Biogeosciences Discuss.,
https://doi.org/10.5194/bg-2018-156-RC2, 2018

Biogeosciences Discussions

Interactive comment on "Calibrating a process-based forest model with a rich observational dataset at 22 European forest sites" *by* David Cameron et al.

Anonymous Referee #2

Received and published: 31 May 2018

In general, this paper is a bit long and not focused enough as to what are the main hypotheses that you want to test, what you found and why it all matters. (There's a lot of text devoted to what you did, and what happened, but it's not crafted into a coherent story.) Even the question of pooled vs. un-pooled calibration across sites has some mixed results, and it's also not clear that the "richer" dataset used here as compared to previous studies helped you to answer questions that couldn't be answered previously.

Other major and minor comments are listed below, followed by comments on tables & figures.

Major comments

The abstract is currently too long and meandering and should be shortened and focused. Some terminology is also unclear until after reading the paper, e.g. "single dataset calibrations". It's not clear what datasets you're referring to. The abstract also suggests the use of Bayesian hierarchical calibration to realize the benefits of both the pooled and un-pooled approaches. Since this method is not specifically applied in this study, the authors should clarify that this method may be useful in future work. (Is there any chance to include a small case study with Bayesian hierarchical calibration, perhaps in SM? Or just do a test and describe in a few sentences the results in the main text?)

It's not clear if the same observational data is first used for calibration, and then for validation (model-data differences). Wouldn't this be a non-independent use of the data for validation/ evaluation of the model? It appears to be so given your statement on page 16, line 6 that "Most obviously, reducing model-data differences is most effective when the data used in the comparison are also included in the calibration". The authors need to better justify this double use of the data for both calibration and validation.

The term "rich dataset" is somewhat over-used since this is a vague term. Should specify quantitatively how many more ecosystem variables and sites you are using than was used in prior studies. Also, consider making the paper title more specific in terms of the dataset used for calibration. What is it that you can do with this dataset and the BASFOR model that wasn't possible in previous studies?

Throughout the results sections (4 to 6) , there are a large number of statements that describe modeling results without offering any explanation as to why the phenomenon occurred. Please keep in mind that the reader knows less about this topic than you do, even if you discuss the results more generally in the Discussion section. You don't need to describe every aspect of your results, only the important ones that help to tell a story. If there is no significance in a particular result, consider removing the statement. Some examples below:

* p. 11, lines 13 to 19, entire paragraph starting with: “For the low frequency variables, model outputs at sites NL-Spe and IT-Ren generally moved furthest away from the observations, relative to the prior mode. . .”

* p. 12, line 1: “There is also a much stronger relationship between the number of observations included in the calibration and the reduction in uncertainty from prior to posterior than was true for RMS deviation changes after calibration. It is also noted that uncertainty reduction was less where the variance in time of the model output increased significantly from prior to posterior.”

* p. 12, lines 21 to 24: paragraph starting with “Not all the calibrations improved the comparison against observations. Indeed, the “ET” calibration increases model differences. . .”

* p. 13, lines 8 to 10: “The “ET” calibration is most likely to increase differences with observations after calibration. . .” Physiologically, why would an ET-only calibration reduce above-ground C sequestration?

The authors should describe the rationale for looking at the two different metrics of model quality, 1) the normalized RMS deviation (ratio between posterior and prior) and 2) the range of the 95th to 5th quantiles (ratio between posterior and prior). What different sorts of information is each metric providing? In terms of rationale, it appears that 1) gets at model fit to data, whereas 2) is more of the mathematical representation of uncertainty. What does it mean when model uncertainty goes down, but model fit to data gets worse? The authors should discuss this, e.g. in last paragraph of page 15, although this is later touched upon in section 7.1.2. Also, the word “uncertainty” is used loosely throughout the paper but it should really be clarified early on in the methods and then used in a precise way going forward.

In general, there is a lot of detail in the Methods section which could be moved to the supplemental material to help improve the flow of the paper. For example, most of section 2.4, especially sections 2.4.5 and 2.4.6 on the optimization procedure, are

details that detract from the main story of the paper. Also, in the Methods section, it makes more sense to first present the model (i.e. BASFOR) and then describe the observational data used to calibrate and validate it, i.e. move sections 2.1.2 and 2.1.3 after 2.3 (model driving/ input data).

What is the rationale for looking at the 6 different datasets for calibration? If you have all this data, wouldn’t you typically use all of it (both high and low-frequency)? Are you trying to make recommendations for cases when other researchers have less data available to them? The authors should try to better explain the rationale for these tests, rather than just trying out a bunch of stuff and then reporting the results.

When discussing the relative benefits of pooled vs. un-pooled calibrations, please keep in mind that it matters where you are planning to apply the model. If the model is mainly used at the sites included in the calibration, un-pooled parameters may make more sense. Also, if the areas you are extrapolating to can be well-represented by one or a few sites included in the calibration, un-pooled parameters optimized for just those sites may be best. If areas where the model is applied cannot be neatly classified into types as represented by the calibration sites, then a pooled model that captures more variability across ecosystems may be better. In this study, it looks like you don’t apply the model to sites not included in the calibration; so therefore, this consideration is not discussed, but likely relevant for “real-world” applications of the model.

If this study had access to a richer dataset with more variables describing different parts of the ecosystem than previous studies using BASFOR, how do the MAP parameters from the “all data” calibration compare to those from previous studies? Can you learn something about ecosystems and/ or model performance from your best model?

Minor comments

Should try to simplify & clarify terminology throughout the paper when comparing prior vs. posterior (i.e. MAP) and come up with shorter names for the various metrics you use to assess model quality. It’s not easy to read or follow now. For example, in the

caption to Figure 4: “Ratio of the posterior to the prior of the range of the 5th to the 95th quantile”. Maybe simplify to something like: “uncertainty reduction from posterior to prior”, and then refer to the section where this is explained in detail. It also would be useful to provide 2 simple equations in the methods section (2.6) to clarify what each of the two uncertainty metrics (normalized RMS deviation and the range of the 5th to the 95th quantile) are showing.

There are lots of long sentences without commas throughout the paper, which makes it hard to understand the flow of the sentences and arguments. For example, p. 11, line 27: “Of these model outputs for NL-Loo was one of the closest sites to observations prior to BC.” Also, page 12, lines 18-20: “Model derived GPP was not included in the calibration so in this case the largest decreases are found when all the calibration data are included although as might be expected inclusion of NEE is the next most important dataset for decreasing model and data differences”. Many more examples throughout the paper.

p. 5, line 16: “the (weather) data were replicated backwards in time based on the available time series. . .” Can you please explain how you replicate weather data backwards in time, and the potential modeling errors associated with this replication?

p. 5, lines 24-25: “A figure showing the stand history reconstructions used is given in the supplementary material”, I don’t see this in the SM. p. 6, lines 24-25: “We chose the Beta distribution.” Why? Should try to offer brief justification for modeling choices.

Why did the authors choose to use site-specific priors for the water retention curve, rooting depth and initial soil and litter carbon values (p. 7, lines 11-12)? Was it availability of data, or knowledge of ecosystem variability?

p. 8, lines 6-7: Why is a 30-day average of model output more reliable than daily? Why not 3-monthly or annual? 30 days seems kind of arbitrary.

p. 8, line 8: Remind us what sigma is again. Is this measurement error?

p. 8, line 24: “In this study, we made three kinds of calibration”. Later, you refer to the multi-site tests with 6 different calibration datasets. Can you clarify here how the three kinds are related to the 6 datasets? Will help reader to follow along later.

Does the pooled calibration include pine, spruce and deciduous sites? (It seems like it based on page 9, line 14: “. . . identical to the default calibration labelled ‘All’ above”). Why not restrict this pooled test to just the pine sites? Why did you choose pine for this pooled vs. un-pooled set of tests?

Page 9, line 27: Please explain what “burn-in” is.

Sections 3.1 and 3.2 (page 10), isn’t it pretty obvious that the calibration would bring the posterior closer to the observations than the prior? How representative are the two sites that you chose in terms of the calibration performance of other sites?

p. 16, line 24: “In general, calibration data was less effective in reducing model-data differences when there were larger inconsistencies present.” This is a pretty vague sentence, inconsistencies between what and what?

p. 17, line 18: “Further, large reductions in uncertainty also occurred when the observations included in the calibration had a tendency to increase model-data differences.” Shouldn’t this be to decrease model-data differences? Shouldn’t model uncertainty theoretically go down when model/ data fit improves?

p. 19, lines 12-13: “suggesting a strong relationship between the underlying processes as represented in the model.” This seems like a self-evident statement that the ecosystem variables included in the model are likely correlated with one another. Can you be more specific here?

Tables and Figures

It would be good to reduce the number of figures in the paper, e.g. by combining high and low frequency variables into a single plot (i.e. Figs 5/6, 7/8, 9/10, 11/12), and then indicating high vs. low frequency using background shading or labels. The authors

may also consider just showing one metric or another rather than both in the main text (moving the alternative metric to SM). Given a "storyline" in the paper, how does each figure help to illustrate an aspect of this story? The figure captions should also indicate what are "good" values for the metrics being shown, or otherwise, how to interpret them, i.e. that values lower than 1 imply some amount of uncertainty reduction.

Figures 1 and 2: how did you choose which sites to show? Shouldn't the posterior be closer to the observations by definition? Maybe combine Figures 1 and 2, and/or move to SM? These figures could also use a legend. Should also specify which calibration is being shown. Is this "all data"?

Figures 3, 4 and 13: It would be helpful to show means for each row and column in these plots, at least across sites for each variable. Figures 3 and 13 are also small and hard to read. For Figure 3, consider flipping each sub-figure and then stacking the two up and down? Might make more space for bigger labels. Again, which calibration are you looking at?

Figures 9 and 10: where is the spread in the boxplots coming from? Is this the spread across sites? Should specify this in caption, and also that you're looking specifically at pine sites here.

Table 3: this is mentioned in text after Table 4, therefore the tables should be re-ordered. Also, please include the full name for each variable, which will help reader to follow along in the text. It would also be nice to show the timescale and the start and end years for each variable. Why is GPP in bold?

Table 4: Do the "measurement available" columns refer to the previous columns? Should clarify.

Table 5: What is the sigma value referred to here? Should define. Also, why are Rh, NO2 and NO included here, but not in any of the results?

Supplemental material

Why is Figure 2 in between Tables 1 and 2?

Figures 3, 5 and 6: what do you mean by "ancillary observations" in the figure captions? How are these different from ordinary "observations", e.g. in Figure 4 caption?

Interactive comment on Biogeosciences Discuss., https://doi.org/10.5194/bg-2018-156, 2018.