iEMSs International Congress on Environmental Modelling and Software

9th International Congress on Environmental Modelling and Software
Fort Collins, Colorado, USA, Mazdak Arabi, Olaf David, Jack Carlson, Daniel P. Ames (Eds.)
https://scholarsarchive.byu.edu/iemssconference/2018/

MINT: Model Integration Through Knowledge-Powered Data and Process Composition

Yolanda Gil[1], Kelly Cobourn[2], Ewa Deelman[1], Chris Duffy[3], Rafael Ferreira da Silva[1], Armen Kemanian[3], Craig Knoblock[1], Vipin Kumar[4], Scott Peckham[5], Lucas Carvalho[6], Yao-Yi Chiang[1], Daniel Garijo[1], Deborah Khider[1], Ankush Khandelwal[4], Minh Pahm[1], Jay Pujara[1], Varun Ratnakar[1], Maria Stoica[5], Binh Vu[1]

[1]University of Southern California
[2]Virginia Tech
[3]The Pennsylvania State University
[4]The University of Minnesota
[5]University of Colorado, Boulder
[6]University of Campinas

gil@isi.edu

Abstract: Major societal and environmental challenges require forecasting how natural processes and human activities affect one another. Model integration across natural and social science disciplines to study these problems requires resolving semantic, spatio-temporal, and execution mismatches, which are largely done by hand today and may take more than two years of human effort. We are developing the Model INTegration (MINT) framework that incorporates extensive knowledge about models and data, with several innovative components: 1) New principle-based ontology generation tools for modeling variables, used to describe models and data; 2) A novel workflow system that selects relevant models from a curated registry and uses abductive reasoning to hypothesize new models and data transformation steps; 3) A new data discovery and integration framework that finds and categorizes new sources of data, learns to extract information from both online sources and remote sensing data, and transforms the data into the format required by the models; 4) New knowledge-guided machine learning algorithms for model parameterization to improve accuracy and estimate uncertainty; 5) A novel framework for multi-modal scalable workflow execution. We are beginning to annotate models and datasets using standard ontologies, and to compose and execute workflows of models that span climate, hydrology, agriculture, and economics. We are building on many previously existing tools, including CSDMS, BMI, GSN, WINGS, Pegasus, Karma, and GOPHER. Rapid model integration would enable efficient and comprehensive coupled human and natural system modeling.

***Keywords**:* Model integration, semantic workflows, ontologies, reasoning, automated planning, machine learning, model metadata, data catalogs, model catalogs.

1 INTRODUCTION

Major societal and environmental challenges require forecasting how natural processes and human activities affect one another. This requires integrating highly heterogeneous models from separate disciplines, including geosciences, agriculture, economics, and social sciences. Model integration requires resolving semantic, spatio-temporal, and execution mismatches, which are largely done by hand today. It is also challenging to locate appropriate models and to find data at the resolution needed for each scenario and region. In addition, models are often designed or calibrated to approximate the real phenomena under study, which can lead to poor performance. Composing models to enable end-to-end simulations and executing them with large-scale data requires coordinating many requirements.

Unfortunately, model integration is extremely time consuming and integrating two or three models from different disciplines can take months or years. There are tools such as repositories to find models (Peckham et al 2013), software for regridding data to address mismatches of time and space scales, data representation standards, and model coupling for execution interleaving (BMI 2018). However, these tools address only slices of the process and are not well integrated, so model integration is largely done by hand. (Laniak et al 2013) call for the "development of standards for publishing data and models in forms suitable for automated discovery, access, and integration."

The paper describes initial work to develop an end-to-end approach that uses artificial intelligence to assist modelers by automating and improving important aspects of model integration. Building on our extensive prior work in artificial intelligence and modeling, we are developing the MINT (Model INTegration) framework that incorporates three key innovations: 1) Semantic technologies to address model and data discovery and to bridge the heterogeneity of model requirements and assumptions; 2) Automated planning to resolve data mismatches by including data transformations; and 3) Machine learning to improve model parameterization through the extraction of more accurate data from remote sensing and other sources and search optimization.

The paper begins with an overview of MINT and its core approach. Section 3 gives an overview of the semantic framework used to characterize models and data. Section 4 presents our work on using automated planning to generate complete workflows that include necessary data transformations to execute models. Section 5 describes our work on machine learning to extract data from remote sensing sources and to efficiently set model parameters. We also give overview of ongoing and future work.

2 A KNOWLEDGE-POWERED APPROACH FOR MODEL INTEGRATION

Our approach to model integration is to capture extensive knowledge about models, data, and the modeling process. Figure 1 gives an overview of the main components of our approach. The left side of the figure illustrates how models and data are characterized using semantic techniques to create a Semantic Model Catalog, a Semantic Data Catalog, and data transformation services. The right side of the figure illustrates that those catalogs and services are used in workflow planning to assist users to create workflows that include models and data transformations. The rest of this section gives an overview of the different components of MINT, which will then be described in detail in the rest of the paper.

To characterize data and models, we rely on principled ontology development to represent modeling variables. A key challenge in model integration is to understand what model variables mean and their relationship to the data used, which often involves reading lengthy model documentation. We use ontologies that rely on principles and patterns to create structured names for modeling variables so that it is easier to identify what each model variable means, as well as to make correspondences across models. We build on our prior work on the Geoscience Standard Names (GSN) (Peckham 2014), which covered a variety of domains in geosciences and have been used to integrate models. We are extending GSN to develop the MINT ontologies that include socio-political and economic modeling.

An aspect of model integration that takes significant effort is finding existing models, understanding how they work, and configuring and calibrating them with the data that is available. To facilitate this, we capture rich model metadata that describe characteristics of models that are important to a scientist. In prior work, we developed the OntoSoft ontology for scientific software metadata and the OntoSoft software metadata registry (Gil et al 2016). OntoSoft already contains more than 600 entries, including models from the Community Surface Dynamics Modeling System (CSDMS) (Peckham et al 2013). We are extending this prior work to create the MINT Model Catalog that will describe model invocation functions and data pre-processing workflows, and use the MINT ontologies to represent model variables and processes. The MINT Model Catalog will also include agricultural and economic models.

Another aspect of model integration that is very time consuming is finding and preparing data to calibrate models and to run scenarios. A key aspect of our approach is to represent the content of data sources using the MINT ontologies. To do this, we build on our prior work on Karma (Knoblock and Szekely 2015), a semi-automated framework to annotate data sources that uses machine learning to predict what the data represents. We are extending this work to handle time-series data. Remote

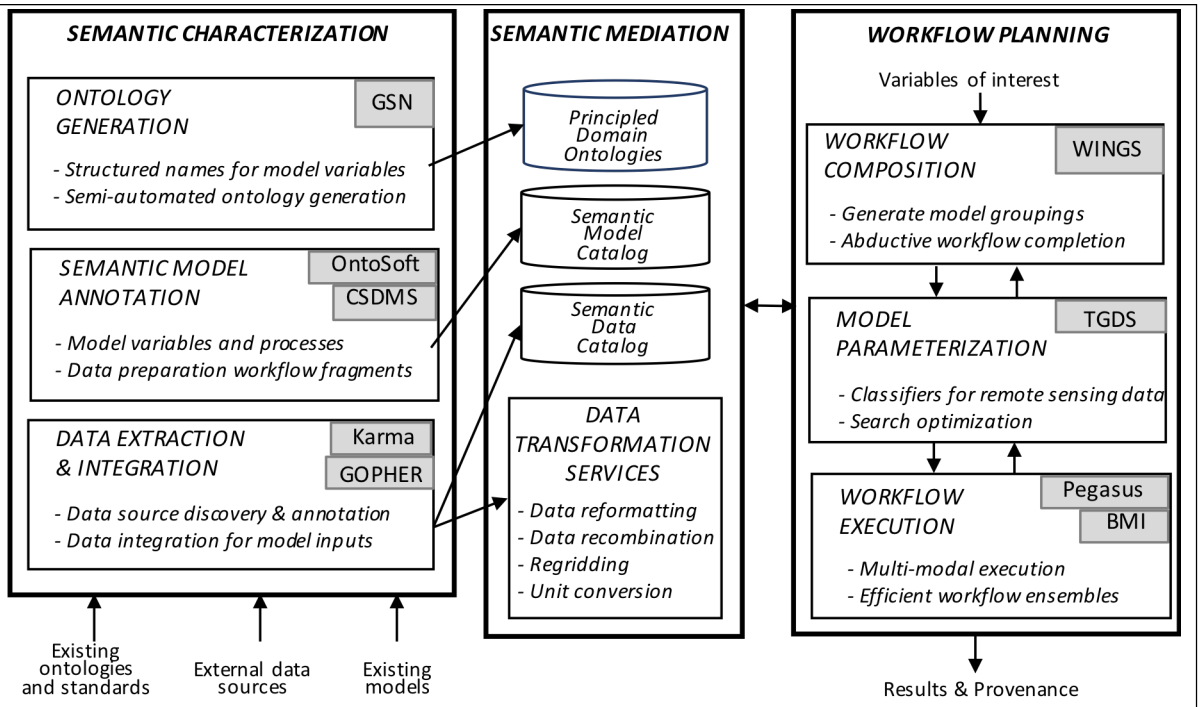

Figure 1. Overview of our use of artificial intelligence techniques to support model integration, highlighting in grey the components that we have developed to date. Semantic approaches (shown on the left) are used to characterize data and models in order to create semantic data and model catalogs as well as data transformation services (shown in the middle). Automated planning is used to support the creation and execution of workflows (shown on the right). Machine learning is used to generate information from remote sensing data sources, to extract data from Web sites and documents, and to optimize model parameter search (shown on the right).

sensing observations are often a great source of data to produce more accurate models. We build on our prior work on GOPHER (Karpatne et al 2016) that used a variety of machine learning techniques to extract data about water and land use dynamics. We are extending this work to extract new kinds of data, particularly population accounts and urban growth. The result of this work will be the MINT Data Catalog and data transformation services.

The MINT workflow planning component will assist a user in the creation of workflows that integrate several models. A user would start by specifying variables of interest, which would be used by the system to retrieve relevant models. We use automated planning to reason about the models and add appropriate data transformations services. We build on our prior work on the WINGS semantic workflow system (Gil 2014), which propagates the requirements of each workflow step, adds new components when needed, and ensures the generation of valid workflows. We are extending this work to enable users to simply specify variables of interest and use them to select valid combinations of models.

Model calibration is very challenging due to the large size of the parameter space. We build on our prior work on Theory-Guided Data Science (TGDS) (Karpatne et al 2017), which uses physics constraints and other knowledge to guide machine learning algorithms for model parameterization.

To explore simulation scenarios, the user would specify different input conditions to run the workflows. This requires executing many workflows at scale. We will use our Pegasus workflow system (Deelman et al 2015). Pegasus can manage the execution of workflow steps in distributed resources, move the data where the execution will take place, and recover from execution failures. To support the concurrent execution of models, we will leverage our prior work on the CSDMS Basic Model Interface (BMI) (BMI 2018) which provides an API for models and a framework for model coupling (Peckham et al 2013). We are integrating BMI with Pegasus, and extending Pegasus to coordinate the execution of large collections of workflows for model calibration and scenario exploration. The MINT workflow planning

component will generate a complete provenance record for workflow executions, so that users can explore the predicted results and estimate their uncertainty.

MINT uses existing components that are all open source, and will focus on models that are also open source. MINT uses open standards, in particular the W3C RDF, SPARQL, and PROV standards.

3 A SEMANTIC FRAMEWORK TO DESCRIBE AND REASON ABOUT MODELS AND DATA

An important challenge in model integration is addressing semantic mismatches between models and data. Although semantic technologies have been used by others to address this issue, we take a unique approach. First, we use principled ontologies with structured descriptions of modeling variables. Second, we modularize and describe modeling software with a methodology that facilitates model composition and data transformations. Third, we use semi-automated tools to characterize data and to create data transformation services. We describe these three aspects in turn.

3.1 Developing Principled Ontologies for Modeling Variables

A major challenge in integrating models is understanding model variables, processes, and assumptions. For example, a model may refer to "streamflow" and another to "discharge" and it may take some time to understand that they refer to the same physical variable. Although standards and ontologies have been created for specific domains, mapping variables across them remains an open problem.

Our approach is to develop general principles and turn them into patterns to create names for model variables, processes, and assumptions. In prior work, we developed a cross-domain ontology called the Geoscience Standard Names (GSN) (Peckham 2014). The GSN ontology was designed to serve as a semantic mediation hub and is based on very general principles that have been shown to apply to a wide variety of science domains including oceanography, atmospheric science, hydrology, glaciology, sea ice, geomorphology, general physics, continuum mechanics, thermodynamics, electricity and magnetism, seismology and environmental chemistry. GSN also includes standards for assumptions that models make, such as the Navier-Stokes equation for fluid dynamics. Development of the GSN ontology itself required several years and the expert human knowledge of multiple scientists with strong backgrounds in math, physics and engineering. This work was informed by community meetings with experts in several different science domains. It also included the analysis of several large controlled vocabularies and the variable names of numerous representative resources (e.g. models from different science domains). We are using GSN to characterize data and models in geosciences. Figure 2 shows an example of the principles and patterns in GSN to characterize model variables.

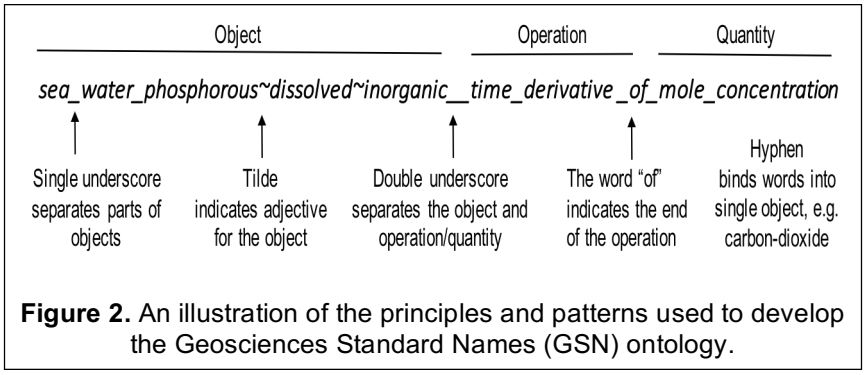

Figure 2. An illustration of the principles and patterns used to develop the Geosciences Standard Names (GSN) ontology.

We are extending our work on GSN to develop principled ontologies in other domains, particularly for social and economic modelling, using a general methodology. The first step requires gathering the domain vocabulary and common terms for variables from documents. Next, we determine the functional and conceptual categories of those terms by mapping them onto an upper ontology (e.g., quantity, process, phenomenon, etc.). At this point, term labels may be standardized by identifying common patterns and synonyms. Once the terms are properly mapped, the standard names generation engine uses the defined ontological structure to assemble standard names.

3.2 The MINT Model Catalog

Model repositories, such as CSDMS, provide a single access point to find and often execute models. However, to use a model one must investigate and understand how to use it. The OntoSoft software

metadata registry (Gil et al 2016) was developed to capture extensive information that is needed by scientists to understand how models work. Most of that information is available, but is scattered in publications, manuals, code documentation, and web sites (Essawy et al 2017). Having this information organized in a catalog saves scientists a lot of time in understanding and comparing models.

For the MINT Model Catalog, we need additional information that will enable the workflow planning system to select and compose models, as illustrated in Figure 3. These requirements are based on our detailed analysis of several distinct hydrology models. First, we need to represent the variables in each model and their dependency graph. We have used GSN to map variables for two models: the Penn State Integrated Hydrologic Model (PIHM) (Qu and Duffy 2007), which has more than 60 variables, and TopoFlow (Peckham et al 2017), with has more than 100 variables. Second, we need to represent explicitly the processes and methods used in a model. The figure illustrates two processes for TopoFlow (infiltration and snowmelt) and a method for each (energy balance and Richards 1D respectively). Third, we need to represent how the model variables are mapped to input and output files. These include their file structures and formats, spatial and temporal grids, values and units. We are working on representing common geoscience formats such as NetCDF and GRIB. Fourth, we need to represent distinctly the model invocation functions that correspond to different combinations of processes and methods when using a model. For example, the figure illustrates an invocation of TopoFlow for two processes (infiltration and snowmelt), but TopoFlow can also be used with a third process of subsurface flow in a saturated zone which would be a different invocation with new input data required. Finally, we are capturing common data pre-processing steps as workflow fragments. In addition to PIHM and TopoFlow, we are characterizing Cycles agriculture model (Cycles 2018) and the MODFLOW family of models (MODFLOW 2018) and the associated FloPy software in terms of data preparation requirements (Carvalho et al 2017). We also plan to include in the catalog economic models for natural resources that combine biophysical and socioeconomic data (e.g., (Cobourn et al. 2011)).

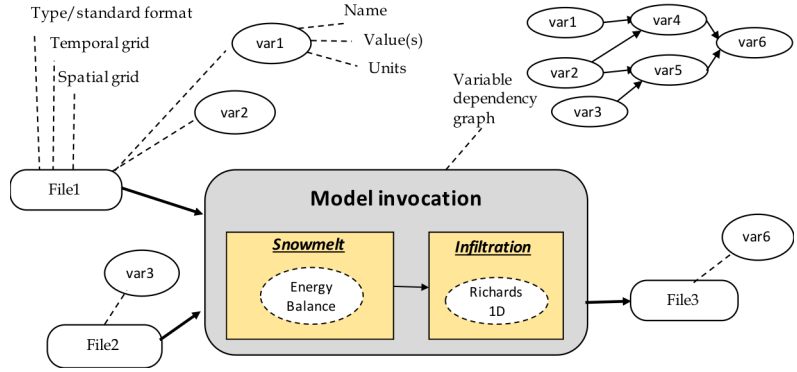

Figure 3. Representing the semantics of model invocation in the MINT Model Catalog.

3.3 The MINT Data Catalog

A key challenge in creating and executing the workflows is identifying the required data and transforming that data into the required format to run the models. Such data can be extracted from a variety of sources including databases, CSV files, online services, tables, remote sensing data, and other types of geospatial layers, such as maps and vector layers. As part of the extraction process, the information also needs to be aligned to the GSN ontology so that the models can make use of this information. Finally, we also need to deal with the various types of semantic mismatch that arise between what data is available and what data is needed by the quantitative models, which may require techniques such as regridding, temporal interpolation, unit conversions, filling in missing values, converting between different syntactic formats, and applying other types of transformations.

To address these challenges, we are developing semi-automatic approaches for source discovery, data extraction and alignment, and semantic mismatch resolution. The goal is build the MINT Data Catalog by finding the relevant data, extracting and aligning it to the GSN ontology, and then automatically transforming the data in the Data Catalog into whatever format is required by a given model.

We plan to build the MINT Data Catalog by starting with a large repository of existing databases and remote sensing data. The catalog will have metadata to describe the coverage and content of each source. For new sources, the system will need to extract the relevant data and produce a semantic description. The exact approach to this problem depends on the specific type of information. For structured data sources, such as CSV files, databases, or even web forms, we will build on our previous work in Karma (Knoblock and Szekely 2015), which provides a semi-automatic approach to mapping a structured source to a domain ontology using machine learning techniques. For time-series data, we are developing techniques to understand and semantically annotate such data. And for remote sensing data, we are developing techniques to automatically extract the relevant features from that type of source.

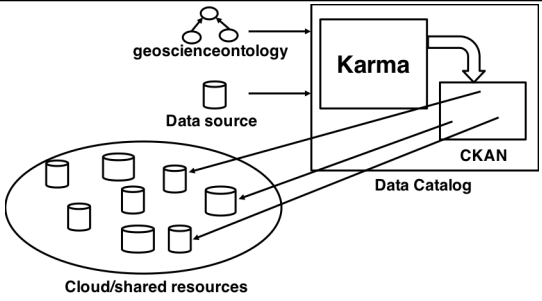

Figure 4. An illustration of the MINT Data Catalog.

Semantic mismatch resolution will be done through data transformation services to transform the data into the form required for each specific model. We plan to do this by defining local-as-view (LAV) queries, which we have used in our previous work on information integration (Knoblock et al 2001), that define the precise data required as input including the units, level of aggregation, joins, and so on. The task of the semantic mismatch module is to automatically produce the data transformation plan that will map from the data that is either in the MINT Data Catalog or produced by another step in the workflow into the required format. To fill in gaps where the required transformations are not available, we plan to develop a new capability that learns transformation using a combination of online tables and rule induction techniques to automatically create transformations, building on our past work on supervised learning of data transformations (Wu and Knoblock 2016).

In situations where there is no data in the catalog to support a given model, we plan to develop source discovery techniques. Given a specific geographic region and a type of information required for a model, the system will search the Internet for relevant datasets, online services, and data in knowledge graphs (such as Linked Data). To focus the search and find similar data, we will exploit any sample data that has been used to populate the initial MINT Data Catalog. There is previous work on performing source discovery to find web pages relevant to a domain. Our work will focus on sources discovery techniques for finding geospatially targeted data. We plan to do this using search phrases constructed from other climate-related sources as well as using terms from the geographic region of interest.

4 AUTOMATED PLANNING FOR DATA AND MODEL COMPOSITION

We use automated planning to reason about the semantic descriptions of data and models just described, and assist users to compose them into workflows that include data transformations. We are also developing a new workflow execution engine to support a range of model coordination modalities.

4.1 Workflow Generation through Data and Model Composition

To generate workflows composed of diverse models and the necessary data transformation steps, we use the WINGS semantic workflow system (Gil 2014) to reason about data characteristics and model requirements available in the MINT Model and Data Catalogs. While a traditional workflow simply represents dataflow among software components, a semantic workflow also represents the characteristics of the input and output datasets for each software step and any constraints in those datasets or parameters to the step. WINGS includes workflow reasoning algorithms that propagate those constraints for automated workflow elaboration, workflow matching, provenance and metadata generation, workflow validation, and interactive assistance.

A user would interact with the MINT throughout the workflow planning process. The user starts by specifying some variables of interest, which would indicate the scope of the problem and the level of detail required of the models. For example, a user may be interested in precipitation, crop yields, and

land use in a region. Those variables are then mapped to the MINT ontologies, and used to select relevant models. Several models may be available in the Model Catalog to generate any given variable, so MINT would generate several possible model groupings. Each model grouping would represent the initial skeleton for a workflow, which would then be expanded using the data pre-processing workflow fragments specified in the Model Catalog. This results in an initial workflow template. WINGS will then reason about the requirements of each model and add any data conversion steps needed to transform model outputs into the format required by other models. We are extending this work to use abductive reasoning and machine learning to create new models for variables of interest to the user that are not generated by any existing model. Once a complete and valid workflow is generated, the user can specify different scenarios and run the resulting workflows as described in the next section.

As users create workflows for different modeling scenarios and regions, MINT will acquire a growing collection of workflows that integrate diverse models. We will extend our prior work that uses machine learning and graph mining to learn workflow fragments that represent common combinations of models and data preparation steps (Garijo et al 2014).

4.2 Multi-Method Scalable Workflow Execution

Workflows are typically represented as directed-acyclic graphs (DAGs), where the outputs of a job (a node in the graph) are input for subsequent jobs. The DAG paradigm fits many integrated model execution requirements, where each model runs to completion and its output is used by other models. We will capitalize on our Pegasus workflow system (Deelman et al 2015) to enable scalable workflow execution in distributed, heterogeneous environments. In some cases, we will need to support coupled models, i.e., the concurrent execution of models with continuous data exchanges or transformations (when data is not in the expected format). We will leverage our prior work on the CSDMS Basic Model Interface (BMI) (BMI 2018) that provides a standardized framework-independent API for models. BMI is easy to implement and provides all information needed to deploy a model in multiple model coupling frameworks. We are developing a novel workflow engine that will combine DAG and BMI capabilities.

Model parameterization will trigger a large number of executions of a single workflow with different variables or data, i.e., a collection of workflows. The system should execute them efficiently, provide feedback from the executions, and allow users to adjust priorities at runtime. We are planning to develop new approaches for managing the execution of collections of workflows building on our prior work on dynamic workflow partitioning and data reuse (Chen et al 2016).

5 MACHINE LEARNING FOR MODEL PARAMETERIZATION

To improve model parameterization, we use machine learning to extract dynamic data from remote sensing sources to search efficiently the parameter space.

Advances in machine learning in conjunction with the growing volumes of remote Earth observation data offer a great opportunity for extracting useful information that can feed as inputs into models. However, the ability to extract such information is limited by the significant challenges posed by the data, including violation of standard independent and identically distributed statistical assumptions (due to auto-correlation in the data across space and time), paucity of labeled data, multi-source and multi-scale data, and large volume. Our prior work on GOPHER (Karpatne et al 2016; Khandelwal et al 2017) showed that machine learning approaches hold great promise for addressing many of these challenges and advancing the state-of-the-art in monitoring ecosystem resources. However, substantial advances are needed to make these techniques effective for remote sensing data globally. Our research aims to address these challenges and transform the state-of-the-art in land use land cover (LULC) change detection for spatio-temporal geoscience data. As an example, detecting buildings and other infrastructure using high resolution imagery can provide valuable inputs to population models that aim to forecast population growth and its spatial distribution in a region.

Model Parameterization is another area where we are using machine learning. We are building upon the paradigm of theory-guided data science (TGDS) that we have recently formulated (Karpatne et al 2017). This approach introduces scientific consistency as an essential component for learning generalizable models. We plan to use TGDS to develop novel methods for model parameterization, where both physics and data science are used in a synergistic manner in hybrid-physics-data models.

5 CONCLUSIONS

This paper presents a novel approach to assisting users with cross-disciplinary model integration, using on artificial intelligence techniques to accelerate scenario analysis. These techniques include semantic representations of model requirements and data characteristics, automated planning to generate workflows that include data transformation steps, and machine learning to improve the efficiency and accuracy of various stages of the modeling process. We are implementing the MINT framework for model integration scenarios that involve climate, hydrology, agriculture, and economic modeling.

ACKNOWLEDGMENTS
This work was funded by the Defense Advanced Research Projects Agency with award W911NF-18-1-0027, and the National Science Foundation with award ICER-1440323. We thank our collaborators, in particular Bakinam Essawy, Rosa Filgueira, Jon Goodall, Claudia Medeiros, and Suzanne Pierce.

REFERENCES

Basic Model Interface (BMI) Forum on GitHub. 2018. https://github.com/bmi-forum.

Carvalho, L. A. M. C.; Essawy, B. T.; Garijo, D.; Medeiros, C. B.; and Gil, Y. 2017. Requirements for Supporting the Iterative Exploration of Scientific Workflow Variants. Proceedings of the ACM Workshop on Capturing Scientific Knowledge (SciKnow).

Chen, W., R. Ferreira da Silva, E. Deelman, and T. Fahringer, 2016. Dynamic and Fault-Tolerant Clustering for Scientific Workflows. IEEE Transactions on Cloud Computing, (4)1, pp. 49-62.

Cobourn, K.M., H.J. Burrack, R.E. Goodhue, J.C. Williams, and F.G. Zalom. 2011. Implications of Simultaneity in a Physical Damage Function. Journal of Environmental Economics and Management, 62(2): 278-289.

Cycles. 2018. http://plantscience.psu.edu/research/labs/kemanian/models-and-tools/cycles

Deelman, E., K. Vahi, G. Juve, M. Rynge, S. Callaghan, P. J. Maechling, R. Mayani, W. Chen, R. Ferreira da Silva, M. Livny, and K. Wenger, 2015. Pegasus: A Workflow Management System for Science Automation. Future Generation Computer Systems, vol. 46, pp. 17-35.

Essawy, B. T., Goodall, J. L., Xu, H., and Gil, Y. 2017. Evaluation of the OntoSoft Ontology for Describing Legacy Hydrologic Modeling Software. Environmental Modelling and Software, 92.

Garijo, D., Corcho, O., Gil, Y., Gutman, B. A., Dinov, I. D., Thompson, P., and Toga, A. W. 2014. Fragflow: Automated fragment detection in scientific workflows. Proceedings of the IEEE Conference on e-Science.

Khandelwal, A., A. Karpatne, M. Marlier, J. Kim, D. Lettenmaier, and V. Kumar. 2017. An approach for global monitoring of surface water extent using MODIS data. Remote Sensing of Environment, 202.

Karpatne, A., Jiang, Z., Vatsavai, R. R., Shekhar, S., & Kumar, V. 2016. Monitoring land-cover changes: A machine-learning perspective. *IEEE Geoscience and Remote Sensing Magazine*, *4*(2), 8-21.

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., S. Shekhar, N. Samatova, and Kumar, V. 2017. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, *29*(10), 2318-2331.

Knoblock, C.A., and Szekely P. 2015. Exploiting semantics for big data integration. AI Magazine.

Knoblock, C.A., Minton, S., Ambite, JL, Ashish, N., Muslea, I., Philpot, A., and Tejada, S. 2001. The ARIADNE approach to web-based information integration. International Journal of Cooperative Information Systems 10(1/2):145–169.

Laniak, G.F., G. Olchin, J. Goodall, A. Voinov, M. Hill, P. Glynn, G. Whelan, G. Geller, N. Quinn, M. Blind, S. Peckham, S. Reaney, N. Gaber, R. Kennedy and A. Hughes. 2013 Integrated environmental modeling: A vision and roadmap for the future. Environmental Modeling and Software, 39, 3-23.

MODFLOW. 2018. US Geological Survey. https://water.usgs.gov/ogw/modflow/

Peckham, S.D. 2014. The CSDMS Standard Names: Cross-domain naming conventions for describing process models, data sets and their associated variables. Proceedings of the 7th Intl. Congress on Env. Modelling and Software. International Environmental Modelling and Software Society (iEMSs).

Peckham, Scott D., Eric WH Hutton, and Boyana Norris. 2013. A component-based approach to integrated modeling in the geosciences: The design of CSDMS. Computers and Geosciences, 53.

Peckham, S.D., M. Stoica, E.E. Jafarov, A. Endalamaw, W.R. Bolton. 2017. Reproducible, component-based modeling with TopoFlow, a spatial hydrologic modeling toolkit. Earth and Space Science, 4(6).

Qu Y., C. J. Duffy. 2007. A semidiscrete finite volume formulation for multiprocess watershed simulation, Water Resources Research, 43(8).

Wu, B., and Knoblock, C.A. 2016. Maximizing correctness with minimal user effort to learn data transformations. Proceedings of the 21st ACM International Conference on Intelligent User Interfaces.