Anonymous Referee #2

Received and published: 18 May 2018

> This manuscript presents a methodology for quantifying the "predictability" of land atmosphere fluxes of water, energy and carbon across 155 eddy covariance sites, with the goal of helping to better interpret comparisons between these observations and output from land surface models. This idea has considerable merit, and could be of interest to a large number of land surface model developers, and other synthesizers of eddy covariance data sets. Unfortunately, in its current form it is difficult to extract the most important information, as there is insufficient emphasis on what might be valuable, and much material is included which is not relevant. Overall, it needs to be much more focused, and the authors need to concentrate on: (i) predictability and their predictability metric; (ii) the models used, whilst greatly streamlining the hypotheses, given the inconclusiveness of the majority of the analysis.

We thank the reviewer for their insightful comments, and are pleased they see merit in our approach. We had attempted to answer each individual comment below, and we think that in doing so, we have made some substantial improvements to the paper.

We note in particular the general concern with the lack of clarity around the definitions of predictability and uniqueness. We have edited the manuscript in multiple places in an attempt to clarify our meanings here, and have separated the use of the words "predictability" and "uniqueness": "predictability" is only used for the general sense of over-all predictability, and "uniqueness" is only used to refer to the metric described in the paper.

We also note the reviewer's concern about the affect of using multiple different models, and we have added a number of clarifications and extra analyses to the paper, which are described below.

**General Comments**

1. As the authors acknowledge, "there is no single definition of predictability" – and this is a key challenge to this paper. The introduction needs to address this much earlier than the bottom of page 3, so there aren't several pages of text discussing something that has not been defined, or at least how it is being treated in the context of this manuscript.

We have changed this sentence to read:

*Predictability can broadly be defined as the ability to reproduce a property of a system, given only knowledge of variables that are causally related to that property.*

We think that that this is an all-encompassing definition of predictability. What is less defined is how to measure predictability - it is possible that there is no single metric that will cover all aspects of predictability. The following paragraph already covers some of those in an effort to illustrate that idea. We don't think that the broad concept of predictability, the meaning of which is fairly self-evident in the word itself, really needs to be spelled out more than this earlier in the introduction.

2. It seems that the authors are treating predictability as the inverse of "uniqueness", which is characterized as the deviation between a globally optimized model, versus

a locally optimized model. This needs additional clarification and justification. With this definition, the predictability is inherently model dependent, rather than some intrinsic property of the site alone. This is always the case possibly, but needs to be spelt out. It also highlights the importance of the models.

We are treating non-uniqueness as one aspect of predictability.

Predictability is an aspect of the relationship between flux datasets and meteorological datasets (and possibly other datasets). There is most likely no way to characterise the strength of that relationship other than by modelling it using data mining/statistical techniques. Information theory techniques might work conceptually, for example we could use something along the lines of mutual information for relationships between single met variables and single flux variables, but multivariate mutual information is a underdeveloped technique at best, and also does not capture transformations of variables (such as the impact of historical variable values) that are possible when modelling.

Yes, the metrics are somewhat model dependent, but the models we have used are structurally data agnostic, and the driving variables were chosen specifically because they add predictive value to simulations. In any case, predictability is necessarily a relative measure, and as long as the same procedure is used across all sites, it can be informative regardless of model choice. We also hope that the changes that we have made to Figure 1, and each of the GAM plot figures helps to alleviate concerns about model sensitivity potentially having a major impact on the qualitative results of the study.

3. A reliance on this uniqueness as a proxy for predictability seems like it might have drawbacks. Consider the NEE plots in Figure 1. There are a group of sites with relatively high "uniqueness", and thus low predictability, but with a global RMSE less than 7, which is lower than for a large fraction of sites indicating a better mean performance. Are these sites more or less predictable? It seems that this can only be quantified through combing the uniqueness metric with the mean performance into a single metric, but uniqueness is discussed in isolation throughout the results section.

Again, uniqueness is one aspect of predictability. On reading through the paper, it is clear that we were not always rigorous in our terminology. We have now been through the manuscript, and have made sure that every case where "predictability" is used, that it's actually in the general sense, and that it is always clear that uniqueness is simply a major component of predictability. We use "predictability metrics" to refer to all combinations of mean/uniqueness and RMSE/Corr/Overlap metrics.

4. This can be addressed by combining the metrics with appropriate weights. The authors say this is not done due to the difficulty in combining the different metrics, but given the lack of information in the metrics other than RMSE, and the apparent requirement to combine uniqueness and mean performance, this should be reconsidered, with at least these two components of each metric combined.

The problem with combining metrics is that you lose a substantial amount of nuance. This is particularly the case when one metric has a strong, but meaningless trend in it - for example RMSE over mean temperature (this is obviously not meaningless in the broad sense, but it is not useful for predictability assessment, except as something to be aware of). On top of this, a sum (weighted or

not) suffers from the problem that the mean metric is linear, while the uniqueness metric is radial, and so the uniqueness metric dominates at small RMSE values, and the mean dominates at large RMSE values.

In our results, there are very few cases where the mean performance metric has any clear patterns in it that are either not explainable in this way, or not already visible in the uniqueness metric. As such, while mean metric performance is important to take into account, uniqueness is the more important and interesting component of predictability, as far as comparing FLUXNET sites goes.

5. What is the additional information gained from switching from Cartesian to polar coordinates? Would not a simple mean of the global and local models, and the normalized difference between then suffice?

The benefit of using polar coordinates is that uniqueness is orthogonal to mean performance. Using the angle between the metrics is almost equivalent to using the 1 - log(local/global), except that the normalisation is problematic for metrics that often have values below 1. The angle method is the same for all metrics, normalisation would not be.

6. This manuscript relies heavily on previous work (Best et al, 2015, Haughton et al., 2018). Indeed, it is not possible to understand much about the models with out consulting this closely. Given how dependent the predictability metrics are on these models, some further description of them is required here.

The models are relatively simple, and there is a not a lot more to describe, but we have expanded the model description somewhat. This point is substantially similar to one of Reviewer 1's comments, and we copy that here:

The beginning of the 4th paragraph of the Methods already describes the cluster-plus-regression models - they are conceptually simple, and there is not a lot more to say, but we have split this paragraph in two, and changed the first part to add some description of the use of the cluster-plus-regression models:

*This procedure is model-agnostic, and we have used models in the framework developed in Best et al. (2015) and Haughton et al. (2018), because they are conceptually simple, but able to fit complex functional relationships. These models (listed in Table 1) include some simple linear regressions, as well as cluster-plus-regression models. The cluster-plus-regression models consist of a K-means clustering over meteorological driving data, and then an independent linear regression between drivers and fluxes at each cluster. These cluster-plus-regression models can fit arbitrary functional forms between predictor and response variables, when using a high enough cluster count (k), and given enough data. The models are not perfectly deterministic, since K-means convergence is dependent on cluster initialisations, but the variance in the results is small (see supplementary material, Haughton et al., 2018), and unlikely to substantially affect our results substantially. Our use of an ensemble of models at each site further mitigates this problem. The ensemble also allows us to overcoming the problems of the simpler models failing to capture behavioural nuances, and of the more complex models failing to train at some sites due to insufficient data (described below).*

We also moved the note about which fluxes are modelled to the top of the methods section.

We have added the long-form of each long-term/short-term model name to the table.

7. Indeed, it's very unclear why multiple models are being used at all? What is the benefit of doing this rather than using the single "best" model?

The "best" models - the most complex ones - often fail locally due to not having enough training data. On the other hand, the simple models clearly only capture a subset of the behaviours at any given site. By using an ensemble of models, we also minimise model-related variance. We have modified what is now the 3rd paragraph of the methods section to address these points:

*This procedure is model-agnostic, and we have used models in the framework developed in Best et al., 2015 and Haughton et al., 2018, because they are conceptually simple, but able to fit complex functional relationships. These models (listed in Table 1) include some simple linear regressions, as well as cluster-plus-regression models. The cluster-plus-regression models consist of a K-means clustering over meteorological driving data, and then an independent linear regression between drivers and fluxes at each cluster. These cluster-plus-regression models can fit arbitrary functional forms between predictor and response variables, when using a high enough cluster count (k), and given enough data. The models are not perfectly deterministic, since K-means convergence is dependent on cluster initialisations, but the variance in the results is small (see supplementary material, Haughton et al., 2018), and unlikely to substantially affect our results substantially. Our use of an ensemble of models at each site further mitigates this problem. The ensemble also allows us to overcoming the problems of the simpler models failing to capture behavioural nuances, and of the more complex models failing to train at some sites due to insufficient data (described below).*

8. Although the authors suggest they want to leave this for future work as it is "substantially more complex", it seems at least some examples are required to explain how the predictability metrics are sensitive to the models, and how this can be interpreted when discussing specific sites. For example, it seems that many semi-arid sites are characterized by these models to have high uniqueness, but what would happen if soil moisture was included in the model?

We have updated Figure 1 to now include short_term243, in order to show how much the change in QC affects a more complex model. The paragraph describing it and the figure and caption have been modified to read:

*The uniqueness and mean performance metrics are shown for RMSE in Figure 1 for the S_lin and short_term243 models to illustrate how to interpret later figures:* uniqueness *is the angle measured clockwise from the origin (the optimal metric value) and the 1:1 line (equal local and global performance), and* mean performance *is the average performance of the local and global simulations, given by the distance of each point from the origin. Each point is a different site. Figure 1 also illustrates the differences between the results when the local training data is identical to the testing data, and when it differs due to mismatch between the meteorological and flux QC flags between training and testing. In each panel, the blue points indicate the local and global RMSE values used for the simulation in the remainder of the study. The tail from each point indicates where these values would have been if the same QC data that was used for training was used for evaluation (meteorological + flux QC, instead of just Flux QC. The tail points are strictly at or below the 1:1 line (as the empirical fit is optimised for RMSE locally, but not globally). The flux-only QC evaluated blue points can shift, and some lie very slightly above the 1:1 line. Tails pointing towards the origin indicate that these simulations' mean RMSE is worse than it would be using the training QC. Tails pointing clockwise indicate the these simulations appear to be less unique under RMSE than they would be using the training QC. Perhaps surprisingly, the differences for the simpler model appear much more variable, but we also note that most of the larger discrepancies result in similar changes using the global and local evaluation, meaning the bias is mostly in the mean performance, and less so in the uniqueness metric. We considered the option of using the training QC flags for*

*the evaluation period, however this would result in different models having very different evaluation periods.*

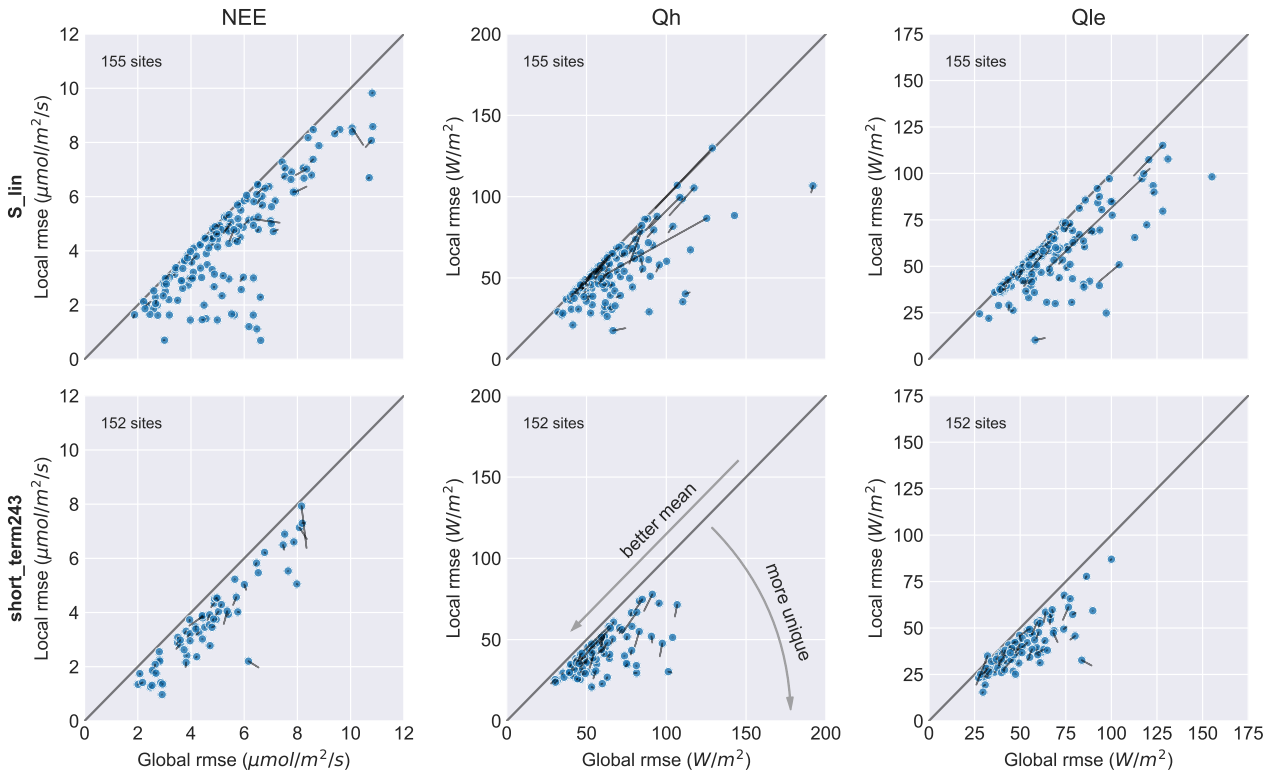


Figure 1: RMSE values for the global (x-axis) and local (y-axis) model simulations. Columns show the three fluxes, the forst row shows data for S_lin, the second row for short_term243. The tails of each point show where the local and global RMSE values would be if the same QC flags were used for training and evaluating (the intersection of meteorological and flux QC flags). Tails pointing toward the zero in each axis indicate the model would have performed better using these QC flags. In other words, a tail pointing towards the origin means that our evaluation method has a bias toward worse mean RMSE, and a tail pointing clockwise from the origin indicates that our method has a bias towards lower uniqueness.

We have also updates all of the GAM plots (Figures 2, 3, 5, 8, 9, 10) now include separate GAM fits for the longer-term models and the linear models. We added the following text to the paragraph describing Figure 2:

*We note that the two subset GAM plots for linear and longer-term models describe a similar pattern in each metric in most panels, here and in later plots. The main differences seem to be largely to do with the more complex models' ability to capture more of the variance: the mean performance of these models under each metric is better (and the linear models' worse) than the mean, and the uniqueness is higher for the Corr and Overlap metrics, but quite similar for RMSE.*

Unfortunately soil moisture is not reliably available in the FLUXNET dataset, which is a large part of the reason this paper and Haughton et al (2018) did not include it. The humidity and rainfall lags in the long_term243 model act as a proxy for soil moisture, however, the long_term243 model

did not run successfully locally at enough sites to provide an adequate picture of predictability contingent on those proxies.

9. Whilst uniqueness as defined here is certainly a useful metric to assess flux sites by, and to help interpret comparisons between observations and land surface model output, it is unclear it represents something like the inverse of "predictability". In fact, a contrary argument could be made that the sites that exhibit large reductions in model error when optimized with local data are the most predictable. Whilst for those sites that don't see model improvement when just local data are used this lack of sensitivity might also be interpreted as a lack of predictability, particularly for sites with low mean performance. In this context, a predictable site is one where given more information, model skill increases, and whilst at an unpredictable site specific information does not increase skill.

This was a very thought-provoking comment, and provided us with the impetus to try to set out the relationship between predictability and uniqueness and mean performance more clearly.

We have added the following discussion of the relationship between uniqueness and predictability and corresponding figure to the beginning of the discussion:

*In our exploration of meteorological predictability, which we characterised using both uniqueness and mean performance for each metric (described in the Methods), we have mostly focussed on uniqueness of behaviour of sites. This metric is the most novel component of this study, and is, we think, the most interesting aspect of relative site predictability. However, it is also less intuitive than mean performance. It is worth reiterating that uniqueness is* not *the direct inverse of predictability, and under certain conditions it can actually be correlated with overall predictability. In an effort to clarify uniqueness in an intuitive way, we illustrate the relationships between meteorological drivers and fluxes, and site-specific (non-meteorological) modulators of these relationships in Figure 11.*

*Figure 11 provides a schematic for understanding how mean performance and uniqueness interact as components of predictability. In all cases, we assume that the observations include some noise, or unpredictable components of variability. In the left column, we present a case where there is a strong universal relationship between meteorological forcings and flux observations (in this case for example, a linear trend), in the right column, this relationship is weak or non-existent. In the top row, we have the case where the local site conditions modulates the behaviour of the fluxes (in this case for example, a simple bias, but it could also be a complex non-linear relationship), and in the bottom row, there is no such site-related modulation. In the right column, where there is no meteorological driver relationship, we can see that uniqueness is correlated with performance: in the absence of site modulators (bottom right) there is no predictability (as all variability is due to noise) and both models perform poorly. In the presence of site modulators (top-right), predictability is higher and uniqueness is also higher (because the local model performs better). When there* are *strong meteorological drivers (left column), predictability is higher in the case where site modulators are weak (bottom left) because the global model is able to perform well. In this case, sites with strong site modulators are less predictable because the global performance is worse, and uniqueness would clearly be anti-correlated with "predictability". It is clear that there* are *strong relationships between meteorological drivers and fluxes (see Best et al., 2015, Haughton et al., 2018), and so the inferences made in this study mostly fall in the left column. This suggests that in real-world situations uniqueness is mostly anti-correlated with predictability, but not perfectly, and so we emphasise that uniqueness alone is not an adequate proxy for predictability.*

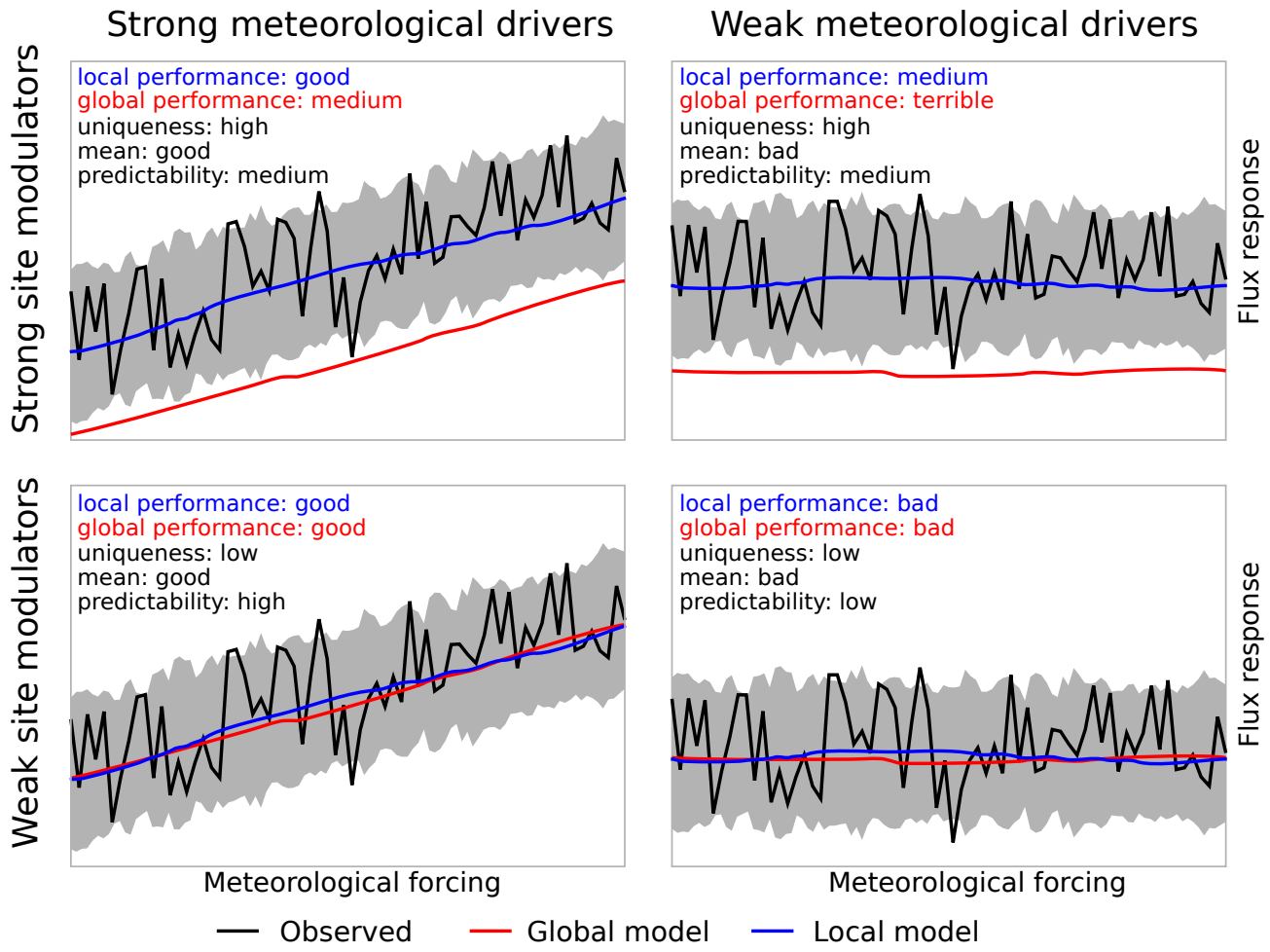


Figure 2: Schematic of sources of variability and how they affect predictability, mean performance, and uniqueness. The black line represent the flux observations, and the grey ribbon, the unpredictable internal variability or noise in the system. "Meteorological drivers" indicate universal physical relationships between meteorological forcings and fluxes, here we have indicated this using a simple linear trend as an example (red and blue lines). "Site modulators" are characteristics of the site (soil or vegetation properties, storage pools, geography, or data problems) that modulate the meteorological -> flux relationship such that it is different to the relationships observed globally - here we have used a simple bias as an example.

10. Both sections 2.2 and 3.2 read as an overly long laundry list of "everything we tried". It will be easy to greatly increase the overall focus of the manuscript by addressing this. Given the lack of conclusiveness regarding the majority of the hypotheses about determinants of predictability, a brief note that they were considered and findings were inconclusive is all that is required.

We have significantly reduced the information in both of these sections. See our response to Reviewer 1's comment on Pg 9-11.

**Specific Comments**

P1 L1-8. Rather like the manuscript as a whole, the abstract needs much more focus. Emphasis specific detail, not background information and motivation.

We have re-written the abstract to include less background and more results. It now reads:

*The FLUXNET dataset contains eddy covariance measurements from across the globe, and represents an invaluable estimate of the fluxes of energy, water and carbon between the land surface and the atmosphere. While there is an expectation that the broad range of site characteristics in FLUXNET result in a diversity of flux behaviour, there has been little exploration of how predictable site behaviour is across the network. Here, 155 datasets from the Tier 1 of FLUXNET2015 were analysed in a first attempt to assess individual site predictability. We defined site* ***uniqueness*** *as the disparity in performance between multiple empirical models trained globally and locally for each site, and used this along with the mean performance as measures of predictability. We then tested how strongly uniqueness was determined by various site characteristics, including climatology, vegetation type, and data quality. The strongest determinant of predictability appeared to be that drier sites tended to be more unique. We found very few other clear predictors of uniqueness across different sites, and in particular found little evidence that flux behaviour is well discretised by vegetation type. Data length and quality also appeared to have little impact on uniqueness. While this result might relate to our definition of uniqueness, we argue that our approach provides a basis for site selection in LSM evaluation, and invite critique and development of the methodology.*

P4 L13-21. Not methods.

We have moved this section into the introduction, and have adjusted the surrounding text to suit this change.

P8 Fig 1. Don't understand the need for colored dots?

As stated in the caption, "Colours simply serve to identify sites, and allow clearer comparison between the top and bottom rows".

P12 L29. Yes, you might want to [QUANTIFY] that

Whoops! We changed this to (< -5°C).

P15 Fig 4. Presumably it is the mean values that are being plotted here?

Yes, we added a parenthetical remark to the figure caption to note this. The first line of the caption now reads:

*Predictability metrics for mean annual temperature vs mean annual precipitation (mean across models).*

P15 Fig 4. Seems like NEE needs a different scale?

Yes. Since the relative differences are of interest, not the absolute values, instead of giving NEE a different scale, we have multiplied the NEE mean RMSE by 10, and added a note to that effect to the caption.

P16 Fig 5. The two sites with an aridity index higher than 3 haven't been excluded.

This was left over from a previous version of the figure. In any case, we have removed Aridity from the paper and put it in the supplementary material, and have adjusted the caption to suit.