

Interactive comment on “Soil respiration compartments on an aging managed heathland: can model selection procedures contribute to our understanding of ecosystem processes?” by G. R. Kopittke et al.

G. R. Kopittke et al.

vanloon@uva.nl

Received and published: 19 February 2013

We appreciated the extensive and thoughtful comments by the reviewer a lot. They led to several improvements to the original manuscript. Some remarks or requests by different reviewers were related. So, before answering the questions and issues raised by reviewer #2, we will discuss our view on these general issues and explain our approach in making adjustments to the original manuscript.

Questions regarding model reporting

C8323

All three reviewers made remarks about the types of model details that were reported. Reviewer 1 commented particularly with regards to details of the fittest model, such as including qq plots and information about residual variance. Reviewer 3 indicated he/she had doubts about the low RMSE values for one model (the GLMM-T model) in which the functional form was almost identical to another model (Selsted). In addition, some elaboration was requested by both reviewers on the decisions made within the modeling process or modeling details. The remarks by both reviewers were very appropriate and we have tried to deal with these points by including the requested additions in a new table (Table 4: residual variance), in a new appendix (Appendix D: diagnostic plots) and several sections within the text.

The mixed models used in this study are able to make predictions for different location conditions, depending on one's knowledge of the location for which a prediction is to be made. A paragraph has been added in the methods section which explains that mixed models can make 'location specific' predictions (treating the location factors as 'fixed effects'), or general predictions for a not previously visited location (treating the location factors as 'random effects'). The first predictive mode uses more model parameters and leads to lower prediction errors. We generated location-specific predictions with the GLMM in the original manuscript for the calibration data (where it was possible) but not for the validation data. However, due to the remark by Reviewer 3, we realize that it is better to generate non-location-specific predictions with the mixed models for both the calibration and validation data. This makes it possible to compare the RMSE across all models and calibration/validation sets and avoids confusion.

Questions related to the validation procedure and error criteria

Each of the reviewers requested an adjustment with regard to model performance criteria and calibration/validation schemes: Reviewer 1 suggested a comparison of different calibration – validation methods (e.g. cross-validation, or calibrate using all data with no separate validation). Reviewer 2 suggested the calculation and application of the Akaike information criterion for model selection. Reviewer 3 suggested that the Nash

C8324

criterion be used instead of a RMSE. We think that each of these suggestions is appropriate and interesting by itself. However, including all of these additional statistics in the results and discussing them accordingly would lead to a very long manuscript with the risk of losing focus. We have therefore chosen to include an additional appendix (Appendix C) which considers a number of additional calibration / validation methods and additional error metrics. This Appendix includes a summary of the additional methods, the results from an additional model calibration (where all data is included), additional model validation (cross-validation), and two additional error criteria (the Nash criterion and a recent modification to it). The additional methods and goodness-of-fit criteria all resulted in very similar outcomes to the results already included in the original manuscript. Appendix C is referred to from the relevant places in the results and discussion sections.

As indicated by many studies (e.g. Wilmott et al., 1985; Legates and McCabe, 1999), every model performance criterion measures something different, and the choice for a certain criterion is dependent on the model aim, i.e. which aspect(s) of the system should the model especially reproduce well.

We do not have an opinion about the most suitable criterion to evaluate the prediction of yearly soil respiration (and think the best criterion is context dependent), but in our original manuscript chose to use RMSE, since this statistic has been frequently used in recent soil respiration literature (e.g. Chen et al. 2010; Migliavacca et al. 2011; Keenan et al. 2012). After calculating additional error metrics it turned out that the correlation between the various error metrics is very high and we still choose to use the RMSE as a sufficient statistic for the purposes of our study.

Chen, X., Post, W., Norby, R. and Classen, A.: Modeling soil respiration and variations in source components using a multi-factor global climate change experiment, *Climatic Change*, 107, 459-480, 2011.

Keenan, T. F., Davidson, E., Moffat, A. M., Munger, W. and Richardson, A. D.: Using

C8325

model-data fusion to interpret past trends, and quantify uncertainties in future projections, of terrestrial ecosystem carbon cycling, *Global Change Biol.*, 18, 2555-2569, 2012.

Legates, D.R., and McCabe Jr., G. J. : Evaluating the Use of "Goodness-of-Fit" Measures in Hydrologic and Hydroclimatic Model Validation, *Water Resour. Res.*, 35(1), 233-241, 1999
Migliavacca, M., Reichstein, M., Richardson, A. D., Colombo, R., Sutton, M. A., Lasslop, G., Tomelleri, E., Wohlfahrt, G., Carvalhais, N., Cescatti, A., Mahecha, M. D., Montagnani, L., Papale, D., Zaehle, S., Arain, A., Arneth, A., Black, T. A., Carrara, A., Dore, S. and Gianelle, D.: Semiempirical modeling of abiotic and biotic factors controlling ecosystem respiration across eddy covariance sites, *Global Change Biol.*, 17, 390-409, 2011.

Willmott, C.J., Ackleson, S.G., Davis, R.E., Feddema, J.J., Klink, K.M., Legates, D.R., O'Donnell, J., and Rowe, C.M. : Statistics for the evaluation and comparison of models, *J. Geophys. Res.*, 90, 8995–9005, 1985

General remarks by Reviewer 2

A) Empirical models are used to represent the variability of observed data and therefore they are not a mathematical description of an ecosystem process; which is the main objective of process-based models. That said, the authors need to make a better case in this manuscript for the use of empirical model selection procedures as a tool for understanding ecosystem processes. How these empirical fits describe the ecosystem processes? A clear example about how these empirical models cannot represent ecosystem processes is that they fail to represent the extreme values reported on 21 March 2012. See the large differences between measurements in Figure 4 and model simulations in Figure 9. The extreme values presented in Figure 4 (which likely represent an important flux at the annual scale, and are driven by key ecological processes as discussed in page 16251 lines 26-28) were purposely excluded from the analysis because these empirical models cannot simulate this variability in the observed data

C8326

(i.e., they cannot simulate ecosystem processes). Therefore, I encourage the authors to think and discuss on the applicability and limitations of empirical models for understanding ecosystem processes as stated in the title. At the minimum the word “empirical” should be included in the title. A model is a partial representation of reality and ideally it should reflect the underlying biophysical mechanisms to truly represent and contribute to our understanding of ecosystem processes. Empirical models, such as the ones used in this study, do not translate into mathematical equations the ecosystem processes. At their best these models represent the patterns and magnitudes of the measured variables and it is interpreted that they could provide insights about processes. Furthermore, empirical models cannot be used to extrapolate observations outside the range of measured values and therefore lack of predictive power. Thus, the main question is: If empirical models do not represent ecosystem processes per se (as they are developed to fit observed data); how can empirical model selection procedures contribute to our understanding of ecosystem processes? This is the main goal of the manuscript but I respectfully believe it needs to be developed with more clarity along the manuscript.

The reviewer provides his/her view on the difference between empirical and process-based models, and encourages us to discuss the applicability and limitation of empirical models to understand ecosystem processes. We agree that more can be said about the nature of the models that were used in this study and in the revised manuscript we have included the discussion in the one-before last paragraph of discussion section 4.3.

With respect to the request to add the term ‘empirical’ to the title: we did as requested and the word empirical has been included.

Next, we like to respond to several of the statements by the reviewer in more detail below. “Empirical models are used to represent the variability of observed data and therefore they are not a mathematical description of an ecosystem process; which is the main objective of process-based models”. . . .(and later in the second paragraph):. . .

C8327

“Empirical models, such as the ones used in this study, do not translate into mathematical equations the ecosystem processes. At their best these models represent the patterns and magnitudes of the measured variables and it is interpreted that they could provide insights about processes.”

We would like to argue that the spectrum of different models or modeling approaches is not well represented by the dichotomy between the terms ‘empirical’ and ‘process-based’. Even though dichotomies are often used to simplify discussions and there are many dichotomies to characterize scientific models (deterministic versus stochastic, applied versus theoretical, Lagrangian versus Eulerian, reductionistic versus holistic, dynamic versus static, causal versus correlative, etc.; see Jorgensen & Bendoriccio, 2001), we don’t think this simplification is particularly helpful when discussing the soil respiration models used in our study. There are two reasons why it is, in our opinion, not appropriate to label the models in our study as ‘not process based’: Exponential functions can be seen as steady-state solutions to a dynamic decay process. We will explain this point a bit further. When considering exponential equations in the context of soil respiration (e.g. of the form $x(t) = x_0 e^{-kt}$), where x_0 is the value of x at the start of the decay process), these equations have in fact been derived as solutions to a first order linear process: $(dx(t))/dt = -kx(t)$, representing linear decay or outflow from a linear reservoir. Now, this first order differential equation is generally seen as an example of a process-based model, because it represents the state-variable of interest and defines its dynamics to a well-defined decay/outflow-constant. But at the same time, the closed-form solution to this differential equation (exponential function) holds the same information and is in our view a process-based model as well. Obviously, in a heterogeneous environment, it is not possible to obtain steady state solutions. And applying a static equation is based on a quasi steady-state assumption. However, we think this assumption does not make the model more or less process-based (one could discuss about the validity of a steady state assumption under certain conditions – but that applies to dynamic and static models alike and is in our view not related to a model being process-based or not). Many models which have been labeled as

C8328

being 'process-based' in the literature, e.g. Gerten et al (2008), Kutsch et al (2010) or Keenan et al (2012)), contain the same exponential models that we are fitting as components in a more comprehensive ecosystems-model. In our study, influential factors like soil moisture or photosynthesis are measured and used as boundary conditions or model input, while in those models they are often (but not always) being coupled and calculated internally. The status of a model as being process-based or non-process-based is not influenced by the decision to place it (or not) within a larger framework as a sub-model, nor the decision to specify input to this sub-model as a boundary condition or the output from another sub-model.

On the basis of these two points we disagree with the reviewer's statement that "Empirical models, such as the ones used in this study, do not translate into mathematical equations the ecosystem processes": the models included in our study are solutions to differential equations of well accepted representations of ecosystem processes and as such mathematical equations of these ecosystem processes themselves; and many other studies represent the same processes in the same way, while also labeling them as process-based. Having said this, we do agree with the reviewer that modeling soil respiration dynamically can be very useful and may lead to different insights than the static models used in this study. This is an issue which deserves to be explored further in future research.

Next, we think that the reviewer understands something different with the term 'empirical' than we do. We would define an 'empirical model' as: 'a model which can be compared to real observations and in which (some) parameters are derived by using observations from the physical world'. It appears that the reviewer defines an empirical model as 'one where parameters cannot be related to fundamental ecological processes or constants' (based on the quote in the previous paragraph). So we completely agree that our model is empirical when we use our definition, but not when we use the reviewer's definition. We would contrast an empirical model with a theoretical model, rather than a process-based model (defining a theoretical model as a model

C8329

which would only contain a priori defined parameters, would not involve parameter calibration, and for which the output would not necessarily be compared to observations from the physical world).

Having specified how we disagree on some points with the reviewer, we also like to express our agreement where he/she asks to more elaborately discuss the applicability and limitations of the models we tested for understanding ecosystem processes and more specifically: to provide a clear answer to the question in the title. We agree this aspect was not fully considered in the original manuscript and has now been expanded in the revised manuscript. As an answer to the question in the title: we think we did learn something about ecosystem functioning through the model-selection process. 1) We think that (for the system and scale of our study) some process or variable besides photosynthesis, plant biomass, PAR, root biomass, or microbial biomass (which are all not important) is driving soil respiration together with soil temperature, 2) In the absence of these important covariates, soil temperature explains soil heterotrophic and autotrophic respiration just as well as more complex quasi steady-state models. For our future experimental work this model based on soil temperature can be used for predictive purposes and act as a null model against which future models can be compared. In the revised manuscripts this is stated in the conclusion section.

Gerten, D., Luo, Y., Le Maire, G., Parton, W. J., Keough, C., Weng, E., Beier, C., Ciais, P., Cramer, W., Dukes, J. S., Hanson, P. J., Knapp, A. A. K., Linder, S., Nepstad, D. A. N., Rustad, L. and Sowerby, A.: Modelled effects of precipitation on ecosystem carbon and water dynamics in different climatic zones, *Global Change Biology*, 14, 2365-2379, 2008. Keenan, T. F., Davidson, E., Moffat, A. M., Munger, W. and Richardson, A. D.: Using model-data fusion to interpret past trends, and quantify uncertainties in future projections, of terrestrial ecosystem carbon cycling, *Global Change Biol.*, 18, 2555-2569, 2012.

Kutsch, W., Persson, T., Schrumph, M., Moyano, F., Mund, M., Andersson, S. and Schulze, E.: Heterotrophic soil respiration and soil carbon dynamics in the deciduous

C8330

Hainich forest obtained by three approaches, *Biogeochemistry*, 100, 167-183, 2010.

There is another point where we differ in opinion with the reviewer. The reviewer notes: "A clear example about how these empirical models cannot represent ecosystem processes is that they fail to represent the extreme values reported on 21 March 2012. See the large differences between measurements in Figure 4 and model simulations in Figure 9. The extreme values presented in Figure 4 (which likely represent an important flux at the annual scale, and are driven by key ecological processes as discussed in page 16251 lines 26-28) were purposely excluded from the analysis because these empirical models cannot simulate this variability in the observed data (i.e., they cannot simulate ecosystem processes)."

We fully agree with the reviewer that understanding the response of ecosystems to extreme peaks, which result from extreme freeze periods are important, and could be interesting for formulating hypotheses about changes to the frequency of extreme climatic events. To do this however, we would need to collect more data over at least several years. However, we disagree with the notion that not being able to simulate a certain process is a characteristic for a certain class of models (in this case 'empirical models'). In our opinion a model is designed for a certain purpose (e.g. prediction, understanding, communication) and matched (in scale and detail of the process-descriptions) to the available data. Our models were designed to produce the best-possible autotrophic and heterotrophic soil respiration predictions for the study site, with representations of temperature, soil moisture and plant activity influences in our model. Any nutrient-related processes were deliberately excluded as our observational data for these processes was insufficient. From this perspective, it made sense to compare model-results to the cases that the model represented. At the moment we assume or conclude that a certain event or period is to an important degree determined by variations in nutrient availability – it does not make sense to describe that event/period with our model. We think this logic applies to any model evaluation case. Moreover, we think that in this particular case we are observing an extreme frost event.

C8331

Many models have problems with extreme events (and following from this we currently see an increased research effort on extreme events: droughts, floods, heat waves etc.).

B) The authors used the RMSE as a statistic for model selection (Figure 8), but generally for model evaluation there are several parameters that are important to consider. These are the RMSE, the correlation coefficient and the standard deviation that could be represented in a table or in a diagram (Taylor, 2001). Furthermore, models that have different number of parameters are generally evaluated via the maximum likelihood approach known as the Akaike's Information Criterion (AIC) (Burnham and Anderson, 2002). I strongly suggest including this statistic when evaluating models that have different numbers of parameters.

This comment relates to a suggestion by Reviewer 1 to include cross-validation and by Reviewer 3 to include the Nash-criterion for model assessment. We have given a general response to all three of these comments at the start of the review comments. Here we add a response specifically about the AIC criterion.

We did in fact also calculate AIC values during modeling and, as already foreseen by the reviewer, the low AIC values corresponded with the low RMSE values. We also considered including AIC in reporting model performance in our study when preparing the original manuscript, but eventually decided not to do so because the AIC is not defined for all the models included in our study (this is further explained below).

The reviewer refers to the publication by Taylor, K. E. (2001) 'Summarizing multiple aspects of model performance in a single diagram'. Indeed, this paper provides a very insightful discussion on the use of multiple statistics for model or model-data comparison. We agree with the reviewer that the AIC (as well its adjustments like AICc, and other information criteria like BIC, DIC) are very useful, as these combine the calculation of fitness with a penalty for the use of (additional) parameters. However, this approach has especially added value when evaluating a model on calibration data only. When involving a (cross-)validation scheme, measuring the model performance on the

C8332

independent data automatically determines the generalization error and measures essentially the same as the AIC (or its variants). As a matter of fact, Stone (1977) showed that asymptotically Leave-one-out-cross validation is equivalent to the AIC.

The AIC value as a model-criterion has the disadvantage that it is based on a model likelihood, and this likelihood is not always defined (or cannot always be calculated). In the context of our study we use a penalized quasi-likelihood (PQL) implementation to fit the LMM and GLMM models (Breslow & Clayton 1993, Wolfinger & O'Connell 1993). The likelihood is not defined for such models, hence an AIC cannot be calculated. An additional property of the AIC (which it shares with all other information criteria as well as RMSE) is that its value can only be compared with respect to exactly the same observational data (taking different data leads to different values for these criteria). Sometimes it is desirable to use criteria that can be compared between studies or data sets like e.g. an R^2_{adj} or a Nash efficiency (see comment by Reviewer 3).

Breslow, N.E. and Clayton, D.G.: Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* 88: 9-25, 1993

Stone M.: Asymptotics for and against cross-validation. *Biometrika* 64 (1): 29-35, 1977
Wolfinger R. and O'Connell M. : Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation* 48: 233-243, 1993.

C) Importantly, I do not think that including the AIC and other statistics will change the results as the "best" empirical model includes only soil temperature (arguably the simplest empirical model and the one with the highest parsimony). Thus, if the simplest model is used to describe this system, then it is difficult to justify such a complex model selection to interpret such a simple system. This argument relates back to my main comment about the applicability of empirical models to describe ecosystem processes. Why is needed to go through the workflow provided if temperature alone explains the best most of the variability observed? What does this means in terms of ecosystem processes? This comment is debatable and I encourage the authors to discuss it in a

C8333

revised version of this manuscript.

This study aims to calculate annual C loss from soil respiration from three ages of heathland. To do this, a work flow method to select a soil respiration model was presented. This is a process which is generally not being adopted in soil respiration research (or at least, it is not reported in the literature). In our view, the field would benefit from such an approach in two ways: to make studies more reproducible and to reduce publication bias (see Dieleman et al. 2011). Publication bias does in this case not only refer to the tendency to publish cases where elevated respiration is demonstrated, but also results which show the influence of certain environmental variables. This argument was not yet included in the manuscript. But in the revised manuscript we have included it.

The model which resulted from the selection process used soil temperature alone, which is arguably the simplest model. The reviewer questioned the decision to use a work flow method, when the 'best' model was such a simple model. This outcome was not known prior to the start of the study and it certainly was not expected, based on our hypotheses. In addition, the outcome of a simple model does not mean that the underlying reasons are simple. The model selection process highlighted that none of the seven additional variables that were measured (air temperature, soil moisture, biomass, photosynthesis, microbial biomass, root biomass, or PAR) improved the model fit (as alternatives or in addition to soil temperature). As soil moisture, biomass, photosynthesis, microbial biomass, root biomass and PAR were all not highly correlated to soil temperature (air temperature was not used in combination with soil temperature but as alternative), the fact that these variables did not explain additional soil respiration variance, leaves us with one of these potential explanations: a) temperature is indeed the only important driver in this system (and at this scale), b) there are additional drivers, however these are not represented by the variables included in this study (and we will have to continue searching for different explanations). Considering the structure in the model-residuals, we think there are indeed important drivers in ad-

C8334

dition to temperature, and we hope to find these by observing different variables when conducting new experiments at the research site in the future.

The interpretation of our model results as well as possibilities for observing different variables in follow-up research are extensively discussed in section 4.3 of our revised manuscript.

Dieleman, W.I.J., and Janssens, I.A.: Can publication bias affect ecological research? A case study on soil respiration under elevated CO₂. *New Phytologist*, 190 (3): 517-521, 2011 <http://dx.doi.org/10.1111/j.1469-8137.2010.03499.x>

D) The manuscript is quite long in its present form. The main objective is to propose a model selection protocol to find the best empirical model that fit the observed variables (not the ecosystem processes) and to interpolate in time the autotrophic and heterotrophic components of soil respiration. However, the reader can be easily distracted with discussions about extensive justification on the experimental design (arguably pseudoreplication), discussion of data not presented clearly (some will be presented in further studies (e.g., page 16250 line 6; page 16276 line 19), and discussion of global change implications (page 16276 and abstract), which some may be moved to supplementary information. Maybe a way to focus and shorten the paper is to better discuss model selection procedures and its implications/limitations to our understanding of ecosystem processes; which is the title of this paper.

We agree that the manuscript is long and made a number of number of changes in the revised manuscript to shorten it (e.g. removal of C balance comment from abstract, incorporating RMSE results into a new Table 4, condensing section 3.5, removing some of the discussion of treatment effects, removing the diurnal variation results and associated discussion). However, we do not wish to separate the modeling discussion from the processes occurring in the heathland or the practical outcomes of a modeling process. The aim of this paper was to take field observation data of soil respiration and predict annual C losses. This is a goal of many soil respiration papers, when consider-

C8335

ing C feedback loops and global climate change. However, not many of these papers use a model selection process to generate this number. How do they know that the variables they are measuring contribute to the data explanation? How do they know that the model is better or worse than other potential models? This paper aims to incorporate an outcome (C cycles) that is important to biogeochemists and soil scientists and bring it into closer contact with the world of modeling by using a model selection process.

E) The results of the modeling selection does not support the hypothesis presented in page 16244 lines 10-14). I agree on this hypothesis, but it is not clear how the authors interpret that a multi-level modeling approach provides different results than those expected. Is this site specific? How these results from empirical models are interpreted in terms of ecosystem processes? A more clear discussion about this hypothesis is needed to link it with the main objective and title of this paper. Furthermore, where is the clear answer to the question posed in the title? The discussion is somewhere in the manuscript but it is not clear and easy to understand in the current form of the manuscript.

The main aim of this paper was to take field soil respiration data and, using of a model chosen through a modeling selection procedure, to calculate annual soil C losses (page 16243 lines 25-29). Based on the known relationship between soil microbial decomposition, soil temperature and soil moisture, it was hypothesized that soil temperature and soil moisture would be significant variables for the RH model. In addition, based on the belowground allocation of plant metabolites contributing to root processes, it was hypothesized that soil temperature, soil moisture and a measure of plant activity would contribute significantly to the RS models for all three communities. These sentences have been clarified to link the microbial and root processes to the hypothesized significant variables (last paragraph of the introduction section).

During the model selection process, it became clear that the hypotheses were not supported: that is, soil moisture was not significant in the 'best' RH models and nei-

C8336

ther were soil moisture nor plant activity variables significant in the 'best' RS model. However, the model selection process did indicate that an important covarying factor was lacking from the variables that had been measured. This indicated that another ecosystem variable besides plant photosynthesis, plant biomass, PAR, root biomass, or microbial biomass was important when trying to explain soil respiration in this system. In the absence of this important covarying factor, soil temperature sufficed to explain the seasonal variation observed in the ecosystem. This has been included in a Conclusion section at the end of the paper, to clarify the link between the aim, the hypotheses, the model selection process and the outcomes.

Comments in detail (we numbered the comments, to make cross-referencing easier)

1) Introduction. Which is the difference between "components" of soil respiration and "compartments" of soil respiration? These definitions are used interchangeably across the manuscript and needs to be clarified or unified. We used the two terms as synonyms. By slightly rewording the text, we removed both 'compartments' and 'components' from the text.

2) Line 3 page 16240 – it is unclear the statement "modeling tools which generally include model variables. . ." Please define the difference between modeling tools, model variables, and empirical models in the main text. The reviewer requested clarification on the definition of modeling tools and model variables (Line 3 page 16240). This has been addressed by removing the reference to modeling tools, as the same underlying message can be equally conveyed through the use of the words models or modeling (depending on the context) (also see comment 4). As explained under remark A, we removed the term empirical from the text (and a discussion on the degree at which the models identified in our study represent ecosystem processes adequately is included in section 4.3.

3) Page 16241 line 24. Define which are the "underlying drivers". This is ambiguous. The major drivers of total soil respiration have been defined by a number of studies to include abiotic factors, such as temperature and soil moisture and include biotic factors,

C8337

such as gross primary productivity (Bahn et al., 2010a; Davidson and Janssens, 2006; Trumbore, 2006). The text on page 16241 line 24 has been amended to clarify the link between this paragraph and the following paragraph.

4) Page 16242 line 19. Define what "modeling tools" mean. This is ambiguous. The use of the term "modeling tools" has been removed to avoid ambiguity and the term 'models' or 'modeling' (as appropriate) has been implemented throughout the document (Page 16242 Line 19 of the original document) (also see comment 2).

5) Page 16242 line 28. Which is the difference between a "process-based" model and "more empirical" model? This is definition should be clear in this manuscript as the models used are empirical but used to interpret ecosystem processes. Could you give an example of a process-based model and a "more empirical" one for soil respiration? In fact, we would like to avoid the discussion about empirical versus process-based modeling in the manuscript. We think it would make the manuscript longer and less focused. But we do want to share our view and discuss it in our comments. Apart from this location (p. 16242, line 28), there is one other location where we use the term 'empirical' (p 16254 line 4 - to indicate that we also place the Selsted model in this model category); and in hindsight the choice for these terms is unfortunate. In the revised manuscript we have therefore replaced the terminology (the sentence on page 16242 now reads: 'The degree to which soil respiration models could be modeled with a priori chosen parameter values or required calibration did depend on both the spatio-temporal scale at which the models were to be applied and the available environmental data (Keenan et al., 2012). Hence there is no need to define the terms 'empirical' and 'process based' in the manuscript anymore. However, still a discussion on the way by which the models identified in our study can be used to learn more about ecosystem processes is included in our manuscript (as requested in remark A, see also our reply to remark A).

6) Page 16243- Describe what "multi-level modeling" approach is, how it is being used and how are you applying it in the study. This is somehow in the text but I suggest

C8338

to be clearly explained in the introduction. How a multi-level modeling is being used for longitudinal and clustered in space data of soil respiration? How this relates to the current study? Why is better than other approaches? A multi-level model describes the relation between one or more response variables measured at one scale as affected by by predictors measured at that same scale and by predictors that are working at a different spatial or temporal scale (other terms that are used to denote the same type of model are mixed effect, random effect and hierarchical model). Multi-level modeling is a complex and quickly evolving field, about which a considerable literature exists. We believe that the paper by Qian et al. (2010), as cited in our manuscript, provides a good entry point for multi-level models in an ecological context.

In response to the request by the reviewer, we have added a sentence in the introduction where we explain the gist of using a multi-level analysis: 'Hereby the measurements which are collected for the same observation unit are explicitly assumed to be dependent, which leads to a more realistic estimate of the effective degrees of freedom (and consequently more realistic confidence bounds) than when assuming independent observations.'

7) Page 16244- How stands be of "similar age", but at the same time have a younger and older stand? I think it is better to be clear about the ages and on the interpretation of young and older. Then in line 25 of the same page the authors say that the stands are of "different ages". This is confusing. . . maybe just use the real ages of the stand to avoid ambiguity in the interpretation of the results. The introduction has been amended on Page 16244 to clarify the sentence regarding stand age. Within each community age, the Calluna plants were of similar age, but each community had been last cut at a different time.

8) Page 16245. True replication within natural ecosystems is extremely difficult, thus the interpretation of the results and how we treat statistical analyses has to be carefully done. This is of critical importance when extrapolating results in time and space thus I suggest to be very careful when interpreting the results outside the main objective. For

C8339

example I suggest removing section 4.4. This is not part of your main objective and has assumptions that are not clearly discussed. We agree with the reviewer that true replication is difficult in long term field trials. In fact, we thought to have emphasized this aspect by explicitly mentioning on page 1645 that in our research 'a quasi-experimental design was used, in which groups were selected upon which the variables were tested but where randomization and replication processes were not possible (Campbell and Stanley, 1966).' And we are well aware that we should therefore be careful to extrapolate our findings. However, comparison of the results from our experiment with those found elsewhere (first part of section 4.4) were not meant to be an extrapolation, but instead an attempt to place our results in a useful context and make it accessible for the wider community. In the second part of section 4.4 we emphasize that the statements with respect to the effect of vegetation age on C sequestration are hypothetical ('Based on the presented data and a preliminary assessment of other fluxes within the system, it is hypothesized that..'). In the revised manuscript, we have therefore not removed section 4.4.

9) Page 16245 lines 4-10 this is repetitive from introduction. The repetition associated with describing the management practices of the heathland (16245 lines 4-10) has been removed to improve readability of the document.

10) Page 16245 lines 16 what do you mean by nutrient-poor. Explain nutrient-poor with respect to which threshold. Also give soil texture data for all your sites. This is arguably an important parameter for estimating soil respiration in space but it is not included in any of your empirical models. The reviewer commented that soil texture is arguably an important variable for estimating soil respiration. This is an interesting question and indeed, a number of soil respiration models include this as a variable, particularly where a number of geographic locations are being compared. To enable better comparison with other studies, soil texture is now included in Table 1 (bottom row). However, it was not included as a model variable within this study because soil texture shows very little variation between plots (which are only meters apart, see Figure 1) and would not have

C8340

improved model fit. In the introduction, the nutrient status of the site was referred to as being 'poor'. The nutrient status has been defined in other studies on this site, and therefore a reference has now been included in the introduction to guide the reader to this information (Van Meeteren et al, 2008).

van Meeteren, M.J.M., Tietema, A., van Loon, E.E. and Verstraten, J.M.: Microbial dynamics and litter decomposition under a changed climate in a Dutch heathland, *Applied Soil Ecology*, 38, 119-127, 2008.

11) Page 16247 line 16. 10 cm of buffer zone from a trench seems to be somehow short. Do you expect any "boundary effects", how realistic were these measurements considering potential effects (if any). The placement of the soil respiration collars inside the plots was described in the methodology as being a minimum of 10cm from the plot boundary, although the majority of the measurement locations were actually at least a 20cm buffer zone. The soil inside the plot was deliberately not disturbed during the digging of the trench, although it is possible that as roots were severed there would have been some minor disturbance in the soil. There was a four month data exclusion period at the beginning on the trial to remove the effect of any of these minor soil disturbances. A 10 cm buffer was therefore considered adequate.

12) Page 16247 lines 24-27. There is no guarantee that the plots were similar before the trenching as no measurements were done before trenching (i.e., initial conditions of the experiment). According to the authors measurements started 3 days after trenching. . .please discuss the implication that no pre-treatment spatial heterogeneity was evaluated. How this could affect the interpretation of the results? The reviewer commented that there was no guarantee that the plots were similar prior to trenching as spatial heterogeneity was not evaluated. However, pre-trenching assessments were undertaken on the three community ages and the overall interpretation of the RH results is not considered to be affected. These pre-trenching assessments included an assessment of community carbon stocks prior to trenching (Page 16244 Line 8; and Page 266 line 10) to assess differences between community ages. There were no

C8341

significant differences in the soil C stock (microbial food source) to 10cm soil depth between the ages of vegetation. Additionally, soil respiration measurements were obtained prior to trenching although these are not reported or discussed in this paper. It was not considered necessary to include these results in the paper as the trial layout was described to be in a nested design (Page 16255 and Page 16291 Figure 1 caption) which therefore has a control plot (total soil respiration) for every trenched plot (heterotrophic respiration). This trial design allowed for analyses of variance to be undertaken on the control plots and the results are deemed to represent the ambient conditions of the heathland (ie represent the variability of the pre-trenched plots). Given that the paper is already of considerable length and there was a reported absence of C stock difference, this discussion was not included in the paper.

13) Page 16248 lines 1-7. The study seems a little bit rushed. The authors let the trenched plots to "stabilize" only for 4 months and only observations after 21 September were included in the analysis. Is four months "enough" for decomposition of all that organic material to properly interpret Rh? Probably not and it is probably the reason why the authors did not find differences in heterotopic respiration across the sites. This potentially rapid organic matter decomposition must be clearly discussed in light of interpretation of Rh in time and space. The authors respectfully disagree with the reviewers comment that the study seems a little rushed. In fact, our own perception is quite contrary. A period of months of thoughtful discussion, preparation and planning preceded the fieldwork, involving several experts from the INCREASE project (EU project - <http://www.increase-infrastructure.eu/>); and since the onset of the fieldwork, data analysis and discussion on the (then preliminary) results took place. We'd like to highlight two details, illustrating the level of planning that went into this research.

Firstly, this study excluded four months of data where results indicated that soil disturbance provided a flush of CO₂ that was not associated with RH (ie a treatment effect). Some studies have not excluded data in this initial period although there is recognition that this process may be occurring (Moyano 2007). Other studies indicate

C8342

this period can take 2 days {{1638 Edwards, 1991;}}, several months {{1581 Comstedt 2011;}} or a year {{1637 Silver 2005;}}. On our trial, the length of the exclusion period was based on analysis of the trenched plots to determine when CO₂ variability decreased and the time when there was no statistical difference between the plots. Additionally root biomass was examined in April 2012 to assess root decomposition (as described on page 16248 lines 1 to 5 in the original manuscript – this is unchanged in the revised manuscript).

Secondly, we compared the Trenched plots respiration (RH) with the RH measurements obtained from an adjacent long term trial, in which vegetation was removed from three plots in September 2009 and measured since April 2011. There were no statistically significant differences between the areas cut in 2009 and the Trenched plots for the time period September 2011 – August 2012. This indicated that the Trenched results were indicative of RH and the exclusion period was sufficient to remove the CO₂ flush data associated with the trenching method. This discussion was not included in the paper, considering the length of the paper.

Given these two aspects, the Trenched plot results are considered to be RH and allow for interpretation of soil microbial respiration from a one year period.

14) Page 16249. Photosynthesis measurements are not clearly discussed and the authors state that they are evaluated in more detail in another publication. Why is this not discussed and which publication is this? This paper focuses on soil respiration and using a model selection process (with a number of measured variables) to calculate annual C losses from the soil. Plant activity variables are commonly included in soil respiration models but are not commonly assessed for parameter significance. This paper is long already and additional discussion on the plant C fluxes is outside the scope of this paper. It is planned that an additional paper will be written that uses a similar modelling process to the one described here, to investigate plant and ecosystem fluxes. However, the reviewer makes a good point that this unspecified paper should not be referred to in the current paper. This sentence has been amended to remove this

C8343

comment and following reviewer #1 comments, much of this section has been moved to a supplementary section S1.

15) Section 2.7.2. Please include the validation of the soil moisture model. This is important for the interpretation of the results and for validation of the empirical models that used this input variable. We agree with the reviewer. In the revised manuscript, performance by the soil moisture model is shown graphically in the supplement where the soil moisture model is described (S2). In the text, the numerical performance is mentioned in section 2.7.2.

16) Section 2.7.4. As discussed earlier: The authors used the RMSE as a statistic for model selection (Figure 8) but generally for model evaluation there are several parameters that are important to consider. These are the RMSE, the correlation coefficient and the standard deviation that could be represented in a table or in a diagram (Taylor, 2001). Furthermore, models that have different number of parameters are generally evaluated via the maximum likelihood approach known as the Akaike's Information Criterion (AIC) (Burnham and Anderson, 2002). I strongly suggest including this statistic when evaluating models that have different numbers of parameters. See our answer under general comment B. We agree with the reviewer that the AIC is a useful criterion, but have decided not to report it in this study because it cannot be calculated for some models.

17) Page 16256 line 15. Please define "parameters reasonableness". Which is the basis of the thresholds that were selected? Parameter reasonableness was determined based on known respiration rates at R₀ (measurements in the field) and realistic values, based on first principles (mainly variables forced to be positive for physical reasons, e.g. the constant 'a' must be greater than zero in the GLMM model to describe the soil moisture (M) relationship. This sentence has been amended to make it clearer that this is the definition for parameter reasonableness. (see also comment 20)

18) Results section. The authors used a series of statistical analyses to test for the

C8344

observational data (section 2.7.1). I encourage including the analysis used, the statistic (e.g., F, t) and the p-value when referring to statistical differences. So far it is only reported the p value but it is hard to track which test was used to evaluate the differences. This will help the readers better understand the text. In the revised manuscript we included the F-values in the results section, as requested.

19) Page 16258 lines 23-27 (and associated discussion). I feel that the diurnal variation does not add much to the manuscript and is quite weak as only few manual measurements were performed. To simplify the manuscript I suggest removing these results and the associated discussion. The diurnal variation results, Figure 5 and associated discussion points have been removed (Page 16258 lines 23-27). These results are interesting from the perspective of comparing actual diurnal patterns with modeled output patterns; however, Reviewer #2 is correct that for the sake of shortening the paper, this can be removed.

20) Page 16261 lines 13-14 and 27-8. Please explain how the authors arrive to the conclusion that the parameters of the models are considered reasonable. Which were the assumptions or thresholds? Parameter reasonableness is defined in the methodology being parameters which were feasible according to literature values and experience. The reasonableness of parameters were defined for basal respiration rate (R_0 0 to < 0.5 (for the Selsted and GLMM models), $a > 0$ and $c > 0$ (GLMM models) or $a < 0$ and $c < 0$ (LMM models) (this was mentioned in lines 15-20 on p. 16256 of the original manuscript, and still is in the revised manuscript).

21) Section 3.7. There are multiple ambiguities in this paragraph. What do “was so similar”, “were so small” and “the most attractive Rs model choice” mean? Please be clear and avoid using terms that are meaningless such as those described in section 3.7. The words have been changed to clarify the meanings and indicate that model RMSE, model parameter values and predictions were used.

22) Page 16266 lines 10-14 and page 16269 lines 20-23. These lines seem contradic-

C8345

tory. First the authors say that there were no differences in patterns of soil moisture and then they state that there were differences. Please clarify. The soil moisture patterns were similar between the trenched plots, in which heterotrophic respiration was measured (Page 16266 lines 10-14) and this was regardless of community age. There was however, a difference observed in the soil moisture patterns when comparing between Trenched and Untrenched plots (page 16269 lines 20-23) and this is not in contradiction of the previous statement. This difference is hypothesized to be a result of root removal in the trenched plots, as discussed in Section 4.2.

23) Remove section 4.4 Section 4.4 has been modified to emphasize the speculative nature of the hypothesis. However, part of the aim of this study was to determine the annual soil C loss from the site. Therefore, the authors believe that it is an important part of the study, and one in which should remain, although in a modified format.

24) Table 1. Include soil texture and soil bulk density. Soil texture is included in Table 1— and discussed under comment 10 above.

25) Figure 5. Consider removing it. (Discussed above)

26) Figure 6. Include validation of soil moisture data. Validation of soil moisture data is included, however not in Figure 6, this is further explained under comment 15

27) Figure 7. Please indicate what the error bars mean and the small letters. One can imagine what they mean but it is always advisable to be clear. The error bars are standard errors of the mean (as defined in section 2.7.1). For clarity this bar and letter information has been included in the Figure caption.

28) Figure 8. Consider including AIC. See our answer to general comment B.

Please also note the supplement to this comment:

<http://www.biogeosciences-discuss.net/9/C8323/2013/bgd-9-C8323-2013-supplement.zip>

C8346

C8347