

BERAC ASKD Challenge

The BER Advisory Committee in its report “BER Virtual Laboratory: Innovative Framework for Biological and Environmental Grand Challenges” recommends creating a BER Virtual Laboratory that would be developed by integrating and strategically expanding BER resources[1]. A key goal of the Virtual Laboratory is to transition BER’s research program from one associated with distributed datasets, specific process knowledge, and individual component models to one that provides a predictive understanding of key couplings and feedbacks among natural-system and anthropogenic processes across scales. In other words, this transition involves moving beyond the investigation of “parts” to an understanding of integrated environmental systems behavior.

Key to achieving this BER Virtual Laboratory goal is development of a BER-base User Facility: Biological and Environmental Research Knowledgebase that provides the computational curation, modeling, analytical and visualization tools needed to translate disparate and multi-scale measurements from field observatories plus process understanding from the advanced genomic and environmental measurement facilities and elsewhere into new knowledge. A logarithmic expansion of Kbase (which enables community investigation of systems biology), the BER-base would enable quantification of dynamic and multi-scale interactions among: gene and protein functions in natural systems, terrestrial system behavior, and climate systems. Many of the BERAC Virtual Laboratory goals align closely with the ASKD program and so this challenge document is extracted from the BERAC report.

BER-base data assimilation and simulation tools would aid development of the next generation of advanced theory and the integration of information spanning molecular to global scales, enabling predictions about environmental changes that will inform policy. Building upon and integrating BER’s existing knowledge discovery infrastructure, BER-base would develop strategies for linking heterogeneous databases and for federation and exchange of information obtained from field observatories, the advanced genomic and environmental measurement facilities, and other resources. The resulting environment for multiscale knowledge discovery would provide innovative capabilities in distributed data discovery, visualization, analysis, and uncertainty quantification. Also envisioned for BER-base is the development of advanced system component models.

BER-base should advance the following key characteristics:

- Database Linkages. Because a single database cannot serve all the needs of the BER community, strategies should be developed and implemented for federation and exchange of data collections among resources, including data repositories, databases, and knowledgebases. Furthermore, new approaches such as cloud computing may make seamless access to public data feasible for users and provide data and analysis tools in a scalable fashion (an

- approach used by KBase). A major challenge is determining how best to maintain and curate these databases over the long term.
- **Assimilation of Data and Knowledge into Models.** BER-base would facilitate integration of heterogeneous data and knowledge derived from that data (e.g., improved parameterizations and model parameters) into models via intuitive interfaces that advance discovery and knowledge development. The goal is to accelerate the use of these data—available from the advanced genomic and environmental measurement facilities, field observatories, climate models, and other resources—to provide a comprehensive analysis of biological and environmental systems, thereby advancing system understanding and model predictions and fidelity.
 - **Multiscale and Advanced System Component Models.** BER-base would both advance individual system component models and develop new approaches for bridging computation and natural phenomena representation across vast temporal and spatial scales, from molecular to global. The new approaches are expected to benefit from novel mathematical constructs and process-based theory and understanding. Once developed, the simulation capabilities could be used to identify the greatest sources of model uncertainty, critical thresholds and tipping points, and sensitivities in system response, all of which can, in turn, drive and prioritize process investigations and observations.
 - **Knowledge Discovery.** A key BER-base component is development of a knowledge discovery environment that provides innovative capabilities in distributed data discovery, visualization, and analysis (including uncertainty). This component will leverage existing BER investments, such as KBase.

Specific BER-base Recommendations

- **Link to Heterogeneous Databases.** Many of the most interesting and important future applications of BER datasets will require integrating terrestrial subsurface and land surface, marine, atmospheric, and biological (e.g., organisms, genomic, and molecular) information. Linking this information will involve specifically designing measurement sites appropriate for the eventual integration of datasets that cross system boundaries. Critical measurements that may not appear interesting for specific communities but could be important for understanding processes occurring at the interfaces must be identified and given priority. Common metadata standards and data collection and management protocols need to be developed and adopted, and data should be organized geospatially using geoinformatics.
- **Develop Multiscale Simulation Frameworks and Data Assimilation Tools.** These tools are needed to facilitate hypothesis testing, assimilate multiscale data, and assess many fundamental issues pivotal to sustainable environmental and energy strategies that involve processes and their couplings ranging from molecular and cellular levels to the ecosystem scale. A variety of simulation approaches are needed, depending on the question

asked, the data available, and the scales of interest. Advances are needed to assimilate heterogeneous (streaming) datasets into simulations and thus effectively use all the different types of data that are and will continue to become available for systems understanding. Although simulation frameworks may include components that vary from application to application, they likely will have common approaches, architectures, features, modules, and couplers that can be used to advance different aspects of multiscale biological and environmental system predictions.

- **Develop Advanced System Component Models.** As part of the BER-base effort, strengthening individual system component models and the linkages between different models is imperative. Examples of needed improvements include incorporating cell function into reactive transport models, coupling subsurface and watershed biogeochemical simulators to land models, and advancing DOE climate models as outlined below. First, only a few studies have attempted to incorporate mechanistic microbial function into subsurface biogeochemical reactive transport models. Examples include the use of *in silico* (Fang et al. 2011) to ecotrait-based methods (Bouskill et al. 2012). Significant efforts are needed to advance these methods to allow simulation of cellular, organismal, or community responses to environmental fluxes; their impacts on the field environment; and the effects of field-scale biogeochemical, hydrological, and atmospheric fluxes on microbial community functioning. Considerable challenges lie in (1) developing and testing such coupled models; (2) determining the level at which microbial function and functional groups can and should be represented in reactive transport models; (3) representing microbe-microbe, microbe-plant, and microbe-mineral interactions; and (4) determining how to parameterize microbial functioning in a tractable yet representative manner. Second, capabilities for simulating water flows within a watershed have improved over the last decade in the hydrological community. However, the development of mechanistic reactive transport models that also can simulate biogeochemical reactions coupled to these flows up to the watershed scale is in the early stages. Moreover, the coupling of reactive watershed models with land surface models is a frontier area. Improved understanding and methodologies are greatly needed to represent hydrological and biogeochemical connectivity across scales in a manner that honors landscape characteristics, hydrological boundaries, lateral and vertical fluxes and transformations, and subsurface heterogeneity and processes that influence biogeochemical cycling.
- **Extend and Link Knowledge Discovery Tools.** Methods are needed for analyzing interaction networks, from molecular to global scales, to gain insight into environmental and climate effects on terrestrial parameters. These methods might include tools to measure and visualize feed-forward and feed-back mechanisms controlling nutrient and signal exchange among organisms within an ecosystem. Other questions could relate to understanding the impact of various stressors on interactions between cells

and organisms or the role of aerosols in clouds. Such knowledge discovery is needed to increase the information content of individual databases so that the data are more effectively used for science applications and integrated with models. Although information is implicitly contained within datasets, the gulf between data, information, and actionable knowledge is broad and can be spanned only by developing and implementing technologies such as relational databases embedded in process models of the system within which the data have been collected. Future process models need to explicitly include linkages to databases and knowledgebases that form the basis for those models and therefore can be updated, improved, or removed as scientific understanding advances or changes. In particular, scientific and modeling applications that use BER data require information systems that synthesize multiple datasets of the same parameter; develop gridded data; provide essential metadata; and offer information on the accuracy, precision, and uncertainty of different levels of processed data. A framework that explicitly recognizes the needs of data curation, integration, refinement, and abstraction is essential. Each modeling community needs high-quality data products at the right level of abstraction and the provenance necessary for curation. BER's infrastructure should include open data repositories, databases, and knowledgebases, the latter containing the highest level of curated data that can be built on by others. A key issue in this regard is development of controlled vocabulary ontologies and semantic-oriented search tools to provide a broadly useful index and searching capabilities for finding records according to their scientific meaning, rather than syntax.

References

1. "BER Virtual Laboratory: Innovative Framework for Biological and Environmental Grand Challenges," A Report from the Biological and Environmental Research Advisory Committee, February 2013. Available at http://science.energy.gov/~media/ber/berac/pdf/Reports/BER_VirtualLaboratory_finalwebLR.pdf