratory (JPL), and GeoSemantic Web, http://sweet.jpl.nasa.gov/ontology/]. The interface needs to focus on frequent query types via a set of simple questions that direct the user to common sources of data. For more sophisticated users, the tool also needs to accept key-word information, relying on metadata to discover available sources of data.
2. *Extraction*. One of the main challenges faced by environmental scientists today is to access rapidly increasing numbers of data collections of different types and from different sources. These data are collected from different institutions and individual researchers and are available at different scales. Formats vary from point observations to spatiotemporal data, from satellite data to ground-based sensor networks, and from images to simulation outputs. Also, data are collected, stored, and accessed differently in different communities. For example, many institutions provide web access to their data sets. However, users need to navigate several web pages before reaching the data of interest. Following this step, the data may be accessible through HTTP download, FTP transfer, or even by cutting-and-pasting from a web

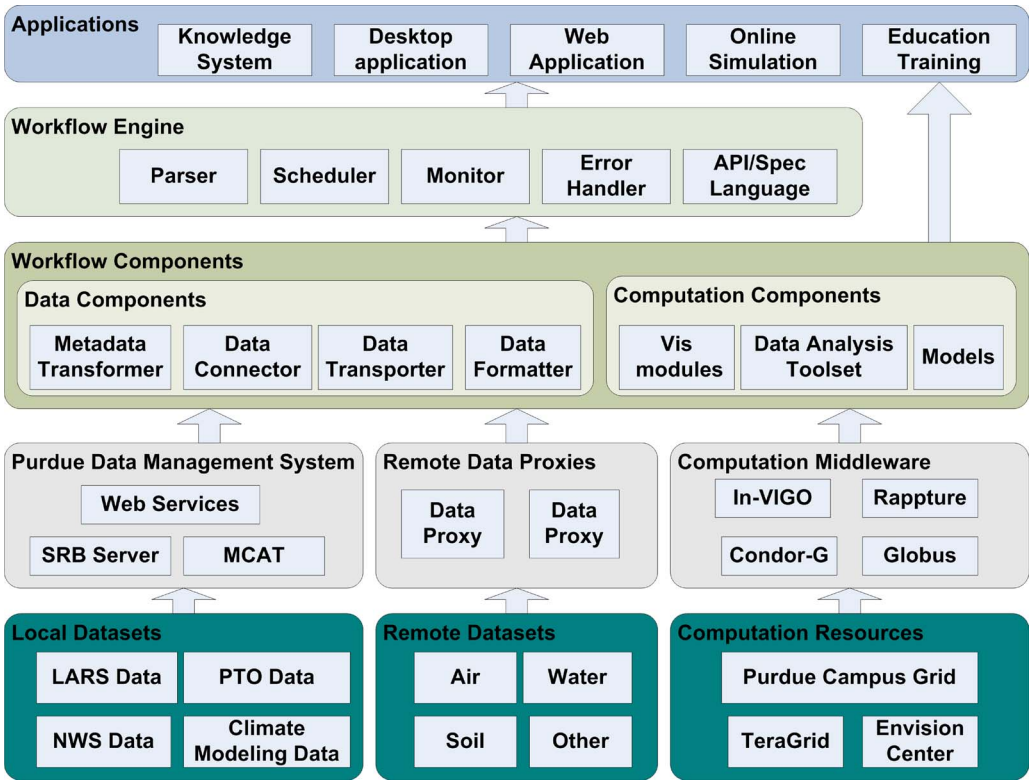

Fig. 2. Interdisciplinary environmental research workflow system

page. These data extraction operations can create obstacles for researchers.

Ultimately, standard, machine-readable, and semantically rich metadata and their processing history are needed to support users who do not have or need knowledge of data collection and management details. C4E4 will help researchers make effective use of data from different sources, ultimately increasing data usability and reusability. Starting from the operating multidisciplinary data framework, an open, extensible infrastructure that supports easy access to and scalable integration of data from heterogeneous environmental domains will be developed. This integration will include the incorporation of external data sets managed by different organizations directly or indirectly using customized data adaptors, unifying various access interfaces for remote collections, and integrating data sets with differing formats.

3. *Transformation*. The C4E4 framework will develop a user-friendly workflow system to help scientists focus on scientific questions without being hampered by the complicated and tedious tasks of understanding the underlying software and hardware systems. Researchers can use it to compose high-level experiments from small tasks using distributed data and tools. Example tasks include data retrieval from distributed data sources, data calibration that converts data formats, the assimilation of heterogeneous data sets from different disciplines, data feeding to tools, and receiving outputs that are fed to a postprocessing component such as a visualization toolkit.

   Using the workflow environment, a researcher who is interested in performing localized model calibration need only identify the data sources s/he is interested in, connect them to the model to be calibrated, set the execution conditions using a graphical interface, and then click the "Run" button. Behind the scenes, the workflow system can fetch the data from different sources, transform the data format to fit the requirements of the model, execute the model under the conditions provided by the user, and generate results. The user will be able to monitor progress of the workflow, change parameter settings interactively on the basis of results from previous runs, and even ingest the newly localized model into the infrastructure as a module for future use.

The architecture design of the workflow system is shown in Fig. 2. It consists of multiple layers. At the bottom are the distributed data and computation resources available to the system. On top of it are the multidisciplinary data management system and other middleware systems that manage and provide access to the resources. A collection of software building blocks performs basic tasks using the middleware interfaces, including data/metadata extraction, ingestion, transformation, modeling, and visualization. Several such software components are already developed to access local data sets via web services leveraged by local data portals and third-party applications (CUAHSI 2005; Zhao et al. 2007). In addition, current simulation tools (such as the ones developed for NanoHUB using the Rappture toolkit and grid middleware using In-VIGO and Condor-G) can be extended to build and enable the data visualization, data analysis, and modeling components (Frey et al. 2001). These software building blocks can be dynamically connected together into end-to-end scientific service pipelines using the workflow composer provided by the workflow environment. These scientific experiments can then be executed using the workflow run-time engine. At the application layer, customized applications can be developed to invoke and monitor the workflows constructed.

The workflow engine will include a specification language/API that supports workflow constructs such as conditions, loops, sequential and parallel patterns, a task scheduler, a status monitor, a parser, and a component for failure detection and handling. This workflow system will be extensible, so that new tasks and modules or new simulation models, tools, and data sources can be plugged into the system as needed. Users may also "save" their workflows and high-level modules that consist of small tasks. They may build more complex workflows on top of existing ones, and they may share any of these with other researchers.

4. *Output and Knowledge Creation*. The system needs to support a variety of output capabilities including web pages, visualization, model output, and XML-formatted data sets, and the ability to spatially overlay data from multiple sources. This may include a representative set of tools, including conversion tools, models, data assimilation models, statistical analysis tools, model recalibration algorithms, and GIS engines. A system such as this will enable the recalibration of regional models for local conditions.

Beyond the cyber component, the successful implementation of C4E4 will greatly depend on partnerships with agencies and researchers who are owners of data and develop unified interfaces to access their data sets. An increasing number of universities and institutions deploy SRB as the middleware to manage data collections. For example, the CUAHSI HIS is collaborating with the USGS, the National Climatic Data Center (NCDC), and the EPA to provide point observations through the hydrologic data access system. C4E4 will leverage such existing collaborations. For agencies that cannot partner with C4E4 or other similar activities, and opt to provide HTTP or FTP access to data, a simple connector component will be developed to harvest the metadata, register the data as an HTTP URL or FTP data source through the SRB data engine, and ingest the metadata to the SRB Metadata Catalog (MCAT) server. The data can then be made accessible through SRB using SRB tools and interfaces, leveraging SRB's internal support of HTTP and FTP data sources.

For each data source to be integrated into C4E4, metadata for the data will be harvested and converted. Remote sensing data have metadata embedded in the header. Climate modeling data have internal metadata that specify the variables computed and the dimensions of data values, as well as the processing history and model information. Other observational data will be coupled with metadata that specify the instruments and procedures used to generate the data. The Federal Geographic Data Committee has defined geospatial metadata standards for environmental data in geographic information systems. For disciplines without widely adopted metadata standards, an ongoing need exists to work with the monitoring community to define practical metadata schemas that will be used as an internal standard to solve the data interoperability issue. Adaptors can convert metadata from different data sources to be compliant with the corresponding metadata standard.

Visualization Capabilities

Novel tools are available which allow the three-dimensional visualization and exploration of both model outputs (e.g., WRF, microphysical cloud models) and measured data (e.g., Doppler, satellite, sensor). C4E4 will combine photorealistic visualization of atmospheric data with more traditional visualization (e.g., particle tracing and glyph rendering) to allow the simultaneous visualization of multiple data fields. Such a combination will allow the interactive exploration of atmospheric data on desktop PCs, harnessing the incredible power of recent PC graphics processor units (Riley et al. 2003, 2004, 2006). In summary, C4E4 will have multiscale, multifield visualization tools that can incorporate novel, effective, photorealistic, and illustrative visualization techniques; fuse observational and model data; scale from microphysical to mesoscale to planetary models; create effective multiscale visual representations, and produce an environment for comprehensive, and efficient visual analysis.

Examples of Potential C4E4 Applications

The research questions of greatest relevance for environmental management are hardly ever crisp, single-issue queries. They come in clumps, very much like the bodies of data that must be interrogated to correctly pose and resolve them. The following presents some examples of scenarios that describe how C4E4 can be useful for end-to-end explorations.

Watershed managers are concerned about sediment, nutrient, and contaminant loads at the outlet of a watershed of interest (Flanagan et al. 2003; Duris et al. 2004). Field-scale managers are concerned about these loads at the subwatershed or farm scale. Managers may make substantially different decisions depending on the boundaries of their jurisdiction. A best management practice (BMP) that is effective at the farm scale may be completely redundant at a larger scale (Arabi et al. 2006). After a major rainfall, managers want to assess how the BMPs in place have helped or hindered the reduction of sediment and nutrient losses. C4E4 capabilities enable them to search databases to establish the existing BMPs in that watershed and to find any previous studies related to the efficacy of different BMPs. Managers can then search for an existing model that can be used directly or modified as needed to make the assessment on the basis of current data. Given the previous and current data, C4E4's knowledge tools may suggest appropriate calibrations for the measurements and may calculate levels of uncertainty associated with model predictions.

Researchers wonder if it is possible to establish statistically defensible spatiotemporal linkages between pesticide applications and high rates of birth defects over the past decade (Garry et al. 2002; Greenlee et al. 2003). Using C4E4, they may access a database of public water supply systems and another for nitrate and pesticide data for Indiana drinking water. A C4E4 model can then be adapted to develop a geospatial data map of the areas served by the drinking water systems, on which the researchers may overlay nitrate and pesticide exposure events as well as birth-defect occurrences. Correlations may be derived from the visual data and sharpened by testing against the original data.

Indiana water authorities want to develop an early warning system to determine the likelihood of exceeding total chlorotriazine (TCT) concentrations for the remainder of any given year. TCT is a measure of the concentration of atrazine (a herbicide) and three products along its degradation pathway. The C4E4 environment enables authorities to forecast wet and dry periods for the remainder of the year. This may be combined with data on the amount of corn planted, dates of planting, and dates and amounts of atrazine application (estimated from sales of atrazine). Such data may be gathered from county extension and watershed management personnel. C4E4 can then set up the data for use in the

web-based National Agricultural Pesticide Risk Analysis (NAPRA) model (see Table 1) that can estimate TCT levels, expected ranges, and associated probabilities across the areas in question. This will enable anticipation of and triage for excessive TCT levels.

Quantification of contaminants' residence times and fluxes requires persistent sampling programs. Intermittent measurements can miss the true time behavior of interacting processes. Even the best sampling programs (e.g., USDA's recon surveys) may resolve only weekly changes. Although these data allow for a crude estimation of exposure concentrations, they do not reflect the true temporal concentration variability that many chemical and biological systems exhibit, with high-concentration events of particular note. The sampling of concentrations in parts per billion or trillion, requiring elaborate chemical postprocessing, is difficult to automate. Although work goes on to automate processes at sampling sites, however, can ways to obtain more sensitive and better resolved information from the existing data be found?

In one scenario, elevated levels of *E. coli*, long monitored by the Indiana Department of Environmental Management, evidences clear signals of fecal contamination. Researchers want to pinpoint the source or sources of *E. coli* contamination: are they failing septic tanks or sewage treatment facilities, domestic animals, livestock, or wildlife? The C4E4 user begins by searching for databases containing differentiable characteristics of fecal contamination from various sources. S/he finds the characteristics of water samples taken at various locations in the watershed and tries to provide a mapping of plausible sources within the watershed. C4E4 will make it possible for this mapping to be compared with existing data on the location and products of treatment plants, concentrated animal feeding operations (CAFOs), and other sources.

CAFOs distributed widely within Indiana produce manure which is generally used as cropland fertilizer. Does the total product of the CAFOs in any area exceed the agronomic rate, the amount of manure that can be used by plant life per acre of field? C4E4 users can quickly locate all of the CAFOs in the state and examine the numbers of livestock in each to obtain an expected loading rate. Raw rates can be converted to expected nitrogen and/or phosphorus loadings and then this converted rate can be combined with georeferenced models which map soil types, land uses, and average water table depths. The risk percentages of nitrates leaching into the groundwater at any given location can be derived, and high expectations can be checked against well water quality data, also accessible via C4E4.

Another scenario speaks to climate monitoring. In a globally warming world, climate change affects meteorological events, which in turn control the fate of such pollutants as black carbon emitted from coal-fired power plants that are, in turn, carried into the atmosphere by large-scale wildfires. Atmospheric scientists want to assess the effects of climate-perturbed meteorology and the increased availability of carbon for deposition on regional air quality. They want to understand how these changes interact with present-day sulfate aerosol distributions to change cloud albedos. What will be the consequences of increased black carbon availability for surface temperature trends in regions prone to sulfate-rich air masses in a globally warmed world? What will be the consequences of the precipitation scavenging of the redistributed sulfate aerosols? C4E4 and its links to the TeraGrid enable scientists to design and perform numerous simulation studies that link global general circulation regimes to very fine-scale aerosol microphysics and regional air quality.

Moreover, the skills acquired in modeling these atmospheric consequences might easily be applied to the accidental or deliberate release of toxins into the atmosphere. An urgent situation could arise requiring the quick assessment of potential damage in order to determine an immediately needed course of action. Once the pollutant is identified, C4E4 users can access data on transport characteristics and medical data on toxicity and symptomology. With postrelease weather data, wind patterns and high-exposure areas can be mapped using GIS techniques. Such information, if accessed and developed rapidly, can help in accurately alerting and advising emergency management and hospital personnel.

As these scenarios suggest, in many situations, knowledge acquired on a small scale can be difficult to interpret in larger scale contexts. Likewise, events occurring on a small scale may demonstrate extremely nonlinear behavior that is important but invisible in larger scale analyses. C4E4 capabilities will be key to the preservation of meaningful information as the scope of data analysis widens.

A similar spatially sensitive environmental impact is noted for regional watersheds. For instance, hypoxia in large receiving waters may result from agricultural practices aggregated across multiple watersheds. However, remediation strategies may only have been attempted on the smaller scale of brooks or ponds. Can or should such strategies be scaled up? C4E4 users can interrogate data on all scales and aggregate them to visualize and estimate their combined effects and infer the ecological functioning of larger bodies of water. They may find, for example, that links also exist among urbanization (increased sewage), climate change, and hypoxia in addition to the agricultural etiology. Such findings will in turn affect cost/benefit estimates for the scaling of agricultural remedies.

In another context, data on the placement of tiled drainage systems over portions of the USGS National Water-Quality Assessment (NAWQA) watersheds is sparse, yet the hydrology of small plots with such drainage has been well studied. Do the preferential flow paths of water over large areas confound the conclusions of such studies? That is, at the watershed scale, how can multiple responses be integrated? With C4E4 modeling strategies, such questions may result in an estimation of the predictability of the interaction of manmade drainage systems with natural drainage. Such estimations might figure importantly into large improvements in watershed management.

Initial Efforts

Our initial efforts for realizing the C4E4 vision (see Zhao et al. 2007, for details) have focused on the use of a process-based distributed watershed model, the SWAT, over the St. Joseph Watershed in northern Indiana. Our objective in setting up SWAT for St. Joseph Watershed is to create a base model that potential users can use to evaluate the effect of different BMPs and land use changes on watershed hydrology and water quality. Users will have options to change land use management scenarios via the interface, run the model, and evaluate potential benefits with respect to the base model run.

The 280,000 ha St. Joseph River Watershed, located in northeast Indiana, northwest Ohio, and south central Michigan, is a Source Water Protection Initiative watershed. The main stream of the watershed is the St. Joseph River, approximately 100 mil long, which runs in a NE–SW direction and joins the Maumee River at Fort Wayne (Fig. 3). The St. Joseph River is the main source of drinking water for approximately 200,000 residents in

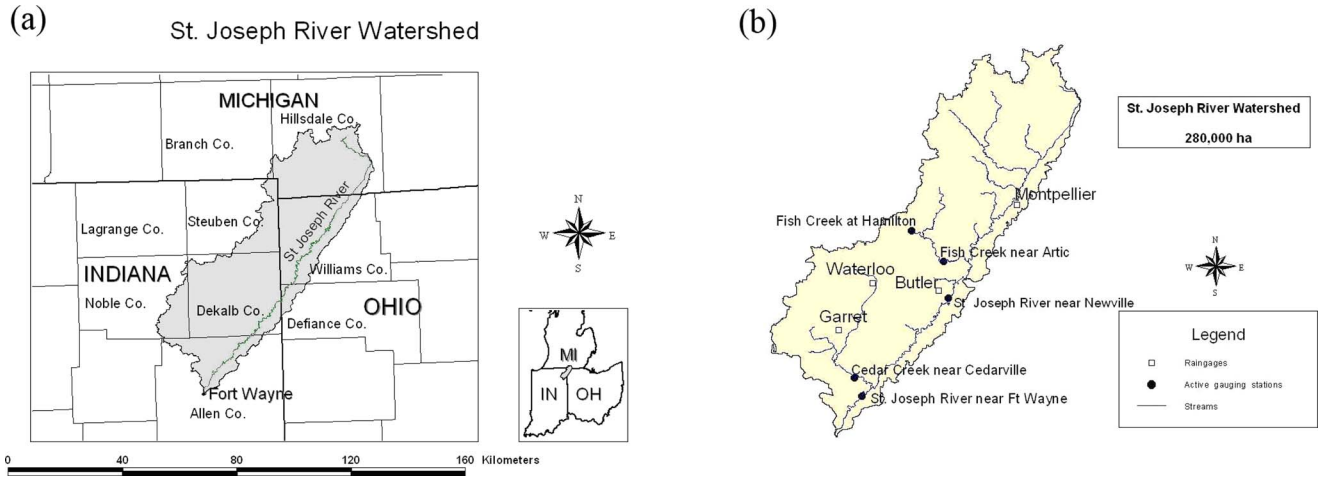

Fig. 3. Maps of the St. Joseph Watershed showing (a) geographical location of the watershed; (b) USGS gauges and NCDC stations

Fort Wayne. A number of environmental groups/community interfaces (see http://www.sjrwi.org/) are active in this watershed. Since 1995, agricultural chemicals have been detected in the St. Joseph River at Fort Wayne. Peak levels of atrazine (an herbicide) exceeding 3 ppb, the EPA drinking water standard, have been reported at different sites in the watershed between 1995 and 1998 by a network of environmental groups (Environmental Working Group) and the St. Joseph River Watershed Initiative (SJRWI).

SWAT is a process-based distributed-parameter watershed model, developed by the USDA to quantify the impact of land management practices in complex watersheds with varying soils, land use, and management conditions over a long time period (Neitsch et al. 2002). Major components of the model include weather, surface runoff, return flow, percolation, evapotranspiration, transmission losses, pond and reservoir storage, crop growth and irrigation, groundwater flow, reach routing, nutrient and pesticide loads, and water transfer. It is currently only available for MS Windows platform. In order to run SWAT on the TeraGrid Linux resources, source code of SWAT 2005 was ported to Linux using Intel Fortran 90 compiler.

The web services interfaces used in the SWAT workflow are implemented using Apache Axis API. The web services interfaces invoked by the SWAT workflow modules are briefly described in Table 2. For flexible and extensible design, debugging, and future support, the JOpera workflow engine and run-time environment has been incorporated into the C4E4 architecture (see Fig. 4). With minimal coding effort, a prototype SWAT modeling pipeline is constructed on this framework that accepts details of a SWAT simulation, runs it on the TeraGrid Condor cluster, fetches the output, transforms, plots, and publishes the result, and finally sends an e-mail notification to the user.

Because SWAT is a conceptual model, calibration of its parameters to reproduce observed streamflow is the first step in setting up the model to make future predictions. Model calibration involves three steps: (1) data organization and preprocessing; (2) watershed delineation and its discretization into subwatersheds and hydrologic response units (HRUs are basic calculation units composed of unique combination of land use and soil type); and (3) preparation of input files followed by parameter calibration using optimization algorithms. The final C4E4 architecture is expected to provide functionalities for all three steps including execution of SWAT for future predictions. Because model calibration can take anywhere from a few days to several weeks depending on the number of subwatersheds/HRUs, calibration period, and number of targeted parameters, our initial efforts are focused on Step 3 to leverage TeraGrid's parallel computing resources. As a first step, the writers were successful in implementing the SWAT autocalibration routine on TeraGrid, thus running multiple calibrations in parallel. C4E4 portal has a window that uses web

Table 2. Web Services Involved in SWAT Workflow Modules

Step	Interface name	Description
1	*submitJob*	This interface composes a SWAT simulation job based on the input parameters provided by the caller, and submits the job to the Globus Condor job manager running on the Purdue TeraGrid gatekeeper using Globus GRAM Java API. It returns when the job completes its execution in the TeraGrid Condor pool. The output of the job is archived in a tar file and sent back to the submission node.
2	*extractOutput*	This interface extracts the specific target output files out of the tar file generated in the first step based on the simulation information the user is interested in. For example, in the case of total amount of precipitation or surface runoff contribution to streamflow, the output file *output.std* will be extracted.
3	*getData*	This interface parses the extracted output file and transforms the specific simulation information into a form readable for gnuplot, a portable command line interactive plotting utility.
4	*gnuplot*	This interface converts the data in the transformed result file into two-dimensional graphs using gnuplot Java library. The plot data are stored in portable network graphics (PNG) file format and can be viewed or downloaded through a web server.
5	*sendMail*	This interface receives as input the URL to the plot and sends it in an e-mail to the user so that s/he can view the result online.

Note: Steps 3 and 4 can be invoked multiple times depending on the number of simulation field values that the user is interested in analyzing. Adapted from Zhao et al. (2007).

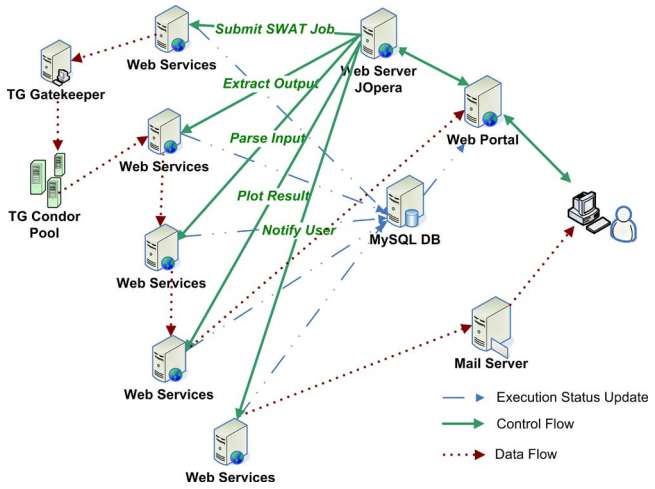

Fig. 4. Distributed SWAT workflow execution diagram

services listed in Table 2 to accept input files for SWAT calibration and users are notified after the job is complete (Fig. 4).

To create a base model for the St. Joseph Watershed, six watershed configurations, each involving two resolutions of soil data (SSURGO and STATSGO) resulting in twelve configurations in all, were used to calibrate a set of fourteen parameters. In addition, sixteen configurations involving SSURGO and STATSGO were created for one of the subwatersheds, Cedar Creek, within the St. Joseph Watershed (Table 3). These configurations that include subwatersheds at different scales were created to analyze the effect of spatial scale and soil data on model performance including variability in calibrated parameters. Twenty eight calibrations, with an average computation time of 2 weeks for simulating 7 years of daily streamflow data, would take almost a year on a single computer. However, using the C4E4 framework, this task was accomplished in 3 weeks by performing the calibration simulations in parallel. It was found that the more sensitive parameters (e.g., SCS curve number) exhibit less variability and less sensitive parameters (e.g., soil hydraulic conductivity) exhibit greater variability (Fig. 5) among the configurations. In addition, different subwatershed configurations and soil data resolution do not show significant effect on model performance in terms of Nash-Sutcliffe coefficient (Table 3).

Our initial efforts using SWAT demonstrate C4E4's capability to support computational needs in environmental modeling efforts. Researchers can focus on application questions instead of being hindered by model run time and computational effort. Even inexperienced users can configure parameters through a user-friendly web interface, launch experiments, and view results online.

Education and Training

The success of C4E4 will largely depend on educating the scientific and applications communities and providing training to potential users and stakeholders. C4E4 can provide a powerful learning environment supported by and supportive of an enthusiastic, engaged learning community as part of ongoing formal and informal environmental education. Learning modules can be built in compliance with emerging e-learning standards and specifications, making such modules easy for broad dissemination. Course modules can embrace active learning and team teaching techniques that can be incorporated not only into undergraduate and graduate education, but also into training sessions for K–12 teachers. Installation of Access Grid nodes and associated cyberinfrastructure can enable collaborative environmental learning and research between different institutions.

Table 3. Watershed Configurations for St. Joseph and Cedar Creek, and Their Model Performance during Calibration and Validation Phase

	% CSA	Soil data	Model code	N	Calibration (1993–1999) R^2_{NS}	Validation (2000–2003) R^2_{NS}
St. Joseph	0.5	SSURGO	A0.5	97	0.65	0.66
	1.0		A1.0	58	0.61	0.59
	2.0		A2.0	36	0.66	0.67
	3.0		A3.0	24	0.59	0.61
	5.0		A5.0	12	0.46	0.60
	7.0		A7.0	10	0.60	0.61
	0.5	STATSGO	B0.5	97	0.60	0.62
	1.0		B1.0	58	0.66	0.66
	2.0		B2.0	36	0.66	0.61
	3.0		B3.0	24	0.61	0.66
	5.0		B5.0	12	0.43	0.50
	7.0		B7.0	10	0.52	0.61
Cedar Creek	1.5	SSURGO	C1.5	41	0.69	0.54
	2.0		C2.0	23	0.68	0.56
	2.5		C2.5	17	0.67	0.58
	3.0		C3.0	17	0.70	0.54
	4.0		C4.0	17	0.70	0.55
	5.0		C5.0	15	0.68	0.56
	7.0		C7.0	9	0.69	0.56
	10.0		C10.0	7	0.70	0.58
	1.5	STATSGO	D1.5	41	0.73	0.61
	2.0		D2.0	23	0.75	0.62
	2.5		D2.5	17	0.75	0.62
	3.0		D3.0	17	0.75	0.61
	4.0		D4.0	17	0.75	0.61
	5.0		D5.0	15	0.75	0.59
	7.0		D7.0	9	0.74	0.60
	10.0		D10.0	7	0.75	0.60

Note: CSA refers to critical threshold area used and delineat stream network; N refers to number of subwatersheds; R^2_{NS} refers to Nash–Sutcliffe coefficient.

Conclusions

C4E4 is envisioned as a prototype that enables a broad community of researchers to ask and answer environmental questions at local, state, national, and even global scales. It aims to become particularly useful to the participants in various national environmental observatory and infrastructure projects (NEON, GEON, CUAHSI, CLEANER, and others), who will be able to share data and access resources of multiple scientific computational grids, including the TeraGrid, through the C4E4 portal. Large-scale computation and the development of advanced cyberinfrastructure will play an important role in forging new collaborations within an especially diverse environmental science community. As the C4E4 framework outlined in this paper is not location specific, it will be broadly applicable for many research problems at different locations and scales, including those that require an

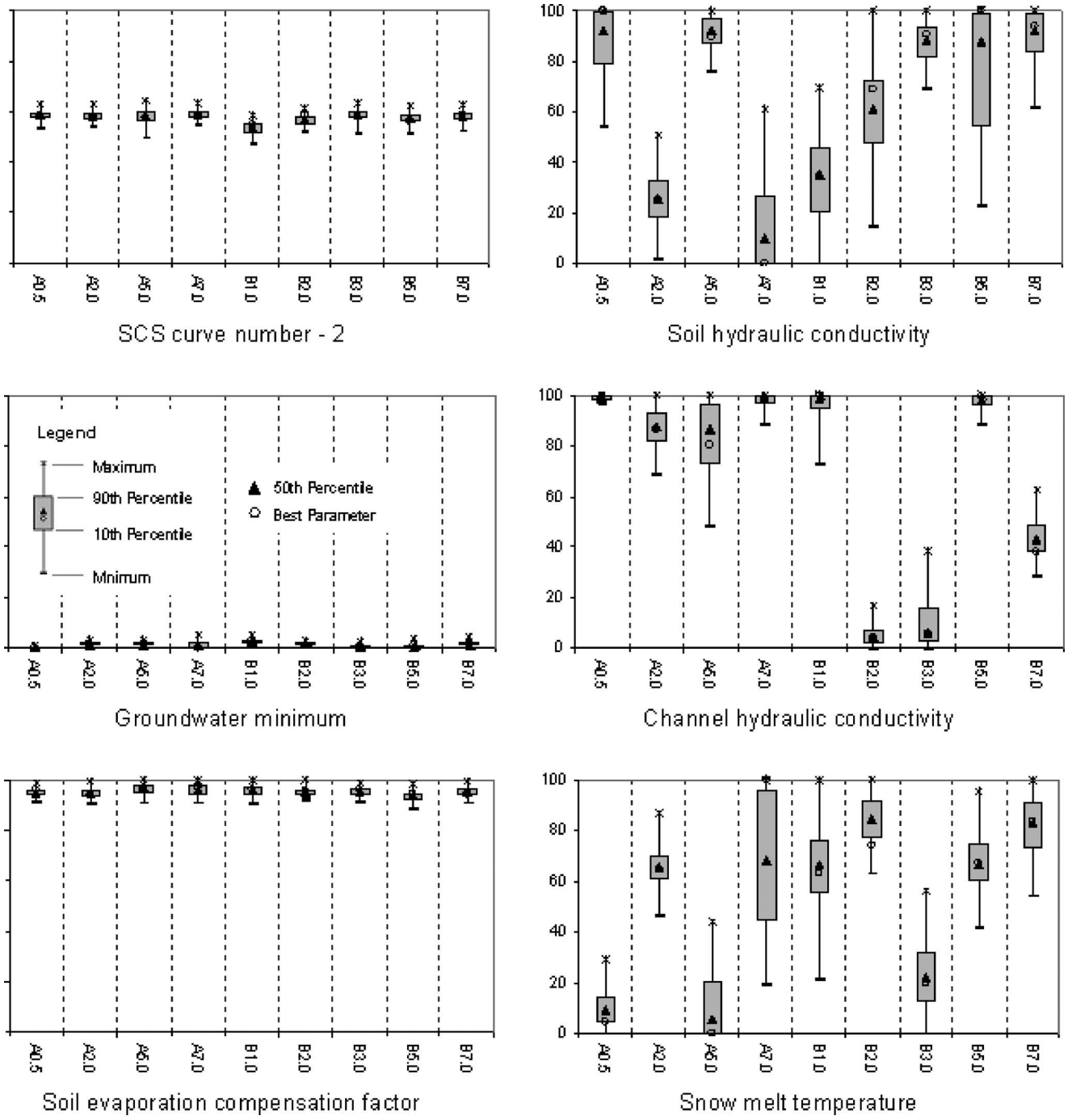

Fig. 5. Selected parameter sets from autocalibration. *X*-axis refers to selected model codes from Table 3, and *y*-axis shows normalized parameter range obtained from autocalibration.

interdisciplinary approach. C4E4 in conjunction with existing resources such as TeraGrid will allow integration and collaborative linkages among local, regional, and national efforts.

Acknowledgments

C4E4 project is supported by the National Science Foundation under Grant No. 0619086. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the writer(s) and do not necessarily reflect the views of the National Science Foundation.

Appendix. NanoHUB Infrastructure

The NSF-funded Network for Computational Nanotechnology (NCN), centered at Purdue University, is a research site connecting theory, experimentation, and computation by supplying online simulation and educational content services remotely through the web. The NCN NanoHUB is the interface to the cyberinfrastructure and the defining deliverable of NCN. It puts data, simulation tools, and research-grade software, as well as educational materials, in the hands of users ranging from nanoelectronic researchers to K–12 educators and students.

The NanoHUB has established itself as a model of service-

oriented science through its easy-to-use online simulation tools and variety of educational and research materials. The website, http://www.nanohub.org, supplies free accounts to users who may then access various programs, run simulations, and view results through web browsers without having to download and install software on their local computers. The web site also enables users to share their tools, contribute online courses, tutorials, and seminars, and participate in discussions with peers. Educators can also post teaching materials and homework assignments.

Usage of NanoHUB resources, including tools and learning materials, and user profiles (e.g., research, industry, K–12), are tracked and analyzed to help improve the quality of the online services and content. The educational materials include learning modules, entire courses, tutorials, and seminars that are used extensively.

Two key NanoHUB middleware developments will also be deployed in C4E4. In-VIGO (In Virtual Grid Organization) allows any NanoHUB simulation to seamlessly access local computing resources and, more important, resources on the computing grids, like the TeraGrid supercomputers without the need to understand how to gain access to these resources. A major issue in web-enabling applications has been that users often spend significant amounts of time developing graphical user interfaces (GUIs) for different applications. NanoHUB has developed a software toolkit called Rappture (Rapid Application Infrastructure) to enable easy creation of user-friendly GUIs for legacy and new applications without a significant burden on the application developer. Rappture provides a simple API for use in the application to describe inputs (e.g., temperature, windspeed) and outputs (e.g., two-dimensional image, line graph). NanoHUB software engineers also develop Rappture I/O wrappers for legacy tools. The "rappturized tools" can then be readily deployed and shared online. Rappture will be used in C4E4 to enable various simulation tools and applications to be accessible over the web.

References

Arabi, R. S., Govindaraju, R. S., Hantush, M. M., and Engel, B. (2006). "Role of watershed discretization on evaluation of long-term impact of best management practices on water quality." *J. Am. Water Resour. Assoc.*, 43(2), 513–528.

Baru, C., Moore, R., Rajasekar, A., and Wan, M. (1998). "The SDSC storage resource broker." *Proc., CASCON'98*, Toronto.

Clendenon, C. J., and Beaty, J. E., eds. (1987). *Water resource availability in the St. Joseph River Basin, Indiana*, Assessment No. 87–1, D6 Division of Water Resources, Division of Water, Indianapolis.

CUAHSI. (2005). "Hydrologic information system status report version 1.0." ⟨http://www.ncar.ucar.edu/cyber/cyberreport.pdf⟩ (Oct. 28, 2008).

Domenico, B., Caron, J., Davis, E., Kambic, R., and Nativi, S. (2002). "Thematic real-time environmental distributed data services, (THREDDS): Incorporating interactive analysis tools into NSDL." *J. Digital Inf.*, 2(4), 15–35.

Duris, J. W., Reeves, H. R., and Kiesler, J. L. (2004). "Atrazine concentrations in stream water and streambed sediment pore water in the St. Joseph and Galien River basins, Michigan and Indiana, May 2001–September 2003." USGS, Washington, D.C., *Open-File Rep. 2004-1326*.

Engel, B., Lim, K. J., and Navulur, K. C. S. (2007). "The role of geographical information systems in groundwater engineering." *The handbook of groundwater engineering*, J. W. Delleur, ed., CRC, New York, 30-1–30-17.

Flanagan, D. C., Livingston, S. J., Huang, C. H., and Warnemuende, E. A. (2003). "Runoff and pesticide discharge from agricultural watersheds in NE Indiana." *ASAE Paper No. 03-2006*, American Society of Agricultural Engineers, St. Joseph, Mich.

Frey, J., Tannenbaum, T., Livny, M., Foster, I., and Tuecke, S. (2001). "Condor-G: A computation management agent for multi-institutional grids." *Cluster Compu.*, 5(3), 237–246.

Garry, V. F., Harkins, M. E., Erickson, L. L., Long-Simpson, L. K., Holland, S. E., and Burroughs, B. L. (2002). "Birth defects, season of conception, and sex of children born to pesticide applicators living in the Red River Valley of Minnesota, USA." *Environ. Health Perspect.*, 110(3), 441–449.

Goolsby, D. A., et al. (1999). "Flux and sources of nutrients in the Mississippi-Atchafalaya River Basin." *Topic 3 Rep. for the Integrated Assessment on Hypoxia in the Gulf of Mexico, NOAA Coastal Ocean Program Decision Analysis Series No. 17*, NOAA Coastal Ocean Office, Silver Spring, Md.

Goolsby, D. A., Battaglin, W. A., Aulenbach, B. T., and Hooper, R. P. (2001). "Nitrogen input to the Gulf of Mexico." *J. Environ. Qual.*, 30(2), 329–336.

Greenlee, J. S., Arbuckle, A. R., and Chyou, P. H. (2003). "Risk factors for female infertility in an agricultural region." *Epidemiology*, 13(4), 429–436.

Holtschlag, D. J., and Nicholas, J. R. (1998). "Indirect groundwater discharge to the Great Lakes." *Open-File Rep. No. 98-579*, USGS, Washington, D.C.

National Center for Atmospheric Research (NCAR). (2003). "Cyberinfrastructure for environmental research and education." *Rep.*, NSF sponsored workshop, ⟨http://www.ncar.ucar.edu/cyber/cyberreport.pdf⟩ (Oct. 28, 2008).

National Science Foundation (NSF). (2006). "Sensors for environmental observatories." ⟨www.nsf.gov⟩ (Oct. 28, 2008).

Neitsch, S. L., Arnold, J. G., Kiniry, J. R., Williams, J. R., and King, K. W. (2002). "Soil and water assessment tool theoretical documentation, version 2000." Grassland, Soil and Water Research Laboratory, Agricultural Research Service, Temple, Tex.

Novotny, J., Russell, M., and Wehrens, O. (2004). "Gridsphere: A portal framework for building collaborations." *Concurrency Comput.: Pract. Exper.*, 16(5), 503–513.

Riley, K., Ebert, D., Hansen, C., and Levit, J. (2003). "Visually accurate multi-field weather visualization." *Proc., 14th IEEE Visualization Conf. (VIS'03)*, IEEE.

Riley, K., Ebert, D. S., Kraus, M., Tessendorf, J., and Hansen, C. (2004). "Efficient rendering of atmospheric phenomena." *Proc., Eurographics Symp. on Rendering 2004*, Springer, 375–386.

Riley, K., Song, Y., Kraus, M., Levit, J., and Ebert, D. (2006). "Visualization of structured nonuniform grids." *IEEE Comput. Graphics Appl.*, 26(1), 24–33.

Sgouros, T. (2004). *OPeNDAP user guide, version 1.14*, ⟨http://www.opendap.org/user/guide-html/guide.html⟩.

U.S. EPA. (2000). "National primary drinking water regulation-regulated contaminants." Title 40 Code of Federal Regulations, Part, 141, Subpart 0, App. A, 336–538.

USGS. (2000). "Nitrogen in the Mississippi Basin—Estimating sources and predicting flux to the Gulf of Mexico." *USGS Fact Sheet No. 135-00*, Washington, D.C.

Zhao, L., et al. (2007). "Interweaving data and computation for end-to-end environmental exploration on the teraGrid." *Proc., TeraGrid 2007 Conf.*, Madison, Wis.

Zhao, L., Park, T., Kalyanam, R., Lee, W., and Goasguen, S. (2006), "Purdue multidisciplinary data management framework using SRB." *SRB Workshop*, SDSC, San Diego.