Geosci. Model Dev. Discuss.,
doi:10.5194/gmd-2016-242-RC1, 2016

Geoscientific Model Development Discussions

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 #1

Received and published: 22 December 2016

The manuscript proposes to automatically derive model structures using Gene Expression Programming (GEP) introduced by Ferreira (2001). The authors apply GEP to different components of terrestrial CO_2 fluxes measured in an 80 year old deciduous oak plantation in the Alice Holt forest in SE England. The goal is to compare automatically derived model structures with predictions by other machine learning methods and from other published models of ecosystem respiration.

The paper is in the scope of the journal and the topic could be interesting for a broad audience of geoscientists.

In the present form, I cannot recommend publishing and ask the authors to thoroughly review their manuscript taking the below mentioned points into account. Additionally,

the manuscript would benefit from a proofreading by a native speaker.

Major comments

1. The goals stated in the introduction are scattered over page 3 (ll. 3–4, ll. 23–25, ll. 28–30). Please state them clearly at the end of the introduction.
2. GEP is the key part of the manuscript. It is not a standard modelling framework and needs a clear introduction. In the present form, Section 2.1 is difficult to understand for someone not familiar with GEP. Please define clearly what is a gene, a chromosome and an expression tree and how they are related. Use examples for illustration. The original paper by Ferreira (2001) is written for a broad readership and can serve as an example. How are the mathematical statements coded in chromosomes evaluated to generate predictions?
3. The use of the fitness measures is inconsistent throughout the manuscript. In section 2.2 you derive a composite fitness measure CEM and state that this is your final normalized form of the fitness function (eq. 2.3). However, later in the results you report MEF or MEF+NP (that was never property introduced). Explain clearly which function was used to measure the fitness. Also p. 8 l. 4–5 shows that CEM is apparently not your final fitness function.
4. What were the functions that were coded in GEP and could thus form algebraic expressions? How did you chose them?
5. Section 3.1.1: You state the the machine learning methods (Artificial Neural Networks, Support Vector Machines, Random Forests and Kernel Ridge Regression) were used without tuning the hyperparameters. I have a serious objection here. While some of the hyperparameters could be safely set to default values, others have to be tuned and do affect the performance of those models (e.g. the

cost parameter of Support Vector Machines). I recommend that you consult the technical literature here and tune hyperparameters for a fair comparison. A good point to start is the book by Kuhn and Johnson (2013).

6. Which predictors did you use for the machine learning methods on the artificial data?

7. p. 9 l. 29–32 You state that you log-transformed the fluxes before modelling and back-transformed the model structures. Did you also back-transform the predictions? At least in standard regression, back-transformations need particular attention. When back-transforming from the log transformation, the variance of the residuals has to be considered in order to avoid a bias. Please explain what and how you back-transformed. How did you take care of a possible bias?

8. Fig 8 shows a lot of dynamics in residuals from the GEP approach. Because you are dealing with time series, reporting MEF only is not satisfactory. A more in depth comparison of the different models at different time scales is appropriate (e.g. Mahecha et al., 2010). Which temporal patterns can be well reproduced by the different models?

9. From Fig 10 we learn that the machine learning algorithms performed better than GEP. In Section 5.2 you state that GEP underestimates high fluxes as do the published semi-empirical models. So what is the advantage of using a GEP approach? What can we learn from it? I suggest that you restructure your discussion such that this aspect becomes really clear. In the present form Section 5.2 is somehow lost.

Detailed comments

- p. 3 l. 14 Explain briefly symbolic regression here and in more details in the method section (p. 4 ll. 9ff).

- p. 4 l. 14–17 You state that the "variables and functions are subsequently mapped to a set of characters", then that the "mapping process generates sets of strings. . . " And then in the next sentence "the mapped letters are randomly combined . . . ". This is confusing. State clearly what is the alphabet used to map functions and variables. They cannot be randomly combined: a binary function has to have two inputs, for example, and this is taken care of in the coding sequence. The initial chromosomes are generated randomly, however, the genes must be valid mathematical expressions.

- p. 4 l. 32 explain individual.

- p. 5 l. 1 How is the hyperparameter tuned?

- p. 5 l. 8–9 How is the population diversity related to stochastic bias?

- p. 6 l. 2 and eq. 2.2 inconsistent names: SE or S[P]?

- Give more details on the calculation of the permutation entropy (Bandt and Pompe, 2002). A reader not familiar with the method should be able to understand what you calculated.

- Eq. 2.3. I don't understand the last term in your derivation of CEM. Why $1 - SE$? The permutation entropy varies between 0 and $log(n!)$, n being the order of permutation ($n = 4$ in your case). Did you normalise SE by its maximum?

- Is CEM maximized or minimized?

- p. 6 l. 22 Why are model parameters constant values? This term for an entity being optimized is confusing.

- p. 7 l. 26 and Tab1: You never explained head and tail of genes.

- p. 9 l. 25 Explain briefly how the Singular Spectrum Analysis works and give references to the original publications (Broomhead and King, 1986, for example).
- p. 10 l. 1–2 I don't understand how your split you data in training and test data sets. According to p. 8 l. 21 you have two years of hourly observations. So what are the 500 target time steps and why are there 613 time steps in total? How did you calculate the subsets?
- p. 12 l. 20–22 What do you mean by a "component of R_{eco} not seen in the training procedure"? Which components were not modelled?
- p. 14 l. 10 Which water reservoir do you refer to? Soil water? Then reservoir is misleading.
- p. 16 l. 13 You state that GEP is not prone to overfitting. How did you analyse this?
- What are the error bars in Fig3(a), (b) and fig4 (c)?
- Fig3(c) is not necessary.
- Fig12 is never discussed in the text.

References

Bandt C and Pompe B. 2002. Permutation entropy: a natural complexity measure for time series. *Physical review letters*, **88**(17): 174102.

Broomhead D and King GP. 1986. Extracting qualitative dynamics from experimental data. *Physica D: Nonlinear Phenomena*, **20**(2-3): 217–236.

Ferreira C. 2001. Gene expression programming: A new adaptive algorithm for solving problems. *Complex Systems*, **13**(2): 87–129.

Kuhn M and Johnson K. 2013. *Applied predictive modeling*. Springer.

Mahecha MD, Reichstein M, Jung M, Seneviratne SI, Zaehle S, Beer C, Braakhekke MC, Carvalhais N, Lange H, Le Maire G, and Moors E. 2010. Comparing observations and process-based simulations of biosphere-atmosphere exchanges on multiple timescales. *J. Geophys. Res.*, **115**(G2): G02003–.

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