Geosci. Model Dev., 10, 3519–3545, 2017
https://doi.org/10.5194/gmd-10-3519-2017

Geoscientific Model Development
Open Access
EGU

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

Iulia Ilie[1], Peter Dittrich[2,3], Nuno Carvalhais[1,4], Martin Jung[1], Andreas Heinemeyer[5], Mirco Migliavacca[1], James I. L. Morison[8], Sebastian Sippel[1], Jens-Arne Subke[6], Matthew Wilkinson[8], and Miguel D. Mahecha[1,3,7]

[1]Max Planck Institute for Biogeochemistry, Department Biogeochemical Integration, Hans-Knoell-Str. 10, 07745 Jena, Germany
[2]Bio Systems Analysis Group, Institute of Computer Science, Jena Centre for Bioinformatics and Friedrich Schiller University, 07745 Jena, Germany
[3]Michael Stifel Center Jena for Data-Driven and Simulation Science, 07745 Jena, Germany
[4]CENSE, Departamento de Ciéncias e Engenharia do Ambiente, Faculdade de Ciéncias e Tecnologia, Universidade NOVA de Lisboa, Caparica, Portugal
[5]Department of Environment, Stockholm Environment Institute, University of York, York, YO105NG, UK
[6]Biological and Environmental Sciences, School of Natural Sciences, University of Stirling, Stirling, UK
[7]German Centre for Integrative Biodiversity Research (iDiv), Deutscher Platz 5e, 04103 Leipzig, Germany
[8]Forest Research, Alice Holt Lodge, Farnham, Surrey, GU10 4LH, UK

Correspondence to: Iulia Ilie (ilie@bgc-jena.mpg.de) and Miguel D. Mahecha (mmahecha@bgc-jena.mpg.de)

Received: 29 September 2016 – Discussion started: 7 November 2016
Revised: 31 July 2017 – Accepted: 21 August 2017 – Published: 25 September 2017

Abstract. Accurate model representation of land–atmosphere carbon fluxes is essential for climate projections. However, the exact responses of carbon cycle processes to climatic drivers often remain uncertain. Presently, knowledge derived from experiments, complemented by a steadily evolving body of mechanistic theory, provides the main basis for developing such models. The strongly increasing availability of measurements may facilitate new ways of identifying suitable model structures using machine learning. Here, we explore the potential of gene expression programming (GEP) to derive relevant model formulations based solely on the signals present in data by automatically applying various mathematical transformations to potential predictors and repeatedly evolving the resulting model structures. In contrast to most other machine learning regression techniques, the GEP approach generates "readable" models that allow for prediction and possibly for interpretation. Our study is based on two cases: artificially generated data and real observations. Simulations based on artificial data show that GEP is successful in identifying prescribed functions, with the prediction capacity of the models comparable to four state-of-the-art machine learning methods (random forests, support vector machines, artificial neural networks, and kernel ridge regressions). Based on real observations we explore the responses of the different components of terrestrial respiration at an oak forest in south-eastern England. We find that the GEP-retrieved models are often better in prediction than some established respiration models. Based on their structures, we find previously unconsidered exponential dependencies of respiration on seasonal ecosystem carbon assimilation and water dynamics. We noticed that the GEP models are only partly portable across respiration components, the identification of a "general" terrestrial respiration model possibly prevented by equifinality issues. Overall, GEP is a promising tool for uncovering new model structures for terrestrial ecology in the data-rich era, complementing more traditional modelling approaches.

1 Introduction

One prerequisite to understand and anticipate the global consequences of anthropogenic climate change is an accurate

quantitative description of the terrestrial carbon cycle (Bonan, 2008; Heimann and Reichstein, 2008; Luo et al., 2015). However, the description of the mechanisms underlying the total terrestrial efflux of CO_2 (S. Peng et al., 2014), often referred to as "terrestrial ecosystem respiration" (R_{eco}), varies across the scientific literature and existing global models. This is partly because R_{eco} does not originate from a single process but is the sum of fluxes from different autotrophic and heterotrophic respiration processes that operate across different temporal and spatial scales and compartments (e.g. soil depths). Hence, it is experimentally very difficult to disentangle the main abiotic and biotic factors driving respiratory processes at the ecosystem level (Trumbore, 2006) and to derive suitable models for the individual respiration processes. In the rest of the paper we use the term "model" as an equivalent of "response functions", i.e. some analytic description of how environmental drivers influence ecosystem fluxes.

Traditionally, respiration models have been based on some theoretical considerations, but largely remain empirical in nature (e.g. Reichstein and Beer, 2008; Gilmanov et al., 2010; Hoffmann et al., 2015). Conventional model building (Fig. 1) is primarily hypothesis driven and capitalizes both on some understanding of the system and reported scaled experiments (Migliavacca et al., 2012; Richardson et al., 2008). Gupta et al. (2012) describe this common paradigm of model development as a four-step approach involving observational, conceptual, mathematical and computational phases (see also e.g. Bennett et al., 2010; Williams et al., 2009). During the observational phase, the system under scrutiny is monitored and observations are assembled, ideally representing process responses to hypothesized driving variables. Based on these observations, a conceptual model is proposed, which subsequently guides the formulations of mathematical representations of the system states and dependencies. The mathematical description then provides the basis for computational models that are used for simulations (Jakeman et al., 2006). Model–data integration may additionally lead to iterative structural revisions or parameter optimizations (Williams et al., 2009). This conventional approach to model development is also characteristic of different kinds of ecological model building, including the development of biogeochemical models (Williams et al., 2009).

We explore the possibility of reverse engineering offering an automated alternative to model development for predicting terrestrial carbon fluxes (Fig. 1). In reverse engineering, the work flow is fundamentally different (Bongard and Lipson, 2007), comprising a database set-up phase, a computational phase, a mathematical phase and a conceptual phase (Gupta et al., 2012). The rationale behind reordering the key phases is firstly to minimize the human influence and perception biases that might shape the formulation of new hypotheses, and secondly to increase the chance of novel model structures automatically emerging from the available data and that would not be so obvious from a direct analysis. Reverse engineering aims at identifying some mathematical representation of a system that is to a large degree independent of a priori conceptualizations: in the current case, the respiratory response of terrestrial ecosystems to environmental drivers. Reverse engineering leaves the model construction up to an algorithm and is therefore a way to empirically learn from observations with minimal user input.

Of course, expert knowledge still has a large influence on the modelling process, as only a certain set of variables can be measured and an even smaller subset is indeed available for model development, which includes the restriction to a certain plausible number of time lags, and hence full objectivity of automatic model development cannot be truly achieved. Furthermore, expert knowledge comes into play when the algorithm is set for running, by tuning the set of parameters according to the problem needed to be solved and as well during the observation collection and during the final decision on whether the solution returned by the algorithm actually makes sense at all and whether it can be used further. Nevertheless, we believe that by shifting the moment when the analyst makes the decision regarding the selected model, a larger degree of objectivity in modelling is achieved.

Reverse engineering is close to machine learning based regression techniques, where various candidate model formulations and specifications are explored in order to minimize the prediction error. The fundamental difference from typical model building is that reverse engineering typically provides a symbolic regression, that is, the resulting structures are ideally directly readable as mathematical functions (i.e. response functions) and can be interpreted. The readable character of the returned solutions allows us to consider the applicability of the derived structures in other system domains (Ashworth et al., 2012).

Here, we focus on the gene expression programming (GEP, Ferreira, 2001) reverse engineering approach. GEP is an evolutionary algorithm that constructs mathematical response functions. In its essence, GEP basically converges to a solution after rejecting a large number of potential regression models over a certain number of evolutionary steps. Due to its structural design, GEP can be applied in a wide range of empirical modelling problems (Y. Peng et al., 2014; Khatibi et al., 2013; Traore and Guven, 2013), including (soil) hydrology (Fernando et al., 2009; Hashmi and Shamseldin, 2014). To the best of our knowledge the potential of GEP has not yet been explored for modelling biogeochemical fluxes in terrestrial ecosystems.

We seek to understand as well whether automating model development can provide new insights into understanding the dynamics of terrestrial respiration processes. We base our study on data from a long-term monitoring experiment of R_{eco} components, i.e. above-ground respiration, root respiration, mycorrhiza respiration, and soil autotrophic and soil heterotrophic respiration. The monitoring was done separately but in a time-synchronized way over 2 years and is described in detail by Heinemeyer et al. (2012).

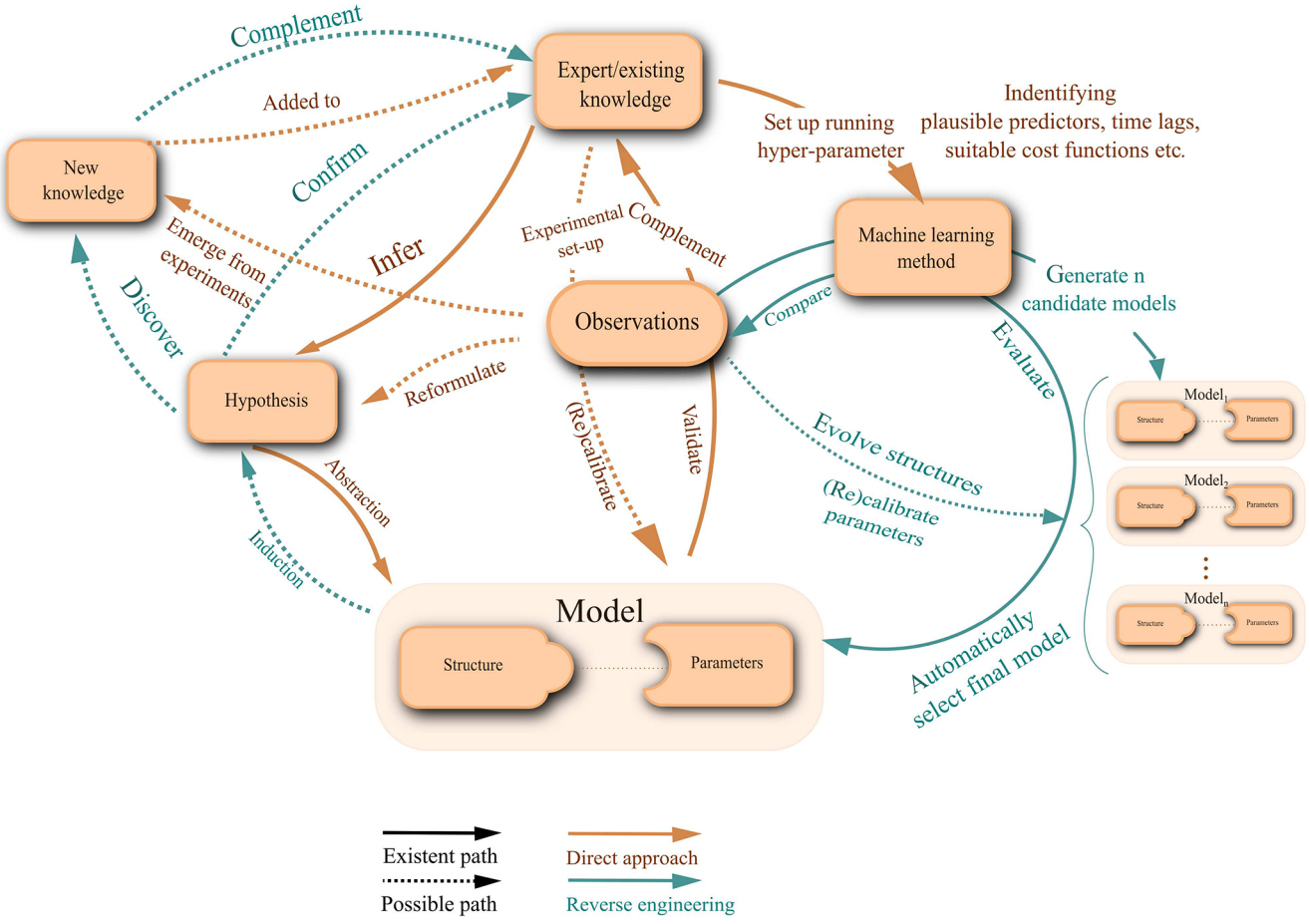

Figure 1. Direct approach and reverse engineering in model development for describing dynamical systems. Existing and possible steps needed in the process of building a model. For the direct approach, the process starts with the building of a hypothesis from existing knowledge. The hypothesis is then the subject of abstraction and is summarized in a mathematical model that has two components: the structure and the parameters. The mathematical model can be translated into a computational form that will generate predictions. Depending on how well the predicted values manage to recreate the available observations, the model's parameters are calibrated or, if the general trends are missed, there might be a need for structural reformulation. On the other hand, in the reverse engineering approach, a machine learning method is used to generate a set of candidate models that are then compared with the available observations and which according to the prediction capacity may have to go through structural changes by automatic evolution or through a final parameter adaptation. From the set of evolved models, the best model in terms of prediction capacity is chosen and its structure will be the basis for hypothesis building, as an expert would try to explain why a specific structure was automatically evolved and whether the structure of the model can be explained from the studied system-intrinsic processes. If that is the case, and the structure has not emerged randomly, the conclusions can be compared with the existing knowledge which can be reconfirmed, or new aspects of the studied system might be brought to light.

The fundamental question addressed in this paper is whether regression models can be constructed more objectively by leaving the task of proposing a final regression model to an algorithm rather than directly to an analyst. The need for human intuition during the actual process of constructing a regression model becomes reduced, and the input of expert knowledge shifts towards identifying input variables, parameters, a suitable cost function and model plausibility.

With the current study we investigate as well whether automatically derived model structures differ substantially from models conventionally used in the study of R_{eco} and its components or whether they are consistent with established theory. The separation of R_{eco} into its components also allowed us to test the portability of individual model structures across different respiration components. In this sense, we investigate whether a generic "respiration" response can be derived or whether specific formulations for a range of respiration components are required.

Study structure

First, we introduce the GEP methodology and explore its performance for symbolic regression types of problems us-

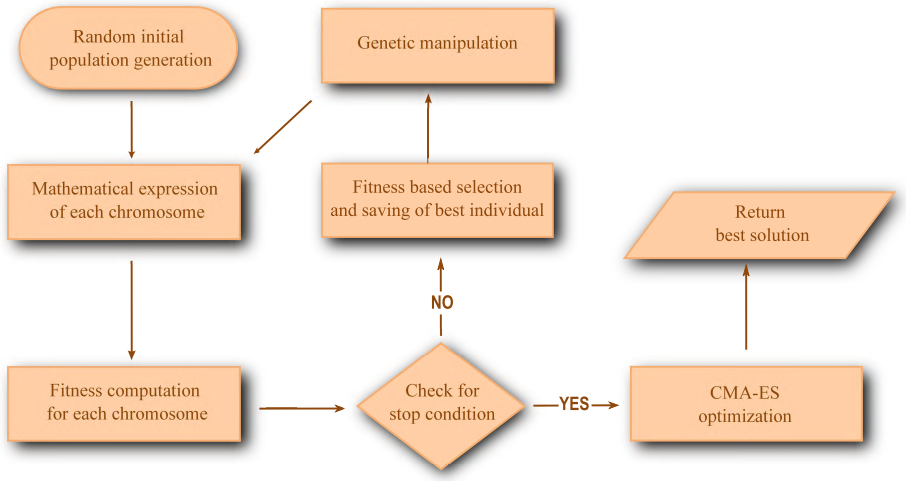

Figure 2. The work flow used in solving symbolic regression problems with GEP. The process of evolving an optimal solution from observations starts with randomly generating a set number of evolution individuals called chromosomes. The chromosomes are composed of genes that are sets of strings encoding expression trees that can be translated into mathematical expressions in the subsequent step. Following the mathematical expression comes the evaluation of each emerging individual (model) against the target variable values and for each one a fitness value is assigned. If the stopping criterion has not been reached (e.g. best fitness possible, highest number of generations allowed, convergence) the best individual in terms of fitness is saved and the remaining set of chromosomes are selected for genetic manipulation. When the stop criterion is reached, the parameters of the best chromosome is calibrated against the training data with an optimization approach, the CMA-ES, and the best solution is returned.

ing an artificial experiment under varying degrees of noise contamination designed to resemble R_{eco}. Second, we apply GEP to model the various respiration observations provided by Heinemeyer et al. (2012).

The observational record provided by Heinemeyer et al. (2012) is exceptional, because measurements of soil or ecosystem respiration that are typically only integrated are here continuously and regularly measured, and the components measured offer a perfect test case for the GEP methodology.

For both the artificial experiment and real-world observations, we systematically confront the prediction error of GEP with other state-of-the-art machine learning regression approaches. In addition, we adjust the modelling approach such that the objective function (or fitness function) not only accounts for absolute or relative error, but also reduces structure in the residuals. The discussion focuses on the comparison of the various GEP-derived models, their equifinality, and performance compared to widely used literature models.

2 Method

We rely on the GEP method (Ferreira, 2001) which automatically constructs model structures based on a set of given observations. As the models we want to obtain are mathematical structures, their construction can be achieved by solving a symbolic regression (Kotanchek et al., 2013) type of problem. That is, we are not only interested in determining an optimal set of parameters for a known regression, but here, we want to discover the symbolic form of the regression itself by identifying the most important predictors and their functional transformations. The general GEP approach in solving symbolic regressions is presented in the following section and is illustrated in Fig. 2.

2.1 Gene expression programming, GEP

The process of finding the most suitable model structure based on the signal present in data in GEP starts with an initial generation of n possible model structures (Fig. 3a). These can be called evolution individuals and in GEP they are known as "chromosomes". The chromosomes are composed of a fixed number of "genes" that are connected by a binary mathematical operator. Each gene is encoded in a string with a fixed length that contains specific characters that map to either a set of possible predictors, e.g. $A = \{a, b\} \rightarrow A_m = \{x_1, x_2\}$, or a set of their possible functional transformations, e.g. $F = \{+, -, L, E\} \rightarrow F_m = \{$addition, substraction, logarithm, exponential$\}$ (see Fig. 3a).

The choice of input functions used for applying mathematical transformations to the predictors depends on the type of problem we try to solve with GEP. When the problem is a symbolic regression type of problem, as here, most often a set of primitive functions is proposed, such as addition, multiplication, or exponential. More complex functions could increase model complexity too much and risk overfitting. However, if there are already known functional transformations of certain predictors that could be part of the final desired solu-

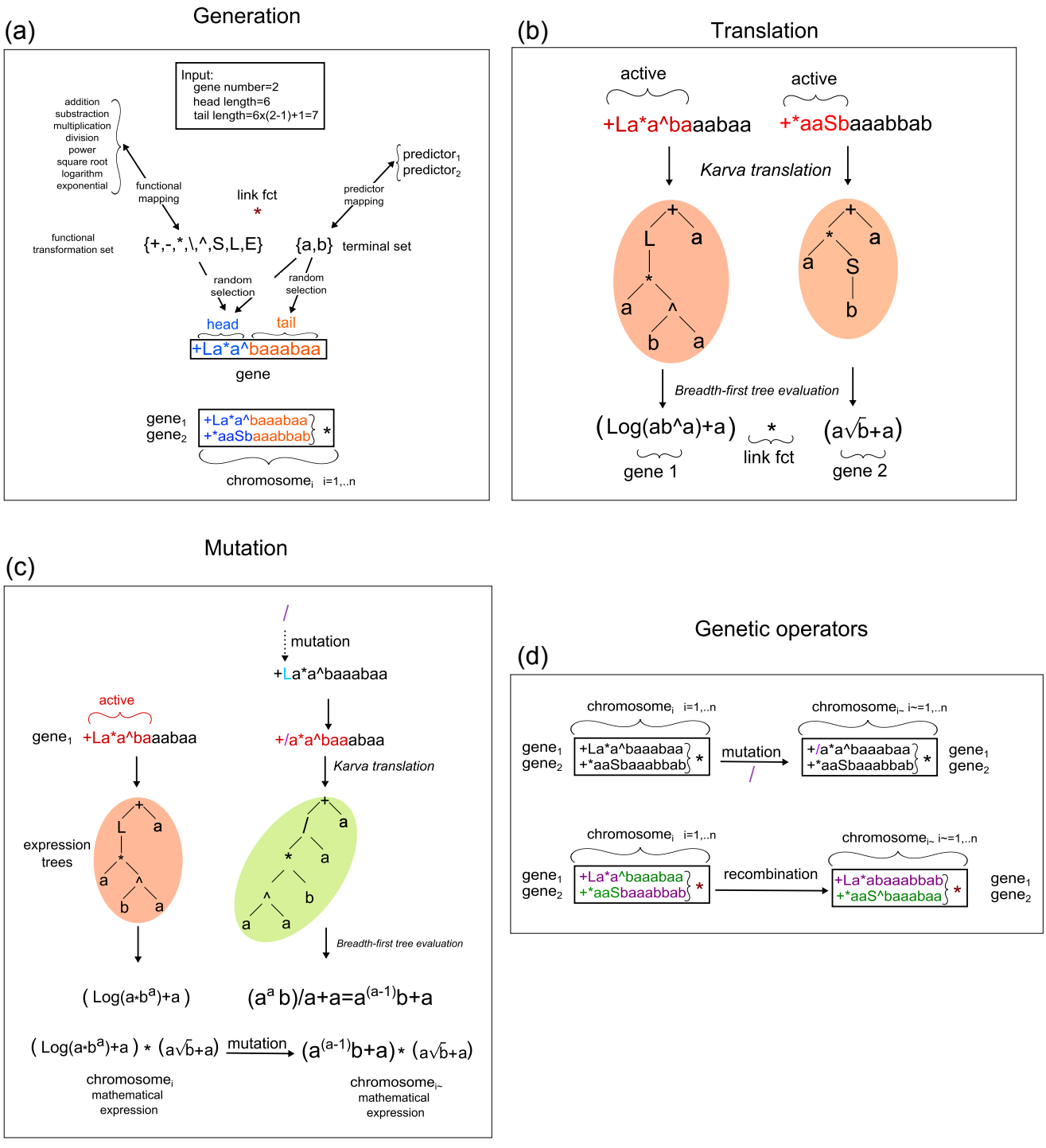

Figure 3. GEP evolution process components. **(a)** Initial random generation of genes for creating chromosomes, the individuals evolved by GEP. **(b)** GEP internal translation process from strings to expression trees and mathematical expressions. **(c)** Changes made in the mathematical expression when applying the mutation operator to the genes of a GEP individual. **(d)** Types of genetic operators for changing the GEP evolution individuals.

tion, the user can define a new function and introduce it in the set of input functions.

All genes are made up of a "gene head", containing a combination of characters mapping to both predictors and functional transformations, and a "gene tail", with characters that map only to predictors. The gene length is given by $g_l = h_l + t_l$, where $t_l = (f_{max} - 1) \times h_l + 1$, with g_l as gene length, h_l head length, t_l tail length and f_{max} the maximum parity of a functional transformation.

As in biology evolution, regardless of the actual length, the GEP genes have active sections of variable length called "open reading frames" (ORFs) that can encode various expression trees which can be evaluated into mathematical expressions (Ferreira, 2006). The lengths of the ORFs are determined only after the encoded expression trees are translated using an internal reading language (see Fig. 3b). Ferreira (2001) argues that the power of GEP lies in its use of fixed length linear strings for representing trees (ET) of varied shapes and sizes that simplify the evolutionary process and help reach a final solution faster.

The total number of chromosomes generated over each evolution step make up the GEP population. The evolution steps are also known as "generations". The maximum number of generations allowed to run until reaching a solution is often used as a stopping criterion.

One of the crucial components of model development within an evolutionary algorithm is the selection process. In GEP, the chromosomes can be translated into mathematical expressions that can be evaluated, and a distance between the current structure based predictions and the original target is computed. The measures are known as "fitness values" and are assigned to all the chromosomes in the population at each generation by means of a predefined fitness function. The evolution of the final solution with GEP is done based on optimizing the fitness function values after each generation, usually by minimizing prediction error, but more complex criteria can be taken into account as well.

Once all the fitness values have been computed and assigned, the chromosomes in a generation are sorted from best to worst fit.

If no stop criteria have been met, preparations for the reproduction of new chromosomes for the next generation are made. The chromosome with the best fitness value is reproduced unchanged in the first position of the new generation. To fill the remaining $n - 1$ positions, chromosomes are selected from the entire population for the new generation with a tournament procedure $n - 1$ times.

In tournament selection, two chromosomes are randomly selected from the entire population and the individual with the better fitness value goes through.

To ensure that novel material is introduced in the pool of possible model structures, $n - 1$ newly selected chromosomes are subject to genetic operators, such as mutation, recombination, transposition and inversion as presented in Fig. 3d, which can fully change the encoded mathematical expressions (see Fig. 3c).

Once the population of chromosomes is ready for the new generation, the evolution procedure is repeated until a stop criterion is reached, such as best fitness achieved, maximum number of unimproved generations reached, or time limit.

The hyper-parameter needed for a GEP run, i.e. the set of all parameters that need to be fixed before a GEP run is performed, has either components with recommended default values, especially for the genetic operator rates considered when applying the available genetic operators (Ferreira, 2006), or has components for which the values have been established empirically after experience in working with the GEP approach. The latter typically depend on the requirements of the problem to solve.

Such is the case for setting the length of the gene head or the number of genes in a chromosome that can be lower if the interest is in obtaining more compact solutions, with larger values possibly leading to a fast expansion of solution length which can easily overfit the initial target. When the lengths of the chromosomes are kept too low, the structures in the population can converge too soon to a unique solution that might lack the ability to capture meaningful signals present in the training data, due to low diversity of the encoded expression trees.

Another important component of the hyper-parameter to fix is the mutation rate, which is one of the genetic variation operators. When the mutation rate is too large, it can become disruptive and lead to loss of information acquired along the previous evolutionary time steps, reducing the general convergence of the GEP run. Conversely, if the rate is too low, relevant structures may not be constructed in the given time limit.

The current implementation of the GEP approach does not contain an explicit population diversity management component which could increase the confidence that a certain solution did not just appear by chance but was actually selected over a larger pool of possible model structure types. In order to reduce stochastic bias and avoid getting stuck in local optima that would produce overfitted results, we chose the practical approach of multi-start (multiple runs with the same settings) as proposed by Ferreira (2006).

The version of the GEP method presented in this paper was implemented by the first author in the C++ language and is freely available upon request. All the experiments reported in this work were executed on a cluster running SuSE SLES 11 SP1 and StorNEXT (global file system running on the IO nodes) and that contains 868 CPU cores, 14.5 TB RAM, and 1.2 PB file space. The large performance capacity of the cluster allowed for multiple parallel runs and speed in reaching the final solutions.

2.2 Fitness measure

In our study, the fitness measure is reported in terms of the Nash–Sutcliffe modelling efficiency (MEF) coefficient (Nash and Sutcliffe, 1970; Bennett et al., 2010) which is often used in the context of quantifying the performance of terrestrial biosphere models (Mitchell et al., 2009; Migliavacca et al., 2015). The MEF is computed as

$$\mathrm{MEF} = 1 - \frac{\sum_{i=1}^{n} (o_i - p_i)^2}{\sum_{i=1}^{n} (o_i - \overline{o})^2} \tag{1}$$

where o_i is the observed value at step i, p_i is the predicted value at step i and $\overline{o}$ is the mean of observed values. MEF values range between $-\infty$ and 1, where an MEF value of 1 corresponds to the case where the predicted and observed values are identical. A negative MEF value means that the predictions are worse than the mean of the observations in recreating the observed signal. $\mathrm{MEF} = 0$ indicates that the model predictions are as good as a prediction by $\overline{o}$.

During the GEP learning process, however, we use the $(1 - \mathrm{MEF})$ measure as we want to minimize the fitness function values.

Although the MEF metric offers a straightforward interpretation, it does not take the number of parameters of the models into account. In real-world applications, it might be desirable to derive models with fewer parameters if those are not (much) worse in terms of prediction capacity than models with a higher number of free terms. Thus, we include in our cost (fitness) function a normalized term related to the number of parameters (ratio of the current number of parameters to the maximum number of possible parameters given the GEP run settings).

Moreover, any systematic pattern in the model residuals needs to be reduced as the latter should ideally only represent uncorrelated noise. To meet this criterion, we complement the fitness function with a term related to the information content (entropy) in the residual time series. Entropy values would be maximized for data without structure (i.e. white noise), and lower entropy values would be obtained for structured data, e.g. correlated stochastic or deterministic processes (Rosso et al., 2007). The information content in a time series is typically quantified by the Shannon entropy (SE, Shannon, 1948), i.e. a term of the form

$$\mathrm{SE}(X) = -\sum_{i=1}^{N} p_i \ln\left[p_i\right]. \tag{2}$$

Here, $X = \{p_i;\, i = 1, \ldots, N\}$ denotes a probability distribution with $\sum_{i=1}^{N} p_i = 1$ and N possible states.

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

As our aim is to minimize structure in the residuals, the temporal order becomes important. In recent years, the Bandt–Pompe approach has become popular, because it directly takes time sequences 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 (Bandt and Pompe, 2002).

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 it allows us to disentangle various possible states of the system that are then encoded in the probability distribution (see e.g. Zanin et al., 2012, for a review of the method and applications).

The single parameter that needs specification is the window length. This parameter is fixed to $n_{\mathrm{demb}} = 4$ throughout the entire paper following previous work on ecosystem gross primary productivity dynamics by Sippel et al. (2016).

The final normalized form of the fitness function further used in our work is

$$\mathrm{CEM} = \sqrt{(1 - \mathrm{MEF})^2 + \left(\frac{P}{P_{\max}}\right)^2 + (1 - \mathrm{SE})^2}, \tag{3}$$

$$P_{\max} = n_{\mathrm{g}} \times l, \tag{4}$$

where CEM stands from here on for "complexity corrected efficiency in modelling", P is the number of parameters present in a model structure, $P_{\max}$ is the maximum number of parameters possible for each individual from a GEP run set-up, n_{g} is the number of genes in a chromosome and l is the length of a gene.

To assess the effect of adding the entropy component for the residuals in the CEM fitness function, we introduce as well a fitness measure containing elements regarding only the MEF and the number of parameters.

$$\mathrm{MEF} + \mathrm{NP} = \sqrt{(1 - \mathrm{MEF})^2 + \left(\frac{P}{P_{\max}}\right)^2} \tag{5}$$

For all experiments reported in this paper, the optimization is done by minimizing the CEM fitness function values. The best value that can be reached for all presented fitness functions is 0.

2.3 Parameter optimization

The GEP algorithm does not have a specific treatment of constants in the building of model formulations, but mutations can change both the model structure and constants. However, the scaling of constant values (model parameters) might be a decisive factor in adequately determining the fitness of a formulation. Without this, a model structure might be discarded regardless of potentially being a very powerful candidate. Furthermore, model parameters are often very informative regarding a system's sensitivity to some modifications of the drivers. These aspects have led to the addition of a final parameter optimization step at the end of each GEP run.

In order to obtain an optimal set of parameters for the GEP-extracted model structures, an approach that would be applicable in a large set of generated search spaces was necessary. Here we use the covariance matrix adaptation evolution strategy (CMA-ES, Hansen et al., 2003) for optimization. The CMA-ES is a stochastic optimization algorithm that seeks to minimize a fitness function by estimating and adapting a covariance matrix according to a sampling from a multivariate normal distribution (Beyer and Schwefel, 2002; Auger and Hansen, 2005). According to Hansen (2006), one of the main arguments in favour of the CMA-ES approach is that it has shown good results even in the case of ill-posed problems (Kabanikhin, 2008), which may very well be the case for some of the GEP structures that are automatically generated.

The CMA-ES version used for the final step of optimization is the Hansen Python implementation found at https://pypi.python.org/pypi/cma.

3 Experimental design

To explore the possibility of using GEP in developing relevant model structures for describing the terrestrial carbon fluxes, two case studies were designed: firstly, an experiment based on artificially generated data to better understand and present the general properties and capacities of GEP. Secondly, we explored the use of GEP on real measurements of various respiratory flux components monitored continuously over 2 years in an oak forest (Heinemeyer et al., 2011).

3.1 Artificial experiments

These experiments were designed to explore whether our implementation of the GEP method is suitable for symbolic regression types of problems, and how robust/vulnerable it is across various signal-to-noise ratios. We explored a set of functions with increasing levels of non-linearity to generate data points.

$$f(x_1) = 2x_1 + 1 \tag{6}$$

$$f(x_1) = x_1^2 + 3x_1 + 5 \tag{7}$$

$$f(x_1) = e^{x_1} + 1 \tag{8}$$

$$f(x_1) = e^{-x_1} - x_1 \tag{9}$$

$$f(x_1) = x_1^2 - 4\sin(x_1) \tag{10}$$

$$f(x_1) = x_1^3 + 6x_1^2 + 11x_1 - 6 \tag{11}$$

$$f(x_1, x_2) = x_2 x_1 \tag{12}$$

$$f(x_1, x_2) = x_2 x_1 - 3\cos(x_1) \tag{13}$$

$$f(x_1, x_2) = 2x_1^2 + 3x_2^2 \tag{14}$$

$$f(x_1, x_2, x_3) = 2x_1^2 + 3x_2^2 + 2\sin(x_3) \tag{15}$$

Two-thousand data points were randomly generated with $x_1 \in [1, 20]$, $x_2 \in [1, 5]$, and $x_3 \in [1, 100]$, and all the functional transformations were done based on the same initial set of 2000 data points. Out of the 2000 data points, 1000 data points were used for training, while 1000 data points were reserved for validation. The GEP settings used for each of the 20 runs are given in Table 1. If a returned structure was identical to the originally prescribed function or if $(1 - \mathrm{MEF}) \leq 10^{-5}$ at validation, the retrieval of the original structure was considered to be a success. To allow the approaches to do an automatic feature selection, all three variables, x_1, x_2, and x_3, were used for learning and validation for all 10 functions in the benchmark set.

To investigate the capacity of GEP to reconstruct a simple model used in the ecology field as well, we introduced as well an artificial test for the "Q_{10}" model that is used in the field for simulating the response of ecosystem respiration to change in air temperature of 10 °C at a reference temperature of 15 °C. The formulation we used for the "Q_{10}" model is

$$R_{\mathrm{eco}} = 2^{(0.1T_{\mathrm{air}} - 1.5)} \tag{16}$$

with R_{eco} the ecosystem respiration flux and T_{air} the air temperature. Again, we generated 2000 data points for both predictor and target and we used half for training 100 runs and half for validation. The modelling capacity of the best structure in terms of fitness value at validation is reported.

In order to investigate the response of the GEP approach to noise-contaminated data, we simulated Gaussian noise that scales with signal amplitude as often observed in the case of terrestrial ecosystem (Lasslop et al., 2012) and soil respiration (Lavoie et al., 2015) fluxes. The signal-to-noise ratio (SNR, measured as ratios of standard deviations) was varied between 10 and 1 in six steps.

For each of these functions and SNR levels, we sampled 100 validation data points 10 times; 20 GEP runs were performed on the 1000 training data points and the GEP model structure with the highest mean MEF value over the 10 validation sets was chosen.

As the choice of fitness function was crucial for the construction of structures in a GEP type of approach, we also investigated in one experiment the effects of minimizing the CEM values (Eq. 3) as opposed to using only MEF (Eq. 1) or MEF + NP (Eq. 5) as a fitness function.

Table 1. GEP settings.

Parameter	Artificial data	Real observations
Number of chromosomes	2000	2000
Number of genes	3	2
Head length	5	6
Functions	$+, -, /, *, x^y, \sqrt{}$, ln, exp, sin, cos	$+, -, /, *, x^y, \sqrt{}$, ln, exp
Terminals	x_1, x_2, x_3	GPP_s, T_{Air}, T_{-10}, SWC
Link function	+	+
Max run time	1200 s	1800 s
Fitness function	CEM	CEM
Selection method for replication	tournament (Coello and Montes, 2002)	tournament
Mutation probability	0.2	0.2
IS and RIS transposition probabilities	0.05	0.05
Two-point recombination probability	0.3	0.3
Inversion probability	0.05	0.05
One-point recombination probability	0.4	0.4

Alternative machine learning methods

The prediction performance of the best GEP-derived models based on the data in Sect. 3.1 was compared with the prediction performance of four commonly used state-of-the-art machine learning methods (MLMs), i.e. artificial neural networks, ANNs (Yegnanarayana, 2006), support vector machines, SVMs (Hearst, 1998), random forests, RFs (Breiman, 2001) and kernel ridge regressions, KRRs (Hoerl and Kennard, 1970).

The toolboxes and settings used for generating the predictions by the ANN and KRR methods are described by Tramontana et al. (2016) and found in the "simple R" regression toolbox (Lazaro-Gredilla et al., 2014). The predictions of the SVM were obtained by using the "LIBSVM" library (Chang and Lin, 2011) from the "SimpleR" regression toolbox where the regularization term, the insensitivity tube (tolerated error) and a kernel length scale were automatically adjusted during each run. Lastly, the RF predictions were obtained after running the MATLAB statistics toolbox implementation with default settings. The hyper-parameters of all MLMs were estimated to avoid overfitting during each run as presented in Sect. S6 of Tramontana et al. (2016).

All the present machine learning approaches have been applied to the same training data sets as those used for building the GEP models, and their predicted values were compared with the validation sets used for determining the best GEP solution.

3.2 Measured ecosystem CO_2 fluxes

In the second experiment we assessed the possibility of reverse-engineering model structures R_{eco} and its components based only on real measured data. Specifically, we explored GEP-derived model structures for various components of terrestrial ecosystem respiration fluxes measured in an 80-year old deciduous oak plantation in the Alice Holt forest in south-eastern England as described in Heinemeyer et al. (2012) and Wilkinson et al. (2012).

3.2.1 Alice Holt in situ data

The Alice Holt data set contains observations of R_{eco} and the total influx of CO_2 to the ecosystem as mediated via photosynthesis (gross primary production, GPP) and various soil respiration components.

R_{eco} and GPP were estimated from eddy covariance measurements of the forest net CO_2 exchange (NEE, Eq. 17) and were obtained from a micro-meteorological measurement tower at the same site that reports half-hourly integrals of NEE with the eddy covariance (EC) methodology (Moncrieff et al., 1997). The Reichstein et al. (2005) procedure was used for gap-filling and separation of NEE into GPP and R_{eco}. Given that R_{soil} is a fraction of R_{eco}, above-ground respiration can be calculated as the difference between R_{eco} and R_{soil}. For an in-depth description of other site conditions and measurements, see Heinemeyer et al. (2012).

A multiplexed chamber system was used for separately measuring soil respiration (R_{soil}) and its components, using a continuous sampling method at fixed locations during 2 years at an hourly resolution. In order to partition the R_{soil} flux into its components, mesh bags that are not penetrable by roots but allow for mycorrhizal hyphae development were installed. Deep steel collars were applied to stop both root and mycorrhizae in-growth. As a result, root respiration (R_{root}) is given by the difference of R_{soil} and the respiration recorded in the mesh bag chambers, mycorrhiza respiration (R_{myc}) is given by subtracting the steel collar flux from the mesh bag chamber flux, and the soil heterotrophic respiration (R_{soil_h}) is given by the CO_2 efflux at the steel collar chambers. Lastly, soil autotrophic respiration (R_{soil_a}) is estimated as the sum of R_{myc} and R_{root} (Eqs. 18 and 20).

The above-ground respiration (R_{above}) was given as well and was estimated by difference (Eq. 17). Additionally, direct measurements of soil moisture (SWC), air temperature, surface temperature, and soil temperature taken at 2, 10 and 20 cm depths are present in the data set.

$$R_{\text{eco}} = \text{NEE} + \text{GPP} \tag{17}$$

$$R_{\text{above}} = R_{\text{eco}} - R_{\text{soil}} \tag{18}$$

$$R_{\text{soil}_\text{a}} = R_{\text{root}} + R_{\text{myc}} \tag{19}$$

$$R_{\text{soil}} = R_{\text{soil}_\text{a}} + R_{\text{soil}_\text{h}} \tag{20}$$

The computation of R_{above} as the difference between R_{eco} and R_{soil} might be highly uncertain because of the different techniques used to compute the two respiration components, the completely different footprints, and the typical high flux underestimation and low flux overestimation of R_{eco} from EC (Wehr et al., 2016). The limitations of the separation of R_{eco} into its components and the uncertainty of the estimates are further discussed by Heinemeyer et al. (2011, 2012) and Wilkinson et al. (2012).

3.2.2 Data processing

We used the following candidate driver variables: soil volumetric moisture measurements, air temperature (from micro-meteorological stations), temperatures at different soil depths, and GPP. A number of recent studies have shown a tight linkage between GPP and R_{soil}, reflecting dynamics of respiratory substrate supply to roots and mycorrhizal fungi from recently assimilated C in plants. (Moyano et al., 2008; Mahecha et al., 2010; Migliavacca et al., 2011, amongst others). We use GPP obtained from EC measurements at the site, but acknowledge the conceptual problem that R_{eco} and GPP were derived from the same observations of NEE. In order to minimize the potential spurious correlation between R_{eco} and GPP as well as redundancy of possible GPP influence with the meteorological drivers, we considered low-frequency variability of GPP only (i.e. low-pass filtered modes of GPP which correspond to variability beyond a 60-day periodicity only; see Mahecha et al., 2010). Singular spectrum analysis (SSA, Broomhead and King, 1986) as described and implemented by Buttlar et al. (2014) was used to obtain a smooth GPP signal. The seasonal cycle was extracted with the SSA method as the assumption is that GPP affects mainly the seasonality of the respiration, while the variability at the high frequency is assumed to be more related to meteorological drivers (e.g. temperature, Mahecha et al., 2010). The SSA method is a tool used mainly in time series analysis with the purpose of decomposing a time series signal into its independent sum components, such as trends, seasonality and high-frequency components based on a singular value decomposition of trajectory matrices computed after embedding the time series (Buttlar et al., 2014).

To reduce the skewness and the search space that the GEP evolution would have to cover in order to construct valuable solutions (Keene, 1995), we log-transformed the seven target respiration data sets (see Fig. S1 in the Supplement) and applied a back-transformation when reporting the respective model structures. Manning (1998) and Newman (1993) show that when regressions are built based on log-transformed targets, the back-transformation of the regressions to non-transformed target needs to include a bias correction that refers to the residuals of the log models.

As such, if the log model is $\log y = \alpha x + \epsilon$, the back-transformation to y should not simply be $y = e^{(\alpha x)}$, but should include a correction of the bias induced by ϵ, and depending on the distribution of the residuals, the back-transformation can be

- $y = e^{(\alpha x + 0.5\sigma_\epsilon^2)}$, when the residuals are log-normal distributed;
- $y = e^{(\alpha x)} E\,(e^{\epsilon})$, where E is the mean of the sample, when the residuals show heteroscedasticity, as was the case for most of the residuals computed for the GEP models as seen in Fig. S2; and
- $y = e^{(\alpha x)}$ if no bias correction is desired, or a naive approach.

The time series used for the candidate driver observations remain unchanged.

3.2.3 GEP set-up

For each combination of respiration target and possible drivers, 50 subsets of 500 target time steps each were randomly selected and used for the training of GEP models using the settings found in Table 1. The 50 subsets of the remaining 113 time steps are used for cross-validation and the model with the lowest average validation CEM value is finally selected for each respiration type. For all runs the observations are given as records of daily mean values.

We were particularly interested in determining the general character of each extracted model with respect to the different respiration fractions. We therefore re-optimized the parameters of all extracted model structures when applying one extracted model as the candidate function for a different respiration term. For example, the model formulation extracted for R_{eco} is re-calibrated for all the other types of respiration, creating six parameter sets (one for each respiratory flux) per equation. To cross-validate parameter sets, we computed performances for each train–validation data set pair and report averaged MEF values.

As in the artificial example, we compared the returned GEP solution prediction performance with that of other common MLMs such as SVN, KRR, ANNs, and RF. All methods were used to generate 50 subsets of 113 prediction values, after training on the 50 subsets of 500 time steps of observations presented at the start of Sect. 3.2.3. Then, a mean MEF value was computed for all methods for all respiration components and the best mean MEF values were reported

Table 2. Respiration model formulations commonly used in the environmental science community.

Model	Formulation	Reference
Arrhenius	$a \times e^{-E_0/RT}$	Lloyd and Taylor (1994)
Q_{10}	$\phi_1 \times \phi_2^{\left(\frac{T-T_{\text{ref}}}{10}\right)}$	Reichstein and Beer (2008)
Water Q_{10}	$\phi_1 \times \phi_2^{\left(\frac{T-T_{\text{ref}}}{10}\right)} \times \frac{\text{SWC}}{\text{SWC}+\phi_3} \times \frac{\phi_4}{\text{SWC}+\phi_4}$	Richardson et al. (2008)
LinGPP	$(R_0 + k_2\text{GPP}) \times e^{E_0\left(\frac{1}{T_{\text{ref}}-T_0} - \frac{1}{T_{\text{A}}-T_0}\right)} \times \frac{\alpha k+\text{SWC}(1-\alpha)}{k+\text{SWC}(1-\alpha)}$	Migliavacca et al. (2011)
ExpGPP	$\left[R_0 + R_2\left(1 - e^{k_2\text{GPP}}\right)\right] \times e^{E_0\left(\frac{1}{T_{\text{ref}}-T_0} - \frac{1}{T_{\text{A}}-T_0}\right)} \times \frac{\alpha k+\text{SWC}(1-\alpha)}{k+\text{SWC}(1-\alpha)}$	Migliavacca et al. (2011)
addLinGPP	$R_0 \times e^{E_0\left(\frac{1}{T_{\text{ref}}-T_0} - \frac{1}{T_{\text{A}}-T_0}\right)} \times \frac{\alpha k+\text{SWC}(1-\alpha)}{k+\text{SWC}(1-\alpha)} + k_2\text{GPP}$	Migliavacca et al. (2011)
addExpGPP	$R_0 \times e^{E_0\left(\frac{1}{T_{\text{ref}}-T_0} - \frac{1}{T_{\text{A}}-T_0}\right)} \times \frac{\alpha k+\text{SWC}(1-\alpha)}{k+\text{SWC}(1-\alpha)} + R_2\left(1 - e^{k_2\text{GPP}}\right)$	Migliavacca et al. (2011)

a, E_0, ϕ_1, ϕ_2, ϕ_3, ϕ_4, R_0, R_2, k, k_2 and α are model parameters that can be optimized.

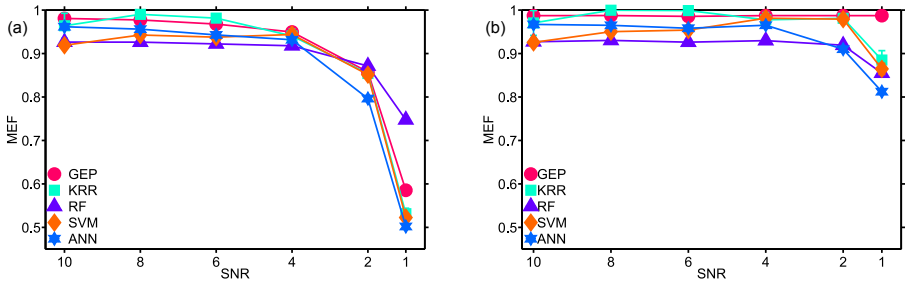

Figure 4. Effect of adding noise to the original signal on the prediction capacity for GEP, KRR, RF, SVM and ANN. The first panel contains the evolution of mean modelling efficiency (MEF) values from 20 independent runs for each increasing level of noise. MEF is computed after learning from a data set of 200 data points and validating against 1000 data points containing noise. The second panel shows the evolution of mean MEF values from 20 independent runs for each increasing level of noise where MEF is computed after learning from a data set of 200 data points and validating against noise-free 1000 data points generated from Eq. (14).

and compared with those of the GEP-extracted models. The comparison is done in terms of MEF as a number of model parameters were not available and CEM could not be computed.

3.2.4 GEP in the context of other known ecological models: real observational data

A comparison was done between the GEP-built models and some common literature respiration models with different structures and driving variables that were also optimized using CMA-ES. The optimization was performed for each respiration data set and its candidate drivers and parameters (Table 2). The structures and prediction performances of the GEP models were then compared with those of the optimized literature models.

4 Results

4.1 Artificial experiments

In the first artificial experiment the GEP approach is used to verify whether it can reconstruct prescribed functions. Following the training of the 20 independent GEP runs, the initial functions were successfully reconstructed for all 10 equations defined in Sect. 3.1.

For the Q_{10} model artificial test, the following structure was finally selected:

$$R_{\text{eco}} = 0.35 \times 2.5^{(0.01 T_{\text{air}})}, \quad (21)$$

with a validation MEF value > 0.99.

MEF values for the GEP-extracted models and for the predictions generated by ANN, RF, KRR and SVM are illustrated in Fig. 4. These MEF values were obtained through cross-validation against independent yet equally noise-contaminated data points (the SNR values are given on the x axis in reverse order to visualize the increase in noise levels). There is a clear pattern of decreasing MEFs with in-

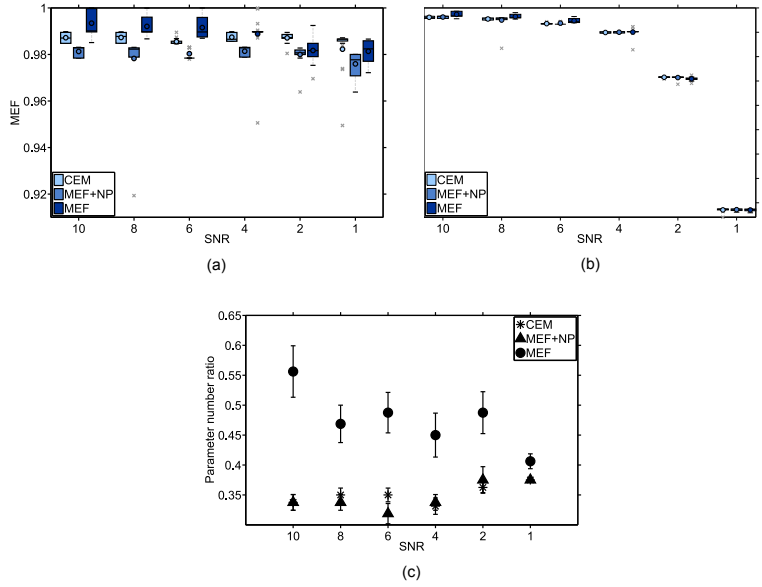

Figure 5. Effects on modelling performance and parameter number caused by choice of fitness function during GEP training for artificial noisy data generated by Eq. (14), where MEF is defined in Eq. (1) and CEM is defined in Eq. (3). **(a)** Mean MEF when validated against noisy data after 20 GEP runs with different fitness functions. **(b)** Mean MEF when validated against noise-free data after 20 GEP runs with different fitness functions. **(c)** Ratio of predicted number of parameters to true number of parameters after 20 GEP runs with different fitness functions.

creasing noise contamination. This was expected, as none of the methods should fit the noise added to the signal.

Figure 4b shows MEF values equivalent to Fig. 4a but applied to noise-free data points of the validation set, in order to compare GEP outputs to the "true" structure underlying the artificial data set. In this set-up, the MEF values remained relatively constant across SNR values above 2. When the SNR level was set to 1, predictions for all investigated machine learning methods, except for GEP predictions, show decreased fitness, with MEF values decreasing to a minimum of 0.8.

In order to verify the effects of changing the fitness function from MEF to CEM, we compare the distributions of MEF values for all runs for all studied SNRs. Figure 5 exemplifies outputs for Eq. (14); Fig. 5a shows a drop in the prediction capacity of the GEP models with noise increase for all types of fitness functions when compared with noise-infused data. This contrasts with the reduced MEF assessed against original data, where a slight drop in MEF with noise increase for the MEF optimization structures was seen, and where the CEM optimized structures show stability in MEF with noise. The new CEM leads to a reduced number of returned parameters compared to MEF (Fig. 5c), as well.

4.2 Measured ecosystem CO_2 fluxes

Applying GEP to the Alice Holt data set yielded a series of model structures for each respiration type. The returned model structures after bias corrected back-transformation are illustrated in Eqs. (22)–(28).

$$R_{\mathrm{eco}} = 1.2\log\left(T_{-10}\right)^{0.8} \times e^{\left(\frac{\mathrm{GPP_s}}{T_{-10}}\right)}, \tag{22}$$

$$R_{\mathrm{above}} = 1.1\mathrm{SWC}^{0.3} \times e^{(0.1\mathrm{GPP_s})}, \tag{23}$$

$$R_{\mathrm{soil}} = 0.04e^{\left(1.1T_{-10}^{0.4}+1.6\mathrm{SWC}\right)}, \tag{24}$$

$$R_{\mathrm{root}} = 1.1e^{\frac{0.9\mathrm{GPP_s}-6.8}{T_{-10}}}, \tag{25}$$

$$R_{\mathrm{myc}} = 0.001T_{-10}^{1.2} \times e^{\left(1.6T_{-10}\right)^{\mathrm{SWC}}}, \tag{26}$$

$$R_{\mathrm{soil_a}} = 0.01e^{\left(0.8T_{-10}^{0.6}+2.6\mathrm{SWC}\right)}, \tag{27}$$

$$R_{\mathrm{soil_h}} = 0.8e^{\frac{0.6\mathrm{GPP_s}-2.4}{T_{-10}}}, \tag{28}$$

where $\mathrm{GPP_s}$ is gross primary production that has been smoothed using the SSA method with a 60-day window; T_{-10} is soil temperature measured at 10 cm depth; and SWC is volumetric soil water content. The corresponding cross-validation MEF values are given in Table 3, indicating a range of capacities for GEP models to represent different respiration types.

Whilst GEP-derived models may differ between respiration types, there are a number of equivalent models for different respiration components. R_{soil} and $R_{\mathrm{soil_a}}$ were described by identical model structures (but distinctive parameter values), and R_{root} and $R_{\mathrm{soil_h}}$ were described by similar (but not identical) models. Overall, the most common selected drivers were T_{-10}, SWC and GPP.

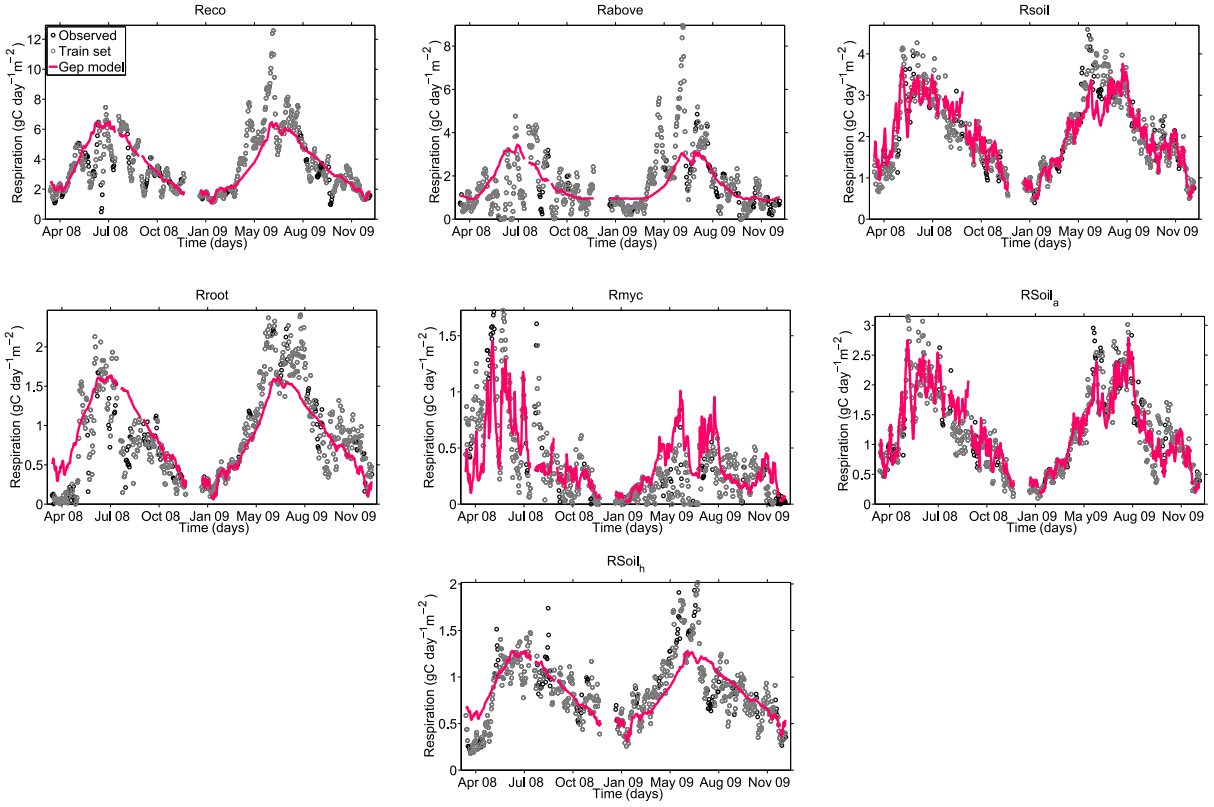

Figure 6. Observed and predicted outgoing CO_2 fluxes; 613 time steps of daily averaged CO_2 effluxes for 2 years at the Alice Holt oak forest site. The predicted values are generated with the models automatically built by the GEP approach with the settings given in Table 1 for the following types of respiration, R_{eco}, R_{above}, R_{soil}, R_{root}, R_{myc}, R_{soil_a}, and R_{soil_h}, and back-transformed with a smear term bias correction. The models are given in Eqs. (22)–(28).

Table 3. Modelling performance for all extracted model structures after cross-validation over 90 cases.

Respiration type	MEF	σMEF	Equation
R_{eco}	0.57	0.13	(22)
R_{above}	0.31	0.23	(23)
R_{soil}	0.79	0.04	(24)
R_{root}	0.59	0.08	(25)
R_{myc}	0.39	0.28	(26)
R_{soil_a}	0.82	0.05	(27)
R_{soil_h}	0.52	0.08	(28)

The highest performance in terms of MEF value was recorded for R_{soil_a} and for R_{soil}, that is, 0.82 and 0.81 respectively. The lowest capacity of process representation, with an MEF value of 0.28, was recorded for R_{above} (Table 3), possibly because this specific component would need to include active versus inactive periods determined by dormancy and leaf fall (i.e. seasonality in this deciduous forest). A comparison of the predicted values and observed fluxes for all types of respiration can be seen in Figs. 6 and 8. Figures 7 and 9 show the effects of the three different types of bias correction on the global signal reconstruction and prediction capacity with MEF values computed in a cross-validation manner. For all respiration types, except R_{soil}, doing the second type of bias correction with a smear term improved the prediction capacity. Although for R_{soil} it seems that doing no bias correction gives a higher MEF value, we chose to keep the model including the smear term.

In order to explore the capacity of the GEP models generated for the R_{eco} components to recreate the larger, across-compartmental summed fluxes, we summed the predictions of the models and compared them with the original fluxes (Fig. 10). Based on a modelling performance comparison of the models defined as sum models of the initial GEP models trained on the component fluxes with the original GEP models trained on the summed fluxes, we found no significant differences after performing a Student's t-test ($h = 0$, $p = 0.5$). However, we found that the total number of parameters is much larger for the sum models. This can be a result of the GEP approach eliminating the "low impact" drivers due to complexity pressure. We can see as well that the sensitivity of the sum fluxes to certain drivers can strongly manifest itself only in certain components, which is why the drivers

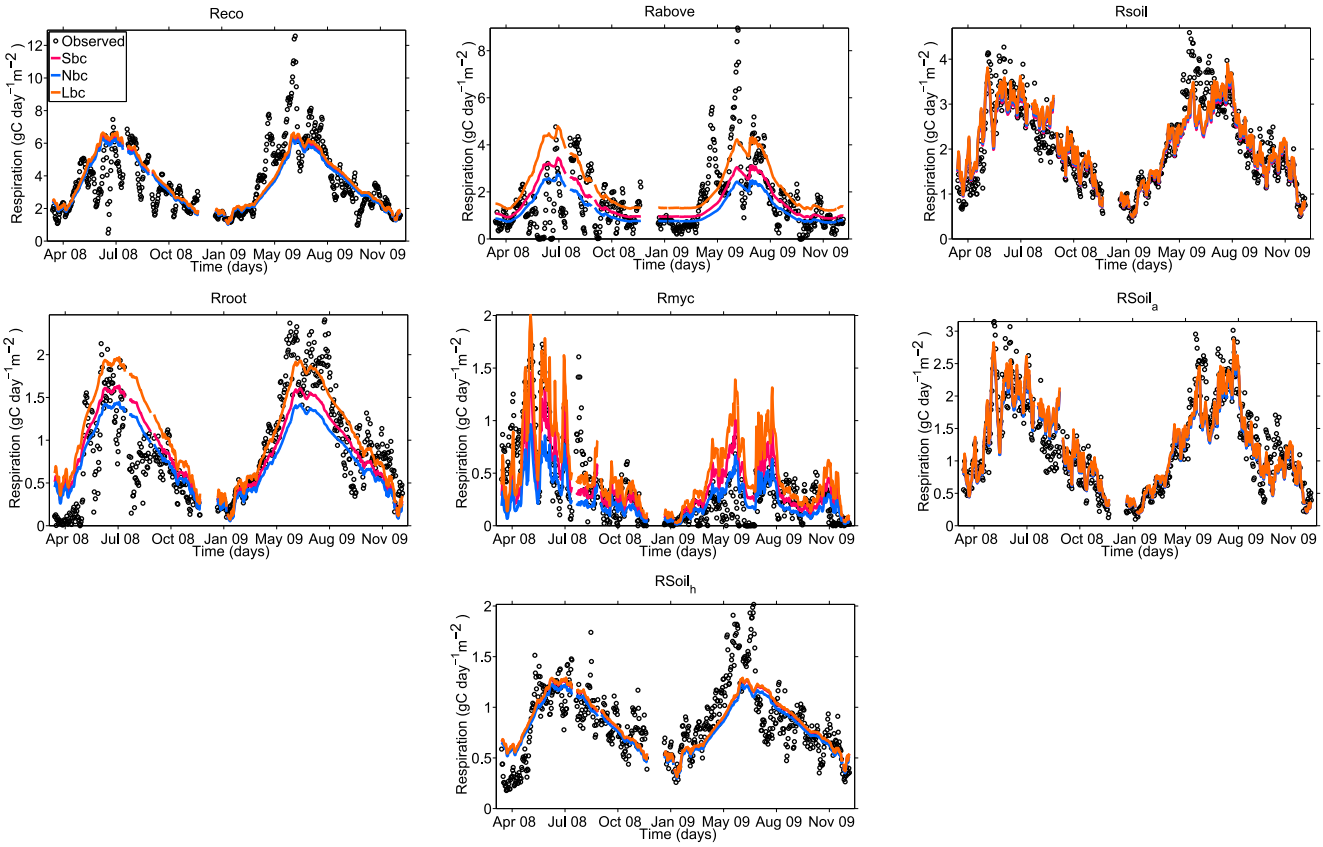

Figure 7. Observed and predicted outgoing CO_2 fluxes; 613 time steps of daily averaged CO_2 effluxes for 2 years at the Alice Holt oak forest site. The predicted values are generated with the models automatically built by the GEP approach with the settings given in Table 1 for the following types of respiration, R_{eco}, R_{above}, R_{soil}, R_{root}, R_{myc}, R_{soil_a}, and R_{soil_h}, and back-transformed with three types of residual **b**ias **c**orrection terms: **s**mear term, **n**aive, and **l**og-normal term.

only get selected in the models built for those specific components.

The residuals depict some remaining patterns (Figs. 11 and S3) and the null hypothesis of normal distribution was rejected for all seven respiration component residuals at the 5 % significance level with the one-sample Kolmogorov–Smirnov test. Hence, we might expect additional information that could be extracted from the residuals. In order to check whether the remaining structure was missed in the first training routine because of imposing a multiplicative form in the models by log-transforming the target data, we performed GEP runs on the residuals and combined the models. The improvement in overall modelling performance is minimal, yet model structures become overly complex. The capacity of the GEP approach to retrieve new information from the residuals is illustrated in Fig. 13 in comparison with that of the other MLM presented in Sect. "Alternative machine learning methods". When correlation values were computed between the candidate drivers and the residuals, no significant linear correlations were found (Figs. S5 and S6).

4.2.1 Model transferability

We investigated the capacity of each extracted model structure (Eqs. 22–28) to represent a component of R_{eco} not seen in the training procedure. This was done by means of new CMA-ES optimization steps. The new prediction performances are illustrated in Table 4.

After optimization, none of the structures show an overall best MEF for all the R_{eco} components (i.e. we clearly cannot identify an optimal general model). However, we identify certain model structures that tend to perform overall better than others. This is the case for the R_{myc} model (Eq. 26). It can also be seen that after the individual model optimizations, the structures for R_{eco} and that for R_{soil_a} have similar prediction capacities.

The prediction capacity of the GEP-generated models in the context of other commonly utilized MLMs was assessed as well. KRR, ANN, SVM and RF were used for generating 113 predicted data points as described in Sect. 3.2 (Fig. 12). The prediction performances of GEP, KRR, ANN, SVM and RF are shown in Fig. 13. Figure 13a contains the average MEF values computed for all MLM methods' predicted val-

Table 4. Average validation MEF performance for all extracted model structures when re-optimized against all other respiration CO_2 flux observations.

Trained for/opt. for	R_{eco}	R_{above}	R_{soil}	R_{root}	R_{myc}	R_{soil_a}	R_{soil_h}
R_{eco} (Eq. 22)	0.57	0.27	0.77	0.58	0.10	0.68	0.42
R_{above} (Eq. 23)	0.56	0.31	0.69	0.44	0.07	0.60	0.46
R_{soil} (Eq. 24)	0.50	0.20	0.79	0.47	0.38	0.82	0.39
R_{root} (Eq. 25)	0.23	0.27	0.57	0.59	0.01	0.65	0.51
R_{myc} (Eq. 26)	0.54	0.22	0.82	0.50	0.39	0.84	0.52
R_{soil_a} (Eq. 27)	0.50	0.20	0.79	0.47	0.38	0.82	0.39
R_{soil_h} (Eq. 28)	0.55	0.26	0.76	0.56	0.06	0.67	0.52

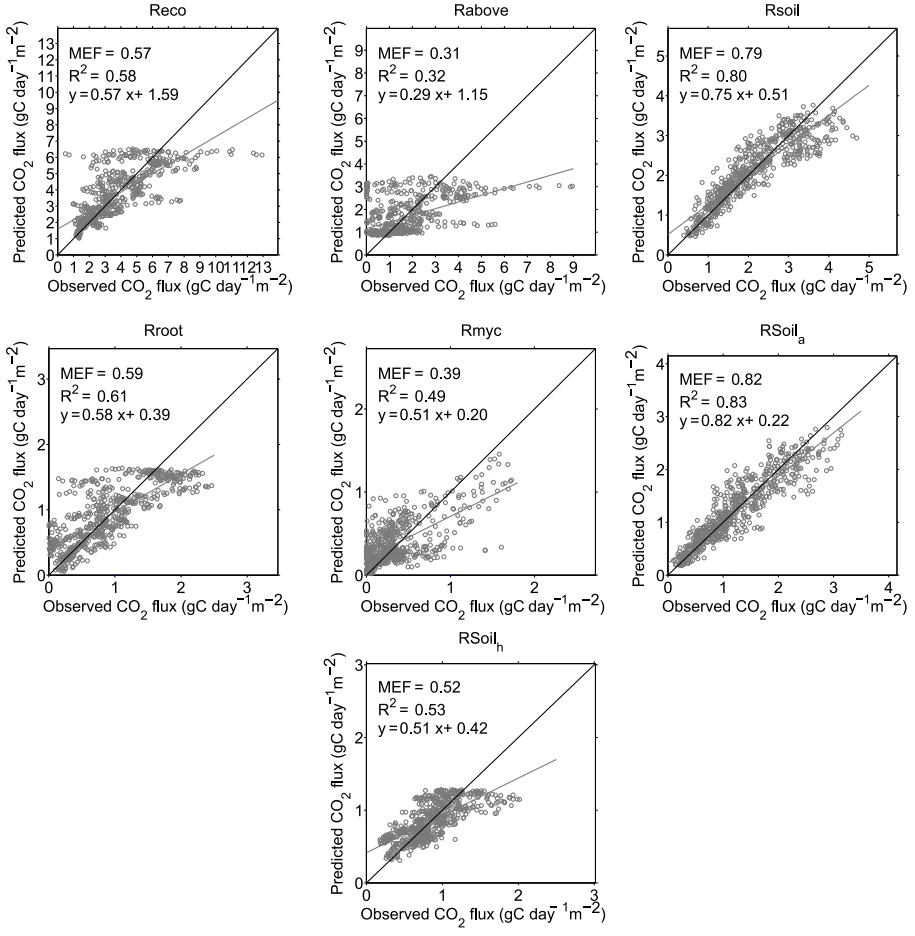

Figure 8. Observed and predicted outgoing CO_2 fluxes; 613 time steps of daily averaged CO_2 effluxes for 2 years at the Alice Holt oak forest site. The predicted values are generated with the models automatically built by the GEP approach with the settings given in Table 1 for the following types of respiration, R_{eco}, R_{above}, R_{soil}, R_{root}, R_{myc}, R_{soil_a}, and R_{soil_h}, and back-transformed with a smear term bias correction. The models are given in Eqs. (22)–(28).

ues when compared to the original observations for R_{eco}, R_{above}, R_{soil}, R_{root}, R_{myc}, R_{soil_a}, and R_{soil_h}. For all other cases, the performance is in the same range for all methods, but the GEP-derived models have the lowest mean MEF values. Figure 13b shows that when all MLMs were trained on the residuals obtained from comparing the GEP outputs with the observations, the GEP approach has the lowest capacity to capture new relevant signals and is strongly outperformed by the rest of the MLM, indicating that the amount of information retrievable by GEP with the current fitness and settings is limited and captured already in the first run.

Table 5. Average validation MEF performance for CMA-ES optimized selected literature model formulations when compared with respiration CO_2 flux observations.

Model formulation	R_{eco}	R_{above}	R_{soil}	R_{root}	R_{myc}	R_{soil_a}	R_{soil_h}
Arrhenius	0.41	0.15	0.65	0.50	0.07	0.61	0.38
Q_{10}	0.47	0.19	0.69	0.52	0.09	0.62	0.46
Water Q_{10}	0.50	0.20	0.79	0.55	0.40	0.81	0.43
LinGPP	0.55	0.25	0.74	0.57	0.17	0.70	0.49
ExpGPP	0.58	0.30	0.76	0.57	0.20	0.72	0.54
addLinGPP	0.55	0.27	0.73	0.56	0.12	0.67	0.48
addExpGPP	0.56	0.27	0.73	0.54	0.20	0.69	0.49

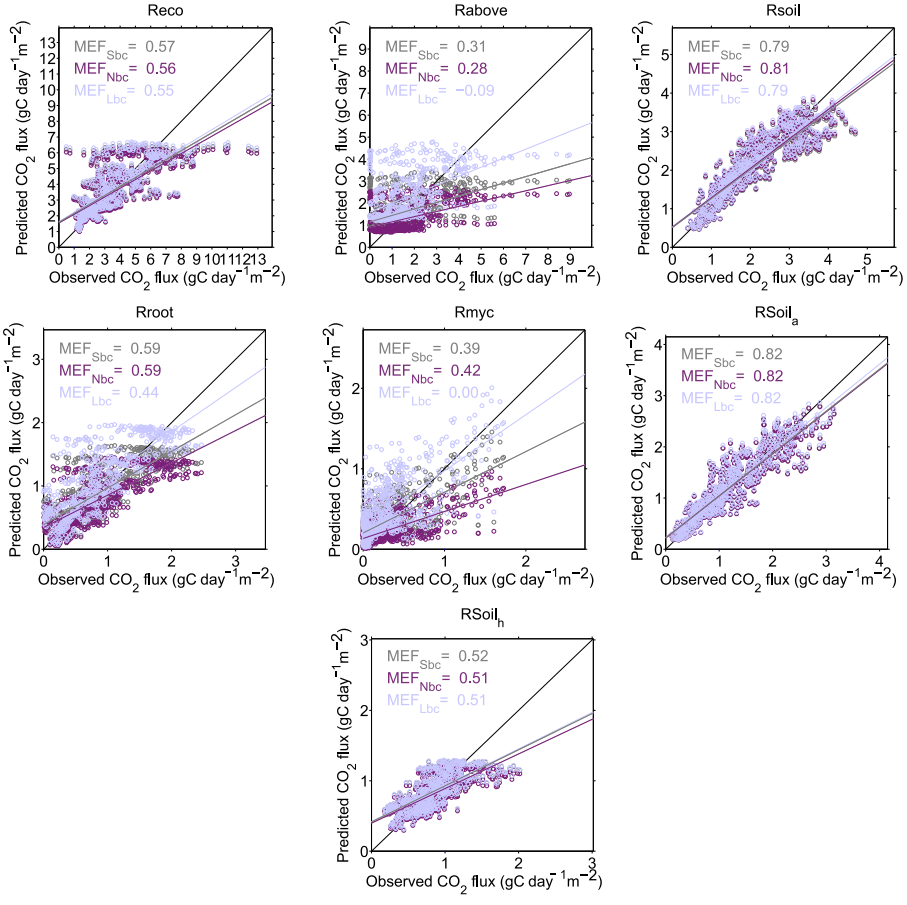

Figure 9. Observed and predicted outgoing CO_2 fluxes; 613 time steps of daily averaged CO_2 effluxes for 2 years at the Alice Holt oak forest site. The predicted values are generated with the models automatically built by the GEP approach with the settings given in Table 1 for the following types of respiration, R_{eco}, R_{above}, R_{soil}, R_{root}, R_{myc}, R_{soil_a}, and R_{soil_h}, and back-transformed with three types of residual **b**ias **c**orrection terms: **s**mear term, **n**aive, and **l**og-normal term. The figure contains the MEF values for each type of bias correction in each respective colour.

4.2.2 Comparing with literature models

Lastly, the GEP-generated models were compared with some of the most commonly used literature models for describing respiration. The resulting MEF values obtained after individual parameter optimization using the CMA-ES procedure for each literature model are given in Table 5. The literature model structure that performed best overall in terms of prediction capacity measured as MEF is the WaterQ_{10} model (Fig. 14). Figure 14 shows as well that certain types of respiration are easier to represent by all models, including the models GEP generated, whilst other types of respiration are poorly predicted by all models. Nevertheless, for all respira-

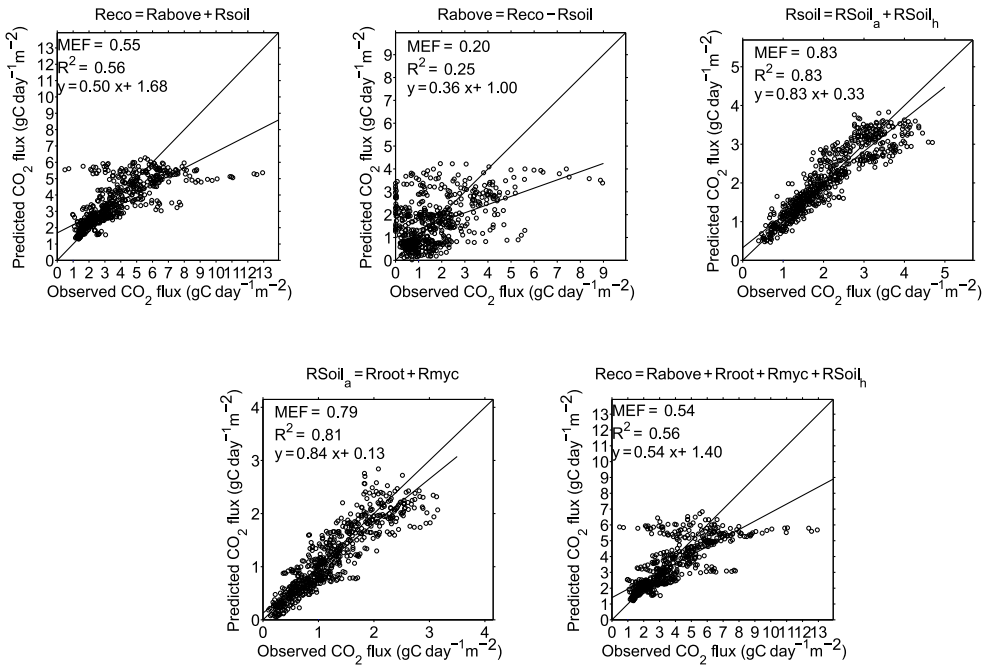

Figure 10. Observed versus predicted R_{eco} component fluxes, where predicted values are computed as derived fluxes based on the GEP models given in Eqs. (22)–(28) that were trained on 500 data points (d.p.) of daily mean values of various R_{eco} components.

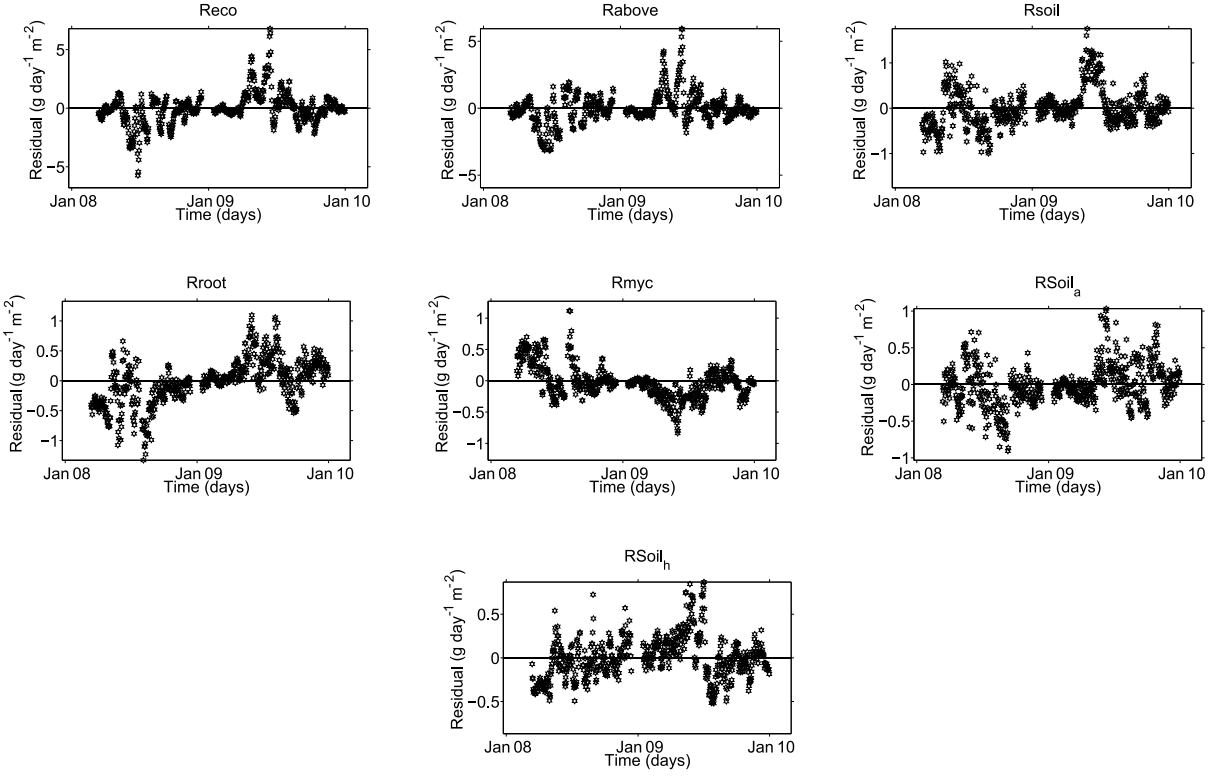

Figure 11. Residuals computed for smear term bias corrected back-transformed GEP models for various types of CO_2 respiration fluxes after training against log-transformed targets with the settings given in column 2 of Table 1.

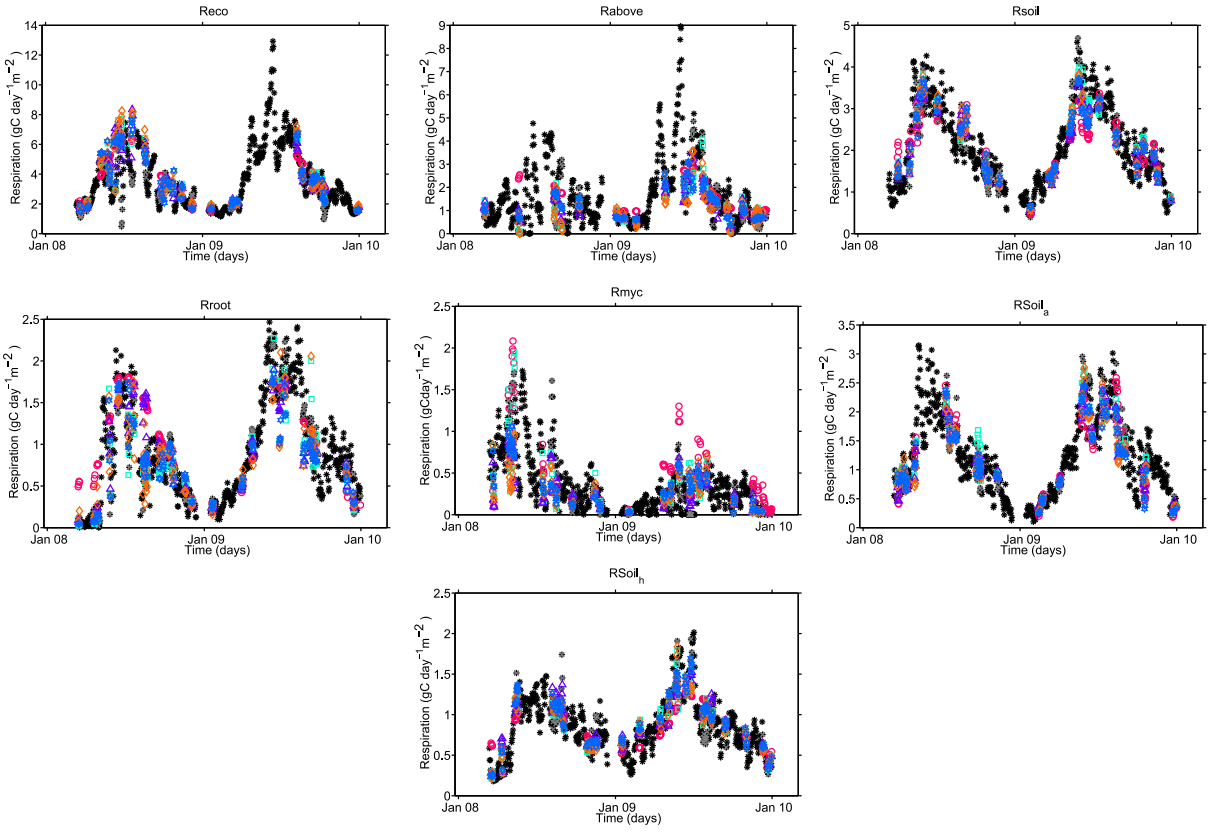

Figure 12. Observed CO_2 fluxes and one set of 113 predicted values given by the some common machine learning methods (MLMs) after training on 500 data points and after smear term bias corrected back-transformation.

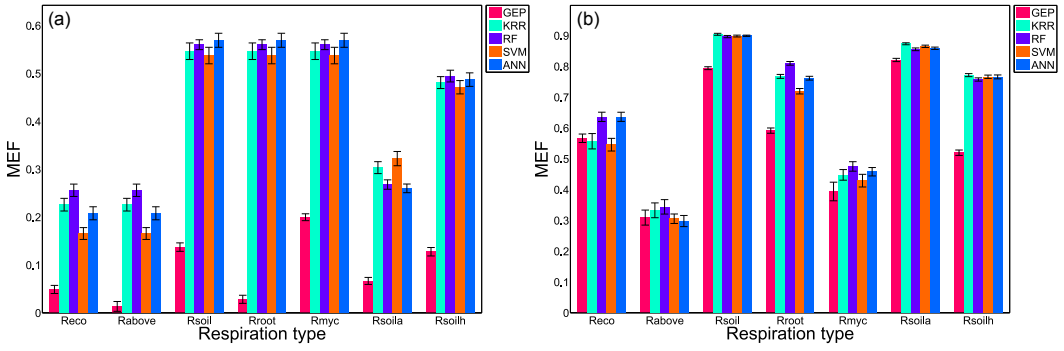

Figure 13. Machine learning methods (MLM) prediction performance for all respirations components **(a)** and for the residuals **(b)** resulting from the GEP trained models after smear term bias corrected back-transformation. The MEF values obtained for validation by all the MLM methods for R_{eco}, R_{above}, R_{soil}, R_{root}, R_{myc}, R_{soil_a}, R_{soil_h}.

tion types, the highest MEF values are generally recorded by the GEP models.

As the studied literature models performed best in modelling R_{soil}, we focus on contrasting GEP model results with literature model outcomes for this ecosystem respiration component. Of all models included, the GEP model and Q_{10} model including SWC dependency captured seasonal variability best, but no model satisfactorily represented short-term CO_2 flux variations (Fig. 15a). All models show the largest range of residuals for the months May to July in 2008, and June/July in 2009 (Fig. 15b), with the two best-performing models (GEP and WaterQ_{10}) having the narrowest range of absolute residuals. Monthly mean average errors (MAEs) indicate as well a systematic underestimation of soil CO_2 efflux in the first year (Fig. S4).

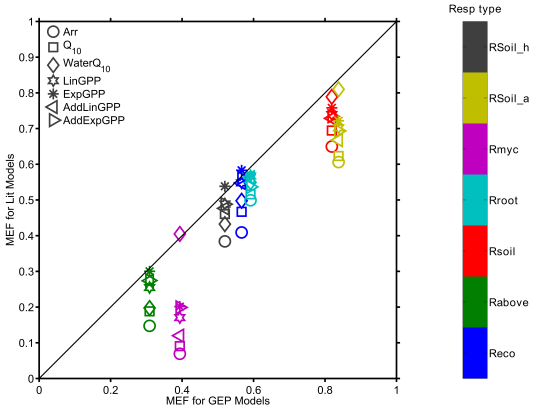

Figure 14. MEF validation values for literature models and for the best GEP model in terms of MEF at each respiration level. Each R_{eco} flux component is shown in a separate colour.

5 Discussion

5.1 On the GEP method

In this work, the primary reason for the artificial experiments was to obtain a better understanding of the capacity of GEP to solve symbolic regression types of problems. We put an emphasis on GEP performance in the presence of noise. This aspect was important, given that monitoring data from terrestrial ecosystem CO_2 effluxes are typically contaminated by sometimes substantially large random uncertainties and measurement noise. In the case of NEE flux measurements, Lasslop et al. (2008) and Richardson et al. (2008) show that the measurement error typically scales with the magnitude of the flux, leading us to simulate that type of situation by adding noise that scales with signal to an already known function, Eq. (14). The results show that all the studied methods are stable in the presence of noise in the training set. These results increase our confidence in the predictions generated by studied machine learning methods; in particular, GEP-derived modes can tolerate SNRs of 1. Considering that the SNR in the R_{eco} observations (if noise is only considered as a random error) is probably larger than 4, which is where the curve starts decreasing in Fig. 4, the noise presence in the data should not influence the automated model construction process and the real signals should be accurately captured when data uncertainties follow the pattern described here. On the other hand, for R_{soil} and other CO_2 fluxes measured with other techniques, the magnitude and the distribution of the uncertainty can be different (Ryan and Law, 2005; Pérez-Priego et al., 2015), and we cannot state what the response of the present MLM is in the presence of different types of uncertainties and measurement noise.

Our findings illustrate that the selection of CEM over MEF as a fitness function for optimization has a minor effect on the global mean MEF (Fig. 5). We also notice that due to applying constraints on the presence of structures in the residuals and the length of the parameter vector, the final mean number of parameters is lower when CEM is chosen.

Limitations

One of the critical aspects in our work is that GEP, as implemented here, can only represent and derive "$n \rightarrow 1$" types of response functions. We are not able to generate model structures that encode e.g. system-intrinsic dynamics like feedback loops, which are expected from our current understanding of biogeochemical cycles in terrestrial ecosystems (Ehrenfeld et al., 2005; Friedlingstein et al., 2006). Hence, we believe GEP is suitable for e.g. understanding and describing the sensitivities and non-linear responses to changes in hydro-meteorological drivers, but fails to represent more complex carbon or soil water dynamics. Pools and pool transfers cannot be introduced currently in the input, unless the inflow–outflow equations are known and can be included in the set of functions that can participate in the evolution.

Lagged responses can only be detected if the number of lags from a driver is correctly included in the input, which already implies sufficient knowledge of their existence and behaviour. Whilst in the current implementation of the GEP algorithm, shifts in conditions and responses cannot be encoded or detected, these could be addressed with the inclusion of a conditional operator in the set of functions encoded in the GEP evolution individuals.

Nevertheless, it would be fair to mention that the same limitations can affect the results of the other MLM and empirical models presented in this paper. A clear advantage ANN, RF and SVM have though over the GEP symbolic regression construction is the fact that when the target variable presents a skewed distribution, log-transforming of the target data is recommended for regression types of methods, such as GEP (Keene, 1995), whereas there is no effect on the prediction capacity of the other MLM as far as we are aware. Moreover, such a log-transformation needs a back-transformations that might induce a bias if the right correction is not performed (Manning, 1998). For these reasons, in cases where less steps in obtaining predictions are desired and no mathematical expression of the models needed to obtain the predictions are needed, non-GEP approaches might be recommended.

5.2 The value of GEP for modelling ecosystem respiration fluxes

We automatically generated a series of model structures to describe terrestrial CO_2 respiration fluxes (Eqs. 22–28) with the GEP approach. Most of these structures (five out of seven) were of rather low complexity, requiring only four free parameters and allowing for further interpretation. The most complex structure is found for the R_{myc} representation, which is in line with previous findings (Shi et al., 2012).

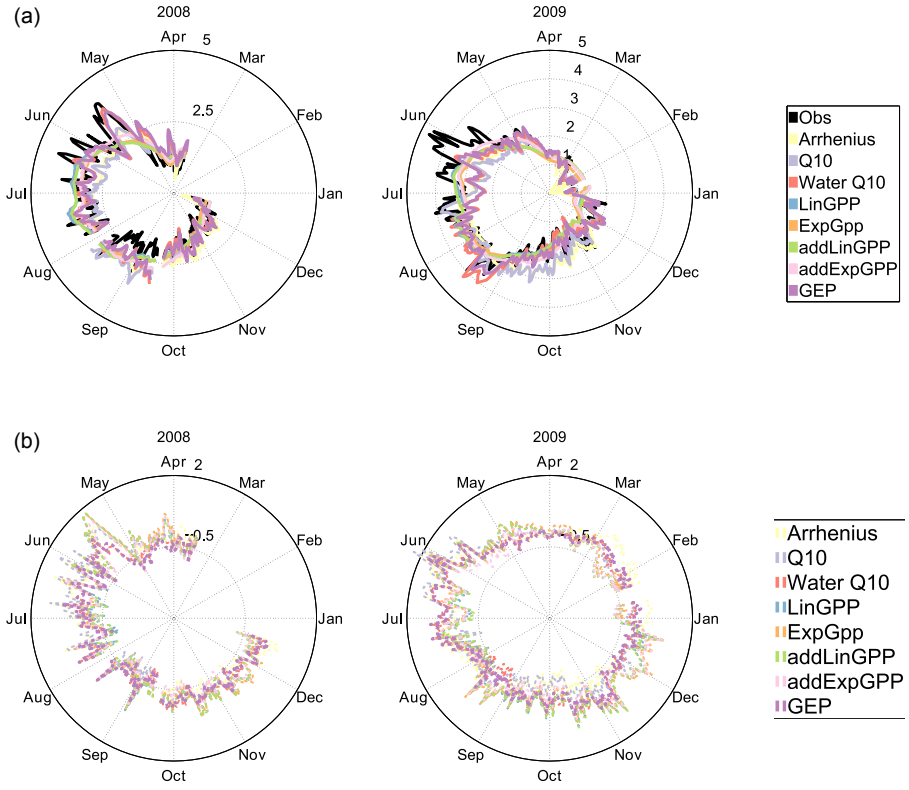

Figure 15. Daily R_{soil} fluxes **(a)** illustrated in the context of the 2 studied years and residual values **(b)** of the total soil daily CO_2 outgoing fluxes as simulated by the investigated literature models and the GEP emerged model after smear term bias corrected back-transformation. The fluxes shown here are the real flux measured at the site and the predicted fluxes generated according to the GEP model and some of the models used in the environmental science community. The centre of the plots in the second row is -1. The scale of the fluxes is given in $gC\,m^{-2}\,day^{-1}$.

Interestingly, the models derived for R_{eco} and R_{soil} are structurally very similar. That is also the case of R_{root} and heterotrophic respiration, where the difference lies in the set of parameters and the added presence of an intercept in the formulation of the R_{soil_h} model. This finding suggests a consistency in the response of the R_{soil} components to their drivers, considering that the separation of the R_{soil} into its components might still lack accuracy (e.g. Hanson et al., 2000; Kuzyakov, 2006; Subke et al., 2006; Heinemeyer et al., 2011).

When we compared the GEP-derived models with the community established semi-empirical models from a structural point of view, we found that they shared some key features for temperature dependencies of CO_2 fluxes, which are typically captured by exponential relationships, but reveal some previously unconsidered dynamics as well.

A major difference was in the response of the respiration components to SWC, where the GEP models often chose SWC as one of the drivers. Moreover, the GEP models often contained an exponential dependency, i.e. there are only certain parts of the signal that are strongly sensitive to varying SWC. We believe that the exponential dependency of terrestrial ecosystem respiration components on SWC is a very intuitive pattern that has not yet been reported in the literature and requires further exploration.

Another difference we found was the strongly seasonal response of the respiration components to GPP, possibly as a proxy to light and vegetation availability which were not included in the set of candidate predictors.

Considering that GEP identified plausible models that are very different structurally from previously reported semi-empirical models, still yielding equivalent or better modelling performance, the validity of the conventional semi-empirical models can be questioned. Nevertheless, we do believe that there is a need for more in-depth analysis for determining whether the GEP described processes make actual biological sense and whether the selected drivers and their interactions represent true processes and responses.

5.3 Data quality

During our study, it was apparent that the highest MEF values were obtained for all the studied methods in the case of the respiration types that had direct measured observations and were not derived. It might be the case that when fluxes are

obtained from derivations, the measurement error will also increase, and the partition of a clear signal existing in the observations is not sufficient for constructing a good model with GEP.

5.4 High-frequency variability

All GEP-generated models underestimated the high respiration fluxes (Fig. 8) and typically did not capture the fast responses. This phenomenon was in some cases a systematic pattern, and sometimes affected only certain times of the year. Similarly, semi-empirical models struggled to adequately simulate CO_2 flux peaks and in some cases monthly flux averages (Fig. 15).

A more in-depth comparison of all the GEP and conventional respiration models, based on a timescale-dependent assessment of model–data mismatch (Mahecha et al., 2010), could help to further elucidate the problem and clarify some of the strengths and weaknesses of the different modelling approaches, especially when seasonal mismatches appear. Nevertheless, a detailed timescale-dependent assessment is beyond the scope of this study, and for such an analysis, the current time series are simply too short.

The question is whether the GEP method lacks the ability to build models that correctly represent the processes and their fast dynamic responses, or whether the candidate drivers and the observations used for their representation are simply not sufficient for generating representative models. In the end, the response of R_{soil} and R_{eco} to external drivers might be too complex to describe solely with the currently available measurements and with the selected drivers.

We believe that the consistent underestimation of fast responses was partly due to surface moisture affecting litter decomposition and fungal activity, as soil moisture was only monitored over the average 8 cm surface, with the top few centimetres most likely presenting the highest activity and partly due to some potential processes/drivers like lags between GPP and respiration (Hölttä et al., 2011) or phenology (Migliavacca et al., 2015) that were not specifically included in the learning process.

Another explanation for missing some of the (high flux) variability could be in our choice of fitness function. As we decided to penalize during the learning process for structures with many parameters, it is likely that some structures were eliminated early on during this process, even though they may be well suited for describing a given process from a modelling efficiency point of view. However, this is a case of trade-off between a good fit and structural simplicity, and in our approach, we decided that simplicity of structure, i.e. the possibility of interpretation, is a very important asset.

We explored as well the possibility of the underestimation of the carbon flux variability being caused by the log-transformations applied to the observations. It could have been the case that the log-transformations excluded interesting components of the model structures by forcing the method to build multiplicative models. Nevertheless, when the GEP was run again on the residuals, without log-transforming, no new meaningful information was retrieved, indicating that multiplicative models were sufficient for reconstructing the R_{eco} components present in this study.

5.5 Equifinality

Table 4 shows that when optimizing the parameters for all structures, the prediction performance becomes similar, which leads to the question of equifinality of dynamical systems, where different models that try to capture their structure might have different formulations but represent the same response.

A critical question for the applicability of any ecosystem model is whether the model structure is more important than the parameterization of a given "best" model. For this question to be addressed, however, we need a larger sample of ecosystem types representative of different types of responses where we can explore the importance of the obtained structures and their parameter sets.

5.6 GEP models in the context of other machine learning methods

The comparison of GEP-generated models and machine learning methods showed a narrow range of predicted fluxes (Fig. 13). The analysis of training all the MLMs on the GEP residual output showed that the GEP approach is not able to retrieve any new meaningful structural components, but that the remaining MLMs are much better at reconstructing the signal left in the residuals. This indicates that although the GEP is actually a reliable MLM when it comes to reconstructing the underlying R_{eco} fluxes and is not prone to overfitting, it could be that the current set-up of the GEP is not sufficient for an exhaustive description of those fluxes, or that it might be overly strict on the complexity of models compared to other MLMs. The GEP approach has, nevertheless, the benefit of producing mathematical model structures that can be the basis for future interpretation.

6 Conclusions and outlook

Overall, our results suggest that the GEP approach is a potentially powerful tool of reverse engineering, particularly helpful for building ecological models when there is a minimum of a priori system understanding. We exemplified this conceptually using artificial data, but also show that GEP always yields as good or better results compared to conventionally used models in the case of ecosystem respiration. Based on data from a long-term monitoring site of different respiratory fluxes, and using GEP as a reverse engineering tool, we found new structures for modelling R_{eco} components. The GEP-derived models outperform conventionally used models and generally differ by the way temperature and GPP but also

SWC are interpreted, indicating that conventional respiration models might have to be revised. At the same time, we found that when the GEP-derived models are mutually compared, there are sufficient structural particularities for each terrestrial respiration type so as to not allow for the formulation of a general R_{eco} law. More research is needed on a larger set of sites to identify widely usable models and for their interpretation. A particular matter of concern is the apparent equifinality of selected model structures, indicating that many response functions are yielding predictions of almost similar quality. A study of multiple sites would enable an investigation of whether specific ecosystem types result in similar model structures, or whether response functions apply across contrasting ecosystem types.

The current study has also revealed methodological aspects that could be improved. In particular, we found the inclusion of a parameter optimization step very helpful to further test the transferability of model structures. But this approach could be potentially integrated into the GEP evolution. More specifically, we think that the next development of GEP could include the parameter optimization as an intermediate step before selection during each evolution generation (Ilie et al., 2017). In this way, a model structure could be chosen according to not only the current state of parameters, but also according to its potential, and convergence to a global solution might be achieved faster.

Code and data availability. All code and data used to produce the results of this paper can be provided upon request by contacting Iulia Ilie or Miguel D. Mahecha.

Appendix A: Glossary

expression tree	binary tree used to represent algebraic expressions
gene	set of characters of fixed length that encodes an expression tree
chromosome	individual used in automatically evolving an optimal solution comprised of a set of genes that are connected with a binary operation (e.g. $+ \times -$)
GEP	**g**ene **e**xpression **p**rogramming, machine learning method that evolves chromosome structures with the purpose of minimizing a cost function
genetic operator	operator that produces changes in the structure of a chromosome and the expression tree it encodes by altering the strings representing composing genes (e.g. mutation, inversion, recombination)
evolution	the process of producing an optimal solution by GEP through
generation	time step of an evolution
genetic operator rate	probability of a genetic manipulation occurring during a generation
population	total set of chromosomes that participate at a certain step in the evolution of an optimal solution in the GEP approach
CMA-ES	covariance matrix adaptation evolutionary strategy
MLM	**m**achine **l**earning **m**ethod that can produce predicted values based on a training set
ill-posed problem	a problem for which the solutions might not be unique or unstable, also known as an inverse problem
reproduction	process of generating new individuals for a new generation starting from the present-generation individuals after they go through structure modification and fitness-based selection
individual	GEP entity that is a component of a population during a certain step of the evolution process. Also known as a chromosome.
gene head	initial section of the string that comprises a GEP gene, containing a combination of characters that map to predictors and possible functional transformations
gene tail	end section of the string that comprises a GEP gene, containing only characters that map to predictors
solution	finally selected model structure resulting from a GEP run
hyper-parameter	set of parameters which need to be set for the runs of a machine learning approach

The Supplement related to this article is available online at https://doi.org/10.5194/gmd-10-3519-2017-supplement.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We thank Markus Reichstein for all the useful comments and suggestions.

This work was supported by the International Max Planck Research School for global Biogeochemical Cycles (IMPRS-gBGC), Jena, by the European Union's H2020 research and innovation programme project BACI, grant agreement 640176, and by NOVA grant UID/AMB/04085/2013. The Alice Holt Forest GHG Flux site is funded by the UK Forestry Commission.

The article processing charges for this open-access publication were covered by the Max Planck Society.

Edited by: Sandra Arndt
Reviewed by: two anonymous referees

References

Ashworth, J., Wurtmann, E. J., and Baliga, N. S.: Reverse engineering systems models of regulation: Discovery, prediction and mechanisms, Curr. Opin. Biotechnol., 23, 598–603, https://doi.org/10.1016/j.copbio.2011.12.005, 2012.

Auger, A. and Hansen, N.: A restart CMA evolution strategy with increasing population size, 2005 IEEE Congress on Evolutionary Computation, 2, 1769–1776, https://doi.org/10.1109/CEC.2005.1554902, 2005.

Bandt, C. and Pompe, B.: Permutation entropy: a natural complexity measure for time series, Phys. Rev. Lett., 88, 174102, https://doi.org/10.1103/PhysRevLett.88.174102, 2002.

Bennett, N. D., Croke, B. F., Jakeman, A. J., Newham, L. T. H., and Norton, J. P.: Performance evaluation of environmental models, in: 2010 International Congress on Environmental Modelling and Software Modelling for Environment's Sake, 1–9, http://scholarsarchive.byu.edu/iemssconference/2010/all/247/ (last access: September 2017), 2010.

Beyer, H.-G. and Schwefel, H.-P.: Evolution Strategies, Natrual Computing, 1, 3–52, 2002.

Bonan, G. B.: Forests and climate change: forcings, feedbacks, and the climate benefits of forests, Science, 320, 1444–1449, https://doi.org/10.1126/science.1155121, 2008.

Bongard, J. and Lipson, H.: Automated reverse engineering of nonlinear dynamical systems, P. Natl. Acad. Sci. USA, 104, 9943–9948, https://doi.org/10.1073/pnas.0609476104, 2007.

Breiman, L.: Random forests, Mach. Learn., 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Broomhead, D. and King, G. P.: Extracting qualitative dynamics from experimental data, Physica D, 20, 217–236, https://doi.org/10.1016/0167-2789(86)90031-X, 1986.

Buttlar, J. V., Zscheischler, J., and Mahecha, M. D.: An extended approach for spatiotemporal gapfilling: Dealing with large and systematic gaps in geoscientific datasets, Nonlin. Processes Geophys., 21, 203–215, https://doi.org/10.5194/npg-21-203-2014, 2014.

Chang, C.-C. and Lin, C.-J.: Libsvm, ACM T. Intell. Syst. Technol., 2, 1–27, https://doi.org/10.1145/1961189.1961199, 2011.

Coello, C. A. and Montes, E. M.: Constraint-handling in genetic algorithms through the use of dominance-based tournament selection, Adv. Eng. Inform., 16, 193–203, https://doi.org/10.1016/S1474-0346(02)00011-3, 2002.

Ehrenfeld, J. G., Ravit, B., and Elgersma, K.: Feedback in the plant-soil system, Annu. Rev. Environ. Resour., 30, 75–115, https://doi.org/10.1146/annurev.energy.30.050504.144212, 2005.

Fernando, D., Shamseldin, A. Y., and Abrahart, R. J.: Using gene expression programming to develop a combined runoff estimate model from conventional rainfall-runoff model outputs, in: IMACS/MODSIM Congress, July 2009, 13–17 July 2009, Cairns, Australia, 748–754, 2009.

Ferreira, C.: Gene expression programming: a new adaptive algorithm, in: The 6th Online World Conference on Soft Computing in Industrial Applications, Complex Systems, 13, 87–129, 2001.

Ferreira, C.: Gene expression programming: mathematical modeling by an artificial intelligence, in: vol. 21, 2nd Edn., Springer-Verlag, Berlin, Heidelberg, https://doi.org/10.1007/3-540-32849-1, 2006.

Friedlingstein, P., Cox, P., Betts, R., Bopp, L., von Bloh, W., Brovkin, V., Cadule, P., Doney, S., Eby, M., Fung, I., Bala, G., John, J., Jones, C., Joos, F., Kato, T., Kawamiya, M., Knorr, W., Lindsay, K., Matthews, H. D., Raddatz, T., Rayner, P., Reick, C., Roeckner, E., Schnitzler, K.-G., Schnur, R., Strassmann, K., Weaver, A. J., Yoshikawa, C., Zeng, N., Friedlingstein, P., Cox, P., Betts, R., Bopp, L., von Bloh, W., Brovkin, V., Cadule, P., Doney, S., Eby, M., Fung, I., Bala, G., John, J., Jones, C., Joos, F., Kato, T., Kawamiya, M., Knorr, W., Lindsay, K., Matthews, H. D., Raddatz, T., Rayner, P., Reick, C., Roeckner, E., Schnitzler, K.-G., Schnur, R., Strassmann, K., Weaver, A. J., Yoshikawa, C., and Zeng, N.: Climate–Carbon Cycle Feedback Analysis: Results from the C_4 MIP Model Intercomparison, J. Climate, 19, 3337–3353, https://doi.org/10.1175/JCLI3800.1, 2006.

Gilmanov, T. G., Aires, L., Barcza, Z., Baron, V. S., Belelli, L., Beringer, J., Billesbach, D., Bonal, D., Bradford, J., Ceschia, E., Cook, D., Corradi, C., Frank, A., Gianelle, D., Gimeno, C., Gruenwald, T., Guo, H., Hanan, N., Haszpra, L., Heilman, J., Jacobs, A., Jones, M. B., Johnson, D. A., Kiely, G., Li, S., Magliulo, V., Moors, E., Nagy, Z., Nasyrov, M., Owensby, C., Pinter, K., Pio, C., Reichstein, M., Sanz, M. J., Scott, R., Soussana, J. F., Stoy, P. C., Svejcar, T., Tuba, Z., and Zhou, G.: Productivity, Respiration, and Light-Response Parameters of World Grassland and Agroecosystems Derived From Flux-Tower Measurements, Rangeland Ecol. Manage., 63, 16–39, https://doi.org/10.2111/REM-D-09-00072.1, 2010.

Gupta, H. V., Clark, M. P., Vrugt, J. A., Abramowitz, G., and Ye, M.: Towards a comprehensive assessment of model structural adequacy, Water Resour. Res., 48, W08301, https://doi.org/10.1029/2011WR011044, 2012.

Hansen, N.: The CMA Evolution Strategy: A Comparing Review, Stud. Fuzzin. Soft Comput., 192, 75–102, https://doi.org/10.1007/3-540-32494-1, 2006.

Hansen, N., Müller, S. D., and Koumoutsakos, P.: Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES), Evolut. Comput., 11, 1–18, https://doi.org/10.1162/106365603321828970, 2003.

Hanson, P. J., Edwards, N. T., Garten, C. T., Andrews, J. A., Hanson, P. J., Edwards, C. T. G., and Andrews, J. A.: Separating root and soil microbial contributions to soil respiration: A review of methods and observations, Biogeochemistry, 48, 115–146, https://doi.org/10.1023/A:1006244819642, 2000.

Hashmi, M. Z. and Shamseldin, A. Y.: Use of Gene Expression Programming in regionalization of flow duration curve, Adv. Water Resour., 68, 1–12, https://doi.org/10.1016/j.advwatres.2014.02.009, 2014.

Hearst, M. A.: Support vector machines, IEEE Intell. Syst. Appl., 13, 18–28, https://doi.org/10.1109/5254.708428, 1998.

Heimann, M. and Reichstein, M.: Terrestrial ecosystem carbon dynamics and climate feedbacks, Nature, 451, 289–292, https://doi.org/10.1038/nature06591, 2008.

Heinemeyer, A., Di Bene, C., Lloyd, A. R., Tortorella, D., Baxter, R., Huntley, B., Gelsomino, A., and Ineson, P.: Soil respiration: Implications of the plant–soil continuum and respiration chamber collar-insertion depth on measurement and modelling of soil CO_2 efflux rates in three ecosystems, Eur. J. Soil Sci., 62, 82–94, https://doi.org/10.1111/j.1365-2389.2010.01331.x, 2011.

Heinemeyer, A., Wilkinson, M., Vargas, R., Subke, J. A., Casella, E., Morison, J. I. L., and Ineson, P.: Exploring the overflow tap theory: Linking forest soil CO_2 fluxes and individual mycorrhizosphere components to photosynthesis, Biogeosciences, 9, 79–95, https://doi.org/10.5194/bg-9-79-2012, 2012.

Hoerl, A. E. and Kennard, R. W.: Ridge Regression: Biased Estimation for Nonorthogonal Problems, Technometrics, 12, 55–67, https://doi.org/10.1080/00401706.1970.10488634, 1970.

Hoffmann, M., Jurisch, N., Albiac Borraz, E., Hagemann, U., Drösler, M., Sommer, M., and Augustin, J.: Automated modeling of ecosystem CO_2 fluxes based on periodic closed chamber measurements: A standardized conceptual and practical approach, Agr. Forest Meteorol., 200, 30–45, https://doi.org/10.1016/j.agrformet.2014.09.005, 2015.

Hölttä, T., Mencuccini, M., and Nikinmaa, E.: A carbon cost-gain model explains the observed patterns of xylem safety and efficiency, Plant Cell Environ., 34, 1819–1834, https://doi.org/10.1111/j.1365-3040.2011.02377.x, 2011.

Ilie, I., Mahecha, M. D., Jung, M., Carvalhais, N., and Dittrich, P.: Evolving compact symbolic expressions by a GEP CMA-ES hybrid approach, Genet. Program. Evolvab. Mach., in preparation, 2017.

Jakeman, A. J., Letcher, R. A., and Norton, J. P.: Ten iterative steps in development and evaluation of environmental models, Environ. Model. Softw., 21, 602–614, https://doi.org/10.1016/j.envsoft.2006.01.004, 2006.

Kabanikhin, S. I.: Definitions and examples of inverse and ill-posed problems, J. Inverse Ill-Posed Probl., 16, 317–357, https://doi.org/10.1515/JIIP.2008.019, 2008.

Keene, O. N.: The log transformation is special, Stat. Med., 14, 811–819, https://doi.org/10.1002/sim.4780140810, 1995.

Khatibi, R., Naghipour, L., Ghorbani, M. A., Smith, M. S., Karimi, V., Farhoudi, R., Delafrouz, H., and Arvanaghi, H.: Developing a predictive tropospheric ozone model for Tabriz, Atmos. Environ., 68, 286–294, https://doi.org/10.1016/j.atmosenv.2012.11.020, 2013.

Kotanchek, M. E., Vladislavleva, E., and Smits, G.: Symbolic Regression Is Not Enough: It Takes a Village to Raise a Model, in: Genetic Programming Theory and Practice X, Springer Science + Business Media, New York, 187–203, https://doi.org/10.1007/978-1-4614-6846-2, 2013.

Kowalski, A. M., Martín, M. T., Plastino, A., Rosso, O. A., and Casas, M.: Distances in Probability Space and the Statistical Complexity Setup, Entropy, 13, 1055–1075, https://doi.org/10.3390/e13061055, 2011.

Kuzyakov, Y.: Sources of CO_2 efflux from soil and review of partitioning methods, Soil Biol. Biochem., 38, 425–448, https://doi.org/10.1016/j.soilbio.2005.08.020, 2006.

Lasslop, G., Reichstein, M., Kattge, J., and Papale, D.: Influences of observation errors in eddy flux data on inverse model parameter estimation, Biogeosciences, 5, 1311–1324, https://doi.org/10.5194/bg-5-1311-2008, 2008.

Lasslop, G., Migliavacca, M., Bohrer, G., Reichstein, M., Bahn, M., Ibrom, A., Jacobs, C., Kolari, P., Papale, D., Vesala, T., Wohlfahrt, G., and Cescatti, A.: On the choice of the driving temperature for eddy-covariance carbon dioxide flux partitioning, Biogeosciences, 9, 5243–5259, https://doi.org/10.5194/bg-9-5243-2012, 2012.

Lavoie, M., Phillips, C. L., and Risk, D.: A practical approach for uncertainty quantification of high-frequency soil respiration using Forced Diffusion chambers, J. Geophys. Res.-Biogeo., 120, 128–146, https://doi.org/10.1002/2014JG002773, 2015.

Lazaro-Gredilla, M., Titsias, M. K., Verrelst, J., and Camps-Valls, G.: Retrieval of Biophysical Parameters With Heteroscedastic Gaussian Processes, IEEE Geosci. Remote Sens. Lett., 11, 838–842, https://doi.org/10.1109/LGRS.2013.2279695, 2014.

Lloyd, J. and Taylor, J. A.: On the temperature dependence of soil respiration, Funct. Ecol., 8, 315–323, 1994.

Luo, Y., Keenan, T. F., and Smith, M. J.: Predictability of the terrestrial carbon cycle, Global Change Biol., 21, 1737–1751, https://doi.org/10.1111/gcb.12766, 2015.

Mahecha, M. D., Reichstein, M., Carvalhais, N., Lasslop, G., Lange, H., Seneviratne, S. I., Vargas, R., Ammann, C., Arain, M. A., Cescatti, A., Janssens, I. A., Migliavacca, M., Montagnani, L., and Richardson, A. D.: Global convergence in the temperature sensitivity of respiration at ecosystem level, Science, 329, 838–840, https://doi.org/10.1126/science.1189587, 2010.

Manning, W. G.: The Logged dependent variable, heteroskedasticity, and the retransformation problem, J. Health Econ., 17, 283–295, https://doi.org/10.1016/S0167-6296(98)00025-3, 1998.

Migliavacca, M., Reichstein, M., Richardson, A. D., Colombo, R., Sutton, M. A., Lasslop, G., Tomelleri, E., Wohlfahrt, G., Carvalhais, N., Cescatti, A., Mahecha, M. D., Montagnani, L., Papale, D., Zaehle, S., Arain, A., Arneth, A., Black, T. A., Carrara, A., Dore, S., Gianelle, D., Helfter, C., Hollinger, D., Kutsch, W. L., Lafleur, P. M., Nouvellon, Y., Rebmann, C., Humberto, R., Rodeghiero, M., Roupsard, O., Sebastià, M. T., Seufert, G., Soussana, J. F., and Michiel, K.: Semiempirical modeling of abiotic and biotic factors controlling ecosystem respiration

across eddy covariance sites, Global Change Biol., 17, 390–409, https://doi.org/10.1111/j.1365-2486.2010.02243.x, 2011.

Migliavacca, M., Sonnentag, O., Keenan, T. F., Cescatti, A., O'Keefe, J., and Richardson, A. D.: On the uncertainty of phenological responses to climate change, and implications for a terrestrial biosphere model, Biogeosciences, 9, 2063–2083, https://doi.org/10.5194/bg-9-2063-2012, 2012.

Migliavacca, M., Reichstein, M., Richardson, A. D., Mahecha, M. D., Cremonese, E., Delpierre, N., Galvagno, M., Law, B. E., Wohlfahrt, G., Andrew Black, T., Carvalhais, N., Ceccherini, G., Chen, J., Gobron, N., Koffi, E., William Munger, J., Perez-Priego, O., Robustelli, M., Tomelleri, E., and Cescatti, A.: Influence of physiological phenology on the seasonal pattern of ecosystem respiration in deciduous forests, Global Change Biol., 21, 363–376, https://doi.org/10.1111/gcb.12671, 2015.

Mitchell, S., Beven, K., and Freer, J.: Multiple sources of predictive uncertainty in modeled estimates of net ecosystem CO_2 exchange, Ecol. Model/, 220, 3259–3270, https://doi.org/10.1016/j.ecolmodel.2009.08.021, 2009.

Moncrieff, J., Massheder, J., de Bruin, H., Elbers, J., Friborg, T., Heusinkveld, B., Kabat, P., Scott, S., Soegaard, H., and Verhoef, A.: A system to measure surface fluxes of momentum, sensible heat, water vapour and carbon dioxide, J. Hydrol., 188-189, 589–611, https://doi.org/10.1016/S0022-1694(96)03194-0, 1997.

Moyano, F. E., Kutsch, W. L., and Rebmann, C.: Soil respiration fluxes in relation to photosynthetic activity in broad-leaf and needle-leaf forest stands, Agr. Forest Meteorol., 148, 135–143, https://doi.org/10.1016/j.agrformet.2007.09.006, 2008.

Nash, J. and Sutcliffe, J.: River flow forecasting through conceptual models part I – A discussion of principles, J. Hydrol., 10, 282–290, https://doi.org/10.1016/0022-1694(70)90255-6, 1970.

Newman, M. C.: Regression Analysis of Log-Transformed Data – Statistical Bias and Its Correction (Short Communication), Environ. Toxicol. Chem., 12, 1129–1133, https://doi.org/10.1002/etc.5620120618, 1993.

Peng, S., Ciais, P., Chevallier, F., Peylin, P., Cadule, P., Sitch, S., Piao, S., Ahlström, A., Huntingford, C., Levy, P., Li, X., Liu, Y., Lomas, M., Poulter, B., Viovy, N., Wang, T., Wang, X., Zaehle, S., Zeng, N., Zhao, F., and Zhao, H.: Benchmarking the seasonal cycle of CO_2 fluxes simulated by terrestrial ecosystem models, Global Biogeochem. Cy., 29, 46–64, https://doi.org/10.1002/2014GB004931, 2014.

Peng, Y., Yuan, C., Qin, X., Huang, J., and Shi, Y.: An improved Gene Expression Programming approach for symbolic regression problems, Neurocomputing, 137, 293–301, https://doi.org/10.1016/j.neucom.2013.05.062, 2014.

Pérez-Priego, O., López-Ballesteros, A., Sánchez-Cañete, E. P., Serrano-Ortiz, P., Kutzbach, L., Domingo, F., Eugster, W., Kowalski, A. S., Sánchez-Cañete, E. P., Serrano-Ortiz, P., Kowalski, A. S., López-Ballesteros, A., Domingo, F., Kutzbach, L., Eugster, W., and Pérez-Priego, O.: Analysing uncertainties in the calculation of fluxes using whole-plant chambers: random and systematic errors, Plant Soil, 393, 229–244, https://doi.org/10.1007/s11104-015-2481-x, 2015.

Reichstein, M. and Beer, C.: Soil respiration across scales: The importance of a model-data integration framework for data interpretation, J. Plant Nutr. Soil Sci., 171, 344–354, https://doi.org/10.1002/jpln.200700075, 2008.

Reichstein, M., Falge, E., Baldocchi, D., Papale, D., Aubinet, M., Berbigier, P., Bernhofer, C., Buchmann, N., Gilmanov, T., Granier, A., Grünwald, T., Havránková, K., Ilvesniemi, H., Janous, D., Knohl, A., Laurila, T., Lohila, A., Loustau, D., Matteucci, G., Meyers, T., Miglietta, F., Ourcival, J. M., Pumpanen, J., Rambal, S., Rotenberg, E., Sanz, M., Tenhunen, J., Seufert, G., Vaccari, F., Vesala, T., Yakir, D., and Valentini, R.: On the separation of net ecosystem exchange into assimilation and ecosystem respiration: Review and improved algorithm, Global Change Biol., 11, 1424–1439, https://doi.org/10.1111/j.1365-2486.2005.001002.x, 2005.

Richardson, A. D., Mahecha, M. D., Falge, E., Kattge, J., Moffat, A. M., Papale, D., Reichstein, M., Stauch, V. J., Braswell, B. H., Churkina, G., Kruijt, B., and Hollinger, D. Y.: Statistical properties of random CO_2 flux measurement uncertainty inferred from model residuals, Agr. Forest Meteorol., 148, 38–50, https://doi.org/10.1016/j.agrformet.2007.09.001, 2008.

Rosso, O. A., Larrondo, H. A., Martin, M. T., Plastino, A., and Fuentes, M. A.: Distinguishing Noise from Chaos, Phys. Rev. Lett., 99, 154102, https://doi.org/10.1103/PhysRevLett.99.154102, 2007.

Ryan, M. G. and Law, B. E.: Interpreting, measuring, and modeling soil respiration, Biogeochemistry, 73, 3–27, https://doi.org/10.1007/s10533-004-5167-7, 2005.

Shannon, C. E.: A Mathematical Theory of Communication, Bell Syst. Tech. J., 27, 379–423, 1948.

Shi, Z., Wang, F., and Liu, Y.: Response of soil respiration under different mycorrhizal strategies to precipitation and temperature, J. Soil Sci. Plant Nutr., 12, 411–420, https://doi.org/10.4067/S0718-95162013005000053, 2012.

Sippel, S., Lange, H., Mahecha, M., Hauhs, M., Gans, F., Bodesheim, P., and Rosso, O.: Diagnosing the dynamics of observed and simulated ecosystem gross primary productivity with time causal information theory quantifiers, PLoS ONE, 11, e0164960, https://doi.org/10.1371/journal.pone.0164960, 2016.

Subke, J.-A., Inglima, I., and Francesca Cotrufo, M.: Trends and methodological impacts in soil CO_2 efflux partitioning: A metaanalytical review, Global Change Biol., 12, 921–943, https://doi.org/10.1111/j.1365-2486.2006.01117.x, 2006.

Tramontana, G., Jung, M., Schwalm, C. R., Ichii, K., Camps-Valls, G., Ráduly, B., Reichstein, M., Arain, M. A., Cescatti, A., Kiely, G., Merbold, L., Serrano-Ortiz, P., Sickert, S., Wolf, S., and Papale, D.: Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms, Biogeosciences, 13, 4291–4313, https://doi.org/10.5194/bg-13-4291-2016, 2016.

Traore, S. and Guven, A.: New algebraic formulations of evapotranspiration extracted from gene-expression programming in the tropical seasonally dry regions of West Africa, Irrig. Sci., 31, 1–10, https://doi.org/10.1007/s00271-011-0288-y, 2013.

Trumbore, S.: Carbon respired by terrestrial ecosystems – recent progress and challenges, Global Change Biol., 2, 141–153, https://doi.org/10.1111/j.1365-2486.2006.01067.x, 2006.

Wehr, R., Munger, J. W., McManus, J. B., Nelson, D. D., Zahniser, M. S., Davidson, E. A., Wofsy, S. C., and Saleska, S. R.: Seasonality of temperate forest photosynthesis and daytime respiration, Nature, 534, 680–683, https://doi.org/10.1038/nature17966, 2016.

Wilkinson, M., Eaton, E. L., Broadmeadow, M. S. J., and Morison, J. I. L.: Inter-annual variation of carbon uptake by a plantation oak woodland in south-eastern England, Biogeosciences, 9, 5373–5389, https://doi.org/10.5194/bg-9-5373-2012, 2012.

Williams, M., Richardson, A. D., Reichstein, M., Stoy, P. C., Peylin, P., Verbeeck, H., Carvalhais, N., Jung, M., Hollinger, D. Y., Kattge, J., Leuning, R., Luo, Y., Tomelleri, E., Trudinger, C. M., and Wang, Y.-P.: Improving land surface models with FLUXNET data, Biogeosciences, 6, 1341–1359, https://doi.org/10.5194/bg-6-1341-2009, 2009.

Yegnanarayana, B.: Artificial neural networks, Prentice-Hall of India Pvt. Ltd, New Delhi, 2006.

Zanin, M., Zunino, L., Rosso, O. A., and Papo, D.: Permutation Entropy and Its Main Biomedical and Econophysics Applications: A Review, Entropy, 14, 1553–1577, https://doi.org/10.3390/e14081553, 2012.