**Response to 'gmd-2014-179 comments', Anonymous Referee #1.**

***Referee #1:*** *Abstract is the place to clearly concisely show what you have done, why do you think the work is important. What are your results and why are they significant? The abstract is somehow not clear and concise enough to me; may be also confuse the potential readers. For example, the first paragraph basically says: the Bayesian approach is widely used to calibrate forest model, which has already been well accepted (hundreds of published studies). No need to spend entire paragraph to clarify this point. Please consider reconstruct your abstract to be as concise as possible.*

**Authors:** We thank the reviewer for raising this point. We agree the abstract is not concise enough and we will rewrite it in the revised version of the manuscript, including the new results and conclusions we will find as a consequence of the several changes suggested in the review process, and considering the reviewer's comments including but not limiting to the size of the abstract, main novelties of this work and their impact.

***Referee #1:*** *The presentation is not complete. For example, the author started the introduction with the definition of GPP and followed by the observation of GPP, GPP modeling and model calibration. Lots of important information are missing, including but not limiting to: (1) Besides Eddy Covariance GPP data, MODIS-GPP is another famous GPP product; (2) Eddy Covariance network only measure NEP, GPP is derived from their model; (3) There are several other ways to model GPP besides LUE model (e.g., Farquhar 1980 type model). (4) Dislike LUE model, in Farquhar model GPP associated parameters have physical meaning, thus they are relatively easy to infer from observations.*

**Authors:** We agree with the points raised by the reviewer, and we will adjust the introduction consequently according to these points. We are also well aware that there are several ways to simulate GPP, and Farquhar (1980) is one of the most commonly found in forest modelling. However it is not free of disadvantages (Yin et al., 2004; Van Oijen et al., 2004): the Farquhar model parameters have no physical meaning at the canopy scale since they are chloroplast parameters with at best some validity up to the leaf level, but not more. Its parameters are also not easy to infer: A-Ci measurements with leaves sampled from all across the canopy are needed, with young leaves having much higher values of $Vc_{max}$ and $J_{max}$ than old ones. We will include a more

complete presentation of GPP in the revised introduction, including advantages and disadvantages of the most common methods to estimate it.

***Referee #1:*** *In the second part of introduction, the author presented the idea of Bayesian Calibration. It worth to mention that Bayesian calibration is not necessary rely on MCMC method. Bayesian approach relying on adjoint method is also an effective calibration method (Zhu 2014). Also it worth to mention other important type of ecosystem model calibration method: Kalman filter (Gao 2011).*
*And the author need to justify the reason why they decide to use MCMC methods, given that other two types of calibration methods (adjoint method and kalman filter) could be much more efficient (e.g., adjoint method is a local optimization method, while this study needs a global optimization method? I believe the authors have their own reasons).*

**Authors:** We will add a paragraph in the introduction of the revised manuscript, referring to the two techniques mentioned by the reviewer and justifying our choice to apply a MCMC-based method. These reasons will read as follows: "The data assimilation techniques mentioned above are special cases of Bayesian calibration (Wikle & Berliner 2007), where a prior probability distribution for parameters is specified and updated using Bayes Theorem. However, in contrast to our MCMC approach, the old data assimilation methods – though computationally efficient - require assumptions of linearity and Gaussian distributions that are restrictive and inappropriate in the case of the highly nonlinear models that we study here. Therefore such methods are common in state estimation of computationally demanding models such as GCMs, but they are not common in parameter estimation of ecosystem models. Our MCMC method allows for any type of prior and posterior distribution, including asymmetric and multimodal ones, in contrast to the other methods. Moreover, the sample from the posterior distribution generated by MCMC represents the full posterior probability distribution, in contrast to the adjoint method which only provides an estimate of the mode, and uncertainties can only be assessed fully with such global methods, not local ones."

***Referee #1:*** *The purpose of model calibration is to improve the posterior model predictability. This study only presented the calibrations, but miss the posterior model evaluation.*

**Authors:** We carried out a posterior model evaluation for the approaches that resulted in proper convergence. We decided not to include this result in the paper since its main focus is on Bayesian calibration. The model results were insensitive to the algorithm used or to the procedure applied.

We will include the results of model evaluation in the revised version of the manuscript, focusing on the differences (or similarities) between the different calibration procedures.

***Referee #1:*** *One common approach is that: the model should be first calibrated at one EC tower site and then apply to another site that has the same plant function type. The cross-site evaluation is necessary to ensure the efficacy of model calibration. I suggest that the author should apply the posterior model parameters derived by different calibration methods to another tower site, in order to fairly compare the goodness of different calibration methods.*

**Authors:** The reviewer is correct in describing the common approach to forest modelling. We did indeed implement this approach in a recently published paper (Bagnara et al. 2014). The latter work focuses on the same model analysed here and tests it on several EC sites. The focus of the present study is on evidencing potential issues in calibrating a simple but highly non-linear model, characterized by a commonly applied mathematical structure. We will refer to Bagnara et al. (2014) in the Discussion of the revised manuscript, focusing on the impact of our results to their findings, but we think a model validation on different EC sites is beyond the scope of this paper.

***Referee #1:*** *Broad impact of this study is not well discussed. It is not clear to me how their findings interest our molding community and facilitate future studies in terms of forest model calibration.*

**Authors:** We thank the reviewer for raising this very important point, and we will emphasize the impact of our findings in the revised manuscript. We think our results are important because they focus on issues that have never been discussed before in the field of forest modelling: there are no studies on the difficulties in calibrating this kind of models, which are widely applied to forest research and management. Several well accepted studies and models could be affected by this kind of issues, and we are stressing the need of a more careful approach to calibration to solve potential problems which have been rarely mentioned before.

***Referee #1:*** *Another issue worth discussing is that the parameter calibration could only reduce model parameter uncertainty, however, is not able to constrain model structure uncertainty. There are two LUE models with different model structure used in this study, which might provide insight into the uncertainty in model structure.*

**Authors:** We will apply the procedure described in Van Oijen et al. (2013), based on the ratio of posterior model probabilities, to assess the importance of model structure uncertainty. The results and discussion about this interesting point will be included in the revised manuscript.

***Referee #1: Minor comments:***
*Page 6998 Line 2: Remove in very different forest all over the world. Do you mean different forest functional type? Line 4: "easy to use" is not a rigid scientific term. Define it more appropriately. Maybe "pragmatic"? Line 13: what does "user-friendly" mean? Line 19: this sentence needs to be rephrased. Line 22: calibration did not*

**Authors:** As stated in the previous point, the abstract will be entirely rewritten according to the reviewer's comments and to the changes to the revised manuscript. These comments will be taken into account while doing so, replacing " *very different forest all over the world* " with "different biomes and PFTs all over the world" and defining "easy to use" and "user-friendly" as in Landsberg & Waring (1997), that is models available to a broader public then models with strictly research and academic purposes, as they are based on a few equations, few parameters, and they do not require high computational power nor lot of data to be run.

***Referee #1:*** *Page 6999 Line 2: terrestrial ecosystem carbon balance*
*Line 11. Cite paper here Reichstein 2005*

**Authors:** We will include these last two comments in the revised manuscript.

***Referee #1:*** *Page 7000 Line 13: sentence needs to be rephrased.*

**Authors:** The sentence will be rephrased as: "There are several LUE-based models in the existing literature: a few examples are C-Fix (Veroustraete et al., 1994), 3PG (Landsberg and Waring, 1997), Prelued (Mäkelä et al., 2008a), and the model described in Horn and Schulz (2011b)."

***Referee #1:*** *Line 18: compared with*
*Line 21 daily time step, based on*
*Page 7001 Line 1: The Bayesian model calibration approach*

**Authors:** We will include these last three comments in the revised manuscript.

***Referee #1:*** *Line 15: The efficiency of the MCMC technique highly depends on the model structure. Is it true in general? How about other factors?*

**Authors:** The dependency of the MCMC efficiency on the model structure has been proven, among others, by Gilks & Roberts (1996) and Browne et al. (2009). The effect of model structure on efficiency of Bayesian calibration is the main point of the paper, as we show that commonly used models with a simple but highly nonlinear structure can be hard to calibrate. Other important factors are the uninformative prior distributions (Hartig et al., 2012) and the heavy use of empirical parameters in the model formulation, as discussed in the submitted version of the manuscript. We will discuss on the importance of these factors more in detail and include the previous references in the revised version of the manuscript.

***Referee #1:*** *Page 7002 Line 26: Why only use one year data? Lavarone site has multiple-year data (2000-2006). Perhaps, it is a good practice to use part of the data as calibration dataset, and use the rest for model validation.*

**Authors:** We believe that one year of data is more than enough to successfully calibrate a 6-parameters model. Therefore we saw no need to run a longer calibration on several years that would be heavier in terms of computational power required. We did use a different part of the available dataset to evaluate the model performances, and this results will be added to the revised manuscript as stated in the previous comments.

***Referee #1:*** *Page 7005 Line 4: Why do you chose 0.3*GPP as a upper bound of GPP data uncertainty? Any reference or reasons?*

**Authors:** Very few examples can be found in the literature of uncertainty estimates of daily GPP. Moreover, these are not consistent across studies: Mo et al. (2008) set daily uncertainties on GPP as 15% of its value, while Duursma et al. (2009) estimated them to be 5% of GPP. We set them to 30% of GPP as done by Williams et al. (2005), as a conservative estimate for calibration purposes, also to be sure that the information content of the data was not overestimated.

***Referee #1:*** *Page 7011 Line 16: "multiplicative structure of Prelued was probably the main factor responsible for the difficulty in the calibration." Is it true?*

**Authors:** The reviewer raises a good point. We don't have enough evidence to blame the mathematical structure of Prelued as the only or main responsible for the difficulties in calibration, but we consider also several other factors, like poor a-priori knowledge on the parameter values and their empirical nature. It is crucial that these points are clear to the reader, therefore we will emphasize them in the Discussion section of the revised manuscript

***Referee #1:*** *(1) First of all, photosynthesis (GPP) is a very complex biology process, a certain level of model complexity is needed. The difficulty of model calibration might be simply due to the fact that LUE model is too simple (model structure) to capture the GPP response to environmental changes.*

**Authors:** If that were true, the model would have difficulties in reproducing the data, even after calibration, on the same site and period of simulation, which is not the case. Including the model evaluation will show that, even for algorithms that did not reach proper convergence, the model results are pretty good. The model does not have difficulties in reproducing GPP, so its structure does not seem to be inadequate for that purpose.

***Referee #1:*** *(2) Multiplicative structure is common in other GPP models (such as abovementioned Farquhar model), there is no evidence that the multiplicative structure hinders model calibrations.*

**Authors:** That relates to the point already mentioned above: the fact that there is no evidence of that does not mean it does not exist, only that it has not been studied so far. Such behavior from a simple model was very surprising as we did not expect such difficulties in calibration, but it gave us the possibility to study an issue that, to our knowledge, has never been studied before in the field of forest modelling. That is the main novelty of this work. Moreover, as the model by H&S had the same problems, this work raises an important question: can we really trust the well-accepted models with similar structure developed so far, or are they all affected by the same calibration issues we described? We will add and highlight this concept in the revised manuscript, as it is one of the main central points of the paper and it should be emphasized to be brought to the reader's attention.

***Referee #1:*** *Page 7013 line 14: Any suggestion of future development of LUE model? At least, based on your findings, the LUE model needs a better mathematical structure? Which structure should it be?*

**Authors:** Our recommendation, that will be included in the Conclusions of the revised paper, is that a more complicated structure should be applied to LUE-models. For example, including Prelued as a module in a more structured model could reduce the difficult in calibration (like PRELES, successor of Prelued ), also to avoid estimating only one variable very complicated as GPP.
It should also be pointed out that this kind of models does not allow to compare model estimates against actual data: GPP is not measured, it is derived from NEP. So NEP should be the model output against which the calibration should be performed, and it should be included in LUE models via combination with a respiration model.
Another important point relates to the empirical nature of the parameters: when possible, a large use of parameters with no physical or physiological meaning should be avoided, in order to rely on the physiological basis of GPP as much as possible.

**References:**


Yin, X., Van Oijen, M., & Schapendonk, A. H. C. M. (2004). Extension of a biochemical model for the generalized stoichiometry of electron transport limited C3 photosynthesis. *Plant, Cell & Environment*, 27(10), 1211-1222.

Van Oijen, M., Dreccer, M. F., Firsching, K. H., & Schnieders, B. J. (2004). Simple equations for dynamic models of the effects of CO2 and O3 on light-use efficiency and growth of crops. *Ecological Modelling*, 179(1), 39-60.

Wikle, C. K., & Berliner, L. M. (2007). A Bayesian tutorial for data assimilation. *Physica D: Nonlinear Phenomena*, 230(1), 1-16.

Bagnara, M., Sottocornola, M., Cescatti, A., Minerbi, S., Montagnani, L., Gianelle, D., & Magnani, F. (2014). Bayesian optimization of a light use efficiency model for the estimation of daily gross primary productivity in a range of Italian forest ecosystems. *Ecological Modelling*, in press.

Van Oijen, M., Reyer, C., Bohn, F. J., Cameron, D. R., Deckmyn, G., Flechsig, M., Härkönene S., Hartig F., Huth A., Kiviste A., Lasch P., Mäkelä A., Mette T., Minunno F. & Rammer, W. (2013). Bayesian calibration, comparison and averaging of six forest models, using data from Scots pine stands across Europe. *Forest Ecology and Management*, 289, 255-268.

Landsberg, J. J., & Waring, R. H. (1997). A generalised model of forest productivity using simplified concepts of radiation-use efficiency, carbon balance and partitioning. *Forest Ecology and Management*, 95(3), 209-228.

Gilks, W. & Roberts, G. (1996). "Strategies for Improving MCMC." In W Gilks, S Richardson, D Spiegelhalter (eds.), Markov Chain Monte Carlo in Practice, p. 89-114. Chapman & Hall, Boca Raton, FL.

Browne, W. J., Steele, F., Golalizadeh, M., & Green, M. J. (2009). The use of simple reparameterizations to improve the efficiency of Markov chain Monte Carlo estimation for multilevel models with applications to discrete time survival models. *Journal of the Royal Statistical Society. Series A, (Statistics in Society)*, 172(3), 579–598. doi:10.1111/j.1467-985X.2009.00586.x

Hartig, F., Dyke, J., Hickler, T., Higgins, S. I., O'Hara, R. B., Scheiter, S., & Huth, A. (2012). Connecting dynamic vegetation models to data–an inverse perspective. *Journal of Biogeography*, 39(12), 2240-2252.

Mo, X., Chen, J.M., Ju, W., Black, T.A., 2008. Optimization of ecosystem model parameters through assimilating eddy-covariance flux data with an ensemble Kalman filter. *Ecol. Modell.* 217, 157–173. doi:http://dx.doi.org/10.1016/j.ecolmodel.2008.06.021.

Duursma, R., Kolari, P., Permki, M., Pulkkinen, M., Mäkelä, A., Nikinmaa, E., Hari, P., Aurela, M., Berbigier, P., Bernhofer, C., Grünwald, T., Loustau, D., Mölder, M.,Verbeeck, H., Vesala, T., 2009. Contributions of climate, leaf area index and leaf physiology to variation in gross primary production of six coniferous forests across Europe: a model-based analysis. *Tree Physiol*., 29(5), 21-639. doi:http://dx.doi.org/10.1093/treephys/tpp010.

Williams, M., Schwarz, P. A., Law, B. E., Irvine, J., & Kurpius, M. R. (2005). An improved analysis of forest carbon dynamics using data assimilation. *Global Change Biology*, 11(1), 89-105.

**Response to 'Comments to gmd-2014-179', Anonymous Referee #2.**

***Referee #2:*** *This paper compares 3 MCMC methods for 2 simple GPP models, to examine the convergence of the posterior parameter distribution. The conclusion that simple models' advantage is limited due to the difficulty in parameter tuning even for new MCMC methods is important, and could be considered for publication on GMD.*
*However, the current version has some significant problems, which should be fixed before going to the next stage. In particular, my main concern is generality of the results.*
*The experiments were carried out only for one case. Discussion on the application of the results for other sites and for other type of forests is needed.*

**Authors:** In a recent paper we focused on the same model and tested it on several EC sites, implementing a DEMC algorithm and a very high number of iterations (Bagnara et al., 2014). The focus of the present study is on evidencing potential issues in calibrating a simple but highly non-linear model, characterized by a commonly applied mathematical structure, using one EC site as a case study. We will refer to Bagnara et al. (2014) in the Discussion of the revised manuscript,

focusing on the impact of our results to their findings, but we think a model validation on different EC sites is beyond the scope of this paper.

***Referee #2:****It is also needed to discuss the influence of GPP uncertainty, i.e. effect of changing the term y_j in eq. (7).*

**Authors:** A paragraph describing the importance of data uncertainties on the calibration procedure will be added in the revised version of the manuscript. This is: " The uncertainties around the data are of primary importance for the effectiveness of the calibration. If the data are uncertain, i.e. become less informative, then the likelihood distribution in parameter space becomes more uniform. As a consequence, every proposed new candidate parameter vector will have similar likelihood as the current parameter vector, so the likelihood ratio will always be very close to 1 and the candidate vector will always be accepted unless its prior probability is low. This very high acceptance rate will slow down the effective exploration of parameter space as the random walk loses direction, slowing down the identification of the convergence region as a direct consequence. On the other hand, if data uncertainties are too small, i.e. if the data are considered too informative, the likelihood ratio will be always close to 0, causing a very low acceptance rate. This would cause the MCMC to move very slowly through parameter space, again resulting in a delayed identification of the convergence region."

***Referee #2:*** *In addition, I also have some concerns in methods:*
*- The trials and errors in determining the appropriate initial conditions, the scale and the orientation of the sampling (for MHRW and AM) should be described in detail. Otherwise we can not evaluate how effective DEMC is. Is it always promised that MHRW and AM have similar posteriors as DEMC, or it was just by chance?*

**Authors:** After around 50 trials, we set the scale and orientation for MHRW and AM to the most promising values we tested. Therefore, we believe that the MHRW and AM algorithms are as effective as they can be for this particular model and data, so that the effectiveness of the three algorithms can be compared. It must be also pointed out that whatever combination of scale and orientation we used as the best one, there can be no evidence that there is not a better one we did not try. An extensive test of different combinations of scale and orientation, and a detailed comparison of algorithms in terms of effectiveness, are beyond the scope of this paper.

It is always promised that MHRW and AM have similar posteriors distributions as DEMC, since all these algorithms are proven to lead to a representative sample from the posterior distribution. If the posterior samples differed, it would mean then at least one algorithm had not yet converged, and this is a confirmation of the reaching of convergence in the correct region of the parameter space. We will strengthen this point in the revised version of the manuscript.

***Referee #2:-*** *In the two-step method, rather than using a linear regression, to sample considering the coefficient of correlation in the proposal distribution looks more reasonable.*
*Discussion on what you lose by taking a linear regression is needed.*

**Authors:** The reviewer raised an interesting point. The linear regression approach that we used has the advantage of reducing the dimensionality of the proposal distribution, lowering the number of parameters, and addressing a possible over-parameterization. On the other hand, this approach assumes a perfect correlation between the parameters, e.g. we sample one parameter and calculate the value of the correlated one leaving no room for variation. If that was so, the coefficient of correlation would be 1. In our opinion, we lose a source of variation in the parameter values choosing a linear regression over a sampling from a modified distribution which takes into account the coefficient of correlation, but we achieve our aim of a significant simplification of the sampling procedure during the second step.

***Referee #2:*** *[Specific comments]*
*Title: current one may be too general. I may recommend something like "Bayesian calibration of a simple forest model with a multiplicative mathematical structure: a case study with : : :".*

**Authors:** The title will be rewritten as "Bayesian calibration of a simple forest model with a multiplicative mathematical structure: a case study with a Light Use Efficiency model in an alpine forest".

***Referee #2:****Page 6998 Abstract: Introduction part is too long. The first two paragraphs should be shortened and the third one should be more in detail (e.g., consider including one of the conclusions, recommendation of DEMC).*

**Authors:** As required also by the reviewer #1, we will rewrite the abstract in the revised version of the manuscript, including the new results and conclusions we will find as a consequence of the several changes suggested in the review process, and considering the reviewer's comments.

***Referee #2:****Page 6999 Lines 4-10: Eddy-covariance is more ground-based observation method than remote sensing. Thus it is a bit strange for me to mention EC just after remote sensing without any words. It also may be helpful to add advantage and disadvantage of remote sensing and EC.*

**Authors:** We will rewrite this section, as requested also by reviewer #1. We will mention EC before remote sensing and describe briefly their main differences: as described in Baldocchi et al. (1996), the scale and detail of the measurements are the main differences between these two methods. EC is anon-invasive ground technique where GPP is derived from NEP measurements taken at very high frequency (usually 20 Hz), therefore it allows continues measurements on a very high temporal resolution. On the other hand, it has several theoretical assumptions (Burba & Anderson, 2010) that can seriously limit its application in topographically complex environments, the costs for the setup of the EC systems are high, and its estimates are limited to the footprint of the EC tower. The estimate of GPP via remote sensing (through sensors on aircrafts or satellites) has the clear advantage of covering very wide areas and is not as site-specific as EC. It allows estimates of GPP on larger scales (up to global), but needs to be validated by ground measurements in order to ensure the reliability of the data. We will refer to Baldocchi et al. (1996) and Baldocchi (2014) for a more complete comparison of these two methods.

***Referee #2:****Page 6999 Line 11: Better to add a notation that difference of GPP and Re is the carbon balance (relating to Line 2).*

**Authors:** We will add in the revised manuscript a notation stating that the difference between GPP and Re is the Net Ecosystem Exchange (NEE). They are major components of the C balance, and we will refer to Nagy et al. (2006) and Chapin III et al. (2006) for a more detailed description of all the major components and of the methods to estimate them.

***Referee #2:****Page 7001 Lines 15-16: Add literature (or other basis) for "The efficiency of the MCMC technique is highly dependent on the model structure."*

**Authors:** The dependency of the MCMC efficiency on the model structure has been proven, among others, by Gilks & Roberts (1996) and Browne et al. (2009). We will refer to those studies in the revised manuscript.

***Referee #2:****Page 7001 Lines 19-21: Do you think "use of very long chains" is a good method? So why you stick to the speed of convergence in this study?*

**Authors:** Geyer (1992) proposed the use of long chains to monitoring the reaching of convergence, and we believe it is the easiest method to ensure the reaching of proper convergence, but not the fastest one. Given the computational time required for the calibration with a very high number of iteration, we tried to find different and faster solutions to this issue, that would allow to calibrate a model such as Prelued without losing the speed that constitutes one of the main advantages of a simple model. We were unable to find proposal algorithms or model reparameterizations that allowed the MCMC to converge with shorter chains than in the simple MHRW, making the use of long chains the most effective method to ensure the reaching of proper convergence.

***Referee #2:****Page 7001 Lines 21-22: Describe what "more efficient algorithms" are like.*

**Authors:** The reviewer raises a good point. There are several papers on MCMC efficiency, and often they refer to very different things. For example, ter Braak (2006) calculates efficiency considering the mean square errors of different algorithms, but it can also be considered as the proper sampling from a posterior distribution (thus related to the acceptance rate). In this particular study, we considered efficiency as the capability of the algorithm to identify the convergence region minimising the number of model evaluations, i.e. maximising the speed of convergence. We will reformulate the sentence including this definition of "efficiency" in the revised manuscript.

***Referee #2:****Page 7002 Line 16: How multiple chains learn scale and orientation from each other?*

**Authors:** We will refer in the revised manuscript to Ter Braak (2006), where the DEMC algorithm is presented and described in detail. This paragraph will read as follows: "the scale and orientation of the jumps in DEMC automatically adapt themselves to the variance-covariance matrix of the target distribution. It is precisely this that each point in the population learns in DEMC from the

others, nothing more and nothing less. Neither the location nor the fitness of the other points is used in the proposal scheme."

***Referee #2:****Page 7002 Lines 19-20: Add a notation that calculation time is shortened, but the total computational resource needed is not reduced by DEMC.*

**Authors:** The following sentence will be added in line 20: "Although the DEMC algorithm is more computationally efficient, and its implementation can reduce the time needed for calculations, the total computational resource needed are not reduced by its use."

***Referee #2:****Page 7004 Lines 19-24: Why you did not use MODIS's fAPAR product? http://modis.gsfc.nasa.gov/data/dataprod/dataproducts.php?MOD_NUMBER=15*

**Authors:** The MODIS's fAPAR product for the site of Lavarone showed unrealistic variations, which seemed to be unrelated to a possible seasonal trend and were far too high for an evergreen coniferous forest. The NDVI product, on the other hand, did not show such unrealistic variation and we considered it to be more representative of the real situation on the field. Moreover, the NDVI product is available at a higher spatial resolution, which allowed to include in the input data only values read from the footprint of the EC tower, without including neighboring patches of grassland, which clearly affected the fAPAR data.

***Referee #2:****Page 7004, Line 25: Do you mean you used the data of 292 days (of one point)? Describe calibration process more in detail.*

**Authors:** The reviewer is correct. We will rewrite the sentence as follows: "Therefore, we used 292 days for calibration, each one consisting of one data point."

***Referee #2:****Page 7005 Line 21: Only the initial condition is different in the 100 pairs? Describe how the initial condition for each chain was determined.*

**Authors:** The reviewer is correct, only the initial starting point is different in the 100 chains. We will include the following sentence: "The initial starting point of each chain is randomly sampled from the prior distribution at the beginning of the calibration. This is the only difference in the starting condition of the 100 chains."

***Referee #2:****Page 7006 Lines 8-9: Are there any specific reasons why description of GPP and the units in LUE and APAR are different from Eq. 1?*

**Authors:** As stated on line 5, the following equations refer to the model by Horn and Shultz (2011b). In that particular model, GPP is calculated slightly differently from Prelued and LUE and APAR are expressed with different units. We decided to use the original units of measurement in the model and transform our data accordingly.

***Referee #2:****Page 7007 Line 9: Tabulate the parameters and their ranges like Table1, as it is not clear which rows in Table 2 of Horn and Schulz (2011a) are used.*

**Authors:** We will add a statement in the revised manuscript, making clear that we used all of the rows in table 2 of Horn and Shulz (2011a). Each represents the parameterization for one particular site, therefore we used the parameter values in all sites to build the prior distribution for our calibration. We will add a table in the revised manuscript with the information on the prior distribution we built for the model by Horn and Shulz (2011b).

***Referee #2:****Page 7007 Lines 11-12: Describe the basis for the re-parameterization you applied here. The result indicates the re-parameterization itself is not effective, or just your way of re-parameterization is not appropriate?*

**Authors:** We were looking for a way to change the meaning of the parameters, and therefore the model structure, in order to reduce the issue of slow convergence. Unfortunately the possibilities for re-parameterization are extremely limited given the simple structure of the model. Our way of re-parameterization was not effective, which does not mean that re-parameterization in general is not effective, but given the simplicity of the model we changed the parameter meanings as much as possible and we are confident that our way of re-parameterization was appropriate.

***Referee #2:****Page 7008 Lines 10-11: "For the DEMC algorithm, only the chain with maximum loglikelihood was chosen for this purpose." Describe why you look at the best one, not the average. In presenting the posterior distribution for DEMC, you present the result of the best chain, or that of all chains?*

**Authors:** In presenting the posterior distribution for the DEMC, we present the results of the best chain only.
The MCMC algorithm samples the vectors of candidate parameters from a multivariate distribution, and they result in a joint posterior distribution. The values of the parameters in each vector are not independent from one another and must be considered together for every purpose. Therefore, it is not possible to consider the average of the parameter values in all the chains without altering the posterior distribution. However, the reviewer makes a good point: instead of mixing in the individual parameter values, in the revised manuscript we will mix in the whole parameter vectors instead, since they can be considered to be a different sample from the posterior distribution. This approach would allow us to use a lot of information now discarded.

***Referee #2:****Page 7008 Lines 18-21: and Fig 2: Note and discuss some exceptions like for DEMC (blue line).*

**Authors:** The exceptions mentioned by the reviewer are due to the final rearrangements of the figures for the submission process. The procedure described above for the DEMC algorithm will result in new figures which will be described in detail in the revised manuscript.

***Referee #2:****Page 7008 Lines 22-24: it looks strange, as Fig 2 shows different results in , X_0, S_max for DEMC from other methods. Describe why the optimized values for those parameters (in Table 2) are almost same in DEMC too.*

**Authors:** We disagree with the reviewer. Fig.2 shows the same posterior for DEMC as other methods concerning X0, while concerning Smax the convergence region for the DEMC is slightly (but not significantly) different. However, in table 2, the value for Smax is lower for DEMC that for the other algorithms (12.21 for DEMC, 13.28 for MHRW, 12.91 for AM), and not the same.

***Referee #2:****Page 7009 Lines 17-20: Give comments on exceptions: LUE for MHRW and AM, and T_opt/W_i for DEMC.*

**Authors:** We will include the following explanation for the LUE parameter: “both in MHRW and AM, the chain for the LUE parameter is still exploring a wide range of the parameter space. There is no convergence, therefore the prior distribution is not narrowed enough and the posterior distribution is different.”

As stated in a previous point, also the exception of Topt and Wi for the DEMC algorithm are likely due to the final rearrangements of the figures for the submission process. As stated above, the new DEMC procedure will result in new figures which will be described in detail in the revised manuscript.

***Referee #2:****Page 7010 Section 3.1.4: Present the coefficients of correlations (Table 3 shows for 10ˆ6 iteration case, but how about those for 10ˆ4 and 10ˆ5 iterations?) and coefficients in linear regression used here.*

**Authors:** The coefficients of correlation at $10^4$ and $10^5$ iterations were not calculated. We based the second step of the calibration on the correlations between parameters found during the $10^6$ iterations first step, since it was the only one that gave reliable results. Based on that we removed 2 parameters.
The coefficient of the linear regression for the second step were calculated on the appropriate first step. We will add a table for the coefficient of each linear regression in the revised manuscript.

***Referee #2:****Page 7010 Section 3.1.4: Is the linear relationship you get here by chance, or results of over-parameterization?*

**Authors:** The reviewer raises a good point. The very high correlation coefficients between some of the parameters (>= 0.9) clearly indicates a linear relationship between them. In most of the cases a linear relationship between parameters is a result of over-parameterization, especially when the parameters are empirical and therefore not necessary for a physical or physiological reason. In our case, the parameters that resulted to be correlated have similar role in the model structure: $\beta$ and $\gamma$ are both involved in the response to APAR, while X0 and $S_{max}$ are both involved in the response to temperature. Given their similar role and their empirical nature it is very likely they are redundant and not strictly necessary, which is why we believe that the linear relations we found are a result of over-parameterization.

***Referee #2:****Page 7010 Section 3.1.4: Add discussion on the comparison with the result of the 10ˆ6 iteration case in the single-step method.*

**Authors:** We thank the reviewer for bringing this interesting possibility to our attention. We will add a paragraph on this comparison in the revised manuscript, considering also the results from the

new DEMC procedure described above. It will also be linked to the evaluation of model results requested by the reviewer#1 and will include a discussion on the similarities and differences between the posterior distributions of the parameters that are present in both calibrations.

***Referee #2:****Page 7011 Line 16: Why can you say "possibly the main factor"? The slower convergence for the LUE model indicates different possibility.*

**Authors:** The structure of the model by Horn and Shulz is less multiplicative than Prelued, but not much. It still relies on several multiplications and could have the same structure-related issues then Prelued. We will reformulate the sentence in the revised manuscript as follows: " the multiplicative structure of Prelued was likely one of the factors responsible for the difficulties in the calibration, but is unlikely to be the only one".

***Referee #2:****Page 7012 Line 22: Present the result to support "this did not result in better model performances over all".*

**Authors:** This point has been raised also by the reviewer #1. We carried out a posterior model evaluation for the approaches that resulted in proper convergence, which we decided not to include in the paper given its main focus on Bayesian calibration. The model results were insensitive to the algorithm used or to the procedure applied. We will include the results of model evaluation in the revised version of the manuscript, focusing on the differences (or their absence) between the different calibration procedures.

***Referee #2:****Page 7012 Lines 25, 28: Describe the trials and errors you did for MHRW and AM before starting calibration more in detail (see general comment too).*

**Authors:** As stated above, an extensive test of different combination of scale and orientation for the algorithms we used is beyond the scope of this paper. We do not think it would add any useful information to the reader, and that it is not necessary to the comparison of the effectiveness of the different algorithms.

***Referee #2:****Page 7021 Table 3: Test the statistical significance and show the results. Also, highlighting*

*the different sign case may not be so useful, as the difference of 0.006 and -0.021 is not significant (both of them indicate no correlation).*

**Authors:** We agree with the reviewer about highlighting the different sign of the coefficient, and we will remove the highlighting in the revised version of the manuscript. However, we do not think that the tests for statistical significance will add any useful information to the reader: since the models are deterministic, correlations of exactly zero between parameters are impossible, unless one parameter has zero impact on model output. The only relevant information for this study is how important these correlations were, in order to improve the model structure removing redundant parameters.

***Referee #2:****Page 7023 Fig. 1: It is hard to get useful information from the figures for DEMC. How about presenting the average and the range of uncertainty (e.g., standard deviations) for 100 chains (or presenting the best one?). Same for other figures too.*

**Authors:** We agree with the referee that figures should be improved to assure their readability. We will follow the procedure for DEMC described above, mixing in all the parameter vectors as different samples from the same posterior, and we will re-arrange the figures in a clearer way according to the new results.

***Referee #2:****Page 7024 Fig. 2: Is it no problem that sometimes red lines are invisible? Also check if the ranges of y axis are appropriate with DEMC for X_0 and S_max (see blue lines).*

**Authors:** The reviewer is right, the red lines are sometimes invisible because they overlap with the others (blue ones especially), in a few cases perfectly. The range of the y axis has been calculated on the data: reducing it would cut the upper part of the distributions, while increasing it keeping the dimension of the figures fixed would squeeze them and make them even less visible. We will improve the readability of the figures, also with the addition of an appendix or supplementary material, in the revised version of the paper.

***Referee #2:*** *[Technical corrections]*
*Page 6999 Line 7: Eddy-covariance -> "Eddy-Covariance (EC)" and then use EC for later parts.*
*Page 7003 Line 5: FL_j, FS_j, FD_j are included as F_ij in Eq.1? If so specify i=L,S,D,*
*and replace FL_j, FS_j, FD_j with F_Lj, F_Sj, F_Dj.*

*Page 7005 Line 9: (Mäkelä et al., 2008a) -> Mäkelä et al. (2008a). In some parts you cite "Mäkelä et al. (2008a)", but in the reference list there is only one Mäkelä et al. (2008).*

**Authors:** We will rewrite EC as suggested, change the mathematical notation and check the references throughout all the revised manuscript.

***Referee #2:****Page 7005 Line 26: Do you mean ZF is used as Ts in eq (9)?*

**Authors:** The reviewer is correct. We will add the following sentence in page 7007 line 4: "ZF calculated in Eq. (11) is therefore used as Ts in Eq. (9)"

***Referee #2:****Page 7007 Line 1: What is "(-)" after _?*

**Authors:** It identifies a dimensionless parameter, in contrast with all the other parameters for which the units of measurements have been reported, in brackets, while describing their role.

***Referee #2:****Page 7007 Line 18: "faster the convergence" -> "the (a?) faster convergence,"*

**Authors:** The sentence will be reformulated as "a faster reaching of convergence"

***Referee #2:****Page 7010 Line 6: 3.1 -> 3.1.1?*

**Authors:** We will replace 3.1 with 3.1.1

***Referee #2:****Page 7010 Lines 6, 8: _ -> X_0?*

**Authors:** That is correct, the correlation is between X0 and Smax. We will check the manuscript and replace τ with X0 wherever necessary in the revised version.

***Referee #2:*** *Page 7010 Lines 9-10: Markov Chain Monte Carlo -> MCMC*
*Page 7011 Line 4: "6-parameters empirical model" -> "6-parameter empirical model"*
*Page 7019 Table 1: Add that the distribution is uniform.*

**Authors:** We will include the last three comments in the revised manuscript

**References:**


Bagnara, M., Sottocornola, M., Cescatti, A., Minerbi, S., Montagnani, L., Gianelle, D., & Magnani, F. (2014). Bayesian optimization of a light use efficiency model for the estimation of daily gross primary productivity in a range of Italian forest ecosystems. *Ecological Modelling*, in press.

Baldocchi, D., Valentini, R., Running, S., Oechel, W., & Dahlman, R. (1996). Strategies for measuring and modelling carbon dioxide and water vapour fluxes over terrestrial ecosystems. *Global change biology*, 2(3), 159-168.

Burba, G., & Anderson, D. (2010). A brief practical guide to eddy covariance flux measurements: principles and workflow examples for scientific and industrial applications. Li-Cor Biosciences.

Baldocchi D (2014) Measuring fluxes of trace gases and energy between ecosystems and the atmosphere – the state and future of the eddy covariance method. *Global Change Biology* 20(12):3600-3609.

Nagy, M. T., Janssens, I. A., Curiel Yuste, J., Carrara, A., & Ceulemans, R. (2006). Footprint-adjusted net ecosystem CO2 exchange and carbon balance components of a temperate forest. *Agricultural and forest meteorology*, 139(3), 344-360.

Chapin III FS, et al. (2006) Reconciling carbon-cycle concepts, terminology, and methods. *Ecosystems* 9:1041-1050.

Gilks W, Roberts G (1996). "Strategies for Improving MCMC." In W Gilks, S Richardson, D Spiegelhalter (eds.), *Markov Chain Monte Carlo in Practice*, p. 89-114. Chapman & Hall, Boca Raton, FL.

Browne, W. J., Steele, F., Golalizadeh, M., & Green, M. J. (2009). The use of simple reparameterizations to improve the efficiency of Markov chain Monte Carlo estimation for multilevel models with applications to discrete time survival models. *Journal of the Royal Statistical Society. Series A, (Statistics in Society)*, 172(3), 579–598. doi:10.1111/j.1467-985X.2009.00586.x

Geyer, C. J. (1992). Practical markov chain monte carlo. *Statistical Science*, 7, 473-483.

Ter Braak, C. J. F. (2006). A Markov Chain Monte Carlo version of the genetic algorithm Differential Evolution: easy Bayesian computing for real parameter spaces. *Statistics and Computing*, 16(3), 239-249.

Horn, J. E., Schulz, K.(2011a). Identification of a general light use efficiency model for gross primary production, *Biogeosciences*, 8, 999–1021, doi:10.5194/bg-8-999-2011.

Horn, J. E., Schulz, K.(2011b). Spatial extrapolation of light use efficiency model parameters to predict gross primary production, *J. Adv. Model. Earth Syst*., 3, M12001,5. doi:10.1029/2011MS000070.