Anonymous Referee #1

Received and published: 17 May 2018

> The manuscript presents a method for measuring the "uniqueness" of sites based on the ratio of model performance when trained at site level compared to when trained across all sites. I find the rational and motivation for the study to be very relevant, which looks to quantify and empirically examine the the experience that everyone who has worked with diverse datasets such as FLUXNET knows, that it is easier to get good model performance as some sites than others. However, I find the manuscript in it's current state to be rather unfocused, needing more synthesis to focus on key hypothesis and findings and focusing on what the metrics can explain (not what they can't explain) and what is most useful for the users of FLUXNET and other datasets. There are some good outlines of possible avenues for analysis on page 24, lines 18-24, which are discounted as being too complex. While I can appreciate that this synthesis work is complex, the lack of a clear message really hinders the usefulness of the paper as is.

We are glad the reviewer found the concepts in the paper relevant, and thank them for their comprehensive comments. We agree that the paper needs a tighter focus, and have found the reviewers comments valuable in making those improvements. The reviewer also presents many useful ideas for extended analysis to make the method more rigorous, and we have attempted to implement as many ideas as we could, without expanding the paper dramatically.

We note in particular the Reviewer's concerns about site and model selection, and the possibility of variance in the results. We have added some analysis to more adequately assess the impact of the multiple models, and made a number of other changes in an attempt to add some clarity to the paper.

Responses to individual comments are provided below.

Some possible ideas:

> Sensitivity of the metric such as within the empirical ensemble: how dependent is the metric on model used, QC of predictor variables in the empirical models, site selection when calculating the metric: does it change drastically if the global run is only performed on a subset of sites (how many sites are needed?).

We have extended the analyses of each hypothesis by adding two separate GAM plot lines to each GAM figure, one for each of two subsets of models: the two simple linear models, and the models with longer-term averages included (short_term243, long_term243). These serve to show that while there are certainly differences between the models in terms of performance - the more complex models do perform substantially better in the mean, and also pick up more uniqueness - the relative differences between the sites for the different model types remains qualitatively the same in all cases. We have added the following to the paragraph describing the GAM plots:

We have also fitted two other GAM models using subsets of the model ensemble: In each such plot, the red line represents a GAM fit using only the linear regression models (S_lin, ST_lin), and the purple line represents only the models with lagged input variables ("Longer models" - short_term243, and long_term243). These serve to show any differences in the predictability metrics that are contingent on model complexity, non-linearity, or input variables.

We have also added text to figure captions and results text where relevant, in particular to the first GAM plot. We added the following to the paragraph describing this plot:

We note that the two subset GAM plots, here and in later plots, describe a similar pattern in each metric in most panels. 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.

We don't think it would be useful to run separate analyses for different sets of QC flags, as discussed below in response to the comment on Pg 7, line 10.

The Global run should not change drastically based on a random subset of sites. The nature of the data is similar at each site globally, and we are using regressions with few parameters, relative to the number of data points, so there is very little chance of over-fitting the global dataset.

> Based on your analysis, what are the most and least unique sites with respect to each flux?

This question is answered in Figure 11, and we have modified it further with numerical values to make the differences between sites explicit. In writing this paper, we considered making site recommendations, but ultimately decided that such statements could end up being more misleading than useful. Predictability is potentially a very useful metric to have available when deciding on sites to use for a study, but ultimately each study will have different aims, and other factors such as vegetation representation and climatic zone will also play a large decision in site selection, which will be tailored for those aims. We have added a sentence to this effect in the second paragraph of the last discussion section:

We intentionally avoid recommending a particular set of most or least predictable sites, as the suitability of a given set of sites for a particular study is going to be dependent on many factors.

> Possibly framing the analysis as a variable selection framework to predict uniqueness.

This study is not a variable selection framework. Haughton et al. (2018) provided the variable selection framework that this study uses.

Abstract

> Nearly half of the abstract is motivation, while containing no tangible results or discussion.
>
> Pg 1, line 12: "A number of hypotheses potentially explaining site predictability were then tested..." This is very vague, could at least give number of hypothesis, or focus on those that are most important.

We agree, and have reduced the amount of motivation, and included some more specifics about results in the abstract, which 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 with 30 minute temporal resolution 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 little evidence that flux behaviour was 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 useful basis for site selection in LSM evaluation, and invite critique and development of the methodology.*

Introduction

> Pg 2, line 3: "Perhaps surprisingly, the predictability of a site is rarely considered when choosing sites to evaluate models." I would argue this is the key motivation, but also with the caveat that predictability is likely considered (either explicitly or implicitly) but not quantified and often not discussed.

We agree that this was poorly worded, and have changed the sentence to read:

Perhaps surprisingly, the predictability of a site is rarely considered explicitly when choosing sites to evaluate models.

> Pg 2, lines 5-12: While I can appreciate that the study is motivated from a LSM perspective, they are not evaluated in the manuscript, making this paragraph unnecessary. Furthermore, predictability may be useful in other contexts such as empirical upscaling of fluxes (Tramontana et al 2016).

This is a good point. We have removed the irrelevant historical assessment, and focussed on the more important point that PFTs are an unknown factor, as far as predictability goes. The paragraph now reads:

Modern land surface models (LSMs) attempt to describe the exchange of energy, water and, more recently, carbon, by explicitly representing the soil-vegetation continuum (Pitman, 2003). Common to virtually all LSMs is an assumption that flux behaviour variations between biomes, given similar driving conditions, can be explained by a small sample of structural and physiological parameters, grouped as plant functional types (PFTs). As a result, land modellers have sought observations from locations characteristic of these broad PFTs to develop and evaluate models. However, the actual practical representativeness of PFTs of the underlying vegetation properties has only recently begun to be investigated (e.g. Alton, 2011), and no explicit empirical assessment of PFTs as a driver of predictability has been undertaken.

> Pg 2, lines 35-36: "...were not able to identify any obvious patterns in model performance across sites." patterns with respect to what?

With respect to differences between sites. However, we have decided to remove this sentence. It was a distraction from the main message of the paragraph.

> Pg 3, lines 8-24: Is there an indication that any of these studies would potentially have a more/less predictable subset of sites? Not to pick on any one study, but are there any indications they could have a selection bias that would benefit from the uniqueness metrics you are proposing?

We do not know, and we would prefer not to speculate, as this is not really knowable without replicating those studies on other datasets. However, this is an interesting question, and we have added the following sentence to that paragraph to address it:

Whether or not any of the studies mentioned above are biased by a lack of consideration for predictability is unknown, because this this was not part of the selection process for the sites chosen.

> Pg 4, line 3: Does the method presented here not have the same assumption? For example, if the empirical model was a random number generator the RMSE between local and global would be the same and uniqueness would always be 0?

Yes, it would, on average. We agree the that the metric is only useful when used with models that have some predictive power. The Kaboudan (2000) metric is also univariate, so we have changed that sentence to read:

Kaboudan (2000) provides another univariate predictability metric.

> Pg 4, lines 7-11: Here the outline tells that the manuscript will both be an exploration analysis ("investigate several hypothesis"), but then promises a sound theoretical basis for site selection. It would be useful to outline how the hypothesis you explore will lead to concrete methods that are useful to modelers, because at current state the manuscript requires a fair amount of digging in order to get any idea as to what sites will be more or less predictable.

We disagree that this section promises anything - it simply sets out aims. However, in an attempt to improve clarity, we have changed the last sentence of the introduction to read:

This will allow expectations of model performance to be better defined by providing a priori *estimates of local predictability based on site characteristics. We hope this can provide some mitigation of the potential for ad-hoc site selection to shape judgement of how well LSMs perform.*

Methods

> Pg 4, line 24: The empirical models really need to be explicitly describe in the manuscript, seeing as they are the basis for calculating your metric. Furthermore, is it necessary to use this particular suite of models or will any empirical model do?

We removed the parenthetical remark about the models from this paragraph, so that the paragraph is model-agnostic.

See also our response to Pg 5, line 10-25 comment.

> Pg 4, line 29: This is also a subset as you do not have infinitely many sites. How robust is the metric to site selection. How variable is the metric when performed on one subset of sites to the next?

As noted above in response to the first comment, we expect the global simulation to be very stable, due to the nature of the models, and the size of the dataset. The metric itself should be more stable due to being an average across multiple models. The metric will vary between sites as the local performance varies, but this is the point of the exercise.

> Pg 5, line 10-25: It seems this work is very reliant on the previous works (Best et al. (2015), Haughton et al. (2016), and Haughton et al. (2018)) and as such these studies should be outlined more. In the current state, one would need to read the previous three papers to understand the core methodology presented in this manuscript. For instance, there is no citation for the cluster-plus-regression methodology. Furthermore, the long_term and short_term notations are never introduced.

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.

> Pg 6, Table 1: Did the long_term_729 model never have a successful simulation? In which case it was never actually used? Also, for the simple models it seems that negative uniqueness happens quite frequently for corr and overlap, [unreadable in PDF]

This is correct. To save on confusion, we have removed all mention of the long_term729 model from the paper, and adjusted text to suit.

The second part of this comment was truncated in the review PDF, so we may have missed part of it. Yes, these metrics suffer more from negative uniqueness with the simple models.

We have looked more into the cause of the overlap in correlation, and it appears that most of the problem in correlation is very small - all of the negative correlation uniqueness are $>$ -10e-14 for S_lin. We have set a negative uniqueness threshold of -1e-8 instead of 0 in all cases, and updated the table to reflect that, and added a note to the table caption.

Some are down to -1.7e-2 for ST_lin which can be accounted for by the different weights on the two input variables.

For the Overlap metric it is possible for the global model to produce a stronger trend, e.g. because the training dataset includes more diversity relative to noise. This would result in a higher variance, which would counteract the smoothing effect of the linear regressions, and produce a better overlap score for the global model, resulting in negative uniqueness. We added the following sentences to the paragraph that mentions the negative uniqueness section of the table:

We note that ST_lin Corr has a relatively large number of negative uniquenesses, which can be accounted for by better estimates of the S and T variable coefficients in the global model. There are also negative Overlap uniquenesses in the linear models, which is likely due to the global model training resulting in a stronger trend, and thus a higher variance, counteracting the fact that empirical models are generally smoothers.

> Pg 6, line 1: How would one interpret the "Mean performance" metric? Is there an advantage of this method compared to simply taking the arithmetic mean of local and global RMSE rather than the distance from the origin?

It is the same thing. We have changed the sentence to read:

Mean performance: the arithmetic mean of the local and global metric at each site, defined by distance from the origin.

> Pg 7, line 10: Does this suggest that the uniqueness metric, when using the different QC flags between the training and prediction runs, is combining both the effects of information content of the predictor variables and the gap filling? As you state the difference can be large with the complex models, as such this should be reported. Would it make sense to always use the same QC flags for your analysis?

This is a problem that we struggled with. Yes, it is combining the effects of predictor variables and gap-filling. We added this sentence to the end of the paragraph:

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.

We have also updated Figure 1, and adjusted the caption to reflect this. Now, row 1 shows all of the data that was in both rows, using "tail" lines to join the dots. We have added another row using short_term243 as an example of a more complex model. This figure now more clearly shows the variance added to the RMSE due to the discrepancy between the flux-only and met+flux QC flags. We have updated the paragraph in text to describe this, and the implications shown in the graph:

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.

The updated figure and caption is included here.

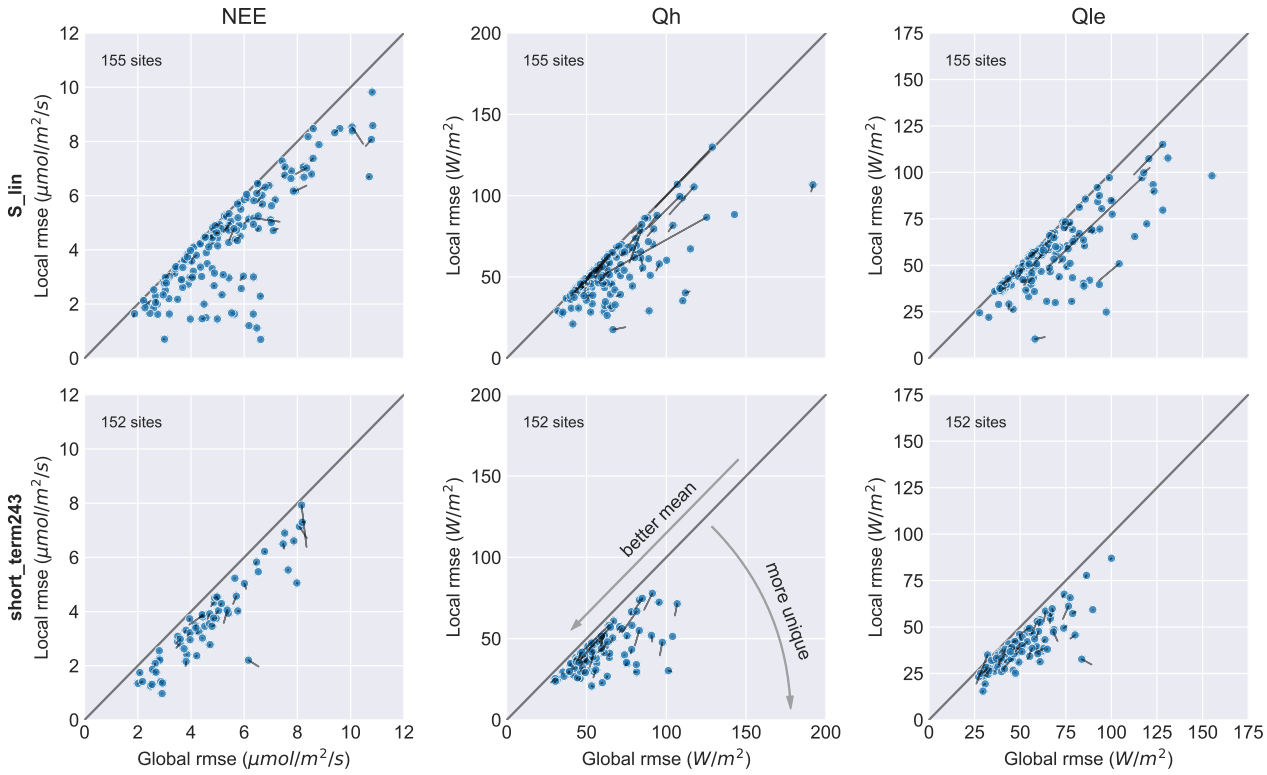

Figure 1: RMSE values for the global (x-axis) and local (y-axis) model simulations. Columns show the three fluxes, the forst row shows data for S_lin, the second row for short_term243. The tails of each point show where the local and global RMSE values would be if the same QC flags were used for training 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.

> Pg 7, line 17: Possibly using a combined summary statistic could simplify the procedure a bit? e.g. Gupta et al 2009

As noted in Gupta et al. (2009), their combined statistic still suffer from a pareto optimality problem,

in that there are multiple not-strictly-worse choices for weighting any combination. Additionally, such combined metrics reduce transparency about where problems are coming from. For these reasons, we felt it was more sensible to keep the metrics separate.

> Pg 7, line 24: Was there any attempt to prevent over-fitting, such a cross validation. How resistant is the cluster-plus-regression model to over-fitting?

Given enough data, the models are not sensitive to over-fitting. However, over-fitting on the local data is effectively part of the methodology. That is, the question is, how much does the functional form of the local data differ from the global data. That difference may be due to measurement error or physical site differences (these are not separable from a data perspective).

Of course, there is the possibility of noise in the local training data skewing the metric for individual sites. Our choice to use a minimum sample size of 5*n_vars per cluster for the cluster-and-regression models limits over-fitting to a large degree, while being small enough to not exclude all sites. On top of this, we are using multiple empirical models. These models should have minimal biases, and those biases should cancel when taking the multi-model mean. We are also looking for patterns across 100-155 sites in each analysis, using a GAM smoother, which also reduces the effects of individual model error on the patterns in the metrics substantially.

> Pg 8, lines 1-10: Again, a sensitivity analysis of how model and QC selection effect the metric would give users more confidence in the metrics.

We have included separate GAM interpolations, as noted in response to a comment above. We hope that this is satisfactory.

> Pg 8, line 8: While these caveats have been raise, really some effort to test the impacts of each should be included in the manuscript. I think it would be more beneficial to show the sensitivity of the metric rather than the extensive hypothesis testing.

This is a valid point, and we have moved a number of the results sections into the supplementary material, see the next point.

Of the caveats, the training testing mismatch is already shown in Figure 1, and the difference between metrics is shown in Figure 2. We can not see a way of testing the impact of the model failures, as complex models for those sites will not run successfully.

> Pg 9-11: While I appreciate the thorough analysis, I feel that the paper is lacking focus and comes across as sort of a data-dump. Especially given the fact that many of the results are inconclusive. Possibly focusing on some of the most promising hypothesis and moving many others to the supplemental material, or removing and simply mention that they were tested and the results were inconclusive. In all honesty I had to take a break from reading the paper after finishing this section.

This is a fair assessment. We have have moved a number of results sections into the supplementary material. The sections that we removed are:

- Aridity: partial duplication of the Mean Precip/Budyko sections, and less informative. We added a sentence to the Budyko Framework section to note the addition to supplementary.
- Interannual variability, Diurnal ranges, and Seasonality: These are kind of addenda to the Mean Temp/Precip section, and are not very informative. We added a sentence noting their presence in the supplementary material to the Mean Temp/Precip paragraph.

- Remoteness: This is already somewhat evident in the map figure. We have left the text alone, other than to note the presence of this figure in the supplementary material.

  Pg 12, line 5: Is there a reason for using the CRU mean annual temperature and not the mean annual temperature from the sites themselves?

The reason is that the CRU data spans a longer period, and is therefore more representative as a site characteristic MAP/MAT. Some sites have short record lengths and variable climates and so a years MAT/MAP (noting likely data gaps), would not be instructive. This was already partially explained at p.8, l.20 in the original submission, but we have modified the sentence to read:

For example, if the site only has a short dataset measured over a particularly wet or dry period, or if a site has a strong seasonal pattern in the quality of the temperature data, the mean would be less representative of the general site characteristics than a longer-term dataset.

> Pg 12, line 29: Seems an editing note got left in "(QUANTIFY?)". It would probably be a good idea to quantify what you mean by cooler.

Whoops! We changed this to "(< -5°C)".

> Pg 13, Figure 2: There seems to be some anti-correlation between the uniqueness and mean metrics, particularly with the RMSE of NEE. Is this likely just spurious?

Part of the reason the RMSE mean is lower at lower sites is that these sites are less energetic, and there is less variance in the fluxes, so modelling errors are correspondingly larger. This is part of the reason that we chose to stick mostly with the RMSE uniqueness for later figures in the paper, as it is less confounded, as well as being more immediately understandable compared to the other two metrics.

> Pg 14, line 4: Typo: "a possible a lower", furthermore, it would be beneficial to be more exact, as the results often report that there may be patters.

Thank you, fixed.

We have changed the text in the paragraph describing the Figure 2 to make it clearer that by "patterns" we mean patterns in over-all behaviour of the metrics contingent on the predictors:

In Figure 2, we see that there are some consistent behaviour in the predictability metrics, which might indicate that mean annual temperature is a driver of predictability, but in general any consistency in the behavioural patterns are not strong. For instance, for RMSE uniqueness (first row), we see a slight increase in uniqueness (or lack of predictability) in sites that are cooler (< -5°C), as well as sites that have a mean annual temperature around 20°C, for both NEE and Qle. That pattern is less distinct in Qh. There is a stronger trend in RMSE mean (second row) for Qle, but this is likely largely due to the fact that warmer sites naturally tend to have larger heat fluxes. It seems surprising that Qh does not exhibit the same behaviour, since it is more directly related to temperature. Correlation uniqueness (third row) and mean (4th row) shows a similar pattern to RMSE uniqueness for NEE and Qle, where cooler sites and sites around 20°C tend to be harder to predict well. Patterns in Overlap uniqueness (5th row) and mean (6th row) are less clear, but there may be a slight indication of higher uniqueness around 20°C for NEE, and possibly a lower distribution predictability at higher average temperatures. Note that the negative Overlap uniqueness values are largely due to the fact that regression models do not perform particularly well on extreme values (as indicated in Best et al., (2015). We note that the two subset GAM plots for linear and

longer-term models describe similar behaviour 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.

> Pg 15, Figure 4: The use of two colormaps with overlapping colors can be confusing, giving a false indication that the yellow in both plots is related.

The use of separate colour schemes is simply to indicate that the plots are showing different variables. There are only a limited number of potential colour schemes available when aiming to maximise accessibility. In both cases, darker is higher, so we don't think that this is a real problem.

> Pg 16, line 9: Again, using CRU for precipitation data when you have site level data seems curious.

See reply to the comment on Pg 12, line 5.

> Pg 16, line 18: Is high diurnal temperature range not related to mean temperature? Can you differentiate this signal from that seen in Figure 2?

Actually, higher diurnal temperature range is more directly correlated with a lower mean precipitation/humidity. Since we have moved this figure to supplementary, and it does not form a core part of the paper, we have simply added a note to this effect in the figure caption, and hope that this suffices.

> Pg 17, Figure 6: Deviance from Budyko curve is never explicitly defined.

We added an extra sentence to the text paragraph describing the figure, and modified the following sentence:

We also calculate a "Budyko deviance", which is simply the difference between the actual and predicted values on the Budyko plot, normalised by the predicted values, such that sites falling further above the Budyko curve have a positive deviance. There does not appear to be any pattern in predictability for NEE or Qle as a function of deviance from the Budyko curve (see second row in Figure 5), however there does seem to be some trend toward higher uniqueness for Qh for sites further above the Budyko curve.

> Pg 18, line 2: "met forcing", met. is an abbreviation.

Replaced with "meteorological"

> Pg 18, line 9: "Shrubland and Savannah, and Grass", => Shrubland, Savannah, and Grass?

The "Grass" is the start of the next clause. We have replaced the comma with a semicolon:

. . . more unique than all Forest types, Shrubland and Savannah; and Grass also tends to be more unique than . . .

> Pg 22, line 2: parentheses has no close.

Fixed.

> Pg 22, line 8: This is an example of a concrete example which give more confidence in the metric, yet it is given little attention compared to other analysis which are relatively inconclusive. If other examples exist possibly they could be highlighted.

There are other examples, although none that we can see as as obvious as the Metolius sites. We have extended the paragraph to read:

There are interesting differences within clusters of FLUXNET sites, for example the US Metolius sites (US-Me1, US-Me2, US-Me6) are similarly unique for Qh and Qle, but US-Me1 is substantially more unique for NEE, and this site was measured for two years after a fire that killed all trees at the site (Law, 2016). This gives some indication that our uniqueness metric does indeed have bio-physical meaning. A similar though less distinct pattern can be seen in the CA-SF sites in Saskatchewan - the CA-SF3 site was burnt much more recently than the other two. There is also a notable gradation in Qle predictability in the UCI burn sites (CA-NS), that correlates with time since the last burn. There are likely other comparisons that can be drawn with sites not included in Tier 1, and an extended year-by-year analysis might also pick up land use changes related to cropping, for example.

> Pg 23, Figure 15: Could this figure be organized in a way that gives more information, such as ordering by uniqueness or grouping by PFT? In the current state it would maybe be more useful as a table with actual numbers.

The name-ordering provides a useful look-up table for sites, as well as grouping neighbouring sites in many cases. Ordering by uniqueness (of one variable? Of the average across variables?) would highlight the most and least unique-sites, but we are explicitly trying to avoid this, as discussed in response to the second comment.

We have added numbers to the plot, so it is now effectively a coloured table. In doing so, we encountered two bugs, which are now fixed, namely that the last site, ZM-Mon, had been dropped off the table due to a rounding error, and more seriously that each panel of the figure had been plotted with an independent colour scale.

The figure is now substantially different (in particular, the 2nd, 4th, 5th, and 6th panels are substantially lighter), although none of the text needed changing, since none of the comparison we made were affected by this change.

> Pg, 24, line 5-7: I don't follow your logic here. I am not sure how the lack of a strong trend in Figure 12 provides support to the methodology. Also, I would not conflate the proximity of one tower to other towers with biome representativeness.

This is a fair point, we have removed the sentence.