In the following, reviewers' comments are in black, whilst our responses are in red. The text added in the revised version of our manuscript is in italics.

Anonymous Referee #2

(A) General comments :This paper describes the evaluation of a revised version of the ORCHIDEE model, incorporating representations of the carbon (C) and nitrogen (N) (and water) interactions. This paper comes 9 years after the initial publication of a first version of a C-N version of the ORCHIDEE model (Zaehle & Friend, 2010; hereafter ZF10). As stated by the authors, this version of the ORCHIDEE model is very similar to the one already published by ZF10, with several modifications (listed from P3L30 ("Page 3 Line 30") to L4L11 and on P5L4-9). These modifications are mostly not mathematically described in the paper. Different from ZF10 that evaluated fluxes simulated over a set of European forests, this paper provides an evaluation of the revised ORCHIDEE over GPP data acquired across the globe (using both Fluxnet data and a machine-learning product predicting GPP across the globe "MTE-GPP"). After this initial evaluation, the paper presents sensitivity analyses (SA) aiming at inferring the role of simulated C-N coupling on the centennial dynamic of simulated GPP.
When evaluating a revised version of a model, one needs two references: (1) ground-truth data and (2) a previous version of the model from which the one we are evaluating has been developed. Both are mandatory to provide a thorough evaluation of a revised version of a model, and conclude as whether or not the developments have indeed improved the model.
We thank the reviewer for their view on model evaluation but the listed prerequisites narrow the definition of model evaluation to what is generally considered a benchmark. Contrary to the reviewer's view, the literature shows a much richer practice which reflects differences in objectives across models and model developments. For instance, looking at the temporal or spatial dynamic of a model or performing parameter sensitivity analysis are also valid ways of evaluating model behaviour and could even be more insightful than a comparison against ground truth data.

In the study under review, the objective was not to compare the former trunk version (r3977, without a nitrogen cycle) against the version presented in the manuscript and including a nitrogen cycle and the C/N interactions. The nitrogen cycle is a new functionality (compared to the former trunk version) and given its link to ecological theory we certainly want to keep it even if this implies a loss of model skill for few pools or fluxes. That's the reason why we did not focus on model evaluation against a former model version but look more precisely at the model response to the coupling/decoupling of the C and

N cycles. However, we understand that knowing how does the current model version compare to the former trunk version is an outcome of the model evaluation that could be expected by some readers. In the revised manuscript, we propose to show in a supplementary figure (Figure S1, see below) - similar to Figure 1 - the GPP model/data comparison at the site level for the former trunk version (r3977).

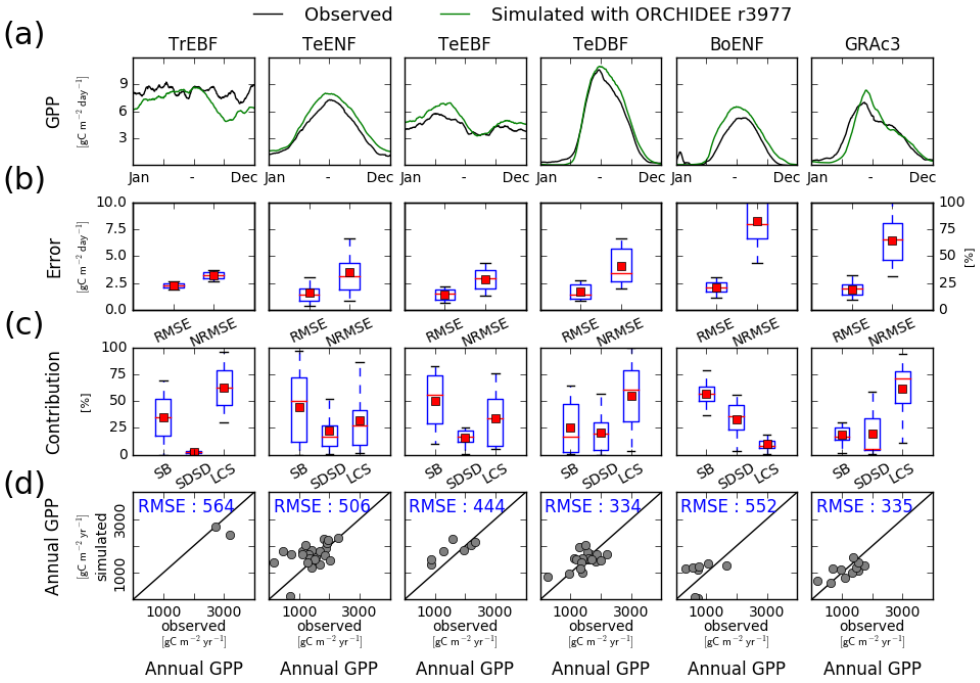


Figure S1 : Site-level evaluation of ORCHIDEE r3977 (ie without N cycle) simulations against Fluxnet observations. (a) Vegetation-class mean seasonal variations of GPP, (b) Root Mean Square Error (RMSE) and Normalized Root Mean Square Error (NRMSE) of simulated daily variations of GPP per vegetation class, (c) Attribution of the Mean Square Error (MSE) of the daily variations of GPP to model errors on mean value (SB), standard deviation (SDSD) or correlation (LCS) (Kobayashi and Salam, 2000) and (d) simulated vs. observed Annual mean GPP at site-level. On panels (b) and (c), the box extends from the lower (25 %) to upper quartile (75 %) values of the data, with a red line at the median and a red square at the arithmetic mean. The whiskers extend from the box to show the range of the data within 1.5 × (25–75 %) data range.

Page 14 line 22 of the initial manuscript, we add the following paragraph for describing how revision 3977 (without the N cycle) compare to rev 4999 (with the N cycle) in terms of model/data agreement at Fluxnet sites:

“In order to analyse how ORCHIDEE r4999 performs compared to r3977 (original version without the nitrogen cycle), we evaluated the GPP simulated by ORCHIDEE r3977 against Fluxnet observations (Fig. S1). The model/data agreement for r3977 was comparable to the one for r4999 but slightly better. In particular, NRMSE of the simulated daily GPP flux (Fig. S1b) and the RMSE of

the simulated annual mean GPP (fig. S1d) were lower in r3977 compared to r4999, for Temperate Evergreen Needleleaved and Broadleaved Forests and Temperate Deciduous Broadleaved Forest sites. Especially for Temperate Evergreen Needleleaved and Broadleaved Forests sites, the lower mean NRMSE of the simulated daily GPP at the PFT level for r3977 was due to a narrower range of NRMSE values at site level (whisker boxes are narrower), indicating that the NRMSE was not systematically lower at all sites but only at some specific ones."

We also propose to extent figure 8 with a third panel where we map the difference between the annual mean GPP simulated by the former trunk version of ORCHIDEE (ref3977) and the annual mean GPP computed by the MTE-GPP product over 2000-2010;

Last regarding the GPP evaluation, we propose to add on Figure 9 (see below), the evolution of the global mean and latitudinal band mean GPP simulated by the former trunk version in addition to the GPP simulated by the revision 4999 (S1-CNdyn configuration) and the GPP estimated by the MTE-GPP product.

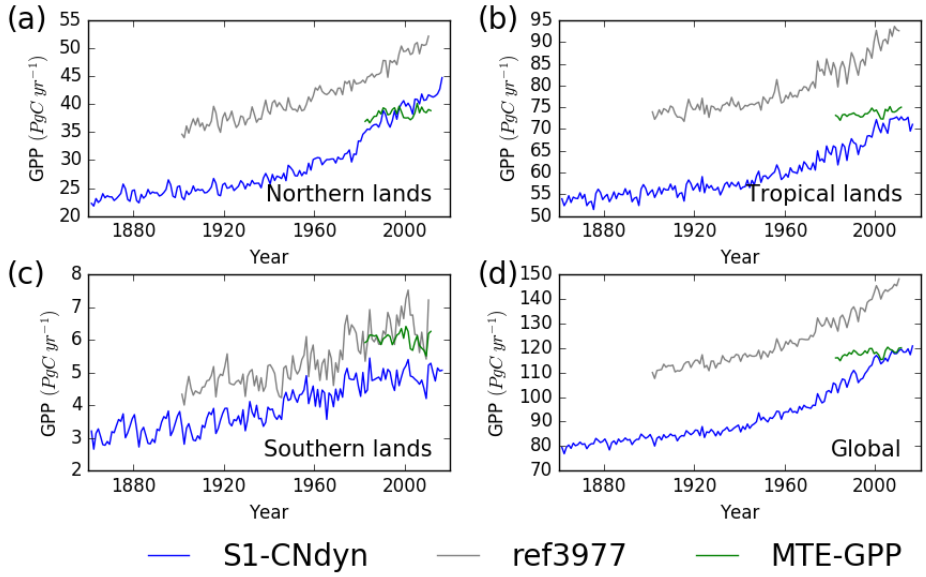
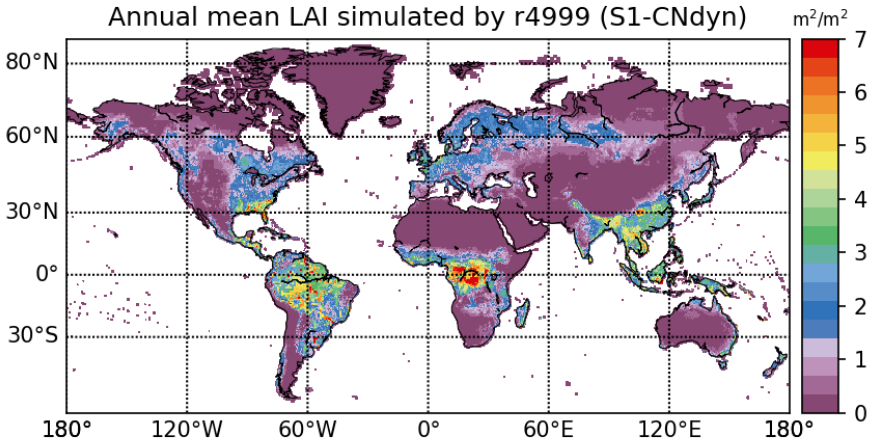


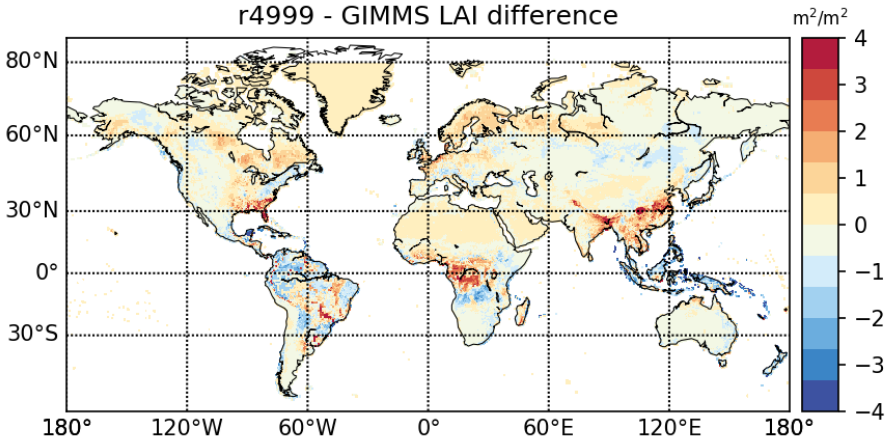
Figure 9: Evaluation of GPP from ORCHIDEE against the observation-based MTE-GPP product for four regions. Time evolution of the annual mean GPP (PgC yr^{-1}) estimated by ORCHIDEE r4999 (in blue) and ORCHIDEE r3977 (in grey) and by the observation-based MTE-GPP product (in green) for (a) Northern lands (>25°N), (b) Tropical lands (<25°N and >25°S), (c) Southern lands (<25°S) and (d) all lands.

In addition to the GPP evaluation, we now add an evaluation of the LAI at global scale (see below). Consequently, new figures have been created similar to Figures 8 and 9 but for LAI (respectively Figure S3 and S4).

(a)



(b)



(c)

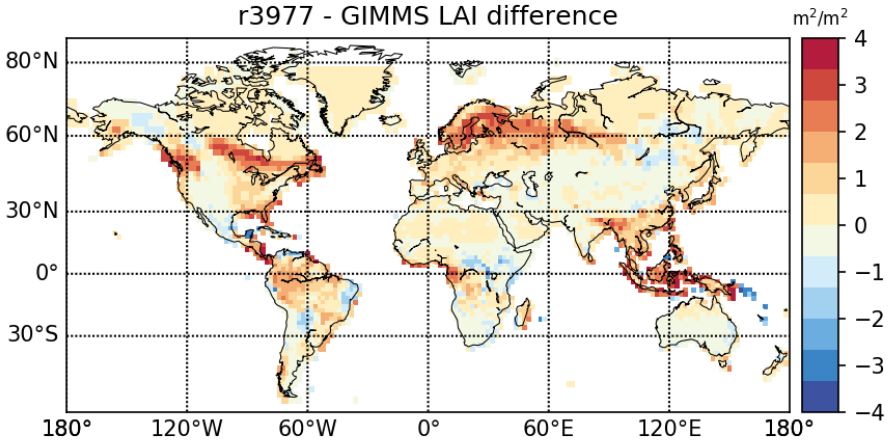


Figure S3: Global scale evaluation of ORCHIDEE against the observation-based GIMMS product. (a) Global distribution of the simulated annual mean LAI by ORCHIDEE r4999 (m^2 m^{-2}) over 2001-2010; (b) Global distribution of the difference between the simulated annual mean LAI by ORCHIDEE r4999

and the GIMMS product; (c) Global distribution of the difference between the simulated annual mean LAI by ORCHIDEE r3977 and the GIMMS product

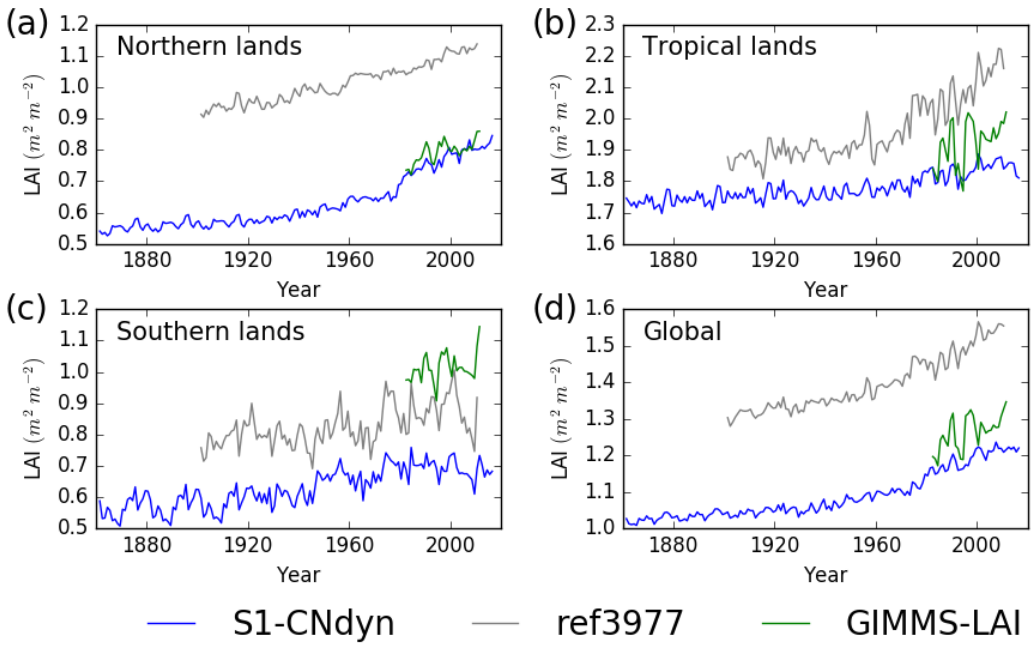


Figure S4: Evaluation of LAI from ORCHIDEE against the observation-based GIMMS product for four regions. Time evolution of the annual mean LAI (m^2 m^{-2}) estimated by ORCHIDEE r4999 (in blue) and ORCHIDEE r3977 (in grey) and by the observation-based GIMMS product (in green) for (a) Northern lands (>25°N), (b) Tropical lands (<25°N and >25°S), (c) Southern lands (<25°S) and (d) all lands.

The following sentences have been added page 17 line 16 of the initial version to present the results about the LAI global distribution and mean annual values averaged per latitudinal regions:
"Similarities between the simulated global distributions and biases in GPP and LAI (compare Fig. S3a to Fig. 8a, and Fig. S3b to Fig. 8b) suggest that the bias in GPP originates from the bias in LAI rather than from more fundamental issues with the calculation of GPP. The model/data agreement for LAI when averaged per latitudinal band is comparable to the one for GPP, with a good agreement for the Northern and Tropical lands and model underestimation in the Southern lands."

The following sentences have been added page 17 line 16 of the initial version to compare performances of the rev3977 and rev4999 at simulating LAI and GPP at global scale.
"The agreement between the modelled and observed annual mean LAI and GPP summed over three latitudinal bands as well as at the global scale was higher for r4999 (ie S1-CNdyn simulation) compared to r3977 without the

nitrogen cycle (see Fig. 9 and S4). r3977 systematically overestimated LAI and GPP for any region, except for the Southern lands where r3977 provided similar values than the GIMMS and MTE-GPP products, respectively. Compared to GIMMS and MTE-GPP products, gridded annual mean LAI and GPP values simulated by r3977 were overestimated in the Northern lands with biases exceeding those found in r4999. On the opposite, biases of the r4999 were higher than those of r3977 in the tropical regions, in particular in Central Africa (see Fig. 8 and S3).”

As regards ground-truth data:
- the model is evaluated against GPP time series. This is indeed an important flux, for which the model needs be evaluated. However, we are here dealing with a coupling of C and N cycles in the model. Evaluating the model against C flux data is clearly not enough. I know that N data are much less common than C data (e.g. Vicca etal. 2018), but the effort has already been made in earlier versions of ORCHIDEE (seeZF10 for instance). Hence I expect at least a minimal evaluation of this new version against some N data;

Focusing on GPP flux was a deliberate choice, which we thought was motivated by exactly the same arguments as the reviewer:

- N data are much less common than C data. Model evaluation against such data would be mostly anecdotal (or at least very partial) which goes against the objective of a global scale model such as ORCHIDEE.
- The effort of looking at N data has already been made in an earlier version of ORCHIDEE. In this respect, we think that the model version we present here has not changed sufficient with regard to the way the N dynamics are modelled compared to ZF10 (in which the requested evaluation has been presented).

Additionally, the extensive dataset of carbon fluxes from the Fluxnet network has so far, however, not been used for evaluating any of the ORCHIDEE versions with a nitrogen cycle. Rather than reproducing ZF10 we engaged into a relatively original study of which the main findings are presented in the manuscript under review. Thus, the manuscript combines an extensive evaluation of the model GPP at FluxNet sites with an evaluation of the impact of nitrogen limitation on GPP under atmospheric CO2 increase. Such combination also directly contributes to the novelty of the manuscript.

We agree that this choice was not sufficiently motivated in the initial manuscript. Consequently, we propose to add the following sentence in the revised manuscript, Page 3 line 16 of the initial manuscript:
“While the OCN model (Zaehle and Friend, 2010), the predecessor of ORCHIDEE r4999, has already been evaluated over a restricted set of sites for which C and N data are available, the extended C flux dataset from the

Fluxnet network has so far not been used for an in-depth evaluation of an ORCHIDEE version that includes the N cycle and the C/N interactions."

- the N cycle also impacts respiration. Since the Fluxnet data include both daytime and nighttime (i.e. respiration) fluxes, I see no good reason for the authors not to evaluate the model ability to simulate respiration fluxes;
In addition to the estimate of GPP flux, the partitioning of the NEE flux measured at site provides an estimate of the Total ecosystem respiration, which includes the autotrophic respiration and the heterotrophic respiration by soil microorganisms. Because the heterotrophic respiration is highly dependent of the long-term site history in terms of land use - which we can not account for in our modelling set-up at the site level -, direct comparison of the modelled total ecosystem respiration with the one based on site measurements will include a systematic bias. For this reason, we did not evaluate the modelled ecosystem respiration against site data as neither a good nor a poor match between the data and simulation could lead to a robust conclusion concerning model performance.

- since part of the sensitivity analysis implies simulated transpiration fluxes, I also expect to see some comparison of simulated evapotranspiration against flux tower data.
In the revised manuscript, we propose to add a supplementary figure (Figure S2) where we summarise the model/data agreement for the latent heat flux.

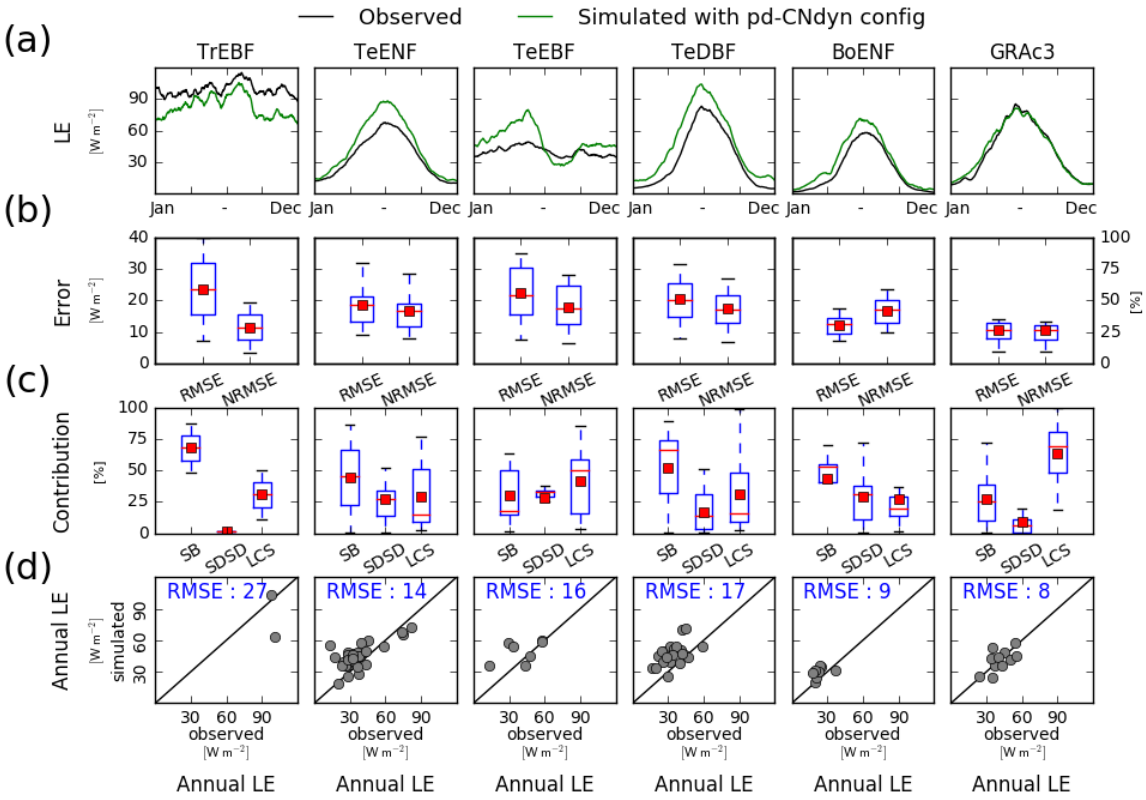


Figure S2 : Site-level evaluation of ORCHIDEE r4999 simulations against Fluxnet observations. (a) Vegetation-class mean seasonal variations of Latent

Heat flux (LE; W m^{-2}), (b) Root Mean Square Error (RMSE; W m^{-2}) and Normalized Root Mean Square Error (NRMSE; %) of simulated daily variations of LE per vegetation class, (c) Attribution of the Mean Square Error (MSE) of the daily variations of LE to model errors on mean value (SB; %), standard deviation (SDSD; %) or correlation (LCS; %) (Kobayashi and Salam, 2000) and (d) simulated vs. observed Annual mean LE at site-level (Wm^{-2}). On panels (b) and (c), the box extends from the lower (25 %) to upper quartile (75 %) values of the data, with a red line at the median and a red square at the arithmetic mean. The whiskers extend from the box to show the range of the data within 1.5 × (25–75 %) data range.

Page 14 line 22 of the initial manuscript, we add the following paragraph for describing the model/data agreement at fluxnet sites for the latent heat flux:
"Because the GPP flux is intimately linked to the transpiration flux through stomatal control, a site-level evaluation of the pd-CNdyn simulation has been performed against site-level observations of the latent heat (LE) flux (Fig. S2), an energy flux to which transpiration contributes to, as does the soil evaporation. Overall, the model performed better at simulating LE variations than variations in GPP. This was particularly true when looking at the NRMSE of the simulated daily flux, which never exceeded 50% as a mean average score at the PFT level for LE, while it went to values up to 75% for GPP for some PFTs (BoENF and GRAc3)."

To this respect, P3L7 is misleading stating that the paper includes a "evaluation of simulated gross carbon uptake and transpiration by plants.". I see no evaluation against transpiration data in the paper.
Although we now add a model evaluation for the the latent heat flux in the manuscript, we prefer removing "transpiration" and keeping only "gross carbon uptake", as it is the key variable we are focus on, in the manuscript. Thus, we rephrased as followed:
"evaluation of simulated gross carbon uptake by plants."

As regards comparison with previous versions of the model: when reading the paper, I cannot evaluate how the model modifications affected the model prediction accuracy. As said above, there are two groups of modifications listed by the authors: group 1(p. 3-4 of the manuscript) seems to be overlooked by the authors, while group 2 (p.5: modifications in the photosynthesis scheme and in the photosynthesis-N coupling) appear more important (i.e. the authors refer to them later in the paper). If the authors think group 2 would significantly impact the simulations, I expect to see a model comparison confronting simulations from a former (e.g. O-CN?) and the current model version. Since two main modifications are mentioned (modification of the photosynthesis scheme and modification of the

photosynthesis-N coupling), I expect to see how both independently impact the model output.

Based on these two points (partial model evaluation against ground-truth data and lack of comparison with model previous versions to evaluate the impact of model modifications), I think the paper in its current version is not ready for publication.

We understood this point that is the summary of the concerns detailed by the reviewer. We hope that we were able to discuss the referee's concerns point-by-point and list the main proposed changes:

- Including model vs data evaluation for GPP, LAI and LE through in-text changes and supplementary figures
- Including model vs model evaluation for GPP, and LAI through in-text changes and supplementary figures

(B) Additional comments:

P2L18, replace "is plentiful" by "is non-limiting provided adequate mineral nutrition in the future,"
We thank the reviewer for this suggestion but modifying the sentence as proposed will be in conflict with the rest of sentence where we state that "it remains questionable whether sufficient nutrients, in particular nitrogen, will be available". We propose the following change:
"Even if atmospheric [CO_2] will be plentiful in the future, it remains questionable whether sufficient nutrients, in particular nitrogen, will be available to fully sustain the increase of primary production associated solely to the rise of [CO_2]."

P2L21, replace "will" by "would"
This will be corrected in the revised manuscript.

P3L7 "thorough"
This will be corrected in the revised manuscript.

P5L30 eq. 2: on which data were the parameters fitted ? On GPP data? These parameters are very sensitive, please be precise.
We will give more details in the revised manuscript. We propose to modify the sentence as follows:
"where $a_{rJ,V}$ and $b_{rJ,V}$ are fitted parameters of the relationship between observation-based values of $J_{max,ref}/Vc_{max,ref}$ and t_{growth}, and equal to 2.59 [-] and -0.035 [$°C^{-1}$], respectively"

P8: How were equations 11 and 12 parameterized? Fitted on which data?

Equations 11 and 12 are empirical functions adapted from Zaehle and Friend (2010), whom parameters have been adjusted to match the shape of the equation 21 of the Supplementary Material of Zaehle and Friend (2010).
We will add this information in the revised manuscript:
"$a_{D_{max}}$ and $b_{D_{max}}$ are two empirical parameters set to 1.6 and 4.1, respectively, in order to best fit the original function (eq. 21 of the Supplementary Material of Zaehle and Friend (2010))."