Scientific workflows for parameter-setting support in environmental modelling

N. Radosevic*[1], M. Duckham[2], and G.J. Liu[3]

[1, 2, 3]RMIT University, Australia

*Email: nenad.radosevic@rmit.edu.au

Abstract

Parameter uncertainty, restricted information and insufficient transparency of many environmental models are often limiting factors in their use and application. The aim of this work is to demonstrate how scientific workflows can be designed and implemented to address these limitations in use and application of environmental models. Specifically, this work shows an application of scientific workflows and their potential in delivering a transparent, reproducible, integrated and automated tool for a parameter setting decision support. The results show how parameter uncertainty can be captured and managed using a freely-available and open-source scientific workflow management system (SWFMS). Specifically, scientific workflows integrate environmental modelling and a machine learning technique to discover and select appropriate parameter settings. This work shows a case study with the particular example of environmental modelling for estimating solar radiation potential of building rooftops. In general, an appropriate user support and management of parameter uncertainty help to broader usage and warrantability of environmental models and their outputs, both in research related work and in interdisciplinary analysis for decision making.

Keywords: scientific workflows, environmental modelling, decision support, parameters.

1 Introduction

In general, a problem related with setting parameter values can be related to many limiting factors in environmental modelling. For example, restricted information and lack of transparency ("black-box") of environmental models may affect their use and practical functionality. A limited knowledge and insufficient transparency of models together with their complexity are some of restricting factors which may enhance error propagation in the modelling (Ascough et al., 2008; Refsgaard et al., 2007).

Another restricting and relevant factor is parameter uncertainty in the modelling. Parameter uncertainty refers to uncertainty related to imprecision and inaccuracy of determining and selecting parameter setting values Walker et al. (2003). In general, parameter settings for environmental models can be derived from related literature, real observations or based on calibration or estimation methods (Gallagher and Doherty, 2007; Maier et al., 2008). However, for novice users selecting

*

appropriate parameter settings can be a challenging task. A study by Refsgaard et al. (2007) showed that user's experience played a significant role in the final results of an environmental modelling. For instance, lack of domain knowledge related to model inputs imposed difficulties for novice users to choose a correct decision in the modelling.

One of the ways to address this problem is to design and implement a transparent and reproducible decision support tool to help inexperienced users in use and application of environmental models. This study explores usefulness and practical applicability of scientific workflows in providing an assistance for selecting appropriate parameter settings. Specifically, this work demonstrates how transparent, reproducible, integrated and automated method such as the scientific workflows can be used to provide users support and to widen usage and warrantability of environmental models.

The remainder of the paper is organized as follows. To address this problem, section 2 provides introduction to scientific workflows and scientific workflow management system (SWFMS). Following section 3 looks at background of the specific environmental model for predicting solar radiation potential (the Solar Analyst). Section 4 demonstrates a case study with an example of scientific workflows designed to integrate environmental modelling of solar radiation with decision tree learning to find appropriate parameter settings. The last section 5, concludes the paper.

2 Scientific workflows

Scientific workflows are a procedure or method to execute complex scientific analysis which predominantly has a goal to capture and document all steps involved in preparing, processing and presenting scientific data (Ludscher and Goble, 2005). According to Gil et al. (2007); Granell et al. (2013), and Kitzes et al. (2017) scientific workflows improve reproducibility, transparency, documenting, flexibility, management, integration, automation, and computing of complex scientific studies. Improvement in reproducibility, transparency and documenting is illustrated through an open access, detailed description and self-documenting of scientific workflows. Different computational assets such as multiple scientific data sets, models, parameters, scripts and programming tools can be easily integrated with a flexibility to automate every step in scientific workflows. These characteristics can improve understanding of environmental models and provide a decision support to prevent errors in the outputs they generate.

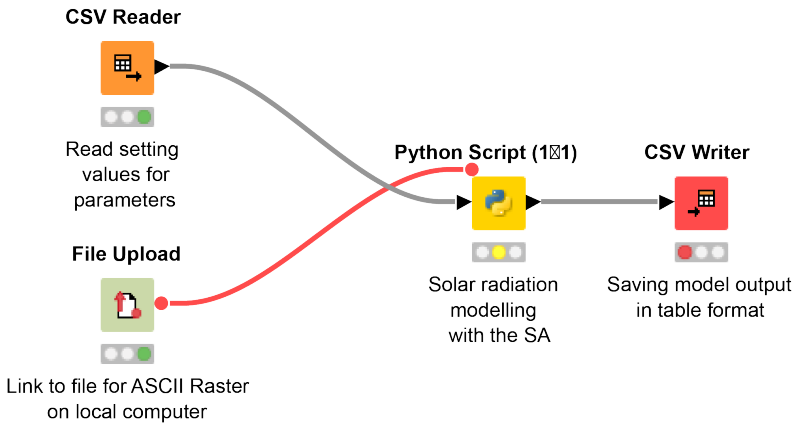

Figure 1: A simple KNIME scientific workflows for environmental modelling of solar radiation

The design and implementation of scientific workflows is supported in freely-available and open source SWFMS such as Kepler, VisTrails and KNIME which provide a user-friendly graphic interface for easier design, implementation, execution and documenting. Figure 1 illustrates a simple example of scientific workflows designed in KNIME (Bakos, 2013; Berthold et al., 2009). Each node in this scientific workflows presents a task or a data process. A state for each node is illustrated with the traffic lights whether the task has been processed (green), is ready to be processed (yellow) or has not been processed yet (red).

Application of scientific workflows has a potential in many interdisciplinary scientific studies. For example, a geospatially enabled scientific workflows can be designed for an observation and modelling different physical systems. In their review, Granell et al. (2013) stated that scientific workflows provides a unique set of tools with elements of a component-based support for environmental modelling. Zyl et al. (2012) designed scientific workflows as a system support for wildfire spatiotemporal analysis to examine improvements in accessing, implementing and integrating geospatial data into the scientific workflows. Another study conducted by Kaster et al. (2005) involved application of scientific workflows for agricultural environmental modelling. The system support provides an open and transparent tool for documenting, monitoring and assessing impacts of fertilizing process on environment (soil and water). In the next section we introduce one of the most commonly used solar radiation models, the Solar Analyst (SA).

3 Solar Analyst

Environmental modelling of solar radiation is important, both in many scientific disciplines (engineering, geospatial, and environmental) and for industry related projects (renewable energy and agriculture). For example, modelling solar radiation potential in cities plays a significant role in urban planning and development of sustainable and livable cities. Thus, many studies related to modelling solar radiation emerged in the past (Kodysh et al., 2013; Li et al., 2015; Redweik et al., 2013; Santos et al., 2014). This model uses an upward-looking viewshed algorithm to calculate solar radiation potential for any given locations by including sky obstruction from surrounding objects (Rich et al., 1994; Fu and Rich, 1999). Figure 2 illustrates an upward looking hemispherical viewshed, and a yellow line defines a boarder between visible and obstructed sky for a given location. The area of interest is defined by a Digital Surface Model (DSM) as the main spatial data input. Consequently, the SA runs a viewshed analysis and estimates solar radiation potential in each raster cell of the DSM.

Figure 2: Upward looking hemispherical viewshed (Fu and Rich, 2000)

Three main groups of parameters are used in this model:

1. spatial;
2. temporal; and
3. radiation.

Spatial parameters such as latitude, topography of DSM, sky size and number of azimuth directions for calculating viewhsed relate to a location and spatial context of an area of interest. Temporal parameters (hours, days, months and year) define a time period of solar radiation exposure. Radiation parameters such as diffuse and transmission proportion determine an intensity of incoming solar radiation, and azimuth and zenith directions define granularity of distribution for a diffuse component of incoming solar radiation. These are some of the most significant parameters of this model, and other parameters can be found in work of Fu and Rich (2000, 1999).

The popularity of this model among modellers with various level of expertise including inexperienced modellers raises questions about user support, parameter uncertainty and warrantability of outputs that the SA generates. The next section demonstrates an example of scientific workflows designed to support parameter-setting for modelling rooftop solar radiation potential with the SA.

4 Design and implementation of scientific workflows

Figure 4 depicts an example of the scientific workflows for parameter-setting support. The design of this scientific workflows can be divided into two major parts:

1. modelling of solar radiation using different parameter settings; and
2. decision tree learning for appropriate parameter setting.

The following subsections provide deeper understanding about the design and implementation of this scientific workflows.

4.1 Modelling of solar radiation using multiple parameter settings

This part of the scientific workflows is designed to estimate solar radiation with different parameter settings. Specifically, a user have freedom and flexibility to use multiple parameters with different values and to investigate their applicability for modelling. The scientific workflows uses three main inputs to run the SA:

1. a CSV file with different setting values for parameters;
2. a DSM in raster format for an area of interest (Figure 3); and
3. a geographical latitude of an area of interest.

Table 1 shows the most important and difficult parameters for setting their values. A brute-force technique to explore appropriate parameter setting with all $4^6 = 4096$ parameter setting combinations is used to generate parameter inputs for the model. Parameter setting combinations are iteratively imported, one at a time into a workflow loop (Loop start, Python Script and Loop End, Figure 4). This workflow loop automates modelling of solar radiation and generates a data

set for each parameter setting combination. Python Script as a central node of the workflow loop runs the SA for each iteration.

Parameter	Setting 1	Setting 2	Setting 3	Setting 4
Sky size	250	350	450	550
No. calculation directions	8	16	24	32
No. azimuth divisions	8	16	24	32
No. zenith divisions	8	16	24	32
Diffusion proportion	0.1	0.2	0.3	0.4
Transmission proportion	0.4	0.5	0.6	0.7

Table 1: Parameters and their setting values

Generated solar radiation data sets from this workflow loop are aggregated together with all parameter setting combinations into a new data set. This data set is classified with Numeric Binner node into three different categories ("Too Low", "Expected" and "Too High") based on difference from a ground truth data observed by meteorological stations in local area."Too Low" classifies data bellow the "Expected" range, while "Too High" is considered above the "Expected" range. The observed ground truth data is captured with no obstruction from surrounding objects and because of that it is equal to the maximum value of the "Expected" range.

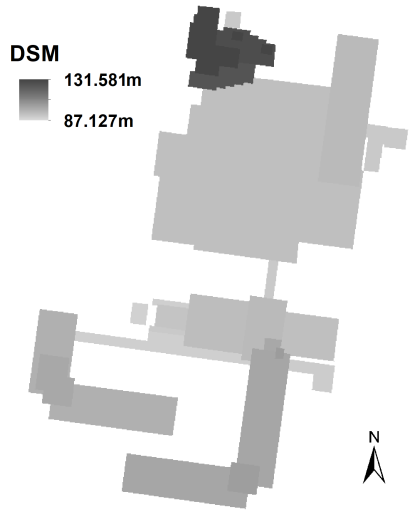

Figure 3: DSM of building rooftops

4.2 Decision tree learning for finding appropriate parameter setting

The first part of this scientific workflows successfully integrates multiple computational assets such as raster and text data, and R and Python scripts. The second part introduces a decision tree model which demonstrates workflows' ability to incorporate multiple models. Decision tree learning is a type of machine learning which iteratively categorizes a data set based on its most important features (Russell and Norvig, 2002). In this scientific workflows the decision tree model uses two algorithms ID3 and C4.5 (Quinlan, 1986, 1993). Both algorithms classify data set by indicating at each tree node the attributes that have the highest Shannon's information gain (Shannon and Weaver, 1949).

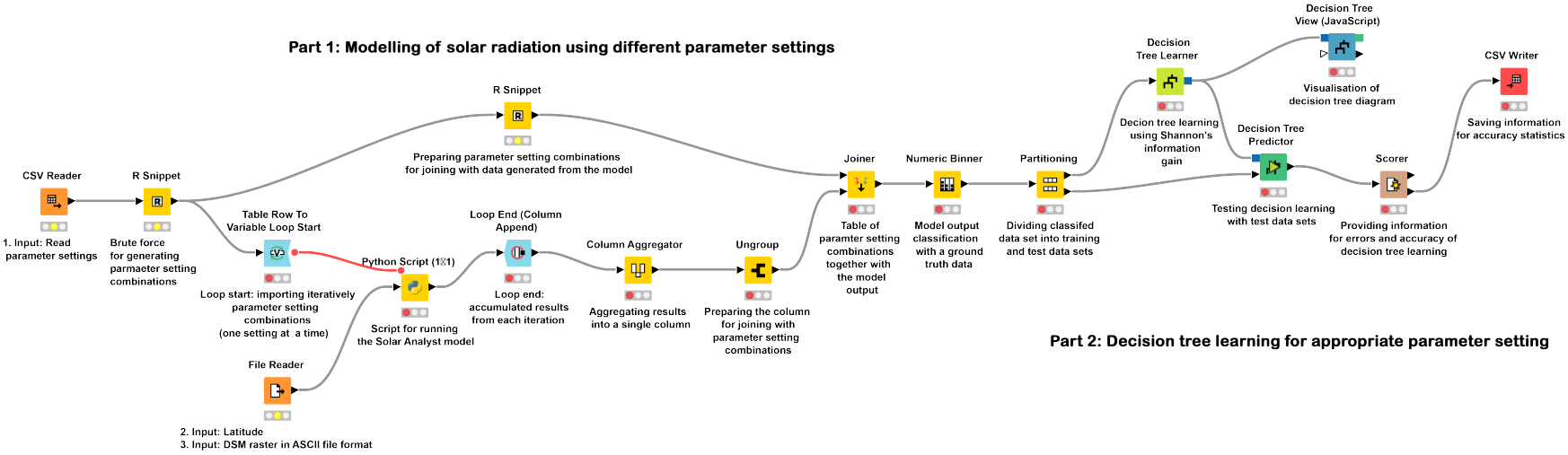

Figure 4: Scientific workflows for parameter-setting support

In this case, decision tree learning can be applied to classify different parameter settings based on the information gained from the output that the model generates. Specifically, decision tree learning classifies parameter settings based on information gained from three categories of the SA output ("Too Low", "Expected" and "Too High"). Decision tree learning classifies and detects appropriate parameter settings based on category "Expected", while inappropriate parameter settings are selected by other two categories "Too Low" and "Too High". The validation of the decision tree learning is tested by Decision Tree Predictor node and calculated by a Scorer node depicted in Figure 4.

4.2.1 Decision tree diagram

Figure 5 illustrates a decision tree diagram and each internal node demonstrates a test based on a parameter setting. Each branch represents the outcome of this test. A yellow node presents the end of each branch showing one of three categories of the output from the model. Tree branches ending with "Radiation: Expected" indicate appropriate settings for parameters. Along these branches novice users can select appropriate settings for model parameters.

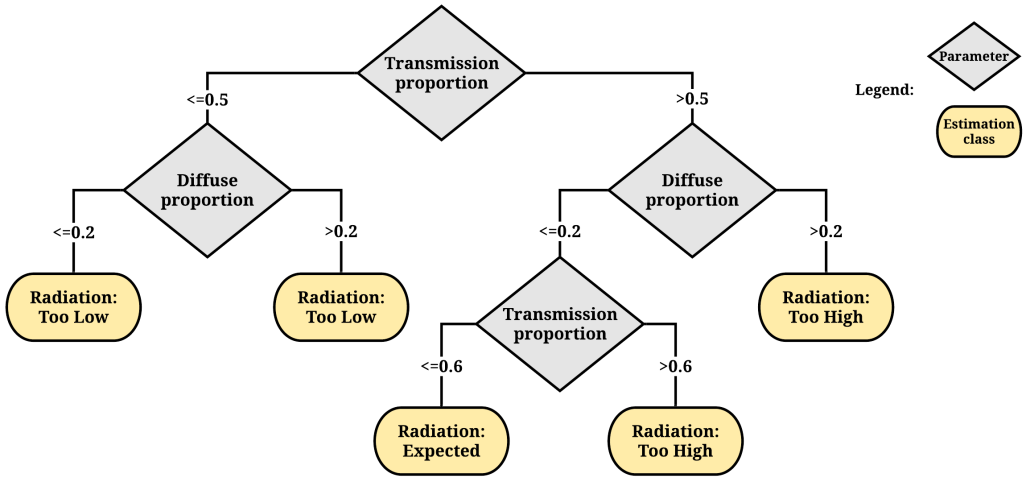

Figure 5: Decision tree diagram for parameter-setting support

For instance, Figure 5 illustrates transmission proportion as a root node greater than 0.5 (right-hand branch) versus transmission proportion of 0.5 or less (left-hand branch) as being the parameter setting related with the highest information gain in the training data set. A general rule is that parameter settings placed closer to the root node have stronger influence on the model output. This decision tree diagram includes two out of six tested parameters. This decision tree also indicates that settings for parameters such as sky size, number of calculation directions, number of azimuth and zenith divisions are associated with the lowest information gain in the training data set. Thus, decision tree learning excludes them from the decision tree diagram. In other words, from six selected parameters, the most significant are transmission and diffuse proportion. This is a useful information for novice users and their decision making process of selecting appropriate parameter settings.

4.3 Discussion

The application of scientific workflows in this work demonstrates an example of transparent, reproducible, self-documented, automated, integrated and flexible support tool for parameter setting. Transparency, reproducibility and self-documenting assist inexperienced modellers for deeper understanding of environmental models and modelling process. In this example description and open access for each task (node) improves understanding about each step for delivering parameter setting support. The automation of modelling with the workflow loop improves time efficiency and reduces error for generating scientific data with environmental models. Integration of multiple computational assets such as different data sets, programming languages and models provides multi-functionality of scientific workflows transforming them into efficient and practical tool. The decision tree learning provides a graphical support for user's decisions to correctly select parameter setting (Figure 5). Decision tree learning have demonstrated accuracy of 91.5% for predicting three different categories of the model output. Ultimately, the usefulness and practical applicability of this scientific workflows goes across multiple properties from transparency, documenting and integration of multiple tasks and different computational assets to an provider of assistance for user's decisions in setting parameters of the model.

5 Conclusion and further work

Limiting factors such as lack of transparency ("black-box" models), limited information and parameter uncertainty can affect use and application of environmental models. This may be specifically the case with novice and inexperienced users. This study shows a design and implementation of the scientific workflows as a decision support tool for parameter setting in environmental modelling. Specifically, user's decisions are supported by a decision tree diagram providing an interactive and transparent guidance for setting parameters. In addition to that, addressing this problem using scientific workflows helps users to widen understanding of environmental models. On the other hand, a potential disadvantage can be a complexity of scientific workflows design because they may involve multiple data flow streams with complex structure.

Our future work may include design and development of different scientific workflows which are able to provide improvement in reproducibility of models and modelling process. For example, a KNIME scientific workflows to improve transparency of black-box models by designing and implementing a modular version of the same model.

6 Acknowledgements

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

7 References

Ascough, J. C., H. R. Maier, J. K. Ravalico, and M. W. Strudley
2008. Future research challenges for incorporation of uncertainty in environmental and ecological decision-making. *Ecological Modelling*, 219(3):383–399.

Bakos, G.
2013. *KNIME Essentials*. Packt Publishing.

Berthold, M. R., N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, K. Thiel, and B. Wiswedel
2009. Knime - the konstanz information miner: Version 2.0 and beyond. *SIGKDD Explor. Newsl.*, 11(1):26–31.

Fu, P. and P. M. Rich
1999. Design and implementation of the solar analyst: an arcview extension for modeling solar radiation at landscape scales. In *Proceedings of the Nineteenth Annual ESRI User Conference*.

Fu, P. and P. M. Rich
2000. The solar analyst 1.0 user manual. Report, Helios Environmental Modeling Institute.

Gallagher, M. and J. Doherty
2007. Parameter estimation and uncertainty analysis for a watershed model. *Environmental Modelling & Software*, 22(7):1000–1020.

Gil, Y., E. Deelman, M. Ellisman, T. Fahringer, G. Fox, D. Gannon, C. Goble, M. Livny, L. Moreau, and J. Myers
2007. Examining the challenges of scientific workflows. *Computer*, 40(12):24–32.

Granell, C., S. Schade, and N. Ostlnder
2013. Seeing the forest through the trees: A review of integrated environmental modelling tools. *Computers, Environment and Urban Systems*, 41:136–150.

Kaster, D. S., C. B. Medeiros, and H. V. Rocha
2005. Supporting modeling and problem solving from precedent experiences: The role of workflows and case-based reasoning. *Environmental Modelling & Software*, 20(6):689–704.

Kitzes, J., D. Turek, and F. Deniz
2017. The practice of reproducible research.

Kodysh, J. B., O. A. Omitaomu, B. L. Bhaduri, and B. S. Neish
2013. Methodology for estimating solar potential on multiple building rooftops for photovoltaic systems. *Sustainable Cities and Society*, 8:31–41.

Li, Z., Z. Zhang, and K. Davey
2015. Estimating geographical pv potential using lidar data for buildings in downtown san francisco. *Transactions in GIS*, 19(6):930–963.

Ludscher, B. and C. Goble
2005. Guest editors' introduction to the special section on scientific workflows. *SIGMOD Record*, 34(3):3–4. Cited By :33 Export year: 18 July 2017.

Maier, H., J. A. II, M. Wattenbach, C. Renschler, W. Labiosa, and J. Ravalico
2008. Chapter five uncertainty in environmental decision making: Issues, challenges and future directions. In *Environmental Modelling, Software and Decision Support*, A. Jakeman, A. Voinov, A. Rizzoli, and S. Chen, eds., volume 3 of *Developments in Integrated Environmental Assessment*, Pp. 69 – 85. Elsevier.

Quinlan, J.
1993. *C4.5: Programs for machine learning.* San Francisco, CA: Morgan Kaufmann.

Quinlan, J. R.
1986. Induction of decision trees. *Machine Learning*, 1(1):81–106.

Redweik, P., C. Catita, and M. Brito
2013. Solar energy potential on roofs and facades in an urban landscape. *Solar Energy*, 97:332–341.

Refsgaard, J. C., J. P. van der Sluijs, A. L. Højberg, and P. A. Vanrolleghem
2007. Uncertainty in the environmental modelling process - a framework and guidance. *Environmental Modelling & Software*, 22(11):1543–1556.

Rich, P. M., R. Dubayah, W. A. Hetrick, and S. C. Saving
1994. Using Viewshed models to calculate intercepted solar radiation: applications in ecology. *American Society for Photogrammetry and Remote Sensing Technical Papers*, Pp. 524–529.

Russell, S. J. and P. Norvig
2002. *Artificial Intelligence: A Modern Approach*, 2nd edition. Upper Saddle River, NJ: Prentice-Hall.

Santos, T., N. Gomes, S. Freire, M. C. Brito, L. Santos, and J. A. Tenedrio
2014. Applications of solar mapping in the urban environment. *Applied Geography*, 51:48–57.

Shannon, C. and W. Weaver
1949. *Mathematical Theory of Communication.* Urbana, IL: University of Illinois Press. Republished 1963.

Walker, W., P. Harremos, J. Rotmans, J. van der Sluijs, M. van Asselt, P. Janssen, and M. K. von Krauss
2003. Defining uncertainty: A conceptual basis for uncertainty management in model-based decision support. *Integrated Assessment*, 4(1):5–17.

Zyl, T. L. v., A. Vahed, G. McFerren, and D. Hohls
2012. Earth observation scientific workflows in a distributed computing environment. *Transactions in GIS*, 16(2):233–248.