

Response to Reviewer 3

Abramowitz and others critique methodologies used for model comparisons using outputs of the PLUMBER2 experiment. Honestly I like the conversational tone and general idea, but a few statements need more context and clarity. It's heartening to see that the modeling community is coming around to the idea that the best model for one site is just a simple model trained on a different site, and that the seemingly endless parameterization, boundless model complexity, and questionable spin up procedures is often overkill, at least when predicting fluxes over short time periods like half-hourly measurements to annual (or interannual) sums.

The time taken to give your considered feedback is much appreciated - we clearly agree! We have endeavoured to respond to each point raised below. Some involve changes to the manuscript, and others are simply points of clarification.

The abstract would be stronger if more quantitative. The first part of the paragraph beginning line 48 is an example (although the point in the latter part of the paragraph is extremely interesting).

As suggested, we have amended the abstract to be more quantitative, starting with the paragraph starting on line 48:

“Predictions from 7 out-of-sample empirical models are used to quantify the information available to land models in their forcing data, and so the potential for land model performance improvement.”

“In all but two cases, latent heat flux and net ecosystem exchange of CO₂ are better predicted by land models than sensible heat flux, despite seeming to have fewer physical controlling processes.”

We appreciate these are only two minor instances (although they include the suggested change), but feel that including more detailed specific results here would detract from the ability to present the broader findings of this work.

80: are these really shortcomings or is the best available approach at the time? As we learn more, many things become a shortcoming in retrospect but this might not be a fair comparison as innovation proceeds.

Yes, no disagreement, but given this is more or less the same team of authors we're happy to wear the criticism!

94: is the point here that pixel-scale estimates can never be validated?

'Never' is perhaps a little strong, since this is clearly dependent on resolution, observational constraint, and the heterogeneity of the landscape within the grid cell. One could imagine a heavily instrumented 1km x 1km grid box that could be 'validated' (although I dislike this term, as it is categorical), or alternatively, gridded evaluation products that truly attempted to quantify observational uncertainty in their estimates (rather than, say, the spread of available products). Our point here is really just that at this point in time, model constraint is much closer to being achievable at the site scale.

109: *this is a defined problem in and of it is related to the subpixel heterogeneity problem*

Yes, heterogeneity will always be a challenge for any gridded simulation, as it means that model simulations must represent emergent characteristics rather than be an explicit representation of the surface. But this section is discussing whether it is reasonable to evaluate models that are *inherently designed for gridded scale simulation* at the higher resolutions of their intended applications.

178: *some of the sites may have time-evolving data but there isn't an easy structure in flux network data compilations that make it easy to include it*

Yes, understood. It's also not likely to be available at a high enough proportion of sites for this kind of broad-scale modelling experiment.

195: *why wind speed, was friction velocity or the standard deviation of vertical velocity (σ_w) not considered for data quality control?*

Wind speed was used as a simple, easily available proxy. It does not need to be perfect, as it is just being used to understand whether or not the apparent poor performance of land models was in fact due to aggregate metrics being dominated by low turbulence periods... it did not make any qualitative difference.

222: *I'm happy to see that the authors critiqued energy balance closure-based "corrections" (quotes intentional) to data and not surprised to find that adjusting data lowered its quality. The flux and modeling community really needs to stop adjusting flux data because the lower-frequency flux transporting eddies responsible for much of the lack of closure needn't be strongly related to turbulent fluxes from the footprint.*

Yes, we agree 'corrections' should not be applied universally when the basis for these is not proven to be sound - it will only make prediction of 'observed' fluxes more convoluted and difficult. We do believe that some version of the machine learning approach here could form the basis of a metric that might go part way to discerning whether a given 'correction' is appropriate or not.

449: *'there is no a priori reason'...yes there's a very strong a priori reason because both Q_{le} and NEE are dominated by stomatal function (transpiration and GPP). There should be a strong relationship.*

Yes, sure, we have amended this text to read:

"Given the expectation that NEE is likely to be strongly dependent on site history, and that we could not reliably include this information in the modelling protocol or account for it in this plot, there is no a priori reason to expect a clear relationship across all sites here, beyond both fluxes being dependent on stomatal function."

455: *regarding this, were sites always compared against NEE measurements that include subcanopy storage for forest systems, or were eddy flux (F_c) measurements used as a surrogate for NEE? This is a great approximation for non-forested ecosystems but there will be a large bias, especially when partitioning GPP and RE, in many forests if storage flux isn't considered.*

Yes, eddy flux measurements were used, but we did not partition GPP and RE.

Figure 5: Why wasn't the random forest model included? From personal experience, XGBoost often emerges as the best fit for flux data, which is most similar to RF.

The random forest model is included. It is in the top row of subpanels, labelled "RF_raw" and "RF_eb". Perhaps the reviewer is asking why RF was not used as the highest performing reference model, since XGBoost worked well for them...? The answer, as noted in the text and shown in the figures, is that the LSTM outperformed the RF model out-of-sample, most likely because it includes internal states.

Why is color faded in some Fig 6 subpanels?

Figures 5 and 6 both have the first 10 subpanels faded. As noted in the caption "*The first 10 panels (faded) show empirical or physical benchmark models.*"

55: I'm not entirely convinced that this is a lower bound estimate. For a real lower bound wouldn't one want a method rooted in information theory that can estimate a theoretical predictability? I know that this is more of a thing for meteorology and perhaps not pertinent to this discussion, but perhaps one day it might be (e.g. 'predictive power' from Schneider & Griffies 1999: https://journals.ametsoc.org/view/journals/clim/12/10/1520-0442_1999_012_3133_acffps_2.0.co_2.xml)

We assume this refers to the use of 'lower bound' on line 559. The reviewer is correct that we could work harder to produce a better empirical model that could predict fluxes with more fidelity, and so provide a stricter, higher estimate of the lower boundary of predictability - this will always be true. Whether considered in an information theory framework or not, the process in this case will always be empirical rather than analytical, in a mathematical sense. That does not change the fact that 'true' predictability is higher than our best empirical model here – it remains a lower bound estimate.

Fig 8 I'd like to think that almost all wetland sites will have an evaporative fraction greater than 1; considering these as a unique case could provide even more context.

Interesting point, we've now mentioned wetland sites in two places in this discussion. Actually none of the sites described as wetland have an evaporative fraction greater than 1. But we would only really expect this if they were in a water-limited climate. Cold area wetlands presumably do not have the available energy to evaporate more water than is precipitated. Text changes are:

"A significant proportion of sites had Qle fluxes larger than incident rainfall, and since this is something that most LMs will be structurally prohibited from replicating (with the possible exception of wetlands), we explore why this might be the case, and whether the issue has biased our overall conclusions about LM performance."

"Of the sites in Fig. 8 with water evaporative fraction greater than 1, only one is irrigated (ES-ES2) and none are wetland sites."

676: I'm not against the statement that the correction is categorically incorrect, per se, but feel that the authors are stating that it's categorically incorrect for the incorrect reason. Between unmeasured storage flux terms including photosynthetic energy flux, mesoscale meteorological

motions, and the fact that advective flux does not necessarily follow the Bowen ratio correction, there are many physical, i.e. not modeling reasons for why the correction is incorrect.

We do not provide any physical reasoning for why the correction is categorically incorrect, just point out that we have a metric that can show it is. That does not make our metric “the incorrect reason”.

As an analogy, if a bath is overflowing because the plug is left in, it is not incorrect to state that we can measure the rate or amount of water flowing over the side. Many true statements can be made about this situation - for example that we can identify that there is a problem because the water is flowing out - that are not about identifying the cause.

760: only if the errors are independent

Yes, of course. It is, as we stated, a ‘crude analogy’. We have now added to this:

“A very crude statistical analogy might be that if we have a model with one process that is right 90% of the time, the model is 90% accurate. But if we have a model with 10 serial processes that are right 90% of the time, the model is $0.9^{10} = 35\%$ accurate (although only if errors are independent).”