Ref.: bg-2018-179
Biogeosciences
Response to reviewers

Dear Dr Keenan,

Thank you for the chance to respond to reviews. We found both reviews insightful and constructive. We have attempted to implement as many suggestions and comments from both reviewers as possible. Details are included in the individual responses to each reviewer in the interactive discussion.

Responses to reviewers and a pdf of the differences between this and the last version is attached below.

Cheers

ned haughton, Gab Abramowitz, Martin De Kauwe and Andy Pitman.

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.

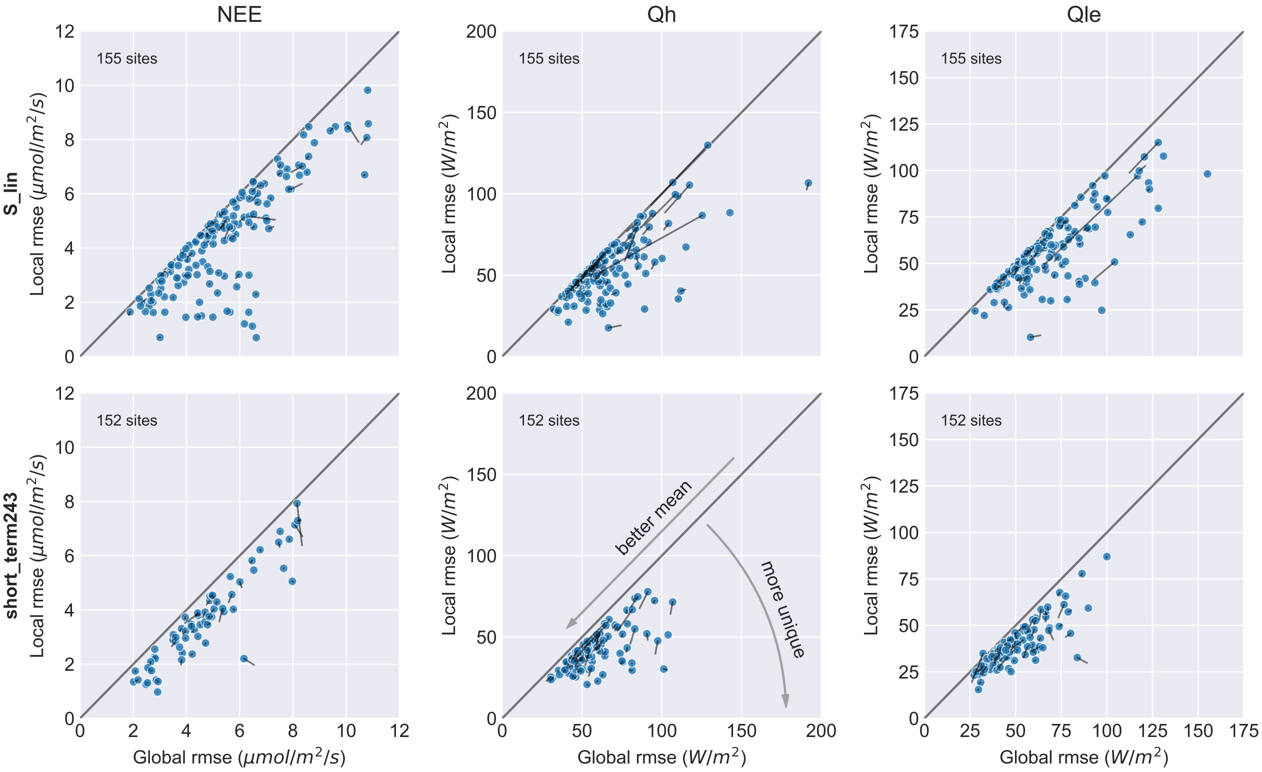

Figure 1: RMSE values for the global (x-axis) and local (y-axis) model simulations. Columns show the three fluxes, the forst row shows data for S_lin, the second row for short_term243. The tails of each point show where the local and global RMSE values would be if the same QC flags were used for training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

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.*

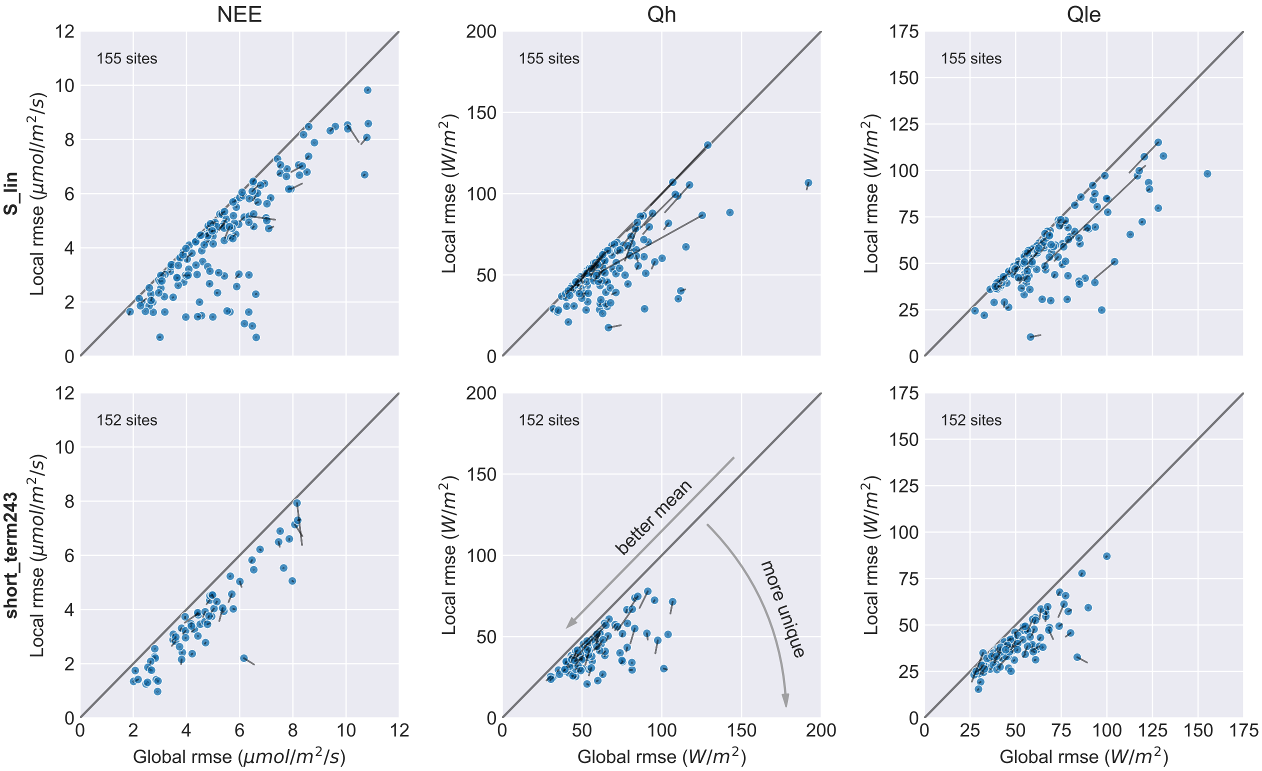

Figure 1: RMSE values for the global (x-axis) and local (y-axis) model simulations. Columns show the three fluxes, the forst row shows data for S_lin, the second row for short_term243. The tails of each point show where the local and global RMSE values would be if the same QC flags were used for training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

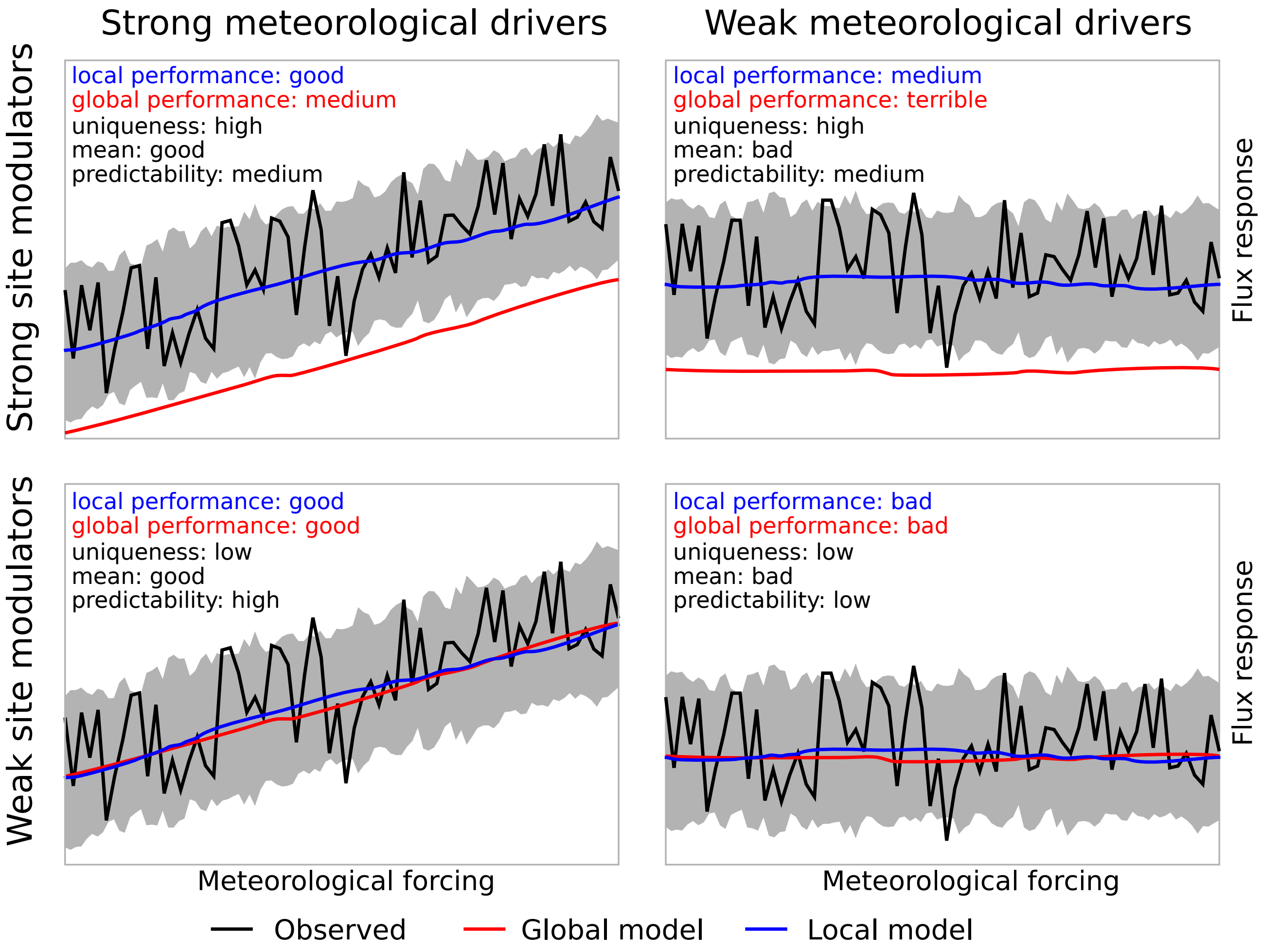

Figure 2: Schematic of sources of variability and how they affect predictability, mean performance, and uniqueness. The black line represent the flux observations, and the grey ribbon, the unpredictable internal variability or noise in the system. "Meteorological drivers" indicate universal physical 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.

Does predictability of fluxes vary between FLUXNET sites?

Ned Haughton[1], Gab Abramowitz[1], Martin G. De Kauwe[1], and Andy J. Pitman[1]

[1]Climate Change Research Centre, UNSW Australia

Correspondence to: Ned Haughton (ned@nedhaughton.com)

Abstract. 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. ~~Aside from intrinsic interest in this fundamental question, understanding site predictability would be useful for land surface model (LSM) evaluation in setting a priori expectations of model performance. It would also provide a clear rationale for selecting particular FLUXNET sites for model development, evaluation and benchmarking.~~ 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. ~~Predictability was defined using the disparity between the ability to simulate fluxes at a site given specific knowledge of the site, and the ability to simulate fluxes given general land surface specifications. We then examined predictability using performance metrics including RMSE, correlation, and probability density overlap, and~~ We defined site **uniqueness** as the disparity in performance between multiple empirical models trained globally and locally for each site~~. A number of hypotheses potentially explaining site predictabilitywere then tested~~, 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, ~~data quality and site characteristics~~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~~including~~, and in particular little evidence that flux behaviour ~~is~~ was well discretised by vegetation ~~types~~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 ~~is sound and~~ provides a useful basis for site selection in LSM evaluation, and invite critique and development of the methodology.

1 Introduction

The land surface is a key component of the climate system, as it provides feedbacks to atmospheric conditions via the exchange of heat, moisture, and carbon fluxes. These surface-atmosphere exchanges are contingent on the characteristics of the soil and vegetation. However, these interactions between the atmosphere and land are not uniform, for example in hot, mesic environments net primary productivity (vegetation productivity) becomes less sensitive to the effect of climate (Bonan, 2015; e.g. increasing precipitation Chapin III et al., 2011; Del Grosso et al., 2008; Gillman et al., 2015; Huston and Wolverton, 2009; Schuur, 2003). Across the globe, variability in the productivity-climate relationship suggests that the behaviour of some ecosystems must be more predictable than others. Intuitively, the behaviour of ecosystems that experience marked stochasticity in precipitation (e.g. ecosystems that rely on monsoonal rains for growth), would likely be harder to predict than ecosystems that experience relatively consistent year-to-year conditions (e.g., the boreal zone, the wet tropics or desert regions). Furthermore, whilst vegetation patterns are broadly understood at global scales (e.g. the Köppen climate classification, Kottek et al., 2006; and Whittaker's biome classification, Whittaker, 1962), at local and regional scales, plants exhibit local scale adaptations to their environment (e.g. soils and topography). Taken together, these relationships between climate and local factors point to a spectrum of site predictability. Perhaps surprisingly, the predictability of a site is rarely considered explicitly when choosing sites to evaluate models~~.~~, and never quantified.

~~Within the climate science community, 30-years of research originating from Dickinson et al. (1986) has transformed how~~ Modern land surface models (LSMs) attempt to describe the exchange of energy, water and, more recently, carbon~~. The community has moved from no explicit representation of vegetation in climate models (e.g. Manabe, 1969), through highly empirical second generation models (which included an explicit simulation of heat and moisture fluxes and storage in soil and vegetation), to models that attempt to explicitly represent~~ , by explicitly representing the soil-vegetation continuum (~~see review by~~ 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.

Land surface modellers often use FLUXNET data to evaluate their models, and to tease out weaknesses, with the goal of model improvement. In the 1990s and 2000s, when only a handful of flux tower site measurements were available, the LSM community gravitated to these datasets, hence observations taken at Cabauw (The Netherlands), Harvard Forest (USA), and near Manaus (Brazil) were widely used. Over the last two decades, direct measurement of land surface fluxes and meteorological variables has rapidly expanded, as new flux towers are installed and existing towers continue to gather data. FLUXNET 2015 (Fluxdata.org, 2018), as a synthesis of these measurements, represents a rich source of information about the exchange of carbon, water and energy. The freely available release (Tier 1), encompasses over 150 sites and includes over 500 site-years of high temporal, quality-controlled data. These data provide an unparalleled opportunity to improve our observationally-based understanding of land-atmosphere exchanges of carbon, water and energy. They are also particularly useful for LSM evaluation since both the necessary driving variables (meteorological variables) and prediction variables (energy, water and carbon fluxes) are reported at a spatio-temporal scale relevant to LSMs. As a consequence, land surface modellers have developed tools to enable the FLUXNET 2015 data to be used routinely (e.g. Ukkola et al., 2017). However, with hundreds of site datasets now freely available, site choice for model evaluation varies widely among the land surface community, with no common strategies for site selection. FLUXNET sites differ in many ways: in data record length (from less than one to greater than twenty years); in climate regime; and in soil and vegetation characteristics. Their similarity to each other also varies – FLUXNET is not evenly distributed over the globe, and has higher density in more densely populated and wealthy regions, such as Western Europe, and the north-east of the United States, with particularly heavy representation of temperate forests.

Despite obvious distinctions between sites in FLUXNET defined by precipitation regime, temperature, seasonal snow cover and indeed PFT type, it is not immediately clear which of the 150-plus freely available sites are most useful for model evaluation. One might assume that given the diversity of sites, some are easier to simulate than others, and it seems sensible to assume that the choice of sites could have an impact on insight gained from model evaluation at these sites. However, assumptions about the predictability of different sites have not been explicitly tested. ~~For example, in recent multi-model evaluation and benchmarking experiments, where multiple FLUXNET sites were used, Best et al. (2015), Haughton et al. (2016) and Haughton et al. (2018) were not able to identify any obvious patterns in model performance across sites.~~ The lack of quantification of predictability means that site selection for evaluation is potentially susceptible to confirmation bias. That is, a modeller might unconsciously choose sites that are easier for their model to simulate, rather than selecting sites based on their instructiveness for identifying flaws in a model. For example, consider the implications of evaluating a model against ten FLUXNET sites that happen to be the least predictable in comparison to evaluation against the ten most predictable sites. In the former case, a modeller might become disillusioned with the apparent lack of skill of a potentially good model, while in the latter case a modeller might become overconfident concerning the skill of a poor model.

This issue of site predictability has been ignored in historical flux-model comparisons, where modelling groups have generally not tried to explicitly justify their choice of sites, or have based their reasoning around issues such as data availability or length of record. Chen et al. (1997) chose the Cabauw site for a multi model intercomparison because it was considered relatively easy to simulate. Several authors chose longer (multi-year) sites (Balsamo et al., 2009; Lawrence et al., 2011; Wang et al., 2011). Some evaluation papers explicitly sought to sample a range of PFTs (Bonan et al., 2014; De Kauwe et al., 2015). Many highlighted choices based on the availability of gap-filled data (Krinner et al., 2005; Slevin et al., 2015; Wang et al., 2011). A few papers highlighted the high natural variability of a site (Balsamo et al., 2009), or a high degree of climate differences between sites (Wang et al., 2011). Others highlighted the quality of specific sites and some provided evidence for this decision based on energy closure (Napoly et al., 2017). In contrast to the often detailed explanation for why a specific model or parameterisation is chosen, the defence of specific evaluation data sets often lacks a coherent rationale. Most commonly, "high quality" or "longer" data sets are selected. A longer data set may sample more years, but a single month of data from another site might provide more information regarding a specific phenomenon (e.g. the response to a drought or a heatwave). Sampling more PFTs might be valuable, but might also bias results if the selected sites fall within a similar behavioural regime not well discretised by PFT. In short, it would be useful to be able to make clear, evidence-based statements about the relative predictability of different sites, based on meteorological patterns or local site characteristics. This would allow modellers to make informed site selection choices for model development/evaluation that maximise coverage of diverse site behaviours, and ultimately help to reduce uncertainty in model projections. 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.

~~There is no single definition of predictability, but it~~ 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. Predictability of a system should therefore also encompass the capacity to predict changes in the property of interest, given changes in the drivers of the system, for example differing flux responses in wet and dry periods. Differences in predictability between sites might be due to many factors, including, but not limited to:

- variability of meteorology (e.g. strong seasonality in precipitation compared with low variability, large seasonal cycles in incoming radiation compared to small seasonal cycles, and stochastic events);
- complexity or consistency of the site itself (e.g. orographic effects, managed land use including different irrigation and cropping patterns, vegetation and soil structures);
- broader scale impacts (e.g. climate type, regional aridity, teleconnections to major oceanic drivers, landscape heterogeneity, geological basins);
- technically sourced variance (quality of instrumentation, assumptions and application of eddy covariance methodology, post-processing).

In this context we might envisage predictability to be the degree to which a “perfect” model could accurately estimate measured fluxes at a site, given appropriate meteorological variables and relevant site characteristic information. Of course, we lack a perfect model, and the accuracy of our observational data is always limited by measurement error, and noise in the system being measured. As such, any practical measure of predictability will be limited in accuracy, but this does not mean that it cannot still be useful.

Some predictability metrics do exist: Colwell (1974) defines a predictability metric based on constancy in time and contingency on season but this metric only captures one aspect of performance – temporal correlation. Abbas and Arif (2006) also proposed a number of time series predictability metrics, but these are only useful in univariate time-series prediction, where the forecast is made only given knowledge about the predicted variable itself, rather than knowledge of other predictor variables, as is the case with flux prediction from meteorological variables. ~~There are also model-class specific performance metrics, such as the Genetic Programming predictability metric presented in~~ Kaboudan (2000) ~~, but such metrics rely on the assumption that the model is suitable for predicting the data in question~~provides another univariate predictability metric.

Since existing predictability metrics are not suitable to our problem, below we detail a new metric of site predictability and analyse the FLUXNET 2015 sites according to their predictability. To do this, we applied a suite of empirical models to predict fluxes at the 155 flux tower sites with half-hourly data included in the Tier 1 FLUXNET 2015 release. We also investigate several hypotheses that might explain the variation in site predictability in different locations. Finally we attempt to provide a sound theoretical basis for site selection for LSM development and model intercomparison projects. This will allow ~~*a priori*~~ expectations of model performance to be better defined ~~, as well as mitigate~~ 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.

2 Methods

~~Differences in predictability between sites might be due to many factors, including, but not limited to:~~

- ~~variability of meteorology (e.g. strong seasonality in precipitation compared with low variability, large seasonal cycles in incoming radiation compared to small seasonal cycles, and stochastic events);~~
- ~~complexity or consistency of the site itself (e.g. orographic effects, managed land use including different irrigation and cropping patterns, vegetation and soil structures);~~
- ~~broader scale impacts (e.g. climate type, regional aridity, teleconnections to major oceanic drivers, landscape heterogeneity, geological basins);~~
- ~~technically sourced variance (quality of instrumentation, assumptions and application of eddy covariance methodology, post-processing).~~

~~We focus on the first of these~~ This study focuses on predictability of three key fluxes: net ecosystem exchange (NEE), sensible heat (Qh), and latent heat (Qle). We focus predictability as a consequence of meteorological drivers only, and ask whether predictability at a specific site can be understood in terms of the differences in flux behaviour given particular site and meteorological conditions, relative to the flux behaviour that would be expected at other sites given the same conditions. We do this by training a suite of empirical models ~~(based on the models described in Haughton et al., 2018)~~ to predict fluxes ~~,~~ based on meteorology, ~~at~~ twice for each FLUXNET site~~twice~~. First we train the empirical models using all of the available data from all of the available sites at once ("global training"), to characterise the general expected flux behaviour given a specific set of meteorological conditions. Then we re-train the models using only data from the individual site in question ("local training"). The globally and locally trained versions of the models are then used to make predictions at each FLUXNET site, and their performances are compared, using a range of performance metrics. Any improvement in performance by the locally trained model over the globally trained model is an indication of driver-flux relationships that are unique to the site in question (note that this may include systematic errors in measurement). Since such a site exhibits relationships between drivers and fluxes that are not broadly shown at other sites, we argue this site has lower predictability than a site that acts more similarly to the global behaviour.

To quantify this, we plot the local and global metric values as Cartesian coordinates, then convert them to polar coordinates (see Figure ~~??~~??). The origin represents the best possible performance metric value, so distance to the origin represents the mean site performance across the global and local simulations. The degree to which each point drops below the 1:1 line will be our definition of uniqueness~~, or lack of predictability~~. To illustrate, imagine a model that perfectly represented all relevant process and fully utilised all of the available information in the input data to make the best possible prediction. This model could be used to assess site predictability based on the residual sum of squares against observations, and this metric value could be compared across different sites. No such model exists of course, and we therefore use empirical models to assess the predictability of the data while minimising assumptions about the functional form of any relationships between variables. For further discussion of why empirical models are suitable for estimating the information available in FLUXNET data, see Best et al. (2015) and Haughton et al. (2016).

~~In particular,~~ This procedure is model-agnostic, and we have used models in the framework developed in Best et al. (2015) and Haughton et al. (2018), ~~to predict net ecosystem exchange (NEE), sensible heat (Qh), and latent heat (Qle)~~because they are conceptually simple, but able to fit complex functional relationships. These models ~~included~~ (listed in Table ??) 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).

Models used various combinations of meteorological driving variables: down-welling shortwave radiation (S), surface air temperature (T), relative humidity (H), wind speed (W), and precipitation (P). Models also used a number of derived variables, including difference in air temperature and specific humidity since dawn (dT and dQ, respectively), as well as lagged averages of each meteorological forcing (e.g. lH10d indicates a 10-day lagged average of H), where the lags were chosen pseudo-optimally, and lagged differences from the current time step (e.g. lT6hM indicates a 6-hour lagged average of air temperature minus instantaneous air temperature). Haughton et al. (2018) showed that each of these driving variables we use here added predictive power to the models, and had relatively low correlation, avoiding problems of collinearity. Models referenced below follow a standard naming scheme that indicates the structure of the model, for example, *S_lin* indicates a linear regression using only shortwave down, while *STHdT_lS30d_km243* would indicate a cluster-plus-regression model with 243 clusters, with shortwave down, air temperature, relative humidity, temperature difference since dawn, and a 30-day lagged average of shortwave down as inputs. None of the models are provided with site characteristic data (e.g. geographic, soil, or vegetation information) as we want to use the models to test the effects of these characteristics on predictability. A complete list of the empirical models is shown in Table **??**.

To run the models, we converted the raw FLUXNET 2015 Tier 1 data (only sites with half-hourly data, 155 in total), using the FluxnetLSM tool developed by Ukkola et al. (2017). In all cases, the empirical models are trained only on high-quality non-gap-filled data, according to Quality Control (QC) flags from FLUXNET 2015 and FluxnetLSM. The models are then run on all available data (including gap-filled data, to maximise the time coverage of empirical models with time-lagged drivers), and evaluated only on time steps with non-gap-filled data. We then plotted each metric value for each site in a scatter plot, with the global value on the x-axis, and the local value on the y-axis. We decomposed that information into:

1. Mean performance: the ~~mean metric value over~~ arithmetic mean of the local and global ~~simulation~~ metric at each site, defined by distance from the origin. Higher is worse in most metrics, including root mean square error (RMSE), but

	Successful simulations				Negative uniqueness		
model name	global	NEE	Qh	Qle	rmse	corr	overlap
S_lin	155	155	155	155	2	~~206~~ 0	120
ST_lin	155	155	155	155	1	20	117
STH_km27	152	152	152	152	3	4	~~50~~ 49
STH_km81	152	146	150	149	3	4	39
STH_km243	152	108	133	131	3	3	25
STHW_km81	152	142	148	147	2	3	23
STHW_km243	152	88	126	123	2	2	11
short_term243	(STHWdTdQ_lT6hM_km243)						
	152	65	98	97	1	1	9
long_term243	(STHWdTdQ_lS30d_lR30d_lH10d_lT6hM_km243)						
	152	3	12	12	~~2 2 0 long_term729 152 0~~ 0	0	~~0 0 0~~ 1

Table 1. number of sites that models ran successfully at, for global training (columns 1), the number successful local simulations for each variable (columns 2-4), and the number of cases of negative uniqueness, indicating that the local model performed worse than the global model, for each of the three metrics (columns 5-7). Three sites (CA-Man, DE-RuR, and DE-RuS) did not include relative humidity, and so all models including that variable failed, including the global model. The uniqueness calculations are sometimes, particularly in the case of correlation, prone to numerical error, and so the threshold we use here is -1e-8 instead of zero.

lower is worse in the case of Pearson's correlation coefficient (Corr) and Perkins' distribution overlap metric (Overlap, Perkins et al., 2007).

2. "Uniqueness": the angle below the 1:1 line. Uniqueness is calculated as $\frac{4}{\pi}arctan(\frac{x-y}{x+y})$, such that if, for example, RMSE is 0 locally and some positive value globally, uniqueness will be 1.

Note that because the best possible result for some metrics is 1 (e.g. Corr and Overlap), in those cases we subtract the value from 1 such that the best result is 0 before calculating the uniqueness, so that it can be interpreted the same way across metrics - that positive numbers indicate better local performance. We avoid transforming metrics for mean performance, so that metrics are in their standard units.

In general, this definition of uniqueness ranges from -2 to 2, and is strictly between -1 and 1 for metrics that only have values on one side of "best" (e.g. RMSE is positive definite, Corr is always less than or equal to 1), but in most cases should lie between 0 and 1. A model's uniqueness is 0 if the local and global simulations perform equally well, between 0 and 1 if the local model performs better than the global model, and negative if the local model performs worse that the global. Negative values are unusual, and indicate that the local meteorological forcing provides insufficient useful information to increase performance, and that the local model has failed in a spurious way (discussed in more detail below). The number of negative uniqueness values for each metric and each model is shown in the last three columns of Table **??**, out of a maximum of 10 x 155 = 1550 cases. 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.

The uniqueness and mean performance metrics are shown for RMSE in Figure ~~??~~ ?? for the S_lin ~~empirical model~~ 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 ~~??~~ ?? also illustrates the differences ~~in a simple case~~ 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. ~~Both rows of figures show local versus global values, but the first row uses only flux variable QC flags for evaluation, and so can include part of the period for which there are poor QC meteorology values. The second row uses exactly~~ 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 ~~flags (and so time steps) in both training and testing phases (as a proof of concept). As a result, in the second row, the~~ 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 ~~. In the first row, the points shift slightly~~(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. ~~This difference can be exacerbated for more complex models~~ 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.

Best et al. (2015) used the concept of ranking over multiple performance metrics, and then aggregating over rankings to arrive at a single value that represented a broad concept of performance for each model. This methodology is extremely useful for model evaluation using FLUXNET site datasets. However, due to the very different distributions of results for the different metrics (discussed below), we avoided aggregating over metrics and instead examined a set of key metrics separately for their ability to capture independent aspects of performance. The metrics we chose were RMSE, as it provides an overview of model accuracy in relevant units, Pearson correlation (Corr) as a measure of temporal correlation, and Perkins' distribution overlap metric (Overlap), as it gives a measure of the match between the observed and modelled distributions.

2.1 Caveats

In an idealised experiment, even if we exclude the possibility of over-fitting, the locally-trained model should *always* perform better than the globally trained model to some degree. This is because the local model is predicting the same data that it is

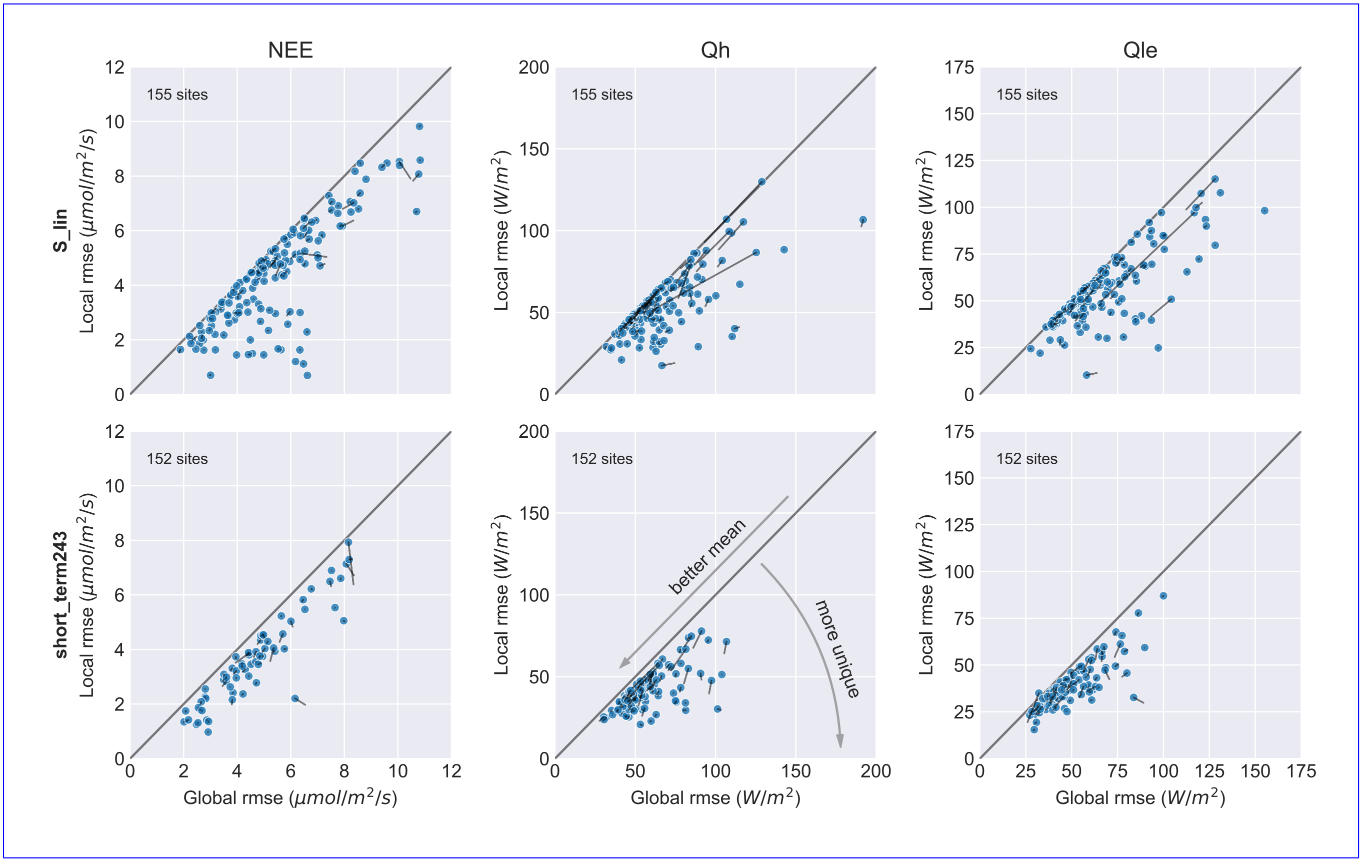

Figure 1. RMSE values for the global (x-axis) and local (y-axis) model simulations. Columns show the three fluxes~~. The top ~~, the first row ~~uses ~~shows data for S_lin, the second row for short_term243. The tails of each point show where the local and global RMSE values would be if the ~~flux ~~same QC flags were used for training ~~,~~ and ~~only ~~evaluating (the intersection of meteorological and flux QC flags~~for evaluation (the method used throughout the paper~~). ~~In ~~Tails pointing toward the ~~second row, ~~zero in each axis indicate the ~~meteorological ~~model would have performed better using these QC flags~~are also used for evaluation, such that the training and testing data are identical~~. ~~Colours simply serve to identify sites~~In other words, ~~and allow clearer comparison between ~~a tail pointing towards the ~~top and bottom rows~~origin means that our evaluation method has a bias toward worse mean RMSE, and a tail pointing clockwise from the ~~blue line is ~~origin indicates that our method has a ~~1:1 line~~bias towards lower uniqueness.

trained on, and should capture any behaviour that is site-specific (that is, it is being tested in-sample). However, there are a number of factors that might prevent this from happening.

First, a model may require a substantial amount of data to avoid over-fitting, and some sites may not provide enough data to train the model locally. For example, very few sites had enough data to adequately train the long_term243 ~~or long_term729 models~~ model from Haughton et al. (2018), ~~each of which have~~ which has 10 input variables (S, T, H, W, dT, dQ, lS30d, lP30d, lH10d, lT6hM) and ~~hundreds of clusters. As such, these models would potentially require~~ 243 clusters. This model potentially requires hundreds of non-gap-filled data samples at each cluster to obtain a reliable linear regression estimate (so ~10^4 samples in total). ~~These~~ The more complex models often fail to run locally, or run successfully but produce erroneous results (e.g. due to too few samples to obtain reliable regression results for a K-means cluster - this problem is described in detail in the Supplementary Material in Haughton et al., 2018). To mitigate this problem, we modified the models from Haughton et al. (2018) to ensure that each cluster always contained a number of samples at least 5 times greater than the number of input variables. When clustering failed, it was re-attempted a further 9 times, and if that was not successful, the model was excluded. See Table **??** for details on how many models ran successfully for each variable.

Second, as noted above, the training and testing data for the local model are in practice nearly always different, because the QC flags for the flux variables being evaluated against do not correspond perfectly with the meteorological forcing variable QC flags. Models are trained only on data that has good meteorological and flux QC flags for all relevant variables. However, simulations are evaluated on a larger subset of all data – those time steps marked as good QC for the flux variable alone. The motivation for doing this is to ensures that all of the different empirical models are evaluated on the same number of time steps. So, for instance, with the S_lin model predicting Qle at a particular site, the number of time steps with good S *and* good Qle QC flags might be only 80% of the time steps with good Qle QC flags only. Consequently, the model will be trained only on the 80% of period that it is tested on. This problem is exacerbated for models with more inputs and for models with lagged average inputs, which will usually be trained on substantially smaller subsets of data than they are evaluated on.

Lastly, "performance" is dependent on metric, and so performance will only be strictly better locally for metrics that are optimised by the regression-based structure used in the empirical models. For instance, ordinary least squares linear regression optimises RMSE in the training dataset, so assuming the training and evaluation datasets are identical, then the RMSE of the local model will be strictly not worse than the RMSE of the global model. However, metrics which assess model performance in terms of distribution, such as the distribution overlap metric or temporal correlation, may occasionally show that the local model performs worse than the global model, even when the local model is clearly better under RMSE. This is particularly pertinent in the context of a generally useful predictability metric.

These caveats are worth keeping in mind, but in the majority of the results below, they do not play a particularly large role. We are confident that our predictability metrics are satisfactory for a first attempt to estimate site predictability.

2.2 Hypothesis testing

Once we have a predictability metric, we can generate a number of hypotheses about what might determine predictability at different sites. Below we list several hypotheses, many of which intersect, and so in some of these cases we also mapped some predictability metrics against two hypothetical predictability sources.

Mean annual temperature and precipitation Sites with higher mean temperature tend to be those closer to the equator, and tend to have a smaller annual temperature cycle range. All other things being equal, we might therefore expect warmer sites to be more consistent over time, have a more constant response to meteorological forcing, and therefore be more predictable. Sites with higher average precipitation would be expected to have fewer drier periods, more consistently available soil moisture, and higher humidity resulting in a damped daily temperature cycle, and are therefore likely to be more predictable.

For these hypotheses, the FLUXNET site data is not always adequate, as the mean may not be perfectly representative of the true climatology of the site. 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, ~~this would introduce a bias~~the mean would be less representative of the general site characteristics than a longer-term dataset. For this reason, we calculated mean annual temperature and precipitation from the half-degree CRU TS4.01 data (Harris and Jones, 2017), using data from 1961-2016, and using the nearest neighbouring grid cell.

~~**Aridity** Arid sites tend to have higher precipitation variability, with fewer, heavier rain events, and longer dry periods (Donat et al., 2016). We would expect that flux predictability would be lower at arid sites. For this hypothesis, we used an aridity index based on mean annual precipitation from CRU TS4.01, and the energy-only estimate for potential evapotranspiration (PET) from Milly and Dunne (2016), based on net radiation and ground heat flux (PET $= 0.8(\overline{Rnet} - \overline{Qg})$) from FLUXNET, such that the aridity index (AI) = mean precipitation/PET. We assumed Qg=0 where sites did not provide Qg (which is approximately true on long time scales).~~

Budyko curve deviations The Budyko curve (Gerrits et al., 2009) plots an evaporative index against a dryness index, with the expectation that sites should, in the long term, fall along a function of dryness that is both energy and water limited. Sites that fall further from the Budyko curve may indicate data errors, or hydrological uniqueness (for example, rapid drainage, or external water sources), or that the data in question is not long enough to adequately capture and account for long-term internal variability. Whatever the cause of divergence from the Budyko expectation, we would expect that more divergent sites would be more difficult to predict.

~~**Interannual variability** Sites heavily influenced by longer term climate patterns, such as decadal scale ocean oscillations, are less likely to have all of their relevant patterns captured within the period of FLUXNET measurement, and so potentially contain systematic biases. We compared the interannual variability between sites for both T and P, using the CRU TS4.01 data. We calculated the coefficient of variance (CoV) for annual means of temperature (K), and precipitation (mm/year). We would expect that as IAV increases (shown by greater CoV), predictability would decline.~~

~~**Diurnal ranges** Sites with large diurnal ranges have stronger rates of change between daily peaks and troughs, and these are likely to make prediction harder. Faster changes in temperature, for example, can cause rapid changes in relative humidity,~~

~~which is a major driver of latent heat flux. We used the BioClim (WorldClim, 2016) mean diurnal temperature range using the nearest neighbouring grid cell for each site.~~

~~**Seasonality** Larger differences between winter and summer conditions would likely lead to lower predictability, since we would expect flux behaviour at such sites to be more diverse over the course of the year. This would also affect the relative influence of time varying factors, e.g. timing of snow melt, or vegetation phenology. For model and site combinations where the training and testing data is more disjointed, this might also lead to lower predictability due to the non-training testing data diverging more in behaviour. Since about 55% of sites in Tier 1 are less than 5 years long, we used the BioClim variables (WorldClim, 2016) to compare seasonality between sites. We investigated: isothermality - the ratio of diurnal temperature range to annual temperature range; temperature seasonality - the standard deviation of monthly average temperatures, normalised by the annual average in K; temperature annual range; precipitation seasonality; precipitation of wettest quarter; and precipitation of the driest quarter.~~

Vegetation type The FLUXNET 2015 sites are categorised by International Geosphere-Biosphere Programme (IGBP) vegetation types. There is a widely held assumption that different vegetation types behave differently in response to similar meteorological forcings (although this was assumption was questioned by Alton, 2011), and this presumably also applies to the overall predictability of a site. We grouped IGBP vegetation types into 5 major groups:

- **Evergreen Forest**: Evergreen Broadleaf Forests, Evergreen Needleleaf Forests (49 sites).
- **Deciduous Forest**: Deciduous Broadleaf Forests, Deciduous Needleleaf Forests (16 sites).
- **Mixed Forest**: Mixed Forests (7 sites).
- **Crop**: Cropland/Natural Vegetation Mosaics, Croplands (15 sites).
- **Grass**: Grasslands (29 sites).
- **Shrubland**: Barren or Sparsely Vegetated, Closed Shrublands, Open Shrublands (11 sites)
- **Savanna**: Savannas, and Woody Savannas (13 sites).
- **Wetland**: Permanent Wetlands (15 sites).

Other IGBP vegetation types not represented in FLUXNET 2015 Tier 1 included Snow and Ice, Unclassified, Urban and Built-Up Lands, and Water Bodies. We then compared the performance metrics across these groups.

~~**Geographic remoteness** We are training the global models on all available sites, but FLUXNET sites are not evenly distributed over the land area of the globe. As such, we might expect that sites that have many other similar sites in the global dataset would have their behaviours more adequately captured by a globally trained model. To investigate whether more geographically unique sites were less predictable, we mapped the sites by uniqueness, and also compared uniqueness by average distance to all other sites.~~

Energy balance closure Wilson et al. (2002) showed that FLUXNET sites often have a problem closing their energy balance. Net incoming radiation (Rnet) does not match the total energy accounted for by the heat fluxes (Qh, Qle, and Qg) and changes in heat storage, on average having an imbalance of around -20% at each site, but ranging from -60% to +20%. Since this imbalance pertains to boundary conditions, which are all measured (sometimes with the exception of Qg, although that can be assumed to be too small to account for the difference on a long enough time scale), the imbalance indicates some problem with either the measurement system, or the eddy covariance methodology. We would assume that sites with worse energy imbalances are likely to be more difficult to predict. We calculated the energy closure gap as the energy_gap = mean(Rnet - Qh - Qle - Qg) (we used Qg=0 for sites missing Qg), and also compared sites by normalised energy gap, using abs(1 - energy_gap/Rnet). Note that this is not the exact formulation used by Wilson et al. (2002), but it serves the same purpose – to identify energy closure imbalances.

Record length Since many of the longer-term or rarer behaviours mentioned above are more likely to be captured adequately in site datasets that span longer periods, we should expect that longer sites would be more predictable. On top of this, site principal investigators are likely to become more familiar with problems with their sites, equipment, or methods, and more likely to be able to find solutions to those problems over time, and so we should expect that data quality should improve in longer site datasets. We examined the number of years in the dataset as a predictor for uniqueness.

Gap-filling ratio Some bad data is likely to make it through quality assurance procedures, and such bad data would make prediction more difficult. It is not clear how one would tell such data in most cases, unless patterns are obvious. We visually inspected the time series plots produced by FluxnetLSM for each relevant variable, for each site, and saw no obvious problems within the data periods marked as good QC. However, some proxy for data quality may be possible, and in particular sites with more high quality data may indicate better instrumentation or procedures, and less likelihood of having bad data marked as high quality. We compared sites by the proportion of data marked as good QC to total data, averaged over all variables, separately for meteorological and flux variables.

We also looked at various other hypotheses, including the impact of aridity index; daily, seasonal, and interannual variability in temperature and precipitation, and geographic remoteness. The majority of these had few interesting trends or other patterns, and so we have excluded them from the paper. They are included in the Supplementary Material.

We note that some determinants of predictability could not be calculated for some sites. For example, a number of sites have no non-gap-filled data for precipitation, and so mean annual precipitation can not be calculated, and neither can dependent determinants, such as aridity index. In such cases, the sites are omitted from individual analyses.

3 Results

3.1 Viability of the "predictability" metrics

First, we show how the uniqueness and mean performance metrics vary across all models and sites for RMSE, Corr and Overlap. Figure **??** shows each of the 3 metric pairs (rows) for each of the three fluxes (columns), and how those metrics vary with mean annual temperature in the CRU TS4.01 dataset. Here the uniqueness and mean performance values are similar to those explained in Figure ~~??~~??, but use more complex models in addition to S_lin (listed in Table ??). Note that uniqueness values less than zero indicate that the local model is not performing better than the global model, as noted above.

- Row 1 shows the RMSE uniqueness of each site, with more unique sites having higher values.
- Row 2 shows the mean of the RMSE of the global and local simulations for each site. For this metric, one might expect that sites that are more difficult to predict would have higher values, but note that sites with more available energy will generally tend to have larger fluxes and so higher RMSE values, regardless of uniqueness.
- Row 3 shows correlation uniqueness. Like RMSE uniqueness, higher values indicate lower site predictability. Note that there are a large number of zero values for this metric, because for instantaneous linear regression models, correlation is always identical (or inverted) between global and local models, since they are using the same input data, and so uniqueness is always 0.
- Row 4 shows the mean correlation with observed values for local and global simulations - sites with a low correlation are more difficult to predict (at least by these models). Note that there are a few simulations with 0 mean correlation – these are cases where linear regressions had global and local gradients with opposite sign, resulting in an exactly opposite correlation. In those cases, the zero does not indicate that the global and local simulations had low correlation.
- Row 5 shows the Overlap uniqueness. Higher values indicate sites for which the local Overlap was better than the global Overlap, and negative values indicate the global model performed better in terms of Overlap.
- Row 6 shows the mean model-obs Overlap values of global and local models, and lower values indicate a site that is harder to model in terms of Overlap (Overlap=1 indicates that the model's flux distribution is identical to the observed distribution).

All plots have a fitted generalised additive model (GAM) line, added to help indicate trends in the site means. It is estimated using the pyGAM package (Servén, 2018), using 8 splines, and plotted with a 95% confidence interval. 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.

In Figure ??, we see that there are some ~~patterns~~ consistent behaviour in the predictability metrics, which might indicate that mean annual temperature is a driver of predictability, but in general ~~these~~ 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 (~~QUANTIFY?~~<-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 ~~pattern~~behaviour, since it is more directly related to temperature. Correlation uniqueness (third row) and mean (4th

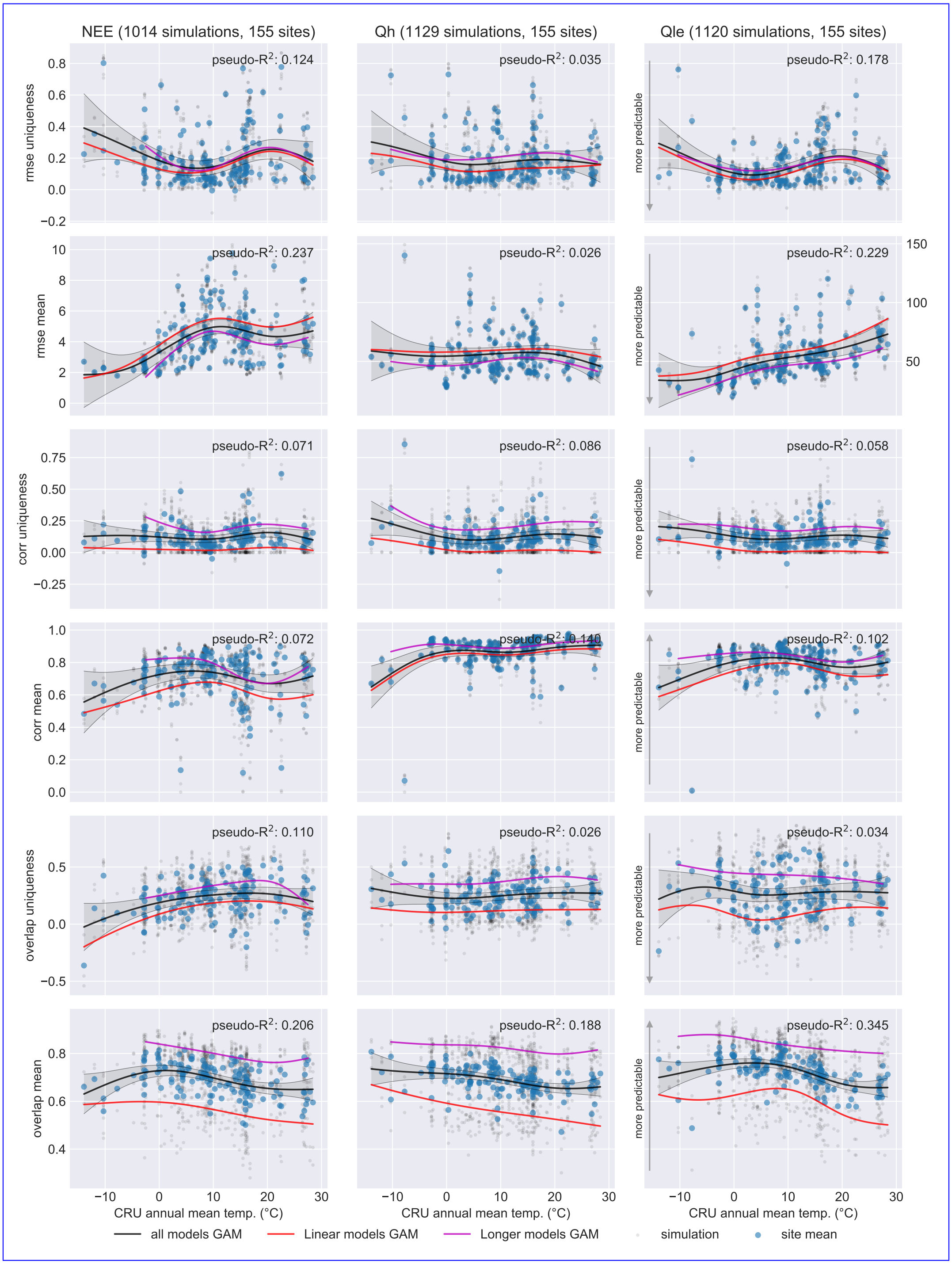

Figure 2. Predictability metrics for mean annual temperature, for all models. The three columns represent the three fluxes, NEE, Qh, and Qle. The six rows show RMSE uniqueness, RMSE mean, correlation uniqueness, correlation mean, Overlap uniqueness, and Overlap mean. Grey points are individual simulation values, blue points are site means across all empirical models. The black line represents a GAM fit of the site means across models, and the ribbon represents a 95% confidence interval. The red line is a GAM fit using only the linear models (S_lin, ST_lin), and the purple line is a GAM fit using only the models with lagged variables (short_term243, long_term243). Note that the mean RMSE for NEE is an order of magnitude smaller than for Qh and Qle, and so we have used a different scale for NEE in the second row (Qh and Qle scale indicated on the right).

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 ~~a possible~~ possibly a lower distribution ~~predictability~~ uniqueness 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.

3.2 Determinants of predictability

Since there are a large number of hypotheses to test, only a selection of the most interesting results is shown here. We have also opted to show only the RMSE uniqueness in many plots, since its interpretation is the most straightforward, given the regression based nature of the empirical models, and since in many cases it correlates with some of the other metrics. Methods and plots for other hypotheses tested are included in the Supplementary Material, along with further details of some of the results presented below (including plots of the other 4 predictability metrics). As some determinants are not available for some sites, the number of site and model combinations in each analysis is noted in each figure title. The figures below use the same methodology as the Figure **??**.

3.2.1 Predictability as a function of energy and water

The three fluxes we investigate are clearly dependent on the availability of both water and energy. The availability of water is largely defined by precipitation, and temperature provides a proxy for the amount of energy available. We show the RMSE uniqueness for mean precipitation in Figure **??**. There appears to be some trend associated with precipitation indicating that the driest sites are more unique for all fluxes, particularly for NEE and Qle.

In Figure **??**, the RMSE uniqueness and RMSE mean are plotted as a scatter plot of mean annual temperature and mean annual precipitation. There appears to be some interaction between the two variables, with drier sites with a mean temperature around 20°C showing the highest uniqueness. As in Figure **??**, there is also some indication of higher RMSE mean for warmer sites in all fluxes.

~~Figure **??** shows RMSE uniqueness for aridity index. The pattern shown for each flux, and particularly NEE and Qle, is quite similar to that for mean precipitation in Figure **??**, with more very arid sites being less predictable.~~

~~RMSE uniqueness for Aridity Index. Two sites with an Aridity Index greater than 3 are excluded for readability, and 32 sites failed the Aridity Index calculation.~~

Figure **??** shows how the sites sit in the Budyko framework. The first row shows the sites on a standard Budyko diagram, with actual evaporation divided by mean annual precipitation on the y-axis, and potential evaporation divided mean annual precipitation on the x-axis. Theoretically, a site should fall just below the solid blue line, but location can be affected by available

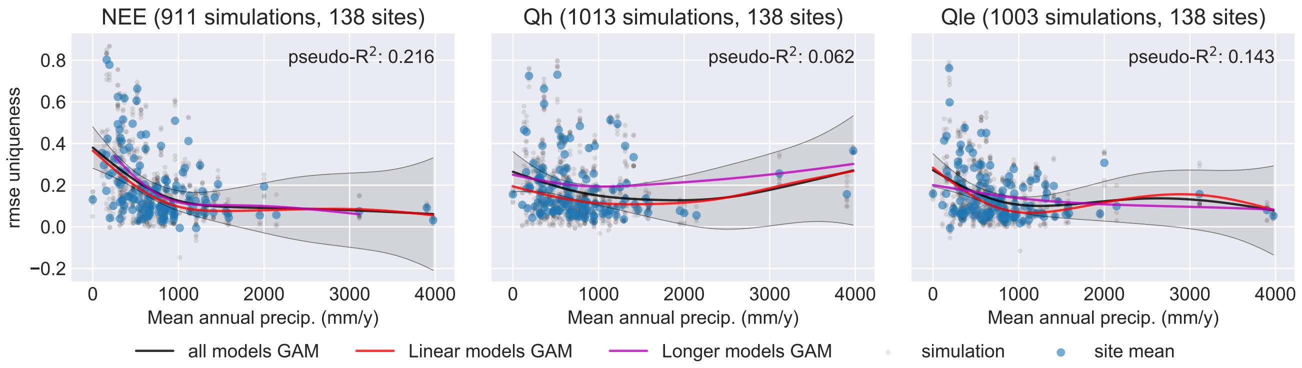

Figure 3. RMSE uniqueness for mean annual precipitation.

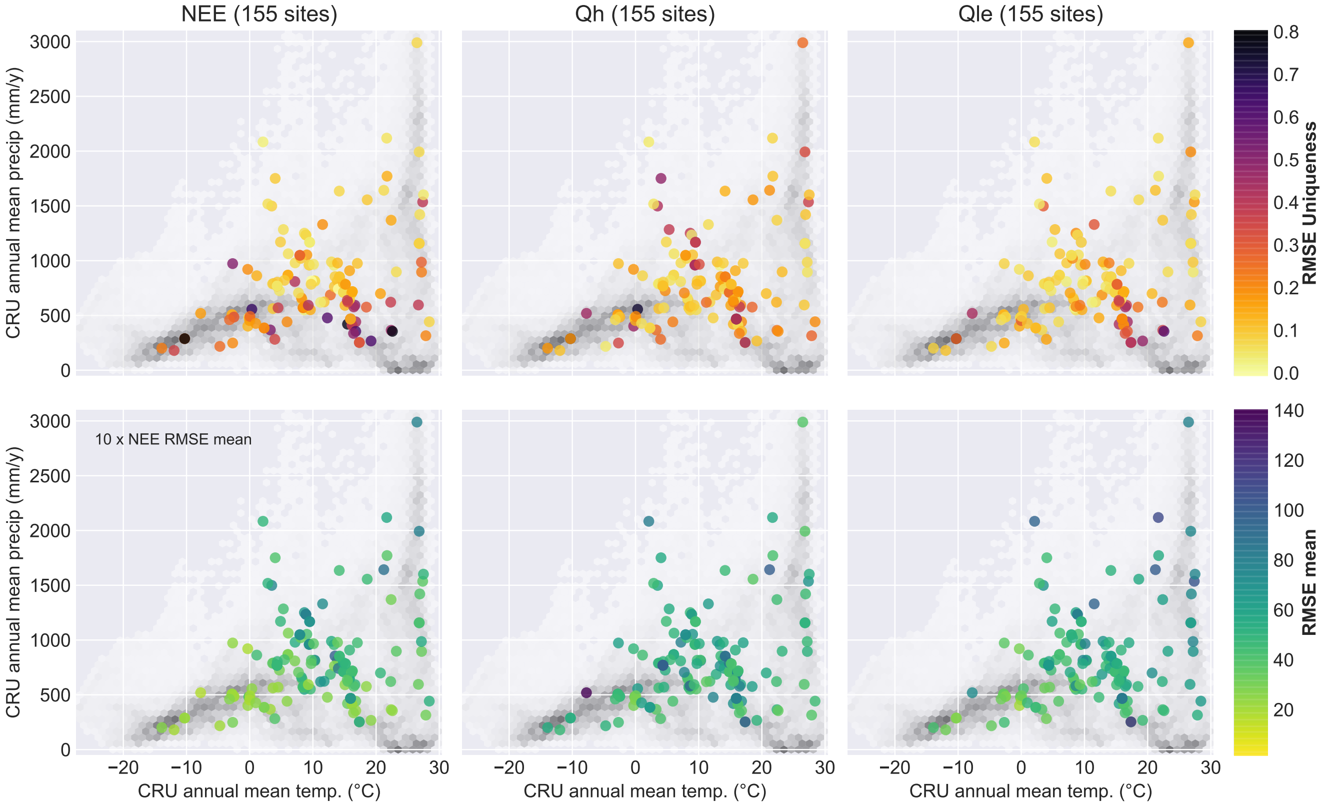

Figure 4. Predictability metrics for mean annual temperature vs mean annual precipitation (mean across models). The top row is RMSE uniqueness (darker colours indicate a more unique, less predictable site), and the bottom row is RMSE mean performance (darker colours indicate higher over-all RMSE). The grey underlying hexbin plot indicates the global distribution of mean precipitation and precipitation from the CRU dataset for all grid cells over land, to give an indication of the representativity of these sites.

water (e.g. inflow, or precipitation in the period before the measurement period), or the method of estimating potential evaporation. There do not appear to be strong patterns in the potential evapotranspiration uniqueness (see Supplementary material) and actual evapotranspiration appears to have some weak patterns (greater NEE uniqueness at sites with lower evaporation, and the opposite for Qh and Qle, see Supplementary Material), although these are not particularly clear in the Budyko diagrams in the first row. 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~~ uniqueness for NEE or Qle as a function of deviance from the Budyko curve (~~shown in the~~ see second row in Figure **??**), however there does seem to be some trend toward higher uniqueness for Qh for sites further above the Budyko curve~~(positive deviance)~~. Note that one site (AU-Lox) is excluded from this plot, as its values are too large (AET/MAR of 11.77, and a PET/MAR of 10.72). Its RMSE uniqueness values are 0.352 for NEE, 0.476 for Qh, and 0.438 for Qle. This site and the other sites with AET/MAR values over 2 are all Wetland sites, and as such are likely to have surface water available from upstream run-off in quantities far exceeding that due to precipitation alone.

3.2.2 ~~Predictability as a function of site variability~~

~~Variability of forcing variables is a major component in the predictability of fluxes. In general, we might expect higher variability to lead to lower predictability. Here we examine predictability at various time scales. Figure **??** shows the RMSE uniqueness over the interannual variability of temperature, and Figure **??** shows the same for precipitation, using the CRU TS4.01 data. There does not appear to be a strong trend in increased predictability with higher interannual temperature variability Figure **??**. However, there does appear to be a clear trend toward higher uniqueness at sites with stronger interannual variability in precipitation for NEE and for Qle (Figure **??**).~~

~~RMSE predictability by temperature interannual variability, calculated from the coefficient of variation in the CRU TS4.01 annual means.~~

~~RMSE predictability by precipitation interannual variability, calculated from the coefficient of variation in the CRU TS4.01 annual means.~~

~~Other modes of variability descending in scale include intra-annual variability, such as annual range, or variance of monthly values (seasonality); means of particular seasons; and diurnal ranges, as well as mixed-scale measurements, such as isothermality (ratio of diurnal range to annual range of temperature). Measures of each of these for both temperature and precipitation are included in the BioClim data, and plots of uniqueness as a function of each variable are included in the Supplementary Material. We do not include them here because, for the majority of cases, there appears to be no clear patterns of note. The exception includes some increase in RMSE uniqueness in NEE, and perhaps also for Qle, for sites with a higher diurnal temperature range (Figure **??**).~~

~~RMSE uniqueness for temperature diurnal range.~~

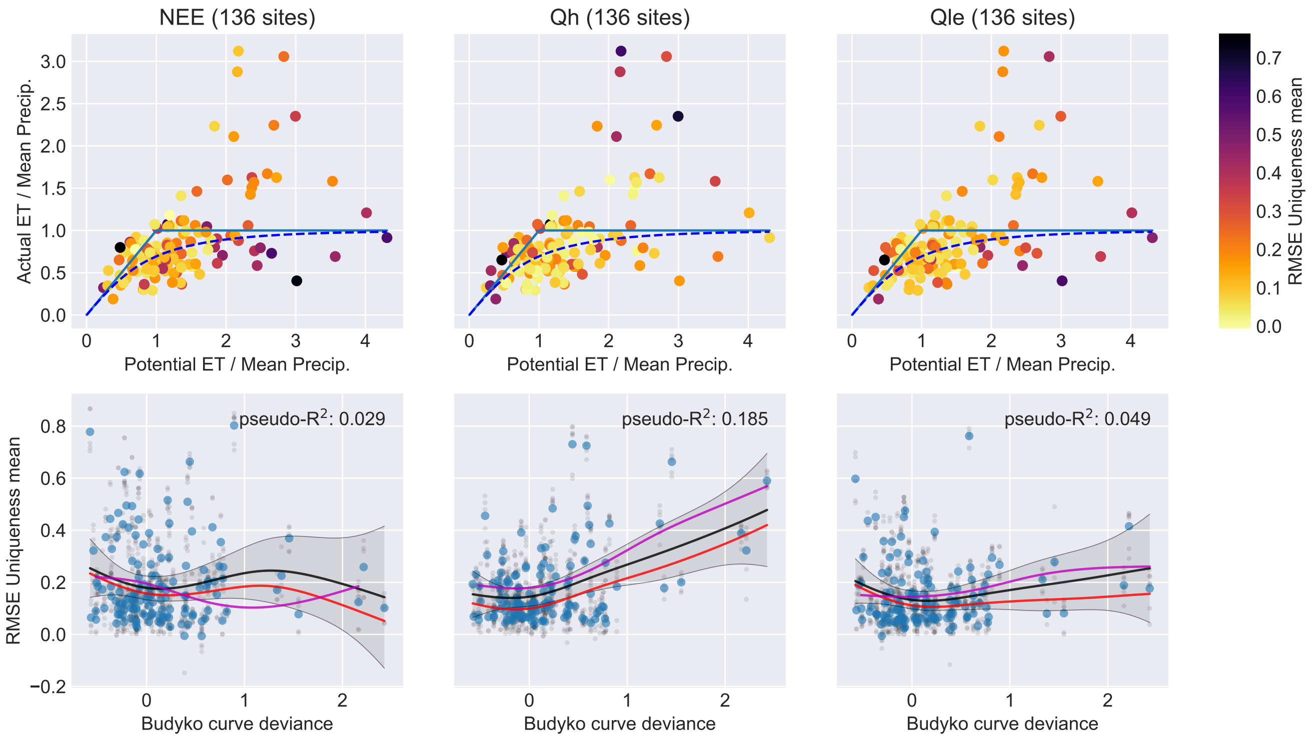

Figure 5. RMSE uniqueness for the Budyko analysis. AET and PET come from the FLUXNET 2015 data, mean precipitation comes from CRU TS 4.01. In the first row, colour indicates RMSE uniqueness, averaged across models, where darker colours are more unique. The solid line represents the theoretical energy and water limitations, and the dotted line represents the Budyko curve (Gerrits et al., 2009). The second row shows sites' deviance from the Budyko curve, normalised by the Budyko expectation for the site (sites > 0 lie above the curve in the first row).

3.2.2 Predictability as a function of vegetation characteristics

Vegetation type is a defining characteristic of different sites, and we would expect different vegetation types to behave differently, reflecting both their adaptations to their environment as well as their response to the ~~met~~ meteorological forcing. In 470 particular, we would expect the behaviour of some vegetation types to be more predictable than others. Figure **??** shows the RMSE uniqueness relative to grouped vegetation type (see methods). While there are some differences in uniqueness by vegetation type, few are significant. The main significant differences in RMSE uniqueness (Tukey's honest significant difference test of means across models per site, $p < 0.05$) are:

- For NEE, Shrubland sites tend to be more unique than all other vegetation types.

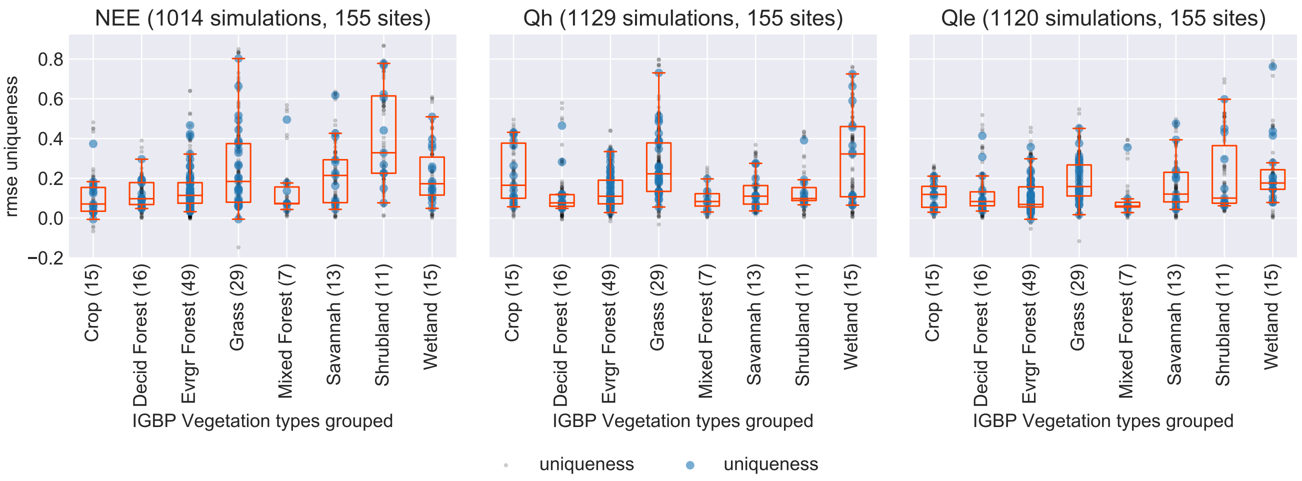

Figure 6. RMSE uniqueness for vegetation type (grouped, see Methods).

- For Qh, Wetlands are more unique than all Forest types, Shrubland and Savannah~~,~~; and Grass also tends to be more unique than Evergreen and Deciduous Forests, and Savannah.
- For Qle, Wetlands and Grasses tend to be more unique than Evergreen Forests.

However, there is still substantial overlap between even these groups, and the differences between the vegetation type groups are even less distinct when compared over the other five predictability metrics (see Supplementary Material).

3.2.3 Predictability as a function of geography

Globally, FLUXNET sites are not evenly distributed, both in space, and in climate regime. Figure **??** shows RMSE uniqueness for NEE as mapped globally, and averaged across models for each site. Given that the models are trained on all sites globally, and those sites are not evenly distributed around the globe (Figure **??**) we might expect that sites less well represented (more remote) would be more unique. In Figure **??**, there is a hint that more remote sites might be more unique for NEE. Such a pattern is not obvious in the maps for Qh or Qle (see Supplementary Material). To confirm this, we plotted uniqueness by remoteness~~(~~, defined as the average distance from a site to all other sites ~~) in Figure ??~~(shown in Figure 15 in the Supplementary material). There is a indeed a weak trend towards uniqueness at more remote sites for NEE, but not for Qh and Qle. There are no strong patterns evident in remoteness for any variable for any of the other predictability metrics (see Supplementary Material).

~~RMSE Uniqueness by remoteness (average distance to all other sites)~~

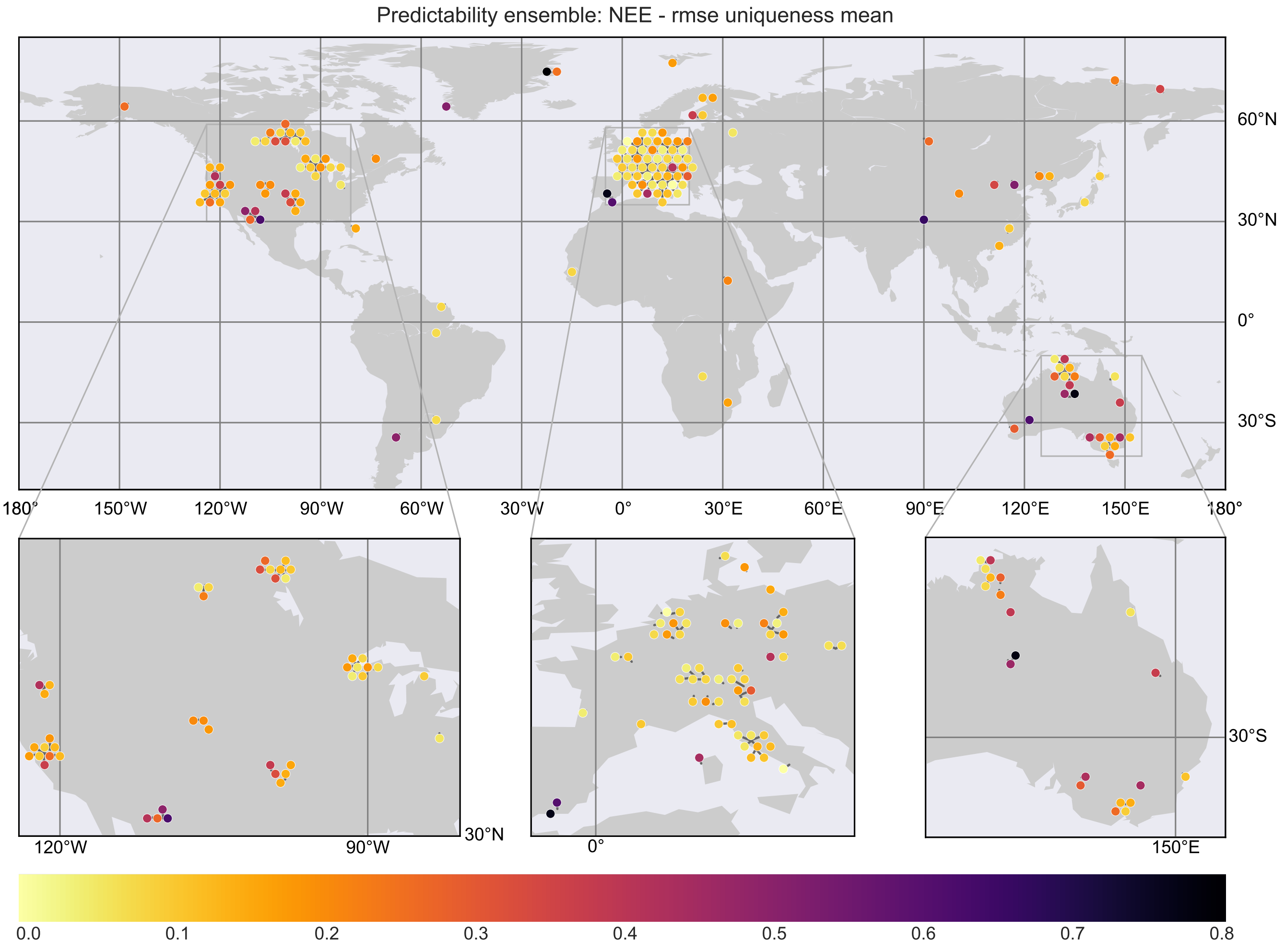

Figure 7. Map of NEE predictability - RMSE uniqueness, averaged across models, darker colours are more unique for NEE. In this map, sites are moved to avoid overlap, and a black line joins the site do to its' original location. This way the map gives a better idea of density of FLUXNET in different regions.

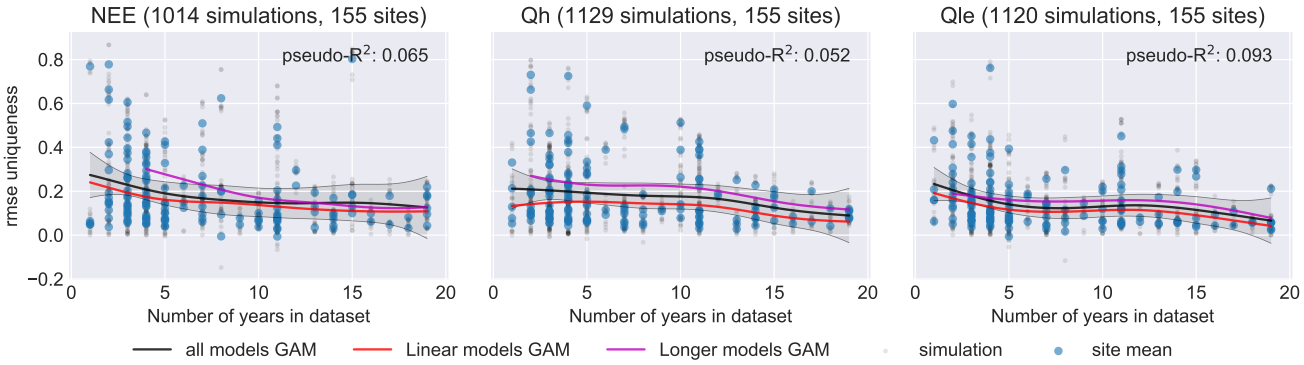

Figure 8. RMSE uniqueness for number of years in dataset.

3.2.4 Predictability as a function of data quality

There are a number of ways that data quality might affect uniqueness. We investigated the energy closure problem in FLUXNET by comparing predictability as a function of the actual energy closure imbalance, as well as the energy closure imbalance normalised by Rnet. While the energy closure problem in FLUXNET is perhaps one of the most obvious candidates for a determinant of a site's predictability, there does not appear to be a strong pattern in the data for RMSE uniqueness in either plot (nor for any of the other predictability metrics, see the Supplementary Material).

The number of years in the dataset is another obvious candidate determinant of predictability. There does seem to be a weak trend towards shorter sites being more unique, particularly for NEE and Qle (Figure **??**). This may be due to longer sites biasing the global training data such that the global model is more like their local models (and hence they appear less unique). This weak trend is somewhat visible in the other predictability metrics (see Supplementary Material, for example in mean Corr, and mean Overlap), but in each case is not strong enough to be significant.

Although the number of years gives a broad scale view of the amount of data in a dataset, it does not tell the whole story. For example, one 2-year site might contain almost a whole 2 years worth of good QC data, while another might contain less than a single year. As such, we also examined the ratio of good QC data to bad QC data at each site. Figure **??** shows the good QC ratio for the flux data combined . Like many of the other potential determinants of predictability, we did not find any clear patterns.

3.3 Predictability summary

While we have shown that predictability is affected to some degree by various factors (dryness and some vegetation types in particular), it is useful to be able to have an overview of the entire dataset. Figure **??** shows the mean RMSE uniqueness for each of the three fluxes, sorted alphabetically by FLUXNET site code. Here we see that uniqueness is somewhat consistent across variables at each site - Pearson correlation coefficients between variables are: NEE-Qh: 0.113, NEE-Qle: 0.536, Qh-Qle: 0.456.

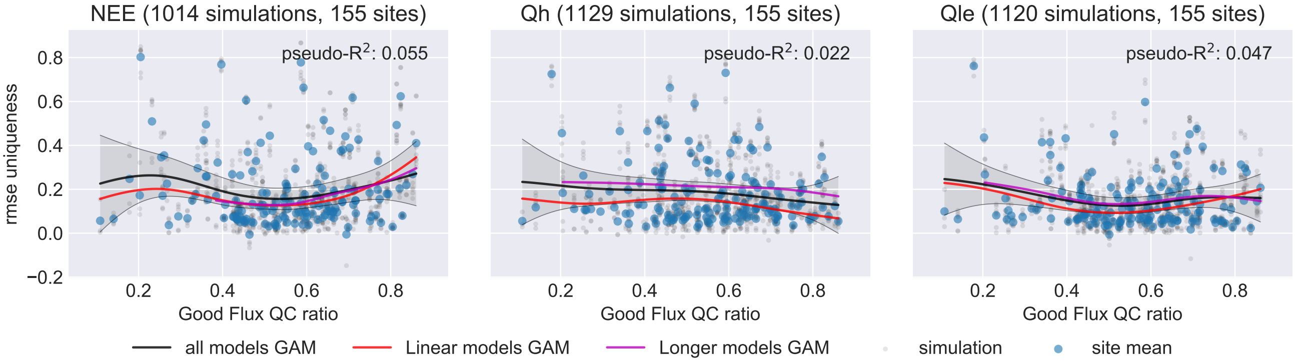

Figure 9. RMSE uniqueness for good QC ratio in flux forcings.

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 RMSE uniqueness 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.

4 Discussion

In this paper we applied a suite of empirical models to the 155 flux tower sites with half-hourly data included in Tier 1 release of FLUXNET. Our ~~aims were~~ aim was to explore how predictability varied across sites contingent on meteorological forcings, and then to use this insight into predictability to develop a more systematic approach to guide site selection in model evaluation exercises.

In our exploration of meteorological predictability, which we characterised as both uniqueness and mean performance for each metric, 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 over-all predictability. In an effort to clarify how uniqueness works in an intuitive way, we have laid out the relationships between meteorological drivers and site-specific drivers in Figure **??**.

Figure **??** 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 component of variability. In the left column,

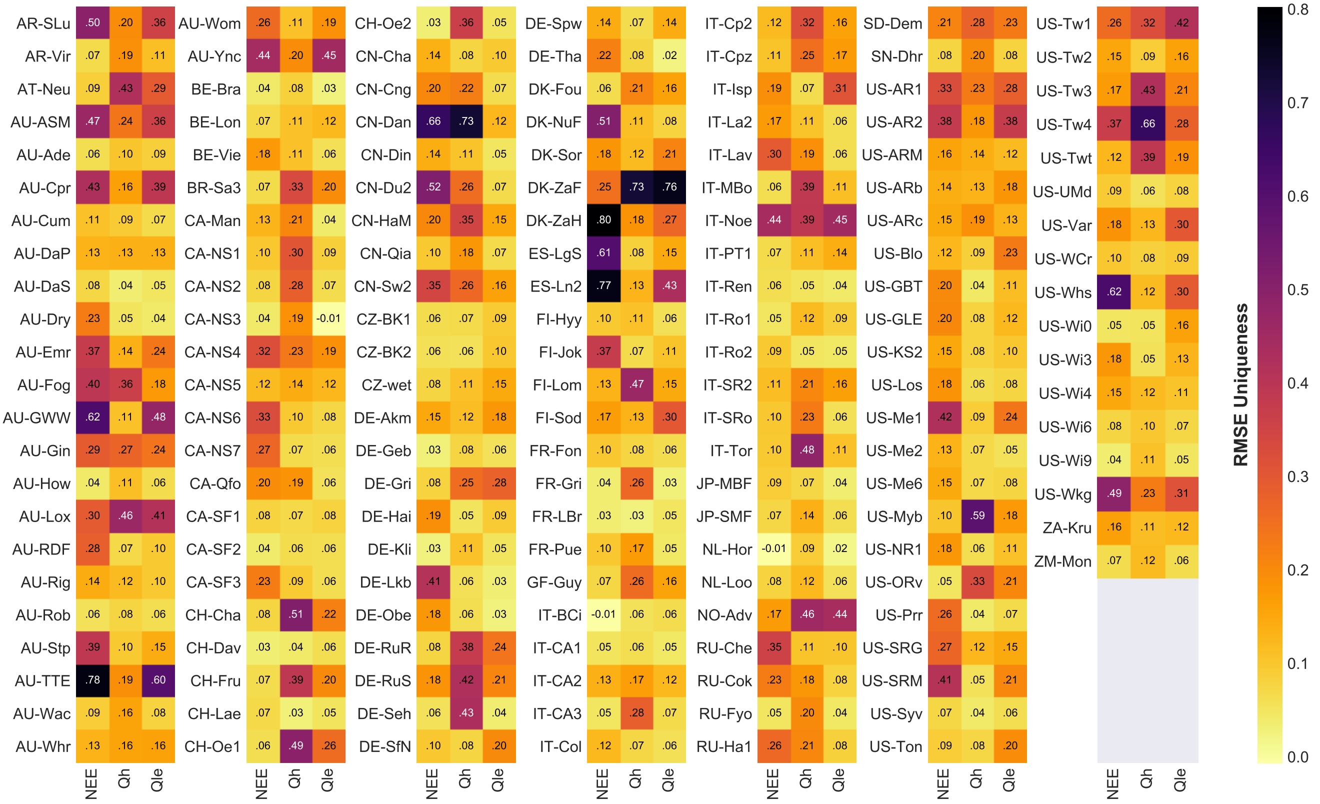

Figure 10. RMSE uniqueness mean across models for each flux at each site, in alphabetical order. Darker colours indicate ~~mode~~ more unique sites for each flux.

we have the 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 there is some local site condition that 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 modules 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 *is* a strong meteorological driver (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 correlated with "predictability". It seems 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

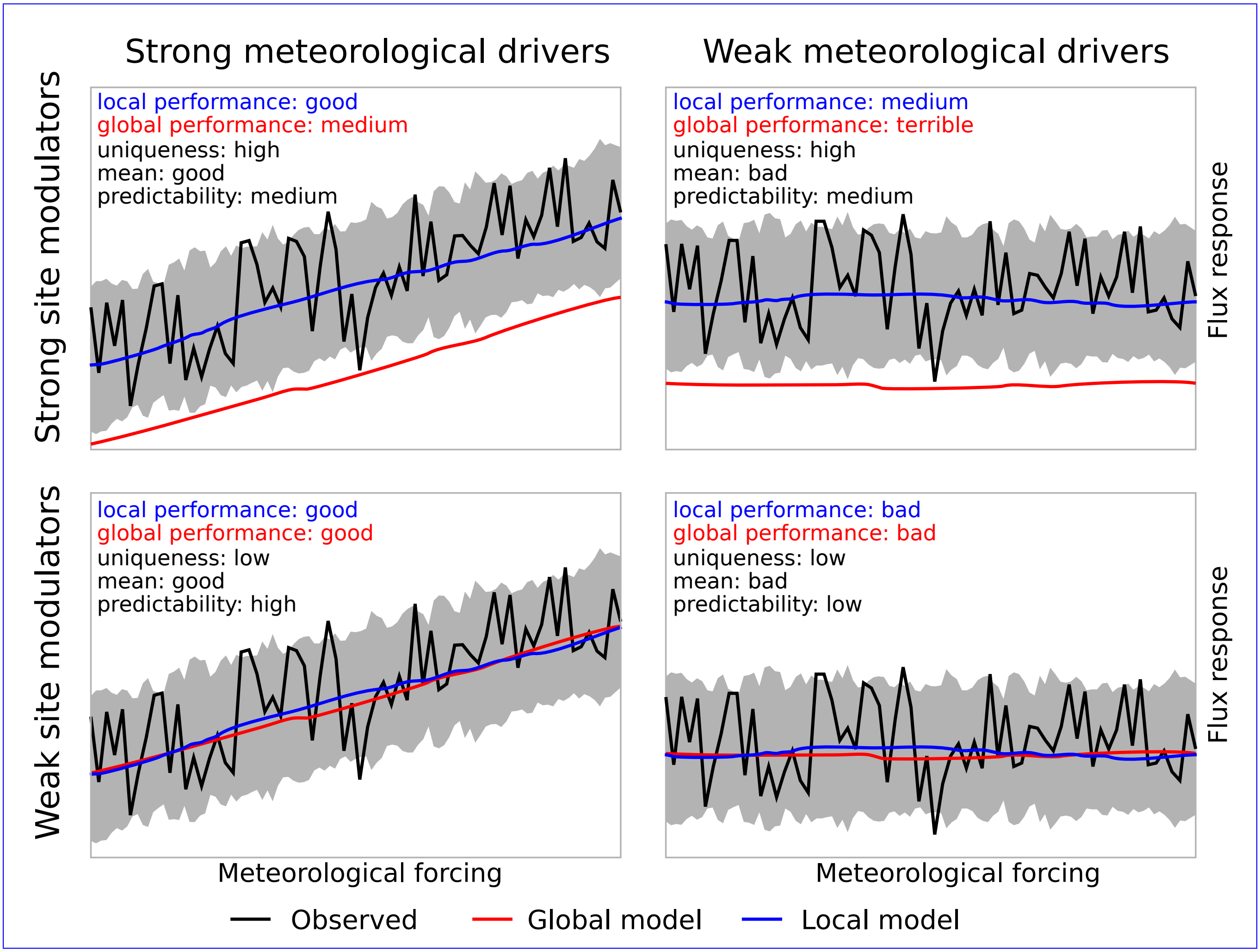

Figure 11. Schematic of sources of variability and how they affect predictability, mean performance, and uniqueness. The black line and grey ribbon represent the flux observations, and the unpredictable internal variability or noise in the system. "Meteorological drivers" indicate universal physical relationships between meteorological forcings and fluxes, here we have indicated this using a simple linear trend as an example. "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.

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.

4.1 Site predictability

Our multi-site analysis points to marked variability in predictability. For example, it appears that sites in warmer, drier climates tend to be more unique for all fluxes (Figure **??** ~~, Figure **??** and Figure **??**~~and Figure **??**), and sites with a large diurnal temperature range tend to be more unique, particularly for NEE, and to a lesser extent for Qle (~~Figure **??**~~See Figure 6 in the Supplementary Material). On the other hand, potential determinants that we expected to have quite strong effects on predictability did not appear to do so, for instance mean temperature (Figure **??**), dataset length (Figure **??**), and major vegetation types (Figure **??**). There are several reasons why this might have been the case.

First, the assumption that vegetation type is a major driver of flux behaviour may be wrong. It is perhaps more likely that the widely used approach of analysing FLUXNET sites grouped by a small number of discrete plant functional types is too simplistic, as opposed to exploring differences at a species level, or relating differences to a spectrum of plant traits, plant life spans and metabolism (Kattge et al., 2011; Reich et al., 1997; Wright et al., 2004). Despite widely acknowledged issues with this PFT approach (Alton, 2011; Pavlick et al., 2013; Van Bodegom et al., 2012), this analysis framework is still used, partly because this is the relevant interpretation metric that LSMs use, but also because the necessary information to dig deeper into site differences along these lines is still lacking. Whilst datasets do exist - for example TRY (Kattge et al., 2011), GLOPNET (Wright et al., 2004), LEDA (Kleyer et al., 2008), and ECOFLORA (Fitter and Peat, 1994) - often these are not freely available and the existing ancillary information relating to vegetation available via FLUXNET is minimal, which impedes analyses in this direction. This point was eloquently demonstrated by Konings and Gentine (2016), who used data from the AMSR-E satellite to characterise global variations in isohydricity (the degree to which plants regulate their stomata as leaf water potential declines). When they categorised their analysis on a PFT level, differences between sites and species were no longer distinct. This remains an avenue ripe for future analysis.

Second, our predictability metrics (RMSE, Corr, Overlap uniqueness and mean) may not be appropriate. There may be systematic biases that inhibit our estimate of predictability due to over-representation of particular biomes, or because measurement periods were not representative. ~~However, the analysis in Figure **??** indicates that there is not a strong trend towards more remote sites being harder to predict, providing some support to our methodology.~~ Our results did indicate a weak trend toward higher uniqueness in sites with shorter measurement periods (see Figure **??**), however, a single year of flux data represents a substantial amount of useful data. Short datasets may also be particularly useful if they happen to include rare events that are not well represented in other datasets, such as regional droughts or heatwaves. We nevertheless openly invite constructive arguments against our predictability metric proposal, identification of flaws in the process, or alternative definitions of site predictability or uniqueness.

Should we expect stronger patterns of predictability? In our view, there are strong arguments to support the utility of the FLUXNET data for analyses of predictability. We know that meteorological data measured at flux tower sites does contain a great deal of information about the measured fluxes (Best et al., 2015; Haughton et al., 2018). Indeed the information contained in the meteorological data about fluxes was very much consistent across sites and this was key to the success of those experiments. So we know that the empirical models used here, which follow a very similar methodology, are capturing the relationships between the meteorological forcing and the predicted fluxes relatively well. MDK: support this?

One way we might improve upon ~~this~~ our analysis is by focusing on the differences in performance or uniqueness between models with similar structure, but with extra forcing variables. This would tell us something about the predictability contingent on that variable. For instance, if a model such as STH_km243 (a 243-centre cluster and regression on shortwave down, air temperature, and relative humidity, see Table **??**) performs substantially better at a class of sites than an ST_km243 model (the same, but missing relative humidity), then we can ~~say~~ infer that predictability at those sites may be contingent on information in the humidity data. This analysis is substantially more complex, and so we have left it for future work. The code used to run these models is freely available at https://github.com/naught101/empirical_lsm, version 1.1 was used for this paper.

4.2 Model evaluation

Our second major aim was to develop a more systematic approach for LSM evaluation underpinned by differences in site predictability. Recent work has already illustrated the benefits of defining benchmark levels of performance for a given metric, at a given site (Best et al., 2015; Haughton et al., 2016). The empirical analysis of site predictability we presented goes one step further, effectively quantifying the additional benefit to model performance that site-specific information can provide in the form of the locally trained empirical models.

Land surface modellers will usually rationalise why a particular module was selected to represent a physical process, or why a specific atmospheric model was used. Given the new information presented in this paper, we suggest that a thorough rationale for why specific FLUXNET sites were used should be explicit in future publications. Importantly, we note that we could not provide evidence that would support site choices based on PFT (Figure **??**), data length (Figure **??**), quality control (Figure **??**) and so these really do not seem legitimate ways to rationalise choice of sites. 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 will be dependent on many factors. Instead, we recommend that the predictability of the site is one element for choosing sites, but the process of selecting sites should be more rigorous and reported whether or not this recommendation is followed.

How might this site-specific information be used? Perhaps most obvious would be the clustering of sites, based on their predictability, for use in model evaluation and benchmarking exercises. In Figure **??**, we provided some initial guidance to the LSM evaluation community. Here, sites shown in darker colours are sites that exhibit unusual meteorological-flux relationships for a given flux. These are the sites that are likely to present more of a challenge for process-based LSMs to simulate. On the flip side, lighter coloured sites follow commonly observed patterns of behaviour, so good LSM performance at these should be less surprising, and is perhaps less of an achievement. What is important is that modellers should know if the sites they are evaluating their models against are relatively predictable, or unpredictable. Our results, and Figure **??** in particular, gives modellers a tool that can form the basis of a strategy to choose sites, a defence if they choose unpredictable sites and do poorly, and a challenge if they choose more predictable sites and do well. We suspect that the best ~~general~~ strategy for model evaluation would be to pick a set of sites that includes both very predictable sites, as well as very unpredictable sites, with a distribution informed by the determinants of predictability presented above.

Of particular note in Figure **??**, but of interest beyond LSM evaluation, is that predictability can be markedly different for different surface fluxes at the same site. For example, we see a number of sites with high NEE and Qle uniqueness, and low Qh uniqueness (e.g. AU-TTE, AU-Ync, ES-Ln2, US-Whs, US-Wkg), and other sites with high Qh uniqueness (e.g. CH-Cha, IT-MBo, IT-Tor, US-Myb). We also see some neighbouring sites with extremely different predictability responses for different fluxes (e.g. DK-ZaF, a wetland site, has very high NEE uniqueness, while the neighbouring DK-ZaH, a heath [grasslands] site, has high Qh and Qle uniqueness). MDK: is there a table with these sites full names? If so refer to it This is evident in other figures where there uniqueness patterns are not shared between fluxes (for example the differences between Qh and the other fluxes in ~~Figures **??**and **??**~~Figure **??**). This provides new justification for different site selection strategies depending on the processes being evaluated.

MDK: should this paragraph come above the diversity one? Seems to follow better there as there is some repetition?

Our analysis may understandably lead to modelling groups gravitating toward evaluating their models only against a specific sub-sets of FLUXNET sites. We do not think that this is a desirable outcome, and thus have not provided a suggestion of specific sites to use. Indeed care must be taken when evaluating models on small groups of FLUXNET sites due to the greater need to consider the various intricacies of site-specific behaviour. When models are evaluated against a large number of sites, an argument can be advanced that unique site behaviour may average out in the noise. If analysis approaches like ours were to lead to small groups of sites being used to evaluate models, greater care would be needed to capture an adequate diversity of site characteristics. For example, it may be that sites we determine to be unique are simply those that have undergone a disturbance event (e.g. clear felling, fire, wind storms, etc.), or are subject to management (e.g. cropping, irrigation). With improved information about site characteristics (e.g. time since last disturbance), these issues could be avoided. A major advance that would be useful to the LSM community would be the systematic publishing of metadata characterising each site in the FLUXNET data.

Finally, the logical next extension of our work is to evaluate a suite of LSMs at the sites deemed to be most and least predictable, in order to understand the extent to which site predictability translates into model skill. Such an analysis will of course need careful consideration of the kinds of site eccentricities noted above, noting that information about these eccentricities is not as commonly available as flux and meteorological data. Nevertheless, work of this kind will ultimately help refine how this predictability metric is best utilised in model evaluation strategies.

5 Conclusions

In this study, we applied a novel methodology to characterise the predictability of surface fluxes at sites within the FLUXNET2015 dataset. We had two key aims: first, we sought to explain why predictability varied across the 155 FLUXNET sites, with the expectation that we would find patterns in predictability along gradients such as ~~aridity~~mean precipitation, vegetation type, or in relation to various bioclimatic metrics, both annually and seasonally. Whilst we did show that the 155 FLUXNET sites vary strongly in their predictability, we did not find ~~especially~~ strong patterns in predictability, with the possible exception of ~~aridity~~drier sites, which appear to be substantially more unique. We acknowledge that we might have missed some relevant determinants of predictability, or some transformation of, or interaction between the determinants that we did have available. If we could incorporate these, a clear pattern of predictability might emerge.

Our second aim was to propose a more systematic approach to site selection for model evaluation, underpinned by differences in site predictability. While we found fewer patterns in predictability that we expected, we nevertheless now have a basis on which to define *a priori* expectations of model performance. We suggest that careful choice of FLUXNET sites based on predictability may avoid modellers incorrectly judging their models negatively (via choice of very unpredictable sites) or positively (via choice of very predictable sites). While further work based on this predictability metric is required before a complete rationale for site selection is obvious, we now have a basis on which to develop such a strategy. As a first step, we strongly encourage modelling groups to explain why they choose specific sites for evaluation because, thanks to the FLUXNET community, a lack of availability of data is no longer a reason for site selection.

References

Abbas, S. R. and Arif, M.: New Time Series Predictability Metrics for Nearest Neighbor Based Forecasting, in 2006 IEEE International Multitopic Conference, pp. 100–105., 2006.

Alton, P. B.: How useful are plant functional types in global simulations of the carbon, water, and energy cycles?, J. Geophys. Res., 116(G1), doi:10.1029/2010JG001430, 2011.

Balsamo, G., Beljaars, A., Scipal, K., Viterbo, P., van den Hurk, B., Hirschi, M. and Betts, A. K.: A Revised Hydrology for the ECMWF Model: Verification from Field Site to Terrestrial Water Storage and Impact in the Integrated Forecast System, J. Hydrometeor., 10(3), 623–643, doi:10.1175/2008JHM1068.1, 2009.

Best, M. J., Abramowitz, G., Johnson, H. R., Pitman, A. J., Balsamo, G., Boone, A., Cuntz, M., Decharme, B., Dirmeyer, P. A., Dong, J., Ek, M. B., Guo, Z., Haverd, V., van den Hurk, B. J. J., Nearing, G. S., Pak, B., Peters-Lidard, C. D., Santan, J. S., Stevens, L. E. and Vuichard, N.: The Plumbing of Land Surface Models: Benchmarking Model Performance, J. Hydrometeor, 16(3), 1425–1442, doi:10.1175/JHM-D-14-0158.1, 2015.

Bonan, G.: Ecological Climatology: Concepts and Applications, 3rd ed., Cambridge University Press, National Center for Atmospheric Research, Boulder, Colorado., 2015.

Bonan, G. B., Williams, M., Fisher, R. A. and Oleson, K. W.: Modeling stomatal conductance in the earth system: Linking leaf water-use efficiency and water transport along the soilplantatmosphere continuum, Geosci. Model Dev., 7(5), 2193–2222, doi:10.5194/gmd-7-2193-2014, 2014.

Chapin III, F. S., Matson, P. A. and Vitousek, P.: Principles of Terrestrial Ecosystem Ecology, 2nd ed., Springer-Verlag, New York. [online] Available from: //www.springer.com/gp/book/9781441995032 (Accessed 3 April 2018), 2011.

Chen, T. H., Henderson-Sellers, A., Milly, P. C. D., Pitman, A. J., Beljaars, A. C. M., Polcher, J., Abramopoulos, F., Boone, A., Chang, S. and others: Cabauw experimental results from the project for intercomparison of land-surface parameterization schemes, Journal of Climate, 10(6), 1194–1215 [online] Available from: http://journals.ametsoc.org/doi/abs/10.1175/1520-0442(1997)010/%3C1194:CERFTP/%3E2.0.CO;2 (Accessed 10 April 2015), 1997.

Colwell, R. K.: Predictability, Constancy, and Contingency of Periodic Phenomena, Ecology, 55(5), 1148–1153, doi:10.2307/1940366, 1974.

De Kauwe, M. G., Kala, J., Lin, Y.-S., Pitman, A. J., Medlyn, B. E., Duursma, R. A., Abramowitz, G., Wang, Y. P. and Miralles, D. G.: A test of an optimal stomatal conductance scheme within the CABLE land surface model, Geosci. Model Dev., 8(2), 431–452, doi:10.5194/gmd-8-431-2015, 2015.

Del Grosso, S., Parton, W., Stohlgren, T., Zheng, D., Bachelet, D., Prince, S., Hibbard, K. and Olson, R.: Global potential net primary production predicted from vegetation class, precipitation, and temperature, Ecology, 89(8), 2117–2126, 2008.

~~Dickinson, E., Henderson-Sellers, A., Kennedy, J. and Wilson, F.: Biosphere-atmosphere Transfer Scheme (BATS) for the NCAR Community Climate Model, doi:, 1986.~~

~~Donat, M. G., Lowry, A. L., Alexander, L. V., O'Gorman, P. A. and Maher, N.: More extreme precipitation in the world's dry and wet regions, Nature Climate Change, 6(5), 508–513, doi:, 2016.~~

Fitter, A. H. and Peat, H. J.: The Ecological Flora Database, Journal of Ecology, 82(2), 415–425, doi:10.2307/2261309, 1994.

Fluxdata.org: FLUXNET2015 Dataset, Fluxdata [online] Available from: http://fluxnet.fluxdata.org/data/fluxnet2015-dataset/ (Accessed 8 January 2018), 2018.

Gerrits, A. M. J., Savenije, H. H. G., Veling, E. J. M. and Pfister, L.: Analytical derivation of the Budyko curve based on rainfall characteristics and a simple evaporation model, Water Resour. Res., 45(4), W04403, doi:10.1029/2008WR007308, 2009.

Gillman, L. N., Wright, S. D., Cusens, J., McBride, P. D., Malhi, Y. and Whittaker, R. J.: Latitude, productivity and species richness, Global Ecology and Biogeography, 24(1), 107–117, doi:10.1111/geb.12245, 2015.

Harris, I. and Jones, P.: CRU TS4.01: Climatic Research Unit (CRU) Time-Series (TS) version 4.01 of high-resolution gridded data of month-by-month variation in climate (Jan. 1901- Dec. 2016), Centre for Environmental Data Analysis, 623–642 [online] Available from: http://dx.doi.org/10.5285/58a8802721c94c66ae45c3baa4d814d0 (Accessed 20 March 2018), 2017.

Haughton, N., Abramowitz, G., Pitman, A. J., Or, D., Best, M. J., Johnson, H. R., Balsamo, G., Boone, A., Cuntz, M., Decharme, B., Dirmeyer, P. A., Dong, J., Ek, M., Guo, Z., Haverd, V., van den Hurk, B. J. J., Nearing, G. S., Pak, B., Santanello,

J. A., Stevens, L. E. and Vuichard, N.: The Plumbing of Land Surface Models: Is Poor Performance a Result of Methodology or Data Quality?, J. Hydrometeor., 17(6), 1705–1723, doi:10.1175/JHM-D-15-0171.1, 2016.

Haughton, N., Abramowitz, G. and Pitman, A. J.: On the Predictability of Land Surface Fluxes from Meteorological Variables, Geosci. Model Dev., 11(1), 195–212, doi:10.5194/gmd-11-195-2018, 2018.

Huston, M. A. and Wolverton, S.: The global distribution of net primary production: Resolving the paradox, Ecological Monographs, 79(3), 343–377, doi:10.1890/08-0588.1, 2009.

Kaboudan, M. A.: Genetic Programming Prediction of Stock Prices, Computational Economics, 16(3), 207–236, doi:10.1023/A:1008768 2000.

Kattge, J., Díaz, S., Lavorel, S., Prentice, I. C., Leadley, P., Bönisch G., Garnier E., Westoby M., Reich P. B., Wright I. J., Cornelissen J. H. C., Violle C., Harrison S. P., Van BODEGOM P. M., Reichstein M., Enquist B. J., Soudzilovskaia N. A., Ackerly D. D., Anand M., Atkin O., Bahn M., Baker T. R., Baldocchi D., Bekker R., Blanco C. C., Blonder B., Bond W. J., Bradstock R., Bunker D. E., Casanoves F., Cavender-Bares J., Chambers J. Q., Chapin Iii F. S., Chave J., Coomes D., Cornwell W. K., Craine J. M., Dobrin B. H., Duarte L., Durka W., Elser J., Esser G., Estiarte M., Fagan W. F., Fang J., Fernández-Méndez F., Fidelis A., Finegan B., Flores O., Ford H., Frank D., Freschet G. T., Fyllas N. M., Gallagher R. V., Green W. A., Gutierrez a. G., Hickler T., Higgins S. I., Hodgson J. G., Jalili A., Jansen S., Joly C. A., Kerkhoff a. J., Kirkup D., Kitajima K., Kleyer M., Klotz S., Knops J. M. H., Kramer K., Kühn I., Kurokawa H., Laughlin D., Lee T. D., Leishman M., Lens F., Lenz T., Lewis S. L., Lloyd J., Llusià J., Louault F., Ma S., Mahecha M. D., Manning P., Massad T., Medlyn B. E., Messier J., Moles a. T., Müller S. C., Nadrowski K., Naeem S., Niinemets Ü., Nöllert S., Nüske A., Ogaya R., Oleksyn J., Onipchenko V. G., Onoda Y., Ordoñez J., Overbeck G., et al.: TRY a global database of plant traits, Global Change Biology, 17(9), 2905–2935, doi:10.1111/j.1365-2486.2011.02451.x, 2011.

Kleyer, M., Bekker, R. M., Knevel, I. C., Bakker, J. P., Thompson, K., Sonnenschein M., Poschlod P., Van Groenendael J.M., Klimeš L., Klimešová J., Klotz S., Rusch G.M., Hermy M., Adriaens D., Boedeltje G., Bossuyt B., Dannemann A., Endels P., Götzenberger L., Hodgson J.G., Jackel A-K., Kühn I., Kunzmann D., Ozinga W.A., Römermann C., Stadler M., Schlegelmilch J., Steendam H.J., Tackenberg O., Wilmann B., Cornelissen J.H.C., Eriksson O., Garnier E. and Peco B.: The

LEDA Traitbase: A database of life-history traits of the Northwest European flora, Journal of Ecology, 96(6), 1266–1274, doi:10.1111/j.1365-2745.2008.01430.x, 2008.

Konings, A. G. and Gentine, P.: Global variations in ecosystem-scale isohydricity, Global Change Biology, 23(2), 891–905, doi:10.1111/gcb.13389, 2016.

Kottek, M., Grieser, J., Beck, C., Rudolf, B. and Rubel, F.: World Map of the Köppen-Geiger climate classification updated, Meteorologische Zeitschrift, 259–263, doi:10.1127/0941-2948/2006/0130, 2006.

Krinner, G., Viovy, N., de Noblet-Ducoudré, N., Ogée, J., Polcher, J., Friedlingstein, P., Ciais, P., Sitch, S. and Prentice, I. C.: A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system, Global Biogeochem. Cycles, 19(1), GB1015, doi:10.1029/2003GB002199, 2005.

Law, B.: AmeriFlux US-Me1 Metolius - Eyerly burn, AmeriFlux; Oregon State Univ., Corvallis, OR (United States)., 2016.

Lawrence, D. M., Oleson, K. W., Flanner, M. G., Thornton, P. E., Swenson, S. C., Lawrence, P. J., Zeng, X., Yang, Z.-L., Levis, S., Sakaguchi, K., Bonan, G. B. and Slater, A. G.: Parameterization improvements and functional and structural advances in Version 4 of the Community Land Model, J. Adv. Model. Earth Syst., 3(1), M03001, doi:10.1029/2011MS00045, 2011.

~~Manabe, S.: Climate And The Ocean Circulation 1: I. The Atmospheric Circulation And The Hydrology Of The Earth's Surface, Mon. Wea. Rev., 97(11), 739–774, doi:, 1969.~~

~~Milly, P. C. D. and Dunne, K. A.: Potential evapotranspiration and continental drying, Nature Climate Change, 6(10), nclimate3046, doi:, 2016.~~

Napoly, A., Boone, A., Samuelsson, P., Gollvik, S., Martin, E., Seferian, R., Carrer, D., Decharme, B. and Jarlan, L.: The interactions between soilbiosphereatmosphere (ISBA) land surface model multi-energy balance (MEB) option in SURFEXv8

2: Introduction of a litter formulation and model evaluation for local-scale forest sites, Geosci. Model Dev., 10(4), 1621–1644, doi:10.5194/gmd-10-1621-2017, 2017.

Pavlick, R., Drewry, D. and Kleidon, A.: Quantifying the importance of plant functional diversity for ecosystem functioning and resilience under scenarios of climate change (Invited), AGU Fall Meeting Abstracts, 32, B32D–02 [online] Available from: http://adsabs.harvard.edu/abs/2013AGUFM.B32D..02P (Accessed 20 March 2018), 2013.

Perkins, S. E., Pitman, A. J., Holbrook, N. J. and McAneney, J.: Evaluation of the AR4 Climate Models' Simulated Daily Maximum Temperature, Minimum Temperature, and Precipitation over Australia Using Probability Density Functions, J. Clim., 20(17), 4356–4376, doi:10.1175/JCLI4253.1, 2007.

Pitman, A. J.: The evolution of, and revolution in, land surface schemes designed for climate models, Int. J. Climatol., 23(5), 479–510, 2003.

Reich, P. B., Walters, M. B. and Ellsworth, D. S.: From tropics to tundra: Global convergence in plant functioning, PNAS, 94(25), 13730–13734 [online] Available from: http://www.pnas.org/content/94/25/13730 (Accessed 20 March 2018), 1997.

Schuur, E. A. G.: Productivity and global climate revisited: The sensitivity of tropical forest growth to precipitation, Ecology, 84(5), 1165–1170, doi:10.1890/0012-9658(2003)084[1165:PAGCRT]2.0.CO;2, 2003.

Servén, D.: pyGAM: Generalized Additive Models in Python, [online] Available from: https://github.com/dswah/pyGAM (Accessed 15 March 2018), 2018.

Slevin, D., Tett, S. F. B. and Williams, M.: Multi-site evaluation of the JULES land surface model using global and local data, Geosci. Model Dev., 8(2), 295–316, doi:10.5194/gmd-8-295-2015, 2015.

Ukkola, A. M., Haughton, N., De Kauwe, M. G., Abramowitz, G. and Pitman, A. J.: FluxnetLSM R package (v1.0): A community tool for processing FLUXNET data for use in land surface modelling, Geosci. Model Dev., 2017(9), 3379–3390, doi:10.5194/gmd-10-3379-2017, 2017.

Van Bodegom, P. M., Douma, J. C., Witte, J. P. M., Ordoñez, J. C., Bartholomeus, R. P. and Aerts, R.: Going beyond limitations of plant functional types when predicting global ecosystematmosphere fluxes: Exploring the merits of traits-based approaches, Global Ecology and Biogeography, 21(6), 625–636, doi:10.1111/j.1466-8238.2011.00717.x, 2012.

Wang, Y. P., Kowalczyk, E. A., Leuning, R., Abramowitz, G., Raupach, M. R., Pak, B., van Gorsel, E. and Luhar, A.: Diagnosing errors in a land surface model (CABLE) in the time and frequency domains, J. Geophys. Res., 116(G1), doi:10.1029/2010JG001385, 2011.

Whittaker, R. H.: Classification of natural communities, Bot. Rev, 28(1), 1–239, doi:10.1007/BF02860872, 1962.

Wilson, K., Goldstein, A., Falge, E., Aubinet, M., Baldocchi, D., Berbigier, P., Bernhofer, C., Ceulemans, R., Dolman, H., Field, C., Grelle, A., Ibrom, A., Law, B. E., Kowalski, A., Meyers, T., Moncrieff, J., Monson, R., Oechel, W., Tenhunen, J., Valentini, R. and Verma, S.: Energy balance closure at FLUXNET sites, Agricultural and Forest Meteorology, 113(14), 223–243, doi:10.1016/S0168-1923(02)00109-0, 2002.

~~WorldClim: WorldClim Version2, onlineAvailable from: (Accessed 9 January 2018), 2016.~~

Wright, I. J., Reich, P. B., Westoby, M., Ackerly, D. D., Baruch, Z., Bongers, F., Cavender-Bares, J., Chapin, T., Cornelissen, J. H. C., Diemer, M., Flexas, J., Garnier, E., Groom, P. K., Gulias, J., Hikosaka, K., Lamont, B. B., Lee, T., Lee, W., Lusk, C., Midgley, J. J., Navas, M.-L., Niinemets, U., Oleksyn, J., Osada, N., Poorter, H., Poot, P., Prior, L., Pyankov, V. I., Roumet, C., Thomas, S. C., Tjoelker, M. G., Veneklaas, E. J. and Villar, R.: The worldwide leaf economics spectrum, Nature, 428(6985), 821–827, doi:10.1038/nature02403, 2004.

Competing interests.

Acknowledgements. “We acknowledge the support of the Australian Research Council Centre of Excellence for Climate System Science (CE110001028) and the ARC Centre of Excellence for Climate Extremes (CE170100023). This work used eddy covariance data acquired and shared by the FLUXNET community, including these networks: AmeriFlux, AfriFlux, AsiaFlux, CarboAfrica, CarboEuropeIP, CarboItaly, CarboMont, ChinaFlux, Fluxnet-Canada, GreenGrass, ICOS, KoFlux, LBA, NECC, OzFlux-TERN, TCOS-Siberia, and USCCC. The ERA-Interim reanalysis data are provided by ECMWF and processed by LSCE. The FLUXNET eddy covariance data processing and harmonization was carried out by the European Fluxes Database Cluster, AmeriFlux Management Project, and Fluxdata project of FLUXNET, with the support of CDIAC and ICOS Ecosystem Thematic Center, and the OzFlux, ChinaFlux and AsiaFlux offices. A full list of Fluxnet sites and citations is included in the Supplementary material. We would also like to thank two anonymous reviewers for insightful comments that helped improve the paper.”