Comment on bg-2022-109
Anonymous Referee #2

Referee comment on "Assimilation of multiple different datasets results in large differences in regional to global-scale NEE and GPP budgets simulated by a terrestrial biosphere model" by Cédric Bacour et al., Biogeosciences Discuss., https://doi.org/10.5194/bg-2022-109-RC2, 2022

Summary of manuscript:

The manuscript describes a set of data assimilation experiments applied to the ORCHIDEE Terrestrial Biosphere Model. Three data streams were used to optimize model parameters: NEE measurements from flux sites, atmospheric CO2 observations, and satellite derived NDVI values (at select pixels). These three data streams were used for data assimilation separately, in pairs, and all together (either simultaneously or in a step-wise procedure). Parameters related to photosynthesis, phenology, and respiration were optimized in various combinations appropriate to the data streams being employed. Broadly, the authors found that simultaneous data assimilation resulted in better bias reduction and potentially less overfitting, and that calibrating the initial soil carbon stocks has a large influence on the data-model agreement (particularly for observed CO2 concentrations).

General comments:

Data assimilation in the context of terrestrial biosphere models is challenging—this manuscript is a detailed attempt at meeting this challenge, which seems valuable.

I think this manuscript could benefit from clearer overarching justification. Is the data assimilation exercise presented here purely for technical exploration? How will it serve broader goals related to C cycle forecasting and model uncertainty analysis? What are the next steps after this technical advance? Other broad questions might be worth briefly addressing:
- To what extent could optimization disguise biases that result from missing processes or the underlying model structural assumptions (e.g., how soil carbon is simulated, see below)?

- To what extent is overfitting identifiable? Is there a way to quantify overfitting quantitatively in this context, or is it being defined qualitatively here?

- How might this data assimilation procedure compare to other procedures (e.g., MCMC)?

In addition to these broad questions I have a major technical concern related to the treatment of soil carbon.

Optimizing the initial soil carbon pools without any actual constraints on soil carbon seems far from ideal. Why not use actual observations of soil carbon stocks or respiration? The link between soil carbon and the variables that are being used for data assimilation (e.g., atmospheric CO2) is very indirect.

I also wonder if it might be more appropriate to optimize the rate parameters and transfer coefficients in the soil-carbon sub model within the TBM, rather than adjusting the initial stocks. Optimizing the initial stocks assumes that mis-match between the modeled and observed values is due to the fact that soil carbon pools are not at steady state. It certainly seems plausible that the steady state assumption doesn’t hold, but there are plenty of other reasons why modeled soil carbon stocks might not match observations (e.g., the rate parameters and transfer coefficients in the soil carbon model aren’t optimal, or more likely the DATCENT-type soil carbon model is only a crude approximation of soil carbon biogeochemistry). Optimizing initial stocks sweeps all of these issues under the rug, so to speak, implicitly making the argument that all of the error related to soil carbon is due to the fact that transience during the spin up period has been ignored. Carbon inputs and soil respiration might be at least as important (see for instance this paper: https://doi.org/10.5194/bg-10-1717-2013).

The authors partly acknowledge these issues (lines 835-853), but do not fully explain why soil data can't be used for data assimilation, or outline the pitfalls associated with optimizing initial pools rather than inputs or rate parameters. If the approach used here is retained, it must be clearly explained.

Also, it might be worth noting that DAYCENT-type first order soil carbon models (which are standard in terrestrial biosphere models) are all structurally deficient, in that they cannot capture feedbacks related to microbial activity. There has been a great deal of experimentation in the last decade aimed at addressing this deficiency (e.g., the MIMICS,
MEND, and CORPSE biogeochemical models, which are all "microbially explicit"). At some level, DAYCENT-type models can only be right for the wrong reasons, no matter how they are optimized.

While I understand that full optimization using soil carbon datasets would fall outside the scope of this paper, it would be a good idea to compare the optimized initial carbon stocks with observations at a regional scale as a qualitative reality check. The ISCN or ISRIC-WoSIS soil profile databases or the gridded Harmonized World Soil Database or Soilgrids data products might be useful for regional ground truthing.

Detailed comments

Title: consider striking “different”, since “multiple datasets” strongly implies different datasets.

41-43: Is “correction of the initial carbon stocks” a general problem for TBMs, or a problem in this context given how spin-up was dealt with?

Line 47: Suggestion: replace “grown” with “increased”.

Lines 95-96: What does this mean: “considered potential incompatibilities between the model and the observations”?

Lines 109-110: This point about Gaussian errors seems important. How does it affect the analyses presented later in the paper?

Lines 130-132: Is it OK to tune priors or prior uncertainties? This seems antithetical to the definition of a prior.

Line 230: Were the pixels selected entirely at random given these constraints?

Line 256: The notation in the equation may be confusing to some readers: the superscript
indicating transposition (t) might be mistaken for an exponent. Perhaps clarify what this superscript mean in the text, and/or use a different notation (the symbol ` or an upper case T might be less confusing).

Line 261: Does this approach really account for uncertainty in the model structure? Perhaps elaborate in another sentence, as it is not obvious why this would be the case.

Line 270: “the calculation of using” is confusing syntax, consider rephrasing

Lines 291-292: Land use is important, but this statement is unreferenced and there are certainly other sources of variation (and other potential drivers of model-data mismatch).

Line 366: “of and of” … missing word?

Lines 515-517: Could this indicate overfitting? Tuning the soil carbon pools could be masking other sources of disagreement between model and data.

Line 588: Why no transient simulations in this study? Is this a computational constraint?

Lines 832-834: This point needs some elaboration. How does correcting the CO2 trend bias hinder evaluation of photosynthesis?

Lines 847-848: The steady state or the non-steady state approach causes bias? This sentence is ambiguous.

Lines 849-851: What are the inconsistencies that limit application of soil carbon data? This point needs a far stronger justification and should be made earlier in the paper, given that soil carbon pools are being tuned without reference to soil observations.