Why the climate modelling community needs flux tower data

Gab Abramowitz

ARC Centre of Excellence for Climate System Science
University of New South Wales

What I want to do…

1. Evaluate the ability of land surface models* to accurately simulate atmospheric fluxes at a range of spatial scales

2. Do this in a way that allows us to tell what might cause any problems

… I'll try to argue that flux tower data is *the* key to doing this.

* i.e. coupled to a climate model => must work globally (without local calibration); half hourly time step; no local data beyond vegetation type and height; sensible, latent NEE fluxes

Confirmation holism

Lenhard & Winsberg (2010): Holism, entrenchment, and the future of climate model pluralism *Studies in History and Philosophy of Modern Physics*

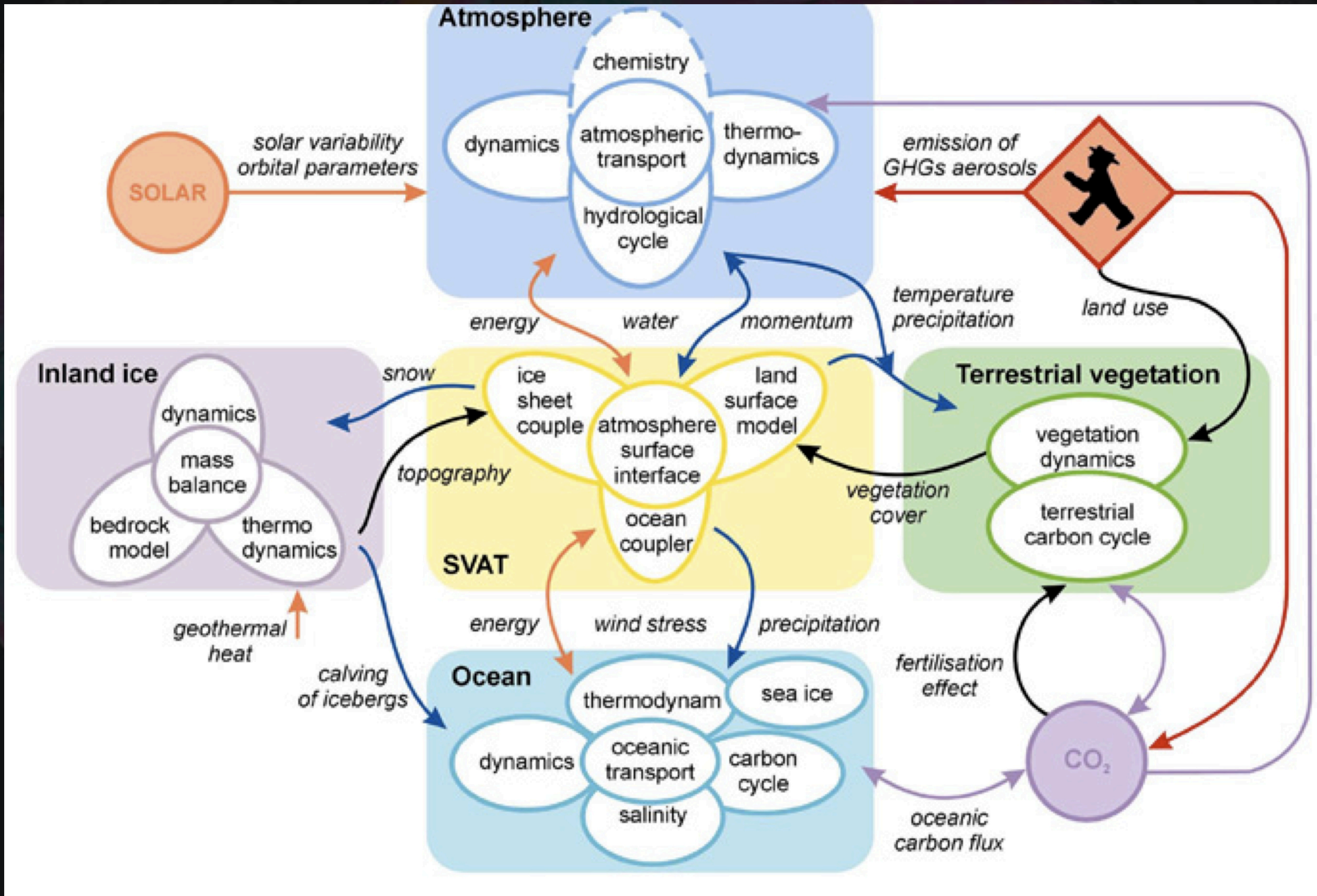
"Confirmation holism… is the thesis that a single hypothesis cannot be tested in isolation, but… depend(s) on other theories or hypotheses. It is always this collection of theories and hypotheses as a whole… that confront the tribunal of experience."

- ''If the predicted phenomenon is not produced, not only is the questioned proposition put into doubt, but also the whole theoretical scaffolding used by the physicist'' (Duhem 1954).
  - The experiment tells that there is something wrong, but does not tell where the error comes from, hence the doubt is necessarily holistic.

- "problems of understanding what features of our models are responsible for their best and worst qualities"
- "it is impossible to tell [categorically], by any method, where to locate the sources of the failures of our models to match known data."

Confirmation holism

Lenhard & Winsberg (2010): Holism, entrenchment, and the future of climate model pluralism *Studies in History and Philosophy of Modern Physics*



Confirmation holism causes: fuzzy modularity

"That means, one parametrization is tested on the basis of the parameterizations that are already part of the concrete model under construction. That contributes to what we call path-dependency: the next modeling step is influenced by the accumulated effects of the previously implemented steps. And it creates a 'fuzzy' kind of modularity: normally, modules are thought to stand on their own. In this way, modularity should have the virtue of reducing complexity. In our present case, however, the modules (parameterizations) are interdependent and therefore lack this virtue."

Lenhard & Winsberg (2010): Holism, entrenchment, and the future of climate model pluralism *Studies in History and Philosophy of Modern Physics*

Confirmation holism causes: kludging

kludge | klʌdʒ, kluːdʒ | (also **cludge**) informal

noun

an ill-assorted collection of parts assembled to fulfil a particular purpose.

- Computing a machine, system, or program that has been badly put together, especially a clumsy but temporarily effective solution to a particular fault or problem.

verb [with obj.]

improvise or put together from an ill-assorted collection of parts.

Hugh had to kludge something together.

ORIGIN 1960s: invented word, perhaps influenced by **BODGE** and **FUDGE**.

Confirmation holism causes: kludging

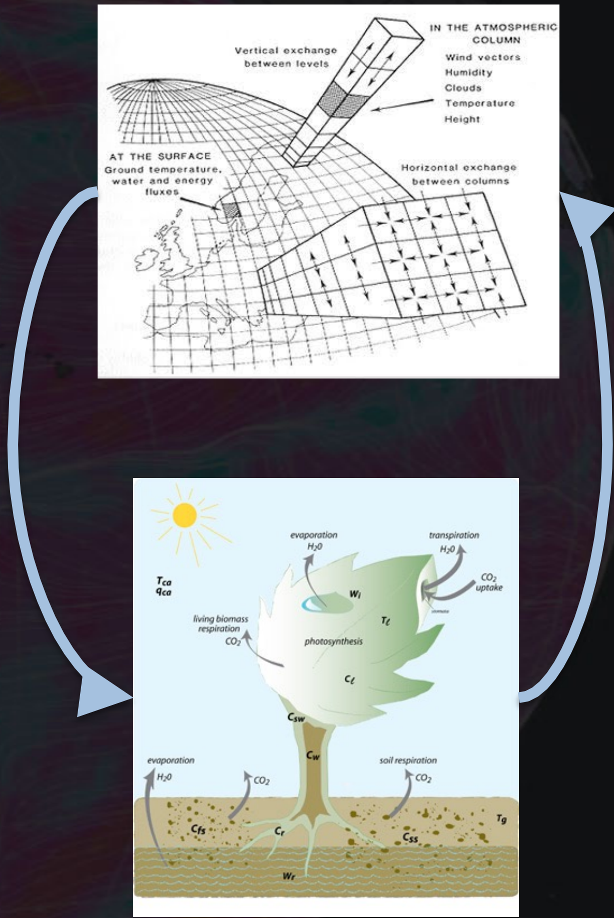
"A kludge is built to optimize the performance of the overall model as it exists at that particular time, and with respect to the particular measures of performance that are in use right then. There is no guarantee that an implemented kludge is optimal in any general sense."

"Kludging plays a central role in the construction of complex models. When modifications are made to a complex model, and are shown to improve model performance, there is often a mixture of principled and unprincipled steps involved in the modification. And so when some new elements are added to a model, and improve model performance, it is often impossible to know if this happens because what has been added has goodness-of-fit on its own, or merely because, in combination with the rest of the model, what is achieved on balance is an improvement. "

Lenhard & Winsberg (2010): Holism, entrenchment, and the future of climate model pluralism *Studies in History and Philosophy of Modern Physics*

Evaluating LSMs in gridded coupled simulations

- Confirmation holism most evident
- Results are a function of the quirks of individual component models AND the way in which these components interact
  - E.g. some LSM behaviour evident only when coupled to a certain boundary layer model
- Predicted variable values are an emergent property of the feedbacks and sensitivities of the entire modelling system, and may not reflect the LSM being evaluated

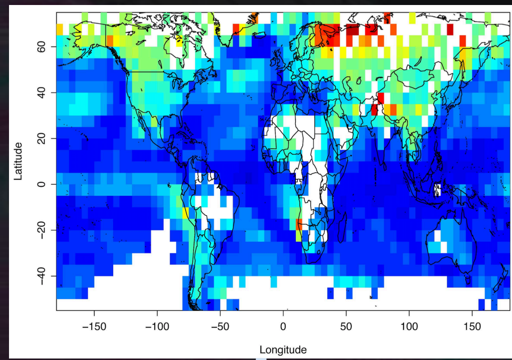


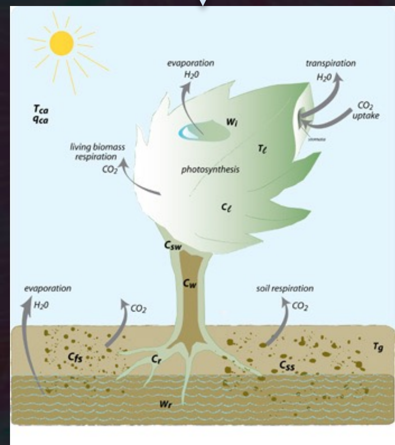
Can we categorically: *Evaluate the ability of land surface models to accurately simulate atmospheric fluxes at a range of spatial scales*

No.

Evaluating LSMs in gridded offline simulations

- Prescribe meteorological forcing data (SWdown, LW down, Tair, Hum, Wind, Pr)
  - Reanalysis or a collection of observationally-based interpolated products
  - Typically daily - require a weather generator for hourly / half-hourly fluxes
- Chaotic / emergent phenomena are no longer an issue: no feedbacks
- Is poor LSM performance due to forcing or LSM?
  - E.g. average precip may be < average ET for some products



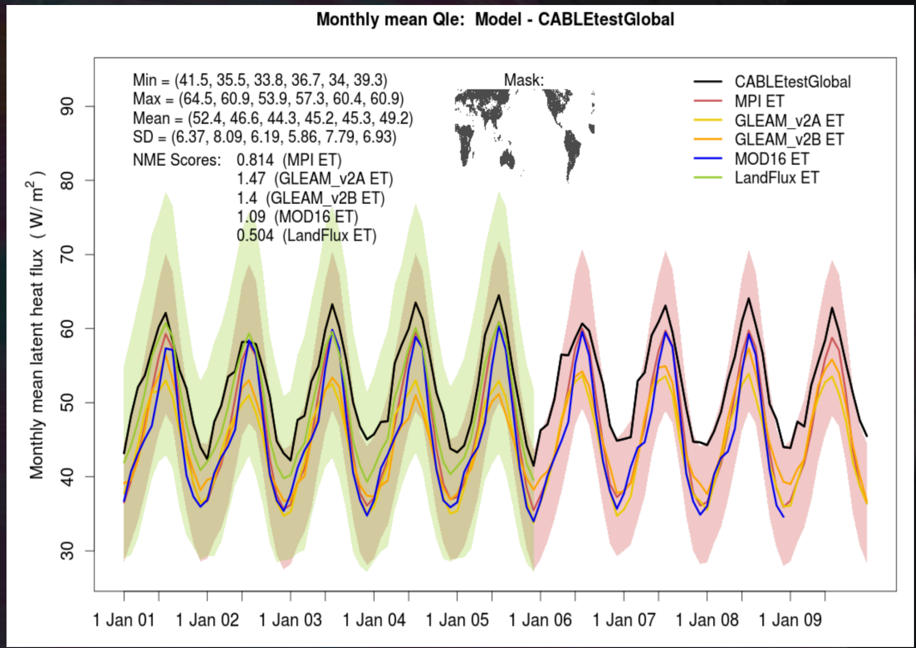


Can we categorically: *Evaluate the ability of land surface models to accurately simulate atmospheric fluxes at a range of spatial scales*

No.

Offline or coupled: gridded evaluation data

- Typically based on remote-sensing
  - Frequency of overpass – representative?
  - Viewing angle
  - Missing data from cloudiness
  - Depth of measurement (e.g. soil moisture)
  - Monthly time step
  - Competing products show big differences
- Compensating LSM processes



Can we categorically: *Evaluate the ability of land surface models to accurately simulate atmospheric fluxes at a range of spatial scales*

No.

Offline or coupled: parameter identifiability

- Very few of the 30-50 parameters that LSMs require are identifiable at global or regional scales
  - Aggregating assumptions about process dependence – veg & soil ‘types’
  - Reduces spatially varying surface information to 2-3 dimensions

- What is the cost of this lack of heterogeneity?

Can we categorically: *Evaluate the ability of land surface models to accurately simulate atmospheric fluxes at a range of spatial scales*

No.

Confirmation holism in gridded simulations

- Parameter identifiability
- Temporal averaging is necessary for evaluation
- Compensating processes within the LSM (offline and coupled)
- Discordant relationships between LSM and host model (coupled)

We can note that gridded evaluation products and LSMs disagree, but…

We cannot say why - inability to attribute errors

(Trial and error changes to parametrisation does not resolve this problem)

Evaluation at flux tower sites

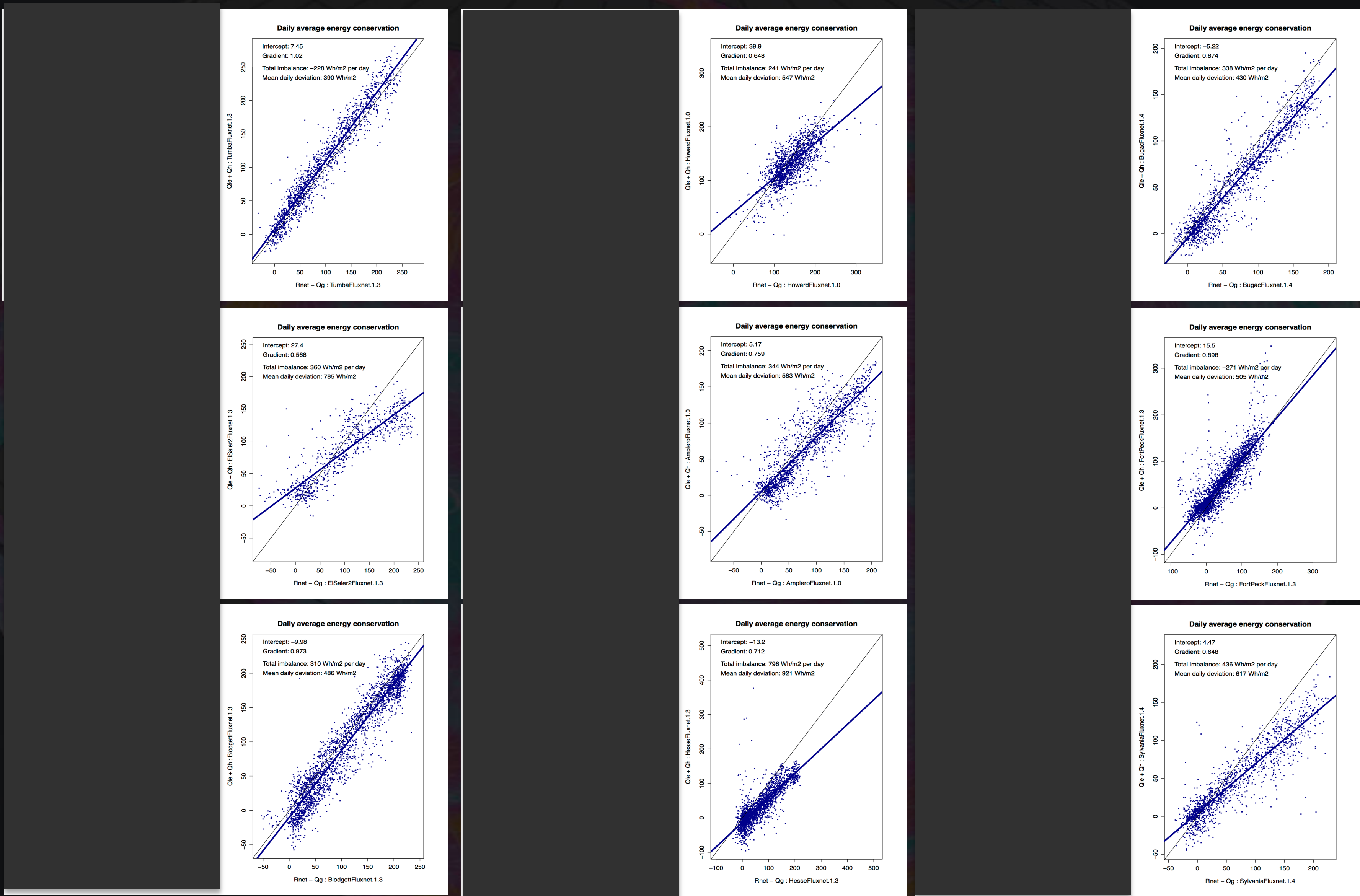
- Measurement of all meteorological forcing variables
- Measurement of atmospheric fluxes
- Very high temporal resolution, giving representative values when averaged up to time step size of a LSM (e.g. hourly)
- A significant proportion of LSM parameters are directly measurable
- Coverage across biomes internationally, 100s sites
- Diversity in individuals taking measurements and processing – independent samples
- Appropriately sited flux tower has fetch of order 1km^2 – LSM length scale?

Can we categorically: *Evaluate the ability of land surface models to accurately simulate atmospheric fluxes at a range of spatial scales*

Maybe yes...

Evaluation at flux tower sites: conservation

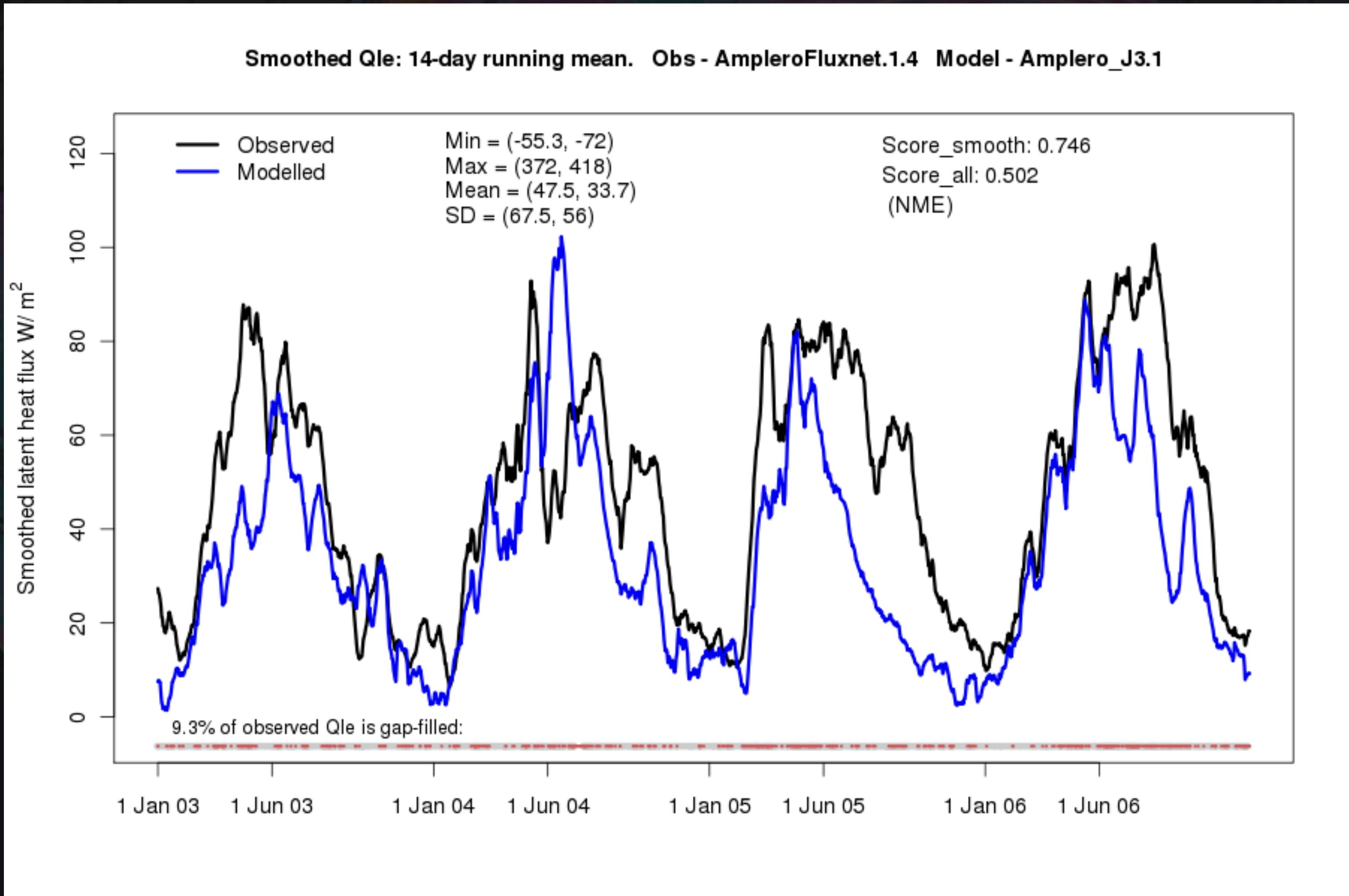
Sensible + latent heat flux



Rnet - ground heat flux

Evaluation at flux tower sites: an example

"This is a great simulation of latent heat flux"



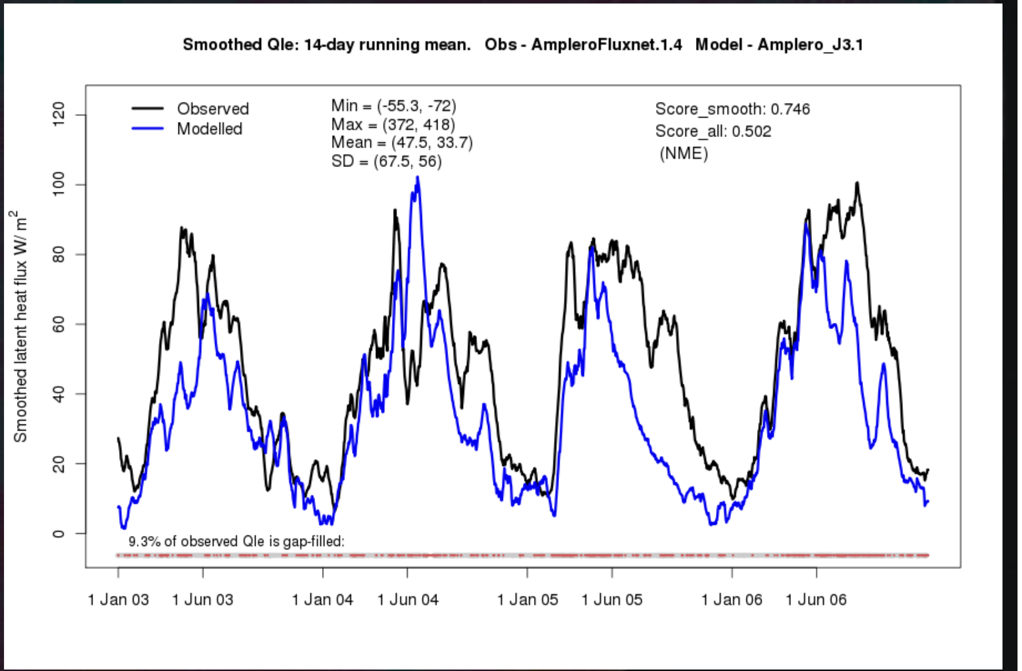
Evaluation at flux tower sites: an example

How well should we expect a LSM to predict latent heat (LH) flux at the Amplero site?

1. Take several (19) flux tower sites other than Amplero
2. Train a linear regression between shortwave radiation and LH flux
3. Use these regression parameters to predict LH at Amplero using site met

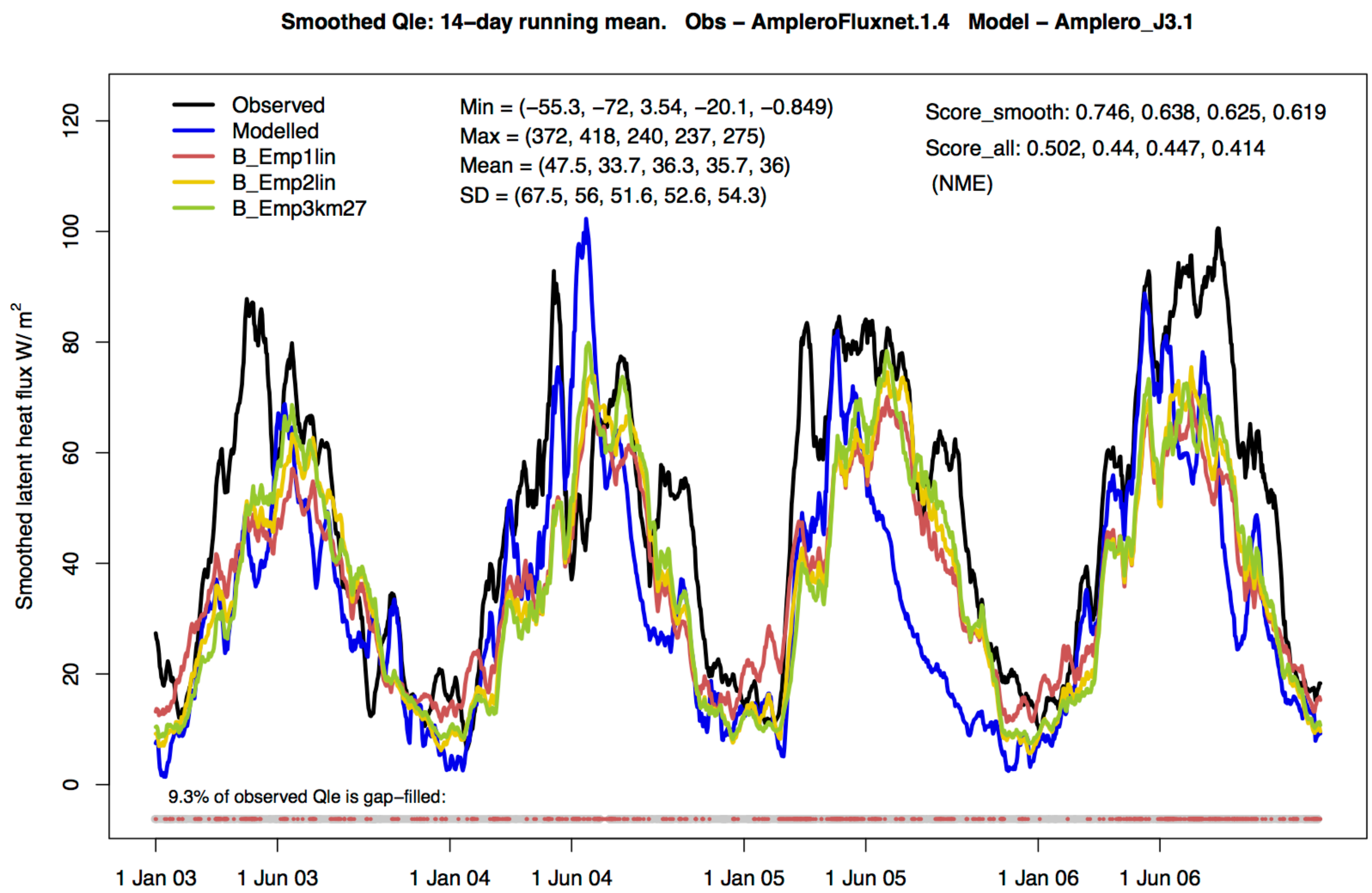
This will tell us:

- The extent to which LH is predictable from SWdown - just 1 model input variable
- How a very simple functional relationship would represent LH in our usual diagnostics
- How predictable LH at Amplero is, out-of-sample – is this site unusually difficult?



Abramowitz, 2012, GMD

Evaluation at flux tower sites: an example



Abramowitz, 2012, GMD

Evaluation at flux tower sites: an example

- We can use empirical models can quantify the amount of information in the met data about fluxes (at the same time step size as the LSM)

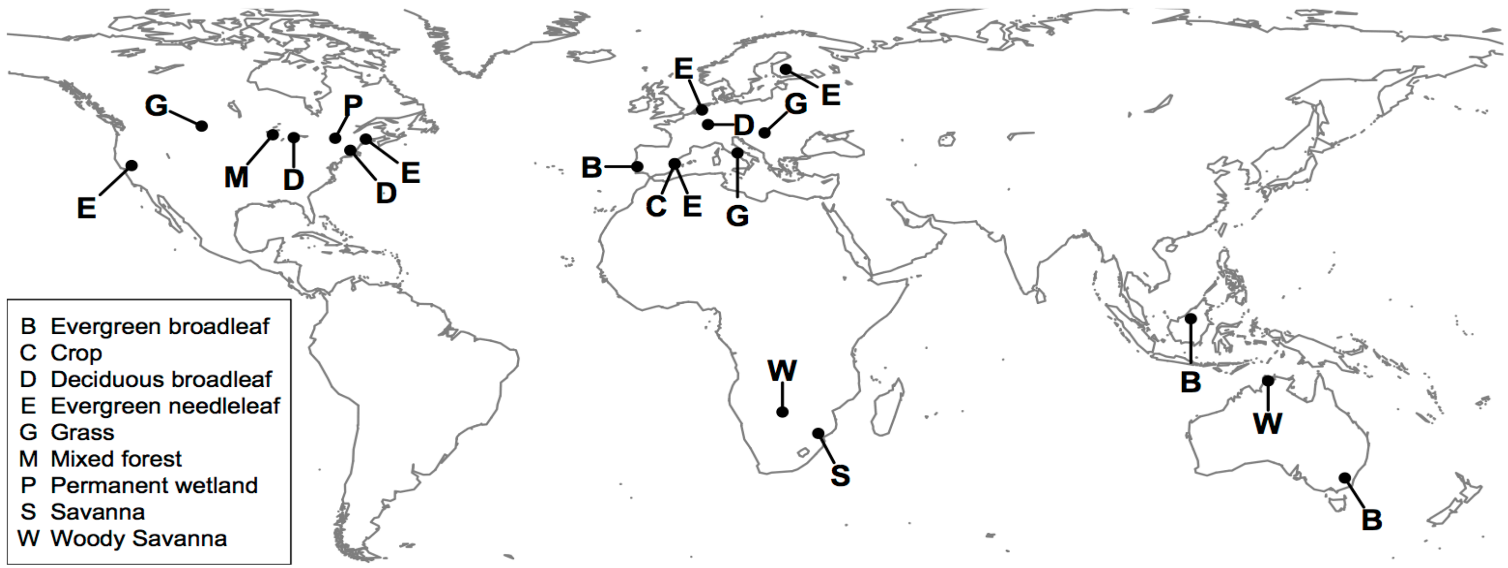
- It gives us a way to quantify how well we should *expect* a LSM to perform

- It provides a LSM-like time series, and so provides benchmark performance levels in any chosen metric

- To make the benchmark appropriate, we can control:
  - The amount of information given to empirical model (i.e. how many / which model inputs)
  - The complexity of the empirical model (linear regression, ANNs, cluster+regression, etc)
  - The relationship between the training and testing sets (extent of out-of-sample test)

Expanded example: The PALS Land sUrface Model Benchmarking Evaluation pRoject (PLUMBER)

- 20 Flux tower sites; latent and sensible heat
- 4 metrics: bias, correlation, SD, normalised mean error
- 9 LSMs, 15 LSM versions
- LSMs given: veg type, veg height, tower height (as in gridded application)
- Benchmarks: two 'physical' – PM and Manabe bucket; 3 empirical



Best et al, 2015, J Hydromet.

The three empirical benchmarks in PLUMBER

- All 3 empirical models relate met forcing and a flux and are trained with data from sites other than the testing site (i.e. out of sample)

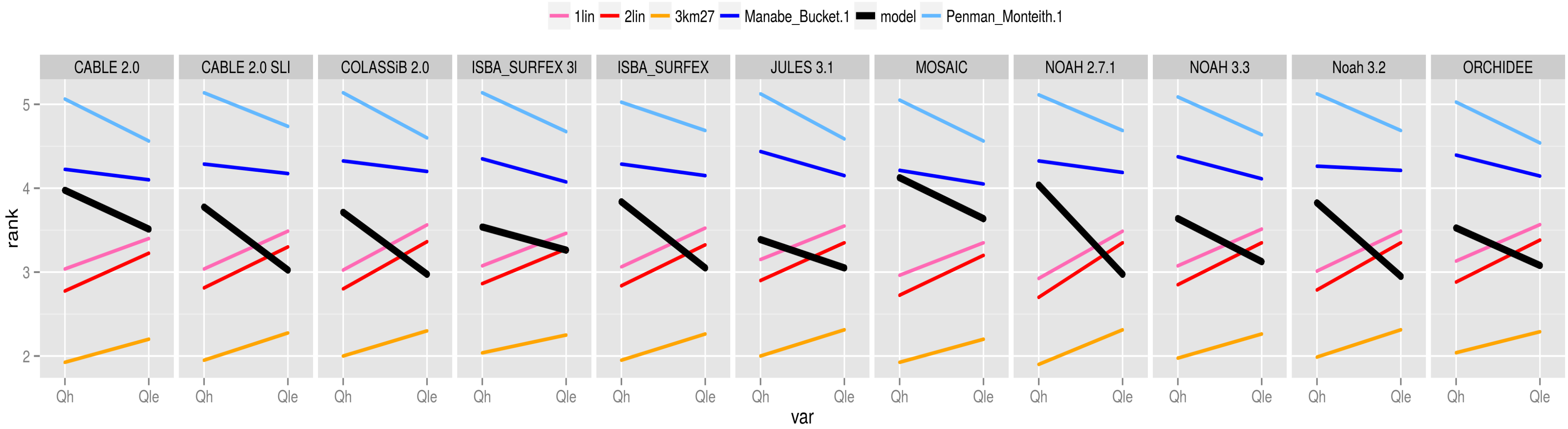
- They are each created for LE, H:
  - "1lin": linear regression of flux against downward shortwave (SW)
  - "2lin": as above but against SW and surface air temperature (T)
  - "3km27": non-linear regression – 27-node k-means clustering + linear regression against SW, T and relative humidity at each node

- Gap-filled data are NOT included in training

- All are instantaneous responses to met variables with no knowledge of vegetation type, soil type, soil moisture or temperature, C pools.

PLUMBER results

Best et al, 2015, J Hydromet.



Vertical axis is the rank of each LSM (black) against the 5 benchmarks, averaged over:

- 20 Flux tower sites (IGBP vegetation types)
- 4 metrics (bias, correlation, SD, NME)

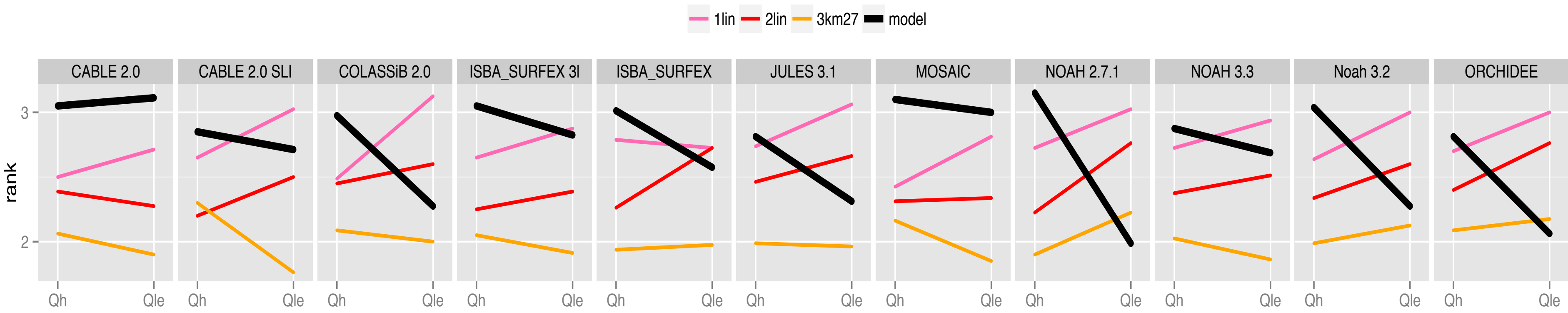
- On average, LSMs outperform Penman-Monteith and Manabe bucket implementations
- On average, LSMs sensible heat prediction is worse than an out-of-sample linear regression against downward SW radiation
- For all fluxes, models are comfortably beaten by out-of-sample regression against Swdown, Tair and RelHum

PLUMBER results – methodology?

- Lack of flux tower energy conservation advantaging empirical models?
- Time scale – daily, monthly, seasonal rather than per time step performance?
- Time of day – diurnal biases in flux tower favouring empirical models?
- Poor LSM initialisation?
- Are ranks not representative of metric values?
- Biased by metric choice?
- Biased by site choice?

Haughton et al, J Hydromet, in review

PLUMBER results – why? Not energy conservation.



- Constrain each empirical model to have the same sum of (latent + sensible) heat flux as the LSM at every time step
  - Each empirical model then effectively has the same Rnet and ground heat flux as the LSM it's being compared to – and conserves energy.
- Results are mixed but the regression against SWdown, Tair and RelHum still comes out on top, especially for sensible heat flux.

Haughton et al, J Hydromet, in review

Difficult questions for land surface modellers

- Is over-parameterisation is hurting (calibration of unconstrained parameters inhibits predictive capacity)? Should we have 3-4 parameter global LSMs (i.e. the dimension of surface data available at that scale)?
- Has the drive to add more processes into LSMs (often based on sparse data sets) led to intractable modelling systems with relatively poor accuracy?
- What does to mean to say we have a “physically based” model of a natural system if we don’t have enough data to build an empirically based model?
  – Conceptually consistent does not imply physically consistent
  – Working well at a few sites does not necessarily imply “physical”, especially if calibrated

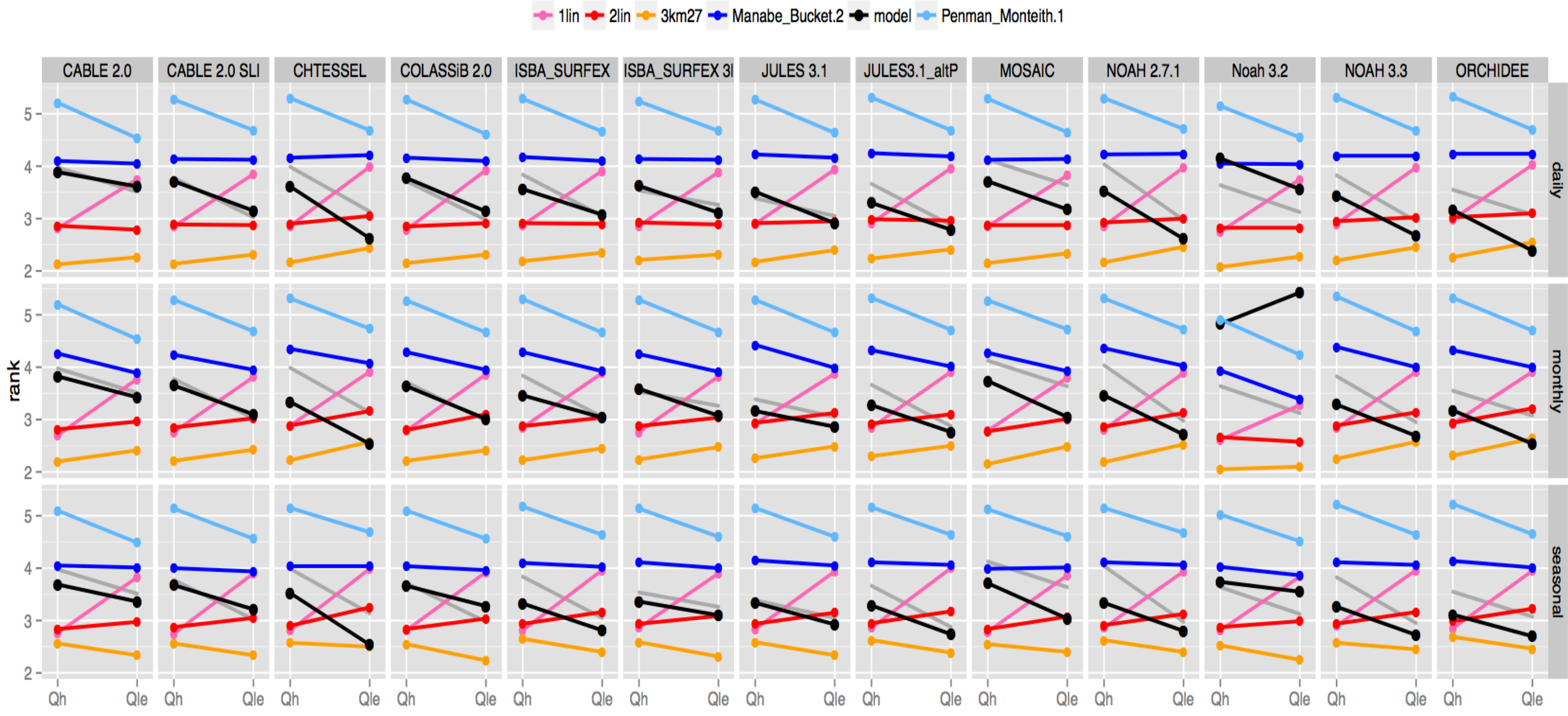
Why does this matter?

- This shows that the information in meteorology about atmospheric fluxes is consistent across flux tower sites
  - Despite many site PIs, measurement and data processing approaches
- Energy conservation at flux tower sites is not insurmountable
- Measurements of radiation, meteorology, soil moisture, soil temperature, radiation, carbon pools, vegetation properties, soil properties, atmospheric fluxes:
  - Flux tower data offers the possibility of avoiding "fuzzy modularity" and confirmation holism… the only example we have?
- The LSM community is extremely lucky to have flux tower data
  - We would not be able to know LSMs were performing poorly otherwise

References

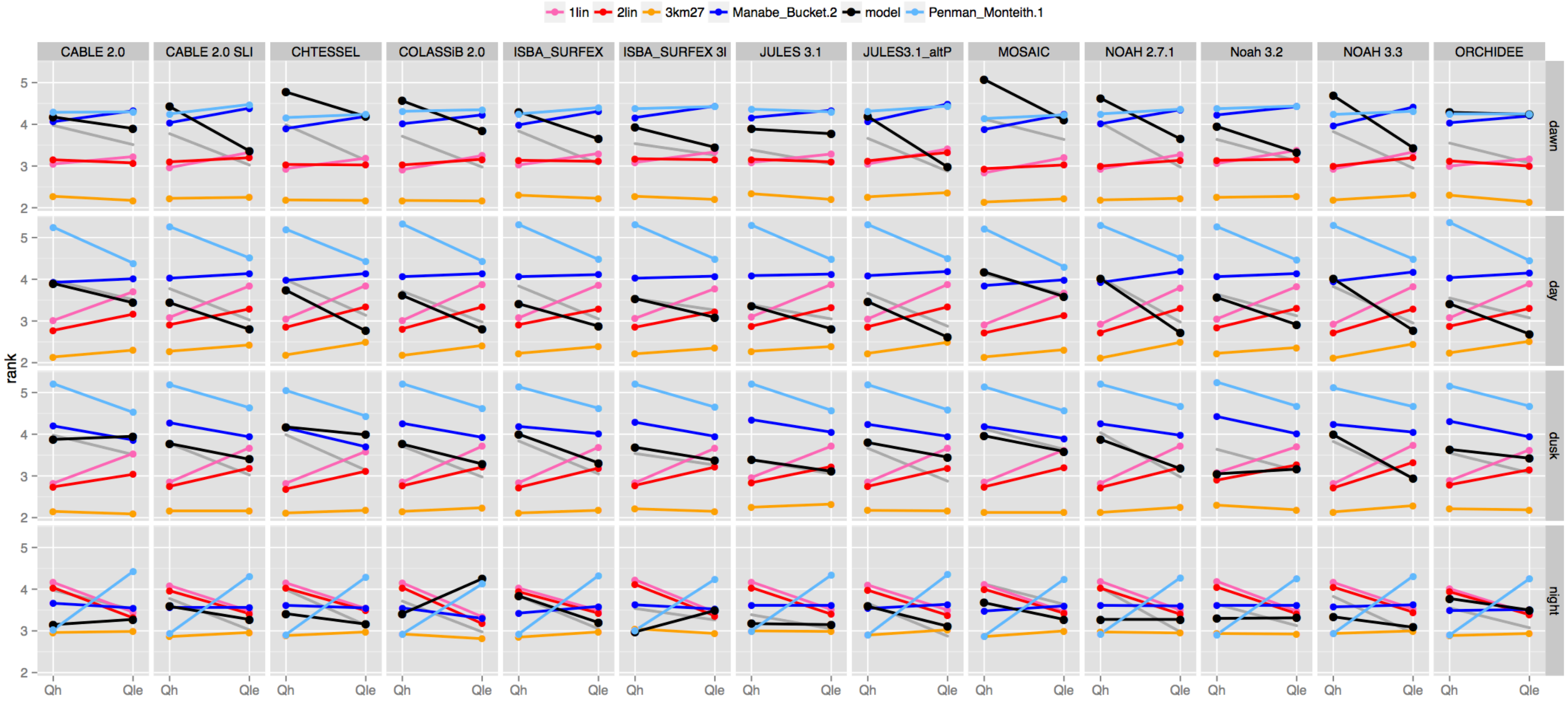
- Lenhard, J. and E. Winsberg (2010), Holism, entrenchment, and the future of climate model pluralism, *Studies in History and Philosophy of Modern Physics*, 41, 253–262.
- Abramowitz, G. (2012) Towards a public, standardized, diagnostic benchmarking system for land surface models, *Geoscientific Model Development*, 5, 819-827
- Best, M.J., G. Abramowitz, H.R. Johnson, A.J. Pitman, G. Balsamo, A. Boone, M.Cuntz, B. Decharme, P.A. Dirmeyer, J. Dong, M. Ek, Z. Guo, V. Haverd, B.J.J van den Hurk, G.S. Nearing, B. Pak, C. Peters-Lidard, J.A. Santanello Jr, L. Stevens, N. Vuichard (2015) The plumbing of land surface models: benchmarking model performance, *Journal of Hydrometeorology*, 16, 1425-1442.
- Haughton, N., 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.J. van den Hurk, G. S. Nearing, B. Pak, J.A. Santanello Jr., L.E. Stevens, N Vuichard The plumbing of land surface models: why are models performing so poorly? *Journal of Hydrometeorology*, in review.

PLUMBER results – timescale?



Haughton et al, J Hydromet, in review

PLUMBER results – time of day?



Haughton et al, J Hydromet, in review

PLUMBER results – initialisation?

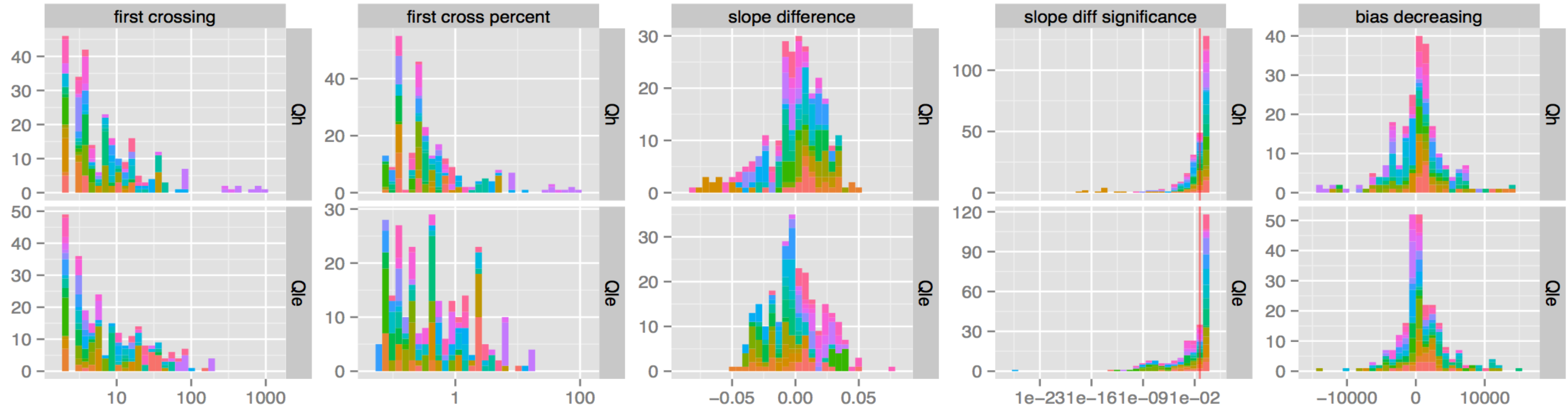
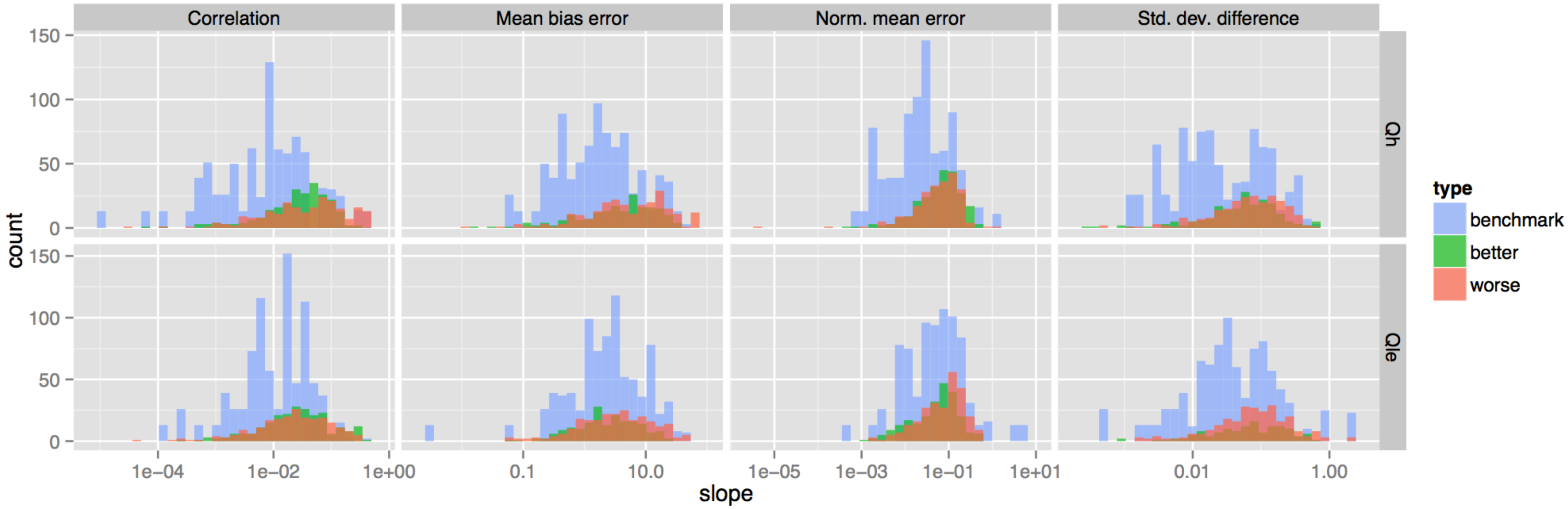


FIG. 7. Model spin-up metrics, based on daily averages, from all LSMs at all sites. From left to right: 1) day at which the simulated series crosses the observed series; 2) as previous, but as a percentage of the time series; 3) difference in the slopes of linear regressions of simulated and observed 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.

Haughton et al, J Hydromet, in review

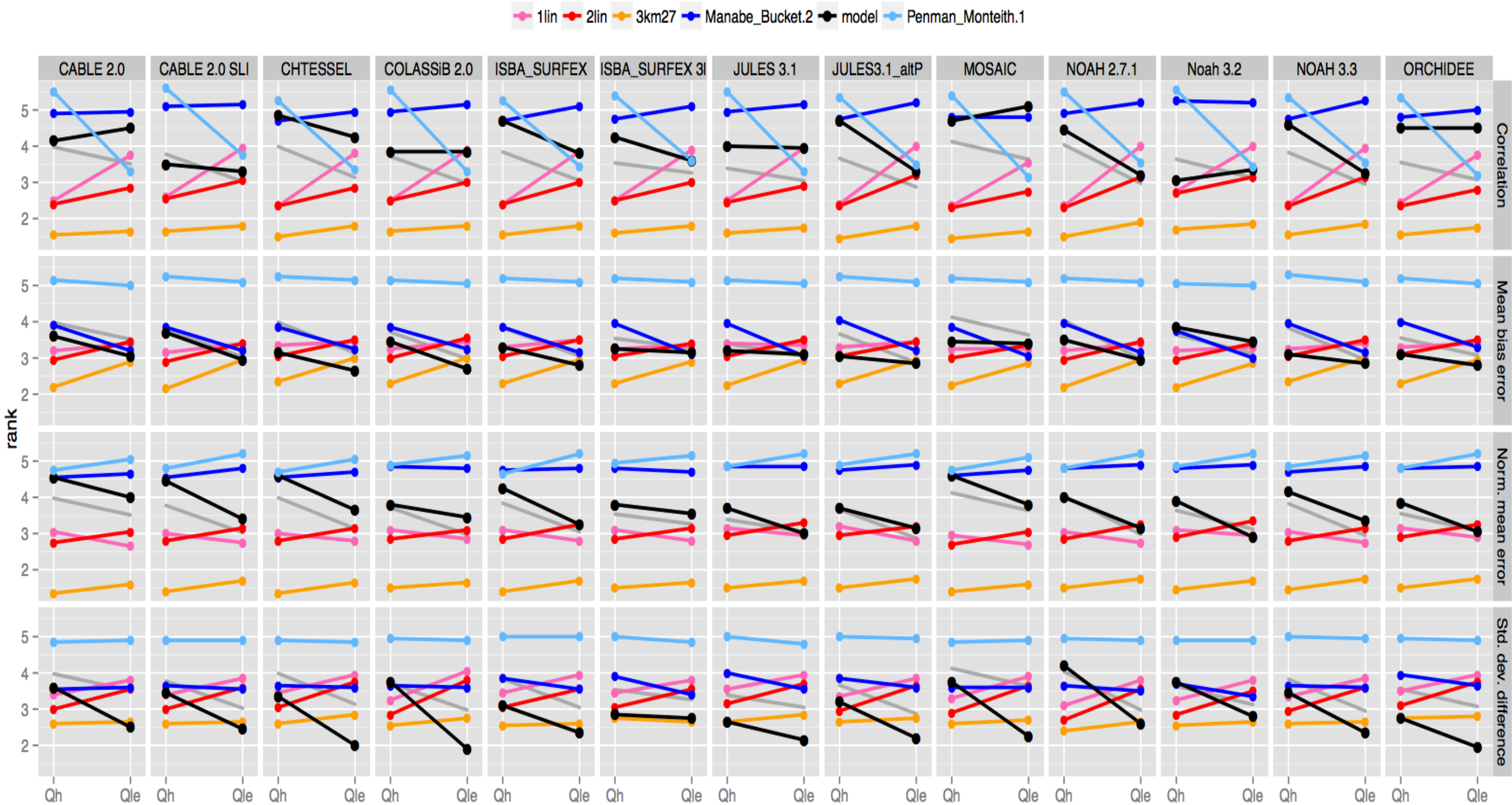
PLUMBER results – ranks vs metric values?



Haughton et al, J Hydromet, in review

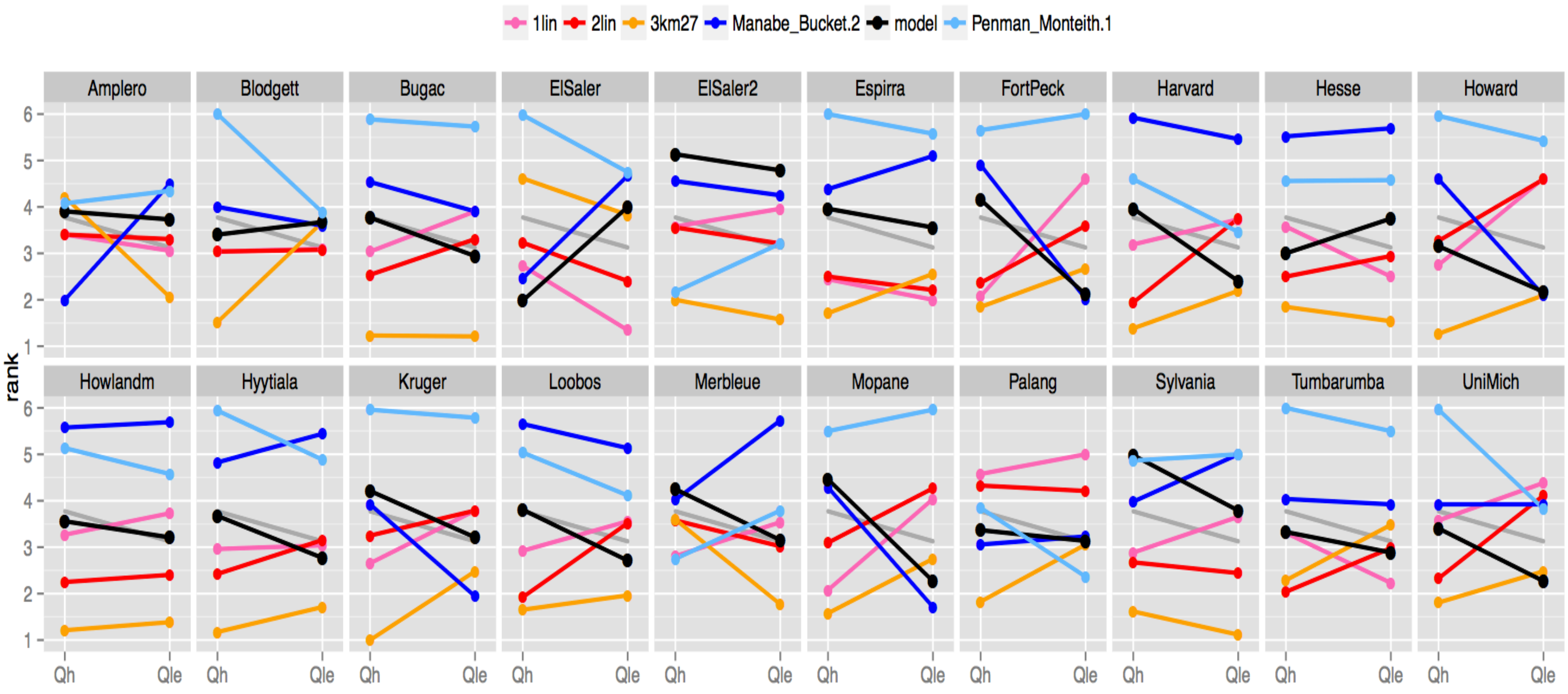
FIG. 3. Histograms of differences between metric values for benchmarks and models with neighbouring ranks, for all models at all sites. Values are calculated by taking the difference of the metric value for each model from the model ranked next-worst in for each LSM, Fluxnet site, metric, and variable. 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 results – metric?



Haughton et al, J Hydromet, in review

PLUMBER results – sites?



Haughton et al, J Hydromet, in review