Author response to interactive comments on "Optimal model complexity for terrestrial carbon cycle prediction"

Reviewer: The paper by Famiglietti et al uses a suite of data assimilation (parameter optimization) experiments that encompasses models of varying degrees of complexity together with different datasets included in the assimilation to test to what degree model complexity impacts model forecast skill. This is motivated by the general, but not widely tested, assumption that increasing model complexity (and in doing so, the number of parameters) may increase model realism but decrease model predictive skill.

Crucially in this DA context, they use a complexity metric that accounts for both the model structural complexity and the information content of the data that are used to optimize the model parameters.

They show that when unconstrained by data, models of intermediate complexity have the highest skill, thus demonstrating a trade-off between complexity and skill. However, they nicely demonstrate that when constrained by data, models of higher complexity also achieve high forecast skill; thus, confronting models against data (i.e. calibrating parameters, or constraining parameters by some other method) is a prerequisite for increasing the complexity of model structure.

This is important work, particularly given the global terrestrial biosphere model community is still striving to increase the representation of different processes that are deemed necessary to realistically simulate the impacts of climate and environmental change. However, the same community is not investing heavily in implementing DA to constrain uncertainty in their models. This study proves that the two should come hand in hand.

I thought the study was well designed and executed, and the results clearly described and nicely discussed. I only have a few thoughts and suggestions.

We thank the reviewer for their positive comments, which we believe will improve the clarity of the manuscript. We address each comment inline below (author response shown in blue).

Reviewer: In Section 2.5.1 I would explicitly state why you are using this complexity metric (i.e. that it links both the model complexity and number of parameters but also the information content of data), instead of the other model complexity metrics that are available. This is stated multiple times elsewhere, but I think it would be useful here as well.

We agree with the reviewer and have added text to Section 2.5.2 to reflect this suggestion: "The effective complexity of each model run links model structure (i.e., process representation) and number of parameters to the information content of assimilated data. It was computed using a principal component analysis (PCA) on the posterior parameter space."

Reviewer: I'm not 100% sure what point I'm trying to make with this comment so bear with me, but I found myself wanting to dig into more of the nuances of Figure 4, especially in relation the different processes that are included in the model and the level of detail for each process. You do explore more about the of the differences related to the type of data that are included when you talk about Fig 6, and about the differences across sites with Fig 7. So I found myself wondering about the impact process representation (e.g. ACM v1 vs v2). But I appreciate that's beyond the scope of this study. I guess it just may be worth pointing out (the relatively obvious point) that this framework could also be used to determine which exact model representation is most useful for representing a given dataset – e.g. whether to include a water cycle, or not etc.

We thank the reviewer for bringing up this point. This is also of interest to us and we are planning to investigate this in future publications. Nonetheless, we have added text in the Discussion (Section 4.3) to address the reviewer's suggestion: "Finally, while beyond the scope of this study, future work will investigate the linkage between specific processes or process representations (e.g., the inclusion or exclusion of water cycling) and predictive performance to better parse ecological controls on the complexity–skill relationship."

Minor comments

Reviewer: It would be good to add the prior to Fig. 3, just to see how well the DA system is doing.

We have added a supplementary figure showing what the reviewer suggests. The supplementary figure, which is also included on the next page, compares NEE predictions produced using model parameters drawn from their prior distributions (blue) to NEE observations (red). For ease of comparison, this figure's panels directly align with the model–site combinations shown in Figure 3.

Reviewer: Unless I am misunderstanding the histogram intersection metric you have described in 2.5.1 would not range between 0 and 1. Perhaps you mean the normalized intersection metric?

The reviewer is correct—the intersection metric is normalized. We thank them for noting that this was unclear in the text. We have added clarification in Section 2.5.1: "In our case, p was the histogram of predicted NEE or LAI ensembles for a given timestep and q was a discretized Gaussian distribution with mean and standard deviation equivalent to the observed NEE or LAI value and its error, respectively. We normalize the metric by $\sum_{i=1}^{n} p_i$ so that it is bounded between 0 (no overlap) and 1 (identical distributions)."

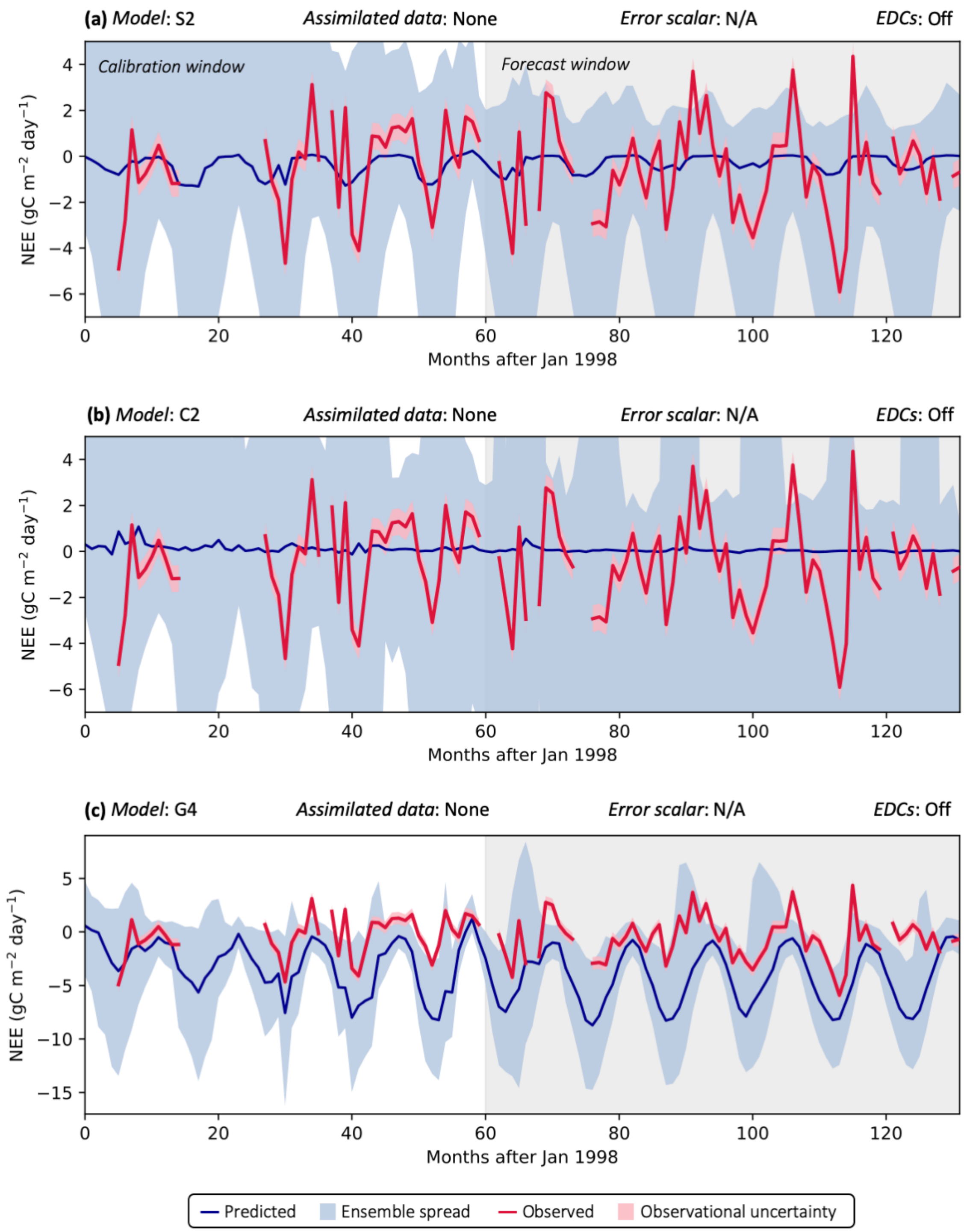

Fig. S1 ***(to be included in revised manuscript)*****: Example model runs parameterized strictly using prior distributions at the FR-LBr site. For comparison, panels correspond directly to the models shown in Fig. 3. The calibration window—the first 5 years of the record—is shown in white and the forecast window is shaded gray. The ensemble spread (blue shading) encapsulates the 5th-95th percentile of runs.**