Earth Syst. Dynam. Discuss.,
https://doi.org/10.5194/esd-2018-19-RC2, 2018

Earth System Dynamics
Discussions

Interactive comment 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: 7 August 2018

In this paper the authors take the CARDOMOM + DALEC Bayesian calibration system an apply it specifically to the arctic using a number of regional-scale data products. Once the model is fit to data, it is then used to assess carbon pools and benchmark global vegetation models. The scale and scope of the analysis is quite impressive – building up their system to this point was clearly a lot of work and the attempt to synthesize multiple data constraints at a regional scale is really important, especially for a highly influential and understudied region like the arctic.

That said, I do have a few high level concerns about what the authors have done. The easiest of these to address is that the details of what was actually done was insuffi-

cient and teasing out important high-level facets of CARDOMOM are left to the reader tracking down earlier papers. Particularly important is to clarify whether DALEC is calibrated independently for every pixel, in some sort of spatially correlated manner, or with a single parameterization for the two PFTs across the whole region. My recollection from earlier papers made me think the first (independent fits), but in reading the results it is hard to distinguish parameter uncertainty from parameter spatial heterogeneity. The authors need to be more explicit about this. Likewise, the authors need to be more clear about whether this is really a data assimilation system, or if it's just a calibration system. This matters because in DA (e.g. EnKF) the analysis provides a formal synthesis of observations and process understanding, but in a calibration system your estimated states are ultimately just a forward model run. To me, it feels like the authors are treating a forward model run as if it were a reanalysis product. If this is true what the authors did is still valuable but they should be more open about this and the limitations of this approach.

Second, in light of the earlier point about reanalysis vs forward simulation, I am really uncomfortable about the author's use of their model as benchmark for other models. This is particularly true given the non-trivial biases in some of the verification (biomass) and validation (GPP, Rh) analyses and the lack of independent validation of a number of the other processes in the model (e.g. turnover). I think this manuscript could stand alone without the GVM component.

Third, I'm really concerned about how the authors assimilate these derived data products. There's not really any discussion of how the observation errors in the data and process error in the model are treated. There's not any discussion of how the authors handled the non-independence of spatial pixels in these data products. Indeed the authors seem to treat data products as if they are truly data, which likely results in an overestimation of the true information content in the data. For example, if I have 10 observations I can Krige a map that has 10k grid cells, but my true sample size remains 10 not 10k and any data assimilation system needs to reflect that.

Detailed Comments:

L126: 1) Is calibration really data assimilation? 2) inclusion of process error?

L135: What is the actual underlying sample size? Derived data products can massively conflate the actual information content. Errors in these data products are hugely autocorrelated and that observation uncertainty is not captured correctly in these products. Also, many of these constraints are not data (GPP, LAI, biomass) but just different models.

L144: 500 samples per chain? That's way too small. Also, what's the effective sample size after accounting for autocorrelation? I'd recommend the authors shoot for an effective sample size around $\sim$5000 total, which likely will require a much larger total number of samples given their reliance on Metropolis-Hastings. Not stated explicitly whether this is one global parameter set or one per grid cell? My memory from Bloom et al 2016 is the latter.

L146: A 90% CI is typical. Reason for not 95% norm?

L154: This isn't independent of the calibration product

L164: should really include the 95% CI in addition to the interquartile

L169: You can't compare a complex model against a (mis)calibrated simple model and call it a benchmark. Especially true if you're looking at the marginal distributions of indirectly inferred latent variables.

L177: If looking at the historical period, why weren't models run under reanalysis meteorology rather than GCMs?

L184: Drop this whole paragraph – it's a bit confusing to give a summary of the results before presenting the results without making it clear that this is a summary of highlights. Right now it just feel like you're going though the results really quickly without much explanation.

L187: A 28% bias against the data that the model was calibrated to seems like a pretty big problem.

L190: "This mismatch is important in the context of FLUXCOM, as noted" what do you mean "as noted" you never noted anything

L203: "and marginally varied across tundra" I don't understand what you mean here

L209: Distinguish tundra and taiga. These numbers don't seem plausible for tundra

L211: That the tundra numbers are so close to the taiga numbers doesn't seem correct. How well do these numbers validate against direct field data (not derived data products)?

L216: A transit time of 4.3 years in the woody tissues of a spruce tree seems really fast give their lifespan. How does this compare to field data (e.g. isotopes)

L217: The CI on the SOM is really large (essentially 10-1000 years). Is this just the prior?

L228: This results needs additional explanation with regards to what this test statistic applies to. You calibrated a mechanistic model via MCMC, this isn't a t-test. What specifically changed that much?

L234: What do you mean priors, isn't this the data?

L257: This is almost the exact same sentence as L190

L295: Is this statement that CARDOMOM is more sensitive biomass than soil C really fair? In one case you're comparing whether a data constraint is included at all, while in the other your comparing different derived data products, which are likely relying on similar underlying raw data. I think for this to be fair you would want to include a version where you don't have any soil C constraint.

L308: There's a 28% bias in biomass, how is that "good agreement". The Discussion

seems to be missing the critical point that if a model is faced with multiple constraints and can't reconcile them then there's either inconsistencies in the data, structural errors in the model, or both. And why is there no comparison to LAI and GPP constraints? Also, there seems to be no discussion of how observations error in the data are derived/treated and how you're handling the process error in the model (is this a fit parameter or just ignored).

L312: I haven't looked into the details of the Jung 2011 product vs the Jung 2017 product, but I'm skeptical that these are independent. Would be good to state more explicitly what each product is upscaling to generate GPP (FLUXNET? SIF?). If they're both FLUXNET-based then they're not independent if they're just applying different algorithms to upscale the same underlying data.

L314: "One difference between these two models is. . ." What two models?

L317: I'd recommend making this sentence the start of the next paragraph

L319: How do you know that the issue is only one of scale difference, and not some other error in the model or DA system? What could you do to confirm this (e.g. run with local drivers)?

L326: This error in timing is an example of why it might be better to run a system that performs both state and parameter data assimilation, rather than just parameters.

L328: It's a bit surprising that you're running a model in the arctic that doesn't include snow or permafrost. I see that this point is in the Discussion, but it seems really important to be more upfront about this earlier in the paper, as it's a pretty limiting assumption and should lead to greater caution in how confidently you interpret the results. It also begs the question as to why you didn't couple CARDOMOM to a more sophisticated land model for this analysis.

L365: But is there any direct field constraint (e.g. isotope data)

Interactive comment on Earth Syst. Dynam. Discuss., https://doi.org/10.5194/esd-2018-19, 2018.