Jun 17th, 3:40 PM - 5:20 PM

Exploring Environmental Model Catalogs

Ilva Zaslavsky
San Diego Supercomputer Center, University of California San Diego, zaslavsk@sdsc.edu

Thomas Whitenack
San Diego Supercomputer Center, University of California San Diego, twhitenack@sdsc.edu

David Valentine
San Diego Supercomputer Center, University of California San Diego, valentin@sdsc.edu

International Environmental Modelling and Software Society (iEMSs)
7th Int. Congress on Env. Modelling and Software, San Diego, CA, USA
Daniel P. Ames, Nigel W.T. Quinn and Andrea E. Rizzoli (Eds.)
http://www.iemss.org/society/index.php/iemss-2014-proceedings

Exploring Environmental Model Catalogs

Ilya Zaslavsky, Thomas Whitenack, David Valentine
San Diego Supercomputer Center, University of California San Diego,
{zaslavsk|twhitenack|valentin}@sdsc.edu

Abstract: As part of the NSF EarthCube cross-domain interoperability roadmap and the CINERGI (Community Inventory of EarthCube Resources for Geoscience Interoperability) projects, we assembled and analysed several catalogs of environmental models. The goals of this work included analysis of patterns of pre-model data flows from multiple domains and consistent cataloguing models across multiple model collections using a common set of metadata fields. This paper presents status updates and initial analytical results along these goals. We explored model registries assembled by NOAA and EPA, as well as model descriptions created within CSDMS, ESMF and TESS projects. These model metadata collections were visualized and analysed using an online faceted search interface. In each case, model metadata differed across collections depending on the scope, objectives and methods used to assemble each inventory. This heterogeneity precludes efficient model linking or compatibility analysis, as well as model discovery across collections. The paper focuses on the first steps in exploring and visualizing such inventories, discussing and comparing the respective metadata models, and analysing cross-domain pre-model data flows.

Keywords: geoscience models; catalogs; interoperability; EarthCube.

1 INTRODUCTION

Integrated environmental modeling (IEM), which emphasizes new approaches to cross- disciplinary modeling and decision-making, has been the focus of recent conceptual and methodological analysis (e.g. Laniak et al., 2013). One of its key guiding principles is reliance on a modular approach to model creation, where independently developed models from different disciplines can be combined into more complex environmental modeling systems. This, in turn, requires that the modules are consistently described and registered, and can be invoked from other modules via an agreed upon application programming interface (API). Indeed, several examples of such integrated modeling platforms have been developed (CSDMS, 2011; Hill et al., 2004; Gregersen et al, 2007; Granel et al, 2013), which defined architecture, metadata descriptions and APIs for integration of model components within a host modeling framework. There is also an ongoing effort to standardize such interfaces (e.g. OGC, 2013.)

Several integrated modeling platforms are assembling collections of model components compatible with the host frameworks. However, multiple challenges remain. They include defining common model descriptions and converging to standards for publishing model components, to eventually enable their automated discovery, access and integration. Also, a large number of environmental models have been developed outside such frameworks. They are often difficult to find and evaluate. One of the reasons is that these legacy models may use different and often incompatible variable names, units, data formats, spatial objects or gridding schemes, time-stepping schemes, coordinate systems, resolutions, etc. Another reason is that information about these model characteristics is not typically organized in model catalogs with standard

metadata, and remains inaccessible. A necessary first step in implementing the vision of integrated environmental modeling is understanding the diversity and complexity of existing models, by organizing and comparing available registries of models and model components. Having uniform catalogs of model resources would facilitate model analysis and further model development, help avoid duplication, and make it easier to discover model components and assess their compatibility. Several such initial registries of model components are reviewed in the next section of the paper. It is followed by an initial comparison of the available inventories, with specific focus on metadata used to characterize the models. In most cases these characteristics can be used for model discovery and assessment, but fall short of a more complete model representation for the purposes of automatic integration of model components. We then present an application of model registry analysis focused on assessing cross- disciplinary data flows for a range of models. The latter analysis is done as a methodological exercise to demonstrate connections between different disciplinary domains within geosciences, developed as part of an EarthCube roadmap project [EarthCube, 2012]. The conclusion reviews the results and outlines unresolved issues.

2 AVAILABLE MODEL REGISTRIES

We have examined several registries of models and model components. The registries have different content, since model inventories have been assembled by different research groups, for different purposes and with different level of effort. They derived from both academic projects and government agency efforts, and followed different procedures for catalog construction. Models may be included in the catalogs following a systematic inventory protocol, or be self-submitted; the inventories may emphasize different domains or different aspects of modelling work (models themselves, or modelling activities, or model components), and hence have different model metadata requirements and schemas. An additional serious challenge is varying understanding of what a "model" is in different fields and disciplines, and different decisions about inventory scope with respect to what types of models shall be included. Yet we find that some of the key dimensions of model description are used across different model inventories. They include domain, purpose and scientific basis of the model, technical specification, and author's contact information – but may also extend to implementation details, types of inputs and outputs, and organizational issues. In the case of environmental models, spatial characteristics are also present, typically. The registries we examined are briefly reviewed below.

2.1 CSDMS model registry and repository

The Community Surface Dynamics Modeling System (CSDMS) is an architecture, a set of conventions and protocols, a collection of models and a community of experts that focus on modeling earth surface processes. At the time of writing, the CSDMS web site lists 170 models and 55 related modeling tools, and includes terrestrial, coastal, hydrological, marine, climate and carbonate models (CSDMS, 2011). CSDMS includes a detailed metadata scheme for describing models and model components. The main rubrics include:

- Summary: model type, spatial dimensions and extent, model domain, short and long model description, keywords.
- Contact: name, type of contact (e.g. "model developer"), address, email, phone, fax.
- Technical specifications: platforms, programming language, optimal processor setup (e.g., "single processor"), development status, code availability and URL, license; memory requirements, typical run time, etc.

- Input and Output: parameter descriptions and formats, pre- and post- processing and visualization software.
- Process: description of key physical processes represented by the model, parameters and equations, spatial and temporal scales, and numerical limitations.
- Testing: available calibration and test data sets, and ideal data for testing.
- Other: collaboration plans, model web site, forum, and manual.
- Component Info: compliance with selected model interfaces, model version and DOI.

In addition, each model description includes download statistics, and may include references to publications, visualizations, and other supporting files. While the model metadata template is extensive, many of the fields in the module questionnaire are optional, and remain blank, especially in rubrics that reflect testing, processes, inputs and outputs, and some of the technical specs. Also, the registry includes both actively used and continuously developed models, and older models that have not been used widely or are no longer in use. In our early attempt to evaluate activity associated with CSDMS-registered models, we used an internet search engine (Google) to record counts of hits associated with model names: the numbers ranged from several dozen for older models to hundreds of thousands for such models as WRF, MITgcm, SPARROW, SWMM, and Delft3D. Specialized tools and visual interfaces for assessing metadata content and completeness (such as the one described in section 3), along with a system for curation of submitted metadata and organizing models based on their actual usage would help fill in the gaps, especially for the models that are likely to be downloaded and re-used by geoscience researchers.

2.2 EPA model registry

EPA Models Knowledge Base includes computational models developed, used or supported by EPA Offices, and, as of this writing, lists 187 models (EPA, 2014.) Model descriptions include several rubrics:

- General information: model identifier, overview and description, contact information, a link to its home page, model version and lifecycle phase, and expected users of the model.
- Model use: technical requirements to run the model (hardware, operating system and software), model inputs and outputs, user's guide and related technical guidance documents, user support contact, and expected user qualifications.
- Model science: conceptual basis and science behind the model, its structure and equations, information about the model evaluation, scope (e.g. spatial and temporal scales) and relevant case studies.
- Model attributes: a collection of tags associated with the model that support model registry browsing and search based on general model type (e.g. deterministic vs stochastic, descriptive vs normative), statutes and regulations it addresses, pollutant type (physical, chemical, biological), pollution source type (point, area or mobile), media (ground, air, water, ecosystem), exposure characteristics, and on indicators of human health, ecological impact or different types of damages.
- Interdependencies: interfaces with other information resources, and subsystems.

As in the CSDMS case, a number of model metadata fields are not populated in the EPA registry. Interesting additions are the ability to exclude model resource in the Termination life cycle phase, convenient browsing based on model tags listed in the

Model Attributes section, and direct connection between modeling activities and normative documents and regulations. There are several models that are listed in both CSDMS and EPA registries (e.g. the NASA AOM: Atmosphere-Ocean Model or HSPF: Hydrological Simulation Program - FORTRAN), – though their descriptions in both registries significantly differ. Such lack of consistency in model descriptions across registries makes it difficult to detect duplicate model records.

2.3 NOAA compilation of environmental modelling activities

NOAA inventory of environmental modeling activities was developed in 2008-2009, and included 427 consistently described modeling projects (Leonardi, 2014). The key metadata fields can be organized as follows:

- Model or modeling activity: name and acronym.
- Contact information.
- Brief description of model or activity, including geographic coverage.
- Status or type of activity, with a controlled vocabulary (CV) including terms such as operational use and R&D use.
- Modeling discipline: a CV of 12 discipline names ("physical oceanographic", "hydrologic", "ecosystem", etc.) and "other".
- Modeling activity purpose: assessment, nowcast/forecast, etc.
- Products and services supported by the modeling activity.
- User community description
- Computational platform
- Types of model input (remote sensing, in situ, etc.)
- Funding: level of investment, funding from non-NOAA sources. NOAA investments in external activities.
- Partner organizations (e.g. universities, agencies such as FEMA).
- Requirements and gaps for future modeling effort.

The inventory represented a one-time snapshot of modeling activities, compiled through a self-reporting arrangements rather than via a strictly administered survey. While it contains fewer metadata fields than CSDMS or the EPA models knowledge base, the registry is consistently populated. Because of its focus on modeling activities rather than on models, several inventory records may point to a single model. For example, the FVCOM (Finite Volume Coastal Ocean Model), also listed in CSDMS, is referred to from a number of NOAA modeling activity pages. In essence, the inventory of modeling activities in several cases deals with model instances developed for specific locations rather than with general models. This makes it difficult to integrate the inventory with other model inventories, but provides a useful and unique perspective on model activity metadata and inventory compilation trade-offs.

2.4 ESMF model components

The Earth System Modeling Framework (ESMF, 2014) is a flexible software infrastructure for composing high-performance coupled modeling systems primarily for climate and numerical weather prediction. The modeling systems are composed from ESMF components, which are functionally-defined software units communicating with other similar units over a specific API. The components may represent physical domains, coupler functionality, or other functions. The online inventory of ESMF components and modeling systems includes 91 entries, with the following metadata characteristics:

- Component name, institutions of lead author and ESMF wrapper author (if applicable.)
- Infrastructure and superstructure use, which refer to ESMF features used by the component, and the context in which the component is run.
- ESMF conventions used.
- Code, ESMF version, and general reference (URL).

The choice of metadata descriptions here reflects the inventory's IEM perspective, where specific focus is on component interoperability within the ESMF infrastructure. While several ESMF components appear in other registries as well (for example, MITgcm, ROMS and WRF also appear in the CSDMS registry), the IEM focus results in significantly different technical specifications for ESMF-wrapped components as compared to original models. This, in turn, makes it difficult to integrate this inventory in a joint model registry system.

2.5 Model registry assembled by the TESS project

A database of existing models for local ecosystem management has been collected within the Transactional Environmental Support System (TESS, 2014) project, funded though the European Commission's Seventh Framework Programme. The models are mostly focused on sustainable ecosystem services and health of ecosystems at small scales. While the website suggests that the database is not yet ready, it already represents an impressive collection of 393 models, characterized with the following rubrics and fields:

- General model information: name, acronym, description, application area, subject, and ecosystem service management focus area.
- Technical description: modeling paradigm and approach; computing platform, operating system and language.
- Geographic applicability area and time horizon.
- User friendliness.

This database provides an interesting example of a large model metadata catalog assembled within a research and development project. Yet, as most inventories that rely on self-submission of records, the collection has uneven coverage and unknown completeness, and many non-mandatory fields are left blank.

2.6 Visualization of Model Catalogs

We have experimented with visualizing several model collections, using Silverlight PivotViewer application (e.g. see http://maxim.ucsd.edu/tessmodels3/ for the TESS model inventory, and http://maxim.ucsd.edu/esmfmodels/ for the ESMF components, described above). The viewer implements visual faceted search over the model collections, and supports generation of bar charts for any metadata facet and model subset. In particular, this visual approach to exploration of model catalogs allows users to quickly identify gaps in model descriptions and examine distributions of models on multiple criteria, to assess catalog completeness and model metadata completeness and consistency. Figure 1 provides a snapshot of the visual interface with a bar chart showing distribution of TESS models by computer platform. Figure 2 shows a closer view of the same inventory, at the zoom level where model descriptions become visible.

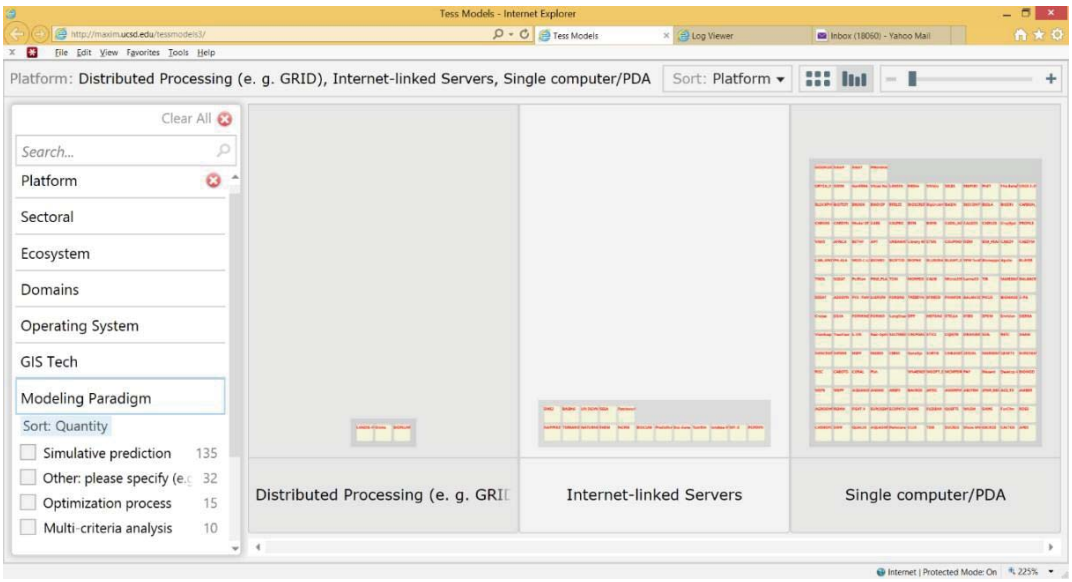

Figure 1. Visual interface for browsing model catalogs, implemented with Silverlight PivotViewer

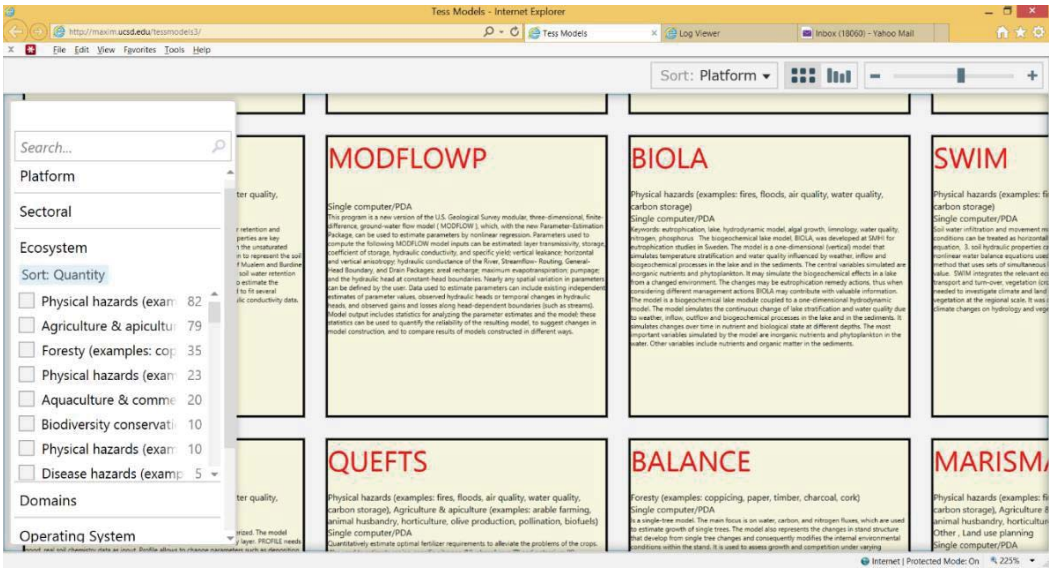

Figure 2. Zoomed in view for browsing model descriptions

3 PRELIMINARY COMPARISON OF MODEL CATALOGS

As mentioned above, the model inventories have been created by different groups, for different purposes, and with different scopes and methodologies. The table below combines metadata descriptions from several model catalogs, in an attempt to reconcile some of the semantic differences. The inventories we review are still developing, and our goal is not to point to perceived deficiencies but to examine diversity of approaches offered by different systems. The comparison is complicated by different semantics of model descriptions, and, in many cases, by the choice of free text descriptions over controlled vocabularies. Both features are indicative of an early phase of model registry development. In proposing the model annotation fields in this table, we try to capitalize

on semantic commonalities across the existing model inventories, while pointing to differences between them and, in a few cases, suggesting characteristics that appear to be missing from all of them. The comparison table is preliminary.

	CSDMS	NOAA	TESS	EPA	ESMF
Content	170 models, 55 tools	427 modeling activities	393 models for local ecosystem management	187 computational models	91 components or coupled modeling systems
Summary information					
Model name and alias	*	*	*	*	*
Model category	*	*	*	*	*
Purpose		*	*	*	
Model type	*		*	*	*
Model domain	*		*	*	
Description (short and/or long)	*	*	*	*	
Keywords	*			*	
DOI	*				
Recommended citation(s)	*			*	
Contact information	*	*	*	*	*
Science use case					
Processes modeled (w/equations)	*		*	*	
Key model assumptions	*			*	
Applicability limits, numerical limitations	*			*	
Spatial dimension	*			*	
Spatial extent, resolution/discretization	*		*	*	
Time-stepping scheme and horizon	*		*	*	
Vertical coordinates					
General references	*			*	*
Technical Specification					
Supported platforms	*	*	*	*	*
Programming languages	*		*		
Required compute resources	*		*	*	
Costs (development,		*			
Coupling design					*
Inputs and outputs					
Input parameters and format(s)	*	*	*	*	
Forcing, initial conditions, boundary conditions					
Output parameters and format(s)	*			*	
Vocabularies/ontologies for parameters	*				
Data quality requirements for inputs					
Uncertainty estimates for output					
Resolution of output					
Pre-processing software	*				

Post-processing or visualization software	*			*	
Lineage and Compatibility					
Framework compliance	*			*	*
Standards compliance					
Model lineage					
Incorporated models	*			*	
Can be coupled with	*			*	*
Development Status					
Current status, version	*	*			*
Model web presence (URL)	*		*	*	
Development start and end	*				
Plans for further development		*			
Lifecycle management				*	
Documentation and validation					
Model documentation and instructions	*			*	
Results interpretation guidelines				*	
Testing/validation procedure	*			*	
Testing/validation results				*	
Calibration and test datasets	*				
Availability					
Ownership and control of the model		*		*	
Module and code availability	*			*	
License	*				
User-friendliness			*		
Usage information and community					
User community description		*		*	
Partner organizations		*			
Investments in external activities (e.g. grants)		*			
Community annotation on the models, voting	*				
Model use/popularity	*				
Intended use and constraints on use		*		*	
Citations to derivative work	*			*	
Cross-linking	*	*			

4 CROSS-DOMAIN DATA FLOWS: PATHWAYS THROUGH EARTHCUBE

One of the approaches explored in the EarthCube Cross-Domain Interoperability roadmap (EarthCube, 2012) was understanding cross-domain connections by using model catalogs. To implement it, we annotated models in one of the catalogs (the TESS catalog) by specifying which disciplines each model draws data from. Once models are

annotated with domains from which they draw data, it is possible to count models in each domain that require data from each other domain. This would point to pairs of domains that have consistent and intense data exchanges, and would therefore require robust data exchange interfaces. An initial example of such an online connectivity map ("a map of EarthCube pathways"), built from annotated models, is shown in Figure 3. In this application, models are considered as proxies of cross-domain research scenarios. It is difficult to compile large collections of research scenarios, but using model registries instead may support a more efficient cross-domain connectivity analysis. Such a connectivity analysis and visualization, when triangulated with user surveys and research use case inventories, would help define cross-domain infrastructure requirements and development priorities.

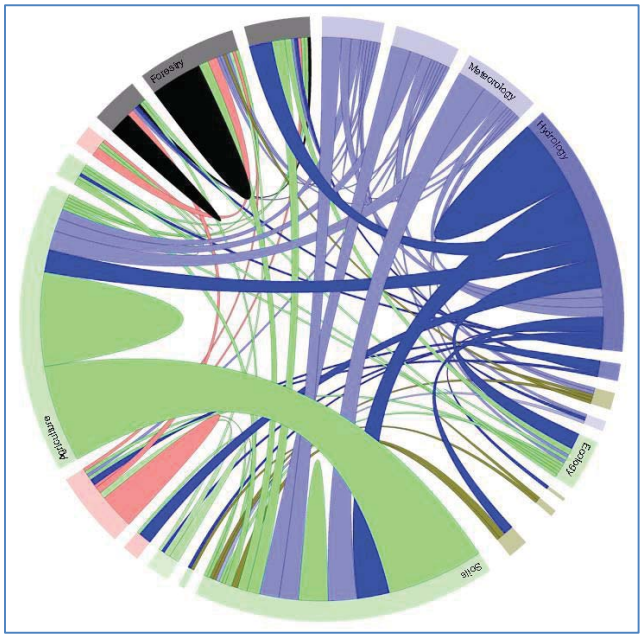

Figure 3. A snapshot of an online interactive visualization showing connections between domains as derived from annotated environmental models in the TESS model registry. Sector sizes reflect the number of models in each domain, and connections show relative numbers of "data pathways" between the domains.

5 CONCLUSIONS

We presented an initial exploration of different catalogs of environmental models. Model semantics are complex, and different inventories have defined different model metadata templates depending on a number of factors. These factors include scope and intent of the inventory (e.g. internal assessment of modeling activities in case of NOAA, general community model inventory with the focus on consistent inter-model connections in case of CSDMS, registry of components and systems for model coupling in case of ESMF), domain focus (e.g. ecosystems in case of TESS, mostly surface processes in case of CSDMS), procedures for assembling and curating the content. Because of these differences, and lack of respective standards, relating metadata descriptions across catalogs is challenging. Additional complications arise from reliance on free-text metadata descriptions and the resultant lack of semantic consistency: for example, a lot of information can be derived from text analysis of extended model descriptions, but it is difficult to relate it to other fields in a consistent manner. Yet, as our initial comparison shows, it is possible to identify common dimensions across catalogs, at least with respect to summary and contact information, and several technical specification fields.

We hope that this work would continue by exploring additional issues such as tracing model lineage, establishing best practices for model catalogs, and eventually standardizing semantics of some of the descriptive fields. This would lead to more efficient model discovery and eventually model interoperability within the emerging integrated environmental modeling paradigm. It would also support more efficient analysis of cross-domain data flows and influence priorities of large infrastructure development initiatives such as EarthCube.

6 ACKNOWLEDGMENTS

Support from NSF under awards "EAR-1238420: EAGER: Readiness of Disciplinary Data Systems for Cross-Domain Interoperability within a Standards-Based EarthCube Reference Framework" and "ICER-1343816: EarthCube Building Blocks: Community Inventory of EarthCube Resources for Geoscience Interoperability (CINERGI)" is gratefully acknowledged. We appreciate efforts of model registry creators referenced in the paper. We also thank members of the EarthCube cross-domain interoperability roadmap team for valuable discussions.

7 REFERENCES

CSDMS, 2011. Community Surface Dynamics Modeling System, CSDMS. http://csdms.colorado.edu/ (last accessed March 20, 2014).

EarthCube, 2012. EarthCube Roadmap prepared by the Cross-Domain Interoperability test Bed Group. PI: I. Zaslavsky. Version 1.1, Aug. 2012. http://workspace.earthcube.org/sites/default/files/files/document-repository/EarthCube%20Cross-Domain%20Interoperability%20Roadmap.pdf (last accessed March 20, 2014).

EPA, 2014. Models Knowledge Base. http://www.epa.gov/crem/knowbase/ (last accessed March 20, 2014).

ESMF, 2014. ESMF Components and Modeling Systems. http://www.earthsystemmodeling.org/components/ (last accessed March 20, 2014)

Granell, Carlos, Sven Schade, and Nicole Ostländer. 2013. "Seeing the Forest through the Trees: A Review of Integrated Environmental Modelling Tools." *Computers, Environment and Urban Systems* 41: 136–50.

Gregersen, J., P. Gijsbers, and S. Westen. 2007. "OpenMI: Open Modelling Interface." Journal of Hydroinformatics 9 (3): 175–91.

Laniak, Gerard F., Gabriel Olchin, Jonathan Goodall, Alexey Voinov, Mary Hill, Pierre Glynn, Gene Whelan, Gary Geller, Nigel Quinn, and Michiel Blind. 2013. "Integrated Environmental Modeling: A Vision and Roadmap for the Future." *Environmental Modelling & Software* 39: 3–23.

Leonardi, Alan, 2014. NOAA Environmental Modeling Activities (personal communication).

Hill, Chris, Cecelia DeLuca, Max Suarez, and Arlindo da Silva. 2004. "The Architecture of the Earth System Modeling Framework." *Computing in Science & Engineering* 6 (1): 18–28.

Jakeman, Anthony J., Olivier Barreteau, Mark E. Borsuk, Sondoss Elsawah, Serena H. Hamilton, Hans JøRgen Henriksen, Sakari Kuikka, Holger R. Maier, Andrea Emilio Rizzoli, and Hedwig Van Delden. 2013. "Selecting among Five Common Modelling Approaches for Integrated Environmental Assessment and Management." *Environmental Modelling & Software* 47: 159–81.

OGC, 2011. The OGC and OpenMI Association to advance computer modelling standards. http://www.opengeospatial.org/pressroom/pressreleases/1450 (last accessed 3/20/2014)

TESS, 2014. Transactional Environmental Support System (TESS), Database of models for local ecosystem management. http://www.tess-project.eu/models/ (last accessed March 20, 2014).