Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2017-153-RC2, 2017

Geoscientific Model Development Discussions

Interactive comment on "On the Predictability of Land Surface Fluxes from Meteorological Variables" *by* Ned Haughton et al.

Anonymous Referee #2

Received and published: 27 August 2017

General Comments:

This article follows on and extends Best et al. 2015 to create new empirical benchmarks that can be used for evaluating the performance of land surface models. The authors choose empirical models that maximize the amount of available information in the forcing data that is supplied to LSMs. A main result of the study is the production of an ensemble of benchmarks that the authors hope will be used by the LSM community to benchmark land surface models in the future.

The article is very well-written and the topic is certainly of interest to the land surface modeling community. The results are robust and will become increasingly pertinent as our land surface models continue to grow in complexity. There are parts of the manuscript that rely too heavily on pointing the reader to other papers, rather than

explaining the necessary finding from the source that is relevant to this manuscript. While this issue is certainly not severe enough to prevent publication, I believe extending the explanation a bit more in several areas of the manuscript would go far towards improving the readability and clarity. I offer specific comments related to this below.

Specific Comments:

Page 1, Line 24: Please define or expand upon 'confirmation holism'.

Page 2, Line 28: It seems like an unsubstantiated opinion to say that the selection of empirical models in Best was a ad-hoc. Please justify, rephrase, or remove.

Page 5, Line 14: Ockham's Razor approach – please define this briefly for unfamiliar readers, or remove entirely if unnecessary.

Page 7, Line 14: "The linear models out-perform the other models in most cases for NEE under the distribution metrics." I don't see this in Figure 2. The linear models have the highest rank (4 and $\sim$4.2) for distribution in NEE, which indicates the worst performance, correct?

Figure 3: Which metrics are being used here (e.g., all, common, extremes, etc.)? I'm guessing all metrics, but I don't think it's explicitly stated. Please note this.

Page 7, Line 28: "at both resolutions" – the use of 'resolutions' in this context is confusing. Is this referring to the clustering? Please clarify.

Page 8, line 8: The 10 day lag of H was chosen, but it seems like the 7 day lag could have been a good choice also. Figure 4 (Lagged RelHumidity) shows that the 10 day lag of H gives the best performance for Qle, but the 7 day of H gives the best performance for Qh and NEE changes little between 7, 10, and 30 day lags. Was the 10 day lag of H chosen (over 7 days) because it shows the best overall performance in any variable (i.e., Qle)?

Page 15, Line 28-29: "this indicates that our newer and more complex benchmarks are

adding substantial performance improvements over the PLUMBER benchmarks." This is a little unclear - the benchmarks themselves are better (have better rankings) or the LSMs perform better as compared to the new benchmarks?

Page 19, Line 1: Please define the Parerto principle.

Page 19, Line 15-29: I appreciate this discussion paragraph because I was asking myself exactly that question – and the answer was clearly articulated.

Technical Corrections:

1) Page 2 Line 34 – might 'be' narrowed down. 2) Page 13, line3 – show to shown 3) Page 13, line 17: gradation to graduation? 4) Page 15, line 33: most complex benchmark, 3km27.

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-153, 2017.