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/

Towards Model Integration via Abductive Workflow Composition and Multi-Method Scalable Model Execution

Rafael Ferreira da Silva[1], **Daniel Garijo**[1], **Scott Peckham**[2], **Yolanda Gil**[1], **Ewa Deelman**[1], **Varun Ratnakar**[1]
[1]*University of Southern California, Information Sciences Institute*
[2]*University of Colorado, Boulder*
rafsilva@isi.edu, dgarijo@isi.edu, scott.peckham@colorado.edu, gil@isi.edu, deelman@isi.edu, varunr@isi.edu

Abstract: Workflows provide a solid foundation to address model integration challenges. Integrated models may be simply chained, or they may need to run in an interleaved (tightly-coupled) fashion. Data exchange formats may significantly differ (e.g., scale), and data transformations may be required to convert available data into the formats required by the models. In this work, we are creating the MINT (Modeling INTegration) environment for workflow composition and execution by extending the well-established workflow composition (WINGS) and execution (Pegasus) systems with a framework for model coupling for execution interleaving (EMELI/BMI). WINGS provides a semantic workflow system that can represent and propagate constraints to validate workflows, while Pegasus enables distributed workflow execution across infrastructures and provides automated data management and fault-tolerance. BMI provides standardized, noninvasive, and framework-independent API for models. Models for integration will be selected from a Model Catalog based on variables of interest (and built on ontologies of standard variable names). Via abductive reasoning, MINT will assess the viability of workflows by hypothesizing data transformation tasks for converting available data into the formats required by the models. Data transformation services will generate multi-step scripts for accommodating the hypothesized data transformation tasks. MINT's multi-method scalable model execution will then enact the execution of the tight model coupling (using EMELI/BMI) and independent model chaining applying, when needed, the required transformations. The MINT integrated modeling environment would facilitate and accelerate modeling analysis by generating new data transformations via abductive reasoning, and by providing scalable execution of chaining or tightly-coupled models.

***Keywords**:* Model integration; scalable model execution; distributed workflow execution; semantic workflows.

1 INTRODUCTION

The Model INTegration (MINT) project [Gil et al. 2018] envisions the development of a framework for resolving semantic, spatio-temporal, and execution mismatches in model integration. MINT will provide tools to facilitate/automate the discovery of models and data to enact the model execution. In this work, we describe the novel approach for workflow composition and execution for MINT, which encompass the generation of semantic workflows using a model and a data catalogs enriched with ontologies. By using abductive reasoning and machine learning, we plan to derive data transformations when models are incompatible, or data is unavailable in the expected format. The output workflow will be implemented and executed in a distributed, heterogeneous infrastructure, and will enable the execution of both chained and coupled models.

2 PRIOR WORK

Semantic Workflow Framework – In order to generate workflows that integrate diverse models from different disciplines, we will capitalize on the recent advances of the WINGS semantic workflow system [Gil 2014, Gil et al. 2011]. 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, data-driven adaptive workflow customization, parallel data processing, workflow validation, and interactive assistance. WINGS is the basis for the MINT workflow generation framework.

Model Metadata Registry – 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 2015; Gil et al 2016; OntoSoft 2018] was developed to capture extensive information that is needed by scientists to understand how models work. Most of that information is available, is scattered in publications, manuals, code documentation, and web sites [Essawy et al 2017]. OntoSoft is the basis for the MINT Model Catalog, which will describe model invocation functions and data pre-processing workflows and use principled ontologies to represent model variables and processes.

Workflow Management System – We will capitalize on the well-established Pegasus [Deelman et al. 2015] workflow management system. Pegasus provides the necessary abstractions for scientists to create workflows, allow for transparent execution of these workflows on a range of computer platforms, and implements resource management strategies that automate computational job distribution and execution, data management, fault-tolerance, monitoring, among others. Since its inception 17 years ago, Pegasus has become an integral part of the production scientific computing landscape in several scientific communities, including astronomy, bioinformatics, biology, climate modeling, earthquake engineering and science, material science, and neuroinformatics.

Model Coupling Framework – We will leverage the EMELI (Experimental Modeling Environment for Linking and Interoperability) [Peckham 2014] modeling framework, which is designed to couple reusable component models to create new, composite models through the use of the CSDMS BMI (Basic Model Interface) [Peckham et al. 2013] for self-description and model control. A model implementing BMI provides the framework complete, fine-grained control of the model, and make the model self-describing (e.g., by exposing, for example, its list of variables, types, and units), so that the framework can use BMI functions to communicate with the underlying wrapped model (and vice-versa) and enable model coupling.

3 MINT WORKFLOW GENERATION

To generate workflows composed of diverse models and the necessary data transformation steps, we will build on the WINGS semantic workflow system to reason about data characteristics and model requirements available in the MINT Model Catalog. First, the MINT Model Catalog will include explicit representations of the variables in each model and their dependency graph. We have already represented the variables for two hydrological models: the Penn State Integrated Hydrologic Model (PIHM) [PIHM 2018], which has more than 60 variables, and TopoFlow [Peckham et al. 2017], which has more than 100 variables. Second, we will include explicit representations of the processes and methods used in a model. For example, TopoFlow can model infiltration and snowmelt processes, and can use different methods for each process. Third, we will represent how the model variables are mapped to input and output files. This includes their file structures and formats, spatial and temporal grids, values and units. We are working on representing common geosciences formats such as NetCDF. Fourth, we will represent distinctly the model invocation functions that correspond to different combinations of processes and methods when using a model. For example, an invocation of TopoFlow a third process of subsurface flow in a saturated zone in addition to infiltration and snowmelt 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 the Cycles agriculture model (http://plantscience.psu.edu/research/labs/kemanian/models-and-tools/cycles) and the MODFLOW family of models (https://water.usgs.gov/ogw/modflow) and the associated FloPy software. We also plan to include in the catalog economic models for natural resources that combine biophysical and socioeconomic data.

A user would start the workflow generation process by specifying some variables of interest (Figure 1.a). For example, a user may be interested in precipitation, crop yields, and land use in a region. These variables indicate the scope of the problem and the level of detail required of the models. The variables specified by the user are then used to select relevant models based on the model variables specified in the MINT Model Catalog. Several models may be available in the catalog to generate any given variable, so MINT would generate several possible *model groupings* (Figure 1.b) 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 (Figure 1.c). 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.

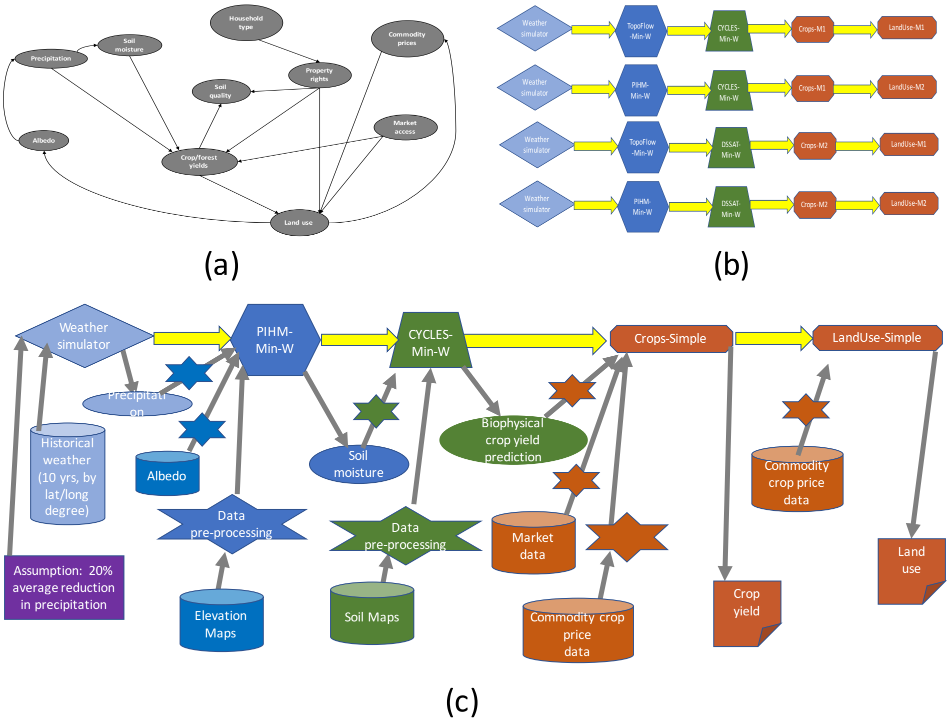

Figure 1. Example of MINT's workflow composition and abduction process.

4 MINT WORKFLOW EXECUTION

The MINT project targets the execution of model chaining and model coupling workflows in a distributed, scalable, and reliable manner. In model chaining (Figure 2-*left*), there is no execution interleaving between models, i.e., models are executed independently, and dependencies in the graph are only executed when the parent model tasks have been completed. In this model, data transformations may occur between model executions (represented as additional tasks to the workflow graph) – the subsequent models will only start their execution once the transformations have been completed. Model chaining workflows are implemented as Pegasus workflows for execution on distributed computing infrastructures.

In model coupling (Figure 2-*right*), in contrast, data exchanges occur during the execution of tightly-coupled models. The challenge is how to seamless share data among models in a coherent manner, i.e., the data produced by one model is in the required format (units, steps, etc.) by consumer. While EMELI/BMI enables model interoperability, it assumes that data is always produced/consumed in the expected format. Therefore, we are extending EMELI/BMI to support data transformations to be performed within the communication channel. Note that since these transformations may require significant processing time/power and involve large datasets or access to heterogeneous infrastructures, it is important to schedule such operations (or the required infrastructure) in advance and efficiently so that they would not become a bottleneck for the workflow execution.

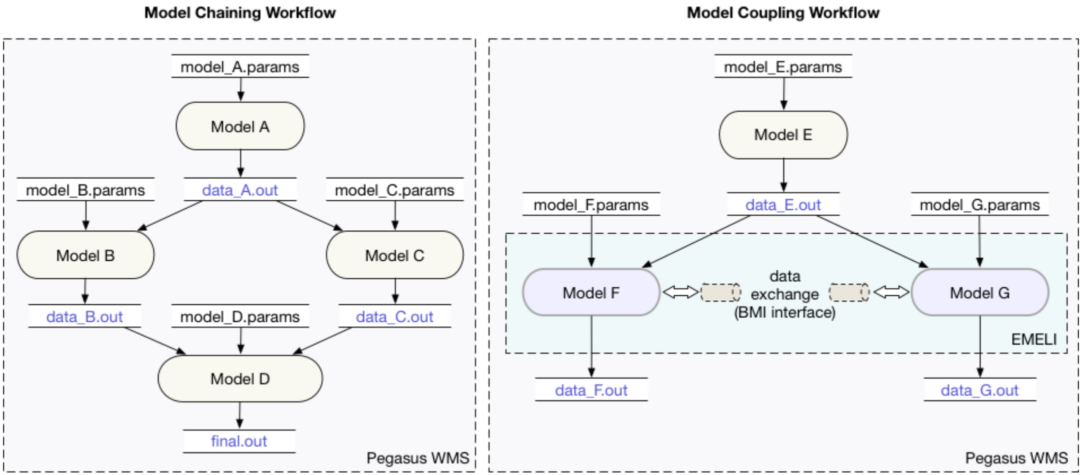

Figure 2. Illustrative example of a model chaining workflow (left) and a model coupling workflow (right) using MINT tools for workflow execution.

ACKNOWLEDGMENTS

This project is funded by the Defense Advanced Research Projects Agency under award W911NF-18-1-0027. We would like to thank our MINT project collaborations, in particular Kelly Cobourn, Chris Duffy, Bakinam Essawy, Jon Goodall, Armen Kemanian, and Suzanne Pierce.

REFERENCES

Deelman, E., Vahi, K., Juve, G., Rynge, M., Callaghan, S., Maechling, P. J., Mayani, R., Chen, W., Ferreira da Silva, R., Livny, M., and Wenger, K., 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., Gil, Y., 2017, Evaluation of the OntoSoft Ontology for Describing Legacy Hydrologic Modeling Software, Environmental Modelling & Software, vol. 92.

Gil, Y., Cobourn, K., Deelman, E., Duffy, C. Ferreira da Silva, R., Kemanian, A., Knoblock, C., Kumar, V., Peckham, S. D. et al., 2018, MINT: model integration through knowledge-powered data and process composition, 9th International Congress on Environmental Modelling and Software.

Gil, Y., Garijo, D., Mishra, S., Ratnakar, V., 2016, OntoSoft: A Distributed Semantic Registry for Scientific Software, In Proceedings of the Twelfth IEEE Conference on eScience, Baltimore.

Gil, Y., Ratnakar, V., Garijo, D., 2015, OntoSoft: capturing scientific software metadata, In Proceedings of the Eighth ACM International Conference on Knowledge Capture.

Gil, Y., 2014, Intelligent workflow systems and provenance-aware software, In Proceedings of the Seventh International Congress on Environmental Modeling and Software, San Diego, CA.

Gil, Y., Ratnakar, V., Kim, J., Gonzalez-Calero, P. A., Groth, P., Moody, J., Deelman, E., 2011, Wings: intelligent workflow-based design of computational experiments, IEEE Intelligent Systems, vol 26(1), pp. 62-72.

OntoSoft, 2018, http://www.ontosoft.org

Peckham, S.D., Hutton, E.W.H., Norris, B., 2013. A component-based approach to integrated modeling in the geosciences: the design of CSDMS, Computers & Geosciences: Modeling for Environmental Change, vol. 53, pp. 53-12.

Peckham, S. D., 2014. Emeli 1.0: an experimental smart modeling framework for automatic coupling of self-describing models, 11th International Conference on Hydroinformatics.

Peckham, S.D., Stoica, M., Jafarov, E.E., Endalamaw A., Bolton, W. R., 2017, Reproducible, component-based modeling with TopoFlow, a spatial hydrologic modeling toolkit, Earth and Space Science, special isssue: Geoscience Papers of the Future, American Geophysical Union.

PIHM: The Pennsylvania Integrated Hydrology Model, 2018, http://www.pihm.psu.edu.