20th International Congress on Modelling and Simulation, Adelaide, Australia, 1–6 December 2013
www.mssanz.org.au/modsim2013

Development of complex scientific workflows: towards end-to-end workflows

D.J. Penton [a], A. Freebairn [a], R. Bridgart [a], N. Murray [a] and T. Smith [a]

[a] *CSIRO Land and Water, Australian Capital Territory*
Email: dave.penton@csiro.au

Abstract: The analysis of water planning options on environmental assets relies on combining mathematical models from several disciplines. The growing complexity of these modelling tasks increases the potential for mistakes and misinforming stakeholders and the public. Through better capture of provenance information (audit trails), scientific workflow tools improve the transparency of model interactions, which increases our confidence in the modelling results. However, scientific workflow tools can be complex to use and increase software development costs. Consequently, they have not had widespread adoption.

This paper examines progress toward end-to-end scientific workflows. These end-to-end workflows link heterogeneous data sources through to reports that evaluate ecosystem function. The aim of end-to-end workflows is to provide a system that can evolve as understanding progresses, data services come online and reporting requirements change.

The case study for this investigation is Little Rushy Swamp. Located in the Barmah forest near Echuca, Little Rushy Swamp supports a range of bird life. Over the past century, human regulation of the flows of the River Murray has changed the timing and frequency of flooding, which has caused the deterioration in Red Gum (*Eucalyptus camaldulensis*) health and wetland habitats. In recent years, the Australian Government has bought water licenses with the aim of improving the health of riparian ecosystems. The Government has used their water licenses to provide significant environmental flooding for the Barmah Forest. This case study shows how to link a model that characterises wetland hydrology in a way that supports ongoing reporting requirements that may be necessary for monitoring or management.

The workflow was designed using tools from the Hydrologists Workbench, which aims to ease the development of complex automated workflows in Trident. In the workflow, we use a simple water balance model to understand how the water level of Little Rushy Swamp varies under different climatic periods. The water balance model has daily inputs. The inputs to the model include data from text files, web feature services and outputs from other simple numerical models. The results are analysed using R-based statistics and ArcGIS-based geoprocessing. The workflow exports resulting images to Microsoft SharePoint using its web service interface. Using links, the images are incorporated into a Microsoft Word document. The Word document updates when the images on SharePoint update, providing an end-to-end workflow.

We find that the end-to-end workflow for Little Rushy Swamp addresses a number of challenges that exist in the integrated environmental modelling space. In particular, we establish that the workflow provides a valuable tool for incorporating new or revised datasets and methods. However, the benefits of such workflows are limited by the availability of web service data feeds that use consistent data formats. Further work should be directed towards handling of uncertainty by workflow systems.

Keywords: *Scientific Workflows, Integrated Water Resource Management*

1. INTRODUCTION

Modellers face an extraordinary challenge: how to provide policy-makers with accurate information about how water resources and dependent ecosystem will behave under a changing climate and under different policy interventions (Laniak et al., 2013). The challenge is extraordinary because of the desire to:

- integrate models from many different disciplines,
- provide guidance about the uncertainty of results,
- incorporate observations from a variety of sources, and
- continually update models with new information for adaptive management.

Integrating models from different disciplines is a challenge because each model has its own model conceptualisation with associated tools and platforms. For problems where a tight coupling is required, several authors demonstrate successful implementations of OpenMI, e.g. Mansour et al. (2013) using Pipestrelle and Castronova et al. (2013b) using HydroDesktop. Castronova and Goodall (2013a) show that OpenMI computing overheads are negligible in a desktop application using .NET 4.0. The overheads are considerable when adapting OpenMI to web-based systems, according to Goodall et al. (2011) for service-oriented architectures and according to Castronova et al. (2013b) for web processing services.

Providing guidance about the uncertainty of model results is a challenge because of the computation required to test model sensitivity to input data and parameter ranges. Model agnostic uncertainty tools such as DREAM allow reasonable statistical rigour to parameter estimation problems (Vrugt et al. 2009); however, they require running a model many times. Even when parameter uncertainty is established, changing (non-stationary) climate can affect a models ability to make accurate predictions (Vaze et al. 2010). In such cases, practitioners should revisit the results when new information becomes available.

Incorporating observations from a variety of sources is a challenge because of unknown data quality, data lineage and format. Recently the Open Geospatial Consortium presented the WaterML 2.0 standard for exchange of hydrological observations, e.g. time-series of water levels at a stream gauge. However, most modelling exercises still involve exchanging hydrologic data as comma-separated files. Increasingly data is available via a website or as a web service, such as WaterOneFlow (Ames et al., 2009). Similarly, Baber and Li (2010) report using Kepler to integrate sensor observation services into scientific workflow systems.

Continually updating models with new information for adaptive management is a challenge because technical details of modelling are often lost and assumptions poorly documented. Michener (2012) describes building a scientific workflow for managing Kruger National Park in South Africa using an iterative adaptive management approach. However, in many cases, no one has kept the actual models and input data. When they are kept, there can be further confusion around the version of model codes and input data that was used.

Scientific workflows have been proposed as a solution for some of these challenges. The primary purpose of scientific workflow tools such as Kepler (Altintas et al. 2004) and Taverna (Oinn et al. 2004) is to apply gridded computing resources to difficult and often data intensive problems. However, scientific workflows also aid scientific reproducibility, record keeping (provenance trails), result processing and sharing (Gil et al., 2007, Fitch et al. 2011). Towards reproducibility, Saint and Murphy (2010) propose a model of an end-to-end workflow for examining the effect of environmental changes on local watersheds. In the end-to-end workflow, Kepler runs a series of hydrological models to transform inputs from web-based data sources into outputs delivered with web-based data sources. The aim of such end-to-end workflows is to provide a system that can evolve as understanding progresses, data services come online and reporting requirements change.

This paper examines the idea of an end-to-end workflow. The inputs to the workflow are realistic heterogeneous data feeds. The outputs of the workflow are reports. Section 2 introduces the environmental asset of interest - Little Rushy Swamp - and a simple water balance model that we use to represent it. Section 3 describes the incorporation of this model into a workflow that generates inputs for reporting. Section 4 discusses the effectiveness of the workflow in addressing integrated modelling challenges.

2. LITTLE RUSHY SWAMP MODEL

Located in the Barmah forest near Echuca in Victoria, Little Rushy Swamp supports a range of bird life. Over the past century, human regulation of the flows of the River Murray has changed the timing and frequency of flooding, which has caused the deterioration in Red Gum (*Eucalyptus camaldulensis*) health and wetland habitats. In recent years, governments have bought water licenses with the aim of improving the health of riparian ecosystems. Water managers have used these water licenses to supplement the flooding of Barmah forest to support bird-breeding events.

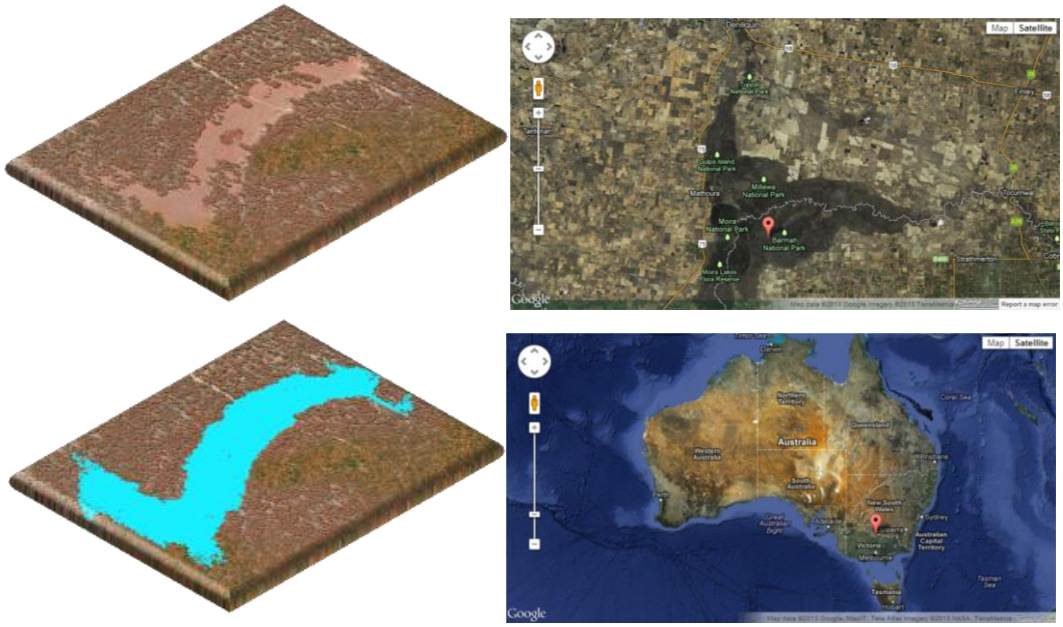

Figure 1. Right: Barmah Forest, Victoria, Australia (Google, TerraMetrics 2013). Top left: Aerial photograph of Little Rushy Swamp (700x300m). Bottom left: wetland extent when filled.

Little Rushy Swamp is a small ephemeral wetland that is 13 hectares in size. The wetland fills during high flows of the River Murray. Overton et al (2010), Penton et al (2007) and Womersley and Arrowsmith (2009) describe various methods that have been applied for modelling the connection of the river to the floodplain in the Barmah Forest. For the purpose of this paper, we will consider the wetland to fill whenever the gauge at Tocumwal (409202) upstream of the Barmah Forest is above 22GL/day.

The workflow represented Little Rushy Swamp with a simple water balance model (coded in C#). The simple water balance model calculated the wetland level from stream-flow, rainfall, evaporation and hydraulic conductivity (groundwater) as shown below:

$$\Delta\, Storage\ (ML) = Inflows\ (ML/d) + Precipitation\ (ML/d) - Evaporation\ (ML/d) - Groundwater\ (ML/d)$$

Table 1 shows Little Rushy Swamp's bathymetry from interpolating 2m resolution LIDAR (Light Detecting And Ranging) final returns (see also Figure 2). Flight lines and reed beds did not appear to distort the LIDAR response in this region. The LIDAR point density was consistent across the wetland, suggesting that there were no water returns. The aerial photography flown simultaneously would seem to confirm this.

Table 1. Bathymetry of Little Rushy Swamp from LIDAR Digital Terrain Model

Height (m)	Volume (ML)	Surface Area (ha)
0.00	0.00	0.00
0.20	0.00	0.01
0.35	0.15	0.34
0.50	2.19	3.04
0.65	11.00	8.74
0.80	27.06	12.08
0.90	38.90	13.18

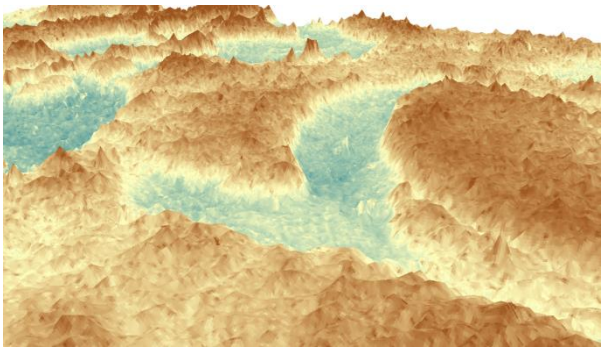

Figure 2. Little Rushy Swamp Digital Terrain Model coloured from lowest point (blue) to highest point (brown). The sill is around 0.9 metres above the deepest point.

The Deniliquin rainwater gauge (st. 74128, 41km north, similar elevation) provided rainfall. The model approximated evaporation at Kyabram (st. 080091, 48km south, similar elevation) using a cosine function – this was for demonstrating a different workflow input. The model did not consider local conditions such as humidity, temperature, wind-speed and vegetation, which influence the actual evaporation rate. Rainfall on the surrounding catchment and its influence on soil water table and local runoff were not considered. Instead, the wetland model calculated the flux in storage due to rainfall and evaporation based on the surface area.

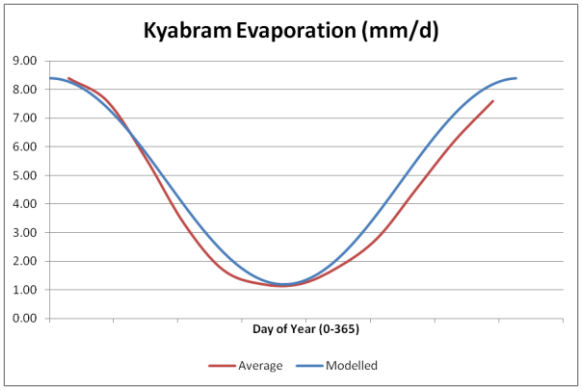

Figure 3. Comparison of average monthly evaporation rate to $evap = \cos\left(\frac{dayOfYear}{365} * 2\pi\right) * 3.6 + 4.8$

Groundwater depths, soil types and infiltration rates were not available at a resolution suitable for the size of the wetland. Instead, the model contained a groundwater loss based on the hydraulic head of the wetland. The model contained a hydraulic conductivity term of 10^{-5} cm/s based on an assumption of clay-type soil properties. The groundwater level was assumed to be constant and equal to the lowest point on the swamp.

3. LITTLE RUSHY SWAMP WORKFLOW

This study used the Trident scientific workflow system with tools from the Hydrologists Workbench. Fitch et al. (2010), Cuddy and Fitch (2010), and Box (2010) describe the tools, principles and proposed governance of Hydrologists Workbench. The scientific workflow system runs a series of codes (Activities) sequentially using .Net as the lingua franca for components. Where codes are not written in .Net, the Hydrologists Workbench provides a user interface to generate .Net wrappers to the codes. As shown in **Figure 4**, the Little Rushy Swamp workflow contains activities for model input, execution and report generation.

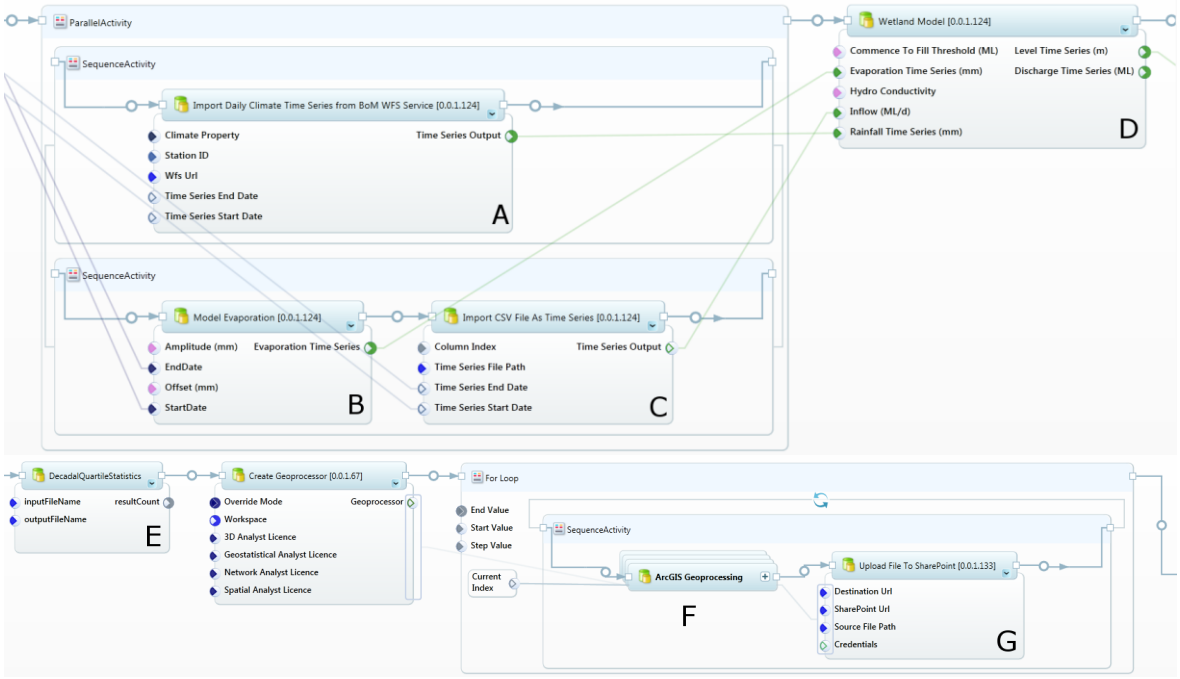

Figure 4. The workflow involves the execution of three data import activities (A, B, C), wetland model execution (D), statistical analysis (E) and result preparation (F - Geoprocessing, G - upload to SharePoint).

The temporal inputs to Little Rushy Swamp model include rainfall (mm/day), stream-flow (ML/day) and evaporation (mm/day). The scalar inputs to the model are commence-to-fill threshold (22,000 ML/day) and hydraulic conductivity (10^{-5} cm/s). The demonstration system provided temporal input to the wetland model by importing data from a comma separated file, connecting to a web feature service and running another C# model. The web feature service delivering meteorological data such as precipitation was a prototype Spatial Information Services Stack (SISS) from the Bureau of Meteorology (as in Iwanaga et al. (2013)). The C# activity interfaced the SOAP interface of the web feature service, which was running on Linux.

It would be unreasonable to expect policymakers or stakeholders to interpret the raw time series of modelled wetland levels. Instead, depth charts are prepared as shown in **Figure 5**. First, the time series were aggregated into decades then presented as quartile level values as shown in **Table 2**. The statistical routines were written in R. Since R is not a native .NET language, a C# wrapper was constructed using the activity generator described by Fitch (2011).

Table 2. Little Rushy Swamp Statistics

Decade	Quantile	Predicted wetland level (m)
1990s	0	0.10
1990s	25	0.39
1990s	50	0.64
1990s	75	0.89
1990s	100	1.11
2000s	0	0.00
2000s	25	0.25
2000s	50	12.08
2000s	75	13.18
2000s	100	0.90

For each of these rows in the table we devised a set of ArcGIS GeoProcessing commands to calculate the depth of inundation in the wetland. We used an approach similar to Penton et al. (2007) - the elevation of the water surface in Australian height datum was compared the digital terrain model.

The result of the workflow's spatial operations is a set of images shown in **Figure 5**. The workflow uses the SharePoint SOAP web interface to save the images to a SharePoint library. Separate from the workflow, the authors wrote this Modelling and Simulation paper using Microsoft Word's image links. When the workflow writes new images to the SharePoint library, the Word document updates on load.

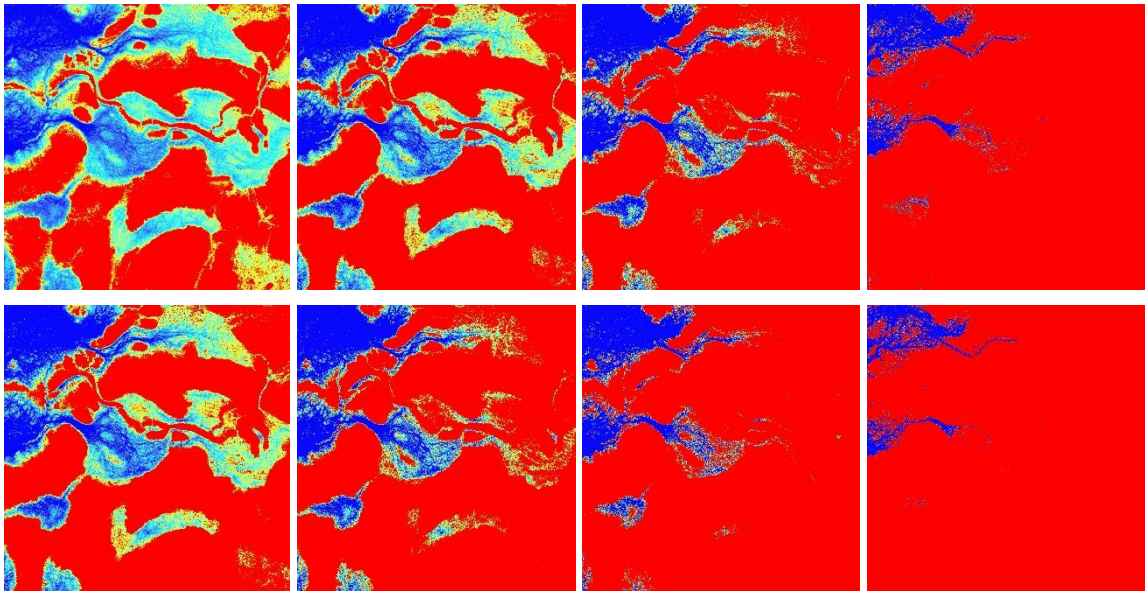

Figure 5. The images above show the inundation depth for 75th, 50th, 25th and 0th percentile flooding during the 1990s (top) and 2000s (bottom). In the images, blue represents deep inundation (up to 1 metre) and red represents no inundation. The results are illustrative and do not represent real wetland levels.

4. DISCUSSION

Little Rushy Swamp workflow presents a subset of the complexity of many integrated environmental modelling exercises. The workflow integrates models from different disciplines (hydrology and climatology), though the workflow does not contain any feedbacks or time-step dependencies. The workflow enables easily swapping or upgrading components based on better knowledge or understanding. In particular, the visual programming interface of Trident makes the major components of the analysis easily understandable. The Hydrologists Workbench activity generator component for R (Fitch et al. 2011) provided sensible mechanisms for generating and importing the code necessary for the cross-language interaction. The integration of Microsoft Word and SharePoint provides a new mechanism for aligning reports with the best available data. Further, the writing and editing of reports can occur, to some extent, independent of the modelling.

The workflow was limited in its ability to extend the temporal period because it incorporates static comma-separated value datasets. WFS servers or similar data-feeds with appropriate caching are clearly superior. However, their supply is dependent on data providers upgrading their services. The provenance of data automatically incorporated from web services needs some consideration. For example, river operators revise gauging station flows when improved rating table information becomes available.

The major limitation of this exercise was the treatment of uncertainty. While we described individual sources of uncertainty, the sensitivity of the model to that uncertainty was not established. The workflow, as presented, can calculate the results using different inputs and parameters; however, the workflow cannot present the results of a suite of possible inputs or parameters (though there is no conceptual limitation).

5. CONCLUSIONS

This paper examined progress toward end-to-end scientific workflows. These end-to-end workflows link heterogeneous data sources through to reports that evaluate ecosystem function. The aim of end-to-end workflows is to provide a system that can evolve as understanding progresses, data services come online and reporting requirements change. The focus has been solutions for complex scientific workflows that involve real world modelling.

Little Rushy Swamp workflow contains a simple physical model that provides information on the current state of inundation in a small wetland with respect to modelled previous wetland states. This paper has not covered the dynamics of ecosystem processes, or how the inundation reflects suitability as a habitat for bird breeding events. Although we derived statistics and inundations for Little Rushy Swamp, we did not validate these modelled outputs against any measurements. Given the uncertainties involved, we doubt the results would actually match reality.

However, an acknowledgement of the model's inaccuracies is the primary motivation for describing the problem as a scientific workflow. The nature of workflows is that components can be easily swapped or upgraded based on better knowledge or understanding.

ACKNOWLEDGMENTS

We would like to acknowledge the funding of this work from the Water for a Healthy Country Flagship. We appreciate the contribution of the WIRADA alliance with the Bureau of Meteorology in developing many of the tools presented here. The SISS4BOM project that provided the web feature service for meteorological data. We thank the Murray Darling Basin Authority for the provision and supply of LIDAR data for the Murray River and associated floodplains. We also appreciate the peer reviewers for their time and constructive criticism of the work presented here.

REFERENCES

Altintas, I., Berkley, C., Jaeger, E., Jones, M., Ludascher, B., & Mock, S. (2004, June). Kepler: an extensible system for design and execution of scientific workflows. In Scientific and Statistical Database Management, 2004. Proceedings. 16th International Conference on (pp. 423-424). IEEE.

Ames, D. P., Horsburgh, J., Goodall, J., Whiteaker, T., Tarboton, D., & Maidment, D. (2009, July). Introducing the open source CUAHSI Hydrologic Information System desktop application (HIS Desktop). In *MODSIM09, Modelling and Simulation Society of Australia and New Zealand* (pp. 4353-4359).

Baber, C., & Li, Y. (2010, March). Sensor information framework: Using workflow to integrate distributed sensor services. In *IEEE SoutheastCon 2010 (SoutheastCon), Proceedings of the* (pp. 60-63). IEEE.

Box, P. (2010). Hydrologists workbench: A governance model for scientific workflow environments. *In International Congress on Environmental Modelling and Software Modelling, Fifth Biennial Meeting, Ottawa, Canada.*

Castronova, A. M., & Goodall, J. L. (2013a). Simulating watersheds using loosely integrated model components: Evaluation of computational scaling using OpenMI. *Environmental Modelling & Software*, *39*, 304-313.

Cuddy, S. & Fitch, P. (2010). Hydrologists Workbench – a hydrological domain workflow toolkit. *In International Congress on Environmental Modelling and Software Modelling, Fifth Biennial Meeting, Ottawa, Canada.*

Castronova, A. M., Goodall, J. L., & Ercan, M. B. (2013b). Integrated modeling within a hydrologic information system: an OpenMI based approach. *Environmental Modelling & Software*, 39, 263-273.

Fitch, P., Perraud, J. M., Cuddy, S., Seaton, S., Bai, Q., & Hehir, D. (2011). The Hydrologists Workbench: more than a scientific workflow tool. In Proceedings, *Water Information Research and Development Alliance Science Symposium*.

Gil, Y., Deelman, E., Ellisman, M., Fahringer, T., Fox, G., Gannon, D., ... & Myers, J. (2007). Examining the challenges of scientific workflows. *Computer*,*40*(12), 24-32.

Goodall, J. L., Robinson, B. F., & Castronova, A. M. (2011). Modeling water resource systems using a service-oriented computing paradigm. *Environmental Modelling & Software*, *26*(5), 573-582.

Iwanaga, T., El Sawah, S., & Jakeman, A. (2013). Design and implementation of a Web-based groundwater data management system. *Mathematics and Computers in Simulation*.

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

Mansour, M., Mackay, J., Abesser, C., Williams, A., Wang, L., Bricker, S., & Jackson, C. (2013). Integrated Environmental Modeling applied at the basin scale: linking different types of models using the OpenMI standard to improve simulation of groundwater processes in the Thames Basin, UK. *In: MODFLOW and More 2013: Translating Science into Practice, Colorado, USA, 2-5 June 2013*.

Michener, W. K., & Jones, M. B. (2012). Ecoinformatics: supporting ecology as a data-intensive science. *Trends in ecology & evolution*, *27*(2), 85-93.

Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., ... & Li, P. (2004). Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, *20*(17), 3045-3054.

Overton, I., Penton, D., Doody, T., Saintilan, N., & Overton, I. (2010). Ecosystem response modelling in the River Murray. *Ecosystem Response Modelling in the River Murray, edited by N. Saintilan, and I. Overton*, 243-263.

Penton, D. J., & Overton, I. C. (2007). Spatial modelling of floodplain inundation combining satellite imagery and elevation models. In *MODSIM 2007 International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand* (pp. 1464-1470).

Saint K. and Murphy S. (2010). End-to-End Workflows for Coupled Climate and Hydrological Modeling. *In International Congress on Environmental Modelling and Software Modelling, Fifth Biennial Meeting, Ottawa, Canada.*

Vaze, J., Post, D. A., Chiew, F. H. S., Perraud, J. M., Viney, N. R., & Teng, J. (2010). Climate non-stationarity–Validity of calibrated rainfall–runoff models for use in climate change studies. *Journal of Hydrology*, *394*(3), 447-457.

Vrugt, J. A., Ter Braak, C. J., Gupta, H. V., & Robinson, B. A. (2009). Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling?. *Stochastic Environmental Research and Risk Assessment*, *23*(7), 1011-1026.

Womersley, T. & Arrowsmith, C. L. (2009). Barmah-Millewa Hydrodynamic Modelling Model Re-calibration, Report J727/R01 Rev 3, *Water Technology*.