

## A knowledge-and-data-driven modeling approach for simulating plant growth: A case study on tomato growth



Xing-Rong Fan<sup>a</sup>, Meng-Zhen Kang<sup>b,\*</sup>, Ep Heuvelink<sup>c</sup>, Philippe de Reffye<sup>d</sup>, Bao-Gang Hu<sup>a</sup>

<sup>a</sup> National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

<sup>b</sup> State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

<sup>c</sup> Horticulture and Product Physiology Group, Wageningen University, PO Box 630, 6700 AP Wageningen, The Netherlands

<sup>d</sup> Cirad-Amis, TA 4001, Ave Agropolis, 34398 Montpellier Cedex 5, France

### ARTICLE INFO

#### Article history:

Received 24 December 2014

Received in revised form 1 June 2015

Accepted 3 June 2015

Available online 25 June 2015

#### Keywords:

Data-driven model

Knowledge-driven model

GreenLab

Knowledge-and-data-driven model

Model integration

Plant growth modeling

### ABSTRACT

This paper proposes a novel knowledge-and-data-driven modeling (KDDM) approach for simulating plant growth that consists of two submodels. One submodel is derived from all available domain knowledge, including all known relationships from physically based or mechanistic models; the other is constructed solely from data without using any domain knowledge. In this work, a GreenLab model was adopted as the knowledge-driven (KD) submodel and the radial basis function network (RBFN) as the data-driven (DD) submodel. A tomato crop was taken as a case study on plant growth modeling. Tomato growth data sets from twelve greenhouse experiments over five years were used to calibrate and test the model. In comparison with the existing knowledge-driven model (KDM, BIC = 1215.67) and data-driven model (DDM, BIC = 1150.86), the proposed KDDM approach (BIC = 1144.36) presented several benefits in predicting tomato yields. In particular, the KDDM approach is able to provide strong predictions of yields from different types of organs, including leaves, stems, and fruits, even when observational data on the organs are unavailable. The case study confirms that the KDDM approach inherits advantages from both the KDM and DDM approaches. Two cases of superposition and composition coupling operators in the KDDM approach are also discussed.

© 2015 Elsevier B.V. All rights reserved.

## 1. Introduction

Plants, like other bio-systems, are highly complex and dynamic systems. Modeling plant growth dynamics is a great challenge for scientists in all related fields who are progressively improving models and generating new ones for a vast variety of applications. Modeling approaches vary in a number of aspects (e.g., the scale of interest, the level of description, the integration of environmental stresses, etc.). With respect to the degree of domain knowledge (e.g., basic physical, chemical and biological principles), Todorovski and Džeroski (2006) and Atanasova et al. (2008) considered two basic modeling approaches, namely, “knowledge-driven” and “data-driven” modeling. The knowledge-driven modeling approach relies mainly on the given domain

knowledge. In contrast, the data-driven modeling approach is capable of formulating a model solely from the given data without using any domain knowledge.

In general, a model that can learn from data without using any domain knowledge is called a data-driven model (DDM), for example, artificial neural networks (Recknagel, 2001; Daniel et al., 2008), support vector machines (Pouteau et al., 2012), fuzzy methods (Gutiérrez-Estrada et al., 2013), generalized linear models and generalized additive models (Guisan et al., 2002; Zhang et al., 2005). The DDM also includes the radial basis function network (RBFN), which is one of the most popular neural network models and widely used for function approximation, time series prediction, and nonlinear regression (Buhmann, 2003). Among these methods, they have many desirable characteristics, such as imposing fewer restrictions on assumptions, the ability to approximate nonlinear functions, strong predictive abilities, and the flexibility to adapt to the inputs of a multivariate system. However, data-driven models (DDMs) encounter difficulties in retaining the

\* Corresponding author. Tel.: +86 10 82544502; fax: +86 10 82615087.

E-mail address: [mengzhen.kang@ia.ac.cn](mailto:mengzhen.kang@ia.ac.cn) (M.-Z. Kang).

physical explanations or structural knowledge of a physical system because they are usually considered black-box models, and their parameters do not generally represent physical parameters in a physical system. Hence, DDMs are also called “non-parametric models”.

A model that is derived from domain knowledge is called a knowledge-driven model (KDM), also known as a physically based (Solomatine and Ostfeld, 2008) or mechanistic model (Todorovski and Džeroski, 2006); For plants, knowledge-driven models (KDMs) include process-based models (PBMs) (Vos et al., 2007; de Reffye et al., 2009). Early PBMs for plant growth concerned plant functioning in relation to environmental conditions, especially biomass production and its partitioning. More recently, a new generation of PBMs, often known as functional–structural plant models (FSPMs), has emerged, which incorporates previously neglected aspects, such as the interactions among plant structure (e.g., shape and orientation of organs), the function of organs (e.g., leaf photosynthesis), and the environment (e.g., light) and the feedback between biomass acquisition and its allocation for both plant development and growth. To date, FSPMs have been regarded as potential tools for predicting and simulating plant growth and structural development (Renton, 2013).

The GreenLab model is a generic, mechanistic functional–structural plant model (FSPM), integrating the knowledge of the underlying processes of plant architecture and physiological functioning. The model, in its discrete version, was introduced by de Reffye and Hu (2003) and was studied in the case of tomato crops by Dong et al. (2008) and Kang et al. (2011); its key advantage over other plant models, which are commonly limited to simulation, is its parametric identification (Christophe et al., 2008). Because of the mathematical formalism of the GreenLab model, all model parameters can be identified using inverse methods from measurement data (Zhan et al., 2003; Guo et al., 2006). Although KDMs integrate domain knowledge in modeling and contain physically interpretable parameters, they often have poor predictive ability and do not deal with situations associated with adding and/or missing variables or data. For example, the GreenLab model cannot be effectively applied in a wide range of environmental conditions in which one environmental variable (e.g., solar radiation, temperature) is missing and does not take certain environmental data (e.g., the carbon dioxide concentration, the planting date and the weight at planting date) into account even when these data are available.

To take advantage of both the KDM and DDM approaches, studies on integrating these two types of modeling approaches have been conducted (Džeroski and Todorovski, 2003; Hu et al., 2009; Qu and Hu, 2011; Czop et al., 2011; Ran and Hu, 2014). Investigations on the successful application of this integrated approach especially deserve greater attention in the ecological sciences (Todorovski and Džeroski, 2006; Atanasova et al., 2008; Qu and Hu, 2009; Matsunaga et al., 2013). Among these methods, grammar or rules constructed by domain-specific knowledge were embedded into the DDM to select a candidate model that fits the data best. Unlike the above methods, our main interest is to propose a novel knowledge-and-data-driven modeling (KDDM) approach for simulating plant growth that integrates the knowledge-driven theoretical approach to modeling with the data-driven modeling. A tomato crop was taken as a case study on plant growth modeling. The GreenLab model was adopted as the knowledge-driven submodel and the radial basis function network (RBFN) as the data-driven submodel. The two types of submodels were integrated using a two-way coupling connection. Next, two versions of the KDDM based on the superposition and composition coupling operators were developed. Finally, the validity and usefulness of the KDDM approach in application of modeling the dynamics of plant growth processes from real data sets were illustrated.

## 2. Materials and methods

### 2.1. Models

#### 2.1.1. Radial basis function network (data-driven model)

Radial basis function (RBF) networks typically have three layers: an input layer, a hidden layer with a non-linear RBF activation function and a linear output layer. The input can be modeled as a vector of real numbers  $\mathbf{x} \in \mathbb{R}^n$ . The output of the network is then a scalar function of the input vector,  $\mathbf{f}_d$ , is given by Eq. (1):

$$\hat{y} = \mathbf{f}_d(\mathbf{x}, \theta_d) = \Phi(\mathbf{x})\theta_d, \quad (1)$$

where  $d$  is the subscript associated to the DDM (i.e., RBFN),  $\hat{y}$  is the output of the RBFN,  $\theta_d = [w_1, \dots, w_h]^T$  represent the weights of the network,  $h$  is the number of neurons in the hidden layer, and  $\Phi(\mathbf{x}) = [\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \dots, \phi_h(\mathbf{x})]$ . In this work, the multiquadric RBF  $\phi_j(\mathbf{x}) = \sqrt{1 + ||\mathbf{x} - \mathbf{c}_j||^2/\sigma_j^2}, j = 1, 2, \dots, h$ , was taken as the RBF activation function, where  $\mathbf{c}_j \in \mathbb{R}^n$  are the RBF centers and  $\sigma_j$  controls the width of the RBF.

#### 2.1.2. GreenLab model (knowledge-driven model)

The GreenLab model is a generic functional–structural plant model simulating the dynamics of plant organogenesis, biomass production and allocation (Yan et al., 2004; Guo et al., 2006; Dong et al., 2008; Kang et al., 2011). At each time interval, called growth cycle (GC), the complete formulation of biomass production of a plant,  $Q(i)$ , is given by Eq. (2):

$$Q(i) = E(i) \cdot r \cdot S_p \left( 1 - \exp \left( -\frac{1}{S_p \cdot slw} \sum_{j=1}^{\min(i, t_a)} [N_b(i-j+1) \cdot \left( \sum_{k=1}^{\min(j, t_x^b)} P_b(k) \frac{Q(i-j+k-1)}{D(i-j+k)} \right)] \right) \right), \quad (2)$$

where  $i$  (GC) is the observed phyllochron expressed in thermal time;  $E(i)$  is the average potential biomass production at growth cycle  $i$ , which depends on microclimatic conditions during plant growth (e.g., temperature, wind speed, relative humidity and solar radiation, etc.);  $r$  is the water use efficiency;  $S_p$  is a characteristic surface area related to plant crown projection, modulated by the effects of self-shading and neighbor competition that is related to plant density;  $slw$  is the specific leaf weight, which is assessed directly from the data;  $t_a$  is the blade functioning duration;  $t_x^b$  is the blade expansion duration;  $N_b(i)$  is the number of leaves produced by the plant at growth cycle  $i$ ;  $P_b(k)$  is the sink strength of the blade of age  $k$ ; and  $D(i)$  is the demand of all expanding organs at growth cycle  $i$ , which is the sum of all the individual organ sink strengths, calculated according to Eq. (3):

$$D(i) = \sum_o \sum_{j=1}^{\min(i, t_x^o)} N_o(i-j+1) P_o(j), \quad (3)$$

where  $o$  = indices for organ type (blade, b; petiole, p; internode, i; fruit, f);  $t_x^o$  is the expansion duration of organ type  $o$ ;  $N_o(i)$  is the number of organs type  $o$  at growth cycle  $i$ ; and  $P_o(i)$  is the sink strength of organ type  $o$  of age  $i$ , calculated according to Eq. (4):

$$P_o(i) = P_o f_o(i), \quad (4)$$

where  $P_o$  is the relative sink strength of organ type  $o$ , indicating the competitive ability of a certain type of organ  $o$  to accumulate biomass from the common pool, and  $f_o(j)$  is a sink variation function

**Table 1**

Sink and source parameters of the GreenLab model.

Parameter	Definition	Units
$P_p, P_i, P_f^a$	Organ sink strength [cf. Eq. (4)]	–
$S_p$	Projection area of a plant [cf. Eq. (2)]	cm <sup>2</sup>
$r$	Water use efficiency [cf. Eq. (2)]	mg cm <sup>-2</sup> mm <sup>-1</sup>

<sup>a</sup> p, petiole; i, internode; f, fruit.

of the organ type  $o$  of age  $j$ , described by a discrete beta function (Kang et al., 2011). The blade strength  $P_b = 1$  was a normalized reference for all other organ sink strengths.

For crop plants grown in a wide range of environmental conditions, the potential evapotranspiration (PET) is the average potential biomass production ( $E$ ) driving biomass acquisition in GreenLab (Ma et al., 2007; Dong et al., 2008; Kang et al., 2011). The daily PET (mm day<sup>-1</sup>) is often calculated using the FAO-24 radiation method (Jensen et al., 1990) as defined in Eq. (5):

$$\text{PET} = a + b \left( \frac{\Delta}{\Delta + \gamma} \cdot R_s \right), \quad (5)$$

where the PET of a growth cycle (mm GC<sup>-1</sup>) is summed from daily PET values, with the duration varying with daily temperature and the phyllochron per growth cycle;  $\Delta$  (kPa °C<sup>-1</sup>) is the slope of the vapor pressure curve related to daily temperature;  $R_s$  (MJ m<sup>-2</sup> day<sup>-1</sup>) is solar radiation;  $\gamma$  (kPa °C<sup>-1</sup>) is a psychometric constant, which is set to 2.45;  $a$  (mm day<sup>-1</sup>) is -0.3; and  $b$  is an adjustment factor, which is set to 1.065. The wind speed and the relative humidity within the greenhouse are respectively set to 0 and 100% due to their relatively small effects on PET. Note that PET is a function of several environmental variables, primarily daily temperature, solar radiation, wind speed and relative humidity, not all of which are explicitly given in Eq. (5).

The GreenLab model, described by a set of recurrent equations, is able to quickly compute the biomass accumulation of an organ at each growth cycle under different environmental stress conditions. For simplicity, the GreenLab model can be also rewritten as Eq. (6):

$$\hat{y} = f_k(x, \theta_k), \quad (6)$$

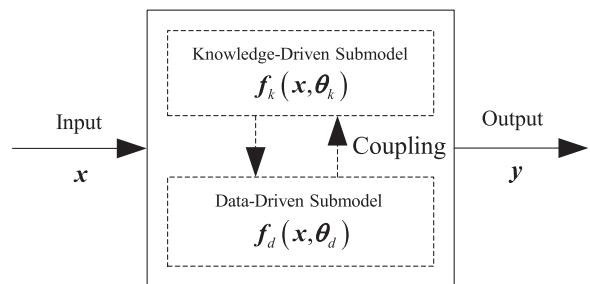
where  $x$  represents the environmental variables related to plant growth,  $\hat{y}$  denotes the output of the GreenLab model (e.g., the total dry weight and the dry weights of different types of organs),  $f_k$  is the function associated with the KDM (i.e., GreenLab model),  $k$  is the subscript associated with the KDM,  $\theta_k$  is a vector of model parameters, including the organ sink strength ( $P_p, P_i, P_f$ ) and two source parameters ( $S_p, r$ ) in the function Eq. (2) controlling plant biomass production (Table 1).

### 2.1.3. Knowledge-and-data-driven modeling approach

This paper proposes a KDDM approach for simulating plant growth that primarily consists of two submodels, as schematically shown in Fig. 1. The upper part of Fig. 1 represents the “knowledge-driven (KD)” submodel, which is derived from all available domain knowledge, including all known relationships derived from physically based or mechanistic models. The lower part of Fig. 1 represents the “data-driven (DD)” submodel, which is constructed solely from data without using any domain knowledge. This model can be expressed by the following mathematical formula

$$\hat{y} = f(x, \theta) = f_k(x, \theta_k) \oplus f_d(x, \theta_d), \quad \theta = (\theta_k, \theta_d), \quad (7)$$

where  $x \in \mathbb{R}^n$  and  $\hat{y} \in \mathbb{R}^m$  are input and output vectors, respectively;  $f$  is a function for a complete model relation between  $x$  and  $\hat{y}$ ; and  $f_k$  and  $f_d$  are the functions associated with the KD and DD submodels, respectively.  $\theta \in \mathbb{R}^p$  is the parameter vector of the function  $f$ , and  $\theta_k$  and  $\theta_d$  are the parameter vectors associated with the



**Fig. 1.** Schematic diagram of the knowledge-and-data-driven model (KDDM), which consists of the “knowledge-driven (KD)” submodel and “data-driven (DD)” submodel (Hu et al., 2009; Ran and Hu, 2014).

functions  $f_k$  and  $f_d$ , respectively. In  $p = p_d + p_k$ ,  $p_d$  and  $p_k$  are the number of model parameters  $\theta_k$  and  $\theta_d$ , respectively. The symbol “⊕” represents a coupling operation between the two submodels.

A two-way coupling connection between the two submodels is applied, which provides flexibility to represent various forms of the two interacting submodels, as implemented in Eq. (7). In general, there is no generic approach to designing the coupling connections. The actual configuration of the coupling is more problem dependent and can be quite complicated because it greatly depends on the form in which domain knowledge is available; domain knowledge may be presented in the form of constraint functions [Table 1 in Hu et al. (2009)], grammar (Todorovski and Džeroski, 2006), rules (Matsunaga et al., 2013), and even physically based models (Czop et al., 2011). Considering the wide variety of coupling connections in the KDDM approach, we limited our coupling connections to two simple and common coupling operators, namely, “superposition” and “composition” (Thompson and Kramer, 1994; Hu et al., 2009). The mathematical expressions of these operators are given by the following:

$$\text{Superposition : } f(x, \theta) = f_k(x, \theta_k) + f_d(x, \theta_d), \quad (8)$$

$$\text{Composition : } f(x, \theta) = f_k(x, f_d(x, \theta_d), \theta_k). \quad (9)$$

The schematic diagrams of the KDDM based on the superposition and composition coupling operators are shown in Fig. 2.

The KDDM based on the superposition coupling operator (abbreviated as KDDM.Sup), together with Eqs. (1), (6) and (8), can be written as

$$\hat{y} = f(x, \theta) = f_k(x, \theta_k) + \Phi(x)\theta_d, \quad (10)$$

where  $x$  represents the environmental variables related to plant growth and  $\hat{y}$  is the output of the model. In Eq. (10), the KDDM.Sup used the RBFN (i.e., DD submodel) to predict the residuals not explained by the GreenLab model (i.e., KD submodel).

The KDDM based on the composition coupling operator (abbreviated as KDDM.Com), together with Eqs. (1), (6) and (9), can be written as

$$E = \Phi(x)\theta_d, \quad (11)$$

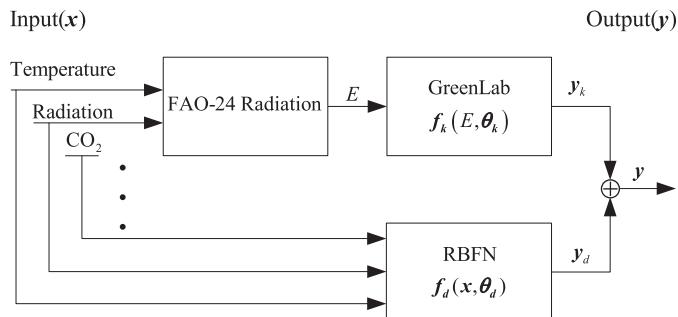
$$\hat{y} = f(x, \theta) = f_k(E, \theta_k).$$

In Eq. (11), the KDDM.Com adopted the RBFN (i.e., DD submodel) to quantify the average potential biomass production ( $E$ ) with all environmental variables ( $x$ ). Once the model parameters  $\theta$  were determined, given  $x$ , the RBFN output ( $E$ ) directly affected the GreenLab model output.

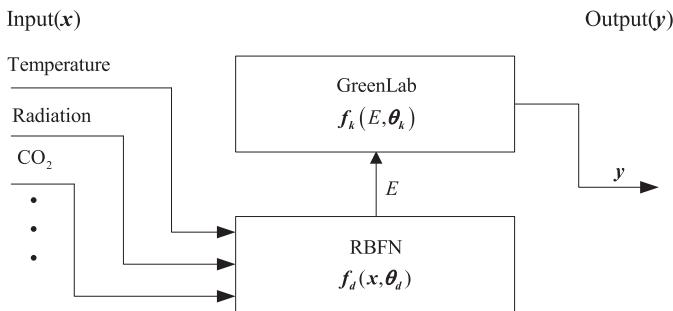
### 2.2. Plant materials and measurements

Data from twelve growth experiments with indeterminately growing tomatoes (*Lycopersicon esculentum* ‘Counter’) were collected over varying seasons across five years in greenhouses at

(a)



(b)



**Fig. 2.** The knowledge-and-data-driven model (KDDM) with two cases of coupling: (a) superposition coupling operator and (b) composition coupling operator.

the Department of Horticulture (Wageningen, the Netherlands) (Heuvelink, 1995). The dry weights of leaves (including petioles), stems and individual fruit trusses were collected destructively from three to eight tomato plants every 6–22 days, and the numbers of leaves, stems and fruits per truss were recorded. The total dry weight (in total, 151 data points) was calculated from the weight of the components. In addition, several daily environmental variables were recorded during the entire growth cycle of the tomato crop (Table 2).

### 2.3. Parameter estimation

With respect to the set of observed data included in the model, "A" and "P" are used to distinguish between the models (Table 3). "A" represents models in which all the observed data were used, including the dry weights of different types of organs and the total dry weight (the sum of the organ dry weight); "P" represents models in which partial observed data (i.e., only total dry weight) were used. For each model, the mean square error (MSE) optimization criterion is used to minimize the difference between the observed values and the predicted values,  $J(\theta)$ , is given by Eq. (12):

$$J(\theta) = \frac{1}{2}(\mathbf{y} - \hat{\mathbf{y}})^T(\mathbf{y} - \hat{\mathbf{y}}), \quad (12)$$

**Table 2**

Description of variables.

Variable	Definition	Units
$x_1$	Planting date	day
$x_2$	Weight at planting date	$\text{g m}^{-2}$
$x_3$	Daily global radiation	$\text{MJ m}^{-2} \text{ day}^{-1}$
$x_4$	Daily temperature	°C
$x_5$	Daily carbon dioxide concentration ( $\text{CO}_2$ )	$\mu\text{mol mol}^{-1}$
$y$	Total dry weight	$\text{g m}^{-2}$
$y_o$	Dry weights of different types of organs	$\text{g m}^{-2}$

**Table 3**

The degree of observed data included in different models.

Model	Independent variables/dependent variables	Degree of observed data included
RBFN(A)	$(x_1, x_2, x_3, x_4, x_5)/y_o$	$y, y_o$
RBFN(P)	$(x_1, x_2, x_3, x_4, x_5)/y$	$y$
GreenLab(A)	$(x_3, x_4)/y$	$y, y_o$
KDDM.Sup(A)	$(x_1, x_2, x_3, x_4, x_5)/y$	$y, y_o$
KDDM.Sup(P)	$(x_1, x_2, x_3, x_4, x_5)/y$	$y$
KDDM.Com(A)	$(x_1, x_2, x_3, x_4, x_5)/y$	$y, y_o$
KDDM.Com(P)	$(x_1, x_2, x_3, x_4, x_5)/y$	$y$

where  $\theta$  represents the model parameters, and  $\mathbf{y}$  and  $\hat{\mathbf{y}}$  represent the observed and predicted values, respectively.

A two-stage learning algorithm for training the RBFN was chosen. In the first stage, the centers and widths of the multiquadric RBFs were determined using k-center and simple heuristic relationships (Zhu and Zhang, 2000). In the second stage, the weights of the RBFN were estimated via the gradient descent method or least squares algorithm. All the observed data were applied to identify source-sink parameters in the GreenLab model using the generalized least squares (GLS) method, as described in more detail by Zhan et al. (2003) and Guo et al. (2006).

Because of the coupling operation between the two submodels, the KDDM approach may have some unidentifiable parameters (i.e., parameters that cannot be determined uniquely) even if the parameters of each submodel are respectively identifiable (Ran and Hu, 2014). To address its parametric identification, parameter estimation for the KDDM approach is performed in two stages: initially, the parameters of the GreenLab model (i.e., KD submodel) are either identified through the GLS method when all the observed data are available or obtained from empirical knowledge or published experimental data on the genotype-specific crop when only total dry weight is given; then, the parameters of the RBFN (i.e., DD submodel) are estimated. This approach is reasonable because the GreenLab model is able to characterize plant growth and biomass allocation based on a set of relatively stable parameters for the genotype-specific crop; the level of parameter variability for the genotype-specific crop is low (Ma et al., 2007; Kang et al., 2011). Once the parameters of the GreenLab model were determined, the remaining task of the KDDM approach was to determine the appropriate setting of the RBFN's parameters. After minimizing the performance measure  $J(\theta)$ , as in Eq. (12), the weights  $\theta_d$  of the DD submodel were estimated using the Levenberg–Marquardt algorithm (Moré, 1978).

### 2.4. Criteria for evaluating the model performance

The Bayesian information criterion (BIC) is a criterion for model selection to evaluate the performance of different models (Burnham and Anderson, 2002). The value of BIC is calculated according to the following equation (Venables and Ripley, 2002)

$$\text{BIC} = N * \log \left( \frac{\sum_{l=1}^N (\mathbf{y}_l - \hat{\mathbf{y}}_l)^2}{N} \right) + p * \log(N), \quad (13)$$

where  $N$  is the number of data points,  $\mathbf{y}_l$  represents the observed values,  $\hat{\mathbf{y}}_l$  represents the predicted values, and  $p$  is the number of model parameters.

**Table 4**

The model selection procedure for the RBFN(A) to modeling tomato crop growth processes using a 12-fold cross-validation strategy.

No.	Num. of para.	BIC	Dry weight of leaves		Dry weight of stems		Dry weight of fruits		Total dry weight	
			RMSE <sub>tr</sub>	RMSE <sub>te</sub>	RMSE <sub>tr</sub>	RMSE <sub>te</sub>	RMSE <sub>tr</sub>	RMSE <sub>te</sub>	RMSE <sub>tr</sub>	RMSE <sub>te</sub>
1	2 × 3	1209.98	26.33	26.95	10.65	10.99	45.11	45.81	71.14	74.00
2	3 × 3	1202.13	21.38	22.26	8.84	9.33	41.29	41.35	65.56	68.74
3	4 × 3	1200.34	19.88	20.69	7.89	8.37	40.13	39.94	61.75	64.23
4	5 × 3	<b>1199.37</b>	19.14	20.25	7.26	8.04	38.14	39.31	58.34	<b>62.47</b>
5	6 × 3	1211.14	19.00	20.38	7.01	7.99	37.91	39.91	57.71	62.48
6	7 × 3	1226.25	18.95	20.44	6.99	8.02	37.95	40.45	57.78	63.17
7	8 × 3	1240.72	18.90	20.47	6.98	8.04	37.84	40.77	57.71	63.65
8	9 × 3	1255.56	18.83	20.43	6.95	8.04	37.85	41.22	57.72	64.17
9	10 × 3	1270.42	18.80	20.41	6.93	8.02	37.81	41.31	57.73	64.39
10	11 × 3	1285.46	18.80	20.44	6.93	8.03	37.87	41.39	57.78	64.43
11	12 × 3	1300.92	18.76	20.42	6.90	7.98	37.86	41.55	57.93	65.46

The lowest BIC and RMSE<sub>te</sub> values for the total dry weight are represented in bold.

The root mean square error (RMSE) is the standard criterion which measures the distance between the predicted and observed values and is given by Eq. (14):

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{l=1}^N (\mathbf{y}_l - \hat{\mathbf{y}}_l)^2}. \quad (14)$$

The modeling efficiency (EF) is a dimensionless quantity which measures the overall goodness of fit between the predicted and observed values and is given by the following formula (Baey et al., 2013)

$$\text{EF} = 1 - \frac{\sum_{l=1}^N (\mathbf{y}_l - \hat{\mathbf{y}}_l)^2}{\sum_{l=1}^N (\mathbf{y}_l - \bar{\mathbf{y}}_l)^2}, \quad (15)$$

where  $\bar{\mathbf{y}}_l$  is the mean of observed values.

For each model, a 12-fold cross-validation strategy was used during the model selection procedure. Twelve experimental data sets were circularly partitioned into training and testing sets. One experimental data set was retained as the testing data set, and the remaining eleven data sets were used as training data sets. In total, twelve training root mean square errors (RMSE<sub>tr</sub>), twelve testing root mean square errors (RMSE<sub>te</sub>) and twelve BIC values were obtained during the learning procedure. The results were then averaged to produce the mean errors and BIC values for performance comparison.

### 3. Results

The RBFN, GreenLab model and KDDM (including KDDM\_Sup and KDDM\_Com) were used to model the dynamical plant growth process of tomato crops. The tomato growth data sets from twelve greenhouse experiments were used to calibrate and test these models.

#### 3.1. Results from individual models

##### 3.1.1. Results from the RBFN

For the RBFN(A) and RBFN(P), the average RMSE<sub>tr</sub>, RMSE<sub>te</sub> and BIC values with respect to the different numbers of parameters are given (Tables 4 and 5). The RBFN(A) 4 and the RBFN(P) 4 had the lowest BIC and RMSE<sub>te</sub> values for the total dry weight, respectively. Compared with the RBFN(A), the RBFN(P) had a lower BIC value (1150.86) and a lower RMSE<sub>te</sub> value ( $62.34 \text{ g m}^{-2}$ ) for the total dry weight. Both RBFN(A) and RBFN(P) under the BIC criterion obtained the best generalization performance, and the best number of hidden nodes for them was 5. Moreover, the total number of parameters of the RBFN(A) was three times that of the RBFN(P) because the former was a three-output RBF network, whereas the

latter was a single-output RBF network. Note that the total dry weight for the RBFN(A) was the sum of the network outputs (i.e., organ dry weight), for which RMSE<sub>tr</sub> and RMSE<sub>te</sub> are given (Table 4).

##### 3.1.2. Results from the GreenLab model

For GreenLab(A), the parameter estimation procedure was repeated twelve times, with each of the twelve experimental data sets used exactly once as the testing data set. The training errors, testing errors, estimated parameter values and their respective standard deviations (Std.) and coefficients of variation (CV) are given (Table 6). The CV values for parameters of the GreenLab model varied between 1.95 and 3.03, which indicates that the five source-sink parameters exhibited very little variation for the genotype-specific tomato crop (*L. esculentum* 'Counter'). Although the CV value of RMSE<sub>tr</sub> was low, the CV value of RMSE<sub>te</sub> in this model was very high, which indicates that the performance of the GreenLab model is lowly reliable. According to values of organ sink parameters, the average ratio of biomass partition from leaves (including petioles) to stems to fruits was 1.432:0.643:3.527 or 26:11:63.

#### 3.2. Results from the KDDM approach

##### 3.2.1. Model selection for KDDM

For the KDDM\_Sup(A) and KDDM\_Com(A), the average RMSE<sub>tr</sub>, RMSE<sub>te</sub> and BIC with respect to the different numbers of parameters are given (Table 7). The KDDM\_Sup(A) 4 and the KDDM\_Com(A) 15 had the lowest BIC and RMSE<sub>te</sub> values, respectively. The KDDM\_Sup(A) and KDDM\_Com(A) under the BIC criterion both obtained the best generalization performance, and the best number of hidden nodes for the RBFN (i.e., DD submodel) in the KDDM\_Sup(A) and KDDM\_Com(A) were 5 and 16, respectively.

**Table 5**

The model selection procedure for the RBFN(P) to modeling tomato crop growth processes using a 12-fold cross-validation strategy.

No.	Num. of para.	BIC	Total dry weight	
			RMSE <sub>tr</sub>	RMSE <sub>te</sub>
1	2	1190.26	71.14	73.99
2	3	1172.54	65.56	68.75
3	4	1161.57	61.90	64.42
4	5	<b>1150.86</b>	58.51	<b>62.34</b>
5	6	1152.35	57.79	62.36
6	7	1157.50	57.83	62.84
7	8	1162.31	57.81	63.25
8	9	1167.28	57.81	63.63
9	10	1172.38	57.85	63.85
10	11	1177.55	57.90	63.95
11	12	1183.18	58.05	64.69

The lowest BIC and RMSE<sub>te</sub> values are represented in bold.

**Table 6**

The parameter estimation procedure for the GreenLab(A) to modeling tomato crop growth processes using a 12-fold cross-validation strategy.

Expo.	Parameters of the GreenLab model					Total dry weight	
	$P_p$	$P_i$	$P_f$	$S_p$ (cm <sup>2</sup> )	$r$ (mg cm <sup>-2</sup> mm <sup>-1</sup> )	RMSE <sub>tr</sub>	RMSE <sub>te</sub>
1	0.425	0.633	3.469	1158.711	0.166	78.80	59.01
2	0.428	0.638	3.508	1157.971	0.166	76.70	75.21
3	0.423	0.631	3.547	1132.218	0.168	72.66	128.31
4	0.435	0.648	3.517	1172.124	0.168	73.12	80.33
5	0.435	0.648	3.511	1151.715	0.167	77.41	67.20
6	0.427	0.637	3.522	1183.146	0.164	76.01	71.41
7	0.424	0.630	3.567	1166.277	0.164	75.85	88.45
8	0.446	0.666	3.615	1194.565	0.170	64.51	141.80
9	0.436	0.650	3.473	1165.134	0.167	77.34	35.40
10	0.419	0.620	3.351	1130.649	0.171	78.46	31.28
11	0.449	0.669	3.598	1196.886	0.164	78.09	23.70
12	0.439	0.652	3.642	1264.283	0.158	71.62	104.31
Mean	0.432	0.643	3.527	1172.807	0.166	75.05	75.54
Std.	0.009	0.015	0.077	35.571	0.003	4.07	36.69
CV (%)	2.083	2.333	2.183	3.033	1.947	5.42	48.57

Note:  $P_b$  was set to 1 as reference, data were not listed here.

**Table 7**

The model selection procedure for the KDDM approach to modeling tomato crop growth processes using a 12-fold cross-validation strategy.  $p_k$  is the number of parameters of the KD submodel, and  $p_d$  is the number of parameters of the DD submodel.

No.	Num. of para. ( $p_k + p_d$ )	KDDM_Sup(A)			KDDM_Com(A)		
		BIC	Total dry weight		BIC	Total dry weight	
			RMSE <sub>tr</sub>	RMSE <sub>te</sub>		RMSE <sub>tr</sub>	RMSE <sub>te</sub>
1	5+2	1179.43	68.48	66.80	1180.15	68.60	66.92
2	5+3	1179.69	67.33	66.70	1156.03	61.76	67.27
3	5+4	1180.59	66.36	65.90	1161.06	61.79	67.27
4	5+5	<b>1172.99</b>	63.42	<b>65.72</b>	1165.76	61.74	67.32
5	5+6	1174.09	62.55	66.44	1168.52	61.24	68.94
6	5+7	1179.05	62.55	66.75	1171.04	60.72	67.50
7	5+8	1183.93	62.54	67.11	1147.05	54.69	58.61
8	5+9	1188.81	62.53	67.49	1163.94	57.11	59.67
9	5+10	1193.79	62.54	67.71	1153.57	54.03	58.16
10	5+11	1198.91	62.59	67.81	1162.04	54.72	59.80
11	5+12	1203.57	62.53	68.46	1165.99	54.55	58.69
12	5+13	1208.76	62.59	68.26	1155.24	51.55	54.39
13	5+14	1212.73	62.39	67.96	1158.12	51.17	54.07
14	5+15	1217.40	62.31	68.08	1168.86	52.26	55.59
15	5+16	1220.31	61.85	68.12	<b>1144.36</b>	46.98	<b>50.26</b>
16	5+17	1224.60	61.71	67.94	1149.59	47.03	50.38
17	5+18	1229.37	61.67	68.08	1154.07	46.96	50.52

The lowest BIC and RMSE<sub>te</sub> values are represented in bold.

Compared with the KDDM\_Sup(A), the KDDM\_Com(A) had a lower BIC value (1144.36) and a lower RMSE<sub>te</sub> value (50.26 g m<sup>-2</sup>). Fig. 3 shows a plot of the regression curves for the 151 data points with respect to the different planting dates from the KDDM\_Com(A) under the BIC criterion.

For the KDDM\_Sup(P) and KDDM\_Com(P), the parameter values of the GreenLab model (KD submodel) obtained from Table 4 in Dong et al. (2008) were set to  $P_p = 0.54$ ,  $P_i = 0.62$ ,  $P_f = 3.28$ ,  $S_p = 1/(11.1 \times 0.77)\text{m}^2$ , and  $r = 1/0.28\text{g m}^{-2}\text{mm}^{-1}$ . Here, only main results of the KDDM\_Sup(P) and KDDM\_Com(P) with the lowest BICs are presented. The KDDM\_Sup(P) and KDDM\_Com(P) had the lowest BIC values (1151.96 and 1140.90, respectively) and RMSE<sub>te</sub> values (62.31 g m<sup>-2</sup> and 51.29 g m<sup>-2</sup>, respectively). The optimal number of hidden nodes for the RBFN (DD submodel) were 5 and 16, respectively. The results reveal that the KDDM\_Com(P) had lower BIC and RMSE<sub>te</sub> values than the KDDM\_Sup(P).

### 3.2.2. Organ dry weight from KDDM\_Com

In addition to the total dry weight, the KDDM\_Com was able to predict the dry weights of different types of organs. The root mean square error (RMSE) and the modeling efficiency (EF) for the dry weights of different types of organs from the KDDM\_Com(A) and

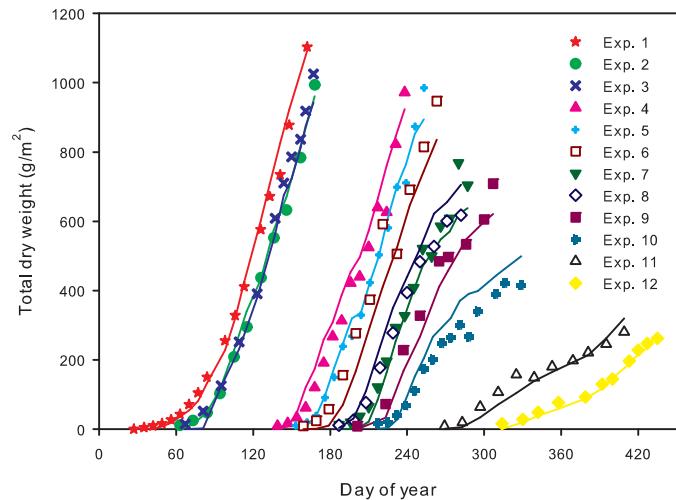
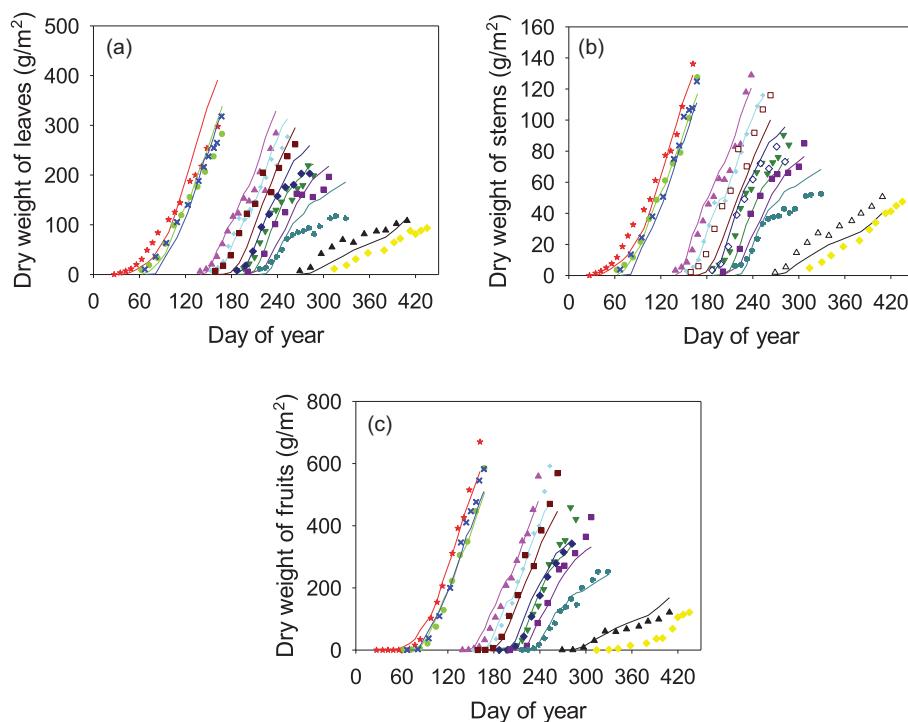


Fig. 3. Regression curves of the total dry weight for the KDDM.Com(A) from twelve greenhouse experiments with different planting dates (January 1 designated day 1). The data points from twelve experiments were taken from Heuvelink (1995).



**Fig. 4.** Prediction outputs of dry weights for three types of organs – (a) leaves, (b) stems, and (c) fruits – from KDDM.Com(P), which used the partial observed data (i.e., only total dry weight) in the training phase. Parameters of the KD submodel (i.e., GreenLab model) in the KDDM.Com(P) model were obtained from Dong et al. (2008). The legend is the same as in Fig. 3.

the KDDM.Com(P) over all greenhouse data sets are given (Table 8). The KDDM.Com(A) had the lowest RMSE value for the dry weights of leaves and fruits and had approximately the same RMSE value for the dry weight of stems in comparison with the KDDM.Com(P). Fig. 4 gives predictions of the KDDM.Com(P) regarding the dry weights of different types of organs from twelve greenhouse experiments with respect to the different planting dates.

### 3.2.3. Average potential biomass production from KDDM.Com

In addition to the FAO-24 radiation method adopted by the GreenLab model, the KDDM.Com can provide an alternative method for calculating the average potential biomass production ( $E$ ) by the RBFN submodel. Fig. 5 shows values of  $E$  based on the FAO-24 radiation method and KDDM.Com(A). Overall, the trends of two methods agreed, and their values of  $E$  ranged from 0.1 to  $27.5 \text{ mm}^{-2} \text{ GC}^{-1}$ . The results of the two methods suggest that the values of  $E$  in late autumn and winter (Exps. 10, 11 and 12) were much lower than in the other experiments. In comparison with the FAO-24 radiation method, the KDDM.Com resulted in larger values of  $E$  in Exps. 11 and 12 during the entire growth cycle of the tomato crop.

**Table 8**

The RMSE and EF for the dry weights of different types of organs from KDDM.Com(A) and KDDM.Com(P) over all greenhouse dataset. A: using all the observed data; P: using the total dry weight only.

		Dry weights of different types of organs		
		Leaves	Stems	Fruits
KDDM.Com(A)	RMSE	<b>25.71</b>	10.30	<b>31.43</b>
	EF	<b>0.896</b>	0.907	0.967
KDDM.Com(P)	RMSE	30.66	<b>9.92</b>	34.44
	EF	0.852	<b>0.914</b>	0.960

The lowest RMSE and the highest EF values for the different types of organs are represented in bold.

### 3.2.4. Quantification of the four variables of KDDM.Com(A)

The contributions of the four inputs used to obtained output from the KDDM.Com were quantified (Table 9). The results indicate that the full model ( $x_1, x_2, x_3, x_4, x_5$ ) produced the best result under the BIC criterion, the model ( $x_1, x_2, x_4, x_5$ ) achieved nearly the same performance as the model ( $x_2, x_3, x_4, x_5$ ), the variable combination ( $x_1, x_3, x_4, x_5$ ) was comparable to the four variable-combination models ( $x_1, x_2, x_3, x_5$ ) and ( $x_1, x_2, x_3, x_4$ ).

### 3.3. Results of model performance evaluation

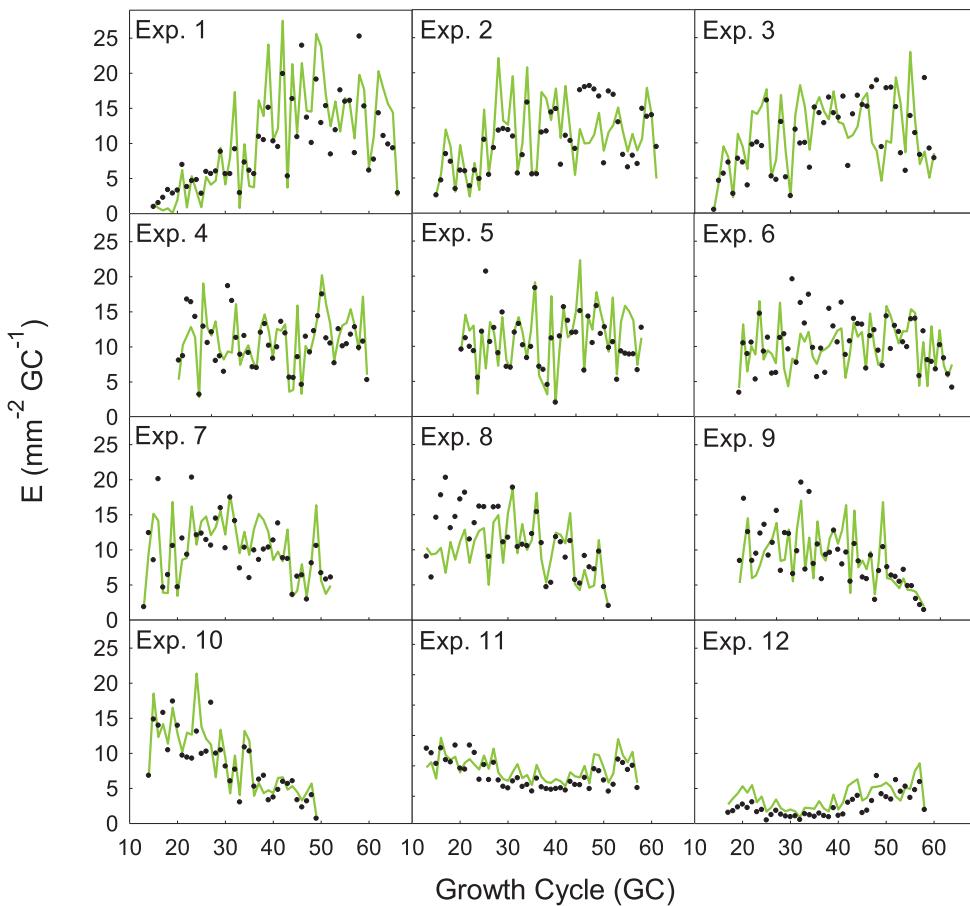
To compare the performance among the individual models (i.e., RBFN and GreenLab) and the KDDM approach, the average RMSE<sub>tr</sub>, RMSE<sub>te</sub> and BIC values from their respective best models are assessed respectively (Table 10). For the individual models, the RBFN had a lower BIC, and lower mean and standard deviation values of RMSE<sub>te</sub> than the GreenLab model. In comparison with individual models, the KDDM approach with a lower BIC value decreased not only the mean of RMSE<sub>te</sub>, which indicates improved accuracy, but also the standard deviation, which suggests an improved capacity to generalize.

**Table 9**

The average RMSE<sub>tr</sub>, RMSE<sub>te</sub> and BIC values of KDDM.Com(A) when one independent variable is missing.

Input variables	BIC	Total dry weight	
		RMSE <sub>tr</sub>	RMSE <sub>te</sub>
$x_1, x_2, x_3, x_4, x_5$	<b>1144.4</b>	<b>46.98</b>	<b>50.26</b>
$x_1, x_2, x_3, x_5$	1159.6	49.63	52.48
$x_1, x_2, x_4, x_5$	1175.8	62.11	69.67
$x_2, x_3, x_4, x_5$	1172.5	61.24	67.94
$x_1, x_3, x_4, x_5$	1162.5	51.72	54.02
$x_1, x_2, x_3, x_4$	1161.2	50.24	53.74

The lowest BIC, RMSE<sub>tr</sub> and RMSE<sub>te</sub> values are represented in bold.



**Fig. 5.** Values of the average potential biomass production ( $E$ ) based on the FAO-24 radiation method and KDDM.Com(A). Solid dot, FAO-24 radiation method; solid line, KDDM.Com(A).

#### 4. Discussion

##### 4.1. The characteristics of individual models

As far as individual models (i.e., RBFN and GreenLab) are concerned, the GreenLab model integrates knowledge of the underlying processes of plant development and growth and can characterize plant growth and biomass allocation based on a set of relatively stable parameters (Table 6), a finding that is in accordance with the results of Dong et al. (2008) and Kang et al. (2011). The five source-sink parameters in the GreenLab model have their own physical interpretation. Specifically, the average projection area of the tomato crop ( $S_p$ ) was  $0.1173 \text{ m}^2$ , less than the inverse of the planting density ( $1/2.1 \text{ m}^2$ ) (Heuvelink, 1995), which indicates that the tomato crop can not form a closed canopy.  $r$  was estimated

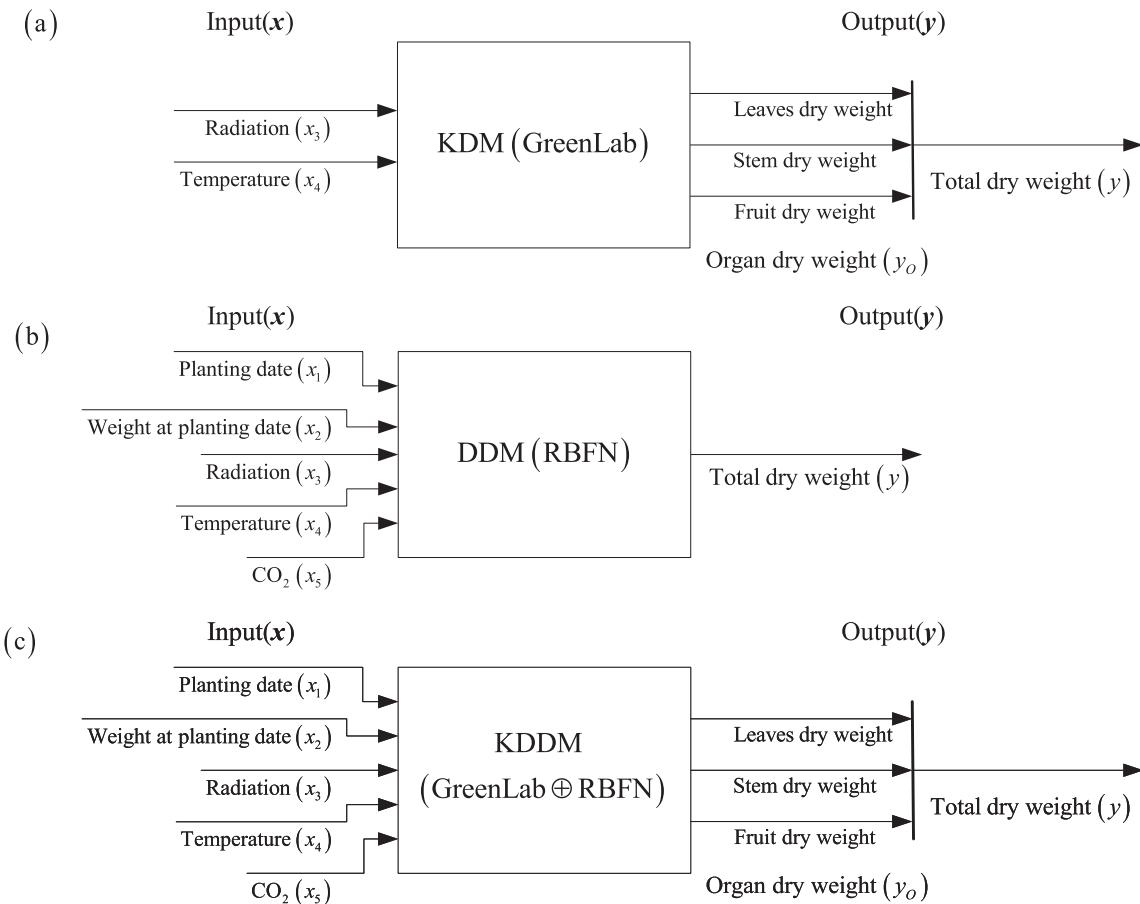
to be  $0.166 \text{ mg cm}^{-2} \text{ mm}^{-1}$ , which reveals the average water use efficiency during the whole growth cycle of tomato crop. Results on the organ sink parameters suggest that approximately 26% of the total dry weight was distributed to the leaves, 11% was distributed to the stems, and 63% was distributed to the fruits (See Section 3.1.2), which indicates that sink parameters are capable of indicating the ability of different types of organs to compete for biomass. This conclusion is basically consistent with the experimental observations of Heuvelink (1995). In addition to the physical interpretation of model parameters, another key characteristic of the GreenLab model is that it is able to compute the dry weights of different types of organs. However, the GreenLab model generally exhibits poor accuracy in predicting the total dry weight and cannot utilize the planting date ( $x_1$ ), the weight at planting date ( $x_2$ ) or the daily carbon dioxide concentration ( $x_5$ ) maximally to improve the

**Table 10**

Results of model performance evaluation for RBFN, GreenLab, KDDM and GAFNs under the BIC criterion. A: using all the observed data; P: using the total dry weight only.  $p_k$  and  $p_d$  are the same as in Table 7.

Model	Num. of para. ( $p_k + p_d$ )	BIC	Total dry weight	
			RMSE <sub>tr</sub> (mean/std.)	RMSE <sub>te</sub> (mean/std.)
RBFN(A)	0 + 15	1199.37	58.34/2.14	62.47/26.28
RBFN(P)	0 + 5	1150.86	58.51/2.15	62.34/25.93
GreenLab(A)	5 + 0	1215.67	75.05/4.07	75.54/36.69
KDDM.Com(A)	5 + 16	1144.36	46.98/2.01	<b>50.26</b> /22.76
KDDM.Com(P)	5 + 16	1140.90	46.40/2.02	<b>51.29</b> / <b>21.82</b>
GAFNs(P) (Qu and Hu, 2009)	0 + 13	<b>1016.33</b>	<b>43.99</b> / <b>1.89</b>	54.95/25.59

The lowest mean and standard deviation values of RMSE<sub>tr</sub> and RMSE<sub>te</sub> are represented in bold.



**Fig. 6.** The inputs and outputs of each approach for modeling the plant growth process: (a) KDM (GreenLab), (b) DDM (RBFN), (c) KDDM (GreenLab  $\oplus$  RBFN).

model performance even when these data are available (Table 10); this shortcoming is due to the model's reliance on daily global radiation ( $x_3$ ) and the daily temperature ( $x_4$ ) (Dong et al., 2008; Kang et al., 2011) (see Fig. 6a).

The main characteristics of the RBFN are that it has high predictive accuracy and can offer a high degree of flexibility to utilize environmental data maximally. The experimental results on the total dry weight suggest that the RBFN under the BIC criterion is suitable for this problem because it exhibited the best generalized performance of the models considered (Tables 4 and 5), and has a better predictive accuracy than the GreenLab model because it decreased the mean and standard deviation of RMSE<sub>te</sub> (Table 10). Unlike the GreenLab model, the RBFN can utilize the planting date ( $x_1$ ), the weight at planting date ( $x_2$ ) and the daily carbon dioxide concentration ( $x_5$ ) by means of adding or reducing its number of input nodes to improve the model performance when these data are available. However, the RBFN lacks transparency with respect to physical comprehension and explanations of a physical system and cannot predict the dry weights of different types of organs when observational data on the organs are unavailable (see Fig. 6b). Furthermore, its model complexity (e.g., the total number of parameters) increases dramatically with the number of output nodes when the dry weights of different types of organs are used as its output (Table 4).

#### 4.2. Advantage of composition over superposition concerning the coupling operator in the KDDM approach

Two cases of the KDDM approach based on the superposition and composition coupling operators were applied to modeling the plant growth respectively. Encouragingly, the KDDM.Com

shows higher accuracy in predicting the total dry weight than the KDDM.Sup (see Section 3.2.1), which suggests that the composition coupling operator in the KDDM approach is more reasonable than the superposition coupling operator for modeling tomato crop growth. As shown in Fig. 2, the KDDM.Sup adopted the RBFN as the DD submodel to compensate for nonlinear deviations of the KD submodel's outputs and considered only the interactions between partial environmental variables and plant growth. By contrast, the DD submodel in the KDDM.Com was used to quantify the average potential biomass production with all environmental variables first; it then directly affected the outputs of the KD submodel. In this way, the DD submodel can effectively compensate for the unknown part and error of the KD submodel due to uncertainty during plant growth. The KDDM.Com took the interactions among all environmental variables and plant growth fully into account. The experimental results suggest that the effect of the environment on the crop yield depended on the plant growth process during which the environmental variables played an important role.

#### 4.3. Comparison of the KDDM with the existing KDM and DDM

In comparison with the existing KDM and DDM (i.e., RBFN and GreenLab), the KDDM approach exhibits better predictive performance for the total dry weight (Table 10). The structure of the KDDM enables the model to utilize domain knowledge (including physically based models) and environmental variables in an optimal way to achieve better performance (Fig. 6c). On the one hand, in the conventional application of the DDM, when there are no observational data of the organ dry weight, such predictions cannot be obtained. In principle, the DDM does not have this capability. Although both the KDM and the KDDM have this capability, the

**Table 11**

Calculation methods for the average potential biomass production ( $E$ ) under different environmental stress conditions.

Types	Methods	Environmental stress conditions	References
I	$E(i) = \text{constant}$	Without any environmental stress	Zhan et al. (2003)
II	$E(i) \propto \text{PAR}(i)^a$	Single environmental variable (light-limiting)	de Reffye et al. (2009)
	$E(i) \propto Q_w(i)^b$	Single environmental variable (water-limiting)	Wu et al. (2012)
III	$E(i) = a + b[R_s \cdot \Delta]/(\Delta + \gamma)$ $E(i) = f_d(\text{temperature, light, CO}_2, \dots)$	Multiple environmental variables (temperature, light and humidity) Multiple environmental variables (temperature, light, $\text{CO}_2$ , etc.)	Ma et al. (2007) Present work

<sup>a</sup> PAR( $i$ ) is the amount of incident photosynthetically active radiation at growth cycle  $i$ .

<sup>b</sup>  $Q_w(i)$  is the amount of soil water content at growth cycle  $i$ .

KDM generally exhibits poor accuracy regarding organ dry weight, whereas the KDDM is able to improve its accuracy through empirical knowledge of sink-source parameters, as evidenced by the results that the KDDM.Com(P) exhibits a high predictive ability for the dry weights of different types of organs, especially for the dry weight of fruits (Table 8). On the other hand, the KDDM approach, compared with the GreenLab model, can provide a more accurate evaluation of the average potential biomass production ( $E$ ) because it may utilize more environmental variables (i.e., the planting date, the weight at planting date and the daily carbon dioxide concentration) (Fig. 5). By contrast, the GreenLab model, which often adopts the FAO-24 radiation method, cannot provide an accurate evaluation of  $E$ , especially when environmental variables are insufficient.

Furthermore, the variable contributions of four inputs used to obtain output from the KDDM.Com were quantified (Table 9). Compared with the models including global radiation ( $x_3$ ), the model ( $x_1, x_2, x_4, x_5$ ) produced the worst result, which suggests that global radiation ( $x_3$ ), which influences plant growth processes is an important environmental variable. Fig. 3 shows that tomato crops grew slowly in late autumn and winter (Exps. 10, 11 and 12) due to reduced radiation. The full model ( $x_1, x_2, x_3, x_4, x_5$ ) has a better prediction accuracy than the four variable-combination models ( $x_2, x_3, x_4, x_5$ ), which indicates that the planting date ( $x_1$ ) accounts for seasonal effects and can be considered to be the seedling date of the crops. In fact, Fig. 3 shows that total dry weight is highest for tomatoes planted between March and June (Exps. 3, 4 and 5) and lowest for those planted after mid-September (Exps. 11 and 12). In particular, the effect of the planting date ( $x_1$ ) on crop growth was greatly magnified when in combination with the global radiation ( $x_3$ ); this effect was evidenced by the fact that the two four-variable combination models ( $x_1, x_2, x_4, x_5$ ) and ( $x_2, x_3, x_4, x_5$ ) performed significantly worse than the full model and the other four variable-combination models. The results suggest that all four variable-combination models have a much better prediction accuracy than the GreenLab model (Tables 9 and 10). Therefore, in addition to ability to utilize domain knowledge and environmental variables maximally, another main characteristic of the KDDM approach is that it can be effectively applied even when one environmental variable (e.g., the planting date, the global radiation) is missing. The case study confirms that the KDDM approach inherits advantages from both the KDM and DDM approaches.

In addition, we compared our method with generalized associative functional networks (GAFNs) (Qu and Hu, 2009), which can be regarded as a simple case of the KDDM approach from the point of view of model structure. The main difference between the two methods is the configuration of coupling due to the availability of domain knowledge in different forms. The experimental results suggest that although the performance, according to the BIC criterion, was not as good as that of the GAFNs with more parameters (Table 10), the KDDM approach decreased the average and standard deviation of RMSE<sub>te</sub>, which indicates that the performance of the KDDM approach is more reliable; it was even able to predict the dry weights of different types of organs.

#### 4.4. A new method for calculating average potential biomass production

One of the important challenges for plant growth modeling is model flexibility of dealing with environmental variables. Model flexibility refers to the ability to flexibly adopt to variable changes in a model. This feature is desirable in either a mechanism-based study or a real application. In a mechanism-based study, one may need to focus on a specific mechanism from which the related environmental variables are considered. However, in different application scenarios, modelers are required to design a model from the given environmental variables which may be different due to the measurement facility. GreenLab provides model flexibility through the calculation of a function  $E(i)$ , which is the average potential biomass production during growth cycle  $i$ . According to different environmental stress conditions during plant growth, methods for calculating  $E$  in GreenLab are divided into three types (Table 11): (I)  $E(i)$  is set to a constant without any environmental stress; (II)  $E(i)$  is assumed to be proportional to a non-optimal single environmental variable under the conditions that other environmental variables are assumed to be optimal; and (III)  $E(i)$  is the potential evapotranspiration (PET) under multiple environmental stress conditions. These methods can be regarded as mechanism-based methods and are focused on the specific mechanism from which the related environmental variables are considered. In this paper, the KDDM.Com approach provides a new method for calculating  $E$  from the given environmental variables by means of the RBFN submodel (see III in Table 11). That is, it can not only utilize the given environmental variables maximally for improved performance, but also be effectively applied when one environmental variable (e.g., light, temperature) is missing (see Section 4.3).

#### 5. Conclusion

This paper presents a knowledge-and-data-driven modeling (KDDM) approach for simulating plant growth. The results of the application of the KDDM approach to a case study on tomato growth revealed several benefits: (a) The KDDM approach is able to preserve physically interpretable parameters (e.g., the source-sink parameters) and has explanatory power in predicting plant growth. (b) The proposed KDDM approach exhibits high accuracy in predicting the dry weights of leaves (including petioles), stems and fruits even when observational data on the organs are unavailable; this characteristic can greatly improve data collection efficiency (only requiring measures of the total dry weight). (c) The DD submodel in the KDDM approach can effectively compensate for the unknown part and error of the KD submodel due to uncertainty during plant growth. (d) This approach can not only maximally utilize domain knowledge and ecological data to improve model performance, but it also effectively addresses situations associated with adding and/or missing variables or data.

Furthermore, the present study provides a promising advance regarding plant growth modeling using the GreenLab model. A new version, which can be called GreenLab.KDDM, is able to take

advantage of data-driven modeling approaches while maintaining the physically based model as the core component. Although the experimental results on tomato data sets confirm a superior predictive power using the GreenLab\_KDDM, a more extensive study of this novel model is needed. The GreenLab\_KDDM should be further developed as a generic tool for use in real-world applications for other crops and plants.

## Acknowledgements

This work is supported in part by China 863 Program (#2012AA101906-2), and the National Science Foundation of China (#31170670, #61273196). Special thanks go to Dr. Hanbing Qu for providing his computer code developed in his paper and many valuable comments and suggestions for the present work. We are grateful to the anonymous reviewers and editors for the comments on the manuscript.

## References

- Atanasova, N., Todorovski, L., Džeroski, S., Kompare, B., 2008. Application of automated model discovery from data and expert knowledge to a real-world domain: lake Glumso. *Ecol. Model.* 212, 92–98.
- Baey, C., Didier, A., Lemaire, S., Maupas, F., Courn'ede, P.H., 2013. Parametrization of five classical plant growth models applied to sugar beet and comparison of their predictive capacity on root yield and total biomass. *Ecol. Model.* 290, 11–20.
- Buhmann, M.D., 2003. Radial Basis Functions: Theory and Implementations. Cambridge University Press, pp. 11–29.
- Burnham, K.P., Anderson, D.R., 2002. Model Selection and Multi-model Inference: A Practical Information-Theoretic Approach. Springer, pp. 284–288.
- Christophe, A., Letort, V., Hummel, I., Cournède, P.H., de Reffye, P., Lecœur, J., 2008. A model-based analysis of the dynamics of carbon balance at the whole-plant level in *Arabidopsis thaliana*. *Funct. Plant Biol.* 35, 1147–1162.
- Czop, P., Kost, G., Slawik, D., Wszołek, G., 2011. Formulation and identification of first-principle data-driven models. *J. Achiev. Mater. Manuf. Eng.* 44, 179–186.
- Daniel, J., Andrés, P.U., Héctor, S., Miguel, B., Marco, T., 2008. A survey of artificial neural network-based modeling in agroecology. In: Prasad, B. (Ed.), Soft Computing Applications in Industry. Springer, pp. 247–269.
- de Reffye, P., Heuvelink, E., Guo, Y., Hu, B.G., Zhang, B.G., 2009. Coupling process-based models and plant architectural models: a key issue for simulating crop production. In: Cao, W.X., White, J.W., Wang, E. (Eds.), Crop Modeling and Decision Support. Springer, pp. 130–147.
- de Reffye, P., Hu, B.G., 2003. Relevant qualitative and quantitative choices for building an efficient dynamic plant growth model: GreenLab case. In: Hu, B.G., Jaeger, M. (Eds.), First International Symposium on Plant Growth Modeling, Simulation, Visualization and Applications (PMA). Tsinghua University Press/Springer, Beijing, China, pp. 87–107.
- Dong, Q.X., Louarn, G., Wang, Y.M., Barczi, J.F., de Reffye, P., 2008. Does the structure-function model GreenLab deal with crop phenotypic plasticity induced by plant spacing? A case study on tomato. *Ann. Botany* 101, 1195–1206.
- Džeroski, S., Todorovski, L., 2003. Learning population dynamics models from data and domain knowledge. *Ecol. Model.* 170, 129–140.
- Guisan, A., Edwards, T.C., Hastie, T., 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecol. Model.* 157, 89–100.
- Guo, Y., Ma, Y.T., Zhan, Z.G., Li, B.G., Dingkuhn, M., Luquet, D., de Reffye, P., 2006. Parameter optimization and field validation of the functional-structural model GreenLab for maize. *Ann. Botany* 97, 217–230.
- Gutiérrez-Estrada, J.C., Pulido-Calvo, I., Bilton, D.T., 2013. Consistency of fuzzy rules in an ecological context. *Ecol. Model.* 251, 187–198.
- Heuvelink, E., 1995. Growth, development and yield of a tomato crop: periodic destructive measurements in a greenhouse. *Sci. Hortic.* 61, 77–99.
- Hu, B.G., Qu, H.B., Wang, Y., Yang, S.H., 2009. A generalized-constraint neural network model: associating partially known relationships for nonlinear regressions. *Inform. Sci.* 179, 1929–1943.
- Jensen, M.E., Burman, R.D., Allen, R.G., 1990. Evapotranspiration and Irrigation Water Requirements. American Society of Civil Engineers, New York, NY, pp. 332–360.
- Kang, M.Z., Yang, L.L., Zhang, B.G., de Reffye, P., 2011. Correlation between dynamic tomato fruit-set and source-sink ratio: a common relationship for different plant densities and seasons? *Ann. Botany* 107, 805–815.
- Ma, Y.T., Li, B.G., Zhan, Z.G., Guo, Y., Luquet, D., de Reffye, P., Dingkuhn, M., 2007. Parameter stability of the functional-structural plant model GreenLab as affected by variation within populations, among seasons and among growth stages. *Ann. Botany* 99, 61–73.
- Matsunaga, F.T., Rakocovic, M., Brancher, J.D., 2013. Modeling the 3d structure and rhythmic growth responses to environment in dioecious yerba-mate. *Ecol. Model.* 290, 34–44.
- Moré, J.J., 1978. The Levenberg–Marquardt algorithm: implementation and theory. In: Watson, G.A. (Ed.), Numerical Analysis. Springer, pp. 105–116.
- Pouteau, R., Meyer, J.Y., Taputuarai, R., Stoll, B., 2012. Support vector machines to map rare and endangered native plants in Pacific islands forests. *Ecol. Inform.* 9, 37–46.
- Qu, H.B., Hu, B.G., 2009. Variational learning for generalized associative functional networks in modeling dynamic process of plant growth. *Ecol. Inform.* 4, 163–176.
- Qu, Y.J., Hu, B.G., 2011. Generalized constraint neural network regression model subject to linear priors. *IEEE Trans. Neural Netw.* 22, 2447–2459.
- Ran, Z.Y., Hu, B.G., 2014. Determining structural identifiability of parameter learning machines. *Neurocomputing* 127, 88–97.
- Recknagel, F., 2001. Applications of machine learning to ecological modelling. *Ecol. Model.* 146, 303–310.
- Renton, M., 2013. Aristotle and adding an evolutionary perspective to models of plant architecture in changing environments. *Front. Plant Sci.* 4, 284.
- Solomatine, D.P., Ostfeld, A., 2008. Data-driven modelling: some past experiences and new approaches. *J. Hydroinform.* 10, 3–22.
- Thompson, M.L., Kramer, M.A., 1994. Modeling chemical processes using prior knowledge and neural networks. *AIChE J.* 40, 1328–1340.
- Todorovski, L., Džeroski, S., 2006. Integrating knowledge-driven and data-driven approaches to modeling. *Ecol. Model.* 194, 3–13.
- Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S, fourth ed. Springer, pp. 107–132.
- Vos, J., Marcelis, L.F.M., Evers, J.B., 2007. Functional-structural plant modelling in crop production: adding a dimension. *Frontis* 22, 1–12.
- Wu, L., Le Dimet, F., de Reffye, P., Hu, B.G., Cournède, P.H., Kang, M.Z., 2012. An optimal control methodology for plant growth – case study of a water supply problem of sunflower. *Math. Comput. Simul.* 82, 909–923.
- Yan, H.P., Kang, M.Z., de Reffye, P., Dingkuhn, M., 2004. A dynamic, architectural plant model simulating resource-dependent growth. *Ann. Botany* 93, 591–602.
- Zhan, Z.G., de Reffye, P., Houllier, F., Hu, B.G., 2003. Fitting a functional-structural growth model with plant architectural data. In: Hu, B.G., Jaeger, M. (Eds.), First International Symposium on Plant Growth Modeling, Simulation, Visualization and Applications (PMA). Tsinghua University Press and Springer, Beijing, China, pp. 108–117.
- Zhang, L.J., Gove, J.H., Heath, L.S., 2005. Spatial residual analysis of six modeling techniques. *Ecol. Model.* 186, 154–177.
- Zhu, M.X., Zhang, D.L., 2000. Study on the algorithms of selecting the radial basis function center. *J. Anhui Univ. (Nat. Sci.)* 24, 72–78.