*International Environmental Modelling and Software Society (iEMSs)*
*2010 International Congress on Environmental Modelling and Software*
*Modelling for Environment's Sake, Fifth Biennial Meeting, Ottawa, Canada*
*David A. Swayne, Wanhong Yang, A. A. Voinov, A. Rizzoli, T. Filatova (Eds.)*
*http://www.iemss.org/iemss2010/index.php?n=Main.Proceedings*

# Hydrologists Workbench – a hydrological domain workflow toolkit

**Susan M Cuddy[a] and Peter Fitch[a]**
[a] *CSIRO Land and Water, Canberra, Australia. susan.cuddy@csiro.au*

**Abstract:** Large-scale hydrological modelling exercises are becoming routine to address concerns about future water availability. These involve coupling multiple models to simulate the water cycle; time is of the essence; and the questions to be answered (and associated indicator metrics) are multi-dimensional. As these exercises become more complex and the volume of data increases exponentially, automating management of the flow of data through models, accessing relevant tools, while ensuring auditability and compliance, become essential. The Hydrologists Workbench (HWB) is being developed to meet this need and its shape and content are informed by recent large-scale sustainable (water) yield modelling exercises in Australia. Built on commercial off-the-shelf scientific workflow software which provides the workflow, audit and governance utility, it draws together public domain and proprietary hydrological, statistical and GIS toolkits with tailored workflows to provide an extensible portal for the provision and management of (one-off or routine) modelling exercises. This paper describes the intent, design structure and current state of development of the HWB and uses the example of a reporting workflow that executes a series of data transformations to produce maps, tables and plots for a monthly water situation report. The paper concludes by identifying key challenges that have emerged, and evaluates progress to date against a priori design objectives.

***Keywords****:* scientific workflows; hydrological modelling

## 1. INTRODUCTION AND BACKGROUND

### 1.1 Project background

The Hydrologists Workbench (HWB) as a product and a project developed from the authors' experiences with several large-scale modelling exercises in Australia (from 2007 to present). These studies (the Murray-Darling Basin Sustainable Yields project [CSIRO, 2007-2008] and its follow-on projects in Tasmania, south-west Western Australia and Northern Australia) were commissioned by the Australian Government to predict the likely impact of future climate on Australia's water resources, and were undertaken by Australia's national research organisation, CSIRO. In Australia, the availability, quality and management of water are significant national issues. Australia's water resources are under threat from a number of factors including climate change, over-allocation, and increasing consumptive demands. Knowing how much water is available, together with knowing how much water is needed for the environment and consumptive uses, provides the basis for establishing sustainable diversion limits for water resources across the country.

The suite of Sustainable Yields projects coupled climate models (how much rain), catchment water yield models (how much runoff), operational river system models provided by state water agencies, groundwater models, and water accounting models to quantify basin water availability in a suite of reporting products. Models were stitched together within a framework built on the fly by expert software engineers [Yang 2009]. While adequate for the modelling exercise, such a system cannot be expected to (and did not) provide an extensible platform for undertaking further studies. The tasks involved

were extensive, complex, time-consuming and composed of many manual steps. The generation of a typical reporting product required the collection and storage of large volumes of input files (for model calibration), control of execution of complex integrated sets of models, creation and curation of thousands of modelling results (files, databases), modelling and geo-processing artefacts (map and chart templates), and final archive and metadata tagging (repeated for every sub-basin within the project areas). These tasks could be easily envisaged as large scientific workflows, coupled with business and reporting workflows to manage the approval and publishing activities.

An equally significant initiative of the Australian Government in 2007 was to make the Australian Bureau of Meteorology (the Bureau) responsible for compiling, managing and disseminating Australia's water resources information; and making that information as comprehensive and accessible as weather and climate information through conducting periodical water resources assessments, producing annual national water accounts, and providing national water availability forecasts. This is an enormous responsibility and workload; and requires large-scale modelling exercises and reporting processes that are commensurate with those undertaken for the Sustainable Yields projects. With these shared experiences and needs, the Bureau and CSIRO formed a Water Information Research and Development Alliance (WIRADA) to address the Bureau's operational needs through CSIRO's R&D expertise in water and information sciences.

The Hydrologists Workbench project was established within WIRADA and the objectives of the project were defined in response to the Bureau's needs (and Sustainable Yields learnings). The HWB is also the name of the key deliverable of the project – a workflow toolkit, customised to meet the needs of hydrological modellers and hydrological reporting, allowing the integration and reuse of key processes. This paper describes HWB progress to date. The rest of this section describes evaluation criteria and functional requirements. Section 2 is a short discussion on the current implementation. Section 3 describes the case study and some of the challenges that it presented. The paper concludes with an assessment of progress against evaluation criteria.

### 1.2 Case for investment in scientific workflow tools

Current practices in hydrological modelling exhibit a heavy reliance on individuals and manual steps – leading to lack of resilience, poor use of valuable skills, and increased risk of non-reproducible errors. Information is dispersed depending on the culture of the modelling team. There is a heterogeneity of toolsets and capabilities with islands of specialisation (e.g. R, Excel, Perl, Matlab, Python, Fortran, C#) with manual transmission in between. There are divergent and non-conforming data formats and idiosyncratic access arrangements; and basic traceability is reliant on 'off-line' documentation (if documentation exists at all). These practices were all exhibited in the Sustainable Yields projects. Insights gained from being part of these projects, together with the similar needs of the Bureau, confirmed the need for a tool that would support the rapid construction and execution of computationally demanding integrated modelling and reporting systems within an environment governed by business rules and protocols. Scientific workflow software, as a technology for composing and executing chains of scientific processes, appeared to meet this need; and the HWB project was initiated between CSIRO and the Bureau to undertake the necessary research and development to apply this technology.

An investment in a particular technology by itself cannot guarantee improvement in practice, as there are so many other influences on modes of working (e.g. disciplinary culture, team cohesion, deadlines, organisational business rules). Significant investment needs to be supported by a convincing case identifying realisable benefits. Up-front costs associated with such an investment (research and development, training, etc.) need to be compared to the 'hidden' costs of current work practices, especially the costs of high manual handling associated with data manipulation, versioning and archiving, and the unmitigated risk of weak or no audit trails. The four realisable benefits promoted to justify investment in the HWB were that adoption would:

- facilitate investigative research and foster reproducibility

- result in more robust and efficient modelling exercises
- reduce the manual handling and tedium, resulting in less error and faster turn-over for research workflows
- improve efficiency and reduce duplication.

Progress against these evaluation criteria is summarised in the final section of this paper.

### 1.3 Hydrologists Workbench (HWB) Objectives

The project brief was to develop an integrated hydrological modelling and reporting platform, drawing upon the new technologies of web services and scientific workflows to provide a solution to integration and workflow composition for automation. Most importantly, HWB was to be built on an existing scientific workflow application. Such a platform should provide:

- the ability to discover and visualise spatio-temporal data from Bureau databases and other sources
- interfaces to external services and toolsets that manipulate and visualise hydrological data
- interfaces to existing hydrological models to allow integrated modelling and reporting tasks to be easily generated and executed
- the ability to save workflows to be rerun at a later date, thus providing the benefits of repeatability, transparency and auditability.

Benefits to the Bureau would include:

- enabling Bureau hydrologists to select and integrate components of other tools, workbenches (e.g. ArcGIS) and modelling applications; thus reducing dependency on a single workbench or modelling application
- providing a means to automate repetitive tasks, providing consistency and rigour in modelling and reporting tasks
- reducing the number of GIS and software engineers required to support integrated hydrological modelling.

It was envisaged that the interfacing of HWB to models and data would be achieved through interface tools, thus exposing the models and data for HWB users. Libraries of tools useful for hydrological modelling, e.g. geo-processing and visualisation, would be built. Importantly, the HWB must provide for the inclusion of scripts (R and Python) written by Bureau staff to provide geo-processing and statistical processing.

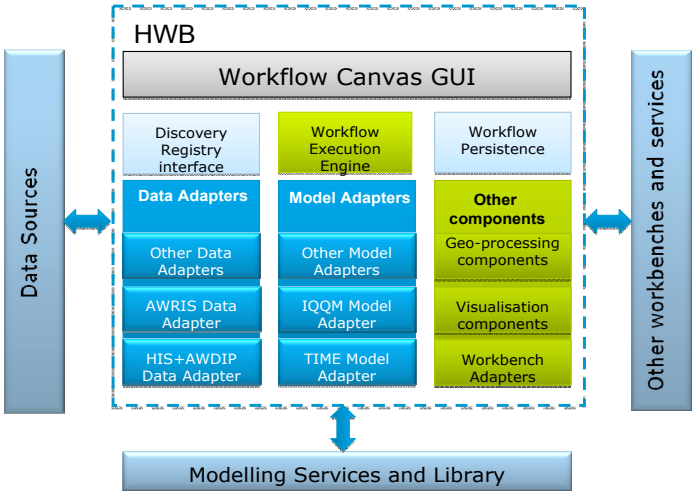
A further design imperative was that HWB comply with international and national standards and protocols, such as those being developed within WIRADA, including the Australian Hydrological Geospatial Fabric [Atkinson et al. 2008] and the Water Data Transfer standards [Walker et al. 2009] projects.

HWB would also need to be mindful of governance (of the platform and its components) requirements, both from within the Bureau and across CSIRO and the Bureau, to ensure that HWB would fulfil the objectives of reusability, accountability and auditability. An HWB governance framework is posed in Box [2010].

Other non-functional requirements address deployment considerations (lifecycle management) and usability of the product. The requirement to support the production of Bureau data products raises significant and sufficient architectural and implementation challenges to drive the R&D.

### 1.4 Linkages with Data Service and Modelling Initiatives

There are several hydrological data service and modelling initiatives underway internationally and in Australia; and the HWB needs to take advantage of these. Figure 1 is an early view of the kind of functionality to be provided through HWB.




AWRIS – Australian Water Resources Information Service – database under construction within the Bureau to ingest and serve water resources data [BoM, 2009]; HIS – Hydrologic Information Service [CUAHSI, 2009]; AWDIP – Australian Water Data Infrastructure Project to develop interoperability standards and protocols [BRS, 2009]; IQQM – River system model widely used in eastern Australia; TIME – a software development framework for creating models [Rahman et al. 2003]

**Figure 1**. An early view of the functionality to be provided through HWB, showing links to external data service and modelling initiatives

## 2. IMPLEMENTATION

The HWB project is nearing the end of the 2nd year of a five-year programme. The first year was dedicated to exploring the potential of various workflow platforms and determining their suitability for hydrological modelling; and was largely driven by the researchers within the team (see Guru et al. [2009] for details). The second year has focussed on an application case study (described below), working with Bureau staff to identify opportunities and impediments to individual and organisational adoption of the technology. Return on investment must be demonstrated at this stage to ensure ongoing support for the project.

### 2.1 Choice of Scientific Workflow Platform

Several platforms were trialled with Kepler [Ilkay et al. 2004] adopted to develop demonstration workflows. Subsequently Trident [Microsoft Corporation, 2009] was adopted. Trident is based on Windows Workflow Foundation and is part of the .NET framework. Trident provides a graphical user interface for building workflows, which are composed of activities (coded as .NET classes). These activities and workflows are managed via a registry which can be local or shared through a central registry. Being a .NET product was advantageous because many of the hydrological tools being developed within WIRADA and CSIRO are .NET applications. The selection was supported by a technical evaluation (e.g. flexibility, robustness, software interoperability) that established that Trident met the requirements for use as the core technology for HWB. More details on Trident from a HWB perspective are discussed by Perraud et al. [2010] and Box [2010].

## 3. CASE STUDY

Monthly reporting of the water situation in selected catchments across Australia was nominated by the Bureau to prototype use of the HWB. The models and analyses to produce the reports were already encoded in Python and R scripts, and it was hoped that this would facilitate rapid implementation. The workflows are to become part of a suite of scheduled monthly tasks, managed by operators other than those who have developed the workflows.

Meetings were held with Bureau staff responsible for production of the reports (and the writing of the scripts) to clarify the processing steps. Figure 2 shows a flow chart of the monthly water situation reporting, comprising:

- accessing and sub-setting gridded rainfall, soil moisture, streamflow and groundwater data based on spatial and temporal qualifiers
- aggregation at specified spatial (sub-catchment) and temporal (monthly) scales
- calculating statistics for the subsetted data
- producing artefacts (maps and graphs)
- transferring these artefacts to other products that combine with interpretive text for reporting.

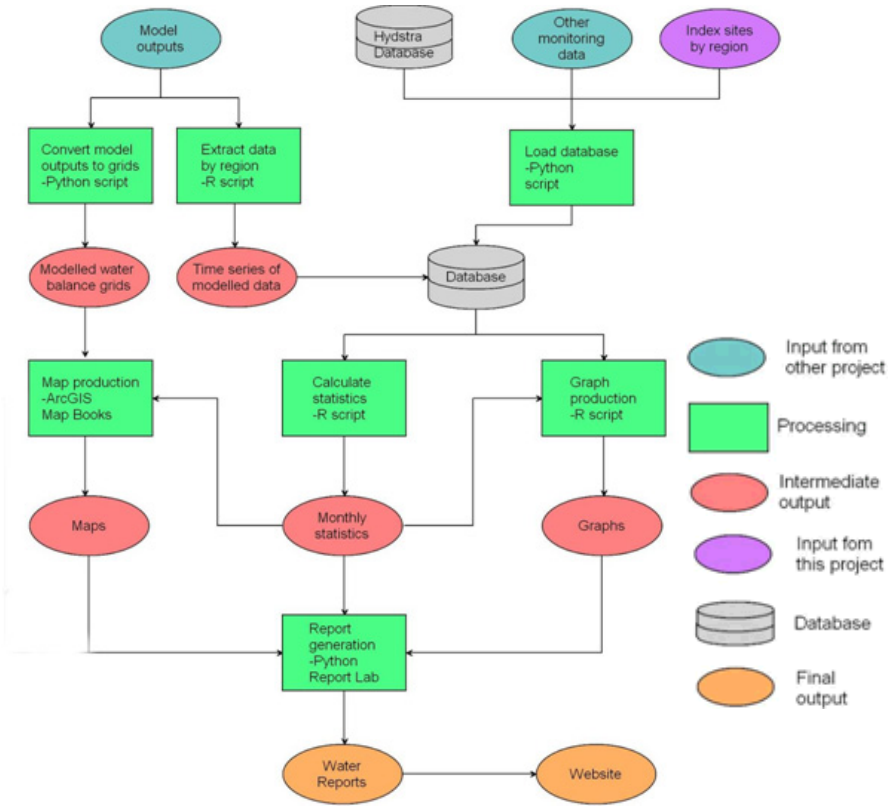



**Figure 2**. Flow chart of basic processes in the monthly Water Situation Report workflow, showing the role of Python and R scripts (reproduced courtesy of Bureau of Meteorology)

The report has been decomposed into separate workflows for each attribute (being rainfall, soil moisture, streamflow and groundwater), with the rainfall workflow being the first to be composed. This composition exposed many behavioural, technical and organisational challenges, some of which are discussed in the following section.

## 4. CHALLENGES

A preface to this discussion on challenges is an admission that scientific workflow technologies were new to the project team. Thus, it was a learning year, both in coming to grips with the capabilities (and limitations) of the Trident product, and in approaching the design of process integration/coupling differently from the more traditional approaches to design and build of modelling packages (to which the team was well accustomed). For example, there were no user interfaces to build, and no databases to design; this signalled a reduction in control of the design process that took some getting used to.

### 4.1 Technical Challenges

The most significant technical challenges encountered were how best to:

- reuse existing scripts (with minimal refactoring)
- interface with ESRI's ArcGIS
- interface with classes of TIME models
- utilise Trident registry databases for managing development, testing and deployment
- package and distribute workflows
- specify Trident activity requirements and document solution
- compose workflows
- balance flexibility with usability in the design of Trident activities and workflows.

While we made headway on all of these, progress on the first two of these challenges is described in this sub-section (paper length precluding discussion on all).

*Reusing existing scripts*
Existing Python and R scripts contained the controls to partition and manage the flow of data and processes. Rewriting the scripts was not a preferred option as a key HWB objective is to support interoperability of programming languages. A simple solution would have been to write Trident activities that 'wrapped' the existing scripts. However, this would result in Trident activities that were not reusable, and reusability is another key objective. Part of the learning exercise was to pull the scripts apart to understand the processes and refactor as smaller, potentially reusable Trident activities. During this process conventions for writing scripts were developed.

*Interface with ESRI's ArcGIS*

Spatial subsetting and zonal statistics are common processes in assessing regional water resources. The water situation report uses existing grids of rainfall, soil moisture, streamflow and groundwater levels, over the preceding month, to report on water availability. A library of Python scripts had been written to perform these tasks, calling ArcGIS functions. Initially these were invoked by writing customised Trident activities that 'wrapped' the script. A better solution was to provide full access to the ArcGIS commands from HWB (each command having its own Trident activity). Composite ArcGIS tools which string together commonly used geoprocessing sequences (e.g. Select, Mask, Clip) have been written, wrapped by customised geoprocessing workflow fragments. This has proved to be an excellent integration solution which we believe is applicable to integration with other external workbenches.

### 4.2 Behavioural Challenges

Two significant behavioural challenges arose. The first was the need to demonstrate to the script developers (who are effectively end-users of HWB) the value of using workflows rather than scripts for run control. The perceived overheads were justified in terms of the workflow approach providing transparency of process (ie the visualisation of the processing sequence was appealing), and provenance tracking of data and run execution.

The second was the separation of the role of developer (in this case the programmers who coded the Trident activities) from that of workflow composer. This was intentional as the workflow composer would typically not be a skilled programmer and a key goal of HWB is to reduce the dependence on programmers, i.e. the underlying Trident activities had to be coded in such a way that they were meaningful to and usable by others – always a challenge.

### 4.3 Governance Challenges

Governance challenges are considerable, but not insurmountable. Trident's use of a central registry for managing workflows and activities, and its ability to track the provenance of workflows and the outputs of their execution, are useful. However, within an operational environment such as the Bureau, substantial investment is required to establish a governance framework to manage development, production and deployment of activities, workflows, input data and output products – without compromising being able to use HWB as a 'sandbox' for scientific experimentation. A theoretical governance framework for HWB has been proposed by Box [2010], recognising that governance arrangements are ultimately determined by the cultural practices and obligations for quality control/assurance of the organisation. In the case of the Bureau, and the production of water resources assessment reports, every aspect will be the subject of extensive public scrutiny and strong governance is of utmost importance.

## 5. DISCUSSION AND CONCLUSIONS

*Progress against evaluation criteria*

Challenges, such as those described above, were expected. The reimplementation of the water situation report demonstrated very clearly that the adoption of scientific workflow technology is not just a technological issue – it also requires behavioural and organisational change. At the start of the project, we posed four criteria that would need to be met to justify the investment in workflow technology – that they would facilitate investigative research and foster reproducibility; result in more robust and efficient workflows; reduce the manual handling and tedium; and improve efficiency and reduce duplication. Have we met these criteria? The answer would be – not yet. While the project team was well equipped to solve many of the technical challenges, the emergent behavioural and organisation challenges require longer term investment to resolve, and must be done in partnership with the end-user, the Bureau. However, the project has demonstrated enough promise that we feel confident in our choice of technology as capable of meeting the Bureau's scientific modelling and reporting needs.

*Learnings*

The design objective of reuse and adapt (with re-invent a last resort) has had a profound effect on the relationship between the HWB professional programmers and Bureau scientists who write code (and scripts). Script writers rarely have formal programming training, yet their knowledge as encoded in their models and scripts form the core of the activities on which the workflows have been built. Any perceived reduction in coding quality is more than adequately compensated by the benefits of ownership and understanding of the underlying models and scripts by the scientists themselves.

Understanding how best to construct workflows (e.g. granularity and complexity of individual activities) is like writing good code – it comes with experience and sharing. Writing workflows that are understandable, usable and re-usable is definitely an acquired art, in fact Gil et al. [2007] describe it as a 'black art'.

*Current State of HWB and Future Plans*

The vision of building libraries of tools (delivered as activities) has been constrained to date by the need to build tools to meet the Bureau's immediate reporting needs. The workflows to produce these reports are well advanced and are becoming operational within the Bureau. Over the next year, the composition of workflows will transfer to the Bureau, allowing the development team to build the functional components envisaged in Figure 1.

Our ambition for HWB is that it is well aligned with the skills and culture within the Bureau such that it becomes the technology of choice for scientific experimentation and water resources assessment. We believe that the development of the rainfall reporting workflow, with the challenges met and those overcome, has successfully demonstrated the capability of HWB and the potential of the technology; and look forward to building HWB's repertoire together with the Bureau and the wider scientific modelling community.

## ACKNOWLEDGMENTS

WIRADA is managed within the CSIRO Water for a Healthy Country Flagship and the authors thank WIRADA management for its continuing support of the project and its objectives. The authors are indebted to the HWB project team within CSIRO and the Water Division at the Bureau of Meteorology for their contributions of background material and diagram to this paper; and the anonymous reviewers of this paper.

## REFERENCES

Atkinson, R., R. Power, D. Lemon, R. O'Hagan, D. Dovey, and D. Kinny, The Australian Hydrological Geospatial Fabric – Development Methodology and Conceptual Architecture. Water for a Healthy Country Flagship, CSIRO, Canberra, Australia, 2008.

BoM, Australian Water Resources Information System (AWRIS), http://www.bom.gov.au/water/about/publications/document/InfoSheet_3.pdf, 2009.

Box, P., Hydrologists workbench: a governance model for scientific workflow environments, [this conference], 2010.

BRS, The Australian Water Data Infrastructure Project, http://www.daff.gov.au/brs/water-sciences/ground-surface/awdi-project, 2009.

CSIRO, Water Availability in the Murray-Darling Basin. 18 regional reports, and 1 whole-of-Basin report. CSIRO Water for a Healthy Country Flagship, Australia, http://www.csiro.au/partnerships/MDBSYReports.html, 2007-2008.

CUAHSI, CUAHSI Hydrologic Information System: 2009 Status Report, http://his.cuahsi.org/documents/HISOverview2009.pdf, 2009.

Gil, Y, P.A. Gonzalez-Calero, and E. Deelman, On the Black Art of Designing Computational Workflows, Proc. 2nd Workshop Workflows in Support of Large-Scale Science (WORKS07), ACM Press, pp 53–62, 2007.

Guru, S.M. M. Kearney, P. Fitch and C. Peters, Challenges in Using Scientific Workflow Tools in the Hydrology Domain, In: Proc MODSIM09, pp 3514–3520, 2009.

Ilkay, A., C. Berkley, E. Jaeger, M. Jones, B. Ludäscher and S. Mock, Kepler: An Extensible System for Design and Execution of Scientific Workflows, Proc. SSDBM'04, 2004.

Microsoft Corporation, Project Trident: An Introduction, Microsoft Project Trident: A Scientific Workflow Workbench Version 1.0a – July 9, 2009.

Perraud, J-M, Q. Bai, and D. Hehir, On the appropriate granularity of activities in a scientific workflow applied to an optimization problem, [This conference], 2010.

Rahman, J.M., S.P. Seaton, J-M. Perraud, H. Hotham, D.I. Verrelli and J.R. Coleman, It's TIME for a New Environmental Modelling Framework. In: Proc. MODSIM03, http://mssanz.org.au/modsim03, pp 1727–1732, 2003.

Walker, G., P. Taylor, S. Cox, and P. Sheahan, Water Data Transfer Format (WDTF): Guidling principles, technical challenges and the future. In: Proc. MODSIM09, pp 4381–4387, 2009.

Yang A., An integrated catchment yield modelling environment, In: Proc. MODSIM09, http://mssanz.org.au/modsim09, pp 3577–3583, 2009.