Vol. 55: 253–265, 2013
doi: 10.3354/cr01138

CLIMATE RESEARCH
Clim Res

Published online February 6

Statistical emulators of a plant growth simulation model

Padmaja Ramankutty[1,2,3,*], Megan Ryan[1,2], Roger Lawes[4], Jane Speijers[3], Michael Renton[1,2,4]

[1]School of Plant Biology and Institute of Agriculture, and [2]Future Farm Industries CRC, The University of Western Australia, 35 Stirling Hwy, Crawley, Western Australia 6009, Australia
[3]Department of Agriculture and Food, 3 Baron-Hay Court, South Perth, Western Australia 6151, Australia
[4]CSIRO Ecosystem Sciences, Private Bag 5, Wembley, Western Australia 6913, Australia

ABSTRACT: A new methodology was developed to create statistical emulators from an existing complex plant growth simulation model, in order to provide simple models for predicting perennial pasture production based on meteorological data. The goal was to find emulators that could fit simulation data and provide reliable predictions of new simulation output. Four types of statistical emulators were compared — linear models, non-linear models, linear splines and cubic splines — which varied in complexity due to the use of different functional forms to fit main effects and different methods to determine model parameters. Two methods of model validation were used to ensure that the emulator selected provided good fits to the initial simulation output and gave adequate predictions of future simulation output. We illustrated the various techniques of model development and validation using the Agricultural Production Systems sIMulator (APSIM) plant growth model. The effects of rainfall amount, rainfall frequency, temperature and radiation on APSIM-generated biomass production of the perennial pasture species lucerne *Medicago sativa* were used as a case study. We conclude that, in this case, the best statistical emulator is the cubic spline. The novel approach illustrated here provides a rigorous, flexible and powerful means of developing and testing emulators for any complex simulation model.

KEY WORDS: Statistical emulator · Linear mixed effects model · Model validation · Cubic splines · APSIM · *Medicago sativa* · Biomass production

1. INTRODUCTION

Mechanistic computer simulation models that represent the bio-physical processes driving plant growth can help understand the interactions between the processes involved and weigh the benefits of different management options (Bouman et al. 1996, Moore et al. 1997, Brisson et al. 2003, Jones et al. 2003, Keating et al. 2003, van Ittersum et al. 2003). Such mechanistic models use computer simulation to represent many of the physical and biological processes involved in plant growth, thereby increasing the clarity and understanding of how these processes and their underlying mechanisms are involved in predicting plant development. This is the strength of complex plant growth models: by taking a relatively mechanistic approach, these physical and biological processes are represented at a relatively high level of detail and realism.

However, there are disadvantages to using such a mechanistic and complex approach to modelling plant growth. Their mechanistic detail can make them inaccessible to people who have limited understanding of the bio-physical processes involved. Pre-

*Email: ramanp01@student.uwa.edu.au

diction of yield or production using these types of models requires extensive inputs such as detailed quantitative information, first on plant physiology and phenology for each new species, and then on soil characteristics and meteorological data for each new location. Furthermore, such models may consist of several connected modules that are often developed by different people with particular areas of expertise. As such, the models can lack transparency, which can make it difficult for even experienced agronomists to modify them for a new species or a new environment without specialised training. Modification can be especially challenging in situations where not all of the requisite information is available or where significant structural changes are required because of a change in understanding of the driving processes (Holzworth et al. 2010). Complex simulation models are relatively computationally expensive, taking longer to run than simpler models, which can be significant when models need to be run repeatedly. Therefore, it is desirable to create simplified versions of complex growth models that retain their benefits, but without these disadvantages.

Various methods have been used to simplify complex plant growth models. Gibbons et al. (2010) developed a method that generates simplified models by combining existing model variables; Iglesias et al. (2000) developed a multiple linear regression model and used it to conduct a spatial analysis of climate change impacts on national wheat production in Spain; Iizumi et al. (2009b) used a Bayesian approach to estimate model parameters and quantify uncertainty of yield estimates; Iizumi et al. (2009a) created a lookup table from numerous simulation outputs to reduce the simulation time in integrated impact assessment models; Lobell & Burke (2010) assessed the performance of statistical models against more complex models; and Brooks et al. (2001) developed meta-models of a wheat-growth simulation model.

However, in this study we developed a method for simplifying complex growth models based on creating simple statistical models, or statistical emulators, which accurately predict a particular outcome of a computer simulation model using only its principal drivers as inputs (Conti & O'Hagan 2010). Emulators are designed to eliminate much of the complexity inherent in the simulation model to provide greater clarity and transparency in the modelling process. While being less complex in design than the simulation model, they can still provide reasonably accurate predictions of the response of interest. To illustrate this methodology, we chose to use the Agricultural Production Systems sIMulator (APSIM) as the complex growth model (Keating et al. 2003), and emulate the biomass production for the perennial pasture species lucerne *Medicago sativa* predicted by APSIM based on meteorological and soil-type input data (Probert et al. 1998, Robertson et al. 2002). To ensure the robustness of the resulting emulators, we developed and compared statistical emulators for APSIM-generated lucerne biomass production based on 4 levels of model complexity, and then used statistical model validation to determine which of the 4 emulators provides the best model in terms of fit and predictive ability.

2. METHODS

2.1. Case study

A dataset of lucerne production (dry biomass) was generated from APSIM using weather data from the Badgingarra Weather Station, Western Australia (30.34°S, 115.54°E) taken from the Australian Bureau of Meteorology database (www.bom.gov.au/climate/data) and described in Jeffrey et al. (2001). This file contained historical daily information of rainfall (mm), radiation, Radn (MJ m^{-2}), and maximum and minimum temperatures (°C) for ~100 yr (1889–2006). Using APSIM simulation, lucerne was planted on 15 May 1889 (early autumn) and allowed to mature for the first 11 yr until monitoring was started on 1 December 1900. The initial growth for 11 yr was to ensure that only production of well-developed plants was assessed. For the purposes of this study of developing a statistical emulator, we chose to run all simulations in a simple sandy soil. For full details of the dataset used, see the figures in the technical report available at www.michaelrenton.info/publications.

Daily average temperature, Temp, was calculated from this weather data as: $T_{ave} = (T_{min} + T_{max}) \div 2$, where T_{min} and T_{max} are the daily minimum and maximum temperature; the dataset consisting of these readings, along with the resulting biomass production generated by APSIM, was consolidated into 2-monthly readings (using total rainfall, average radiation and temperatures, and total biomass production) from 1 December 1900 to 30 September 2006 to give 641 data points, 6 for each year. It was also standardised by multiplying each reading by [61 ÷ (number of days in 2-monthly period)] so that each 2-month period had 61 d. The resulting dataset was split to obtain 2 sets of almost equal size. One set, known as the Fitting Set, was used to develop the statistical emulators, while the other set, known as the Validating Set, was used independently to determine the predictive ability of these statistical emulators. Each set had >300 data points, thus ensuring the reliability of the models developed.

Initially the data were split on the basis of time, with Set 1 containing all the values from 1 December 1900 to 30 November 1953, and Set 2 containing the remaining values. However, since the weather data prior to the 1960s were based solely on interpolations from nearby weather stations, it was noted that this data split could result in differences due to background causes such as climate change between the Fitting and the Validating Set, which could then contribute to invalidation of the fitted models. So another method of data splitting was also considered. To minimise any loss of the correlation between years that existed in the original dataset, the second method of data splitting consisted of dividing the data into groups of 4 consecutive years and these groups were numbered sequentially. Set 3 contained all the odd-numbered groups and Set 4 contained the even-numbered groups. Models were then developed using each set in turn as the Fitting Set with the corresponding remaining data forming the Validating Set.

2.2. Emulator construction

All statistical modelling was carried out using R (R Development Core Team 2009). A preliminary analysis was conducted, based on plotting production predictions against one varying driving variable, while all other driving variables were held constant. Results from this preliminary analysis are included in the technical report (available at www.michaelrenton.info/publications), and indicated that main effects of rainfall amount (RAmt) and Radn on APSIM-generated biomass production could be modelled using sigmoid curves, while a unimodal function could be used to model the main effect of Temp. However, examination of the data indicated that simpler models might also provide a reasonable fit to the relationships between APSIM-generated lucerne biomass production and either RAmt or Radn, as most of the variation in biomass appeared to be linear. This linear relationship between production and Radn was also found in Iizumi et al. (2010) for the response of paddy rice yield to solar radiation bias. There was also evidence of effects of rainfall frequency in days (RFreq) and interactions between all these terms. However, it was unclear what functional forms these effects might take.

2.2.1. Linear emulator

The linear emulator included the linear main effects of RAmt, RFreq and Radn, both linear and quadratic effects of Temp, as well as linear interactions between these variables. These models were fitted using the R 'lm' function (Chambers 1992), which gives a best fitted linear regression model based on the method of least squares. The fitted full linear emulator included all main effects and all interactions and was of the form:

$$\begin{aligned} y = {} & a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5x_4^2 + a_6x_1x_2 + \\ & a_7x_1x_3 + a_8x_1x_4 + a_9x_1x_4^2 + a_{10}x_2x_3 + a_{11}x_2x_4 + \\ & a_{12}x_2x_4^2 + a_{13}x_3x_4 + a_{14}x_3x_4^2 + a_{15}x_1x_2x_3 + \\ & a_{16}x_1x_2x_4 + a_{17}x_1x_2x_4^2 + a_{18}x_1x_3x_4 + a_{19}x_1x_3x_4^2 + \\ & a_{20}x_2x_3x_4 + a_{21}x_2x_3x_4^2 + a_{22}x_1x_2x_3x_4 + a_{23}x_1x_2x_3x_4^2 \end{aligned} \tag{1}$$

where y = APSIM-generated lucerne biomass production, x_1 = RAmt, x_2 = RFreq, x_3 = Radn, x_4 = Temp, and a_i, for i = 1...23, are regression coefficients.

2.2.2. Non-linear emulator

In the non-linear emulator, non-linear parametric models were fitted to the relationships between dependent and independent variables. These non-linear models were fitted using the 4-parameter logistic function (SSfpl) for RAmt and the 3-parameter logistic function (SSlogis) for Radn because an initial examination indicated that the lower asymptote was zero for Radn but non-zero for RAmt. A simple linear model was fitted for RFreq and a quadratic polynomial for Temp. All first-order interactions of these variables were included in the model as interactions between linear effects. This model was referred to as the fitted full non-linear emulator. Higher order interactions were not investigated for non-linear models due to problems with convergence of the iterative processes involved in determining the least squares estimates of the parameters and the belief that they would not be significant given the variability in the response variable. Non-linear modelling was carried out using the R 'nls' function (Crawley 2007, R Development Core Team 2009), which determines the non-linear (weighted) least-squares estimates of the parameters of a non-linear model. The fitted full non-linear model was of the form:

$$\begin{aligned} y = {} & a_0 + \mathrm{SSfpl}(x_1) + a_2x_2 + a_3x_1x_2 + \mathrm{SSlogis}(x_3) + \\ & a_4x_1x_3 + a_5x_2x_3 + a_6x_4 + a_7x_1x_4 + a_8x_2x_4 + \\ & a_9x_3x_4 + a_{10}x_4^2 + a_{11}x_1x_4^2 + a_{12}x_2x_4^2 + a_{13}x_3x_4^2 \end{aligned} \tag{2}$$

where $\mathrm{SSfpl}(x_1) = a_1 + \frac{(b_1 - a_1)}{1 + \exp[(k_1 - x_1)/m_1]}$

and $\mathrm{SSlogis}\,(x_3) = \frac{b_2}{1 + \exp[(k_2 - x_3)/m_2]}$

and a_1, b_1, k_1, k_2 are model parameters to be fitted.

2.2.3. Spline emulators

One means of incorporating more flexibility into the emulator, to account for non-linear variation in the response, is to base the emulator on semi-parametric regressions or p^{th}-order splines. A spline is a piecewise polynomial function that has knots at various points along the real axis where segments join (Smith 1979, Wand 2000). A p^{th}-order spline is made up of pth-order polynomials that join smoothly up to the $(p - 1)$th derivative at each knot (Friedman 1991). In particular we examined linear splines and cubic splines. A linear spline, also called a piecewise linear model, is composed of line segments that join to one another to form a continuous function. Linear splines do not join smoothly at the knots; they are not continuously differentiable. The R 'segmented' function (Muggeo 2008), which fits regression models with broken-line relationships, was used to obtain parameter values for the linear spline by an iterative process. Interaction effects were not examined in this case, since their inclusion would have resulted in models that were too complicated and unwieldy to be useful. The fitted linear spline emulator containing all main effects was of the form:

$$y = a_0 + f_1(x_1) + f_2(x_2) + f_3(x_3) + f_4(x_4) \tag{3}$$

where $f_j(x_j) = a_j + b_j x_j + \sum_{i=1}^{n_j} c_i (x_j - \psi_i)_+$

ψ_i are knots

a_j, b_j, c_j are regression coefficients

and $(x_j - \psi_i)_+ = \begin{cases} (x_j - \psi_i), & (x_j - \psi_i) \geq 0 \\ 0, & (x_j - \psi_i) < 0 \end{cases}$

Cubic splines are constructed from segments of third-order polynomials which are constrained to blend smoothly where segments join. Cubic splines were fitted to the main effects of RAmt, RFreq, Radn and Temp while all interactions of RAmt, RFreq, Radn and Temp were fitted as linear effects. These models, known also as linear mixed effects models (Verbyla et al. 1999), were fitted using the R 'asreml' function (Gilmour et al. 2007), which estimates variance components under a general linear mixed model by residual maximum likelihood (REML). This function allocated 2 components to each spline; a linear trend component and a component which represented the variability about that trend. In this procedure the linear trend was fitted as a fixed effect and variability about that trend was fitted as a random effect. It is commonly assumed that cubic splines have a tendency to overfit, and thus may not provide good predictions (Wold 1992). Methods to avoid overfitting include selection of the number of knots (Friedman 1991). We therefore tested several possible knot numbers and chose to use the minimum number that was sufficient to ensure that the shape of the curve did not change markedly with a further increase in the number of knots. In this case, 12 knots was sufficient.

The fitted full cubic spline emulator was of the form:

$$\begin{aligned} y &= a_0 + a_1x_1 + a_2x_1^2 + a_3x_1^3 + a_4x_2 + a_5x_2^2 + a_6x_2^3 + \\ &a_7x_3 + a_8x_3^2 + a_9x_3^3 + a_{10}x_4 + a_{11}x_4^2 + a_{12}x_4^3 + \\ &\sum_{i=1}^{4}\sum_{k=1}^{K_i} b_{ik}(x_{ik} - \psi_{ik})_+^3 + a_{13}x_1x_2 + a_{14}x_1x_3 + a_{15}x_1x_4 + \\ &a_{16}x_2x_3 + a_{17}x_2x_4 + a_{18}x_3x_4 + a_{19}x_1x_2x_3 + a_{20}x_1x_2x_4 + \\ &a_{21}x_1x_3x_4 + a_{22}x_2x_3x_4 + a_{23}x_1x_2x_3x_4 \end{aligned} \tag{4}$$

Table 1 summarises the models that were fitted to each independent variable.

2.2.4. Model selection

The emulators described above represent the initial Full Models for each emulator. Model selection was then carried out based on backward selection from this Full Model. Each term of the Full Model was then omitted sequentially, starting with the highest order term(s), to create several nested submodels (not shown). Each submodel differed from its preceding neighbour by the elimination of only 1 term. Adjacent pairs of models of this sequence (starting with the Full Model) were compared using analysis of variance (ANOVA), or REML in the case of cubic spline models. This comparison was used to determine if the term dropped from the model was significant ($p < 0.05$). If so, then the preceding neighbour model was taken as the Simplified Full Model. For the spline models, all variance components were examined and the random model was simplified where possible on the basis of likelihood ratio tests prior to examination of the fixed effects as outlined above. The Simplified Full Model produced for each of the emulator functions is in the Results section.

Table 1. Emulators fitted to each independent variable in this study. RAmt: rainfall amount; RFreq: rainfall frequency; Radn: radiation; Temp: temperature

Independent variable	Linear emulators		Non-linear emulators			Spline emulators	
	Linear	Quadratic	Linear	Quadratic	Logistic	Linear	Cubic
RAmt	×				×[a]	×	×
RFreq	×		×			×	×
Radn	×				×[b]	×	×
Temp	×	×	×	×		×	×

[a]4-parameter logistic; [b]3-parameter logistic

Many plant growth models have been constructed with a limited set of variables and it may therefore be possible to exclude some of the variables from the Simplified Full Model when constructing other candidate models for the emulator functions. For this reason, a number of models that include subsets of the independent variables were also considered. French & Schultz (1984) suggested that potential plant growth throughout southern Australia is influenced mainly by rainfall, with ~70 % of total production variability being accounted for by this factor alone. The preliminary analysis given in the supplement indicated that RAmt may have more effect on APSIM-generated lucerne biomass production than RFreq. Therefore one model chosen for assessment contained the main effect of RAmt only (ID = 1 in Table 2) while another model contained the main effects of both RAmt and RFreq (ID = 2 in Table 2) but excluded Temp and Radn. All other models contained the effects of both RAmt and RFreq as well as various combinations of the other independent variables (Radn, Temp, $Temp^2$) to create 5 more models (ID = 3–7 in Table 2).

Table 2. Potential emulator models fitted using each emulator function. RAmt: rainfall amount; RFreq: rainfall frequency; Radn: radiation; Temp: temperature; (✓) inclusion; (×) exclusion

ID	Independent variables in model	Linear	Non-linear	Linear spline	Cubic spline
1	RAmt	✓	✓	✓	✓
2	RAmt, RFreq	✓	✓	✓	✓
3	RAmt, RFreq, Radn	✓	✓	✓	✓
4	RAmt, RFreq, Temp	✓	✓	✓	✓
5	RAmt, RFreq, Radn, Temp	✓	✓	✓	✓
6	RAmt, RFreq, Temp, $Temp^2$	✓	✓	×	×
7	RAmt, RFreq, Radn, Temp, $Temp^2$	✓	✓	×	×
8	Full Model	✓	✓	×	✓

2.3. Emulator validation

Statistical model validation is used to ensure that a statistical model not only provides a good fit to existing data but also gives adequate predictions of future observations; in the case of a statistical emulator, existing data is the data from the simulation model used to build the emulator and future observations are potential future predictions of the simulation model (Bastos & O'Hagan 2009). Model validation can involve both graphical and numerical techniques. Graphical methods can include plots of emulator predictions against new simulation-generated output and scatter plots of residual errors against fitted values from the emulator model but these provide only a subjective evaluation (Mayer & Butler 1993). Numerical methods such as bias, determination coefficient (R^2) statistic, and root mean square error provide a more quantitative means of validation, but each method tends to focus on only a limited aspect of the relationship between emulator model data and simulation model data (Walther & Moore 2005).

In this study, validation of the emulator models shown in Table 2 was carried out in 2 ways (Snee 1977). The first method (Validation Method 1) used mean squared prediction error (MSPE), which measures the prediction capability of the model parameterised by the Fitting Set, on the Validating Set (Neter et al. 1989, Stevenson et al. 2005):

$$\text{MSPE} = \sum_{i=1}^{n^*} \frac{(Y_i - \hat{Y}_i)^2}{n^*} \qquad (5)$$

where Y_i is i^{th} response value from the Validating Set, $\hat{Y}_i$ is predicted value of Y_i using the model parameterisation from the Fitting Set and n^* is the number of observations in the Validating Set.

If the value of this statistic is close to the mean square error (MSE) from the Fitting Set then the MSE is considered to give an indication of the predictive capability of the fitted model (Neter et al. 1989).

The second method of model validation (Validation Method 2) involved the calculation of a prediction sum of squares (PRESS) statistic for the fitted model from each dataset. The PRESS statistic is an adjustment to the residual sum of squares (SSE) to account for the leverage of each data value (Allen 1974). The leverage of a data value, h_{ii} with $0 \le h_{ii} \le 1$, is the amount of influence that the value has in defining the shape of the fitted model (Wei et al. 1998):

$$\text{PRESS} = \sum_{i=1}^{n} \left(\frac{e_i}{1-h_{ii}}\right)^2 \quad (6)$$

where e_i is the ordinary residual, h_{ii} is the i^{th} leverage value calculated as the i^{th} diagonal element of the hat matrix (i.e. the n × n matrix formed by the least squares estimates of the parameter values of the fitted model) except in the case of models fitted using the nls function when $h_{ii} = \sum_{k=1}^{p} q_{ik}^2$, $i = 1 \ldots N$, where $Q = [q_{ik}]_{N \times p}$ is the orthogonal matrix of the QR decomposition of the Jacobian matrix of model parameters (St Laurent & Cook 1992, 1993, Rivals & Personnaz 2004). (q_{ik} is the element in the i^{th} row and j^{th} column of the matrix Q which has N rows and p columns.)

Thus data values with high leverage increase the PRESS value. PRESS values that are close to the corresponding SSE values imply that the MSE can be used as an indicator of the predictive ability of the model (Neter et al. 1989).

We calculated MSPE:MSE and PRESS:SSE ratios for each potential emulator model with fitting and validating set combinations as shown in Table 3. This table simply defines a notation that is then used in Table 4 for clarity. The closer these ratios are to 1 for each of the 4 datasets, the greater the predictive ability of the corresponding model.

Thus from each emulator function, the best model(s) in terms of both fitting ability and predicting ability were those which, across all 4 datasets, had low MSE values, and MSPE:MSE and PRESS:SSE ratios close to 1.

The Akaike information criterion (AIC) (Akaike 1974) and the Bayes information criterion (BIC) (Schwarz 1978) were used as guides for comparison of the selected non-nested models (Hastie et al. 2009). The model with the lowest AIC and BIC was considered to be the best in terms of predictive ability (Hastie et al. 2009). Thus the model that gave the best emulation of APSIM-generated biomass production was determined. The variance components for the random effects and the coefficients of the fixed effects of this model, along with their associated standard errors, were then obtained by fitting this model using the entire dataset originally generated from APSIM for this study.

Table 3. Ratios calculated using the different Fitting Set and Validating Set combinations for each validation method

Fitting Set	Validating Set	Validation Method 1	Validation Method 2
1	2	$\left(\frac{\text{MSPE}}{\text{MSE}}\right)_{\text{Set1}}$	$\left(\frac{\text{PRESS}}{\text{SSE}}\right)_{\text{Set1}}$
2	1	$\left(\frac{\text{MSPE}}{\text{MSE}}\right)_{\text{Set2}}$	$\left(\frac{\text{PRESS}}{\text{SSE}}\right)_{\text{Set2}}$
3	4	$\left(\frac{\text{MSPE}}{\text{MSE}}\right)_{\text{Set3}}$	$\left(\frac{\text{PRESS}}{\text{SSE}}\right)_{\text{Set3}}$
4	3	$\left(\frac{\text{MSPE}}{\text{MSE}}\right)_{\text{Set4}}$	$\left(\frac{\text{PRESS}}{\text{SSE}}\right)_{\text{Set4}}$

3. RESULTS

In this section, all models have been shown using symbolic representation (Wilkinson & Rogers 1973) for convenience and clarity. Simplified Full Models for each emulator type are shown below:

Linear:
1 + RAmt + Radn + Temp + Temp^2 + RAmtRadn + RAmtTemp + RAmtTemp^2 + RadnTemp + RadnTemp^2;

Non-linear:
1 + SSfpl(RAmt) + RFreq + RAmtRFreq + SS logis (Radn) + RAmtRadn + Temp;

Cubic spline:
1 + lin(RAmt) + spl(RAmt) + RFreq + lin(Radn) + spl(Radn) + lin(Temp) + spl(Temp) + RAmtRFreq + RAmtRadn + RAmtTemp + RFreqRadn + RFreqTemp + RadnTemp + RAmtRFreqRadn + RAmtRadnTemp;

where AB represents the product of the terms A and B; ABC represents the product of the terms A, B and C; for {A, B, C} ⊂ {RAmt, RFreq, Radn, Temp, Temp^2}.

Note that there was no simplification of the full model for the linear spline emulator since interactions were not included in the model, and elimination of first order effects was considered in the other models listed in Table 2.

Table 4. Regression results from model fitting and validation. Models in **bold**: acceptable validation statistics and low MSE. RAmt: rainfall amount; RFreq: rainfall frequency; Radn: radiation; Temp: temperature; (–) undefined due to observations with high leverage values

Emulator function	Independent variables in model	MSPE:MSE				PRESS:SSE				MSE (× 10^6)			
		Set 1	Set 2	Set 3	Set 4	Set 1	Set 2	Set 3	Set 4	Set 1	Set 2	Set 3	Set 4
Linear	RAmt	0.96	1.06	0.97	1.04	1.01	1.01	1.01	1.01	1.55	1.48	1.55	1.49
	RAmt, RFreq	0.96	1.06	0.97	1.04	1.02	1.02	1.02	1.02	1.55	1.48	1.55	1.49
	RAmt, RFreq, Radn	0.96	1.06	0.97	1.05	1.02	1.02	1.02	1.02	1.54	1.47	1.54	1.47
	RAmt, RFreq, Temp	**1.01**	**1.09**	**0.98**	**1.05**	**1.02**	**1.02**	**1.02**	**1.02**	**0.72**	**0.69**	**0.73**	**0.70**
	RAmt, RFreq, Radn, Temp	**1.06**	**1.09**	**0.95**	**1.07**	**1.03**	**1.03**	**1.03**	**1.03**	**0.59**	**0.58**	**0.62**	**0.58**
	RAmt, RFreq, Temp, $Temp^2$	**0.93**	**1.13**	**0.90**	**1.15**	**1.03**	**1.03**	**1.03**	**1.03**	**0.61**	**0.55**	**0.62**	**0.55**
	RAmt, RFreq, Radn, Temp, $Temp^2$	**0.98**	**1.14**	**0.87**	**1.18**	**1.04**	**1.04**	**1.03**	**1.04**	**0.48**	**0.45**	**0.51**	**0.43**
	Simplified[a] Full Model	1.27	1.05	1.07	1.11	1.07	1.08	1.07	1.08	0.19	0.21	0.21	0.20
Non-linear	RAmt	0.97	1.03	1.05	0.98	1.03	1.02	1.03	1.03	1.21	1.17	1.17	1.21
	RAmt, RFreq	0.97	1.03	1.06	0.97	1.03	1.03	1.03	1.03	1.20	1.17	1.14	1.21
	RAmt, RFreq, Radn	0.97	1.03	1.05	0.97	1.05	1.04	1.05	1.04	1.20	1.17	1.16	1.21
	RAmt, RFreq, Temp	0.97	1.03	1.05	0.97	1.04	1.04	1.04	1.04	1.21	1.17	1.16	1.21
	RAmt, RFreq, Radn, Temp	0.96	1.05	1.04	0.99	1.06	1.05	1.06	1.05	1.27	1.17	1.21	1.24
	RAmt, RFreq, Temp, $Temp^2$	0.99	1.03	1.07	0.96	1.05	1.05	1.05	1.05	1.33	1.20	1.29	1.36
	RAmt, RFreq, Radn, Temp, $Temp^2$	0.98	1.02	1.05	0.98	1.07	1.07	1.07	1.07	1.39	1.34	1.35	1.40
	Simplified[a] Full Model	1.04	0.97	1.18	0.88	24.98	29.66	29.70	13.78	1.15	1.16	1.09	1.26
Linear spline	RAmt	0.96	1.11	1.09	1.00	1.03	1.03	1.03	1.03	1.20	1.14	1.14	1.19
	RAmt, RFreq	1.03	1.05	1.19	0.96	1.07	1.06	1.08	1.06	1.01	1.00	0.97	1.04
	RAmt, RFreq, Radn	1.03	1.04	1.20	1.00	1.10	–	1.11	1.09	0.56	0.99	0.53	1.04
	RAmt, RFreq, Temp	2.46	5.43	0.98	2.16	1.11	1.10	1.11	1.09	0.37	0.36	0.41	0.36
	RAmt, RFreq, Radn, Temp	1.44	3.31	3.09	1.96	–	1.13	–	1.13	0.18	0.16	0.18	0.17
Cubic spline	RAmt	0.97	1.03	1.03	0.97	1.03	1.03	1.03	1.03	1.20	1.18	1.17	1.22
	RAmt, RFreq	1.06	1.00	1.09	0.93	1.05	1.05	1.06	1.04	1.02	1.07	1.02	1.12
	RAmt, RFreq, Radn	1.02	1.26	1.08	0.97	1.14	1.18	1.16	1.14	0.37	0.33	0.35	0.37
	RAmt, RFreq, Temp	**1.21**	**1.07**	**0.91**	**1.18**	**1.13**	**1.10**	**1.10**	**1.12**	**0.36**	**0.39**	**0.42**	**0.36**
	RAmt, RFreq, Radn, Temp	**1.10**	**1.12**	**0.85**	**1.16**	**1.13**	**1.13**	**1.13**	**1.13**	**0.17**	**0.17**	**0.19**	**0.16**
	Simplified[a] Full Model	1.30	1.27	1.01	1.18	1.17	1.17	1.19	1.17	0.16	0.15	0.17	0.16

[a]As specified at the beginning of the Results section

Table 4 showed that MSE ranged from 0.15 × 10^6— for the simplified full cubic spline emulator model when fitted using Set 2 — to 1.55 × 10^6 for the 2 simplest linear emulator models (the one including RAmt and the one including RAmt and RFreq) when fitted using either Set 1 or Set 3. Ratios for MSPE:MSE ranged from 0.85 to 5.43 and ratios for PRESS:SSE ranged from 1.01 to 29.70. The simplified full model for each of the emulator functions had at least one high validation ratio (MSPE:MSE or PRESS:SSE). In the case of the non-linear emulator function, this model also had high MSE values.

Of the 8 linear emulator models considered, only those models containing the effect of temperature had acceptable MSE values. The models that also included a quadratic term for temperature had considerably lower MSE values and also good validation statistics.

The non-linear emulator models had high residual errors as shown by their high MSE values, indicating that the models did not fit the original data well.

The 3 simplest linear spline emulator models had high MSE values while the remaining models had low MSE values but some high validation ratios (>2.0). The 2 simplest cubic spline emulator models had high MSE values. Of the 4 remaining models, the one including RAmt, RFreq, Radn and Temp, but excluding interactions, had both low MSE values and good validation statistics.

Of the original 27 emulators considered (Table 4), 6 were chosen as the best models in terms of both fitting ability and predicting ability (Table 5). It is clear that the cubic spline models had considerably lower AIC and BIC values than the linear emulator models (Table 5). This indicates that the cubic spline models selected from Table 4 had a better trade-off between accuracy and model complexity (as measured by the number of parameters) than the linear models, and were the most efficient predictive models among the alternatives in Table 5. In particular, the cubic spline emulator with 4 inde-

Table 5. Potential emulator models with acceptable validation statistics and low MSE. Models in **bold** had the lowest AIC and BIC values across all datasets. RAmt: rainfall amount; RFreq: rainfall frequency; Radn: radiation; Temp: temperature

Emulator function	Symbolic representation of model	AIC Set 1	AIC Set 2	AIC Set 3	AIC Set 4	BIC Set 1	BIC Set 2	BIC Set 3	BIC Set 4	All data R^2_{adj} (%)
Linear	1 + RAmt + RFreq + Temp	5203	5269	5384	5094	5222	5288	5403	5113	67.2
	1 + RAmt + RFreq + Radn + Temp	5142	5215	5331	5036	5165	5238	5354	5059	72.5
	1 + RAmt + RFreq + Temp + $Temp^2$	5152	5201	5333	5021	5174	5223	5356	5043	73.0
	1 + RAmt + RFreq + Radn + Temp + $Temp^2$	5079	5133	5269	4948	5105	5159	5295	4974	78.3
Cubic spline	**1 + lin(RAmt) + spl(RAmt) + lin(RFreq) + spl(RFreq) + lin(Temp) + spl(Temp)**	**4426**	**4505**	**4616**	**4341**	**4464**	**4543**	**4655**	**4379**	**82.7**
	1 + lin(RAmt) + spl(RAmt) + lin(RFreq) + spl(RFreq) + lin(Radn) + spl(Radn) + lin(Temp) + spl(Temp)	**4179**	**4241**	**4363**	**4083**	**4227**	**4289**	**4411**	**4131**	**92.2**

pendent variables (RAmt, RFreq, Radn and Temp) was the single most efficient predictive emulator since it consistently had the lowest AIC and BIC values across all datasets.

Each of the 6 emulators shown in Table 5 was now fitted to the whole dataset, permitting comparisons between fitted values from the model and actual values generated from APSIM. Fig. 1 confirms that the emulator that best fits the original simulation data is the cubic spline with 4 independent variables. None of the models fitted particularly well for very low levels of simulated biomass, but the cubic spline emulator with 4 independent variables showed the least deviation from the line of best fit. In addition, this emulator was the only one that fitted well at high levels of simulated biomass.

The presence of clusters of points, with some close to the line of best fit and some further away from this

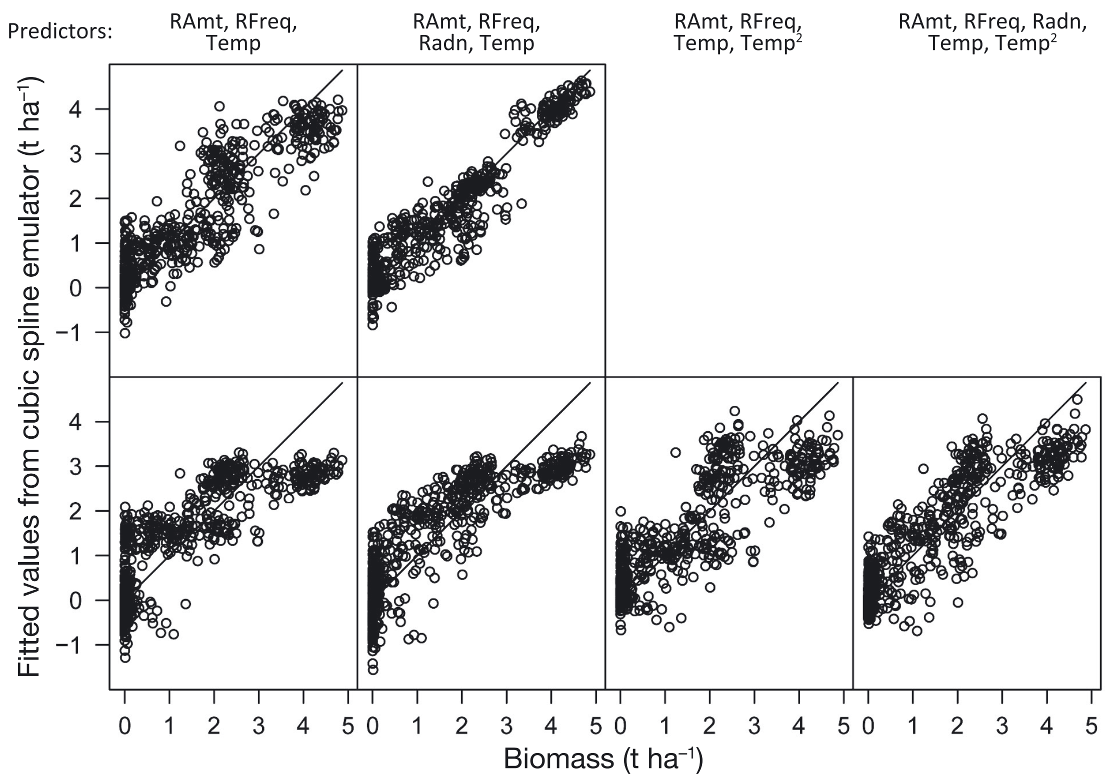

Fig. 1. Comparison of fitted values from the 6 emulators against APSIM-generated 2-monthly lucerne harvested biomass at Badgingarra in sandy soil. $y = x$ line is shown in each panel for reference. RAmt: rainfall amount; Radn: solar radiation; RFreq: rainfall frequency; Temp: temperature

line, as seen in each of the panels, suggests that all of the emulators fit well in certain seasons but not in others. Figs. 2 & 3 and Figs. S1–S4 (in the supplement at www.int-res.com/articles/suppl/c055p253_supp.pdf) compared APSIM-generated biomass and predicted biomass from each of the selected emulator models with the associated rainfall separately for each season.

Examination of the 4 linear emulator graphs (Fig. 2 & S1–S3) showed that, in general, predictions from linear emulators did not fit the simulation data very well in any of the months, particularly in the predominantly wet months (June–September) when there was high biomass production. Emulator fits improved with increasing complexity of the linear emulator (Table 4). This was most noticeable in the drier months from December to March, but was also present in the April–May period. Comparing Figs. S2 & S3 shows that the inclusion of radiation effects appeared to slightly decrease the emulator fit in the drier months from October to May (the predictions of APSIM and the emulators match better in Fig. S2

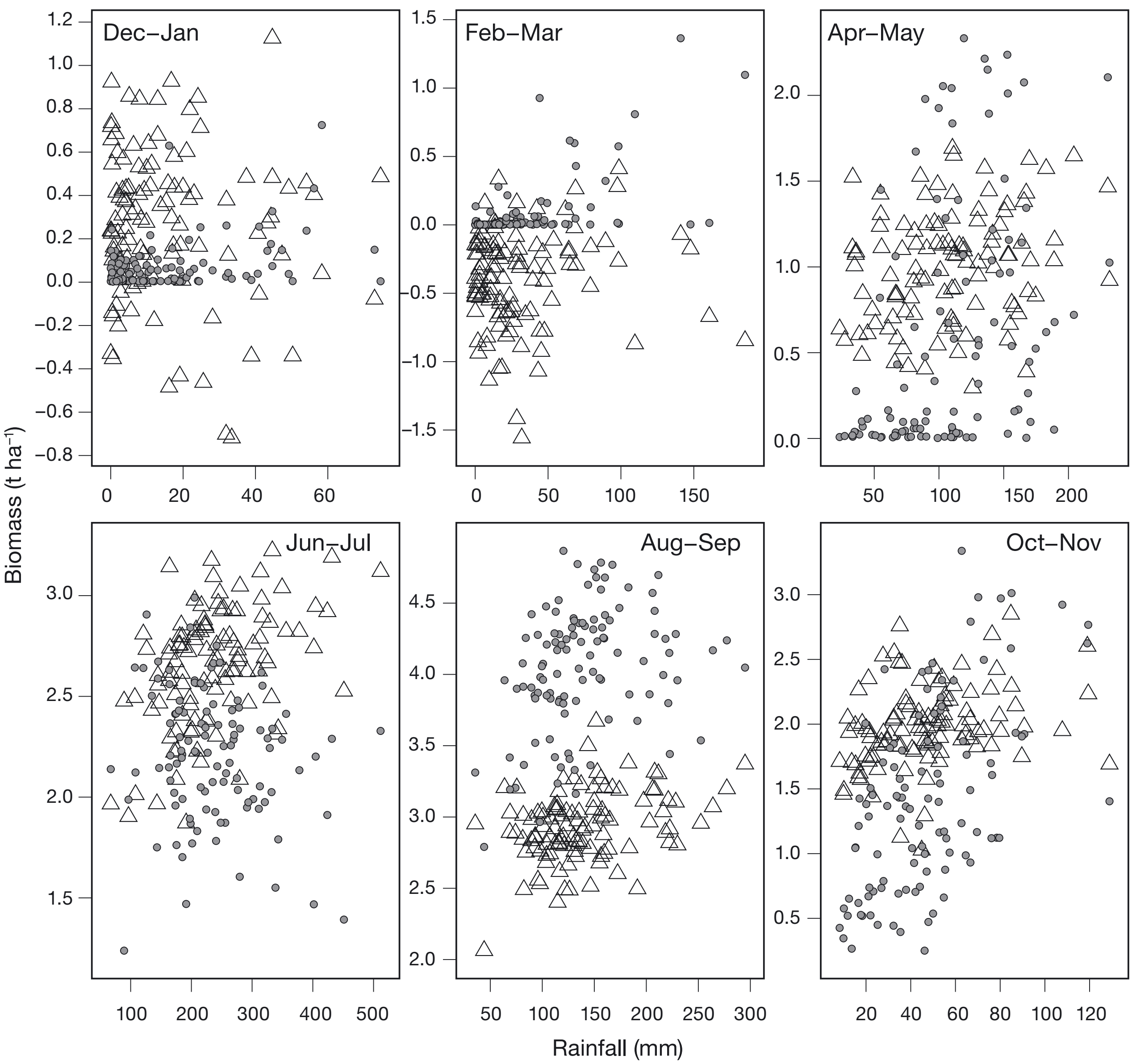

Fig. 2. Comparison of APSIM-generated 2-monthly biomass (actual, ●) and predicted biomass from the linear emulator that included rainfall amount, rainfall frequency, radiation and temperature (predicted, Δ), with associated 2-monthly rainfall for each season

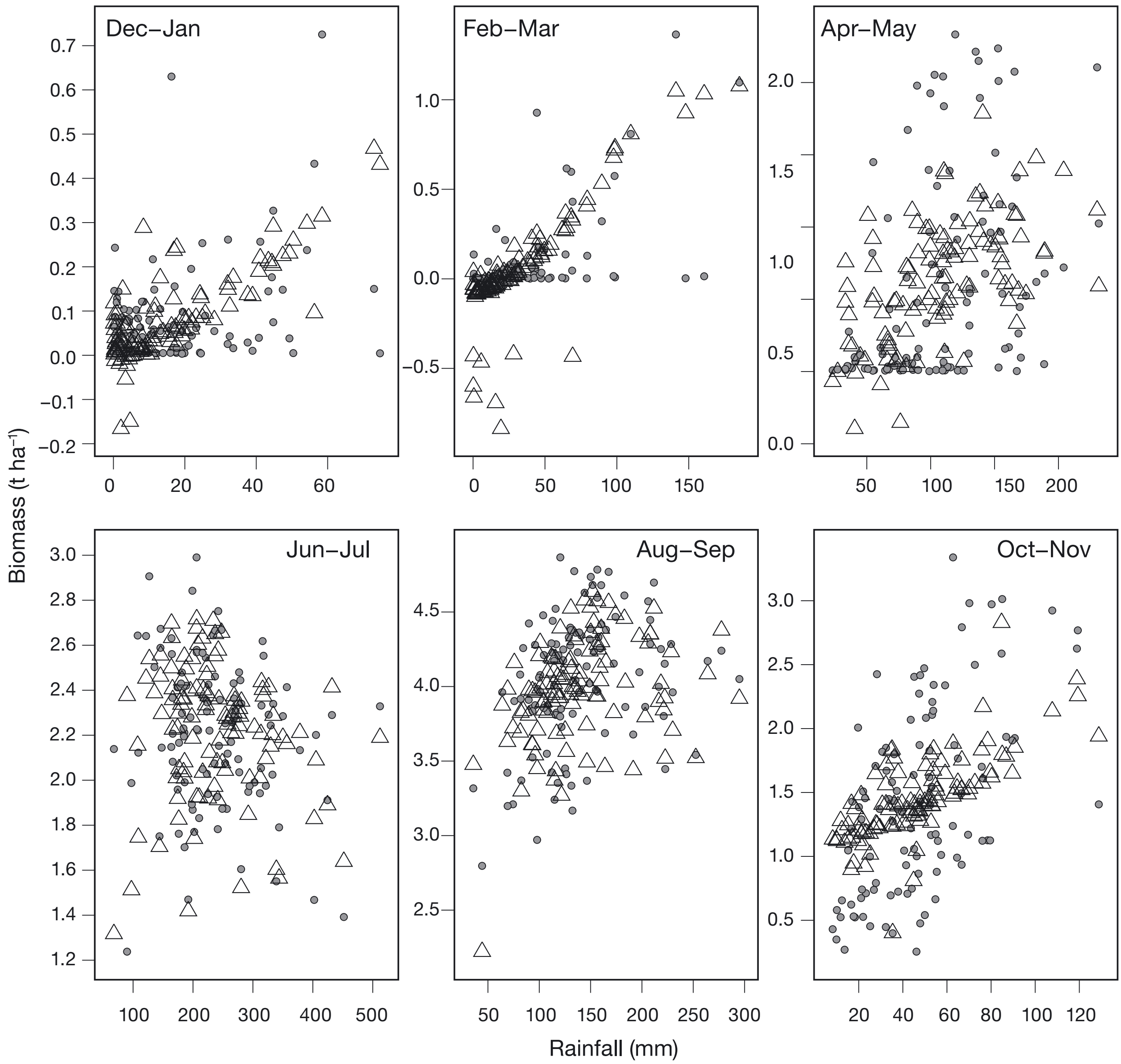

Fig. 3. Comparison of APSIM-generated 2-monthly biomass (actual, ●) and predicted biomass from the cubic spline emulator that included rainfall amount, rainfall frequency, radiation and temperature (predicted, Δ), with associated 2-monthly rainfall for each season

than Fig. S3 in the panels for October, December, February and April).

In contrast, the 2 cubic spline emulator graphs (Figs. 3 & S4) show that overall predictions from these emulators fit the simulation data well. Increasing the complexity of the emulator increased the model fit throughout the year in both wet and dry periods. The notable exception was during October–November, when the inclusion of radiation in the emulator model appeared to decrease the model fit.

4. DISCUSSION AND CONCLUSION

The cubic spline emulator that included RAmt, RFreq, Radn and Temp was the best emulator in terms of both fitting ability and predicting ability. This model provided reasonable fits to existing data as indicated both analytically through MSE and R^2 statistics and graphically through diagnostic plots. This model had good predictive ability as assessed using 2 separate statistical techniques.

The superiority of this model is somewhat surprising in light of the tendency of cubic spline models to overfit, in comparison to simple linear models (Wold 1992). However, the use of 2 methods of model validation and, in particular, the use of the MSPE statistic to provide external validation using a different dataset from that used for model fitting has confirmed that the chosen emulator model did not overfit the data, probably due to the relatively small number of knots used.

Although the emulators had reasonable predictive ability, particularly the cubic spline emulator, their predictions were far from perfect. There are several possible reasons for this. The preliminary analysis given in the technical report (available at www.michaelrenton.info/publications) showed that when artificially generated weather files were used, with rain at regular intervals over a full year, and one weather variable was varied while all other variables were kept constant, then there was a relatively simple relationship between the varying climate variable and the yield predicted by APSIM. This shows that the differences between biomass predictions from the emulators and APSIM evident in this paper (Figs. 2, 3 & S1–S4) are likely due to variability in temperature, radiation and particularly rainfall across the year, both between and within 2-month periods. For example, the consistent overestimation of biomass production in August–September by the linear emulators relative to APSIM (Fig. 2 & S1–S3), and the corresponding over-prediction in April–May and June–July, is likely due to the fact that APSIM accounts for the effect of rainfall that falls before the current 2-month period and is stored in the soil to be used by plants within the current 2-month period. On the other hand, the emulators only based their predictions on rainfall for the current 2-month period, and so could not account for the fact that growth in August–September is more likely to benefit from stored soil moisture than growth in April–May or June–July. The effect of variation in temporal rainfall patterns within 2-monthly periods and their effect on soil water is accounted for in the soil-water model APSIM, where large, less-frequent rain events will have a more beneficial effect on growth than smaller, more-frequent events, that will tend to have more loss to evaporation, or very large and infrequent rain events where the soil profile's water holding capacity is exceeded and water is lost to deep drainage; this is the likely cause of much of the smaller discrepancies between predictions of APSIM and the emulators. This is accounted for in the current emulators with the RFreq variable, but only in a simple and partial way. Other sources of error include limitations in the model form. For example, the linear and spline functions did not constrain production to be non-negative, while APSIM does. Also, the fact that the spline models performed better than the linear models indicates that the linear functional form is too restrictive. It was interesting to note that the inclusion of the extra term radiation in the emulator model appeared to decrease the model fit for both linear and cubic models in particular periods. As the inclusion of an extra term cannot decrease overall model fit, and Table 4 indicates that in most cases the addition of the radiation term actually improved model fit overall, this decrease in model fit for some periods must have been accompanied by a greater increase in model fit for other periods. We note that the periods where fit appeared to decrease were those with low production, where small changes were much more obvious in the plots, and conclude that adding radiation must improve fit in the high production periods at the cost of a small decline in fit in low production periods.

The benefits of the approach to developing statistical emulators for complex plant growth models outlined in the paper are numerous. The ease of fitting and statistically validating the models, as well as the transparency of the process, make it equally usable by non-statisticians and statisticians. The emulator can be used to produce relatively quick but quite accurate estimates of biomass production under different weather conditions. At any point in time, the emulator can be used to predict future biomass production over a period of time based on an average weather scenario, or on recent historical data combined with average weather scenarios for the near future. It can also be used to quickly produce a probability distribution of possible production results over a period of time, based on a collection of historical weather records, in order to quantify uncertainty in a seasonal biomass production forecast. Another advantage is that the simplicity of the statistical emulator that is developed facilitates its incorporation into more complex models that require biomass production estimates, such as models simulating agroecological dynamics of farming systems over a number of years (Lawes & Renton 2010).

There are some limitations in this study, some of which could be addressed in future work. The statistical emulators developed here were designed to predict growth and production over 2-monthly periods, and would thus not be well suited for assessing the impact of short periods (for example daily, weekly or even biweekly) of weather on crop growth. The results in this study show that simply constraining biomass predictions to a minimum of zero would improve the emulators. They also show that accounting for soil water storage across the boundaries of consecutive 2-monthly periods (by including a lag effect for rainfall in some way) would improve the emulators, as would better accounting for variability in patterns of rainfall within 2-month periods; methods to achieve these aims should be developed in future. However, in general, emulators are not good at capturing extreme values in calibration/validation data (Iizumi et al. 2009a), so some discrepancies between emulators and simulation model are always likely to remain. The linear spline emulators examined here did not include any interaction terms due to the fact that interactions in this type of model would have increased the complexity of the models beyond the scope of this study. While the preliminary analysis given in the supplement (www.int-res.com/articles/suppl/c055p253_supp.pdf) showed trends in the data that are consistent with models that contained mixtures of several modelling functions, including non-linear models for RAmt and for Radn, these mixture models were also excluded from this study. Since non-linear modelling methods involve iterative processes to find parameter values, the fitting of parameter values would become more difficult, and the complexity of the resulting mixture model would also be increased. Nonetheless, future work could consider creating more complex emulators such as those described above. Other methods for constructing statistical emulators, such as using Bayesian analysis to create Gaussian process emulators (O'Hagan 2006), or classical modelling methods such as polynomial regression (Kleijnen 2006) along with Fourier models and splines (Barton 1998) could also be considered in future.

The emulator developed in this study was built and tested for its ability to predict biomass production for a particular species, at a single site, on a particular soil type. There are many potential uses for such an emulator, and there is no reason why a number of emulators could not be built in a similar fashion for different species, sites or soil types as required. However, expanding the current approach to enable development of emulators for a wider range of soil types, species or locations would increase its usefulness and power. Further work could also investigate methods for using real data as it becomes available to calibrate models initially created using simulated data, in order to increase emulator prediction accuracy.

We conclude that the approach demonstrated in this study is a powerful means for creating and testing relatively simple emulators of complex growth models that can make similar predictions to the complex models, but with much improved computational efficiency and transparency.

Acknowledgements. We thank Art Diggle for his assistance in hastening the completion of this paper. We thank Future Farm Industries CRC for their financial support.

LITERATURE CITED

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Automat Contr 19:716–723

Allen DM (1974) The relationship between variable selection and data augmentation and a method for prediction. Technometrics 16:125–127

Barton RR (1998) Simulation metamodels. In: Medeiros DJ, Watson EF, Carson JS, Manivannan MS (eds) Proc 1998 Winter Simulation Conf, IEEE Press, Piscataway, NJ, p 167–174

Bastos LS, O'Hagan A (2009) Diagnostics for Gaussian process emulators. Technometrics 51:425–438

Bouman BAM, van Keulen H, van Laar HH, Rabbinge R (1996) The 'School of de Wit' crop growth simulation models: a pedigree and historical overview. Agric Syst 52:171–198

Brisson N, Gary C, Justes E, Roche R, Mary B (2003) An overview of the crop model STICS. Eur J Agron 18: 309–332

Brooks RJ, Semenov MA, Jamieson PD (2001) Simplifying Sirius: sensitivity analysis and development of a metamodel for wheat yield prediction. Eur J Agron 14:43–60

Chambers JM (1992) Linear models. In: Chambers JM, Hastie TJ (eds) Statistical models. Wadsworth & Brooks/Cole, Pacific Grove, CA, p 96–144

Conti S, O'Hagan A (2010) Bayesian emulation of complex multi-output and dynamic computer models. J Statist Plann Inference 140:640–651

Crawley MJ (2007) The R Book. John Wiley & Sons, Chichester

French RJ, Schultz JE (1984) Water use efficiency of wheat in a Mediterranean-type environment. I. The relation between yield, water use and climate. Aust J Agric Res 35: 743–764

Friedman JH (1991) Multivariate adaptive regression splines. Ann Stat 19:1–67

Gibbons JM, Wood ATA, Craigon J, Ramsden SJ, Crout NMJ (2010) Semi-automatic reduction and upscaling of large models: a farm management example. Ecol Model 221:590–598

Gilmour AR, Gogel BJ, Cullis BR, Welham SJ, Thompson R (2007) ASReml user guide release 2.0. VSN Int, Hemel Hempstead

Hastie T, Tibshirani R, Friedman J (eds) (2009) Model assessment and selection. In: The elements of statistical learning. Springer, New York, NY, p 219–260

Holzworth DP, Huth NI, de Voil PG (2010) Simplifying environmental model reuse. Environ Model Softw 25: 269–275

Iglesias A, Rosenzweig C, Pereira D (2000) Agricultural impacts of climate change in Spain: developing tools for a spatial analysis. Glob Environ Change 10:69–80

Iizumi T, Yokozawa M, Nishimori M (2009a) Development of impact functions on regional paddy rice yields in Japan

for integrated impact assessment models. J Agric Meteorol 65:179–190
Iizumi T, Yokozawa M, Nishimori M (2009b) Parameter estimation and uncertainty analysis of a large-scale crop model for paddy rice: application of a Bayesian approach. Agric For Meteorol 149:333–348
Iizumi T, Nishimori M, Yokozawa M (2010) Diagnostics of climate model biases in summer temperature and warm-season insolation for the simulation of regional paddy rice yield in Japan. J Appl Meteorol Climatol 49:574–591
Jeffrey SJ, Carter JO, Moodie KB, Beswick AR (2001) Using spatial interpolation to construct a comprehensive archive of Australian climate data. Environ Model Softw 16:309–330
Jones JW, Hoogenboom G, Porter CH, Boote KJ, Batchelor WD (2003) The DSSAT cropping system model. Eur J Agron 18:235–265
Keating BA, Carberry PS, Hammer GL, Probert ME and others (2003) An overview of APSIM, a model designed for farming systems simulation. Eur J Agron 18:267–288
Kleijnen JPC (2006) White noise assumptions revisited: regression metamodels and experimental designs in practice. In: Perrone LF, Wieland FP, Liu J, Lawson BG, Nicol DM, Fujimoto RM (eds) Proc 38th Conf Winter Simulation, Monterey, CA, p 107–117
Lawes R, Renton M (2010) The Land Use Sequence Optimiser (LUSO): a theoretical framework for analysing crop sequences in response to nitrogen, disease and weed populations. Crop Pasture Sci 61:835–843
Lobell DB, Burke MB (2010) On the use of statistical models to predict crop yield responses to climate change. Agric For Meteorol 150:1443–1452
Mayer DG, Butler DG (1993) Statistical validation. Ecol Model 68:21–32
Moore AD, Donnelly JR, Freer M (1997) GRAZPLAN: decision support systems for Australian grazing enterprises. III. Pasture growth and soil moisture submodels, and the GrassGro DSS. Agric Syst 55:535–582
Muggeo VMR (2008) Segmented: an R package to fit regression models with broken-line relationships. R News 8: 20–25
Neter J, Wasserman W, Kutner MH (1989) Applied linear regression models. Richard D Irwin, Burr Ridge, IL
O'Hagan A (2006) Bayesian analysis of computer code outputs: a tutorial. Reliab Eng Syst Saf 91:1290–1300
Probert ME, Robertson MJ, Poulton PL, Carberry PS, Weston EJ, Lehane KJ (1998) Modelling lucerne growth using APSIM. Proc 9th Aus Agron Conf, Aus Soc Agron
R Development Core Team (2009) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna
Rivals I, Personnaz L (2004) Jacobian conditioning analysis for model validation. Neural Comput 16:401–418
Robertson MJ, Carberry PS, Huth NI, Turpin JE and others (2002) Simulation of growth and development of diverse legume species in APSIM. Aust J Agric Res 53:429–446
Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464
Smith PL (1979) Splines as a useful and convenient statistical tool. Am Stat 33:57–62
Snee RD (1977) Validation of regression models: methods and examples. Technometrics 19:415–428
St Laurent RT, Cook RD (1992) Leverage and superleverage in nonlinear regression. J Am Stat Assoc 87:985–990
St Laurent RT, Cook RD (1993) Leverage, local influence and curvature in nonlinear regression. Biometrika 80: 99–106
Stevenson DE, Feng G, Zhang R, Harris MK (2005) Physiological time model of *Scirpophaga incertulas* (Lepidoptera: Pyralidae) in rice in Guandong Province, People's Republic of China. J Econ Entomol 98:1179–1186
van Ittersum MK, Leffelaar PA, van Keulen H, Kropff MJ, Bastiaans L, Goudriaan J (2003) On approaches and applications of the Wageningen crop models. Eur J Agron 18:201–234
Verbyla AP, Cullis BR, Kenward MG, Welham SJ (1999) The analysis of designed experiments and longitudinal data by using smoothing splines. J R Stat Soc Ser C Appl Stat 48:269–311
Walther BA, Moore JL (2005) The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. Ecography 28:815–829
Wand MP (2000) A comparison of regression spline smoothing procedures. Comput Stat 15:443–462
Wei BC, Hu YQ, Fung WK (1998) Generalized leverage and its applications. Scand J Stat 25:25–37
Wilkinson GN, Rogers CE (1973) Symbolic description of factorial models for analysis of variance. J R Stat Soc Ser C Appl Stat 22:392–399
Wold S (1992) Nonlinear partial least squares modelling. II. Spline inner relation. Chemom Intell Lab Syst 14:71–84

Editorial responsibility: Mikhail Semenov, Harpenden, UK

Submitted: April 25, 2012; Accepted: September 27, 2012
Proofs received from author(s): January 3, 2013