

Review of Yang et al. “CARDAMOM-FluxVal Version 1.0: a FLUXNET-based Validation System for CARDAMOM Carbon and Water Flux Estimates”

Reviewer 1

General Comments

1.1 Yang et al present a consistent validation framework (including a suite a model evaluation metrics) for a terrestrial biosphere model data assimilation system (CARDAMOM) that can be used to systematically test different versions of the model in addition to various DA system configurations (observation record length etc). The introduction really nicely lays out the motivation for developing a rigorous DA-validation framework and the analyses presented provide a useful demonstration of how that framework can be used to answer the key questions posed in lines 90-95. This is important work that all TBM DA groups need to be routinely performing, and as such this paper should serve as a guiding framework or benchmark that I believe will be of wide use to the TBM modeling community.

I have very few comments or suggestions – this paper could be published as is. The following are requests for minor clarifications or questions based on curiosity.

Author response: We thank the reviewer for their positive comments to our study. We have revised the manuscript thoroughly to address all the comments raised by the reviewer, and the response to each comment is listed below (author responses shown in blue).

Minor specific comments

1.2 Line 104-105: Am I missing something here because “one solely trained by satellite and inventory datasets, and the other constrained by 50% of data at each FLUXNET site” is not the same as what is presented in the CARDAMOM Analysis 1 and 2 (lines 136-139)? [Updated comment] This actually only becomes clear once we’ve read the first paragraph of section 2.2. It might be useful to make this more clear at lines 104-105 so readers know what is coming up.

Author response: We agree with the review’s suggestion, and revised the text in the manuscript to make it clear to understand. Now it reads, “one solely trained by satellite and inventory datasets, and the other additionally trained with 50% of GPP, NEE and ET observations at each FLUXNET site”.

1.3 Line 119: I am wondering why the authors didn’t use the FLUXNET site met forcing but the ERA-Interim grid instead? I wasn’t sure based on reading lines 119 and 155 if this was the case, but it became clear in Section 4.1. Why use a coarser scale forcing when that could introduce biases and the gap-filled site met forcing are readily available (although I appreciated this caveat is acknowledged in the discussion)? (In fact, the FLUXNET 2015 site data have been gap-filled using ERA-I based on the method of Vuichard et al. (2015)). Other caveats related to the

different spatial resolutions (and need to aggregate to 1km) for the satellite data used in the assimilation could also be addressed in Section 4.1.

Author response: We thank the reviewer for pointing this out. We chose to use reanalysis-based forcing data, since gap-filled site meteorological datasets are only available for the timespan of the FLUXNET GPP, NEE and ET measurement record, while re-analysis is available for the entire 2001-2015 CARDAMOM analysis period. To address the reviewer's remarks, we have made the following changes:

- (1) In the methods section, we now explicitly state that we use a reanalysis-based 2001-2015 forcing dataset at all FLUXNET sites. We also clarify that we chose to implement CARDAMOM for the entirety of 2001-2015 across all sites in order to exclude the effect of varying CARDAMOM simulation lengths in our subsequent results.
- (2) In the discussion section, we now clarify that gap-filled site meteorological datasets are available for the timespan of the FLUXNET2015 GPP, NEE and ET measurement record (as detailed in Pastorello et al., 2020), and that these can be preferable for CARDAMOM FLUXNET experiments where the analysis window is confined to the FLUXNET measurement record.
- (3) In order to facilitate the use of gap-filled site meteorological forcings in subsequent CARDAMOM-FLUXVAL analyses, we have now added technical guidelines on how to replace selected drivers in the manuscript supplement (section S2, Code Implementation) and reference these in section 4.2.

1.4 Line 150: might be good to put the original spatial resolution in Table S1.

Author response: We have followed the reviewer's suggestion and added the original spatial resolution in Table S1.

1.5 Fig. 7: really interesting that there seems to be a dip in model performance with record lengths of 2-3 years, especially for the C fluxes. Any thoughts as to why that might be? Is it again related to the types of sites/land cover types with that record length?

Author response: In the results section, we now state that (i) for the sites with record lengths of 2-3 years, the percentage of the non-forest PFT (grassland) is higher than other year ranges, and (ii) given that forested sites overall outperform non-forested sites, we speculate that the lack of forest sites in the "2-3 years" category is the likely cause of the relative dip in model performance.

1.6 Line 36: Minor comment but I think there maybe the latest GCB budget paper might be a better reference here than Le Quere et al. (2020) here given this paper is focused on the impact of COVID-19 on the global C sink (latest GCB budget is Friedlingstein et al. (2020))?

Author response: We agreed with the reviewer and changed the reference to Friedlingstein et al. 2020.

1.7 Line 39: If the authors would like an updated version of Friedlingstein et al. (2014) see Arora et al. (2020).

Author response: We thank the reviewer for the updated the reference. It has been updated in the manuscript. We changed the Friedlingstein et al. (2014) to Arora et al. (2020).

1.8 Line 41: I'm not sure I'd use the Reich or the FLUXCOM references for "improve modelling of the Earth's climate system and reduce uncertainties" – especially in the context of terrestrial biosphere modeling as these references refer to a scaled-up flux data product?

Author response: we expanded "modelling" to include "empirical modelling or data-driven predictions" and make sentence more comprehensive. We also changed "climate system" to "key components of the land surface and Earth system" in the manuscript.

1.9 Certain references are given as the discussion (preprint) version of the manuscript and not the final accepted paper and some references are missing.

Author response: We have cleaned the references and added the missing ones in the manuscript as suggested by the reviewer.

1.10 Table S4: I am guessing for the initial conditions you mean "at time $t=0$ " – i.e. you only update these at the beginning of the assimilation window? Also, I'm probably missing something but what does "at T, P " mean?

Author response: Yes, the initial conditions mean "at time $t=0$ " in the model setting. And the "at T, P " means the time-averaged reference temperature and precipitation. We have added an overbar symbol to correctly denote this in Table S4, and we have updated the table caption to better describe the aforementioned terms.

1.11 Figure 1: I guess the tables of metrics are just for the prediction window, or for the whole timeseries?

Author response: Yes, all the tables of metrics are only for the prediction window; we have updated the figure caption to clarify this.

Reviewer 2:

2.1 This study, “CARDAMOM-FluxVal Version 1.0: a FLUXNET-based Validation System for CARDAMOM Carbon and Water Flux Estimates” by Yang et al. provides benchmarking for the performances of the CARDAMOM-DALEC modeling framework across FLUXNET (2015 release) sites. The manuscript focuses on the performances of GPP, NEE and, ET fluxes with only RS-based constraints (A2) and with additional FLUXNET observations of the same variables (A1). The manuscript shows that the performance of the model improves, in general, across all variables and temporal scales. The added CARDAMOM-FluxVal workflow, therefore, definitely brings in a useful additional data stream to the CARDAMOM-DALEC framework, and may thus be a valuable baseline for scientific studies of a similar nature.

Nevertheless, the study and manuscript would benefit by being a bit more comprehensive (with additional analyses or detailed discussions) about the following aspects.

Author response: We thank the reviewer for summarizing our study and providing the positive comments and suggestions. We have revised the manuscript and added additional analysis and discussions. We have addressed all the comments and attached the point-to-point responses below.

2.2 As mentioned in the limitations, the use of nearest gridded (a half degree) ERA-Interim (hereafter, ERA-I) data for forcing the DALEC model is a major limitation. But, the rationale and necessity of using the gridded data instead of potentially using the tower measurements are not clearly described. It may be that it was not possible because there are gaps in eddy tower-based meteorological measurement, but this has to be mentioned and discussed. Also, one can provide a meta-analysis on comparison of (the available) meteorological variables from the tower and the corresponding ERA-I estimates for the sites. Such analysis would provide information on whether the gridded data are representative of the ecosystem micro-climate or whether they can be used at all. Additionally, it seems that Pastorello et al (2020) also provide the gap-filled meteorological variables. Was this option evaluated as well?

Author response: We thank the reviewer for pointing this out. In response to reviewer 1 (comment 1.3), we clarify that we opted to use ERA-Interim re-analysis data, since gap-filled site meteorological datasets are only available for the timespan of the FLUXNET measurement record, while re-analysis is available for the entirety of the 2001-2015 CARDAMOM analysis period at each site. To highlight the potential use of gap-filled site meteorology in subsequent CARDAMOM efforts, we now (1) clarify that gap-filled site-level meteorological data is available to use for CARDAMOM-FLUXVAL analyses temporally confined to the FLUXNET measurement record (Pastorello et al., 2020), and (2) include technical guidelines on replacing CARDAMOM-FLUXVAL ERA-interim data with gap-filled FLUXNET meteorology in the manuscript supplement, in order to facilitate their use in subsequent efforts. See response to reviewer 1 (comment 1.3), for details on the aforementioned changes.

2.3 The baseline remote sensing constraints: The paper mentions that the MODIS LAI was aggregated from the original 1 km resolution. It was unclear if it was aggregated to a half degree or some other resolution. For the site level simulation, the 1 km resolution data would probably be the closest to the normal footprint of eddy towers. So, I suggest discussing why such aggregation would be needed. Perhaps, aggregating the LAI to a coarser resolution may make it more consistent with the ERA-Interim climate at the same resolution. This may explain why the baseline A2 simulations are already performing quite well. One specific question would be, “does the half-degree forcing reproduce the LAI variability at 1 km resolution or does the LAI have to be aggregated as well?”

Author response: We thank the reviewer for bringing this up; we did in fact use the nearest 1km MODIS LAI retrieval, and did not aggregated LAI to 0.5 degree. We have now reworded the methods text (section 2.2) to clarify that for each FLUXNET2015 site, MODIS 500m resolution LAI data was aggregated to the 1km x 1km area surrounding each FLUXNET site.

2.4 Use of the gap-filled observation: It was unclear why the gap-filled variables from FLUXNET were used. Any particular reasons to use gap-filled constraints were not given, and the cost metric (likelihood) can be calculated only using the time steps which have the observations. The manuscript would benefit by having an explanation of how/why we pick the right observational data variable when optimizing model parameters. In my opinion, the gap-filling itself can be a source of uncertainty. For example, is there a systematic pattern between the fraction of gap-filled observation and the (lack of) improvement in model performance?

Author response: In the revised manuscript, we now clarify that we use the monthly GPP, NEE and ET provided in the FLUXNET2015 monthly resolution dataset, and we further filter these using the FLUXNET2015 quality check flags; we therefore only calculate the model likelihood where and when GPP, NEE and ET observations are available. For completeness, we also clarify in section 2.2 that the FLUXNET 2015 monthly resolution estimates used in our analysis are derived from gap-filled hourly or half-hourly resolution data (Pastorello et al., 2020); this step is necessary to minimize sub-monthly representation errors in the monthly flux averages.

2.5 Consideration of observational uncertainty: In the likelihood function, each data stream has an error scaling (σ , sigma). It is mentioned that it represents error across model and data. But it is unclear how these values were set. Was it based on the observational uncertainty (available for some of the FLUXNET variables)? How does σ (sigma) affect the parameter inversion or a contribution of a particular data stream to the total likelihood?

We prescribed ABGB and LAI uncertainty values based on previous CARDAMOM efforts (Bloom et al., 2020); NEE sigma values were also based on previous NEE error characterizations (Famiglietti et al., 2021; Papale et al., 2006), and further scaled to $1\text{gC}/\text{m}^2/\text{day}$, as we found the cost function was otherwise insensitive to other datasets. For lack of better knowledge, we used a trial-and-error approach to choose the largest GPP and ET uncertainty values that provided model efficiency (MEF) values comparable to NEE (Figure 2). We now include a description of observational uncertainty choices in the revised text and Table S1.

For completeness: we note that FLUXNET2015 reported observation uncertainties are overall considerably smaller than our assumed model-data residual errors; we therefore implicitly assume that our uncertainty choices predominantly represent either model structural error and/or additional uncharacterized observation errors (Bloom et al., 2020; Famiglietti et al., 2021).

As highlighted in section 4.3, uncertainty choices are a determinant of model performance, and we advocate for using CARDAMOM-FLUXVAL to investigate these in subsequent efforts. We now also highlight previous CARDAMOM work quantifying the relative importance of error scaling in CARDAMOM flux predictions (Famiglietti et al., 2021).

2.6 PFT-level comparison: The manuscript presents a PFT-level analysis of the model performances. As the parameters are optimized per site (I assume), it is unclear if the performance should be associated with PFT at all. Are there PFT-specific parameters or are the relatively poorer performance in non-forest PFT indicative of model structural shortcomings? Also, would climate-based segregation reveal anything interesting? ET would be a much valuable constraint in a moisture-limited climate than in an energy-limited one.

Author response: We thank the reviewer for pointing out the issue of PFT-related analysis. Yes, we optimized the parameters per site, and there for the performance in theory should not be PFT-dependent. However, there could be the following reasons:

Due to the FLUXNET data quality issues in each site, and the large uncertainty often existing in the tropical forest sites and/or nonforest sites with heterogeneous landscape, we found the performance differences based on PFT. However, our PFT-level analysis may also indicate model structure challenges, as suggested by the reviewer. Similarly, we anticipate that categorizing performance in climate space can provide further quantitative insights on potential model structural shortcomings.

We have modified section 4.2, Limitations of FLUXNET validation approach, where we now (i) highlight the possibility of model structural shortcomings alongside with the observational uncertainty for different PFT, and (ii) clarify that projecting model performance in climate space can be useful for further characterizing model shortcomings.

2.7 Parameter uncertainty: The manuscript presents an analysis on parameter uncertainty and model performances across sites, but falls short in addressing a more basic question of whether including additional FLUXNET data-stream helps reduce the uncertainty of estimated parameters within a site. For example, the parameters related to (radiation and) water may be better constrained when an additional data stream of ET is introduced. The manuscript can be better in this aspect. Table S4 should be extended to include the optimized parameter values and uncertainty ranges for both A1 and A2 experiments. As the study clearly states in the introduction and motivation, additional data streams can not only help in improving the performance of the model but also potentially reduce parameter uncertainty and identify model structure errors. The first aspect is well covered, but discussion on the last two aspects would be equally useful as well.

Author response: We agree with the reviewer that the additional data streams could impose constraints on model parameters as well. While the posterior ranges of the optimized parameters vary from site to site, we did find a consistently reduced uncertainty in a few estimated parameters. Following the reviewer's comment, we have added an additional figure (Fig. S5) showing the improved ranges for a selected number of parameters from A2 to A1 experiments. We did not expand Table S4 as the reviewer suggested because (i) the range variations of different parameters can be large from site to site, and (ii) we did not find any universal patterns in estimated parameters, likely due to the considerable site-to-site variability. Although a more detailed investigation of parameter posterior ranges is beyond the scope of our manuscript, we highlight that PFT or climate-space aggregation of parameter constraints are worth investigating in subsequent efforts.

References:

Bloom, A. A., Bowman, K. W., Liu, J., Konings, A. G., Worden, J. R., Parazoo, N. C., Meyer, V., Reager, J. T., Worden, H. M., Jiang, Z., Quetin, G. R., Smallman, T. L., Exbrayat, J.-F., Yin, Y., Saatchi, S. S., Williams, M., and Schimel, D. S.: Lagged effects regulate the inter-annual variability of the tropical carbon balance, 17, 6393–6422, <https://doi.org/10.5194/bg-17-6393-2020>, 2020.

Famiglietti, C. A., Smallman, T. L., Levine, P. A., Flack-Prain, S., Quetin, G. R., Meyer, V., Parazoo, N. C., Stettz, S. G., Yang, Y., Bonal, D., Bloom, A. A., Williams, M., and Konings, A. G.: Optimal model complexity for terrestrial carbon cycle prediction, 18, 2727–2754, <https://doi.org/10.5194/bg-18-2727-2021>, 2021.

Papale, D., Reichstein, M., Aubinet, M., Canfora, E., Bernhofer, C., Kutsch, W., Longdoz, B., Rambal, S., Valentini, R., Vesala, T., and Yakir, D.: Towards a standardized processing of Net Ecosystem Exchange measured with eddy covariance technique: algorithms and uncertainty estimation, 3, 571–583, <https://doi.org/10.5194/bg-3-571-2006>, 2006.

Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y.-W., Poindexter, C., Chen, J., Elbashandy, A., Humphrey, M., Isaac, P., Polidori, D., Ribeca, A., van Ingen, C., Zhang, L., Amiro, B., Ammann, C., Arain, M. A., Ardö, J., Arkebauer, T., Arndt, S. K., Arriga, N., Aubinet, M., Aurela, M., Baldocchi, D., Barr, A., Beamesderfer, E., Marchesini, L. B., Bergeron, O., Beringer, J., Bernhofer, C., Berveiller, D., Billesbach, D., Black, T. A., Blanken, P. D., Bohrer, G., Boike, J., Bolstad, P. V., Bonal, D., Bonnefond, J.-M., Bowling, D. R., Bracho, R., Brodeur, J., Brümmer, C., Buchmann, N., Burban, B., Burns, S. P., Buysse, P., Cale, P., Cavagna, M., Cellier, P., Chen, S., Chini, I., Christensen, T. R., Cleverly, J., Collalti, A., Consalvo, C., Cook, B. D., Cook, D., Coursolle, C., Cremonese, E., Curtis, P. S., D'Andrea, E., da Rocha, H., Dai, X., Davis, K. J., De Cinti, B., de Grandcourt, A., De Ligne, A., De Oliveira, R. C., Delpierre, N., Desai, A. R., Di Bella, C. M., di Tommasi, P., Dolman, H., Domingo, F., Dong, G., Dore, S., Duce, P., Dufrière, E., Dunn, A., Dušek, J., Eamus, D., Eichelmann, U., ElKhidir, H. A. M., Eugster, W., Ewenz, C. M., Ewers, B., Famulari, D., Fares, S., Feigenwinter, I., Feitz, A., Fensholt, R., Filippa, G., Fischer, M., Frank, J., Galvagno, M., Gharun, M., Gianelle, D., et al.: The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data, 7, 225, <https://doi.org/10.1038/s41597-020-0534-3>, 2020.