Common Sense In Model Testing

[1]Huth, N. and [1]D. Holzworth

[1]CSIRO Sustainable Ecosystems, E-Mail: Neil.Huth@csiro.au

Keywords: *Modelling; APSIM; Validation; Verification; Testing*

EXTENDED ABSTRACT

Before any model can be used with confidence it must be tested to assess its fitness for the given task. Discussions of model goodness tend to often revolve around the various definitions of several core concepts such as "validation" or "verification". For the purpose of this paper, we will describe model evaluation as consisting of two main areas: usefulness and reliability. For a model to be useful, it must reflect the behaviour of the real world being simulated with an acceptable level of accuracy. A model is said to be reliable if the implementation of the calculations involved reproduce the conceptual model of the system to be simulated. In this paper, we will describe the approach to model evaluation employed by the APSIM model development team and explain how this process is used to address the various issues raised regarding model evaluation and testing.

APSIM is a complex systems' model and thus requires a large effort for testing. A testing process has been developed consisting of four basic types of tests. Unit and system tests are targeted at model reliability and stability. The validation and sensibility tests are designed as an attempt to address the issue of model usefulness.

Unit tests provide tests of the reliability of the execution of a particular implementation in capturing the desired behaviour of a specified model. These are targeted at the smallest units of model code. System tests are constructed in much the same way as unit tests but evaluate the reliability of larger collections of code, or in fact entire simulations. Where possible, the results of these complex simulations are compared to desired outcomes. Where this is not possible, results are compared against previous results in order to identify changes in model behaviour. Validation tests incorporate the traditional elements of model testing. The results of simulations of experimental conditions are compared against the recorded observations. When model functionality is changed formal statistics may be employed to determine model usefulness. Minor changes to model code are more simply evaluated by once again comparing against previously acceptable results.

These well-known testing methods, whilst effective for some things, are not necessarily effective at ensuring that a model faithfully captures the underlying process driving the system in question and thus will be applicable in any new situation where a model may be employed. We have attempted to address this deficiency via the use of sensibility tests. In these tests, model responses to various stimuli are evaluated against the regularly observed system responses that are often captured in the notion of "common sense". The scientific literature is full of instances where the behaviour of complex systems is reduced to a set of basic emergent behaviours. Dynamic models, when not developed from these relationships, can be quickly evaluated against such relationships.

Three case studies are demonstrated in which key emergent properties such as productivity, product quality, resource use efficiency and hydrological balance of wheat cropping systems simulations are all evaluated against documented relationships. In each case the usefulness of the model in describing these key processes is easily identified via simple graphical comparison.

The authors want to be very clear that we are not suggesting sensibility tests as a replacement for the more established formal methods. The process described here makes great use of traditional model validation and software quality acceptance tests. However, as the usage of modeling systems continues to grow, the requirement for a greater breadth of testing will necessitate the use of more extensive yet efficient testing procedures. We have found sensibility tests useful in evaluating the ability of models to capture the underlying processes in real-world situations.

1. INTRODUCTION

The topic of model evaluation has long attracted significant debate amongst members of the scientific community. In many ways, there is as much debate over the meaning of terms such as "testing", "validation", "verification" and "calibration" as there is on the actual process of evaluation itself. Prisley and Mortimer (2004) collated a large collection of definitions for various terms and highlighted the ways in which they have been defined in the literature. In this paper, we will describe the approach to model evaluation employed by the APSIM model development team and explain how this process is used to address the various issues raised regarding model evaluation and testing.

For the purpose of this paper, we will describe model evaluation as consisting of two main areas: usefulness and reliability. For a model to be useful, it must reflect the behaviour of the real world being simulated with an acceptable level of accuracy. This behaviour may be taken to consist not only in reproducing observable system states, but the processes that lead to them. A model is said to be reliable if the implementation of the calculations involved reproduce the conceptual model of the system to be simulated. Whilst such a simple division of all the issues pertaining to evaluation can have its problems, it can allow model developers to communicate adequately enough in order to leave the semantic debate behind. It is immediately worth noting that both terms are context dependant. A model may be useful in one context and not in another. The tests employed, therefore, will also have to reflect the differing contexts of model application.

APSIM (Keating et al, 2003) is a component-based modeling framework developed for the simulation of agricultural and natural systems. It emerged out of the work of various scientific research teams and its early focus was on providing a simulation capability for scientists to use in distinct agricultural situations. In recent years, however, the model has been increasingly used for informing policy development and even for real-time information support for land-managers (Hochman et al, 2005). This change in model application has led to a change in the focus in model testing.

It is easy to see how a growing user-base for a model, including users making real-time decisions, can place a greater importance on the need for testing for model reliability. Model results need to execute when required, and according to specification, for a wider array of model executions. Modern software techniques have evolved to address this problem and can be readily employed for the model in question. It is not so obvious, however, just how one maintains that a model is, and remains, useful to this growing suite of model users. Rykiel (1996) suggests that the most common problem with the validation of ecological models is that of identifying what constitutes the evaluation criteria. Similarly, there is a need for model developers to understand the things that would make a model useful to a model user.

The APSIM model development team has attempted to address these issues in a formal model testing methodology that explicitly addresses both reliability and usefulness. Whilst many of the elements of this evaluation procedure are common-place, this paper will highlight the use of "sensibility tests" in helping to evaluate model usefulness.

2. THE APSIM TESTING METHODOLOGY

APSIM is a complex systems' model and thus requires a large effort for testing. This is even more important given the ongoing development of the model. Model evaluation is never complete. For this reason, an automated testing process has been employed to detect changes in model reliability or usefulness whether intended or not (the "ripple effect"). This closely follows the software development approach of Jeffries (2001) that developers "test everything that could possibly break, using automated tests that must run perfectly all the time".

Each day, or more frequently during intensive model development, the entire code-base is constructed from a version control system and compiled to create an instance of the model. This executable model is then tested against a large suite of tests. Where possible, the results are automatically compared against a known correct answer for that test. Where a complex system is being simulated, and the result cannot be determined *a priori* without the model, the results are compared against the previous execution of the model in order to detect possible undesirable changes in behaviour.

This suite of tests consists of four basic types of test. Unit and system tests are targeted at model reliability and stability. The validation and sensibility tests are designed as an attempt to address the issue of model usefulness.

2.1. Unit Tests

The APSIM Development team follows an agile software process where the emphasis is on the running code rather than lots of documentation. One of the tenets of agile software processes is the unit test. This is the lowest form of test where individual functions and methods are tested for correctness. Each function is called many times with different inputs, with the outputs carefully compared against known correct results. In fact, the "Extreme Programming" approach (Jeffries 2001), dictates that these tests should be created before any lines of code are written.

Whilst this is difficult to achieve on a large existing code base, and in complex models like APSIM, some unit testing is better than none. When combined with the higher level tests outlined below, they provide a degree of robustness that is important in the overall context.

2.2. System Tests

System tests aggregate the functions and methods into higher level units and test these. For APSIM, these modules are at the level of whole crop modules or water balance modules. There are a suite of tests that exercise these modules, driving the crop modules with very low and very high amounts of water and nitrogen. They are designed to try and break the module in question, to see how the module performs in extreme situations.

In a sense they are very much like unit tests but operate at a much higher level.

2.3. Validation Tests

Validation tests are traditionally used extensively among agricultural modelers to gauge how well a model performs against observed field and laboratory measurements. The APSIM testing suite has many validation results for many crop modules that are compared automatically. Charts of predicted and observed data are generated and tools have been developed to compare these plots against known good plots. However, like all the tests in the suite, it is the automated nature of the testing that adds value. It is the change in behaviour that is being detected by the automated process.

2.4. Sensibility Tests

Like validation tests, sensibility tests evaluate model usefulness. The objective is to see how the model performs in different real world scenarios. Sensibility tests differ from validation tests in that the comparison of model output is made against more subjective, local experts feeling for what the model should do. Examples from local agronomists include statements like "under those conditions the model should have a median yield of x tonnes/ha with a range of y to z.". These tests are a way of making sure the model performs in situations where validation / observed data are not available. In a sense it is about ensuring the model performs adequately outside the conditions in which it was built.

At the same time, however, their use in testing the correctness of the underlying model design should not be underestimated. Model development can be misleading if modellers forget that the "hard" data available for model testing is only a small snapshot of system behaviour. Seibert and McDonnell (2002) showed that by incorporating "soft" data into model development one might obtain a better process representation of a catchment's hydrology. This might safeguard against being "right for the wrong reasons" due to fitting a model to a small set of observations. One might become "less right" but for the right reasons. Similarly, Pastres et al (2004) found that whilst traditional testing of their seagrass model gave an adequate description of the available model calibration data, the model failed to capture the known trends in seagrass evolution over a longer time frame. These authors also suggest simple ways of testing that model results are reasonable, even in the absence of "hard" data.

3. SENSIBLITY TEST CASE STUDIES

The APSIM Wheat module simulates the growth, development and resource use of wheat crops. Its current design builds upon the wheat model of Wang et al (2003) but is implemented in a common crop modeling framework (Robertson et al 2002, Wang et al 2002) The following sensibility tests have been used with the APSIM wheat model to test its usefulness for model users and its fidelity to the real world it attempts to describe.

3.1. Grain Yield and Quality in response to water and nitrogen supply

Wheat production in the Northern grain-growing regions is strongly linked to the supply of water and nitrogen. A useful model of these systems must be able to capture the effect of both of these resources on not only grain yield but also the quality of that grain, usually described by its nitrogen or protein content. Increased nitrogen supply should not only increase production but

also increase grain nitrogen content. A similar story should exist for water supply. However, a tradeoff does exist for farmers in drought affected years. Lowering of wheat yields should result in an increase in grain protein levels thus increasing the quality of the product and the price received for it. For a model to be useful to a land manager in such systems, such complex interactions must be captured by the model. The reproduction of such emergent properties of the production system is a good sensibility test for the model.

Dalal et al (1996) describe a long term trial at Warra (26°47'S, 150°53'E) in Queensland. investigating the effects of various management factors on wheat production. These included a range of nitrogen application rates for a range of wet and dry seasons. The trial data suggest that the yield potential for high nitrogen supply is 4000-5000 kg/ha whereas under nitrogen limitation yields would rarely exceed 2000 kg/ha. However, when drought conditions prevail, water limitation can reduce yield to approximately 1000 kg/ha irrespective of nitrogen supply. Reduction in yield due to water shortage will however increase grain nitrogen concentrations from 1.5-2.0% to 2.5% under low nitrogen supplies and to just over 3.0% when nitrogen supply is high.

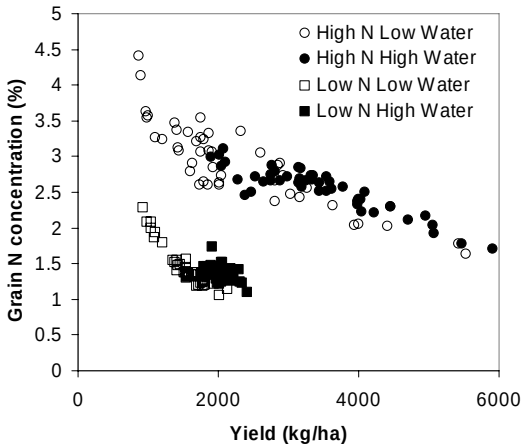

Figure 1. The emergent relationship between wheat grain nitrogen content and yield of the APSIM simulations for a range of soil conditions at Warra. Qld.

Figure one shows the results of a simulation configured to reproduce the conditions of this particular trial. Simulations for 1957 to 2002 were initialized at sowing each year to capture a matrix of high and low levels of both soil water and nitrogen. One can see from the simulation results that the model is able to capture the important emergent behaviour of the production system in terms of both grain yield and quality in response to various interacting resource supplies. That being so, the model developer can gain confidence that the model would be useful to a model user for issues involving these interactions. On the other hand, the grain nitrogen contents under extreme drought conditions might be said to appear less sensible. Grain nitrogen can appear to reach values higher than expected under drought conditions. This might bring the model utility into question for such conditions. The benefit of this sensibility test is clearly demonstrated in its ability to highlight model behaviour in extreme circumstances that might not appear in the formal model validation datasets, but that certainly exist in the real world in which the model will be applied.

3.2. Nutritional Constraints on Water Use Efficiency

The link between crop production and rainfall has long been understood. In many regions of Australia where water is limiting strong relationships exist between water supply (rainfall + stored soil moisture) and production (French and Schultz, 1984). This is often described using the term "water use efficiency". The efficiency with which a crop can make use of available moisture is sometimes regulated by nutrition. Anderson (1992) for example, showed trends in the relationship between water supply and production for conditions of high and low nitrogen supply. Once again, if these basic emergent trends can be also seen in a simulation of such systems, a model developer will gain confidence in the model's ability.

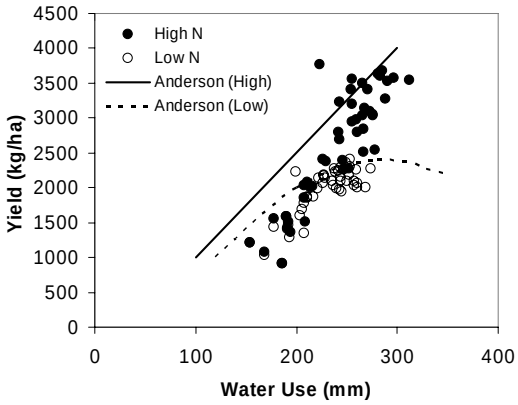

Figure 2. Comparison of the emergent relationship of wheat yield vs water use with or without nutrient constraint simulated using APSIM against that described by Anderson (1992).

Figure two illustrates the basic emergent relationships from Anderson (1992) for various levels of water and nitrogen supply across a range of sites in South Australia. Long term simulations for Roseworthy Agricultural College (-34°30'S,

138°41'E) under low and high nitrogen supply show that the model appears to behave sensibly under a range of growing conditions. Whilst a direct comparison is difficult unless one was to simulate the exact same fields used in the original, a simpler test for emergent relationships can be done in order to gain confidence that the basic processes behind the real system are being captured in the model. It is worth noting that Farré et al (2004) performed a similar test to evaluate the response to water supply of their model of lupin growth.

3.3. Hydrologic Balance of the Cropping System

The salinisation of landscapes is among the greatest challenges to the dryland farming systems of Australia (NLWRA, 2000). Much of the problem is brought about by the change in the hydrological balance of our landscapes as a result of the introduction of agricultural crops. Numerous studies have been held to quantify catchment response to rainfall for various vegetation types. Zhang et al (2001) were able to summarise a great deal of the information gathered using a simple model incorporating the main factors of rainfall, evaporative potential and soil water holding capacity.

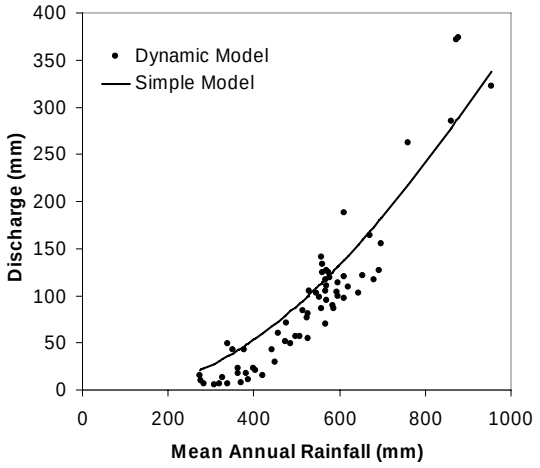

Figure 3. Comparison of the emergent relationship of catchment discharge vs annual rainfall for wheat cropping systems across a range of eastern Australian rainfall sites for APSIM (results from Keating et al, 2002) and a simple model (Zhang et al, 2001).

When Keating et al (2002) investigated the hydrological characteristics of various cropping systems they chose to test the sensibility of the simulated results by comparing them to the emergent behaviour of the many datasets captured in the simple model. The fact that the dynamic model followed the general trends provided confidence in the model in a much simpler fashion than comparison to each of the many datasets on their own. Once this confidence had been established the authors were able to utilize the power of the dynamic systems model to investigate various management interventions.

4. CONCLUSIONS

The authors want to be very clear that we are not suggesting sensibility tests as a replacement for the more established formal methods. These other testing procedures have been described above to show that sensibility tests are a useful compliment to these other methods. The process described here makes great use of traditional model validation and software quality acceptance tests. However, as the usage of modeling systems continues to grow, the requirement for a greater breadth of testing will necessitate the use of more extensive yet efficient testing procedures. We have found sensibility testing valuable in testing for the ability of models to capture the underlying processes in real-world situations, not just the observable phenomena in scientific experiments. They can provide a greater breadth to a model test suite at very little cost and when looking to evaluate a model in terms of its usefulness, prove valuable by focusing on the common sense emergent relationships regularly used by model users.

5. ACKNOWLEDGMENTS

The authors would like to acknowledge the other members of the APSIM software development team over the last few years that have assisted in the development of the current process.

6. REFERENCES

Anderson, W.K. (1992), Increasing Grain Yield and Water Use of Wheat in a Rainfed Mediterranean Type Environment. *Australian Journal of Agricultural Research* 43, 1-17.

Dalal, R.C., W.M. Strong, E.J. Weston, J.E. Cooper, K.J. Lehane, A.J. King, and C.J. Chicken (1996), Sustaining productivity of a Vertisol at Warra, Queensland, with fertilizers, no-till, or legumes. I. Organic matter status. Australian *Journal of Experimental Agriculture* 35, 903-913.

Farré, I., M.J. Robertson, S. Asseng, R.J. French, and M. Dracup (2004), Simulating lupin development, growth, and yield in a

Mediterranean environment. *Australian Journal of Agricultural Research*. 55,863-877.

French, R.J. and J.E. Schultz. (1984), Water Use Efficiency of Wheat in a Mediterranean-type Environment. I The Relation between Yield, Water Use and Climate. *Australian Journal of Agricultural Research* 35, 743-764.

Hochman, Z., H. van Rees, P.S. Carberry, D. Holzworth, N.P. Dalgliesh, J. Hunt, P.L. Poulton, L.E. Brennan, T. Darbas, J. Fisher, S. van Rees, N.I. Huth, A. Peak, and R.L. McCown (2005), Can access to a cropping system simulator help farmers reduce risk in drought prone environments? Proceedings of the InterDrought-II Congress, Rome, Italy.

Jeffries, R., A. Anderson and C. Hendrickson (2001), Extreme Programming Installed, Addison-Wesley, ISBN: 201-70842-6.

Keating, B. A., P.S. Carberry, G.L. Hammer, M.E. Probert, M.J. Robertson, D.P. Holzworth, N.I. Huth, J.N.G. Hargreaves, H. Meinke, Z. Hochman, G. McLean, K. Verburg, V.O. Snow, J.P. Dimes, D.M. Silburn, E. Wang, S.D. Brown, K.L. Bristow, S. Asseng, S.C. Chapman, R.L. McCown, D.M. Freebairn and C.J. Smith (2003), An overview of APSIM, a model designed for farming systems simulation. *European Journal of Agronomy* 18, 267-288.

Keating, B. A., D.S. Gaydon, N.I. Huth, M.E. Probert, K. Verburg, C.J. Smith, and W.J. Bond (2002), Use of modelling to explore the water balance of dryland farming systems in the Murray-Darling Basin, Australia. *European Journal of Agronomy* 18, 159-169.

NLWRA. (2000), Dryland Salinity. Australia's dryland salinity assessment 2000. National Land and Water Resources Audit ISBN 0 642 37106 7. Nov 2000

Pastres, R., D. Brigolin, A. Petrizzo and M. Zucchetta (2004), Testing the robustness of primary production models in shallow coastal areas: a case study. *Ecological Modelling*. 179, 221-233.

Prisley, S. P. and M.J. Mortimer (2004), A synthesis of literature on evaluation of models for policy applications, with implications for forest carbon accounting. *Forest Ecology and Management* 198, 89-103.

Robertson, M.J., P.S. Carberry, N.I. Huth, J.E. Turpin, M.E. Probert, P.L. Poulton, M. Bell, G.C. Wright, S.J. Yeates, and R.B. Brinsmead (2002), Simulation of growth and development of diverse legume species in APSIM. *Australian Journal of Agricultural Research*, 53, 429-446.

Rykiel Jr., E.J. (1996), Testing ecological models: the meaning of validation. *Ecological Modelling*. 90, 229-244.

Seibert, J. and J.J. McDonnell (2002), On the dialog between experimentalist and modeler in catchment hydrology: Use of soft data for multicriteria model calibration. *Water Resources Research*. 38, art. no. 1241.

Wang, E., M.J. Robertson, G.L. Hammer, P.S. Carberry, D.P. Holzworth, H. Meinke, S.C. Chapman, J.N.G. Hargreaves, N.I. Huth and G. McLean (2002), Development of a generic crop model template in the cropping system model APSIM. *European Journal of Agronomy*, 18, 121-140.

Wang, E., E.J. van Oosterom, H. Meinke, S. Asseng, M.J. Robertson, N.I. Huth, B.A. Keating, and M.E. Probert (2003), The new APSIM-Wheat model - performance and future improvements. Solutions for a better environment: Proceedings of the 11th Australian Agronomy Conference Geelong, Victoria. 2003.

Zhang, L., Dawes W.R., Walker G.R., (2001), Response of mean annual evapotranspiration changes at catchment scale, *Water Resources Research*, Volume: 57, 701-708.