Biogeosciences Discuss.,
https://doi.org/10.5194/bg-2018-179-RC1, 2018

Biogeosciences Discussions

Interactive comment on "Does predictability of fluxes vary between FLUXNET sites?" *by* Ned Haughton et al.

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.

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?).

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

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

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.

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.

- 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).

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

- 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?

- 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?

- 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.

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?

- 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?

- 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$_{t}ermandshort_{t}ermnotationsareneverintroduced.$

- Pg 6, Table 1: Did the long$_{t}erm729modelneverhaveasuccessfulsimulation?Inwhichcaseitu$

- 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?

- 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?

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

- 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?

- 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.

- 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.

- 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.

- 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?

- 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.

- 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?

- 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.

- 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.

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

- 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?

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

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

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

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

- 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.

- 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.

- 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.

Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. Journal of Hydrology 377, 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Tramontana, G., Jung, M., Schwalm, C.R., Ichii, K., Camps-Valls, G., Ráduly, B., Reichstein, M., Arain, M.A., Cescatti, A., Kiely, G., Merbold, L., Serrano-Ortiz, P., Sickert, S., Wolf, S., Papale, D., 2016. Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms. Biogeosciences 13, 4291–4313. https://doi.org/10.5194/bg-13-4291-2016

Interactive comment on Biogeosciences Discuss., https://doi.org/10.5194/bg-2018-179, 2018.