Geosci. Model Dev. Discuss.,
doi:10.5194/gmd-2016-242-RC2, 2016

Geoscientific
Model Development
Discussions
Open Access

Interactive comment on "Reverse engineering model structures for soil and ecosystem respiration: the potential of gene expression programming" *by* Iulia Ilie et al.

Anonymous Referee #2

Received and published: 24 December 2016

Review of "Reverse engineering model structures for soil and ecosystem respiration: the potential of gene expression programming"

In this manuscript Ilie et al. explore the use of "gene expression programming" (GEP) to select empirical models for soil and ecosystem respiration. The authors make a case that GEP is a technique for reverse engineering model structures by elucidating underlying mechanisms, rather than depending on hypothesis-driven experiments to identify these mechanisms.

I have several concerns about the conceptual framework the authors used to present GEP. I am convinced that GEP is an interesting and worthwhile approach to automate model selection. However, I think it is over-reaching to suggest that GEP can

'reverse engineer' model development. It seems to me that the value of GEP is simply to automate the process of exploring a large number of regression models. I am not convinced that GEP reorganizes the model development process, because regression already is often the first step in model development. Further, I find that the claim that GEP minimizes human influence and perception bias to be strong, as the authors seemingly arbitrarily select the driving variables for the model, regardless of how the model's functional form is derived. From other work we know that selecting a single soil temperature at 5 cm soil depth can give a very different model from selecting a temperature from 15 cm soil depth (Graf et al., Biogeosciences doi:10.5194/bg-5-1175-2008). Similarly selecting to use VWC rather than a parameter like matric potential could be the difference between being able to predict rapid increases in flux with rainfall and not.

In the end, the functions selected by GEP suffer from the same problems as previously-used formulae shown in Table 2. All of these functions tend to underestimate large fluxes ("hot spots" and "hot moments"). While the form of the functions may hold-up from training datasets to prediction datasets, the specific parameterizations often do not. I believe the authors have done a good job discussing limitations of GEP, and empirical approaches in general, in section 5.1.1. We know biogeochemical fluxes integrate multiple pools, reservoir dynamics and lags, and these are difficult to detect using semi-empirical models. The largest gains recently in representing soil respiration have come from simulating enzyme kinetics and solute diffusion (e.g. DAMM model) as well as simulating microbial growth dynamics. These advances have come from implementing expert knowledge, not from expediting regression model selection.

Overall I would recommend that this manuscript be rejected in the current form, and the authors re-evaluate the presentation of the GEP method both in terms of creating certainty within the biogeosciences community that the approach is effective and accessible, as well as readily applicable to field data as was demonstrated with the data from Alice Holt. As was mentioned, I believe the GEP method has considerable potential, but as the manuscript is currently written my concern is that it will pass unnoticed by the community as a whole due to poor accessibility rather than scientific merit.

General Comments: I do not agree with Figure 1, that model development starts with expert knowledge. Expert knowledge does not come about on its own, but comes from observations, and regressions are critical to making sense of observations. By helping to identify which variables among a large number of potential explanatory variables correlate to a phenomenon, regression-type analyses lead to the second step in the scientific process: manipulative experiments to confirm hypothesized cause-and-effect relationships. Demonstrating cause-and-effect relationships limits the number of processes that need to be represented in models. I am not convinced that GEP provides a short-cut to this process.

Section 3.1 and 4.1, which outline artificial experiments with the GEP method could be strengthened considerably if the authors were to us a simple, mechanistic model of soil or ecosystem respiration rather than a seemingly random set of algebraic expressions. Using such a respiration model would allow the authors to attempt to recover the model basis functions and, if successful, enhance the reader's confidence with respect to the data from the site at Alice Holt.

I am concerned about the evaluations of GEP presented in Figs. 3 and 4. Fig 3 compares alternate machine learning techniques by comparing the MEF of the final model selected by each approach. It seems to me also important to compare the actual model structures, not just the fitness score. Did all the techniques recover the original models? If not, is variation in the MEF meaningful?

Figure 4c suggests that GEP was only able to recover about 30-55% of the correct number of parameters. If so, it seems GEP did NOT do a good job of recovering the original models.

Another major concern is the exercise shown in figure 7. The authors have examined whether summing predicted component fluxes gives predicted total fluxes that resemble observations. This is an interesting idea, but ultimately not that useful for two reasons:

1) The observed fluxes were not independently measured, e.g. Rauto was not measured independently, but was calculated by measuring the total flux (Rsoil) minus RH. I think you want to test whether all the variability simulated for the components can explain the variability observed for the total flux, but you don't have a measure of the component fluxes independent from the total flux. 2) We would like to see that the predictions for total flux are no worse than the predictions for the component fluxes. But in several cases the prediction for component fluxes are pretty poor. E.g. Predictions for RECO won't turn out any better than predictions for Rabove, which themselves were poor. That's not so interesting.

The manuscript is figure heavy, consider condensing figures or removing. For example can Figures 5 and 9 be combined in an effective way? Are there other figures that may be unnecessary to the reader if they were described in the text or in a table?

Specific comments: Abstract is long, introduces a lot of terminology. Consider distilling to the most important take-homes, and make more approachable for a general audience. p.3 l. 8. The rationale for reordering should also be to try more options, things that people might miss p. 3. L. 30. Why would we expect the functions to be portable across scales? Provide an ecological justification, otherwise this is not an interesting or useful exercise. p. 3. L. 22-35. When reading initially I found it difficult to understand what hypotheses the authors were testing. I think all of this information is there but needs to be re-organized to make it stand out to the reader. p.4 ll. 5. No need to introduce the conclusions. Consider shortening this to reduce repetition. 2.1 This section was not clearly written, I suggest more careful editing by co-authors. Please avoid including extra words in parantheses, they add complexity without clarity. p.4 ll.15. Is the process of mapping operations to strings relevant to model fitting? I don't think so. Either this is excessive detail about the internal workings of GEP, or you need to explain how this is relevant. p. 4. L. 20, what do you mean by "solution" The final selected model? Or the respiration predicted by that model? "Genes" and "chromosomes" should be presented in quotations initially. p. 4 l. 30 I think you can shorten this paragraph to one sentence, simply state that in each generation, the best variants of a chromosome are determined by a fitness function described below. p. 4 l. 32, what is an individual? Do chromosomes make up individuals? p. 5, l. 1 What is a hyper-parameter? Again, please try to avoid parenthetical phrases in this paragraph. p. 5, ll. 12 “upon request” rather than “on demand”. p. 5, l. 11-14 most of this information doesn’t appear useful, for example, does it actually matter that the cluster had 51 nodes? If someone ran it on a cluster with 12 nodes would it also work but be slower? Either explain the relevance of these details or remove them. p. 5, ll. 31 Consider omitting “derived from information-theoretic considerations”. p. 6, ll. 20-25. I didn’t understand the reason for this additional optimization. This sounds very much like ordinary regression model selection; does this undermine the unique value of GEP? p. 6, ll. 27 Scaling noise with signal amplitude: This is good to include! This has been shown for soil respiration too (Lavoie et al. 2015, JGR-Biogeosciences, doi: 10.1002/2014JG002773) Section 3.2.1 The first two paragraphs are repetitive in describing computation of GPP.Consider omitting or shortening the section on soil flux measurements, since these methods were reported previously. Section 3.2.4 This paragraph can be removed to shorten. Figure 3c, consider omitting. It is repetitive, and the manuscript already has a large number of figures. p.12, l. 7 Sentence starting “We find that the global modelling performance. . .” Please reword, I don’t understand this statement. Figure 12, is there a reason that this is presented in a polar plot? It seems on first glance that it could equally be presented as a 4-pane set of cartesian time series plots.

Interactive comment on Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-242, 2016.