



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



Xing-Rong Fan^a, Meng-Zhen Kang^{b,*}, Ep Heuvelink^c, Philippe de Reffye^d, Bao-Gang Hu^a

^a National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

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

^c Horticulture and Product Physiology Group, Wageningen University, PO Box 630, 6700 AP Wageningen, The Netherlands

^d 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 (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 (PBM) (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 $\mathbf{x} \in \mathbf{R}^n$. The output of the network is then a scalar function of the input vector, f_d , is given by Eq. (1):

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

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 $\mathbf{c}_j \in \mathbf{R}^n$ are the RBF centers and σ_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

Download English Version:

<https://daneshyari.com/en/article/6296575>

Download Persian Version:

<https://daneshyari.com/article/6296575>

[Daneshyari.com](https://daneshyari.com)