Inductive Revision of
Quantitative Process Models

Nima Asgharbeysi,¹ Stephen Bay,¹ Pat Langley,¹ and Kevin Arrigo²

¹Computational Learning Laboratory
   Center for Study of Language and Information
   Stanford University, Stanford, CA 94305 USA

²Department of Geophysics
   Stanford University, Stanford, CA 94305 USA

Abstract. Most research on computational scientific discovery has focused on developing an initial model, but an equally important task involves revising a model in response to new data. In this paper, we present an approach that represents candidate models as sets of quantitative processes and that treats revision as search through a model space which is guided by time-series observations and constrained by background knowledge cast as generic processes. We demonstrate our system's ability on two different scientific domains and associated data sets. We also discuss its relation to other work on model revision and consider directions for additional research.

1 Introduction and Motivation

Most research on computational scientific discovery has emphasized the generation of entirely new models to describe or explain data. However, human scientists spend much of their time not developing new accounts of phenomena but rather revising and improving an existing model that already has credibility. If we desire computational tools that help scientists in practice, we should examine seriously issues that arise in the revision of scientific models.

There are other excellent reasons for focusing on model revision rather than generation. Most approaches to scientific discovery carry out search through the space of hypotheses or models, which can be quite large. By starting from an existing candidate, we can reduce the effective size of this space, making some tasks tractable that would otherwise be too difficult. This approach also places constraints on the search mechanism that can reduce variance and thus decrease chances of overfitting the data. Finally, model revision gives the domain user more control over the region of the model space explored, and increases the chances that he will find the new model comprehensible.

In this paper, we report one approach to the problem of scientific model revision. Unlike earlier work on this topic, we focus on the modification of quantitative process models, a representation of knowledge that we believe offers additional advantages related to search and interpretability. We review this formalism briefly in the section that follows, along with RPM, an algorithm that
revises an initial process model in response to time-series data. After this, we report experimental results on two environmental domains, one involving water behavior in a Danish fjord and the other concerning phytoplankton growth in an Antarctic sea. In closing, we discuss related work on model revision and outline directions for future research.

2 Process Models and their Revision

In this section, we review our previous work on process models and their induction from time-series data and background knowledge. After this, we present our approach to revising models within the process-modeling framework.

2.1 Quantitative Process Models

Scientific models are often stated formally as sets of equations, but they are also described informally in terms of the processes that determine those equations. We have developed the formalism of quantitative process models to encode both aspects of scientific knowledge. In this framework, a model consists of a set of processes, each of which specifies one or more equations that represent causal relations among variables. These are cast as algebraic equations for instantaneous effects or differential equations for changes over time. Processes can include threshold conditions on variables that characterize when they are active.

A process model specifies not only a set of processes, but also the variables which they connect. A given variable may be labeled as observable, meaning it is present in the data, or it may play the role of a theoretical term that serves mainly to link processes. Each variable may also be labeled as exogenous, in that it influences other variables but is not influenced in return, or as endogenous, which means it appears in the left side of one or more equations. Our framework is a quantitative variant of Forbus' (1984) qualitative process theory, from which we have borrowed many ideas.

Table 1 shows a process model for the aquatic ecosystem of the Ross Sea in Antarctica, which is based on an earlier differential equation model developed by the Earth scientist member of our team (Arrigo et al., 2003). The model includes four observables - the exogenous variables light and ice factor and the endogenous concentrations of phytoplankton (phyto) and nitrate. The variable ice factor represents the fraction of the sea not covered by ice. The first process characterizes loss of phyto due to miscellaneous sources, along with an increase in residue, whereas the second specifies growth of phyto as a function of its current concentration and its growth rate. The third process concerns the decrease in nitrate due to uptake by phytoplankton. Next, growth limitation indicates that growth rate is the maximum rate \( r_{\text{max}} \) times the minimum of two theoretical terms, nitrate rate and light rate, which are functions of nitrate and light. The light attenuation process states that ice factor influences the light reaching the sea. A final process specifies two parameters that occur across processes.

We can utilize a process model of this sort, together with initial values, to simulate the model's behavior over time and thus predict values for each endogenous variable. We have implemented a simulator that determines which
Table 1. A process model for the Ross Sea ecosystem.

```plaintext
model Ross_Sea_Ecosystem;
variables phyto, residue, nitrate, light, growth_rate, effective_light, ice_factor,
    nitrate_rate, light_rate, n_fo_c_ratio, r_max;
observable phyto, nitrate, light, ice_factor;
exogenous light, ice_factor;
process phyto_loss;
equations d[phyto,t,1] = -0.1 * phyto;
d[residue,t,1] = 0.1 * phyto;
process phyto_growth;
equations d[phyto,t,1] = growth_rate * phyto;
process phyto_uptakes_nitrate;
equations d[nitrate,t,1] = -1 * n_fo_c_ratio * growth_rate * phyto;
process growth_limitation;
equations growth_rate = r_max * min(nitrate_rate, light_rate);
process nitrate_availability;
equations nitrate_rate = nitrate/(nitrate + 5);
process light_availability;
equations light_rate = effective_light/(effective_light + 50);
process light_attenuation;
equations effective_light = light * ice_factor;
process global_parameters;
equations n_fo_c_ratio = 0.204;
r_max = 0.23;
```

processes are active on each time step and then invokes standard methods for solving ordinary differential equations to generate predicted trajectories. Later we present trajectories that a revised version of this model predicts for phyto and nitrate, along with observations for the same variables.

Process models provide an explanation of observations in that they offer a causal account in terms of processes and equations that are familiar to domain specialists. For example, Table 2 presents some generic processes relevant to aquatic ecosystems that serve as background knowledge. These differ from specific processes in that they do not commit to particular variables or parameter values. However, they can indicate constraints, such as stating that the variable P in the generic process grazing must have type p_species and that its coefficient gamma must fall between 0 and 1. The table also specifies that the same parameter must appear in multiple equations within some processes.

In a previous paper (Langley et al., 2003), we proposed the task of inducing process models like the one in Table 1 from time-series data and from background knowledge like the generic processes in Table 2. In order to address this task we developed an initial algorithm, called IPM, that carries out exhaustive
search through the entire space of process models. The system then selects the parametrized model which produces the best score on an evaluation criterion that incorporates both error and model complexity.

Experiments with IPM produced encouraging results on real-world data collected from batteries on the Space Station, but our studies with environmental models, which had motivated our research on process model induction, dealt only with synthetic data and involved a target model with only five processes. The availability of both new environmental data sets and codification of additional ecosystem processes has encouraged us to extend the IPM framework and evaluate its behavior in this new context.

### 2.2 The RPM Revision Algorithm

Although IPM produced promising results, it had drawbacks that limited its applicability. The system constrained its search space by utilizing background knowledge about generic processes, but the space of models could still be large. Also, IPM provided no way to guide the search toward models a scientist might find more plausible. As argued earlier, model revision seems an appropriate response to both issues, so we developed an extended system, RPM, that adopts this approach to process model induction. This revision module is a key component in an integrated environment that we are developing to aid scientists in developing and improving their models. This environment assumes that models are cast as sets of quantitative processes and that generic processes are available as background knowledge.

The RPM algorithm requires the user to specify four inputs. These include: an initial model that encodes beliefs about the processes that are most likely
involved; a set of constraints representing allowed changes to the initial model that specify which initial processes should be fixed, can be removed, or have their parameters changed; a set of generic processes that may be added to the initial model; and observations to which the revised model should be fit. The initial model constitutes the user’s best guess about the processes that are present in the system, whereas the allowed changes indicate his areas of uncertainty. Combined with the candidate additions, these provide RPM with a heuristic that guides search toward parts of the model space that are consistent with domain knowledge.

As output, the algorithm generates a set of revised models that are sorted by their distance from the initial model and presented with their mean squared error on the training data. The distance between a revised model and the initial model is defined as the number of processes that are present in one but not in the other. This output format lets one observe the trade-off between performance of revised models and their similarity to the initial model. This leaves the user to determine the best compromise between the factors and to select an appropriate model from the suggested set.

The RPM system operates in two main stages. The first involves searching through the model space and finding all model structures that are consistent with the specified constraints, including user-approved changes. The system first generates all instantiations of the user-recommended generic processes that satisfy constraints on variable types; these become candidates for addition to the model. Next, RPM carries out search through the space of model structures, using the initial model as the start state. The search method utilizes two operators: adding a process from the set of instantiated generic processes and removing a process from the initial model. The current implementation uses breadth-first search so that models closer to the initial model are considered first. The algorithm also performs sanity checks on each candidate that ensure it forms a single connected graph and includes all observable variables. The result is a set of revised model structures that explain relations among the variables.

The second stage determines, for each model structure, the parameter values for new processes and ones allowed to change. To this end, RPM utilizes a combination of the Levenberg-Marquardt method interleaved with randomized jumps. The search algorithm starts by selecting a random initial point that falls within the parameter ranges specified in the generic processes. The algorithm then attempts to optimize the parameters with the Levenberg-Marquardt routine until it converges to a local optimum. RPM then generates several new candidates by positing random jumps along the dimensions of the parameter vector. If a jump leads to lower error, it moves to that point and returns to the Levenberg-Marquardt method; otherwise, the system repeatedly generates new candidates and gradually increases the jump size. However, if RPM observes no improvement after a specified number of iterations (we used a maximum of 20), it restarts the entire process from a new random initial point. We have found that this parameter-fitting method gives enough flexibility to produce reasonable matches to time series from a variety of domains.
model Fjord_Height_Dynamics;
variables H, hsea, n, Q, Wdir, Wvel, WaterFlow, hf, hw;
observable H, hsea, n, Q, Wdir, Wvel;
exogenous hsea, n, Q, Wdir, Wvel;
process waterFlowThroughGates;
equations WaterFlow = -1000 * 0.5 * n * (H - hsea);
process freshWaterInput;
equations WaterFlow = Q;
process flowHeightRelation;
equations d[hf,t,1] = 86400 * WaterFlow / A(H);
process windForcingSine;
equations d[hw,t,1] = 0.05 * Wvel * sin(Wdir);
process totalHeight;
equations H = hf + hw;

3 Experimental Evaluation

We decided to evaluate RPM by applying it to two actual modeling problems: predicting the height of water within a Danish fjord and modeling phytoplankton dynamics in the Antarctic Ocean. Naturally, we were interested in whether the revised models constituted improvements over the initial ones and in whether they generalized well to separate test sets.

3.1 Water Dynamics in Ringkobing Fjord

The Ringkobing Fjord, on the Danish west coast, is a shallow body of water that is fed by tributaries from the land and that exchanges water with the North Sea through a narrow channel on its western edge. A barrier with 14 gates has been constructed across this channel in order to regulate water flow between the fjord and the sea. Officials would like to predict in advance the water level at the gates, so they can be ready to open or close them as needed.

When the gates are closed, the dominant influences on the water level H inside the gates to the fjord are the wind direction Wdir and wind speed Wvel, as well as the inflow Q of fresh water per second. When the gates between the fjord and the sea are open, the difference between H and sea level (hsea) is also an important factor. All of these influences are incorporated in the initial model presented in Table 3.\(^1\)

However, domain experts are uncertain of some model components, such as the wind forcing function, so we provided our system with four other generic processes shown in Table 4. The first one incorporates the cosine of wind direction
\(^1\) Todorowski (2003) describes a similar initial model that is incomplete. He notes the domain expert stated that wind affects height through an unknown function, which we have instantiated as the windForcingSine process in the table.
Table 4. Additional generic processes for Ringkøbing Fjord.

<table>
<thead>
<tr>
<th>Process Name</th>
<th>Variables</th>
<th>Parameters</th>
<th>Equations</th>
</tr>
</thead>
<tbody>
<tr>
<td>windforcing_cosine</td>
<td>Wvel {speed}, Wdir {direction}, h {sublevel}</td>
<td>b [-0.1,0.1]</td>
<td>[d[h,t] = b \times Wvel \times \cos(Wdir)]</td>
</tr>
<tr>
<td>windforcing_cosine_damped</td>
<td>Wvel {speed}, Wdir {direction}, h {sublevel}</td>
<td>b [-0.1,0.1], c [-0.1,0.1]</td>
<td>[d[h,t] = b \times c \times Wvel \times \cos(Wdir) - h]</td>
</tr>
<tr>
<td>windforcing_sine_damped</td>
<td>Wvel {speed}, Wdir {direction}, h {sublevel}</td>
<td>b [-0.1,0.1], c [-0.1,0.1]</td>
<td>[d[h,t] = b \times c \times Wvel \times \sin(Wdir) - h]</td>
</tr>
<tr>
<td>windforcing_simple_damping</td>
<td>h {sublevel}</td>
<td>c [0.1], k [0.1]</td>
<td>[d[h,t] = -k \times (h - c)]</td>
</tr>
</tbody>
</table>

In an effort to capture influences of the wind’s other components, three other processes reflect the idea that, as wind pushes water to one side of the fjord, gravitational potential energy builds up in the water and causes it to resist.

We have access to almost one year’s observations of the fjord, sampled every five hours, for all of the variables in the initial process model. To evaluate RPM in this domain, we treated the first 1100 observations as a training set and used the remaining 551 data points as a test set to measure generalization error. We provided the initial model, the generic processes, and the training data to RPM, which searched the revision space and returned the best model at each distance from the initial one. The system considered 32 distinct model structures, which took about 12 hours on a Linux PC with a 2.8 GHz Pentium 4 processor.

Figure 1 summarizes the results by plotting the training and test error against the distance from the initial model for the revisions suggested by RPM. We measured this distance as the symmetric difference between the processes in the initial and revised models. The curves reveal that error decreases initially as one moves away from the initial model, but after a distance of two the errors increase again. This is not surprising if the ‘true’ model falls at an intermediate distance from the initial one. More interesting is that the curves for training and test set error are similar, which suggests that our model-revision method is not overfitting the training data, at least in this domain.

Inspection suggests that the revisions RPM proposes are physically reasonable. The best model at distance one has the process windforcing_cosine added, which means that the east-west component of the wind (measured in polar notation) has an effect on the water height. The next revision adds the process windforcing_simple_damping, which incorporates resistance to the wind force due to the gravitational force.
The best revised model’s predictions for water height are shown in Figure 2, with points to the left of the vertical line denoting the training set and those to its right the test set. As the figure shows, RPM predicts the qualitative behavior of the fjord dynamics in a reasonable way. In particular, the trajectory tracks the high peaks of the water height very well. These results indicate that the system can revise a model in ways that extrapolate effectively to new observations.

For comparison purposes, we also ran IPM on the same set of processes and training data for Ringkobing Fjord. The IPM algorithm took about six times longer than RPM to finish investigating its model space, which consisted of 217 different model structures. The best-scoring model found by IPM had lower mean squared error on the test data (0.0081) than the best induced by RPM (0.0099), but it fared much worse in terms of qualitative prediction. In fact, this model did not track any of the major variations in water height because it predicted a nearly flat trajectory over time.

### 3.2 Population Dynamics in the Ross Sea

The Ross Sea in Antarctica has been the focus of many studies (Arrigo et al., 2003) because it has a relatively simple food web in comparison to open ocean systems, and thus provides a tractable starting point for modeling efforts. The most important organism in this ecosystem is phytoplankton, which undergoes repeated cycles of population increase and decrease.

Incorporating domain knowledge from our team’s Earth scientist (Arrigo), we developed the initial process model shown in Table 1, which relates resources such as light and nutrients to phytoplankton growth. Although the model encodes much existing knowledge about how the variables interact, uncertainty about several components led us to consider alternative generic processes like those in Table 2. These include mechanisms for zooplankton grazing on phytoplankton, nitrate remineralization, and residue loss. Because zooplankton was not measured, RPM treats it as an unobserved theoretical variable.

We have two sets of daily measurements of phytoplankton and nitrate concentrations in the Ross Sea, as well as light levels and ice coverage, each spanning...
Fig. 2. Observed water height in Ringkøbing Fjord over a year, along with heights predicted by the revised model.

188 days for two consequent years. We suspected that the differences in the two trajectories were caused by the higher ice coverage in the second year. We ran RPM using the first year’s data as the training set, utilizing the generic processes from Table 2 as background knowledge and letting the system revise all parameters in the initial model from Table 1. Figure 1 plots the training and test errors against distance from the initial model for the revisions suggested by RPM. The minimum error for both curves occurs with four structural revisions, although the test curve does not follow the systematic U shape of the training curve.

Detailed inspection by the Earth scientist on our team revealed that RPM’s revisions were ecologically plausible. The best model at distance one added the process nutrient remineralization, which involves restoring nitrate ions into the water from the residue of dead phytoplankton. The best model at distance two also added the process residue loss to remineralization, which is a conservation term that balances mass transfer from residue to nitrate ions. The next two revisions added grazing and Ivlev rate, which together model zooplankton grazing on phytoplankton.

Figure 3 shows the observed trajectories for both the training and test years, along with predictions from the revised model with the best training score. The predicted trajectories have the right qualitative shape for both years, but the fit is much weaker for the test set. However, the mean squared error for this model was 2.54 on the training set and 54.06 on the test set, as compared with 22.89 and 888.18 for the initial model structure from Table 1 with revised parameters. Thus, RPM’s search through the space of model structures produced a candidate that generalized to the second year substantially better than did the original.

Nevertheless, the predicted curves for the two years are very similar, which suggests that the model’s behavior is not sensitive to their differences in the amount of ice, as we had anticipated. To better understand this effect, we trained
Fig. 3. Observed phytoplankton and nitrate concentrations in the Ross Sea, along with predictions from the revised process model, for training (left) and test (right) data.

RPM on the second year’s data and tested on the first year. The structure of the best-scoring model was nearly the same, but the parameters were very different. Moreover, the model followed the the second year trajectory closely but not that for the first year. This suggests the parameter-fitting method is overfitting the training data, and thus overlooking relations that would improve generalization.

We also trained the original IPM system on the first year’s data and tested its results on the second year. The program took about 20 hours to explore 121 different model structures, considerably more time than RPM, which took five hours to consider 40 candidate structures. The best-scoring models produced by the two system had very similar structures and comparable error rates. IPM had the same tendency to overfit the Ross Sea data as did the model revision system.

The results from this study show again that, by revising an initial model instead of learning from scratch, RPM can use relatively little search to find alternative models that improve greatly on the original yet remain consistent with background knowledge. However, in this domain, the parameter estimation method appears to overfit the training data, and addressing this issue is an important direction for future research.

4 Related and Future Work

Computational methods for model revision are certainly not a new idea. Early research focused on supervised learning for classification tasks, but supported modification of models with theoretical terms and offered a general framework from which we have borrowed. For instance, Ourston and Mooney’s (1990) EITHER utilized search operators for adding and deleting rules, which correspond to our operators for adding and removing processes. They also included ones for adding and removing conditions on rules, which are quite different from our scheme for parameter revision.

A different tradition has explored methods for revising qualitative causal models of scientific phenomena. Early systems altered their models incrementally, but later work has utilized nonincremental methods with simpler representations of causal connections. Recent examples include Bay et al.’s (2003)
method for revising qualitative models of gene regulation, which carries out greedy search guided by candidate models’ fits to the data. Bryant et al. (2001) report a different approach that uses abductive logic programming to extend a qualitative model of metabolic control.

A parallel line of research has addressed the revision of quantitative models. Chown and Dietterich (2000) dealt with improvement of the parameters in a complex ecosystem model, whereas Glymour et al.’s (1987) TETRAD system revised the structure of models cast as sets of linear equations. Other work by Bay et al. (2002) and by Todorovski (2003) has focused on revising the structure and parameters of nonlinear models, often drawing on knowledge about the domain to constrain search. These two efforts dealt with differential equation models for time-series data, and thus come closest to the work reported here. The main advance is that our approach introduces the notion of processes, which Todorovski has also recently explored. These provide a useful framework for encoding domain knowledge that constrains search and produces more interpretable results.

Naturally, we also intend to extend our revision algorithm itself on various fronts. One drawback of the current system is that parameter estimation takes 99.9 percent of the computation time, which limits the number of models with different structure it can consider effectively. Preliminary studies with a hierarchical method for multiple shooting (Horbelt et al., 2001) have shown promise for speeding this component. Also, our results for the Ross Sea suggest that overfitting is occurring with the parameter estimates. We plan to investigate several methods for mitigating this problem, including averaging parameter estimates with statistical resampling techniques and incorporating uncertainty explicitly in the estimated parameters used for simulation. We should also introduce guards against overfitting in the search for model structures.

Finally, RPM’s current reliance on exhaustive search lets it handle only relatively small model spaces. Future versions should replace this approach with a heuristic method that scales to more complex models and to more extensive changes. We should also extend the framework to support revision of models with subsystems, which ecosystem modelers often utilize when dealing with complex domains. This would reduce search by drawing on knowledge about likely subsystems rather than isolated processes, and thus further improve scaling ability.

5 Concluding Remarks

In this paper, we reported an approach to computational scientific discovery that focuses on the revision of existing models rather than on their construction. Unlike earlier work in this area, we utilized the formalism of quantitative process models, which support explanatory accounts of continuous time series in terms of unobservable variables and processes. Our specific algorithm, RPM, lets users specify an initial model, a data set, and a set of allowed revisions. The latter can include specific processes that may be deleted, specific processes for which the parameters may be altered, and generic processes that may be added. We demonstrated this system’s abilities in two environmental domains, one involving changes in water height in a Danish fjord and another concerning population dynamics in the Ross Sea.
Our experimental results were generally encouraging. In both domains, we showed that RPM found meaningful revisions that had substantially lower error than the original model. Moreover, we found that these models were generally as accurate as those produced by IPM, which composes models entirely from generic processes, but were obtained with less search and in less time. However, we also identified some limitations of the system that future work should address. In summary, our approach builds on previous research in model revision and scientific discovery, but extends their ideas in ways that are useful for fields like Earth science, which focus on quantitative models of dynamical systems.

References


