Combined Break-out Group 3 Report Backs

Wednesday Morning

Simulation and Modeling

Research issues in Simulation and Modeling for Eco-Informatics

- **Coupling diverse models** that use different assumptions or definitions, or that represent a wide range of spatial and temporal scales
- Addressing **Values in Design**: eco-informatics and modeling both have strong value implications. How can we design, build, and use models in a way that approaches these value/system interactions in a principled way and that makes sense to the diverse set of stakeholders?
- Investigating new or improved **visualizations** for model results, but also model structure, assumptions, processes and influences
- **Representing error and uncertainty:** in particular, how do errors and uncertainty propagate through diverse coupled models, operating at widely different spatial and temporal scales; and how to best communicate this information to decision-makers
- Handling **Large data sets** and related performance challenges
- Developing an **open-source modeling infrastructure**, that is flexible and reusable, along with the social practices that sustain it

Sustaining Research on Ecological Simulations and Models

- Proposal: develop an open-source modeling infrastructure, that is flexible and reusable, along with the social practices that sustain it
  - Supports a community of practice and helps expand it
  - Possible analog: the R statistical system
  - Need good descriptions of the models so that people can evaluate what parts are suitable for their application; useful and up-to-date documentation and tutorials
  - To help sustain this, include contributing to the shared community infrastructure as one of the criteria by which proposals to other NSF programs will be evaluated
- Scenarios of use
  - Allow researchers to conveniently experiment with a new model or change an existing one. Plug into an existing infrastructure (and possibly data)
  - Facilitate comparing alternate models
  - When feasible, other stakeholders (decision-makers, students, …) should be able to access models via a web interface

Data Quality

Group members:
Larry Sugarbaker
Sherry Pittam
Kevin Gergely
Craig Palmer – presenter Dec 14 PM
Julia Jones – scribe, presenter Dec 15

Objectives of morning Dec 15 session:

I. Refine research issues

- Prioritize? Or Categorize? We chose categorize.

- General problem of data quality in decision-making: How to determine and communicate uncertainty to decision-makers in studies integrating multiple data sources?
- Overarching research question: Does uncertainty associated with data quality and synthesis really have an influence on policymaking and plan implementation?

Category 1: Issues for individual studies in which diverse data sources are combined

Knowledge of points where error is introduced

- a. Develop methods of reducing the introduction of error when datasets are (created and) combined
- b. Develop methods for error measurement and logging at each stage
- c. Develop methods for characterizing relationships among errors – additive, multiplicative, averaging
- Associating errors with alternative decisions/actions

Category 2: Issues for sharing of data

Can metadata become a part of the dataset?

j. What happens to metadata when multiple sources are integrated?

k. How can metadata management be automated once it is created?

l. How can data standardization help the process of combining metadata from multiple sources?

m. Can open-source tools be developed for mapping data content standards to one another?

II. Make a scenario or tell a story, or describe how research cycle can be sustained? We chose stories. Stories are numbered to correspond to the list of research issues above.

1. Story: the SEEK project focuses on extracting knowledge from one study design to apply to another, e.g. workflow diagram approach of the SEEK project. This approach could serve as the template for examining potential sources and types and magnitudes of errors. We are not sure whether SEEK is trying to answer questions 1a,b,c.

1.a.. Sherry (OSU): Can standardized tools be developed for data entry? Larry (NatureServ): The problem is about developing intuitive data entry tools that are mapped to standardized data models. Larry's group have a proposal that has been repeatedly submitted to NSF but not funded on this. NCEAS (Matt Jones) has tried but has had disappointing results. Two approaches: one (NCEAS) involves the use of a questionnaire to provide specifications for designing a form, but the technology available to generate standardized forms is too crude. An alternative approach (Larry's group) is to develop a set of tools that are mapped to standardized database structures. These design frames could be selected and tailored by users, and would be available as templates on a website.

1b. NSF could develop and publish metadata standards across all grants, instead of just for certain programs. By far the most advanced work is being done by the Federal Geographic Data Committee within the USGS, including a biology standard. Metadata standards are well developed and in use by the LTER information managers, and these standards are used in internal reviews of LTER projects. NBII is making a very big push to require metadata using the FGDC standards for its projects. It would be important to pull Viv Hutchison into this discussion.

2. a.,b, c. Julia Jones. Data harvesters collect existing databases – e.g. the Long-term Ecological Research network's Clim-DB, Hydro-DB. One could ask how existing data-harvesters answer questions 2 a,b,c.

Overarching research question: Does uncertainty associated with data synthesis really have an influence on policymaking and plan implementation? Studies could be done of decision-makers perceptions of the value of science findings made from synthesized or integrated data. For example, data harvesters such as Clim-DB and Hydro-DB have generated publications from combined datasets, which are (perhaps) being used by land managers or decision makers in the Forest Service and NOAA. This work could be extended by examining how syntheses of datasets are used by decision-makers and how apparent and important the errors were to decision-makers. Specifically, the research question is: how is the increase in power associated with data synthesis balanced by the increase in uncertainty associated with the ways in which the errors were combined? An extension of this work could examine how synthetic studies stand up in courts of law in comparison with other forms of "expert testimony."

- A great example of a possible study would be a follow up on the President's Forest plan and how the data synthesis and uncertainty affected the ability of a plan to be implemented. Craig Palmer and others are involved in a 10-year review of the President's Forest plan but this does not (yet) include an analysis of the effect of uncertainty and its consequences.

2.d. Larry Sugarbaker. Problem: Automated mapping of various vegetation classification standards. XML based products are emerging to create these mappings but the matching has to be done by hand. A solution could be creating an exchange tool; it would involve two innovations: (1) develop a set of definitions that define the relationships among the categories that are matched, and (2) publish this as open source code. Research challenge: How general can these tools become to be applied to a wide range of ecological datasets?

Integration and Ontologies (combined discussion of groups)

Ontology/info integration breakout group

Lois, Frank, Molly, Bruce, Ed, Tim, Vivian, Anne

The Fish Story

- Rainbow trout were reclassified as salmon around early 90s by scientists/taxonomists/biologists at XXX
- Why? Differentiae suggested rainbow trout were salmon
- They were the experts recognized by IDIS (Integrated Dynamic Info System)
- Therefore, the reclassification had the force of law
- Who cared?
  - Fisheries in WA: treaties gave WA Indians the right to fish salmon in perpetuity; suddenly they could catch rainbow trout as they wanted, which affected the fisheries
  - Anglers cared
  - The trout: their sustainability was affected
- Lesson: "No indicator is innocent!". A seemingly small decision in the advancement of science had a major effect due to the existing circumstances
- How did society respond? (Check Google)
- What can IT do?
  - The scientists should have been warned about the effects
  - Policymakers should have been warned
  - See research agenda following

Research themes 1: topic clusters

- Ethical issues:
  - Data collection, ontologization decisions, and modeling / processing embody **value judgments**: what do you highlight and what do you sideline? How do you reason about these issues? Computational support for approaches and methods of reasoning about complex issues (e.g., Precautionary Principle) within BDE; implications of presentation/interface design decisions; etc.
  - **Data sharing** and hiding:
  - **Teaching** ethics of IT for BDE: ethics-based SIM City?
- Transparency (across data, assumptions, models, metadata, etc., all anchored on the data)
  - Ability to determine, for each datum, all associated/encapsulating/definitional/interpreted information
  - Propagation of effects of decisions

Topics continued 2

- Semantics of BDE:
  - Defining the meanings of BDE concepts; operationalizing these concepts in domains and tasks; verifying theoretical semantics work; supporting data standardization; finding, capturing, managing, and sharing semantics; providing semantic services for others
- Application of knowledge and practice in new domains:
  - Do existing practices work in general? Codification and formalization of approach; development of methodology (protocols, processes, testing, etc.)…
  - Finding and re-using existing work (e.g., from Health — can the environment be treated as a patient?)

Topics continued 3

- Quality Control:
  - Data uncertainty (formal measures)
  - Accuracy and reliability
  - Change / recency
  - Trustworthiness of source/creator quality ratings, pedigree
- Centrality of location to BDE info:
  - GIS; flexible media (integrated displays of space and other characteristics across media); overlaying different spatial frameworks of interpretation (polygons) for different people/models; topological relationships among regions, etc.; ontological definitions for space
- Collaboration:
  - Support for collaborative decision-making: sharing data; dynamic decision-making; sharing simulations of decisions; training of value of collaborations
  - Community building technology