City University of New York (CUNY)
CUNY Academic Works

International Conference on Hydroinformatics

8-1-2014

Challenges In The Simultaneous Development And Deployment Of A Large Integrated Modelling System

Matthew Paul Stenson

Peter Fitch

Dave Penton

Jamie Vleeshouwer

Ben Leighton

See next page for additional authors

Follow this and additional works at: http://academicworks.cuny.edu/cc_conf_hic

Part of the Water Resource Management Commons

Recommended Citation

Stenson, Matthew Paul; Fitch, Peter; Penton, Dave; Vleeshouwer, Jamie; Leighton, Ben; Perraud, Jean-Michel; Warren, Garth; and Collins, Daniel, "Challenges In The Simultaneous Development And Deployment Of A Large Integrated Modelling System" (2014). *CUNY Academic Works.*
http://academicworks.cuny.edu/cc_conf_hic/31

This Presentation is brought to you for free and open access by CUNY Academic Works. It has been accepted for inclusion in International Conference on Hydroinformatics by an authorized administrator of CUNY Academic Works. For more information, please contact AcademicWorks@cuny.edu.

Authors

Matthew Paul Stenson, Peter Fitch, Dave Penton, Jamie Vleeshouwer, Ben Leighton, Jean-Michel Perraud, Garth Warren, and Daniel Collins

This presentation is available at CUNY Academic Works: http://academicworks.cuny.edu/cc_conf_hic/31

11th International Conference on Hydroinformatics
HIC 2014, New York City, USA

CHALLENGES IN THE SIMULTANEOUS DEVELOPMENT AND DEPLOYMENT OF A LARGE INTEGRATED MODELLING SYSTEM

MATT STENSON (1), PETER FITCH (2), DAVE PENTON (2), JAMIE VLEESHOUWER (1), BEN LEIGHTON (3), JEAN-MICHEL PERRAUD (2), GARTH WARREN (2), DANIEL COLLINS (4), QIFENG BAI (2), ALBERT VAN DIJK (2, 5), NEIL VINEY (2), LUIGI RENZULLO (2)
(1): CSIRO Land and Water, Dutton Park, Brisbane, QLD, Australia
(2): CSIRO Land and Water, Black Mountain, Canberra, ACT, Australia
(3): CSIRO Land and Water, Highett, Melbourne, VIC, Australia
(4): CSIRO Scientific Computing, Kensington, Perth, WA, Australia
(5): Fenner School of Environment & Society, Australian National University, Canberra, ACT, Australia

Many of our natural resource management issues cannot be adequately informed by a single discipline or sub-discipline, and require an integration of information from multiple natural and human systems. As we are unable to observe and monitor more than a few important indicators there is a strong reliance on supplementing observed information with modelled information. Following a period of record drought in the 1990's, the Australian government recognised the need for better quality, more integrated, and nationally consistent water information. The Australian Water Resources Assessment system (AWRA) is an integrated hydrological modelling system developed by CSIRO and Australian Bureau of Meteorology (the Bureau) as part of the Water Information Research and Development Alliance (WIRADA) to support the development of two new water information products produced by the Bureau. This paper outlines the informatics, systems implementation and integration challenges in the development and deployment of the proto-operational AWRA system. A key challenge of model integration is how you access and repurpose data, how you reconcile semantic differences between both models and disparate input data sources, how you translate terms when passing between often conceptually different modelling components and how you ensure consistent identity between real world objects. The rapid development of AWRA and simultaneous transfer to an operational environment also raised many additional challenges, such as supporting multiple technologies and differing development rates of each model component, while still maintaining a working system. Additionally the continentally sized model extent, combined with techniques relatively new to the hydrologic domain, such as data assimilation and continental calibration, have introduced significant computational overheads. While an in-house fit for purpose operational build of AWRA is currently under development within the Bureau, the research challenges undertaken early in AWRA's development still hold many valuable lessons. We have found that the use of file standards such as NetCDF, services-based modelling, and scientific workflow technologies such as 'The WorkBench' combined with strong model governance has mostly reduced the burden of system development and deployment and exposes some important lessons for future integrated modelling and systems integration efforts.

BACKGROUND TO AWRA

During the late 1990's and early into the new century Australia was experiencing an extended period of below average rainfall. Faced with an impending crisis both within major agricultural regions such as the Murray-Darling Basin and most major urban centres, the Australian government initiated a series of national water reforms, known as the Water Act of 2007. One of the major areas targeted was water information, with the hope that higher quality, better integrated water information would allow early warning of potential issues through the generation of timely national scale water information products such water assessments and water accounts. As part of the 2007 Water Act, Australia's Bureau of Meteorology (the Bureau) was given mandate to become the home for national water information, and charged with producing the new annual national water resource assessments, and national water accounts. To support these new information products the Australian Water Resources Assessment system (AWRA) (Stenson, Fitch et al. 2011[7]) was developed through the strategic Water Information Research and Development Alliance (WIRADA) between the Bureau and Australia's Commonwealth Scientific and Industrial Research Organisation (CSIRO). The AWRA system is a near real-time, hind casting, hydrologic modelling system that seeks to provide an assessment of water dynamics across the whole Australian continent. Quantitative assessments of water fluxes and storage are informed by observed and modelled data. Typically a water assessment is synthesized from data that is extensive across space, time, and type. This paper covers the informatics, systems implementation and integration research techniques trialled as part of the development of the proto-operational implementation of AWRA, and discusses many of the lessons learnt. An in-house fit for purpose operational build of AWRA is currently under development within the Bureau.

THE NEED, DESIGN CONSTRAINTS, AND GUIDANCE FROM THE CLIENT

By making the Bureau the custodian of water information within Australia, the Water Act of 2007 mandated that all organisations provide water information to the Bureau, and in turn the Bureau would publish a nationally consistent view of this data, as well as producing several new water information products; the National Water Account (NWA) and the National Water Resources Assessment (NWRA). While both these products would be retrospective, the sparse nature of observed data, which is often incomplete in both space and time, meant a data model fusion approach of observed and modelled data was necessary. Additionally water storages such as soil moisture are not widely monitored so need to be modelled to give a complete picture on water movement, use, and storage. The AWRA modelling system was developed from the ground up using data model fusion approaches (Van Dijk et al., 2011 [8]) to support the NWA and NWRA information products. Model requirements included;

- Best estimates of all water stocks and flows within and between landscape, surface water, groundwater, rural and urban water systems.
- National scale with local detail
- Longest time period possible, sub-annual reporting and current relevance
- Transparent and high quality (benchmarked and peer reviewed) estimation methods
- Consistent with measurements
- From observation and modelled data, using state-of-the-art science and technology

In addition to the requirements listed above, AWRA had a series of functional and non-functional requirements related to the system itself;

- Single modular modelling system
- The system should also be able to be run in continuous mode i.e. a near real time system with automated data feeds
- Assimilation and use of relevant data where appropriate
- Model calibration (i.e. parameter value optimisation) and benchmarking (i.e. evaluation of estimates against independent observations)
- Simultaneous development and transfer to operations

Each of these requirements brings its own challenges. Particular challenges were introduced by the following requirements:

1. modular software design;
2. near real time production;
3. data assimilation;
4. complex model calibration; and
5. simultaneous development and transfer into operations.

Modular systems

Modular systems require defined and stable interfaces between components, which is difficult to achieve when both the system and each of the model components is under continuous development and deployment to operations. For models to integrate they need to communicate, tight coupling reduces modularity and flexibility, and often restricts the mixing of technologies. Loose coupling requires a slow and somewhat messy exchange of files. Additionally, semantic meaning, underlying concepts and dimensions can all differ between model components. For example, ‘runoff’ in the Landscape model corresponds to ‘inflows’ in the River model, where the first is a spatially distributed quantity (with units mm d^{-1}) and the second is a spatially explicit subset of point flow volumes (with units m^3 d^{-1}). These types of differences can require costly transforms.

Near real time systems

Near real time systems require stable operational data feeds which can be especially problematic with new model components as they often make use of non-operational data such as remotely sensed products. Additionally, data often comes from a wide range of sources, requiring significant work to compile. It can be provided in different formats, with differing levels of quality assurance, and often not well described with metadata. Semantic meanings can also vary subtly between data providers making repurposing and use problematic.

Data assimilation

Data assimilation can be both computationally and data transfer intensive. If using an ensemble technique will require many tens to hundreds of model instances to be run and managed at each time step.

Calibration

Like data assimilation, calibration can be extremely numerically expensive, requiring many model instances to be run and managed in controlled experiments. In the case of AWRA a spatially lumped calibration over 300 unimpaired catchments (one model instance for each) by 10,000 model runs by 10,000 time steps by 17 free parameters creates a huge computational challenge (Vleeshouwer, Perraud et al. 2013[9]). A gridded calibration is as above but works at the level of the grid cells for the 300 unimpaired catchments, so >9,000 model instances rather than 300. Additionally rigorous validation and benchmarking is needed to assess the results of model changes and calibration experiments and requires both tools and human workflows to be developed and supported.

The Transfer of Research to Operations

Development of model components and processes, and updates to forcing data and parameters while simultaneously transferring to an operational environment can cause many issues with model versioning, technology stack requirements, deployment testing and support. Additionally many of the technologies used by either the domain specialists developing the models (often due to familiarity), or the system developers (for stability and scalability) are not supported within operational environments, due to the paramount need for stability and system up time which is greatly supported by a narrow and well supported technology stack. The path to operations needs to be clearly defined early in the project, but we have found this is not always possible.

THE AWRA SYSTEM

AWRA (Stenson, Fitch et al. 2011[7]) is a modular scientific workflow with two main model components (Figure 1);

1. AWRA-LG (Van Dijk, Bacon et al. 2011[8]) is the landscape component of AWRA. It is a gridded water balance model with continental coverage and operates on an approximate 5km grid, covering almost 270,000 model instances at a daily time step, with a temporal period of 1911-present day. The groundwater component of AWRA (Crosbie, Peeters et al. 2011[1]) simulates components such as flood recharge, surface water groundwater interactions and groundwater equilibration between AWRA-LG model cells.
2. AWRA-R (Lerat, Paydar et al. 2011[4]) is the river modelling component of AWRA. It is a node link network model that uses a schematised system to define diversions, inflows and outflows in a standardised way. AWRA-R presently covers approximately 600 river reaches at a daily time step with a temporal period of 1970 to present day.

A more detailed description of the overall model can be found in Vaze, Viney et al. 2013[12].

The proto-operational version of the AWRA workflow is executed by Delft-FEWS (Werner, Schellekens et al. 2013[11]) which has provided a high degree of modularity and configurability.

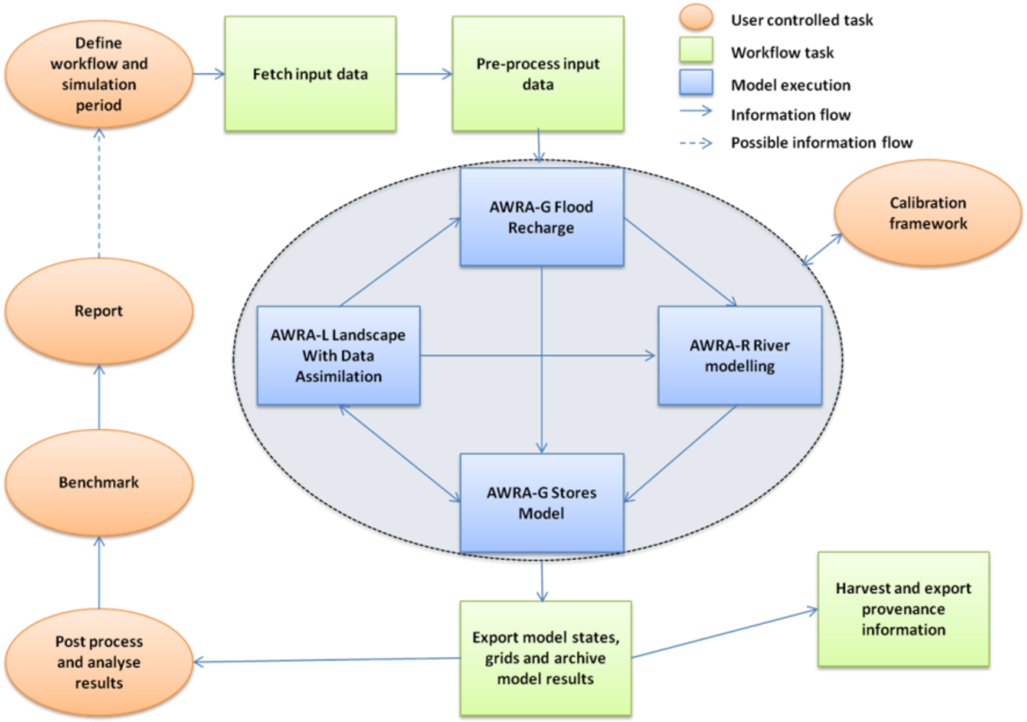

Figure 1. The AWRA workflow showing workflow (green), modelling (blue) and human tasks (orange).

CHALLENGES AND RESPONSES

Each of the five challenges outlined above was addressed during the development and deployment of AWRA;

1. Modular systems

Modularity was addressed through the use of scientific workflows, Delft-FEWS (Werner, Schellekens et al. 2013[11]) for the operational system, and The Workbench (Fitch, Perraud et al. 2011[2]) for the calibration system. These scientific workflows:

1. encourage a clear *separation of concerns* which addressed one of the main design briefs around modularity.
2. aid in *re-use and re-purposing* of model components. This is especially useful with regards to some of the pre- and post-processing tools.
3. through their inherent modularity, allowing for *staggered development* by multiple teams, by allowing different model development teams to plug new versions of the models into the larger AWRA workflow.
4. aid in the *capture of provenance information* by allowing provenance harvesting tools to be developed at a workflow rather than model level. This enables provenance information to be captured at a more realistic granularity, and at a lower cost to the model developers.

5. proved an *effective scenario management* tool. This was especially useful during model testing and calibration where many full historic AWRA runs would be performed, tracked, and reported on.
6. An additional benefit of using Delft-FEWS was its use of an internally consistent data model, and associated transformational tools. This gave the various components of the workflow a semantically resolved point of truth.

2. Near real time production

The main concern with the second challenge was with establishing automated data feeds, both from the various agencies around the country providing measured data to the Bureau, and quality controlled data feeds from Bureau services into the AWRA system itself, in addition to some considered data feeds from organisations abroad (e.g., satellite products). The data transfer and ingestion to the Bureau from data providers was handled through the development and use of a water data transfer standard, WDTF2 (Walker, Taylor et al. 2009[10]). The ingestion from the Bureau into the model itself was only partially managed because the Bureau's AWRIS data warehouse did not become available at the time expected. The risk around data feeds was reduced through the use of standard data transfer formats such as NetCDF with embedded machine readable metadata. At the time of writing most of the data feeds remain manually assembled (Figure 2) as opposed to the automated feeds that should drive the system in the future (Figure 3). Data standards have an agreed file format to lower the cost of developing tools for ingestion and manipulation and use agreed terms to represent the data items, and often are accompanied by published vocabularies. Data services provide a consistent point of truth, they are discoverable and aid in re-use. They also assist in the capture of provenance trails through providing machine readable metadata and unique and resolvable links to data products.

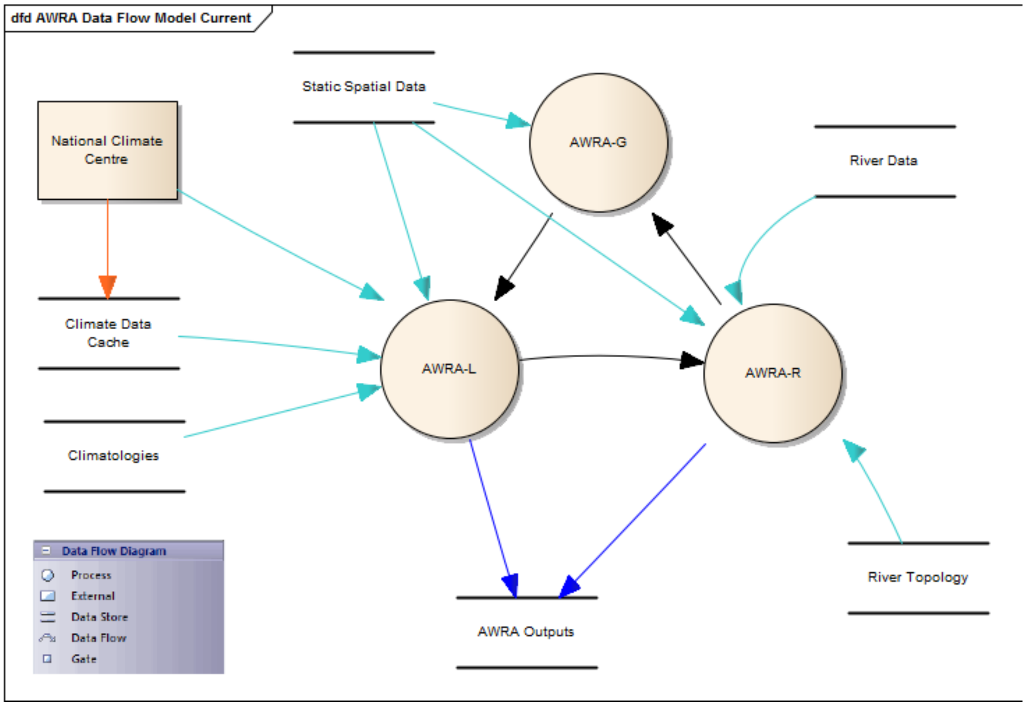

Figure 2. High level representation of data flows within the AWRA system. Note the barred data sources are internal ad hoc rather than operational data sources. Orange arrows are ASCII grids via FTP delivery, teal arrows are binary files via direct transfer, blue arrows are NetCDF export to THREDDS server and black arrows are PI-XML via Delft-FEWS internal data store. After Fitch, Brodaric et al. In Review[3].

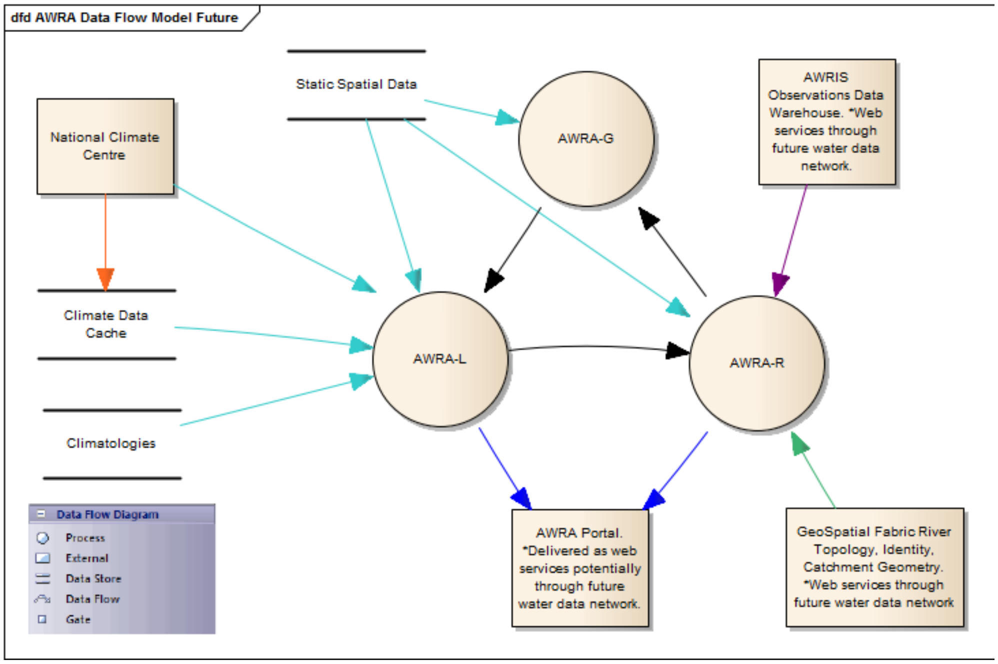

Figure 3. Future high level representation of future idealised data flows for the AWRA system, showing the current ad hoc data streams replaced by operational services. Note the barred data sources are internal ad hoc rather than operational data sources. Orange arrows are ASCII grids via FTP delivery, Teal arrows are binary files via direct transfer, blue arrows are NetCDF export to THREDDS server, green arrows are GML via web services, mauve arrows are WaterML2 via web services and black arrows are PI-XML via Delft-FEWS internal data store. Currently several of the data streams are not available as services and so are manually prepared. After Fitch, Brodaric et al. In Review[3].

3. Data assimilation

The third challenge was handled through the adoption and use of services based modelling. The two most computationally intensive components of the system (AWRA-LG and AWRA–R) were designed and built to run as services called from within the main AWRA workflow. Through the separation of compute from orchestration, these computationally intensive processes could be hosted on more powerful, dedicated hardware, and more instances of the modelling service spawned as required. Also, as the AWRA system is stateless, running model instances as separate services allowed model state to be held warm in the memory of each service instance, eliminating the need to load them at each time step as would have been required if calling a normal executable from within the workflow to only run a time step at a time.

4. Complex model calibration

Challenge four was calibration and benchmarking. As with data assimilation, computational loads are very high. However, unlike data assimilation the workflow requires multiple complete model runs, making it easier to distribute the process across multiple systems, as close coupling between instances is not needed. An additional challenge with calibration is the management and benchmarking of multiple scenarios, each consisting of a combination of changes to model code, parameters or the forcing data. This was handled by using a workflow tool ‘The Workbench’ (Fitch, Perraud et al. 2011[2]), which is built on Microsoft Trident and allows the capture of scenario and provenance information at a calibration experiment level. The Workbench also offloaded the actual calibration tasks to Windows high performance computing resources, and makes use of the CSIRO Metaheuristics toolset (Perraud, Wang et al. 2012[6]). In addition to the calibration workflows developed for AWRA, a set of modular and flexible benchmarking scripts statistically and graphically comparing model estimates to independent observations were developed in R to facilitate structured and repeatable generation of “benchmark cards” of AWRA, allowing comparison against previous versions of the model, as well as against peer models. The benchmarking model code was called as the final step in the calibration workflow to allow rapid and independent analysis of model changes and updates as part of the governance process.

5. Simultaneous development and transfer into operations

The fifth challenge is the management and coordination of model development, and simultaneous transfer to operations. This was an especially difficult task with multiple model development teams, a systems implementation team, and a set of stakeholders within the client organisation with diverse expectations around timelines and requirements. The Transfer of Research to Operations (TROPS) for AWRA was handled through a dedicated governance process (Figure 4). This governance was needed to:

1. translate client needs into science and technology questions;
2. control the flow of research into development, testing and finally operations;
3. encourage discussion, peer review, validation and testing;
4. aid in the simultaneous development of multiple components of an integrated modelling system;

The introduction of a governance mechanicsm did not always sit comfortably with the general management structure of the developer and client organisations, and the separate governance process that existed for the project overall. In particular, project and organisational management processes were sometimes more directive and less consultative than the system governance processes. On occasion this created tension, confusion and inefficiencies, particular where individuals did not have a clear role in the system governance process., but overall it had a positive effect.

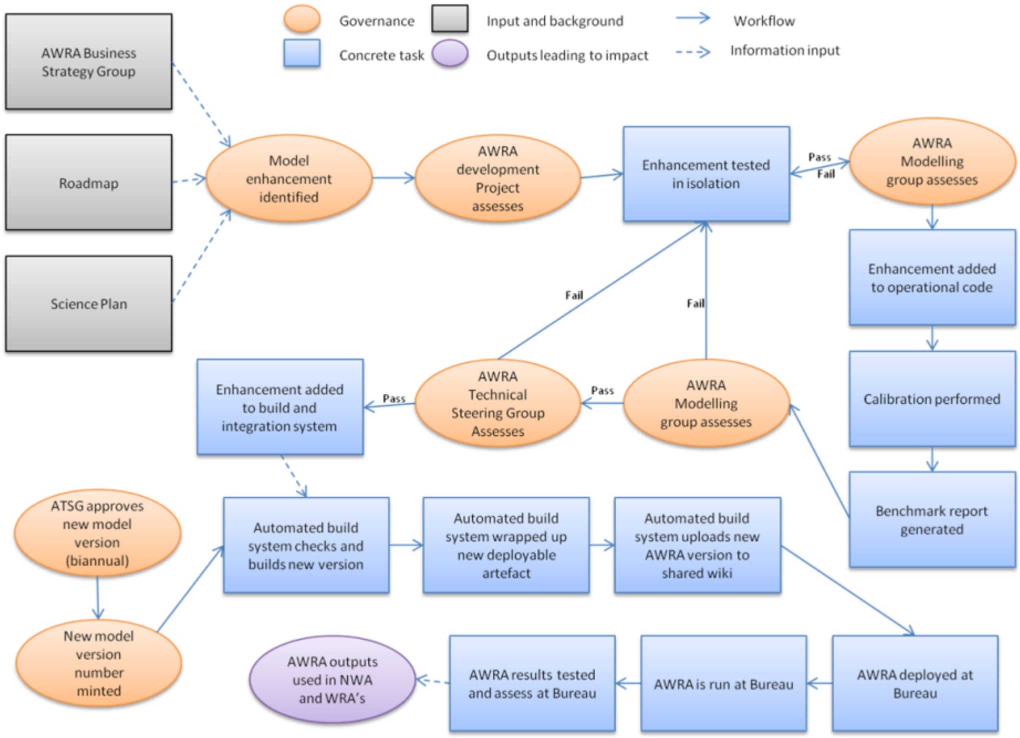

Figure 4. AWRA Governance workflow showing how the system needs are realised through model development and testing, which are then benchmarked and moved into operations.

DISCUSSION AND CONCLUSIONS

This paper has shown an example of developing a large integrated modelling system from a systems implementation perspective, some of the challenges that will likely be faced, and some of the technologies used by the authors to overcome or mitigate them. Some key points are as follows.

1. Data services using standards are crucial but not always available in a timely manner, especially for newly developed data products.
2. Scientific workflows allow a clear separation of development concerns, and assist with deployment, governance and provenance.
3. Services-based modelling allows a clear separation of concerns between orchestration and computation, allowing computation to better be distributed, scaled and managed. However there are overheads both in development (e.g. the translation or restructuring of research code) as well as in run time.
4. Simultaneous operationalisation and research is challenging. In addition to finding the right scientific and technological solutions, the process of transferring from research and development requires strong governance.
5. Fast-paced development meant pragmatic technology choices had to be used at times; these choices were not always supported within the operational environments.

ACKNOWLEDGEMENTS

This work was funded through WIRADA, an alliance of the Bureau of Meteorology and CSIRO's Water for a Healthy Country Flagship.

REFERENCES

[1] Crosbie, R. S., et al. (2011). AWRA-G: A groundwater component to a continental scale landscape model. MSSANZ 2011 Perth.

[2] Fitch, P., et al. (2011). The Hydrologists Workbench: more than a scientific workflow tool. Proceedings, Water Information Research and Development Alliance Science Symposium.

[3] Fitch, P., et al. (In Review). Integrated Groundwater Data Management.

[4] Lerat, J., et al. (2011). Better use of prior information in the calibration of river system models. 20th MODSIM Congress. Adelaide, Australia.

[5] Perraud, J. M., et al. (2013). A balancing act in heterogeneous computing - Developing the AWRA-Landscape data assimilation system. 20th MODSIM Congress. Adelaide, Australia.

[6] Perraud, J. M., et al. (2012). A data model for a sustainable management of parameter sets and optimization results. International Congress on Environmental Modelling and Software Managing Resources of a Limited Planet, Sixth Biennial Meeting. R. Seppelt, A. A. Voinov, S. Lange and B. D. Leipzig Germany.

[7] Stenson, M., et al. (2011). Operationalising the Australian Water Resources Assessment (AWRA) system. Proceedings, Water Information Research and Development Alliance Science Symposium.

[8] Van Dijk, A., et al. (2011). Design and development of the Australian Water Resources Assessment system. Proceedings, Water Information Research and Development Alliance Science Symposium.

[9] Vleeshouwer, J., et al. (2013). Using scientific workflows to calibrate an Australian land surface model (AWRA-L). 20th MODSIM Congress, Adelaide, Australia.

[10] Walker, G., et al. (2009). Water Data Transfer Format (WDTF): Guiding principles, technical challenges and the future. 18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation, Modelling and Simulation Society of Australia and New Zealand and International Association for Mathematics and Computers, in Simulation.

[11] Werner, M., et al. (2013). "The Delft-FEWS flow forecasting system." Environmental Modelling & Software 40(0): 65-77.

[12] Vaze, J., N. Viney, M. Stenson, L. Renzullo, A. Van Dijk, D. Dutta, R. Crosbie, J. Lerat, D. Penton and J. Vleeshouwer (2013). The Australian Water Resource Assessment Modelling System (AWRA). 20th MODSIM Congress, Adelaide, Australia.