Computational Revision of Quantitative Scientific Models

Kazumi Saito,[1] **Pat Langley**,[2] **Trond Grenager**,[2] **Christopher Potter**,[3] **Alicia Torregrosa**[3], and **Steven A. Klooster**[3]

[1] NTT Communication Science Laboratories
2-4 Hikaridai, Seika, Soraku, Kyoto 619-0237 Japan
saito@cslab.kecl.ntt.co.jp

[2] Computational Learning Laboratory, CSLI
Stanford University, Stanford, California 94305 USA
{langley,grenager}@cs.stanford.edu

[3] Ecosystem Science and Technology Branch
NASA Ames Research Center, MS 242-4
Moffett Field, California 94035 USA
{cpotter,lisy,sklooster}@gaia.arc.nasa.gov

Abstract. Research on the computational discovery of numeric equations has focused on constructing laws from scratch, whereas work on theory revision has emphasized qualitative knowledge. In this paper, we describe an approach to improving scientific models that are cast as sets of equations. We review one such model for aspects of the Earth ecosystem, then recount its application to revising parameter values, intrinsic properties, and functional forms, in each case achieving reduction in error on Earth science data while retaining the communicability of the original model. After this, we consider earlier work on computational scientific discovery and theory revision, then close with suggestions for future research on this topic.

1 Research Goals and Motivation

Research on computational approaches to scientific knowledge discovery has a long history in artificial intelligence, dating back over two decades (e.g., Langley, 1979; Lenat, 1977). This body of work has led steadily to more powerful methods and, in recent years, to new discoveries deemed worth publication in the scientific literature, as reviewed by Langley (1998). However, despite this progress, mainstream work on the topic retains some important limitations.

One drawback is that few approaches to the intelligent analysis of scientific data can use available knowledge about the domain to constrain search for laws or explanations. Moreover, although early work on computational discovery cast discovered knowledge in notations familiar to scientists, more recent efforts have not. Rather, influenced by the success of machine learning and data mining, many researchers have adopted formalisms developed by these fields, such as decision trees and Bayesian networks. A return to methods that operate on established scientific notations seems necessary for scientists to understand their results.

Like earlier research on computational scientific discovery, our general approach involves defining a space of possible models stated in an established scientific formalism, specifically sets of numeric equations, and developing techniques to search that space. However, it differs from previous work in this area by starting from an existing scientific model and using heuristic search to revise the model in ways that improve its fit to observations. Although there exists some research on theory refinement (e.g., Ourston & Mooney 1990; Towell, 1991), it has emphasized qualitative knowledge rather than quantitative models that relate continuous variables, which play a central role in many sciences.

In the pages that follow, we describe an approach to revising quantitative models of complex systems. We believe that our approach is a general one appropriate for many scientific domains, but we have focused our efforts on one area – certain aspects of the Earth ecosystem – for which we have a viable model, existing data, and domain expertise. We briefly review the domain and model before moving on to describe our approach to knowledge discovery and model revision. After this, we present some initial results that suggest our approach can improve substantially the model's fit to available data. We close with a discussion of related discovery work and directions for future research.

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 existing scientific models of the Earth's biosphere. One such model, CASA, developed by Potter and Klooster (1997, 1998) at NASA Ames Research Center, 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

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

NPPc is the net plant production of carbon at a site during the year.
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 ($0.5 < W < 1$) for dry regions.
Topt is the average temperature for the month at which MON-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 the annual heat index for a given site.
MON-FAS-NDVI is the relative vegetation greenness for a given month as measured from space.
IPAR is the energy from the sun that is intercepted by vegetation after factoring in 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 solar irradiance for a given month at a site.
SOL-CONVER is 0.0864 times the number of days in each month.
UMD-VEG is the type of ground cover (vegetation) at a site.

original 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 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 submodel, in which the dependent variable, NPPc, represents the net production of carbon. 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 (0.56) 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. A second factor, T1, involves

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

$$\text{NPPc} = \sum_{month} \max(\text{E} \cdot \text{IPAR}, 0)$$
$$\text{E} = 0.56 \cdot \text{T1} \cdot \text{T2} \cdot \text{W}$$
$$\text{T1} = 0.8 + 0.02 \cdot \text{Topt} - 0.0005 \cdot \text{Topt}^2$$
$$\text{T2} = 1.18/[(1 + e^{0.2 \cdot (\text{Topt} - \text{Tempc} - 10)}) \cdot (1 + e^{0.3 \cdot (\text{Tempc} - \text{Topt} - 10)})]$$
$$\text{W} = 0.5 + 0.5 \cdot \text{EET/PET}$$
$$\text{PET} = 1.6 \cdot (10 \cdot \text{Tempc} \ / \ \text{AHI})^A \cdot \text{PET-TW-M if Tempc} > 0$$
$$\text{PET} = 0 \text{ if Tempc} \leq 0$$
$$\text{A} = 0.000000675 \cdot \text{AHI}^3 - 0.0000771 \cdot \text{AHI}^2 + 0.01792 \cdot \text{AHI} + 0.49239$$
$$\text{IPAR} = 0.5 \cdot \text{FPAR-FAS} \cdot \text{MONTHLY-SOLAR} \cdot \text{SOL-CONVER}$$
$$\text{FPAR-FAS} = \min((\text{SR-FAS} - 1.08)/\text{SRDIFF(UMD-VEG)}, 0.95)$$
$$\text{SR-FAS} = - (\text{MON-FAS-NDVI} + 1000) \ / \ (\text{MON-FAS-NDVI} - 1000)$$

the nearness of Topt to a global optimum for all sites, reflecting the intuition that plants which are better adapted to harsh temperatures are less efficient overall. 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, another component of potential evapotranspiration.

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-CONVER, the number of days in that month. FPAR-FAS is a function of MON-FAS-NDVI, which indicates relative greenness at a site as observed from space, and SRDIFF, an intrinsic property that takes on different numeric values for different vegetation types as specified by the discrete variable UMD-VEG.

Of the variables we have mentioned, NPPc, Tempc, MONTHLY-SOLAR, SOL-CONVER, MON-FAS-NDVI, and UMD-VEG are observable. Three additional terms – EET, PET-TW-M, and AHI – 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 An Approach to Quantitative Model Revision

As noted earlier, our approach to scientific discovery involves refining models like CASA that involve relations among quantitative variables. We adopt the traditional view of discovery as heuristic search through a space of models, with the search process directed by candidates' ability to fit the data. However, we assume this process starts not from scratch, but rather with an existing model,

and the search operators involve making changes to this model, rather than constructing entirely new structures.

Our long-term goal is not to automate the revision process, but instead to provide an interactive tool that scientists can direct and use to aid their model development. As a result, the approach we describe in this section addresses the task of making local changes to a model rather than carrying out global optimization, as assumed by Chown and Dietterich (2000). Thus, our software takes as input not only observations about measurable variables and an existing model stated as equations, but also information about which portion of the model should be altered. The output is a revised model that fits the observed data better than the initial one.

Below we review two discovery algorithms that we utilize to improve the specified part of a model, then describe three distinct types of revision they support. We consider these in order of increasing complexity, starting with simple changes to parameter values, moving on to revisions in the values of intrinsic properties, and ending with changes in an equation's functional form.

3.1 The RF5 and RF6 Discovery Algorithms

Our approach relies on RF5 and RF6, two algorithms for discovering numeric equations described Saito and Nakano (1997, 2000). Given data for some continuous variable y that is dependent on continuous predictive variables $x_1, \ldots, x_n$, the RF5 system searches for multivariate polynomial equations of the form

$$y = w_0 + \sum_{j=1}^{J} w_j \prod_{k=1}^{K} x_k^{w_{jk}} = w_0 + \sum_{j=1}^{J} w_j \exp\left(\sum_{k=1}^{K} w_{jk} \ln(x_k)\right), \quad (1)$$

Such functional relations subsume many of the numeric laws found by previous computational discovery systems like BACON (Langley, 1979) and FAHRENHEIT (Żytkow, Zhu, & Hussam, 1990).

RF5's first step involves transforming a candidate functional form with J summed terms into a three-layer neural network based on the rightmost form of expression (1), in which the K hidden nodes in this network correspond to *product units* (Durbin & Rumelhart, 1989). The system then carries out search through the weight space using the BPQ algorithm, a second-order learning technique that calculates both the descent direction and the step size automatically.

This process halts when it finds a set of weights that minimize the squared error on the dependent variable y. RF5 runs the BPQ method on networks with different numbers of hidden units, then selects the one that gives the best score on an MDL metric. Finally, the program transforms the resulting network into a polynomial equation, with weights on hidden units becoming exponents and other weights becoming coefficients.

The RF6 algorithm extends RF5 by adding the ability to find conditions on a numeric equation that involve nominal variables, which it encodes using one input variable for each nominal value. To this end, the system first generates one such condition for each training case, then utilizes k-means clustering to generate

a smaller set of more general conditions, with the number of clusters determined through cross validation. Finally, RF6 invokes decision-tree induction to construct a classifier that discriminates among these clusters, which it transforms into rules that form the nominal conditions on the polynomial equation that RF5 has generated.

3.2 Three Types of Model Refinement

There exist three natural types of refinement within the class of models, like CASA, that are stated as sets of equations that refer to unobservable variables. These include revising the parameter values in equations, altering the values for an intrinsic property, and changing the functional form of an existing equation.

Improving the parameters for an equation is the most straightforward process. The NPPc portion of CASA contains some parameterized equations that our Earth science team members believe are reliable, like that for computing the variable A from AHI, the annual heat index. However, it also includes equations with parameters about which there is less certainty, like the expression that predicts the temperature stress factor T2 from Tempc and Topt. Our approach to revising such parameters relies on creating a specialized neural network that encodes the equation's functional form using ideas from RF5, but also including a term for the unchanged portion of the model. We then run the BPQ algorithm to find revised parameter values, initializing weights based on those in the model.

We can utilize a similar scheme to improve the values for an intrinsic property like SRDIFF that the model associates with the discrete values for some nominal variable like UMD-VEG (vegetation type). We encode each nominal term as a set of dummy variables, one for each discrete value, making the dummy variable equal to one if the discrete value occurs and zero otherwise. We introduce one hidden unit for the intrinsic property, with links from each of the dummy variables and with weights that correspond to the intrinsic values associated with each discrete value. To revise these weights, we create a neural network that incorporates the intrinsic values but also includes a term for the unchanging parts of the model. We can then run BPQ to revise the weights that correspond to intrinsic values, again initializing them to those in the initial model.

Altering the form of an existing equation requires somewhat more effort, but maps more directly onto previous work in equation discovery. In this case, the details depend on the specific functional form that we provide, but because we have available the RF5 and RF6 algorithms, the approach supports any of the forms that they can discover or specializations of them. Again, having identified a particular equation that we want to improve, we create a neural network that encodes the desired form, then invoke the BPQ algorithm to determine its parametric values, in this case initializing the network weights randomly.

This approach to model refinement supports changes to only one equation or intrinsic property at a time, but this is consistent with the interactive process described earlier. We envision the scientist identifying a portion of the model that he thinks could be better, running one of the three revision methods to improve its fit to the data, and repeating this process until he is satisfied.

4 Initial Results on Ecosystem Data

In order to evaluate our approach to scientific model revision, we utilized data relevant to the NPPc model available to the Earth science members of our team. These data consisted of observations from 303 distinct sites with known vegetation type and for which measurements of Tempc, MON-FAS-NDVI, MONTHLY-SOLAR, SOL-CONVER, and UMD-VEG were available for each month during the year. In addition, other portions of CASA were able to compute values for the variables AHI, EET, and PET-TW-M. The resulting 303 training cases seemed sufficient for initial tests of our revision methods, so we used them to drive a variety of changes to the handcrafted model of carbon production.

4.1 Results on Parameter Revision

Our Earth science team members identified the equation for T2, one of the temperature stress variables, as a likely candidate for revision. As noted earlier, the handcrafted expression for this term was

$$T2 = 1.8/[(1 + e^{0.2(Topt - Tempc - 10)})(1 + e^{-0.3(Tempc - Topt - 10)})] \ ,$$

which produces a Gaussian-like curve that is slightly assymetrical. This reflects the intuition that photosynthetic efficiency will decrease when temperature (Tempc) is either below or above the optimal (Topt).

To improve upon this equation, we defined $x = \text{Topt} - \text{Tempc}$ as an intermediate variable and recast the expression for T2 as the product of two sigmoidal functions of the form $\sigma(a) = 1/(1 + \exp(-a))$ and a parameter. We transformed these into a neural network and used BPQ to minimize the error function

$$\mathcal{F}_1 = \sum_{sample} (\text{NPPc} - \sum_{month} w_0 \cdot \sigma(v_{10} + v_{11} \cdot x) \cdot \sigma(v_{20} - v_{21} \cdot x) \cdot \text{Rest})^2 \ ,$$

over the parameters $\{w_0, v_{10}, v_{11}, v_{20}, v_{21}\}$, where $\text{Rest} = 0.56 \cdot \text{T1} \cdot \text{W} \cdot \text{IPAR}$. The resulting equation generated in this manner was

$$T2 = 1.80/[(1 + e^{0.05(Topt - Tempc - 10.8})(1 + e^{-0.03(Tempc - Topt - 90.33})] \ ,$$

which has reasonably similar values to the original ones for some parameters but quite different values for others.

The root mean squared error (RMSE) for the original model on the available data was 467.910. In contrast, the error for the revised model was 457.757 on the training data and 461.466 using leave-one-out cross validation. Thus, RF6's modification of parameters in the T2 equation produced slightly more than one percent reduction in overall model error, which is somewhat disappointing.

However, inspection of the resulting curves reveals a more interesting picture. Plotting the temperature stress factor T2 using the revised equations as a function of the difference Topt − Tempc still gives a Gaussian-like curve, but within the effective range (from −30 to 30 Celsius) its values decrease monotonically. This seems counterintuitive but interesting from an Earth science perspective,

as it suggests this stress factor has little influence on NPPc. Moreover, the original equation for T2 was not well grounded in first principles of plant physiology, making empirical improvements of this sort beneficial to the modeling enterprise.

As another candidate for parameter revision, we selected the PET equation,

$$\text{PET} = 1.6 \cdot (10 \cdot \max(\text{Tempc}, 0) \ / \ \text{AHI})^{A} \cdot \text{PET-TW-M} \ ,$$

which calculates potential water loss due to evaporation and transpiration given an unlimited water supply. By transforming this expression into

$$\text{PET} = \exp(\ln(1.6) + \text{A} \cdot \ln(10)) \cdot (\max(\text{Tempc}, 0) \ / \ \text{AHI})^{A} \cdot \text{PET-TW-M}$$

and replacing the parameter values ln(1.6) and ln(10) with the variables v_0 and v_1, we constructed a neural network and used BPQ for error minimization. When transforming the trained network back into the original form, the equation that resulted was

$$\text{PET} = 1.56 \cdot (9.16 \cdot \max(\text{Tempc}, 0) \ / \ \text{AHI})^{A} \cdot \text{PET-TW-M} \ ,$$

which has values that are very similar to those in the original model's equation.

Moreover, since the RMSE for the obtained model was 464.358 on the training data and 467.643 using leave-one-out cross validation, the revision process did not improve the model's accuracy substantially. However, since the PET equation is based on Thornthwaite's (1948) method, which has been used continuously for over 50 years, we should not be overly surprised at this negative result. Indeed, we are encouraged by the fact that our approach did not revise parameters that have stood the test of time in Earth science.

4.2 Results on Intrinsic Value Revision

Another portion of the NPPc model that held potential for revision concerns the intrinsic property SRDIFF associated with the vegetation type UMD-VEG. For each site, the latter variable takes on one of 11 nominal values, such as grasslands, forest, tundra, and desert, each with an associated numeric value for SRDIFF that plays a role in the FPAR-FAS equation. This gives 11 parameters to revise, which seems manageable given the number of observations available.

As outlined earlier, to revise these intrinsic values, we introduced one dummy variable, UMD-VEG$_k$, for each vegetation type such that UMD-VEG$_k$ = 1 if UMD-VEG = k and 0 otherwise. We then defined SRDIFF(UMD-VEG) as $\exp(-\sum_k v_k \cdot \text{UMD-VEG}_k)$ and, since SRDIFF's value is independent of the month, we used BPQ to minimize, over the weights $\{v_k\}$, the error function

$$\mathcal{F}_2 = \textstyle\sum_{site} (\text{NPPc} - \exp(\sum_k v_k \cdot \text{UMD-VEG}_k) \cdot \text{Rest})^2 \ ,$$

where $\text{Rest} = \sum_{month} E \cdot 0.5 \cdot (\text{SR-FAS} - 1.08) \cdot \text{MONTHLY-SOLAR} \cdot \text{SOL-CONVER}$.

Table 3 shows the initial values for this intrinsic property, as set by the CASA developers, along with the revised values produced by the above approach when

Table 3. Original and revised values for the SRDIFF intrinsic property, along with the frequency for each vegetation type.

vegetation type	A	B	C	D	E	F	G	H	I	J	K
original	3.06	4.35	4.35	4.05	5.09	3.06	4.05	4.05	4.05	5.09	4.05
revised	2.57	4.77	2.20	3.99	3.70	3.46	2.34	0.34	2.72	3.46	1.60
clustered	2.42	3.75	2.42	3.75	3.75	3.75	2.42	0.34	2.42	3.75	2.42
frequency	3.3	8.9	0.3	3.6	21.1	19.1	15.2	3.3	19.1	2.3	3.6

we fixed other parts of the NPPc model. The most striking result is that the revised intrinsic values are nearly always lower than the initial values. The RMSE for the original model was 467.910, whereas the error using the revised values was 432.410 on the training set and 448.376 using cross validation. The latter constitutes an error reduction of over four percent, which seems substantial.

However, since the original 11 intrinsic values were grouped into only four distinct values, we applied RF6's clustering procedure over the trained neural network to group the revised values in the same manner. We examined the effect on error rate as we varied the number of clusters from one to five; as expected, the training RMSE decreased monotonically, but the cross-validation RMSE was minimized for three clusters of values. The estimated error for this revised model is slightly better than for the one with 11 distinct values.

Again, the clustered values are nearly always lower than the initial ones, a result that is certainly interesting from an Earth science viewpoint. We suspect that measurements of NPPc and related variables from a wider range of sites would produce intrinsic values closer to those in the original model. However, such a test must await additional observations and, for now, empirical fit to the available data should outweigh the theoretical basis for the initial settings.

In another approach to revising intrinsic values, we retained the original grouping of vegetation types into sets, with each type in a given set having the same value. We utilized a weight-sharing technique to encode this background knowledge in a neural network. For example, let v_A and v_F be weights corresponding to the SRDIFF values for vegetation types A and F, respectively; to ensure these values remained the same, we treated them as a single weight, say v_{AF}. Here we can see that BPQ calculates the derivative of the error function over v_{AF} as a sum of the individual derivatives over v_A and v_F,

$$\frac{\partial \mathcal{F}_2}{\partial v_{AF}} = \frac{\partial \mathcal{F}_2}{\partial v_A} + \frac{\partial \mathcal{F}_2}{\partial v_F} \, .$$

In the trained neural network, the derivative over v_{AF} becomes zero, but there is no guarantee that each derivative over v_A or v_F will do so. Therefore, we can treat the sum of the absolute values for derivatives over shared weights, like v_A and v_F, as a criterion for the 'unlikeness' among the elements of such a grouping.

Table 4 shows the revised values for the intrinsic property SRDIFF that result from this approach, along with values for the unlikeness criterion defined above.

Table 4. Original and revised values, using the original groupings, for the SRDIFF intrinsic property, along with the frequency and unlikeness for each vegetation group.

vegetation type	A∨F	B∨C	E∨J	D∨G∨H∨I∨K
original	3.06	4.35	5.09	4.05
revised	2.23	3.27	2.54	1.81
frequency	22.4	9.2	23.4	44.9
unlikeness	26.1	0.3	2.3	13.6

As before, the obtained intrinsic values are always lower than the initial ones, and our criterion suggests that the group containing the vegetation types A and F has the least coherence. The RMSE for the revised model was 442.782 on the training data and 449.097 using leave-one-out cross validation, again indicating about four percent reduction in the model's overall error.

4.3 Results on Revising Equation Structure

We also wanted to demonstrate our approach's ability to improve the functional form of the NPPc model. For this purpose, we selected the equation for photosynthetic efficiency,

$$E = 0.56 \cdot T1 \cdot T2 \cdot W \ ,$$

which states that this term is a product of the water stress term, W, and the two temperature stress terms, T1 and T2. Because each stress factor takes on values less than one, multiplication has the effect of reducing photosynthetic efficiency E below the maximum 0.56 possible (Potter & Klooster, 1998).

Since E is calculated as a simple product of the three variables, one natural extension was to consider an equation that included exponents on these terms. To this end, we borrowed techniques from the RF5 system to create a neural network for such an expression, then used BPQ to minimize the error function

$$\mathcal{F}_3 = \sum_{site} \left(\mathrm{NPPc} - \sum_{month} u_0 \cdot \mathrm{T1}^{u_1} \cdot \mathrm{T2}^{u_2} \cdot \mathrm{W}^{u_3} \cdot \mathrm{IPAR}\right)^2 \ ,$$

over the parameters $\{u_0, u_1, u_2, u_3\}$, which assumes the equations that predict IPAR remain unchanged. We initialized u_0 to 0.56 and the other parameters to 1.0, as in the original model, and constrained the latter to be positive. The revised equation found in this manner,

$$E = 0.521 \cdot T1^{0.00} \cdot T2^{0.03} \cdot W^{0.00} \ ,$$

has a small exponent for T2 and zero exponents for T1 and W, suggesting the former influences photosynthetic efficiency in minor ways and the latter not at all. On the available data, the root mean squared error for the original model was 467.910. In contrast, the revised model has an RMSE of 443.307 on the training set and an RMSE of 446.270 using cross validation. Thus, the revised

equation produces a substantially better fit to the observations than does the original model, in this case reducing error by almost five percent.

With regards to Earth science, these results are plausible and the most interesting of all, as they suggest that the T1 and W stress terms are unnecessary for predicting NPPc. One explanation is that the influence of these factors is already being captured by the NDVI measure available from space, for which the signal-to-noise ratio has been steadily improving since CASA was first developed.

These results encouraged us to explore more radical revisions to the functional form for photosynthetic efficiency. Thus, we told our system to consider a form that omitted the three stress factors but that included the four variables – Topt, Tempc, EET, and PET – that appear in their definitions:

$$\mathrm{E} = v_0 \cdot \exp(-0.5 \cdot (v_1 \cdot \mathrm{Topt} + v_2 \cdot \mathrm{Tempc} + v_3 \cdot \mathrm{EET} + v_4 \cdot \mathrm{PET} + v_5)^2) \ .$$

This Gaussian-like activation function satisfies the constraint that E is positive and less than one. Running BPQ to minimize the error function over $\{v_0, \ldots v_5\}$ produced the equation

$$\mathrm{E} = 0.57 \cdot \exp(-0.5 \cdot (-0.04 \cdot \mathrm{Topt} + 0.03 \cdot \mathrm{Tempc} - 0.03 \cdot \mathrm{EET} + 0.01 \cdot \mathrm{PET})^2),$$

where we eliminated the parameter v_5 because its value was -0.003. The RMSE for the revised model was 439.101 on the training data and 444.470 using leave-one-out cross validation, indicating more than five percent reduction in error.

These results are very similar to those from our first approach, which produced a cross validation RMSE of 446.270. In this case, the revised model is simpler in that it defines E directly in terms of Topt, Tempc, EET, and PET, rather than relying on the theoretical terms T1, T2, and W, two of which provide no predictive power. On the other hand, the original form for E had a clear theoretical interpretation, whereas the new version does not. In such situations, the final decision should be left to domain scientists, who are best suited to balance a model's simplicity against its interepretability.

5 Related Research on Computational Discovery

Our research on computational scientific discovery draws on two previous lines of work. One approach, which has an extended history within artificial intelligence, addresses the discovery of explicit quantitative laws. Early systems for numeric law discovery like BACON (Langley, 1979; Langley et al., 1987) carried out a heuristic search through a space of new terms and simple equations. Numerous successors like FAHRENHEIT (Żytkow et al., 1990) and RF5 (Saito & Nakano, 1997) incorporate more sophisticated and more extensive search through a larger space of numeric equations.

The most relevant equation discovery systems take into account domain knowledge to constrain the search for numeric laws. For example, Kokar's (1986) COPER utilized knowledge about the dimensions of variables to focus attention and, more recently, Washio and Motoda's (1998) SDS extends this idea to support different types of variables and sets of simultaneous equations. Todorovski

and Džeroski's (1997) LaGramge takes a quite different approach, using domain knowledge in the form of context-free grammars to constrain its search through a space of differential equation models that describe temporal behavior.

Although research on computational discovery of numeric laws has emphasized communicable scientific notations, it has focused on constructing such laws rather than revising existing ones. In contrast, another line of research has addressed the refinement of existing models to improve their fit to observations. For example, Ourston and Mooney (1990) developed a method that used training data to revise models stated as sets of propositional Horn clauses. Towell (1991) reports another approach that transforms such models into multilayer neural networks, then uses backpropagation to improve their fit to observations, much as we have done for numeric equations. Work in this paradigm has emphasized classification rather than regression tasks, but one can view our work as adapting the basic approach to equation discovery.

We should also mention related work on the automated improvement of ecosystem models. Most AI work on Earth science domains focuses on learning classifiers that predict vegetation from satellite measures like NDVI, as contrasted with our concern for numeric prediction. Chown and Dietterich (2000) describe an approach that improves an existing ecosystem model's fit to continuous data, but their method only alters parameter values and does not revise equation structure. On another front, Schwabacher and Langley (2001) use a rule-induction algorithm to discover piecewise linear models that predict NDVI from climate variables, but their method takes no advantage of existing models.

6 Directions for Future Research

Although we have been encouraged by our results to date, there remain a number of directions in which we must extend our approach before it can become a useful tool for scientists. As noted earlier, we envision an interactive discovery aide that lets the user focus the system's attention on those portions of the model it should attempt to improve. To this end, we need a graphical interface that supports marking of parameters, intrinsic properties, and equations that can be revised, as well as tools for displaying errors as a function of space, time, and predictive variables.

In addition, the current system is limited to revising the parameters or form of one equation in the model at a time, as well as requiring some handcrafting to encode the equations as a neural network. Future versions should support revisions of multiple equations at the same time, preferably invoking the same variants of backpropagation as we have used to date, and also provide a library that maps functional forms to neural network encodings, so the system can transform the former into the latter automatically. We should also explore using other approaches to equation discovery, such as Todorovski and Džeroski's LaGramge, in place of the RF6 algorithm.

Naturally, we also hope to evaluate our approach on its ability to improve other portions of the CASA model, as additional data becomes available. Another test of generality would be application of the same methods to other sci-

entific domains in which there already exist formal models that can be revised. In the longer term, we should evaluate our interactive system not only in its ability to increase the predictive accuracy of an existing model, but in terms of the satisfaction to scientists who use the system to that end.

Another challenge that we have encountered in our research has been the need to translate the existing CASA model into a declarative form that our discovery system can manipulate. In response, another long-term goal involves developing a modeling language in which scientists can cast their initial models and carry out simulations, but that can also serve as the declarative representation for our discovery methods. The ability to automatically revise models places novel constraints on such a language, but we are confident that the result will prove a useful aid to the discovery process.

7 Concluding Remarks

In this paper, we addressed the computational task of improving an existing scientific model that is composed of numeric equations. We illustrated this problem with an example model from the Earth sciences that predicts carbon production as a function of temperature, sunlight, and other variables. We identified three activities that can improve a model – revising an equation's parameters, altering the values of an intrinsic property, and changing the functional form of an equation, then presented results for each type on an ecosystem modeling task that reduced the model's prediction error, sometimes substantially.

Our research on model revision builds on previous work in numeric law discovery and qualitative theory refinement, but it combines these two themes in novel ways to enable new capabilities. Clearly, we remain some distance from our goal of an interactive discovery tool that scientists can use to improve their models, but we have also taken some important steps along the path, and we are encouraged by our initial results on an important scientific problem.

References

Chown, E., & Dietterich, T. G. (2000). A divide and conquer approach to learning from prior knowledge. *Proceedings of the Seventeenth International Conference on Machine Learning* (pp. 143–150). San Francisco: Morgan Kaufmann.

Durbin, R. & Rumelhart, D. E. (1989). Product units: A computationally powerful and biologically plausible extension. *Neural Computation*, *1*, 133–142.

Kokar, M. M. (1986). Determining arguments of invariant functional descriptions. *Machine Learning*, *1*, 403–422.

Langley, P. (1979). Rediscovering physics with BACON.3. *Proceedings of the Sixth International Joint Conference on Artificial Intelligence* (pp. 505–507). Tokyo, Japan: Morgan Kaufmann.

Langley, P. (1998). The computer-aided discovery of scientific knowledge. *Proceedings of the First International Conference on Discovery Science*. Fukuoka, Japan: Springer.

Langley, P., Simon, H. A., Bradshaw, G. L., & Żytkow, J. M. (1987). *Scientific discovery: Computational explorations of the creative processes*. Cambridge, MA: MIT Press.

Lenat, D. B. (1977). Automated theory formation in mathematics. *Proceedings of the Fifth International Joint Conference on Artificial Intelligence* (pp. 833–842). Cambridge, MA: Morgan Kaufmann.

Ourston, D., & Mooney, R. (1990). Changing the rules: A comprehensive approach to theory refinement. *Proceedings of the Eighth National Conference on Artificial Intelligence* (pp. 815–820). Boston: AAAI Press.

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 Biogeochemical Cycles*, *12*, 621–635.

Saito, K., & Nakano, R. (1997). Law discovery using neural networks. *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence* (pp. 1078–1083). Yokohama: Morgan Kaufmann.

Saito, K., & Nakano, R. (2000). Discovery of nominally conditioned polynomials using neural networks, vector quantizers and decision trees. *Proceedings of the Third International Conference on Discovery Science* (pp. 325–329). Kyoto: Springer.

Schwabacher, M., & Langley, P. (2001). Discovering communicable scientific knowledge from spatio-temporal data. *Proceedings of the Eighteenth International Conference on Machine Learning* (pp. 489–496). Williamstown: Morgan Kaufmann.

Thornthwaite, C. W. (1948) An approach toward rational classification of climate. *Geographic Review*, *38*, 55–94.

Todorovski, L., & Džeroski, S. (1997). Declarative bias in equation discovery. *Proceedings of the Fourteenth International Conference on Machine Learning* (pp. 376–384). San Francisco: Morgan Kaufmann.

Towell, G. (1991). *Symbolic knowledge and neural networks: Insertion, refinement, and extraction.* Doctoral dissertation, Computer Sciences Department, University of Wisconsin, Madison.

Washio, T. & Motoda, H. (1998). Discovering admissible simultaneous equations of large scale systems. *Proceedings of the Fifteenth National Conference on Artificial Intelligence* (pp. 189–196). Madison, WI: AAAI Press.

Żytkow, J. M., Zhu, J., & Hussam, A. (1990). Automated discovery in a chemistry laboratory. *Proceedings of the Eighth National Conference on Artificial Intelligence* (pp. 889–894). Boston, MA: AAAI Press.