Improved testing of soil water balance models; an example using APSIM

Brett Robinson[1], Jasim Uddin[1], David Freebairn[1] and David McClymont[2]

[1] - University of Southern Queensland, Toowoomba, Qld, 4350; contact brett.robinson@usq.edu.au
[2] - DHM Environmental Software Engineering Pty Ltd, Toowoomba, Qld, 4350; david@dhmsoftware.com.au

Abstract

Soil water balance models provide input to models of agricultural and natural systems that inform soil and crop management. Independent empirical testing is needed to establish their reliability and accuracy.

We consider and contrast statistical methods used to test models, and highlight some common errors. Our metric of model accuracy is the skill (R^2) in predicting the gain in soil water (mm) from the start to the end of summer fallows. APSIM is benchmarked against Fallow Efficiency (FE), which requires only a single parameter. APSIM was supplied with detailed definitions of the starting conditions, the soil, tillage and meteorological conditions.

FE was poor at predicting fallow gain in soil water. APSIM was much better, simulating the fallow gain in soil water for two tillage treatments at Greenmount (residues burnt and zero till) with R^2 of 0.5 and 0.65 with respect to the observed=predicted line. As expected the lines of best fit (predicted = a + b.observed) had higher R^2. For a mean observed gain in soil water of 108 mm the mean absolute errors across the two treatments were 26 and 47 mm for APSIM and FE respectively. APSIM had no skill (R^2<0) in predicting the amount of extra soil water (mm) stored by the zero till treatment.

Key words
Simulation modelling, statistical testing, fallow, tillage

Introduction

Soil water models combine mathematical functions that represent infiltration and runoff, potential and actual soil evaporation and transpiration, soil water extraction and redistribution, deep drainage and other processes. With several co-dependent processes being simulated, and sub models that are usually developed independently, it would be a mistake to assume that these models are accurate. With many parameter choices affecting the results, it is also a mistake to assume that a user, even a highly skilled user, can obtain accurate and reliable results. There was no adjusting, “tuning” or “calibrating” of any parameters to fit observations.

These models are important. APSIM (Keating *et al*. 2003), PERFECT (Littleboy *et al*. 1989) and HowLeaky (McClymont *et al*. 2015) have been used to simulate crop yields (e.g. Whitbread and Hancock 2008), runoff and erosion (e.g. Thornton *et al*. 2007) and deep drainage (e.g. Robinson *et al*. 2007) respectively. Their soil water balances have all been developed from earlier models such as CREAMS (Knisel 1980). Testing the accuracy of the eminent agronomic model APSIM is therefore an appropriate case to investigate.

How can accuracy be assessed? Scatterplots of observations and model predictions are popular, combined with an r^2 or root mean square error (RMSE). For example, Whitbread and Hancock (2008) found that “A regression of the predicted against observed yields result in an r^2 = 0.66 and RMSE of 0.64 t/ha, n = 35 (Fig. 2a)”. However, soil water presents a special statistical problem. Simulation conditions including soil water are usually set equal to the observations at the start of the simulation, creating an artificial r^2 of 1 at that time. The r^2 declines over time during the simulation period as the artefact of initialisation is diluted. Short simulations such as fallows, with widely varying starting conditions are especially affected by this effect. Unfortunately, these were the conditions under which the soil water balance (SoilWat) in APSIM (Probert *et al*. 1996) was “validated”. In this study we improve model testing by plotting *changes* in observed and predicted soil water, eliminating the artefact of initialisation and subsequent auto-correlation.

Methods

The two statistical approaches (r^2 for the line of best fit and R^2 for the 1:1 line) are applied to a case study from the Greenmount erosion and farming systems experiment. Fallows are simulated to avoid testing the complexities of estimating transpiration and soil water extraction by roots.

(a) Statistical methods

A scatterplot of observed data (x) against predicted or simulated data (y) is a common demonstration of model fit. Perfect correspondence between the two datasets results in the data (x_i,y_i) falling on a line where y = x with a coefficient of determination (r^2) of 1. An r^2 or R^2 value is commonly quoted for the line of best fit between x and y, according to y = *a* + *b*.x. However the desired relationship is y = x. This is an important difference, as we shall demonstrate.

Although sometimes confused, the statistic r^2 is only used for the line of best fit (y = *a* + *b*.x), where it is equal to the correlation coefficient r squared, while R^2 is a generic indicator of goodness of fit that can be applied to any relationship, including our y = x. For all relationships other than the line of best fit r^2 will be greater than R^2. We calculate both the r^2 and R^2 for y = x. We also calculate the mean absolute error (mean|observed-predicted|), which is probably a little easier to interpret than the RMSE.

We have used the standard method of calculating R^2, shown in Equation 1. The parameters are the predicted data $y_{i,}$ with a mean of $\bar{y}$, and the corresponding data for the fitted function f_i.

$$R^2 = 1 - SSerror / SStotal \quad \ldots(1)$$

where $SSerror= \sum i(fi- yi)^2$

and $SStotal= \sum i(yi- \overline{y})^2$

(b) Greenmount fallows

Each of the soil water measurements is the mean of 9 gravimetric samples (5 cm cores) distributed spatially through a treatment in 3 groups of 3 (end, middle, end). The bulk density was measured in each soil layer. The volumetric water contents were downloaded from http://www.howleaky.net/ for 2 types of fallows at Greenmount (Burnt and Zero Till), summarised in Table 1. Fallow length as shown in Table 1 is the number of days between the first post-harvest soil water measurement and the last pre-plant measurement.

Table 1. Characteristics of summer fallows at the Greenmount site (data from http://www.howleaky.net/index.php/library/supersites/97-library/site-summaries/cropping/greenmount/146-greenmount-level-4).

	Burnt treatment		Zero till treatment	
	mean	(range)	mean	(range)
Fallow length (days)	151	(76 - 242)	126	(78 – 199)
Fallow rainfall (mm)	268	(63 - 468)	304	(95 – 305)
Soil water gain (mm)	114	(10 - 210)	103	(36 – 248)

(c) Simulation methods

APSIM accounts for 75% of publications concerning crop simulation in Australia (Robertson and Carberry 2010) and was therefore chosen to represent the cohort of Australian soil water balance models. Version 7.7 (build 11 Dec 2014) was used via the standard user interface. Each fallow was simulated separately, with the simulated soil water being reset to the observed amount on the date of first measurement after the harvest of the wheat crop and observed and predicted gains were calculated from that time. The burning of crop residues was simulated in APSIM using the relevant date and tillage option.

The FE method equates fallow soil water gain to 25% of fallow rainfall. Efficiencies of 20% and 30% were also calculated. The changes in r^2 and R^2 values were small because the predictions change in unison.

(d) A note of caution

The models are being tested by comparison with observations that are not free from error. David Freebairn (pers. comm.) found a standard deviation of 60mm of measured soil water at a small site with apparently uniform clay soil. There are also uncertainties that change through time; the longer that a site is monitored, the wider the likely spread between the observed upper and lower limits, affecting the simulation model parameters. How long should a paddock be monitored to estimate the limits of soil water? A more complete analysis should be made elsewhere. Suffice to say that investigating the veracity of both the observations and predictions is important.

Results

Figures 1 and 2 show the observed and predicted results from the two treatments at Greenmount. Relative to APSIM, the fallow efficiency method is a poor predictor of the gain in soil water.

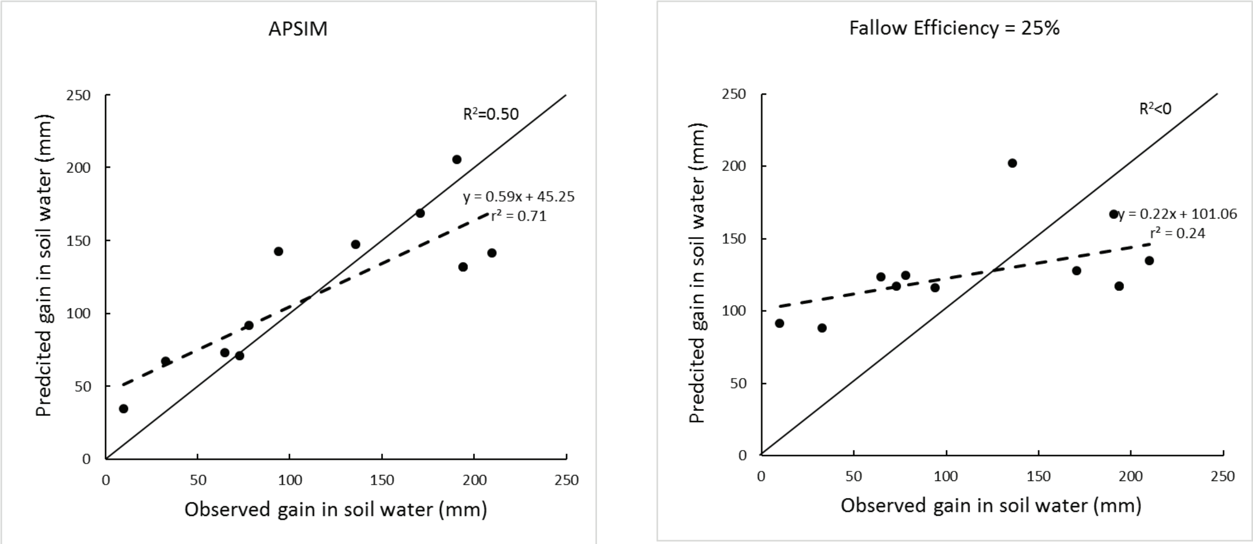

Figure 1. The observed and predicted data for the Burnt treatment.

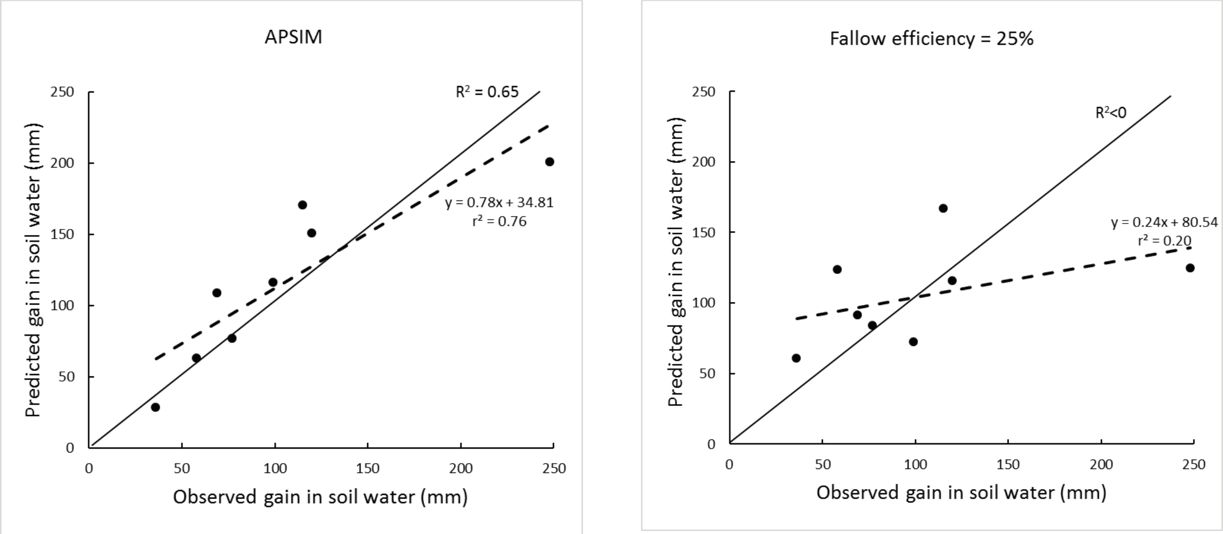

Figure 2. The observed and predicted data for the Zero Till treatment.

Figure 3 shows the observed differences and APSIM's predicted differences between Zero Till and Burnt. The slope of the line of best fit (0.47) is much less than the ideal slope (1.0).

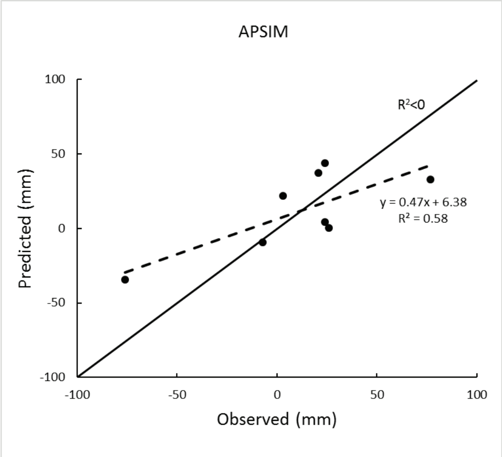

Figure 3. The observed and predicted difference between the Zero Till and Burnt treatments.

Discussion

We have established that the correct metric for judging models is the predicted change in soil water. The correlation between starting and finishing soil water excludes the use of soil water *per* se as a metric. The lines of best fit for APSIM have slopes well below the desired slope of 1 (0.59 and 0.78, Figures 1 and 2). Consequently, there was a considerable difference in the r^2 or R^2 values for the lines of best fit (0.71 and 0.76 for the two treatments) and the observed=predicted line (0.5 and 0.65). The lines of best fit for the FE model have low slopes (0.22 and 0.24, Figures 1 and 2). The R^2<0 for the observed=predicted lines indicate that the mean observation () explains more of the variation in the observed values than the FE predictions.

The artificially high r^2 values for the lines of best fit overestimate the accuracy of both models; APSIM by a moderate amount (R^2 inflated by 21% and 11%) and FE by a large amount (R^2 inflated by 22 and 24%).

For a mean observed gain in soil water of 108 mm in these fallows, the mean errors (absolute) for the two treatments were 26 mm for APSIM and 47 mm for FE. The largest errors (absolute) were 69 mm for APSIM and 127 mm for FE. APSIM had no skill (R^2<0) in predicting the amount of extra soil water (mm) stored by the zero till treatment relative to the burnt treatment. Are APSIM users aware of this lack of skill?

The APSIM model is clearly better in this case than the FE model. This is not too surprising given that evaporation and runoff vary between fallows in ways that FE may not represent. The single, lumped parameter of the FE model appears insufficient to represent the diverse physical processes of soil water storage, while the dozens of parameters of APSIM and similar models can better represent these processes. Further research is required to establish which parts and parameters of APSIM-like models are critical to the task of predicting soil water storage and which, if any, are unnecessary or unnecessarily complicated.

References

Keating BA, Carberry PS, Hammer GL, Probert ME, Robertson MJ, Holzworth D, Huth NI, Hargreaves JNG, Meinke H, Hochman Z, McLean G, Verburg K, Snow V, Dimes JP, Silburn M, Wang E, Brown S, Bristow KL, Asseng S, Chapman S, McCown RL, Freebairn DM and Smith CJ. 2003. An overview of APSIM, a model designed for farming systems simulation. *European Journal of Agronomy* 18:267-288.

Knisel WG. 1980. 'CREAMS: A Fieldscale Model for Chemical, Runoff, and Erosion from Agricultural Management Systems,' USDA, Science and Education Administration, Conservation Report No. 26, Washington, DC.

Littleboy M, Silburn DM, Freebairn DM, Woodruff DR and Hammer GL. 1989. PERFECT, A computer simulation model of Productivity, Erosion, Runoff Functions to Evaluate Conservation Techniques. Queensland Department of Primary Industries, Bulletin QB89005, 119 pp.

McClymont D, Freebairn DM, Rattray DJ, Robinson JB and White S. 2015. Howleaky: Exploring water balance and water quality implication of different land uses. Software V2.17 (http://howleaky.net/)

Probert ME, Dimes JP, Dalal RC and Strong WM. 1996. APSIM SoilWat and SoilN: Validation against observed data for a cracking clay soil. *Proceedings of 8th Agronomy Conference, Toowoomba*.

Thornton CM, Cowie BA, Freebairn DM, Playford CL. 2007. The Brigalow Catchment Study: II. Clearing brigalow (Acacia harpophylla) for cropping or pasture increases runoff. *Australian Journal of Soil Research* 45, 496 – 511.

Robertson M and Carberry P. 2010. The evolving role of crop modelling in agronomy research. *Proceedings of 15th Agronomy Conference,* Lincoln.

Robinson JB, Silburn DM, Rattray D, Freebairn DM, Biggs A, McClymont D, Christodoulou N. 2010. Modelling shows that the high rates of deep drainage in parts of the Goondoola Basin in semi-arid Queensland can be reduced with changes to the farming systems. *Australian Journal of Soil Research* 48, 58 – 68.

Whitbread A and Hancock J. 2008. Estimating grain yield with the French and Schultz approaches vs simulating attainable yield with APSIM on the Eyre Peninsula. *Proceedings of 14th Agronomy Conference* 2008, Adelaide.