Reviewer #1 comments

1. Introduction: The introduction does not provide an overview of currently available comparable models and whether such models already fulfill some of the requirements identified from BETHY studies (L54-59). I believe this is needed to fully evaluate the novelty of the D&B model.

We have added three paragraphs of discussion of three further possible candidate models that all have a history of data assimilation, and all can be run as part of general circulation models (weather forecasting, or earth system models), namely C-TESSEL, JULES and ORCHIDEE, as well as introducing DALEC in this context. The discussion follows the list of requirements, and precedes the introduction of the D&B model:

"There are a number of models that could potentially fulfill those requirements. They range from carbon models incorporated into routine weather forecasting, such as C-TESSEL (Boussetta et al 2013), to highly complex land surface and ecosystem models that can be operated both within earth system models or independently, such as JULES (Best et al 2011, Harper et al 2016) or ORCHIDEE (Traoren et al 2014). Of these, C-TESSEL has probably the strongest track-record for assimilation of satellite data, mainly for the purpose of constraining soil moisture (Scipal et al 2008). However, it does not simulate the mass balance of carbon, despite simulating photosynthesis and respiration, nor can it predict leaf area index, which it requires as input data. C-TESSEL is therefore of limited use when assimilating FAPAR, or variables related to biomass.

By contrast, JULES includes a full set of carbon fluxes and pools (Clark et al 2011). An adjoint version of JULES has been developed to optimise parameters at site level using eddy flux data (Raoult et al 2016). ORCHIDEE includes not only carbon but also nitrogen cycling (Vuichard et al 2019). A data-assimilation framework for ORCHIDEE also exists, which has been successfully employed at site level for the step-wise optimisation of model parameters using remote sensing data (e.g. FAPAR), as well as water and carbon flux observations from the eddy covariance flux networks (Peylin et al 2016).

We note that less complex models such as C-TESSEL are often much better suited for data assimilation than highly complex models, because a simpler structure with fewer parameters, omitting processes not relevant at the time scales of interest, makes optimisation both computationally and mathematically much more more feasible. For example, C-TESSEL and BETHY lack representation of carbon pools (except for leaf area in the case of BETHY) due to a focus on short time scales of up to a few years. This is contrasted by another model, DALEC (Data Assimilation Linked Ecosystem Carbon), which focuses on carbon pools and longer-term processes, but is structurally also simple. DALEC has been developed specifically for assimilating information on C fluxes and pools from satellite observations (Bloom et al 2015), eddy covariance systems (Bloom et al 2015, Famiglietti et al 2021), and biometric data including biomass (Smallman et al 2017, Quegan et al 2019). "

2. L22-24: While I agree with the authors' statement on the reliable characterization of carbon and water fluxes, I think that their justification "we currently lack a robust, spatially and temporally

explicit knowledge of sources and sinks of CO2, or of the drivers of those variations" could be more concise. It is not clear if the authors refer to sources and sinks of CO2 in general or which is what I would expect considering the scope of the article, only those related to vegetation dynamics.

We have amended the statement as follows:

"as we currently lack a robust, spatially and temporally explicit knowledge of the sources and sinks of CO$_2$ ***within the terrestrial biosphere****"*

Additionally, the statement should be underpinned by an overview of the current knowledge. Specifically, it would strengthen their statement to provide information on the following:
- The spatial and temporal resolution and coverage of state of the art models and data products.
- An overview of known and suspected drivers.

We have have attempted to clarify by adding the following discussion:

"The lack of knowledge exists despite of the availability of products of net or gross carbon uptake by terrestrial vegetation, such as those from MODIS with daily and up to 250~m resolution (Zhao eta 2005), or from the Copernicus Global Land Service with a 300~m spatial and 10-day temporal resolution (Swinnen et al 2021). One issue is that those products are no direct observations of carbon fluxes, but rather a combination of remotely sensed information and a set of model assumptions. They thus do not necessarily agree with each other or with the results of ecosystem models (Turner et al 2006, Sun et al 2021). Another issue is that we lack spatially distributed data sets of terrestrial biosphere CO2 sources.

However, in order to identify the drivers of terrestrial carbon sources and sinks, such as vegetation state, soil carbon content of different qualities, temperature, soil moisture, atmospheric humidity, or light availability, we need models that are thoroughly evaluated against reliable observations. If we also want to identify existing carbon sources and sinks and attribute those to certain drivers and processes, we also need to be able to run and evaluate those models at the spatial and temporal resolution of interest. Running models at high spatial and temporal resolution is not an issue in principle, as the model-based data sets referred to above demonstrate. The problem lies in finding suitable observations at high temporal and spatial resolution for terrestrial ecosystem model evaluation and in finding out which model formulations, initial conditions and paramerisations can reproduce those observations. "

3. L28-35: This seems to be a crucial section motivating the development of D&B. However, I do not understand how the need to include primary earth observation data in data assimilation frameworks provides added value for indirect or secondary earth observation data products (L30-35).

As stated already, the issue is that few remote sensing products refer directly to the quantity of interest, i.e. CO2 sources or sinks. Using remote sensing products only indirectly related has then two subsequent purposes: First, it gives us the opportunity to evaluate models and to ensure the produce *the observations as reliably as possible. Second, we can use those model simulations that best reproduce the available observational data to derive what the reviewer calls "secondary earth observation data product".*

The first is explained at the end of the added discussion responding to the previous comment. To explain the second, we have added the following text to the paragraph:

"Data assimilation offers a valuable tool for automatically finding the optimal combination of model initial values, parameters and even input quantities given the observations assimilated, pertinent to certain assumptions about prior values and uncertainties of models and data (Tarantola 2005). While not providing a ready made answer -- it always needs to be assured that the thus optimised model simulations "make sense" -- it can be used to find the model more reliable model and data based estimates of quantities of interest, e.g. carbon fluxes, and serve as tool for evaluating assumptions about the inherent processes driving changes in those fluxes."

4. Section 2.3: I am missing how DALEC and therefore D&B constrain the ratios between the biomass of different plant organs (e.g. leaf to root ratio) and also if allometric relationships are considered somehow. Is this handled implicitly via the regional-scale calibration? This should be included in the model description and potentially the discussion of model limitations."

We explain in the text that "NPP is allocated to each of the four live biomass pools based on fixed site of PFT-specific fractions". Yes, the regional-scale calibration handles the estimation of both allocation ratios and organ lifespans that determines biomass dynamics. However we have not explained how in CARDAMOM we use 'ecological and dynamical constraints' to ensure that allometric relationships (like root:shoot ratio) are kept within ecological realistic bounds.

We have now added the following text to Section 2.3:

"Parameters for the C cycle in D&B use PFT calibrations for DALEC derived using the CARDAMOM model-data fusion approach (Bloom and Williams 2015). CARDAMOM uses ecological and dynamical constraints to ensure that allometric relationships arising from parameter selection (like emergent root:shoot ratios) are kept within ecologically realistic bounds. By calibrating DALEC using both LAI and woody biomass data, a constraint is placed on relevant model parameters to match the measured biomass of these plant organs."

5. Section 4.2: The section is missing the description of the initial conditions for vegetation biomass and soil carbon. If these do not need to be initialized how are initial values determined? Generally, an overview table in the main text or SI containing all variables for which initial conditions need to be prescribed and the respective initial values would in my opinion improve the clarity of the model setup description.

Additionally, the description jumps between sites which reduces reading flow. I would kindly ask the authors to describe the sites after each other.

The initial conditions of the biomass pools are shown in Tables 6 and 7 of the SI, Secton 1.4 "Model setup". We agree that we should have mentioned this in here in the main text.

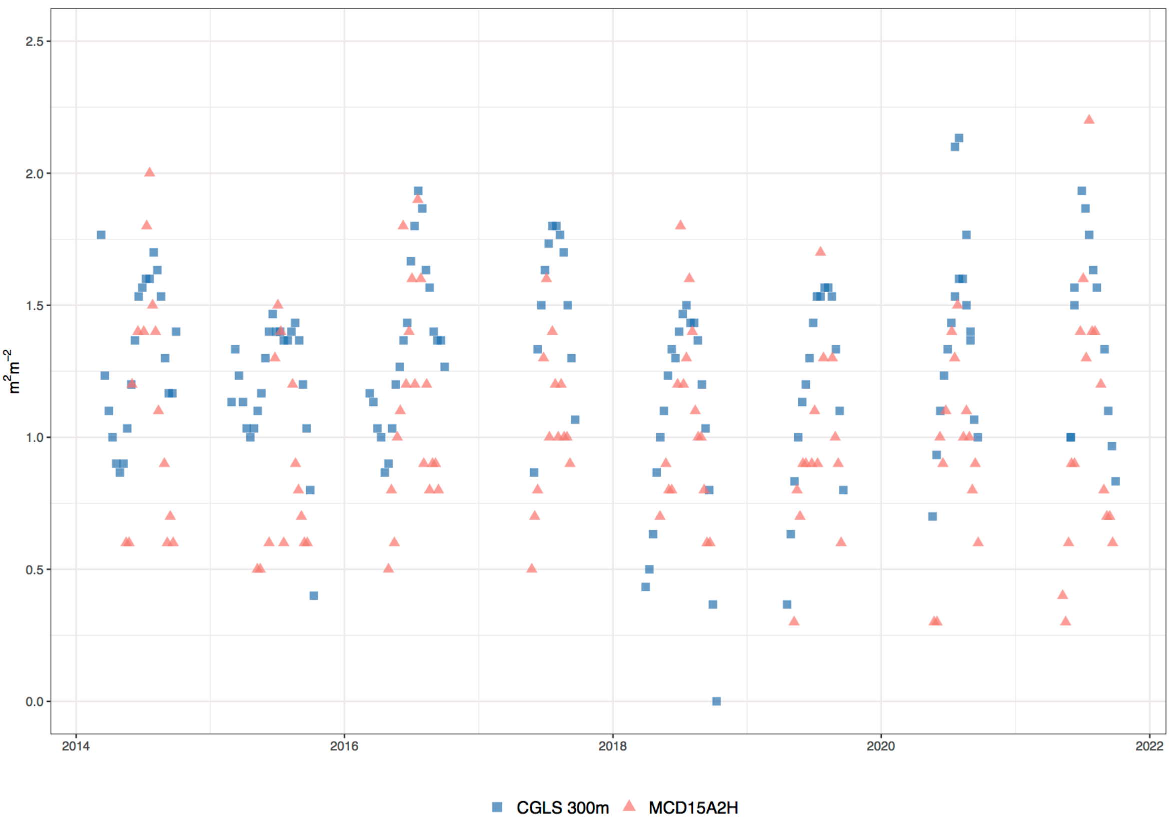
We have restructured Section 4.2 such that in the first paragraph, settings applicable to both sites are described, and then dedicate one paragraph each for each site's particular setup. We also now refer to the tables in the SI that contain the initial conditions.

6. Section 4.3: The description of the evaluation metrics is not detailed enough. For example it is not clear what is meant by multi-year averages of the annual cycle and the multi-year mean and how these differ. From following sections it becomes clear the second is the average of the annual sum but the explanation is missing. The authors should also provide equations for all evaluation metrics.

We have added equations for all metrics.

7. Section 5.1 and 6.3: The authors explain the intra-annual fluctuations of FAPAR by the LAI seasonality (L414f). However, I would expect that the two PFTs evergreen coniferous forest and evergreen shrub should not have a strong LAI seasonality which is confirmed by the observations (L423f). The authors briefly discuss this in the limitations section and relate it to phenology but do not provide a detailed explanation of model behavior (L523-528). I understand that the authors cannot provide a calibrated version of D&B at this point but would like to see how their results relate to eq. 147-149 to fully explain this behavior.

We agree that these two PFTs are not expected to have a strong LAI seasonality. The evergreen coniferous forest and evergreen shrub PFT calibrations for the DALEC component were constrained with seasonal cycles of series of leaf area index (Copernicus Service Information, 2020) from the study domains. The Copernicus product includes a strong seasonal cycle of LAI across northern latitudes, which is unexpected and at odds with in situ knowledge of the ecosystems. Therefore the calibrations of foliage dynamics in the model reproduce the observed cycle from Copernicus data. The outcome is a model calibration which is at odds with the FAPAR product used in the evaluation.



We have included a figure showing the MODIS MCD15A2H LAI product (https://lpdaac.usgs.gov/products/mcd15a2hv061/), which uses observations from both the Terra and Aqua satellites as well as the Copernicus LAI product "Copernicus Global Land Service (CGLS) Collection 300m LAI" (product page: https://land.copernicus.eu/en/products/vegetation/leaf-area-index-300m-v1.0).

In SI Table 6 the relationship with model eq. 147-149 is clarified by the parameter estimates. Particularly see parameter 'clf', which is the reciprocal of the annual leaf loss fraction (and so is proportional to leaf lifespan). Our calibration generated 'clf' values between 1.0-1.5, which represents deciduous ecosystems. Evergreen systems with multi-year needles would have clf>2. We are currently developing processes to correct the bias introduced by the highly seasonal LAI products used in model calibration.

In section 5.1 we now state:

"The level values of the observed FAPAR match the expected behaviour of the largely aseasonal evergreen canopies of the PFTs for the boreal region. The pronounced seasonal cycle of FAPAR in the model runs corresponds to a seasonal cycle in the LAI of the model. The modelled LAI behaviour results from calibration using Copernicus LAI time series which have a strong (and unexpected)

seasonality."

Section 6.3 now states

"While the initial task to match and compare modelled and observed data streams was successfully demonstrated, the results of this study also point at the need to further investigate the representation of the seasonal cycle of LAI in northern evergreen conifer forests and shrubs. Earth observation products for the boreal region show seasonality in LAI that is not consistent with ecological expectation and FAPAR data. The phenology scheme of D\&B has the flexibility to simulate vegetation with a low seasonal variation in LAI, if corresponding information is provided for the prior calibration of the parameters in the phenology scheme. Such information could come from field observations of LAI time series in boreal regions."

In the supplement, the following corrections were made:

"Losses from the foliar pool are linked to specific periods in the annual cycle through parameters for the day of leaf fall (d_fall) and ***period of leaf fall",*** *instead of "labile release"*

"c_lf is the ***reciprocal of*** *annual leaf loss fraction,* ***and so is proportional to*** *leaf life span" instead of "related to"*

8. L541-548: This paragraph is quiet generic and in my opinion applies to process-based models in general. It could in my opinion be extended to highlight how this is different for D&B.

We have added the following sentence:

"The potential advantage of D&B coupled to multiple observation operators is that it allows model testing via multiple data streams, thus providing are more comprehensive model evaluation which makes it less likely the model matches observations while misrepresenting important processes."

In addition, we added that stastical and machine learning models are used as black boxes so the question of being right for the wrong reasons does not even occur.

Minor comments:

1. L25f: The sentence contains some small language issues:

- I suggest to change "terrestrial carbon stores" to "terrestrial carbon storage" because stores has multiple meanings.

- Unclear what "those variations" are. I assume variations in C fluxes and storage but it should be clarified.

-Unclear what is meant by "forcing factors". I assume changing climatic conditions (e.g. temperature change and so on) but it should be defined.

We have replaced "stores" by "pools", and "those variations work and interact" with "variations in carbon fluxes interact".

We have added "(such as climate, land use, \coz fertilisation)".

2. - 7. *These minor comments have been followed as suggested*

8. L150-157: I have several minor issues with understanding:

- Is root water supply capped at field capacity? I assume yes but it is not stated.

- "Actual stomatal conductance are then set such that transpiration is capped at the root supply rate": First, it should be "root water supply rate" not "root supply rate". Second, I would assume that it is capped at the minimum of root water supply rate and demand for transpiration. Can you splease confirm and elaborate this.

- Could you add a reference to the equations of the supply-demand calculation to make it easier to find.

We added ", reaching a maximum with soil water at field capacity"

The sentence has been modified to "Actual photosynthesis and stomatal conductance are then set such that transpiration is downregulated from its potential rate to the rate of maximum root water supply." We have also added a reference to the SI, where the equations and the reference can be found.

9. Section 2.2.: Variable names are introduced in some parts of the sections (e.g. L135-145) but not in the entire section. I think this should either be consistent or the authors should explain why they introduce certain variables in the main text and others only in the SI.

We agree, this paragraph sticks out as listing many variable names that are not later used in the main text. As suggested, we have removed non-esssential introduction of variable names here for consistency.

10. - 12. *These minor comments have been followed as suggested*

13. L193 "ensure" instead of "insure". Also I do not understand where the "separate calculation of FAPAR at the correct solar zenith angle" has to be performed. Is it within the model or is it a correction of data from observations. If the first, what is the difference to using FAPAR calculations from the model run?

We have added: "The former requires a separate FAPAR calculation but produces slightly more accurate results."

14. L261: Delete "bptj" and please elaborate how the parameters were chosen. Was this part of the calibration or an expert assessment?

We have replaced "chosen" by "calibrated such that the model reproduces".

15. *Done as suggested.*

16. L303, 442 and other occurrences: Here the authors refer to Sodankylä as the boreal site deviating from their so far consistent terminology. Similarly they sometimes refer to Majadas de Tietar as the savannah site.

We have kept switching between both, as this is explained in Section

17.-22. *Done as suggested.*

23. L373ff: This sentence is quiet long and I do not fully understand its meaning. E.g. "[...] measured NEE [...] that are not reproduced by the measurements". I believe one of these should refer to simulations and not measurements.

Split into two sentences and dhanged to "… are not reproduced by the model. Such fluctuations are also found in the observation-derived TER flux."

24. L383f: What is the reason for the overestimation under favorable conditions?

We added ", while overestimating photosynthetic capacity."

25. *Done as suggested.*

26. L391: This is true for GPP (2.11 vs 2.25 modelled and 3.39 observed) but not for NEE (1.88 vs -0.09 modelled and -0.05 observed). Neither the order of magnitude is similar nor is the difference to the model mean smaller. I also do not understand the significance of this results. Please elaborate.

We are sorry, this should have read "RMSE of GPP and TER …". We have also modified the last sentence of this section to: "While the high r^2 suggests that the model reproduces the interannual variability of the net carbon fluxes well for this site, the rather high RMSE suggests that daily-to-day variations are less well captured."

27. Section 6: When referring to the results references to the respective sections are missing.

We did so in order to interrupt the flow of the text. We believe it is easy enough to identify the relevant sections.

28. *Done as suggested.*

29. L565f: I am not sure that this can be generalized. The process model may also not be able to match observations for a specific variable within reasonable bounds of the parameter space if the process is implemented but its formulation is not universal and therefore not applicable to the context of the experiment. So you cannot say that a process is missing but only that either a process is missing

or the formulation of processes used is not suitable.

We have changed “by missing process” to “by missing or unsuitable process representation”.

30., 31. *Done as suggested.*

32. Fig. 3, 4, 6 and 7: You could consider adding the 5th to 95th percentile values to illustrate inter-annual fluctuations. This would in my opinion also underpin your results where you compare inter- and intra-annual match between observed and simulated data.

As there are only six years, we have instead plotted the highest and lowest value found for each specific day of year within the time series.

33. Fig. 4, 5, 7, 12, 13, 14 and 15: Captions are missing the color scale.

We are not sure we understand, but the meaning of the colors are explained by the legends shown in the figures.

34. *Done as suggested.*

35. SI Title of section 1.2.4 only refers to evaporation but section describes also transpiration.

Title changed to “Evapotranspiration from vegetation”. Thank you for spotting this.