Geography Compass 2/6 (2008): 2040–2061, 10.1111/j.1749-8198.2008.00160.x

A Perspective on Cyberinfrastructure for Water Research Driven by Informatics Methodologies

Peter Bajcsy*
National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign

Abstract

This article presents a perspective on cyberinfrastructure (CI) for water research driven by informatics methodologies. This perspective is motivated by the fact that CI for water research should increase scientific productivity of a single investigator and dispersed teams in their exploratory studies of experimental observations and theoretical simulations. In the digital era and within the context of CI, the scientific productivity is often determined by the efficiency of data-centric and collaboration-centric activities. These activities follow domain-specific methodologies and informatics approaches aiming at extracting information and knowledge from raw data. We present the concepts of data-centric and collaboration-centric activities supported by concrete examples, and outline the challenges and requirements on CI driven by these informatics activities. Then, we describe a set of common activities cutting across multiple earth science disciplines, and discuss some solutions that already exist or might have to be developed in order to support informatics methodologies. Our perspective emphasizes the importance of exploratory science that is frequently present in informatics methodologies. The contribution of this article is in illuminating the CI development from a perspective that bridges daily activities of many water research scientists with the CI components and functionality.

1 Introduction

Many engineering and scientific fields including earth sciences have always relied heavily on observations and measurements of natural phenomena and learning from the observations. In many cases, one could simplify the learning process as a closed loop that goes from sensors and instruments to data (observations and measurements) to information and to knowledge as illustrated in Figure 1. In the digital era, observations and measurements become digital raw data that are understood in this context as anything that can be ingested by a computer. Raw data are processed to extract the meaning of raw data that is narrow in scope and it has a simple organization. We refer to the extracted meaning as information. From information, one derives knowledge which is an interpretation of information that is broad in scope and it is orderly synthesized. Finally, the knowledge is used

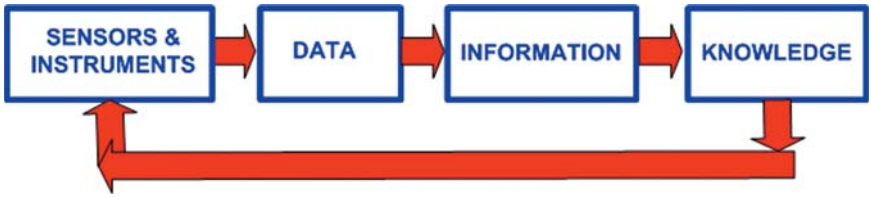

Fig. 1. Typical flow of a learning process in earth sciences.

for optimal spatial placement of sensors, temporal sampling or variable selection in order to support the end applications. One such example could be a decision support system for sustainability of rural and urban environments or for emergency management. This process of going from raw data to information and knowledge has been referred to as informatics science.

1.1 CHALLENGES IN LEARNING

Over the past couple of decades, there have been several developments in informatics science and this learning process that have posed new challenges. First, many sensor and instruments advancements led to new types of data and exponentially increasing amounts of data (Balazinska et al. 2007). The challenge is in *the lack of software tools for dealing with these voluminous data sets*, where the tools would be designed for organizing, storing, preserving, retrieving, browsing, processing, and visualizing data. It is frequently the size that prevents us from learning because it has become cost- and time-prohibitive to visually inspect and analyze terabytes of data everyday. Second, with the dynamic evolution of our societies and the global nature of all economies, the process of learning from observations has become hindered by *the lack of software and hardware solutions needed to support time-critical learning*. Supporting the learning process is critical in order to resolve the constantly competing objectives of economic growths and sustainability of rural and urban environments on earth. In the highly dynamic environments on earth, the rate of learning is time-critical and a lack of solutions for efficient learning leads to gaining temporally obsolete knowledge of lesser value when addressing the sustainability issues. Learning depends not only on theoretical knowledge but also on observations as explicitly stated in several water research-related vision papers (Kirchner 2006; Newman et al. 2006). Third, without acquiring more measurements and without analyzing data, we would not be able to support global economies, to manage the impact of humans on their environment and to address concerns of global warming and alike. The challenge is in conducting *data analyses at global spatial scale* for time-critical applications since the competing objectives are not anymore constrained to a small spatial location and require time-critical responses. Finally, the learning process is not a work of a single investigator or a small team of investigators but it has become a collaborative effort of multiple

teams, agencies, communities and states all over the globe. The challenge lies in providing a basic *infrastructure for enabling collaborative scientific efforts* that are characterized by distributed sensing, distributed human expertise and distributed computational (hardware and software) resources.

1.2 INTRODUCTION TO CYBERINFRASTRUCTURE

To cover the aforementioned challenges in earth sciences and many other fields, the word cyberinfrastructure (CI) has been coined to describe sensor/instrumentation and computational hardware, networking hardware and software supporting many functions of the hardware, such as communication, data management, analyses and syntheses, security, and so on. CI has the potential of having a revolutionary impact of computational technologies on scientific and technological progress due to the emergence of the Internet, ubiquitous (cloud) computing, mobile technologies, and advanced sensing. There is a plethora of definitions, community-specific interpretations, and reports describing the future of CI (Atkins 2003; Finholt and Van Briesen 2007; Maidment 2005; National Science Foundation [NSF] 2006). One of the generally accepted definitions of CI is: 'Cyberinfrastructure refers to infrastructure based upon distributed computer, information and communication technology. If infrastructure is required for an industrial economy, then we could say that cyberinfrastructure is required for a knowledge economy' (Atkins 2003). A critical distinction between CI and earlier terms, such as collaboratories and grids, is the recognition that CI is infrastructure that will become vital in addressing the key research and development challenges and increasing scientific productivity of the 21st century. Despite its promise, CI is as yet far from ubiquitous and is still considered by many to be difficult to use productively.

In some sense, CI has been in development since the advent of the computer and the Internet and has been benefiting from the exponential increases in computing, storage, and networking capabilities ever since. Inflection points such as the development of the Mosaic Web browser in 1993 at the National Center for Supercomputing Applications and the emergence of the World Wide Web have helped shift our perception of CI from simply providing cycles and storage to understanding that its true power includes organization of information and coordination of efforts. Over the past decade, there have been developments such as grids, portals, and social networking sites, and community databases that have helped drive these concepts of community-scale sharing and organization of computation, data, and expertise. For instance, it has been documented in these articles (Daigle and Cuocco 2002; Myers et al. 2003; Spencer et al. 2006) that CI will enable qualitatively new approaches to scientific research in addition to the quantitative improvements leveraging hardware advances. Most recently, the emergence of Web 2.0 technologies has shown how core capabilities created by service providers can rapidly be

'mashed-up' (Ankolekar et al. 2007) and customized to support the needs of specific projects and communities.

1.3 WHY CYBERINFRASTRUCTURE-BASED SCIENCE?

Although one might agree with the learning objective in science, the challenges in sciences and learning, and the advancements of CI (or information technology) as described before, there is still a remaining question: why CI-based science? The answer can be found in (i) the rate of changes of scientific problems and (ii) the formidable complexity of scientific problems. In this era of human-induced change, the rate of changes of environmental conditions is very fast, and solving-related scientific problems often become important for adequate land and water management. Although the problems, such as flood prediction in ungauged basins and the development of generalized scaling theory (Gupta et al. 2007; Wagener et al. 2007), are of great practical significance in engineering hydrology at multiple scales, they cannot be fully explored without analyzing large amounts of complex real-time observations. One can quickly conclude that it is impossible to observe, analyze, predict, and manage in real-time complex watersheds where real people live unless CI enables the water sciences to deal with volume, velocity, and complexity of predictions.

The second part of the answer lies in the complexity of finding solutions to scientific problems. The complexity of scientific solutions stems from the complex instrumentation providing observables about phenomena, and from the complexity of underlying phenomena (and hence the complexity of prediction and decision support models). Furthermore, there is additional complexity in linking instrumentation providing dynamic observables with the prediction and decision support models that are also evolving over time. For example, in a set of problems related to sustainability of rural and urban environments and emergency management, spatially and temporally varying observations flow into models and then the models determine the spatial and temporal samples of next observations in a closed loop system. Such complex problems, including large volumes of observations, computationally demanding models, and closed loop paradigms, cannot be solved without the help of CI and the involvement of multiple community experts. It is the complexity of problems related to the role of water in physical and biological systems that has suggested a paradigm shift in hydrologic education from specializations in intra-disciplinary sub-areas, such as 'surface water hydrology', 'ground water hydrology', or 'wetland hydrology', to cross-disciplinary sub-areas, such as 'eco-hydrology', 'hydro-pedology', or 'hydro-geomorphology' (Kumar 2007). Nonetheless, the paradigm shift to cross-disciplinary education has to be still complemented by collaborations of spatially distributed multiple community experts with various expertise in order to find solutions to scientific problems. Thus, CI and its support for collaborative efforts and

community interactions have become essential to making scientific progress and enabling community science.

1.4 CYBERINFRASTRUCTURE FOR WATER RESOURCE COMMUNITIES

Several communities conducting water-related research have formed consortiums to bring together information technology experts, and domain scientists and practitioners. The goal of these communities is to design and build CI for water-related disciplines with the funding provided by the NSF. For example, the communities of hydrologists and environmental engineers formed institutional umbrellas such as the Consortium of Universities for the Advancement of Hydrologic Science[1] (CUAHSI) and Collaborative Large-scale Engineering Analysis Network for Environmental Research[2] (CLEANER) project for the WATer and Environmental Research Systems Network[3] (WATERS Network). Other National Science Foundation (NSF), National Institutes of Health (NIH), and Department of Energy (DOE) funded projects have contributed to prototyping examples of CI-enabled environments that could be used for water research, for instance, the NSF Network for Earthquake Engineering Simulation (NEES), the NSF National Virtual Observatory (NVO), the NSF National Ecological Observatory Network (NEON), NIH Biomedical Informatics Research Network (BIRN), or the DOE Scientific Discovery through Advanced Computing (SciDAC) projects just to mentioned a few. While the communities building software for CI are numerous, the human resources are limited. Thus, sharing human resources, establishing and following technology standards, and coordination of efforts in meeting informatics requirements have become eminent.

1.5 WHY CYBERINFRASTRUCTURE DRIVEN BY INFORMATICS METHODOLOGIES?

For many communities, the ultimate applications of CI are in the areas of monitoring, management (Jonoski and Popescu 2004) (e.g., emergency or development), preservation, restoration (Marlin and Darmody 2005), prediction, and validation just to name a few. It is the community of scientists working in multiple research areas (called domain scientists) that often define requirements for the development of the enabling capabilities of CI. Similarly, it is the community of information technology experts that design and develop CI components, select the components of interest and assemble them into solutions to meet domain specific needs. The expectations on CI are that researchers and practitioners would be *more productive and effective* in their research and engineering tasks, as well as their *learning would be enabled over voluminous data sets, at global spatial scales for time-critical applications and in collaboration with large distributed teams of investigators*.

In this overview, we present a perspective on CI for water research that is driven by informatics methodologies. We use the term 'water research' in the title and in the rest of this article to refer to those earth sciences

that deal with water. Water is the link between many earth science disciplines and is viewed as the key to many conflicting economic and environmental sustainability issues (Team 2008). We use the term 'informatics methodologies' to refer to all techniques in water research that deal with observations in digital format and follow the learning process depicted in Figure 1.

The perspective on CI that is driven by informatics methodologies is based on our many years of experience while working with several water research communities. We have experienced several 'disillusions' among CI researchers and water resource researches in the past due to the mismatch between what CI would deliver at a fixed cost and what the water resource community would expect from CI (and would view as advancements of hydrologic/environmental science). Therefore, we describe a very pragmatic perspective on CI in this article. We suggest scrutinizing the daily informatics activities of water resource scientists and then considering the potential for increased efficiency of our daily activities by driving the future CI development.

One has to be aware that the CI on its own has very little intelligence and cannot deliver new science without human creativity reflected in various methodologies. Nonetheless, CI has the power of speed (versus humans) when it comes to informatics tasks, such as inspection of large volumes of data, complex computation, or finding 'information needle in a digital haystack'. It is our belief that the advancements of hydrologic/environmental science will only come indirectly through the use of CI but not directly by designing an autonomous CI. Thus, our intention is to encourage the readers to review their daily scientific activities and hopefully turn to CI for the efficiency improvements.

The logical structure of the article follows the above perspective. The example methodologies attempt to illustrate scenarios of daily informatics tasks in section 2 and the challenges encountered while performing them in section 3. In section 4, the daily activities are broken into data-centric and collaboration-centric activities in order to describe the commonality of these activities and to suggest driving CI development with the requirements arising from the most common needs (and efficiency improvements). The sections on solutions supporting common data-centric and collaboration-centric activities provide pointers where to look for CI components of interest. Section 5 summarizes how scientists performing exploratory science using informatics methodologies would benefit from CI supporting the aforementioned activities.

2 Informatics Methodologies in Water Research

In order to make scientists in water research more productive, one has to identify the types of daily activities. From a perspective of a researcher (especially in earth science disciplines), many activities are evolving around understanding phenomena based on observations and modeling. Activities

are typically related (i) to designing methodologies on how to extract information and knowledge from raw observations (raw data) and (ii) to communicating the results of various methodologies to others. We decomposed these scientific activities into data-centric and collaboration-centric activities. We use the term *collaboration* rather than *communication* because in order to increase research productivity collaboration reflects better the desired goal of communication.

The data-centric activities of going from raw data/measurements to information and knowledge have been also referred to as 'informatics' activities (Bajcsy et al. 2005). From the informatics perspective, any software for CI becomes a part of an X-informatics system, where X stands for a specific application domain. Some common uses for X are bio, hydro, geo, medical, document, astro, or sensor (Abbott 1991; Bajcsy 2006; Bajcsy et al. 2004; Price et al. 2000). In all these instances, classes of the X-informatics systems are typically characterized by methodologies of steps or activities that have to take place to conduct experiments and arrive to knowledge. Thus, the research and development of software for CI is therefore driven by a set of requirements for building X-informatics solutions following common methodologies. Broad community infrastructure can be build to support specific research areas by providing common informatics frameworks. These frameworks would be decoupled from the specific components developed by individual researchers and should provide simple mechanisms for inclusion of new methodologies and research studies, no matter what the specific area of study might be.

The collaboration-centric activities range a lot and deserve special attention from a social science view point. We focus more on the spectrum of activities and the forms of collaborations than on the social aspects of collaboration. The fundamental premise of any successful scientific collaboration is sharing of ideas, data, software, hardware, and so on. Sharing is an activity and it has many forms. In the context of informatics methodologies, sharing of all informatics research components enables replication of research results and expedient scientific progress. CI is well suited to provide some of the mechanisms for sharing research components that would lead to increasing scientific productivity.

In the next section, we outline a few examples of methodologies to better understand top-level and low-level data-centric and collaboration-centric activities.

2.1 EXAMPLE METHODOLOGIES

Scientific data sets can come from sensors, instruments, or visual observations, as well as from simulations and modeling efforts. Data-centric activities consist of many steps that follow the top-level informatics flow shown in Figure 1 and methodologies of interest to a particular group of scientists. The informatics flow presented in Figure 1 is elaborated in the work by

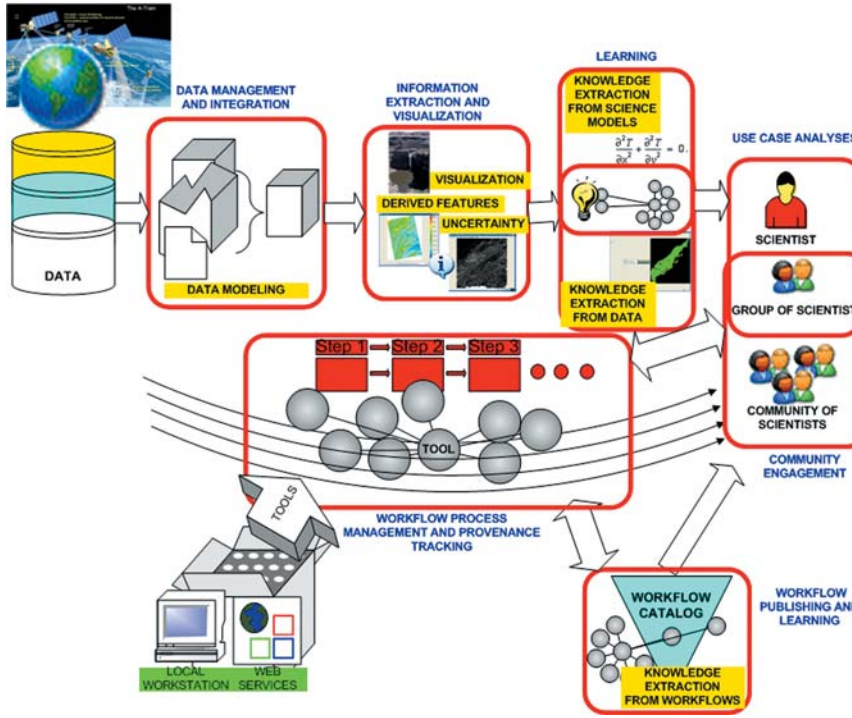

Fig. 2. An example of top-level data-centric and collaboration-centric activities.

Bajcsy (2006). It consists of data understanding, data representation, registration and georeferencing, data fusion, feature extraction, feature selection, analysis, and decision support components. For interested readers, each of these components is described in a separate chapter of the book (Bajcsy 2006), Section IV.

Figure 2 shows another description of the top-level data- and collaboration-centric activities that might be found in water research and could be used to motivate CI requirements. The figure presents a sequence of processing steps that start with data management and integration, information extraction, and visualization, followed by machine learning (data-mining and pattern recognition types of analyses and syntheses). The execution of these processing steps is supported by a workflow that links computational tools residing on local or remote computers, manages executions, and tracks provenance information about exploratory investigations. Multiple workflows are stored, re-used, and re-purposed (modified) whether for publication or learning purposes. The users performing these data-centric activities are individual researchers, groups of investigators, or communities of scientists and practitioners while sharing data, tools, and workflow catalogs.

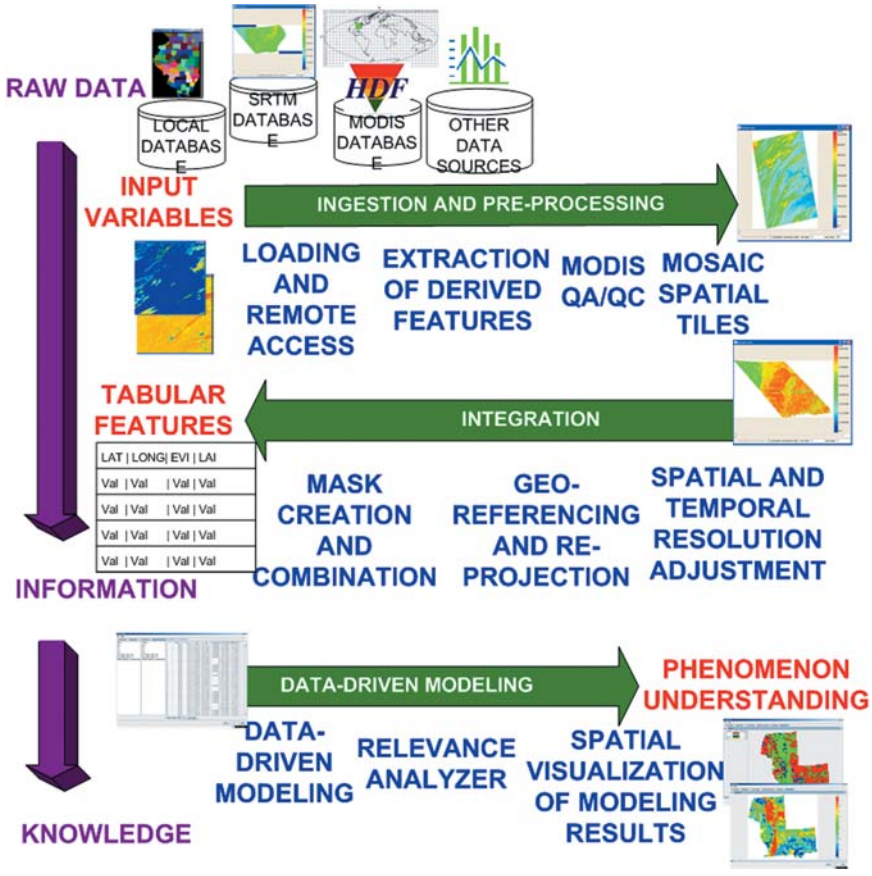

Fig. 3. An example of low-level data-centric activities for data-driven analyses of spatially varying relevance of input (independent) variables with respect to output (dependent) variables.

The top-level scientific activities can be mapped into multiple specific low-level scientific activities as illustrated in Figures 3 and 4. Figure 3 shows an example of a sequence of data-driven analysis steps. In the example, the spatial variance of greenness is studied, where climate and terrain characteristics are the independent variables in the analyses. One would like to document qualitatively and quantitatively why precipitation might be the most relevant variable for predicting vegetation in Florida while slope might be the most important in Colorado. In this case, the low-level data-centric activities in Figure 3 start with gathering, accessing, and loading remote sensing and airborne imagery, and ground measurements and vector files defining regions of interest (e.g., eco-regions). A subset of variables, such as slope, curvature, or aspect, might have to be extracted from loaded elevation maps, or categories of land use and land cover might be extracted by clustering satellite/airborne imagery. Another subset of remotely sensed

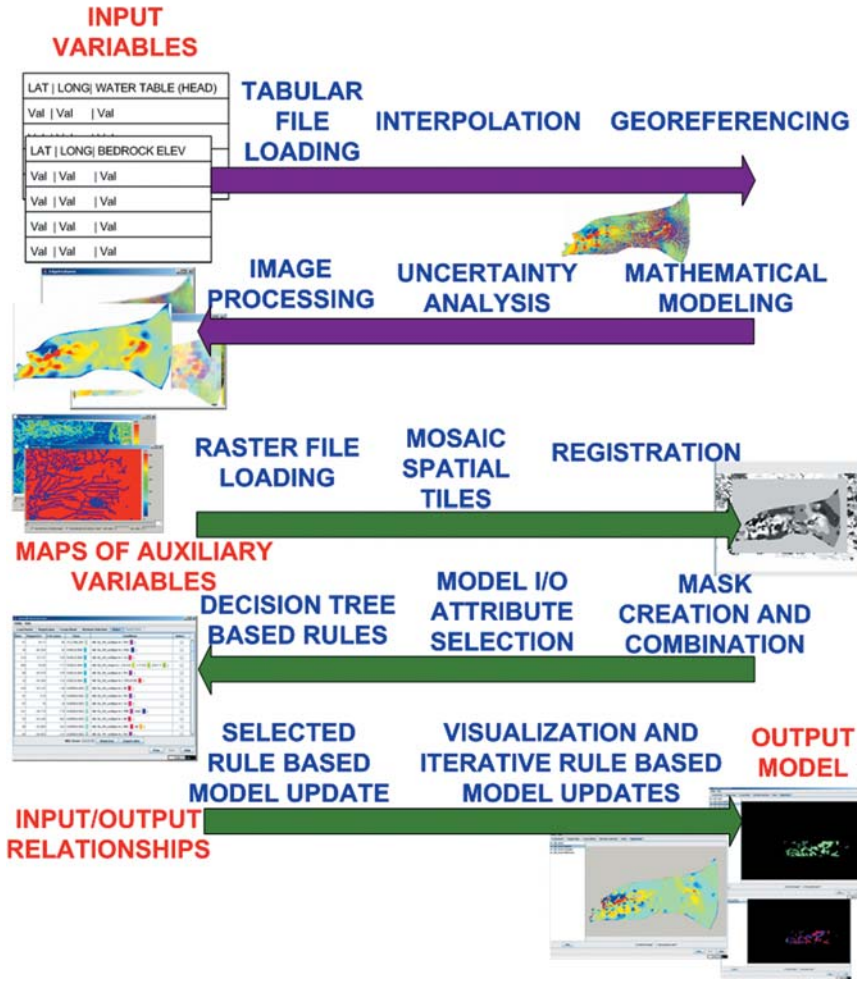

Fig. 4. An example of low-level data-centric activities for optimal integration of physics-based (mathematical) model predictions with auxiliary variables absent in physics-based models but relevant to the underlying phenomena.

images might have to go through quality control (QA) and quality assurance (QA) because cloud coverage during the acquisition impacts the quality of measurements. It should be mentioned that the sub-setting step in Figure 3 might involve integration with other data sets, as well as additional masking operations to obtain any desired subset of measurements. The additional integration might be with vector files defining eco-regions or a set of water stations, with land use maps to select locations meeting land category requirements, or with elevation maps to select locations meeting continuous variable requirements. A particular implementation of this methodology and the choices for incorporating all functionality could

be found in the articles by Bajcsy et al. (2006a, 2007). The collaboration-centric activities are hidden not only in data and software sharing for testing multiple hypotheses using the same methodology (e.g., for vegetation greenness prediction or algal biomass prediction; Bajcsy et al. 2006b, 2007) but also in sharing ideas and expertise across multiple scientific disciplines.

Figure 4 shows another example of low-level scientific activities. In this case, the objective is to perform optimal integration of physics-based (mathematical) model predictions with auxiliary variables absent in physics-based models but relevant to the underlying phenomena. Similar to the previous example, the sequence can be explained by an example-scientific problem relevant to modeling groundwater recharge/discharge (R/D) rates. It is well know that R/D rates can only be derived from a set of indirect measurements that are costly. The objectives of groundwater scientists are to minimize the cost of measurements and to maximize the accuracy of R/D rate predictions. These objectives could be achieved by incorporating auxiliary variables, and by optimizing data-preprocessing parameters in conjunction with measurement sampling. To meet these objectives, the data-driven activities in Figure 4 start with preparation of input variables, such as hydraulic conductivity, water table elevation, and bedrock elevation. In the flow, image processing is applied to compensate for the artifacts introduced by physics-based modeling (e.g., noise removal due to the accuracy limitations of the physics-based model at fine spatial resolutions) and the rule-based, machine learning (or data-driven) relates auxiliary variables with R/D rate predictions as described in the previous example. The last set of data-centric activities (bottom arrow in Figure 4) is to select subsets of derived rules, apply them to old R/D rate prediction, and visualize the new R/D rates. These steps can be repeated according to the expert's visual assessment of R/D rate accuracy and expert's considerations of plausibility of machine-derived rules. The optimal set of interpolation, georeferencing, image processing, masking (sub-setting), and decision tree parameters can be obtained by running the entire analysis multiple times, and by including the cost and measurement sampling variable into the analysis. Although the informatics methodology helps with extracting the relationships among auxiliary variables and the predicted R/D rate (if-then rules in this case), the most accurate result is likely obtained by a collaboration of multiple experts. Thus, the problem of sharing input data, software, parameter settings, and final results (R/D rate maps and applied rules) remains to be worked out among the collaborators. More details on the methodology described above and some of these early analyses can be found in the articles by Lin et al. (2006, 2007).

3 Challenges in Supporting Informatics Methodologies

The next section concentrates on general challenges related to research on water resources. When working with scientific data, common challenges

would be typically related to (i) sharing of potentially large volumes of data, (ii) computational and bandwidth requirements related to data volumes and streaming data rates, (iii) heterogeneity of data, software, and hardware technologies, (iv) management of time-critical data exchanges and analysis execution, (v) curation of data and preservation of scientific analyses, and (vi) complexity of data–software–hardware interfaces during data sharing, interactive data manipulations, and configurations of data-driven analyses. These challenges might occur at different time and location when scientists perform informatics tasks.

The challenges in performing informatics tasks lead to requirements on CI. The requirements are clearly motivated by the need to remove the burden from a single scientist or a team of scientists, and increase individual and collective scientific productivity. For example, given sufficient computational, storage, and networking resources, scientists should not be concerned with data volume and data rate because the software for CI would dynamically allocate resources on demand (Foster and Kesselman 1999; Karo et al. 2001). In addition to seamless allocations of computational and networking resources, software for CI must address the above data-centric challenges by automation. Automation could be applied to common data-centric activities (like integration and curation) in order to increase the productivity of scientists. When common data-centric activities have any time-critical aspect, then a new requirement on job scheduling arises and is addressed based on available resources with synchronous and asynchronous executions. Finally, the user interfaces to data catalogs, to implementations of multiple methodologies, and to services provided with execution of methodologies are critical in building CI. The easy-to-use interfaces lower the adoption barriers in water research and other scientific fields, as illustrated in the articles by Kooper et al. (2007) and Marini et al. (2007).

As the informatics methodologies are executed, one might neglect the challenges of conducting exploratory science. It is very common in earth sciences to explore historical data, global geospatial changes over time from spatially and temporally partitioned data, indirect geospatial variables, or data with heterogeneous geospatial, temporal, and spectral sampling rates. In the majority of these exploratory studies, there is an inherent uncertainty about the source of data, acquisition parameters, metadata information (e.g., about geo-referencing), variable transformations (e.g., indirect variables or un-calibrated variables), or 'approximating' transformations (e.g., interpolations or extrapolations in space, time, and spectrum). Thus, exploratory studies frequently involve verification, validation, and confidence evaluations of the obtained results. In addition, many of the exploratory studies are about comparing multiple implementations of methodologies, a range of parameters associated with software execution or data products. One could also view these activities as computer-assisted decisions and the associated software as computer-assisted decision support.

Driven by the exploratory science, one of the requirements for building software for CI is the integration of data products, and various information sources into data catalogs with an easy-to-use user interface to support computer-assisted scientific explorations and discoveries. The challenges of building a water research-related data catalog with a simple portal gateway could be illustrated by the Hydrologic Information System (HIS) (Maidment 2005, 2008). The design and implementation of HIS as a CI component supporting exploratory science takes a tremendous CUAHSI community effort. The CUAHSI community (originally consisting primarily of hydrologists) is also evolving together with the communities of environmental engineers (CLEANER and later WATERS Network) and the engineers interested in the coupling between chemistry, biology, and geology at the surface of the earth (Critical Zone Exploration Network[4]). The community effort is about reaching consensus on data sharing, data formats, data types, locations of data catalog components, long-term data stewardship, and preservation issues. Similarly, the implementation of HIS, software and hardware upgrades, maintenance and support of access/retrieval features for heterogeneous and large volumes of data involve a community of computer-savvy technologists on its own.

4 From Informatics Methodologies to Cyberinfrastructure for Water Research

4.1 COMMON DATA-CENTRIC ACTIVITIES

Based on the examples of data-centric activities in the presented methodologies, one could immediately identify a few common steps for scientists that could be addressed by CI. The common steps frequently include the raw data to information operations according to our definitions of raw data and information described in the Introduction section.[5] These operations could be divided into activities related to raw data (i) access, (ii) curation, (iii) integration, and (iv) selection, followed by information visualization and inspection. We have provided examples of remote data access, QA/QC (curation) operations, interpolation and conversions supporting integration, and mask-based sub-setting operations supporting selection in section 2.1.

The raw data to information operations could be also classified as (i) representation, (ii) conversion, (iii) validation/verification, and (iv) relationship operations. The operations from the first two categories (i and ii) aim at resolving different dissemination forms of raw data (e.g., single file or spatial tiles, single or distributed hardware), raw data formats, organizations in data formats, data structures representing loaded data, and variable physical units. The operations from the last two categories (iii and iv) answer queries about quality, spatial/temporal/spectral relationships of entities, and variable relationships. These operations often support data/information selection queries needed for browsing and exploration of

observations. It should be mentioned that in order to support selection queries, one has to combine operations from multiple categories. For instance, sub-setting a raster image by a selected eco-region boundary would be supported by conversion of eco-region boundary points to the coordinate system of the raster image and finding the inside/outside relationship for each image pixel with respect to the eco-region boundary.

Although the raw data to information operations are excellent first round candidates for inclusion into CI, there are other data-centric operations that are widely used in water research and would qualify for inclusion into CI in the second round. These operations are needed for data-driven modeling ranging from simple statistical correlations to complex machine learning models of phenomena (Bowdena et al. 2005a,b). There is also a gap in bridging theoretical models with data-driven models. If data-centric activities for bringing together mathematical models with data-driven models would be supported by CI, then many of the theoretical simulation models could be validated, updated, and expanded as the phenomena are changing over time.

4.2 SOLUTIONS SUPPORTING COMMON DATA-CENTRIC ACTIVITIES

While we have identified common data-centric activities, it is hard to find CI solutions for water research that would meet the expectations of data-centric activities. Tools for executing data-centric activities can be found in many software packages, either commercial or open source codes. The existing codes are written in multiple programming languages and compiled for specific operating system platforms. The data formats and code heterogeneity require additional conversion tools to support execution of sequences of data-centric activities that might not exist. The middleware[6] software solutions that would enable creation, execution, and modification of these sequences, also called workflows or scripts, are probably the closest CI solution for supporting data-centric activities. The challenges of creating workflow environments (or process management systems) have been documented in the NSF workshop on Challenges in Scientific Workflows (Gil et al. 2007).

Process management systems and specifically scientific workflow management systems have been originally designed for automation of procedures, linking stand-alone codes and creating flows of data governed by rules. The simplest workflows are scripts of batch files that describe a sequence of tool executions. Driven by complex informatics problems and various methodologies, there are several dimensions along which current workflow technologies have grown (Bajcsy et al. 2005). These dimensions include hierarchical structure and organization of software, heterogeneity of software tools and computational resources, usability of tool and workflow interfaces (e.g., workflow by example), community sharing of workflow and data fragments and publications, user friendly security and provenance,

built-in fault-tolerance, etc. It is apparent that workflow environments address immediately several of the informatics requirements described in previous sections. There is a plethora of existing workflow technologies, although the features of workflows designed for scientific communities are different from those designed for business communities. The primary difference is in responding to the large data volume and data rate requirements in sciences. Among the available scientific workflows, one could list CI (Kooper et al. 2007), Kepler (Ludäscher et al. 2006), D2K (Welge et al. 1999), OGRE, Ensemble Broker (Alameda et al. 2006), ArcGIS ModelBuilder (ESRI ArcGIS), SciFlo, DAGMan, CCA,[7] or Taverna.[8] One could also consider several software packages for scientific computations as workflow environments, for instance, Matlab, Mathematica, Python, or R software.

One should be aware of the fact that the level of technical knowledge in the water research community can greatly vary among its members. For example, in a survey on technology user adoption conducted for the WATERS cyberinfrastructure plan (Finholt and Van Briesen 2007), Excel was the most popular software (88%) followed by ArcGIS and Matlab as the most commonly used software packages. Despite the long list of scientific workflow management system, none of them was mentioned in the survey. We believe that the reasons lie in (i) the difficult integration of any software tool into a workflow environment, (ii) the lack of support for scripting together stand-alone tools running on multiple operating systems, (iii) the missing 'glue' software to convert outputs of one tool into compatible input formats of another tool, and (iv) the use of inappropriate metaphors for workflow composition that do not actively support the end user (Marini et al. 2007). In general, it is not obvious at this point how much each of the above reasons contributes to the low adoption of workflow systems in water communities.

The pros and cons of the existing workflows can be directly related to the above reasons of low technology adoption and to the spectrum of functionalities each workflow can offer. For example, Kepler or D2K would focus on a visual programming paradigm for creating workflows (drag and drop exiting tools, and then link them into a sequence) and the execution of a final workflow. In comparison, CI would focus on an exploratory paradigm for creating workflows (add one tool to a workflow and explore the outcome of executing the tool) and the execution of step-by-step workflows, where one step could be as complex as defined by a user. Similarly, while most of the scientific workflow systems can create workflows from Web services and stand-alone applications, none of them has solved the problem of automatic conversions of outputs of one tool into compatible input formats of another tool. This problem is typically addressed by saving and reading inputs/outputs to and from disk in a standard format, which introduces a computational overhead. Unfortunately, commercial software packages with workflow support, such as Matlab,

Mathematica, or Adobe Photoshop, provide only very limited support for including any external codes and rarely come with data structure conversion utilities to enable easy linking between the 'in-house' commercial code and the external code. On the other side, commercial software packages offer a large library of functionalities that have a high value for doing complex analyses. Thus, the choice of a technology for performing data-centric activities depends on the level of technical (software and hardware) knowledge of a user and the functionalities needed for conducting the activities. As long as suppliers of software use propriety (and often undocumented) formats to store information (the problem of open standards), conversion and integration will inherently remain a problem and complex data-centric activities will have to be supported by multiple software packages preferably with process management capabilities.

4.3 COMMON COLLABORATION-CENTRIC ACTIVITIES

In contrary to data-centric activities, it is much harder to reach an agreement on the common collaboration-centric activities supporting the execution of the presented informatics methodologies. Any kind of collaboration requires communication and at least some level of data/software/hardware sharing. However, a true collaboration involves sharing ideas and human resources that are the key in conducting novel science. The ideas and their implementations are also used as the merit criteria in evaluating scientists. Thus, sharing of ideas, data, software, and so on, is a very sensitive activity where trust and fairness play the key roles. Therefore, the common collaboration-centric activities should always be considered in terms of privacy and ownership of the activities while enabling community-wide collaborations. These might be seemingly contradictory considerations but should be handled by CI.

Enabling collaboration-centric activities is about designing software for social interactions that cover a broad range of generations, expertise, and seniority of scientists in multiple communities characterized by their own cultures. In the digital era, the Internet has become one of the communication media in addition to the more standard face-to-face meetings, phone conferences/meetings, or more advanced WebEx-based meetings with shared computer desktops or the ACCESS Grid-based meetings[9] with shared video and audio streams. We focus on common collaboration-centric activities which would be supported over the Internet since those could be supported by CI. The Internet technologies are not stagnant and have been evolving over the past decades significantly. With the Web 2.0 Internet technologies, the Web would change from static file sharing and browsing communication medium to interactive communication medium. For collaboration-centric activities, wikis and science portals (or gateways) have been viewed as the medium for scientific communication on the Internet.

Wiki[10] is an interface to open editing of Web pages using any Web browser. It is software running on a server with security configurations allowing multiple access hierarchies. Wikis are considered as predecessors of science portals and are very useful for content creation and sharing over the Internet. In comparison, science portals are the points of access on the Internet through which information and services are delivered to a user (a client) from central or distributed computational resources (servers). Although there exist multiple definitions of portals in the literature (Daigle and Cuocco 2002; Jones et al. 2006),[11] the definition above includes multiple purposes of portals. For example, the purpose of sharing supercomputing resources like the Tera-Grid resources, or the purpose of sharing scientific publications, presentations, data and metadata like in many digital library systems, or the purpose of communication via email, blog, chat room, or Skype (video and audio). The agreement on common collaboration-centric activities supported by the gateways has to be reached within each community based on its needs, financial resources, and unwritten social communication protocols.

4.4 SOLUTIONS SUPPORTING COMMON COLLABORATION-CENTRIC ACTIVITIES

Supporting collaboration-centric activities in gateways can be approached by adopting several wiki or commercial portal solutions, for example, Liferay, Blackboard, Campus Ads, Jenzabar, ORACLE PeopleSoft, GridSphere[12] or uPortal. In the realm of scientific portals, MyExperiment.org[13] adopted some of the successful models developed by popular social networking Web site such as Facebook.com and MySpace.com, to provide a way for scientists to share personal profiles, create collaborations on the fly, and share workflow representations. Nanohub.org[14] mixes delivery of teaching material with access to desktop applications inside the browser via the use of virtual machines and VNC[15] At Nanohub.org, the end user can easily access existing applications without worrying about setting up the software environment. These technologies have matured to provide scalability with the number of users, support for limited data-centric activities and customizable user interfaces for data sharing and browsing. However, they do not support currently executions of informatics methodologies as needed in exploratory science.

One example of such a portal-based system in water research is the already mentioned Hydrologic Information System. Hydrologic Information System is designed as 'a combination of hydrologic data, tools and simulation models that supports hydrologic science, education and practice' (Maidment 2005, 2008). Similarly, the environmental engineers envisioned 'an observatory system that will engender a community-wide shift towards systems-level, CI-enabled research, global research coordination, and bi-directional integration of experiment and simulation' (Finholt and Van Briesen 2007). The environmental engineering community has used the CyberCollaboratory portal (Liu et al. 2007) for collaborations which is built on top of Liferay.

It promotes the role of contexts (social context, geospatial context, provenance, etc.) and the use of the Resource Description Framework (Beckett 2004) to enable standard-compliant sharing of data and metadata. Traditional science gateways such as the Tera-Grid User Portal[16] focus on providing access to data and computational resources and 'usually do not provide extensive social networking interaction or social context' (Liu et al. 2007).

The final remark on supporting common collaboration-centric activities is the realization about the digital format of Web-based communication. Due to the digital format, any Web-based communication can be easily recorded. All researchers and educators agree that automated recording of collaboration-centric activities for educational purposes is highly desirable, as it makes the education process more efficient and provides a provenance trail about the collaborative effort. The portal-based systems should be leveraged when education-driven collaborations are fostered. On the other side, the privacy and ownership of the research-oriented collaboration-centric activities are often of concern to collaborators and, therefore, should be enforced when setting up science portals.

5 Conclusion

In the article, we focused on the current development of CI for water research and the need to drive the CI development partially by informatics methodologies. We argued that in order to increase the scientific productivity in water research, there is a perspective on CI as being the means for supporting daily scientific activities following informatics methodologies. These activities were decomposed into data-centric and collaboration-centric activities. They were illustrated by presenting a few examples of informatics methodologies and their top- and low-level views on corresponding activities. The requirements on CI were derived by exploring challenges in supporting informatics methodologies and by searching for common data-centric and collaboration-centric activities cutting across many informatics methodologies. We believe that this analysis and the future analyses of a similar nature would drive the development of CI functionality in a cost-efficient way. Finally, we outlined several CI components that could be the most appropriate candidates for supporting informatics methodologies. The contribution of this article is in illuminating the CI development from a perspective that bridges daily activities of many water research scientists with the CI components and functionality.

In addition to the presented work, we would like to make a final remark about the distributed nature of CI development and the need for shared CI development due to the limited financial and human resources. With this in mind, it is critical to understand what design principles should govern developments of new software for CI, what software technologies are currently available for building virtual observatories, and how the features of these technologies meet the informatics requirements.

Understanding of the above topics would be the next steps in developing CI driven by informatics methodologies.

From a computer science perspective, the CI development has to overcome the challenges of technology adoption. There are (i) fundamental computer science problems, for example, related to linking and executing heterogeneous software running on multiple operating systems, (ii) limiting specifications and availability of infrastructure (e.g., bandwidth, speed of processors, size of computer memory), and (iii) user interface and access to CI functionalities and capabilities. For instance, one of the current trends is to design Web services (uniform programmatic interfaces) to all existing computational functionalities, orchestrate Web services as workflows and then perform virtualization of operating systems and storage on servers where the Web services are executed. Virtualization refers to the fact that a piece of hardware would run multiple operating system images at the same time and multiple storage systems. This future direction would support data-centric and collaboration-centric activities on the network of servers (also denoted as cloud computing).

Short Biography

Peter Bajcsy is currently with the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign, Illinois, working as a research scientist on problems related to automatic transfer of image content to knowledge within the framework of X-informatics, where X stands for hydro, geo, remote sensing, bio, medical image, and advanced sensing. Bajcsy's scientific interests include image and signal processing, statistical data analysis, data mining, pattern recognition, novel sensor technology, and computer and machine vision. He holds an MS degree from the University of Pennsylvania and a PhD degree from the University of Illinois at Urbana-Champaign (UIUC). He is also an adjunct assistant professor in CS and ECE Departments at UIUC, and an associate director of the Center for Humanities, Arts and Social Sciences in Illinois Informatics Institute at UIUC.

Notes

* Correspondence address: Peter Bajcsy, National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, 1205 W. Clark Street, Urbana, IL 61801, USA. E-mail: pbajcsy@ncsa.uiuc.edu.

[1] http://www.cuahsi.org/
[2] http://cleaner.ncsa.uiuc.edu/home/
[3] http://www.watersnet.org/index.html#
[4] http://www.czen.org/
[5] Raw data come directly off the sensor. Information represents the meaning of raw data that is narrow in scope and it has a simple organization.
[6] The term 'middleware' should be understood as software that translates, integrates and/or mediates connections between two applications.
[7] http://www.cca-forum.org/

[8] http://taverna.sourceforge.net/
[9] http://www.accessgrid.org/
[10] http://wiki.org/
[11] Although IBM Global Education Industry (2000) *Higher education portals: presenting your institution to the world* (see Sue Hoffman, IBM brings strategic business intelligence technology to the college campus, IBM Press Room, Nashville, TN, 10 October 2000; http://www-03.ibm.com/press/us/en/pressrelease/1523.wss, last retrieved on 2 October 2008) is viewed as the original introduction of portal technologies for universities by IBM, the publication has not been preserved since it was an industrial presentation of a new IBM product similar to the follow-up presentation at the EDUCAUSE conference (An EDU Odyssey, 28–31 October 2000, Indianapolis, IN, http://net.educause.edu/EDUCAUSE2001/1341; see http://net.educause.edu/content.asp?page_id=1411&MODE=SESSIONS&Heading=Corporate%20Presentations&Product_Code=E01/CORPPRES%25&Meeting=E01&bhcp=1).
[12] http://www.gridsphere.org/gridsphere/gridsphere
[13] http://myexperiment.org
[14] http://www.nanohub.org
[15] VNC stands for Virtual Network Computing and allows controlling another device from a personal computer.
[16] http://portal.teragrid.org/

References

Abbott, M. (1991). *Hydroinformatics, information technology and the aquatic environment*. Aldershot, UK: Avebury Technical.

Alameda, J., et al. (2006). *Ensemble broker service oriented architecture for LEAD*. The 22nd International Conference on Interactive Information Processing Systems for Meteorology, Oceanography, and Hydrology.

Ankolekar, A., et al. (2007). *The two cultures: mashing up web 2.0 and the semantic web*. The 16th International Conference of World Wide Web, Banff, Alberta, Canada.

Atkins, D. (2003). *Revolutionizing science and engineering through cyberinfrastructure: report of the national science foundation blue-ribbon advisory panel on cyberinfrastructure*. Technical Report.

Bajcsy, P. (2006). Data processing and analysis. In: Kumar, P., et al. (eds) *Hydroinformatics: data integrative approaches in computation, analysis, and modeling*. Boca Raton, FL: Taylor & Francis, pp. 258–378.

Bajcsy, P., et al. (2004). Survey of bio-data analysis from data mining perspective. In: Wang, J. T. L., et al. (eds) *Data mining in bioinformatics*. London: Springer Verlag, pp. 9–39.

——. (2005). *A meta-workflow cyberinfrastructure system designed for environmental observatories*. Technical Report: NCSA Cyber-environments Division.

——. (2006a). *GeoLearn: prediction modeling using large size geospatial raster and vector data*. EOS Trans. AGU 87 (52), Fall Meeting Suppl., Abstract IN41C-06.

——. (2006b). *Visualization and data mining tools applied to algal biomass prediction in Illinois streams*. The 7th International Conference on Hydroinformatics, Nice, France, pp. 926–933.

——. (2007). *GeoLearn: an exploratory framework for extracting information and knowledge from remote sensing imagery*. The 32nd International Symposium on Remote Sensing of Environment Sustainable Development through Global Earth Observations, San Jose, Costa Rica.

Balazinska, M., et al. (2007). Data management in the worldwide sensor web. *IEEE Journal on Pervasive Computing* 6, pp. 30–40.

Beckett, D. (2004). *RDF/XML syntax specification (revised)*, W3C Recommendation [of 10 February 2004]. [online]. Retrieved on 1 September 2008 from http://www.w3.org/TR/rdf-syntax-grammar/

Bowdena, G. J., Dandyb, G. C., and Maier, H. R. (2005a). Input determination for neural network models in water resources applications. Part 1: background and methodology. *Journal of Hydrology* 301, pp. 75–92.

Bowdena, G., et al. (2005b). Input determination for neural network models in water resources applications. Part 2. Case study: forecasting salinity in a river. *Journal of Hydrology* 301, pp. 93–107.

Daigle, S. L., and Cuocco, P. M. (2002). Portal technology opportunities, obstacles, and options: a view from the California State University. In: Katz, R. N. (ed.) *Web portals and higher education technologies to make IT personal*. San Francisco, CA: Jossey-Bass, pp. 109–123.

Finholt, T., and Van Briesen, J. (2007). *WATERS Network Cyberinfrastructure Plan: the WATERS Network Project Office Cyberinfrastructure Committee*. [online]. Retrieved on 2 October 2008 from http://www.watersnet.org/docs/CyberinfrastructurePlan.pdf. Also available at http://www.watersnet.org/plngdocs.html

Foster, I., and Kesselman, C. (1999). *Computational grids. The grid: blueprint for a new computing infrastructure*. San Francisco, CA: Morgan-Kaufman.

Gil, Y., et al. (2007). Examining the challenges of scientific workflows. *Computer* 40, pp. 24–32.

Gupta, V. K., Troutman, B. M., and Dawdy, D. R. (2007). Towards a nonlinear geophysical theory of floods in river networks: an overview of 20 years of progress. In: Tsonis, A. A. and Elsner, J. B. (eds) *Nonlinear dynamics in geosciences*. New York: Springer, pp. 121–151.

Jones, N. B., Provost, D. M., and Pascale, D. (2006). Developing a university research web-based knowledge portal. *International Journal of Knowledge and Learning* 2 (1/2), pp. 106–118. [online]. Retrieved on 2 October 2008 from http://inderscience.metapress.com/media/n9a4b9f0np4kpgba6r8x/contributions/7/k/9/0/7k90fxcwdxkda65v.pdf

Jonoski, A., and Popescu, I. (2004). *The hydroinformatics approach to integrated river basin management*. BALWOIS International Conference on Water Observation and Information Systems for Decision Support, Ohrid, Republic of Macedonia.

Karo, M., et al. (2001). *Applying grid technologies to bioinformatics*. The 10th IEEE International Symposium on High Performance Distributed Computing.

Kirchner, J. W. (2006). Getting the right answers for the right reasons: linking measurements, analyses, and models to advance the science of hydrology. *Water Resource Research* 42, W03S04, doi:10.1029/2005WR004362. [online]. Retrieved on 2 October 2008 from http://www.agu.org/pubs/crossref/2006/2005WR004362.shtml

Kooper, R., et al. (2007). *CyberIntegrator: a highly interactive scientific process management environment to support earth observatories*. Geoinformatics Conference, San Diego, CA.

Kumar, P. (2007). Variability, feedback, and cooperative process dynamics: elements of a unifying hydrologic theory. *Geography Compass* 1, pp. 1338–1360.

Lin, Y.-F., et al. (2006). *Development of point-to-zone pattern recognition and learning utilities for groundwater recharge and discharge estimation*. The Geological Society of America 2006 Annual Meeting, Philadelphia, PA.

——. (2007). *Evaluation of alternative conceptual models using interdisciplinary information: an application in shallow groundwater recharge and discharge*. AGU, Fall Meeting Suppl., Abstract H31G-0738.

Liu, Y., et al. (2007). *Towards a rich-context participatory cyberenvironment*. International Workshop on Grid Computing Environments 2007, Supercomputing Conference 2007, Reno, NV.

Ludäscher, B., et al. (2006). Scientific Workflow Management and the KEPLER System. *Concurrence and Computation: Practice and Experience* 18 (10), pp. 1039–1065.

Maidment, D. R. (ed.) (2005). *Hydrologic information system status report, version 1*. Consortium of Universities for the Advancement of Hydrologic Science. [online]. Retrieved on on 2 October 2008 from http://www.cuahsi.org/docs/HISStatusSept15.pdf

——. (2008). *CUAHSI hydrologic information system: overview of version 1.1*. Consortium of Universities for the Advancement of Hydrologic Science, 12 July. [online]. Retrieved on 2 October 2008 from http://his.cuahsi.org/documents/HISOverview.pdf

Marini, L., et al. (2007). Supporting exploration and collaboration in scientific workflow systems. Eos Trans. AGU 88 (52), 2007 AGU Fall Meeting Suppl., Abstract IN31C-07., December 10–14, San Francisco, CA.

Marlin, J. C., and Darmody, R. G. (2005). Returning the soil to the land, the mud to parks project. *Illinois Steward* 14, pp. 11–18.

Myers, J. D., et al. (2003). Re-integrating the research record. *IEEE Computing in Science and Engineering* 5(3), pp. 44–50.

National Science Foundation. (2006). *NSF's cyberinfrastructure vision for 21st century discovery, NSF cyberinfrastructure council*. Washington, DC: National Science Foundation.

Newman, B. D., et al. (2006). Ecohydrology of water-limited environments: a scientific vision. *Water Resource Research* 42, W06302.

Price, R. K., Solomatine, D. P., and Velickov, S. (2000). *Internet-based computing and knowledge management for engineering services*. The 4th International Conference on Hydroinformatics, Iowa, USA.

Spencer, B. F. Jr., et al. (2006). *Cyberenvironment project management: lessons learned*. 5 September. [online.] Retrieved on 2 October 2008 from http://www.nsf.gov/od/oci/CPMLL.pdf

Team, W. M. (2008). Draft science, education, and design strategy for the WATER and environmental research systems network. WATERS Network February 27, 2008.

Wagener, T., et al. (2007). Catchment classification and hydrologic similarity. *Geography Compass* 1, pp. 901–931.

Welge, M., et al. (1999). *Data to knowledge (D2K): a rapid application development environment for knowledge discovery in database*. Technical Report, National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Champaign, IL.