We thank Reviewer 2 for their generous and useful comments on the paper. We have addressed each comment point-by-point below, and we hope we have sufficiently clarified those areas previously lacking in explanation.

> Page 1, Line 24: Please define or expand upon 'confirmation holism'.

We have expanded this sentence to read:

"*This problem is an example of* confirmation holism, *the idea that a single hypothesis cannot be tested in isolation from auxiliary hypotheses upon which it relies. The efficacy of a model's treatment of a particular process can only be tested within the structure and set of assumptions of that particular model. Observations typically only inform the result of a chain of process representations, so that a confirming result is holistic - we are unable to know whether a performance improvement is because of better representation, or because of compensating biases. Confirmation holism is discussed in-depth in a broader climate modelling context in (Lenhard and Winsberg, 2010).*"

> Page 2, Line 28: It seems like an unsubstantiated opinion to say that the selection of empirical models in Best was a ad-hoc. Please justify, rephrase, or remove.

Gab Abramowitz, the second author of Best et al. 2015 and this paper, was the developer of the empirical models used in PLUMBER. We have added a personal communications reference to the sentence.

> Page 5, Line 14: Ockham's Razor approach – please define this briefly for unfamiliar readers, or remove entirely if unnecessary.

We added a parenthetical remark to the sentence:

Where additional complexity did not substantially improve performance, we took an Ockham's Razor approach - that is, where no clear distinction in performance is evident, prefer parsimony - and used the simpler model.

> Page 7, Line 14: "The linear models out-perform the other models in most cases for NEE under the distribution metrics." I don't see this in Figure 2. The linear models have the highest rank (4 and 4.2) for distribution in NEE, which indicates the worst performance, correct?

Yes. We have removed that part of that sentence.

> Figure 3: Which metrics are being used here (e.g., all, common, extremes, etc.)? I'm guessing all metrics, but I don't think it's explicitly stated. Please note this.

Yes, it's all metrics. We added the following sentence to the figure caption:

All points are rank averages over all sites and all metrics.

> Page 7, Line 28: "at both resolutions" – the use of 'resolutions' in this context is confusing. Is this referring to the clustering? Please clarify.

We replaced “resolutions” with “cluster counts”.

> Page 8, line 8: The 10 day lag of H was chosen, but it seems like the 7 day lag could have been a good choice also. Figure 4 (Lagged RelHumidity) shows that the 10 day lag of H gives the best performance for Qle, but the 7 day of H gives the best performance for Qh and NEE changes little between 7, 10, and 30 day lags. Was the 10 day lag of H chosen (over 7 days) because it shows the best overall performance in any variable (i.e., Qle)?

10-days is better for Qle, and NEE. It is slightly worse for Qh, but we think 10 days is a better compromise, and it also slightly reduces the likelihood of correlation with instantaneous relative humidity. We added this text following the sentence mentioned:

In some cases, there were multiple lags with similar over-all performance gains. For these we chose one by selecting the variant that gave the best compromise performance increase between the three fluxes, as well as preferring lags towards the middle of the spectrum, so as to avoid correlation with instantaneous variables, and to maximise the available training data (longer lags means fewer windows with complete data available).

> Page 15, Line 28-29: “this indicates that our newer and more complex benchmarks are adding substantial performance improvements over the PLUMBER benchmarks.” This is a little unclear - the benchmarks themselves are better (have better rankings) or the LSMs perform better as compared to the new benchmarks?

We changed this sentence and the one before it to:

Despite this latter effect, we still see the LSMs generally falling in the middle of the range of empirical models for Qle under the common metrics. This indicates that our newer and more complex benchmarks are adding substantial non-spurious performance improvements over the PLUMBER benchmarks.

> Page 19, Line 1: Please define the Parerto principle.

We removed this sentence, as per reviewer one.

> Page 19, Line 15-29: I appreciate this discussion paragraph because I was asking myself exactly that question – and the answer was clearly articulated.

Yes, we don’t aim to antagonise the entire land surface modelling community (including ourselves!)

> Technical Corrections:
>
> 1) Page 2 Line 34 – might ‘be’ narrowed down.

Corrected.

> 2) Page 13, line3 – show to shown

Corrected.

3) Page 13, line 17: gradation to graduation?

No, gradation is correct.

4) Page 15, line 33: most complex benchmark, 3km27.

Fixed.