CMAGEP: A new method for automatic model discovery from data and its application to terrestrial ecosystem carbon exchange fluxes

seit 1558

Dissertation zur Erlangung des akademischen Grades
doctor rerum naturalium (Dr. rer. nat.)

vorgelegt dem Rat der Fakultät für Mathematik und Informatik der Friedrich-Schiller-Universität Jena

von Iulia Ilie
geboren am 19.02.1988 in Mangalia, Rumänien

Gutachter:

1. **Prof. Dr. habil. Peter Dittrich (Friedrich-Schiller-Universität Jena)**
2. **Dr. sc. Miguel D. Mahecha (Max Planck Insitute for Biogeochemistry, Jena)**
3. **Prof. Dr.-Ing. habil. Clemens Beckstein (Friedrich-Schiller-Universität Jena)**

Tag der öffentlichen Verteidigung: 06.09.2019

Abstract

Accurately representing and understanding the dynamics driving the global carbon cycle are of strong significance for the study of the Earth System as well as for reliable climate change projections. Model development in the biogeochemistry field traditionally relies on empirical studies and on already established theoretical foundations.

With increased data availability, model development in the field of biogeochemistry has started to open more to the use of machine learning approaches for helping to validate and calibrate the existing model formulations. However, the validity of the studied model structures are not often debated.

This thesis introduces a novel framework for modelling biogeochemistry fluxes by using symbolic regression approaches to automatically generate interpretable mathematical models.

The thesis starts by first illustrating the potential of gene expression programming (GEP) to discover interesting models as mathematical formulas based entirely on real time series data measured at a single monitoring site. The GEP discovered models perform better predictions than already established models in the ecology community. Further, the GEP models have the advantage of being represented as mathematical formulas that can be used similarly to natural laws from the ecology community. Still, the complexity of GEP models makes it difficult to really interpret the described model dynamics.

To tackle model complexity GEP is extended with CMA-ES for performing local parameter optimizations in the evolution process. The resulting algorithm is CMAGEP, a novel system that is a GEP and ES hybrid approach capable of delivering more accurate and more compact solutions compared to standard GEP. Generating compact solutions means that CMAGEP discovers mathematical models that can be more easily interpretable, and that can be more easily combined with already established knowledge.

CMAGEP is successfully used for modelling various carbon fluxes; first it helps discover non-linear dynamics in the carbon cycle at an Arctic site and produce a very compact solution, and secondly, it reveals interesting

and relevant patterns in the underlying processes determining the global terrestrial carbon exchanges.

Considering the important results shown in this extensive interdisciplinary study it becomes clear that by introducing the new CMAGEP system, an important contribution was made to the field of symbolic regression by giving deserved attention to the often neglected aspect of interpretability. Furthermore, the application of CMAGEP in a symbolic regression framework to model terrestrial carbon fluxes helped build novel knowledge in the ecology field, giving this approach a significant potential for other future applications.

Zusammenfassung

Die Dynamik, die den globalen Kohlenstoffkreislauf antreibt, genau darzustellen und zu verstehen, ist von großer Bedeutung für das Studium des Erdsystems und für zuverlässige Prognosen zum Klimawandel. Die Modellentwicklung in der Biogeochemie beruht traditionell auf empirischen Studien und auf bereits etablierten theoretischen Grundlagen.

Mit zunehmender Datenverfügbarkeit hat die Modellentwicklung auf dem Gebiet der Biogeochemie begonnen, sich mehr für den Einsatz von Methoden des maschinellen Lernens zu öffnen, um die bestehenden Modellformulierungen zu validieren und zu kalibrieren. Die Validität der untersuchten Modellstrukturen wird jedoch nicht oft diskutiert.

Diese Arbeit stellt einen neuartigen Rahmen für die Modellierung von biogeochemischen Flüssen vor, indem mithilfe von symbolischen Regressionsansätzen interpretierbare mathematische Modelle automatisch generiert werden.

Die Arbeit beginnt damit, zunächst das Potenzial der Gene Expression Programming (GEP) aufzuzeigen, um interessante Modelle als mathematische Formeln automatisch aus Echtzeit-Zeitseriendaten abzuleiten, die an nur einem Ort gemessen worden sind. Das GEP hat dabei Modelle generiert, die eine bessere Performanz als bereits etablierte Modelle der Ökologie-Community aufweisen. Ferner haben die erzeugten Modelle den Vorteil, dass sie als mathematische Formeln repräsentiert werden, die den Formeln der Ökologie-Community ähnlich sind. Allerdings macht die Komplexität der GEP-Modelle es schwierig, die beschriebene Modelldynamik zu interpretieren.

Im nächsten Schritt der Arbeit wurde GEP um eine lokale Parameteroptimierung mittels der CMA-ES erweitert. Das resultierende CMAGEP-System ist ein GEP- und ES-Hybridansatz, der Lösungen liefert, die im Vergleich zu Standard GEP Kohlenstoffflüsse sowohl genauer als auch auch kompakter beschreiben. Die Generierung von kompakten Lösungen bedeutet, dass mathematische Modelle entdeckt werden, die leichter interpretiert werden können und die sich einfacher mit bereits etabliertem Wissen kombinieren lassen.

Im Anschluss wird CMAGEP erfolgreich zur Modellierung von unterschiedlichen Kohlenstoffflüssen verwendet; Erstens hilft es, nichtlineare Dynamiken im Kohlenstoffkreislauf an einem arktischen Standort zu entdecken und eine sehr kompakte Lösung zu erzeugen, und zweitens offenbart es interessante und relevante Muster in den zugrunde liegenden Prozessen, die den globalen terrestrischen Kohlenstoffaustausch bestimmen.

Betrachtet man die wichtigen Ergebnisse dieser umfangreichen interdisziplinären Studie, so wird deutlich, dass mit der Einführung des neuen CMAGEP Systems ein wichtiger Beitrag zum Bereich der symbolischen Regression mit dem oft vernachlässigten aber bedeutsamen Aspekt der Interpretierbarkeit geleistet wurde. Darüber hinaus trug die Anwendung von CMAGEP zur Modellierung terrestrischer Kohlenstoffflüsse dazu bei, neues Wissen auf dem Gebiet der Ökologie aufzubauen, was diesem Ansatz ein signifikantes Potenzial für andere zukünftige Anwendungen verleiht.

Acknowledgements

I would like to thank my research advisers, my colleagues and all the individuals that helped me with the work presented in this thesis.

I would like to thank my PhD advisor, Miguel Mahecha, for offering me the chance to work in such an interesting interdisciplinary project, and for having patience with a novice in the Biogeochemistry field like I was to catch up on very unfamiliar concepts. Miguel has shown me how to do research at the highest level and has always been an inspiration for all research life aspects.

I especially thank Peter Dittrich for simply stopping by my IMPRS Conference Poster one day and then becoming one of my advisers in more ways than one, and without whom I would have not finished this work.

I give many thanks to Mirco Migliavacca for all the inspiring discussions and advices, as well as to all scientists at the MPI for Biogeochemistry who have helped with the work of this thesis.

Many thanks also given to all my Jena friends and colleagues and especially to Min Jung, Talie, Sven, Fabio and Chirag, who have made my days and evenings fuller and sunnier.

I would especially like to thank my strong mother who has provided inspiration in all aspects of my life and without whom doing research would not have even crossed my mind.

I thank Gaurav for being so supportive during all the aspects of my PhD Student life.

Lastly, but surely not least, I thank Aghi, my sweet puppy who has only made my stay in Jena happy.

Contents

"It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts."–
Arthur Conan Doyle (The Adventures of Sherlock Holmes–1892)

Introduction

1.1 Modelling Biogeochemical Cycles

In Biogeochemistry, some of the more important questions refer to untangling, understanding and accurately representing the multitude of processes involved in the Earth System dynamics (Rounsevell et al., 2014). Answering these questions can partly be supported by building mathematical descriptions designed to capture and simulate natural laws (Kumar et al., 2006). The process of building the necessary mathematical descriptions is known as model development(Šimůnek and Suarez, 1993).

Due to an increase in biogeochemistry observations availability Baldocchi (2008); McCain et al. (2006), model development can be aided and accelerated by data-driven learning and inference, especially in the case of dynamics that are too complex to easily describe and capture using empirical methods and currently known mechanics.

Model development in the field of biogeochemistry, currently mainly consists of validating, calibrating and updating existing model structures and established foundations with real observations (Luo et al., 2015).

In this context, although some dynamics can be well represented by established models or parametric methods are sufficient for accurate modelling, it is possible that for a more comprehensive description of certain responses, especially when the conditions in which they appear are not easily reproducible, previously not considered, possibly non-linear laws might be better fitting Bongard and Lipson (2005).

In order to complement the existing biogeochemistry modelling approaches and to explore the possibility of discovering relevant new knowledge from data only, a reverse engineering framework for model building is proposed (see Chapter 2). In the reverse engineering framework, specifically automated model learning where little or no prior constraints are imposed on the desired model structures, novel non-linear responses can emerge and allow for further inference.

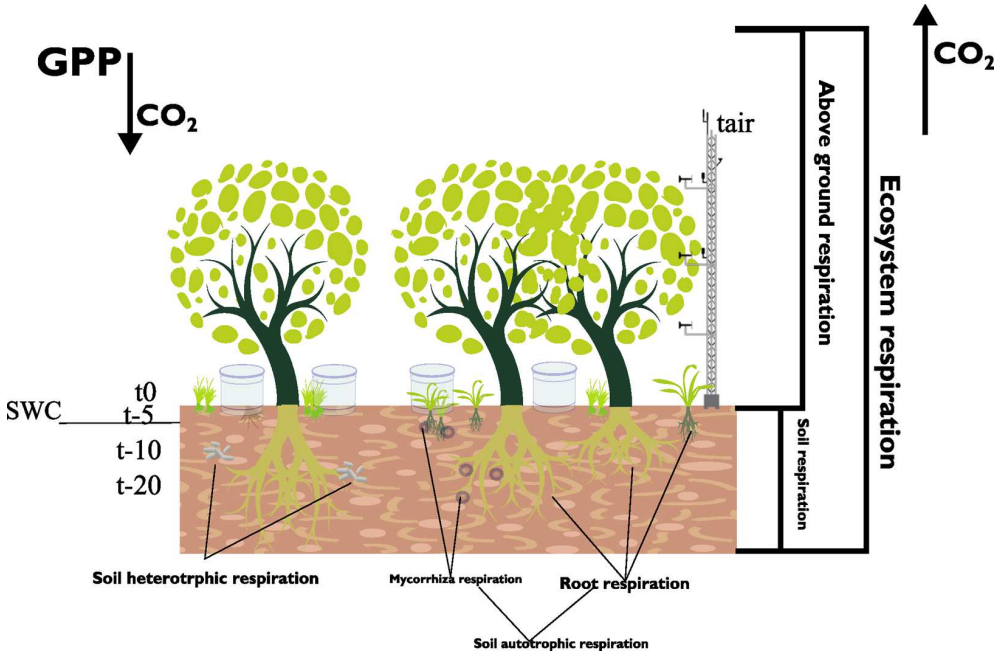


Figure 1.1: **A monitoring site for terrestrial ecosystem biogeochemical cycles.**

One focus of this thesis is studying automated data-driven model development and its to terrestrial ecosystem carbon fluxes, such as terrestrial ecosystem respiration (R_{eco}). The studied terrestrial ecosystem carbon exchanges were continuously monitored by measuring a set of various biotic and abiotic features over a global network Baldocchi (2003). A monitoring site example is illustrated in Fig. 1.1.

1.2 Genetic Programming for Symbolic Regression

All studies of this thesis are based on data structured in sets of real numerical values recorded for an array of features at different time instances (Fig. 1.2). The structure allows for automated discovery of relations between a (sub)set of independent variables to one target variable, otherwise known as candidate drivers and response Vladislavleva et al. (2009).

The relations can be described by Eq. 1.2.1, with $Y = (y_1, y_2, ...y_n)$, a set of responses over n time steps and $X = (x_{i1}, x_{i2}...x_{in})$, $i = 1...m$ a set of m drivers over n time steps.

$$Y(t) = \hat{F}(X(t)) \tag{1.2.1}$$

The modelling problem to solve then is finding $\hat{F}$, the most appropriate mathemat-

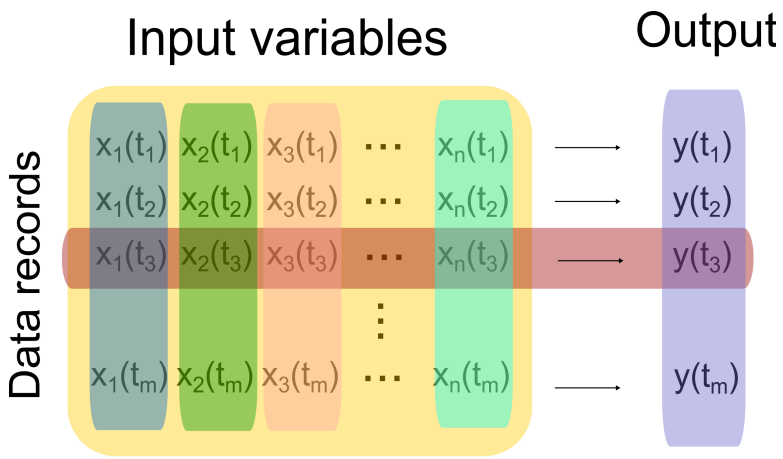


Figure 1.2: **Data organization in automated model development.**

ical formula to describe the response-drivers relation based on possible signals present in available data.

In classic regression, $\hat{F}$ is already assumed to be the weighted sum of the feature set, and the task left is to optimize as accurately as possible the set of weights associated to X to best fit Y. In symbolic regression (SR) there are no assumptions made on the shape and size of $\hat{F}$ with the formula also considered unknown, making the task to solve the discovery of the $\hat{F}$ function formula first, with implicit parameter optimization.

The task of automatically searching for the most appropriate $\hat{F}$ and building its components can be solved by a Genetic Programming (GP) system (Koza, 1994; Langdon and Poli, 2002). GP is an evolutionary algorithm that automatically builds programs meant to solve problems based on prescribed high-level instructions. GP can evolve solutions without the need to specify the shape, structure or size of the desired solution Augusto and Barbosa (2000).

It does this by encoding solutions in expression trees that are called individuals. The encoding allows the evolution process to typically consist of: 1) an initial generation phase, where a specific number of individuals are (randomly) generated, continued by 2) evaluation and fitness based selection, followed by 3) genetic variation, with mutations and cross-over operations between the individuals and their respective expression trees, leading to 4) a novel generation that repeats the evolution steps until a predefined stop criterion is reached and a final solution (set of solutions) is returned.

Poli et al. (2008) propose that GP is most suitable to solve problems when one of more of the following conditions are met:

1. Relations between available features are not well known or understood;

2. The complexity and shape of the solution function are part of the desired solution;

3. An approximate solution is acceptable since GP systems build a solution only from the available data at learning

4. There are existing methods to rank solutions and to test how appropriate the GP returned solution is for the given problem

5. There are no unique analytical or experimental solutions that can completely cover the applicability domain of the given problem, or if there are, these are difficult to apply for specific conditions.

When considering the problem of modelling the response of R_{eco} to candidate drivers with a GP symbolic regression, many if not all conditions are met: the R_{eco} response to external drivers is one of the problems of that is not very well described by the community literature (Friend, 2010; Yvon-Durocher et al., 2012); a completely data driven model is desired for understanding if there are relevant signals to be harvested in measurements and not yet considered, with the final R_{eco} response model having no assumptions regarding structure and shape or size, although there are some physical soundness restrictions and can be applied in a final model selection stage; the goal is not to obtain a single global solution, but to explore the possibility to discover novel or previously not considered components in the response of R_{eco} to external driving factors, model structures can be compared to established models in the field with regards to prediction accuracy; and empirical experiments cannot fully cover the large range of possible responses of R_{eco} to external drivers.

By solving the symbolic regression problem in the GP reverse engineering framework, large part of the freedom of exploration lies in the fact that few or no assumptions are made in relation to the type of distribution present in the data. However, the vast freedom of search for a fitting model structure represents also the largest challenge, as the search space can easily become harder to cover with increasing data, leading to long waiting time for solutions (Poli et al., 2008).

The data-driven aspect of the GP symbolic regression problem for modelling R_{eco} makes approaching it a difficult task, as it will not completely certain if a model structure has emerged only due to the presence of significant signals in studied data or due to the stochastic component of GP. Thus it might be difficult to generalise knowledge gained the GP returned solution structures. It is difficult as well to state that a GP solution is a unique or best solution (Langdon and Poli, 2002), since different mathematical functions show similar behaviour for specific domain values, but might be very different outside of those values. Due to time constraints and different focus of this thesis the equifinality problem is addressed, however only briefly.

The GP variant used to automatically produce regression models in this thesis was the Gene Expression Programming (GEP) Ferreira (2001). GEP introduces to GP an intermediate encoding and translation phase, with strings encoding expression trees, and expression trees encoding mathematical solutions. The extra encoding step mimics genetic genotype-phenotype structure, allowing for more freedom in genetic variation and implicitly easier solution search space cover.

1.2.1 Improving local parameter optimizations: CMAGEP

The interpretation aspect of model development is often disregarded for higher accuracy in prediction capacity. In our work, the goal was not only that of building a mathematical model to use for future simulations, but that of understanding internal dynamic driving the response of terrestrial carbon fluxes. So, being able to easily read and interpret any emerging model from the GEP data-driven model development was necessary.

The GP associated bloat phenomenon Langdon (2000) appears often in the GEP evolutions as well. In order to counter the exponential expansion of solution sizes with generation count, and limit the lengths of the automatically built models, the standard GEP framework where no specific treatment is given to local parametrizations during the evolution process was combined with an evolutionary strategy approach for optimizing the parameters in the top best solutions in an evolution step.

Artificial and real data experiments showed significant decrease in solution complexity, improving the required aspect of interpretability.

The newly proposed algorithm is called CMAGEP, Covariance Matrix Adapted Gene Expression Programming, and is a hybrid of GEP and CMA-ES (Hansen, 2006a), an evolutionary strategy that was chosen due to its capacity to optimise parameter sets for problems without needed a specified function form.

1.3 Code availability

All code used to obtain the results presented in this doctoral thesis is freely available at `https://sourceforge.net/projects/cmagep/`.

1.4 Thesis questions

This doctoral thesis introduces and supports a novel framework for model development, specifically in the biogeochemistry community, where the observations will be the start for the inferences made on the structures of the responses of R_{eco} to external drivers.

Some of the main question sought out to answer with the work presented in this doctoral thesis were:

1. Is it possible to automatically build relevant model structures for Biogeochemical processes, based entirely on observations?

Yes. Chapter 2 shows that GEP generated models perform better than established models from the community and due to the possibility to look at the returned function, we were able to theorize on the inclusion of previously unconsidered terrestrial respiration responses to drivers.

2. Is a genetic programming type of approach a good solution to this problem?

   Yes. Our proposed algorithm shows encouraging results that surpass in prediction performance the currently established models in the biogeochemistry community.

3. How can the complexity of model structures obtained by a genetic programming approach be limited so that interpretability is an option? The bloat of GEP solutions is limited by including a local parameter set optimization for the best solutions of an evolution step by using CMA-ES.

4. Is it possible to quantify the generalisation capacity of a model structure obtained from Biogeochemical flux observations? Possibly. The possibility to re-optimize a model structure for local conditions was studied and the mean prediction performance of the re-optimized model was compared with that of other re-optimized models. Specific model structures seem to have better capacity to cover conditions form other not-trained-for sites. This could be an interesting framework and needs deeper study than the current time limited work.

5. Would such a general model structure be capable to at least capture in a good magnitude global fluxes?

   Yes. A single model is used in Chapter 5 to simulate R_{eco} fluxes based on a small set of features, going from discrete R_{eco} flux cover depending on the distribution of the measuring sites to a continuous global cover. The global yearly sum for the daily simulated flux was in an acceptable range to established estimations. The final model was:

$$R^{g}_{eco}(t) = 0.71 \exp\left(0.45 Sif_{ms}(t) + 0.04 T_{air}(t)\right) \tag{1.4.1}$$

1.5 Most significant contributions

1. I have implemented a C++ software package for standard GEP;

2. GEP was used to build relevant models from real R_{eco} flux data outperforming community established models in prediction accuracy

3. I developed and implemented a novel C++ and Python software package, CMAGEP, containing a hybrid approach between a GEP system and a CMA-ES system

4. CMAGEP generated symbolic regression outperforming standard GEP regressions with regards to prediction over established artificial and real data benchmark problems as well as over a real R_{eco} data case study. More importantly CMAGEP was shown to generate models with significantly fewer parameters, allowing for easier interpretation and novel knowledge discovery;

5. CMAGEP was used to construct much shorter model structures than those originally used for CO_2 fluxes for an Arctic measurement site, helping to better understand the main dynamics in the studied Arctic ecosystem;

6. CMAGEP was used to individually develop models for R_{eco} at 112 real measurement sites; Interesting patterns in model structures emerged and were analysed. The patterns emerging from the 112 obtained model structures allowed to select a single model structure for the global daily R_{eco} flux with individual parameter sets for each of the 112 studied sites. Furthermore, it was possible to select a single parameter set for the unique model structure leading to a single R_{eco} response to external factors model was used for reasonably simulating daily R_{eco} fluxes for all grids of the globe in a specific year.

1.6 Thesis structure

1. **Chapter 2** describes challenges in modelling terrestrial carbon fluxes and illustrates the potential of using only standard GEP in the automated modelling framework to discover new relevant laws for R_{eco} components responses to biotic and abiotic drivers.

   All experiments and results presented in this chapter have been published in: **Ilie, I.**, Dittrich, P., Carvalhais, N., Jung, M., Heinemeyer, A., Migliavacca, M., Morison, J.I.L., Sippel, S., Subke, J.-A., Wilkinson, and M. Mahecha : *Reverse engineering model structures for soil and ecosystem respiration: the potential of gene expression programming. Geoscientific Model Development, 10, 3519-3545, 2017, https://doi.org/10.5194/gmd-10-3519-2017.*

2. **Chapter 3** introduces CMAGEP, a novel hybrid approach of GEP and CMA-ES and illustrate the significant reducing of solutions length with CMAGEP with slight improvements in modelling accuracy over established artificial and real data benchmarks as well as over a real observed R_{eco} case study.

   All experiments and results presented in this chapter will be submitted for publication in: **Iulia Ilie**, Miguel D. Mahecha, Nuno Carvalhais, Martin Jung, Peter

Dittrich: *Evolving compact symbolic expressions by a GEP and CMA-ES hybrid approach. IEEE Transactions on Evolutionary Computation.*

3. **Chapter 4** shows the successful application of CMAGEP to a different domain than that of R_{eco}, but to that of Methane fluxes, with CMAGEP discovering non-linear laws to describe the response of CH_4 to external factors that were not previously considered in the community. The new CMAGEP proposed models contained a significantly lower set of parameters and reached a similar performance to that of very large linear models built with traditional regression methods.

   The CMAGEP experiments and results shown in this chapter were the author's contribution to the following published paper: Min Jung Kwon, Felix Beulig, **Iulia Ilie**, Marcus Wildner, Kirsten Küsel, Lutz Merbold, Miguel D. Mahecha, Nikita Zimov, Sergey A. Zimov, Martin Heimann, Edward A. G. Schuur, Joel E. Kostka, Olaf Kolle, Ines Hilke and Mathias Göckede: *Plants, microorganisms, and soil temperatures contribute to a decrease in methane fluxes on a drained Arctic floodplain. Global change biology 23 (6), 2396-2412, 2016, doi:10.1111/gcb.13558.*

4. **Chapter 5** explores the presence of patterns in the structures of 112 independently built CMAGEP for R_{eco} based on real measurements from a global monitoring network. We find that different climate types have specific responses of R_{eco} to its drivers, but that if a general global model structure is needed, mean best performing model can be selected and used for reasonable simulations during a specific year.

Reverse engineering model structures for soil and ecosystem respiration: the potential of GEP

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 with 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, the potential of gene expression programming (GEP) is explored with respect to deriving 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. The present 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, the responses of the different components of terrestrial respiration at an oak forest in south-east England were explored. The GEP retrieved models are often better in prediction than some established respiration models. Based on their structures, previously unconsidered exponential dependencies of respiration on seasonal ecosystem carbon assimilation and water dynamics. 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.

Highlights

- The current work explores if the process of model building for describing ecosystem CO_2 fluxes can be, to a large extent, automated.
- It is shown that Gene Expression Programming combined with parameter optimization can be a useful algorithm to automatically derive models from ecological time series.
- Alternative models are proposed for the influence of key environmental variables on various respiratory fluxes CO_2 in an oak forest.
- Conventional ecosystem response functions can be revised by new models identified with gene expression programming.

2.1 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, 2016; Heimann and Reichstein, 2008; Luo et al., 2015). However, the description of the mechanisms underlying the total terrestrial efflux of CO_2 (Peng et al., 2014a), 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 remaining manuscript the term "model" is used 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. Gilmanov et al., 2010; Hoffmann et al., 2015; Reichstein and Beer, 2008). Conventional model building (Fig. 2.1) 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 is subsequently guiding 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).

The current Chapter explores the possibility of reverse engineering offering an automated alternative to model development for predicting terrestrial carbon fluxes (Fig. 2.1). 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 for novel model structures to automatically emerge from the available data and that would not be so obvious from a direct analysis. Reverse engineering is aiming at identifying some mathematical representation of a system that is to a large degree independent from 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 even a 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

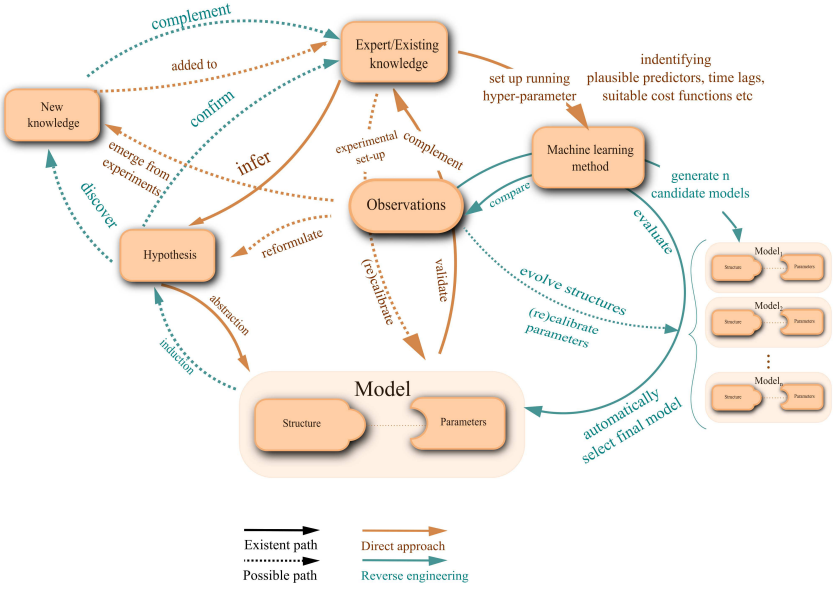


Figure 2.1: **Direct approach and reverse engineering in model development for describing dynamical systems.** Existing and possible steps needed in the process of building a model. For the direct approach, the process starts with the building of hypothesis from existing knowledge, the hypothesis is then subject of abstraction and 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 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 will be 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 into light.

whether the solution returned by the algorithm actually makes sense at all and whether it can be further used. Nevertheless, by shifting the moment when the analyst make 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 to consider the applicability of the derived structures in other system domains (Ashworth et al., 2012).

Here, the focus is 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 amount of evolutionary steps. Due to its structural design, GEP can be applied in a wide range of empirical modelling problems (Khatibi et al., 2013; Peng et al., 2014b; Traore and Guven, 2013), including (soil) hydrology (Fernando et al., 2009; Hashmi and Shamseldin, 2014). To the best of the author's knowledge the potential of GEP has not yet been explored for modelling biogeochemical fluxes in terrestrial ecosystems.

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

The fundamental question addressed in this chapter 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.

The current study investigates as well if automatically derived model structures differ substantially from models conventionally used in the study of R_{eco} and its components or, if 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, the current study investigates whether a generic "respiration" response can be derived, or if specific formulations for a range of respiration components are required.

2.1.1 Study structure

First, the GEP methodology is introduced and its performance is assessed for symbolic regression type of problems using an artificial experiment under varying degrees of noise contamination designed to resemble R_{eco}. Second, GEP is applied 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, the prediction error of GEP with other state-of-the-art machine learning regression approaches are systematically confronted. In addition, the modelling approach is adjusted such that the objective function (or fitness function) accounts not only 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.

2.2 Method

The current relies on the GEP method (Ferreira, 2001) which automatically constructs model structures based on a set of given observations. As the needed models are mathematical structures, their construction can be achieved by solving a symbolic regression (Kotanchek et al., 2013) type of problem. That is, the interest is not only in determining an optimal set of parameters for a known regression, but here, discovering the symbolic form of the regression itself is done 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.

2.2.1 GEP: Gene Expression Programming

The process of finding the most suitable model structure based on signal present in data in GEP starts with an initial generation of n possible model structures (Fig. 2.3, A). 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,

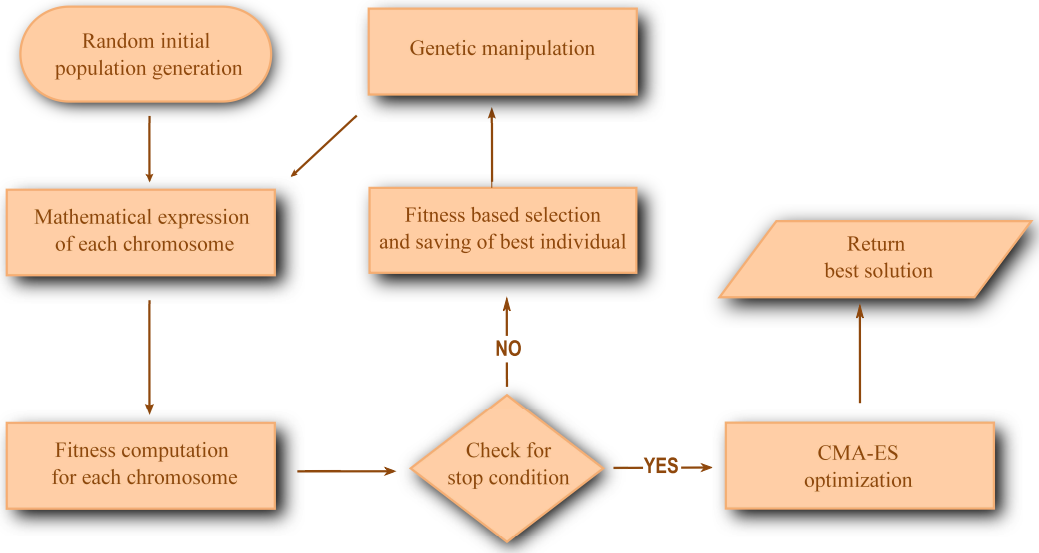


Figure 2.2: **The work flow used in solving symbolic regression problems with GEP.**. The process of evolving an optimal solution from observations starts with randomly generating a set number of evolution individuals called chromosomes. The chromosomes are composed of genes that are sets of strings 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 values is assigned. If the stopping criterion has not been reached (e.g.. best fitness possible, highest number of generations allowed, convergence etc.) 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.

e.g. $F = \{+, -, L, E\} \rightarrow F_m = \{$addition,substraction,logarithm,exponential$\}$, (see Fig. 2.3, A).

The choice of input functions used for applying mathematical transformations on the predictors depends on the type of problem solved 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, exponential and so on. More complex functions could increase model complexity too much and risk over fitting. However if there are already known functional transformations of certain predictors that could be part of the final desired solution, 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} as 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" (ORF) 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. Fig. 2.3, B). Ferreira (2001) argues that, the power of GEP lies in its use of fixed length linear strings for representing expression trees (ET) of varied shapes and sizes that simplifies the evolutionary process, and helps 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 developing 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 has 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. For filling 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, 2 chromosomes are randomly selected from the entire population and the individual with the better fitness value goes through.

For insuring 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. 2.3, D, that can fully change the encoded mathematical expressions (see Fig. 2.3, C).

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 is reached, time limit, etc.

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 looked to solve.

Such is the case for setting the length 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 over-fit the initial target. When the lengths of the chromosomes are kept too low, the structures in the population can convergence 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 that it 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 over-fitted results, the practical approach of multi-start (multiple runs with the same settings) is chosen as proposed by Ferreira (2006).

The version of the GEP method presented in this chapter 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, 1.2 PB file space. The large performance capacity of the cluster allowed for multiple parallel runs and speed in reaching the final solutions.

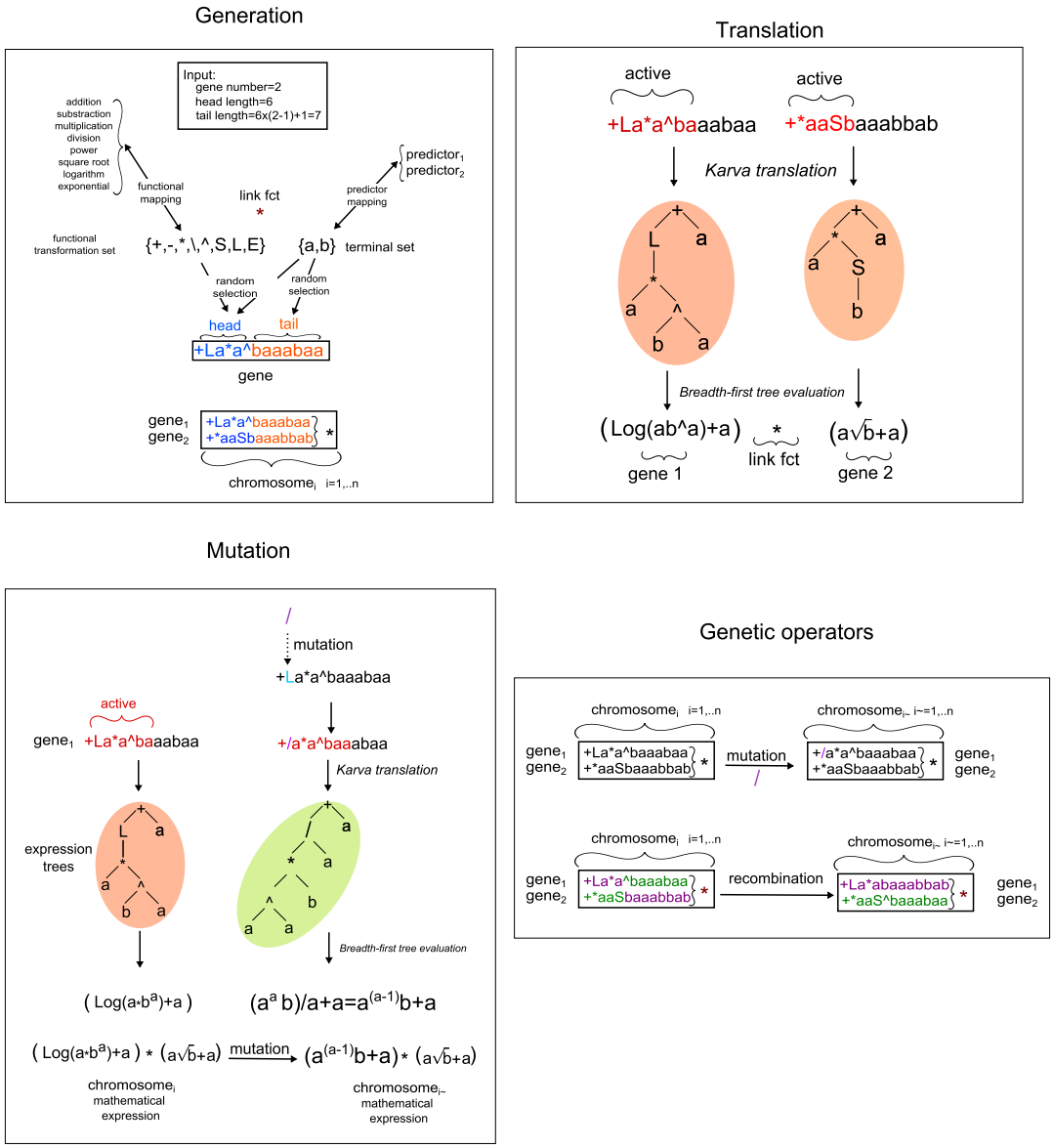


Figure 2.3: **GEP evolution process components.A.** Initial random generation of genes for creating chromosomes, the individuals evolved by GEP. **B.**GEP internal translation process from strings to expression trees and mathematical expressions.**C.** Changes made in the mathematical expression when applying the mutation operator on the genes of a GEP individual.**D.** Types of genetic operators for changing the GEP evolution individuals.

2.2.2 Fitness measure

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

$$\mathrm{MEF} = 1 - \frac{\sum_{i=1}^{n} (o_i - p_i)^2}{\sum_{i=1}^{n} (o_i - \bar{o})^2} \tag{2.2.1}$$

where o_i is the observed value at step i and p_i is the predicted value at step i and $\bar{o}$ 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. MEF=0 indicates that the models prediction are as good as a prediction by $\bar{o}$.

During the GEP learning process however, the (1-MEF) measure was used 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 higher number of free terms. Thus, the cost (fitness) function includes a normalized term related to number of parameters (ratio of current number of parameters to 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, the fitness function includes 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, C. E. Shannon and Shannon (1948)) , i.e. a term of the form

$$\mathrm{S}E(X) \;=\; -\sum_{i=1}^{N} p_i \ln [p_i] \;. \tag{2.2.2}$$

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 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 the 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 allows 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 $n_{demb} = 4$ throughout the entire manuscript 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 this work is:

$$\mathrm{CEM} = \sqrt{(1 - \mathrm{MEF})^2 + (\frac{P}{P_{max}})^2 + (1 - \mathrm{SE})^2} \tag{2.2.3}$$

$$P_{max} = n_g \times l \tag{2.2.4}$$

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.

For assessing the effect of adding the entropy component for the residuals in the CEM fitness function, a fitness measure containing elements regarding only the MEF

and the number of parameters was introduced as well .

$$\text{MEF+NP} = \sqrt{(1-\text{MEF})^2 + (\frac{P}{P_{max}})^2} \tag{2.2.5}$$

For all experiments reported in this chapter, 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.

2.2.3 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, the "Covariance Matrix Adaptation Evolution Strategy" (CMA-ES, Hansen et al. (2003)) is used 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 (Auger and Hansen, 2005; Beyer and Schwefel, 2002). According to Hansen (2006b), 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`.

2.3 Experimental design

For exploring 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, the use of GEP on real measurements of various respiratory flux components monitored continuously over two years in an oak forest (Heinemeyer et al., 2011) was studied.

2.3.1 Artificial experiments

These experiments were designed to explore whether the author's implementation of the GEP method is suitable for symbolic regression type of problems, and how robust/vulnerable it is across various signal to noise ratios. A set of functions with increasing levels of non-linearity were basis to generate data points.

$$f(x_1) = 2x_1 + 1 \tag{2.3.1}$$
$$f(x_1) = x_1^2 + 3x_1 + 5 \tag{2.3.2}$$
$$f(x_1) = e^{x_1} + 1 \tag{2.3.3}$$
$$f(x_1) = e^{-x_1} - x_1 \tag{2.3.4}$$
$$f(x_1) = x_1^2 - 4\sin(x_1) \tag{2.3.5}$$
$$f(x_1) = x_1^3 + 6x_1^2 + 11x_1 - 6 \tag{2.3.6}$$
$$f(x_1, x_2) = x_2 x_1 \tag{2.3.7}$$
$$f(x_1, x_2) = x_2 x_1 - 3\cos(x_1) \tag{2.3.8}$$
$$f(x_1, x_2) = 2x_1^2 + 3x_2^2 \tag{2.3.9}$$
$$f(x_1, x_2, x_3) = 2x_1^2 + 3x_2^2 + 2\sin(x_3) \tag{2.3.10}$$

2000 data points were randomly generated with $x_1 \in [1,20]; x_2 \in [1,5]; x_3 \in [1,100]$ 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 2.1. If a returned structure was identical to the originally prescribed function or if $(1 - \text{MEF}) \leq 10^{-5}$ at validation, the retrieval of the original structure was considered to be a success. For allowing the approaches to do an automatic feature selection, all 3 variables, x_1, x_2, x_3, were used for learning and validation for all 10 functions in the benchmark set.

For investigating the capacity of GEP to reconstruct a simple model used in the ecology field as well, an artificial test was done 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^oC at a reference temperature of 15^oC The formulation used for the "Q_{10}" model

is:

$$R_{eco} = 2^{(0.1T_{air}-1.5)} \tag{2.3.11}$$

with R_{eco} as ecosystem respiration flux and T_{air}, the air temperature. Again, 2000 data points were generated for both predictor and target and using 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, Gaussian noise scaling with signal amplitude was introduced as often as 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, 100 validation data points 10 times were sampled. 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, one experiment investigated the effects of minimizing the CEM values (eq. 2.2.3) as opposed to using only MEF (eq. 3.3.11) or MEF+NP (eq. 2.2.5) as fitness function.

2.3.1.1 Alternative Machine Learning Methods

The prediction performance of the best GEP derived models based on the data in section 2.3.1 was compared with the prediction performance of four commonly used state-of-the-art machine learning methods (MLM), i.e Artificial Neural Networks, ANN, (Yegnanarayana, 2009), Support vector Machines, SVM (Hearst, 1998), Random Forests, RF (Breiman, 2001) and Kernel Ridge Regressions, KRR (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 MLM were estimated to avoid over-fitting during each run as presented in section S6 of Tramontana et al. (2016).

All the present machine learning approaches have been applied on 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.

2.3.2 Measured ecosystem CO_2 fluxes

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

2.3.2.1 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. 2.3.12) 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 two 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} (Eq. 2.3.14 and 2.3.15) .

The above ground respiration (R_{above}) was given as well and was estimated by difference (Eq. 2.3.13). Additionally, direct measurements of soil moisture (*SWC*), air temperature, surface temperature, and soil temperature taken at 2, 10 and 20 cm depth are present in the dataset.

$$R_{eco} = \text{NEE} + GPP \tag{2.3.12}$$
$$R_{above} = R_{eco} - R_{soil} \tag{2.3.13}$$
$$R_{soil_a} = R_{root} + R_{myc} \tag{2.3.14}$$
$$R_{soil} = R_{soil_a} + R_{soil_h} \tag{2.3.15}$$

The computation of R_{above} as 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), Heinemeyer et al. (2012) and Wilkinson et al. (2012).

2.3.2.2 Data processing

The following candidate driver variables were used: soil volumetric moisture measurements, air temperature (from micro-meteorological station), and 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. (Mahecha et al., 2010; Migliavacca et al., 2011; Moyano et al., 2008, amongst others). *GPP* obtained from EC measurements at the site was used, 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, low-frequency variability of *GPP* was considered only (i.e. low-pass filtered modes of *GPP* which corresponds to variability beyond a 60 days 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), the seven target respira-

tion data sets were log-transformed (see Figure 2.16 in supplemental material) and a back-transformation was applied 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 $logy = \alpha x + \varepsilon$, 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:

$y = e^{(\alpha x + 0.5\sigma_{\varepsilon}^2)}$, when the residuals are log normal distributed;

$y = e^{(\alpha x)}E(e^{\varepsilon})$, 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. 2.17 of suppl.;

$y = e^{(\alpha x)}$ if no bias correction is desired, or a naive approach.

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

2.3.2.3 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 2.1. 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.

It was particularly interesting to determine the general character of each extracted model with respect to the different respiration fractions. Therefore the parameters were re-optimized for all extracted model structures by 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, performances were computed for each train–validation data set pair and averaged MEF values are reported.

As done in the artificial example, the returned GEP solutions predictions performances were compared with those of other common MLM such as SVN, KRR, ANNs, and RF. All methods were used for generating 50 subsets of 113 prediction values, after training on the 50 subsets of 500 time steps of observations presented in the start of section 2.3.2.3. Then, a mean MEF value was computed for all methods for all respiration components and the best mean MEF values were reported and compared with those of the GEP extracted models. The comparison is done in terms of MEF as number of model parameters were not available and CEM could not be computed.

Table 2.1: 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 seconds	1800 seconds
Fitness function	CEM	CEM
Selection method for replication	tournament(Coello and Montes, 2002)	tournament
Mutation probability	0.2	0.2
IS and RIS transpositions 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

Table 2.2: 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}	$\phi_1 \times \phi_2^{(\frac{T-T_{ref}}{10})}$	(Reichstein and Beer, 2008)
Water Q_{10}	$\phi_1 \times \phi_2^{(\frac{T-T_{ref}}{10})} \times \frac{SWC}{SWC+\phi_3} \times \frac{\phi_4}{SWC+\phi_4}$	(Richardson et al., 2008)
LinGPP	$(R_0 + k_2 GPP) \times e^{E_0(\frac{1}{T_{ref}-T_0} - \frac{1}{T_A-T_0})} \times \frac{\alpha k + SWC(1-\alpha)}{k+SWC(1-\alpha)}$	(Migliavacca et al., 2011)
ExpGPP	$[R_0 + R_2(1-e^{k_2 GPP})] \times e^{E_0(\frac{1}{T_{ref}-T_0} - \frac{1}{T_A-T_0})} \times \frac{\alpha k + SWC(1-\alpha)}{k+SWC(1-\alpha)}$	(Migliavacca et al., 2011)
addLinGPP	$R_0 \times e^{E_0(\frac{1}{T_{ref}-T_0} - \frac{1}{T_A-T_0})} \times \frac{\alpha k + SWC(1-\alpha)}{k+SWC(1-\alpha)} + k_2 GPP$	(Migliavacca et al., 2011)
addExpGPP	$R_0 \times e^{E_0(\frac{1}{T_{ref}-T_0} - \frac{1}{T_A-T_0})} \times \frac{\alpha k + SWC(1-\alpha)}{k+SWC(1-\alpha)} + R_2(1-e^{k_2 GPP})$	(Migliavacca et al., 2011)

$a, E_0, \phi_1, \phi_2, \phi_3, \phi_4, R_0, R_2, k, k_2$ and α are model parameters that can be optimized

2.3.2.4 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 dataset and its candidate drivers and parameters (Table 2.2). The structures and prediction performances of the GEP models were then compared with those of the optimized literature models.

2.4 Results

2.4.1 Artificial experiments

In the first artificial experiment the GEP approach is used to verify if 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 section 2.3.1.

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

$$R_{eco} = 0.35 \times 2.5^{(0.01 T_{air})} \tag{2.4.1}$$

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. 2.4. 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 for visualizing the increase in noise levels). There is a clear pattern of decreasing MEFs with increasing noise contamination. This was expected, as none of the methods should fit the noise added to the signal.

Figure 2.4,B, shows MEF values equivalent to Fig. 2.4,A, 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 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, the distributions of MEF values were compared for all runs for all studied SNR. Figure 2.5 exemplifies outputs for equation 2.3.10; panel a shows a drop of prediction capacity of the GEP models with noise increase for all types of fitness functions when compared with noise-infused data. This contrasts 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.2.5c), as well.

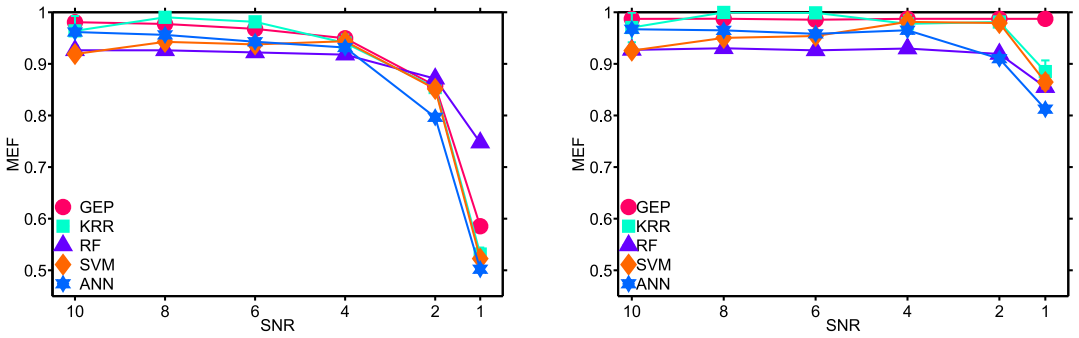


Figure 2.4: **Effect of adding noise to original signal on prediction capacity for GEP, KRR, RF, SVM and ANN**. The first panel contains the evolution of mean modelling efficiency (MEF) values from 20 independent runs for each increasing level of noise. MEF is computed after learning from a data set 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 equation 2.3.10.

2.4.2 Measured ecosystem CO_2 fluxes

Applying GEP on 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 equations 2.4.2-2.4.8.

$$R_{eco} = 1.2\log(T_{-10})^{0.8} \times e^{(\frac{GPP_s}{T_{-10}})} \tag{2.4.2}$$

$$R_{above} = 1.1SWC^{0.3} \times e^{(0.1GPP_s)} \tag{2.4.3}$$

$$R_{soil} = 0.04e^{(1.1T_{-10}^{0.4}+1.6SWC)} \tag{2.4.4}$$

$$R_{root} = 1.1e^{\frac{0.9GPP_s-6.8}{T_{-10}}} \tag{2.4.5}$$

$$R_{myc} = 0.001T_{-10}^{1.2} \times e^{(1.6T_{-10})^{SWC}} \tag{2.4.6}$$

$$R_{soil_a} = 0.01e^{(0.8T_{-10}^{0.6}+2.6SWC)} \tag{2.4.7}$$

$$R_{soil_h} = 0.8e^{\frac{0.6GPP_s-2.4}{T_{-10}}} \tag{2.4.8}$$

where, 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

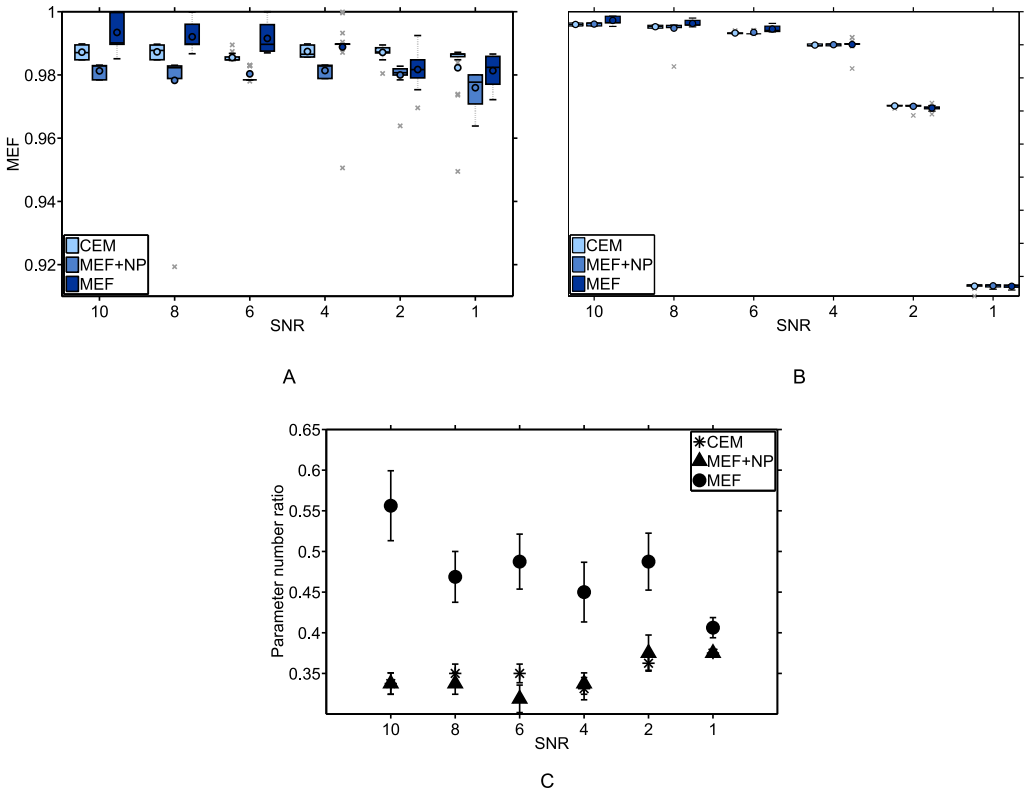


Figure 2.5: **Effects on modelling performance and parameter number caused by choice of fitness function** during GEP training for artificial noisy data generated by equation 2.3.10, where MEF is defined in equation 3.3.11 and CEM is defined in equation 2.2.3. **A**. Mean MEF when validation against noisy data after 20 GEP runs with different fitness functions. **B**. Mean MEF when validation 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.

SWC is volumetric soil water content. The corresponding cross-validation MEF values are given in Table 2.3, indicating a range of capacities for GEP models to represent different respiration types.

Table 2.3: Modelling performance for all extracted model structures after cross validation over 90 cases.

Respiration type	MEF	σMEF	Equation
R_{eco}	0.57	0.13	2.4.2
R_{above}	0.31	0.23	2.4.3
R_{soil}	0.79	0.04	2.4.4
R_{root}	0.59	0.08	2.4.5
R_{myc}	0.39	0.28	2.4.6
R_{soil_a}	0.82	0.05	2.4.7
R_{soil_h}	0.52	0.08	2.4.8

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

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 2.3), 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 Figures 2.6 and 2.7.

Figures 2.8 and 2.9 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} and , 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, the model including the smear term was kept.

In order to explore the capacity of the GEP models generated for the R_{eco} components to recreate the larger, across compartmental summed fluxes, the predictions of the models were summed and compared with the original fluxes (Fig. 2.10).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 Student's t-test (h=0, p=0.5). However, 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. The sensitivity of the sum fluxes to certain drivers

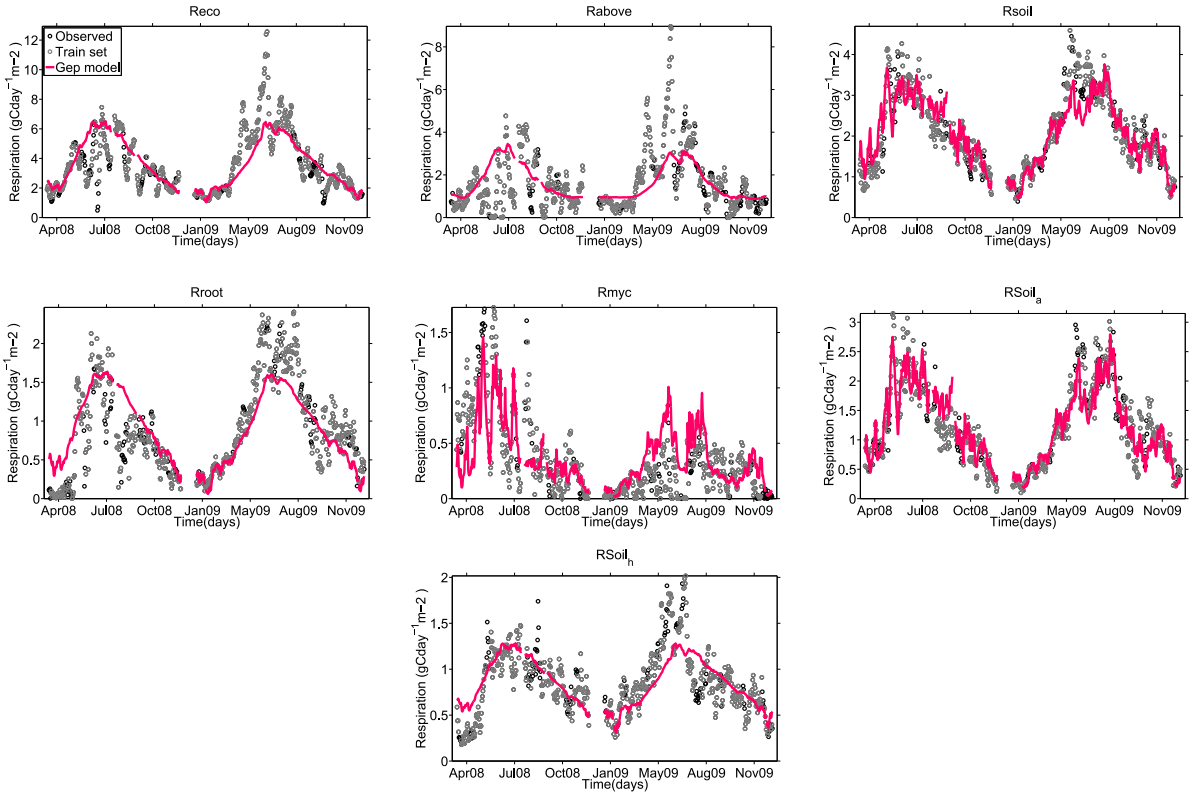


Figure 2.6: **Observed and predicted outgoing CO_2 fluxes.** 613 time steps of daily averaged CO_2 effluxes for two years at the Alice Holt oak forest site. The predicted values are generated with the models automatically built by the GEP approach with the settings given in table 2.1 for the following types of respiration: $R_{eco}, R_{above}, R_{soil}, R_{root}, R_{myc}, R_{soil_a}, R_{soil_h}$ and back-transformed with a smear term bias correction. The models are given in equations: 2.4.2-2.4.8

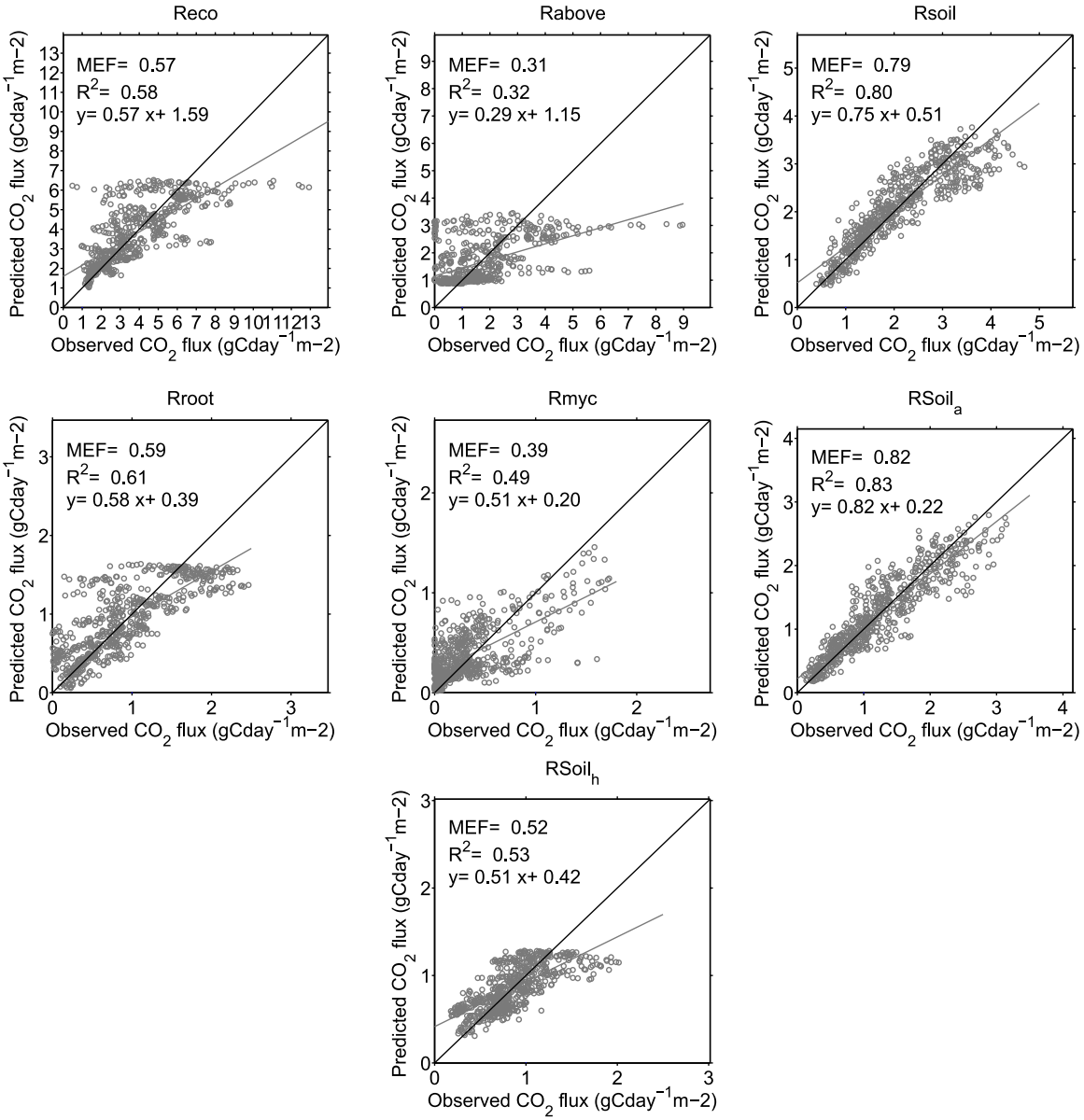


Figure 2.7: **Observed and predicted outgoing CO_2 fluxes.** 613 time steps of daily averaged CO_2 effluxes for two years at the Alice Holt oak forest site. The predicted values are generated with the models automatically built by the GEP approach with the settings given in table 2.1 for the following types of respiration: $R_{eco}, R_{above}, R_{soil}, R_{root}, R_{myc}, R_{soil_a}, R_{soil_h}$ and back-transformed with a smear term bias correction. The models are given in equations: 2.4.2-2.4.8

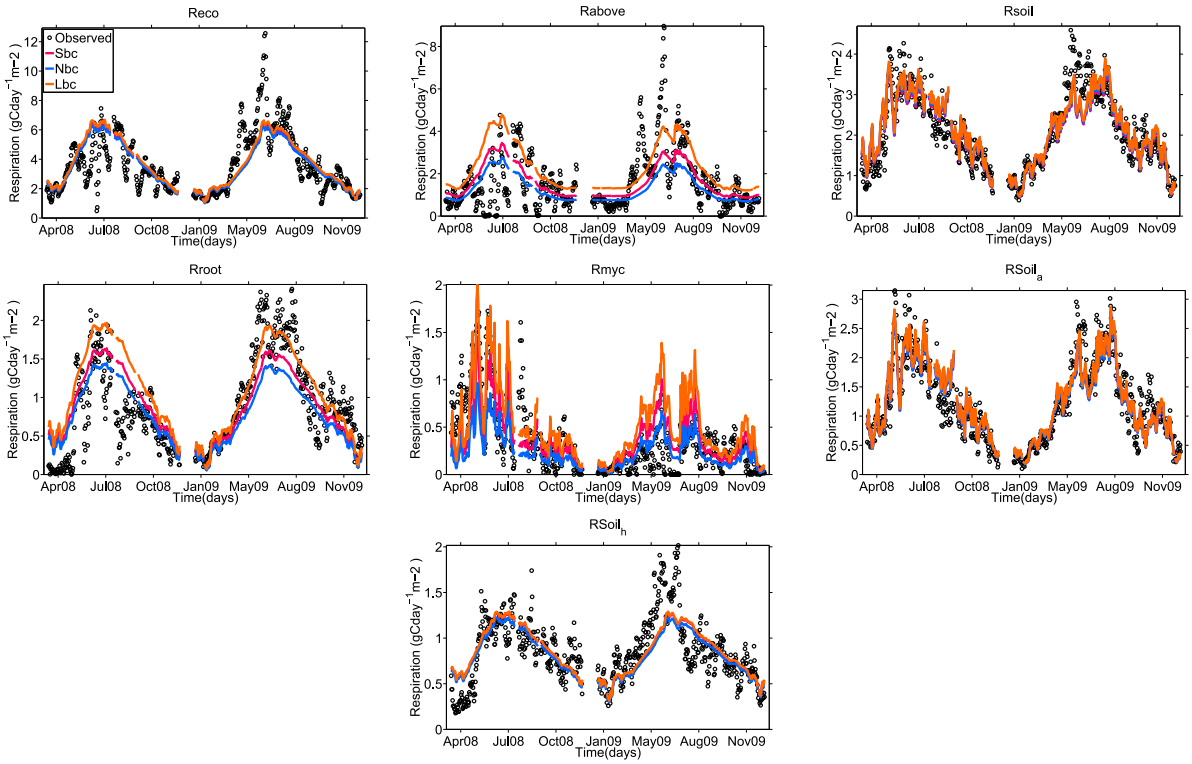


Figure 2.8: **Observed and predicted outgoing CO_2 fluxes.** 613 time steps of daily averaged CO_2 effluxes for two years at the Alice Holt oak forest site. The predicted values are generated with the models automatically built by the GEP approach with the settings given in table 2.1 for the following types of respiration: $R_{eco}, R_{above}, R_{soil}, R_{root}, R_{myc}, R_{soil_a}, R_{soil_h}$ and back-transformed with 3 types of residual **b**ias **c**orrection terms: **s**mear term , **n**aive, and **l**og normal term.

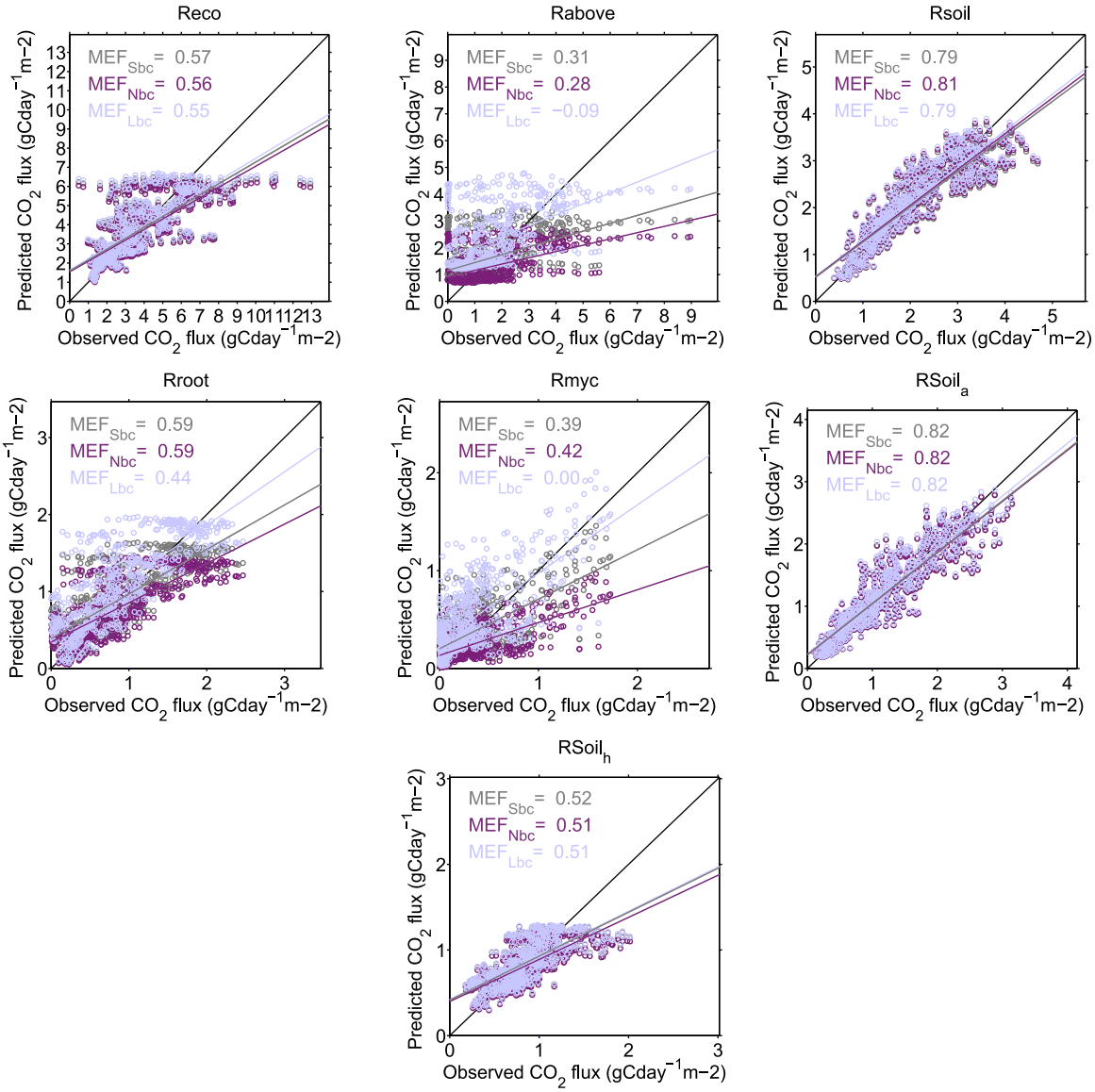


Figure 2.9: **Observed and predicted outgoing CO_2 fluxes.** 613 time steps of daily averaged CO_2 effluxes for two years at the Alice Holt oak forest site. The predicted values are generated with the models automatically built by the GEP approach with the settings given in table 2.1 for the following types of respiration: $R_{eco}, R_{above}, R_{soil}, R_{root}, R_{myc}, R_{soil_a}, R_{soil_h}$ and back-transformed with 3 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.

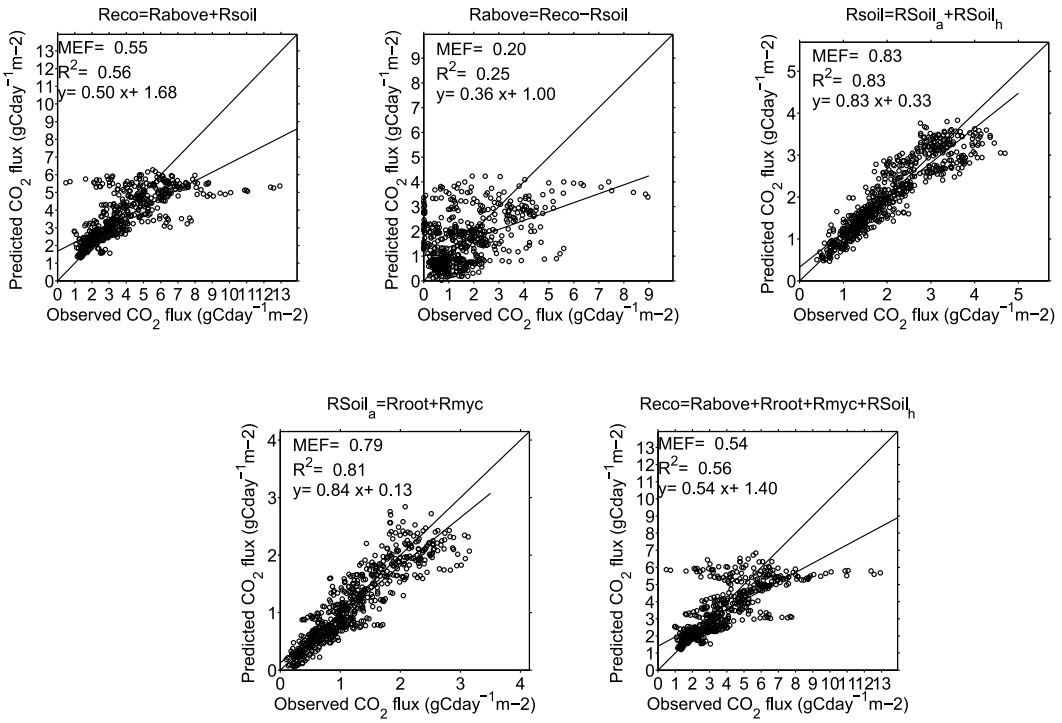


Figure 2.10: Observed versus predicted R_{eco} components fluxes, where predicted values are computed as derived fluxes based on the GEP models given in Eq. 2.4.2-2.4.8 that were trained on 500 d.p of daily mean values of various R_{eco} components.

can strongly manifest itself only in certain components which is why the drivers only get selected in the models built for those specific components.

The residuals depict some remaining patterns (Fig. 2.11 and Fig. 2.18 of suppl.) and the null hypothesis of normal distribution was rejected for all seven respiration component residuals at 5% significance level with the one-sample Kolmogorov-Smirnov test. Hence, additional information that could be extracted from the residuals might be expected. 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, GEP runs were performed 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. 2.15 in comparison with that of the other MLM presented in section 2.3.1.1. When correlation values were computed between the candidate drivers and the residuals, no significant linear correlations were found (Fig. 2.20 and 2.21 of suppl.).

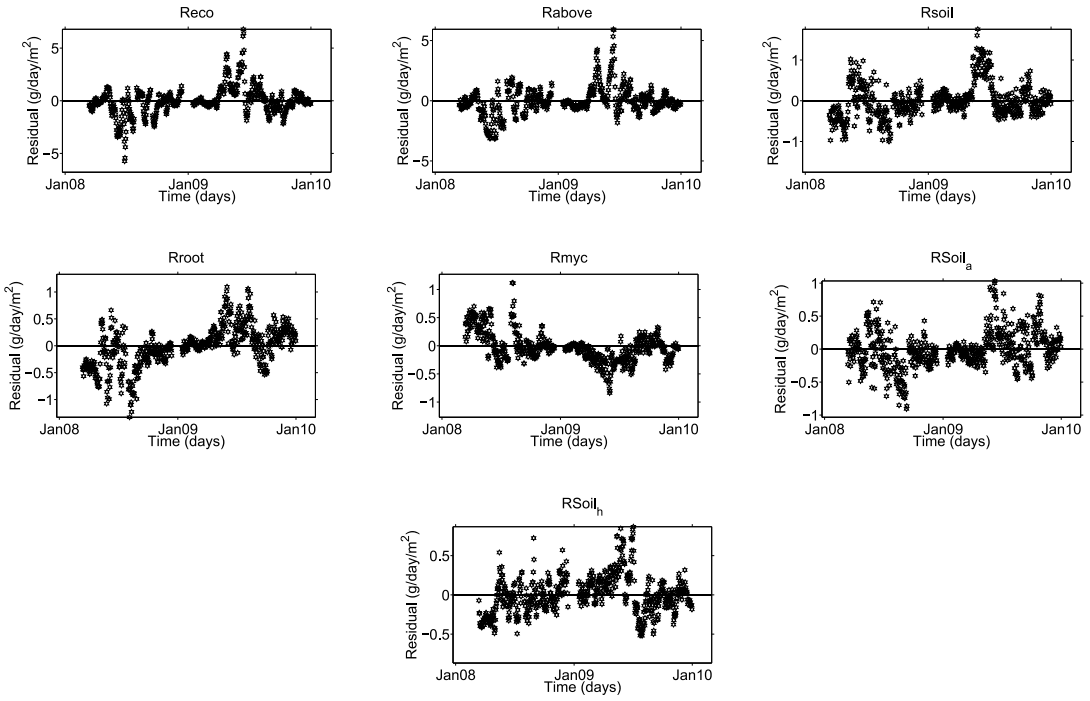


Figure 2.11: **Residuals computed for smear term bias corrected back-transformed GEP models for various types of CO_2 respiration fluxes** after training against log-transformed targets with the settings given in column 2 of Tab. 2.1.

2.4.2.1 Model transferability

The capacity of each extracted model structure (equations 2.4.2-2.4.8) to represent a component of R_{eco} not seen in the training procedure was studied by means of new CMA-ES optimization steps. The new prediction performances are illustrated in Tab. 2.4.

Table 2.4: 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. 2.4.2)	**0.57**	0.27	0.77	0.58	0.10	0.68	0.42
R_{above} (Eq. 2.4.3)	0.56	**0.31**	0.69	0.44	0.07	0.60	0.46
R_{soil} (Eq. 2.4.4)	0.50	0.20	0.79	0.47	0.38	0.82	0.39
R_{root} (Eq. 2.4.5)	0.23	0.27	0.57	**0.59**	0.01	0.65	0.51
R_{myc} (Eq. 2.4.6)	0.54	0.22	**0.82**	0.50	**0.39**	**0.84**	**0.52**
R_{soil_a} (Eq. 2.4.7)	0.50	0.20	0.79	0.47	0.38	0.82	0.39
R_{soil_h} (Eq. 2.4.8)	0.55	0.26	0.76	0.56	0.06	0.67	**0.52**

After optimization, none of the structures show an overall best MEF for all the R_{eco} components (i.e. an optimal general model cannot be clearly identified). However, certain model structures that tend to perform overall better than others were identified. This is the case for the R_{myc} model (eq. 2.4.6). 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 section 3.2 (Fig. 2.12). The prediction performance of GEP, KRR, ANN, SVM and, RF are shown in Fig. 2.15. Panel a contains the average MEF values computed for all MLM methods predicted values when compared to the original observations for $R_{eco}, R_{above}, R_{soil}, R_{root}, R_{myc}, R_{soil_a}, R_{soil_h}$. For all other cases, the performance is in the same range for all methods, but the GEP derived models having the lowest mean MEF values. Panel b shows that when all MLM were trained on the residuals obtained from comparing the GEP outputs with the observations, the GEP approach has the lowest capacity of capturing new relevant signals and is strongly outperformed by the rest of the MLM, indicating that amount of information retrievable by GEP with the current fitness and settings is limited and captured already in the first run.

2.4.2.2 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 Tab. 2.5. The literature model structure that performed best overall in terms of prediction capacity measured as MEF is the $WaterQ_{10}$ model (Fig. 2.13). Figure 2.13 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 respiration types, the highest MEF values are generally recorded by the GEP models.

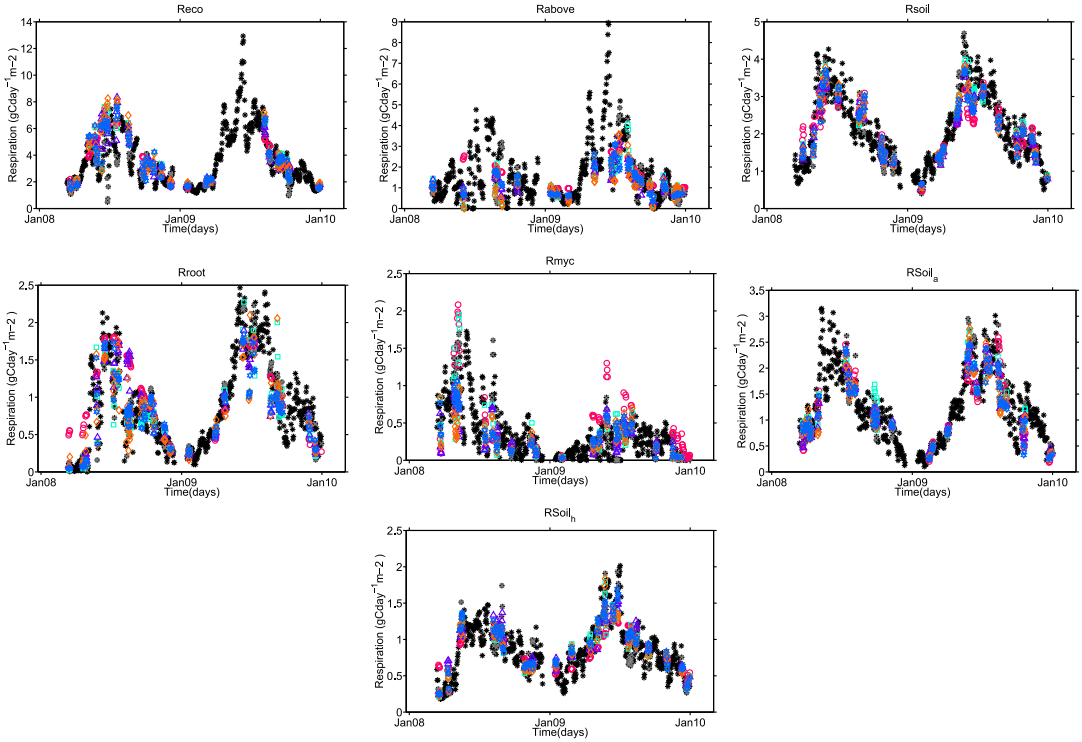


Figure 2.12: **Observed CO_2 fluxes and one set of 113 predicted values given by the some common machine learning methods (MLM) after training on 500 data points and after smear term bias corrected back-transformation.**

Table 2.5: 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

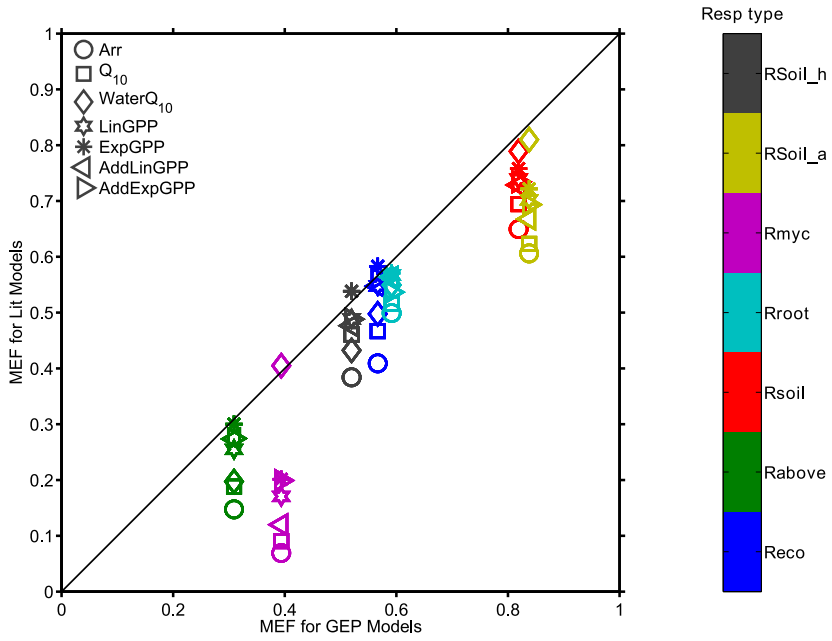


Figure 2.13: **MEF validation values for literature models and for the best GEP model in terms of MEF at each respiration level.** Each R_{eco} flux component is shown in a separate colour.

As the studied literature models performed best in modelling R_{soil}, the focus was on contrasting GEP model results against 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. 2.14, panel a). All models show the largest range of residuals for the months May to July in 2008, and June/July in 2009 (Fig. 2.14, panel b), with the two best-performing models (GEP and $WaterQ_{10}$) having the narrowest range of absolute residuals. Monthly mean average errors (MAE) indicate as well a systematic underestimation of soil CO_2 efflux in the first year (Fig. 2.19 of suppl.).

2.5 Discussion

2.5.1 On the GEP method

In this work, the primary reason for the artificial experiments was obtaining a better understanding of the capacity of GEP to solve symbolic regression types of problems. Emphasis was put 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, equation 2.3.10. The results show that all the studied methods are stable to presence of noise in the training set. These results increase the 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 random error) is probably larger than 4 which is where the curve starts decreasing in Fig. 2.4, 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 (Pérez-Priego et al., 2015; Ryan and Law, 2005), and the response of the present MLM is in the presence of different types of uncertainties and measurement noise cannot be stated.

The present 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. 2.5). Furthermore, it seems that due to applying constraints on the presence of structure in the residuals and the length of the parameter vector, the final mean number of parameters is lower when CEM is chosen.

2.5.1.1 Limitations

One of the critical aspects in this work is that GEP, as implemented here, can only represent and derive "$n \rightarrow 1$" type of response functions. GEP is not able to generate model structures that encode e.g. system-intrinsic dynamics like feedback loops, which are expected from the current understanding of biogeochemical cycles in terrestrial ecosystems (Ehrenfeld et al., 2005; Friedlingstein et al., 2006). Hence, GEP is suitable to e.g. understand and describe 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 chapter. 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 type of methods, such as GEP Keene (1995), whereas there is no effect on the prediction capacity of the other MLM as far as the author is 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.

2.5.2 The value of GEP for modelling ecosystem respiration fluxes

Model structures to describe terrestrial CO_2 respiration fluxes (equations 2.4.2-2.4.8) were automatically generated with GEP. Most of these structures (5 out of 7) were of rather low complexity, requiring only 4 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).

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; Heinemeyer et al., 2011; Kuzyakov, 2006; Subke et al., 2006).

When the GEP-derived models were compared with the community established semi-empirical models from a structural point of view, some key features for temperature dependencies of CO_2 fluxes typically captured by exponential relationships were shared, with some previously unconsidered dynamics revealed 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*. The exponential dependency of terrestrial ecosystem respiration components to *SWC* is a very intuitive pattern that has not yet been reported in the literature, and requires further exploration.

Another found difference 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, there is need for more in-depth analysis for determining whether the GEP described processes make actual biological sense and the selected drivers and their interactions represent true processes and responses.

2.5.3 Data quality

During the present 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 clear signal existing in the observations is not sufficient for constructing a good model with GEP.

2.5.4 High frequency variability

All GEP generated models underestimated the high respiration fluxes (Fig. 2.7) 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. 2.14).

A more in-depth comparison of all the GEP and conventional respiration models, based on a time-scale 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 time-scale 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 can-

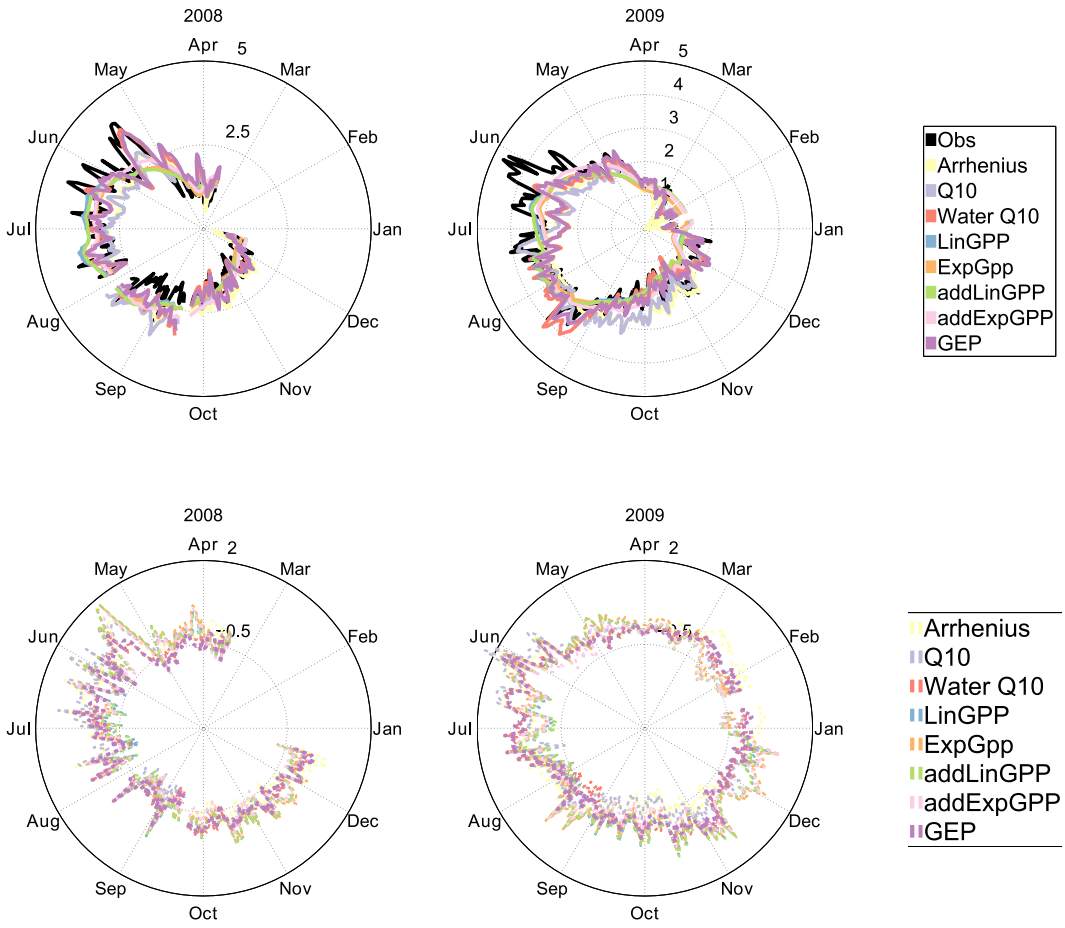


Figure 2.14: **Daily R_{soil} fluxes (A) illustrated in the context of the two studied years and residual values (B) of the total soil daily CO_2 outgoing fluxes as simulated by the investigated literature models and the GEP emerged model after smear term bias corrected back-transformation.** The fluxes 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.

didate 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.

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 the choice of fitness function. As there was penalizing 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 for the current approach the simplicity of structure, i.e. the possibility of interpretation is a very important asset.

The possibility of the underestimation of the carbon flux variability to be caused by the log-transformations applied to the observations was studied. It could be 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.

2.5.5 Equifinality

Table 2.4 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 parametrisation of a given "best" model. For this question to be addressed however, a larger sample of ecosystem types representative for different types of responses is needed where the importance of the obtained structures and their parameter set can be explored.

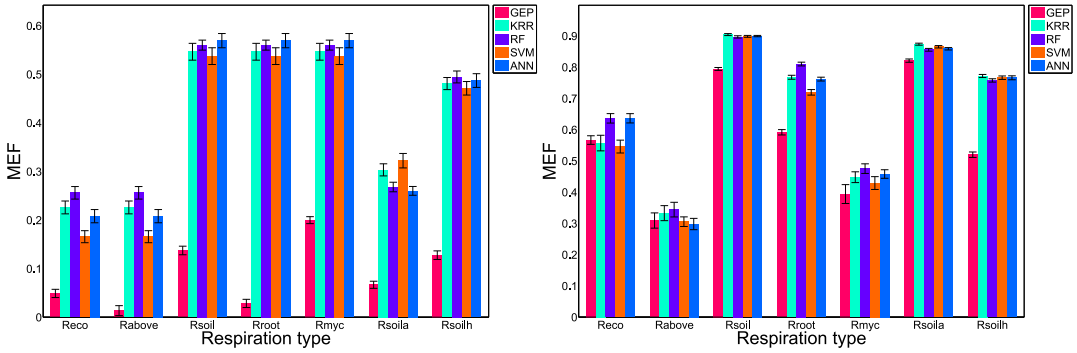


Figure 2.15: **Machine learning methods (MLM) prediction performance for all respirations components (left) and for the residuals (right) resulting from the GEP trained models after smear term bias corrected back-transformation**. The MEF values obtained for validation by all the MLM methods for $R_{eco}, R_{above}, R_{soil}, R_{root}, R_{myc}, R_{soil_a}, R_{soil_h}$

2.5.6 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. 2.15). The analysis of training all the MLM on the GEP residual output showed that the GEP approach is not able to retrieve any new meaningful structural components, but that the remaining MLM 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 reconstructing the underlying R_{eco} fluxes and is not prone to over-fitting, it could be that the current set-up of the GEP is not sufficient for an exhaustive description of those fluxes, or that might be overly strict on complexity of models compared to other MLM. The GEP approach has, nevertheless, the benefit of producing mathematical model structures that can be the basis for future interpretation.

2.6 Conclusions and Outlook

Overall, the 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. The potential of GEP for symbolic

regression was conceptually shown using artificial data. This study also shows that GEP always yields results as good or better than 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, new structures were found 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, when the GEP derived models are mutually compared, there are sufficient structural particularities for each terrestrial respiration type 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, the inclusion of a parameter optimization step was very helpful to further test the transferability of model structures. But this approach could be potentially integrated into the GEP evolution. More specifically, the next development of GEP could include the parameter optimization as an intermediate step before selection during each evolution generation (Ilie et al.). In this way, a model structure could be chosen according to not only the current state of parameters but also on its potential and convergence to a global solution might be achieved faster.

2.7 Supplemental Materials:

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

Table 2.6: Standard error of the MEF at validation values for all MLM for different SNR values when the MEF values are computed against the noisy data.

SNR	GEP	KRR	RF	SVM	ANN
9.82	0.00	0.00	0.02	0.00	0.00
8.18	0.00	0.00	0.02	0.02	0.00
7.01	0.00	0.00	0.02	0.01	0.00
6.14	0.00	0.00	0.02	0.01	0.00
5.45	0.00	0.00	0.02	0.02	0.01
4.46	0.00	0.00	0.02	0.01	0.00
3.27	0.01	0.01	0.02	0.01	0.01
2.73	0.01	0.01	0.02	0.01	0.01
2.34	0.02	0.01	0.02	0.01	0.01
1.96	0.02	0.02	0.02	0.02	0.01
1.75	0.02	0.02	0.02	0.03	0.02
1.40	0.05	0.03	0.02	0.02	0.02
1.23	0.03	0.03	0.02	0.03	0.03
1.09	0.04	0.03	0.03	0.04	0.03
1.00	0.04	0.03	0.02	0.03	0.03

GEP models for all log-transformed respirations types time series, before back-transformation.

$$\log(R_{eco}) = \frac{GPP_s}{T_{-10}} + \log(\log(T_{-10})) \tag{2.7.1}$$

$$\log(R_{above}) = 0.1T_{-10} + 0.4\log(0.8\sqrt{SWC}) \tag{2.7.2}$$

$$\log(R_{soil}) = 1.2T_{-10}^{0.4} + 1.3SWC - 3.1 \tag{2.7.3}$$

$$\log(R_{root}) = 0.9\frac{1.2GPP_s - 8.1}{T_{-10}} \tag{2.7.4}$$

$$\log(R_{myc}) = 1.1\log(1.7T_{-10}) + 1.2T_{-10}^{SWC} - 7.4 \tag{2.7.5}$$

$$\log(R_{soil_a}) = 1.2T_{-10}^{0.5} + 2.5SWC - 4.9 \tag{2.7.6}$$

$$\log(R_{soil_h}) = -0.3 + 0.6\frac{1.1GPP_s - 3.6}{T_{-10}} \tag{2.7.7}$$

Table 2.7: Standard error of the MEF at validation values for all MLM for different SNR values when the MEF values are computed against the clear data.

SNR	GEP	KRR	RF	SVM	ANN
9.82	3e-07	4e-05	2e-02	4e-03	4e-03
8.18	3e-07	6e-05	2e-02	2e-02	2e-03
7.01	3e-07	4e-05	2e-02	1e-02	2e-03
6.14	2e-06	7e-05	2e-02	2e-02	2e-03
5.45	2e-06	1e-04	2e-02	2e-02	4e-03
4.46	6e-06	1e-04	2e-02	2e-02	2e-03
3.27	9e-06	2e-03	2e-02	1e-02	3e-03
2.73	4e-05	4e-04	2e-02	1e-02	6e-03
2.34	4e-05	6e-04	2e-02	9e-03	3e-03
1.96	8e-05	1e-03	2e-02	1e-02	3e-03
1.75	2e-04	8e-04	1e-02	1e-02	5e-03
1.40	8e-04	1e-03	1e-02	2e-02	5e-03
1.23	1e-04	2e-03	1e-02	2e-02	4e-03
1.09	4e-03	3e-03	1e-02	2e-02	5e-03
1.00	7e-04	3e-03	1e-02	5e-02	6e-03

Figure 2.16 in supplemental material illustrates the change in the shape of the PDF estimated for each respiration type after log-transforming. For all time series, the skewness is visibly is reduced.

From Fig. 2.20 and 2.21 is worth mentioning the apparent correlation, although weak in terms of R^2 value, of the R_{myc} residuals with GPP_s, even when this was not chosen as a driver, indicating that the relation was not strong enough for an explicit model inclusion but it could show a dependency to a driver for which GPP_s acts as a proxy such as phenology, or substrate availability. Such weak correlations are present as well between R_{soil} and R_{soil_h} residuals and T_{air}.

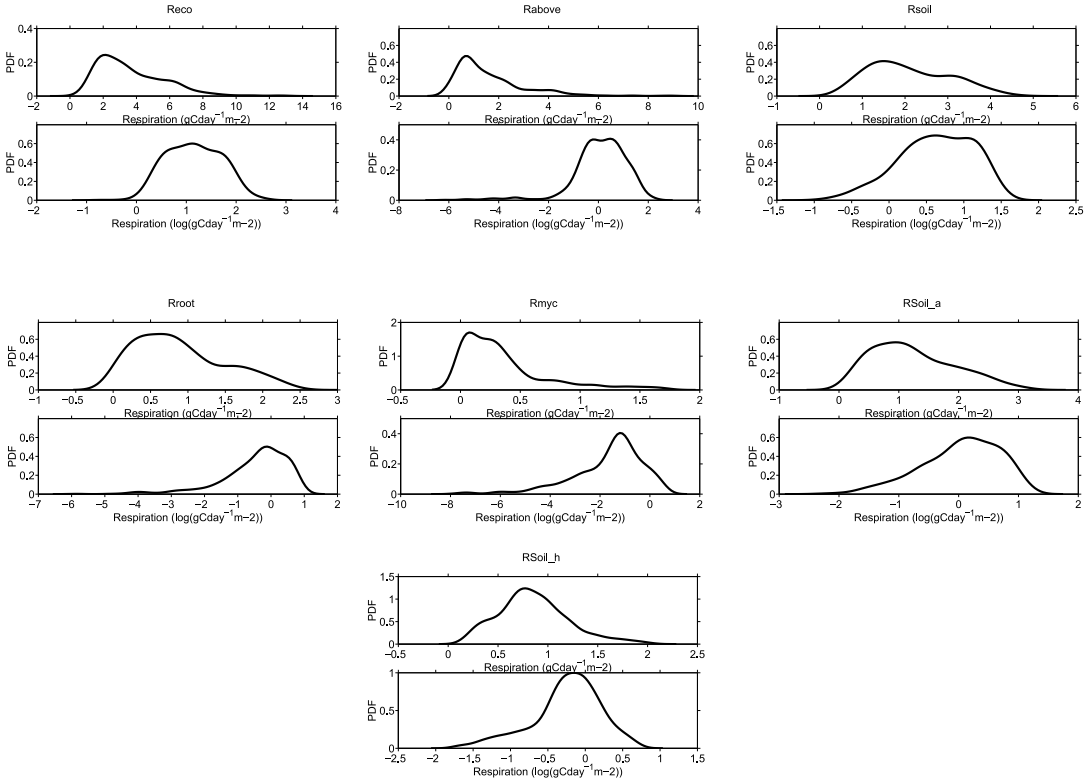


Figure 2.16: **Change in estimated density function of observations before and after log-transforming for all studied respiration types.**

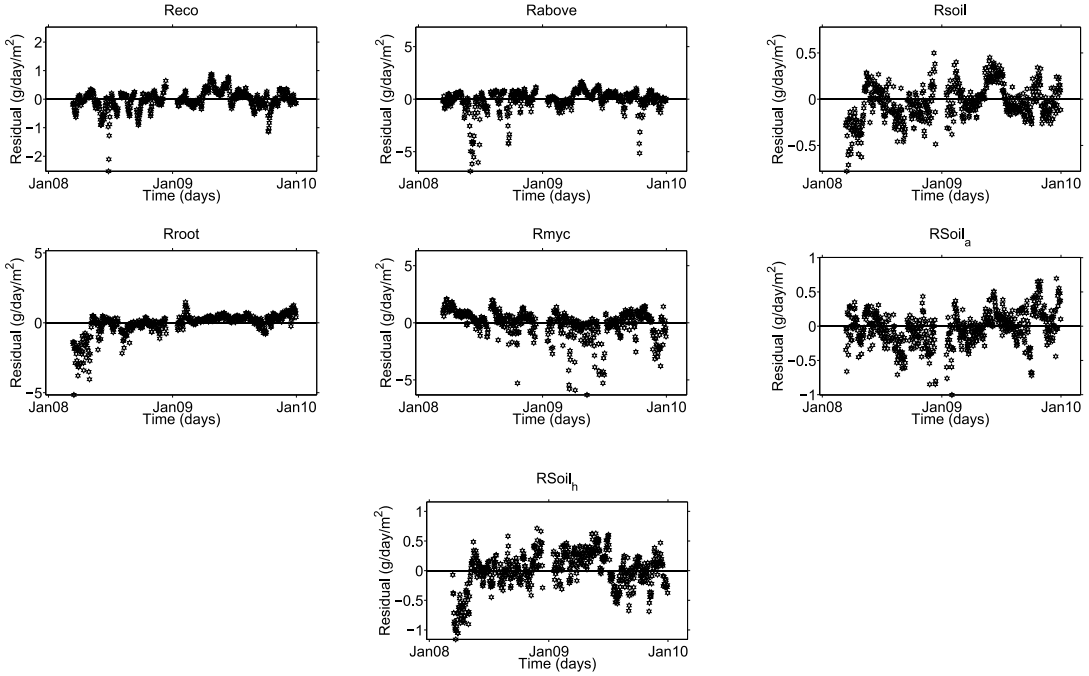


Figure 2.17: **Residuals computed for the GEP models against the log-transformed targets before back-transformation.**

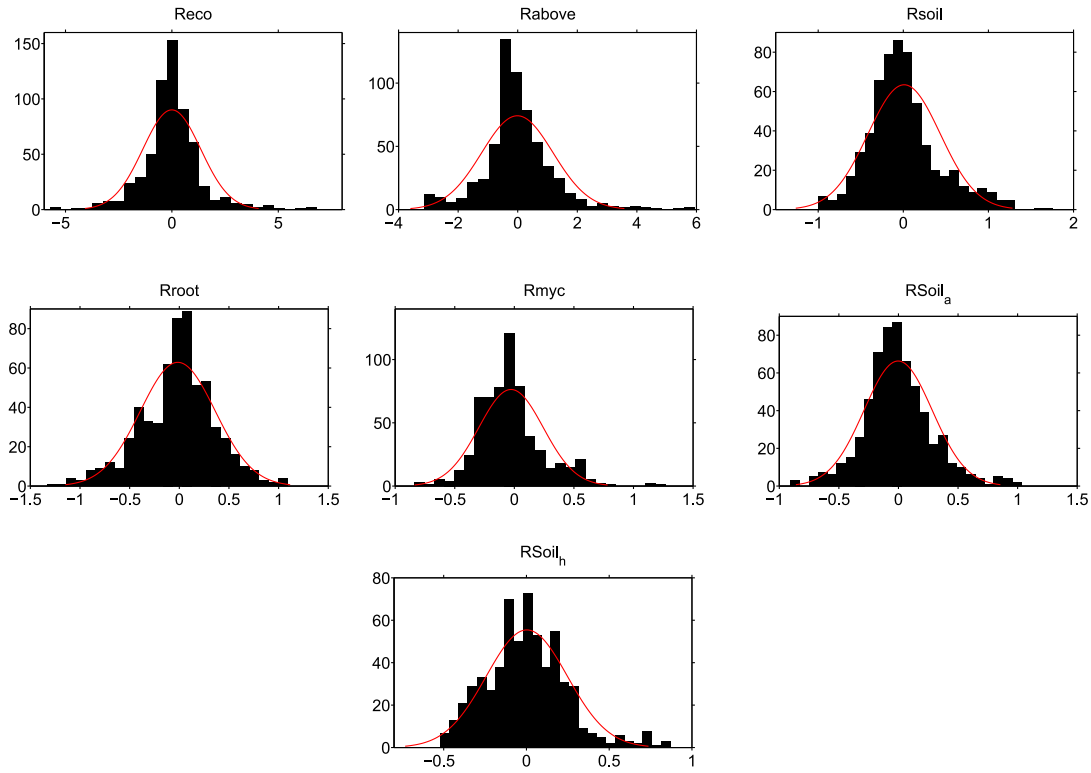


Figure 2.18: **Distributions of the residuals after smear bias correction computed for the GEP models after training on log-transformed data.**

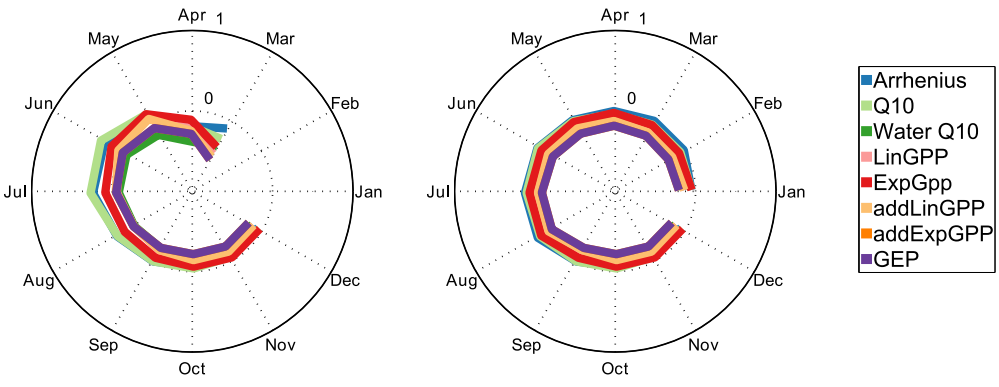


Figure 2.19: **Monthly averaged error values for some literature models for and the GEP generated model** for daily soil CO_2 efflux in the two studied years. The centre of the plots is -1. The scale of the fluxes is given in gC/m^2/day.

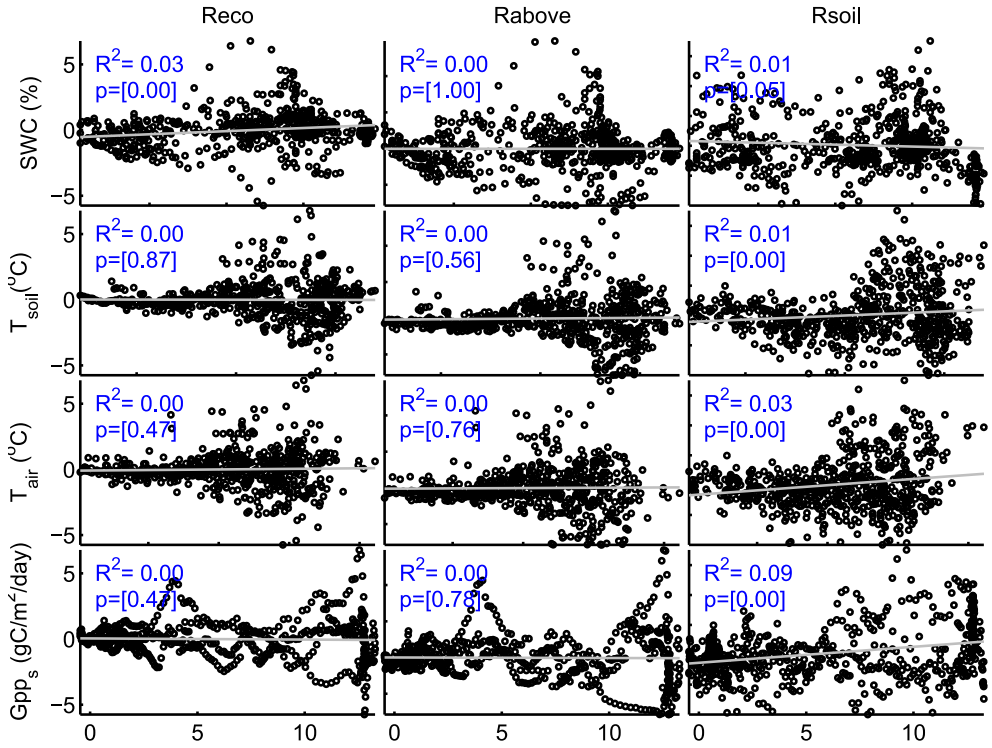


Figure 2.20: **Candidate driver linear correlations with residuals** computed after bias corrected transformation of the GEP models from runs with settings given in Tab 2.1 for R_{eco}, R_{above} and R_{soil}. The drivers are on the X axis and the residuals on the Y axis. The candidate driver is given as title of each row and the type of respiration is given as title of the column.

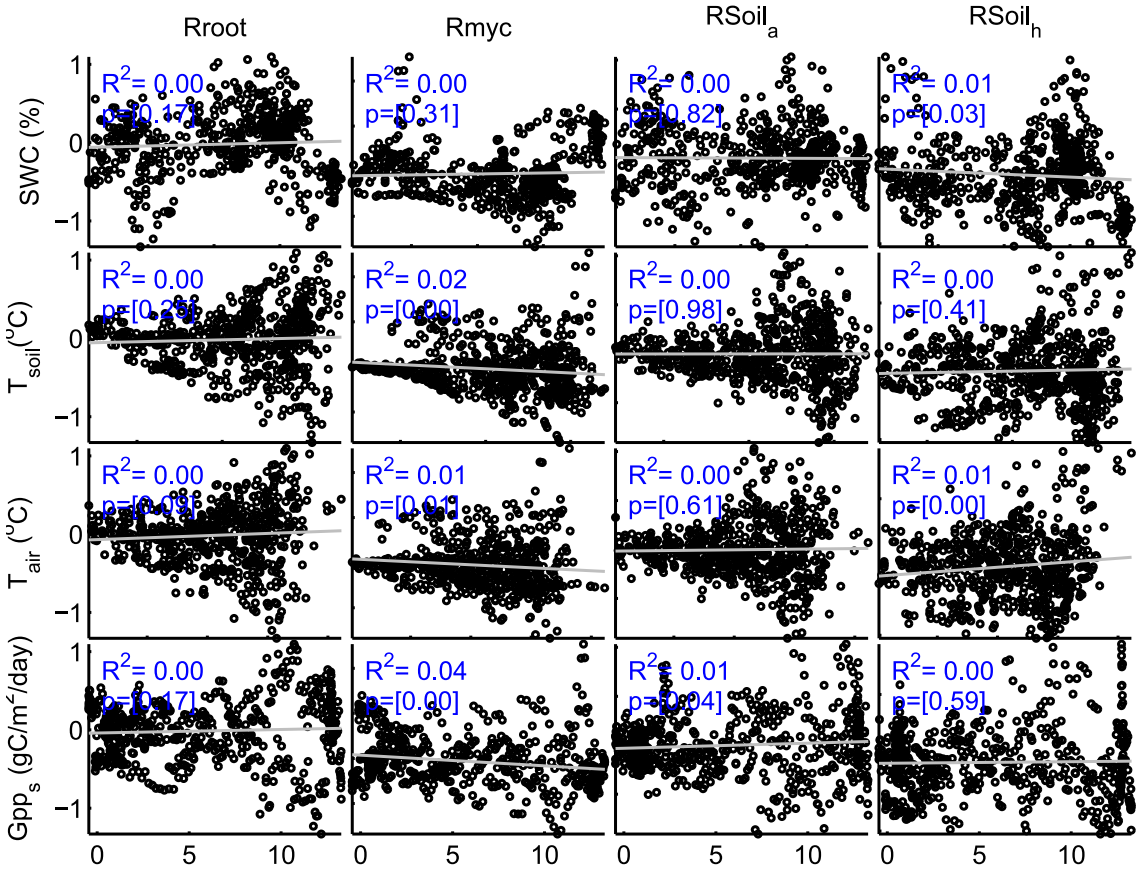


Figure 2.21: **Candidate driver linear correlations with residuals** computed after bias corrected transformation of the GEP models from runs with settings given in Tab 2.1 for R_{root}, R_{myc}, R_{soil_a} and R_{soil_h}.The drivers are on the X axis and the residuals on the Y axis. The candidate driver is given as title of each row and the type of respiration is given as title of the column.

Evolving compact symbolic regressions by a GEP and CMA-ES hybrid approach

Abstract

Gene expression programming (GEP) has been shown to produce good results for symbolic regression problems. However, the GEP proposed mathematical expressions are very often long and difficult to interpret. This chapter describes a novel method that extends GEP with local optimization of the real valued constants in the symbolic expressions by a covariance matrix adaptation evolutionary strategy (CMA-ES), called CMAGEP. The performance of both GEP and CMAGEP approaches is evaluated on a set of 20 artificial test problems, of which 10 are standard problems from the literature and 10 are their variations. Furthermore, two real data sets containing observations of sunspots and soil respiration are considered. It is found that on both training and validation data sets the prediction performance of the new approach is always as good as that of the standard method or improved. More importantly, the CMAGEP proposed symbolic representations are always significantly shorter, with GEP solutions between one and two thirds longer. A comparison with standard implementations of four learning methods (KRR, RF, SVN, ANN) reveals a prediction accuracy in a similar range. It can be concluded that the newly proposed approach is able to deliver relatively compact and interpretable solutions that are at the same time reasonably representative for the data. As such, the present approach can be used for automatically discovering general laws in a symbolic form.

3.1 Introduction

When learning from data, it is sometimes not only desirable to produce sensible predictions but also to obtain a predictor in a symbolic, compact form that can be regarded as a comprehensible law explaining regularities in data. Learning methods like neural networks are good predictors, nevertheless they do not deliver such expressions. Genetic Programming (GP) Koza (1994) type of approaches have been proven promising for obtaining such symbolic forms, although the results are often still difficult to interpret due to bloat and other effects Banzhaf and Langdon (2002); Smith (2000).

Here, I introduce a novel method that combines gene expression programming (GEP) Ferreira (2004) with an evolutionary strategy (ES), specifically covariance matrix adaptation ES (CMA-ES) Hansen (2006a); Hansen et al. (2003), which are state of the art approaches for symbolic regression and real valued parameter optimization, respectively.

GEP is an evolutionary algorithm that evolves computer programs (i.e. mathematical expressions, decision trees, classification rules etc.). As a branch of GP, its internal design makes it a suitable candidate for tackling symbolic regression problemsDanish (2014); Guven and Aytek (2009); Imani et al. (2014). CMA-ES is, on the other hand, black-box type of optimization that samples new solutions during evolutionary generations from a Gaussian multivariate distribution Hansen (2011) by adapting it's internal state variables to perform a natural gradient descentAmari (1998).

CMAGEP is the hybridization of GEP and the CMA-ES approach. CMAGEP allows for CMA-ES optimization of the constants contained in a set number of best individuals, after a given GEP generation count. The optimization of constants is done based on the same fitness function used for selection in the GEP evolution. In the CMAGEP hybrid, the set number of best individuals are re-evaluated after the CMA-ES optimization and get assigned new fitness values based on which become subject for further selection.

The effects of including the CMA-ES constant optimization in the GEP evolution process are studied. I focused on possible changes in proposed structures, in complexity or length of regressions, and on changes in prediction performance of the final solutions. In this framework, a number of artificial and real data test problems were designed based on which the following were found:

- For two artificial data sets generated from 10 well known symbolic problems and 10 derived cases, the prediction accuracy of the CMAGEP method was equal to or higher than the standard GEP approach on training and validation sets, in all 20 studied cases.

- For the same artificial tests it could be seen that the final solution complexity in the CMAGEP case was as much as 60% lower than that of the standard GEP solutions.

- When the optimal time to start the CMA-ES constant optimizations during the CMAGEP evolutions was studied, it was found that the structure length was much lower with an earlier start, but that prediction capacity is only sightly improved.
- The two approaches were evaluated against a well known data set containing the Wolfer Sunspot time series Izenman (1983), where the purpose was to generate a model that predicts the present number of solar spots depending on the previous records with as much as 10 lag variables. For this analysis, the results of my implementations of the GEP and CMAGEP algorithms were also compared with the results returned by the demo version of "GeneXproTools", a commercial software implemented by the creator of GEP. The mean prediction performance of the structures produced by the two GEP approaches was very similar on training sets and higher for my implementation of GEP on validation sets. The same stands for the mean tree size, with commercial GEP solutions having 3 nodes more than those of my implementation. The CMAGEP results outperformed the both GEP implementations in prediction performance as well as solution length.
- Finally, when the two approaches were used to generate models that would simulate outgoing CO_2 soil fluxes depending on soil and air temperature, although the returned models describe similar dynamics, the CMAGEP solution was shorter than the GEP solution, allowing for easier interpretation.

The authors implementation of the algorithms based on which the present results were obtained are available as a package under creative commons licence in a gitHub repository.

3.2 Method

3.2.1 GEP

GEP is an evolutionary system (Fig. 3.2, orange tiles) that, for symbolic regression, automatically constructs a mapping of n independent inputs to 1 dependent target by a random generation of individuals further subjected to selection and evolutionary operators. The most unique feature of GEP is found in the type of encoding its evolution individuals. Introduced by Ferreira (2001, 2006), the GEP system is a complete evolutionary system that includes both a genotype and a phenotype, making it a combination of GP type of approaches and genetic algorithms Goldberg and Holland (1988). The GEP genotype is a set of fixed length linear strings whereas the phenotype is given by

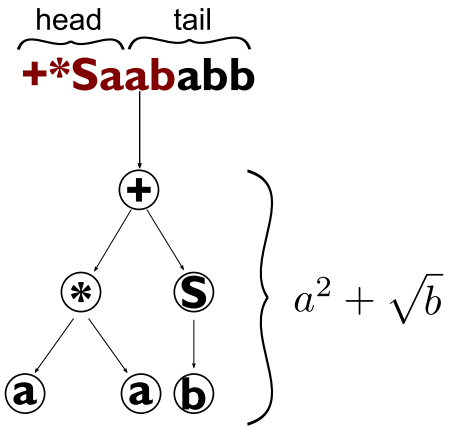


Figure 3.1: **Translation of a GEP gene**, the smallest component of a chromosome, the GEP evolution individual. More than one genes are connected in a chromosome with the help of linking functions. The current gene has a head of 4 characters **+*Sa** and a tail of 5 characters **ababb**. With the help of the GEP internal language, the Karva language, the gene string, +*Saababb is translated into an expression tree like so: each function takes for sub nodes as many characters as it needs that have not been yet used. The process is continued until there are no more functions that have not been associated with their respective components and there are only terminal characters left in the string. In the current example, **+** is a binary function, so it will take as sub nodes the next 2 characters *** S**, * takes**a** and **a** as sub nodes and the tree on this side is complete, after which the unary function **S** only needs **b** to complete the tree. This means that only the red coloured component of the gene is active and translatable into mathematical expressions. The remaining encoded genetic material can only become active during the evolution, by means of genetic manipulation.

encoded expression trees (ET) that can be further translated into mathematical expressions. The decoding of genotype into phenotype is achieved via an internally defined language, called Karva and is further described in Fig. 3.1.

In the GEP context all evolution individuals start as a collection of strings that are called genes. A multitude of linked genes makes a chromosome, the GEP evolution individual. In regression problems, the population of chromosomes is generated based on a set of candidate functional transformations, such as addition, exponential, sine, etc., and a set of candidate predictors, here called terminals, via their prescribed mapping characters. A GEP gene comprises a head and a tail, where the head is of a combination of characters mapping to functions and terminals, and the tail exclusively contains characters mapping to terminals. In order to insure the validity of ET during

translation, the structure of a gene must follow Equation 3.2.1.

$$T = H \times (mP - 1) + 1 \tag{3.2.1}$$

where, T is tail length, H is the head length and mP represents the maximum number of parameters required by any function in the initial function set.

The use of linear strings for encoding ET generates variety in shape and size, allowing for a larger space of exploration while still maintaining a reduced alphabet for encoders. This is because in GEP, the length of the genes in chromosomes does not determine the size of the ET they represent. The ET shape and size is given only by the active part of the gene, also known as the Open Reading Frame (ORF). Ferreira borrowed the concept from natural genetic evolution, where only parts of genes become active at different given times depending on genetic variation. Following this model, in the GEP context , although the genes are of fixed length, the trees they encapsulate can have different shapes and sizes depending on whether an element of the gene is active at the moment of translation or not.

For assessing the performance of expressions encoded by the chromosomes based on target data, a fitness value is assigned to each individual in the population via a fitness function. The fitness of each individual becomes the basis for the selection process, making the fitness function design of utmost importance in the entire evolution process. Al all generation steps, the individual with the best fitness is saved and the n-1 remaining individuals are object of tournament selection for generating offspring for the next generation. Using the replication operator only for the best fit individual and subjecting the rest to genetic variation operators, allows for less fit individuals to be selected and create offspring as well. This aspect of the GEP evolution is highly important as all GEP individuals can become useful over multiple generations, due to genes comprising ORFs and inactive sections. Such a structure makes that some individuals that would not be very fit in a specific evolution step to get elements of their inactive sections activated by genetic operators over the following steps, leading to translations of much better candidate solutions. Another possibility is that by reproducing non-fit individuals, relevant sections of their genetic material is sent in the next generations, and by entering the genes of new individuals can become valuable for the final solution discovery.

Following selection, all individuals, except for the candidate associated to the best fitness value are subject to possible genetic manipulations such as: mutation, inversion, recombination and cross-over. The offspring generated during this process are added, along with the best fit individual, to a new population that will pass through the same evolution cycle as the population in the previous generations. The evolution process is repeated a stop criterion is reached. Some examples of stop criteria are: best possible fitness value, highest number of allowed iterations without change in fitness, maximum run time and so on. Once a stop criterion is reached the individual with the best fitness

value is returned as proposed solution to the symbolic regression problem.

3.2.2 CMA-ES

The Covariance Matrix Adaptation Evolution Strategy Hansen et al. (2003) is black box, stochastic optimization algorithm that seeks to minimize an objective function f, where $f : \mathbb{R}^n \to \mathbb{R}$, by estimating and adapting a covariance matrix $C \in \mathbb{R}^{n \times n}$ based on sampling from a multivariate normal distribution of dimension $\mathbb{R}^{n \times n}$. It is a second order optimization approach that efficiently minimizes objective functions and is widely used in the case of non-linear, non-convex, ill-posed problems Hansen (2006b).

An important feature of CMA-ES is its invariance to linear transformations of the search space Auger and Hansen (2005), giving the same results for an objective function f, where $f : x \in \mathbb{R}^n \to f(x) \in \mathbb{R}$ and on $f_R : x \in \mathbb{R}^n \to f(Rx) \in \mathbb{R}$, where a full rank linear transformation has been conducted. The property of invariance offers the CMA-ES an advantage in searching for solutions in a non-convex non space.

A new aspect of the CMA-ES design in the context of ES algorithms is the possibility to monitor and self-adapt internal state variables such as the evolution paths, the step size and covariance matrix during the process of searching for an optimal solution. By controlling the evolution path and by adapting the mean and the covariance matrix of multivariate normal distribution from which the sampling for a new generation of candidate solutions a steeper learning is achieved. By adapting the step sizes as well, with steps becoming longer in order to cover the search space better when many small steps are made in a similar direction with low fitness gain, a premature convergence is avoided, keeping a higher chance of reaching a global optimal solution.

The CMA-ES algorithm used in this work is based on the Python implementation by Hansen as described in detail in Hansen (2016) and summarized in section 3.6.

3.2.3 CMAGEP, a CMA-ES and GEP hybrid

The GEP based approach proposed in this work for compact symbolic regression discovery is the CMAGEP. In CMAGEP, an optimization of constants step by CMA-ES to the original GEP evolution for a better calibration of the GEP generated solutions. The main reasons for choosing the CMA-ES approach for constant optimizations are *a*) the large domain of applicability, given that regressions that might emerge from the GEP training would have unknown conditions for search spaces of the optimal parameter sets; and *b*) the good success rate of optimization even in the case of non-linear or ill-posed problems , which might very well be the case for some of the GEP evolved individuals.

Because the purpose of introducing the CMA-ES optimization in the GEP evolution was to do a calibration of individuals, and not necessary obtaining the best possible

parameter set for the best possible model structure, I chose to allow for the optimization of constants to take place right before the selection step of the standard GEP version. The optimization location in the evolution process was decided so that the more adjustable solutions have a higher chance of being chosen more often for reproduction. In this manner, the candidate solutions would not be chosen for the performance in a specific state of parametrization, but for their best possible performance.

Very often optimizations that perform a global search, such as CMA-ES does, can be very time and resource costly, so in CMAGEP, two important limiting optimization parameters were introduced: CMA-ES start time t_o, giving the generation count from which individuals are subject to parameter optimization and the number of chromosomes to optimize μ' stating how many individuals to select for CMA-ES constant optimization from the population of candidates after sorting in descending order by fitness values. In order to avoid time spent on overly long CMA-ES searches for optimal parameter sets, a CMA-ES time-out condition was imposed as well, that states a specific time allowed to wait for a CMA-ES result, that when surpassed, stops the CMA-ES, making the GEP individual remain in the unoptimized parametrization state.

Importantly, since the GEP component of CMAGEP generates individuals that map to mathematical expressions not yet in simplified form, i.e, certain combinations of mathematical operations will lead to sections of the expression to be reduced, before optimizing the constants of the μ best individuals, the certain steps need to be performed.

First, the expressions associated to the individuals that need to be optimized are simplified using the "Sympy" Python package. Second, the number and locations of constants to be optimized is determined with the assumption that all functional transformation and all terminals have a constant associated that can be optimized. Lastly, if the constants associated to a function or a terminal are specified in the mathematical expressions, they are added to the set constants that will be further fed as input to CMA-ES, otherwise the default value 1 is added in the set at the location reserved for specific function or terminal.

Once the set of constants to be optimized is built, it is sent to CMA-ES for optimization, and based on the resulting optimized set, the constant values of the best μ individuals are updated. At this step their fitness values are updated as well.

A fully detailed description of the proposed CMAGEP algorithm is given in section 3.6 and in Fig. 3.2.

3.3 Experimental set-up and results

In order to study the effects of introducing the optimization in the GEP evolution on the global fitness performance and solution complexity in newly proposed hybrid algorithm, CMAGEP, two artificial and two real data experiments were conducted.

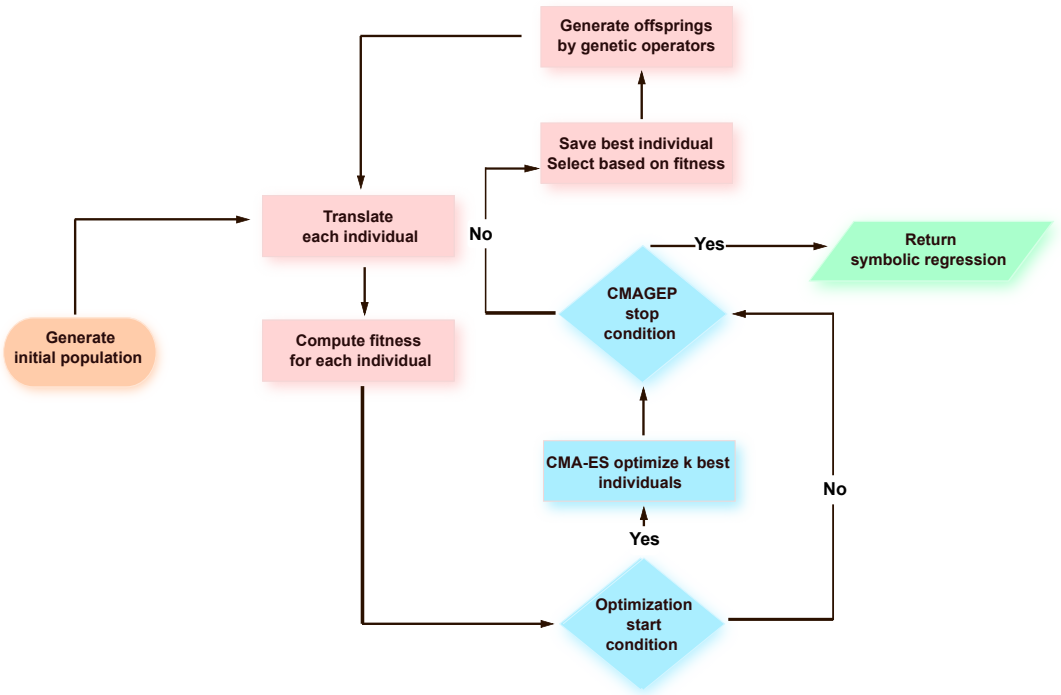


Figure 3.2: **CMAGEP work flow.** The evolution of a symbolic regression by CMAGEP is done as follows: **1.** An initial population of n individuals called chromosomes is generated based on random selection from two sets of characters mapping to possible functional transformations and candidate predictors; **2.** The chromosomes are translated into expression trees and then into mathematical expressions based on the process described in Fig. 3.1;**3.** The mathematical expressions of the chromosomes are evaluated against training data and a fitness value is assigned to each chromosome based on a fitness function; **4.** The population of chromosomes is sorted based on the corresponding fitness values; **5.** Optimization condition is checked and if it is met, **6a.** the best k individuals have their parameters optimized by a CMA-ES; if the condition is not met, **6b.** the CMAGEP stop condition is checked and if met, **7a.** the first individual is returned as solution, otherwise, **7b.** first individual is copied and other n-1 individuals are generated for the next generation after fitness based selection and **8.** genetic manipulation based on the available genetic operators; **9.** Steps 2-9 are repeated for the newly generated individuals until stop conditions are met.

All the experiments presented in this chapter were conducted based on my implementations of the GEP and CMAGEP algorithms and all the runs were done on an HPC cluster, on independent nodes. The GEP implementation is done in the C++ language and compiled under the gcc 4.7 complier. The CMAGEP implementation contains code written in C++ for all the GEP related operations and coded written in Python for the CMA-ES optimization. The results of the optimization are transferred from the Python objects into C++ objects for GEP through the Python/C API Foundation.

3.3.1 GEP benchmark on artificial test functions

In order to asses the prediction performance of the standard GEP approach and the CMA-ES GEP hybrid in the context of symbolic regression, a set of mathematical functions (3.3.1-3.3.10) was built based on the genetic programming community proposals from the GECCO 2013 benchmarking discussions White et al. (2013) and the work by Vladislavleva et. al in Vladislavleva et al. (2009). The set also contains a "V" shaped function (3.3.2) as discussed by Ferreira in Ferreira (2006).

The test for symbolic regression was designed as follows:

- 500 uniformly distributed data points were sampled from the $(0,5)$ interval for 3 independent candidate variables, x_1, x_2, x_3;
- 500 data points were generated for the target variable, f based on the functions defined in Eq. 3.3.1-3.3.10 and the candidate variables x_1, x_2, x_3, ;
- 500 data points were sampled and generated for the independent and dependent variables respectively as described above, for cross validation.

For each of the 10 functions in Eq. 3.3.1-3.3.10, 50 independent runs of the GEP and CMAGEP approaches were performed.

Table 3.1: Settings used for all GEP and CMAGEP runs. Parameters only associated with CMAGEP are given in italic.

Parameter	**Artificial benchmark**	**Sunspots**	**Soil Respiration**
Training sample size	500	80	500
Population size	200	100	1000
Number of genes	4	3	2
Head length	6	12	6
Functions	$+,-,/,*,a^x,\sqrt{x},\ln,\exp,\sin,\cos$	$+,-,/,*,\ln,\exp,$	$+,-,/,*,a^x,\sqrt{x},\ln,\exp$
Terminals	x_1,x_2,x_3	$x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8,x_9,x_{10}$	$GPP_{60},T_{air},T_{soil},SWC$
Link function	+	+	+
Max run time	3600 seconds	1800 seconds	150 seconds
Fitness function	MEF	MEF	AIC
Selection method for replication	tournament Coello and Montes (2002)	tournament	tournament
Mutation probability	0.5	0.2	0.2
IS and RIS transpositions probabilities	0.05	0.05	0.05
Inversion probability	0.05	0.05	0.2
One point recombination probability	0.2	0.4	0.3
Two-point recombination probability	0.2	0.3	0.2
Time to start optimization	*After 0 and 50 and 100 generations*	*10 seconds*	*0 seconds*
Number of chromosomes to optimize	*10*	*10*	*5*
Max. iterations of CMA-ES	*50*	*50*	*50*

Table 3.2: GEP and CMAGEP regression performance statistics based on training and validation data for a **prescribed function set without high precision constants**, after 50 independent runs. The table contains values of mean MEF and mean tree size recorded for all 50 runs during training, as well as other performance measures.

Eq.	training MEF	SE MEF training	validation MEF	SE MEF validation	mean Tree size	SE Tree size
	GEP — CMAGEP	GEP — CMAGEP	GEP — CMAGEP	GEP — CMAGEP	GEP — CMAGEP	GEP— CMAGEP
3.3.1	0.95 — **0.97**	0.00 — 0.01	0.95 — **0.96**	0.01 — 0.01	19 — **13**	1 — 1
3.3.2	1.00 — 1.00	0.00 — 0.00	1.00 — 1.00	0.00 — 0.00	29 — **6**	1 — 0
3.3.3	0.78 — **0.89**	0.03 — 0.01	0.71 — **0.90**	0.09 — 0.01	22 —**8**	1 — 0
3.3.4	1.00 — 1.00	0.00 — 0.00	1.00 — 1.00	0.00 — 0.00	16 — **10**	1 — 0
3.3.5	0.71 — **0.73**	0.01 — 0.01	0.71 — **0.72**	0.02 — 0.01	16 — **12**	1 — 1
3.3.6	0.88 — **0.92**	0.01 — 0.01	0.89 — **0.92**	0.01 — 0.01	15 — 15	1 — 1
3.3.7	0.92 — **0.94**	0.01 — 0.01	0.90 — **0.92**	0.02 — 0.01	19 — **14**	1 — 0
3.3.8	0.73 — **0.74**	0.01 — 0.00	0.69 — **0.71**	0.01 — 0.01	28 — **19**	1 — 0
3.3.9	0.80 — **0.81**	0.01 — 0.01	0.81 — **0.82**	0.01 — 0.01	22 — **15**	1 — 1
3.3.10	0.93 — 0.93	0.01 — 0.01	0.92 — **0.94**	0.01 — 0.01	27 — **21**	2 — 1

$$f(x_1) = \frac{10}{5+(x_1-3)^2} \tag{3.3.1}$$

$$f(x_1) = 4.251x_1^2 + 3.26\log(x_1^2) + 7.8e^{x_1} \tag{3.3.2}$$

$$f(x_1) = e^{(-x_1)}x_1^3\cos(x_1)\sin(x_1)(\cos(x_1)\sin(x_1)^2 - 1) \tag{3.3.3}$$

$$f(x_1) = \log(x_1+1) + \log(x_1^2+1) \tag{3.3.4}$$

$$f(x_1,x_2) = \frac{e^{(-(x_1-1)^2)}}{(1.2+(x_2-2.5)^2)} \tag{3.3.5}$$

$$f(x_1,x_2) = 6\sin(x_1)\cos(x_2) \tag{3.3.6}$$

$$f(x_1,x_2) = \frac{1}{(1+x_1^{(-4)})} + \frac{1}{(1+x_2^{(-4)})} \tag{3.3.7}$$

$$f(x_1,x_2) = (x_1-3)(x_2-3) + 2\sin((x_1-4)(x_2-4)) \tag{3.3.8}$$

$$f(x_1,x_2) = \frac{(x_1-3)^4 + (x_2-3)^3 - (x_2-3)}{((x_2-2)^4+10)} \tag{3.3.9}$$

$$f(x_1,x_2,x_3) = \frac{30(x_1-1)(x_3-1)}{x_2^2(x_1-10)} \tag{3.3.10}$$

The settings for GEP and CMAGEP for all runs are reported in Tab. 3.1 and the fitness function used for selection during the evolution process and for the final validation for both GEP and CMAGEP was the Nash-Sutcliffe modelling efficiency (MEF, Eq. 3.3.11) Nash and Sutcliffe (1970).

$$\text{MEF(o,p)} = 1 - \frac{\sum_{i=1}^{n}(o_i - p_i)^2}{\sum_{i=1}^{n}(o_i - \bar{o})^2} \tag{3.3.11}$$

where o and p are the observed and predicted samples, o_i is the observed value at instance i and p_i is the predicted value for instance i and $\bar{o}$ is the mean of observed values. The MEF values are captured in the interval $(-\infty, 1]$, where 1 is reached when the predicted values are equal to the observed.

The regression performance in terms of MEF values after training and validation and solution length recorded in Tab. 3.2, 3.3, and 3.4. These show that for all functions, the CMAGEP mean MEF performance both at training and validation was at least equal or higher than that of GEP (Fig. 3.3, first panel). For all 10 studied equations, the structures generated by CMAGEP were shorter than the structures returned by GEP, with $\approx$ two-thirds fewer parameters.

Fig. 3.3 illustrates the mean MEF values at validation and best solution tree sizes

Table 3.3: Regression performance statistics on **training** set for best GEP and CMAGEP solutions after 50 runs when **the prescribed function set does not contain high precision constants**.

Eq.	GEP train	CMAGEP train	GEP val	CMAGEP val	GEP size	CMAGEP size
3.3.1	1.00	1.00	1.00	1.00	28	**8**
3.3.2	1.00	1.00	1.00	1.00	15	**5**
3.3.3	0.97	**0.99**	0.96	**0.99**	21	**13**
3.3.4	1.00	1.00	1.00	1.00	28	**8**
3.3.5	0.92	**0.97**	0.92	**0.96**	38	**17**
3.3.6	1.00	1.00	1.00	1.00	**5**	6
3.3.7	0.99	**1.00**	0.99	**1.00**	8	**7**
3.3.8	0.81	**0.83**	0.76	**0.79**	25	**17**
3.3.9	0.93	**0.97**	0.93	**0.95**	26	**14**
3.3.10	0.98	1.00	0.98	**1.00**	21	**13**

Table 3.4: Regression performance statistics on **train and validation** set for best GEP and CMAGEP solutions at validation after 50 runs when **the prescribed function set does not contain high precision constants**.

Eq.	GEP train	CMAGEP train	GEP val	CMAGEP val	GEP size	CMAGEP size
3.3.1	1.00	1.00	1.00	1.00	28	**8**
3.3.2	1.00	1.00	1.00	1.00	15	**4**
3.3.3	0.87	**0.99**	0.98	**0.99**	**10**	13
3.3.4	0.99	**1.00**	1.00	1.00	8	8
3.3.5	0.91	**0.97**	0.92	**0.96**	14	**10**
3.3.6	1.00	1.00	1.00	1.00	8	**6**
3.3.7	0.99	**1.00**	0.99	**1.00**	8	**7**
3.3.8	0.81	**0.82**	0.79	**0.81**	42	**22**
3.3.9	0.88	**0.97**	0.95	0.95	16	**14**
3.3.10	0.97	**0.98**	0.98	**1.00**	39	**13**

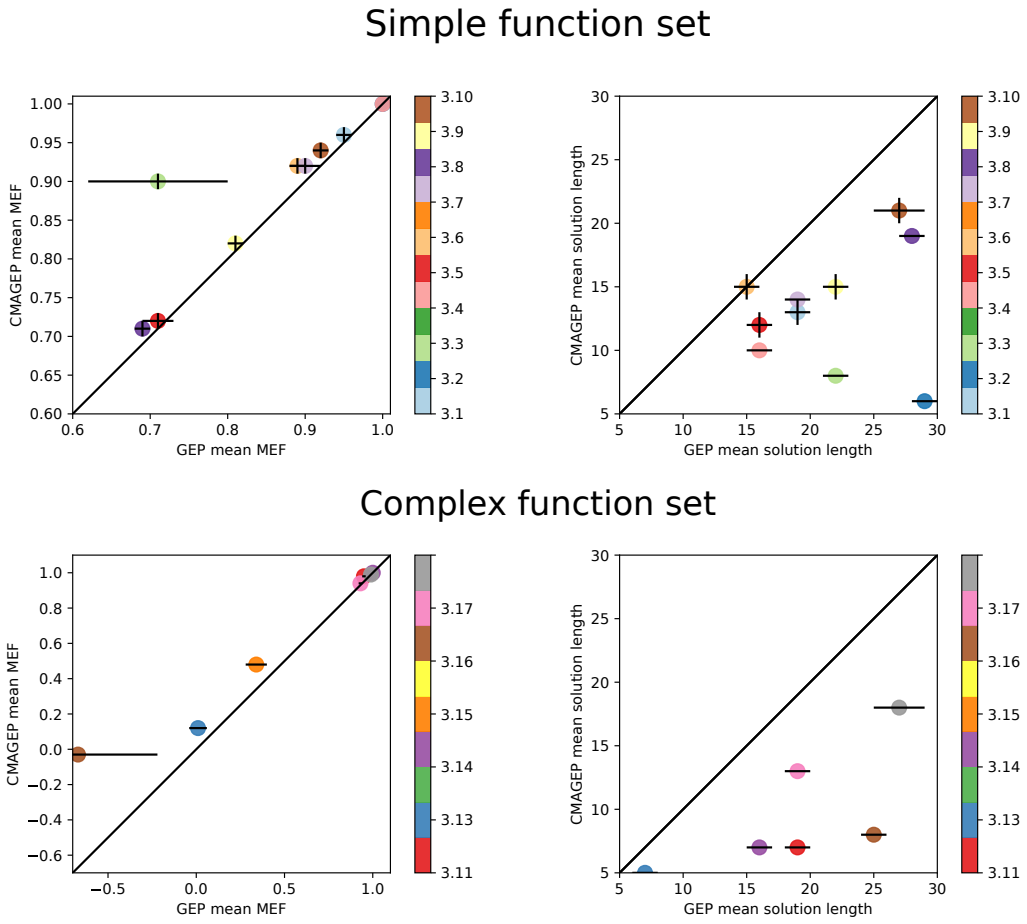


Figure 3.3: GEP vs. CMAGEP regression performance measures on validation data sets for benchmark functions **without (upper panels) and with (lower panels) high precision constants**, after 50 independent runs with settings specified in Table 3.1. Different colours give different equations as described by colour bar.

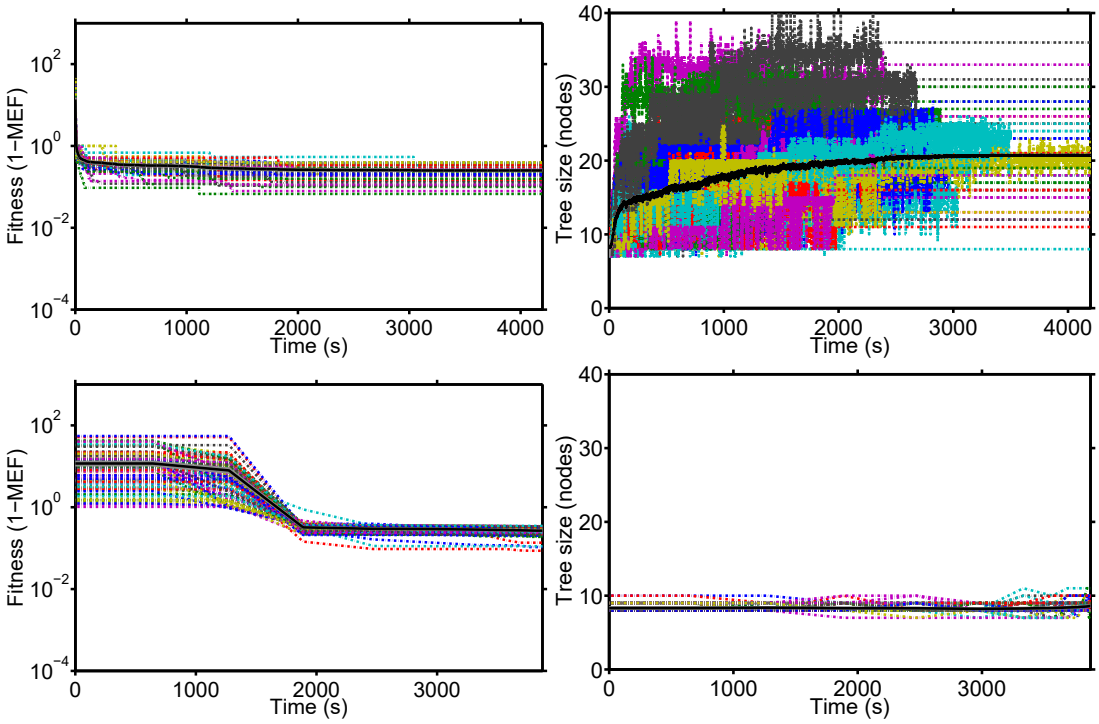


Figure 3.4: **Best GEP (upper panels) and CMAGEP (lower panels) individual-per-run fitness value and solution length (tree size) evolution over runtime**. The evolution is recorded during the training process at 50 independent runs, with each individual run shown in a different colour. Black lines show the mean values. The current panels illustrate the runs on data from prescribed function 3.3.5, lacking high precision constants.

and confirms the improvement in global modelling performance and solution lengths by CMAGEP.

That the MEF validation values were in the same range as the MEF values on the training sets for both approaches, indicates a good capacity of generalization of the resulting regressions.

The evolutions of prediction performance and solution lengths of the best individuals in a generation denoted as tree sizes for 50 runs of both GEP and CMAGEP for equation 3.3.5 from the "simple constants" set are illustrated in Fig. 3.4. The figure shows that for this specific function, the CMAGEP regressions need more time to reach in a similar fitness range with those of GEP, but ultimately reach better fitness scores, likely due to the time spent to retrieve a solution by the CMA-ES optimization. At the same time, the tree sizes of the best solutions of a generation are always smaller for the CMAGEP runs than those of GEP.

The initial set of functions used for generating the regression problems contains very few and simple constant members to be optimized, meaning that the effect of the

CMA-ES optimization might be negligible as the simple constants might have emerge naturally only from the GEP evolution.

To further understand the influence of adding the optimization step in the CMAGEP evolution over the capacity to reconstruct prescribed functions that contain higher precision constants, such constants were added in the Eq. 3.3.1-3.3.10. The new resulting function formulations are given in Eq. 3.3.12-3.3.21 and were the basis for generating new training and validation sets as for the initial function set. The experimental set-up and types of runs presented in section 3.3.1 were repeated for the functions of Eq. 3.3.12-3.3.21 as well.

$$f(x_1) = \frac{10}{5+(3.203x_1-3)^2} \tag{3.3.12}$$

$$f(x_1) = 13.616x_1^2 + 3.26\log(3.203x_1^2) + 7.8e^{3.203x_1} \tag{3.3.13}$$

$$f(x_1) = e^{(-3.203x_1)}3.203x_1^3\cos(3.203x_1)\sin(3.203x_1)(\cos(3.203x_1)\sin(3.203x_1)^2-1) \tag{3.3.14}$$

$$f(x_1) = \log(3.203x_1+1) + \log(3.203x_1^2+1) \tag{3.3.15}$$

$$f(x_1,x_2) = \frac{e^{(-(3.203x_1-1)^2)}}{(1.2+(12.621x_2-2.5)^2)} \tag{3.3.16}$$

$$f(x_1,x_2) = 6\sin(3.203x_1)\cos(12.621x_2) \tag{3.3.17}$$

$$f(x_1,x_2) = \frac{1}{(1+3.203x_1^{(-4)})} + \frac{1}{(1+12.621x_2^{(-4)})} \tag{3.3.18}$$

$$f(x_1,x_2) = (3.203x_1-3)(12.621x_2-3) + 2\sin((3.203x_1-4)(12.621x_2-4)) \tag{3.3.19}$$

$$f(x_1,x_2) = \frac{(3.203x_1-3)^4+(12.621x_2-3)^3-(12.621x_2-3)}{((12.621x_2-2)^4+10)} \tag{3.3.20}$$

$$f(x_1,x_2,x_3) = \frac{30(3.203x_1-1)(0.448x_3-1)}{12.621x_2^2(3.203x_1-10)} \tag{3.3.21}$$

The evolutionary progress is illustrated for Eq. 3.3.18 in the 'complex set" in Fig. 3.5. For this function CMAGEP presents a step behaviour once more, due to the time needed to do the optimizations of the initial evolution step, only that in the case of this function with complex constants, CMAGEP shows better fitness function optimization rate over the entire studied time compared to GEP, with GEP runs displaying local optima behaviour by time 1000 s. The difficulty shown by GEP to reach a good solution is again confirmed by the tree sizes of GEP that are very often ≈ 3 times larger than those needed by CMAGEP to explore the search space.

The regression performance in terms of MEF values after training and validation

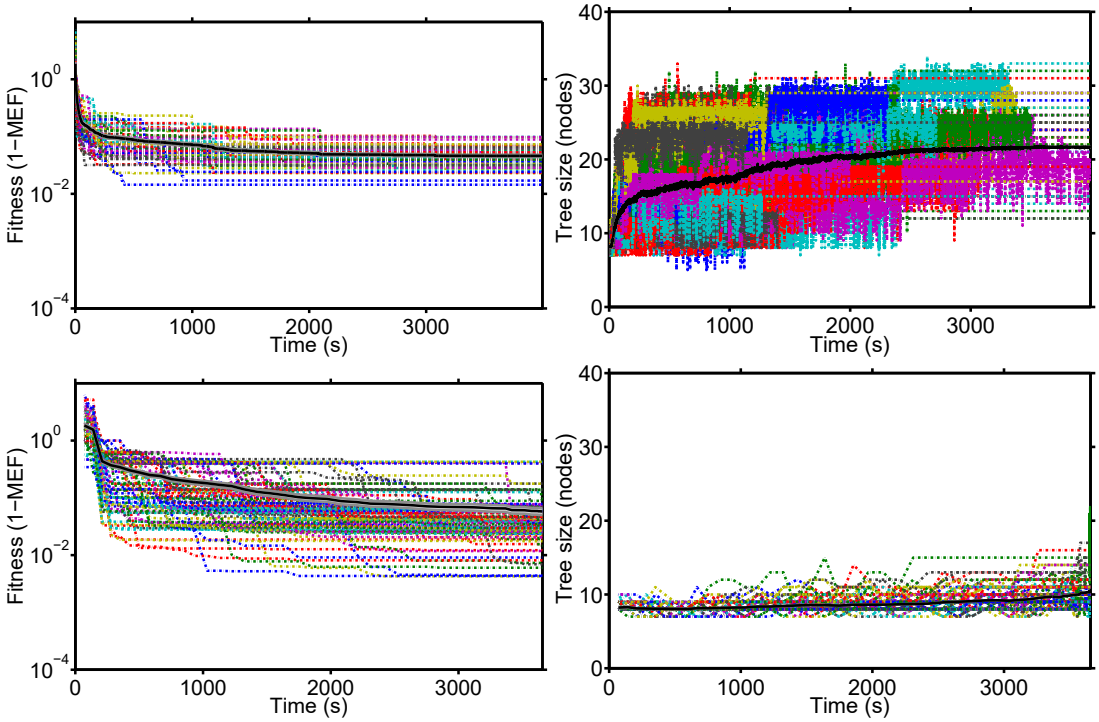


Figure 3.5: **Best GEP (upper panels) and CMAGEP (lower panels) individual-per-run fitness value and solution length (tree size) evolution over runtime**. The evolution is recorded during the training process at 50 independent runs, with each individual run shown in a different colour. Black lines show the mean values. The current panels illustrate the runs on data from prescribed Eq. 3.3.18, containing high precision constants.

Table 3.5: GEP and CMAGEP regression performance statistics based on training and validation data for a **prescribed function set with high precision constants** after 50 independent runs. The table contains values of mean MEF and mean tree size recorded for all 50 runs during training, as well as other performance measures.

Eq.	**training MEF**	**SE MEF training**	**validation MEF**	**SE MEF validation**	**mean Tree size**	**SE Tree size**
	GEP — CMAGEP	**GEP — CMAGEP**	**GEP — CMAGEP**	**GEP — CMAGEP**	**GEP — CMAGEP**	**GEP— CMAGEP**
3.3.12	0.96 —**0.98**	0.00 — 0.00	0.95 — **0.98**	0.01 — 0.0	19 — **7**	1 — 0
3.3.13	1.00 — 1.00	0.00 — 0.00	-187.57 — **0.94**	188.56 — 0.0	25 — **9**	2 — 0
3.3.14	0.08 — **0.21**	0.02 — 0.02	0.01 —**0.12**	0.05 — 0.0	7 — **5**	1 — 0
3.3.15	1.00 — 1.00	0.00 — 0.00	1.00 — 1.00	0.00 — 0.0	16 — **7**	1 — 0
3.3.16	-4.00 — **0.09**	1.66 — 0.02	0.34 — **0.48**	0.06 — 0.0	6 — **4**	1 — 1
3.3.17	0.04 — 0.04	0.00 — 0.01	-0.67 — **-0.03**	0.45 — 0.0	25 — **8**	1 — 0
3.3.18	0.94 — **0.95**	0.00 — 0.01	0.93 — **0.94**	0.01 — 0.0	19 — **13**	1 — 0
3.3.19	0.99 — 0.99	0.00 — 0.01	0.99 — 0.99	0.00 — 0.0	27 — **18**	2 — 0
3.3.20	0.89 — **0.90**	0.01 — 0.02	-8.51 — **0.16**	9.39 — 0.4	22 — **17**	1 — 1
3.3.21	**0.35** — 0.08	0.05 — 0.02	-4278.95 — **0.00**	2674.99 — 0.0	30 — **12**	1 — 0

Table 3.6: Regression performance statistics on **training** set for best GEP and CMAGEP solutions out of 50 runs **with high precision constants in the prescribed function set**.

Eq.	GEP train	CMAGEP train	GEP val	CMAGEP val	GEP size	CMAGEP size
3.3.12	0.99	**1.00**	0.99	**1.00**	22	**8**
3.3.13	1.00	1.00	1.00	1.00	41	**3**
3.3.14	0.85	**0.86**	0.87	0.87	42	**7**
3.3.15	1.00	1.00	1.00	**1.00**	8	**6**
3.3.16	0.99	**1.00**	0.95	**0.99**	26	**14**
3.3.17	0.09	**0.51**	-0.09	**0.44**	41	**6**
3.3.18	0.99	**1.00**	0.98	**1.00**	33	**8**
3.3.19	1.00	1.00	1.00	1.00	34	**7**
3.3.20	0.99	**1.00**	0.99	**1.00**	34	**5**
3.3.21	**0.94**	0.85	-31.88	**0.05**	41	**12**

and solution length were studied for the two GEP approaches over the modified benchmark set (functions 3.3.12- 3.3.21) where constants need higher precision approximation and have longer symbolic representation. The results are reported in Tab. 3.5, 3.6, and 3.7. For all functions, the CMAGEP mean MEF performance both at training and validation was at least equal or higher than that of GEP (Fig. 3.3, third panel). For all 10 studied equations, the structures generated by CMAGEP were once again much shorter than the structures returned by GEP, with at least 2 times fewer parameters. Both GEP and CMAGEP show a significantly higher MEF value on the training set than on the validation set for equation 3.3.21, which can indicate a case of over-fitting, likely due to the extra time needed to capture the high precision constants.

Table 3.7: Regression performance statistics on **validation** set for best GEP and CMAGEP solutions out of 50 runs **with high precision constants in the prescribed function set**.

Eq.	GEP train	CMAGEP train	GEP val	CMAGEP val	GEP size	CMAGEP size
3.3.12	0.99	1.00	0.99	**1.00**	22	**8**
3.3.13	1.00	1.00	1.00	1.00	12	**4**
3.3.14	0.85	0.86	0.87	0.87	42	**7**
3.3.15	1.00	1.00	1.00	1.00	13	**7**
3.3.16	0.98	0.99	0.99	0.99	25	**15**
3.3.17	0.06	0.51	-0.02	**0.44**	25	**6**
3.3.18	0.99	1.00	0.98	**1.00**	33	**8**
3.3.19	1.00	1.00	1.00	1.00	34	**16**
3.3.20	0.97	1.00	0.99	**1.00**	9	**5**
3.3.21	0.12	0.73	0.02	**0.31**	22	**15**

3.3.2 Best starting point for the CMA-ES optimization

Since the number of generations allowed to pass before the CMA-ES optimization starts is a newly introduced CMAGEP parameter that could influence the selection process and that needed further study, the following experiment was devised: the samples generated for functions 3.3.5 and 3.3.18 were used for training and validation of 20 independent runs of CMAGEP with the CMA-ES optimization starting at different generations: 0, 50, 100, 500, 1000.

For understanding if the CMA-ES optimization is actually a part of the learning process during the CMAGEP evolution or, if it would actually be sufficient to only optimize the parameters of the best individual from the last generation, one control case is added to the test, where the entire evolution is only done by GEP and only the last generation of individuals is optimized. The CMAGEP settings for all runs remain unchanged from Tab. 3.1.

The influence of the CMA-ES start point in the two cases of "simple" and "complex" functions sets can be seen in Fig. 3.6 and Fig. 3.7. These showed that although in training it seems that starting the optimization later in the evolution process, the trend is no longer clear on the validation set where all starting points are in a similar prediction performance range, with 0 time showing a slightly higher mean value. More importantly though, it was clear that for the mean number of parameters, the lower the generation start for optimization, the lower the complexity of structures, as seen in the third panels of Fig. 3.6 and Fig. 3.7.

3.3.3 Comparing with other machine learning approaches

To give a good perspective of GEP approaches prediction capacity in the context of other machine learning methods (MLM), a set of four well known approaches such as ANN Yegnanarayana (2009), SVM Hearst (1998) , RF Breiman (2001) and KRR Hoerl and Kennard (1970) were used for generating predictions. The predictions and the performance measures were computed based on artificial data that resulted from the benchmark function set .

The toolboxes and settings used for the predictions of the ANN and KRR methods are described in Tramontana et al. (2016) and are implemented in the "simpleR" regression toolbox Lazaro-Gredilla et al. (2014). The SVM predictions were obtained based on the "LIBSVM" library Chang and Lin (2011) in the "simpleR" toolbox run with default settings, . Lastly, the RF predictions were generated with standard the MATLAB statistics toolbox and default parameter values.

50 sets of predictions were generated by each of the above mentioned methods, after using the same training set presented in the first part of section 3.1. The prediction values are then compared to the original values and to the prediction values generated

by the GEP and CMAGEP structures by means of average MEF.

After the 50 independent runs (see Fig. 3.8) on the "simple" function set of the ANN, KRR, SVM and RF respectively, the distributions of mean MEF values over the 50 validation cases show that the prediction performances of ANN,SVM, and KRR are in a close range or each other, being closely followed by the performances of CMAGEP and GEP and RF.

When learning from samples generated by the function set that contain high precision constants (see Fig. 3.9, an improvement in the general ranking of the GEP based methods was noticeable. This was especially clear in the case of functions 3.16, 3.17 and 3.21 (Fig. 3.12 and 3.13), where is seems that the regression problems have become more difficult to solve than their "simple" counterparts, due to the flattening of some surfaces and inclusion of spikes in the function shape over the studied intervals. In such cases it can be that the internal structure of the GEP individuals that might contain introns is helpful in exploring a possibly rugged fitness landscape more efficiently Ferreira (2006) than for non-GP machine learning approaches.

The highest prediction performance, in terms of mean MEF at validation over 50 runs, was associated with the ANN approach over all studied functions, even when they contained or not high precision constants. At the same time, across all the studied functions, the RF seems to show the prediction performance in the lower range. Nevertheless, the lower prediction performance recorded for the current study does not mean that RF is a less powerful approach, but the lower performance can be attributed to the size of the current chosen learning sample, evidence from other studies showing that the RF approach tends to perform better with larger learning sample sizes Elith et al. (2008).

The functions for which GEP and CMAGEP had a visibly lower average prediction performance that of the non-GP MLM, were Eq. 3.3 and 3.14, which is the complex version of 3.3, with the gap being reduced for 3.14. All the approaches showed the lowest prediction performance for Eq. 3.21, which presents very abrupt changes in convexity in both directions, possibly making the search space too difficult to cover.

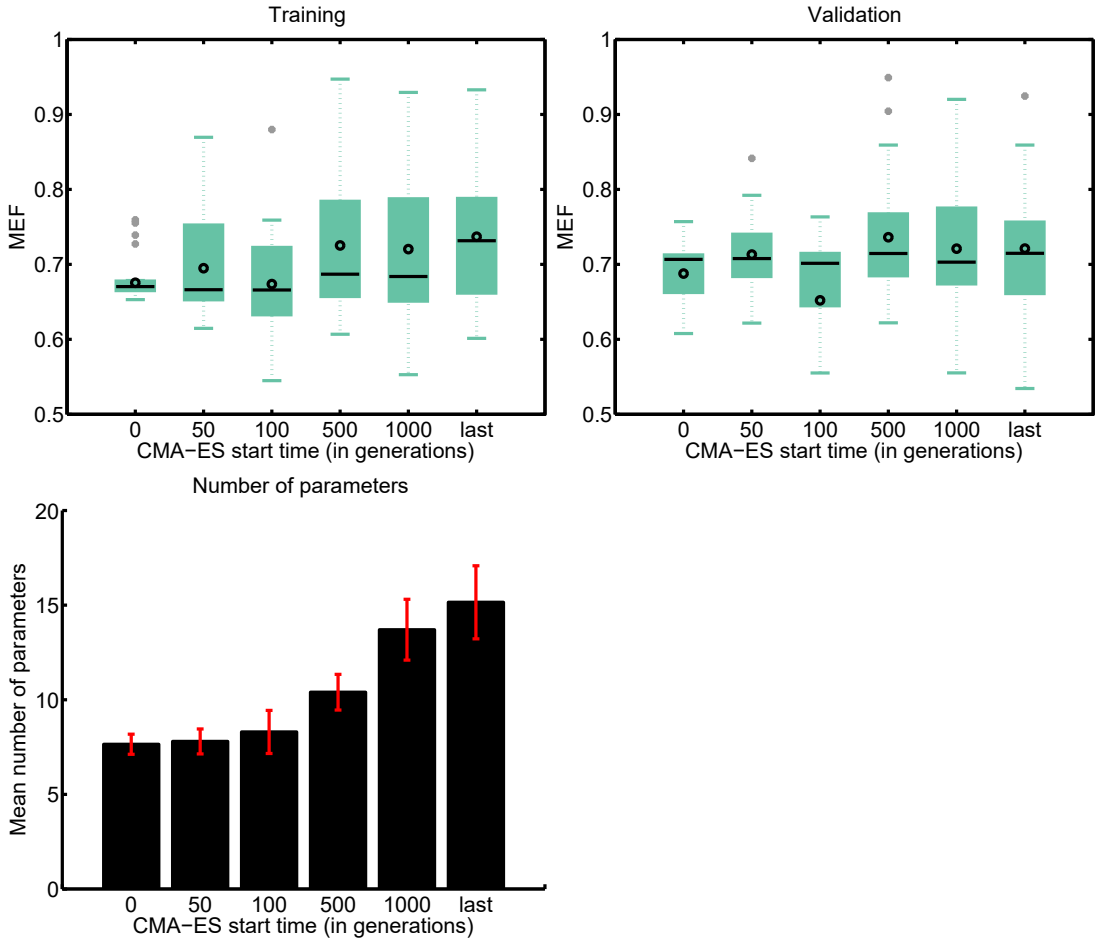


Figure 3.6: **CMAGEP distribution of MEF values and number of parameters** for all solutions based on training and validation data from prescribed function 3.3.5, lacking high precision constants.Values are reported after 20 independent runs based on settings given in Table 3.1 with the CMA-ES optimization starting at different times. The different starting times are given in generations and are shown on the x axis for all panels.

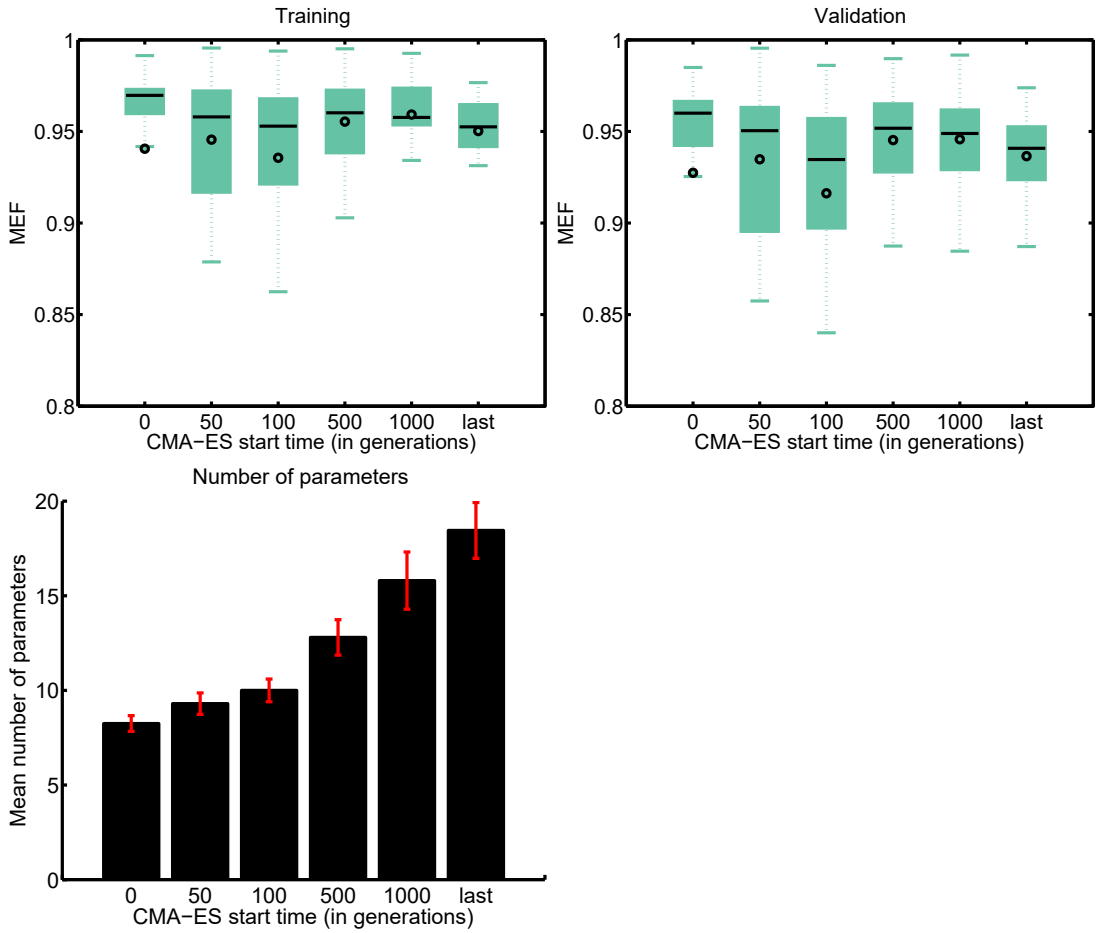


Figure 3.7: **CMAGEP distribution of MEF values and number of parameters** for all solutions based on training and validation data from prescribed function 3.3.18, lacking high precision constants.Values are reported after 20 independent runs based on settings given in Table 3.1 with the CMA-ES optimization starting at different times. The different starting times are given in generations and are shown on the x axis for all panels.

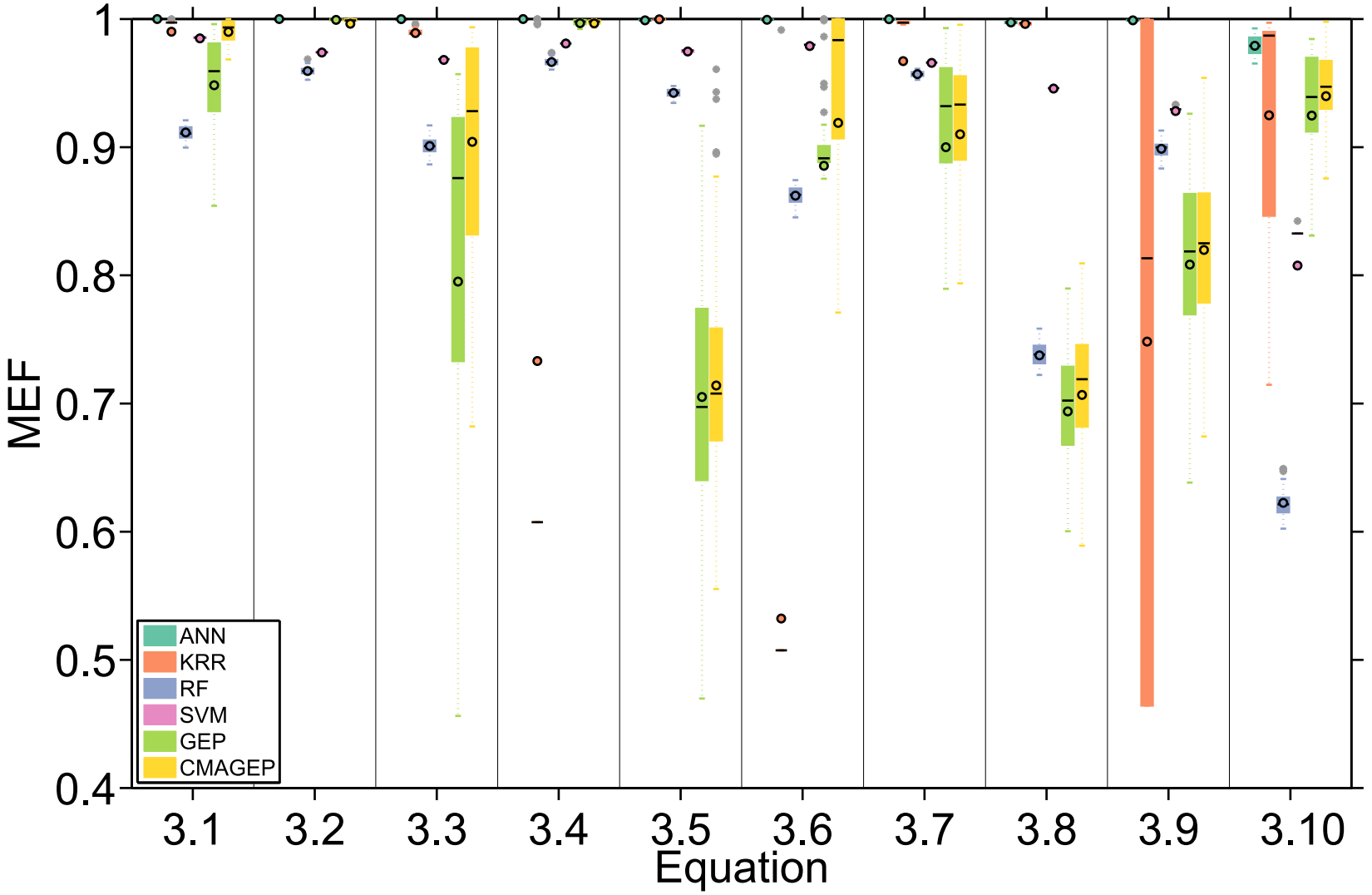


Figure 3.8: Cross validated prediction performance as mean MEF for several machine learning methods (MLM), such as ANN, KRR, RF, SVM, GEP, and CMAGEP after 50 independent runs for a benchmark function set lacking high precision constants.

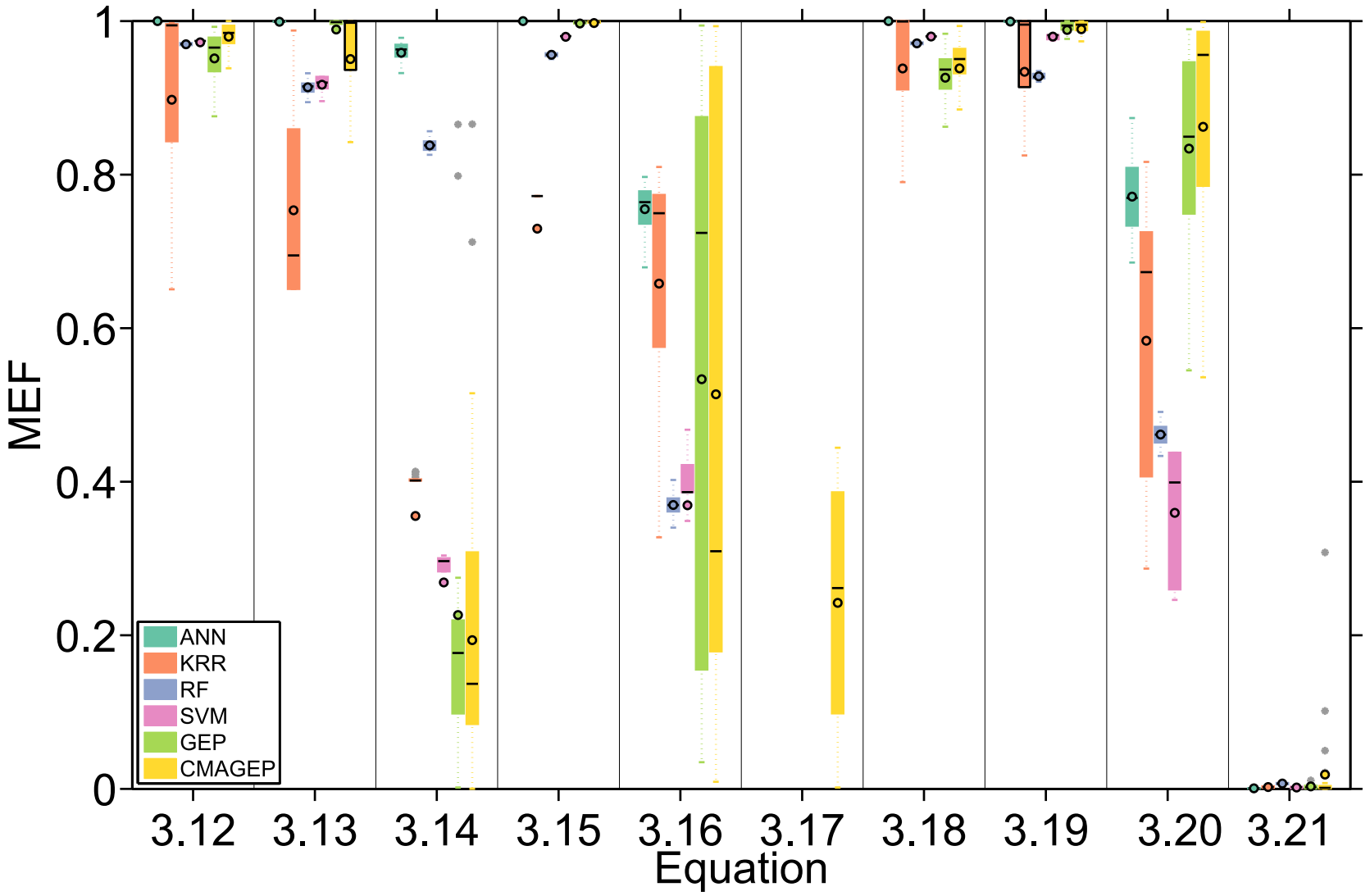


Figure 3.9: Cross validated prediction performance as mean MEF for several machine learning methods (MLM), such as ANN, KRR, RF, SVM, GEP, and CMAGEP after 50 independent runs for a benchmark function set containing high precision constants.

3.3.4 Sunspots and comparing with commercial GEP

With the purpose of studying the prediction performances of CMAGEP against the standard GEP, on real data, in the framework of a well studied problem in the evolutionary algorithm community, the Sunspots data set was used Ferreira (2006). The data set contains 100 observations from the Wolfer time series recorded between 1700 and 1988.

From the first 80 time steps, 10 independent variables are created, with a time lag of one, i.e $t_{-1}, t_{-2} ... t_{-10}$ and a dependent variable t_0. Thus, 80 sets of dependent and independent variables are generated. On this dataset, 100 runs were performed with the settings from 3.1, under the Sunspots column, for the standard GEP and the CMAGEP. The rest of 20 time steps are kept to use for validation.

The standard GEP performance was assessed using my implementation of the GEP algorithm and a commercial implementation by the author of the algorithm, Ferreira. The commercial tool used is the demo version of the "GeneXproTools" (version 5.0.39.02 available at http://www.gepsoft.com/), owned by Gepsoft Limited.

For checking if my implementation of GEP is correct, its performance is compared with a commercially available implementation by GEPSOFT (demo version). Using a standard data set (sunspots) provided in the commercial demo version, it was found that the performances are in a similar range (Tab. 3.8), with a mean MEF value of 0.79 ($\pm$ 0.001 for commercial GEP and $\pm$ 0.003 for my implementation) during learning and a slightly better performance of my GEP implementation during validation of 0.37 ($\pm$ 0.08) compared to 0.29 ($\pm$ 0.007) for the commercial GEP.

When applying the CMAGEP approach on the same data, CMAGEP obtained the highest results with a mean MEF value at training of 0.87 ($\pm$ 0.002) and 0.76 ($\pm$ 0.004) at validation (Tab. 3.8, third column). Furthermore, the tree size of the best individual of all runs on the validation set is much smaller for the CMAGEP approach, solution lengths being approx. half size.

3.3.5 Real observations for soil respiration

Lastly, for assessing the prediction and learning capacities of the two presented approaches, in the case of real ecological observations, a dataset containing 613 measurements of soil respiration (R_{soil}), SSA smoothed with a window of 60 days terrestrial ecosystem gross primary production (GPP_s), air temperature(T_{air}), soil temperature(T_{soil}) and soil water content(SWC) recorded daily was used for automatically constructing a symbolic regression model. In this problem, a functional expression was needed for describing the response of R_{soil} to the candidate drivers, GPP_s, T_{air}, T_{soil} and SWC. The 1 year measurements for all studied fluxes are illustrated in Fig. 3.10.

The measurements were recorded at a single site as detailed in Ilie et al..

Table 3.8: GEP and CMAGEP regression performance statistics on the **Sunspots data set** after 100 independent runs.

Algorithm	commercial GEP	my GEP implem.	CMAGEP
Mean MEF/run train	0.79	0.79	**0.87**
SE MEF/run train	0.001	0.003	0.002
Best MEF train	0.84	0.88	**0.91**
Mean MEF/run validation	0.29	0.37	**0.77**
SE MEF/run validation	0.007	0.08	0.004
Best MEF validation	0.75	0.77	**0.82**
Mean tree size	14	11	**6**
SE tree size	0.61	0.61	0.12
Best train indiv. tree size	21	24	**10**
Best validation indiv. tree size	43	21	**6**
Mean number of generations	15000	14642	41

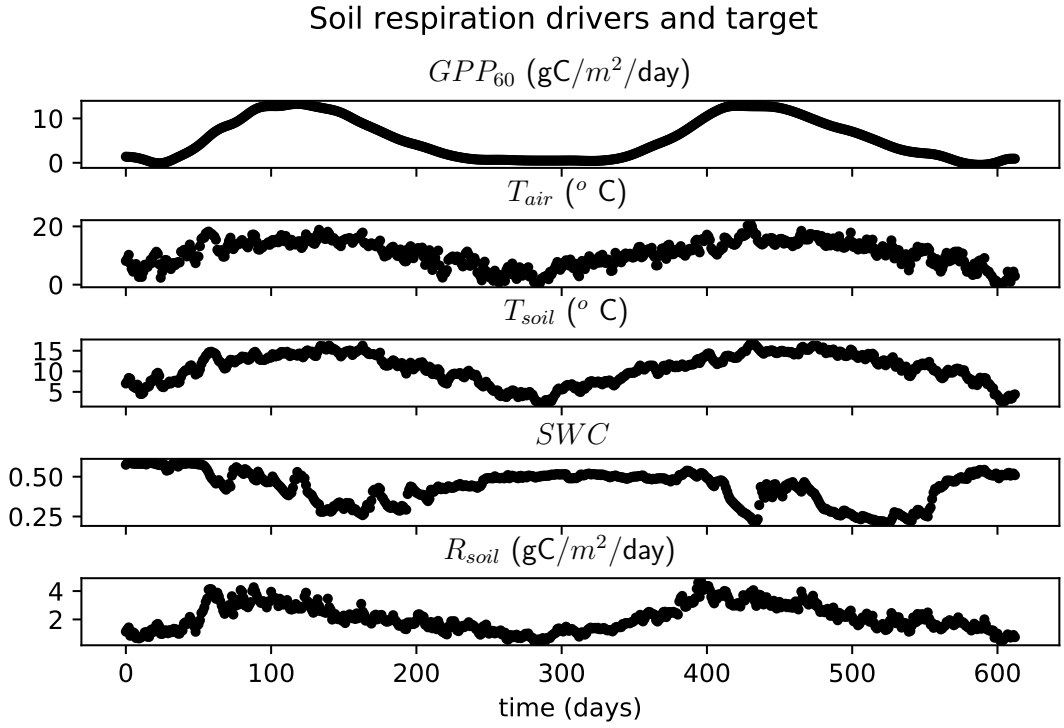


Figure 3.10: Candidate drivers and target variable as time series that were given as input to GEP and CMAGEP runs in real observations experiment describing soil respiration dynamics. The soil respiration flux is given in units of gCO_2/30 min/m^2.

Table 3.9: GEP and CMAGEP regression performance statistics on a data set containing **real measurements of soil respiration** after 20 independent runs with the settings given in Tab. 3.1.

Algorithm	**GEP**	**CMAGEP**
Average MEF best/run training	0.41	**0.74**
SE MEF best/run training	0.04	0.01
Best MEF training	0.84	**0.86**
Average MEF best/run validation	0.09	**0.59**
SE MEF best/run validation	0.18	0.01
Best MEF validation	0.75	**0.77**
Average tree size of best/run	5.5	**3.75**
SE tree size of best/run	0.44	0.28
Best training indiv. tree size	7	**5**
Best validation indiv. tree size	7	**5**

From the total set of data points, 500 data points were randomly sampled and was used for training GEP and CMAGEP models. The train runs were repeated 20 times the settings found in column 3 of Tab. 3.1. The remaining 113 data points left from each sampling were used for computing the cross validated fitness function values for the solutions returned by the GEP and CMAGEP approaches. For cross validation, the GEP and CMAGEP solutions are used to generate predictions for all the test cases and mean fitness function values are computed over the 10 runs.

Finally, the functional expressions of the best solutions for each of the two studied approaches are reported.

After applying the two GEP methods to this real observations dataset, an improved mean performance in modelling efficiency could be observed for the CMAGEP solutions on the training and validation sets with MEF $= 0.74 \pm 0.01$ for CMAGEP and MEF $= 0.41 \pm 0.04$ for GEP. More importantly, the average lengths of the returned solutions were reduced by the CMAGEP approach, with returned solutions having an average of 3.75 ± 0.2 parameters compared to an average of 5.5 ± 0.4 parameters for the GEP solutions. For a more detailed comparison see Tab. 3.9).

The difference in the CMAGEP and GEP modelling is illustrated by the mathematical expression of the final models selected from each of the 20 GEP and CMAGEP returned solutions. The final model selection was done based on mean MEF values at validation.

The two models are given in equations 3.3.22 and 3.3.23.

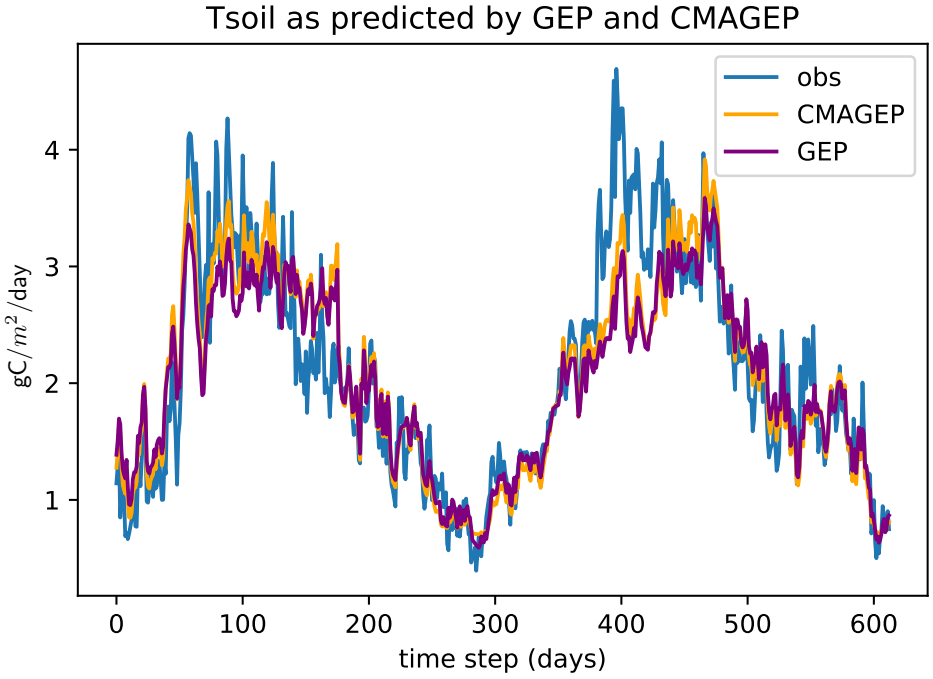


Figure 3.11: **Observed and GEP and CMAGEP predicted soil respiration fluxes.** The observed T_{soil} fluxes are shown as time series. The predicted values are obtained from the GEP and CMAGEP models given in Eq. 3.3.22 and 3.3.23. The models were selected according to fitness after 10 independent GEP and CMAGEP runs. The GEP and CMAGEP runs were performed with the settings given in column 3 of Table 3.1.

$$R_{soil(\text{GEP})} = \exp(SWC + (T_{soil} + SWC + 5.0)^{0.5} - 3.8) \tag{3.3.22}$$

$$R_{soil(\text{CMAGEP})} = \exp(4.7 \times (0.2SWC)^{\frac{6.3}{T_{soil}}} - 0.4) \tag{3.3.23}$$

Although different structurally, the two GEP and CMAGEP models reproduced a similar pattern as seen in Fig. 3.11, where the CMAGEP and GEP predicted soil respiration fluxes were overlaid onto the the real soil respiration flux values measured at the study site. From the fit perspective, the CMAGEP model seems to be better at capturing the higher peaks compared to the GEP model.

3.4 Discussion

3.4.1 GEP benchmark on artificial test functions

In the studies performed on artificially generated data CMAGEP fared better than the GEP approach for reconstructing compact symbolic regressions. One reason for the better results in prediction and learning, as well as smaller complexity of solutions could be that the CMA-ES is efficient in optimizing the parameters of the GEP generated functions, that leads to less need for expansion of structures along the generation count, also known as bloat. This indicates that the small pushes brought by the local parameter optimization can in fact improve the search direction for the genetic evolution of the regressions.

For the majority of test cases, a decrease in the prediction performance was noticeable when the target functions contain complex constants (Eq. 3.3.12-3.3.21). It is possible that by adding the constants in the function formulations, the search space became more complex, and the two approaches may have run into local optima problems, as seen also in Fig. 3.5, where a loss in population diversity is noticeable, and not much fitness is gained after time 2000 for GEP and time 3000 for CMAGEP.

For difficult problems such as Eq. 3.3.14, where the best over-all solution had an MEF of 0.86 and the lowest was negative, with the mean MEF < 0.5 over all validation cases, a multiple start is recommended, or multiple runs approach for GEP and CMAGEP as the range of the solutions can be extended depending on the search stating point, and in order to avoid chance-only driven results, sufficiently large number of runs should be performed.

3.4.2 Best starting point for the CMA-ES optimization

When the effect of the starting point of the CMA-ES optimization was studied, is was apparent that the best time to start the optimization, in terms of MEF performance depends on the type of problem studied, but it seems that having a waiting time for the CMA-ES to start is helpful for the overall performance. However, when the complexity of the solutions is taken into account, it is clear that starting to optimize at the beginning of the learning process, will return solutions that are more compact, without a significant decrease in prediction performance. This naturally brings the known trade-off between prediction performance and model interpretability, and the start time of the optimization should be decided based on the needs of the user.

3.4.3 Comparing with other machine learning approaches

As seen in the current artificial data based study, the MEF values of the best of run solution returned by the GEP based approaches generally spread over a larger range compared to non-GEP MLM, keeping in line with the known stochastic component of multiple run GP approaches. However, the solutions of interest in the GEP approaches for symbolic regression are the solutions that perform best over all runs, and these always perform in a small range of the best solutions of most of the studied non-GP MLM.

That the prediction performance of the solutions returned by the two GEP methods in the context other machine learning methods is very often in the same range can make the results of GEP and CMAGEP more reliable for symbolic regression.

Although the results of GEP and CMAGEP are encouraging for prediction, in the context other MLM, it is important to keep in mind that the comparison done in this study was not completely fair and was done more as a starting guide, since the parameters used for obtaining the KRR, RF, SVN and ANNs predictions would probably be better tuned by experts in the fields, whereas the set-up for the non-GEP runs was done based on default values, the author's experience being in favour of the GEP parameter set tuning.

Nevertheless, the most important aspect of the GEP based approaches for symbolic regression remains the fact that they are capable of building a "readable", and, in the case of CMAGEP , simple mathematical expression that can be further analysed, interpreted and plugged in other systems.

3.4.4 Sunspots and comparing with commercial GEP

In the Sunspots experiment, the commercial implementation of GEP was compared my implementation of standard GEP and the newly proposed CMAGEP. It was interesting to see that the resulting solutions of the commercial implementation and the current implementation were so similar in prediction performance and length for this problem. This indicates that the results generated by my GEP version are not significantly influenced by the implementation environment and portray a good image of the GEP standard performance. The sunspots experiment confirms once more the superiority of the CMAGEP in modelling performance over the standard GEP. The shortening of solution lengths by CMAGEP although there was no explicit parsimony pressure added in the fitness function during the evolution of either CMAGEP or GEP, points to an indirect parsimony pressure being exercised in the CMAGEP. The pressure could appear due the intrinsic property of CMA-ES to give faster solutions to shorter GEP individuals combined with the presence of the CMA-ES time-out parameter.

3.4.5 Real observations for soil respiration

With the equations obtained in section 3.3 it seems that the CMAGEP approach is in line with what GEP can automatically learn for soil respiration and its drivers, from a prediction capacity perspective, however the model structure obtained is less complex, allowing more space for interpretation.

It was interesting to see that although four candidate drivers were given as input, including air and soil temperature and soil water content, the structures that best describes the process in soil respiration chosen from the performed runs is only dependent on GPP, temperature and soil water content. The question that can be raised now is whether GPP soil temperature and soil water content are really the main drivers of soil respiration, or whether the structure describes dependencies on other drivers, but not explicitly.

In Fig. 3.11 both obtained models describe sufficiently well the observed soil respiration flux, except for the higher values of the second section of the studied time, where an underestimation is present. Same patterns were seen in Ilie et al. (2017).

In this work, the experiment on real observations was conducted mainly with the purpose of comparing the two presented approaches and only illustrating the selected model structures, and not for an in-depth analysis of the resulted structures. Such a deeper analysis of the relevance and the applicability of the models and other aspects from the modelling standpoint are further discussed in the author's other work in Ilie et al. (2017).

The fact that both models described a similar flux, even if the structures and internal dynamics are different might indicate that by combining the CMA-ES optimization with GEP, the evolution process is not highly disturbed, but that with the parameter scaling in candidate models a solution is reached faster from a generation number point of view. By managing to reach a final solution faster, in CMAGEP the risk of bloat, commonly observed in genetic programming type of approaches Banzhaf and Langdon (2002); Langdon (2000); Smith (2000) is reduced.

3.5 Conclusion and outlook

Considering the results presented in this chapter, it can be concluded that by adding optimization steps during the evolution of the GEP, the global modelling performance can be improved, but more importantly the returned solutions become more parsimonious and easier for the user to understand. This implies that in real world problems CMAGEP can provide more help to the community, especially if the symbolic regression solutions are needed for further understanding of a system. The CMAGEP approach is recommended for generating simple and concise model structures,for prob-

lems in which the main interest would not be only prediction accuracy also the possibility to interpret.

It would be important for the future development of the approach to investigate whether diversity management would improve the overall learning performance by keeping the GEP generated structures from converging too fast and allowing on the other hand for a larger pool of possible functional structures to pass through the optimization phase.

Another aspect that would need further investigation is a measure of complexity of the returned functions, however not in terms of number of parameters to optimize, but in terms of the degreed of non-linearity in the functions that compose the final solutions as seen in the work of Vladislavleva et al. (2009).

3.6 Algorithms

CMA-ES algorithm

Input

Set default values for $\mu, \lambda, \omega_{i=1\ldots\mu}, c_\sigma, d_\sigma, c_c, c_1$ and c_μ as given in frame 3.6.

Initialization $\mathbf{C} = \mathbf{I}, g = 0, p_c = 0$ and $p_\sigma = 0$ $m \in \mathbb{R}^n$ and $\sigma \in \mathbb{R}_+$

While termination criteria has not been met:

$$g \leftarrow g + 1 \tag{3.6.1}$$

Sampling new population of search points for $k = 1, ..\lambda$:

$$\mathbf{z}_k \sim \mathcal{N}_k(0, \mathbf{I}) \tag{3.6.2}$$

$$\mathbf{y}_k = \mathbf{BDz}_k \sim \mathcal{N}(0, \mathbf{C}) \tag{3.6.3}$$

$$\mathbf{x}_k = m + \sigma y_k \sim \mathcal{N}(m, \sigma^2 \mathbf{C}) \tag{3.6.4}$$

Selection and recombination:

$$\langle \mathbf{y} \rangle_w = \sum_{i=1}^{\mu} \omega_i y_{i:\lambda}, \text{ where} \sum_{i=1}^{\mu} \omega_i = 1, \ \omega_i > 0 \tag{3.6.5}$$

Mean update:

$$m \leftarrow m + \sigma \langle y \rangle_w = \sum_{i=1}^{\mu} \omega_i x_{i:\lambda} \tag{3.6.6}$$

Step size control:

$$p_\sigma \leftarrow (1 - c_\sigma) p_\sigma + \sqrt{c_\sigma (2 - c_\sigma) \mu_{\text{eff}}} \mathbf{C}^{-\frac{1}{2}} \langle \mathbf{y} \rangle_w \tag{3.6.7}$$

$$\sigma \leftarrow \sigma \times e^{\left(\frac{c_\sigma}{d_\sigma}\left(\frac{\|p_\sigma\|}{\mathrm{E}\|\mathcal{N}(0,\mathbf{I})\|}-1\right)\right)} \tag{3.6.8}$$

$$\mathrm{E}\parallel \mathcal{N}(0,\mathbf{I}) \parallel = \sqrt{2}\Gamma\left(\tfrac{n+1}{2}\right) \approx \sqrt{n}\left(1-\tfrac{1}{4n}+\tfrac{1}{21n^2}\right)$$

Covariance matrix adaptation:

Cumulation for **C**:

$$p_c \leftarrow (1-c_c)p_c + h_\sigma \sqrt{c_c(2-c_c)\mu_{\mathrm{eff}}}\langle \mathbf{y}\rangle_w \tag{3.6.9}$$

$$h_\sigma = \begin{cases} 1 & \text{if } \frac{\|p_\sigma\|}{\sqrt{1-(1-c_\sigma)^{2(g+1)}}} < \left(1.4+\frac{2}{n+1}\right)\mathrm{E}\parallel \mathcal{N}(0,\mathbf{I}) \parallel \\ 0 & \text{otherwise} \end{cases}$$

C update:

$$\mathbf{C} \leftarrow (1-c_1-c_\mu)\mathbf{C} + c_1(p_c p_c^T + \delta(h_\sigma)\mathbf{C}) + c_\mu \sum_{i=1}^{\mu} \omega_i y_{i:\lambda} y_{i:\lambda}^T \tag{3.6.10}$$

with, $\delta(h_\sigma) = c_c(1-h_\sigma)(2-c_c) \leqslant 1$

Default strategy parameters

$$\mu_{\text{eff}} = \frac{1}{\sum_{i=1}^{\mu} \omega_i^2} \geqslant 1, \text{ where } \sum_{i=1}^{\mu} \omega_i = 1 \tag{3.6.11}$$

Selection and recombination:

$$\lambda = 4 + \lfloor 3\ln(n) \rfloor, \ \mu = \lfloor \mu' \rfloor, \ \mu' = \frac{\lambda}{2} \tag{3.6.12}$$

$$\omega_i = \frac{\omega_i'}{\sum_{j=1}^{\mu} \omega_j'}, \ \omega_i' = \ln(\mu' + 0.5) + \ln(i), \ i = 1...\mu \tag{3.6.13}$$

Step size control:

$$c_\sigma = \frac{\mu_{\text{eff}} + 2}{n + \mu_{\text{eff}} + 5} \tag{3.6.14}$$

$$d_\sigma = 1 + 2\max\left(0, \sqrt{\frac{\mu_{\text{eff}} - 1}{n+1}} - 1\right) + c_\sigma \tag{3.6.15}$$

Covariance matrix adaptation:

$$c_c = \frac{\frac{\mu_{\text{eff}}}{n} + 4}{n + \frac{2\mu_{\text{eff}}}{n} + 4} \tag{3.6.16}$$

$$c_1 = \frac{2}{(n+1.3)^2 + \mu_{\text{eff}}} \tag{3.6.17}$$

$$c_\mu = \min\left(1 - c_1, \alpha_\mu \frac{\mu_{\text{eff}} - 2 + \frac{1}{\mu_{\text{eff}}}}{(n+2)^2 + \alpha_\mu \frac{\mu_{\text{eff}}}{2}}\right), \ \alpha_\mu = 2 \tag{3.6.18}$$

CMA-ES definitions

λ population size

μ number of selected search points in the population

μ_{eff} is the variance effective selection mass, with $1 \leqslant \mu_{\text{eff}} \leqslant \mu$

$\sigma \in \mathbb{R}_+$, step size

$\omega \in \mathbb{R}_+$, weight coefficients for recombination

$c_c \leqslant 1$, learning rate for cumulation for the rank-one update of the covariance matrix

$c_1 \leqslant 1 - c_\mu$, learning rate for the rank-one update of the covariance matrix update

$c_\mu \leqslant 1 - c_1$, learning rate for the rank-μ update of the covariance matrix update

$c_\sigma < 1$, learning rate for the cumulation for the step-size control

$d_\sigma \approx$ damping parameter for step-size update

$f : \mathbb{R}^n \to \mathbb{R},\ x \mapsto f(x)$, fitness function to minimize

g, generation number

$m \in \mathbb{R}^n$, mean value of the search distribution

$n \in \mathbb{N}$, search space dimension

$\mathbf{I} \in \mathbb{R}^{n \times n}$, $\mathbf{C}^{(g)}$, identity matrix and covariance matrix at generation g

$\mathbf{B}, \mathbf{D}$ come from an eigen decomposition of the covariance matrix $\mathbf{C}$ with $\mathbf{C} = \mathbf{B}\mathbf{D}^2\mathbf{B}^T = \mathbf{B}\mathbf{D}\mathbf{D}\mathbf{B}^T$. $\mathbf{B}$ is an orthogonal matrix with eigenvectors of $\mathbf{C}$ as columns and $\mathbf{D}$ a diagonal matrix with square roots of eigenvalues of $\mathbf{C}$ as diagonal elements

$x_k \in \mathbb{R}^n,\ k = 1...\lambda$ a sample of λ search points.

$\langle \mathbf{y} \rangle_w$ is a step of the distribution mean irrespective of step-size σ.

$x_{i:\lambda} \in \mathbb{R}^n$, the i-th best point of $x_1 ... x_\lambda$, with respect to fitness function f.

$y_k \sim \mathcal{N}_k(0, \mathbf{C})$, for $k = 1...\lambda$, are samples from a multivariate normal distribution with zero mean and covariance matrix $\mathbf{C}$;

CMAGEP algorithm

Start:

initialize the global CMAGEP parameters:
$n, g_n, d_s, f_s, t_s, c_s, f_l, l_{gh}, f, m_r, t_r, i_r, r_r, m_{gen}, m_{rt}, p_o, g_o, i_o, fe_o$
generate initial Ω_1 population of n individuals made of g_n number of genes
:
$i = 0, k = 0$
while $k <= n$**, then:**

$r = 0$

while $r < g_n$**, then**:
for $(\alpha = 1, \alpha <= l_{gh}, \alpha++)$
$g_{i,k,r}[\alpha]$=random selection from $f_s \cup t_s$
endfor
for $(\beta = l_{gh} + 1, \beta <= 2 \times l_{gh} + 1, \beta++)$
$g_{i,k,r}[\beta]$=random selection from t_s
endfor
$r = r + 1$
endwhile

$\Phi_{ik} = \{\{g_{i,k,r} | r = 1...g_n\},\ et_{ik} = \{\quad\},\ st_{ik} = ``\quad",\ st^s_{ik} = ``\quad",\ st^{sc}_{ik} = ``\quad",\ st^o_{ik} = ``\quad",\ P_{ik} = \{\quad\},\ X_{ik} = \{\quad\},\ fv_{ik} = 1000,\ fv'_{ik} = 1000\}$,
where $\{\quad\}$ denotes an empty set and, `` '' denotes an empty string.

$k = k + 1$

endwhile

Evolution loop:

while check for termination criteria ==**False**, **then**:

- translate all individuals q in generation i that are a set genes g_{iqr} linked by link function f_l into expression trees et_{iq}
- obtain the mathematical expression st_{iq} and prediction values X_i for all individuals in the current generation
- evaluate predicted values of the individuals in the current generation against real target data T_d and assign fitness values fv_{iq} based on fitness function f

for$(q = 1, q <= n, q++)$ $et_{iq} = translate(\{g_{iqr} | r = 1...g_n\}, f_l)$
$st_{iq}, X_i = eval(et)$
$fv_{iq} = f(X_i, T_d)$

endfor

sort Ω_i based on cmp

Do CMA-ES optimization

- get simplified mathematical expression st^s_{iq} for each of the best p_o individuals and determine the locations of constants in the expression to optimize st^{sc}_{iq}
- if the individual contains a set of previously optimized constants, return set of optimized constants P_{iq}, optimized expression st^o_{iq}, and optimized fitness value $f'v_{iq}$ using the CMA-ES approach based on fitness function f and the string built from the set of previously optimized constants as initial sample set and their locations
- if the individual to optimize does not contain a set of previously optimized constants, return set of optimized constants P_{iq}, optimized expression st^o_{iq}, and optimized fitness value $f'v_{iq}$ using the CMA-ES approach based on fitness function f and the string built from the current constants determined from the simplified mathematical expression and their locations after applying $determineCst$ function
- evaluate the optimized expression of the chromosome for assigning the optimized fitness value.

for$(q = 1, q <= p_o, q++)$

$st^s_{iq} = simp(\Phi_{iq})$
$st^{sc}_{iq} = detC(\Phi_{iq})$
if$(\Phi_{iq}.P_{iq} != \{\quad\})$
$st^{sc}_{ik} = reconstructS(\Phi_{ik}.st^s_{ik}, \Phi_{iq}.P_{iq})$
$\Phi_{iq}.P_{iq}, st^o_{iq}, f'v_{iq} = L_o(st^{sc}_{ik}, f)$
else
$\Phi_{iq}.P_{iq} = determineCst(st^s_{iq})$
$st^{sc}_{ik} = reconstructS(\Phi_{ik}.st^s_{ik}, determineCst(Phi_{ik}.st^s_{ik}))$
$\Phi_{iq}.P_{iq}, st^o_{iq}, f'v_{iq} = L_o(st^{sc}_{ik}, f)$
endif

endfor

for all individuals that do not belong to the best p_o set, assign current fitness value to optimized fitness value for future optimized fitness value based comparison in the population

for$(q = p_o + 1, q <= n, q++)$
$f'v_{iq} = fv_{iq}$

endfor

sort Ω_i based on optimized fitness values $f'v$ as defined in cmp'

Save best individual and generate chromosomes for new evolution step by **genetic variation**

$\Phi_{(i+1)1} = \Phi_{(i)1}$

for $(k = 2, k <= n, k++)$

do random selection of individuals $\Phi_{i\mu}$ and $\Phi_{i\mu'}$ from Ω_i for tournament based on optimized fitness and apply random genetic variation operators to the best
if$(cmp'(\Phi_{i\mu}, \Phi_{i\mu'}))$ **then**
$\Phi_{(i+1)k} = gV(\Phi_{i\mu})$
else $\Phi_{(i+1)k} = gV(\Phi_{i\mu'})$
endif

endfor

$i = i + 1$

endwhile

return Φ_{i1}
return the best over-all individual for the current evolution

CMAGEP definitions

$i = 1,..m,\ m \in \mathbb{N}$, evolution step, also known as generation

Ω, the set of all possible chromosomes under set conditions

Ω_i, the set of all chromosomes generated at evolution step i

$\langle \Omega_i \rangle = n,\ n \in \mathbb{N}$, population count, fixed for all evolution steps

$g_n \in \mathbb{N}$, number of genes in a chromosome

f_l, link function, that connects all the expression trees associated to the genes of a chromosome into **a single expression tree**

$d_s \in \mathbb{N}$, the size of the data sample used for training, fixed

$g_{i,k,r}, r = 1,..g_n$, gene r of chromosome k at generation i, encoding an expression tree via the Karva language Ferreira (2001)

$g^h_{i,k,r}$, head of gene r of chromosome k at generation i, containing a random array of characters mapping to functions and terminals, with **length** $l_{gh} \in N$**, fixed**

$g^t_{i,k,r}$, tail of gene r of chromosome k at generation i, containing a random array of characters mapping only to terminals, with **length** $2 \times l_{gh} + 1$

$\Phi_{ik},\ i = 1,..m,\ k = 1,..n$, the evolution individual, called chromosome, at generation i, position k

$f : \mathbb{R}^{d_s \times 1} \times \mathbb{R}^{d_s \times 1} \to \mathbb{R}$, **fitness function**, that returns a numerical value based on the value array obtained from mathematically evaluating a chromosome against a given target training data set

f_s, functional transformation set, characters mapping to possible functional transformations

t_s, terminal set, characters mapping to candidate predictors

genetic operators rates:

$m_r \in (0,1)$, mutation rate
$t_r \in (0,1)$, transposition rate
$i_r \in (0,1)$, insertion rate
$r_r \in (0,1)$, recombination rate

Let $\Phi_{ik} := (\{g_{i,k,r} \mid r = 1,..g_n\}, et, et_s, et_{sc}, et_o, X_{ik}, P_{ik}, v_f, v_{f'})$
with:
$\{g_{i,k,r} \mid r = 1,..g_n\}$, the set of genes in chromosome k, at generation i, and cardinal g_n
et_{ik}, the expression tree generated after translating all the strings encompassing the genes of the current chromosome, and applying the link function
st_{ik}, string defining the mathematical expression obtained after parsing the expression tree associated to the chromosome
st^s_{ik}, string defining the simplified mathematical expression associated to the chromosome, after applying the simplification function from the "Sympy" package
st^{sc}_{ik}, string defining simplified mathematical expression associated to the chromosome, after determining the constants to be optimized
st^o_{ik}, string defining optimized mathematical expression associated to the chromosome, after applying the optimization function from the CMA-ES package
$X_{ik} \in \mathbb{R}^{d_s \times 1}$, numerical evaluation array to be compared with target data
$P_{ik} \in \mathbb{R}^{n_c \times 1}$, set of parameters resulted from the local CMA-ES optimization, n_c is the number of constants to be optimized and is determined locally
fv_{ik}, fitness value assigned to the chromosome after evaluating the predictions of the st_{ik} mathematical expression of the current chromosome against known target values, based on the f fitness function
fv'_{ik}, optimized fitness value, obtained after evaluating the optimized mathematical expression, st^o_{ik} of the current chromosome, based on the f fitness function

evolution stop criteria:
m_{gen}, maximum number of generations allowed with no improvement in best fitness
m_{rt}, maximum runtime

optimization parameters:
p_o, number/ percentage of best chromosomes to be optimized
g_o, generation at which optimization can start
i_o, maximum CMA-ES iterations for optimizing a chromosome
fe_o, maximum CMA-ES fitness function evaluations needed for optimizing a chromosome

Important operations:

chromosome comparison:

$$cmp(\Phi_{ik},\Phi_{ij}) = \begin{cases} \Phi ik, \ fv_{ik} > fv_{ij} \\ \Phi_{ij}, \ \text{otherwise} \end{cases}$$

$$cmp'(\Phi_{ik},\Phi_{ij}) = \begin{cases} \Phi ik, \ fv'_{ik} > fv'_{ij} \\ \Phi_{ij}, \ \text{otherwise} \end{cases}$$

simplification:
Let S:={st | st is a string defining a mathematical expression}
$simp : \Omega \rightarrow S$,
$st^s_{ik} = simp(\Phi_{ik}) := Sympy(st_{ik})$, where "Sympy" is a Python symbolic operation package.

determine constants to be optimized and their location:
determineCst : $S \rightarrow \mathbb{R}^{n_c}$, function that determines the constants that can be optimized from the string of the simplified mathematical expression associated with a chromosome
$n_c = count(determineCst(st^s_{ik}))$ number of constants that need to be optimized
$reconstructS : S \times \mathbb{R}^{n_c} \rightarrow S$, function that reconstructs the mathematical expression st^{sc}_{ik} to be optimized based on the determined constants to be optimized and their locations
$st^{sc}_{ik} = reconstructS((\Phi_{ik}.st^s_{ik}, determineCst((\Phi_{ik}.st^s_{ik}))$

local CMA-ES optimization:
$L_o : S \rightarrow \mathbb{R}^{n_c} \times S \times \mathbb{R}$
$P_{ik}, st^o_{ik}, f'_v = L_o(st^{sc}_{ik}, f)$, function that does local CMA-ES optimization for a set of constants in the simplified mathematical expression st^{sc}_{ik} of an individual Φ_{ik} and a fitness function f and returns a set of optimized constants P_{ik}, an optimized mathematical expression st^o_{ik} and and optimized fitness value f'_v

3.7 Supplemental Materials: Visualising Benchmark Test Functions and GEP and CMAGEP reconstructions

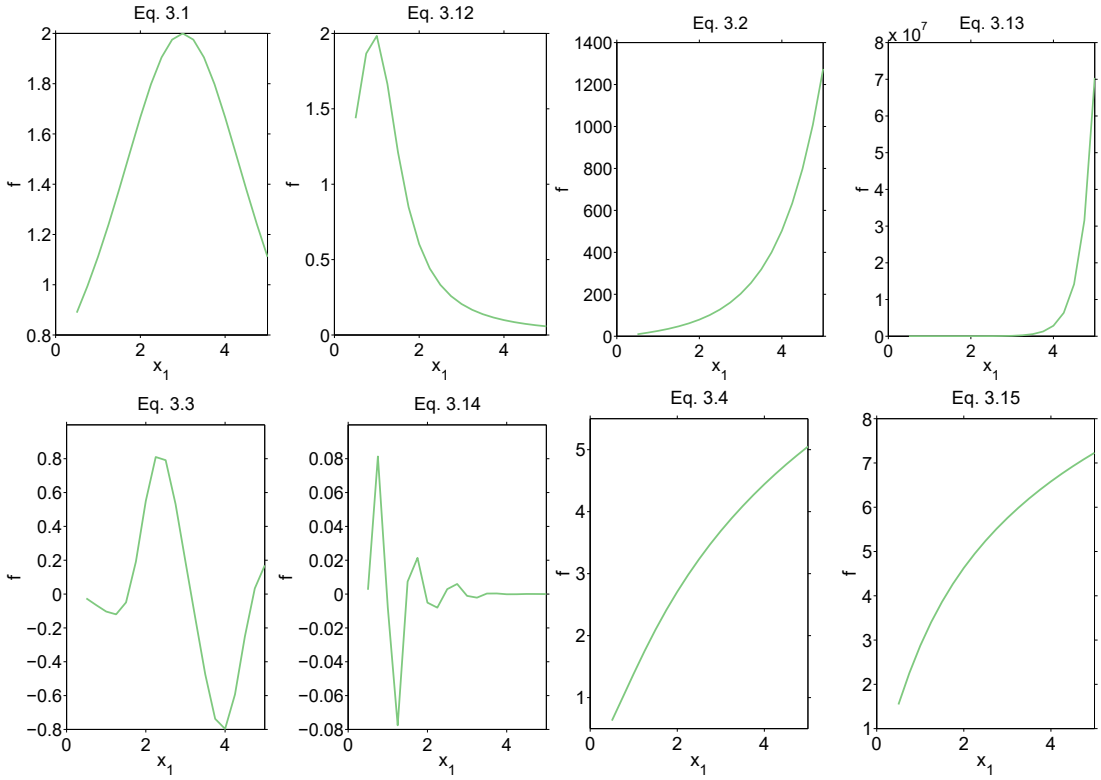


Figure 3.12: Functions used in the artificial benchmark test. The first and third columns show the functions that before adding high precision constants and the second and fourth columns show the changes produced after introducing high precision constants. Part A.

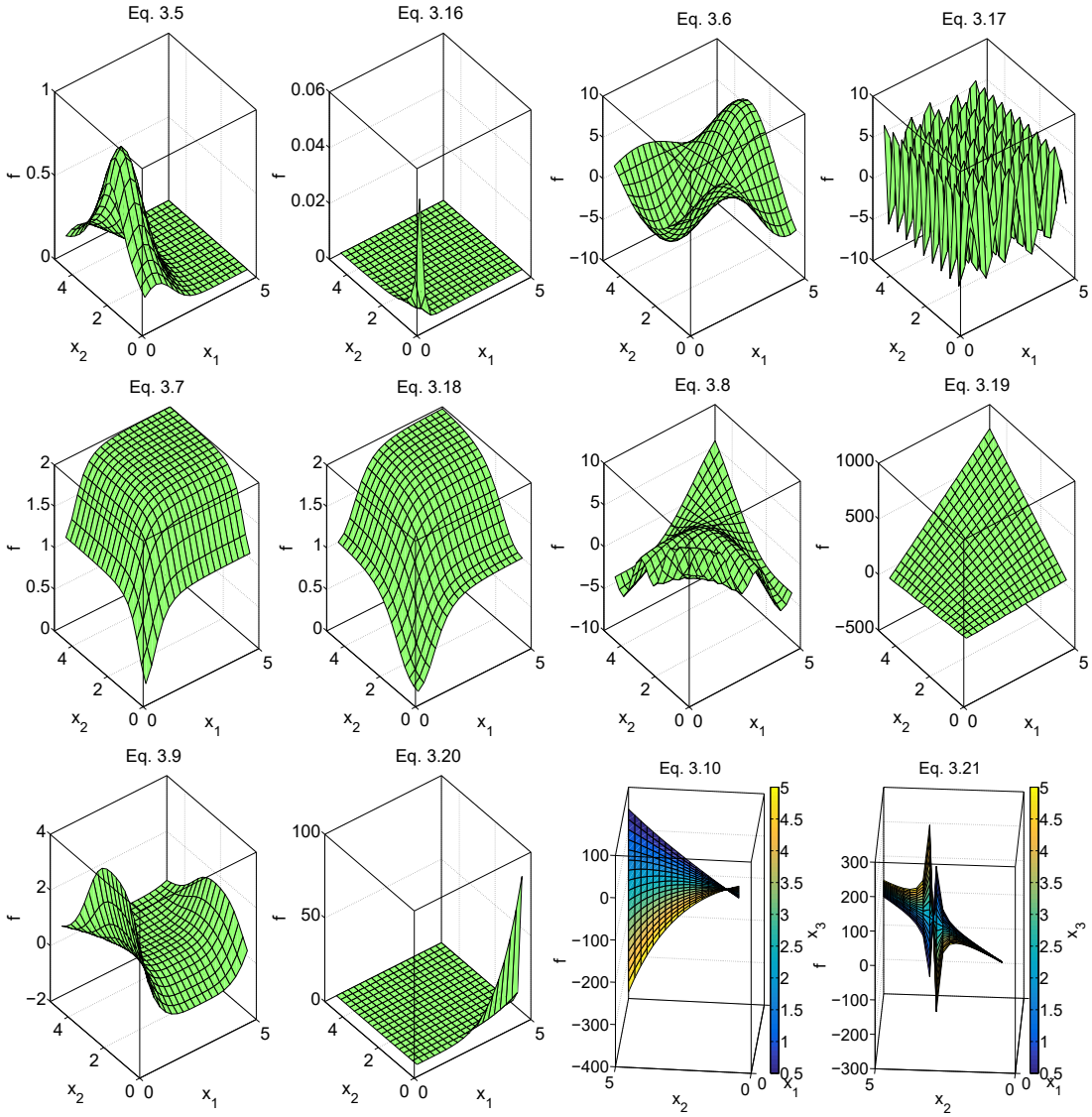


Figure 3.13: Functions used in the artificial benchmark test. The first and third columns show the functions that before adding high precision constants and the second and fourth columns show the changes produced after introducing high precision constants. Part B.

Table 3.10: Best function formulations returned by GEP and CMAGEP after 50 independent runs with settings given in the first column of Table 3.1 for the artificial benchmark function set lacking high precision constants.

Eq.	GEP	CMAGEP
3.3.1	$0.67^{\cos(x_1)} + \log(\sin(x_1)+6)^{0.5} + \cos(\log(\sin(\cos(x_1))+6)) - 0.78$	$\sin(0.97\sqrt{(x_1)} + 0.38\cos(1.33x_1 - 4.06) + 0.62$
3.3.2	$x_1(\log(x_1^2)+6) + 6x_1\log(x_1) + 8e^{(x_1)} - 3$	$12x_1 + 9.9e^{(0.96x_1)} - 13$
3.3.3	${x_1}^{0.5} + e^{(x_1)^{0.25e^{(-{x_1}^2)}}} + \sin(x_1 - \sin(x_1)) - 1.7$	$0.78\sin(3.34\cos(0.73x_1 + 11.84)) + 0.01$
3.3.4	$x_1 + \sin(\sqrt{(x_1)}) + \sin(\sqrt{(x_1)})^{(x_1^{0.5})} - 1.3$	$0.5x_1 - 1.4\sin(1.46\log(3.68x_1 + 3.71)) + 1.2$
3.3.5	$x_1^2.4e^{(-x_1^2)} + 0.2(\frac{0.14x_2}{x_1})^{\frac{x_1}{x_2}} + 0.2e^{(-x_1)}\sin(\log(x_2)) + \frac{0.2\log(x_2)\sin(x_2)}{x_1}$	$3\sin(x_1 - x_2) + 3\sin(x_1 + x_2)$
3.3.6	$6\sin(x_1)\cos(x_2)$	$6\sin(x_1)\cos(x_2)$
3.3.7	$1.1\cos(1.02)^{(-1.01x_1)} + 1.2\cos(1.04x_2)^{(-0.96x_2)} - 0.34$	$0.51\cos\frac{1.63}{x_1} + 0.52\cos\frac{1.55}{x_2} + 1.1$
3.3.8	$(x_1 - x_2)\sin(x_1) + (x_1^{x_1})^{(-0.5)}e^{(\cos(x_1x_2))}$ $+ \cos(\sin(x_2))^{(x_1^2+x_2-x_3)} + e^{(-\sin(x_2))}\sin(x_1)\sin(x_2)$	$2.1\sin(0.38x_1 + 0.59x_2 + 1.52) + 2.9\sin(0.62x_1 - 0.69x_2 + 1.8)$ $-1.9\sin(0.51\cos(1.59x_1 + 1.88x_2))$
3.3.9	$\cos(x_1) + \cos(\sin(x_1)) + \frac{(\cos(x_1)-6)}{(8_2^3+x_1)} + \frac{0.62\sin(x_2+6)}{x_1^2}$	$(-0.73)x_2 + 44\cos(1.528x_3(0.106x_1)^{(1.446x_2)}) - 43$
3.3.10	$(x_1 - x_2)\frac{(\frac{x_1}{x_2} - \log(x_3))}{x_2} + (x_1 - x_2)\frac{(\sin(x_1) - \frac{x_3}{x_2})}{x_2} + \frac{2x_3(-x_1 + \sin(x_1))}{x_2^2}$	$1.3x_1^2\frac{(-0.7x_3+0.67)}{x_2^2} + \frac{0.24}{(1.1x_1+0.57x_2)} + \frac{0.47(-0.76x_2+1.5x_3)}{(x_1x_2)}$

Table 3.11: Best function formulations returned by GEP and CMAGEP after 50 independent runs with settings given in the first column of Table 3.1 for the artificial benchmark function set containing high precision constants.

Eq.	GEP	CMAGEP
3.3.12	$\frac{e^{(\sin(x_1)\cos(x_1)}}{x_1}^{0.5} \sin\left(e^{(x_1{}^{(-x_1)})}\cos(x_1)\right)+\cos(\log(7x_1)-\cos(\log(x_1)))-0.76$	$0.37e^{(1.192(0.67x_1)^{(-0.661x_1)}}\sin(1.401x_1))+0.33\cos(0.604x_1)$
3.3.13	$1.528^{x_1}+\frac{28^{x_1}}{\sin(\frac{3}{x_1})}+0.063x_1^{12}+{x_1}^{2{x_1}^{x_1(\sin(x_1)-7.0)}}$	$7.8e^{(3.203x_1)}+3.2$
3.3.14	$\frac{(\sin(\sin(4x_1))}{e^{e^{(x_1)}}}+\frac{\cos(6+x_1)x_1}{e^{(x_1 4)}}+\frac{\cos(6+x_1)x_1}{e^{(x_1 4)}}+\frac{\cos((x_1+x_1)x_1)}{e^{(x_1 4)}}$	$0.150.37^{(0.9x_1)}\cos(5.455\cos(1.278x_1))-0.003$
3.3.15	$x_1+\sin({x_1}^{0.65})+\sin(\log(x_1))+1.0$	$0.6x_1^{(-1.2)}+3\log(3.83x_1)-1.8$
3.3.16	$e^{(-8x_1-2x_2+\sin(x_1))}+e^{\frac{-8x_1+x_1}{x_2-2x_2}}+1.7\times10^{6^{(-x_1x_2)}}\log(x_3)+1.7\times10^{6^{(-x_1^{0.5})}}x_2^{0.5}$	$\frac{0.02x_1}{x_2}{}^{e^{(x_1)}}+2.0\times0.14x_1^{\frac{0.5}{x_2}(6^{x_1})}$
3.3.17	$\sin(\frac{5}{2x_1})+\sin(\frac{(x_1+x_2)\log(x_1)}{{x_2}^2})+\sin(e^{(-x_3+26)})+\sin(e^{(\sin(-x_1+e^{(x_3)}+5))})$	$3.1\sin(3.174x_1+3.143\cos(6.286x_2)-3.095)+0.25$
3.3.18	$\frac{x_1}{(x_1+x_2+\cos(\cos(x_2)))}+\frac{x_1}{(3.0x_1+\cos(x_1))}+\sin(\log(x_1+\log(x_1+2)))+\frac{\cos(x_2+8)}{x_2}$	$(3.1x_2+1.6\sin(4.37sqrt(x_2))^{0.27}+0.52\sin(1.454\log(0.77x_1))-0.59$
3.3.19	$18x_1x_2+x_2(18x_1-36)+x_2+{x_2}^{{x_1}^{0.5}}+(x_2-3)e^{{x_1}^{0.5}}+\log(x_2^{x_1})$	$15x_1x_2+14x_2(1.7x_1-2.4)-35\sin(0.41\log(1.61x_1))$
3.3.20	$\sin(\cos(\sin(\cos(x_1)))+4)+\cos(4e^{(-x_2)^{\frac{x_1}{2}}})$ $+0.25x_2^{(-x_1)}\log(x_1^{x_1})+\frac{x_2^{(-x_1)}\cos(\cos(x_1))}{x_1}$	$(1.5x_1)^{(-0.97\cos((0.325x_2)^{\frac{(-0.142x_1)}{x_2}})}-0.82$
3.3.21	$\sin(x_1)^{\cos(x_2)^{(-x_1+\cos(x_3))}}+\frac{\cos(x_1)\cos(x_2)}{(x_3\cos(x_3))}+\frac{\log(x_3)\cos(x_1)}{(x_2\cos(x_3))}+\frac{\log(x_3)\sin(x_2)}{x_2(-x_1+x_3)}$	$1.4x_2-0.21\log(1.3\log(1.03x_1+4.9))-0.54+\frac{1.8}{(x_2\sin(1.005x_1))}$

Modelling CH_4 fluxes in an Arctic site using CMAGEP

4.1 Introduction

Rising air and surface temperature values in Arctic regions (Serreze et al., 2000; Serreze and Barry, 2011) can lead to thawing of permafrost layers and melting of ice sheets (Jorgenson et al., 2006; O'Donnell et al., 2011). Such events will most probably cause changes in soil hydrology and plant community compositions as well as CH_4 terrestrial cycle.

Wetlands in the Arctic regions are among the most important sources of CH_4 , accounting for $\approx 10\%$ of the total number of CH_4 emissions (Ciais et al., 2013; Kirschke et al., 2013). Understanding the factors determining the changes in CH_4 fluxes in these ecosystems is necessary for an accurate representation of Arctic wetlands responses to climate change.

Previous studies already indicate that some of the most influential drivers for CH_4 fluxes during the growth periods are water table depth (WTD (Merbold et al., 2009; Sturtevant et al., 2012)), plant community composition (Andresen et al., 2017; McEwing et al., 2015; Morrissey and Livingston, 1992; Tsuyuzaki et al., 2001), air temperature (T_{air}), soil temperatures from different depths (T_{soil}) (Nakano et al., 2000; Tveit et al., 2015) and thaw depth (ThD (Kim, 2015)). However, due to limited number of studies the CH_4 cycle is not yet well understood for the non-growing periods, although there are indications on a higher influence of physical factors such as air pressure (P_a), air and soil temperatures T_{air} and T_{soil} determining the gas exchange during this period (Mastepanov et al., 2013, 2008).

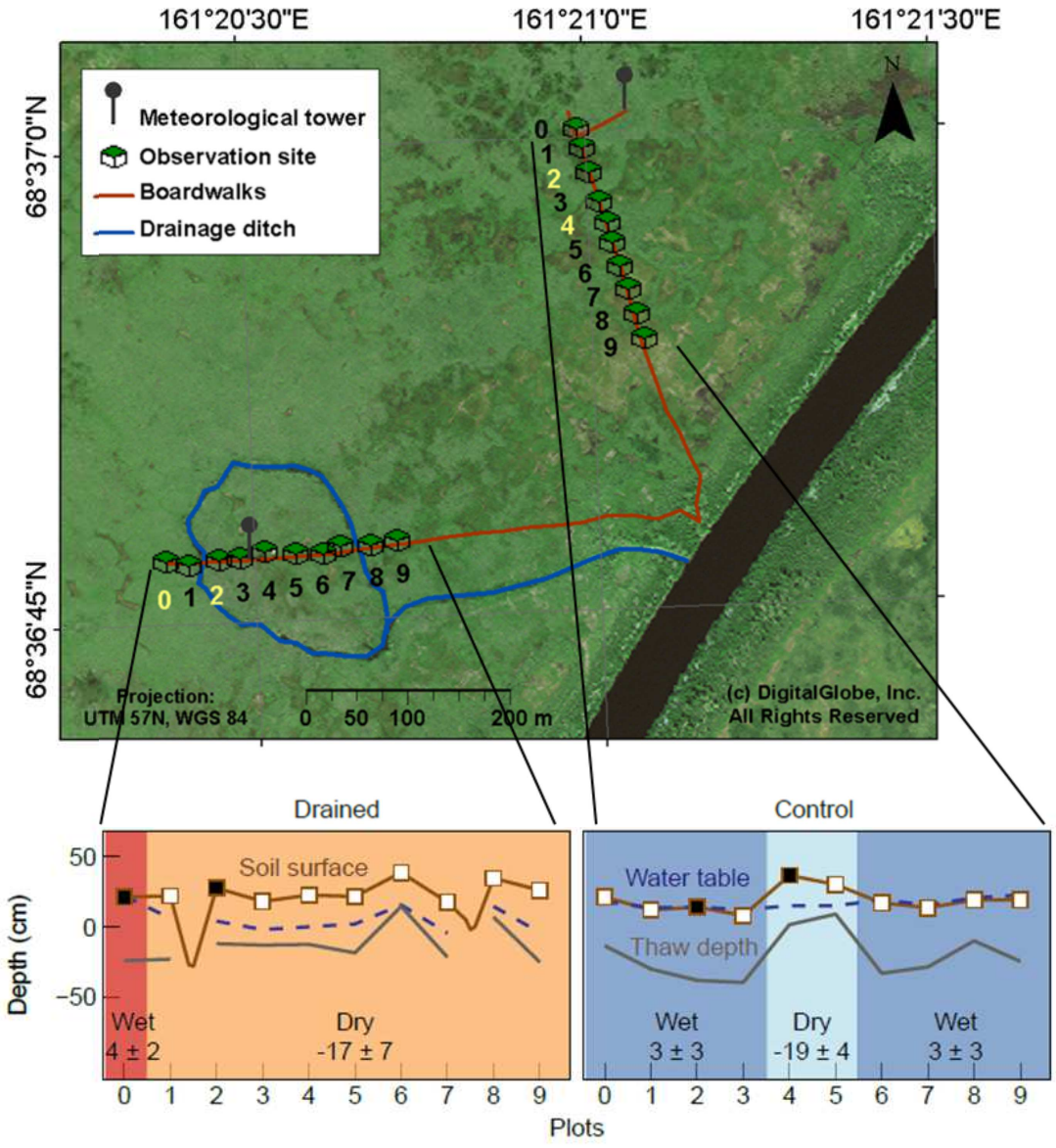


Figure 4.1: Description of measurement conditions of Chersky site in NE Siberia.

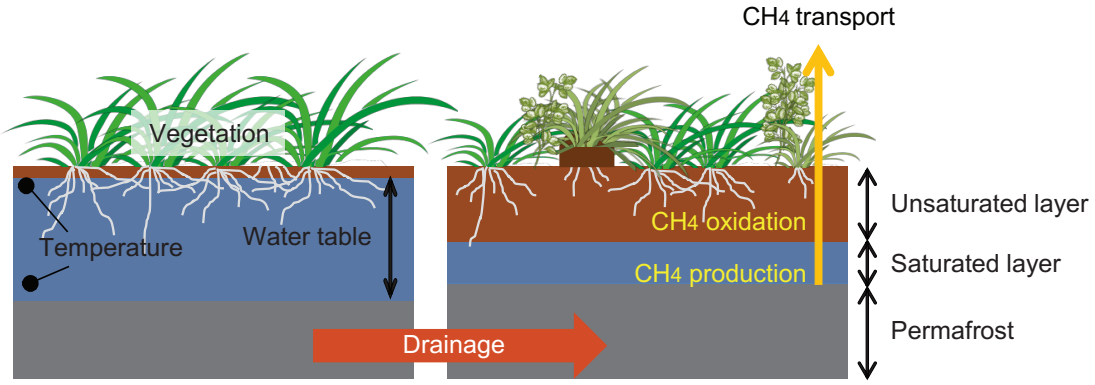


Figure 4.2: Fluxes monitoring at the Chersky site in Russia.

This Chapter investigates the possibility of automatically learning significant models for describing CH_4 fluxes an Arctic site using the CMAGEP approach (Ilie et al.). The CMAGEP retrieved models have explicit structures that could offer new insights to the response of CH_4 flux to candidate drivers and their importance.

4.2 Data and Method

The current study is based on measurements taken at an Arctic floodplain site near Chersky, NE Siberia. CH_4 exchanges were measured and recorded as well possible drivers such as P_a, photosynthetic active radiation (PAR), WTD, ThD, T_{air}, T_{soil} at different depths (in cm, T_{soil5},T_{soil15}, T_{soil25}, T_{soil35}) and Eriophorum. angustifolium (E), Carex. appendiculata (C), and Potentilla palustris (P) abundance in percentage per plots. (Kwon et al., 2016) give a detailed presentation on the measurement procedures.

As the interest was not only in determining the influence of the above mentioned drivers to CH_4 fluxes but also to determine the influence of drainage and the season when the monitoring is done on the laws governing the fluxes, the data was split for the two transects (drained and un-drained), for both summer and winter, giving a total of four transect-season (TS) cases. For all TS the stepAIC R function was applied by Min Jung Kwon to linear models generated with the lm R function (with allowance of interactions) based on all the drivers and some of their functional transformations (square root, logarithm and exponential). The generated models and prediction performances were then reported. (https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/stepAIC.html) Models were automatically bred as well using the CMAGEP approach for each of studied TS. For all TS 70% of the total number of data points were randomly sampled for training, and the remaining 30% were stored for validation. The sampling was repeated

Table 4.1: Variables included in the input of CMAGEP runs for generating models for CH_4 fluxes.

Candidate drivers:
Summer (27 variables):
P_a, PAR, WTD, ThD, T_{air}, T_{soil5},T_{soil15}, T_{soil25}, T_{soil35}, E, C, P,
$\Delta 30P_a$, $\Delta 60P_a$, $\Delta 120P_a$, $\Delta 180P_a$, $\Delta 360P_a$
$\Delta 30T_{air}$, $\Delta 60T_{air}$, $\Delta 120T_{air}$, $\Delta 180T_{air}$, $\Delta 360T_{air}$
$\Delta 30PAR$, $\Delta 60PAR$, $\Delta 120PAR$, $\Delta 180PAR$, $\Delta 360PAR$
Winter (25 variables):
P_a, PAR, T_{air}, T_{soil5},T_{soil15}, T_{soil25}, T_{soil35}, E, C, P
$\Delta 30P_a$, $\Delta 60P_a$, $\Delta 120P_a$, $\Delta 180P_a$, $\Delta 360P_a$
$\Delta 30T_{air}$, $\Delta 60T_{air}$, $\Delta 120T_{air}$, $\Delta 180T_{air}$, $\Delta 360T_{air}$
$\Delta 30PAR$, $\Delta 60PAR$, $\Delta 120PAR$, $\Delta 180PAR$, $\Delta 360PAR$

90 times, thus generating 90 train-validation pairs.

CMAGEP independent runs were performed for each of the 90 train-validation pairs, with the settings given in table 4.3, generating 90 models. Mean Akaike Information Criterion (AIC) values over the 90 validation samples were computed for all 90 generated models and the models with the lowest mean AIC values were selected for each of the transect-seasons and their structures and prediction performances were reported. AIC

To ensure interpretability, the Root Mean Squared Error (RMSE) and Model Bias Error (MBE) were also computed across the entire dataset.

4.3 Results

The models generated by using the stepAIC function:

$$\begin{aligned} \text{CH}_4\ \text{flux}_{ds} = {} & P_a + T_{air} + T_{soil5} + T_{soil15} + T_{soil25} + \\ & T_{soil35} + E + C + WTD + ThD + P_aT_{air} + \\ & P_aT_{soil35} + P_aWTD + P_aThD + T_{soil5}ThD + \\ & T_{soil15}T_{soil35} + T_{soil15}E + T_{soil15}WTD + T_{soil15}ThD + \\ & T_{soil35}E + T_{soil35}ThD + EWTD + CWTD + WTDThD \end{aligned} \tag{4.3.1}$$

$$\begin{aligned}\text{CH}_4\ \text{flux}_{us} = {} & P_a + T_{air} + T_{soil5} + T_{soil15} + T_{soil25} + \\ & T_{soil35} + E + C + P + WTD + ThD + P_aP + \\ & P_aWTD + T_{air}T_{soil15} + T_{air}T_{soil25} + T_{air}E + \\ & T_{air}ThD + T_{soil5}T_{soil35} + T_{soil5}P + T_{soil5}WTD \\ & + T_{soil5}ThD + T_{soil15}T_{soil25} + T_{soil15}E + T_{soil15}P + \\ & T_{soil15}WTD + T_{soil15}ThD + T_{soil25}T_{soil35} + T_{soil25}E + \\ & T_{soil25}C + T_{soil25}P + T_{soil25}WTD + T_{soil25}ThD + \\ & T_{soil35}E + T_{soil35}C + T_{soil35}P + EWTD + EThD + \\ & CWTD + WTDThD\end{aligned} \tag{4.3.2}$$

$$\begin{aligned}\text{CH}_4\ \text{flux}_{dw} = {} & P_a + T_{air} + T_{soil5} + T_{soil15} + T_{soil25} + \\ & T_{soil35} + E + C + P + P_aT_{air} + P_aT_{soil25} + \\ & P_aT_{soil35} + P_aE + T_{air}T_{soil5} + T_{air}T_{soil35} + \\ & T_{air}E + T_{air}C + T_{air}P + T_{soil5}T_{soil25} + T_{soil5}T_{soil35} + \\ & T_{soil5}C + T_{soil5}P + T_{soil15}T_{soil25} + T_{soil15}C + \\ & T_{soil15}P + T_{soil25}T_{soil35} + T_{soil25}P + T_{soil35}C + \\ & T_{soil35}P + EC\end{aligned} \tag{4.3.3}$$

$$\begin{aligned}\text{CH}_4\ \text{flux}_{uw} = {} & P_a + T_{air} + T_{soil5} + T_{soil15} + T_{soil25} + \\ & T_{soil35} + E + C + P + P_aT_{air} + P_aT_{soil5} + \\ & P_aT_{soil15} + P_aT_{soil25} + P_aE + P_aC + \\ & P_aP + T_{air}T_{soil5} + T_{air}T_{soil15} + T_{air}T_{soil25} + \\ & T_{air}T_{soil35} + T_{air}C + T_{air}P + T_{soil5}T_{soil15} + \\ & T_{soil5}T_{soil25} + T_{soil5}T_{soil35} + T_{soil5}C + T_{soil15}T_{soil35} + \\ & T_{soil15}E + T_{soil15}C + T_{soil25}E + T_{soil25}P + T_{soil35}E\end{aligned} \tag{4.3.4}$$

Table 4.2: Modelling performance for all transect-seasons, for stepAIC and CMAGEP generated models.

	DS	**US**	**DW**	**UW**
stepAIC R^2	0.99	0.97	0.91	0.81
CMAGEP R^2	0.90	0.70	0.43	0.08
stepAIC solution length	24	39	30	32
CMAGEP solution length	3	2	3	2

The models generated by using the CMAGEP approach:

$$\text{CH}_4 \text{ flux}_{ds} = \frac{0.09T_{soil35}}{0.15P_a + 4.3C} \tag{4.3.5}$$

$$\text{CH}_4 \text{ flux}_{us} = \frac{0.06E}{P_a} \tag{4.3.6}$$

$$\text{CH}_4 \text{ flux}_{dw} = 30E \times e^{1.3T_{air}} \tag{4.3.7}$$

$$\text{CH}_4 \text{ flux}_{uw} = \frac{0.09}{P_a} \tag{4.3.8}$$

The modelling performance scores for the models generated with the classic stepAIC approach and automatically with the CMAGEP approach are presented in Tab. 4.2.

Fig. 4.3 illustrates the capacity of the CMAGEP model to capture the CH_4 fluxes measured in the summer season in the drained section.

Fig. 4.4 illustrates the capacity of the CMAGEP model to capture the CH_4 fluxes measured in the summer season in the un-drained section.

4.4 Discussion

The models generated with the standard procedure, stepAIC have a higher overall prediction performance compared to the models generated with CMAGEP. However, for all TS it is obvious that the models generated by CMAGEP are much more compact, with far less parameters needed to describe the response of CH_4 to drivers. The fact that CMAGEP achieved to build much more compact solutions allowed for further investigation on whether the structures obtained make actual sense from a physical and biological view and also whether the insights obtained are in line with already observed mechanisms.

Nevertheless it is worth mentioning that when high prediction performance is the purpose for building a model, especially in describing CH_4 flux responses in the Arctic

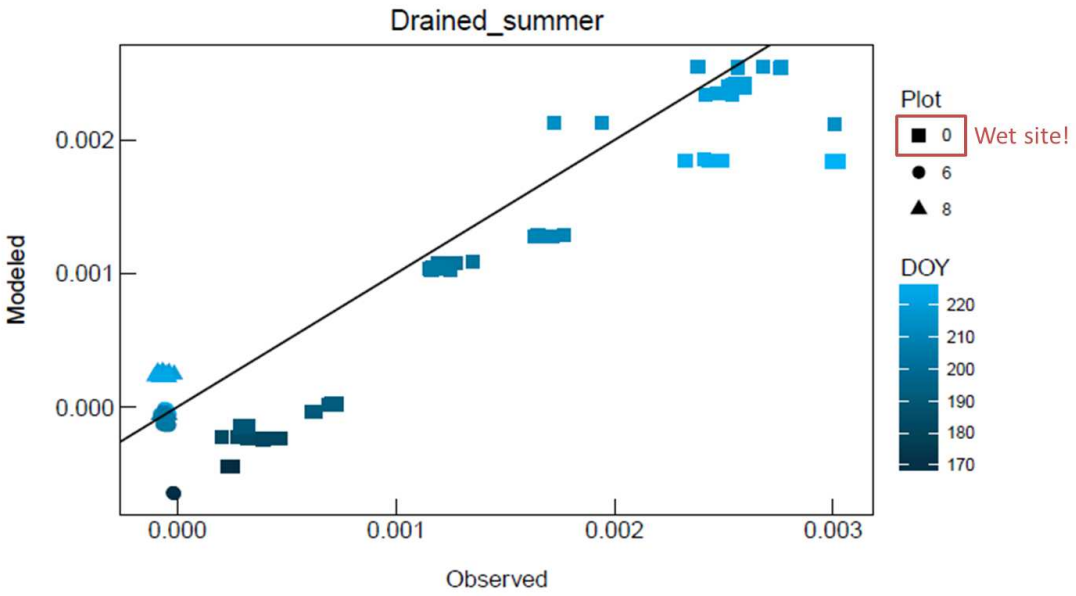


Figure 4.3: Observed and CMAGEP model predicted CH_4 flux at the Drained site in summer season.

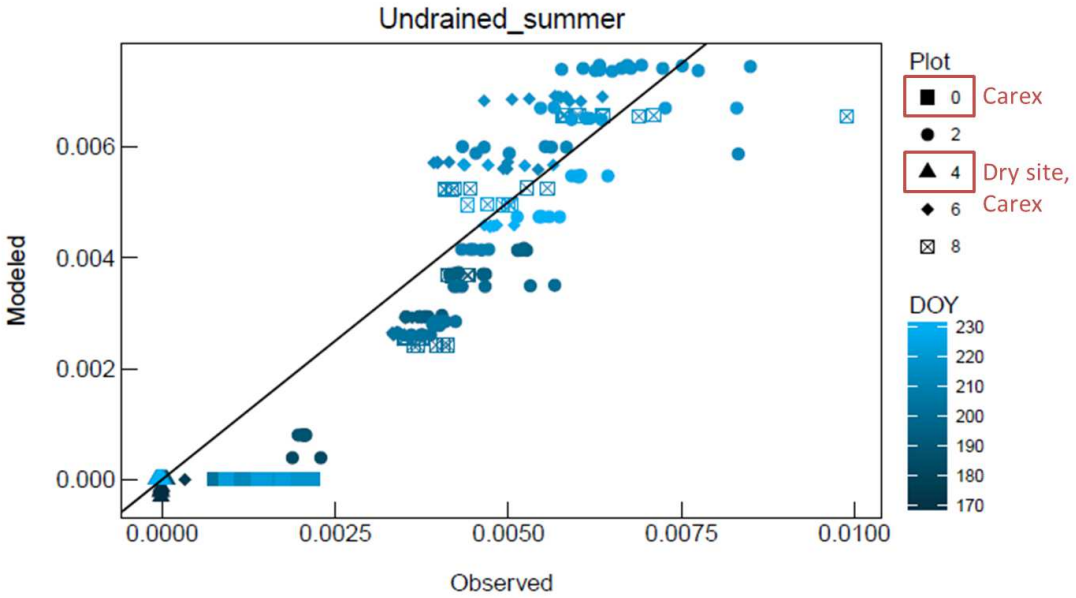


Figure 4.4: Modelled and CMAGEP model predicted CH_4 flux at the Undrained site in summer season.

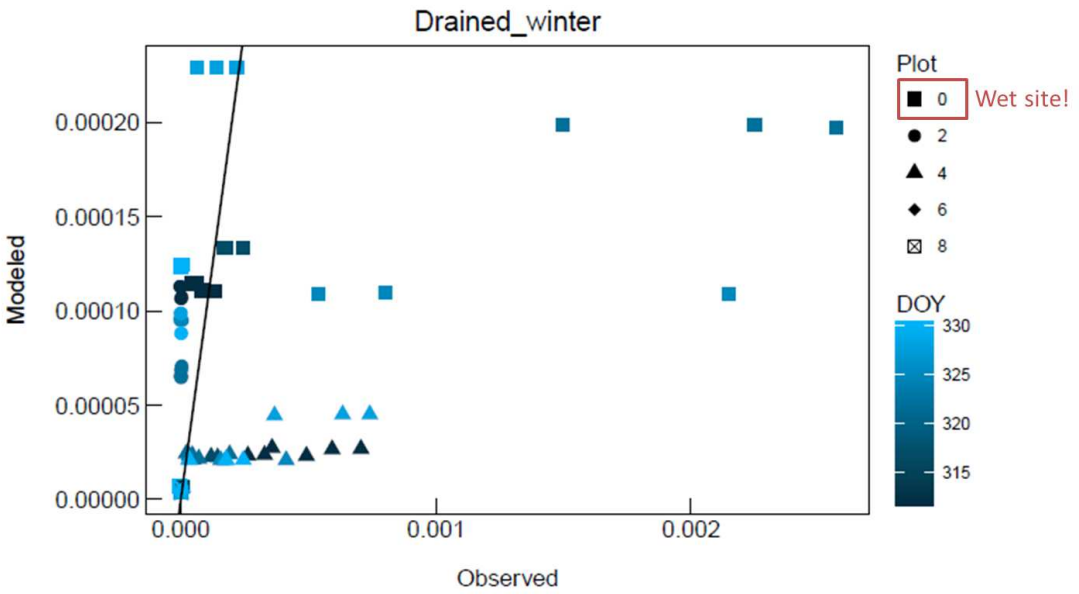


Figure 4.5: Modelled and CMAGEP model predicted CH_4 flux at the Drained site in winter season.

site in the non-growing season, CMAGEP might not be the best solution. On the other hand it is possible that the signals captured in the available data are not sufficient for CMAGEP to construct a relevant model structure for the non-growing season.

The CMAGEP generated models revealed that the growing-season CH_4 flux rates were positively influenced by T_{soil} at deep layers and E cover as well as negatively by P_a and the C (drained: Equation 3; control, Equation 4). In the non-growing season, although the structures of the equations differed from those of the growing season, similar parameters influenced CH_4 flux rates (drained, Equation 5; control, Equation 6).

4.5 Conclusion

CMAGEP was applied on a new real world dataset containing measurements on CH_4 exchanges and new explicit models were generated for describing the relation of some of the most influential inputs to CH_4 fluxes in an Arctic floodplain in different seasons. The models obtained after applying the CMAGEP approach, were much more simple that those obtained through multivariate linear regressions, and for 2 out of the 4 studied cases the loss in prediction accuracy was sufficiently reduced to encourage the further use of the automatically generated models. The results of applying CMAGEP to a specific real world problem encourage us to believe that its applicability can be successfully

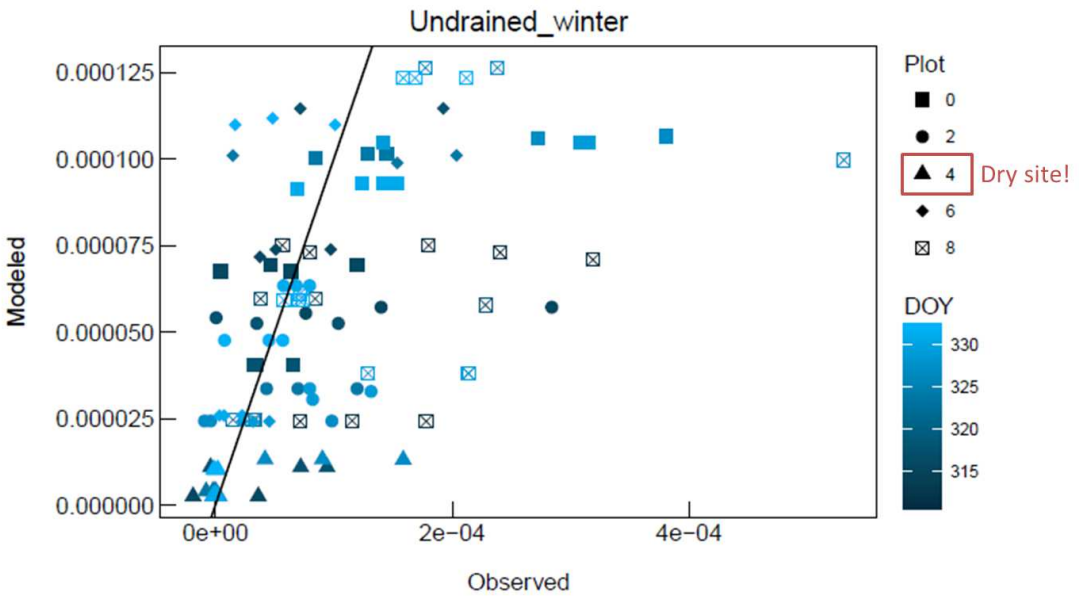


Figure 4.6: Modelled and CMAGEP model predicted CH_4 flux at the Undrained site in winter season.

extended to other problems and that when interpretation is needed, or that when previously not considered non-linear dynamics are present in the studied processes, the CMAGEP for symbolic regression is indeed a suitable modelling framework.

4.6 Author's contribution

The Author has performed all experiments and analysis regarding the CMAGEP for this Chapter. Results from using other regression packages were shared by Min Jung Kwon, who has also graciously shared the figures regarding the physical methane flux experiment set-up and site condition descriptions.

Table 4.3: CMAGEP settings

Parameter	
Number of chromosomes	1000
Number of genes	3
Head length	5
Functions	$+, -, /, *, x^y, \sqrt{}, \ln, \exp$
Terminals	given in table 4.1
Link function	+
Max run time	1800 seconds
Fitness function	AIC
Selection method for replication	tournament(Coello and Montes, 2002)
Mutation probability	0.2
IS and RIS transpositions probabilities	0.05
Two-point recombination probability	0.3
Inversion probability	0.05
One point recombination probability	0.4
Number of individuals to optimize	20
Time to start optimization	0
Maximum CMA-ES iterations	50

Large Scale Automated Discovery of Ecological Respiration models using CMAGEP

5.1 Introduction

Chapter 2 of this thesis explored the potential of a GEP system to automatically discover relevant physical and biological models that describe the response of ecosystem respiration (R_{eco}) components to biotic and abiotic external drivers. It was shown that the author's implementation of the standard GEP system is capable of building "readable" mathematical expression models that both confirm established knowledge in the biogeochemistry field as well as describe novel elements enriching the current knowledge and understanding of R_{eco}.

Although the potential of using GEP constructed model formulations to complement the current understanding of R_{eco} was confirmed (Ilie et al., 2017), it became obvious that the formulas constructed with the standard version of GEP were far too complex to allow drawing clear conclusions regarding the biological or physical soundness of the described models. In order to avoid the typical GP bloat associated with unnecessary evolution steps, improving the solution calibration in the proposed GEP system during the learning phase was necessary.

For improving the interpretability aspect of the solutions proposed by this GEP system, in Chapter 3 and in (Ilie et al.) CMAGEP, a novel GEP and CMA-ES hybrid system, was introduced. CMAGEP generated solutions with improved prediction performance and, more importantly, $\approx 60\%$ shorter when evaluated over an established artificial benchmark and two real-world study cases. The capacity of CMAGEP to improve existing models for carbon exchange was once more confirmed with the work

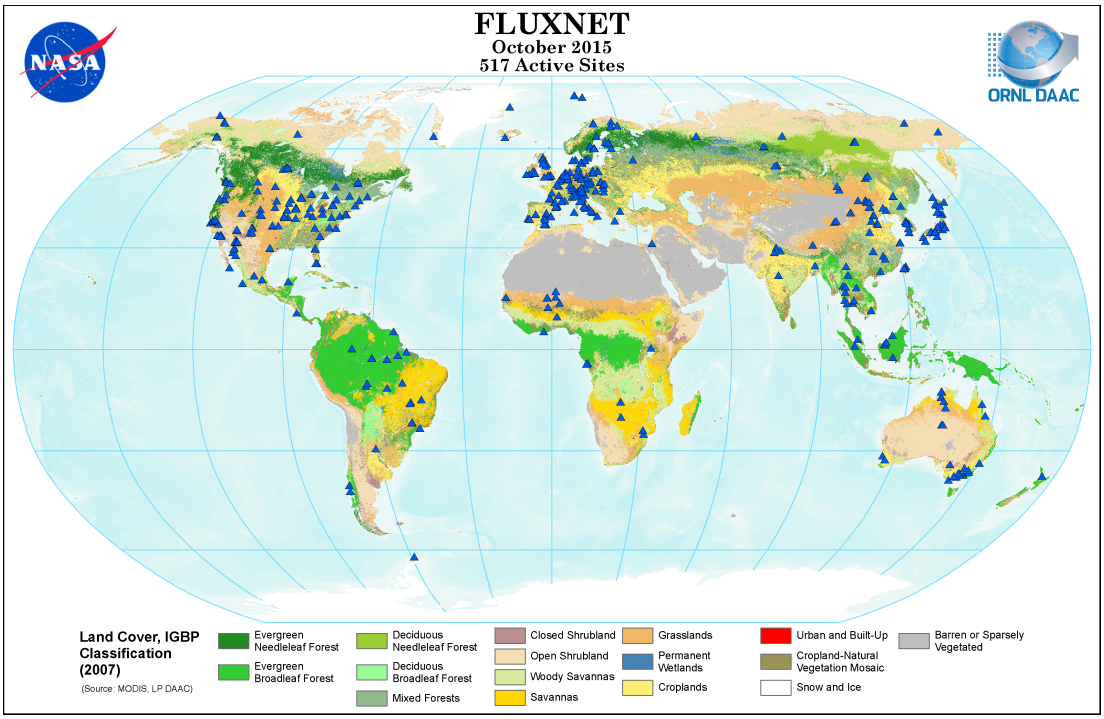


Figure 5.1: World map distribution of FLUXNET sites in 2015. Source: `https://daac.ornl.gov/FLUXNET/guides/Fluxnet_site_DB.html`

shown in Chapter 4 and (Kwon et al., 2016), where the CMAGEP reduced the length of a multivariate linear methane ecosystem exchange model by $\approx 80\%$ after improving model calibration and introducing novel non-linear elements.

Although the above mentioned studies can be considered valuable from the perspective of introducing novel modelling concepts and revealing interesting local CO_2 exchange responses to environmental factors based only on reconstructed signals from measurements, the studies were nevertheless based on single site records of such measurements. In these circumstances it is difficult to make speculations regarding the possibility of expanding the local models to a general R_{eco} model across sites and ecosystem types.

The following study investigates this and other aspects due to access to an important database containing measurements of relevant factors influencing the global terrestrial carbon cycle. The work shown here is based on a set of "open-access" measurements that were recorded at various spatial locations over 112 FLUXNET long-term monitoring sites. The FLUXNET database gathers data from eddy covariance (EC) biogeochemical flux towers globally distributed as shown in Fig. 5.1. Previous work in the field already indicates variation of respiration responses to environmental drivers with latitudinal change (Luo et al., 2015; Mahecha et al., 2010; Shao et al., 2015) and in this chapter such variations are further studied using the novel automated modelling framework, CMAGEP.

Here, one aim is to obtain a more detailed picture of the main drivers of the terrestrial carbon cycle at the single site level as captured in the network of global spatially distributed sites. Furthermore, the present study aims to construct this picture for the larger spatial levels as well, such as the regional and global levels. The presence of distinguishable patterns in the CMAGEP automatically constructed model structures over a multitude of sites is studied. Patterns in the strength of the models in re-capturing signals and generating accurate predictions are explored as well. Once such patterns are revealed, possible links to climate or vegetation type distribution are assessed.

As previously done in Chapter 2, the CMAGEP automatically discovered models were compared for all sites with a set of established literature models, in order to understand if the CMAGEP models managed to generate novel structural understanding or if they re-confirm some of the knowledge regarding the functioning of R_{eco}.

Finally, the possibility of deriving a single CMAGEP model formulation and parametrisation fitting all 112 studied FLUXNET sites is explored as well as the possibility to extrapolate and simulate carbon fluxes over the entire globe for one specific year. Once simulations are generated, the total predicted flux magnitude is compared to results from other independent studies.

The main purpose of the work shown in this chapter is to understand whether an automated modelling framework such as CMAGEP can lead to the discovery of relevant model structures over a large variety of climate and ecosystem type distribution and if these models can capture and reveal interesting patterns of the terrestrial carbon flux over the studied climate and ecosystem types.

5.2 Data and Methods

5.2.1 Data

In recent years the widespread use of the eddy covariance (EC) methodology has led to a large increase in data describing terrestrial land surface exchanges (Baldocchi et al., 2001).

FLUXNET is an international network of EC sites with data processed according to standardized protocols (Luyssaert et al., 2009). The EC time-series data from FLUXNET provide rich insights into exchanges of water, energy and CO_2 across a range of biomes and timescales.

This study is based on data from 112 FLUXNET that contain at least 1 year of daily measurements.

In the available datasets, candidate drivers were defined as instances of sunlight induced fluorescence, (Sif_{ms}) index from the previous day, given as mean seasonal cycle in order to remove high frequency, air temperature at the site (T_{air}), soil temperature

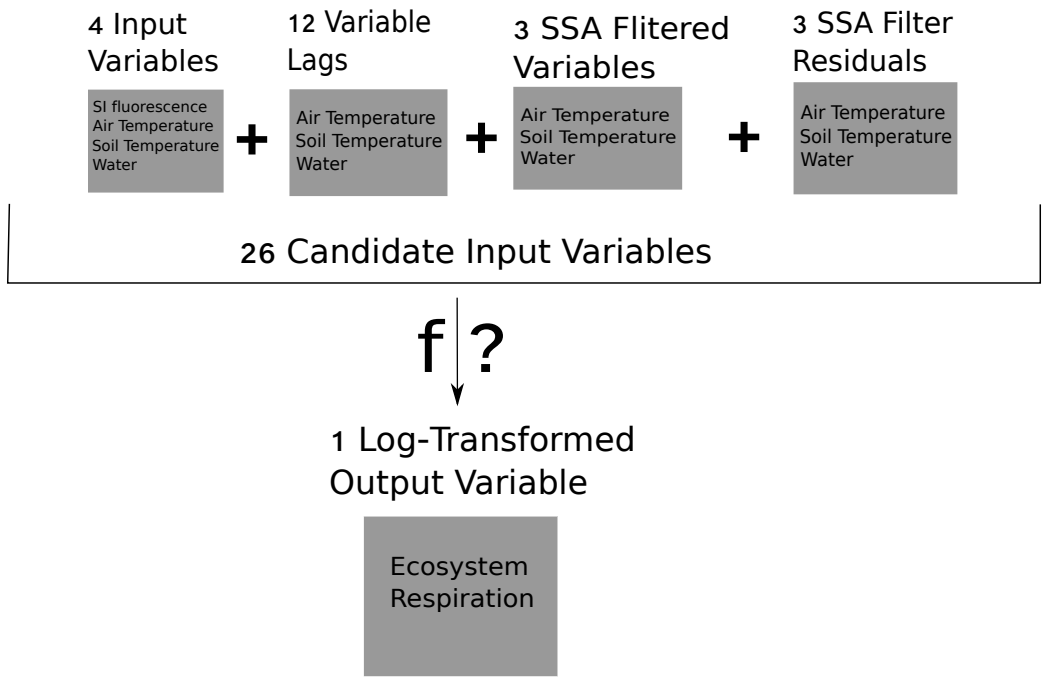


Figure 5.2: Symbolic regression modelling set-up for each CMAGEP run with settings specified in Tab. 5.1 for 112 studied FLUXNET sites.

at the site (T_{soil}), an index for soil water availability *Wai*. The dependent variable, or target, is defined as the R_{eco} flux.

The R_{eco} measurements used in this study were obtained by applying the eddy covariance methodology in order to capture night-time Net Ecosystem Exchange (NEE), followed by flux separation based on the work of (Reichstein et al., 2005). Further details regarding flux measurements for all the FLUXNET used variables are given in (Baldocchi, 2003) and in Chapter 2 of this thesis. The Wai variable is a model product that approximates the soil water availability based on precipitation measurements, obtained as described in Jung et al. (2011). The SIF (Frankenberg et al., 2013) measurement is a satellite product that approximates fluorescence of the surface vegetation and is given as a grid value. For all mentioned candidate drivers lags at 1, 2, 4 and 6 days were generated and added to input, as well as smoothed time series for T_{air}, T_{soil}, *Wai* using a Singular Spectrum Analysis (SSA, Broomhead and King (1986)) filtering over 90 days with Buttlar et al. (2014) implementation as described in (Ilie et al., 2017) and in Chapter 2. Lastly, the high frequency residuals obtained after smoothing are included as independent inputs as well.

The final problem is then modelled as a symbolic regression of 26 candidate variables mapped to 1 target (Fig 5.2).

Previous studies show that R_{eco} tends to exhibit high variability on daily scale,

Parameter	
Number of chromosomes	1000
Number of genes	3
Head length	5
Functions	$+, -, /, \times, x^y, \sqrt{}, \ln, \exp$
Terminals	24 candidate variables–specified in Data Section
Link function	$\times$
Max run time	1800 seconds
Fitness function	CEM
Selection method for replication	tournament(Coello and Montes, 2002)
Mutation probability	0.2
IS and RIS transpositions probabilities	0.05
Two-point recombination probability	0.3
Inversion probability	0.05
One point recombination probability	0.4
Number of individuals to optimize	20
Time to start optimization	0
Maximum CMA-ES iterations	50

Table 5.1: CMAGEP settings for each of the 50 independent runs per site.

which would make learning a regression with a GP based approach difficult, so the target was log-transformed for all sites when training the CMAGEP models repeating the procedure shown in Chapter 2 and (Ilie et al., 2017). For all 112 sites, the observations related to candidate drivers were unaltered and given as input in all independent 50 CMAGEP runs per site.

5.2.2 CMAGEP

The approach chosen for the automatic construction of models is the CMAGEP, a hybrid genetic programming and evolutionary strategy approach as developed and proposed by the author in (Ilie et al.) and in Chapter 3, where details concerning design, implementation and performance are discussed.

All settings used for generating CMAGEP models for all 112 FLUXNET sites are given in Table 5.1.

5.2.2.1 Fitness function

The selection in the present CMAGEP evolutionary process is based on the Corrected for complexity Efficiency of Modelling (CEM, Eq.2.2.3) fitness function, introduced

for the first time by the author in (Ilie et al., 2017).

$$\text{Fitness Function} = \text{CEM}(\phi) = \sqrt{(1-\text{MEF}(\phi))^2 + (\frac{P_{(\phi)}}{P_{max}})^2 + (1-\text{SE}(\phi))^2} \quad (5.2.1)$$

$$\text{MEF}(\phi) = 1 - \frac{\sum_{i=1}^{n}(o_i - p_i)^2}{\sum_{i=1}^{n}(o_i - \bar{o})^2} \quad (5.2.2)$$

$$P_{max} = gN \times h \quad (5.2.3)$$

$$SE(X) \; = \; -\sum_{i=1}^{N} p_i \ln[p_i] \; . \quad (5.2.4)$$

Where, ϕ is an evolution individual, MEF is the Nash–Sutcliffe Modelling Efficiency (MEF) coefficient (Bennett et al., 2010; Nash and Sutcliffe, 1970) (5.2.2), with o_i observed target value at time step i, $\bar{o}$ mean of observed target values, and p_i predicted value at time step i.

$P_{(\phi)}$ and P_{max} are the number of parameters for the current individual and the maximum number of parameters for any individual with the current CMAGEP settings.

gN and h are the number of genes making an individual and the head length.

SE is the normalized Shannon Entropy (SE) value computed for the ordinal patterns of the residuals as defined in the work of (Sippel et al., 2016) and further detailed in (Ilie et al., 2017) with $X = \{p_i; i = 1, \ldots, N\}$ denoting a probability distribution with $\sum_{i=1}^{N} p_i = 1$ and N possible states.

5.2.3 Automated R_{eco} model extraction by CMAGEP: experiment design

Although there were more sites available in the FLUXNET database, after data quality and minimum quantity filtering, 112 measurement sites containing carbon cycle measurements at daily scale of various lengths were selected.

For each site the CMAGEP approach was used to automatically build symbolic regressions to model daily R_{eco} flux.

The CMAGEP SR models were build based on data from each site as follows:

1. The total data set is split into two sets, a training set with 75 % of the total available data instances and a validation set with 25%;
2. The train-validation split is done 50 times, resulting in a total of 50 subset pairs for each site;
3. A CMAGEP model is constructed based on each training set with Table 5.1, resulting in a total of 50 model structures for each site;
4. The fitness function mean is computed over all 50 validation subsets;
5. The structure with the best fitness mean value over all subsets is selected as best structure for the site.

All 50 CMAGEP runs performed for each of the 112 site solutions have a final CMA-ES optimization step, where the maximum iteration number allowed for reaching an optimum is no longer so drastically constricted as in Tab. 5.1.

The most significant structures are selected and reported. Since MEF is a measure allowing for easier understanding and interpretation of goodness of fit and prediction capacity, associated MEF values were computed over the cross validations data sets and reported, and not the associated CEM values as used in the learning process .

5.3 Results

The CMAGEP for SR framework was used to generate models for 112 FLUXNET sites. The 112 generated model structures were subjected to fitness based selection leading to finding a single CMAGEP model structure with 112 parametrisations for 112 local site conditions. Patterns in goodness of fit and mathematical structures of the CMAGEP models were studied and conclusions regarding underlying signals over different climate and vegetation types could be drawn. The CMAGEP models were compared with a set of established models for R_{eco} in the ecology community with the CMAGEP derived models always showing modelling performances in a very close range or surpassing those of established models. Finally, a single CMAGEP model with a unique set of parameters was selected for simulating total daily R_{eco} fluxes globally for one specific year. Remarkably, due to the internal structure of the single CMAGEP model, the simulated daily R_{eco} fluxes did not need input drivers available only at the discrete 1112 FLUXNET sites but were generated over a much finer grid, covering a large global spatial distribution. The total R_{eco} flux was within reasonable orders of magnitude from results in other independent studies (Zscheischler et al., 2017).

5.3.1 CMAGEP models for terrestrial respiration fluxes

Models were generated for describing R_{eco} responses to external, biotic and abiotic drivers for each of the 112 studied FLUXNET measurement sites by applying CMAGEP to real biogeochemical flux measurements with Tab. 5.1 settings.

The CMAGEP generated models showed a large degree of internal structure variation over the 112 studied sites, as well as large variation in prediction capacities, as measured by mean validation MEF values 5.9.

However, since the 112 model structures were only optimized for the local conditions of each site, it would be very difficult to say whether a certain model structure type would appropriately capture general trends of the R_{eco} responses at all sites. With one of the main interests of this study being that of understanding the generality trait of a model structure, that is the capacity of a model built from measurements captured at a certain site to represent the responses of R_{eco} fluxes to candidate drivers at other sites, each individual CMAGEP model structure was re-optimized with CMA-ES for the local conditions of each of the remaining 111 sites not seen during training.

Mean fitness values were computed for each of 112 CMAGEP model structures and their 112 optimized site–parametrisations. The CMAGEP model with the highest mean fitness value over the 112 parametrisations was selected and reported as the best over-all-sites model structure, or the global model structure.

The CMAGEP model structure with highest mean MEF value recorded 0.52 at validation on 112 site–parametrisations and is reported in Eq. 5.3.1. The model structure in Eq. 5.3.1 is called from here onwards the Global Respiration Model Structure (GRMS).

$$\text{GRMS} := R_{eco}(t) = \exp\left(\alpha Sif_{ms}(t) + \beta T_{air}(t) + \gamma\right) \tag{5.3.1}$$

where t is time in days, and α, β and γ are parameters optimized locally by the CMA-ES component of CMAGEP.

Figure 5.3 illustrates the prediction capacity of the individual site parametrisation of the GRMS and that of each of the locally trained CMAGEP and shows that the general trend is quite close to the 1:1 line with very few exceptions from the dry and tropical climate types. Such a similarity in over-all performance indicates that although initially different model structures might have been generated by CMAGEP for each of the 112 sites, there is an underlying signal that can be sufficiently well described by GRMS over all sites when local CMA-ES optimizations are performed.

The similarity in prediction capacity for different structures when local parametrisations are done sufficiently well recalls the results obtained previously in Chapters 2 and 3 regarding the equifinality of models characterizing respiration fluxes. It is possible that although structurally they might seem different the model formulations describe a similar signal over a certain domain with the optimal set of parameters discovered by the CMA-ES component of the CMAGEP.

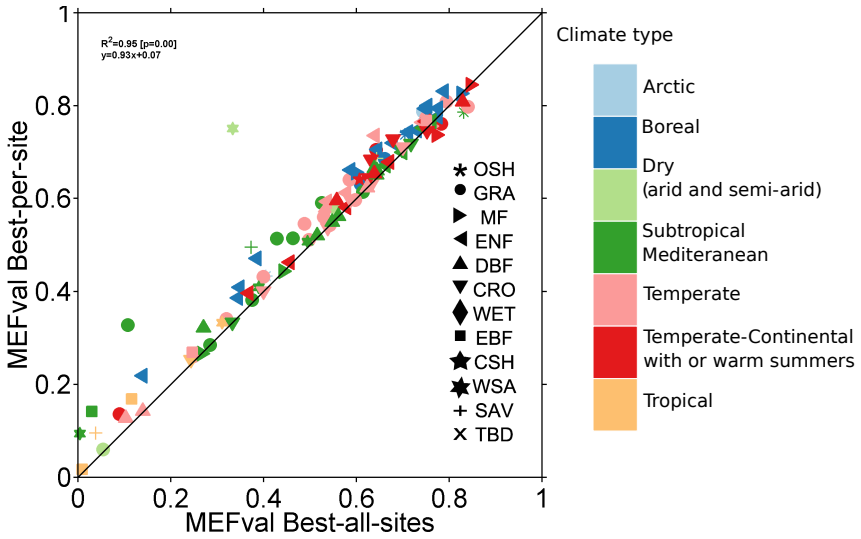


Figure 5.3: Comparing values of mean MEF computed over the validation samples between Best-per-all-sites structure and Best per each site structure over all climate and vegetation PFT types. All 112 CMAGEP models and their 112 parametrisations were obtained after 50 independent runs with the settings given in Table 5.1

5.3.2 Detailed analysis of selected sites

Since the manuscript would become far too dense with the illustration of all CMAGEP models and their capacity to recreate the original R_{eco} signal for each site in time series model fit comparisons, 6 relevant cases of modelling were selected and included in a detailed analysis: a pair taken from the best modelled sites, two averagely modelled sites and two poorly modelled sites.

Figures 5.4 and 5.5 illustrate a selection of modelling situations for terrestrial respiration fluxes. On the first row are illustrated sites that have a good representation of the fluxes by both the locally trained models and the GRMS with local parametrisations, with validation MEF values between 0.79 and 0.84. On the second row sites for which the models record average MEF values between 0.51 and 0.66 and on the third row, site for which it is visible that only the mean values are represented by both the local and global models, with MEF values very close to 0. The Fig. show once more that for certain types of climate and PFTs the CMAGEP models are capable of accurately reconstruct the present signal, especially if a strong seasonality is present, and that for specific sites, just using the mean to model the respiration flux values might be just as useful, such cases being present mostly in tropical climate types.

In Fig. 5.6 and 5.7 the CO_2 flux values observed at the site are compared with the predicted flux values by the best over all sites model with locally optimized parameters

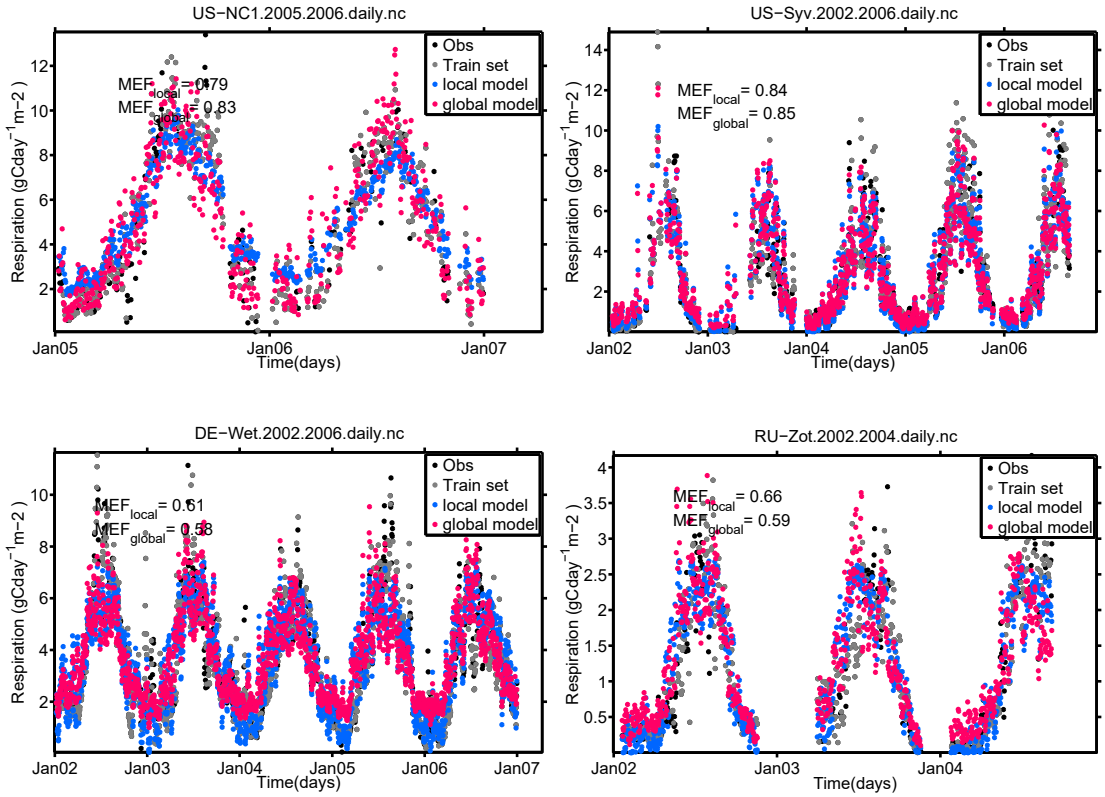


Figure 5.4: (A) Observed daily ecosystem CO_2 outgoing flux, and the fluxes modelled by the best CMAGEP model at the single site and modelled by the CMAGEP model over all-sites in terms of mean MEF. All single site CMAGEP models were selected after 50 independent runs with settings given in Table 5.1. A set of the best, mean and worst modelled sites. The first 2 letters in the titles indicate the country where the site is found and can point to climate type.

for the same set of selected sites as shown in Fig. 5.4 and 5.5. The figures illustrate not the modelling capacity of the GRMS for the selected sites, but also an underestimation of high fluxes, especially for the better modelled sites. The underestimation is even stronger with lower values of *Wai*, meaning higher water stress for the local vegetation.

5.3.3 Patterns in structure types

The interest of the study was not only in determining a unique model structure with a high prediction accuracy, but also in understanding and interpreting the emerging structures over all sites.

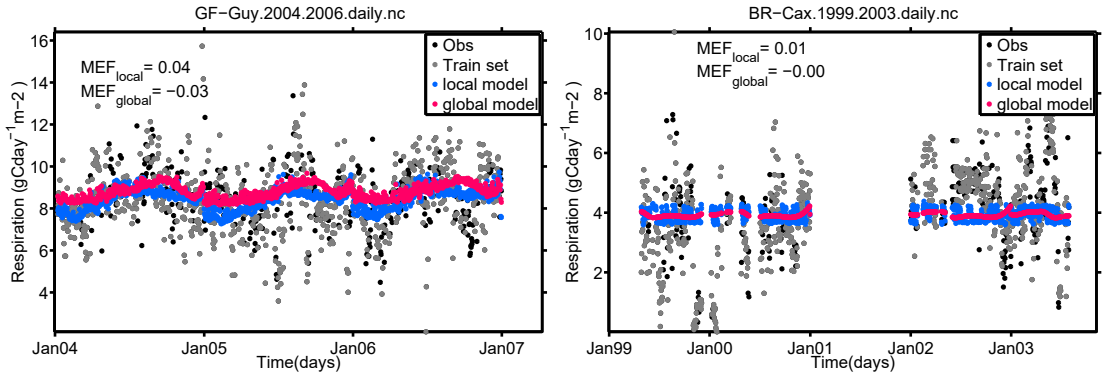


Figure 5.5: (B) Observed daily ecosystem CO_2 outgoing flux, and the fluxes modelled by the best CMAGEP model at the single site and modelled by the CMAGEP model over all-sites in terms of mean MEF.)

In order to visualise the diversity of the model structures performing better globally, a set of 10 best global CMAGEP model structures for R_{eco} responses to environmental drivers was constructed. The set of 10 best models over all sites was built based on selection of mean MEF values computed over 112 FLUXNET sites for the 112 CMAGEP models and their individual parameter sets optimized for the local conditions at each site.

The set of 10 best structures is studied due to the possibility of determining an underlying presence of structural patterns in CMAGEP models for R_{eco} responses to external drivers. and for that purpose the models with the 10 best mean MEF scores were selected and reported in Table 5.2, ordered by mean MEF values over the 112 studied FLUXNET sites.

From the Table 5.2, a clear pattern could be observed, with the structure in equation 5.3.1 appearing in the first 7 out of the 10 reported models, and an exponential response of R_{eco} to Sif_{ms} and T_{air} always being present.

For exploring possible links between CMAGEP extracted model structure types and the climate types associated to the sites where the measurements for which models have been generated, mean MEF values per climate type were computed for each of the 112 models optimized to local site conditions. The mean MEF value for each climate type was also determined for the GRMS and its locally optimized parameter sets.

The local CMAGEP model structure types with the highest mean MEF values per climate type were selected and reported. The selection was done as well for the GRMS and the mean MEF values at each climate type.

When the best structures per climate type were determined based on mean MEF values at validation for all the sites belonging to a certain climate, as reported in Table 5.3, for similar climate types, such as the Arctic and Boreal ones, similar model structures had the best modelling accuracy. It was especially interesting that in these

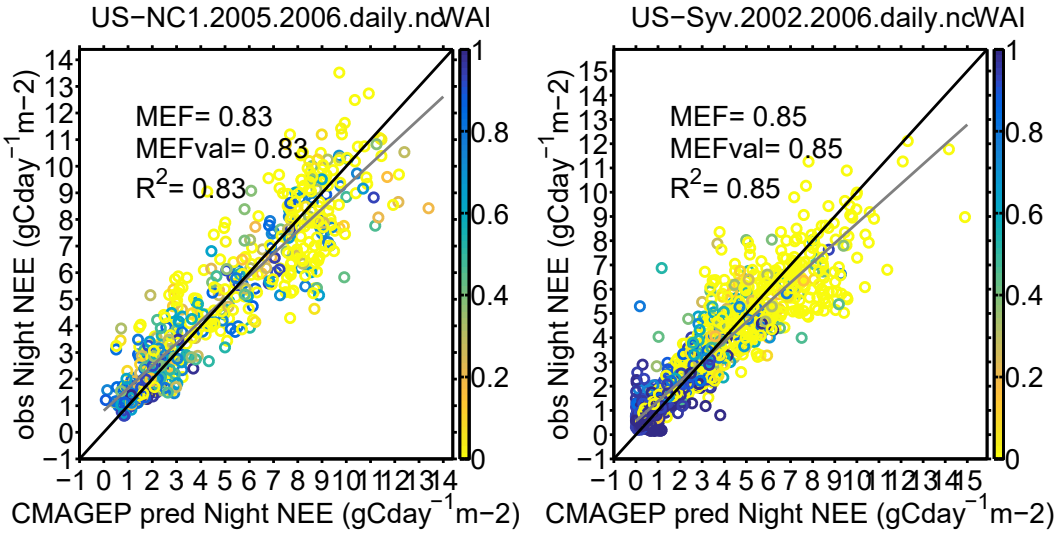


Figure 5.6: (A) Observed daily ecosystem CO_2 outgoing flux against the fluxes modelled by the best CMAGEP model at the single site and modelled by the CMAGEP model over all-sites in terms of mean MEF and their relation to water availability index (WAI). A set of best and averagely modelled sites.

climate types the main features selected were only referring to temperature change and that other drivers did not seem to impact the fluxes as strongly.

Similar structures were obtained for sites from Temperate, Temperate-Continental and Mediterranean climate types, with better prediction accuracy for the model structures for sites from Temperate climate. These sites present the same structure as the global model structure (Eq. 5.3.1). The most complex structures per climate type were selected for climate types that are either water stressed, such as the Dry climate, or in climates that do not show a strong seasonality such as the Tropical climate.

The recorded mean MEF values for the best per climate model and the global model structure are in a close range, with notable significant differences is in Dry climate sites. In the Dry climate sites possibly due to water stress, temperature and vegetation descriptors only might not be sufficient to capture the local dynamics.

Across all climate types the lowest capacity of CMAGEP models to accurately capture R_{eco} flux changes is in the tropics. This is possibly due to a reduced flux seasonality in the tropics that cannot be captured by the candidate drivers that inherently describe seasonal components.

Patterns in possible links of CMAGEP models to climate and PFT combinations were also studied with the previously described selection done at the detailed aggregation of both climate and vegetation type (plant functional type, PFT).

Mean MEF values were computed over the validation sets for all sites in a certain climate PFT type pair. The results were reported in Tab. 5.4 and 5.5.

Although the emerging patterns were no longer clear for the climate PFT pairs, it

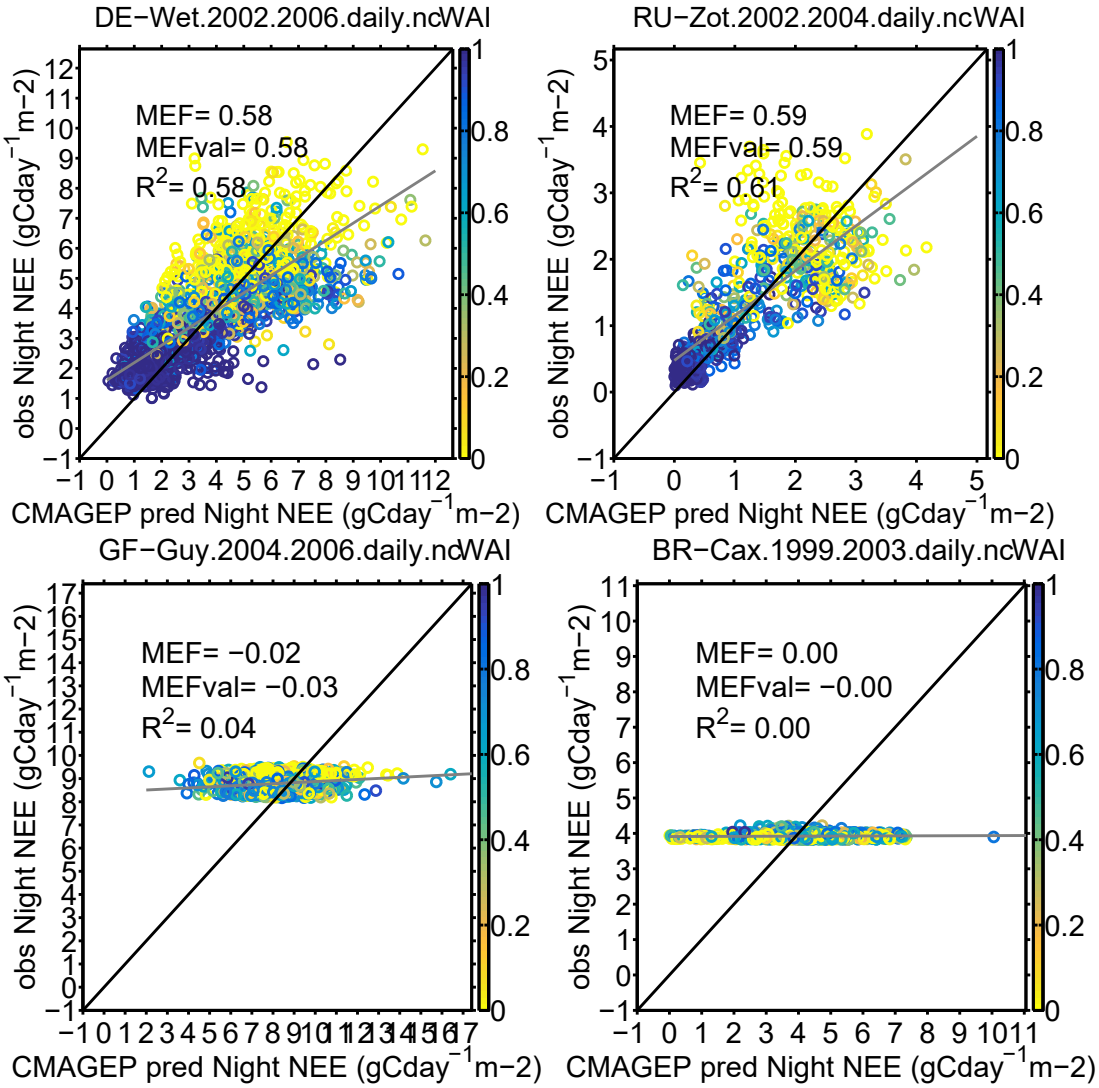


Figure 5.7: (B) Observed daily ecosystem CO_2 outgoing flux against the fluxes modelled by the best CMAGEP model at the single site and modelled by the CMAGEP model over all-sites in terms of mean MEF and their relation to water availability index (WAI). A set of poorly modelled sites.

Table 5.2: Best 10 CMAGEP structures over all sites in terms of mean MEF at validation selected from 112 parametrisation sets for 112 site models generated after 50 independent CMAGEP runs with settings given in Table 5.1.

No.	Structure	$\overline{\text{MEF}}$	σ
1	$R_{eco}(t) = \exp(\alpha Sif_{ms}(t) + \beta T_{air}(t) + \gamma)$	0.52	0.02
2	$R_{eco}(t) = \exp(\alpha Sif_{ms}(t) + \beta T_{air}(t) + \gamma)$	0.52	0.02
3	$R_{eco}(t) = \exp(\alpha Sif_{ms}(t) + \beta T_{air}(t) + \gamma)$	0.52	0.02
4	$R_{eco}(t) = \exp(\alpha Sif_{ms}(t) + \beta T_{air}(t) + \gamma)$	0.52	0.02
5	$R_{eco}(t) = \exp(\alpha Sif_{ms}(t) + \beta T_{air}(t) + \gamma)$	0.52	0.02
6	$R_{eco}(t) = \exp(\alpha Sif_{ms}(t) + \beta T_{air}(t) + \gamma)$	0.52	0.02
7	$R_{eco}(t) = \exp(\alpha Sif_{ms}(t) + \beta T_{air}(t) + \gamma)$	0.52	0.02
8	$R_{eco}(t) = \exp(\alpha Sif_{ms}(t) \times log(\alpha Sif_{ms}(t)) + \beta T_{air}(t))$	0.51	0.00
9	$R_{eco}(t) = \exp(\theta log(\alpha Sif_{ms}(t)) + \beta T_{air}(t) + \gamma)$	0.51	0.00
10	$R_{eco}(t) = \exp(\alpha Sif_{ms}(t) + \beta_4 T_{air}(t-4) + \gamma)$	0.47	0.03

could still be seen that for similar climate and PFT pair types, similar model structures were performing better.

Once more, the mean MEF validation values for the GRMS were in a similar range from MEF values computed for the best structure selected for each climate-PFT type. This was confirmed by the fact that over all climate-PFT combination types the two MEF averages were 0.49 for the GRMS and 0.52 for the best per climate-PFT combination structure respectively.

Table 5.3: Best structures per climate type. Mean MEF values are reported for the best CMAGEP model over all sites in each climate type ($\overline{\text{MEF}}$ and σ) and for the GRMS, the best CMAGEP model structure over all sites ($\overline{\text{MEF}}_x$ and σ_x). All CMAGEP models were obtained after 50 independent runs with Tab. 5.1 settings at each site in the set of 112 studied FLUXNET sites.

Climate type	sites	Structure	$\overline{\text{MEF}}$	σ	$\overline{\text{MEF}}_x$	σ_x
Arctic	2	$exp(\beta T_{air}(t)+\beta_4 T_{air}(t-4)+\gamma)$	0.60	0.02	0.57	0.02
Boreal	21	$exp(\beta T_{air}(t)+\beta_4 T_{air}(t-4)+\gamma)$	0.67	0.02	0.63	0.02
Dry (arid and semi arid)	4	$exp(\beta_4 T_{air}(t-4)-\kappa exp(-\alpha Sif_{ms}(t))+\gamma)$	0.50	0.03	0.39	0.02
SubTropical-Mediterranean	30	$exp(\alpha Sif_{ms}(t)+\beta T_{air}(t)-\gamma)$	0.47	0.02	0.47	0.02
Temperate	30	$exp(\alpha Sif_{ms}(t)+\beta T_{air}(t)-\gamma)$	0.54	0.02	0.54	0.02
Temperate-Continental with hot or warm summers	18	$exp(\alpha Sif_{ms}(t)+\beta T_{air}(t)-\gamma)$	0.63	0.02	0.63	0.02
Tropical	7	$exp(\alpha Sif_{ms}(t)\times log(\beta T_{air}(t))+\omega Wai(t)+\gamma)$	0.10	0.01	0.10	0.01

Table 5.4: Best structures per climate and PFT type (A). Mean MEF values are reported for the best CMAGEP model over all sites in each climate PFT type pair ($\overline{\text{MEF}}$ and σ) and for the GRMS, the best CMAGEP model structure over all sites ($\overline{\text{MEF}}_x$ and σ_x). All CMAGEP models were obtained after 50 independent runs with Tab. 5.1 settings at each site from the set of 112 studied FLUXNET sites.

Climate type	PFT	sites	Structure	$\overline{\text{MEF}}$	σ	$\overline{\text{MEF}}_x$	σ_x
Arctic	GRA	1	$exp(-\kappa exp(-\beta T_{air}(t)) + \gamma)$	0.79	0.00	0.74	0.00
Arctic	OSH	1	$exp(-\omega_2 Wai(t-2)^7 + \beta_4 T_{air}(t-4))$	0.43	0.00	0.40	0.00
Boreal	OSH	2	$exp(\beta T_{air}(t) + \beta_4 T_{air}(t-4) + \gamma)$	0.78	0.01	0.73	0.01
Boreal	MF	2	$exp(\beta T_{air}(t) + \beta_4 T_{air}(t-4) + \gamma)$	0.75	0.05	0.71	0.05
Boreal	GRA	2	$exp(\delta_2 T_{soil}(t-2))$	0.66	0.01	0.63	0.01
Boreal	ENF	15	$exp(\beta T_{air}(t) + \beta_4 T_{air}(t-4) + \gamma)$	0.64	0.07	0.61	0.07
Dry (arid and semi arid)	WSA	1	$exp(-\alpha Sif_{ms}(t) + \omega Wai(t) + \theta log(\alpha Sif_{ms}(t)))$	0.75	0	0.33	0
Dry (arid and semi arid)	SAV	1	$exp(-\alpha Sif_{ms}(t) - \omega Wai(t) + \theta log(\alpha Sif_{ms}(t)) - \gamma)$	0.69	0	0.64	0
Dry (arid and semi arid)	GRA	2	$exp(\beta_4 T_{air}(t-4) - \kappa exp(-\alpha Sif_{ms}(t)) + \gamma)$	0.32	0.12	0.3	0.12
SubTropical-Mediterranean	OSH	1	$exp(\beta T_{air}(t) + \beta_4 T_{air}(t-4) + \gamma)$	0.86	0	0.83	0
SubTropical-Mediterranean	ENF	4	$exp(\alpha Sif_{ms}(t) + \beta T_{air}(t) - \gamma)$	0.72	0.01	0.72	0.01
SubTropical-Mediterranean	CRO	3	$exp(\beta T_{air}(t) + \theta log(\alpha Sif_{ms}(t)) + \gamma)$	0.59	0.07	0.58	0.07
SubTropical-Mediterranean	DBF	7	$exp(\beta T_{air}(t) + \theta log(\alpha Sif_{ms}(t)) + \gamma)$	0.55	0.04	0.55	0.04
SubTropical-Mediterranean	SAV	1	$exp(\alpha Sif_{ms}(t) \times log(\beta T_{air}(t)) + \omega Wai(t) + \gamma)$	0.5	0	0.37	0
SubTropical-Mediterranean	GRA	7	$exp(\beta T_{air}(t) + \theta log(\alpha Sif_{ms}(t)) + \gamma)$	0.4	0.06	0.4	0.06
SubTropical-Mediterranean	MF	2	$exp(-\alpha Sif_{ms}(t) + \beta T_{air}(t))$	0.35	0.04	0.35	0.04
SubTropical-Mediterranean	WSA	3	$exp(\alpha Sif_{ms}(t) + \omega Wai(t) + \beta_4 T_{air}(t-4) - \gamma)$	0.32	0.09	0.3	0.09
SubTropical-Mediterranean	EBF	2	$exp(\alpha Sif_{ms}(t) + \omega Wai(t) + \beta_4 T_{air}(t-4) - \gamma)$	0.05	0.01	0.01	0.01

Table 5.5: Best structures per climate and PFT type (B). Mean MEF values are reported for the best CMAGEP model over all sites in each climate PFT type pair ($\overline{\text{MEF}}$ and σ) and for the GRMS, the best CMAGEP model structure over all sites ($\overline{\text{MEF}}_x$ and σ_x). All CMAGEP models were obtained after 50 independent runs with Tab. 5.1 settings at each site from the set of 112 studied FLUXNET sites.

Climate type	PFT	sites	Structure	$\overline{\text{MEF}}$	σ	$\overline{\text{MEF}}_x$	σ_x
Temperate	MF	1	$exp(\beta T_{air}(t) + \theta log(\alpha Sif_{ms}(t)) + \gamma)$	0.77	0	0.76	0
Temperate	ENF	4	$exp(-\kappa exp(-\beta T_{air}(t)) + \gamma)$	0.68	0.03	0.62	0.03
Temperate	WET	1	$exp(\beta T_{air}(t) + \beta_4 T_{air}(t-4) + \gamma)$	0.59	0	0.53	0
Temperate	CRO	7	$exp(\alpha Sif_{ms}(t) + \beta T_{air}(t) - \gamma)$	0.56	0.03	0.56	0.03
Temperate	GRA	12	$exp(\alpha Sif_{ms}(t) + \beta T_{air}(t) - \gamma)$	0.55	0.06	0.55	0.06
Temperate	EBF	2	$exp(\beta T_{air}(t) + \beta_4 T_{air}(t-4) + \gamma)$	0.53	0.12	0.5	0.12
Temperate	DBF	3	$exp(\alpha Sif_{ms}(t) + \beta T_{air}(t) - \gamma)$	0.29	0.1	0.29	0.1
Temperate-Continental	MF	2	$exp(\beta T_{air}(t) + \beta_4 T_{air}(t-4) + \gamma)$	0.82	0.02	0.81	0.02
Temperate-Continental	CRO	4	$exp(\beta T_{air}(t) + \beta_4 T_{air}(t-4) + \gamma)$	0.68	0.02	0.67	0.02
Temperate-Continental	DBF	3	$exp(\beta T_{air}(t) + \beta_4 T_{air}(t-4) + \gamma)$	0.68	0.05	0.68	0.05
Temperate-Continental	CSH	1	$exp(\alpha Sif_{ms}(t) \times log(\alpha Sif_{ms}(t)) + \beta T_{air}(t))$	0.64	0	0.61	0
Temperate-Continental	ENF	5	$exp(-\kappa exp(-\beta T_{air}(t)) + \gamma)$	0.57	0.05	0.56	0.05
Temperate-Continental	GRA	3	$exp(\beta T_{air}(t) + \theta log(\alpha Sif_{ms}(t)) + \gamma)$	0.53	0.12	0.5	0.12
Tropical	WSA	1	$exp(\alpha Sif_{ms}(t) \times log(\beta T_{air}(t)) + \omega Wai(t) + \gamma)$	0.33	0	0.31	0
Tropical	CRO	1	$exp(\alpha Sif_{ms}(t) \times log(\alpha Sif_{ms}(t)) + \beta T_{air}(t))$	0.25	0	0.24	0
Tropical	SAV	1	$exp(-\alpha Sif_{ms}(t) + \omega Wai(t) + \theta log(\alpha Sif_{ms}(t)))$	0.09	0	0.04	0
Tropical	EBF	3	$exp(\beta T_{air}(t) + \kappa exp(\omega_2 Wai(t-2)) + \theta log(\alpha Sif_{ms}(t)) - \gamma)$	0.05	0.03	0.03	0.03
Tropical	TBD	1	$exp(\beta T_{air}(t) - \omega_2 Wai(t-2) + \gamma)$	0.01	0	0	0

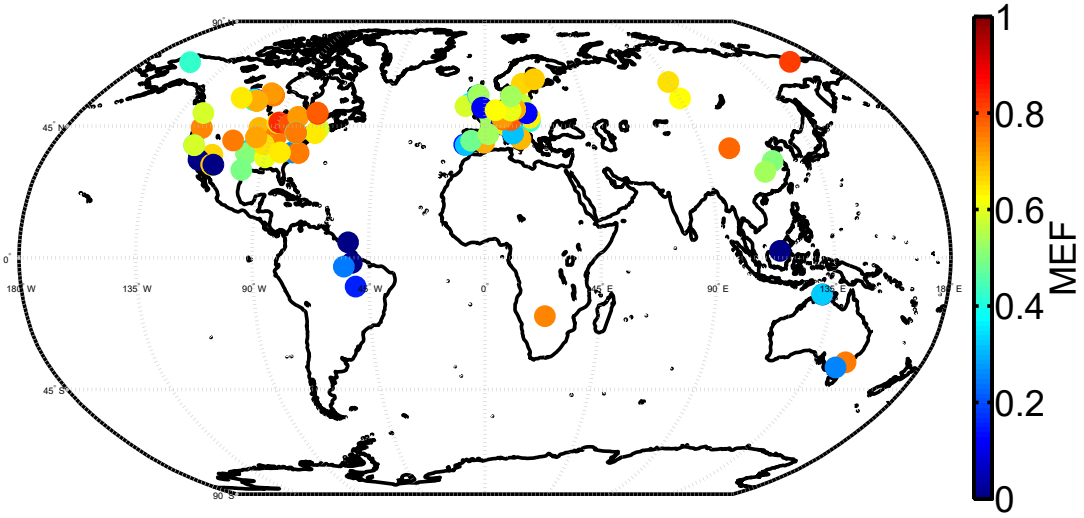


Figure 5.8: World map distribution of MEF at validation for the best structures per each of the 112 studied sites after 50 independent runs of CMAGEP.

5.3.4 Patterns in modelling capacity

In order to investigate if a clear correlation can be made between the capacity of the GEP models to automatically reconstruct the respiration response to possible drivers based on data and local environment conditions at the sites, such as mean annual temperature (MAT), a figure was generated showing the MEF values for each of the 112 CMAGEP models built for the studied sites against the MAT recorded at the site.

A similar figure was generated for the best model structure over all sites by mapping the MEF values for all 112 parametrisations to the corresponding MAT values recorded at the individual sites.

To analyse possible patterns in modelling efficiency relating to climate type and PFT, the figures include information on site classification.

For visualising the global spatial distribution of modelling capacity for the CMAGEP model structures built for the 112 studied FLUXNET sites, and for assessing whether there are sites in certain regions of the world where the respiration flux can be better simulated by the automatically built GEP models, a map of the MEF values corresponding to the structures was produced.

Possible differences in modelling capacities between the 112 structures built independently for each site and the unique solution that performs best over all sites were investigated. Patterns were explored for links between such differences and climate types as well as between modelling capacity differences and climate types–PFT pairs. The best over all 112 studied sites.

When patterns in the capacity of the CMAGEP models to accurately predict the original respiration fluxes for the 112 studied sites were explored, it was found that

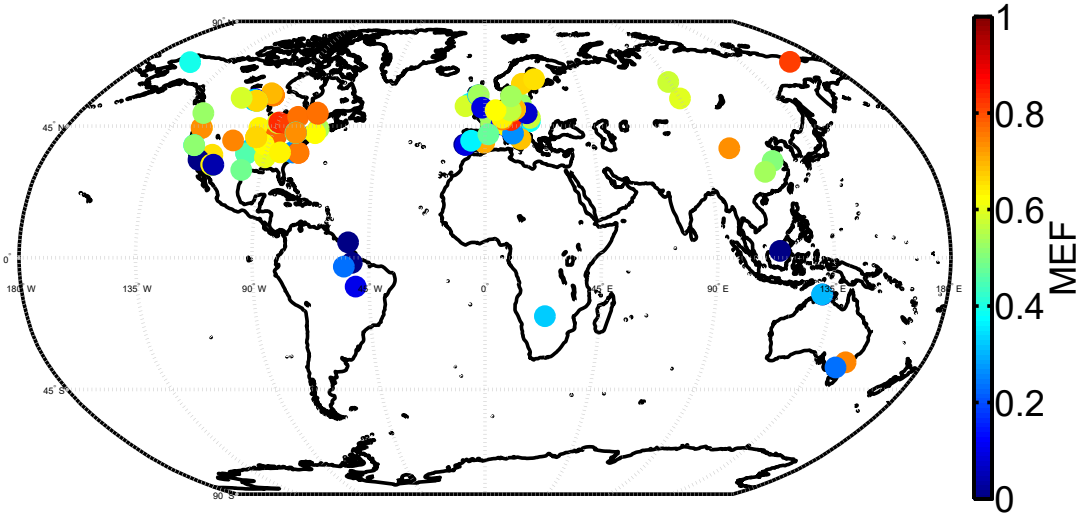


Figure 5.9: World map distribution of MEF at validation for the best structure over all 112 sites after 50 independent runs of CMAGEP.

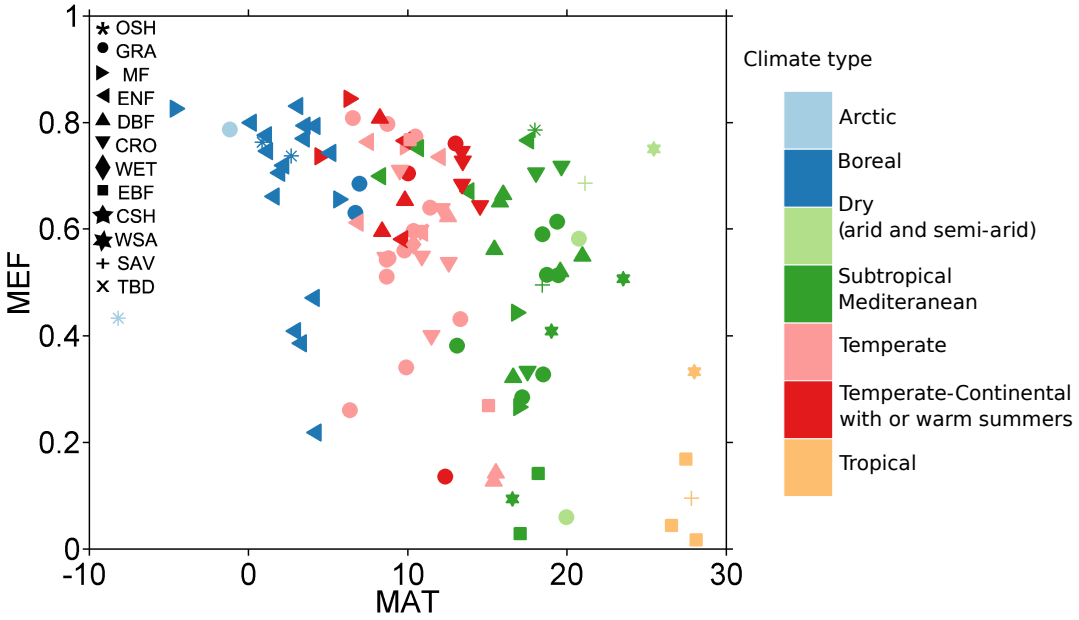
spatially, the sites from the boreal and temperate climates were modelled the best and that the sites in the tropical sites presented the lowest validation MEF values, as seen in Fig. 5.8. A very similar distribution of prediction capacity over the global map could be seen for the 112 parametrisations of the GRMS (Fig. 5.9).

Figures 5.10b and 5.10a confirm the link of modelling capacity with climate type and show furthermore that within a climate type, the respiration fluxes for sites in forests are captured better than others and that the lowest modelling capacities are recorded for sites in grasslands. The same link to climate type, PFT and MEF values is seen when the sites have been aggregated by climate and climate type-PFT (Fig. 5.11). These observed patterns could be related to the magnitude of the respiration fluxes and the seasonality strength in these PFT.

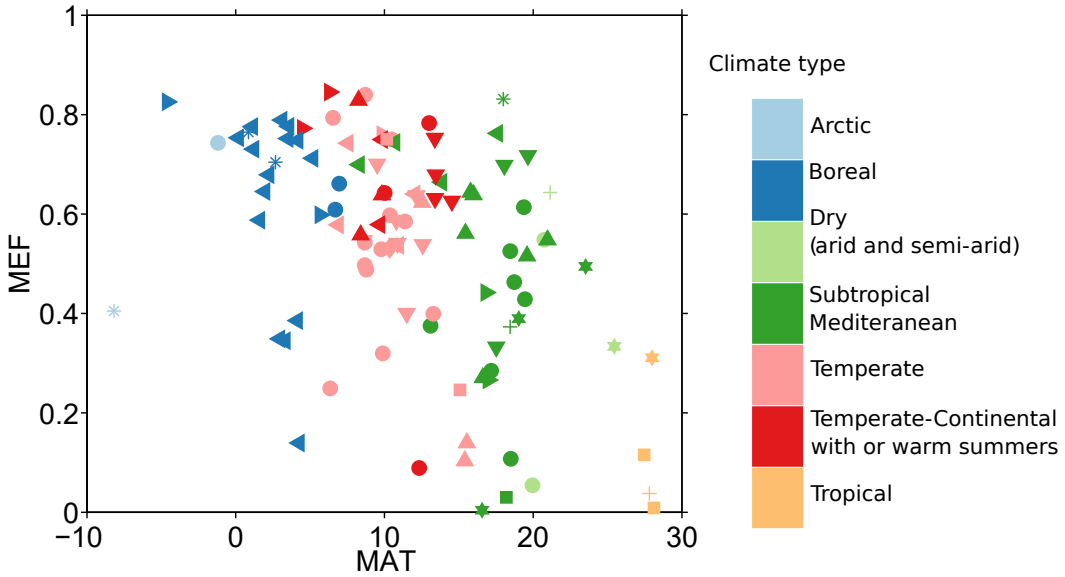
5.3.5 Global solution parametrisations and links to local site environment descriptors

To investigate the main factors determining the parametrisations of the best global solution for the FLUXNET sites, correlations were computed between model structure parametrisations of Eq. 5.3.1 at each site and a set of relevant environment factors measured at the studied sites such as mean annual air and soil temperature (MAT and MAT_{soil}), mean annual precipitation (MAP), mean annual incoming short-wave radiation (Rg), and mean annual vapour pressure deficit (VPD).

Correlation values were reported for the entire range of the measured environment



(a) Best per site CMAGEP structure after 50 independent runs with Tab. 5.1.



(b) Over all-sites CMAGEP best structure after 50 independent runs with Tab. 5.1.

Figure 5.10: **MEF for all sites per climate and vegetation PFT types against environment site descriptors, here Mean Annual Temperature (MAT).**

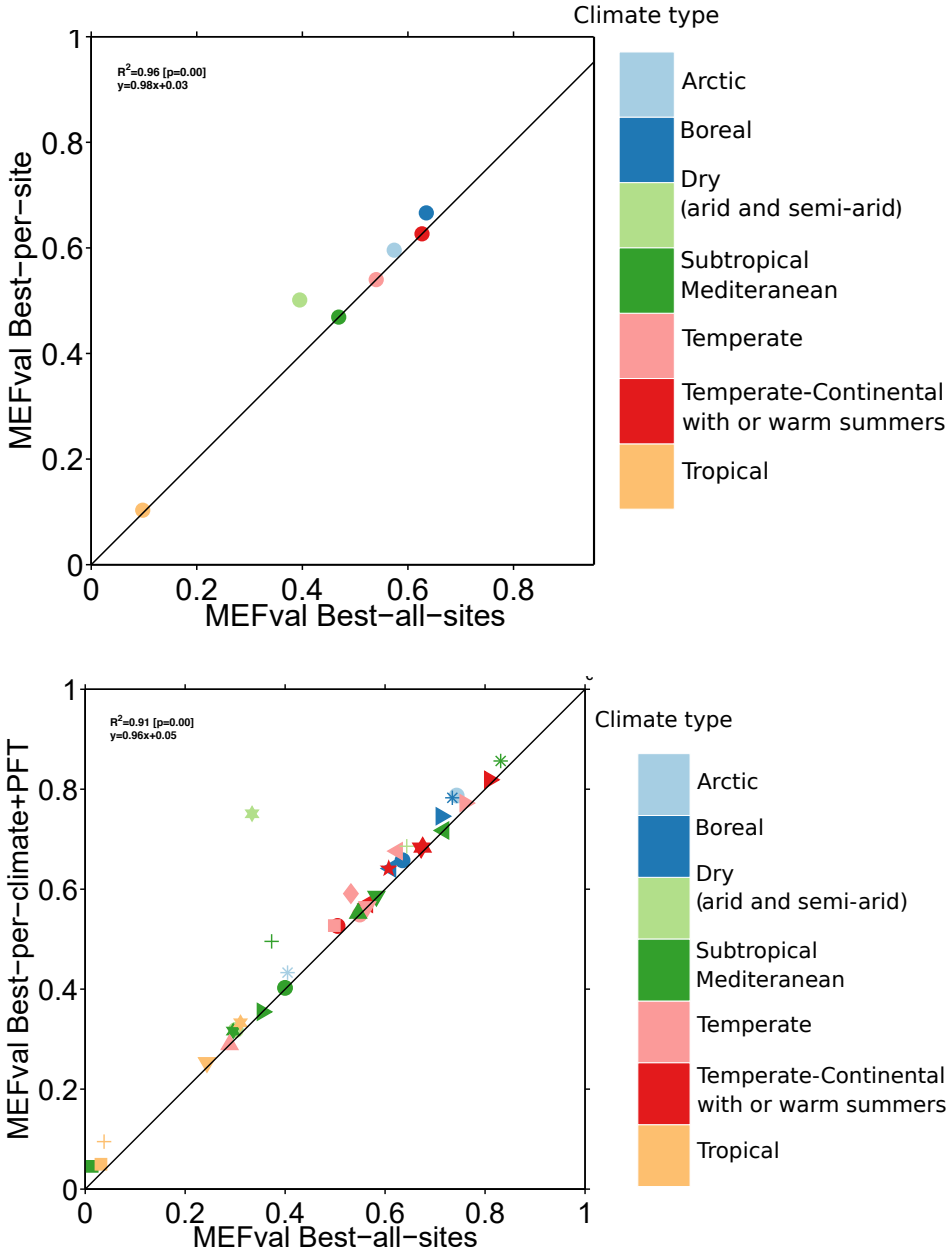


Figure 5.11: Upper: Comparing **MEF validation values aggregated per climate type** between CMAGEP Best per all sites structure and Best per climate types structure. Lower: Comparing **MEF validation values aggregated per climate type and PFT type** between CMAGEP Best per all sites structure and Best per climate type and PFT type structure. The structures were obtained after 50 independent CMAGEP runs with the settings given in Table 5.1.

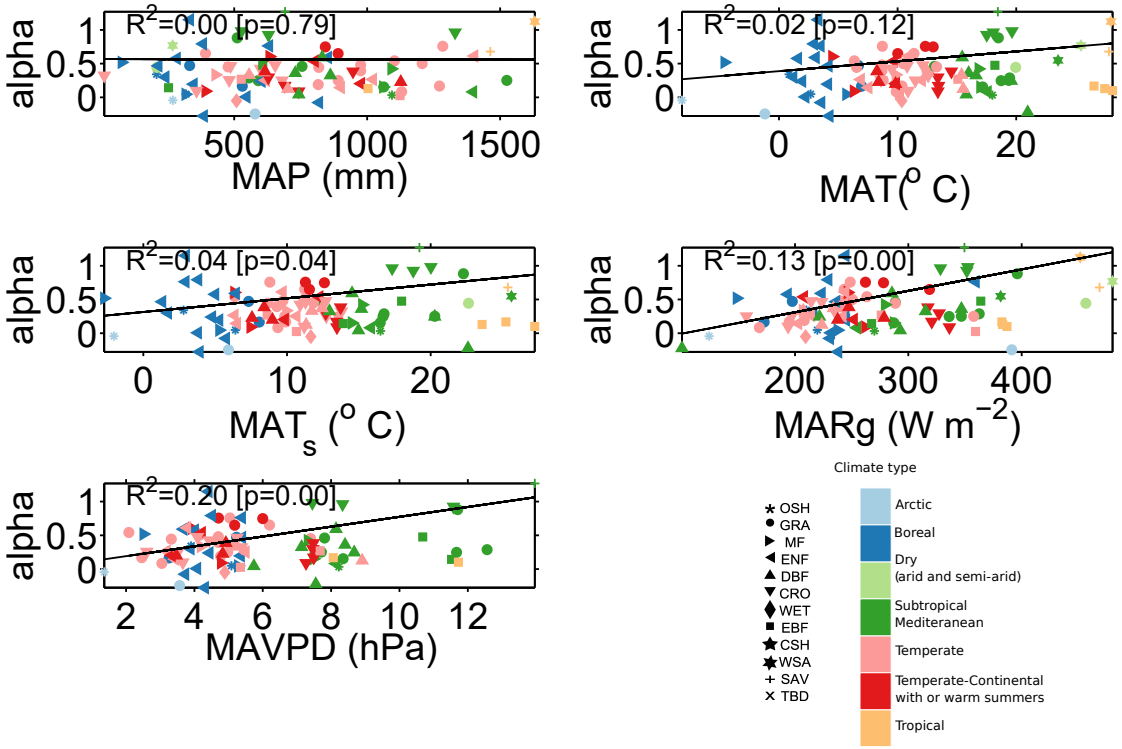


Figure 5.12: α, GRMS (Eq. 5.3.1) parameter associated to Sif_{ms} and its optimized values over 112 sites with CMAGEP with possible correlations to environment site descriptors such as mean annual temperatures, precipitation, water deficit indexes and incoming daylight radiation.

factors as well as by quartiles (Q) splits for all 112 studied sites.

In order to reduce the need to optimize the models locally every time a respiration simulation is required, links were investigated between the values of the GRMS parameter set at each the studied sites and the mean annual values for a set of environmental factors recorded at the studied sites.

For α, parameter associated with changes in Sif_{ms}, Fig. 5.12 shows no significant correlations with the mean annual precipitation and air and soil temperatures values. Low correlations, with R^2 values of 0.13 and 0.2 respectively, are visible however between the recorded α values and the mean annual R_g and VPD values recorded at the 112 sites. The correlations to R_g could point to vegetation dependency to light for photosynthesis and the VPD could be a proxy for the missing water component of the model due to parsimony pressure. The figure shows as well that the higher values of the α parameter are associated to sites coming from subtropical and dry climates.

Similar results were found for β, the T_{air} parameter and mean annual values for the local environmental factors, with R^2 values close to 0 for the precipitation an tem-

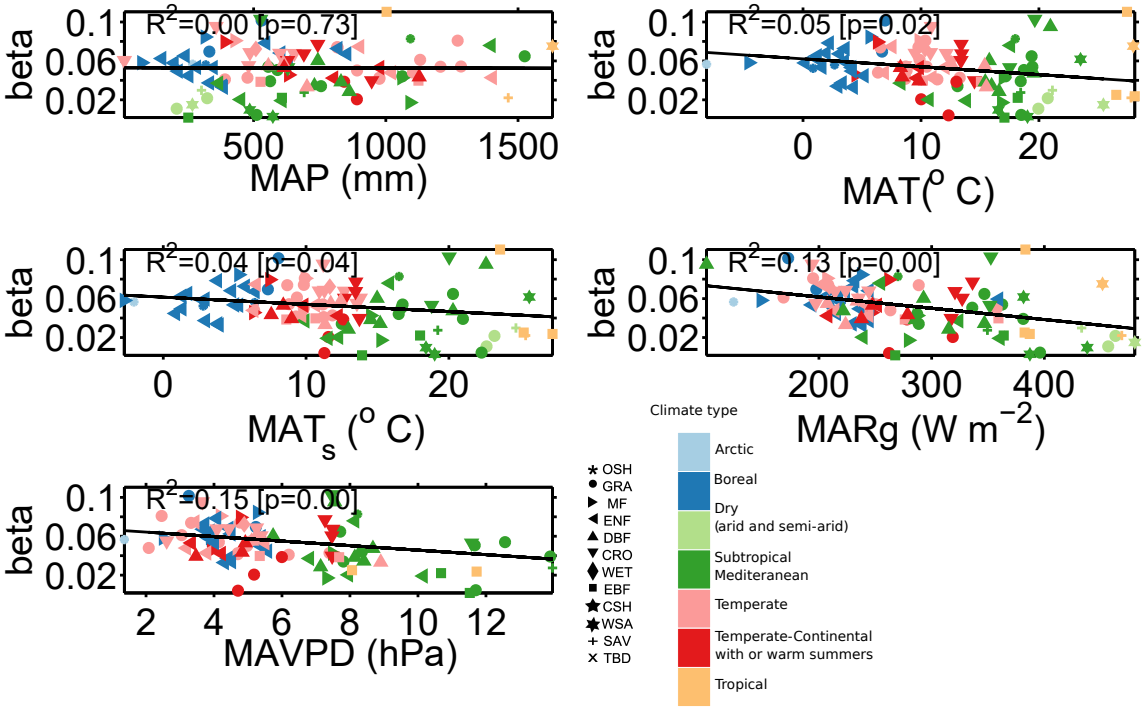


Figure 5.13: β, GRMS (Eq. 5.3.1) parameter associated to T_{air} and its optimized values over 112 sites with CMAGEP with possible correlations to environment site descriptors such as mean annual temperatures, precipitation, water deficit indexes and incoming daylight radiation.

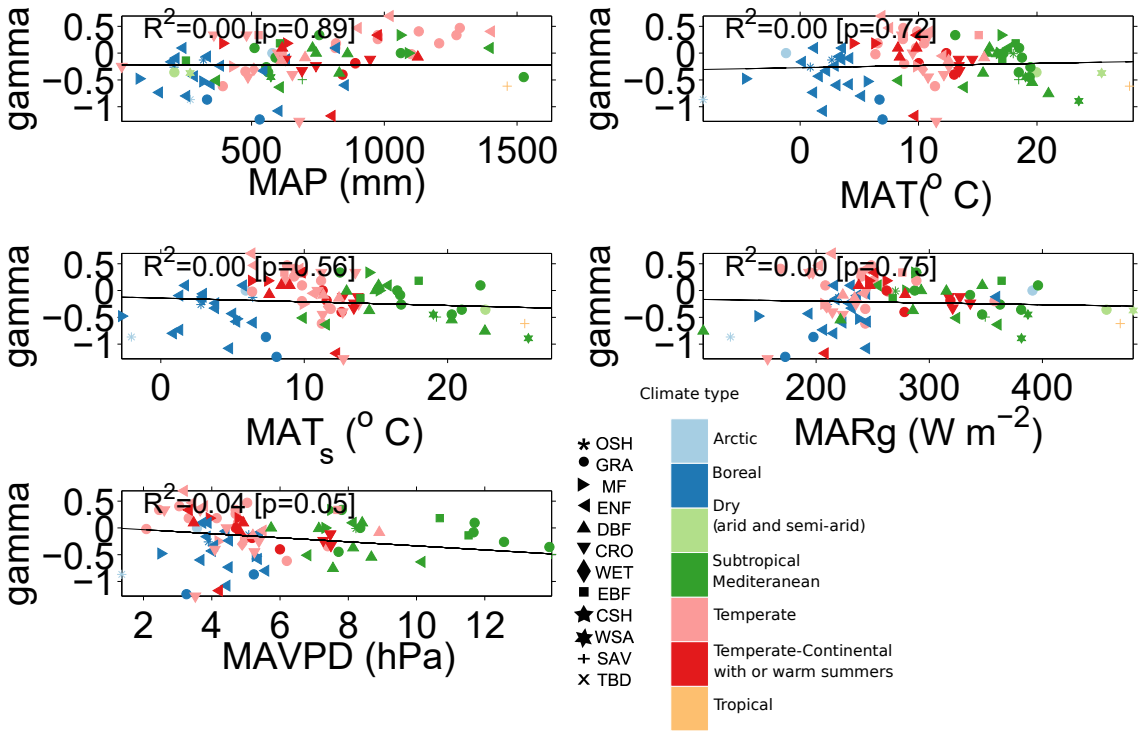


Figure 5.14: γ, GRMS (Eq. 5.3.1) free parameter and its optimized values over 112 sites with CMAGEP with possible correlations to environment site descriptors such as mean annual temperatures, precipitation, water deficit indexes and incoming daylight radiation.

perature indices and low correlations with R^2 values of 0.13 and 0.15 as seen in Fig. 5.13.

No relevant correlations were noticed between γ, the free term of the GRMS and any of the studied environmental indices in Fig. 5.14, where all recorded R^2 values were very close to 0. It is clear however that higher values of the γ parameter are distinctly associated to sites coming from tropical climates.

Figure 5.15 shows the distributions of the α,β and γ parameter values split by Q of the 5 studied environmental markers in order to establish a possible link between the GRMS parameters over 112 studied sites and the specific Q of the environmental markers.

The β parameter shows lower values in the first Q of the T_{air}, others have similar distribution over all Q. The α parameter shows larger values in Q_2 and Q_4 of T_{soil} and low values on Q_1 and Q_3. The β parameter shows lower values in Q_4 of T_{soil}, and similar distribution over all other Q. Similar distributions over all Q for all parameters on MAP, except for α on the 4th Q. γ has higher values over Q_3 of VPD, and α has

similar over all Q with β showing lower values over Q_2 While α shows larger values on the Q_1 of the Rg, the other parameters have similar distributions over all Q.

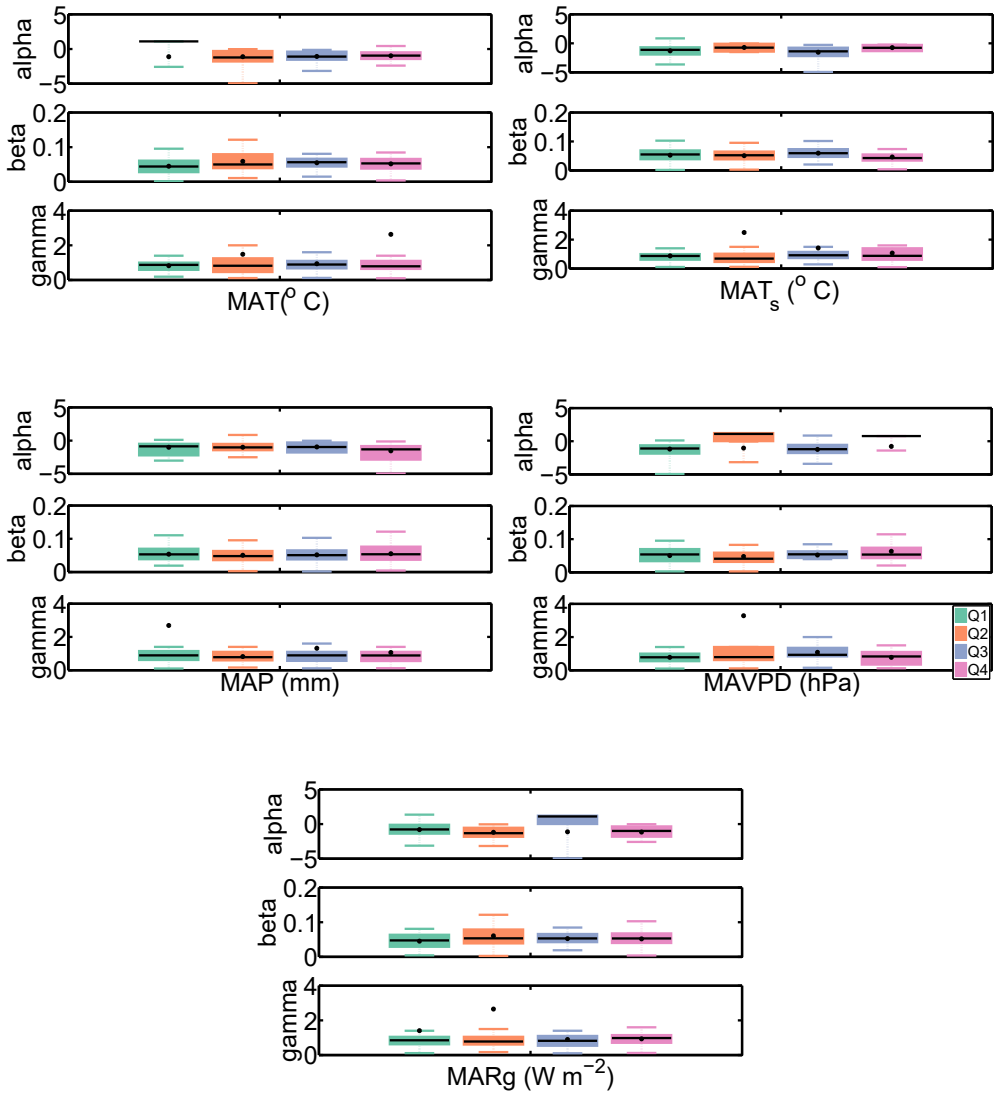


Figure 5.15: Distribution of GRMS parameters vs mean annual values of environment site descriptors.

5.3.6 Comparing with literature established models

In order to assess the goodness-of-fit and prediction performance of models built with CMAGEP for R_{eco} over the 112 FLUXNET sites in the context of established ecology models, the MEF values recorded by the GRMS were compared with MEF values of a

set of established models in the ecology community. These models have been locally optimized using CMA-ES for a fair comparison and are given in Table 5.6 with detailed response descriptions in Ilie et al. (2017).

Prediction capacities in terms of validation values MEF recorded by models commonly used in the biogeochemistry community for simulating and predicting terrestrial CO_2 fluxes were compared with those of the CMAGEP developed GRMS and its local parametrisations at all the 112 studied sites (Fig 5.16 5.17).

The parameters for all the compared models were locally optimized using the CMA-ES approach for fairness.

The models built by CMAGEP usually performed predictions as well and often better compared to the studied literature models at the 112 studied sites, with notable exceptions in sites from subtropical and dry climates. Specifically, the Arrhenius model is clearly outperformed by the GMRS at all 112 sites, the Q_10 model is strongly outperformed by the GRMS model for sites from temperate, subtropical and dry areas, with other sites falling in a similar range of prediction performance. The improvement in performance by GMRMS is even stronger when compared to the Q_10 model that has a water component added. Next, the 4 models presented in the work of (Migliavacca et al., 2011) containing GPP components have a similar prediction performance to each other as well as to the GRMS model parametrisations, with the literature models performing better for some sites from subtropical and dry climates and the GMRS performing better in sites from boreal areas.

Table 5.6: 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}	$\phi_1 \times \phi_2^{(\frac{T-T_{ref}}{10})}$	(Reichstein and Beer, 2008)
Water Q_{10}	$\phi_1 \times \phi_2^{(\frac{T-T_{ref}}{10})} \times \frac{SWC}{SWC+\phi_3} \times \frac{\phi_4}{SWC+\phi_4}$	(Richardson et al., 2008)
LinGPP	$(R_0 + k_2 GPP) \times e^{E_0(\frac{1}{T_{ref}-T_0} - \frac{1}{T_A-T_0})} \times \frac{\alpha k + SWC(1-\alpha)}{k+SWC(1-\alpha)}$	(Migliavacca et al., 2011)
ExpGPP	$[R_0 + R_2(1 - e^{k_2 GPP})] \times e^{E_0(\frac{1}{T_{ref}-T_0} - \frac{1}{T_A-T_0})} \times \frac{\alpha k + SWC(1-\alpha)}{k+SWC(1-\alpha)}$	(Migliavacca et al., 2011)
addLinGPP	$R_0 \times e^{E_0(\frac{1}{T_{ref}-T_0} - \frac{1}{T_A-T_0})} \times \frac{\alpha k + SWC(1-\alpha)}{k+SWC(1-\alpha)} + k_2 GPP$	(Migliavacca et al., 2011)
addExpGPP	$R_0 \times e^{E_0(\frac{1}{T_{ref}-T_0} - \frac{1}{T_A-T_0})} \times \frac{\alpha k + SWC(1-\alpha)}{k+SWC(1-\alpha)} + R_2(1 - e^{k_2 GPP})$	(Migliavacca et al., 2011)

$a, E_0, \phi_1, \phi_2, \phi_3, \phi_4, R_0, R_2, k, k_2$ and α are model parameters.

5.3.7 Unique CMAGEP model for global and yearly simulation

Up-scaling from a terrestrial respiration model structure developed with the help of CMAGEP based on local measurement sites to a model that can be applied at a global level will be difficult from practical and computational reasons.

Although in the previous sections a best over all sites CMAGEP model structure could be proposed with GRMS, the GRMS had 112 unique parametrisations that would not be easily calibrated to all studied FLUXNET sites.

In these conditions, investigating the possibility to select a single parametrisation of GRMS to perform the task of modelling R_{eco} responses to external drivers better than other GRMS parametrisations along all 112 studied sites became an interesting problem. Furthermore, the possibility to apply a unique model structure and parametrisation over the entire globe would be ideal for generating simulated global daily ecological terrestrial respiration fluxes.

To this purpose, mean CEM values were computed over all 112 sites for all 112 GRMS parametrisations. The GRMS parametrisation with the highest over-all sites mean CEM value was selected as the final single CMAGEP model for describing R_{eco} responses to environmental drivers at the global level.

For assessing the capacity of the selected model to accurately capture the R_{eco} flux magnitude and to determine if a single model parametrisation selected based on the over-all sites best fitness function mean value type of approach is worth following, daily terrestrial respiration values were computed for all grids determined by the 360 latitude and 720 longitude lines over one specific year. The daily fluxes were then weighed by the grid area and aggregated over one year for each grid. Finally, a yearly estimated respiration flux aggregated over all grids was computed and compared with established estimations.

A single set of parameters was computed for the GRMS based on mean CEM values, with the unique GMRS further used for extrapolating the R_{eco} fluxes globally, even to grids were no measurements were available. The extrapolation was only done for a specific year, 2006.

The model used for generating the mean R_{eco} daily flux values for all grids over the entire globe is:

$$R^{g}_{eco}(t) = 0.71 \exp\left(0.45 Sif_{ms}(t) + 0.04 T_{air}(t)\right) \tag{5.3.2}$$

The mean daily R_{eco} flux values as predicted by the CMAGEP model for the year 2006 are shown in figure 5.18.

Based on the unique parametrisation of the global CMAGEP model, i.e R^{g}_{eco}, after weighting by grid area, the total predicted terrestrial respiration flux for the year 2006

is

$$R_{eco}(2006) \approx 88 \text{ PgC} \tag{5.3.3}$$

The resulting simulation is found in a sensible range, especially when considering a recent study showing mean yearly global R_{eco} values at ≈ 89 PgC (Zscheischler et al., 2017).

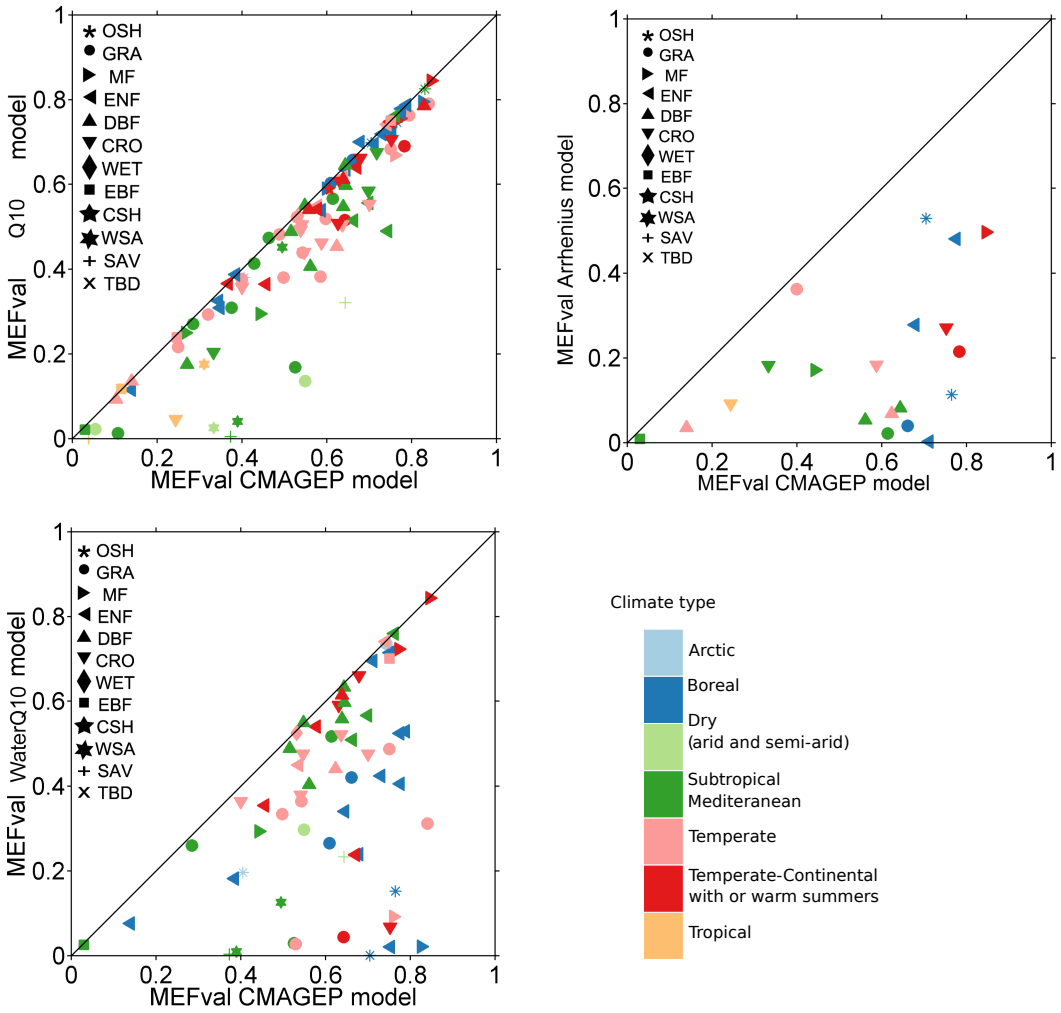


Figure 5.16: Mean MEF values at validation for CMAGEP models vs. mean MEF values at validation for a set of established models in the ecology community, over all 112 studied sites. Part A.

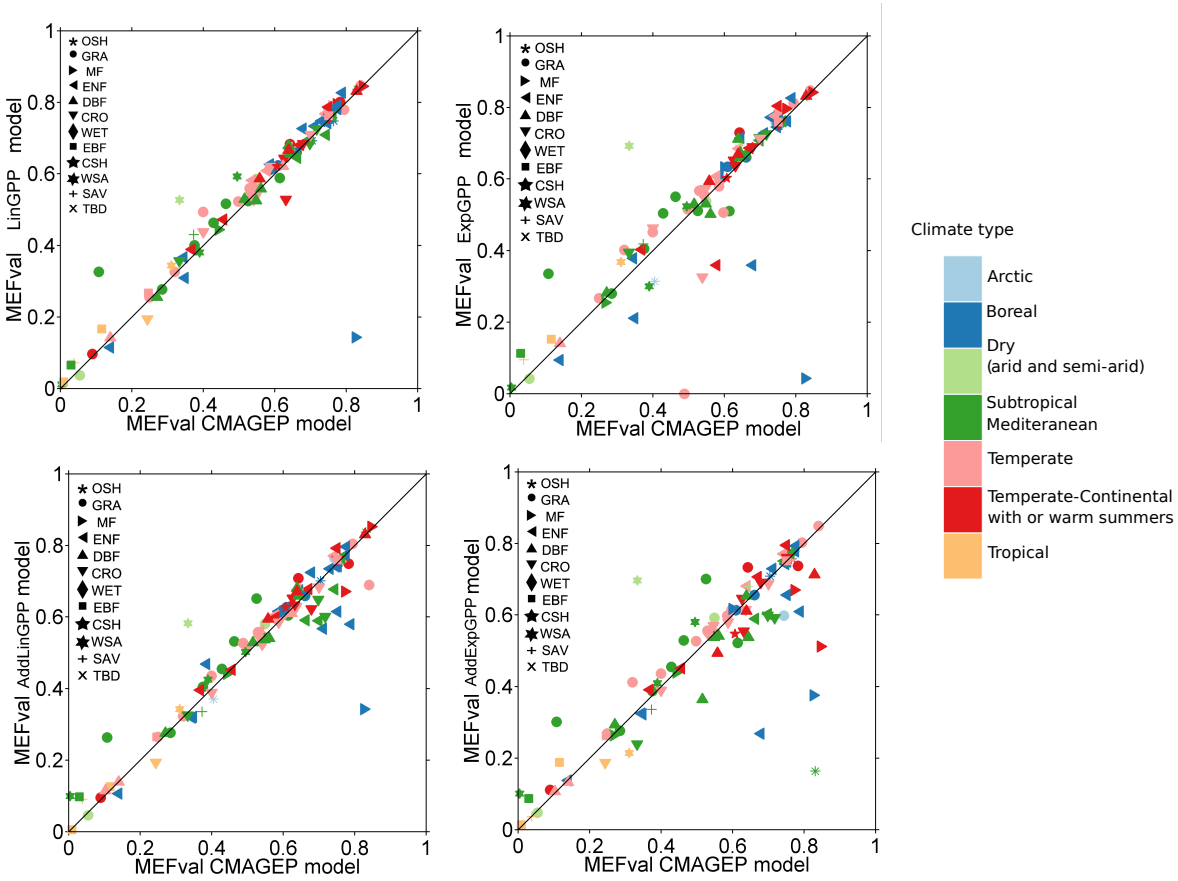


Figure 5.17: Mean MEF values at validation for CMAGEP models vs. mean MEF values at validation for a set of established models in the ecology community, over all 112 studied sites. Part B.

Figure 5.18: CMAGEP modelled annual terrestrial CO_2 efflux per gridded area.

5.4 Discussion

5.4.1 Main remarks on the CMAGEP generated models for R_{eco} fluxes for 112 FLUXNET sites

One of the main results obtained from analysing the models obtained based on CMAGEP evolutions on data from 112 independent measurement sites, was that a clear structural pattern is present in models built for sites from similar climatic zones, differences appearing mainly in local site conditions parametrisations.

This aspect is truly remarkable since it shows not only a possible underlying response of terrestrial ecosystem respiration to drivers across different sites, but also that such a response can be, at least partially, captured by an automated regression technique such as CMAGEP starting from measurements alone.

From the perspective of deploying a model structure with a unique parametrisation to improve the terrestrial respiration component of a global earth system model, it was encouraging to see that a single R_{eco} model structure could be identified to outperform others over the global distribution of sites. That this model represents simple processes is even more important. Of course it was clear that when local estimations are needed, especially in subtropical and dry areas, that the missing Wai component would make a difference in prediction capacity, however the fact that a simple unique model can be sufficiently well recalibrated seems to point to the presence of a general underlying signal of the response of R_{eco} to drivers over the data measured at all sites.

The GRMS model generated the most accurate predictions in the temperate and boreal sites that show strong seasonality in the R_{eco} fluxes, with the lack of temperature seasonality explaining why the GMRS performs least well in the tropical areas, where the model could almost just be replaced by the R_{eco} mean.

When the capacity of CMAGEP models to accurately fit R_{eco} fluxes was studied at the deeper level, an R_{eco} high flux underestimation was present over a large portion of the studied site cases. An R_{eco} high flux underestimation was also present in the study shown in Chapter 2, although in Chapter 2 the relation to water stress was not so clear as in this large scale study. It was especially interesting to see a link between the GRMS missing the high flux values and low WAI values, since the WAI was not specifically chosen as a feature of the GRMS. It could be that the influence the WAI appears only in certain conditions, such as high temperatures, etc. and that these conditions or relations were too expensive for CMAGEP to include in solutions considering the fitness cost penalising for longer solutions. On the other hand it could be that CMAGEP in the current implementation lacks the ability to describe shifts in system conditions, because a conditional operator cannot be included in the initial function set, although

it might be necessary.

A similar shift in system functionality is further described in the work of von Buttlar et al. (2018), where plants are shown to change carbon exchange regime when water and temperature stress are combined. Fig. 5.6 and 5.7 illustrate this relation for the studied FLUXNET sites.

5.4.2 Current CMAGEP implementation limitations

Although the current results are encouraging for the CMAGEP SR framework and its application to biogeochemistry flux modelling, some of the of the CMAGEP limitations for symbolic regression need to be addressed and should be taken in account by future users, or updated by future developers.

The current implementation of the CMAGEP package is limited by the lack of a conditional operator, and that is especially significant when there are shifts in the response of the target to its drivers.

The stochasticity of the GP systems has always been one of the main critiques of the techniques and this plagues CMAGEP as well. The non-deterministic behaviour of CMAGEP makes it challenging to confidently say that one extract model structure is the global optimal solution, especially since the solution is only generated based on the available data.

The CMAGEP does not currently have specific treatment of lags, although including them in the learning process is very relevant. Perhaps only relying on the user to know that a set of lags should be generated and included in the set of candidate drivers is not sufficient and CMAGEP should have the option to automatically generate lags in order to study autoregressive behaviour of the studied signals.

A framework that allows better comparison between the CMAGEP models is recommended, as prediction accuracy is often not sufficient, and the problem of automatically mining for patterns in the functional formulations of the CMAGEP models would prove to be significant, especially from interpretation and knowledge discovery point of view.

The strongest drawback of evolutionary systems such as CMAGEP remains nevertheless the non-scalability (O'Neill et al., 2010), with convergence times being difficult to approximate once data size or fitness function complexity increase drastically. Nevertheless, there are many techniques and parameter settings used by the community to counter this behaviour, such as active learning with strong sub-sampling until a stable solution appears in a generation, etc.

The current CMAGEP implementation addresses this problem by including parameters referring to the maximum allowed time in generations without an improvement in fitness, or even by simply including a stop time in seconds.

5.5 Conclusion and outlook

CMAGEP was successfully applyed to automatically generate models for 112 FLUXNET monitoring sites. After recalibrating each model for the local conditions of all the studied sites, certain model structures were more appropriate for the different studied climates, and interesting pattern emerged.

Although the purpose of the present study was mostly exploratory and the proposal of a unique model formulation for the global terrestrial ecosystem respiration was not one of the initially defined goals, the fact that it was possible to finally select one single model and even generate reasonable simulations for the global yearly aggregated daily R_{eco} flux can be considered a promising start. However, when a single model is desired for extrapolating daily R_{eco} flux at the global level, much more work is needed to understand the links of the local parametrisations to ecosystem conditions and possible correlations to drivers that were not yet considered in this study but that might influence the response of R_{eco}.

The emerging patterns in CMAGEP model structures and parameter sets for R_{eco} responses to external drivers with respect to climate and PFT types were very interesting to analyse and show promise for future studies.

Future work in CMAGEP model structure pattern analysis might include learning from the bulk sites by climate or PFT types followed by comparing the new CMAGEP models with single site CMAGEP generated models.

Other interesting future research directions might include the treatment of the current problem not only as a symbolic regression problem, but as a classification problem. In such a setting, R_{eco}'s response to external drivers discriminating well between the different climate types or even PFTs could be explored.

In order to help the interpretation aspect, this approach might significantly benefit from including Vladislavleva et al. (2009)'s orders of non-linearity as a measure of solution complexity instead of solution length, since such a parsimony pressure would limit the appearance of functions that are very hard to grasp conceptually in the real world modelling such as $\exp(\exp)$.

Due to obvious time and resources limitations, these features will only be addressed in future releases of the CMAGEP package.

Conclusions and Future Work

6.1 Conclusions

The present thesis showed an extensive interdisciplinary study, where the focus has been both on the development of a machine learning approach that allows for the automated discovery of compact symbolic regressions, as well as on the potential of applying such an approach in the biogeochemistry field. In the following paragraphs I will mentions some of the most significant findings of this doctoral project.

A standard GEP C++ package was implemented from scratch and was applied to real terrestrial ecosystem carbon exchange flux observations measured at a single forest site. The GEP models were shown to outperform currently established R_{eco} models in the biogeochemistry field, and displayed lower sensitivity to noise compared to other well known machine learning methods in an artificial data experiment.

The GEP models complemented existing knowledge for the response of R_{eco} to external drivers by describing novel and interesting components of the model structures. Although novel, the GEP model components were still in line with experimental and empirical biogeochemistry studies.

To increase the interpretability of the GEP models for symbolic regression, the CMAGEP was designed and implemented. CMAGEP was then compared to GEP for symbolic regression when applied to two artificial data benchmarks, one established in the GP community, and another containing high precision constants. In both cases CMAGEP outperformed or was equal to GP in terms of accuracy and more importantly, CMAGEP generated ≈ 60 % shorter solutions. The improvements to CMAGEP were more noticeable for the high precision constant funciton set, showing that when constant calibration is needed, the CMA-ES component of the CMAGEP helps reach a solutions optimally.

The current GEP and CMAGEP implementations were compared with a demo

version of a commercial GEP over an established real data benchmark data set, and CMAGEP outperformed both GEP implementations, with GEP being equally accurate for both current and commercial implementations.

When the CMAGEP and GEP models were compared with a set of know machine learning approaches, it was shown that their prediction performances were in a sensible range to the state-of-the-art, with the GP system having the advantage of returning a readable function as well.

CMAGEP was then used to build compact and interesting models for methane terrestrial exchange in an Arctic monitoring site.

Lastly, CMAGEP was used in a large scale modelling experiment for describing R_{eco} responses to external drivers over 112 monitoring sites. The CMAGEP models revealed interesting patterns, and both confirmed known responses as well as introducing new After an extensive analysis on the emerging patterns in CMAGEP model structures, it was possible to select a CMAGEP structure that generalises the R_{eco} functional properties sufficiently well. The CMAGEP model was a simple model with only 2 non-site dependent variables and was easily used to simulate daily R_{eco} fluxes over the entire globe for a specific year. When the fluxes were summed globally and yearly, the total estimated terrestrial ecosystem respiration flux was in a very close range to results from independent studies.

Considering all the above mentioned results it can be stated with confidence that CMAGEP is a promising approach for developing relevant and compact solutions for the response of terrestrial ecosystem respiration to candidate drivers. At the same time it is clear that CMAGEP is not confined to the field of biogeochemistry, but can easily be applied to solve problems in other fields where interpretation and understanding of the modelling process is needed.

6.2 Future Work

Among future research direction, some that seem more interesting include:

- Studying the physical, chemical or biological soundness of the model structures returned by CMAGEP, in order to establish how fit they would be for real world simulation deployment;
- Performing more thorough and exhaustive studies in the matter of independently generated Symbolic Regression formulas equifinality;
- Constructing a framework where the model structures can also be compared from interpretation point of view and not only based on prediction performance;
- Exploring the potential of CMAGEP for classification problems, especially when interpretation is desired;

- Further developing the CMAGEP package to cover some of the limitations mentioned in the thesis Chapters as well as adding learning strategies (e.g. deems, intelligent sub-sampling) to shorten the learning time, since currently time to return a solution scales poorly with increased data availability.

Glossary

chromosome individual used in automatically evolving an optimal solution comprised of a set of genes that are connected with a binary operation (e.g. $+ \times -$). 15

CMA-ES covariance matrix adaptation evolutionary strategy. 22

evolution the process of producing an optimal solution by GEP through. 15

expression tree binary tree used to represent algebraic expressions. 17

gene set of characters of fixed length that encodes an expression tree. 15

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. 17

gene tail end section of the string that comprises a GEP gene, containing only characters that map to predictors. 17

generation time step of an evolution. 17

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, etc.). 17

genetic operator rate probability of a genetic manipulation to occur during a generation. 19

GEP **g**ene **e**xpression **p**rogramming, machine learning method that evolves chromosome structures with the purpose of minimizing a cost function. 14

hyper-parameter set of parameters which need to be set for the runs of a machine learning approach. 19

ill-posed problem a problem for which the solutions might not be unique or unstable, also known as an inverse problem. 22

individual GEP entity that is a component of a population during a certain step of the evolution process. Also known as chromosome. 17

MLM **m**achine **l**earning **m**ethod that can produce predicted values based on a training set. 24

population total set of chromosomes that participate at a certain step in the evolution of an optimal solution in the GEP approach.. 17

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. 17

solution finally selected model structure resulting from a GEP run. 12, 14

Declaration

I hereby declare that this submission is my own work and that all work presented in this thesis as original is so to the best of my knowledge. Some of the research presented in this thesis has previously been published or submitted for publication by the author and where that was the case it is explicitly indicated.

All references and acknowledgements to the work of other researchers have been appropriately given.

List of Figures

List of Tables

Bibliography

Shun-ichi Amari. Natural Gradient Works Efficiently in Learning. *Neural Computation*, 10(2):251–276, feb 1998. ISSN 0899-7667. doi: 10.1162/089976698300017746. URL http://www.mitpressjournals.org/doi/10.1162/089976698300017746. 59

Christian G Andresen, Mark J Lara, Craig E Tweedie, and Vanessa L Lougheed. Rising plant-mediated methane emissions from arctic wetlands. *Global Change Biology*, 23(3):1128–1139, mar 2017. ISSN 13541013. doi: 10.1111/gcb.13469. URL http://doi.wiley.com/10.1111/gcb.13469. 105

Justin Ashworth, Elisabeth J. Wurtmann, and Nitin S. Baliga. Reverse engineering systems models of regulation: Discovery, prediction and mechanisms. *Current Opinion in Biotechnology*, 23(4):598–603, aug 2012. ISSN 09581669. doi: 10.1016/j.copbio.2011.12.005. URL http://www.ncbi.nlm.nih.gov/pubmed/22209016. 14

A. Auger and N. Hansen. A restart CMA evolution strategy with increasing population size. *2005 IEEE Congress on Evolutionary Computation*, 2:1769–1776, 2005. doi: 10.1109/CEC.2005.1554902. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1554902. 22, 63

D.A. Augusto and H.J.C. Barbosa. Symbolic regression via genetic programming. In *Proceedings. Vol.1. Sixth Brazilian Symposium on Neural Networks*, number diagram C, pages 173–178. IEEE Comput. Soc, 2000. ISBN 0-7695-0856-1. doi: 10.1109/SBRN.2000.889734. URL http://ieeexplore.ieee.org/document/889734/. 4

Dennis Baldocchi. TURNER REVIEW No. 15. 'Breathing' of the terrestrial biosphere: Lessons learned from a global network of carbon dioxide flux measurement systems. *Australian Journal of Botany*, 56(1):1–26, 2008. ISSN 00671924. doi: 10.1071/BT07151. URL http://www.publish.csiro.au/?paper=BT07151. 2

Dennis D. Baldocchi. Assessing the eddy covariance technique for evaluating carbon dioxide exchange rates of ecosystems: past, present and future. *Global Change*

Biology, 9(4):479–492, apr 2003. ISSN 1354-1013. doi: 10.1046/j.1365-2486.2003.00629.x. URL http://doi.wiley.com/10.1046/j.1365-2486.2003.00629.x. 3, 118

Christoph Bandt and Bernd Pompe. Permutation entropy: a natural complexity measure for time series. *Physical review letters*, 88(17):174102, apr 2002. ISSN 0031-9007. doi: 10.1103/PhysRevLett.88.174102. URL http://www.ncbi.nlm.nih.gov/pubmed/12005759. 21

W Banzhaf and W B Langdon. Some Considerations on the Reason for Bloat. *Genetic Programming and Evolvable Machines*, 3:81–91, 2002. 59, 89

Neil D. Bennett, Barry F.W. Croke, Anthony J. Jakeman, Lachlan T. H. Newham, and John P. Norton. Performance evaluation of environmental models. *2010 International Congress on Environmental Modelling and Software Modelling for Environment's Sake*, pages 1–9, 2010. URL http://www.iemss.org/iemss2010/papers/S20/S.20.01.Performanceassessmentofenvironmentalmodels-ANTHONYJAKEMAN.pdf. 12, 20, 120

Hans-Georg Beyer and Hans-Paul Schwefel. Evolution strategies – A comprehensive introduction. *Natural Computing*, 1(1):3–52, 2002. ISSN 15677818. doi: 10.1023/A:1015059928466. URL http://link.springer.com/10.1023/A:1015059928466. 22

Gordon Bonan. *Ecological Climatology*. Cambridge University Press, Cambridge, 3rd editio edition, 2016. ISBN 9781107339200. doi: 10.1017/CBO9781107339200. URL http://ebooks.cambridge.org/ref/id/CBO9781107339200. 11

J.C. Bongard and H. Lipson. Nonlinear System Identification Using Coevolution of Models and Tests. *IEEE Transactions on Evolutionary Computation*, 9(4):361–384, aug 2005. ISSN 1089-778X. doi: 10.1109/TEVC.2005.850293. URL http://ieeexplore.ieee.org/document/1492385/. 2

Josh Bongard and Hod Lipson. Automated reverse engineering of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences of the United States of America*, 104(24):9943–9948, 2007. ISSN 0027-8424. doi: 10.1073/pnas.0609476104. 12

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 08856125. doi: 10.1023/A:1010933404324. URL http://link.springer.com/article/10.1023/A:1010933404324. 24, 77

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

01672789. doi: 10.1016/0167-2789(86)90031-X. URL http://linkinghub.elsevier.com/retrieve/pii/016727898690031X. 26, 118

J. V. Buttlar, J. Zscheischler, and M. D. Mahecha. An extended approach for spatiotemporal gapfilling: Dealing with large and systematic gaps in geoscientific datasets. *Nonlinear Processes in Geophysics*, 21(1):203–215, 2014. ISSN 10235809. doi: 10.5194/npg-21-203-2014. 26, 118

C. E. Shannon and C. E. Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal*, Vol. 27(3):379–423, jul 1948. ISSN 00058580. doi: 10.1002/j.1538-7305.1948.tb01338.x. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6773024. 20

Chih-Chung Chang and Chih-Jen Lin. Libsvm. *ACM Transactions on Intelligent Systems and Technology*, 2(3):1–27, 2011. ISSN 21576904. doi: 10.1145/1961189.1961199. URL http://dl.acm.org/citation.cfm?doid=1961189.1961199. 24, 77

P Ciais, C Sabine, G Bala, L Bopp, V Brovkin, J Canadell, A Chhabra, R DeFries, J Galloway, M Heimann, C Jones, C Le Quéré, R B Myneni, Thornton P., and S Piao. Carbon and Other Biogeochemical Cycles. In *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013. 105

Carlos A. Coello and Efrén Mezura Montes. Constraint-handling in genetic algorithms through the use of dominance-based tournament selection. *Advanced Engineering Informatics*, 16(3):193–203, 2002. ISSN 14740346. doi: 10.1016/S1474-0346(02)00011-3. URL http://www.sciencedirect.com/science/article/pii/S1474034602000113. 28, 67, 114, 119

Mohd Danish. Prediction of scour depth at bridge abutments in cohesive bed using gene expression programming. *International Journal of Civil Engineering and Technology (Ijciet)*, 5(11):25–32, 2014. 59

Joan G. Ehrenfeld, Beth Ravit, and Kenneth Elgersma. Feedback in the plant-soil system. *Annual Review of Environment and Resources*, 30(1):75–115, nov 2005. ISSN 1543-5938. doi: 10.1146/annurev.energy.30.050504.144212. URL http://www.annualreviews.org/doi/10.1146/annurev.energy.30.050504.144212. 44

J. Elith, J. R. Leathwick, and T. Hastie. A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4):802–813, jul 2008. ISSN 0021-8790. doi: 10.1111/j.1365-2656.2008.01390.x. URL http://doi.wiley.com/10.1111/j.1365-2656.2008.01390.x. 78

D.A.K Fernando, A.Y Shamseldin, and R.J Abrahart. Using gene expression programming to develop a combined runoff estimate model from conventional rainfall-runoff model outputs, 2009. 14

C Ferreira. Gene expression programming: a new adaptive algorithm. In *The 6th Online World Conference on Soft Computing in Industrial Applications*, 2001. 5, 14, 15, 17, 60, 98

Cândida Ferreira. Gene Expression Programming and the Evolution of Computer Programs. In Leandro N De Castro and Fernando J Von Zuben, editors, *Recent Developments in Biologically Inspired Computing*, chapter 6, pages 82–103. IGI Global, 1st edition, jan 2004. ISBN 159140312X. doi: 10.4018/978-1-59140-312-8.ch005. URL http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-59140-312-8.ch005http://www.gene-expression-programming.com/gep/webpapers/abstracts.asp{#}11. 59

Candida Ferreira. *Gene expression programming: mathematical modeling by an artificial intelligence*, volume 21. Springer-Verlag Berlin Heidelberg, 2 edition, 2006. ISBN 3540262563. doi: 10.1007/3-540-32849-1. URL http://dl.acm.org/citation.cfm?id=1208676http://www.springer.com/us/book/9783540327967. 17, 18, 60, 66, 78, 83

Python Software Foundation. Python/C API Reference Manual. URL https://docs.python.org/2/c-api/. 66

Christian Frankenberg, Joseph Berry, Luis Guanter, and Joanna Joiner. Remote sensing of terrestrial chlorophyll fluorescence from space. *SPIE Newsroom*, feb 2013. ISSN 18182259. doi: 10.1117/2.1201302.004725. URL http://www.spie.org/x92167.xml. 118

P. Friedlingstein, P. Cox, R. Betts, L. Bopp, W. von Bloh, V. Brovkin, P. Cadule, S. Doney, M. Eby, I. Fung, G. Bala, J. John, C. Jones, F. Joos, T. Kato, M. Kawamiya, W. Knorr, K. Lindsay, H. D. Matthews, T. Raddatz, P. Rayner, C. Reick, E. Roeckner, K.-G. Schnitzler, R. Schnur, K. Strassmann, A. J. Weaver, C. Yoshikawa, and N. Zeng. Climate–Carbon Cycle Feedback Analysis: Results from the C 4 MIP Model Intercomparison. *Journal of Climate*, 19(14): 3337–3353, jul 2006. ISSN 0894-8755. doi: 10.1175/JCLI3800.1. URL http://journals.ametsoc.org/doi/abs/10.1175/JCLI3800.1. 44

A. D. Friend. Terrestrial plant production and climate change. *Journal of Experimental Botany*, 61(5):1293–1309, mar 2010. ISSN 0022-0957. doi: 10.1093/jxb/erq019. URL https://academic.oup.com/jxb/article-lookup/doi/10.1093/jxb/erq019. 5

Tagir G. Gilmanov, L. Aires, Z. Barcza, V. S. Baron, L. Belelli, J. Beringer, D. Billesbach, D. Bonal, J. Bradford, E. Ceschia, D. Cook, C. Corradi, a. Frank, D. Gianelle, C. Gimeno, T. Gruenwald, Haiqiang Guo, N. Hanan, L. Haszpra, J. Heilman, a. Jacobs, M. B. Jones, D. a. Johnson, G. Kiely, Shenggong Li, V. Magliulo, E. Moors, Z. Nagy, M. Nasyrov, C. Owensby, K. Pinter, C. Pio, M. Reichstein, M. J. Sanz, R. Scott, J. F. Soussana, P. C. Stoy, T. Svejcar, Z. Tuba, and Guangsheng Zhou. Productivity, Respiration, and Light-Response Parameters of World Grassland and Agroecosystems Derived From Flux-Tower Measurements. *Rangeland Ecology & Management*, 63(1):16–39, jan 2010. ISSN 1550-7424. doi: 10.2111/REM-D-09-00072.1. URL http://www.bioone.org/doi/abs/10.2111/REM-D-09-00072.1. 12

David E. Goldberg and John H. Holland. Genetic Algorithms and Machine Learning. *Machine Learning*, 3(2/3):95–99, 1988. ISSN 08856125. doi: 10.1023/A:1022602019183. URL http://link.springer.com/10.1023/A:1022602019183. 60

Hoshin V. Gupta, Martyn P. Clark, Jasper a. Vrugt, Gab Abramowitz, and Ming Ye. Towards a comprehensive assessment of model structural adequacy. *Water Resources Research*, 48(8):n/a–n/a, aug 2012. ISSN 00431397. doi: 10.1029/2011WR011044. URL http://doi.wiley.com/10.1029/2011WR011044. 12

Aytac Guven and Ali Aytek. New Approach for Stage–Discharge Relationship: Gene-Expression Programming. *Journal of Hydrologic Engineering*, 14(8):812–820, aug 2009. ISSN 1084-0699. doi: 10.1061/(ASCE)HE.1943-5584.0000044. URL http://ascelibrary.org/doi/abs/10.1061/(ASCE)HE.1943-5584.0000044. 59

Nikolaus Hansen. An Analysis of Mutative σ-Self-Adaptation on Linear Fitness Functions. *Evolutionary Computation*, 14(3):255–275, sep 2006a. ISSN 1063-6560. doi: 10.1162/evco.2006.14.3.255. URL http://www.ncbi.nlm.nih.gov/pubmed/16903793http://www.mitpressjournals.org/doi/10.1162/evco.2006.14.3.255. 6, 59

Nikolaus Hansen. *Towards a New Evolutionary Computation*, volume 192 of *Studies in Fuzziness and Soft Computing*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006b. ISBN 978-3-540-29006-3. doi: 10.1007/3-540-32494-1. URL http://www.scopus.com/inward/record.url?eid=2-s2.0-33845271655{&}partnerID=tZOtx3y1http://link.springer.com/10.1007/3-540-32494-1. 22, 63

Nikolaus Hansen. The CMA evolution strategy: A tutorial. In *Vu le*, pages 1–34, 2011. URL http://www.lri.fr/{~}hansen/cmatutorial110628.pdf. 59

Nikolaus Hansen. The CMA Evolution Strategy: A Tutorial. apr 2016. URL http://arxiv.org/abs/1604.00772. 63

Nikolaus Hansen, Sibylle D Müller, and Petros Koumoutsakos. Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). *Evolutionary computation*, 11(1):1–18, jan 2003. ISSN 1063-6560. doi: 10.1162/106365603321828970. URL http://www.ncbi.nlm.nih.gov/pubmed/12804094. 22, 59, 63

P J Hanson, N T Edwards, C T Garten, J A Andrews, C. T. Garten P. J. Hanson, N. T. Edwards, and J. A. Andrews. Separating root and soil microbial contributions to soil respiration: A review of methods and observations. *Biogeochemistry*, 48(1): 115–146, 2000. doi: 10.1023/A:1006244819642. URL http://www.jstor.org/stable/1469555?seq=1{#}page{_}scan{_}tab{_}contents. 45

Muhammad Z. Hashmi and Asaad Y. Shamseldin. Use of Gene Expression Programming in regionalization of flow duration curve. *Advances in Water Resources*, 68: 1–12, jun 2014. ISSN 03091708. doi: 10.1016/j.advwatres.2014.02.009. URL http://linkinghub.elsevier.com/retrieve/pii/S0309170814000323. 14

Marti A. Hearst. Support vector machines. *IEEE Intelligent Systems and Their Applications*, 13(4):18–28, 1998. ISSN 10947167. doi: 10.1109/5254.708428. URL http://ieeexplore.ieee.org/xpls/abs{_}all.jsp?arnumber=708428. 24, 77

Martin Heimann and Markus Reichstein. Terrestrial ecosystem carbon dynamics and climate feedbacks. *Nature*, 451(7176):289–92, jan 2008. ISSN 1476-4687. doi: 10.1038/nature06591. URL http://dx.doi.org/10.1038/nature06591. 11

A. Heinemeyer, C. Di Bene, A. R. Lloyd, D. Tortorella, R. Baxter, B. Huntley, A. Gelsomino, and P. Ineson. Soil respiration: implications of the plant-soil continuum and respiration chamber collar-insertion depth on measurement and modelling of soil CO2 efflux rates in three ecosystems. *European Journal of Soil Science*, 62(1): 82–94, feb 2011. ISSN 13510754. doi: 10.1111/j.1365-2389.2010.01331.x. URL http://doi.wiley.com/10.1111/j.1365-2389.2010.01331.x. 23, 26, 45

A. Heinemeyer, M. Wilkinson, R. Vargas, J. A. Subke, E. Casella, J. I L Morison, and P. Ineson. Exploring the overflow tap theory: Linking forest soil CO 2 fluxes and individual mycorrhizosphere components to photosynthesis. *Biogeosciences*, 9(1): 79–95, 2012. ISSN 17264170. doi: 10.5194/bg-9-79-2012. 14, 15, 25, 26

Arthur E Hoerl and Robert W Kennard. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67, 1970. ISSN 0040-1706. doi: 10.1080/00401706.1970.10488634. URL http://amstat.tandfonline.com/doi/abs/10.1080/00401706.2000.10485983http://amstat.tandfonline.com/doi/abs/10.1080/00401706.1970.10488634. 24, 77

Mathias Hoffmann, Nicole Jurisch, Elisa Albiac Borraz, Ulrike Hagemann, Matthias Drösler, Michael Sommer, and Jürgen Augustin. Automated modeling of ecosystem CO2 fluxes based on periodic closed chamber measurements: A standardized conceptual and practical approach. *Agricultural and Forest Meteorology*, 200:30–45, 2015. ISSN 01681923. doi: 10.1016/j.agrformet.2014.09.005. URL http://linkinghub.elsevier.com/retrieve/pii/S0168192314002160. 12

Teemu Hölttä, Maurizio Mencuccini, and Eero Nikinmaa. A carbon cost-gain model explains the observed patterns of xylem safety and efficiency. *Plant, Cell & Environment*, 34(11):1819–1834, nov 2011. ISSN 01407791. doi: 10.1111/j.1365-3040.2011.02377.x. URL http://doi.wiley.com/10.1111/j.1365-3040.2011.02377.x. 48

I. Ilie, P. Dittrich, N. Carvalhais, M. Jung, A. Heinemeyer, M. Migliavacca, J.I.L. Morison, S. Sippel, J.-A. Subke, M. Wilkinson, and D.M. Mahecha. Reverse engineering model structures for soil and ecosystem respiration: The potential of gene expression programming. *Geoscientific Model Development*, 10(9), 2017. ISSN 19919603. doi: 10.5194/gmd-10-3519-2017. 89, 115, 118, 119, 120, 140

Iulia Ilie, Miguel Mahecha, Martin Jung, Nuno Carvalhais, and Peter Dittrich. Evolving compact symbolic expressions by a GEP CMA-ES hybrid approach. 50, 83, 107, 115, 119

Moslem Imani, Rey-Jer You, and Chung-Yen Kuo. Forecasting Caspian Sea level changes using satellite altimetry data (June 1992–December 2013) based on evolutionary support vector regression algorithms and gene expression programming. *Global and Planetary Change*, 121:53–63, oct 2014. ISSN 09218181. doi: 10.1016/j.gloplacha.2014.07.002. URL http://linkinghub.elsevier.com/retrieve/pii/S0921818114001210. 59

Alan Julian Izenman. J. R. Wolf and H. A. Wolfer: An Historical Note on the Zurich Sunspot Relative Numbers. *Journal of the Royal Statistical Society. Series A (General)*, 146(3):311, 1983. ISSN 00359238. doi: 10.2307/2981658. URL http://www.jstor.org/stable/10.2307/2981658?origin=crossref. 60

a. J. Jakeman, R. a. Letcher, and J. P. Norton. Ten iterative steps in development and evaluation of environmental models. *Environmental Modelling and Software*, 21(5):602–614, 2006. ISSN 13648152. doi: 10.1016/j.envsoft.2006.01.004. 12

M Torre Jorgenson, Yuri L Shur, and Erik R Pullman. Abrupt increase in permafrost degradation in Arctic Alaska. *Geophysical Research Letters*, 33(2):L02503, 2006. ISSN 0094-8276. doi: 10.1029/2005GL024960. URL http://doi.wiley.com/10.1029/2005GL024960. 105

Martin Jung, Markus Reichstein, Hank A Margolis, Alessandro Cescatti, Andrew D Richardson, M Altaf Arain, Almut Arneth, Christian Bernhofer, Damien Bonal, Jiquan Chen, Damiano Gianelle, Nadine Gobron, Gerald Kiely, Werner Kutsch, Gitta Lasslop, Beverly E Law, Anders Lindroth, Lutz Merbold, Leonardo Montagnani, Eddy J Moors, Dario Papale, Matteo Sottocornola, Francesco Vaccari, and Christopher Williams. Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations. *Journal of Geophysical Research*, 116(G3):G00J07, sep 2011. 118

S. I. Kabanikhin. Definitions and examples of inverse and ill-posed problems. *Journal of Inverse and Ill-Posed Problems*, 16(4):317–357, 2008. ISSN 09280219. doi: 10.1515/JIIP.2008.019. 22

Oliver N. Keene. The log transformation is special. *Statistics in Medicine*, 14(8): 811–819, apr 1995. ISSN 02776715. doi: 10.1002/sim.4780140810. URL `http://doi.wiley.com/10.1002/sim.4780140810`. 26, 45

Rahman Khatibi, Leila Naghipour, Mohammad a. Ghorbani, Michael S. Smith, Vahid Karimi, Reza Farhoudi, Hadi Delafrouz, and Hadi Arvanaghi. Developing a predictive tropospheric ozone model for Tabriz. *Atmospheric Environment*, 68:286–294, apr 2013. ISSN 13522310. doi: 10.1016/j.atmosenv.2012.11.020. URL `http://linkinghub.elsevier.com/retrieve/pii/S1352231012010722`. 14

Yongwon Kim. Effect of thaw depth on fluxes of CO2 and CH4 in manipulated Arctic coastal tundra of Barrow, Alaska. *The Science of the Total Environment*, 505:385–389, feb 2015. ISSN 1879-1026. doi: 10.1016/j.scitotenv.2014.09.046. URL `http://www.sciencedirect.com/science/article/pii/S0048969714013680`. 105

Stefanie Kirschke, Philippe Bousquet, Philippe Ciais, Marielle Saunois, Josep G Canadell, Edward J Dlugokencky, Peter Bergamaschi, Daniel Bergmann, Donald R Blake, Lori Bruhwiler, Philip Cameron-Smith, Simona Castaldi, Frédéric Chevallier, Liang Feng, Annemarie Fraser, Martin Heimann, Elke L Hodson, Sander Houweling, Béatrice Josse, Paul J Fraser, Paul B Krummel, Jean-François Lamarque, Ray L Langenfelds, Corinne Le Quéré, Vaishali Naik, Simon O'Doherty, Paul I Palmer, Isabelle Pison, David Plummer, Benjamin Poulter, Ronald G Prinn, Matt Rigby, Bruno Ringeval, Monia Santini, Martina Schmidt, Drew T Shindell, Isobel J Simpson, Renato Spahni, L Paul Steele, Sarah A Strode, Kengo Sudo, Sophie Szopa, Guido R van der Werf, Apostolos Voulgarakis, Michiel van Weele, Ray F Weiss, Jason E Williams, and Guang Zeng. Three decades of global methane sources and sinks. *Nature Geoscience*, 6(10):813–823, sep 2013. ISSN 1752-0894. doi: 10.1038/ngeo1955. URL `http://dx.doi.org/10.1038/ngeo1955`. 105

Mark E Kotanchek, Ekaterina Vladislavleva, and Guido Smits. Symbolic Regression Is Not Enough : It Takes a Village to Raise a Model. In *Genetic Programming Theory*

and Practice X, pages 187–203. Springer Science+Business Media New York, 2013. ISBN 9781461468462. doi: 10.1007/978-1-4614-6846-2. 15

Andres M. Kowalski, Maria Teresa Martín, Angelo Plastino, Osvaldo A. Rosso, and Montserrat Casas. Distances in Probability Space and the Statistical Complexity Setup. *Entropy*, 13(12):1055–1075, jun 2011. ISSN 1099-4300. doi: 10.3390/e13061055. URL http://www.mdpi.com/1099-4300/13/6/1055/. 21

John R. Koza. Genetic programming as a means for programming computers by natural selection. *Statistics and Computing*, 4(2):87–112, 1994. ISSN 09603174. doi: 10.1007/BF00175355. 4, 59

S.V. Kumar, C.D. Peters-Lidard, Y. Tian, P.R. Houser, J. Geiger, S. Olden, L. Lighty, J.L. Eastman, B. Doty, P. Dirmeyer, J. Adams, K. Mitchell, E.F. Wood, and J. Sheffield. Land information system: An interoperable framework for high resolution land surface modeling. *Environmental Modelling & Software*, 21 (10):1402–1415, oct 2006. ISSN 1364-8152. doi: 10.1016/J.ENVSOFT.2005.07.004. URL https://www.sciencedirect.com/science/article/pii/S1364815205001283. 2

Y. Kuzyakov. Sources of CO2 efflux from soil and review of partitioning methods. *Soil Biology and Biochemistry*, 38(3):425–448, 2006. ISSN 00380717. doi: 10.1016/j.soilbio.2005.08.020. 45

Min Jung M.J. Kwon, Felix Beulig, Iulia Ilie, Marcus Wildner, Kirsten Küsel, Lutz Merbold, M.D. Miguel D. Mahecha, Nikita Zimov, Sergey A. S.A. Zimov, Martin Heimann, E.A.G. Edward A. G. Schuur, Joel E. Kostka, Olaf Kolle, Ines Hilke, Mathias Göckede, Iulia Ille, Marcus Wildner, Kirsten Küsel, Lutz Merbold, M.D. Miguel D. Mahecha, Nikita Zimov, Sergey A. S.A. Zimov, Martin Heimann, E.A.G. Edward A. G. Schuur, Joel E. Kostka, Olaf Kolle, Ines Hilke, Mathias Göckede, Iulia Ilie, Marcus Wildner, Kirsten Küsel, Lutz Merbold, M.D. Miguel D. Mahecha, Nikita Zimov, Sergey A. S.A. Zimov, Martin Heimann, E.A.G. Edward A. G. Schuur, Joel E. Kostka, Olaf Kolle, Ines Hilke, and Mathias Göckede. Plants, microorganisms and soil temperatures contribute to a decrease in methane fluxes on a drained Arctic floodplain. *Global Change Biology*, 23(6), nov 2016. ISSN 13541013. doi: 10.1111/gcb.13558. URL http://doi.wiley.com/10.1111/gcb.13558. 107, 116

W. B. Langdon. Quadratic Bloat in Genetic Programming. In *Proceedings of the Genetic and Evolutionary Computation Conference,*, 2000. 6, 89

W. B. (William B.) Langdon and Riccardo Poli. *Foundations of genetic programming*. Springer, 2002. ISBN 9783662047262. 4, 5

G. Lasslop, M. Reichstein, J. Kattge, and D. Papale. Influences of observation errors in eddy flux data on inverse model parameter estimation. *Biogeosciences Discussions*, 5(1):751–785, 2008. ISSN 1726-4189. doi: 10.5194/bgd-5-751-2008. 44

G. Lasslop, M. Migliavacca, G. Bohrer, M. Reichstein, M. Bahn, a. Ibrom, C. Jacobs, P. Kolari, D. Papale, T. Vesala, G. Wohlfahrt, and a. Cescatti. On the choice of the driving temperature for eddy-covariance carbon dioxide flux partitioning. *Biogeosciences*, 9(12):5243–5259, 2012. ISSN 17264170. doi: 10.5194/bg-9-5243-2012. 24

Martin Lavoie, C. L. Phillips, and David Risk. A practical approach for uncertainty quantification of high-frequency soil respiration using Forced Diffusion chambers. *Journal of Geophysical Research: Biogeosciences*, 120(1):128–146, jan 2015. ISSN 21698953. doi: 10.1002/2014JG002773. URL http://doi.wiley.com/10.1002/2014JG002773. 24

Miguel Lazaro-Gredilla, Michalis K. Titsias, Jochem Verrelst, and Gustavo Camps-Valls. Retrieval of Biophysical Parameters With Heteroscedastic Gaussian Processes. *IEEE Geoscience and Remote Sensing Letters*, 11(4):838–842, apr 2014. ISSN 1545-598X. doi: 10.1109/LGRS.2013.2279695. URL http://ieeexplore.ieee.org/document/6595574/. 24, 77

J Lloyd and J a Taylor. On the temperature dependence of soil respiration. *Functional Ecology*, 8(3):315–323, 1994. 28, 140

Y Luo, Trevor F Keenan, and Matthew J Smith. Predictability of the terrestrial carbon cycle. *Global Change Biology*, 21(5):1737–1751, 2015. ISSN 1365-2486. doi: 10.1111/gcb.12766. URL http://www.ncbi.nlm.nih.gov/pubmed/25327167. 2, 11, 116

S. Luyssaert, M. Reichstein, E. D. Schulze, I. a. Janssens, B. E. Law, D. Papale, D. Dragoni, M. L. Goulden, a. Granier, W. L. Kutsch, S. Linder, G. Matteucci, E. Moors, J. W. Munger, K. Pilegaard, M. Saunders, and E. M. Falge. Toward a consistency cross-check of eddy covariance flux-based and biometric estimates of ecosystem carbon balance. *Global Biogeochemical Cycles*, 23(3):1–13, jul 2009. ISSN 08866236. doi: 10.1029/2008GB003377. URL http://www.agu.org/pubs/crossref/2009/2008GB003377.shtml. 117

Miguel D Mahecha, Markus Reichstein, Nuno Carvalhais, Gitta Lasslop, Holger Lange, Sonia I Seneviratne, Rodrigo Vargas, Christof Ammann, M Altaf Arain, Alessandro Cescatti, Ivan a Janssens, Mirco Migliavacca, Leonardo Montagnani, and Andrew D Richardson. Global convergence in the temperature sensitivity of respiration at ecosystem level. *Science (New York, N.Y.)*, 329(5993):838–40, aug 2010. ISSN 1095-9203. doi: 10.1126/science.1189587. URL http://www.ncbi.nlm.nih.gov/pubmed/20603495. 26, 46, 116

Willard G Manning. The Logged dependent variable, heteroskedasticity, and the retransformation problem. *Journal of Health Economics*, 17:283–295, 1998. ISSN 01676296. doi: 10.1016/S0167-6296(98)00025-3. URL Onfile. 27, 45

M Mastepanov, C Sigsgaard, T Tagesson, L Ström, M P Tamstorf, M Lund, and T R Christensen. Revisiting factors controlling methane emissions from high-Arctic tundra. *Biogeosciences*, 10(7):5139–5158, 2013. doi: 10.5194/bg-10-5139-2013. URL http://www.biogeosciences.net/10/5139/2013/. 105

Mikhail Mastepanov, Charlotte Sigsgaard, Edward J Dlugokencky, Sander Houweling, Lena Ström, Mikkel P Tamstorf, and Torben R Christensen. Large tundra methane burst during onset of freezing. *Nature*, 456(7222):628–630, 2008. ISSN 0028-0836. doi: 10.1038/nature07464. 105

Charles McCain, Stanford Hooker, Gene Feldman, and Paul Bontempi. Satellite data for ocean biology, biogeochemistry, and climate research. *Eos, Transactions American Geophysical Union*, 87(34):337, aug 2006. ISSN 0096-3941. doi: 10.1029/2006EO340002. URL http://doi.wiley.com/10.1029/2006EO340002. 2

Katherine Rose McEwing, James Paul Fisher, and Donatella Zona. Environmental and vegetation controls on the spatial variability of CH4 emission from wet-sedge and tussock tundra ecosystems in the Arctic. *Plant and Soil*, 388 (1-2):37–52, jan 2015. ISSN 0032-079X. doi: 10.1007/s11104-014-2377-1. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4372828{%}7B{&}{%}7Dtool=pmcentrez{%}7B{&}{%}7Drendertype=abstract. 105

L Merbold, W L Kutsch, C Corradi, O Kolle, C Rebmann, P C Stoy, S A Zimov, and E D Schulze. Artificial drainage and associated carbon fluxes (CO2/CH4) in a tundra ecosystem. *Global Change Biology*, 15(11):2599–2614, 2009. doi: 10.1111/j.1365-2486.2009.01962.x. URL {%}7B{%}3C{%}7DGoto. 105

M. Migliavacca, O. Sonnentag, T. F. Keenan, a. Cescatti, J. O'Keefe, and a. D. Richardson. On the uncertainty of phenological responses to climate change, and implications for a terrestrial biosphere model. *Biogeosciences*, 9(6):2063–2083, 2012. ISSN 17264170. doi: 10.5194/bg-9-2063-2012. 12

Mirco Migliavacca, Markus Reichstein, Andrew D. Richardson, Roberto Colombo, Mark a. Sutton, Gitta Lasslop, Enrico Tomelleri, Georg Wohlfahrt, Nuno Carvalhais, Alessandro Cescatti, Miguel D. Mahecha, Leonardo Montagnani, Dario Papale, Sönke Zaehle, Altaf Arain, Almut Arneth, T. Andrew Black, Arnaud Carrara, Sabina Dore, Damiano Gianelle, Carole Helfter, David Hollinger, Werner L. Kutsch, Peter M. Lafleur, Yann Nouvellon, Corinna Rebmann, R. Humberto, Mirco Rodeghiero, Olivier Roupsard, Maria Teresa Sebastià, Guenther Seufert,

Jean Francoise Soussana, and K. Michiel. Semiempirical modeling of abiotic and biotic factors controlling ecosystem respiration across eddy covariance sites. *Global Change Biology*, 17(1):390–409, jan 2011. ISSN 13541013. doi: 10.1111/j.1365-2486.2010.02243.x. URL http://doi.wiley.com/10.1111/j.1365-2486.2010.02243.x. 26, 28, 140

Mirco Migliavacca, Markus Reichstein, Andrew D. Richardson, Miguel D. Mahecha, Edoardo Cremonese, Nicolas Delpierre, Marta Galvagno, Beverly E. Law, Georg Wohlfahrt, T. Andrew Black, Nuno Carvalhais, Guido Ceccherini, Jiquan Chen, Nadine Gobron, Ernest Koffi, J. William Munger, Oscar Perez-Priego, Monica Robustelli, Enrico Tomelleri, and Alessandro Cescatti. Influence of physiological phenology on the seasonal pattern of ecosystem respiration in deciduous forests. *Global Change Biology*, pages 363–376, 2015. ISSN 13541013. doi: 10.1111/gcb.12671. 20, 48

Stephen Mitchell, Keith Beven, and Jim Freer. Multiple sources of predictive uncertainty in modeled estimates of net ecosystem CO2 exchange. *Ecological Modelling*, 220(23):3259–3270, dec 2009. ISSN 03043800. doi: 10.1016/j.ecolmodel.2009.08.021. URL http://www.sciencedirect.com/science/article/pii/S0304380009006000. 20

J.B. Moncrieff, J.M. Massheder, H. de Bruin, J. Elbers, T. Friborg, B. Heusinkveld, P. Kabat, S. Scott, H. Soegaard, and A. Verhoef. A system to measure surface fluxes of momentum, sensible heat, water vapour and carbon dioxide. *Journal of Hydrology*, 188-189:589–611, feb 1997. ISSN 00221694. doi: 10.1016/S0022-1694(96)03194-0. URL http://www.sciencedirect.com/science/article/pii/S0022169496031940. 25

L A Morrissey and G P Livingston. Methane emissions from Alaska arctic tundra - an assessment of local spatial variability. *Journal of Geophysical Research-Atmospheres*, 97:16661–16670, 1992. URL {%}7B{%}3C{%}7DGoto. 105

Fernando E. Moyano, Werner L. Kutsch, and Corinna Rebmann. Soil respiration fluxes in relation to photosynthetic activity in broad-leaf and needle-leaf forest stands. *Agricultural and Forest Meteorology*, 148(1):135–143, 2008. ISSN 01681923. doi: 10.1016/j.agrformet.2007.09.006. 26

Tomoko Nakano, Shunich Kuniyoshi, and Masami Fukuda. Temporal variation in methane emission from tundra wetlands in a permafrost area, northeastern Siberia. *Atmospheric Environment*, 34(8):1205–1213, jan 2000. ISSN 13522310. doi: 10.1016/S1352-2310(99)00373-8. URL http://www.sciencedirect.com/science/article/pii/S1352231099003738. 105

J.E. Nash and J.V. Sutcliffe. River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*, 10(3):282–290, apr

1970. ISSN 00221694. doi: 10.1016/0022-1694(70)90255-6. URL http://www.sciencedirect.com/science/article/pii/0022169470902556. 20, 69, 120

Michael C. Newman. Regression Analysis of Log-Transformed Data -Statistical Bias and Its Correction (Short Communication), 1993. 27

Jonathan A O'Donnell, M Torre Jorgenson, Jennifer W Harden, A David McGuire, Mikhail Z Kanevskiy, and Kimberly P Wickland. The effects of permafrost thaw on soil hydrologic, thermal, and carbon dynamics in an Alaskan peatland. *Ecosystems*, 15(2):213–229, nov 2011. ISSN 1432-9840. doi: 10.1007/s10021-011-9504-0. URL http://link.springer.com/10.1007/s10021-011-9504-0. 105

Michael O'Neill, Leonardo Vanneschi, Steven Gustafson, and Wolfgang Banzhaf. Open issues in genetic programming. *Genetic Programming and Evolvable Machines*, 11(3-4):339–363, sep 2010. ISSN 1389-2576. doi: 10.1007/s10710-010-9113-2. URL http://link.springer.com/10.1007/s10710-010-9113-2. 147

Shushi Peng, Philippe Ciais, Frédéric Chevallier, Philippe Peylin, Patricia Cadule, Stephen Sitch, Shilong Piao, Anders Ahlström, Chris Huntingford, Peter Levy, Xiran Li, Yongwen Liu, Mark Lomas, Benjamin Poulter, Nicolas Viovy, Tao Wang, Xuhui Wang, Sönke Zaehle, Ning Zeng, Fang Zhao, and Hongfang Zhao. Benchmarking the seasonal cycle of CO2 fluxes simulated by terrestrial ecosystem models. *Global Biogeochemical Cycles*, pages 46–64, 2014a. doi: 10.1002/2014GB004931. Received. 11

YuZhong Peng, ChangAn Yuan, Xiao Qin, JiangTao Huang, and YaBing Shi. An improved Gene Expression Programming approach for symbolic regression problems. *Neurocomputing*, 137:293–301, aug 2014b. ISSN 09252312. doi: 10.1016/j.neucom.2013.05.062. URL http://linkinghub.elsevier.com/retrieve/pii/S0925231214002598. 14

Oscar Pérez-Priego, Ana López-Ballesteros, Enrique P. Sánchez-Cañete, Penélope Serrano-Ortiz, Lars Kutzbach, Francisco Domingo, Werner Eugster, Andrew S. Kowalski, Enrique P. Sánchez-Cañete, Penélope Serrano-Ortiz, Andrew S. Kowalski, Ana López-Ballesteros, Francisco Domingo, Lars Kutzbach, Werner Eugster, and Oscar Pérez-Priego. Analysing uncertainties in the calculation of fluxes using whole-plant chambers: random and systematic errors. *Plant and Soil*, 393(1-2): 229–244, aug 2015. ISSN 0032-079X. doi: 10.1007/s11104-015-2481-x. URL http://link.springer.com/10.1007/s11104-015-2481-x. 44

Riccardo Poli, William B. (William B.) Langdon, Nicholas F. Mcphee, and John R. Koza. *A Field Guide to Genetic Programming*. [Lulu Press], lulu.com, 2008. ISBN 9781409200734. URL http://www.gp-field-guide.org.uk. 4, 5

Markus Reichstein and Christian Beer. Soil respiration across scales: The importance of a model-data integration framework for data interpretation. *Journal of Plant Nutrition and Soil Science*, 171(3):344–354, jun 2008. ISSN 14368730. doi: 10.1002/jpln.200700075. URL http://doi.wiley.com/10.1002/jpln.200700075. 12, 28, 140

Markus Reichstein, Eva Falge, Dennis Baldocchi, Dario Papale, Marc Aubinet, Paul Berbigier, Christian Bernhofer, Nina Buchmann, Tagir Gilmanov, André Granier, Thomas Grünwald, Katka Havránková, Hannu Ilvesniemi, Dalibor Janous, Alexander Knohl, Tuomas Laurila, Annalea Lohila, Denis Loustau, Giorgio Matteucci, Tilden Meyers, Franco Miglietta, Jean Marc Ourcival, Jukka Pumpanen, Serge Rambal, Eyal Rotenberg, Maria Sanz, John Tenhunen, Günther Seufert, Francesco Vaccari, Timo Vesala, Dan Yakir, and Riccardo Valentini. On the separation of net ecosystem exchange into assimilation and ecosystem respiration: Review and improved algorithm. *Global Change Biology*, 11(9):1424–1439, 2005. ISSN 13541013. doi: 10.1111/j.1365-2486.2005.001002.x. 25, 118

Andrew D. Richardson, Miguel D. Mahecha, Eva Falge, Jens Kattge, Antje M. Moffat, Dario Papale, Markus Reichstein, Vanessa J. Stauch, Bobby H. Braswell, Galina Churkina, Bart Kruijt, and David Y. Hollinger. Statistical properties of random CO2 flux measurement uncertainty inferred from model residuals. *Agricultural and Forest Meteorology*, 148(1):38–50, jan 2008. ISSN 01681923. doi: 10.1016/j.agrformet.2007.09.001. URL http://linkinghub.elsevier.com/retrieve/pii/S0168192307002365. 12, 28, 44, 140

O. A. Rosso, H. A. Larrondo, M. T. Martin, A. Plastino, and M. A. Fuentes. Distinguishing Noise from Chaos. *Physical Review Letters*, 99(15):154102, oct 2007. ISSN 0031-9007. doi: 10.1103/PhysRevLett.99.154102. URL http://link.aps.org/doi/10.1103/PhysRevLett.99.154102. 20

M. D. A. Rounsevell, A. Arneth, P. Alexander, D. G. Brown, N. de Noblet-Ducoudré, E. Ellis, J. Finnigan, K. Galvin, N. Grigg, I. Harman, J. Lennox, N. Magliocca, D. Parker, B. C. O'Neill, P. H. Verburg, and O. Young. Towards decision-based global land use models for improved understanding of the Earth system. *Earth System Dynamics*, 5(1):117–137, feb 2014. ISSN 2190-4987. doi: 10.5194/esd-5-117-2014. URL https://www.earth-syst-dynam.net/5/117/2014/. 2

Michael G. Ryan and Beverly E. Law. Interpreting, measuring, and modeling soil respiration. *Biogeochemistry*, 73(1):3–27, mar 2005. ISSN 01682563. doi: 10.1007/s10533-004-5167-7. URL http://www.springerlink.com/index/10.1007/s10533-004-5167-7. 44

M C Serreze, J E Walsh, F S Chapin III, T Osterkamp, M Dyurgerov, V Romanovsky, W C Oechel, J Morison, T Zhang, and R G Barry. Observational evidence of recent

change in the northern high-latitude environment. *Climatic Change*, 46(1-2):159–207, 2000. ISSN 1573-1480. doi: 10.1023/A:1005504031923. URL http://link.springer.com/article/10.1023/A{%}7B{%}25{%}7D3A1005504031923. 105

Mark C Serreze and Roger G Barry. Processes and impacts of Arctic amplification: A research synthesis. *Global and Planetary Change*, 77(1-2):85–96, may 2011. ISSN 09218181. doi: 10.1016/j.gloplacha.2011.03.004. URL http://www.sciencedirect.com/science/article/pii/S0921818111000397. 105

Junjiong Shao, Xuhui Zhou, Yiqi Luo, Bo Li, Mika Aurela, David Billesbach, Peter D. Blanken, Rosvel Bracho, Jiquan Chen, Marc Fischer, Yuling Fu, Lianhong Gu, Shijie Han, Yongtao He, Thomas Kolb, Yingnian Li, Zoltan Nagy, Shuli Niu, Walter C. Oechel, Krisztina Pinter, Peili Shi, Andrew Suyker, Margaret Torn, Andrej Varlagin, Huimin Wang, Junhua Yan, Guirui Yu, and Junhui Zhang. Biotic and climatic controls on interannual variability in carbon fluxes across terrestrial ecosystems. *Agricultural and Forest Meteorology*, 205:11–22, 2015. ISSN 01681923. doi: 10.1016/j.agrformet.2015.02.007. URL http://linkinghub.elsevier.com/retrieve/pii/S0168192315000362. 116

Z Shi, F Wang, and Y Liu. Response of soil respiration under different mycorrhizal strategies to precipitation and temperature. *Journal of Soil Science and Plant Nutrition*, 12(3):411–420, 2012. ISSN 07189516. doi: Doi10.4067/S0718-95162013005000053. 45

Jiří Šimůnek and Donald L. Suarez. Modeling of carbon dioxide transport and production in soil: 1. Model development. *Water Resources Research*, 29(2):487–497, feb 1993. ISSN 00431397. doi: 10.1029/92WR02225. URL http://doi.wiley.com/10.1029/92WR02225. 2

S; Sippel, H; Lange, MD; Mahecha, M; Hauhs, F; Gans, P; Bodesheim, and OA Rosso. Diagnosing the dynamics of observed and simulated ecosystem gross primary productivity with time causal information theory quantifiers. *PLoS ONE*, 4/2016, in, 2016. 21, 120

P W H Smith. Controlling Code Growth in Genetic Programming. *Advances in Soft Computing*, (1998):166–171, 2000. URL http://citeseer.ist.psu.edu/475882.html. 59, 89

C S Sturtevant, W C Oechel, D Zona, Y Kim, and C E Emerson. Soil moisture control over autumn season methane flux, Arctic Coastal Plain of Alaska. *Biogeosciences*, 9(4):1423–1440, apr 2012. ISSN 1726-4189. doi: 10.5194/bg-9-1423-2012. URL http://www.biogeosciences.net/9/1423/2012/bg-9-1423-2012.html. 105

Jens-Arne Subke, Ilaria Inglima, and M Francesca Cotrufo. Trends and methodological impacts in soil CO2 efflux partitioning: A metaanalytical review. *Global Change Biology*, 12(6):921–943, jun 2006. ISSN 1354-1013. doi: 10.1111/j.1365-2486.2006.01117.x. URL http://doi.wiley.com/10.1111/j.1365-2486.2006.01117.x. 45

Gianluca Tramontana, Martin Jung, Christopher R. Schwalm, Kazuhito Ichii, Gustau Camps-Valls, Botond Ráduly, Markus Reichstein, M. Altaf Arain, Alessandro Cescatti, Gerard Kiely, Lutz Merbold, Penelope Serrano-Ortiz, Sven Sickert, Sebastian Wolf, and Dario Papale. Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms. *Biogeosciences*, 13(14): 4291–4313, jul 2016. ISSN 1726-4189. doi: 10.5194/bg-13-4291-2016. URL http://www.biogeosciences.net/13/4291/2016/. 24, 77

Seydou Traore and Aytac Guven. New algebraic formulations of evapotranspiration extracted from gene-expression programming in the tropical seasonally dry regions of West Africa. *Irrigation Science*, 31(1):1–10, may 2013. ISSN 03427188. doi: 10.1007/s00271-011-0288-y. URL http://link.springer.com/10.1007/s00271-011-0288-y. 14

Susan Trumbore. Carbon respired by terrestrial ecosystems–recent progress and challenges. *Global Change Biology*, 2:141–153, 2006. ISSN 1354-1013. doi: 10.1111/j.1365-2486.2005.01067.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2486.2006.01067.x/full. 12

Shiro Tsuyuzaki, Tomoko Nakano, Shun-ichi Kuniyoshi, and Masami Fukuda. Methane flux in grassy marshlands near Kolyma River, north-eastern Siberia. *Soil Biology and Biochemistry*, 33(10):1419–1423, 2001. ISSN 0038-0717. URL https://www.infona.pl//resource/bwmeta1.element.elsevier-4ad54d6d-c727-3794-948a-fb375a4cff9d. 105

Alexander Tøsdal Tveit, Tim Urich, Peter Frenzel, and Mette Marianne Svenning. Metabolic and trophic interactions modulate methane production by Arctic peat microbiota in response to warming. *Proceedings of the National Academy of Sciences of the United States of America*, 112(19):E2507—-E2516, may 2015. ISSN 1091-6490. doi: 10.1073/pnas.1420797112. URL http://www.pnas.org/content/112/19/E2507. 105

Ekaterina J. Vladislavleva, Guido F. Smits, and Dick den Hertog. Order of nonlinearity as a complexity measure for models generated by symbolic regression via pareto genetic programming. *IEEE Transactions on Evolutionary Computation*, 13 (2):333–349, 2009. ISSN 1089778X. doi: 10.1109/TEVC.2008.926486. 3, 66, 90, 148

Jannis von Buttlar, Jakob Zscheischler, Anja Rammig, Sebastian Sippel, Markus Reichstein, Alexander Knohl, Martin Jung, Olaf Menzer, M. Altaf Arain, Nina Buchmann, Alessandro Cescatti, Damiano Gianelle, Gerard Kiely, Beverly E. Law, Vincenzo Magliulo, Hank Margolis, Harry McCaughey, Lutz Merbold, Mirco Migliavacca, Leonardo Montagnani, Walter Oechel, Marian Pavelka, Matthias Peichl, Serge Rambal, Antonio Raschi, Russell L. Scott, Francesco P. Vaccari, Eva van Gorsel, Andrej Varlagin, Georg Wohlfahrt, and Miguel D. Mahecha. Impacts of droughts and extreme-temperature events on gross primary production and ecosystem respiration: a systematic assessment across ecosystems and climate zones. *Biogeosciences*, 15(5):1293–1318, mar 2018. ISSN 1726-4189. doi: 10.5194/bg-15-1293-2018. URL `https://www.biogeosciences.net/15/1293/2018/`. 147

R. Wehr, J. W. Munger, J. B. McManus, D. D. Nelson, M. S. Zahniser, E. A. Davidson, S. C. Wofsy, and S. R. Saleska. Seasonality of temperate forest photosynthesis and daytime respiration. *Nature*, 534(7609):680–683, jun 2016. ISSN 0028-0836. doi: 10.1038/nature17966. URL `http://www.nature.com/doifinder/10.1038/nature17966`. 26

David R. White, James McDermott, Mauro Castelli, Luca Manzoni, Brian W. Goldman, Gabriel Kronberger, Wojciech Jaśkowski, Una May O'Reilly, and Sean Luke. Better GP benchmarks: Community survey results and proposals. *Genetic Programming and Evolvable Machines*, 14(1):3–29, 2013. ISSN 13892576. doi: 10.1007/s10710-012-9177-2. 66

M. Wilkinson, E. L. Eaton, M. S J Broadmeadow, and J. I L Morison. Inter-annual variation of carbon uptake by a plantation oak woodland in south-eastern England. *Biogeosciences*, 9(12):5373–5389, 2012. ISSN 17264170. doi: 10.5194/bg-9-5373-2012. 25, 26

M Williams, A D Richardson, M Reichstein, P C Stoy, P Peylin, H Verbeeck, N Carvalhais, and M Jung. Improving land surface models with FLUXNET data. *Biogeosciences*, (6):1341–1359, 2009. URL `http://www.biogeosciences.net/6/1341/2009/bg-6-1341-2009.pdf`. 12

B. Yegnanarayana. *Artificial neural networks*. PHI Learning Pvt. Ltd., 2009. ISBN 8120312538, 9788120312531. 24, 77

Gabriel Yvon-Durocher, Jane M. Caffrey, Alessandro Cescatti, Matteo Dossena, Paul del Giorgio, Josep M. Gasol, José M. Montoya, Jukka Pumpanen, Peter A. Staehr, Mark Trimmer, Guy Woodward, and Andrew P. Allen. Reconciling the temperature dependence of respiration across timescales and ecosystem types. *Nature*, 487 (7408):472–476, jul 2012. ISSN 0028-0836. doi: 10.1038/nature11205. URL `http://www.nature.com/articles/nature11205`. 5

Massimiliano Zanin, Luciano Zunino, Osvaldo A. Rosso, and David Papo. Permutation Entropy and Its Main Biomedical and Econophysics Applications: A Review. *Entropy*, 14(12):1553–1577, aug 2012. ISSN 1099-4300. doi: 10.3390/e14081553. URL `http://www.mdpi.com/1099-4300/14/8/1553/`. 21

Jakob Zscheischler, Miguel D. Mahecha, Valerio Avitabile, Leonardo Calle, Nuno Carvalhais, Philippe Ciais, Fabian Gans, Nicolas Gruber, Jens Hartmann, Martin Herold, Kazuhito Ichii, Martin Jung, Peter Landschützer, Goulven G. Laruelle, Ronny Lauerwald, Dario Papale, Philippe Peylin, Benjamin Poulter, Deepak Ray, Pierre Regnier, Christian Rödenbeck, Rosa M. Roman-Cuesta, Christopher Schwalm, Gianluca Tramontana, Alexandra Tyukavina, Riccardo Valentini, Guido Van Der Werf, Tristram O. West, Julie E. Wolf, and Markus Reichstein. Reviews and syntheses: An empirical spatiotemporal description of the global surface-atmosphere carbon fluxes: Opportunities and data limitations. *Biogeosciences*, 14 (15):3685–3703, 2017. ISSN 17264189. doi: 10.5194/bg-14-3685-2017. 121, 142