The plumbing of land surface models: why are models performing so poorly?

N. Haughton G. Abramowitz A. J. Pitman D. Or M. J. Best H. R. Johnson G. Balsamo A. Boone M. Cuntz B. Decharme P. A. Dirmeyer J. Dong M. Ek Z. Guo V. Haverd B. J. van den Hurk G. S. Nearing B. Pak C. Peters-Lidard J. A. Santanello Jr. L. Stevens N. Vuichard

May 21, 2015

What is PLUMBER?

- Benchmarking intercomparison of 8 major LSMs (13 versions)
- benchmarked vs old LSMs & empirical models (trained out of sample)
- 20 Fluxnet sites, 4 common metrics (Cor, Bias, NME, SD)
- Emprical models outperform LSMs most of the time for Qh
- Non-linear empirical benchmark outperforms LSMs for Qle too

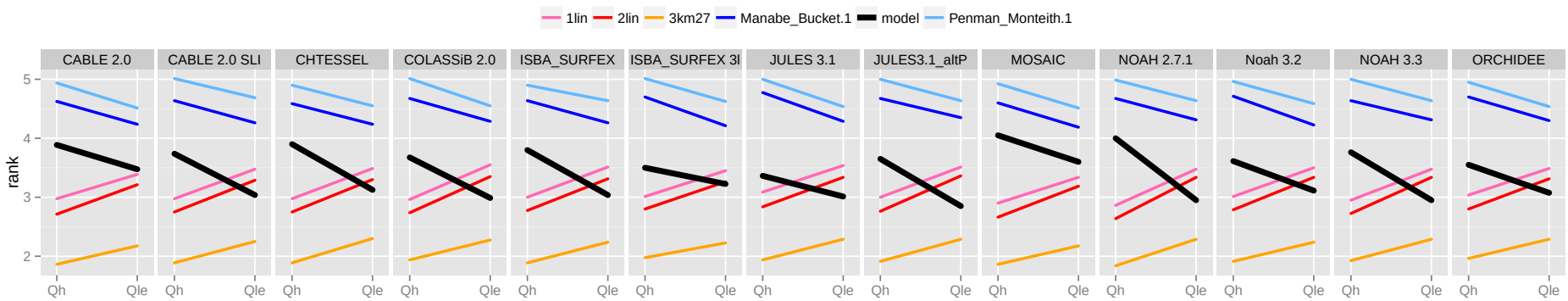


Figure 1: PLUMBER plot: Major columns: different LSMs. Minor columns: sensible heat (Qh) and latent heat (Qle). LSM is black, and various benchmarks are shown in comparison. The vertical axis shows the average performance rank for each model under 4 metrics over the 20 Fluxnet site datasets. In each case, lower is better.

Possible causes

There are three possible categories for the causes of the apparent poor performance seen in PLUMBER:

- The apparent poor performance is due to problems with the PLUMBER methodology;
- The apparent poor performance is due to spurious good performance of the empirical models (e.g. systematic observational error, or empirical models lack of energy conservation constraint); or
- The poor performance is real, and is due to poor representations of physical processes, process order or ability to prescribe appropriate parameter values in LSMs

Possible cause #1: PLUMBER methodology

- Are ranks representative?
- Is aggregation over sites and metrics problematic?
- Do LSMs perform better on longer time scales?
- Are initial conditions a problem?

Are ranks representative of metrics?

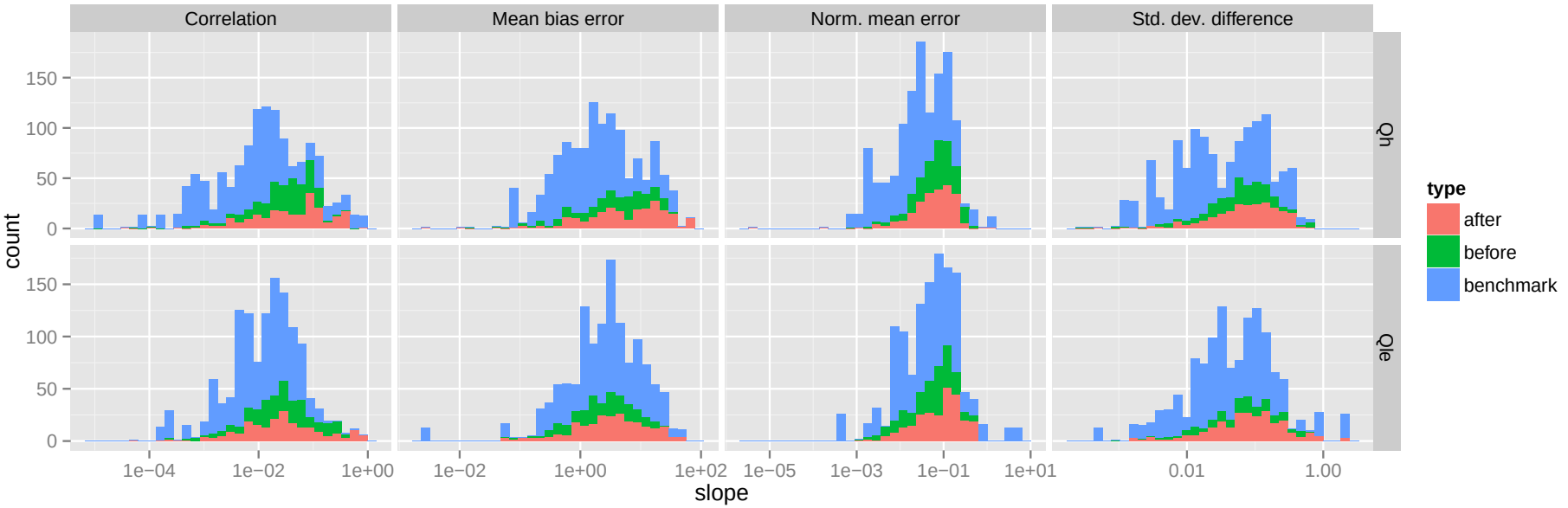


Figure 2: Histograms of differences between metric values for benchmarks and models with neighbouring ranks. Values are calculated by taking the difference of the metric value for each model from the model ranked next-worst in for each LSM, Fluxnet site, metric, and variable. The blue data shows the benchmark-to-benchmark metric differences. The red data show the differences between the LSM and the next worst-ranked benchmark (e.g. if the model is ranked 4, the comparison with the 5th-ranked benchmark). The green data show the difference between the LSM and the next best-ranked benchmark. Because the models are ordered, all differences are positive (correlation is inverted before differences are calculated).

PLUMBER by metric

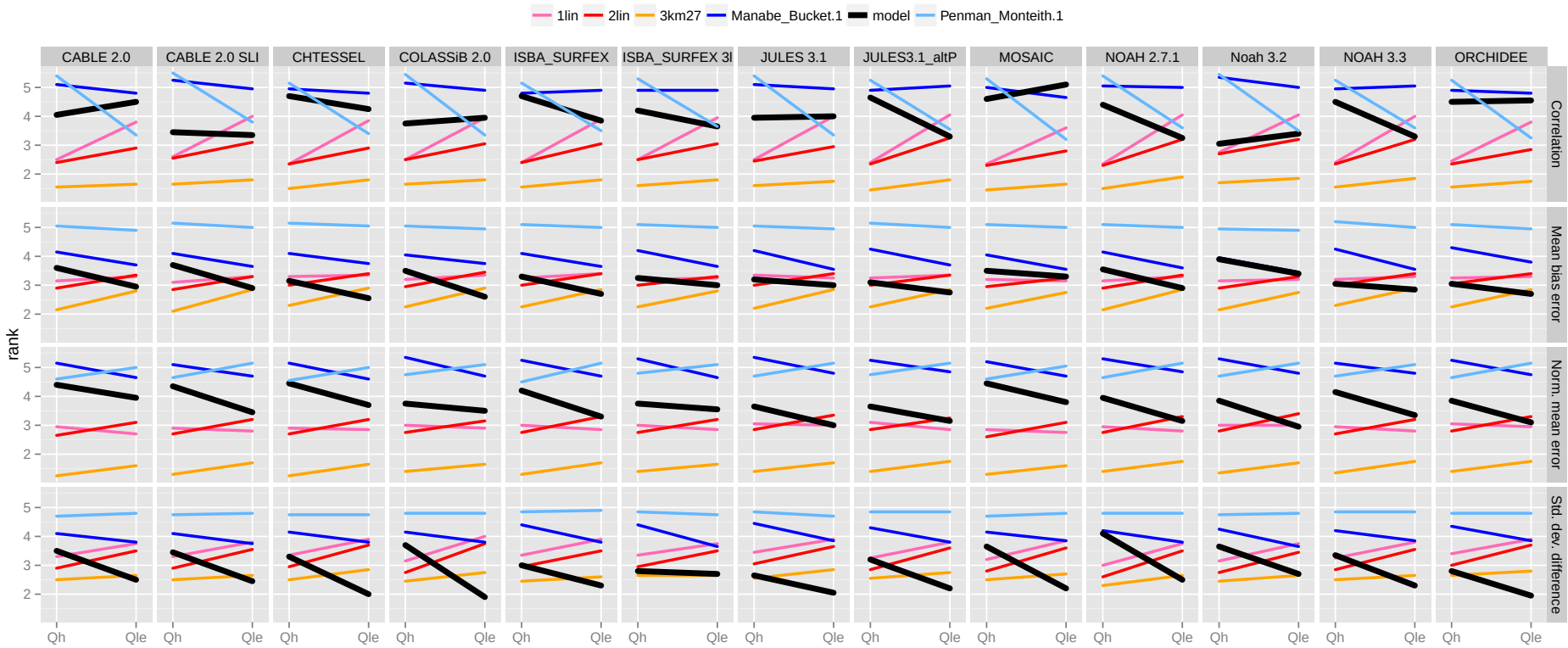


Figure 3: As for Figure 1, but each row represents an individual metric. LSMs perform relatively well under standard deviation difference and mean bias error, and relatively worse under normalised mean error and correlation. There are no clear patterns in the differences between models.

PLUMBER site averages

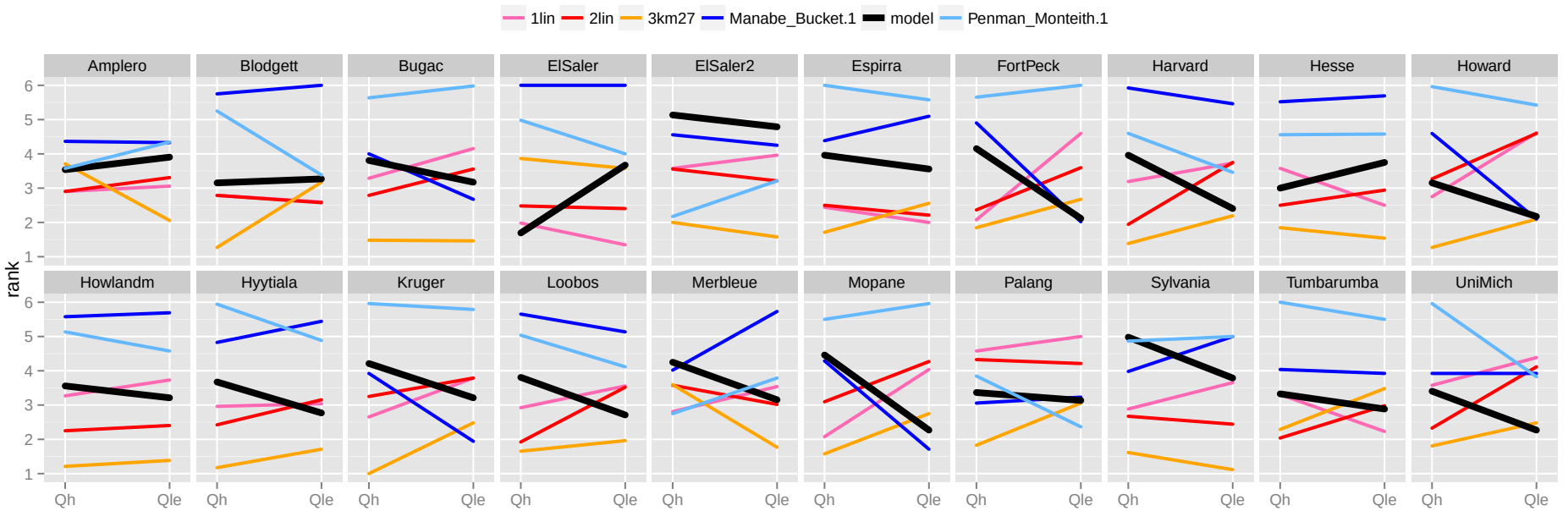


Figure 4: As for Figure 1, but each cell represents the average rank of all LSMs at each individual Fluxnet site. There are no sites where LSMs perform consistently well relative to the benchmarks.

PLUMBER time-averages

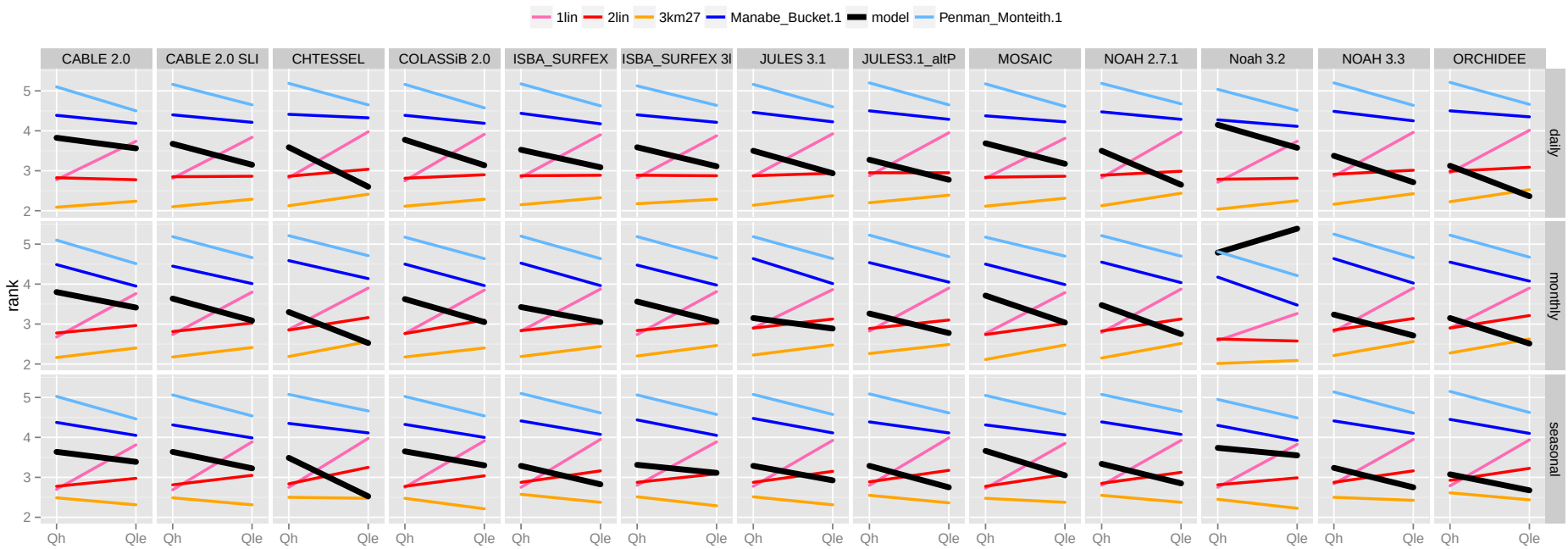


Figure 5: PLUMBER plot, over daily, monthly, and seasonal averages. LSMs actually appear to be performing *worse* under metrics that only take long-term variability into account.

Initial conditions metrics

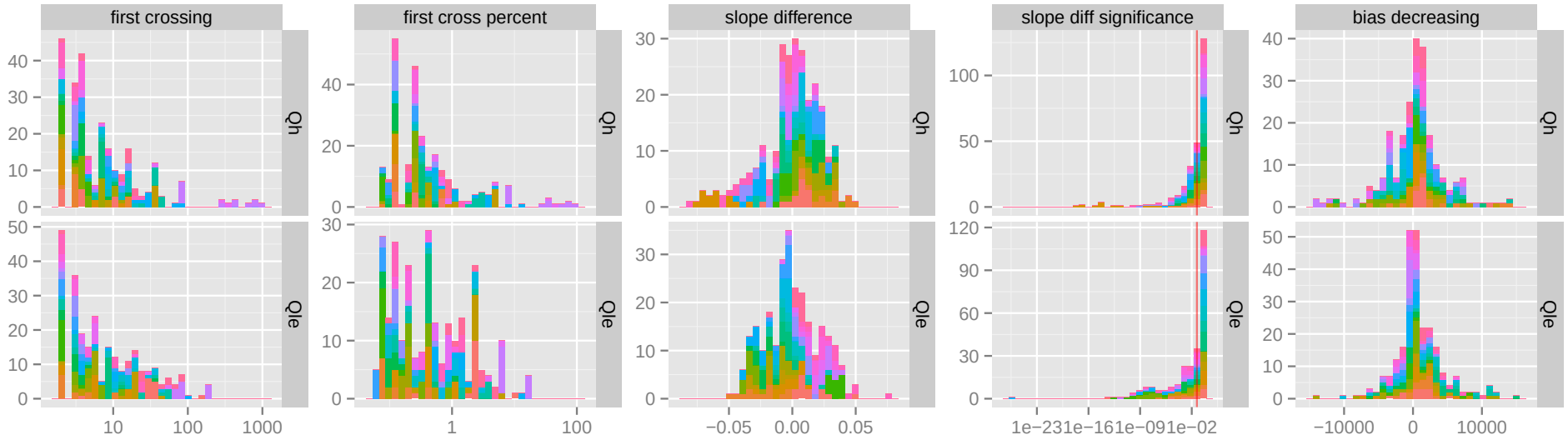


Figure 6: Daily model metrics, from left to right: 1) day at which the simulated series crosses the observed series; 2) as previous, but as a percentage of the time series; 3) difference in the slopes of linear regressions of simulated and observed series over time (W/day); 4) significance of the difference in the previous metric - values left of the red line are significant at the $\alpha = 0.05$ level (~44% of all values); and 5) the rate at which the bias is decreasing, measured by mean(error)/slope(error) - negative values indicate the simulations have a trend toward the observations. Colours indicate the Fluxnet site at which the simulation is run.

Possible cause #2: Spurious empirical model performance

- It's possible the empirical models are out-performing the LSMs due to lack of energy conservation constraint (`Rnet != Qh + Qle`) and/or fitting Fluxnet conservation errors
- We tested this by scaling the available energy of the empirical models to match that of each LSM, and then calulating metrics and ranks
- Emprical models still outperform LSMs for evaporative fraction most of the time

PLUMBER plot - "energy conserving" empirical models

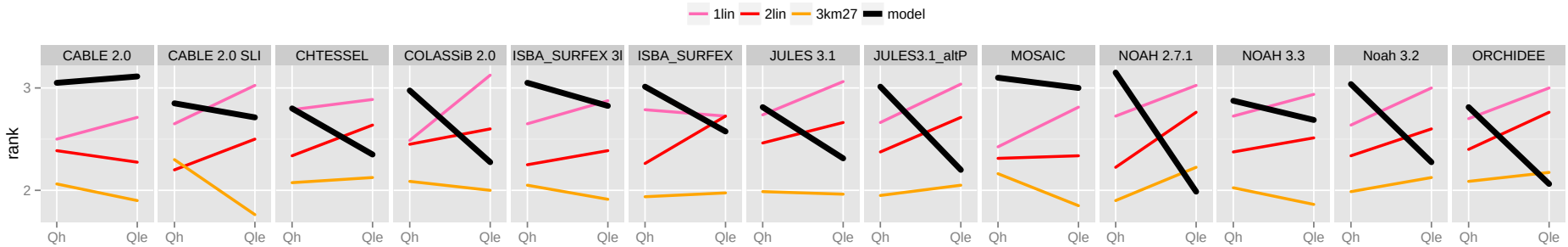
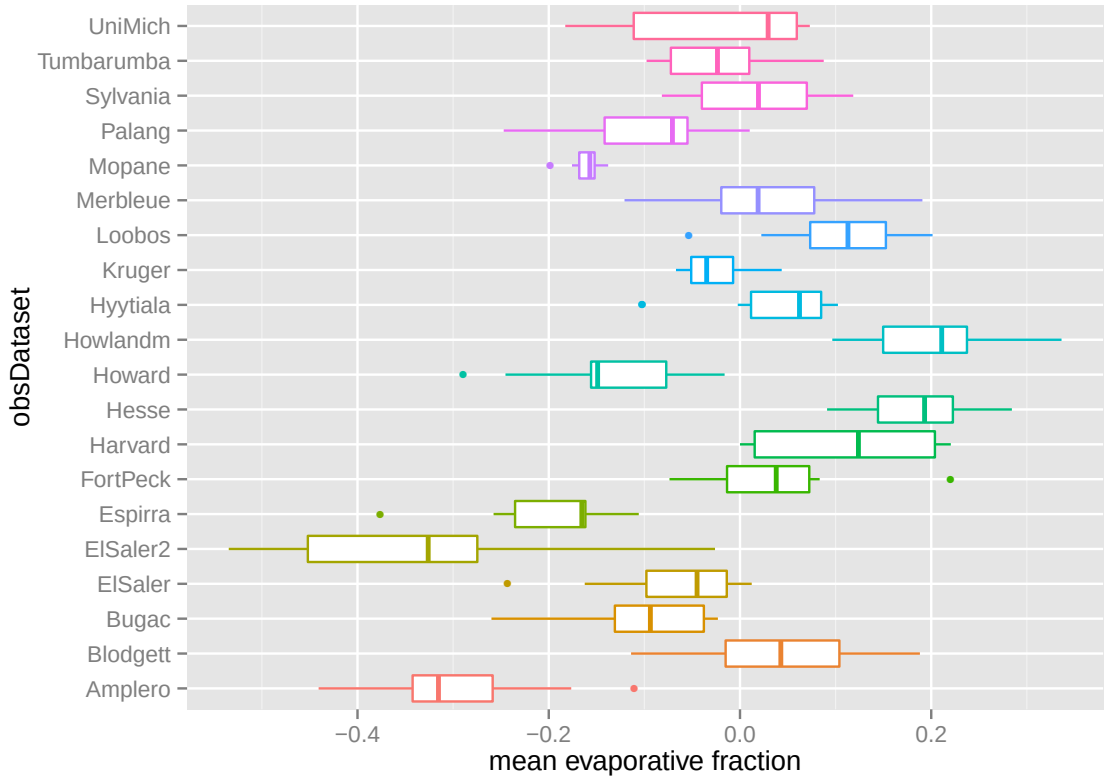


Figure 7: PLUMBER plot, with energy conservation constrained empirical models. Even with the added constraint, the empirical models still largely perform much better than the LSMs for both sensible and latent heat.

- Indicates that the models are not only performing poorly for Qh and Qle, but also the evaporative fraction
- If the problem were due to Fluxnet errors, that would require that there is a consistant bias in the evaporative fraction over all sites (due to the training method for empirical models).

LSM/Fluxnet evaporative fraction biases



Mean biases in daily evaporative fraction for each simulation, grouped by site. While there are sites with strong biases, there is no consistency in biases across sites.

- indicates that the problem does not lie with fluxnet biases: empirical models are trained out-of-sample, and biases are not consistent.

Possible cause #3: Poor model performance

- How do LSMs perform over short time scales?
- Do LSMs perform better at different times of the day?
- How do the LSMs perform as an ensemble?

PLUMBER: High-frequency only

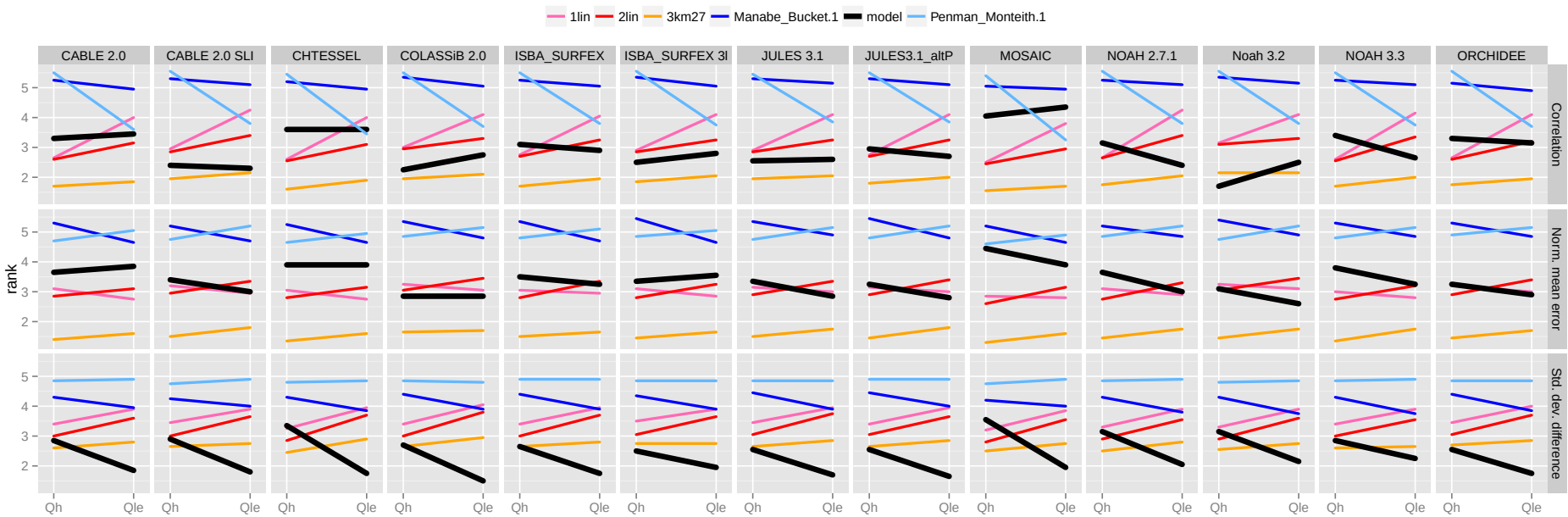


Figure 8: PLUMBER plot, high-frequency response only, by metric - for this plot, LSMs are bias-corrected on a daily basis, and then have the daily cycle in the errors removed. The mean bias error metric is not included because it is trivially 0 due to the bias correction process. Std. dev. difference is also trivially affected by the bias correction.

PLUMBER by time of day

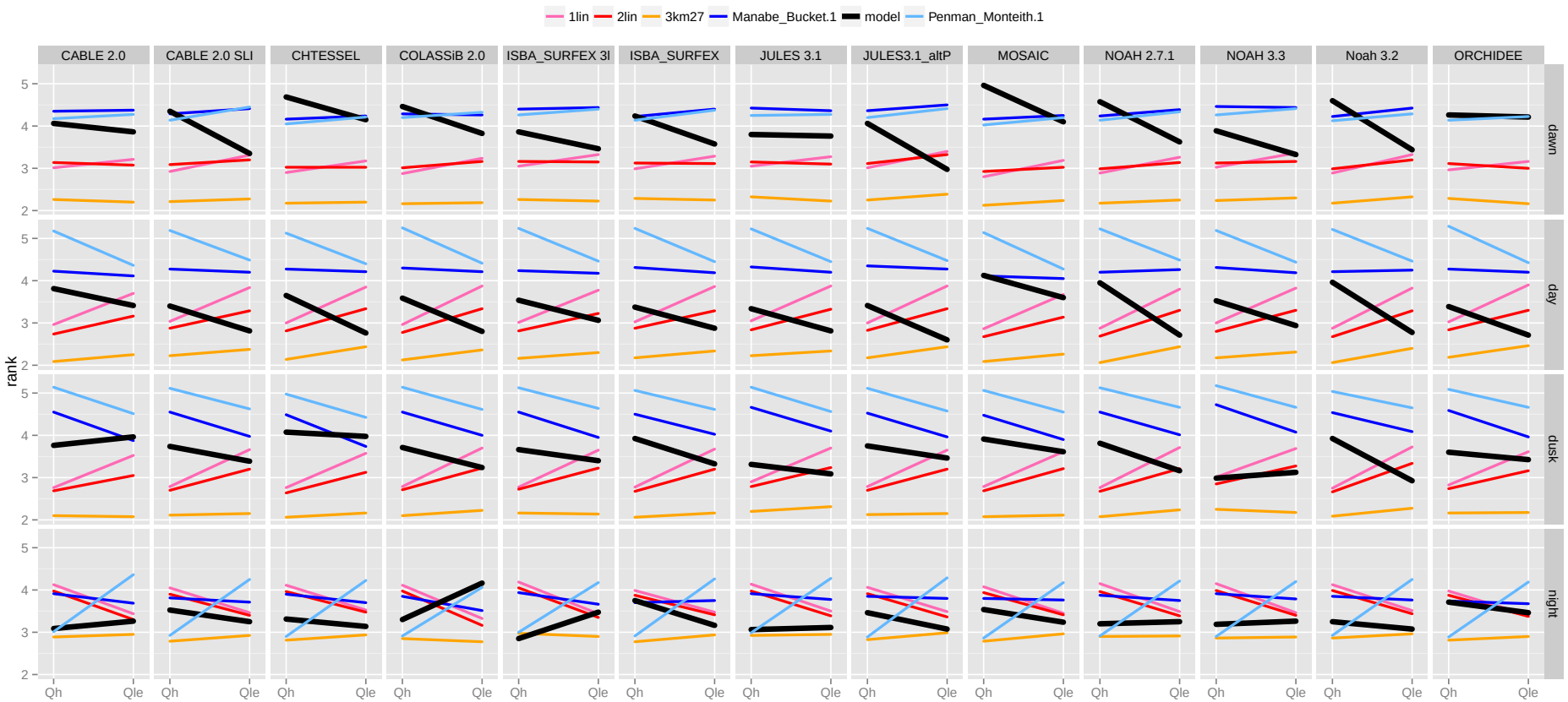
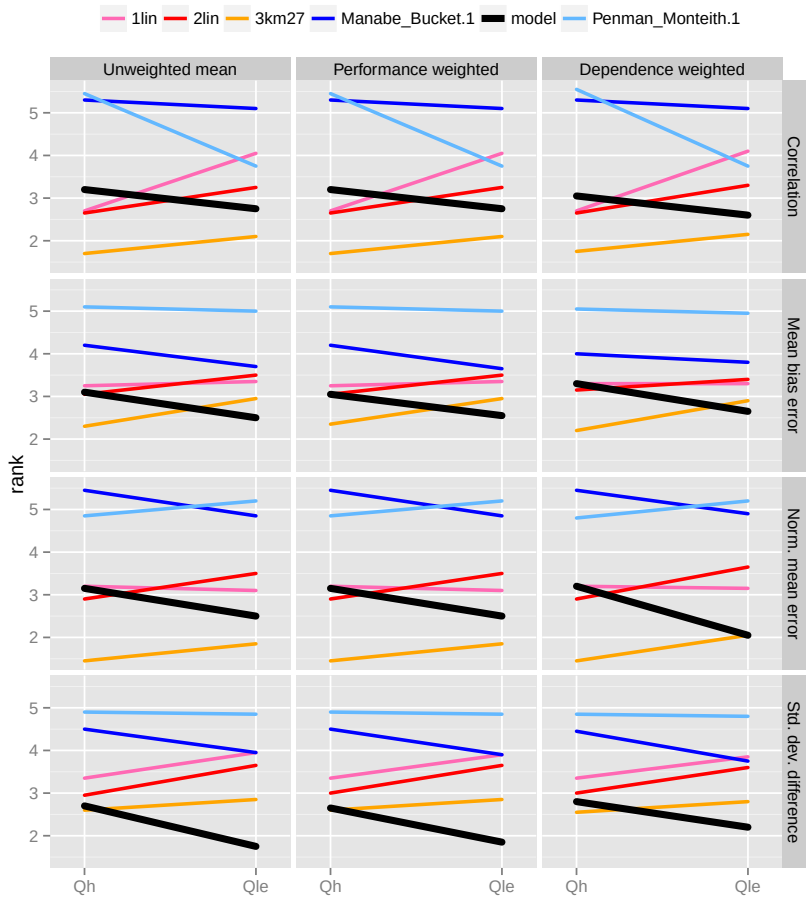


Figure 9: PLUMBER plot, split by daily cycle - the 4 rows represent the 6-hour periods around dawn (3am-9am), noon (9am-3pm), dusk (3pm-9pm), and midnight (9pm-3am).

PLUMBER ensemble means



PLUMBER plot, showing the results for three different means across all LSMs, by metric. In general, we should expect means to perform better under all metrics except the standard deviation metric, as the averaging process acts as a smoother, removing non-correlated noise from the model results.

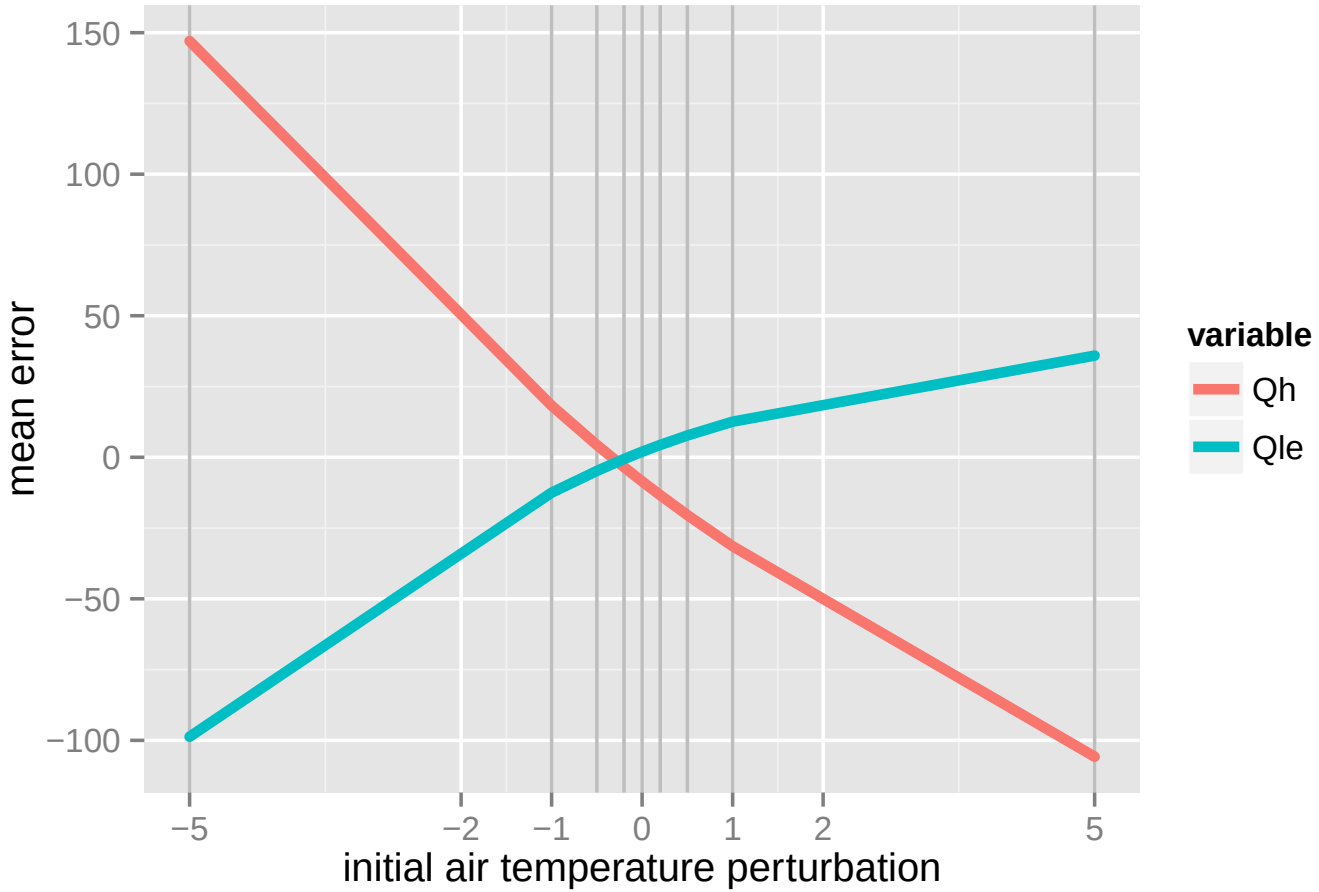
Summary

- Not an obvious methodological problem
- Not obviously due to lack of conservation in benchmarks, or biases in Fluxnet
- Problem appears to be shared across LSMs

Discussion

- Maybe the models are wrong?
  - We're not sure of specific causes
  - One example might be that some derived variables affect the two fluxes differently, biasing the energy partitioning.

CABLE stabilisation feedback



Mean error in Qh and Qle due to perturbation of initial canopy air temperature at each time step, (CABLE at Tumbarumba). The response in Qh to negative temperature perturbations is about 50% stronger than in Qle, and about 3 times stronger for positive perturbations.