Knowledge management in the Earth Systems Sciences: Case study of the co-evolution of watershed and ecosystem patterns

Lawrence Band, Ken Galluppi, Reagan Moore, Arcot Rajaseker, Brian Miles, Antoine de Torcy, University of North Carolina, Chapel Hill

Overview: This white paper addresses the development of earth system science theory within a knowledge management framework. We use a case study focusing on the co-evolution of watersheds and ecosystems, extending from a base of observations, monitoring and experimentation, through information management, modeling and theory development. This case study topic requires the integration and construction of interdisciplinary knowledge from a set of different scientific traditions, observations, analysis and simulation approaches. Explicit and tacit knowledge is included in complex workflows combining information from distinct sources and data models. Cyberinfrastructure development can formalize and streamline data, information and model integration, allowing earth system scientists to leverage consistent, well documented and shareable resources and tools to more efficiently develop and test new theory.

Co-evolution of watersheds and ecosystems has received increased interest and attention over the last decade (e.g. NRC 2010) and integrates coupled development of geomorphic, hydrologic and ecosystem patterns and processes with the interaction of atmospheric, tectonic and anthropogenic systems. From the standpoint of NSF, these fields encompass much of the range of the GEO Directorate, as well as programs in the social sciences and environmental biology. This white paper complements those of Murdoch et al (2011), "A Community Model for Water in the Continental Earth System," and Reinfelder et al (2011), "A Digital Earth's Crust for Modeling Continental Fluid Flow."

Space and time scales of interest in a fully coupled watershed-ecosystem can vary over several orders of magnitude ranging from pore and fracture scale to sub-continental dynamics, and over minutes to millions of years. Typical time and space scales in the component disciplines occupy much narrower ranges, and will set other component dynamics as constant forcing or boundary conditions (e.g. annual climate conditions) or fixed system characteristics (e.g. canopy conditions, topography, soil and rock properties). Direct measurements of energy and mass storage and flux dynamics are possible in some components (atmospheric, hydrologic, ecosystem exchange), while in others they may be indirectly inferred from geochemical or sedimentary signatures of long-term processes.

As such, very different types of instrumentation, direct measurements and inferences are designed across the component disciplines, with separate observational programs, experimental facilities and information management frameworks funded and developed to optimize specific disciplinary needs and questions. This can be viewed as either unnecessary redundancy where significant overlap exists, or a strength to be leveraged as new interdisciplinary questions spontaneously arise that can capitalize on deep knowledge and resources that have developed from the separate investments.

Interdisciplinary knowledge management: Abstracting the specific questions regarding watershed and ecosystem co-evolution as an interdisciplinary knowledge management problem; a framework leveraging human intelligence and technology needs to be developed to synthesize data, information and theory. The synthesis requires a combination of a community of scientists that develop interdisciplinary questions and insight with the leverage of cyberinformatics tools for data and information analysis, modeling and communication. Figure 1 shows a general knowledge hierarchy that extends from existing, disciplinary programs in data collection and management to interdisciplinary theory that can support understanding, prediction and decision making.

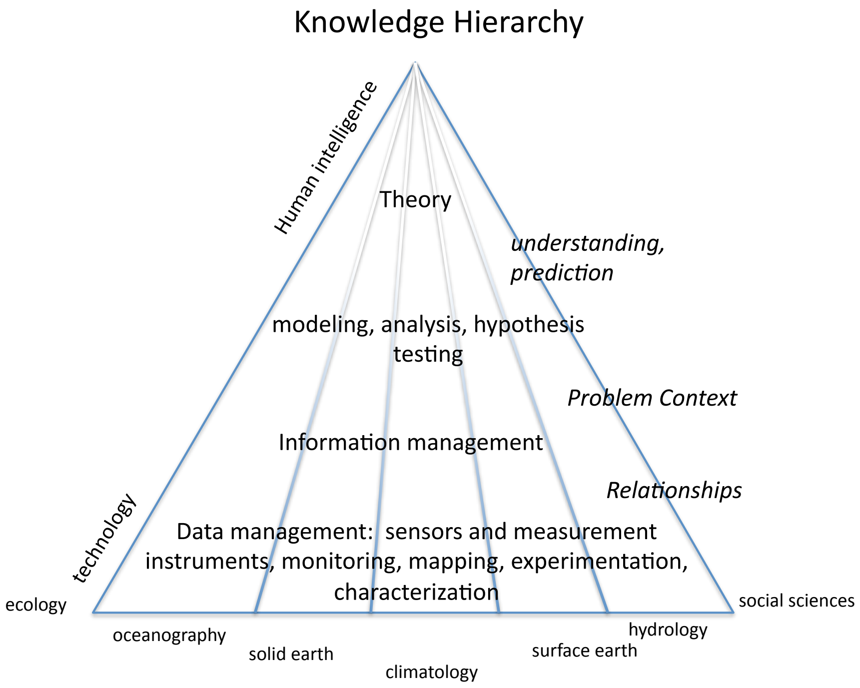

Figure 1: Knowledge hierarchy extending from basic observations and data management through the development and testing of new theory. A key Earthcube challenge is the synthesis and adaptation of existing investment in disciplinary theory and tools into coupled earth system science approaches. Knowledge management should leverage existing disciplinary investments and experience in basic data collection and management, with the goal of accelerating the delivery of higher order interdisciplinary information to earth science (and other) groups developing new theory. Modified from Ackoff (1989)

Moving up the hierarchy, primary observations and data collection from sensors, mapping or surveys (e.g. social science interviews) are collected and managed as individual or small packages of data. In the context of watersheds and ecosystems these may include soil moisture and groundwater measurements, canopy leaf area and species, soil and rock densities, biogeochemistry, mineralogy and hydraulic conductivity. The incorporation of relationships between data including those derived from position within a time series or spatial coordinates, such as drainage rates, soil and vegetation patterns, introduces information management. A problem context or hypothesis that is posed can then be developed into specific types of analyses or modeling studies, with formal hypothesis testing. In the case study we are using, this hypothesis testing may include the coupling and feedback between soil moisture, drainage rates and canopy leaf area patterns along topographic gradients, explicitly coupling water, carbon and energy cycling. Further feedback between below ground canopy carbon allocation, root strength and landslide potential begins to couple longer term soil and geomorphic dynamics. Repeated testing in different regions, new data collection, and refinement of hypotheses develop new theory, which is subject to continued testing and generalization.

Integrating disciplinary information: In the earth systems science, it is rare to develop direct observations across the range of disciplines (arrayed along the base of figure 1) from an integrated sampling or mapping design, with the exception of a few experimental observatory sites (e.g. LTER, CZO). Instead, secondary information developed by different agencies (e.g. USGS NED, NHDplus, river/stream discharge and chemistry, USDA-NRCS soils, EPA National Land Cover Dataset, NWS meteorology) or inferred from remote sensing measurement (MODIS LAI, snow cover) comprise a widely used national spatial data infrastructure. This information is collected with different spatial and temporal support, for different purposes, and is maintained and described with different data models. However, such information is routinely combined and used within earth systems models. An example would be the integration of information from 1:250,000 polygon formats for STATSGO soils information which can include up to 21 different soil types within a single polygon with the 30m (or 10m) NED elevation data for regional watershed models.

The integration of this information up the knowledge hierarchy requires explicit consideration of the information content and uncertainty in each dataset. Much of this workflow is carried out by individual research teams, using separate and often poorly documented methods which either incorporate explicit and tacit knowledge of local context and knowledge of data properties to tailor specific methods based on understanding, or by simple overlay (for spatial data) without explicit consideration of these factors. In the former case, individual or team experience in local field sites, and disciplinary knowledge of the measurements, properties and data models used to represent each data set is critical. This is time consuming, potentially error prone and non-repeatable between different groups, creating a barrier to extending purely technical (fully automatable) workflows up the knowledge hierarchy in figure 1. To advance the types of cross-disciplinary problems posed within this case study, it is necessary to develop the following capabilities within cyberinformatics tools: (1) to codify explicit human knowledge and experience as part of workflows; (2) find methods to formalize and represent tacit knowledge, and (3) to develop collaborative technologies between groups.

Case study - ecohydrological simulation: A subset of the fully coupled watershed-ecosystem coevolution problem can be found in a set of studies that have used temporally fixed topography with high resolution ecosystem and hydrologic energy, water, carbon and nutrient dynamics (e.g. Tague and Band, 2004; Ivanov et al, 2008; and others). An extension of this framework with weathering, soil formation and sediment/soil transport laws would include geomorphic dynamics but is not included here.

One model framework designed for research on the co-evolution of watershed hydrologic and ecological systems is the Regional HydroEcological Simulation System (RHESSys – Band et al, 1993; Tague and Band, 2004). This approach uses a general representation of watershed structure as a hierarchy of landforms, soils, and ecosystems connected by the coupled mass conservation of water, carbon, and nutrients (WCN). Information on topography, canopy cover, soils and hydrology of an area is processed into a nested set of ecosystem patches, hillslopes, and catchments arranged through a connected drainage network, using a workflow that also determines soil physical, and canopy physiologic properties and initial conditions (figure 2) built using information from the USGS (DEM), NASA (canopy NDVI) or other remote sensing information, USDA (soils), and other, diverse data sources. Stream gauge, discharge and chemistry time series and meteorological information can be accessed USGS, NCDC or other records (e.g. USFS, LTER sensors). At present the workflow has been developed as a set of largely manual downloads, reformatting, registration and additional spatial data processing (e.g. transformation of NDVI to LAI) using tools written into the GRASS GIS, and post-processing workflows for analysis of output time series at patch, hillslope or basin scale, spatial time slices or fully updated world files, or to send output into a linked model (landslides, water supply, fire behavior).

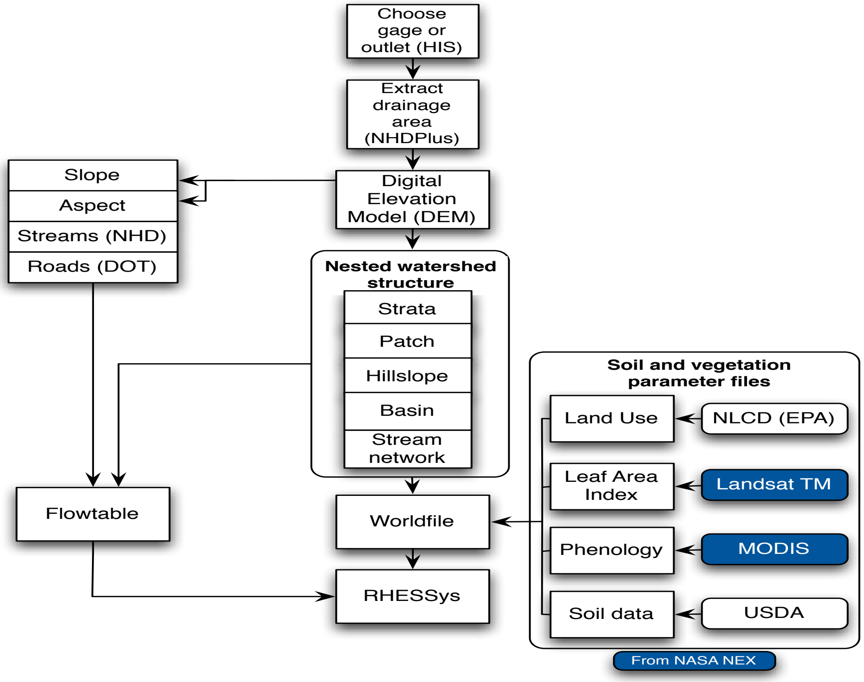

Figure 2: RHESSys workflow to develop nested watershed parameter file (worldfile) containing nested ecogeomorphic object framework, and full (initial) system state.

The nature of the model components linked for the co-evolution of transient ecosystems with the geomorphology and the distributed hydrology of the watershed has the simple constraint that the ecosystem must survive and grow, based on the balance and adjustment of above and below ground biomass, and available water, nutrients and energy. Lack of balance can be developed by unrealistic combinations of climate, soils, topography and standing biomass, which can lead to simulated ecosystem failure or significant adjustment and adaptation of the canopy so that it does not resemble key observed patterns. This is unlike other models that will compute water and carbon cycling, but do not include growth and allocation and the potential for ecosystem failure (e.g. carbon starvation). The practical consideration of this constraint is that simple overlay and combination of natural resources and remotely sensed canopy information typically needs to be augmented by specific decision points and intervention as discussed above, and local context and interdisciplinary knowledge is required. Much of this model is not fully documented and remains as tacit understanding of the linked systems based on field observations and experience. Therefore, this case study provides an example that requires significant improvement in cyberinformatics tools that will improve social interaction between disciplinary scientists, better codify research knowledge and experience, and manage and document the data-to-theory hierarchy including information provenance, access and workflows, and facilitate interoperability and integration of the separate disciplinary data and models represented in figures 1-3 .

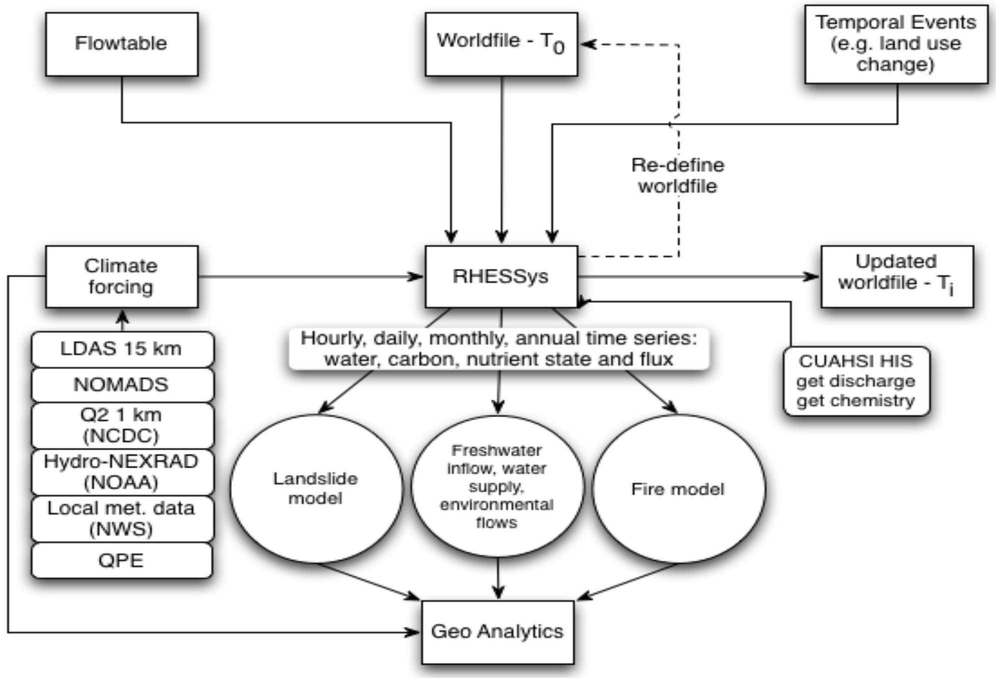

Figure 3: RHESSys workflow (cont'd) with initial worldfile (T_0), climate forcing data (LHS), observed stream discharge and chemistry (from CUAHSI HIS get calls), output of updated full system state (worldfile) and water, carbon, nutrient storage and flux time series and links to additional environmental models.

A prototype version of the workflow presented in figures 2,3 has been implemented within an iRODS framework. Moore et al (2011) have described the integrated Rule Oriented Data System (iRODS) in terms of its advantages for EarthCube goals. The iRODS data grid enables collaborative research through creation of virtual collections across disciplinary information sources that provide a unifying name space for data. The elements of the shared collection can include soft links to data in other repositories and descriptive metadata to support search. Each community develops policies that control the properties of the shared collection, including access constraints, required provenance information, and access services. The community-based policies are used to control procedures that automate administrative tasks, apply desired format transformations, and validate collection properties. The procedures are cast as workflows that accurately capture provenance information (input parameters and files, processing steps, and outcomes). The information needed to re-compute an analysis will be managed as an active object within iRODS, enabling the re-creation of scientific results when any of the underlying data sets are re-calibrated.

In this case study our prototype leverages on the development of CUAHSI Hydrologic Information System to manage all point source time series water information from USGS, NCDC and other sources (his.cuahsi.org) as well as methods developed for other data models. Advantages include integrating across the heterogeneous data models used for the different information sources in RHESSys, implementation of the workflow through the development of specific, linked micro-policies, documentation, archiving and access to the full data and information chain, and the potential to encode local site specific context and knowledge developed through experience as part of conditional workflows. Currently, iRODS manages data access, including use of CUAHSI HIS web services to identify and retrieve hydrologic point time series information, data provenance, and simple workflows to build the RHESSys world files including embedded GRASS GIS functionality, retrieval of forcing and calibration information, and execute model runs. More complete workflows are under development with key implementation steps and research questions addressing the ability to encode basic decision points currently used in manual workflows, and management of geoanalytics tools for higher order visualization and analysis of model predictions.

Generalization of cyberinfrastructure needs for earth systems science: Generalizing from the case study given here, CI is needed to support the creation, codification, and storage of knowledge, as well as making knowledge shareable by groups or organizations. It is accepted that connecting information and knowledge enables scientists to work together to create value by combining different perspectives of a problem. This value is manifested in the creation of new opportunities and the ability to respond to challenges in innovative ways. Purposely managing knowledge opens up multiple pathways for the creation and flow of information that crosses traditional boundaries of disciplines and organizations. A goal of knowledge management, therefore, is more rapid creation and sharing of knowledge, practices, and experiences for exploring new problems or probing issues in new ways. Multidisciplinary teams can therefore be incentivized to form and disband to take advantage of this value. Technologies enable and enhance connections and sharing that support scientists' ability to process knowledge, and share higher levels of knowledge such as discovery, insight, and judgment. Facilitating temporary network of scientists, in a virtual organization, helps to coordinate activities that develop a common understanding to meet a research objective.

Knowledge Networking (figure 4 - adapted from Skyrme ,1999) shows a CI supporting both the codification of knowledge, and its abstraction for sharing. Often, requisite knowledge is lost or duplicated because it is not made explicit or it is not managed. The first capability needed within CI is support for the development, application, and management of both explicit and tacit knowledge. Digital tools, therefore, are needed that help turn tacit knowledge into explicit forms such as documents, processes and practices, databases, analysis tools, and models. These tools need to also support higher levels of knowledge needs such as discovery, analytics, and inference. Cyber tools can codify knowledge in the form of policies, rules, workflows, and specific services – the example of iRODS implementation is given above. Scientists work with knowledge, both tacit and explicit, to explore issues of interest. Having a meta-type description of codification will help make knowledge discovery, accessible and reusable by scientists and students alike.

Scientists work on problems by developing a conceptual understanding or abstraction of important processes or principles. Through collaboration, knowledge is developed at higher levels of thinking but is often inhibited by the need to work with data or information layers, which takes time and resources away from science exploration. Much of this lower level work can be operationalized by CI. Tools are needed therefore to help scientists develop and communicate an abstraction for use by others without the burden imposed by information technologies. Scientists work together by discussing abstractions in order to develop a common understanding. Human social networks are created and disbanded to further science development. The second purpose of CI is to support these social networks depicted in the Skyrme diagram through collaborative strategies, the main focus of which is to deliver the right knowledge, to the right person in ways they can understand. Collaborative strategies allow virtual organization of scientists to work on a common understanding and common purpose but at a higher level of knowledge sharing and not just at data interoperability. Further, these strategies foster higher levels of commitment because they facilitate value creation through more frequent and sustained communications and exchange of tacit knowledge.

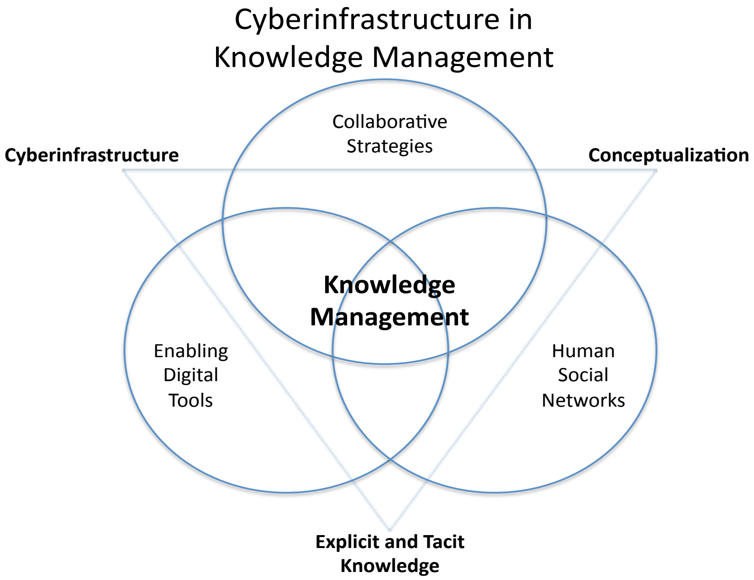

Figure 4: Integration of disciplinary knowledge in the earth sciences and allied disciplines, the conceptualization of knowledge into specific abstract "types", and the development of cyberinfrastructure to connecting individuals with collaborative strategies through facilitated networks. After Skyrme (1999).

References:

Ackoff, R. L., 1989. From Data to Wisdom. Journal of Applies Systems Analysis, Volume 16, 1989 p 3-9.

Band, L.E., P. Patterson, R.R. Nemani and S.W. Running, "Forest ecosystem processes at the watershed scale: 2. Adding hillslope hydrology." Agricultural and Forest Meteorology, 1993, v.63, p.93-126.

Ivanov, V.Y., Bras, R.L., and Vivoni, E.R. (2008). Vegetation-Hydrology Dynamics in Complex Terrain of Semiarid Areas: I. A mechanistic Approach to Modeling Dynamic Feedbacks, Water Resour. Res., 44, W03429, doi:10.1029/2006WR005588.

Moore, R.W., A. Rajasekar, M. Conway, W. Schroeder, M. Wan, 2011. White Paper: National Data Infrastructure for Earth System Science. White Paper submitted to EarthCube project (http://earthcube.ning.com/group/technology-solutions)

Murdoch et al (2011). A Community Model for Water in the Continental Earth System. White Paper submitted to EarthCube project (http://earthcube.ning.com)

NRC, 2010. Landscapes on the Edge: New Horizons for Research on Earth's Surface. National Academies of Science, NRC Board on Earth Sciences and Resources.

Reinfelder et al (2011). A Digital Earth's Crust for Modeling Continental Fluid Flow. White Paper submitted to EarthCube project (http://earthcube.ning.com)

Skyrme, David J. 1999. Knowledge Networking: Creating the Collaborative Enterprise. Butterworth-Heinemann, Boston, MA.

Tague, C.L., L.E. Band, 2004. RHESSys: Regional Hydro-Ecologic Simulation System—An Object-Oriented Approach to Spatially Distributed Modeling of Carbon, Water, and Nutrient Cycling. Earth Interactions 2004 8: 1-42.