Approaches to distributed execution of hydrologic models: methods for ensemble Monte Carlo risk modelling with and without workflows

Trevor Pickett[a], Tim Smith[a], Ben Bulluss [a], Dave Penton[a], Luk Peeters[a], Geoff Podger[a], Susan Cuddy[a]

a) CSIRO Water for a Healthy Country, Australia, CSIRO Land and Water Australia
Email: Trevor.Pickett@csiro.au

Abstract: Recent water reforms in Australia and the release of the Murray-Darling Basin Plan have been supported by climate models and detailed hydrological modelling including river system models. Sensitivity analysis of these river system models provides valuable insights into the often complex and non-linear relationships between uncertainty in input variables and parameters and model outputs. An understanding of these relationships is an important component of assessing the risks in the planning process.

However, a comprehensive sensitivity analysis is computationally intensive, requiring many thousands of simulations to examine a few parameters and may require months of computer time to complete. In this paper we consider a sensitivity analysis of a river system model using new and emergent technologies and discuss the merits of four methodologies for undertaking this analysis. In each case some new tools and techniques have been developed and these are applicable to sensitivity, uncertainty and error analysis of other simulation models.

The Murray-Darling basin is represented by a range of regional river models that are connected together to describe the entire basin. CSIRO recently calibrated regional *Source* models that, when combined, describe all of the Murray-Darling Basin. The Murrumbidgee regional model was selected from this project and subsequently simplified to reduce the runtime while still being representative of the system's behaviour. As part of a risk assessment, the sensitivity of this model was explored.

The sensitivity analysis examined uncertainty in inflows, rainfall, evaporation and groundwater/surface water interaction, via 100,000 simulations and the results can be found in Peeters *et. al.* (2013), submitted to this conference. The four methodologies considered to support this work are:

1. Running all 100,000 simulations on a single computer;
2. Running the simulations using several dedicated machines;
3. Running the simulations using ad-hoc computing resources; and
4. Multi-core execution, where runs are executed on a cluster.

Method 1 was the simplest, but requires the most computer time. Method 2 improved total runtime, but required dedicated computer resources. Method 3 gave reasonable runtimes, did not require dedicated resources, but did require constant monitoring and input. Method 4 was the most complex to configure, but provided very fast runtimes and automated input and output marshalling and cluster job creation and submission.

Methods 1 through 3 used Source's command line interface, while for method 4 the Source model was imported as a workflow activity into Project Trident via 'the Hydrologists Workbench'. Project Trident is a scientific workflow system developed by Microsoft Research and the Hydrologists Workbench is a suite of add-on tools for Trident developed by CSIRO's Water for a Healthy Country Flagship. Using Trident and the Hydrologists Workbench for sensitivity analysis allows the modeler to easily leverage available resources without requiring extensive or complex coding.

Keywords: *Sensitivity Analysis, multi-core execution, river system model*

1. INTRODUCTION

Recent water reforms (NWI, 2004) in Australia and the release of the Murray-Darling Basin Plan (MDBA, 2012) have been supported by climate models and detailed hydrological modelling including river system models. The river system models provide a mechanism to simulate the impacts of climate, demand, infrastructure and management based changes on the river system. A sensitivity analysis of the river system model (analyzing how changes in inputs and parameters affect changes in outputs) provides insight into the model and the underlying uncertainty in outputs. This is of great benefit to the water planner in understanding the risks associated with the model results.

The Murray-Darling basin is represented by a range of regional river models that are connected together to describe the entire basin. CSIRO recently calibrated regional Source models that when combined describe all of the Murray-Darling Basin (Dutta *et. al.*, 2012). The Murrumbidgee regional model was selected from this project and subsequently simplified to reduce the runtime while still being representative of the behaviour. As part of a risk assessment the sensitivity of this model was explored.

Sensitivity analysis of a river system model can:

- Test the robustness of the model results or the underlying system in the presence of uncertainty;
- Provide valuable insights into the often complex and non-linear relationships between input parameters and model outputs;
- Quantify and identify where to reduce uncertainty in the model; and
- Reveal model errors.

Sensitivity analysis is usually undertaken by running the model a (large) number of times while varying model parameters using a specific sampling-based approach. This method has a number of challenges that need to be addressed:

- River system models are often quite complex and have long run times (minutes to hours). Running such models repeatedly can require weeks or months of compute time;
- Sensitivity analysis algorithms are generally linear in nature, therefore significantly large numbers of simulations will be a computationally intensive exercise even for models with very short execution times;
- Simulations may fail because programmatically assigned parameters explore untested areas or because of machine or network failure. Failed runs need to be identified and rerun if required;
- Running large numbers of simulations also have implications for reproducibility as there is often little or no provenance captured.

There are many techniques for overcoming these challenges; one technique is to use scientific workflows. Scientific workflows provide access to distributed computing resources (Altintas *et. al.*, 2004, Oinn *et. al.*, 2004) and capture provenance (Oinn *et. al.*, 2004, Fitch *et. al.*, 2011).

2. IMPLEMENTATION

In this paper we examine a sensitivity analysis of five parameters of a river system model, requiring 100,000 model runs. The values of the input variables were scaled using Sobol sampling (Sobol, 1976). The sampling procedure is covered in detail in a companion paper submitted to this conference (Peeters *et. al.*, 2013) and is not discussed here. The output from the Sobol sampling was a matrix of 100,000 rows, each representing a set of parameter inputs for a single simulation, referred to hereafter as the Input Matrix.

Each row of the Input Matrix (and hence each simulation) was assigned a Globally Unique Identifier GUID) that was incorporated into the names of all input and output files generated by the simulation. Thus by comparing the GUIDs for an input matrix against output filenames one could easily identify failed or incomplete simulations. Figure 1 shows a flow diagram illustrating the logic of this process. The steps in the diagram are as follows:

1. Start by selecting an Input Matrix Row;
2. Check for an output file containing the GUID from this row;
3. If it exists return to start;

4. Else check for an input file containing the GUID;
5. If it doesn't exist create it;
6. Run the simulation;
7. Return to start.

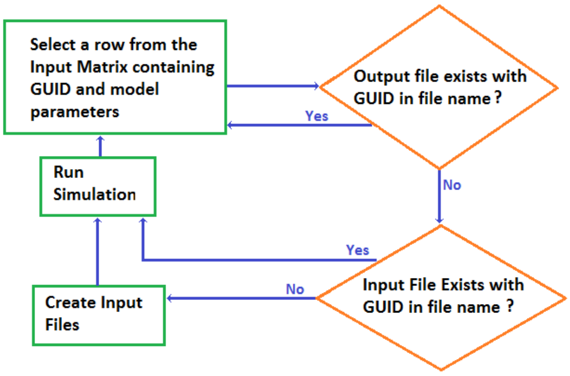
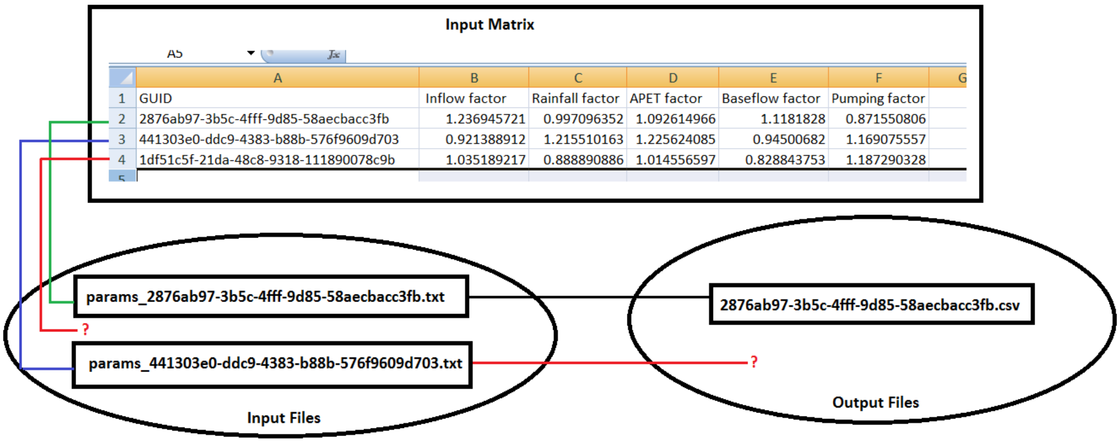

Select a row from the Input Matrix containing GUID and model parameters
Output file exists with GUID in file name ?
Yes
No
Run Simulation
Yes
Input File Exists with GUID in file name ?
No
Create Input Files

Figure 1: Flow diagram for simulation running logic

Figure 2 shows a simplified system state. If the simulations were restarted from this state, row 2 would be selected from the Input Matrix. An output file would be identified and the next row would be selected. The test for an output file for row 3 would indicate no file, the test for an input file would indicate a file. The simulation would be run and the next row would be selected. The test for output and input files for row 4 would indicate neither file existed. The input file would be created, the simulation would be run and the next row would be selected. This would continue until all rows have been selected. The importance of this is that when re-running simulation blocks, only incomplete simulations are re-run, which reduces the number of runs and overall run time.



	A	B	C	D	E	F	G
1	GUID	Inflow factor	Rainfall factor	APET factor	Baseflow factor	Pumping factor	
2	2876ab97-3b5c-4fff-9d85-58aecbacc3fb	1.236945721	0.997096352	1.092614966	1.1181828	0.871550806	
3	441303e0-ddc9-4383-b88b-576f9609d703	0.921388912	1.215510163	1.225624085	0.94500682	1.169075557	
4	1df51c5f-21da-48c8-9318-111890078c9b	1.035189217	0.888890886	1.014556597	0.828843753	1.187290328	
5							

params_2876ab97-3b5c-4fff-9d85-58aecbacc3fb.txt
?
params_441303e0-ddc9-4383-b88b-576f9609d703.txt
Input Files
2876ab97-3b5c-4fff-9d85-58aecbacc3fb.csv
?
Output Files

Figure 2: System state showing the relationship between Input Matrix GUID and model files.

The model used in our test case, the idealized river system model, mirrors the flow regime and management practices of the Murrumbidgee River within Australia's Murray Darling Basin. It was created using the eWater Source model (Welsh et al 2013). Execution time is consistent between runs, and on our machines took approximately 2 minutes.

Source models are generally constructed and executed via Source's graphical interface, but this is unsuitable for running large numbers of simulations as a user must trigger each run. Source models can also be run via a command line interface (Source Command Line) which can be called programmatically and this is the preferred option for running multiple simulations. The Source command line executable can be wrapped in a

Trident workflow activity via the Hydrologists Workbench (Fitch et. al., 2011). This can then be used in a workflow together with activities for starting and terminating the Source server. The total set of simulations is partitioned in smaller blocks (typically of size 2,000 – 5,000). Partitioning serves a dual purpose: firstly it is a simple and efficient way to balance execution load over multiple computers and secondly it is much easier to check and rerun 5,000 simulations than 50,000 simulations. Model inputs (and the GUID) are read from an input matrix file which is also partitioned according to block size. Completed blocks of outputs are moved to a central repository and analysis is performed in a post-processing step using Python scripts written specifically for this purpose.

This paper is an assessment of four methods for running the 100,000 runs required by the sensitivity analysis. In method 1, all simulations are run sequentially on a single computer. In method 2, the set of simulations is split into subsets and distributed and run in parallel on different computers. When any subset on any computer finishes a fresh one is started. Method 3 uses a similar technique to method 2 except that ad-hoc computing resources are used as, and when, available. Methods 1 – 3 execute using a simple batch file approach so that for a block of *n* simulations, the modeler creates a batch file that utilizes the methodology previously described. This ensures that if a subset terminates prematurely and is restarted, only incomplete simulations will be run.

In method 4, we use a multi-core technique where blocks of runs are assigned to nodes on a computer cluster. This technique is executed via the Hydrologists Workbench, a scientific workflow program. A scientific workflow is a sequence of processing steps which, when linked, form a scientific processing task. The Hydrologists Workbench (*ibid.)* is an intuitive platform for creating and executing workflows. It is built upon Microsoft's Trident workflow engine and uses *Activities* to represent each of the processing steps. It extends much of the functionality of earlier versions of the Hydrologists Workbench, including the Activity Generator so that hydrological models such as Source (Welsh *et. al.,* 2013) and MODFLOW (Harbaugh *et. al.*, 2000) can be wrapped as Activities. To perform sensitivity analysis using the Hydrologists Workbench we require two workflows. The first (inner) workflow gathers model inputs, runs the Source model and outputs results. The second (meta) workflow runs the inner workflow a requisite number of times, passing correct inputs and redirecting outputs to the correct location. We also use a purpose built executable which allows Trident workflows to be run without installing Trident on the host machine. This executable is passed an .xoml file (a zipped up set of workflow files) which it unzips and executes on the computer cluster.

3. METHOD INVESTIGATION

This section discusses each of the methods in turn. A summary is presented in Table 1.

Table 1: Summary of each method.

Methods	Approach	Run Time (Calendar Days) for 100,000 runs	Configuration Difficulty	Resources Required	Comments
Single computer	Batch File (semi-automated)	~139	Easy	Single computer	Total run time does not make this viable.
Multiple computers (5)	Batch File (semi-automated)	~ 28	Easy	5 dedicated computers	Run times are better, but it requires dedicated computing resources.
Partitioned Simulation blocks using ad hoc computing resources	Batch File (semi-automated)	~26 – (Based on 84000 runs)	Reasonably easy	Ad hoc computing resources	Run times are reasonable, but constant monitoring is required.
Multi-core (128 nodes)	The Hydrologists Workbench (fully automated)	~6	Training required	Multi-core computing cluster	Run times are very good. Computer cluster accounts must be configured. Initial workflow construction requires careful planning. Once configured provides a repeatable mechanism for running simulations on a computer cluster

3.1. Method 1: All simulations on a single computer

The first method represents the baseline case. It requires that all simulations be run on a single computer. It is relatively easy to configure as we just need to create a single batch file. We only used this method for development purposes with small numbers of simulations. As simulation numbers increased it became clear that this method was unsuitable for our requirements.

Run time for this method is a significant drawback. Given that a single model simulation takes approximately two minutes and we have 100,000 runs total execution time would be almost 139 days. Not only is this an unacceptable time to wait for results, there is no guarantee that we could run simulations for 139 continuous days. Should the block of simulations terminate prematurely or incomplete simulations be detected, then the modeller must rerun the incomplete simulations increasing the total run time. This method is unsuitable for sensitivity analysis requiring large numbers of simulations of with long model run times.

3.2. Method 2: Partitioned simulation blocks on distinct dedicated computers

This method requires that the set of 100,000 simulations be partitioned into blocks and each block is set running on a different computer. Again this is relatively easy to configure. Suppose we have 5 computers available for use and we have partitioned our simulations into blocks of regular size and created batch files for each block. We start our simulations by copying the batch file, the model executables and files for different blocks to each computer and then the batch file is executed on that computer. When a block run has completed, the outputs are counted to determine if there are incomplete runs and if there are the batch file is rerun (as this will only run incomplete simulations). If all simulations are completed, the batch file for a block of simulations that has not been run is copied to that computer and executed.

As would be expected, run time for this method is much better than for Method 1. Since the performance of the simulation algorithm is linear and we are using 5 machines running different simulation blocks at the same time, the expected total run time of approximately 1/5 of that for Method 1, 28 days. While the total execution time is much better there are still problems with this method. Firstly, we need to acquire 5 dedicated machines that we can use. Secondly we need to continually monitor the machines, so that when a block terminates, we can either rerun the batch file or upload and start a new block of simulations as soon as possible.

3.3. Method 3: Partitioned simulation blocks using ad-hoc computing resources

This is the method we used in our original sensitivity analysis runs as dedicated computing resources were unavailable and workbench tools were incomplete. It is similar to Method 2, but rather than utilizing dedicated computing resources it uses ad-hoc computing such as spare computers on the network, and colleagues who were willing to run simulations on their desktops either overnight or on weekends. Varying simulation block sizes (between 1000 to 10000 simulations per block) were used depending on the length of time the resource was available.

This method is not too difficult to configure, model and batch files need to be copied to the host machine and all outputs are written to a central network location. It does require either computer administrator rights or the cooperation of the computer administrator to execute. As this method uses computing resources on an adhoc basis it requires constant monitoring and a significant amount of coordination. Batch files are created on the fly and constant vigilance is required for an optimal use of available computing resources. As one would expect total run times were greater than that of Method 2 as there is not a continuous dedicated computing environment, but we still managed to complete 84000 simulations in ~26 days, a respectable return for using two dedicated machines, 3 part time machines and some minimal CSIRO network computing.

3.4. Method 4: Multicore execution on a cluster

This method introduces a new approach which uses the Hydrologists Workbench and Trident workflows to harness the computing power of a computing cluster. Two workflows are created, one to run a single simulation (the inner workflow) and another to run the first workflow multiple times (the meta-workflow). The inner workflow contains activities that expose the model inputs, run the Source model and pass on the outputs for a single simulation. The meta-workflow performs the required computing steps for a block of simulations. It has activities which:

- Determine the name and GUID of the inner workflow,

- Gather the required inputs for the simulation block and write them as .xml files (one per simulation) in a cluster accessible location,
- Zip the inner workflow files (into a .xoml file) and copy it to an available cluster location,
- Take the XML file names, the inner workflow GUID, the .xoml file location and the Hydrologists Workbench executable location and create computer cluster tasks (one for each simulation),
- Take in user defined computer cluster task parameters and create a task template which determines how much of the available computer cluster resources each task will use,
- Combine user defined cluster and job parameters with task parameters and tasks to create a cluster job script,
- Submits the script as a job on the cluster.

The cluster then runs the simulations in a distributed computing environment using requested resources.

Executing multiple simulations on a computer cluster using Trident and the Hydrologists Workbench has distinct advantages:

- Running the model files on the cluster is easy as they are embedded in a workflow.
- Model input handling for large numbers of model runs is simplified.
- Changing input parameters (re-running the analysis with different parameter values) simply requires a small modification to the meta-workflow.
- The inner workflow can be modified, used and tested independently of the meta-workflow.
- It provides a repeatable, reproducible and easily modifiable framework for model execution.
- The cluster job submission process is simplified.
- Once configured correctly simulation blocks can be run and rerun with a single mouse click.

The method does require a significant amount of effort to initially configure. First time users will need to ensure that they have appropriate cluster permissions and cached passwords and some training for Trident and the Hydrologists Workbench.

There are distinct time advantages to this approach method. Running on the CSIRO Bragg cluster (CSIRO 2013) which has 128 computer nodes but is often heavily used, yielded about a 6 times speed increase over Methods 2 or 3. Having more resources could see a further 2 to 3 times speed increase. This greatly outweighs the small time overheads associated with composing and running workflows.

4. CONCLUSION

Performing a sensitivity analysis of a river system tests model robustness and correctness, reveals information about the relationships between model input variables and parameters, and model outputs and can be used to quantify model uncertainty. However sensitivity analysis is computationally intensive and individual simulations are prone to failure as parameters are selected programmatically.

We have examined four methods for performing Sobol' sensitivity analysis of an idealized river system model using 100,000 simulations. Method 1 is a traditional approach using a batch file to run the simulations on a single computer. Method 2 extends this to use multiple batch files, distinct subsets of the simulations and parallel computing principles to run the simulations on 5 distinct computers. Method 3 is similar to Method 2 except ad-hoc computing resources are used in place of dedicated machines. Method 4 introduces a new approach, running the simulations in a multi-core environment via a scientific workflow constructed and launched from the Hydrologists Workbench.

From our investigation we found that Method 1 has significant run times and is only suitable for a small number of simulations. Method 2 executes much faster than Method 1 but requires dedicated computing resources and monitoring to ensure optimal run times. Method 3 executes just a little slower than Method 2 but requires constant vigilance to maximize resource usage and reduce total run times. Method 4 is clearly the stand-out approach. The benefits of this approach far outweigh the initial workflow configuration difficulties encountered by the first time user. The Hydrologists Workbench provides a stable, repeatable, reproducible, platform for running simulations. It has provenance capture and purpose built workflow activities to simplify the task of computer cluster job submission. The computer cluster provides a true multi-core environment for parallel simulation running and total run times for the sensitivity analysis are many times faster than the other methods. This is an ideal method for this type of sensitivity analysis.

ACKNOWLEDGEMENTS

This work was funded by CSIRO's Water for a Healthy Country Flagship, through its Risk management framework (RMF) for water allocation planning project.

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.

CSIRO (2013) CSIRO GPU Cluster – Bragg. https://wiki.csiro.au/display/ASC/CSIRO+GPU+Cluster+-+Bragg, CSIRO, 8/6/2013.

Dutta, D., Hughes, J., Vaze, J., Kim, S., Yang, A., and Podger, G. (2012). A daily river system model for the Murray-Darling Basin: development, testing and implementation. *34th Hydrology and Water Resources Symposium 2012, 19-22 November 2012*, Sydney, NSW: 1057-1066.

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*.

Harbaugh, A.W., Banta, E.R., Hill, M.C., and McDonald, M.G. (2000). MODFLOW-2000, the U.S. Geological Survey modular ground-water model - User guide to modularization concepts and the Ground-Water Flow Process. *Open-File Report 00-92*. U.S. Geological Survey.

Murray-Darling Basin Authority. (2012). Basin Plan. *http://www.mdba.gov.au/sites/default/files/Basin-Plan/Basin-Plan-Nov2012.pdf*, Murray-Darling Basin Authority, 21/11/2012.

NWI, 2004. Intergovernmental agreement on a National Water Initiative. Council of Australian Governments http://www.environment.gov.au/water/australia/nwi/index.html

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.

Peeters L., Podger G., Smith T. Pickett T. Gao L, Bark R., Qureshi E, & Cuddy S. (2013) Local Model Emulation for Markov Chain Monte Carlo Simulation of a River Management Model. Preprint.

Sobol, I.M. (1976) "Uniformly distributed sequences with an additional uniform property". Zh. Vych. Mat. Mat. Fiz. 16: 1332–1337 (in Russian); U.S.S.R Comput. Maths. Math. Phys. 16: 236–242 (in English).

Welsh, W.D., Vaze, J., Dutta, D., Rassam, D., Rahman, J.M., Jolly, I.D., Wallbrink, P., Podger, G.M., Bethune, M., Hardy, M.J., Teng, J., Lerat, J. (2013). An integrated modelling framework for regulated river systems. Environ. Modell. Softw 39(2013), 81-102.