
Report #1 on "Evaluation of terrestrial pan-Arctic carbon cycling using a data-assimilation system" by Efrén López-Blanco et al.

M. Forkel (Referee)
matthias.forkel@geo.tuwien.ac.at
Submitted on 19 Nov 2018

The authors substantially revised the manuscript and addressed my comments appropriately.
I only disagree how uncertainties where used during the data assimilation. Specifically, the authors state at several places (lines 86, 184-85) that the (biomass) dataset lacks uncertainty or error estimates and hence they used a global uncertainty factor of 1.5 in the cost function.
It is clearly a wrong statement that the biomass maps by Carvalhais et al. (2014) miss uncertainty estimates.
In this dataset, uncertainty was provided based on an ensemble of biomass estimates. This biomass map is also based on the map of forest biomass by Thurner et al. (2014) which also includes a detailed estimate of uncertainties for various vegetation carbon pools.
Please remove the wrong statements about missing uncertainty estimates for the biomass datasets and describe why you did not use these uncertainty estimates or how a potential use could affect your results. With these changes, I'm happy to accept the manuscript for publication.

We apologise for the lack of clarity about uncertainty derivation for the analysis. Here we have adjusted the text on the Introduction section (S1) to remove the sentence

> "However, these products tend to lack clear error estimates."

On S2.2.2, L188-195 we have adjusted the text to:

> "The reported uncertainty on biomass data from Thurner et al. (2014) was +/- 37% at pixel scale. Because of undetermined errors related to tree cover thresholds used in the upscaling, and to reflect unknown model structural error, we slightly inflate the error estimate and use a log-transform(1.5) of ×/÷1.5 (i.e. ×/÷1.5 spans 67% of the expected error). We use the same proportional error for SOC. For MODIS LAI we inflate the proportional error further to log(2) based on well reported biases in this product for evergreen forests (De Kauwe et al. 2011) and the estimated measurement and aggregation uncertainty for boreal forest LAI of 1 m^2 m^{-2} reported by Goulden et al. (2011). The uncertainty assumptions in expression 3 are chosen in lack of better knowledge about the combined uncertainties arising from model representation errors and observation errors:"

In the Discussion we also now review the challenges associated with generating observation and model errors (see response to reviewer 2).

Report #2 on "Evaluation of terrestrial pan-Arctic carbon cycling using a data-assimilation system" by Efrén López-Blanco et al.

Anonymous Referee #2

Received and published: 06 Feb 2019

I'm going to be upfront that I'm very torn about what to recommend with respect to this paper. On the one hand, I acknowledge the incredible amount of work that went into this project and believe that there is important and interesting science coming out of this project. On the other hand, based on the responses to questions raised, it is now clear there are definitely things here that I don't think were done correctly. What complicates this is that many of the things done wrong (especially with respect to model process error) were also done wrong in previous papers on the Bayesian calibration of terrestrial carbon models (both by this team and others). This helps explain such mistakes, but it doesn't justify them, and I worry that continuing to allow papers to make the same mistakes just perpetuates the situation. The crux of the issue is really in how the authors are treating the error term in their likelihood. First, they are ascribing 100% of the error as coming from the observations, and not acknowledging (statistically) that their model is imperfect (though their own Results and Discussion clearly demonstrate that the model is far from perfect). By incorrectly ascribing 100% of the error to observations, and none to process error (model misspecification, stochastic events, unaccounted for heterogeneity), the authors are also missing that (unlike observation error) process error propagates forward into model predictions. This means that modeled fluxes and pools are going to be consistently overconfident by an unknown (but potentially nontrivial) amount. Second, not only do the author ascribe all the error to observations, but they treat that observation error as a known parameter, despite acknowledging that the data products used don't have error estimates. This is a significant departure from standard statistical modeling, where the variance is an unknown fit parameter. For example, when you fit a linear regression the model has three unknown parameters (slope, intercept, sigma) and sigma is virtually never treated as an a prior known quantity. While treating sigma as a known shouldn't have large effects on the mean values of the model parameters (though this is far from guaranteed when dealing with nonlinear models; Jensen's Inequality), more important is that it can have a real effect on the uncertainties about the model parameters. By subjectively choosing the observation error, one is also subjectively choosing the confidence intervals on the parameters. And since in CARDAMOM the only uncertainties that are included in predictions are parameter uncertainties, this also means you are subjectively choosing the uncertainty in the predictive confidence intervals. Ideally, these models should be refit including an unknown, fit model process error, and then that process error should be propagated into predictions/hindcasts. This process error ideally should also be in addition to, not instead of, an observation error (which may not be a known, but may have an informative prior on it)

We recognise the reviewer's concerns about using the correct process for error characterisation in analyses such as that we present here. We agree that our model is not perfect and that identification of process error is critical. We also regret that we did not provide the necessary information on how data uncertainties were derived. We do appreciate the reviewer's concern

about effective error characterisation, and have adjusted the text to reflect this, and to make recommendations about how to address this better.

We did specifically focus on identification of model process error by comparison with independent data (GPP, R_h). Thus, we identified biases in our estimates of LAI, GPP and biomass at landscape scale, and suggest that these likely reflect systematic bias in our photosynthesis model. A next step is to analyse the representation of photosynthesis process error and include this in further analyses. On the other hand, we note that independent evaluation of fluxes at site scale (FLUXNET2015) does not match the GPP bias at landscape pixel scale (FLUXCOM). New site level comparisons (see below) also suggest CARDAMOM produces reasonable or slightly high biased results. We conclude that further investigations into heterogeneity error are required, linked to process error calculation on products such as FLUXCOM as well as our GPP model.

We have adjusted the text (S2.2.2, L188-195) to clearly state that error in the biomass product is reported, and have explained why we have inflated this error in our analysis. We also note that MODIS LAI products have large reported biases, and local observations have important errors, which justifies the larger error we assigned to these data. Our point here is to report an honest overview of uncertainty assumptions used in CARDAMOM:

> "The reported uncertainty on biomass data from Thurner et al. (2014) was +/- 37% at pixel scale. Because of undetermined errors related to tree cover thresholds used in the upscaling, and to reflect unknown model structural error, we slightly inflate the error estimate and use a log-transform(1.5) of ×/÷1.5 (i.e. ×/÷1.5 spans 67% of the expected error). We use the same proportional error for SOC. For MODIS LAI we inflate the proportional error further to log(2) based on well reported biases in this product for evergreen forests (De Kauwe et al. 2011) and the estimated measurement and aggregation uncertainty for boreal forest LAI of 1 m^2 m^{-2} reported by Goulden et al. (2011). The uncertainty assumptions in expression 3 are chosen in lack of better knowledge about the combined uncertainties arising from model representation errors and observation errors:"

We note the reviewer's concerns about making forecasts without properly accounting for model process error. This paper involves an analysis of historical fluxes constrained by contemporary forcing and data. We do not make forecasts or hindcasts, so this criticism is not relevant for this paper.

We have adjusted the text in the discussion (S4.3; L487-501) to reflect the lack of robust knowledge on the interactions between random and systematic biases in the observations, model representation errors and errors in the model drivers:

> "Our approach has used estimated observation error, and inflated this to include unknown errors associated with model process representation. We currently lack any better knowledge of the combined uncertainties arising from model representation errors and observation errors. We acknowledge that all models are an imperfect representation of C dynamics, which generates irreconcilable model-data errors due to the inherent assumptions in model structure. Future analyses should investigate model structural error, using for example error-explicit Bayesian approaches (Xu et al., 2017), or comparing the likelihoods of alternate model structures, of varying complexity. Using multiple sources of data, we have highlighted systematic errors in the model at

landscape scale (Figure 2 and 3) for LAI, GPP and biomass. However, these biases are not consistent for site-scale evaluations. Thus, a next step would be to include explicitly both random and systematic process errors for C fluxes in the data assimilation. These errors could be determined from field scale evaluation of model process representation (Table 2) using e.g. FLUXNET2015 data. We also need to understand better the error associated with landscape heterogeneity of C stocks and fluxes, to upscale from flux tower observations, or direct measurements of LAI, to landscape pixel. This could be achieved by constructing robust observation error models (Dietze, 2017) from field to pixel scale, for e.g. GPP, LAI and foliar N. Evaluation of the sensitivity of C cycling DA analyses to observation error has shown relatively low sensitivity to data gaps and random error on net ecosystem flux data (Hill et al., 2012), but further analyses of error sensitivity are required for multiple streams of stock data."

Additional points of concern:

1) Neither the DALEC2 model nor the CARDAMOM system appear to be publically archived. This means this work can't be reproduced or expanded upon by others. I don't know if such lack of openness is within the letter of the law of this journal, but it's definitely a deviation from the current norms of the community.

We agree that openness is critical to scientific advances. We have submitted the code for DALEC2 on Edinburgh DataShare. We are working to release a community version of CARDAMOM. At present we invite researchers to contact us to gain access to the code.

We have adjusted the text (L517-520):

"**Data and software availability**

CARDAMOM output used in this study is available from Exbrayat and Williams (2018) from the University of Edinburgh's DataShare service at https://doi.org/10.7488/ds/2334. The DALEC2 code is also available on Edinburgh DataShare at https://doi.org/10.7488/ds/2504. Contact MW for access to the CARDAMOM software."

2) As noted in my original review, I'm not comfortable with this system being called data assimilation, at least not with some additional qualifier being added (e.g. "parameter data assimilation") to make it clear that the outputs are deterministic model forward simulations not a reanalysis. To me, calling this data assimilation is like calling linear regression "machine learning." Sure people do it, but it makes the term pretty meaningless.

We disagree; we are using Bayesian parameter calibration of a dynamic model - which is typically referred to as data assimilation or model-data fusion; see "Ecological Forecasting" p. 168, by M. Dietze. However, we adjust our introductory text to improve clarity (S1; L100-104):

"To address these issues we integrate model and data more formally. We apply data assimilation (DA), defined as a Bayesian calibration process for a model of a dynamic

system. DA, through probabilistic parameterisation, supports robust model estimates of C stocks and fluxes consistent with multiple observations and their errors (Fox et al., 2009; Luo et al., 2009; Williams et al., 2005). By following Bayesian methods, the uncertainty on observations weights the degree of data constraint, and the outcome is a set of acceptable parameterisations for a given model structure linked to likelihoods."

3) After clearly diagnosing your photosynthesis scheme (ACM) as being at the root of model biases and compensating errors, the decision to not include any ACM parameters in the calibration (and toss the issue up to a lack of acclimation rather than simple miscalibration) strikes me as odd and I cannot understand why the authors are digging in their heels on this.

We do include an ACM parameter (C_{eff}) in the calibration (and so it is adjusted by the MHMCMC), according to Bloom et al. (2016). We apologise for not making this clear. We consequently have adjusted the Methods text (S2.2.1; L143-145) to read:

> "DALEC2 simulates canopy-level GPP via the Aggregated Canopy Model (ACM; Williams et al., 1997) and the most sensitive ACM parameter, related to canopy photosynthetic efficiency, is included in the CARDAMOM calibration."

4) Similar to (3), since NPP in DALEC is very tightly tied to GPP, and TT = Cstock/NPP, it sure seems like systematic biases in GPP will translate to systematic biases in TT. As noted earlier, I find some of the reported TT estimates to be implausible and don't understand the authors resistance to even considering comparing their results to independent field estimates.

We note that the mean NPP for GVMs across the region is 8% lower than in CARDAMOM, so the regional GVM-CARDAMOM NPP analyses are less different on average than the comparisons of CARDAMOM against data such as FLUXCOM (for GPP). We note that the high latitude TT estimates for CARDAMOM, GVMs (Figure 5) and reported in Carvalhais et al. (2014) are broadly similar. The critical issue we identify is that the spatial differences in NPP and C_{veg} between CARDAMOM and GVMs result in important spatial mismatches in TT estimated by both (compare Figure 5 and Figure 6).

We are confused at the statement that we have "**resistance to even considering comparing their results to independent field estimates**"; we have presented a clear evaluation against multiple independent FLUXNET site data, shown in Figure 4. Nonetheless, we add some further field-based estimates to complement these comparisons in the Discussion (S4.1, L421-435):

> "For a further independent evaluation of CARDAMOM outputs, we compare the tundra and boreal estimates to plot scale flux and stock information. For tundra, Street et al. (2012) calculate growing season GPP estimates of 263-380 g C m^{-2} for *Empetrum nigrum* communities, and 295-386 g C m^{-2} for *Betula nana* communities, which is consistent with the ranges in Figure 1 for tundra. Biomass stocks for Arctic tundra recorded in the Arctic LTER at Toolik Lake range from 105-1160 g C m^{-2} (Hobbie and Kling, 2014), which are consistent with the estimates from CARDAMOM, albeit at the lower end of the model estimates. For boreal forests, Goulden et al. (2011) report annual GPP estimates across a chronosequence of stands, and thus a variation across canopy densities, which varied from 450-720 g C m^{-2} yr^{-1}. These data are consistent with the

span of GPP in CARDAMOM (Figure 1), again best matching the lower end of the model estimates. For the same study, the vegetation C stock estimates varied from 100-5000 g C m^{-2}, consistent with CARDAMOM, and with measurements of 10 to 40-year old boreal stands best matching the CARDAMOM median estimate of ~1500 g C m^{-2}. We conclude from comparisons against site data that CARDAMOM analyses are broadly consistent, with some tendency for CARDAMOM to have a high bias. This comparison is similar to the FLUXNET2015 evaluation of CARDAMOM. But it conflicts with the estimation of low bias from the comparison of CARDAMOM against FLUXCOM GPP and Carvalhais et al. (2014) biomass stock maps. It is possible that the scale differences between site level products and landscape estimates is confusing these comparisons, but there is clearly a need to understand better these inconsistencies in C cycle estimates."

5) The differences between DALEC and observations are greater than the differences between DALEC and the ISI-MIP models, so why are the authors so hard on the ISI-MIP models?

Our key point is that DALEC outputs match the spatial variation in independent (FLUXCOM) and assimilated data (LAI, biomass) well. There may be biases in these comparisons, indicative of model process error and/or upscaling error in the biomass and FLUXCOM products, but CARDAMOM can match the pattern in LAI, biomass, and SOC very well (Figure 2). The poor agreement with ISI-MIP models is with the spatial pattern (Table 3), not with regional median values (Figure 5). From these analyses we note that a reasonable regional estimate is not very useful if patterns are wrong, as this challenges the reliability of ISIMIP models when used for projections. Some models actually match CARDAMOM well, and we noted this clearly. We have edited the text to emphasise these points:

In Results (S3.4, L318-330):

"We used our highest confidence retrievals of NPP, C_{veg} and TT_{veg} (i.e. retrievals including assimilated LAI, biomass and SOC) to benchmark the performance of the GVMs from the ISI-MIP2a project. In this assessment we compared not only their spatial variability across the pan-Arctic, tundra and taiga region (Figure 5), but also the degree of agreement between their mean model ensemble within the 90% confidence interval of our assimilation framework (Figure 6, Table 3). NPP estimates (RMSE = 0.1 kg C m^{-2} yr^{-1}; R^2= 0.44) are in better agreement than C_{veg} (RMSE = 1.8 kg C m^{-2}; R^2= 0.22) and TT_{veg} (RMSE = 4.1 years; R^2= 0.12). The assessed GVMs estimated on average 8% lower NPP, 16% higher C_{veg} and 22% longer TT_{veg} than CARDAMOM across the entire pan-Arctic domain (Figure 5 and 6) on average. Thus, at regional aggregation CARDAMOM analyses agreed more closely with ISI-MIP2a models than with FLUXCOM (51% difference) and with the Carvalhais et al. (2014) biomass data (28% bias).

The poor spatial agreement regarding TT_{veg} between CARDAMOM and ISI-MIP2a (Table 3) is indicative of uncertainties in the internal C dynamics of these models. For instance, the slopes in Table 3 are steep and the R^2 are poor – so there is a substantial disagreement in the spatial pattern, not just a large bias. For ISI-MIP2a comparison R^2 values ranged from 0.03-0.52 for NPP; 0.00-0.31 for C_{veg}; and 0.00-0.24 for TT_{veg}."

In Discussion (S4.3, L449-451):

> "Using CARDAMOM as a benchmarking tool for six GVMs we found disagreements that varied among models for spatial estimates of NPP, C_{veg} and TT_{veg} across the Pan-Arctic (Figure 6) in comparison against CARDAMOM confidence intervals."

Detailed comments:

L60: The authors responses suggested that a more complex calculation of TT was actually performed that relaxed the assumption of steady state. I would include that here (along with the steady state calculation) as I suspect a number of readers (myself included) would prefer to know that you're not relying on a steady state assumption to assess a system that's clearly not in steady state.

The residence time is calculated as per Bloom et al. (2016) equation S8 (SI text, S3 Global State and Process Variables), which specifically accounts for changes in stocks over time. We now adjust the text accordingly in the Introduction (S1) by removing "at steady state" and the Methods (S2.2.2; L202-203):

> "We calculate the transit time for C pools using the approach for non-steady state pools described in Bloom et al. (2016), supplementary information S3."

L160: This line refers to DALEC2 as an 'intermediate complexity' model, but later arguments actually hinge on it being a simple model, and most of us would consider DALEC to really be on the simple end of the process-model spectrum

We have had internal discussions about where on the spectrum of complexity DALEC lies. We have decided that simple models would have only a handful of parameters and few state variables. DALEC has 17 parameters and 6 state variables, so it just qualifies as intermediate. We agree that this is partially a subjective categorisation (now in S2.2.2; L157). We also changed wording in L110, L447, L474, and L478 to keep consistency across the full text.

L171: MODIS LAI reports an uncertainty estimate. How did you aggregate those uncertainties when aggregated the observations? This is nontrivial as neither the MODIS products or MODIS LAI validation papers report anything about the spatial or temporal autocorrelation in the product's errors.

We have adjusted our text to report on MODIS uncertainties (S2.2.2, L191-193):

> "For MODIS LAI we inflate the proportional error further to log(2) based on well reported biases in this product for evergreen forests (De Kauwe et al. 2011) and the estimated measurement and aggregation uncertainty for boreal forest LAI of 1 m^2 m^{-2} reported by Goulden et al. (2011)."

We have also adjusted the discussion to note the challenge for scaling these errors (S4.3, L496-499):

> **"**We also need to understand better the error associated with landscape heterogeneity of C stocks and fluxes, to upscale from flux tower observations, or direct measurements

of LAI, to landscape pixel. This could be achieved by constructing robust observation error models (Dietze, 2017) from field to pixel scale, for e.g. GPP, LAI and foliar N.**"**

L188: Table S2 looks like it just contains a bunch of uniform priors for all other parameters. I think that should be stated here so that readers don't need to find the supplement to learn that. It's perfectly fair, however, to make readers go to the supplement to see the exact numerical values of the priors.

We now include a note (S2.2.1, L143):

> **"**(Table S2; most priors are uniform with broad ranges)**"**

Moreover, we corrected a mistake with C pools units in Table S2. We replaced g C m^{-2} yr with g C m^{-2}.

L194: This sentence states that MODIS doesn't report an uncertainty estimate, but that's not accurate.

The cited statement was removed and we have adjusted (see above) our text to report on MODIS uncertainties (S2.2.2, L191-193):

> "For MODIS LAI we inflate the proportional error further to log(2) based on well reported biases in this product for evergreen forests (De Kauwe et al. 2011) and the estimated measurement and aggregation uncertainty for boreal forest LAI of 1 m^2 m^{-2} reported by Goulden et al. (2011)."

L206: I'm concerned about the way the statistics are being reported here. For example, the RMSE of a model is traditionally based on the model error (difference between the model and the observations). Here, the authors are defining the model's RMSE as the RMSE after applying both a multiplicative and additive bias correction (i.e. the predicted/observed regression). Similarly, the R2 isn't the variance explained by the model, but the variance jointly explained by the model and a linear bias correction to that model. This results in a very optimistic view of the model's actual performance.

We have calculated RMSE following the traditional approach, and we have adjusted the text to clarify this (S2.3, L207-209):

> **"**To assess the degree of statistical agreement we calculated linear goodness-of-fit (slope, intercept, R^2) between CARDAMOM and the two independent datasets and determined RMSE and bias from direct comparison on model-data residuals.**"**

Following the same logic, we have also clarified this in S2.3, L221-223:

> "We performed a point-to-grid cell comparison to assess the degree of agreement between each flux magnitude and seasonality calculating the statistics of linear fit (slope, intercept, R^2) per flux and site between CARDAMOM and FLUXNET2015 datasets and determined RMSE and bias from model-data residuals comparison."

L251: Just want to continue to express my skepticism about some of these pool and flux estimates. For example, in my own experiences in Alaska, the boreal forest has WAY

more than 160% more structural tissue than the tundra. There needs to be some independent plot-scale validation of this.

Independent data from Toolik Lake (tundra) and Boreas (boreal) sites shows the general validity of the CARDAMOM outputs at these intensively studied ecological field sites.

As we presented earlier on, we included the following text in S4.1, L421-430:

> "For a further independent evaluation of CARDAMOM outputs, we compare the tundra and boreal estimates to plot scale flux and stock information. For tundra, Street et al. (2012) calculate growing season GPP estimates of 263-380 g C m^{-2} for *Empetrum nigrum* communities, and 295-386 g C m^{-2} for *Betula nana* communities, which is consistent with the ranges in Figure 1 for tundra. Biomass stocks for Arctic tundra recorded in the Arctic LTER at Toolik Lake range from 105-1160 g C m^{-2} (Hobbie and Kling, 2014), which are consistent with the estimates from CARDAMOM, albeit at the lower end of the model estimates. For boreal forests, Goulden et al. (2011) report annual GPP estimates across a chronosequence of stands, and thus a variation across canopy densities, which varied from 450-720 g C m^{-2} yr^{-1}. These data are consistent with the span of GPP in CARDAMOM (Figure 1), again best matching the lower end of the model estimates. For the same study, the vegetation C stock estimates varied from 100-5000 g C m^{-2}, consistent with CARDAMOM, and with measurements of 10 to 40-year old boreal stands best matching the CARDAMOM median estimate of ~1500 g C m^{-2}".

L258: Likewise, this stem turnover time seems much too fast and needs independent validation. I understand that grid cell to plot- or plant-scale validation isn't perfect, but it's better to report the performance explicitly, and then cushion it based on possible scale mismatch, rather than to ignore whether these estimates are consistent with prior research.

Based on comparison to Carvalhais et al. (2014) TT estimates and to the GPP and C_{veg} estimates reported above, our TT estimates are consistent with independent calculations and their component parts. We understand that TT seem short compared to concepts of stand age. However, litterfall (plant mortality) occurs throughout succession, from all live pools, which means that C turns over faster than age suggests.

L294: typo on "uncertainties"

Corrected.

L313: It would be good to have some sort of quantification of spatial coherence beyond RMSE & R2 (which are nonspatial). Look to the GIS and remote sensing literature for examples of what sort of statistics are available to do this.

There are a number of potential statistics to use. We suggest that our choice of statistics is familiar to biogeochemists and earth system scientists. Coupled with direct mapping of ratios and confidence intervals for visual assessment, we suggest our analysis provides readers with the relevant information on spatial coherence. Adding further statistics is likely to provide only marginal gains, but also increase the intricacy of an already complex paper.

L328: Don't introduce new Methods in the Results. Please document what this analysis is and why you are doing it earlier in the paper.

We agree the reviewer 2 is correct and we have adjusted the text as requested, moving material into the last part of the Methods (S2.4, L235-239):

> "To understand the sources of errors in TT_{veg} calculations, we used CARDAMOM to calculate two hypothetical TT_{veg} (i.e. EXPERIMENT A TT_{veg} = ISI-MIP2a C_{veg} / CARDAMOM NPP and EXPERIMENT B TT_{veg} = CARDAMOM C_{veg} / ISI-MIP2a NPP) and then assessed the largest difference with CARDAMOM's CONTROL TT_{veg}. We estimated the hypothetical TT_{veg} for each pixel in each model, and derived a pixel-wise measure of the contribution of biases in NPP and C_{veg} to biases in TT_{veg} by overlapping their distribution functions."

L378: Consistent with my previous concerns, DALEC appears to be running to fast. That said, this is still a comparison to other models, not to data.

We agree that biases may exist in the CARDAMOM TT estimate, but see above about difference between stand age and TT (L258 comment). Also, note that we are exploring where ISI-MIP2a models lie outside the analysis confidence intervals of CARDAMOM for TT.

L391: Here you say you had a 'strong prior on photosynthesis' but as far as I can tell the photosynthetic parameters were fixed at defaults, not assigned priors. According to Eqn 2, the only 2 parameters assigned non-uniform priors were canopy efficiency (which in Tables 2 and S2 is labeled as a phenology parameter) and autotrophic respiration

As noted before, the canopy efficiency is the calibrated parameter in CARDAMOM for the photosynthesis model ACM; we apologise for confusion in not making this clear before. Now this point is clarified in text (S2.2.2, L143-145):

> "DALEC2 simulates canopy-level GPP via the Aggregated Canopy Model (ACM; Williams et al., 1997) and the most sensitive ACM parameter, related to canopy photosynthetic efficiency, is included in the CARDAMOM calibration."

L397: If you've demonstrated a bias in your photosynthetic model, I'm not sure I agree that this could be resolved with more precise data if you're not updating the parameters in the photosynthetic submodel

Again, we have now clarified that a parameter in the photosynthesis model (canopy efficiency in ACM) is being updated by CARDAMOM.

L427: I fundamentally disagree that models should be benchmarked against highly-derived, model-based data products. But this isn't the central point of the paper and thus I won't hold up this paper over that disagreement.

Every data product used here is in some way model-derived – LAI from MODIS requires a model, biomass from radar and landcover maps also, SOC data from interpolation and machine learning approaches, even in-situ data such as GPP and R_{eco} are separated from NEE using a wide range of partitioning algorithms.

L459: While it's true that brute-force MCMC is not feasible for complex models, but there are other options available that do work with larger models, such emulators (Fer et al 2018 Biogeoscience) and ensemble or particle filters.

We agree that there are a range of alternative approaches beyond MCMC and decided to include a sentence in Discussion including reviewer 2's suggestion (S4.3; L476-477):

> "Other viable options include using emulators (Fer et al., 2018) and particle filters (Arulampalam et al., 2002), but MCMC methods provide the most detailed description of error distributions."

We also re-arranged the following sentences and merged paragraphs to improve clarity (S4.3; L477-486):

> "There remains a strong argument to utilize intermediate complexity models like DALEC2 to evaluate the minimum level of detail required to represent ecosystem processes consistent with local observations, and to allow testing of alternate model structures. And, assimilating further data products, for instance patterns in soil hydrology and snow states across the pan-Arctic from earth observation, could provide useful information on spatio-temporal controls on soil activity and microbial metabolism to constrain below ground processes. This information would need to be tied to process level information on SOM turnover generated from experimental studies, and included in updated versions of DALEC. Thus, more field observations are crucial across the pan-Arctic, specifically on decomposition and TT of SOC (He et al., 2016) and respiratory processes such as partitioning of R_{eco} into R_a and R_h (Hobbie et al., 2000; McGuire et al., 2000), across the growing season and also during wintertime (Commane et al., 2017; Zona et al., 2016)."

L477: For the record, if you didn't fit every grid cell independently then you wouldn't need to upscale/interpolate field observations.

Our point is that critical ecological processes remain poorly understood, and so further field observations are required to constrain these processes. Also, if each pixel had not been treated independently, we would have then relied on PFTs with all their problems (clearly pointed at Introduction and Discussion sections), and which is basically the opposite to what CARDAMOM framework is about.

L495: Where are the DALEC2 and CARDAMOM code repositories?

Following up on reviewer 2 initial concern, we have submitted the code for DALEC2 on Edinburgh DataShare. We are working to release a community version of CARDAMOM.

We have adjusted the text (L517-520):

> "**Data and software availability**
>
> CARDAMOM output used in this study is available from Exbrayat and Williams (2018) from the University of Edinburgh's DataShare service at https://doi.org/10.7488/ds/2334. The DALEC2 code is also available on Edinburgh

DataShare at https://doi.org/10.7488/ds/2504. Contact MW for access to the CARDAMOM software."

Table 2: I find it interesting that, given the papers focus on turnover times, turnover parameters are the least constrained part of the model.

Yes, this is the case, and reinforces the focus on TT in this analysis – we will only improve forecasts of high latitude C dynamics from better understanding TT.
