Testing the generalization of artificial neural networks with cross-validation and independent-validation in modelling rice tillering dynamics

Xiangcheng Mi\textsuperscript{a}, Yingbin Zou\textsuperscript{b}, Wei Wei\textsuperscript{a}, Keping Ma\textsuperscript{a,}\textsuperscript{*}

\textsuperscript{a} Laboratory of Quantitative Vegetation Ecology, Institute of Botany, The Chinese Academy of Sciences, 20 Nanxincun, Xiangshan, Beijing 100093, P.R. China

\textsuperscript{b} Rice Research Institute, Hunan Agricultural University, Changsha 410128, P.R. China

Received 1 May 2003; received in revised form 29 April 2004; accepted 4 June 2004

Abstract

Neural networks (NN) rely on the inner structure of available data sets rather than on comprehension of the modeled processes between inputs and outputs. Therefore, neural networks have been regarded as highly empirical models with limited extrapolation capability to situations outside the range of the training and validation data sets. In this study, the generalization ability of neural networks in predicting rice tillering dynamics was tested and several techniques inducing the generalization ability of neural networks were compared. We compared the performance of cross-validated neural networks with independent-validated neural networks and found that neural networks were able to extrapolate and predict tillering dynamics if the data were within the range of inputs of the training set. An inadequate training set resulted in overfitting of available data and neural networks that were not generalizable. The training set size required to enable a neural network to generalize and predict rice tillering dynamics was found to be at least 9 times as many training patterns for each weight. When a large number of variables are included in the input vector of a neural network with inadequate amounts of training data, we strongly recommend that the dimension of the input vector is reduced using principle component analysis (PCA), correspondence analysis (CA) or similar techniques to decrease the number of weights before the training procedure to improve the generalization ability of the NN. If the amount of training data still is not sufficient after the dimension of the input vector is reduced, regularization techniques, such as early stopping, jittering, and especially the embedment of estimated results by a theoretical model into the training set, should be used to improve the generalization ability of the neural network. The generalization of neural networks presents a wide spectrum of problems, and the proposed approaches are not confined strictly to modelling rice tillering dynamics but can be applied to other agricultural and ecological systems.

© 2004 Elsevier B.V. All rights reserved.

Keywords: Neural networks; Model validation; Tillering dynamics; Rice modelling; Generalization ability

* Corresponding author. Tel.: +86 10 62591431x6223; fax: +86 10 62591781.
E-mail address: mkp@brm.ac.cn (K. Ma).

0304-3800/ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2004.06.015
1. Introduction

All approaches to modelling agroecological systems have particular merits and limitations. Algorithmic models, although often very complex, are not able to simulate actual crop growth processes when site factors such as weeds, diseases, insects, tillage, and nutrient availability, are included in the models in addition to climate, soil water availability and soil nitrogen. This is especially true when the relationships between these variables are only vaguely understood. Because uncertainty and spatial and temporal variability are inherent in agroecosystems, artificial neural networks (NN) have been recommended as a suitable modelling approach that can be combined with an implicit systems analysis to generate appropriate nonlinear model structures for these complex systems. Therefore, NNs are receiving increased attention in agroecology as a powerful, flexible, statistical modelling technique for complex, nonlinear and ill-defined agroecosystems (Lek and Guégan, 1999; Schultz and Wieland, 1997; Schultz et al., 2000). NNs have been successful in predicting crop yields and biomass accumulation in a variety of applications (e.g. Sudduth et al., 1996; Momoto and Hashimoto, 1998; Drummond and Joshi, 2000; Schultz and Wieland, 1997; Schultz et al., 2000).

Until recently, NNs were not a particularly favorite technique for modelling crop systems in spite of their simple model construction and parameter estimation procedure. The main disadvantages of NNs are the large data demands required for achieving the desired robust parameter estimations, and their highly empirical nature and consequent inability to extrapolate to untested data sets. This is because NNs rely on the inner structure of the available data sets rather than on comprehension of the modeled processes (Lek and Guégan, 1999; Schultz and Wieland, 1997; Schultz et al., 2000). It was believed that NNs were not able to make accurate predictions outside the range of the training and validation data sets because of overfitting of the available data sets. Therefore, although neural networks greatly outperformed multiple linear regression models (Lamb et al., 1996; Sudduth et al., 1996; Paruelo and Tomasel, 1997), and the accuracy of the results were very similar to algorithmic simulation models (Schultz and Wieland, 1997), Schultz and Wieland (1997) suggested that neural networks could not compete with algorithmic or other sophisticated models. Hence, in most studies, cross-validation of NNs used in agroecological studies was achieved by analysing and predicting situations from the same agroecological systems (e.g. Findlay and Zheng, 1999; Lek and Guégan, 1999; Schultz and Wieland, 1997). However, Broner and Comstock (1997) established a generalizable NN that could generate fertilizer recommendations for malting barley, and the training set size required to enable their NN to generalize was found to be at least 2000 patterns. Drummond and Joshi (2000) found that when a site-year observation was removed from the training data, the predictive ability of the NNs on that particular site-year was dramatically reduced. This result clearly demonstrated that severe overfitting was occurring and their NN had limited extrapolation capability.

In this study, we attempted to test the generalization ability of NNs in predicting rice tillering dynamics by comparing the performance of cross-validated NNs to independent-validated NNs, and compared techniques for improving generalization ability of NN (e.g. early stopping, jittering, NN incorporated with agroecological knowledge). To address this objective, we developed several NNs with cross-validation and independent validation to simulate the rice tillering dynamics for four sites in Hunan Province and Hubei Province of the Yangtze River Valley, southern China. Also, the performance of cross-validated and independent-validated algorithmic models was provided for background and comparison.

2. Materials and methods

2.1. Field experiments

The experiments were conducted on Zhong-100, an early rice variety, at four sites located in Xiangtan, Ningxiang, Huorong of Hunan Province, and Jianli of Hubei Province in the Yangtze River Valley, southern China. The four sites were located in southern China’s double rice belt, where two sequential rice crops are grown annually. Measurement of soil properties on composite soil samples taken from the four sites on 1999 showed that all sites had similar soil pH, organic matter, soil N supply, phosphorus, potassium, boron and zinc. Zhong-100 was transplanted to the four sites using the same spacing and standard management
practices that typically are employed. Adequate management interventions were taken to control pests and weeds at all sites. Sufficient fertilizer of the same dose was applied at all sites, and water stress was negligible because all sites were irrigated.

Rice tillering data (tillers/m²) was collected at 3 days interval after transplanting from the plots at the four sites during 2000 and 2001, and also in 2002 in Ningxiang, producing a total of 141 patterns for tillering dynamics analysis. The dates of sowing, transplanting, flowering and maturity were recorded at the four sites. Total crop biomass, grain and straw yields were determined at maturity. Climate data was available from four weather stations near each of the four sites. Effective degree-days ($T_n = \sum_{i=1}^{n} (T_i - 10)$, if $T_i \geq 10^\circ C$, $T_i$ is daily mean temperature of $i$th day after transplanting, and $T_n$ is effective degree-days of $n$th days after transplanting), and sum of sun hours were included.

2.2. Rice tillering models

Tiller density is an important indicator of high yields in rice production. Tillering is affected by the rice variety and numerous environmental factors, including temperature, light, nutrient supply, space, and cultural practices. One objective of rice cultivation is to optimise the tillering pattern and panicle number to maximize grain yields. Several algorithmic tillering models have been established that analyze and predict the rice tillering dynamics in the Yangtze River Valley (e.g. Jiang, 1982; Huang et al., 1994, 1996; Wang and Huang, 1997).

2.2.1. Tillering models with one input variable

When sufficient fertilizer and water are supplied and seedlings are transplanted at the same spacing density, temperature and solar radiation are the two key factors affecting rice tillering dynamics (Qi, 1986; Zou et al., 1991b). Because temperature and solar radiation are highly correlated, several tillering models have been established that analyze and predict the rice tillering dynamics in the Yangtze River Valley (e.g. Jiang, 1982; Huang et al., 1994, 1996; Wang and Huang, 1997).

We established a standard three-layered back-propagation NNs as nonlinear function approximations between effective degree-days and tillering dynamics, which we called degree-day-tillering NN. The structures of the NN were defined by an a priori guide: to represent and compute associations between patterns in a single hidden layer enabling network, at least 5–10 training patterns were included for each weight (Goh, 1995). The degree-day-tillering NN had a 1-4-1 structure (input-hidden-output nodes) with effective degree-day-tillering dynamics as input-output variables. The mean square error (MSE), i.e. the sum of the square deviations of the NN outputs from the targets values, was used as the main NN performance criterion during training.

To compare the generalization ability of the degree-day-tillering NN with algorithmic models, we also tested the performance of another algorithmic model with one independent variable, the degree-day statistical model (Jiang, 1982):

$$y = a_0 + (a_f - a_0) e^{-c(T' - d)^2}/T'^2$$  \hspace{1cm} (1)

where $y$ is the tiller number per m²; $a_0$ the transplanted tiller number per m²; $a_f$ the maximum tiller number per m²; $T'$ the effective degree-days after transplanting; and parameters $c$ and $d$ vary per cultivar. The degree-day statistical model predicts tillering dynamics on the basis of the effect of degree-days on tillering dynamics.

2.2.2. Tillering models with multiple input variables

To improve the performance of tillering dynamics networks, we used another NN with a 3-4-1 structure with degree-days, sum of sun hours and days after transplanting as inputs. Because additional factors were taken into consideration in this NN, we called it the improved NN.

In the context of the improved NN, the dimension of the input vector was relatively large for the training set, and the components of the vector, degree-days, sum of sun hours and days after transplanting, were highly correlated and redundant. It was useful in this situation to reduce the dimension of the input vectors, which can be achieved by using principal component analysis (PCA), correspondence analysis (CA) and similar techniques. These techniques reduce the dimension of the input vector by replacing a group of variables with a single new variable. For PCA, this new variable is called the principal component (SAS Institute, 1988). PCA has three effects: it orthogonalizes the components of the input vectors so that they are uncorrelated.
with each other; it orders the resulting orthogonal components (principal components) so that those with the largest variation come first; and it eliminates those components that contribute the least to the variation in the data set (Demuth and Beale, 2001). After performing the operation of PCA using SAS’s Princomp procedure (SAS Institute, 1988), the first and the second components were saved and used as the inputs for the NN with a 2-4-1 structure, which we called the reduced NN.

In order to reinforce the evaluation of the importance of training subset size for the generalization ability of NN, we also constructed another NN in which all three components were included into input vector, which we called three-PCA-components NN, and had a 3-4-1 structure.

To strengthen the evaluation of the generalization ability of NN with multiple input variables, we also tested the performance of another algorithm model: the simulation model of tillering dynamics (SMTL) using the software of rice cultivational simulation optimization-decision making system (RCSODS) (Huang et al., 1994, 1996). In SMTL, the tillering process is simulated by a growth equation that considers tillering before elongation and an extinction equation of tillers after elongation (Gao et al., 1992; Huang et al., 1994, 1996). The tillering growth in numbers during the period from transplanting to elongation is simulated by:

\[ TIL_{i} = TIL_{i-1} + \Delta TIL_{i} \]  

\[ \Delta TIL_{i} = k_{TL1} \times TIL_{i-1} \times \left( \frac{TILM - TIL_{i-1}}{TILM} \right) \]  

where \( TIL_{i} \) and \( \Delta TIL_{i} \) are tiller number per m\(^2\) and tiller increment on the \( i \)th day after transplanting; \( k_{TL1} \), a rate of tillering growth modified by fertilizer, temperature, solar radiation and water. The basic form of \( k_{TL1} \) is given by:

\[ k_{TL1} = k_{TL0} \times FN_{i} \times FT_{i} \times FS_{i} \times FW_{i} \]  

In Eq. (5), \( k_{TL0} \) is a rate of tillering growth under optimum conditions, which has a value between 0.25 and 0.35. The variables \( FN_{i} \), \( FT_{i} \), \( FS_{i} \), and \( FW_{i} \) range from 0.0 to 1.0 and are the function of nitrogen fertilizer, temperature (Eq. (10)), solar radiation (Eq. (8)) and water condition, respectively. Here \( FN \) and \( FW \) were taken to be 1.0. \( TILM \) is the maximum number of tillers at elongation stage, which is estimated by the total leaf number of a given variety, transplanting density, leaf number at transplanting and a tillering ratio. The extinction of tillers after elongation can be computed by:

\[ TIL_{i} = TIL_{i-1} - \Delta TIL_{i} \]  

\[ \Delta TIL_{i} = k_{TL2} \times [(1 - RAP_{i}) - (1 - RAP_{i})^{2}] \times TIL_{i-1} \]  

where \( k_{TL2} \) is an extinction rate, taking a value between 0.1 and 0.2. \( RAP_{i} \) refers to a satisfied degree of photosynthetic products needed for a growing tiller and is estimated from photosynthetic process and related processes (Gao et al., 1992; Huang et al., 1994, 1996). The parameter estimations of the three models were made automatically with the training data subset described in Section 2.3 by implementing numerical optimization procedures by means of programming (e.g. flexible polyhedron search, Marsili-Libelli, 1992; Xu, 1996).

2.2.3. Neural networks with regularization techniques

Because more factors affecting tillering dynamics were considered in the improved NN, which was a larger NN than the degree-day-tillering NN, overfitting was a potential problem. In order to improve the generalization ability of the improved NN, we combined it with two regularization techniques, early stopping and jittering, to establish another NN, which we called regularized NN. Jittering, i.e. addition of a small amount of noise to input patterns at each epoch (Gyngy, 1990), was performed during all training procedures (Gaussian noise at \( \sigma = 0.01 \) was used). Another general method used for improving generalization in this study was early stopping. During the training procedure, the validation set error normally decreased during the initial phase of training, as did the training set error. However, when the network began to overfit the training data, the error on the validation set typically began to rise. When the validation error increased for a specified number of iterations, the training was stopped, and the weights and biases at the minimum of the validation error were returned (Demuth and Beale, 2001).

To improve the generalization of the NN, we also combined the improved NN with another strategy, named metamodel by Scardi (2001), to regularize the neural network. First, results of tillering dynamics were
Subsequently, the accumulated temperature satisfying degrees calculated. Accumulated solar radiation satisfying degrees calculated. Then the mixed training subset \((n=216)\) was used to train the NN, and real data (the same from Section 2.3) were used for testing and validation. We called this NN, metamodel, which was also combined with early stopping.

To incorporate the a priori agroecological knowledge into NNs, we used another NN of 3-4-1 structure with three strategies: (1) Previous studies have shown that in Yangtze River Valley the optimal temperature of tillering is 25–30 °C, the upper limit of tillering stopping is 38–40 °C, and the low limit of tillering stopping is 15–16 °C. Rice tiller density increases rapidly with increasing sun hours, and solar radiation is not a limiting factor when daily global radiation is greater than 20 MJ/m² (Huang et al., 1994; Qi, 1986; Yang, 1990). These research results were expressed as temperature satisfying degrees and accumulated solar radiation satisfying degrees in the following equations (Huang et al., 1994):

\[
RF_i = \begin{cases} 
Q_i/20 & Q_i < 20 \\
1.0 & Q_i \geq 20 
\end{cases} 
\]  
\[
FS_i = RF_i^2 
\]  
\[
TF_i = \begin{cases} 
T_i - 15 & 15 < T_i \leq 30 \\
1.0 & 30 \leq T_i \leq 33 \\
4(T_i - 33) & 33 < T_i \leq 40 \\
0 & 40 \leq T_i 
\end{cases} 
\]  
\[
FT_i = \sqrt{TF_i} 
\]

where \(FT_i\) and \(FS_i\) are temperature satisfying degrees and solar radiation satisfying degrees on the \(i\)th day after transplanting. \(TF_i\) and \(RF_i\) are the tillering function of temperature \(T_i\) and daily global solar radiation \(Q_i\) generated from daily sun hours by the meteorological method described by Guo et al. (1992).

Daily temperature and daily global solar radiation were preprocessed using the above equations, and the accumulated temperature satisfying degrees and accumulated solar radiation satisfying degrees calculated. Subsequently, the accumulated temperature satisfying degrees, accumulated solar radiation satisfying degrees and days after transplanting were used as input variables for this NN. (2) We added an additional penalty term in the MSE calculation, which depended on the deviations of the response surface of the tillering dynamics versus surface accumulated temperature satisfying degrees and accumulated solar radiation satisfying degrees based on known agroecological system. It is well known the pattern of rice tillering dynamics is an irregular parabola: before elongation tiller density increases in a sigmoidal-shaped curve until the maximum tiller density is reached; subsequently, the tiller density gradually declines to a stable tiller density after elongation. Thus, there are two minima and one maximum in the tillering dynamics curve and the final tiller density is greater than initial tiller density in any given site-year. These dynamics can be easily found by looking for those nodes where tillering density estimates were larger (or lower) than the two neighbouring nodes in any given site-year, and thus can be checked throughout the whole training procedure to insure that the model is agroecologically meaningful. If more than one maximum or more than two minima were found, then the penalty term was increased by one unit for each maximum or minimum exceeding these limits through any given site-year. The corrected error \(E^*\) was calculated by the following equation (Scardi, 2001):

\[
E^* = MSE + \text{penalty} \times MSE 
\]

(3) Previous studies in Yangtze River Valley have shown that the relative importance of three factors affecting rice tillering was: degree-days ≥ sum of sun hours > days after transplanting, when sufficient fertilizer and water were supplied (Qi, 1986; Yang, 1990; Zou et al., 1991a,b). The relative importance of the three input variables were calculated at each epoch throughout the training procedure using Garson’s algorithm (Garson, 1991) as modified by Goh (1995). If the relative importance of the three inputs did not conform to the known pattern during the training procedure, the penalty term was increased by one unit, and when the relative importance of days after transplanting was greater than that of degree-days, the penalty term was increased by 2 units. We called this NN, knowledge-incorporated NN, which was simultaneously combined with early stopping and jittering.
2.3. Validation of models

We used two strategies to divide the whole data set into three subsets that were used for training, testing and validating all NNs in our study. The first was a cross-validation procedure (Lek and Guégan, 1999). In this procedure, the whole available data set \( (n = 141) \) was divided randomly into two subsets: the training and testing subsets with the proportion \( \approx 3:1 \) (108:33), and these two subsets were combined to obtain a larger validation subset. Cross-validation indicates training and testing of subsets from the same groups of observations, thus it was reasoned that the models would likely overfit these data. A second validation strategy was using the data from Jianli \( (n = 33, \text{in year 2000 and 2001}) \) as testing subset and using data from other three sites as training subset \( (n = 108) \). The Jianli subset was also combined with the training subset as a larger validation subset. The error of the testing subset in this validation was independent of training procedure, and could be used to test the generalization capacity of the models. Hence, this strategy is independent validation. Jianli is in Hubei Province, to the north of Hunan Province and farther away from the other sites in Hunan Province. Thus the data set collected from Jianli is more appropriate for the generalization testing. In fact, all models in this study had the same validation subset. The training and testing subsets of all the cross-validated models were identical, whereas the training and testing subsets of all the independent-validated models were identical. We reported the results with the smallest validation errors. The goal of our data separation strategies were: (1) to choose the result with the smallest validation error for the NN so that both the training subset and testing subset had the smallest errors; (2) to obtain the same validation subset for the same NN with different validation strategies to ensure that the validation error had a similar affect of regularization on the training procedure of both cross-validated NN and independent-validated NN. The differences in the model performance resulted mainly from differences in the training data when early stopping was used to regularize the NN. All models in this study were parameterized on the training subset of tiller density with different input vectors of different models for different validation strategies. The number of adapted model coefficients and the number of fit-parameters used to calibrate the models onto available data were given in Tables 1 and 2.

### Table 1

<table>
<thead>
<tr>
<th>Validation type</th>
<th>Degree-day statistical model</th>
<th>Degree-day-tillering NN</th>
<th>SMTL Improved NN</th>
<th>Reduced NN</th>
<th>Three-PCA-components NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>CV</td>
<td>R(^2) of validation subset</td>
<td>0.7883</td>
<td>0.7768</td>
<td>0.8028</td>
<td>0.7974</td>
</tr>
<tr>
<td>IV</td>
<td>RMSE of validation subset</td>
<td>87.77</td>
<td>88.09</td>
<td>82.50</td>
<td>84.17</td>
</tr>
<tr>
<td>CV</td>
<td>R(^2) of testing subset</td>
<td>0.786</td>
<td>0.851</td>
<td>0.845</td>
<td>0.874</td>
</tr>
<tr>
<td>IV</td>
<td>RMSE of testing subset</td>
<td>89.03</td>
<td>84.50</td>
<td>83.67</td>
<td>79.31</td>
</tr>
<tr>
<td>CV</td>
<td>Number of parameters in model</td>
<td>4</td>
<td>8</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>IV</td>
<td></td>
<td>4</td>
<td>8</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>CV</td>
<td></td>
<td>4</td>
<td>8</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>IV</td>
<td></td>
<td>4</td>
<td>8</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>IV</td>
<td></td>
<td>4</td>
<td>8</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

CV: Cross-validation; IV: Independent-validation.
Table 2

<table>
<thead>
<tr>
<th>Days after transplanting</th>
<th>Sum of sun hours</th>
<th>Effective degree-days</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sum of sun hours</td>
<td>0.8780</td>
<td>0.9519</td>
</tr>
<tr>
<td>Effective degree-days</td>
<td>0.9519</td>
<td>0.8831</td>
</tr>
</tbody>
</table>

The error back-propagation algorithm was used for all NNs with Neural Network Toolbox 4.0 of MATLAB 6.1 (Demuth and Beale, 2001). It is not appropriate to use an algorithm that converges too rapidly with early stopping, so we used the resilient back-propagation algorithm that usually works well with early stopping for all NNs (Demuth and Beale, 2001). Before using the data set for NN training, normalization of all data sets was typically done to keep the connection weights from becoming too large and swamping the net during training (Eberhart and Dobbins, 1990). We normalized the inputs and targets so that they had zero mean and unity standard deviation, and also the output was trained to produce outputs with zero mean and unity standard deviation. We then converted the network output back into the original units to calculate root mean square error (RMSE) between the output values of models from the target values for three subsets, and $R^2$ between estimated and observed values of the testing subset. Although $R^2$ is an easy and widely applied model performance characteristic, comparing $R^2$ values over different systems and different variables is a dangerous exercise and not very robust over different systems and time periods (Janssen and Heuberger, 1995; Van Wijk et al., 2002). Thus, the comparison of performance between models was mainly based on the comparison of the RMSE between estimated and observed values of different subsets in our study.

3. Results

3.1. Model performance of tillering model with one input variable

The degree-day-tillering NN had a higher performance than the degree-day statistical model (Table 1). Expressed in RMSE of the validation subset, the difference between the models was clear: the performance of cross-validated degree-day-tillering NN and independent-validated degree-day-tillering NN were 6 and 4.45%, respectively, better than the degree-day statistical model with the same validation strategy.

Two cross-validated models had a similar performance to the two independent-validated models (Table 1, Fig. 1), showing that the two models could not only effectively interpolate the available information but also make fairly accurate extrapolations within the range of inputs. In Fig. 1, the response curves of the optimised cross-validated models and optimised independent-validated models were almost identical, and the shape of response curves corresponded to the a priori agroecological knowledge of rice tillering dynamics. RMSE of the testing subset for the independent-validated degree-day statistical model increased by 15.60% over the cross-validated degree-day statistical model, while the RMSE of the testing subset for independent-validated degree-day-tillering NN increased by 20.19% over the cross-validated degree-day-tillering NN. These results clearly demonstrated that slight overfitting was occurring.

In the two models, only one major factor, degree-days, was considered in the degree-day statistical model and degree-day-tillering NN. Simplifying the model structure meant following rather general principles and average behavior patterns (Schultz et al., 2000), which was clearly demonstrated in Fig. 1: all response curves of the optimised models had average behavior patterns. Both independent-validated models could predict the rice tillering dynamics in Jianli by learning the knowledge of rice tillering dynamics from the other three sites. These results indicated that the degree-day-tillering NN had the same power to generalize as the degree-day statistical model in predicting tillering dynamics within the range of inputs of the training subset to some extent. On the other hand, the amount of training data for the degree-day-tillering NN was sufficient to obtain a generalized NN.

3.2. Performance of tillering models with multiple input variables

The SMTL in RCSODS is an algorithmic model based on a large number of dependent agroecological processes of tillering dynamics. Before elongation, the SMTL included the effects of daily temperature and daily global solar radiation on tillering growing, and it considers relationships between photosynthetic production and tillering extinction after elongation.
Fig. 1. Response curves of tillering dynamics (tillers/m²) vs. degree-days estimated by optimised cross-validated models or optimised independent-validated models: (a) from optimised cross-validated degree-day-tillering NN; (b) from optimised independent-validated degree-day-tillering NN; (c) from optimised cross-validated degree-day statistical model; (d) from optimised independent-validated degree-day statistical model. The values of the training data points and the testing data points were from real data.

Compared with the degree-day statistical model, the STML was improved by modelling more factors and processes that affect tillering dynamics. Only two adapted model coefficients were used to calibrate the model to the available data. Both the cross-validated and independent-validated STML provided better performance than the degree-day statistical model using either validation strategy, although slight overfitting occurred in independent-validated STML (Table 1). The slight overfitting of the STML was of course not surprising: this resulted from the effects of site-specific conditions on the tillering dynamics, in other words, from other factors affecting tillering dynamics. These factors, such as fertilizer and water stress, that have a minor influence on tillering dynamics, were deliberately reduced in this study, and major factors and processes that affect tillering dynamics were included. Simplifying the model structure means following rather general principles and average behavior patterns (Schultz et al., 2000). The cross-validated STML could learn the effects of all the simulated site-specific conditions from the training subset of the four sites, and the estimated dynamics followed the average behavior patterns of the four sites; however, the independent-validated STML could not learn the effects of the site-specific conditions on tillering dynamics in Jianli.
from the Jianli subset. The estimated tillering dynamics in Jianli followed the average behavior patterns of the other three sites. Therefore, the slight overfitting of available data for a comprehensive model was unavoidable when not all the factors affecting tillering dynamics could be taken into consideration by the model.

More factors affecting tillering dynamics were included in the improved NN and, at the same time, had a relatively larger structure and more fitting parameters than the degree-day-tillering NN (Table 1). Compared to the cross-validated degree-day-tillering NN, the performance of the cross-validated improved NN was improved dramatically: the performance of the cross-validated improved NN expressed in RMSE of the validation subset was 16.96% better than the degree-day-tillering NN (Table 1). On the other hand, the overfitting of the NN also dramatically increased: RMSE of the testing subset for the independent-validated improved NN increased by 79.56% over that for the cross-validated improved NN. The sharp increase in the RMSE of the testing subset indicated that severe overfitting was occurring. The improved performance of the cross-validated improved NN resulted from both taking more factors into consideration and from doubling the number of fitting parameters from 8 in the degree-day-tillering NN to 16. The severe overfitting of the cross-validated and independent-validated improved NN also was clearly indicated in Fig. 2a and b: the response surfaces in the two figures were devoid of any agroecological meaning as tillering dynamics seemed to be related only to degree-days. The only exception was a small region of the response surface where a low to high range of sum of sun hours made the response surface of the tillering dynamics like a paraboloid (Fig. 2a). The relatively small training subset relative to the number of adapted parameters in the NN resulted in overfitting of the training data yielding networks with little generalization ability. Therefore, it was possible for some cross-validated NNs with inadequate amounts of training data to be overtrained NNs even though they were perfectly cross-validated, as was the case of the cross-validated improved NN in this study.

The dimension of the input vector of the improved NN was relatively large for the training subset, and the three inputs were highly correlated and redundant (Table 2). Therefore, it was helpful to use PCA to reduce the dimension of the input vector and to orthogonalize the components of input vectors. The PCA operation was performed on the input vectors of both the training subset and testing subset (Table 3). The results of the PCA showed that the three input variables had almost an identical effect on the first component, and the sum of sun hours had greater effect on the second component (Table 3). We discarded the third component that contained the least amount of information. The reduced NN was smaller than the improved NN as only the first and second components were included as inputs. Compared to the cross-validated improved NN, the cross-validated reduced NN was similar in terms of the RMSE of the different subsets (Table 1), and the independent-validated reduced NN was similar to the independent-validated SMTL when expressed in terms of the RMSE of the testing subsets (Table 1). When expressed as a response surface of tillering dynamics versus degree-days and sum of sun hours (Fig. 2c and d), the reduced NN provided better results, not only in terms of the RMSE, but also from a biological viewpoint.

It was reasoned that the generalization ability of the reduced NN was derived from both the reduced input vector and orthogonalization of input variables. To test this assumption, we tested the performance of the three-PCA-components NN. Including all three components in the input vector with three input neurons, the cross-validated three-PCA-components NN had a better performance than the cross-validated reduced NN. There were more fitting-parameters in three-PCA-components NN (Table 1) and one more factor

### Table 3

<table>
<thead>
<tr>
<th>Eigenvectors</th>
<th>Eigenvalues</th>
<th>Proportion (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Days after transplanting</td>
<td>0.5819</td>
<td>2.8091</td>
</tr>
<tr>
<td>Sum of sun hours</td>
<td>0.5870</td>
<td>93.64</td>
</tr>
<tr>
<td>Effective degree-days</td>
<td>0.5830</td>
<td>93.64</td>
</tr>
<tr>
<td>Component 1</td>
<td>0.5819</td>
<td>2.8091</td>
</tr>
<tr>
<td>Component 2</td>
<td>0.5630</td>
<td>93.64</td>
</tr>
<tr>
<td>Component 3</td>
<td>0.5830</td>
<td>93.64</td>
</tr>
</tbody>
</table>
Fig. 2. Tillering dynamics vs. degree-days and sum of sun hours at 10 days after transplanting. The response surfaces were obtained by cross-validated NNs and independent-validated NNs: (a) from cross-validated improved NN; (b) from independent-validated improved NN; (c) from cross-validated reduced NN; (d) from independent-validated reduced NN.
than in the cross-validated reduced NN in the three-PCA-components NN, but the independent-validated three-PCA-components NN had much more severe overfitting than independent-validated reduced NN. The RMSE of the testing subset of the independent-validated three-PCA-components NN was 81.81% greater than the RMSE of the testing subset of the cross-validated three-PCA-components NN, so the overfitting of the three-PCA-components NN was similar to the improved NN. By comparing the performance of the three-PCA-components NN and reduced NN in our study, we easily reach the conclusion that the relatively large training data set was the key to the generalization ability of NNs, and the orthogonalization of components of the input vector improved the training procedure.

3.3. Performance of neural networks with regularization techniques

Jittering regularizes NN by providing a virtually unlimited number of artificial training patterns that are similar to, but not exactly the same as, the original one, to ensure the NN avoids local optimization. Early stopping correlates the training procedure with errors of the validation subset. When the validation error increases for a specified number of iterations, the training procedure is stopped to avoid overfitting of training data. Early stopping and jittering were combined into the regularized NN. The performance of the cross-validated regularized NN was similar to the performance of the independent-validated regularized NN in terms of the RMSE of the three subsets (Table 4). The three subsets for both validations had almost identical values of RMSE, which indicated that jittering and early stopping could effectively regularize NN and improve its generalization. But, at the same time, the predictive and training accuracy sharply decreased for both the cross-validated and independent-validated regularized NN, and they had almost similar performances as the degree-day-tillering NN. The improved generalization could also be seen in Fig. 3a and b: the resulting tillering dynamics response surfaces of the cross-validated and independent-validated regularized NN were more ecologically meaningful and the response surfaces of tillering dynamics were positively related to both degree-days and sum of sun hours as compared to Fig. 2a and b. Fig. 3b was slightly overfitted and more generalized than Fig. 3a. The difference between the training and testing subsets of the cross-validated regularized NN was smaller because both of the subsets for the cross-validated NN originated from all four sites, while the difference between the training and testing subsets of the independent-validated regularized NN was greater because the two subsets came from different sites. Thus, before the validation error increased up to a specified number of iterations and the training was stopped, the cross-validated regularized NN was trained with more epochs than the independent-validated regularized NN, which resulted in a lower RMSE for both the training and testing subsets of the cross-validated regularized NN. Thus, the relatively small training subset in the cross-validated improved NN yielded networks with little generalization ability. Standard regularization techniques, jittering and early stopping, could be used to improve the generalization of the NN, but the price for a generalized NN was a decrease in the predictive accuracy of the models.

For the metamodel, both real data and modeled data (from the SMTL) were used to train the NN to help the NN correctly extrapolate (Scardi, 2001). As

<table>
<thead>
<tr>
<th>Validation type</th>
<th>Regularized NN</th>
<th>Metamodel</th>
<th>Knowledge-incorporated NN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CV</td>
<td>IV</td>
<td>CV</td>
</tr>
<tr>
<td>$R^2$ between estimated and observed values of validation subset</td>
<td>0.809</td>
<td>0.806</td>
<td>0.815</td>
</tr>
<tr>
<td>RMSE of validation subset</td>
<td>82.59</td>
<td>83.81</td>
<td>79.74</td>
</tr>
<tr>
<td>RMSE of testing subset</td>
<td>80.54</td>
<td>82.84</td>
<td>74.29</td>
</tr>
<tr>
<td>RMSE of training subset</td>
<td>83.21</td>
<td>84.10</td>
<td>81.33</td>
</tr>
<tr>
<td>Number of parameters in model</td>
<td>16</td>
<td>16</td>
<td>16</td>
</tr>
</tbody>
</table>

CV: Cross-validation; IV: Independent-validation.
Fig. 3. Tillering dynamics vs. degree-days and sum of sun hours or accumulated degree-days satisfying degree (ADDSD) and accumulated solar radiation satisfying degree (ASRSD), at 10 days after transplanting. The response surfaces were obtained by cross-validated NNs and independent-validated NNs: (a) from cross-validated regularized NN; (b) from independent-validated regularized NN; (c) from cross-validated metamodel; (d) from independent-validated metamodel; (e) from cross-validated knowledge-incorporated NN; (f) from independent-validated knowledge-incorporated NN.
seen in Table 4, the performance of the cross-validated metamodel expressed in RMSE of the three subsets decreased when compared to the cross-validated improved NN. However, the metamodel had greater generalization ability and was less overfitted than the improved NN, as indicated by the difference in the RMSE between the testing subsets of independent-validated improved NN and independent-validated metamodel and by comparing Figs. 2a and 3c. The performance of the independent-validated metamodel was better than the independent-validated improved NN and degree-day statistical model (Table 1, Table 4) but similar to the independent-validated SMTL. It acted as a hybrid whose fitness with respect to its ‘knowledge environment’ was better than any one of its parents (Scardi, 2001). Its better generalization originated from the SMTL, and its better performance stemmed from more factors being taken into consideration and more fitting parameters in the metamodel than the degree-day-tillering NN.

For the knowledge-incorporated NN, two input variables were preprocessed based on agroecological knowledge, and a constrained training was performed by applying the error computation penalty term, which depended on the deviation from the above mentioned relative importance of the three inputs and the shape constraints of tillering dynamics curves. Scardi (2001) showed an example of the constrained training results with respect to early stopping and jittering, and, in his example, the constrained approach provided better results than the conventional training procedure both in terms of the mean square error, and also because it was more meaningful from a biological viewpoint. In our study, the cross-validated knowledge-incorporated NN had a similar performance as the improved NN and a better performance than the cross-validated regularized NN and metamodel in terms of the RMSE of the different subsets and the $R^2$ of the validation subset (Table 4), even when regularized by early stopping and jittering. However, the RMSE of the testing subsets for the independent-validated knowledge-incorporated NN sharply increased to 41.84% over the cross-validated knowledge-incorporated NN. Thus, the generalization ability of the knowledge-incorporated NN was much worse than the regularized NN and metamodel, which indicated that severe overfitting was occurring in the knowledge-incorporated NN. The severe overfitting could be seen in Fig. 3e and f, in which the overfitted response surfaces was only partly correlated with the accumulated temperature satisfying degree and accumulated solar radiation satisfying degree. The performance of the knowledge-incorporated NN showed that incorporation of agroecological knowledge could improve the training procedure but make the NN converge too rapidly. Thus, early stopping and jittering did not work well in this study. It is possible to adjust the penalty term so that the knowledge-incorporated NN could converge at an appropriate speed and work well with early stopping and jittering to produce better results both in performance and generalization of the NN. The NN incorporated with agroecological knowledge greatly improved the training procedure, but could not induce a better generalization ability. The generalization ability of the knowledge-incorporated NN would have been worse than the improved NN without jittering and early stopping.

4. Discussion

In this study, NNs provided better performance than the algorithm model mainly because of the greater number of adapted parameters (Table 1) (Van Wijk et al., 2002). The low number of fit-parameters used in algorithm models can be seen as an advantage because they do not need a lot of extra data for parameterization thus increasing the application possibilities (Van Wijk et al., 2002). When considering the independent nature of the parameterization of algorithm models, the performance of the algorithm models is very good (Van Wijk et al., 2002). In contrast, NNs need a larger amount of data to get generalized training results and to validate trained networks.

When the number of input variable increased from one variable in the degree-day-tillering NN to three variables in the improved NN and three-PCA-components NN, the input-output relationship changed from a two-dimensional response curve to a four-dimensional response surface. Neural networks follow the average behavior patterns of training data. For the degree-day-tillering NN, the points of training data in the two-dimensional space were relatively dense. The average behavior of these relatively dense points could represent the general behavior of an agroecological system and guarantee the generalization ability of the NN to extrapolate to other systems. For the improved NN and three-PCA-components NN, the points of train-
ing data in the four-dimensional space were relatively sparse. In some areas of the response surface of the neural network, the average behavior of these few points could not represent the general behavior of an agroecological system, and overfitting of available data occurred in the training procedure of neural networks. Thus, sufficient training data is the key for good generalization ability of neural networks.

In the context of predicting rice tillering dynamics, NNs have the ability to generalize within the range of inputs of the training set. The generalization ability of the NNs was clearly demonstrated by comparing the performance of the degree-day-tilling NN, regularized NN, metamodel, reduced NN and the three-PCA-components NN with cross-validation and independent validation in this study. An inadequate amount of training data resulted in overfitting of available data and NNs with little generalization ability. The key to insure the generalization ability of NNs is to have adequate amounts of data in the training set. There were 7 times as many training patterns for each weight in the training set of the improved NN and three-PCA-components NN that were not generalizable NNs, whereas there were 9 times as many training patterns for each weight in the training set of the reduced NN and 13 times as many training patterns for each weight in the training set of degree-day-tillering NN, both of which were generalizable NNs. Therefore, in the context of this study, it is essential that the input data have at least 9 times as many training patterns for each weight in the training set to get the desired generalization ability of NN. It is extremely time-consuming and labor-intensive for complicated networks to collect such large data sets from crop growth experiments involving countless small plot trials over multiple site-years. The application of NNs might be hindered by the large amount of data needed to acquire a robust NN that allows extrapolation.

When a large number of variables are included in the NN but with inadequate amounts of training data, we strongly recommended that the input vector dimension is reduced by using PCA, correspondence analysis or similar techniques to decrease the number of weights before the training procedure. If the amount of training data is still insufficient after reducing the dimension of input vectors, regularization techniques, such as early stopping, jittering, and especially the embedment of estimated results from a theoretical model into training set, should be used to improve the generalization ability of the NN.

Acting as nonlinear function approximations, NNs have been highly successful in a variety of applications, such as predicting algal bloom dynamics (Lee et al., 2003), predicting forest characteristics (Mossen and Frescino, 2002), predicting aquatic insect species richness (Park et al., 2003), modelling phytoplankton primary production (Scarlet, 2001), and forecasting concentration levels of pollutants (Viotti et al., 2002) and so on. The generalization of neural networks presents a wide spectrum of problems. Broner and Comstock (1997) found that much of the neural network research published in the field of agriculture had inadequate amounts of training and testing data. It is likely that some cross-validated NN research published in the field of ecology and agriculture had inadequate amounts of training data and suffered from overfitting, even if these NNs were perfectly cross-validated, as we demonstrated in the cross-validated improved NN and three-PCA-components NN in this study. The accuracy and usefulness of NNs and the contribution of input variables in the modelling process using methods reviewed by Olden and Jackson (2002) should be interpreted cautiously when there is not sufficient data provided on the parameterization process.

Acknowledgements

We thank Liangzhi Gao for his valuable help in parameterising submodels of tillering dynamics in RC-SODS and for providing some of the original Visual Basic code of RCSODS for this study. We thank Shichiu Liang, Qibing Zhang, Xiaogang Xiong and two anonymous reviewers for their valuable suggestions and comments. Special thanks to Jean-Luc Demuzieux and Ruth E. Sherman for improving the English in our manuscript. This research was financed by National Key Research and Planning Project (G2000046802), Key Innovation Project of CAS (KZCX1-10-05) and National Rice Project (95-001-01-03), part of the “National Ninth Five-Year-Plan”.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi: 10.1016/j.ecolmodel.2004.06.035.
References