We are grateful to the two anonymous Referees for their very constructive comments which we used raising the quality of the manuscript and clarifying some important aspects. Please find below our answers comments (in italics) to the Referees comments (normal formatting):

**Anonymous Referee #1**

This paper presents the modifications brought to the 3D-CMCC ecosystem model that simulates growth of heterogeneous forest ecosystems with the representation of different species, several vertical layers and age classes. The representation of vertical and horizontal heterogeneity is still very rare in most ecological models and this approach is needed within the community. In this study the authors evaluate the model's ability to simulate GPP at site level by comparing the simulations to FLUXnet-derived GPP data with a focus on the time scale for data aggregation, and level of details in the initialization of the model. The paper is well written, however the split between results and discussion makes the 2nd part of the paper more difficult to follow through. Some important aspects of this type of work however seem to be lacking and are described below.

Improvements to the model

From the text, the reader is expecting to see the effects of the changes implemented in the version 5.1 of 3D-CMCC FEM: new carbon allocation, phenological, and autotrophic respiration modules, a 3D-canopy representation, tree carbon-nitrogen allocation and water flows that differentiate 3D-CMCC v5.1 from its previous version. The equations of the modifications brought to the code are shown in the appendix without a lot of details. Moreover, in the description that is given, it is not clear what has been changed from the previous version of 3D-CMCC. For example, in the text, L136 & L176, the authors say that the phenology and the canopy representation have been improved, in appendix A, nowhere is explained the difference between the previous and the new implementation: L 721 'the C flux is still estimated by . . .'. The interest of the work described in this study as a model improvement would be more convincing if the modifications implemented were clearly shown in regards to the previous versions of each element. The authors should explain clearly for each of the improvement what was used in the previous model and how it has been changed.

*We thank Referee #1 for the request to more clearly describe differences between previous and current model version. In the revised manuscript the changes to the model have now been described in more detail where needed. Some of the new parts of the model (e.g. autotrophic respiration) have already been described in other published works and here we refer to them. This served our goal to avoid overloading the manuscript for the benefits of easier readability.*

The validation would also strongly benefit from the comparison of the simulations with the previous versus the new version to see where the addition of details in processes actually improves the model simulations and where some more improvements are still needed.

*We agree that such comparison would be illustrating the improvements to the model well. However this has not been possible because previous model version could not be applied to all sites (e.g. high latitude sites). The previous model version lacks of many fundamental processes that are now included in the new version. To avoid a partial models comparison with the risk of extending the manuscript without giving a complete picture, we preferred not following this suggestion.*

Tuning

The tuning is only quickly mentioned and the paper lacks a clear description (that could be put in appendix) of the tuning process. Also, related to the previous comments, I could not understand if more tuning was done after the changes were implemented, and if so a list of the species-specific tuned parameters with the values in the previous version and in the new version is needed.

*In Collalti et al., 2014 model has been used and parameterized for a turkey oak site. Here, the model has been used for other forest species. The new model parameters are listed in a new table within the manuscript (Supplementary material, Table S3), as suggested by the Referee, highlighting the new ones in respect to the previous version.*

The authors choose to adopt a generic parameterization for species-specific parameters in order to make the model applicable to different sites without prior site-specific calibration. They acknowledge on L453 that this approach yields large uncertainties but do not give any element about how big this uncertainty is. It would be interesting to have an idea of the magnitude of this sensitivity/uncertainty, maybe using a Monte-Carlo approach to propagate parametric uncertainty into the simulations or testing the same simulations with site-specific calibrations to get a feeling of how much is lost through this approach, even if genericity clearly is a choice for future applications.

*We completely agree with the Referee this would be an attractive objective for future applications. However the scope of this manuscript is on the multisite application and the demonstration of the generic power of the new parameterizations. An uncertainty analysis would be another study where one needed to carefully select the sites depending in the availability of test data. The test data availability differs much across the sites considered in this study.*

Validation

The validation of the model is the heart of the paper. The authors try to be exhaustive by using several performance indices, but this section lacks some crucial elements to be fully convincing. For example, many performance indices are calculated and shown but nowhere in the text it is explained or analyzed if they are consistent with each other, if not why and what is the difference in their meaning.

*We have followed these comments and trying to better highlight consistency among indexes, at this regard we changed Bias index with the standardized Mean absolute Bias (MABstd) for consistency with other indexes (see Fig 1). We would anyway focus that most of the several indexes calculated (i.e. r, NRMSE, MABstd, MEF) in the manuscript are described in the result section and discussed. Due to the high number of indexes used, in some cases, only the best and worst results for each index are described and discussed skipping intermediate values. This to avoid making the manuscript too hard and long and to allow it to be completely comprehensive, we have not focused during the explanation on any single index value for any single site.*

In Fig 1, the succession of indices is even harder to follow because all indices do not go in the same direction, i.e. a good model will have a NRME close to 0 but a correlation coefficient close to 1. It would be very useful to find a way to express all indices so that they would all go in the same direction, at least visually on Fig. 1. Also, plotting the daily and monthly indices on the same graphs with different colors would ease the comparison.

*We have modified figure 1 following comments of both Referees including comparison between original dataset and a new Y dataset following Zhao et al. (2012) method for both daily and monthly timescales, respectively. We have included MABstd index as requested from Referee #2 and removing Bias. We would like anyway to point out that a summary graph with correlations and normalized standard deviations are shown in figure 2 for both daily and monthly values, respectively.*

For interannual variability, I found that the bar plots are hard to read. Since it is an evolution with a x-axis with years, maybe line plots would help the reader in understanding the results.

*It is true that somewhat more irregular bar plots are a bit hard to read, but we did not find a better representation and used the way in which relationships are typically presented for better comparison with similar studies. For figure 9 (we think the Referee #1 refers to this figure) we followed Anav et al. (2010) and Keenan et al (2012).*

Also, I regretted that the simulations with varying degrees of complexity in the model initialization were not shown along with the other figures.

*Figures concerning simulations with varying degrees of complexity have been deliberately separated from the main part of the manuscript. Because we think that this concerns a side aspect that demonstrate that a lower degree of model description during the initialization phase gets worse model results.*

The 3D-CMCC model represents carbon, (energy?) and water fluxes. GPP in itself cannot be sufficient to evaluate the relevance of a model in simulating the functioning of an ecosystem. GPP is highly intertwined with other variables and this only appears discretely in the paper when the water stress is suggested to explain the bad performance of the model at the Mediterranean site. It is important to show the outputs of the model as for water and energy variables. I don't know if water vapour variables are available for the 10 fluxnet sites used in the study but latent and sensible heat fluxes must be available. With such data, the validation of the model would result much stronger, else, the authors could still show the model's outputs for water and energy variables to try to explain the water problem at the Mediterranean site and give a less partial overview of the model's performances.

*The 3D-CMCC Model currently doesn't yet include the heat fluxes. We admit that GPP is only one of many aspects of an ecosystem but it is a very important one, because it is linked directly or indirectly with many other different processes e.g. carbon partitioning and allocation, respiration, leaf production, canopy expansion and so on.*

Specific comments & typos

L113 set of annual series *(done)*
L123 answer the following *(done)*
L496 as like as *(done)*
L567 Del Pierre instead of Delpierre *(done)*

Anonymous Referee #2

The manuscript by Collalti et al. presents the developments recently implemented into the 3D-CMCC Forest Ecosystem Model, which relies on the concept of Light Use Efficiency for carbon assimilation. It also presents an evaluation of the Gross Primary Production (GPP) flux simulated by the 3D-CMCC Forest Ecosystem Model against in situ data at 10 European forest sites, from daily to multi-annual time scales. Additional simulations are also presented in order to test if a more in-situ based representation of the forest characteristics can improve the model performances. Overall, the manuscript contains the information needed for understanding the model structure, the simulation set-up and the model evaluation.

However, I would suggest to re-shape the model description section and to provide more information on the model initialization at site level. I have also some serious concerns about the time series analysis performed. These points are developed here below. Model description section:
Most of the model features are described in the Appendix A. Because model development is the main scope of GMD, I would suggest to move the model description in the body of the manuscript, in the 'Materials and Methods' section.

*Thank you for the comment, we moved description of model improvements and their algorithms from appendix section to material and methods. We furthermore inserted mandatory data for model initialization (dbh, tree density, tree height and age) in the site description column in Table 1 and we included a full list of species-related parameters needed for mode runs (Table S3)*

It is mentioned that information regarding model initialization at site level was taken from the BADM database but this information is not reported in the manuscript. It would be valuable to specify all these initial values that are site-specific (DBH, tree height, age and density) and to put these values in an appendix, for instance.

*We agree with Referee #2 and we added information about initial model values for each site on Table 1*

Time-series analysis:
You attend to analyse the GPP at different time-scales (day, month and year) but the way you do it is not appropriate to my opinion. Your daily signal contains information at lower frequencies especially at monthly time-scale (seasonal). Indeed, the largest variations of the daily signal are the seasonal variations, in such a way that the correlation of the daily signal with observation is very similar to the one you get for the monthly time-series (top panels of figure 1).

*We agree with this comments and decided to include a new data analysis using the decomposition approach showed in Zhao et al. (2012) to avoid effects of seasonality on data comparison (line 266-271, see below).*
For the bias, by definition, based on equation (3), the model scores are the same for both daily and monthly time series, except that one is expressed per day and the other per month (compare bottom panels of figure 1).

*We partially agree with comments of the Referee #2, the choice to report both daily and monthly correlations was driven by consistency with literature and providing a benchmark to compare the model results with other models. Furthermore, using monthly data (obtained as a sum of daily values from L4 data) we avoid spikes that are frequently present in the daily values, as the Referee can see in the Taylor diagram daily and monthly data don't follow the same tendency site by site. Furthermore, after adding the Y dataset differences between daily and monthly time-series are even more evident.*

I think working with inter-monthly (IMV) and inter-yearly variabilities (IAV) as you did goes in the right direction but based on equation (4) this is not done properly. When defining IMV, it will be more appropriate to subtract the mean annual value (for each year) from the monthly GPP time-series instead of subtracting the longterm mean (over all the available years, as you wrote). By that way, you do not account for the biases at annual time-scale when analysing the IMV.

*We agree with this reservation, but see the benefit in using equation 4 ($IMV_{(EC \vee MD)i} \vee IAV_{(EC \vee MD)i} = GPP_{(EC \vee MD)i} - avg(GPP)_{(EC \vee MD)}$) also used in Vetter et al. (2008); Keenan et al.*

*(2012) and Balzarolo et al (2014), because it represents a way to quantify the magnitude of variability for both month and year over long period (described by the term avgGPP)*

There are sophisticated time-series decomposition techniques that have been applied to terrestrial carbon flux analysis (for instance, Mahecha et al., 2007) but there are also simple ones that I encourage you to use (for instance see equation (2) of Zhao et al., 2012) in order to perform your analysis on these flux anomalies, only. I would also suggest that you use the Mean Absolute Error (MAE) instead of the mean error for the Bias metric, in such a way that you do not compensate for errors of opposite signs when averaging.

*We agree with Referee #2 and we have included according with the new co-author Zhao also the MABstd index and removing Bias index. Statistical indexes have been computed for both datasets including Y dataset (see above).*

Based on the legends of figure 1 and table S1, I'm expecting to find in the figure 1 the same values than those reported in Table S1 for the correlation, NRMSE, MEF and Bias for daily and monthly time-series. But it is not the case for many sites and metrics, among others: at BE-Bra, MEF and Bias values of the daily GPP flux differ between Figure 1 and Table S1 and the Biases of the monthly time-series are also different at this site. NRMSE of the monthly time-series differ at FR-Hes. In addition, the Bias values reported in Table S1 and Table 3 do not match at IT-Ren (2L_2C) for both the daily and monthly time-series. All this is very confusing. Are there problems in the values that you report on the plot or in the tables, or problems in the legends in case the figure and the tables do not represent the same information?

*Thanks! we are very grateful that Referee #2 found errors in graphs and tables that we have overlooked. We corrected all of them in the revised manuscript*

Other comments You present results in terms of GPP but describe how the model computes autotrophic respiration and other processes not directly linked to GPP calculation. I would suggest you to be consistent by either keeping the description of the respiration flux and showing results in terms of NPP or removing the paragraphs where other processes than C assimilation are described.

*Due to model architecture, the computation of GPP is strongly related to processes like autotrophic respiration. It's then clear how AR strongly and indirectly affects GPP of current and subsequent years (see for example Non Structural Carbon pool). This new part is then relevant in the computation of Carbon flux and it cannot be excluded from model description.*

Page 6899 lines 7-8 you write that "the C/N stoichiometry is constant and depends on species, unfortunately, the model still lacks of an interactive C-N cycle". Based on this information, I don't think it is appropriate to write that 3D-CMCC FEM models nitrogen allocation and represent nitrogen pools.

*We clarify that model doesn't represent nitrogen allocation specifically it couples nitrogen pools with to the dynamic carbon pools through a fixed C to N ratio.*

In the results section, you never try to compare your results with other modelling studies. Other models based on LUE (or not) have also used some of the sites you studied for analysing the GPP flux, like you. It would be good - where possible - to compare your results with these studies, for instance: Balzarolo et al. (2014), Ogutu et al. (2013), Zhao et al. (2012), Yuan et al. (2007)

*To keep the strict focus on the multi-site study, we deliberately decided to not include comparison with other modelling studies for single sites.*

When discussing the results in terms of IAV and the capacity of the model to get the right sign of the IAV, be more critical on your results especially when writing that 'model reproduced well the timing of anomalies in more than half of cases' (page 6883 line 29), keep in mind that with a random selection process you catch the sign of the anomalies in half of the cases, already.

*We agree that our statement was a bit too positive. This was changed in the revised manuscript.*

Do other models perform better than 3D-CMCC FEM in terms of IAV?

*The recent modelling studies that we are aware show unanimously the difficulties of models to explain the large interannual viariability in cases where no obvious triggers like management or climatic extreme are at work (e.g., Keenan et al., 2012; Wu et al., 2013). Our results confirm this.*

*Anav, A., D'Andrea, F., Viovy, N., Vuichard, N., 2010, A validation of heat and carbon fluxesw from high-resolution land surface and regional models. Journal of Geophysical Research, 115, 1-20*

*Balzarolo, M., Boussetta, S., Balsamo, G., Beljaars, A., Miagnan, F., Calvet, J.-C., et al., 2014. Evaluating the potential of large-scale simulations to predict carbon fluxes of terrestrial ecosystems over a European Eddy Covariance network. Biogeosciences , 11- 2661,2678.*

*Keenan, T.F., Davidson, E., Moffat, A.M., Munger, W. and Richardson, A.D., 2012. Using model-data fusion to interpret past trends, and quantify uncertainties in future projections, of terrestrial ecosystem carbon cycling. Global Change Biology, 18, 2555–2569.*

*Moffat, A.M., Papale, D., Reichstein, M., et al., 2007, Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes. Agricoltural anf Forest Meteorology, 147, 209-232.*

*Vetter, M., Churkina, G., Jung, M., Reichstein, M., et al., 2008. Analyzing the causes and spatial pattern of the European 2003 carbon flux anomaly using seven models. Biogeosciences, 5, 561-583.*

*Wu, J. et al., 2013. Modelling the decadal trend of ecosystem carbon fluxes demonstrates the important role of biotic changes in a temperate deciduous forest. Ecological  Modelling, 260, 50-61.*