Geosci. Model Dev. Discuss.,
doi:10.5194/gmd-2016-242-AC1, 2017

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.

Iulia Ilie et al.

ilie@bgc-jena.mpg.de

Received and published: 1 March 2017

1 Response to Reviewer 1

In the following, we denote comments by the reviewer in **bold** and our own reponses in standard fonts.

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 CO2 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.

We would like to thank the reviewer for the evaluation and detailed comments on our manuscript. We further provide responses for the posed questions and details on how we intend to revise the manuscript. Please note that our UK based co-authors had revised the paper, and will again be involved in the submission of the revised manuscript.

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.**
    - The section can and will be re-organized as suggested by the reviewer in the revised manuscript.
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?**

- Thank you four pointing this out here. We will include a figure explaining the process of mapping and evaluation of strings to mathematical functions in the revised version of the manuscript. We also work on describing more carefully what we understand here as “gene”, “chromosome” and an “expression tree”. We agree that this is absolutely key to the readers.

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.**

   - We apologize that we have not been sufficiently clear in our descriptions: CEM=MEF+NP+SE (modelling efficiency +number of parameters+ signal complexity measure) is the final fitness function used for optimizing the solutions for all GEP results presented in this paper. The MEF values are reported for quantifying the model-data misfit which is more natural to “read”.

     More explanations on MEF+NP will be added to the revised manuscript as well. This function is a fitness function similar to CEM, but where the entropy component is missing. This function was introduced in the manuscript in order to better illustrate the effect of each fitness function component for the final GEP solutions performance.

4. **What were the functions that were coded in GEP and could thus form algebraic expressions? How did you chose them?**

   - Usually in genetic programming type of approaches, the identification of input functions depends on the type of problem which we try to solve. If we tackle symbolic regressions, as is the case here, most often a set of primitive

     functions is proposed and sufficient, such as addition, multiplication, exponential and so on. More complex functions could increase model complexity too much and risk overfitting. We will add a more detailed explanation in the revised manuscript.

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 C2 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).**

   - *W*e are sorry for the confusion here: we wrote that “All the runs were performed with default settings” e.g. regarding the choice of their Kernels. But we did, of course allow the hyper parameters to vary and adjusted them in a cross-validation approach as described in Camps-Valls2012.

     The only approach run with default settings was the RF approach from the matlab statistics toolbox implementation.

     The paragraph should say (p8 l9):

     “The toolboxes and settings used for generating the predictions of the ANN and KRR methods are described by Tramontana2016 and found in the “simpleR" regression toolbox Lazaro-Gredilla2014, the predictions of the SVM were obtained by using the “LIBSVM" library Chang2011 from the “simpleR" regression toolbox where the regularization term, the insensitivity tube (tolerated error) and a kernel length scale are automatically adjusted. Lastly, the RF predictions were given by the Matlab statistics toolbox implementation running with default settings. "

Was corrected in the manuscript p8 l9-12.

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

   - Thank you for pointing this aspect out. All the machine learning methods (GEP, KRR, ANN, SVM and RF) learn based on the same input data set for all artificial problems, which contains 3 candidate variables (x_1, x_2 and x_3), which means that all methods are allowed to perform a feature selection as well. We apologize that this was no made clear in the manuscript but we have now corrected that p7 l25-26.

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?**

   - For the GEP solutions, we trained on log-transformed target data. That gave us a set of solutions. But of course, in order to obtain the initial fluxes an exponential function was applied to these solutions. From the exponential functions we obtained predictions which are further compared with the original target data and MEF values were reported. So, yes - we back-transformed the resulting structures.
   - For the remaining machine learning approaches (ANN, SVM, RF and KRR) the exponential is applied directly to the predictions obtained after learning from the log-transformed target and the resulting predicted fluxes are compared with the original target by means of MEF.

   - We don't exactly understand the issue of the bias - it would actually matter during the optimization as the cost-function deals with the log-transformed data. But after back transforming, the data are in original space and the evaluation with the MEF should be fine. This means also that the model selection should be unbiassed.

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?**

   - We agree that MEF is a bit superficial. However, Figure 12 reveals that model-data miss-match is not only an issue of a certain fast time scale, but clearly also occurs on seasonal time scales. The Mahecha2010 approach is very useful if we would be able to additionally deal with e.g. trends etc. But for this kind of analysis the time-series are simply too short.

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.**

   - Thank you for pointing this out. Indeed, this discussion is at the heart of our philosophical approach: We argue that if GEP identifies structurally very different model that, however, yield equivalent model performance, it puts at question the validity of the conventional semi-empirical models. GEP models reveal that certain dynamics that are typically unconsidered in ap-

proaches of this kind, for instance the exponential influence of SWC to respiration components or the seasonal influence of GPP. This section of the discussion will be restructured in the revised manuscript for increased clarification of where we see the added value of such an approach.

Detailed comments

- **p. 3 l. 14 Explain briefly symbolic regression here and in more details in the method section (p. 4 ll. 9ff). C3**

  A symbolic regression is a type of regression where not only the parameters of a known (linear) function are optimized based on data, but where the functional form itself is also constructed based on data as a combination of basic linear and non-linear mathematical functions. Further expanded in the method section.

- **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.**

  The input variables and functions are indeed mapped to characters that are combined into strings which encode the mathematical expressions. The validity of encoded mathematical expression is insured by the internal translation language and by the equation: tail=head*2+1. Thus although each of the sections, head and tail are generated based on random selection from the input characters sets (functions+variables sets for head and variables for tail), there are still rules that

insure validity of mathematical expressions (except for cases where a solution can only be deemed invalid by evaluating the expression, such as division by 0, etc)

- **p. 4 l. 32 explain individual.**

  The individual is a component of the evolution population which encodes a specific mathematical expression. It is the same as chromosome. Added better definition in glossary.

- **p. 5 l. 1 How is the hyper-parameter tuned?**

  The hyper-parameter has either some commonly used default values in the community, especially for the genetic operators rations, or some values that have been empirically established with experience, depending on the problem we are looking at.

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

  Once diversity is insured in the evolution population, we can be more confident that a certain solution does not appear just by chance, as it would have to be good enough to beat a larger pool of solutions.

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

  SE is the name we use for the Shannon entropy. S[P] is changed as well to SE in the manuscript.

- **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.**

  In short, the calculation of an entropy as a measure for randomness from a time series (e.g. Shannon's entropy) requires to determine a probability distribution

that underlies the time series (or dynamical system), which is usually done by a partitioning step (also called phase space reconstruction in other contexts). This is a fundamental step in the methodology, and various methods have been used to arrive at this probability distribution, for instance frequency or histogram-based measures, procedures based on amplitude statistics, or symbolic dynamics (see e.g Kowalski et al 2011 for an overview). In recent years, the Bandt Pompe approach has become popular, because it directly takes sequences in time into account: The technique hence divides the time series into ordinal sequences (i.e. ordinal patterns, or symbolic sequences), and then computes entropy measures directly from the probability distribution of these ordinal patterns Bandt2002. This approach has a number of advantages, namely that it is robust to noise (no sensitivity to numeric outliers) and to trends or drift in the data, it is an (almost) non-parametric method and no prior assumptions about the data are needed (the only parameter that has to be specified is the embedding dimension, i.e. window length), and allows to disentangle various possible states of the system that are then encoded in the probability distribution (see e.g. Zanin2012 for a review of the method and applications). We will describe this method in more detail, and give a few examples of its application in the revised manuscript.

- **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?**

  SE is indeed normalized by its maximum; hence SE varies between 0 and 1, where 1 indicates no correlated structure in the residuals. Furthermore, the best CEM value can take, and towards which the optimized values tend to is 0.

- **Is CEM maximized or minimized?**

Throughout the entire paper, the optimization is done by minimization of the fitness function value.

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

  GEP as a method does not offer a specific optimization of parameters, as it evolves entire mathematical formulations. So until there is a special treatment in terms of optimisation for the parameters, they are considered constants. Once a final solution is reached, a specific optimization algorithm is used for

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

  We apologise for the slip. Added to glossary.

- **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).**

  The SSA method is a very useful tool used mainly in time series analysis with the purpose of decomposing an original time series into the sum of its components, such as trends, seasonality and high frequency components. More details and the references are added to the revised manuscript.

- **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?**

  Thank you for pointing this aspect out. It seems that we have not been clear enough in the description. Data is available with hourly resolution, however, we use daily means for model constructions. So for two years, we should have 732 data points, but after filtering we are left with a gapped set of 613 observations.

Those 613 d.p. are split into two sets of 500 and 113 d.p 50 times. For each of this split we then learn a model and the best over-all at validation is finally selected and presented in the results section. Section is revised for clarity.

- **p. 12 l. 20–22 What do you mean by a "component of Reco not seen in the training procedure"? Which components were not modelled?**

  Each component was separately modelled and a solution is built with GEP. Then, the parameters of each of these solutions are re-calibrated using CMA-ES for the rest of the components for a fair comparison of modelling capacity.

- **p. 14 l. 10 Which water reservoir do you refer to? Soil water? Then reservoir is misleading.**

  Indeed we refer to soil water. We apologize for the confusion and water reservoir has been changed to soil water in throughout the revised manuscript.

- **p. 16 l. 13 You state that GEP is not prone to overfitting. How did you analyse this?**

  This was concluded for the results of the increase of signal to noise ratio exercise, as the MEF values of the solutions reconstructed when compared to original, noise free data do not change significantly with addition of noise.

- What are the error bars in Fig3(a), (b) and fig4 (c)? The error bars are not visible enough at the scale of the plot. A table with the concrete values given in the plots will be added to the revised manuscript.

- **Fig3(c) is not necessary.**

  Removed from manuscript as suggested.

- **Fig12 is never discussed in the text.**

The figure is mentioned in p. 15 l 5. However we agree that it needs more clarification in the manuscript.

References

C. Bandt and B. Pompe. Permutation entropy: a natural complexity measure for time series. Physical review letters, 88(17):174102, apr 2002. ISSN 0031-9007. doi: 10.1103/PhysRevLett.88.174102. URL http://www.ncbi.nlm.nih.gov/pubmed/12005759.

G. Camps-Valls, J. MunËIJ oz-Mar′Äś, L. Go′ mez-Chova, L. Guanter, and X. Calbet. Nonlinear statistical retrieval of atmospheric profiles from MetOp-IASI and MTG-IRS infrared sounding data. IEEE Transactions on Geoscience and Remote Sensing, 50(5 PART 2):1759–1769, 2012. ISSN 01962892. doi: 10.1109/TGRS.2011.2168963.

C.-C. Chang and C.-J. Lin. Libsvm. ACM Transactions on Intelligent Systems and Technology, 2(3):1–27, 2011. ISSN 21576904. doi: 10.1145/1961189.1961199. URL http://dl.acm.org/citation.cfm?doid=1961189.1961199.

M. Lazaro-Gredilla, M. K. Titsias, J. Verrelst, and G. Camps-Valls. Re- trieval of Biophysical Parameters With Heteroscedastic Gaussian Pro- cesses. IEEE Geoscience and Remote Sensing Letters, 11(4):838–842, apr 2014. ISSN 1545-598X. doi: 10.1109/LGRS.2013.2279695. URL http://ieeexplore.ieee.org/document/6595574/.

M. D. Mahecha, M. Reichstein, N. Carvalhais, G. Lasslop, H. Lange, S. I. Seneviratne, R. Vargas, C. Ammann, M. A. Arain, A. Cescatti, I. a. Janssens, M. Migliavacca, L. Montagnani, and A. D. Richardson. Global convergence in the temperature sensitivity of respiration at ecosystem level. Science (New York, N.Y.), 329(5993):838–40, aug 2010. ISSN 1095-9203. doi: 10.1126/science.1189587. URL http://www.ncbi.nlm.nih.gov/pubmed/20603495.

G. Tramontana, M. Jung, C. R. Schwalm, K. Ichii, G. Camps-Valls, B. Ra′duly, M. Reichstein, M. A. Arain, A. Cescatti, G. Kiely, L. Merbold, P. Serrano-Ortiz, S. Sickert, S. Wolf, and D. Papale. Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms. Biogeosciences, 13 (14):4291–4313, jul 2016. ISSN 1726-4189. doi: 10.5194/bg-13-4291-2016. URL http://www.biogeosciences.net/13/4291/2016/.

M. Zanin, L. Zunino, O. A. Rosso, and D. Papo. Permutation Entropy and Its Main Biomedical and Econophysics Applications: A Review. Entropy, 14 (12):1553–1577, aug 2012. ISSN 1099-4300. doi: 10.3390/e14081553. URL http://www.mdpi.com/1099-4300/14/8/1553/.