Evolving Interpolating Models of Net Ecosystem CO_2 Exchange Using Grammatical Evolution

Miguel Nicolau[1], Matthew Saunders[2], Michael O'Neill[1], Bruce Osborne[2], and Anthony Brabazon[1]

[1] Natural Computing Research & Applications Group
University College Dublin, Dublin, Ireland
{Miguel.Nicolau,M.ONeill,Anthony.Brabazon}@ucd.ie
[2] UCD School of Biology and Evironmental Science
University College Dublin, Dublin, Ireland
{Matthew.Saunders,Bruce.Osborne}@ucd.ie

Abstract. Accurate measurements of Net Ecosystem Exchange of CO_2 between atmosphere and biosphere are required in order to estimate annual carbon budgets. These are typically obtained with Eddy Covariance techniques. Unfortunately, these techniques are often both noisy and incomplete, due to data loss through equipment failure and routine maintenance, and require gap-filling techniques in order to provide accurate annual budgets. In this study, a grammar-based version of Genetic Programming is employed to generate interpolating models for flux data. The evolved models are robust, and their symbolic nature provides further understanding of the environmental variables involved.

Keywords: Grammatical evolution, Real-world applications, Symbolic regression.

1 Introduction

Eddy Covariance (EC) techniques are utilised globally to measure Net Ecosystem Exchange (NEE), defined as the net flux of Carbon Dioxide (CO_2) between the atmosphere and the biosphere [9]. NEE represents the balance between photosynthetic carbon uptake and respiratory carbon losses, and is typically measured over 30 minute intervals, which are then summed to give an annual carbon budget. Both short-term information and annual sums are of particular interest to scientists, land managers and policy makers. They allow for a comparison of ecosystem carbon budgets across various land use classes, provide a better understanding of the physiological driving processes, and facilitate an assessment of both inter and intra-annual climatic variability [4].

In order to derive the most accurate annual carbon budget, a complete data set is required; however, average data capture using the Eddy Covariance technique is often as low as 65% [4], due to data loss through equipment failure and routine maintenance. Furthermore, a diurnal bias exists in EC data rejection, due to the limitations of the EC technique at night, when low turbulence conditions occur [1,6,10].

A. Moraglio et al. (Eds.): EuroGP 2012, LNCS 7244, pp. 134–145, 2012.

To augment such fragmented data sets, gap-filling procedures are required to provide a more robust annual dataset [4]. Several gap-filling methodologies are currently employed by the global EC flux community, including linear interpolation, look-up tables, non-linear (semi-empirical) models, artificial neural networks, and multiple imputation techniques [18,9,4,1]. The utilisation of a particular gap-filling methodology is influenced by the experimental site-specific conditions, data availability and the particular end use of the EC data [4], however a particular effort has recently been made within the EC flux community to standardise gap-filling methodologies in order to allow the inter-comparison of different ecosystems, bio-climatic zones and long-term data sets [9]. There is however, a real need to continuously evaluate the accuracy of gap-filling models, which can be difficult to constrain, due to the multiple factors that influence EC measurements. For example, the presence of hysteresis loops in measured daytime NEE data can reduce the ability of semi-empirical light response functions to accurately model daytime NEE [20].

In the work presented here, a grammar-based Genetic Programming system was used to generate interpolating models for NEE data. The results obtained are comparable to the best in the literature [18], and the evolved symbolic models are fine-tunable, and also provide an insight into the effect of different environmental variables. These results highlight once again the real world applicability of evolutionary computation, and genetic programming in general.

The next section introduces the evolutionary algorithm used. Section 3 details the experimental setup, and the results obtained are analysed in Section 4. Finally, Section 5 draws some conclusions and future work directions.

2 Evolutionary Approach

Symbolic Regression is arguably one of the most successful applications of Genetic Programming [12] (GP). The tree structure of GP individuals lends itself to good functional representation and manipulation of sub-expressions, providing solutions that are often very precise, analysable, hand-tunable, and potentially provable.

For the purpose of evolving an EC flux gap-filling model, Grammatical Evolution (GE) [17,22] was used. GE is a grammar-based form of GP [14], which specifies the syntax of solutions in a grammar; this grammar is used to map genotypically evolved strings to syntactically correct phenotypic solutions.

GE performs on par with GP for symbolic regression purposes [17], while its grammar allows for extra control of the syntax of evolved programs, both in terms of biases [16,7] and data-structures used. This allows GE to be applied to a variety of domains, such as Financial Modelling [3], horse gait optimisation [15], wall shear stress analysis in grafted arteries [2], and optimisation of controllers for video-games [19], to name a few.

```
<expr>      ::= + <expr> <expr>
              | * <expr> <expr>
              | x
              | <digit>.<digit>
<digit>     ::= 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9
```

Fig. 1. Example grammar for generation of prefix mathematical expressions

2.1 Mapping Process

To illustrate the mapping process employed in GE, consider the grammar shown in Fig. 1, composed of two *non-terminal* symbols (`<expr>` and `<digit>`) and 14 *terminal* symbols (`+`, `*`, `x`, `.` and `0...9`). Given an integer (genotype) string, such as `(1, 7, 4, 8, 6, 5, 9)`, a program (phenotype) can be constructed, which respects the syntax specified in the grammar.

This works by using each integer to choose productions from the grammar, mapping a given start symbol (typically, the first non-terminal symbol appearing in the grammar) to a sequence of terminal symbols. In this example, the first integer chooses one of the four productions of the start symbol `<expr>`, through the formula $1\%4 = 1$, i.e. the second production is chosen (as the count starts from 0), so the mapping string becomes `* <expr> <expr>`.

The following integer is then used with the leftmost unmapped symbol in the mapping string, so through the formula $7\%4 = 3$ the symbol `<expr>` is replaced by `<digit>.<digit>`, so the string becomes `* <digit>.<digit> <expr>`.

The mapping process continues in this fashion, so in the next step the mapping string becomes `* 4.<digit> <expr>` through the formula $4\%10 = 4$, and through $8\%10 = 8$ it becomes `* 4.8 <expr>`. Finally, the remaining non-terminal symbol is mapped with $6\%4 = 2$, and the final expression becomes `* 4.8 x`, which can then be evaluated.

The evolved strings may not have enough values to fully map syntactic valid programs; several options are available to address this issue, such as reusing the same integer string (in a process called wrapping [17]), assigning the individual the worst possible fitness, or replacing it with a legal individual. In this study, an unmapped individual is replaced by its originating parent.

3 Experimental Setup

3.1 Quality of Data and Input Variables

The calculation of NEE represents the balance between photosynthetic carbon assimilation or gross primary productivity (GPP) and net carbon release through ecosystem respiration (R_{eco}), which can be further sub-divided into autotrophic (R_a) and heterotrophic (R_{het}) components. Daytime NEE data represent the balance between GPP and soil derived R_{het}, while night time NEE data (R_{eco})

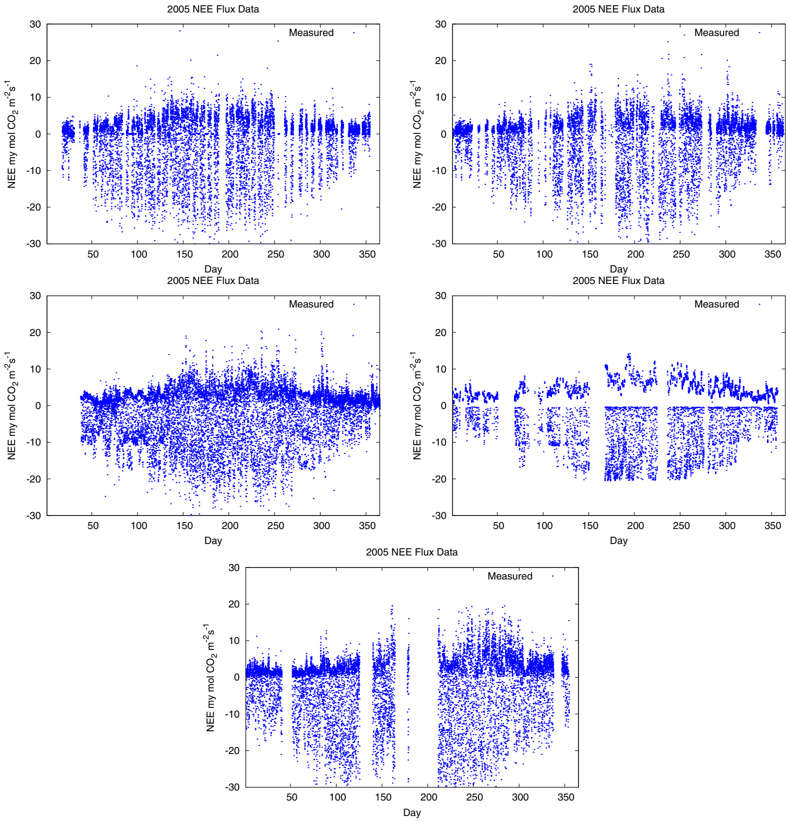

Fig. 2. Observed NEE flux data for the period 2002-2006. Negative NEE values indicate diurnal flux exchanges, whereas positive NEE flux typically occurs at night time.

represent the combined R_a and R_{het} CO_2 efflux from the plant and soil systems combined. The measurement of NEE in this study was made using the closed path EC technique, where fluxes of CO_2 were calculated over 30 minute intervals and the data post-processed and assigned a quality control standard according to the CarboEurope-IP criteria.

Daytime NEE tends to be controlled by both photosynthetic active radiation and air temperature, while R_{eco} is largely a temperature-dependant process. However, even the high quality diurnal flux data show considerable “noise” due to the multiple factors that influence NEE. Figure 2 shows the recorded NEE data for the period 2002-2006. As the annual carbon budget is the typically used unit, and due to annual variations (forest growth and management), each year is treated independently.

As the data is seasonal by nature, the time of day and day of year are used as input variables for model evolution. In order to reduce the linear cumulative numerical weight of these variables ($0 \ldots 23.5$ for time, and $0 \ldots 365$ for day),

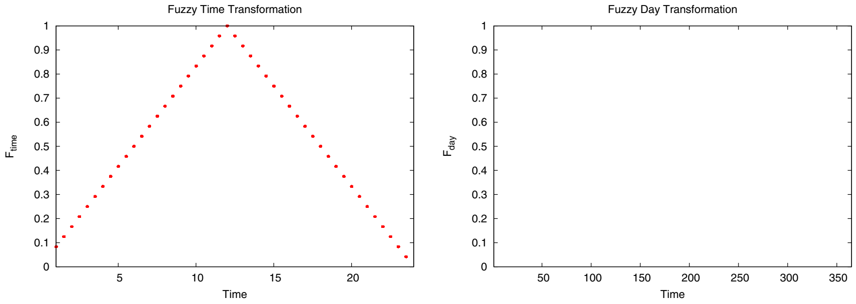

Fig. 3. Fuzzy time and day data transformations used for model evolution

Table 1. Experimental configurations

	Day & Night				Day only				Night only			
	B1	B2	B3	B4	D1	D2	D3	D4	N1	N2	N3	N4
F_{day}, F_{time}, TEMP	x	x	x	x	x	x	x	x	x	x	x	x
PAR	x	x	x	x	x	x	x	x				
sin, cos			x	x			x	x			x	x
TEMP$_{min}$, TEMP$_{max}$, TEMP$_{avg}$		x		x		x		x		x		x
PAR$_{min}$, PAR$_{max}$, PAR$_{avg}$		x		x		x		x				

they were transformed into two fuzzy sets, F_{time} and F_{day}, as seen in previous studies [18]; Fig. 3 shows the fuzzy transformations employed.

Additional meteorological measurements, traditionally used to describe ecosystem carbon flux and model NEE, included air temperature (TEMP), Photosynthetic Active Radiation (PAR), Relative Humidity (RH) and Precipitation (P). Some of these can also exhibit noise in their measurement, and full year-round data is sometimes not available; given the quality of the available data, TEMP and PAR were chosen as meteorological input variables.

Table 1 shows the configurations tested (B1...N4). As diurnal and nocturnal NEE flux dynamics are quite different, models were evolved for either a full dataset, or obtained by combination of separately evolved diurnal and nocturnal models; daytime and night time NEE data were sub-divided based on incident PAR, with data assigned to the daytime data class when $PAR >$ 10 μ mol $m^{-2}s^{-1}$ [13]. Also, given the somewhat regular nature of the data, trigonometric functions were tested in half of the configurations. Finally, some configurations were tested where historical data was used in the function set (PAR$_{min}$, PAR$_{max}$ and PAR$_{avg}$ as the minimum, maximum and average PAR data of the last 24 hours, and likewise for TEMP).

3.2 Evolutionary Setup

Grammar design. The grammars used correspond to the function sets detailed in Tab. 1. They are balanced grammars [7], which helps to control the size of

```
<e> ::= + <e> <e> |  - <e> <e> |  * <e> <e> |  / <e> <e>
      |  + <e> <e> |  - <e> <e> |  * <e> <e> |  / <e> <e>
      |  + <e> <e> |  - <e> <e> |  * <e> <e> |  / <e> <e>
      |  + <e> <e> |  - <e> <e> |  * <e> <e> |  / <e> <e>
      |  + <e> <e> |  - <e> <e> |  * <e> <e> |  / <e> <e>
      |  Fday[i] |  Fhour[i] |  PAR[i] |  TEMP[i] |  <d><d>"."<d>
      |  Fday[i] |  Fhour[i] |  PAR[i] |  TEMP[i] |  <d><d>"."<d>
      |  Fday[i] |  Fhour[i] |  PAR[i] |  TEMP[i] |  <d><d>"."<d>
      |  Fday[i] |  Fhour[i] |  PAR[i] |  TEMP[i] |  <d><d>"."<d>
<d> ::= 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9
```

Fig. 4. Grammar used for setting B1

resulting individuals and thus delaying the onset of bloat; to do so, several equal productions were inserted, to maintain the biases of transformations [16]. Finally, the number of non-terminals was reduced, as this has been shown to help improve the performance of GE [16]. Fig. 4 shows the grammar used for setting B1.

To evolve the integer strings used with GE, a variable-length genetic algorithm was used [8]. The first generation was created using a ramped version of *Sensible Initialisation* [21], resulting in a better spread of initial solutions (albeit not perfect [7]). A “fair” tournament selection was used, where every individual participates at least in one tournament event. Finally, genetic operators were applied only to mapping regions of chromosomes.

Table 2 details the evolutionary parameters used. Note that, since the models evolved for day time and night time are later combined together, the computation effort of their runs doubles that of the runs where a single model is evolved; taking this into account, the population size of the latter (B1. . . B4) is doubled, resulting in a comparable computation effort per generation.

Table 2. Evolutionary Setup

Population Size	500/1000
Generations	50
Derivation-tree Max Depth (for initialisation)	5
Tail Ratio (for initialisation)	50%
Selection Tournament Size	1%
Elitism (for generational replacement)	10%
Crossover Ratio	50%
Average Mutation Events per Individual	1

3.3 Measuring Performance

Evolved models were compared to available NEE data, and the mean squared error between predictions and available data was used as a performance measure. The available NEE data was divided into training and test sets, so as to ascertain

Table 3. Mean squared error (and standard deviation) on test data

	B1	B2	D1 + N1	D1 + N2	N2 + D1	D2 + N2
2002	27.58 (1.49)	26.70 (7.62)	25.32 (0.76)	25.44 (0.83)	27.28 (2.95)	27.40 (3.01)
2003	22.68 (1.50)	25.48 (1.86)	21.15 (1.14)	21.24 (1.08)	21.64 (0.99)	21.78 (0.98)
2004	16.51 (4.74)	19.43 (1.30)	17.54 (1.59)	17.56 (1.55)	19.23 (2.19)	19.00 (1.93)
2005	12.72 (5.23)	22.13 (14.39)	7.13 (2.10)	7.25 (2.09)	8.65 (2.66)	8.66 (2.76)
2006	33.76 (49.94)	25.23 (5.99)	20.15 (2.84)	20.19 (2.78)	24.85 (3.80)	24.77 (3.91)

how well the evolved models generalise to unseen data. In this study, for every four available data points, the first three were used for training, with the fourth used for testing. [1]

4 Results and Analysis

Results using trigonometric functions were on par or worse than the equivalent setups without these functions, and generally produced more complex expressions, so in accordance with the Occam's Razor principle, they were discarded (they are not reported here). Table 3 reports the mean squared error on test data, for all other configurations, averaged over 50 runs. Average minimum error at end of evolution and standard deviation are reported.

The results obtained match the relative quality of different annual data, as could be observed in Fig. 2 (steady improvement of data quality over the years, apart from 2006). Evolving separate daytime and night time models generally provides better performing models. The use of historical data seems to make no difference to the results, but the resulting night time models are more compact on average and were thus preferred. Also note that combined models can be further enhanced, as models obtained in different runs can be matched.

Figure 5 plots the measured NEE flux data for 2005, and the best single (B1) and combined (D1 + N2) models. The difference in performance, particularly for positive NEE values (night time data) is substantial. Also note the occurrence of asymptotes when evolving a single model (Fig. 5 top), suggesting that the use of interval arithmetic [11] might be required to remove these. Figure 6 shows the combined model prediction for April 2005, highlighting both the matching of measured EC and the interpolation of regions with no data recorded.

Figure 7 plots the average training and testing performance over time, for the (D1 + N2) configuration. It can be seen that the model does not overfit the data. Comparison with runs using all the available data for training achieved similar results, suggesting that the use of a 2-set methodology neither hinders nor improves the performance of the obtained models, confirming previous results reported in GP [5] and GE [23].

[1] A more typical division, such as an initial large proportion of data for training and the remaining for testing, is not feasible, given the seasonality of the data, and the uniqueness nature of each year (different models are evolved for different years).

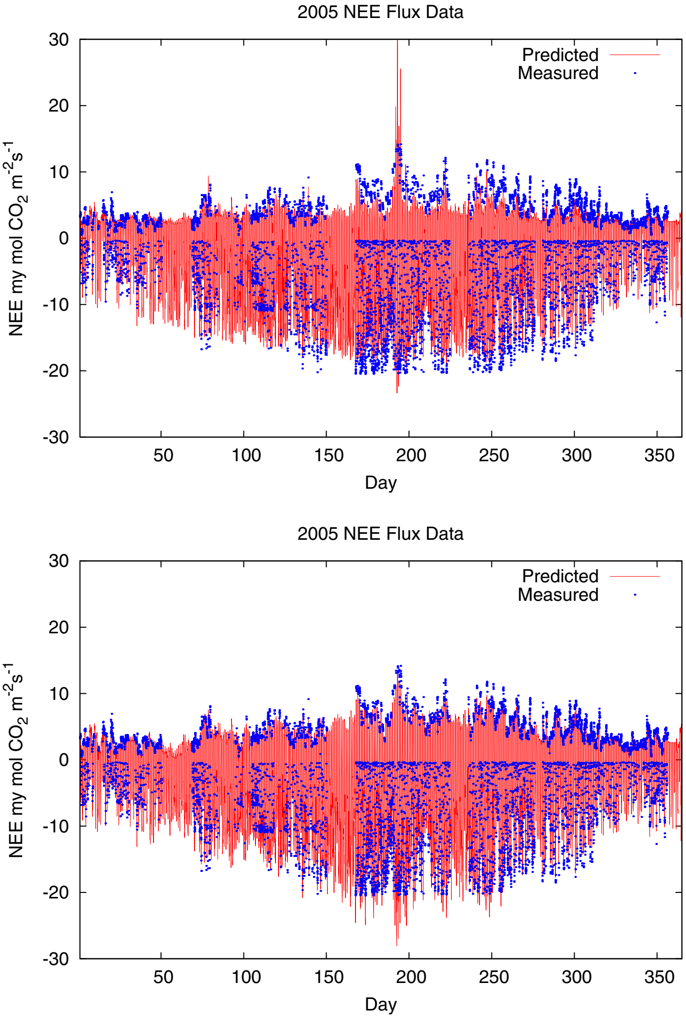

Fig. 5. Eddy value predictions of best full data predictor (top) and combined predictions of the best day time and night time models

The best (D1 + N2) model is shown in Eq. 1. The model has not been simplified; it shows a remarkably compact solution, resulting from the bloat delaying techniques described above, and the relatively short runs (50 generations). Note that night time data only makes use of temperature, showing that the seasonality of this variable is sufficient to match the seasonality of the Eddy values. Similarly, the daytime equation makes no use of F_{time}, showing that daily regularity can be modelled by PAR and $TEMP$. Finally, Fig. 8 shows the correlation between test data and model prediction for this model, including a 1:1 line. The model exhibits a good correlation with measured data, apart from some instances around values close to zero (a mixture of both noisy data and incorrect predictions).

$$\text{eddy} = \begin{cases} \dfrac{\dfrac{93.7}{25.7 + TEMP_{min}} + TEMP}{\dfrac{99.9}{23.6 + TEMP_{avg}}} & \text{if } PAR < 10.0 \\[2em] \dfrac{13.1 + PAR}{\dfrac{TEMP + PAR}{F_{day} - (TEMP + 11.7)} - 21.0} & \text{otherwise} \end{cases} \tag{1}$$

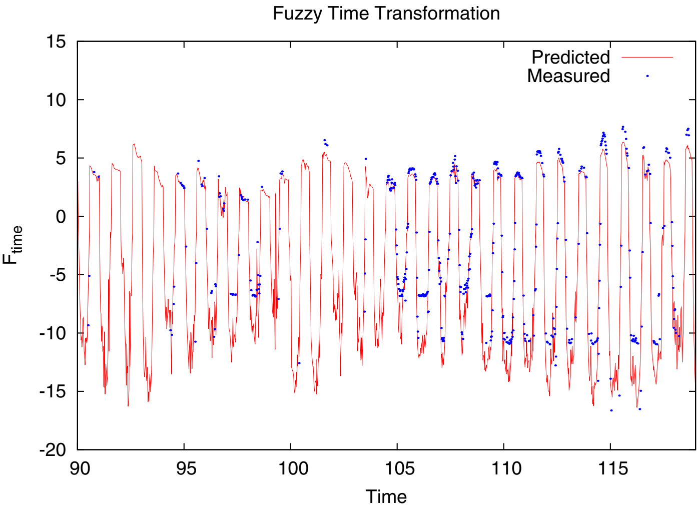

Fig. 6. Measured Eddy value vs. best model prediction, for the month of April 2005

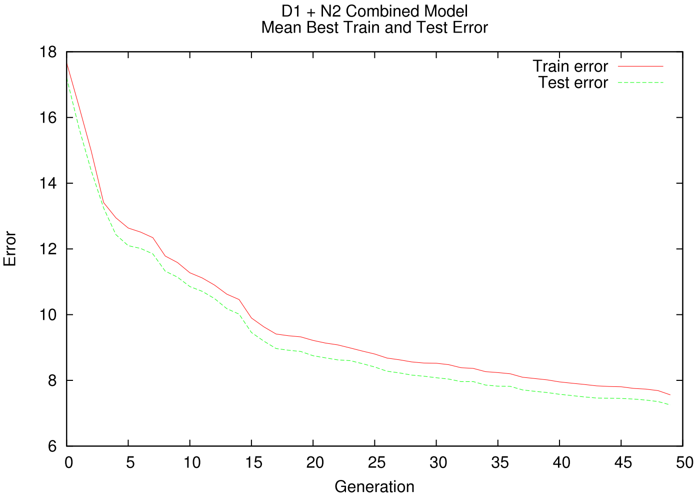

Fig. 7. Training and testing performance for the mean best individual per generation, for (D1 + N2) configuration (averaged across 50 runs)

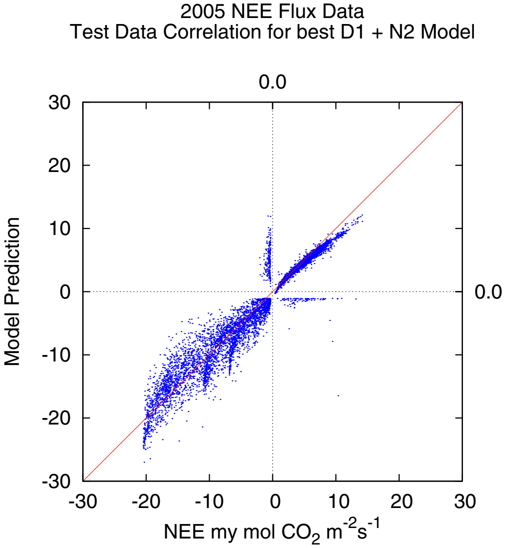

Fig. 8. Correlation between (unseen) test data and model prediction

5 Conclusions

GP in its many flavours has been applied to a multitude of symbolic regression problems over the years, with outstanding results. Yet, in most research fields, standard gap-filling methods such as look-up tables and linear interpolation are still applied as standard. This work presents a collaboration between evolutionary computation practitioners and environmental biologists, in an effort to further highlight the applicability of GP to generate gap-filling models for measured environmental data.

Due to the unique nature of data from different forest sites, proper comparison with other methods is hard to achieve[2]; however, the results obtained seem to be on par with the best in the literature [18]. Not only that, but the use of GP has certain advantages. By providing symbolic models, stating the required input variables, decisions can be made about the required annual measurements, affecting both budget and work force management.

There are plenty of future work directions. The evolved models can have a direct impact on forest management and even policy making, and thus continued efforts to improve their accuracy are ongoing. Another exciting future work direction involves identifying a maximum size of measured data gaps; this allows expensive equipment to be used and rotated across different sites, thus bringing the overall data-gathering costs down. Efforts are ongoing to achieve this.

[2] The data presented here has only been analysed with the current method so far.

Acknowledgements. The authors would like to acknowledge Alexandros Agapitos, for sharing his insightful knowledge of GP and associated techniques. This research is based upon works supported by Science Foundation Ireland under Grant No. 08/IN.1/I1868.

References

1. Moffat, A., et al.: Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes. Agricultural and Forest Meteorology 147, 209–232 (2007)
2. Azad, R.M.A., Ansari, A.R., Ryan, C., Walsh, M., McGloughlin, T.: An evolutionary approach to wall shear stress prediction in a grafted artery. Applied Soft Computing 4(2), 139–148 (2004)
3. Brabazon, A., O'Neill, M.: Biologically Inspired Algorithms for Financial Modelling. Springer, Heidelberg (2006)
4. Falge, E., et al.: Gap filling strategies for defensible annual sums of net ecosystem exchange. Agricultural and Forest Meteorology 107, 43–69 (2001)
5. Gagné, C., Schoenauer, M., Parizeau, M., Tomassini, M.: Genetic Programming, Validation Sets, and Parsimony Pressure. In: Collet, P., Tomassini, M., Ebner, M., Gustafson, S., Ekárt, A. (eds.) EuroGP 2006. LNCS, vol. 3905, pp. 109–120. Springer, Heidelberg (2006)
6. Goulden, M., Munger, W., Fan, S.M., Daube, B., Wofsy, S.: Measurements of carbon sequestration by long-term eddy covariance: methods and critical evaluation of accuracy. Global Change Biology 2, 169–182 (1996)
7. Harper, R.: GE, explosive grammars and the lasting legacy of bad initialisation. In: Proceedings of IEEE Congress on Evolutionary Computation, CEC 2010, July 18-23, Barcelona, Spain, pp. 2602–2609. IEEE Press (2010)
8. Holland, J.H.: Adaptation in Natural and Artificial Systems. University of Michigan Press (1975)
9. Hui, D., Wan, S., Su, B., Katul, G., Monson, R., Luo, Y.: Gap-filling missing data in eddy covariance measurements using multiple imputation (mi) for annual estimates. Agricultural and Forest Meteorology 121, 93–111 (2004)
10. Humphreys, E., Black, T.A., Morgenstern, K., Cai, T., Drewitt, G., Nesic, Z., Trofymow, J.: Carbon dioxide fluxes in coastal douglas-fir stands at different stages of development after clearcut harvesting. Agricultural and Forest Meteorology 140, 6–22 (2006)
11. Keijzer, M.: Improving Symbolic Regression with Interval Arithmetic and Linear Scaling. In: Ryan, C., Soule, T., Keijzer, M., Tsang, E.P.K., Poli, R., Costa, E. (eds.) EuroGP 2003. LNCS, vol. 2610, pp. 70–82. Springer, Heidelberg (2003)
12. Koza, J.R.: Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press (1992)
13. Reichstein, M., et al.: On the separation of net ecosystem exchange into assimilation and ecosystem respiration: review and improved algorithm. Global Change Biology 11, 1424–1439 (2005)
14. McKay, R.I., Nguyen, X.H., Whigham, P.A., Shan, Y., O'Neill, M.: Grammar-based genetic programming - a survey. Genetic Programming and Evolvable Machines 11(3-4), 365–396 (2010)

15. Murphy, J.E., O'Neill, M., Carr, H.: Exploring Grammatical Evolution for Horse Gait Optimisation. In: Vanneschi, L., Gustafson, S., Moraglio, A., De Falco, I., Ebner, M. (eds.) EuroGP 2009. LNCS, vol. 5481, pp. 183–194. Springer, Heidelberg (2009)
16. Nicolau, M.: Automatic grammar complexity reduction in grammatical evolution. In: Poli, R., et al. (eds.) Genetic and Evolutionary Computation Conference (GECCO) Workshops. AAAI (2004)
17. O'Neill, M., Ryan, C.: Grammatical Evolution - Evolutionary Automatic Programming in an Arbitrary Language. Genetic Programming, vol. 4. Kluwer Academic (2003)
18. Papale, D., Valentini, R.: A new assessment of european forests carbon exchanges by eddy fluxes and artificial neural network spatialization. Global Change Biology 9, 525–535 (2003)
19. Perez, D., Nicolau, M., O'Neill, M., Brabazon, A., Yannakakis, G.N.: Evolving Behaviour Trees for the Mario AI Competition Using Grammatical Evolution. In: Di Chio, C., Cagnoni, S., Cotta, C., Ebner, M., Ekárt, A., Esparcia-Alcázar, A.I., Merelo, J.J., Neri, F., Pruess, M., Richter, H., Togelius, J., Yannakakis, G.N. (eds.) EvoApplications 2011, Part I. LNCS, vol. 6624, pp. 123–132. Springer, Heidelberg (2011)
20. Pingintha, N., Leclerc, M., Beasley, J., Durden, D., Zhang, G., Senthong, C., Rowland, D.: Hysteresis response of daytime net ecosystem exchange during drought. Biogeosciences 7, 1159–1170 (2010)
21. Ryan, C., Azad, A.: Sensible initialisation in grammatical evolution. In: Cantú-Paz, E., et al. (eds.) Genetic and Evolutionary Computation Conference (GECCO) Workshops. AAAI (2003)
22. Ryan, C., Collins, J., O'Neill, M.: Grammatical Evolution: Evolving Programs for an Arbitrary Language. In: Banzhaf, W., Poli, R., Schoenauer, M., Fogarty, T.C. (eds.) EuroGP 1998. LNCS, vol. 1391, pp. 83–95. Springer, Heidelberg (1998)
23. Tuite, C., Agapitos, A., O'Neill, M., Brabazon, A.: A Preliminary Investigation of Overfitting in Evolutionary Driven Model Induction: Implications for Financial Modelling. In: Di Chio, C., Brabazon, A., Di Caro, G.A., Drechsler, R., Farooq, M., Grahl, J., Greenfield, G., Prins, C., Romero, J., Squillero, G., Tarantino, E., Tettamanzi, A.G.B., Urquhart, N., Uyar, A.Ş. (eds.) EvoApplications 2011, Part II. LNCS, vol. 6625, pp. 120–130. Springer, Heidelberg (2011)