We thank Reviewer 1 for their and constructive comments. We have attempted to address them all as clearly as possible, point by point below, and hope we have allayed the reviewer's concerns about presentation.

> page 1: Sentence is unclear: imporvement using meteorlogical data alone. Does it mean that models are better using metor. data compared to surface data or that the added information from weather data improves the model

We changed this sentence to read:

We show that substantial performance improvement is possible for empirical models using meteorological data alone, with no explicit vegetation or soil properties, thus setting lower bounds on a priori expectations on LSM performance.

> page 1: I think there are some clear explanations related to non-consistent land-atmosphere interaction behavior

It is not immediately clear from the comment what explanations the reviewer is referring to, but if detail can be provided, we would certainly consider amending this sentence and adding to the discussion. Haughton et al (2016) explored a large number of mooted causes for the PLUMBER findings, and all were negative results.

> page 1: they can be also used in off line mode for hydrology or fluxes for isntance, please modify sentence

We added a sentence following this one:

LSMs are also routinely used in numerical weather prediction and offline hydrological modelling scenarios.

> page 1: maybe mention model complexity

We feel that this is more usefully discussed later in the paper. There is a paragraph devoted to the problem of complexity in LSMs in the discussion (paragraph 4)

> page 2: Mention also model complexity so that current model might not be able to provide good predictive power In fact, in economics model complexity is penalized not in the land surface community We have also moel overcomplexity as we cannot parameterize and test the parameter because of insufficient data

Yes, these are both good points, and something we discussed in the context of PLUMBER in Haughton et al (2016), but we're not sure this is the best place to discuss these points. The issue of complexity in LSMs is covered to some extent in the discussion.

> page 2: cite other papers e.g. Boone et al on AMMA intercomparisons

Added Boone et al (2009)

> page 2: for a given metric

This is already implied from the previous sentence, and we believe adding it would distract from the core message of this sentence.

> page 2: I would here refer to some classical papers such as Lorenz 1963 and 95

We added the following sentence:

More generally, this has been an acknowledged difficulty in numerical modelling for over half a century (Lorenz, 1963).

> page 2: list them

We added them as a parenthetical remark:

The simple empirical models used in Best et al. (2015) (univariate and multivariate linear regressions) have been used for decades, and come with an understanding of their power and limitations.

> page 2: strikeout hyphen in "over-all".

Replaced everywhere.

> page 2: The introduction should be increased to add discussion of work on the information and potential of weather data to estimate surface variables and fluxes starting with Noilhan and Mahfouf 1996, and then more recent work e.g., Salvucci and Gentine 2013, Ridgen et al. 2015, Gentine et al. 2016

We found two papers that might be the Noilhan and Mahfouf (1996) the Referee refers to, but could not clearly see the relevance of the work to this point. Nevertheless we have extended the introduction to make it clear that we are expanding on existing work. The final paragraph of the Introduction now reads:

However, the selection of empirical models used as benchmarks in Best et al. was somewhat ad-hoc (personal communications, 2016). In this paper we attempt to create a framework for assessing the overall predictability of land surface fluxes, by providing a more thorough exploration of the predictive power of empirical models using only meteorological forcing data as inputs. This extends recent work by Salvucci and Gentine (2013), Ridgen et al. (2015), and Gentine et al. (2016). We aim to provide a hierarchy of empirical models that each describe a priori estimates of how predictable land surface fluxes are, by providing a lower bound on best possible performance for a given set of driving variables. These models are able to be used as benchmarks for evaluation of LSMs. We also aim for this set of empirical models to exhibit a diversity of error patterns under different conditions, such that LSM evaluation might be narrowed down to specific failures under particular environmental circumstances (for example, poor performance during drought periods, or at a particular time of day).

> page 3: strikeout "in a global model"

Removed.

> page 3: citation ex Foken et al.

We feel that energy balance closure is ubiquitous in the use of flux tower data, and we did not follow this paper in particular, so have left the text as is.

> page 3: maybe use RH

We use H because it is clearer in the model naming scheme to have single-letter variables. We are aware of the common use of H as sensible heat, but as we use Qh consistently throughout the paper, we do not consider this to be particularly problematic here.

> page 4: strikeout apostrophe !

Fixed

> page 7: strikeout "At the most general level"

We feel that this provides contextualisation for the 4 panels, and so have left it in.

> page 7: Maybe describe those things in more detail but correlation is very high with Qh see FIgure 5. Maybe you should have more detailed sdisscision in the context of also Fig 5

It was not clear to us what was being suggested by the Referee here. We are happy to accommodate suggestions of course, but given this was a "maybe" suggestion, and unclear to us, we left it as is.

> page 8: Mostly because of boundary layer dynamics, see Gentine et al. 2016 and Noilhan and Mahfouf

We have amended this text to read:

Longer lags of S and H help the prediction of Qle, perhaps because these variables act as proxies for soil moisture, ground heat storage or boundary layer dynamics (e.g. Gentine et al 2016)

> page 9: I wonder if this should not ahve come first as the correlation will ionfrm the statistical analysis

We already knew that the fluxes were reasonably highly correlated, and that SWdown was also very highly correlated with each flux (this is the reason for the 1lin model in Best 2015. This plot could not have reasonably been generated prior to the lag analysis. It is only really useful to ensure that proposed driving variables are not highly correlated…

> page 10: You can also mention that many sites do not have longwave and thus better not to use it

We changed the sentence to read:

This may be due to the low quality of L in the datasets, and complete lack of L in over a third of sites (see Figure 4), which would minimise the data available both for training, as well as for evaluation.

page 12: strikeout "a"

Corrected to "an".

page 12: better explain?

We replaced this sentence with the following paragraph:

*We aimed to generate an objectively "best" ensemble that evenly spanned the range of performance in each variable, and maximised behavioural diversity. For example, we might expect that models with instantaneous humidity would exhibit different patterns in their outputs after a rain event than models that do not include humidity. Likewise, models with lagged rainfall averages as drivers should also have a differing behaviour in the period after a rainfall event.

We initially attempted a pseudo-optimisation based ensemble generation approach. …*

page 14: give units for RMSE and compare to total mean value

This would quickly get complicated, due to the diversity of units among metrics. Instead, we used the empty space in the bottom right corner to include distribution plots of the three fluxes at each site, for comparison with appropriate metrics. Each plot also includes a boxplot of the site means for each flux, and the units of each flux is included on the y-axis of each of these plots. We hope the reviewer finds this sufficient.

page 15: strikeout "theoretically"

Fixed.

page 15: strikeout and move "figure 8"

We prefer this format.

page 15: explain better

We added a parenthetical remark:

… that the LSMs are all consistently beaten by even the simplest empirical models (LSMs in black are consistently above S_lin, pink, and ST_lin, red).

page 17: The difference in the Q_le vs other fluxes is really interesting. Why are LSMs doing so well? Is this because of the more pronounced daily cycle, memory because of soil moisture? Maybe elaborate more, for instance H and NEE are more dependent on atmospheric turbulence (strong variations), there is less clear physical constraints (no potential ET for instance)

We added a note at the end of the paragraph discussing figure 8:

It is notable here, as in (Best, 2014), that the LSMs before reasonably well under Qle relative to the other two fluxes. The three fluxes operate very differently, and so it is not clear why this performance difference exists, but it may be due to e.g. tighter constraints on Qle from upper level soil moisture (which is not available to the empirical models), or it may be that the boundary layer turbulence affects Qle less strongly than the other fluxes.

> page 17: using metorological data

We added " using meteorological data alone" to the end of the sentence.

> page 19: strikeout "However, the Pareto principle would suggest that further gains would likely be less substantial and require more effort."

We removed this sentence.

> page 19: you coudl mention site, for isntance soil type, vegetation type… LSMs have this information as well

This is what the "site characteristic data" refers to. We have clarified the sentence:

There is also no doubt that performance could be further increased using similar models with additional site characteristic data, such as soil composition, vegetation structure, and orography.

> page 19: This wasn't really true for lE

This could be interpreted both ways: in Figure 10, individual LSMs are not clearly better than the empirical models in more than a couple of cases, even for Qle. And these empirical models have only a subset of the information provided to the LSMs. This does not inspire confidence in models that are supposed to encode the physics of the systems they are modelling, and we think the point is still valid.

> page 20: yes should have been in the discussion

The point was added above. Since we did not look at site characteristic data, we find it of limited relevance to add detail that would be better discussed in other contexts. We hope that the above clarification satisfies the reviewer, but we are happy to add more relevant detail to the discussion, if the reviewer has particular points they think should be raised.