Available online at www.sciencedirect.com

SCIENCE DIRECT®

Ecological Modelling 170 (2003) 141–154

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecolmodel

Using equation discovery to revise an Earth ecosystem model of the carbon net production

Ljupčo Todorovski [a,*], Sašo Džeroski [a], Pat Langley [b], Christopher Potter [c]

[a] *Department of Intelligent Systems, Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia*
[b] *Institute for the Study of Learning and Expertise, 2164 Staunton Court, Palo Alto, CA 94306, USA*
[c] *Ecosystem Science Branch, NASA-Ames Research Center, Mail Stop 242-4, Moffett Field, CA 94035, USA*

Abstract

Equation discovery approaches to automated modeling from observed data usually derive equation-based models from scratch rather than from an initial model already established in the domain of use. In this paper, we present an approach that uses new or recent observational data to improve an existing equation-based model. The approach is used to reduce the error of the Earth ecosystem model of the net production of carbon in the atmosphere. We revise the initial ecosystem model in two directions. First, we calibrate the values of the constant parameters in the model on new observational data. Second, we allow the use of alternative equation structures for some of the sub-models of the initial model and use our approach to choose among them. Experiments show that both revision of values of the constant parameters and revision of the structures of sub-models can considerably reduce the error of the initial model.

Keywords: Machine learning; Equation discovery; Model revision; Carbon net production

1. Introduction

Machine learning methods can help human experts to discover new and interesting knowledge in collections of measured data. A variety of machine learning methods have been successfully used for computational discovery of knowledge about ecosystems (Džeroski, 2001). Knowledge discovered by machine learning methods is usually expressed in the form of decision trees and rules. Although these formalisms are established as a standard notation in the area of machine learning, they are not widely used in Earth science. The discovered knowledge would be more accessible to Earth scientists if methods for discovery of quantitative laws, expressed in the form of equations, were used.

Equation discovery is the area of machine learning that develops methods for automated discovery of equations from measured data (Langley et al., 1987). Early research in this area focused on the problem of rediscovering known models and laws in different areas of science. More recent work in the area has led to methods capable of new discoveries. Recently developed equation discovery methods have also been successfully applied to different tasks of modeling real-world ecosystems from measured data (Džeroski et al., 1999).

* Corresponding author.
E-mail address: ljupco.todorovski@ijs.si (L. Todorovski).

0304-3800/$ – see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0304-3800(03)00222-9

Despite this progress, there are still some important limitations of these methods. Namely, the state-of-the-art equation discovery methods make use of a very limited portion of the theoretical knowledge available in the domain of interest. An important aspect of the domain knowledge that is neglected by the equation discovery systems are the existing models in the domain. Rather than starting the search with an existing equation-based model, current equation discovery systems always start their search from scratch. In contrast with them, theory revision systems (Ourston and Mooney, 1994; Wrobel, 1996) start with an existing theory and use heuristic search to revise the theory in order to improve its fit to newly acquired observational data. However, theory revision research is mainly concerned with the revision of theories, expressed in propositional or first-order logic.

In this paper, we propose a flexible, grammar-based approach to the task of revising models, based on equation discovery. First, the given existing model is transformed into a grammar. Domain expert can then extend this initial grammar with alternative productions specifying alternative modeling decisions. The extended grammar built in this manner specifies the space of possible revisions of the initial model. Therefore, in the last step, the grammar-based equation discovery method LAGRAMGE (Todorovski and Džeroski, 1997) is applied to search through the space of possible revisions and find the one that fits the newly measured data best. The proposed approach allows for the use of the minimal revision principle: among models of similar goodness of fit to the data, the ones that are closer to the initial model are to be preferred. The use of the proposed approach is illustrated on the problem of revising an equation-based model of the net production of carbon by terrestrial plants in the Earth ecosystem (Potter and Klooster, 1997, 1998, 1999).

The paper is organized as follows. The following section gives a brief review of the CASA Earth ecosystem model. Section 3 gives a brief introduction to equation discovery and a presentation of the grammar-based equation discovery method LAGRAMGE. The grammar based approach to the task of revising equation-based models is presented in Section 4. Section 5 presents the experiments in revising the CASA Earth ecosystem model. The review of the research work related to the one presented in the paper is given in Section 6. The Section 7 summarizes and concludes the paper.

2. A quantitative model of the Earth ecosystem

Data from the latest generation of satellites, combined with readings from ground sources, hold great promise for testing and improving on existing scientific models of the Earth's biosphere. One such model, CASA, developed by Potter and Klooster (1997, 1998, 1999) at NASA Ames, accounts for the global production and absorption of biogenic trace gases in the Earth atmosphere, as well as predicting changes in the geographic patterns of major vegetation types (e.g. grasslands, forest, tundra, and desert) on the land.

CASA predicts, with reasonable accuracy, annual global fluxes in trace gas production as a function of surface temperature, moisture levels, and soil properties, together with global satellite observations of the land surface. The model incorporates difference equations that represent the terrestrial carbon cycle, as well as processes that mineralize nitrogen and control vegetation type. These equations describe relations among quantitative variables and lead to changes in the modeled outputs over time. Some processes are contingent on the values of discrete variables, such as soil type and vegetation, which take on different values at different locations. CASA operates on gridded input at different levels of resolution, but typical usage involves grid cells that are eight kilometers square, which matches the resolution for satellite observations of the land surface.

To run the CASA model, the difference equations are repeatedly applied to each grid cell independently to produce new variable values on a daily or monthly basis, leading to predictions about how each variable changes, at each location, over time. Although CASA has been quite successful at modeling Earth's ecosystem, there remain ways in which its predictions differ from observations, suggesting that we invoke computational discovery methods to improve its ability to fit the data. The result would be a revised model, cast in the same notation as the initial one, that incorporates changes which are interesting to Earth scientists and which improve our understanding of the environment.

Because the overall CASA model is quite complex, involving many variables and equations, we decided

Table 1
Variables used in the NPPc portion of the CASA model

NPPc is the net production of carbon by terrestrial plants at a site
E is the photosynthetic efficiency at a site after factoring various sources of stress
T1 is a temperature stress factor ($0 < T1 < 1$) for cold weather
T2 is a temperature stress factor ($0 < T2 < 1$), nearly Gaussian in form but falling off more quickly at higher temperatures
W is a water stress factor ($0.5 < W < 1$)
topt is the average temperature for the month at which *fas_ndvi* takes on its maximum value at a site
tempc is the average temperature at a site for a given month
eet is the estimated evapotranspiration (water loss due to evaporation and transpiration) at a site
PET is the potential evapotranspiration (water loss due to evaporation and transpiration given an unlimited water supply) at a site
pet_tw_m is a component of potential evapotranspiration that takes into account the latitude, time of year, and days in the month
A is a polynomial function of the annual heat index at a site
ahi is an annual heat index that takes the time of year into account
fas_ndvi is the relative greenness as measured from space
IPAR is the energy intercepted from the sun after factoring in the time of year and days in the month
FPAR_FAS is the fraction of energy intercepted from the sun that is absorbed photosynthetically after factoring in vegetation type
monthly_solar is the average radiation incoming for a given month at a site
SOL_CONV is 0.0864 times the number of days in each month

Table 2
Equations used in the NPPc portion of the CASA model

$$NPPc = \max(0, E \cdot IPAR)$$
$$E = 0.389 \cdot T1 \cdot T2 \cdot W$$
$$T1 = 0.8 + 0.02 \cdot topt - 0.0005 \cdot topt^2$$
$$T2 = 1.1814/((1 + e^{0.2\cdot(TDIFF-10)}) \cdot (1 + e^{0.3\cdot(-TDIFF-10)}))$$
$$TDIFF = topt - tempc$$
$$W = 0.5 + 0.5 \cdot eet/PET$$
$$PET = 1.6 \cdot (10 \cdot \max(tempc, 0)/ahi)^A \cdot pet\_tw\_m$$
$$A = 0.000000675 \cdot ahi^3 - 0.0000771 \cdot ahi^2 + 0.01792 \cdot ahi + 0.49239$$
$$IPAR = FPAR\_FAS \cdot monthly\_solar \cdot SOL\_CONV \cdot 0.5$$
$$FPAR\_FAS = \min((SR\_FAS - 1.08)/srdiff, 0.95)$$
$$SR\_FAS = (1 + fas\_ndvi/1000)/(1 - fas\_ndvi/1000)$$
$$SOL\_CONV = 0.0864 \cdot days\_per\_month$$

to focus on one portion that lies on the model's 'fringes' and that does not involve any difference equations. Table 1 describes the variables that occur in this sub-model, in which the dependent variable, NPPc, represents the net production of carbon by terrestrial plants. As Table 2 indicates, the model predicts this quantity as the product of two unobservable variables, the photosynthetic efficiency, *E*, at a site and the solar energy intercepted, *IPAR*, at that site.

Photosynthetic efficiency is in turn calculated as the product of the maximum efficiency (0.389) and three stress factors that reduce this efficiency. One stress term, *T2*, takes into account the difference between the optimum temperature, *topt*, and actual temperature, *tempc*, for a site. The second factor, *T1*, involves the nearness of *topt* to a global optimum for all sites. The third term, *W*, represents stress that results from lack of moisture as reflected by *eet*, the estimated water loss due to evaporation and transpiration, and *PET*, the water loss due to these processes given an unlimited water supply. In turn, *PET* is defined in terms of the annual heat index, *ahi*, for a site, and *pet_tw_m*, a modifier on *PET* to account for day length at differing locations and times of year.

The energy intercepted from the sun, *IPAR*, is computed as the product of *FPAR_FAS*, the fraction of energy absorbed photosynthetically for a given vegetation type, *monthly_solar*, the average radiation for a given month, and *SOL_CONV*, the number of days in that month. *FPAR_FAS* is a function of *fas_ndvi*, which indicates overall greenness at a site as observed from space, and *srdiff*, an intrinsic property that takes on different numeric values for different vegetation types.

Of the variables we have mentioned, *NPPc*, *tempc*, *ahi*, *monthly_solar*, *SOL_CONV*, and *fas_ndvi*, are observable. Two additional terms—*eet* and *pet_tw_m*—are defined elsewhere in the model, but we assume their definitions are correct and thus, we can treat them as observables. The remaining variables are unobservable and must be computed from the others using their definitions. This portion of the model also contains a number of numeric parameters, as shown in the equations in Table 2.

3. Grammar-based equation discovery

Equation discovery is the area of machine learning that develops methods for automated discovery of quantitative laws, expressed in the form of equations, in collections of measured data (Langley et al., 1987).

The task of equation discovery can be formalized as follows. *Given*: (1) a set of system variables $V = \{v_1, \ldots v_n\}$ of the observed system, including a target variable $v_d \in V$ and (2) a table of observations (measured values) of the system variables; *find* a model M formulated as a set of ordinary algebraic or differential equations defining the target variable v_d. The model M is expected to minimize the discrepancy between the observed values of the target variable v_d and the values of v_d obtained with simulating M.

Equation discovery methods address the above task by decomposing it into two sub-tasks. The first is the model identification problem sub-task where an appropriate structure has to be determined for the equations involved in the model. The second is the parameter estimation sub-task where acceptably accurate values for the constant parameters in the equations are to be determined. For the second sub-task, standard non-linear optimization techniques are used in order to fit the values of the constant parameters against the observed data (Press et al., 1986).

For solving the model identification task, heuristic search through the space of possible equations structures is used. The search is guided using a heuristic function that measures the quality of the current equation structure. The quality is estimated as discrepancy between measured data and data obtained with simulating the equation with the optimized values of the constant parameters. More precisely, the discrepancy is calculated as sum of squared errors (SSE), i.e. sum of squared distances between the observed and simulated values of the dependent variable v_d (Todorovski and Džeroski, 1997).

Note however, that the space of candidate equation structure to be explored during the search is potentially huge. The problem of the huge space of equation structures makes the sub-task of determining the appropriate equation structure very difficult. The equation discovery system LAGRAMGE (Todorovski and Džeroski, 1997) addresses this problem by allowing the user to specify the space of possible equation structures. Thus, a user of LAGRAMGE has an opportunity to tailor the space of equation structures according to the modeling knowledge in the particular domain of interest. The search is then focused to equation structures which make sense from the domain scientist's point of view. The discovered equation can be understood and interpreted better and more easily.

The space of possible equations in LAGRAMGE is specified in the form of a context-free grammar (Hopcroft and Ullman, 1979). A context-free grammar contains a finite set of variables (also called non-terminals or syntactic categories) each of which represents expressions or phrases in a language. The expressions represented by each non-terminal are described in terms of the same and other non-terminals and primitive symbols called terminals. The rules relating the non-terminals among themselves and to terminals are called productions. In the case of equations, a non-terminal symbol represents a set of alternative arithmetical expressions that can appear in equations, while terminals are used to denote system variables, constant parameters and arithmetical operators.

Although the motivation for development of context-free grammars was the syntactic description of natural languages, the formalism is powerful enough to express different aspects of the modeling knowledge in the domain of interest (Todorovski and Džeroski, 1997, 2001; Todorovski et al., 1998). For example, a grammar based on knowledge about typical models of basic population dynamics processes was used in LAGRAMGE for successful modeling of phytoplankton growth in Lake Glumsoe (Todorovski et al., 1998). Ecological modeling domain knowledge was essential for the successful use of equation discovery (automated modeling) on the basis of very sparse measurements taken over a short period of two months.

The successful application of LAGRAMGE in the Lake Glumsoe domain shows the importance of using theoretical modeling knowledge from the domain of interest in the process of equation discovery. An important aspect of domain knowledge that is neglected by current equation discovery methods, including LAGRAMGE, are the existing models in the domain. Equation discovery methods ignore existing models and always start their search for an appropriate model structure from scratch. An alternative approach would be to start with an existing model and try to find an appropriate change of its structure so that the revised model better fits newly collected observational data. As we will show in the following section, the formalism of grammars can be also used to integrate existing models in the process of equation discovery.

4. Grammar-based revision of equation models

To revise an existing model, we follow the LAGRAMGE tradition and we use the formalism of context-free grammars to incorporate an existing model in the process of equation discovery. First, we transform existing model into an initial grammar that generates the initial model only and reflects its structure. Next, the initial grammar is extended adding alternative productions that allow changes of the initial model. The extended grammar specifies the space of possible revisions to the initial model. Then, LAGRAMGE with the extended grammar is used to chose among the possible revisions the one that fits the newly collected observational data best. Finally, minimality of change (MC) heuristic function is implemented in LAGRAMGE to support the MC principle. Each aspect of the approach is explained in detail in the following sections.

4.1. From an initial model to an initial grammar

In a typical setting of revising an existing scientific model, we would only have observational data and a model, i.e. an equation developed by scientists to explain a particular phenomenon. A grammar that would explain how this model was actually derived and provide options for alternative models is typically not available. The above is especially true for simpler models.

However, when the model (equation) is complex, it is only rarely written as a single equation defining the target variable, but rather as a set of equations defining the target variable, which typically contains equations defining intermediate variables. The latter typically define meaningful concepts in the domain of discourse. Often, alternative equations defining an intermediate variable would be possible and the modeling scientist would choose one of these: the alternatives would rarely (if ever) be documented in the model itself, but might be mentioned in a scientific article describing the derived model and the modeling process.

A set of equations defining a target variable through some intermediate variables can easily be turned into a grammar, as demonstrated in Table 3. The presented grammar generates the equations used in the NPPc portion of the CASA model (see Table 2 in Section 2). Each intermediate variable in the NPPc portion of the CASA model has a corresponding non-terminal symbol in the grammar. Each non-terminal symbol has a single production that generates the equation from the CASA model that is used to calculate the corresponding intermediate variable. The terminal symbols of the grammar reflect those CASA variables that are observables (*tempc*, *ahi*, *monthly_solar*, *days_per_month* and *fas_ndvi*) or intermediate variables defined in other portions of the CASA model that are treated as observables (*eet* and *pet_tw_m*).

Table 3
Grammar derived from the equations for the NPPc variable in the CASA model in Table 2

```
NPPc->      max(0,E*IPAR)
E->         0.389*T1*T2*W
T1->        0.8+0.02*topt-0.0005*topt*topt
T2->        1.1814/((1+exp(0.2*(TDIFF-10)))
            *(1+exp(0.3*(-TDIFF-10))))

TDIFF->     topt-tempc
W->         0.5+0.5*eet/max(PET,0)
PET->       1.6*pow(10*max(tempc,0)/ahi,A)
            *pet_tw_m
A->         0.000000675*ahi*ahi*ahi-0.0000771
            *ahi*ahi+0.01792*ahi+0.49239

IPAR->      FPAR_FAS*solar*SOL_CONV*0.5
FPAR_FAS->  min((SR_FAS-1.08)/srdiff,0.95)
SR_FAS->    (1+fas_ndvi/1000)/
            (1-fas_ndvi/1000)
SOL_CONV->  0.0864*days_per_month
```

The grammar generates the equations of the initial model only.

Using the initial grammar from Table 3 only the NPPc portion of the initial CASA model from Table 2 can be generated. In order to allow revisions of the initial model, we should extend the grammar and specify which parts of the model can be changed and specify the allowed changes.

4.2. From an initial to an extended grammar

Having the grammar in Table 3 enables us to specify alternative models through providing additional productions for the non-terminal symbols in the grammar. The additional productions for a non-terminal symbol (or corresponding intermediate variable) specify alternative modeling choices, of which only one will eventually be chosen for the final model. Note however, that the revised model can incorporate

an arbitrary combination of the choices made for individual intermediate variables.

Note that when alternative productions are specified for an intermediate variable, there are no restrictions (at least in principle) on these productions. For example, they can introduce new intermediate variables and productions defining them. They can also specify arbitrary functional forms (in the case of equations). However, they do have to eventually derive (in the context of the entire grammar) valid sub-expressions involving the set of terminal symbols (system variables) associated to the initial model.

A very common alternative production would replace a particular constant value with a generic constant, allowing the equation discovery system to refit its value to the given observational data. That change can be achieved by replacing a terminal symbol representing a constant parameter `v` with a generic symbol `const` that allows for an arbitrary value of the particular constant parameter. For example, consider the production `E->0.389*T1*T2*W`. The alternative production `E->const[:v:]*T1*T2*W` allows for an arbitrary value of the constant parameter (keeping its initial value at `v`). Often, more restricted changes of constant values are desirable: `const[0:v:]` would allow arbitrary non-negative value of the constant parameter. Similarly, a more specific terminal symbol `const[0:v:2*v]` would allow for a 100% relative change of the initial value `v` of the constant parameter.

A slightly more complex alternative production would replace a particular polynomial on the right-hand-side of a production with an arbitrary polynomial of the same (intermediate) variables. For example, consider the production for the non-terminal symbol `T1`. It specifies that the value of the variable *T1* is calculated using a second degree polynomial. We can allow an arbitrary polynomial to be used for calculating *T1* by adding the two alternative productions for the non-terminal `T1` presented in Table 4.

The first alternative production is used to derive the simplest polynomial (of degree zero), which actually defines *T1* as a constant value to be fitted against data. The second production from Table 4 can be repetitively used to derive an arbitrary degree polynomial in the following manner. We start with the initial expression `T1`. Applying the second production to the initial expression once we derive the expression `const+(T1)*topt`. When applying the second production to the newly derived expression, the non-terminal `T1` is replaced by the right-hand side of the production rule, obtaining the expression `const+(const+(T1)*topt)*topt`. At this point, we can decide to apply the first production to generate the terminal expression `const+(const+(const)*topt)*topt`. On the other hand, if we decide to apply the second production once more, we will obtain a third degree polynomial. In general, the degree of the derived polynomial is equal to the number of applications of the second production during the derivation.

In the examples above, we presented several possible extensions of the initial grammar. However, it is also possible to add alternative productions that allow for derivation of an arbitrary arithmetical expression. Therefore, the approach is general in the sense that it can be used to specify arbitrary revisions of the initial model.

Table 4
Two alternative productions that allow an arbitrary polynomial to be used for calculating the value of the intermediate variable *T1*

```
T1->const
T1->const+(T1)*topt
```

4.3. Minimality of change principle

While the approach presented above does take into account the initial model, it may allow for a completely different model to be derived, depending on whether and what kind of productions for alternative models are provided for each of the intermediate variables. It is here that the minimal revision/change principle comes into play: among theories of similar quality (fit to the data), theories that are closer to the initial theory are to be preferred.

The crucial concept that is necessary in order to implement the MC principle is the measure of change or distance between the (potential) revised model and the initial model. Since parse trees are used in LAGRAMGE to represent models, we use a measure of distance between tree structured terms as a measure of distance between models. Thus, the distance measure we use assesses syntactic structural distance, i.e. the amount of change in the structure of the equations of the model.

A common approach to computing distances between strings or tree structured terms is the *editing* approach, leading to *edit distance measure* (Shasha and Zhang, 1997). Following the editing approach, a set of basic edit operations is first defined. The edit operations available for editing trees are relabeling (changing the label), deleting and inserting a node in the tree. Costs are assigned to these operations, depending on the labels of the nodes involved in them. The problem of computing the distance between two tree structured terms $\mathcal{T}_1$ and $\mathcal{T}_2$ is then transformed to the problem of finding a minimal cost sequence of basic editing operations that transforms a tree $\mathcal{T}_1$ into a tree $\mathcal{T}_2$.

Since the models in LAGRAMGE are represented as grammar parse trees, we decided to use a refined edit distance measure, that can be efficiently calculated on parse-trees, presented in (Richter, 1997). Before using the measure, we have to define the costs of the basic edit operation of deletion, insertion, and relabeling. The costs of the basic edit operations are defined as follows:

Deleting a node has a cost of 1.
Inserting a node has a cost of 1.
Relabeling a node has a cost of 1, if the label is actually changed, or 0 otherwise. Note that for non-terminal symbols that denote constant parameters, the actual value of the constant parameter is used a label.

Once we have defined a distance measure between models, we can incorporate it into LAGRAMGE by modifying the MDL heuristic function used in LAGRAMGE to introduce preference toward simpler equations. The MDL heuristics takes into account the complexity of an equation along with its goodness of fit to the data (Todorovski and Džeroski, 1997), i.e.

$$\mathrm{MDL}(M) = \mathrm{SSE}(M) + \frac{l(M)}{10 \cdot l_{\max}} \cdot \mathrm{SSE}(M_0),$$

where $\mathrm{SSE}(M)$ is the sum of squared errors of the current model on the training data, $\mathrm{SSE}(M_0)$ is the error of the simplest model, $l(M)$ is the length of the current model M (in number of terminal symbols) and $l_{\max}$ the length of the most complex equation is the search space. Since the LAGRAMGE search space consists of parse trees with limited depth, the maximal length $l_{\max}$ can be easily computed in advance. Roughly speaking, the second part of the MDL heuristic function of LAGRAMGE adds a penalty for equation complexity to the sum of squared errors.

By analogy to the MDL heuristic, we can define MC heuristic function as follows:

$$\mathrm{MC}(M) = \mathrm{SSE}(M) + \frac{\mathrm{distance}(M, M_0)}{C} \cdot \mathrm{SSE}(M_0),$$

where $\mathrm{distance}(M, M_0)$ is the distance between the current model M and the initial model M_0. Note that the maximal distance is not available as in the case of maximal length for MDL, so we introduce a user defined parameter C. This parameter can be used to trade-off between goodness of fit of the current model and MC with respect to the initial model. Large values of C will diminish the "change penalty" term of the MC heuristic, leading to a preference toward accurate models, not necessarily similar to the initial one. On the other hand, small values of C will increase the "change penalty" term, leading to a preference toward models that are similar to the initial model M_0.

5. Experiments in revising an Earth science model

We illustrate the use of the proposed framework for theory revision in equation discovery on the problem of revising one part of the Earth science CASA model (Potter and Klooster, 1997, 1998, 1999), described in Section 2. The values of the input variables (terminal symbols in the grammar from Table 2) were measured (and/or calculated) for 303 locations on the Earth providing a data set with 303 examples. Measured NPPc was commonly determined by sampling the accumulated biomass amount of the standing vegetation and adjusting for the age of the vegetation community sampled, in order to estimate the yearly NPPc carbon flux.[1]

The training data set used in the experiments of CASA-NPPc model revision consists of 303 data points. Each data point contains measurements of the observed system variables for a distinct location on Earth.

The quality of the revised models is assessed through the discrepancy between the predicted

[1] Data provided by the Global Primary Productivity Data Initiative (GPPDI) NPP Working Groups and Ecosystem Model-Data Intercomparison (EMDI) activity of the International Geosphere Biosphere Program Data and Information System (IGBP-DIS), Oak Ridge National Laboratory.

and observed values of the dependent variable: the smaller the discrepancy, the better the model. The discrepancy is measured using standard root mean squared error (RMSE) measure, calculated as: $\sqrt{\sum_{i=1}^{303}(NPPc_i - \widehat{NPPc_i})^2/303}$, where $NPPc_i$ and $\widehat{NPPc_i}$ are the observed and the predicted value of *NPPc*, respectively. The RMSE of the initial model on the training data set is 517.665.

In order to estimate the error of the revised models on test data unseen during the process of revision, we applied a 30-fold cross validation methodology. Following this methodology, the data set consisting of 303 examples is randomly partitioned into 30 partitions, with approximately the same number of examples (10) in each of them. In each iteration of the cross-validation procedure, 29 out of 30 partitions are used as a training data set for revision of the initial model and the revised model is then used to predict the values of the dependent variable *NPPc* on the remaining partition, unseen during the revision phase. By repeating this iteration thirty times, once for each partition, we obtain 303 predictions of the *NPPc* value for all the data points in the training data set.

5.1. A grammar for the revision of the CASA-NPPc model

As described in Section 4, the given CASA-NPPc model was first transformed into the initial grammar presented in Table 3. In addition, alternative predictions were added to this initial grammar for the four intermediate variables for which experts pointed out that they are not very confident in the equations used to calculate their values. Each of these alternative productions specifies one or more possible revisions of the initial CASA-NPPc model. The complete list of alternative productions added to the initial grammar is given in Table 5. The productions are further discussed below.

Alternative productions for E

`Ec-100` allows a revision of the constant parameter (with the initial value of 0.389) in the equation for the intermediate variable *E*. The alternative production allow for a 100% relative change of the initial value of the constant parameter.

`Es-exp` allows for a replacement of the three terms product from the initial *E* equation (i.e. the product $T1 \cdot T2 \cdot W$) with a product that allows for arbitrary non-negative exponents on these terms (i.e. a product of the form $T1^{c_1} \cdot T2^{c_2} \cdot W^{c_3}$). The initial values of the exponents are set to 1, in which case the product is equivalent to the product in the initial *E* equation.

Alternative productions for T1

`T1c-100` allows for a 100% relative change of the initial values of the constant parameters in the *T1* equation.

`T1s-poly` allows for a replacement of the initial second degree polynomial that defines the value of *T1* with an arbitrary degree polynomial of the variable *topt*. In

Table 5
Alternative productions added to the initial grammar from Table 3

`Ec-100:E->`	`const[_:0:0.389:0.778]*T1*T2*W`
`Es-exp:E->`	`const[_:0:0.389:0.778]*pow(T1,const[_:0:1:])*pow(T2,const[_:0:1:])*pow(W,const[_:0:1:])`
`T1c-100:T1->`	`const[_:0:0.8:1.6]+const[_:0:0.02:0.04]*topt-const[_:0:0.0005:0.001]*topt*topt`
`T1s-poly:T1->`	`const|const+(T1)*topt`
`T2c-100:T2->`	`const[_:0:1.1814:2.3628]/((1+exp(const[_:0:0.2:0.4]*(TDIFF-const[_:0:10:20])))*(1+exp(const[_:0:0.3:0.6]*(-TDIFF-const[_:0:10:20]))))`
`T2s-poly:T2->`	`const|const+(T2)*TDIFF`
`SR_FASc-25:SR_FAS->`	`(1+fas_ndvi/const[_:750:1000:1250])/(1-fas_ndvi/const[_:750:1000:1250])`

Each of them specifies one or more revisions of the initial CASA-NPPc model.

Table 6
The root squared mean error (RMSE) of the revised model, the percentage of relative error reduction (RER) of the RMSE of the revised model when compared to the RMSE of the initial CASA-NPPc model (with RMSE of 517.665) and distance (DIST) of the revised model from the initial one

Alternative production(s)	Training			CV		
	RMSE	RER (%)	DIST	RMSE	RER (%)	DIST
`Ec-100`	458.626	11.40	1	459.212	11.29	1.0
`Es-exp`	442.763	14.47	16	447.456	13.56	16.0
`T1c-100`	458.301	11.47	3	460.352	11.07	3.0
`T1s-poly`	450.265	13.02	46	455.819	11.95	46.0
`T2c-100`	457.048	11.71	3	457.926	11.54	3.0
`T2s-poly`	450.972	12.88	71	463.757	10.41	75.8
`SR_FASc-25`	441.419	14.73	2	441.419	14.73	2.0
All combined	411.627	20.48	60	421.758	18.53	62.6

The RMSE was estimated both on training data and using 30-fold cross-validation.

addition, the maximal depth of the parse trees considered by LAGRAMGE was set to allow a maximal polynomial degree of five.

Alternative productions for T2

`T2c-100` allows 100% relative change of the initial values of the constant parameters in the equation for `T2`.

`T2s-poly` allows for replacement of the initial equation that defines the value of *T2* with an arbitrary degree polynomial of the variable *TDIFF*. Again, the maximal degree of the polynomial was limited to five.

Alternative productions for SR_FAS

`SR_FASc-25` allows for a 25% relative change of the initial values of the constant parameters in the *SR_FAS* equation. The relative change of 25% was used to avoid values of the constant parameters lower than 750, which would cause singularity (division by zero) problems in the equation for *SR_FAS*.

Note, however, that an arbitrary combination of these alternative productions can be added to the initial grammar. If all the alternative productions are added at the same time, then LAGRAMGE will find the most beneficial combination of revisions, i.e. the one that leads to the best revision of the initial model.

5.2. Experimental results

The results of the experiments with the different modeling (revision) alternatives, discussed above are summarized in Table 6.

When we allow only a single of the seven presented alternatives (the first seven rows of Table 6), revising the value of the constant parameters in the equation for calculating *SR_FAS* gives the largest reduction of the error of the initial CASA-NPPc model. The initial values of the parameters (both are equal to 1000) define an almost linear dependence of *SR_FAS* on the observed system variable *srdiff*. The revised values of the constant parameters were equal to 750 (lower bound values), which increase the non-linearity of the dependence. In terms of consistency of the revision with Earth science knowledge, we should note that the Earth scientists' confidence in the range of the *srdiff* variable is low due to the limited terrestrial coverage of the *NPPc* measurements. Therefore, the theoretically based argument for high initial values of the constant parameters in the *SR_FAS* equation is not so strong.

The analysis of the results of the individual structural revisions shows the following. The `T1s-poly` revision replaces the initial second-degree polynomial for calculating *T1* with a fifth degree polynomial. The structural revision `T2s-poly` replaced the complex initial equation structure for calculating *T2* with a fourth degree polynomial. While the initial form of the *T2* equation is fairly well grounded in first

principles of plant physiology, it has not been extensively verified from field measurements. Therefore, both empirical improvements are beneficial.

The most interesting structural revision was the one for the E equation:

$$E = 0.610 \cdot T1^{2.83} \cdot T2^{0.638} \cdot W^{0}$$

The proposed value of 0 for the exponent of the water stress factor W suggests that the water stress factor is not important for predicting the photosynthetic efficiency E. Earth scientists proposed as a possible explanation for this the fact that the influence of the water stress factor is already being captured by the satellite measurements of the relative greenness *fas_ndvi*.

The results of the experiments with searching for the optimal combination of all the alternative revisions are presented in the last row of Table 6. As expected, the optimal combination leads to the maximal relative reduction of the RMSE of more than 20% on the training data and 18.5% when cross-validated. The combination of `Es-exp`, `T1s-poly`, `T2c-100`, and `SR_FASc-25` alternative productions leads to the best revised model, presented in Table 7.

After the initial experiments with the revision of the CASA-NPPc model presented here, Earth scientists that developed the CASA model, decided to change the value of the constant parameter in the E equation from 0.389 to 0.56, independently from our experiments. This change reduces the RMSE of the initial CASA-NPPc model on the training data from 517.665 to 465.213. After rerunning the revision experiments with the new initial CASA-NPPc model, we obtained the results presented in Table 8.

The revisions of the new corrected initial CASA-NPPc model lead to smaller relative reduction of the RMSE. The maximal error reduction of almost 11% on the training data and 9% when cross-validated is obtained when an arbitrary combination of modeling alternatives is allowed. The best revised model, obtained using the combination of `Es-exp`, `T1c-100`, `T2s-poly`, and `SR_FASc-25` alternative productions, is presented in Table 7. Note that the experimental results also show that the reductions, obtained with allowing a single alternative production at a time, sum up, i.e. the error reduction obtained with the combination of alternative productions (almost) equals the sum of the error reductions obtained with individual ones.

Note also that the error of the revision of the corrected model on the training data (414.739) is slightly higher than the error of the best model obtained with revising the initial CASA-NPPc model (411.627, see Table 6). This is due to the problems with the convergence of the method for non-linear optimization of the values of the constant parameters. It is well known that these methods can not guarantee convergence toward the global (or real) optimal values, but can stuck into a local (sub-)optimal values that are closer to the initial values of the constant parameters (Press et al., 1986).

The comparison of the revised models in Table 7 and Table 9 shows that both revised models are similar. Both of them suggest that the W(water stress) segment should be removed from CASA-NPPc model, since it is not important for calculating E. Furthermore, both

Table 7
The revised CASA-NPPc model obtained by allowing an arbitrary combination of modeling alternatives from Table 5

$$NPPc = \max(0, E \cdot IPAR)$$
$$E = 0.312 \cdot T1^{1.36} \cdot T2^{0.728} \cdot W^{0}$$
$$T1 = 3.65 - 0.992 \cdot topt + 0.137 \cdot topt^{2} - 0.00679 \cdot topt^{3} + 0.000111 \cdot topt^{4}$$
$$T2 = 0.818/((1 + \exp(0.0521 \cdot (TDIFF - 10))) \cdot (1 + \exp(0 \cdot (-TDIFF - 10))))$$
$$TDIFF = topt - tempc$$
$$W = 0.5 + 0.5 \cdot eet/PET$$
$$PET = 1.6 \cdot (10 \cdot \max(tempc, 0)/ahi)^{A} \cdot pet\_tw\_m$$
$$A = 0.000000675 \cdot ahi^{3} - 0.0000771 \cdot ahi^{2} + 0.01792 \cdot ahi + 0.49239$$
$$IPAR = FPAR\_FAS \cdot monthly\_solar \cdot SOL\_CONV \cdot 0.5$$
$$FPAR\_FAS = \min((SR\_FAS - 1.08)/srdiff, 0.95)$$
$$SR\_FAS = (1 + fas\_ndvi/750)/(1 - fas\_ndvi/750)$$
$$SOL\_CONV = 0.0864 \cdot days\_per\_month$$

The parts of the models that are not revised are printed in grey.

Table 8
The root squared mean error (RMSE) of the revised model, the percentage of relative error reduction (RER) of the RMSE of the revised model when compared to the RMSE of the (corrected) initial CASA-NPPc model (with RMSE of 465.213) and distance (DIST) of the revised model from the initial one

Alternative production(s)	Training			CV		
	RMSE	RER (%)	DIST	RMSE	RER (%)	DIST
`Ec-100`	458.626	1.42	1	460.5	1.01	0.9
`Es-exp`	443.029	4.77	16	443.032	4.77	16.0
`T1c-100`	458.301	1.49	3	460.799	0.95	3.0
`T1s-poly`	450.265	3.21	46	457.37	1.69	45.8
`T2c-100`	457.018	1.76	3	459.633	1.20	3.0
`T2s-poly`	450.972	3.06	71	461.642	0.77	73.4
`SR_FASc-25`	453.157	2.59	2	455.281	2.13	2.0
All combined	414.739	10.85	104	423.684	8.93	67.4

The RMSE was estimated both on training data and using 30-fold cross-validation.

revised models suggest a lower value (750) for the constant parameter in the *SR_FAS* equation. On the other hand, the models suggest different revisions of the *T1* and *T2* equations.

Finally, in the last series of experiments, we explored the influence of the MC principle on the revised models. For this purpose, we used the MC heuristic function (see Section 4) in LAGRAMGE with 7 different values of the *C* parameter: 32, 64, 128, 256, 512, 1024 and 2048. Recall from Section 4 that the *C* parameter is used to trade-off between goodness-of-fit of the model and MC with respect to the initial model: smaller values lead to a higher preference toward models that are similar to the initial one.

The results of these experiments are summarized in Fig. 1. As expected, the distance of the revised model from the initial model constantly increases as we increase the value of the *C* parameter. The distance is maximal when SSE heuristic function is used, i.e. the MC principle is neglected. The trend of the relative reduction of the error of the initial model, estimated on training data, is the same: it constantly increases and reaches maximum when the SSE heuristic function is used. Thus, the more distant is revised model, the more accurate it is on training data.

The revised model that is most similar to the initial one (i.e. the one found using MC heuristic with $C = 32$) is obtained with revising the values of the

Table 9
The new revised CASA-NPPc model obtained by allowing an arbitrary combination of modeling alternatives from Table 5

$$NPPc = \max(0, E \cdot IPAR)$$
$$E = 0.402 \cdot T1^{0.624} \cdot T2^{0.215} \cdot W^{0}$$
$$T1 = 0.680 + 0.270 \cdot topt - 0 \cdot topt^2$$
$$T2 = 0.162 + 0.0122 \cdot TDIFF + 0.0206 \cdot TDIFF^2 - 0.000416 \cdot TDIFF^3 - 0.0000808 \cdot TDIFF^4 + 0.000000184 \cdot TDIFF^5$$
$$TDIFF = topt - tempc$$
$$W = 0.5 + 0.5 \cdot eet/PET$$
$$PET = 1.6 \cdot (10 \cdot \max(tempc, 0)/ahi)^A \cdot pet\_tw\_m$$
$$A = 0.000000675 \cdot ahi^3 - 0.0000771 \cdot ahi^2 + 0.01792 \cdot ahi + 0.49239$$
$$IPAR = FPAR\_FAS \cdot monthly\_solar \cdot SOL\_CONV \cdot 0.5$$
$$FPAR\_FAS = \min((SR\_FAS - 1.08)/srdiff, 0.95)$$
$$SR\_FAS = (1 + fas\_ndvi/750)/(1 - fas\_ndvi/750)$$
$$SOL\_CONV = 0.0864 \cdot days\_per\_month$$

The parts of the models that are not revised are printed in grey.

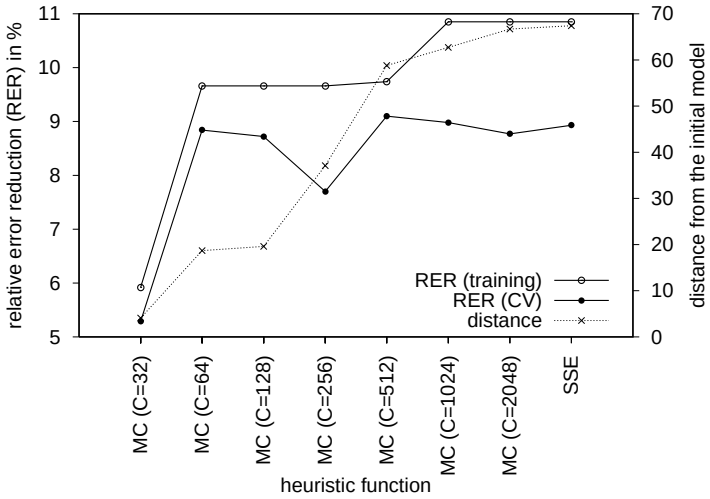

Fig. 1. Relative error reduction and distance from the initial model for revised models obtained using SSE and MC heuristic function with different values of the C parameter.

constant parameters (`Ec-100` and `SR_FASc-25`) of the initial model, leading to an error reduction of 5.29%. This shows that the revisions of the initial equations for E and especially *SR_FAS* are necessary and very important for the error reduction, even if we prefer a minimal change of the initial CASA-NPPc model. The analysis of the second revised model obtained with $C = 64$ gives further support for this claim. Namely, it leads to (almost maximal) error reduction of 8.84% with revising these two equations again. However, in this case, more complex structural revision (`Es-exp`) of the initial E equation has been proposed.

The increasing trend of the performance of the revised models on training data can easily lead to overfitting, especially in cases when arbitrary revisions are allowed. Even in the experiments with a limited set of revision alternatives performed here, we can see that the cross-validated error reduction does not constantly increase. Therefore, models that are closer to the initial one can perform better on test data. In our experiments, the model obtained using the MC heuristic with $C = 512$ (error reduction of 9.10%) slightly outperforms the model obtained using the SSE heuristic (error reduction of 8.93%), when cross-validated. The revised model, obtained using the MC heuristic with $C = 512$, leaves the *T2* equation unchanged and has a structure that is otherwise identical to the structure of the model in Table 9 (the values of the constant parameters in the equations are slightly different). This result shows that the revision of the *T2* equation is not really important for the reduction of the error of the initial CASA-NPPc model.

6. Related work

The research presented in the paper is closely related to several other lines of work.

In the first, Saito et al. (2001) address the same task of revising models based on equations. Their approach is based on transforming a part of the model into a neural network, retraining the neural network on newly measured data, and transforming the trained network back into an equation-based model. The obtained revised models have a considerably smaller error rate than the initial one. Their method gained slightly lower reduction of the initial model error then our method. Another limitation of their method is that it requires some handcrafting to encode the equations as a neural network—the authors state that "the need to translate the existing CASA model into a declarative form that our discovery system can manipulate" is a challenge to their approach. Finally, their method not incorporate the MC principle.

The approach of transforming equation-based models to neural networks and use these for refinement

is similar in spirit to the KBANN approach proposed in Towell and Shavlik (1994). There, an initial theory based on classification rules is first encoded as neural network. Then, the topology of the network is refined and the network is retrained with the newly observed data. Finally, the network is transformed back into rules. However, the application of KBANN is limited to theories and models expressed in a form of classification rules.

Whigham and Recknagel (2000) consider the task of revising an existing model for predicting chlorophyll-a by using measured data. They use a genetic algorithm to calibrate the equation parameters. They also use a grammar-based genetic programming approach to revise the structure of two parts (one at a time) of the initial model. A most general grammar that can derive an arbitrary expression using the allowed arithmetic operators and functions was used for each of the two parts. Unlike the work presented here, Whigham and Recknagel (2000) do not present a general framework for the revision of equation based models. Their approach is similar to ours in that they use grammars to specify possible revisions. However, the grammars they use are too general to provide much information about the domain at hand. Also, they do not incorporate the MC principle in their approach. This can be considered as a weakness of their approach, since genetic programming methods tend to produce large expressions without simplicity bias.

7. Summary and conclusions

In this paper, we have proposed a flexible grammar-based equation discovery approach to the task of revising equation-based models. To support the revision of existing models with equation discovery, we use the transformation principle. First, the given existing model is transformed into an initial grammar that can be used to derive the initial given model only. The non-terminals in the grammar and their productions reflect the structure of the initial model. Domain experts can then focus on the revision process on parts of the model and guide it by providing relevant modeling alternatives that are added to the initial grammar as alternative productions. In this way, the revision process can be interactive, as is quite often the case when revising theories expressed in logic. The method also incorporate the minimality of change principle in a way that allows a trade-off between the goodness of fit of the revised model and its similarity to the initial one.

We have applied our approach to the real-world problem of revising a portion of an existing equation-based model named CASA of the net production of carbon by terrestrial plants in the Earth ecosystem. Experimental results show that small revisions of both the values of the constant parameters and the structure of equations reduce the error of the model considerably (by almost 20%). This improvement is regarded as a non-trivial by Earth scientists that developed the CASA model. Furthermore, the experiments with the improved version of CASA-NPPc model also lead to a revised model that is about 9% more accurate than the initial one. The experiments also show the importance of the minimality of change (MC) principle in model revision from two aspects. First, the use of the MC principle can further reduce the error of the revised model on test data, unused during the process of revision. Second, it can be used to identify the set of most important revisions that lead to largest reduction of the error of the initial model.

References

Džeroski, S., 2001. Applications of symbolic machine learning to ecological modelling. Ecol. Model. 146, 263–273.

Džeroski, S., Todorovski, L., Bratko, I., Kompare, B., Križman, V., 1999. Equation discovery with ecological applications. In: Fielding, A.H. (Ed.), Machine Learning Methods for Ecological Applications. Kluwer, Dordrecht, pp. 185–207.

Hopcroft, J.E., Ullman, J.D., 1979. Introduction to Automata Theory, Languages and Computation. Addison-Wesley, Reading, MA.

Langley, P., Simon, H.A., Bradshaw, G.L., Żythow, J.M., 1987. Scientific Discovery. MIT Press, Cambridge, MA.

Ourston, D., Mooney, R.J., 1994. Theory refinement combining analytical and empirical methods. Artif. Intell. 66, 273–309.

Potter, C.S., Klooster, S.A., 1997. Global model estimates of carbon and nitrogen storage in litter and soil pools: response to change in vegetation quality and biomass allocation. Tellus 49B, 1–17.

Potter, C.S., Klooster, S.A., 1998. Interannual variability in soil trace gas (CO_2, N_2O, NO) fluxes and analysis of controllers on regional to global scales. Global Biogeochem. Cycles 12, 621–635.

Potter, C.S., Klooster, S.A., 1999. Dynamic global vegetation modeling (dgvm) for prediction of plant functional types and biogenic trace gas fluxes. Global Ecol. Biogeogr. Lett. 8, 473–488.

Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterlin, W.T., 1986. Numerical Recipes. Cambridge University Press, Cambridge, MA.

Richter, T., 1997. A new measure of the distance between ordered trees and its applications (technical report). Department of Computer Science IV, University of Bonn, Bonn, Germany.

Saito, K., Langley, P., Grenager, T., Potter, C.S., Torregrosa, A., Klooster, S.A., 2001. Computational revision of quantitative scientific models. In: Proceedings of the Fourth International Conference on Discovery Science. Springer, Berlin, pp. 336–349.

Shasha, D., Zhang, K., 1997. Approximate tree pattern matching. In: Pattern Matching Algorithms. Oxford University Press, London, pp. 341–371.

Todorovski, L., Džeroski, S., 1997. Declarative bias in equation discovery. In: Proceedings of the Fourteenth International Conference on Machine Learning. Morgan Kaufmann, Los Altos, CA, pp. 376–384.

Todorovski, L., Džeroski, S. (2001). Using domain knowledge on population dynamics modeling for equation discovery. In: Proceedings of the Twelfth European Conference on Machine Learning. Springer, Berlin, pp. 478–490.

Todorovski, L., Džeroski, S., Kompare, B., 1998. Modelling and prediction of phytoplankton growth with equation discovery. Ecol. Model. 113, 71–81.

Towell, G.G., Shavlik, J.W., 1994. Knowledge-based artificial neural networks. Artif. Intell. 70, 119–165.

Whigham, P.A., Recknagel, F., 2000. Predicting chlorophyll-a in freshwater lakes by hybridising process-based models and genetic algorithms. Book of Abstracts of the Second International Conference on Applications of Machine Learning to Ecological Modeling, Adelaide University.

Wrobel, S., 1996. First order theory refinement. In: Raedt, L.D. (Ed.), Advances in Inductive Logic Programming. IOS Press, pp. 14–33.