Comments to the authors

Geoscientific Model Development – Discussions
Manuscript ID: gmd-2016-242

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

Authors: Iulia Ilie, Peter Dittrich, Nuno Carvalhais, Martin Jung, Andreas Heinemeyer, Mirco Migliavacca, James I. L. Morison, Sebastian Sippel, Jens-Arne Subke, Matthew Wilkinson, and Miguel D. Mahecha

Major comments

Overall, the quality of the manuscript has improved. The authors addressed most of my concerns. There is still one point missing, however, that I think is important. It concerns that back transformation of the values (p. 12 l. 20–23). In regression, the correct back transformation of $\log y = x\beta + \varepsilon$ is not $e^{x\beta}$ but $e^{x\beta}E(e^{\varepsilon})$, E being the expectation (e.g. Manning, 1998). For a lognormal case (i.e. the residuals in a the regression of the log-transformed variable are normally distributed), one would get $E(y|x) = e^{x\beta+0.5\sigma^2}$. Thus, if one transforms without taking the variance of the residuals into account, the transformed values could be biased. This might affect the MEF part of your fitness function. Please explain how you handled this possible bias.

Also, transformation for Random Forest and SVM are rather unusual. While it might somehow help for SVM if you use a Gaussian (i.e symmetric) kernel, a monotonic transformation for decision trees shouldn't change anything.

I understand that the focus of your work is on GEP. However, you compare with other machine learning algorithms and they have their advantages. And one of the advantages of using Random Forest or SVM, for example, is that they do not assume any normality and thus no transformation is necessary. This advantage should be discussed.

I suggest that the authors take care of the issue of back transformation and either revise their calculation or explain why the back transformation in their case is simply the exponential.

Detailed comments

- Last point in highlights: either introduce GEP in point 2 or avoid the acronym
- p. 5 l. 1,2 What are a set number of genes and a set fixed length? Do you mean fixed/determined? This is confusing as you also use set in the sense of a mathematical set.
- p. 6 l. 1–2 I don't understand this sentence. Is the word 'one' too much?
- p. 6 l. 3 Is the word 'and' too much?

- p. 7 l. 8 What is the hyper-parameter in GEP? The stopping criterion described above? Why does it have components? Does it mean, there are several hyper-parameters?
- p. 7 l. 13 Replace signature by pattern.
- p. 6 l. 17–19 This sentence is broken.
- p. 7 l. 20–27, 28–31, p. 8 l. 1–8 These paragraphs are repetitive. Try to join the description of the Bandt and Pompe approach.
- p. 8 l. 9 . . . of the fitness . . .
- p. 8 l. 14 number
- p. 8 l. 19–20 Which function is minimized? CEM? MEF+NP? Both? On p. 10 l. 17–19 You state that different functions were used for one experiment only. Which function did you use otherwise?
- p. 9 l. 25 Grammar: "each functional values"
- p. 10 l. 25 The MATLAB toolboxes.
- p. 10 l. 26 "SimpleR" or "Simple R"
- p. 10 l. 25–29 For a reader not familiar with those toolboxes, it is not clear whether you tuned the hyper-parameters. It is not enough to cite the tools. Please state clearly whether you tuned the parameters via a cross-validation.
- p. 11 l. 12 can be calculated
- Section 3.2.3 Please add that you use daily mean values.
- p. 14 l. 28 to p. 15 l. 4 I don't understand the point about summing the predictions. Did you also train the GEP models on summed fluxes? Then, this should be mentioned in the material and method section.
- p. 14 l. 31 How did you test for significance?
- p. 15 l. 12–13 Why is it important that there are nor linear correlations between residuals and predictors? You measure randomness of residuals via the permutation entropy.
- Figure 4 in Supplemental Materials is difficult to read. Please consider splitting it.

References

Manning WG. 1998. The logged dependent variable, heteroscedasticity, and the retransformation problem. *Journal of health economics*, **17**(3): 283–295.