



Session 3:

How well can we trust our models,
and how can we be sure?

Session 3a: 13:30 – 15:00


Gab Abramowitz, Beth Ebert, Bellie Sivakumar

Session 3b: 15:30 – 17:00

Dmitri Kavetski, Andrew Frost, Lucy Marshall, Gift Dumedah

Discussion questions

1. The simplest model that explains the observations is necessarily the best model.
2. All models are wrong, but some are still useful.
3. The models are not the main problem, it is the quality of the data and assumptions that go into them.
4. Much more effort is needed to objectively assess the performance of alternative models.
5. We need to stop calibrating our models, it leads to a false sense of security.
6. In circumstances where calibration is essential for a model to be useful, we should just use an empirical model (for example, based on data mining or Bayesian methods).
7. We cannot know whether to trust our models. Therefore multi-model ensembles should be standard operational practice, not just a research endeavour.
8. In the absence of quantitative knowledge of model inter-dependence, ensemble methods are meaningless.
9. Inappropriate values for unconstrained parameters (through calibration or assumption) should remove any trust in predictive ability.
10. Talking about 'physically-based' models is meaningless when there is not enough data to construct an empirical model.



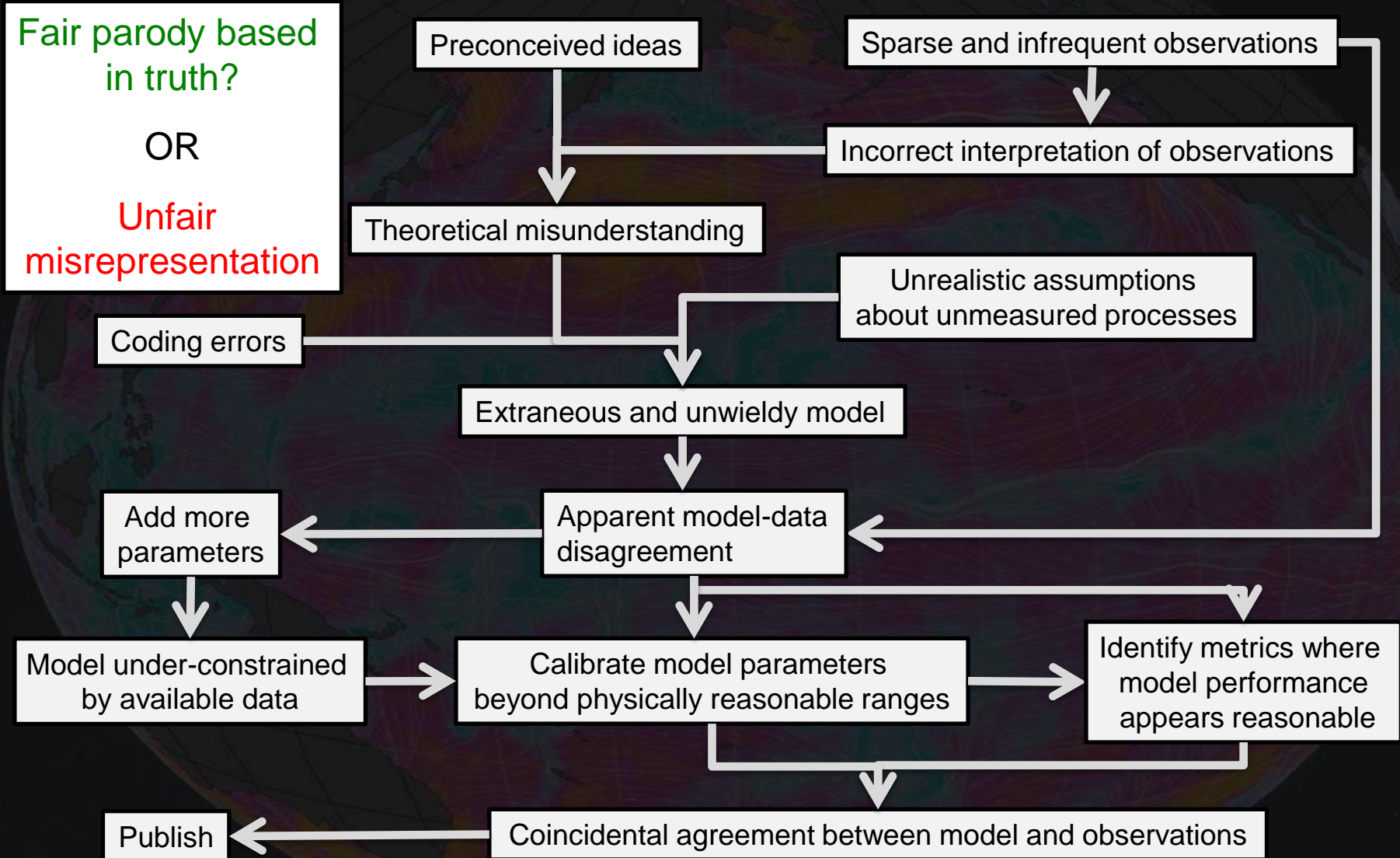
Defining expectations: an approach for quantifying trust in modelling

Gab Abramowitz

Climate Change Research Centre, UNSW and
ARC Centre of Excellence for Climate Systems Science

Ned Haughton and PLUMBER co-authors: M. Best, H. Johnson, A. Pitman, G. Balsamo, A. Boone, M. Cuntz, B. Decharme, P.A. Dirmeyer, J. Dong, M. Ek, Z. Guo, V. Haverd, B. van den Hurk, G. Nearing, B. Pak, C. Peters-Lidard, J. Santanello, L. Stevens, N. Vuichard.

How do we build and use a model?



(after AM Solomon)

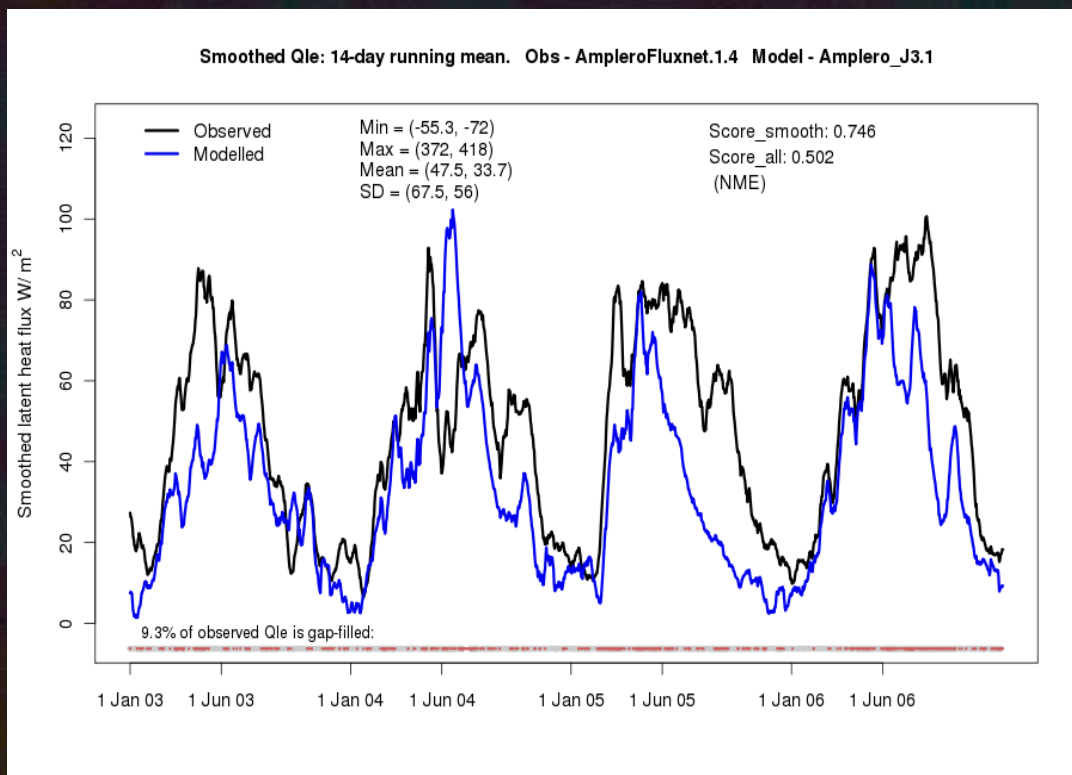
How close are we to this parody?

Think of the model you use - how do you know it's any good?

- Someone else published a paper saying so
- I've done some comparison with observations and it (qualitatively) looks good
- I have benchmark levels of performance that it must meet in prescribed tests, and it does

Define expectations of performance *a priori* – **before** running model

- e.g. previous model version (weak)
- e.g. fit for purpose (stronger / useful – can tell us if a model is “good enough”)
- e.g. utilises information about prediction variables in its inputs well (strong – give us an objective definition of whether a model is “good”)



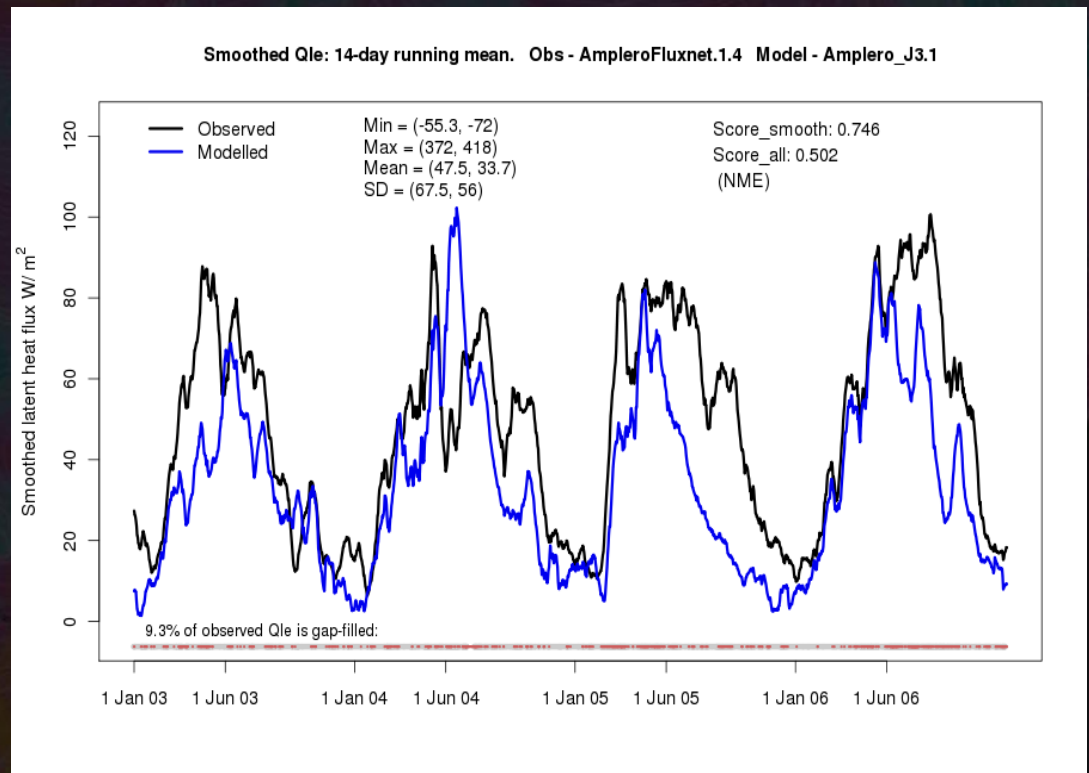
Benchmarking – an example

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

- Several (19) flux tower sites other than Amplerro
- Train a linear regression between shortwave radiation and LH
- Use these regression parameters to predict LH at Amplerro using site SW radiation

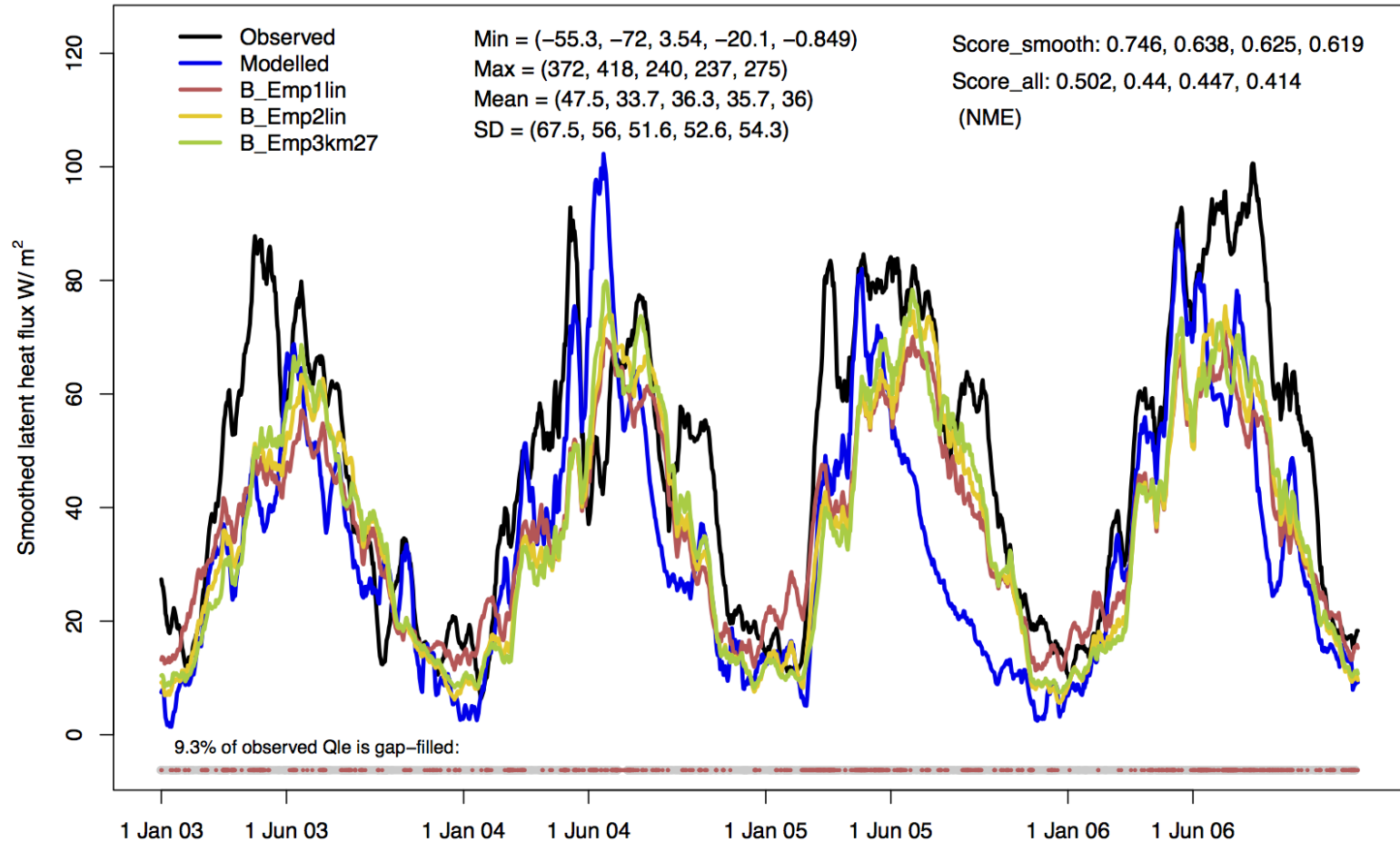
This will tell us:

- 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 Amplerro is, out-of-sample



Benchmarking – an example

Smoothed Qle: 14-day running mean. Obs – AmperoFluxnet.1.4 Model – Ampero_J3.1



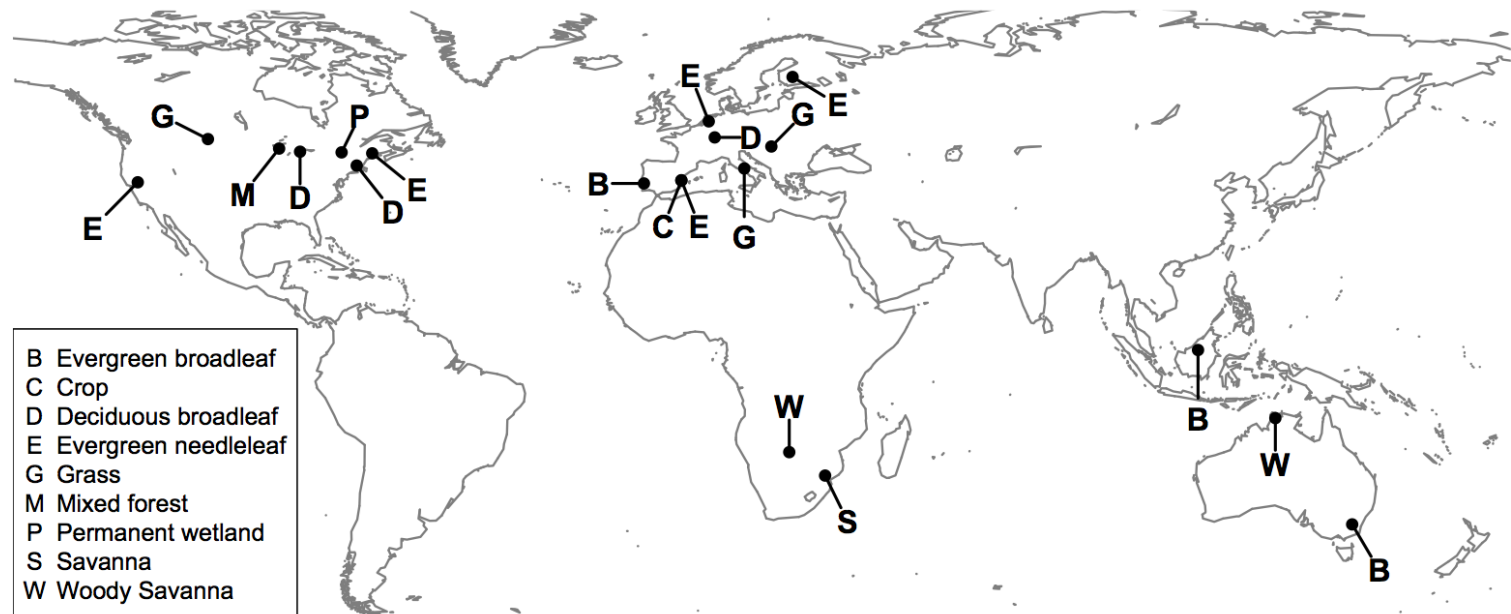
Benchmarking – an example

- Using empirical models (out of sample) as benchmarks can quantify the amount of information available to a model in its inputs about its prediction variables
- It gives one way to quantify how well we should *expect* a model to perform
- It provides a model-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)

It is a better benchmark than “*better than another model*”, since it can answer whether a model is “good” more objectively

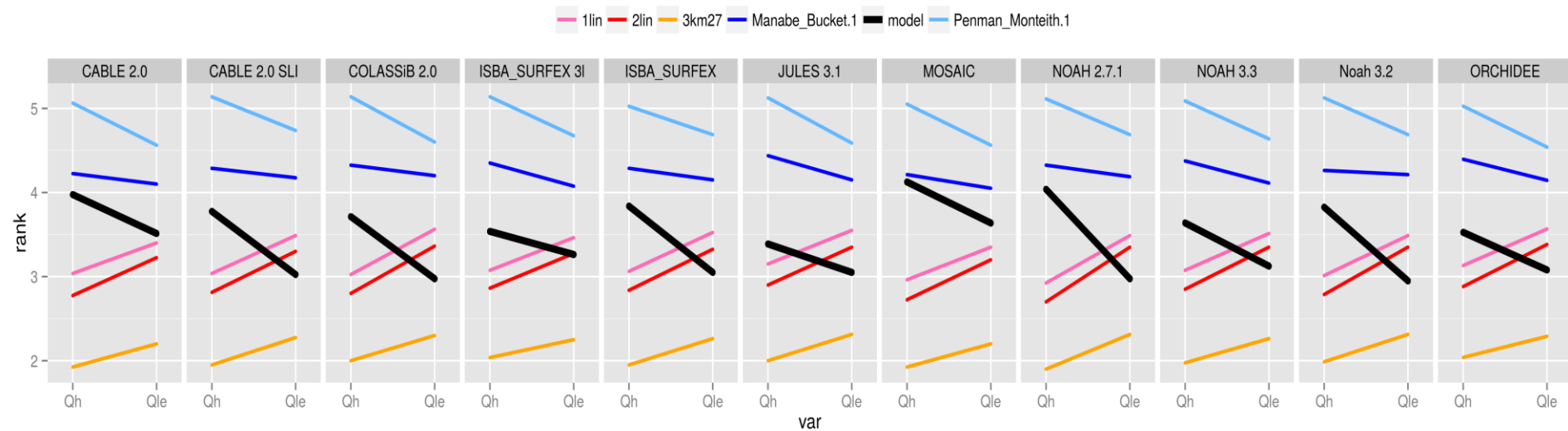
Extended example: The PALS Land sURface Model Benchmarking Evaluation pRoject (PLUMBER)

- 20 Flux tower sites; latent and sensible heat flux
- 4 metrics: bias, correlation, SD, normalised mean error
- 9 LSMs, 15 LSM versions
- Benchmarks: two 'physical' – PM and Manabe bucket; 3 empirical

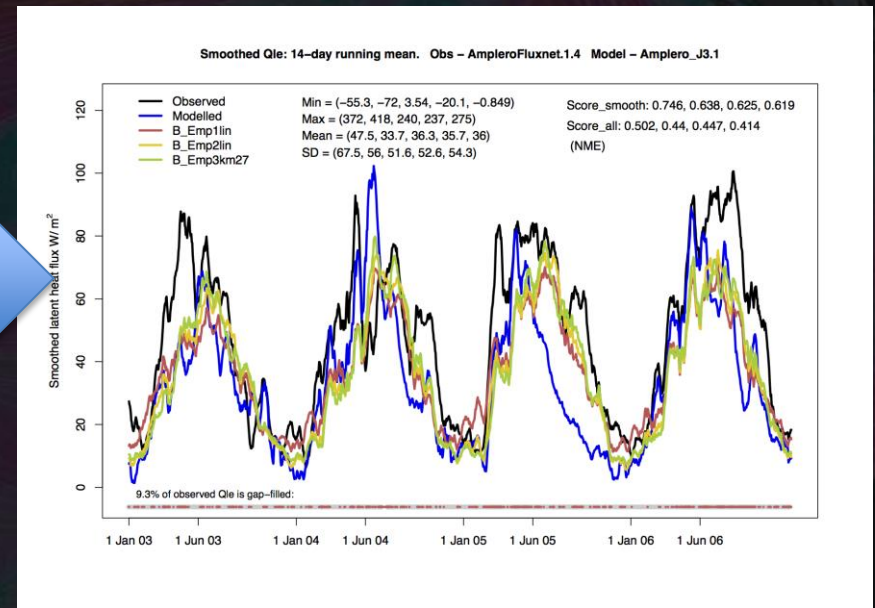
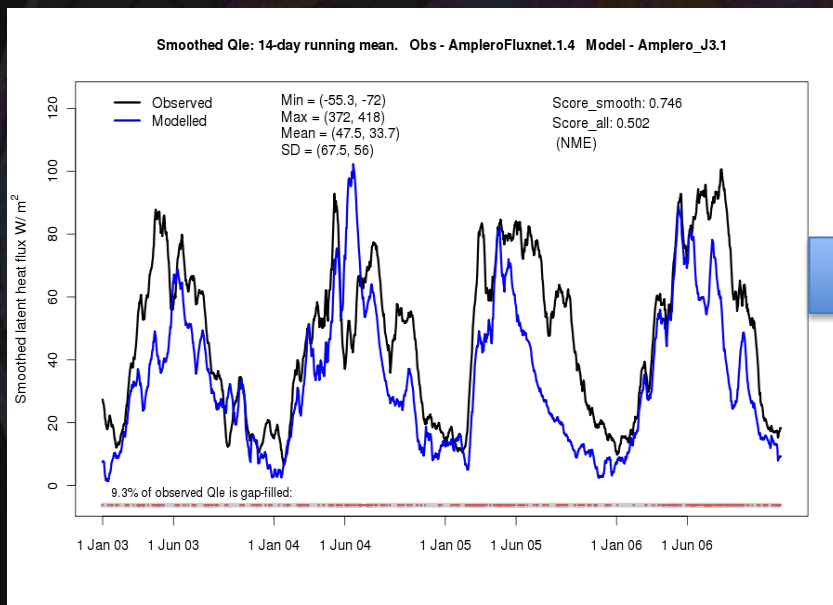


The three empirical benchmarks in PLUMBER

- All 3 empirical models relate net 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
- All are instantaneous responses to met variables with no knowledge of vegetation type, soil type, soil moisture or temperature, C pools.



If we had not tried to quantify information available in met data about fluxes (in this case using empirical models) we would still believe models are doing well!



Quantifying expectations is key to understanding how “good” a model is
=> trust in models

PLUMBER results – why?

1. Flux tower measurements – conservation issues?
2. Is it because the PLUMBER analysis focuses on short timescales?
3. Are flux towers are at the wrong spatial scale?
4. Is the state initialisation inappropriate?
5. Time scale of state variables?
6. Over-parameterisation is hurting – calibration of unconstrained parameters inhibits predictive capacity?
7. LSMs are essentially conceptual models – too many processes not supported by data in the scope of their application

Dissecting the PLUMBER results

Poster: Ned Haughton *Why are land surface models performing so poorly?*

- Energy conservation in flux tower data is NOT the issue
- Looking at longer timescales (where LSMs' states might help) does not change the rank
- We can also try to separate whether:
 - The instantaneous model response is the issue, or
 - Inappropriate magnitude / time scale of model states is the issue

Please go and talk to Ned to find out more....

PLUMBER results – why?

1. Flux tower measurements – conservation issues?
2. Is it because the PLUMBER analysis focuses on short timescales?
3. Are flux towers are at the wrong spatial scale?
4. Is the state initialisation inappropriate?
5. Time scale of state variables?
6. Over-parameterisation is hurting – calibration of unconstrained parameters inhibits predictive capacity?
7. LSMs are essentially conceptual models – too many processes not supported by data in the scope of their application

Questions

- What does it 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?
- How do we know our conceptual representations have any value in the absence of observations that can confirm process representation?
- Has the drive to add more processes into LSMs (often based on sparse data sets) led to intractable modelling systems with relatively poor accuracy?
- Are inappropriate values for the unconstrained parameters (through calibration) actively inhibiting predictive ability?
- Does an incremental improvement in performance through adding / improving a process representation in a LSM matter if SWdown, Tair and humidity provides enough information to comfortably outperform the LSM for latent heat flux, sensible heat flux and NEE?