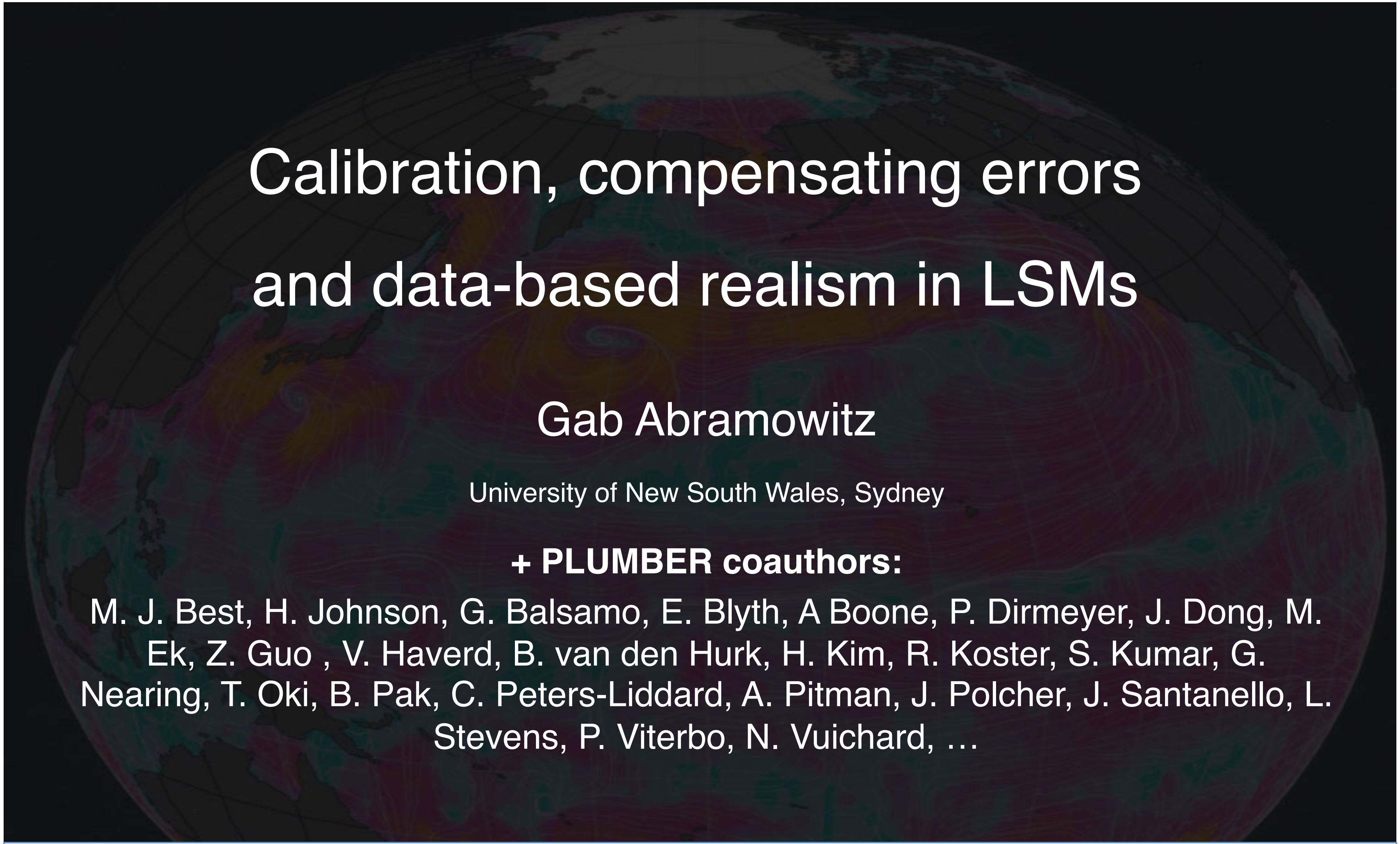


Calibration, compensating errors and data-based realism in LSMs

Gab Abramowitz

University of New South Wales, Sydney

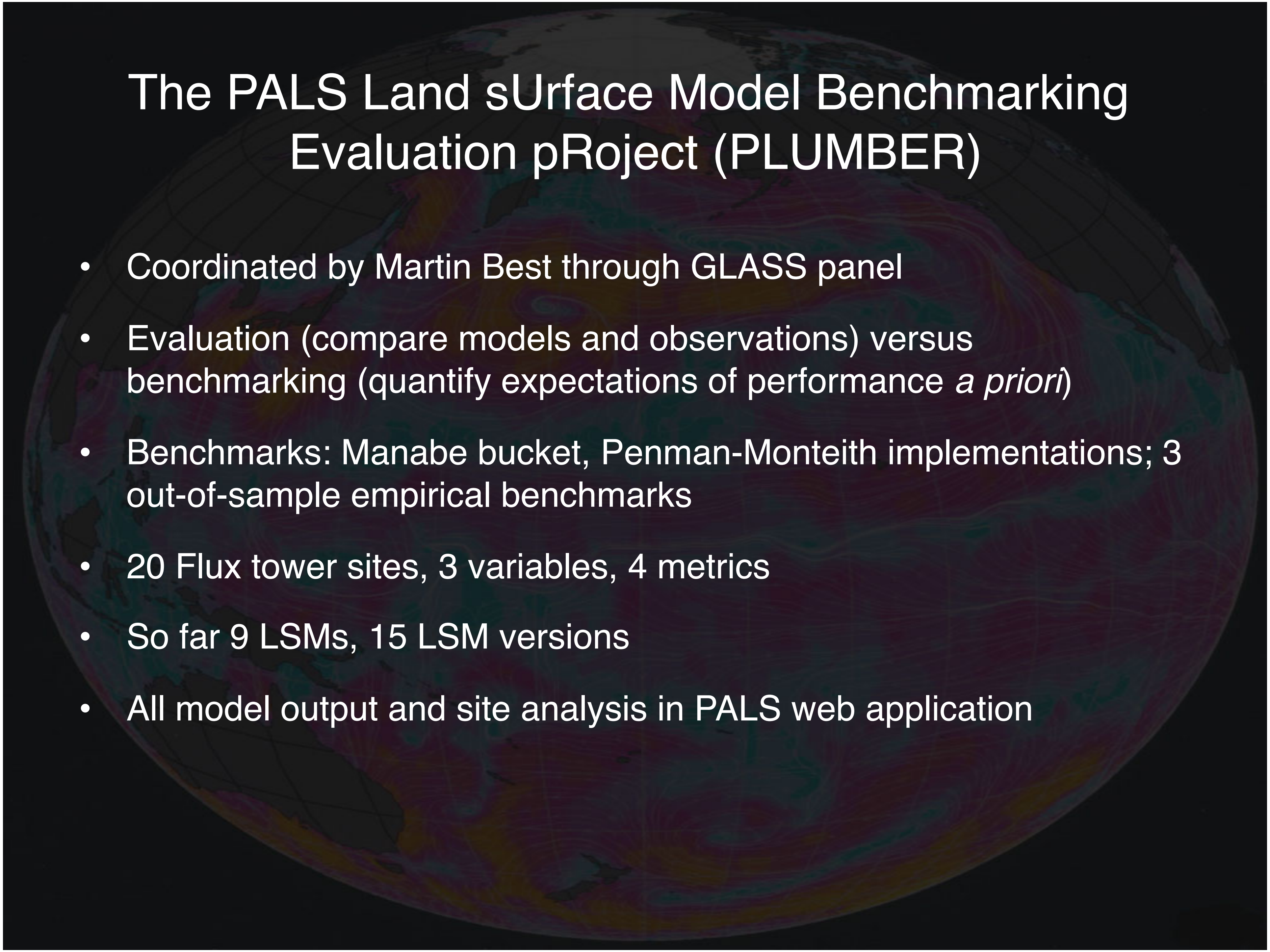
+ PLUMBER coauthors:

M. J. Best, H. Johnson, G. Balsamo, E. Blyth, A Boone, P. Dirmeyer, J. Dong, M. Ek, Z. Guo , V. Haverd, B. van den Hurk, H. Kim, R. Koster, S. Kumar, G. Nearing, T. Oki, B. Pak, C. Peters-Liddard, A. Pitman, J. Polcher, J. Santanello, L. Stevens, P. Viterbo, N. Vuichard, …

UNSW THE UNIVERSITY OF NEW SOUTH WALES

ARC CENTRE OF EXCELLENCE FOR CLIMATE SYSTEM SCIENCE

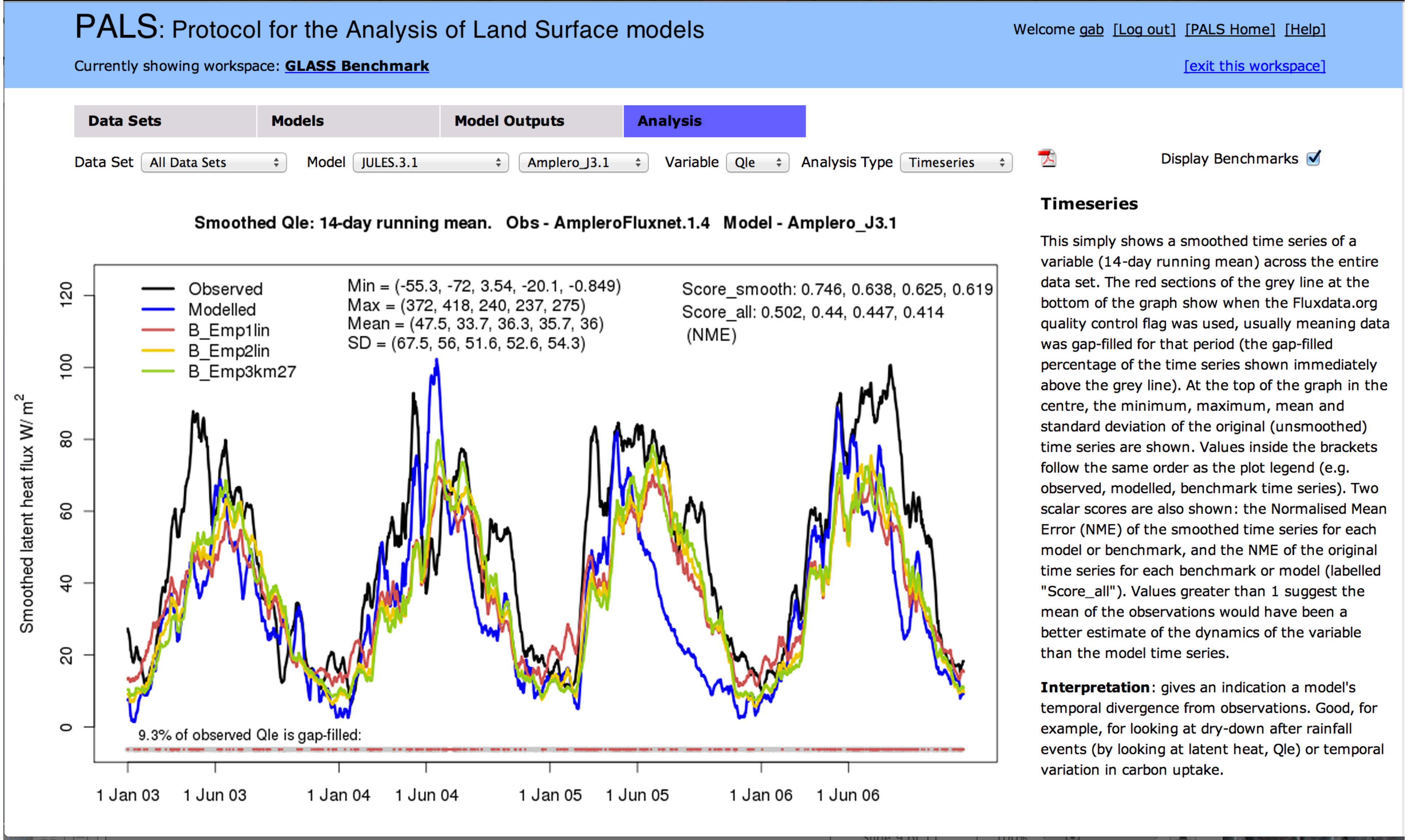
TERN Terrestrial Ecosystem Research Network



The PALS Land sUrface Model Benchmarking Evaluation pRoject (PLUMBER)

- Coordinated by Martin Best through GLASS panel
- Evaluation (compare models and observations) versus benchmarking (quantify expectations of performance *a priori*)
- Benchmarks: Manabe bucket, Penman-Monteith implementations; 3 out-of-sample empirical benchmarks
- 20 Flux tower sites, 3 variables, 4 metrics
- So far 9 LSMs, 15 LSM versions
- All model output and site analysis in PALS web application

PALS



PALS: Protocol for the Analysis of Land Surface models

Welcome gab [Log out] [PALS Home] [Help]

Currently showing workspace: **GLASS Benchmark**

[exit this workspace]

Data Set: All Data Sets | Model: JULES.3.1 | Amplero_J3.1 | Variable: Qle | Analysis Type: Timeseries | Display Benchmarks ☑

Smoothed Qle: 14-day running mean. Obs - AmpleroFluxnet.1.4 Model - Amplero_J3.1

Min = (-55.3, -72, 3.54, -20.1, -0.849)
Max = (372, 418, 240, 237, 275)
Mean = (47.5, 33.7, 36.3, 35.7, 36)
SD = (67.5, 56, 51.6, 52.6, 54.3)

Score_smooth: 0.746, 0.638, 0.625, 0.619
Score_all: 0.502, 0.44, 0.447, 0.414
(NME)

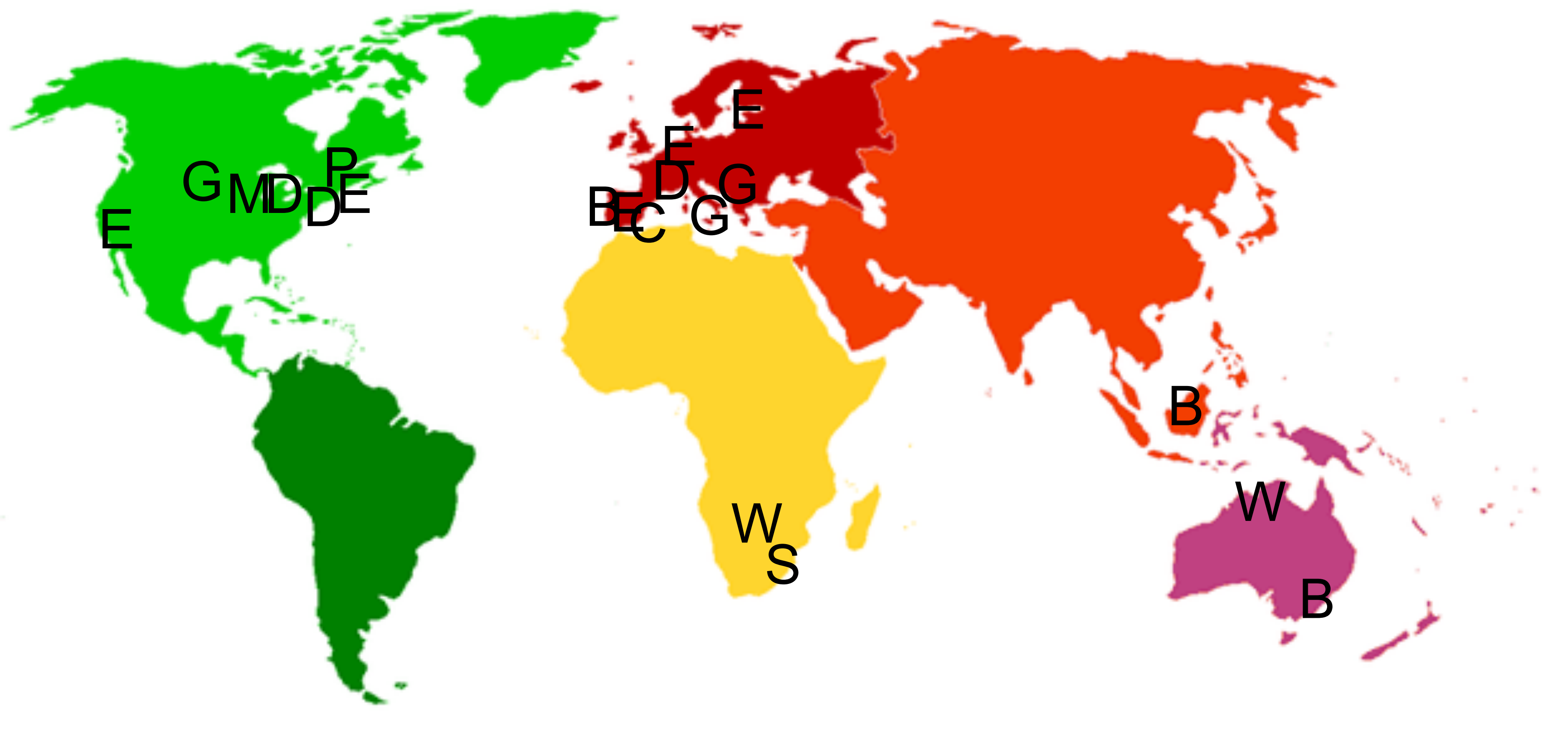
9.3% of observed Qle is gap-filled:

Timeseries

This simply shows a smoothed time series of a variable (14-day running mean) across the entire data set. The red sections of the grey line at the bottom of the graph show when the Fluxdata.org quality control flag was used, usually meaning data was gap-filled for that period (the gap-filled percentage of the time series shown immediately above the grey line). At the top of the graph in the centre, the minimum, maximum, mean and standard deviation of the original (unsmoothed) time series are shown. Values inside the brackets follow the same order as the plot legend (e.g. observed, modelled, benchmark time series). Two scalar scores are also shown: the Normalised Mean Error (NME) of the smoothed time series for each model or benchmark, and the NME of the original time series for each benchmark or model (labelled "Score_all"). Values greater than 1 suggest the mean of the observations would have been a better estimate of the dynamics of the variable than the model time series.

Interpretation: gives an indication a model's temporal divergence from observations. Good, for example, for looking at dry-down after rainfall events (by looking at latent heat, Qle) or temporal variation in carbon uptake.

PLUMBER – 20 flux tower sites

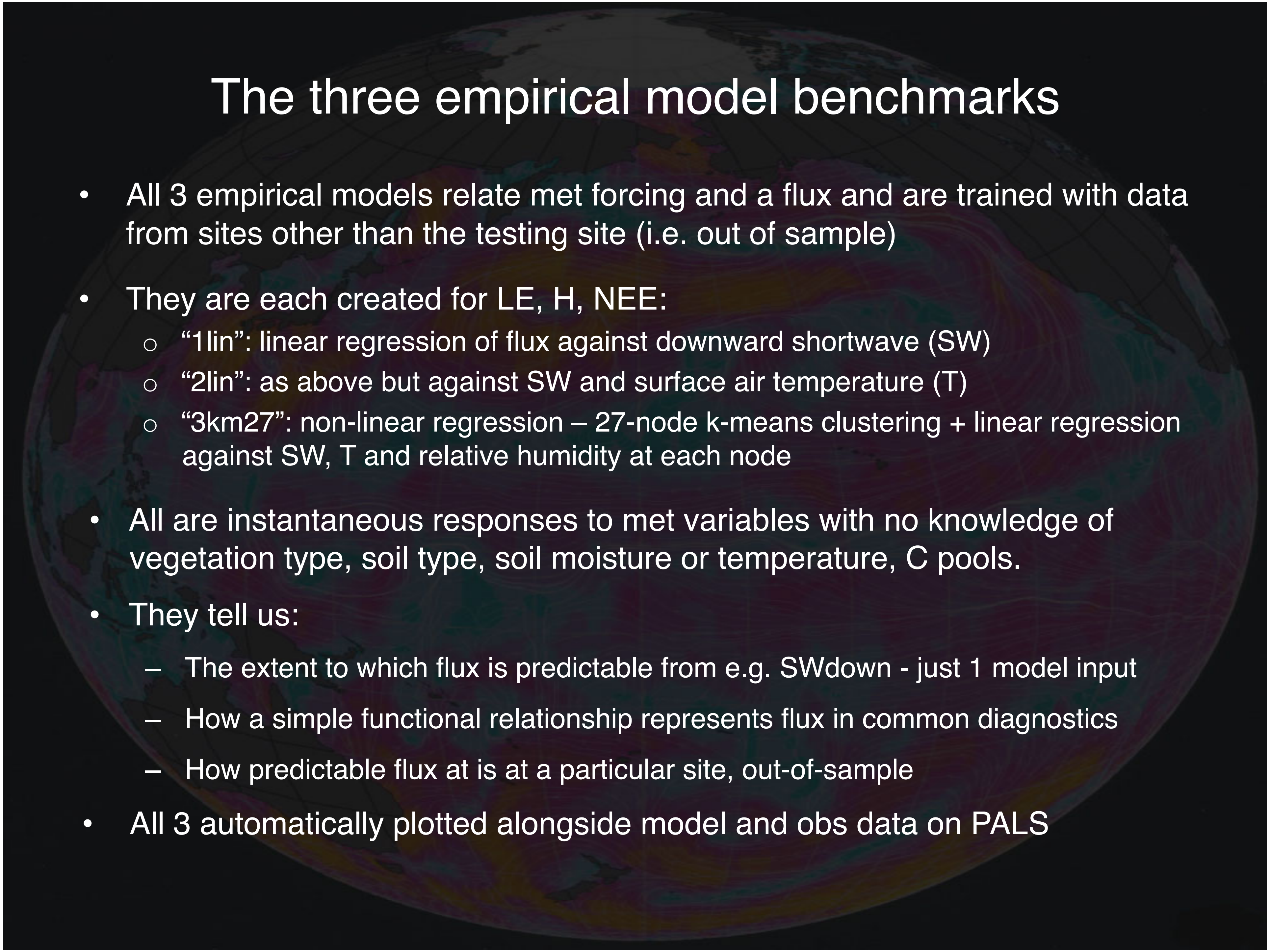


E – Evergreen Needleleaf
B – Evergreen Broadleaf
D – Deciduous Broadleaf
M – Mixed Forest

M – Mixed Forest
G - Grassland
C – Cropland

W – Woody Savanna
S – Savanna
P – Permanent Wetlands

The three empirical model benchmarks



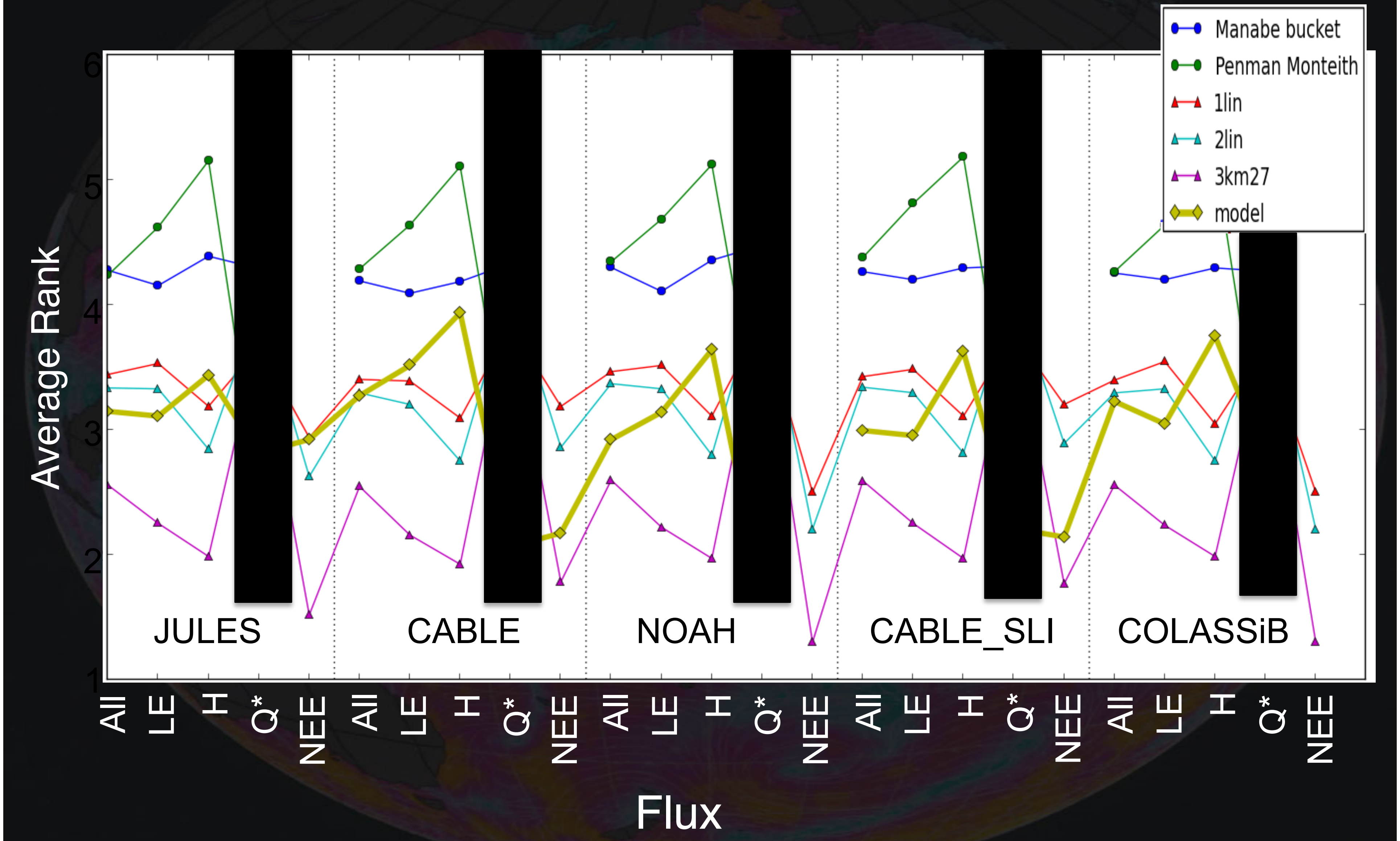
- All 3 empirical models relate met forcing and a flux and are trained with data from sites other than the testing site (i.e. out of sample)
- They are each created for LE, H, NEE:
  - “1lin”: linear regression of flux against downward shortwave (SW)
  - “2lin”: as above but against SW and surface air 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.
- They tell us:
  - The extent to which flux is predictable from e.g. SWdown - just 1 model input
  - How a simple functional relationship represents flux in common diagnostics
  - How predictable flux at is at a particular site, out-of-sample
- All 3 automatically plotted alongside model and obs data on PALS

PLUMBER – variables and metrics

Sensible heat flux (H)	Latent Heat flux (LE)	Net Ecosystem Exchange (NEE)

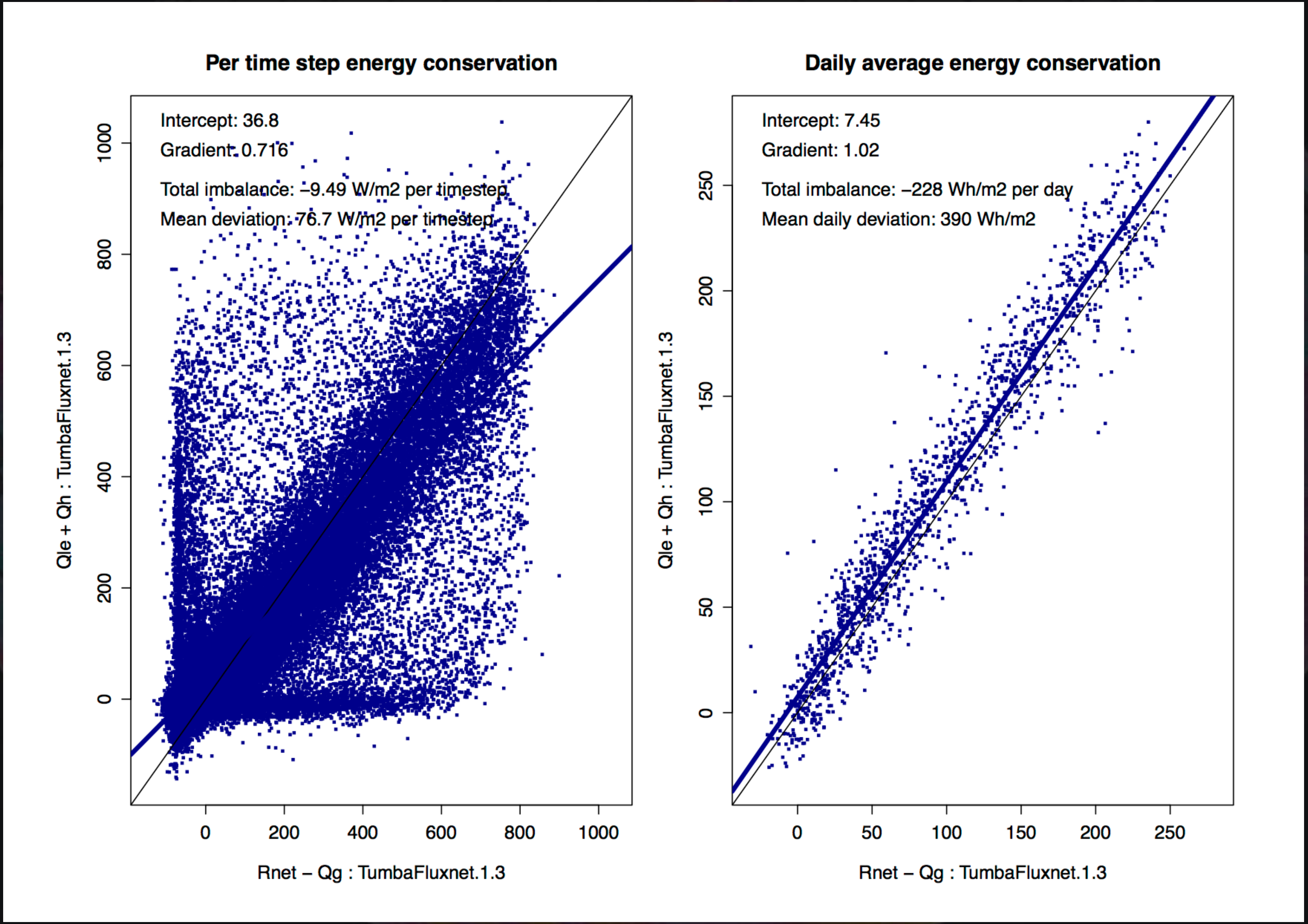
Mean Bias Error	MBE	$\left(\sum_{i=1}^{n}(M_i - O_i)\right)/n$
Normalised Mean Error	NME	$\dfrac{\sum_{i=1}^{n}\lvert M_i - O_i \rvert}{\sum_{i=1}^{n}\lvert \overline{O} - O_i \rvert}$
Standard Deviation	sd	$\sqrt{\dfrac{\sum_{i=1}^{n}\left(M_i - \overline{M}\right)^2}{n-1}}$
Correlation coefficient	r	$\dfrac{n\sum_{i=1}^{n}(O_i M_i) - \left(\sum_{i=1}^{n} O_i \sum_{i=1}^{n} M_i\right)}{\sqrt{\left(n\sum_{i=1}^{n} O_i^2 - \left(\sum_{i=1}^{n} O_i\right)^2\right)\left(n\sum_{i=1}^{n} M_i^2 - \left(\sum_{i=1}^{n} M_i\right)^2\right)}}$

PLUMBER – results (old) – from Martin Best

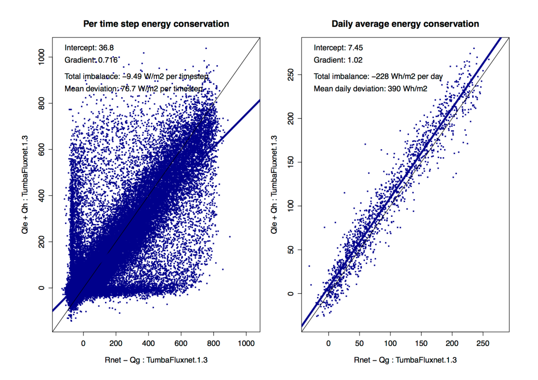


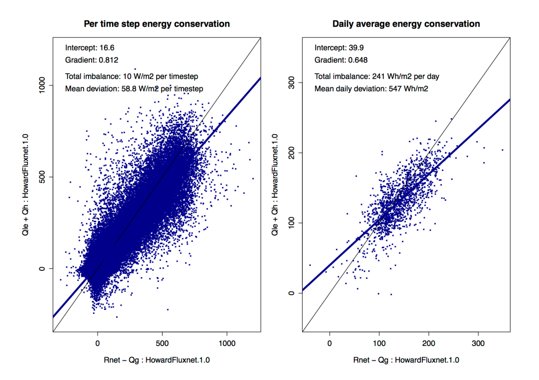
[redacted]
[redacted]
[redacted]
[redacted]
[redacted]

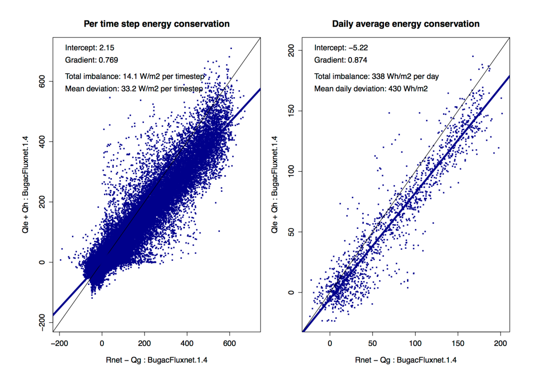
Flux tower systematic bias?

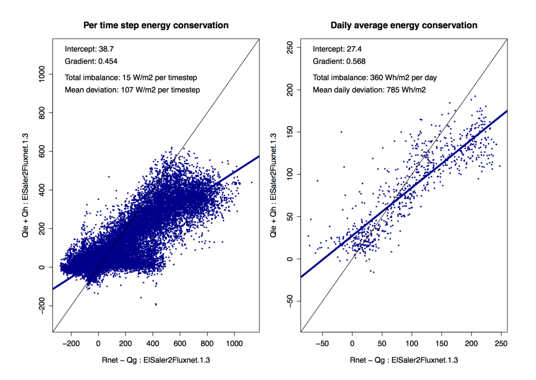


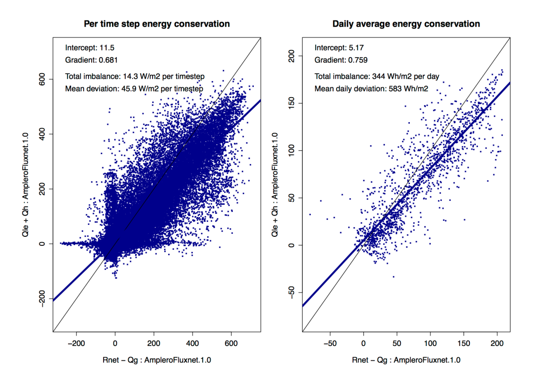
Flux tower systematic bias?

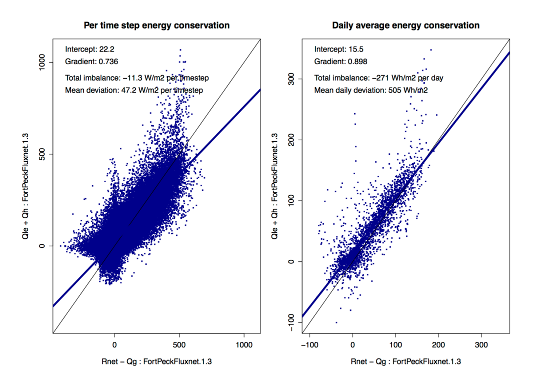


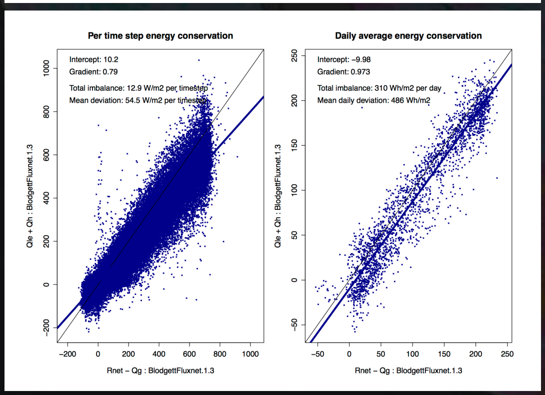


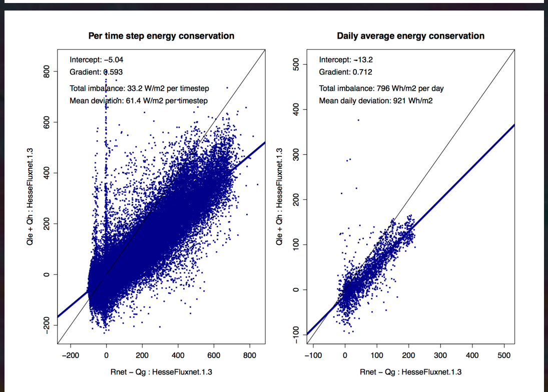


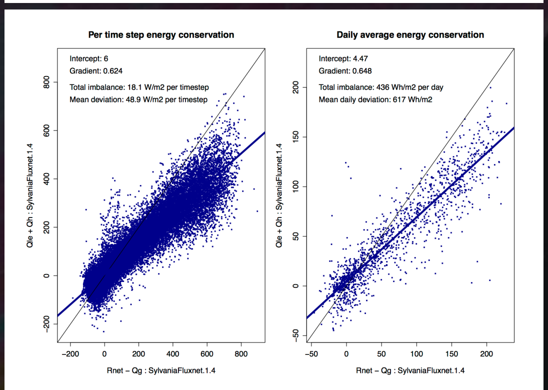








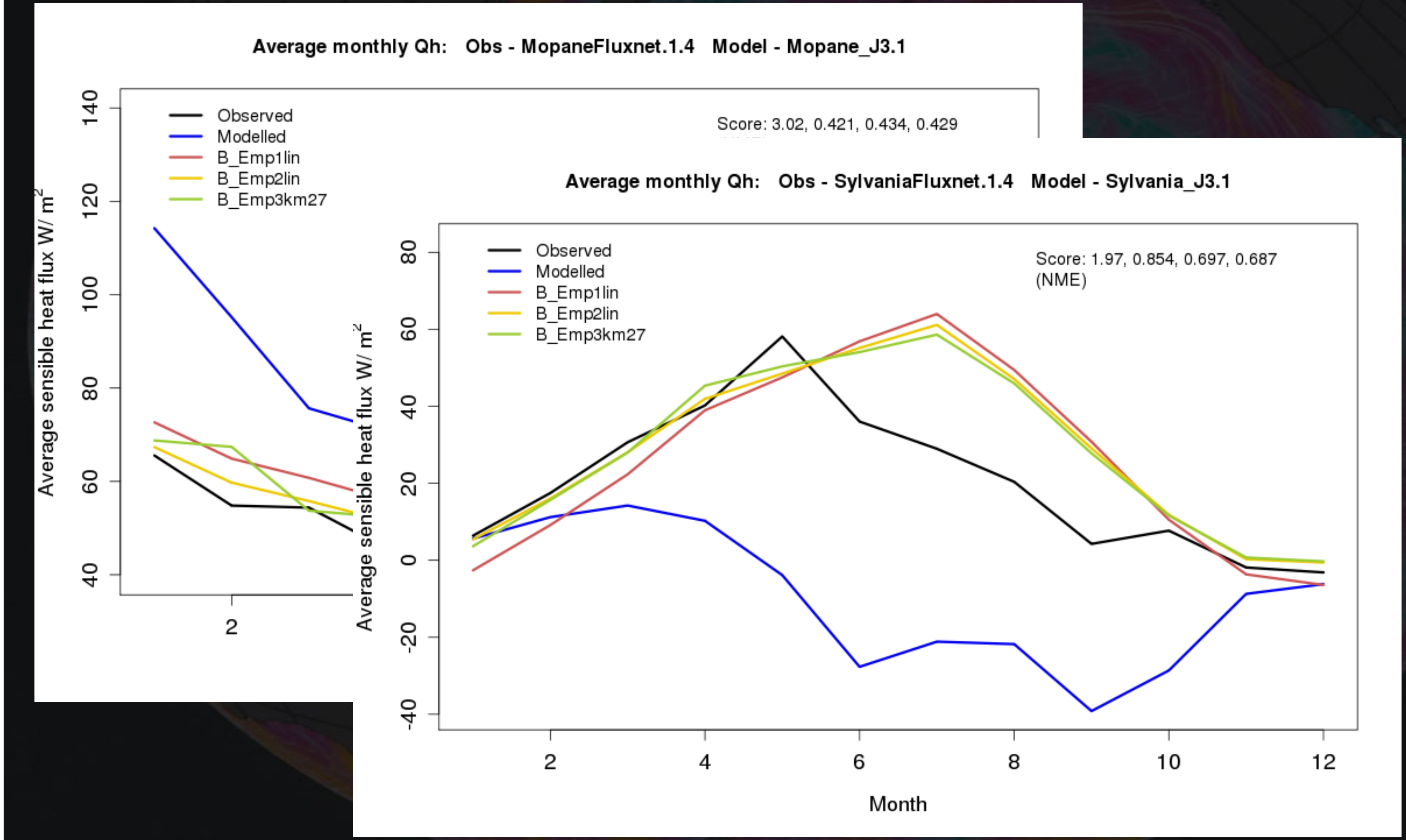




www.pals.unsw.edu.au

Flux tower systematic bias?

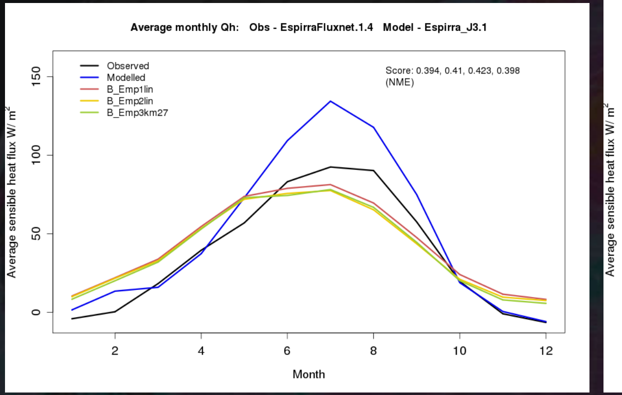
Sensible heat annual cycle examples where benchmarks win – JULES:



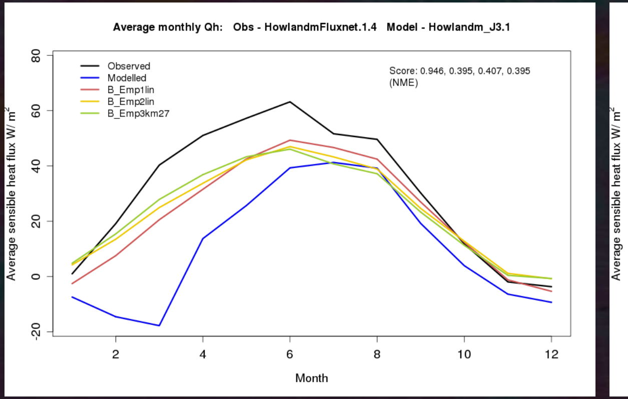
www.pals.unsw.edu.au

Flux tower systematic bias?

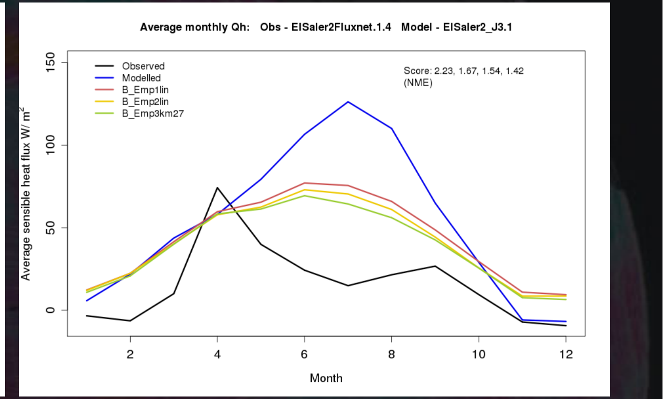
Sensible heat annual cycle examples where benchmarks win – JULES:


Average monthly Qh: Obs - EspirraFluxnet.1.4 Model - Espirra_J3.1

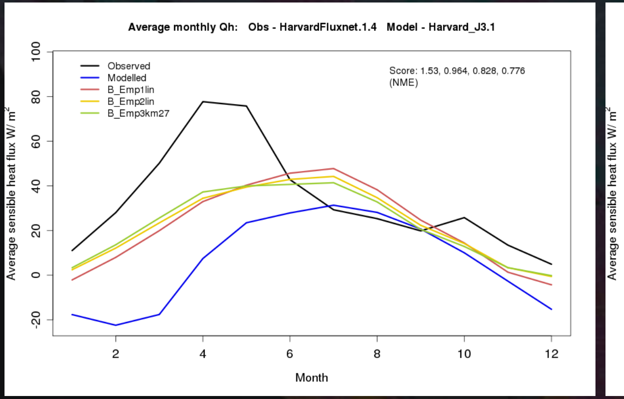
Score: 0.394, 0.41, 0.423, 0.398 (NME)


Average monthly Qh: Obs - HowlandmFluxnet.1.4 Model - Howlandm_J3.1

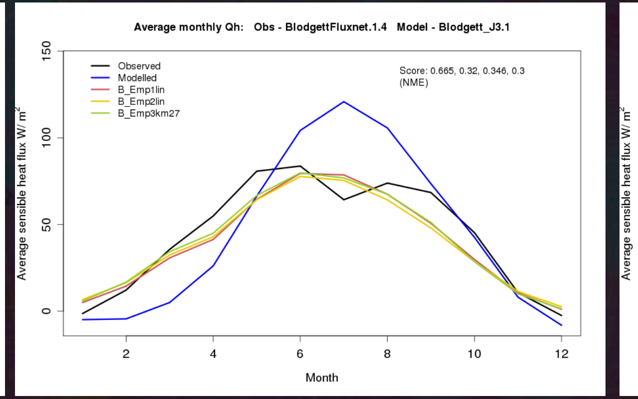
Score: 0.946, 0.395, 0.407, 0.395 (NME)


Average monthly Qh: Obs - ElSaler2Fluxnet.1.4 Model - ElSaler2_J3.1

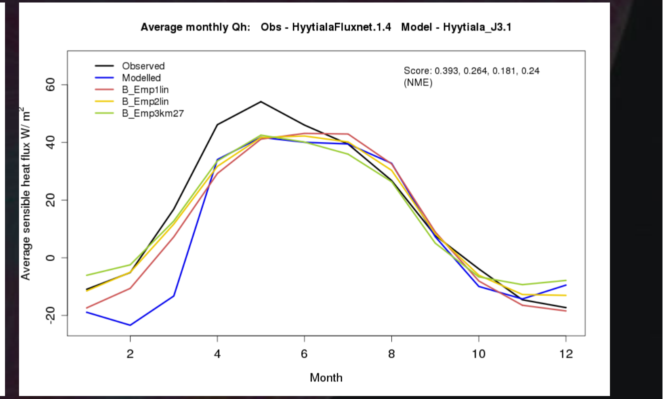
Score: 2.23, 1.67, 1.54, 1.42 (NME)


Average monthly Qh: Obs - HarvardFluxnet.1.4 Model - Harvard_J3.1

Score: 1.53, 0.964, 0.828, 0.776 (NME)


Average monthly Qh: Obs - BlodgettFluxnet.1.4 Model - Blodgett_J3.1

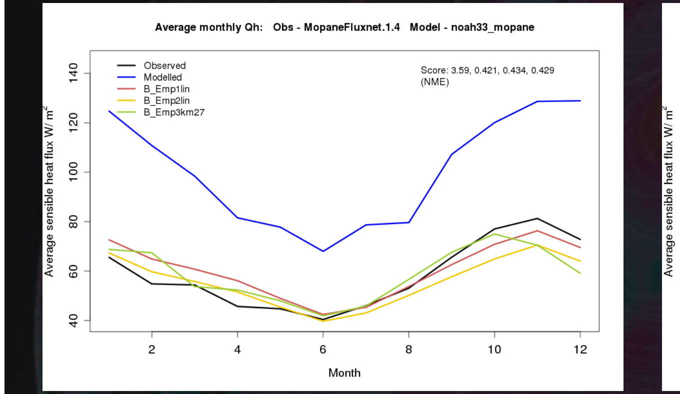
Score: 0.665, 0.32, 0.346, 0.3 (NME)

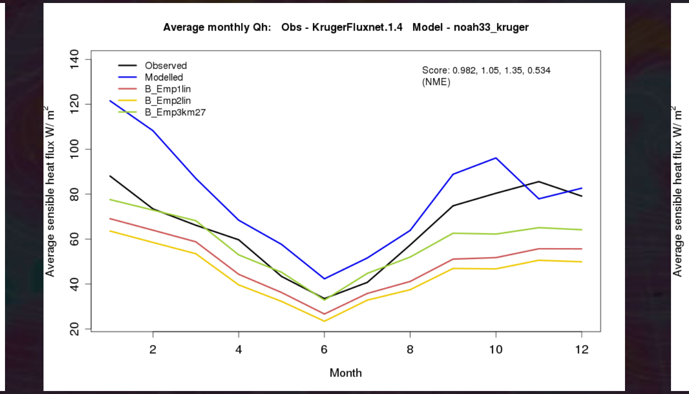

Average monthly Qh: Obs - HyytialaFluxnet.1.4 Model - Hyytiala_J3.1

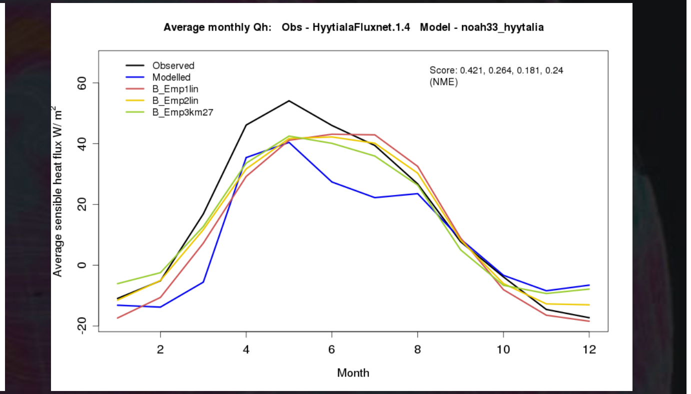
Score: 0.393, 0.264, 0.181, 0.24 (NME)

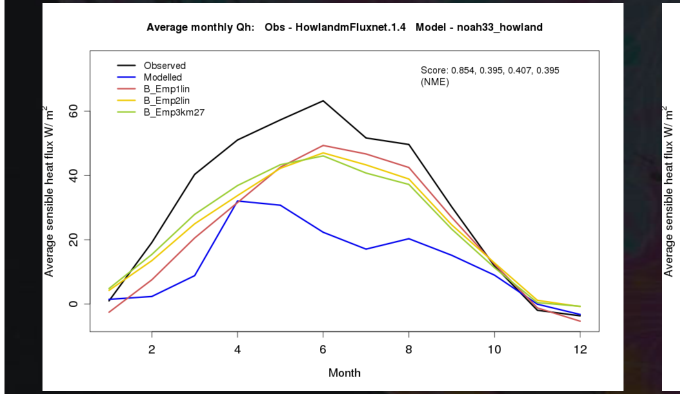
Flux tower systematic bias?

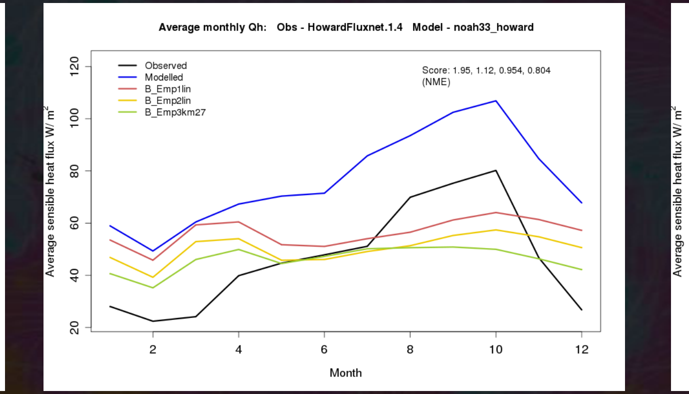
Sensible heat annual cycle examples where benchmarks win – NOAH:

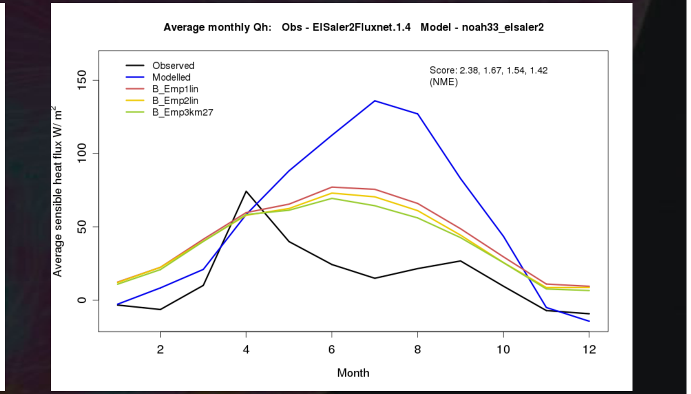








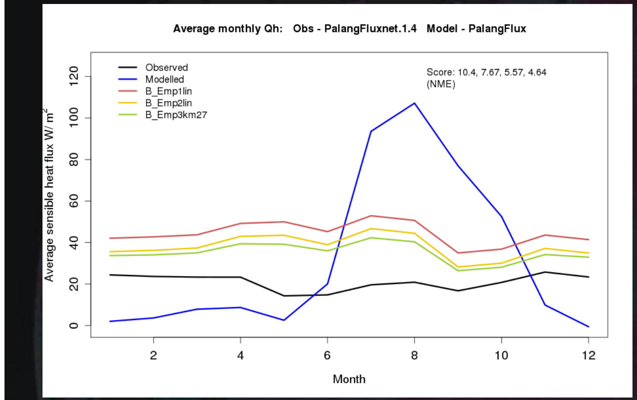


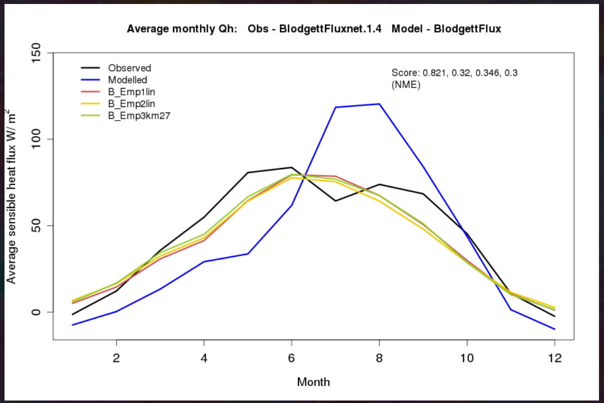


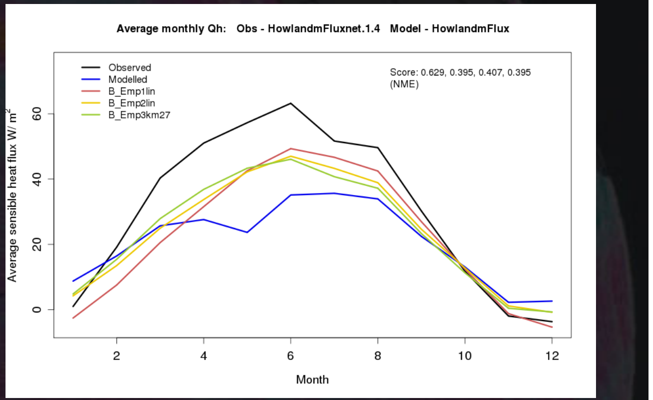
www.pals.unsw.edu.au

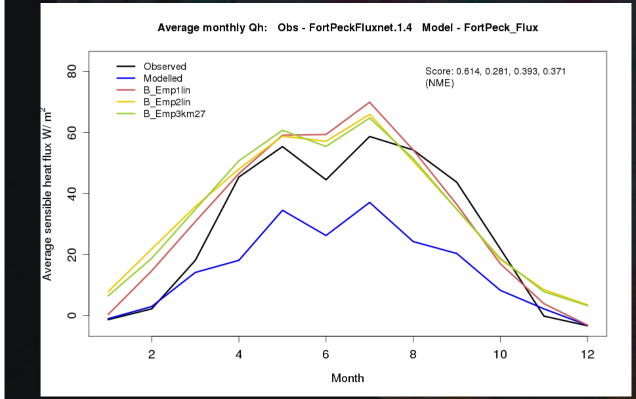
Flux tower systematic bias?

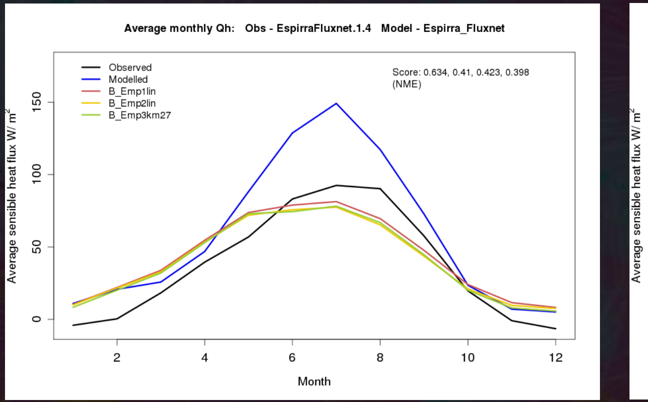
Sensible heat annual cycle examples where benchmarks win – COLASSiB:

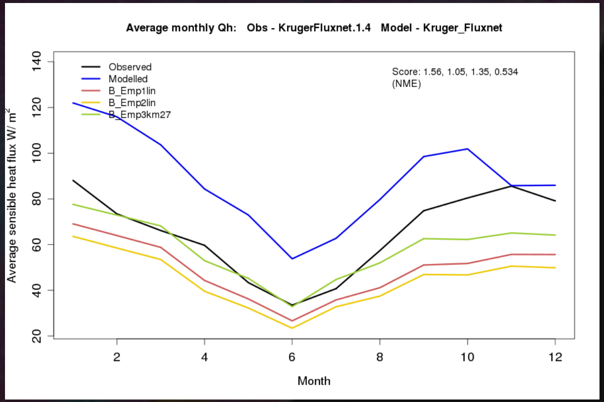


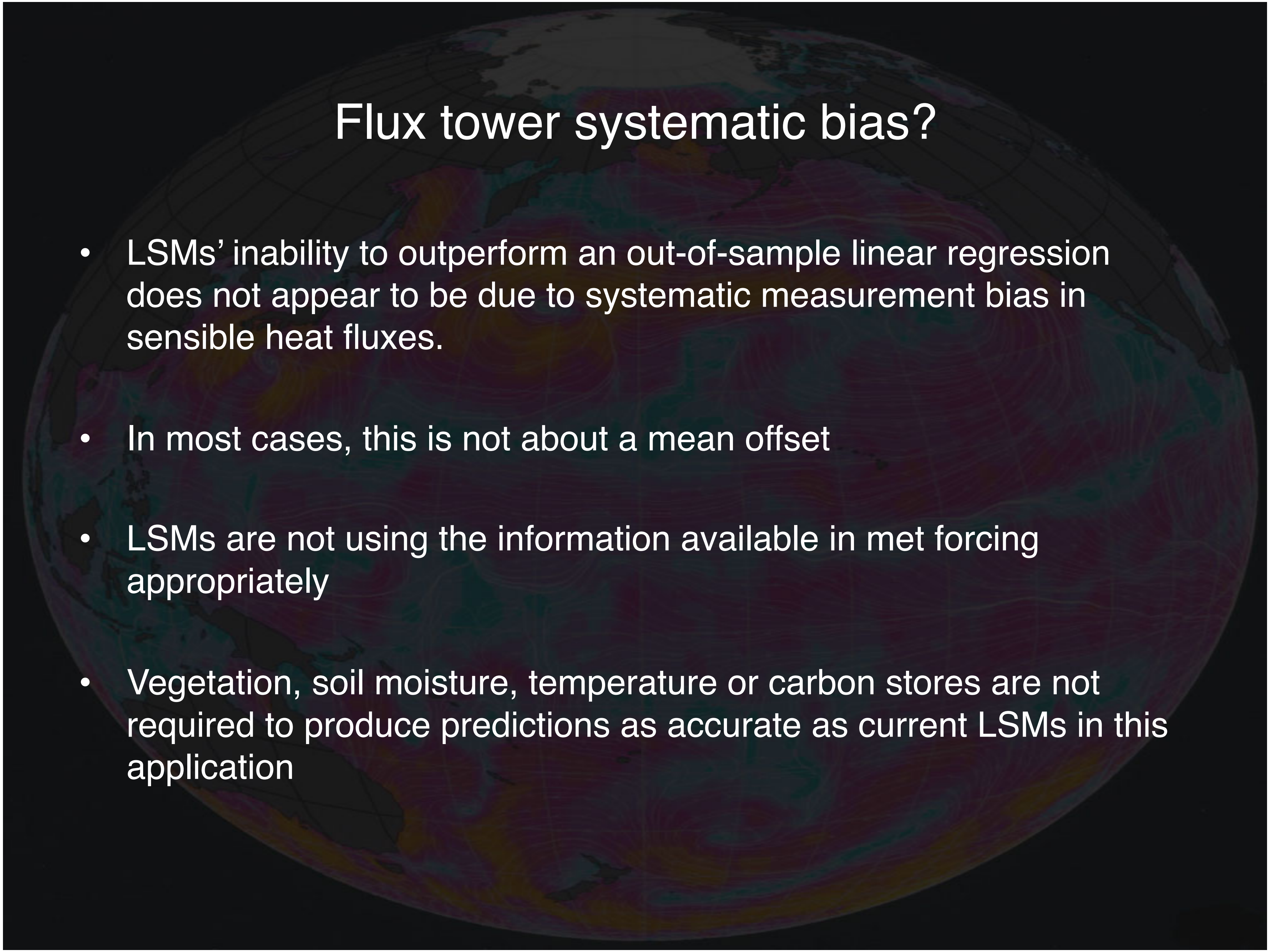












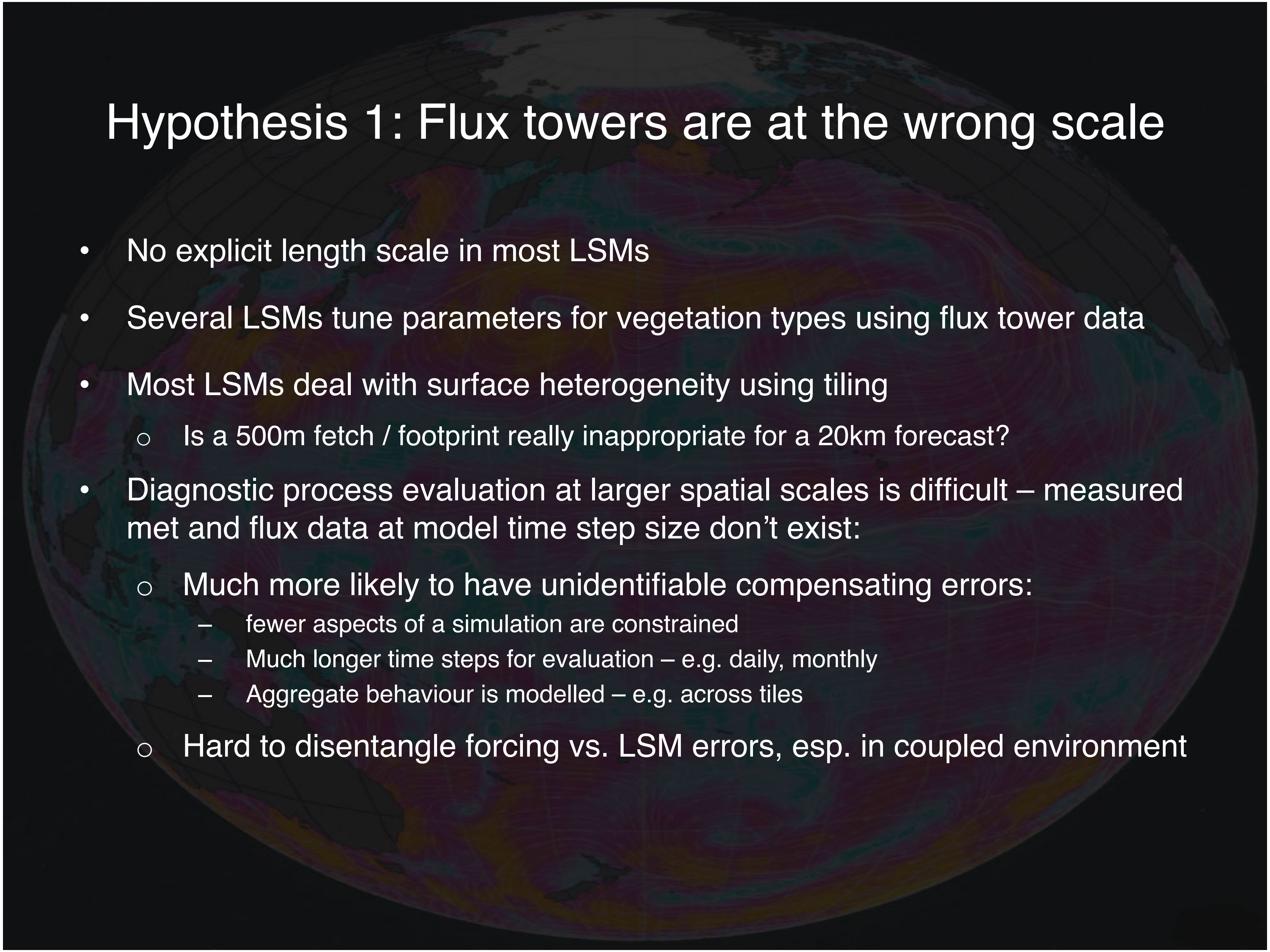
Flux tower systematic bias?

- LSMs’ inability to outperform an out-of-sample linear regression does not appear to be due to systematic measurement bias in sensible heat fluxes.

- In most cases, this is not about a mean offset

- LSMs are not using the information available in met forcing appropriately

- Vegetation, soil moisture, temperature or carbon stores are not required to produce predictions as accurate as current LSMs in this application



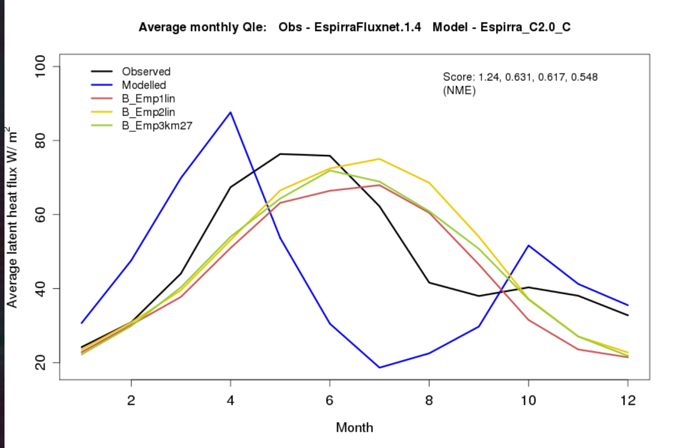
Hypothesis 1: Flux towers are at the wrong scale

- No explicit length scale in most LSMs
- Several LSMs tune parameters for vegetation types using flux tower data
- Most LSMs deal with surface heterogeneity using tiling
  - Is a 500m fetch / footprint really inappropriate for a 20km forecast?
- Diagnostic process evaluation at larger spatial scales is difficult – measured met and flux data at model time step size don’t exist:
  - Much more likely to have unidentifiable compensating errors:
    - fewer aspects of a simulation are constrained
    - Much longer time steps for evaluation – e.g. daily, monthly
    - Aggregate behaviour is modelled – e.g. across tiles
  - Hard to disentangle forcing vs. LSM errors, esp. in coupled environment

Hypothesis 2: State initialisation is inappropriate

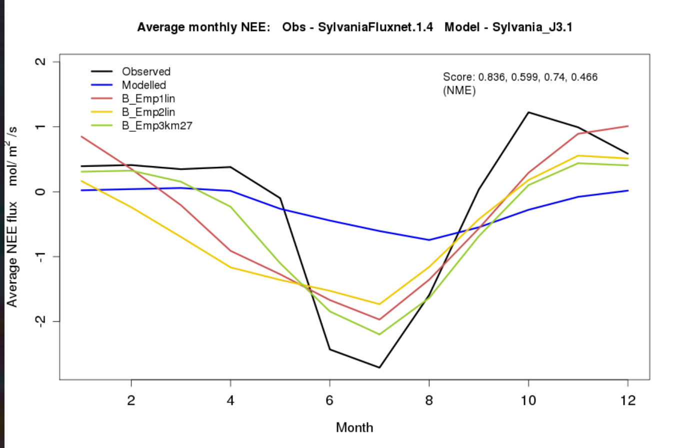
- Repeated spin-up on flux tower met data – no guarantee its representative
- Likely not perfect, but unlikely the major issue – very few cases of flux being consistently too high or too low for all months of average annual cycle:

Latent heat	Too high	Too low	neither
JULES	1	3	16
NOAH	3	1	16
COLA	1	1	18
CABLE	4	0	16

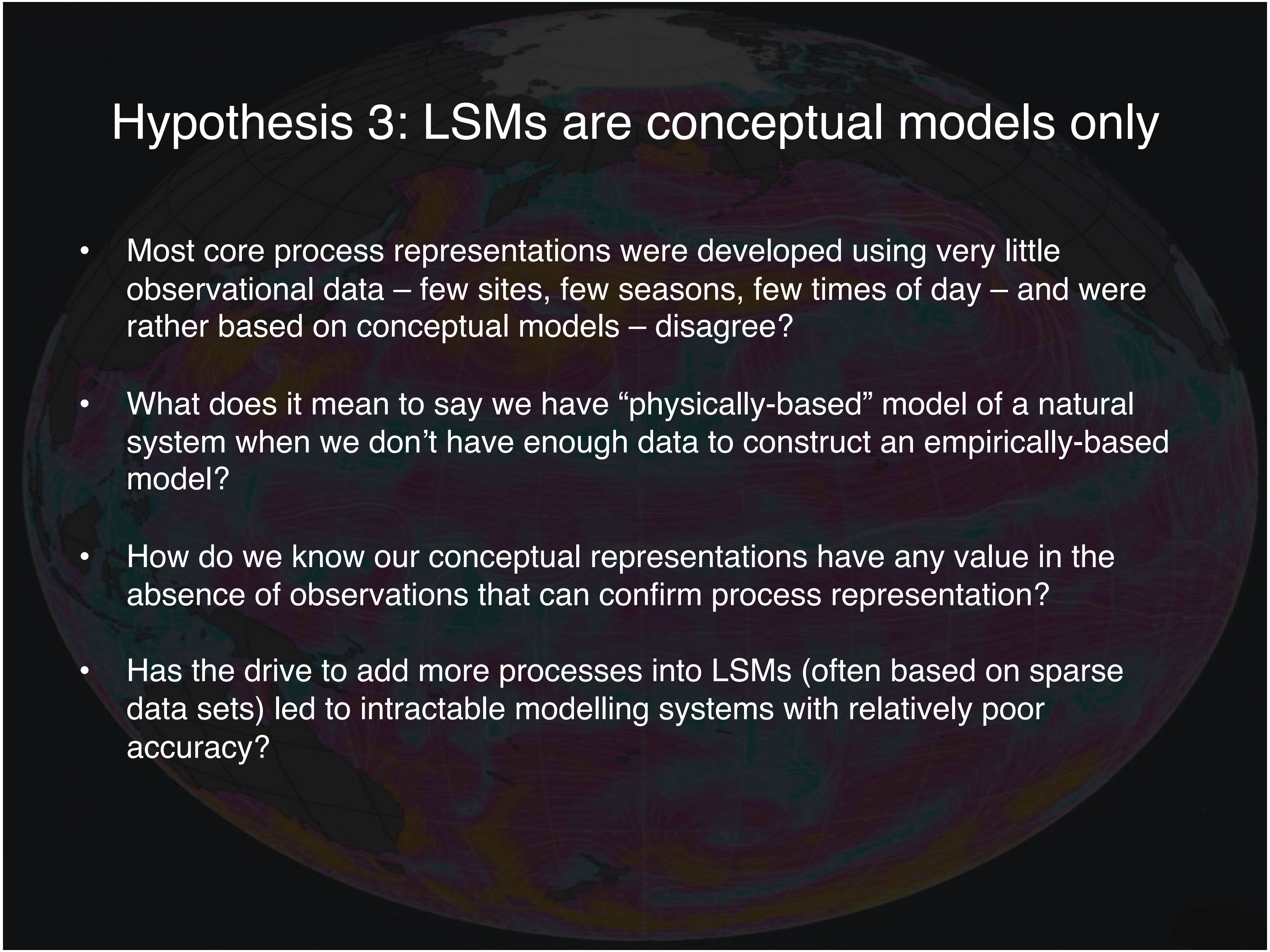

Average monthly Qle: Obs - EspirraFluxnet.1.4 Model - Espirra_C2.0_C

- More likely an issue for NEE?

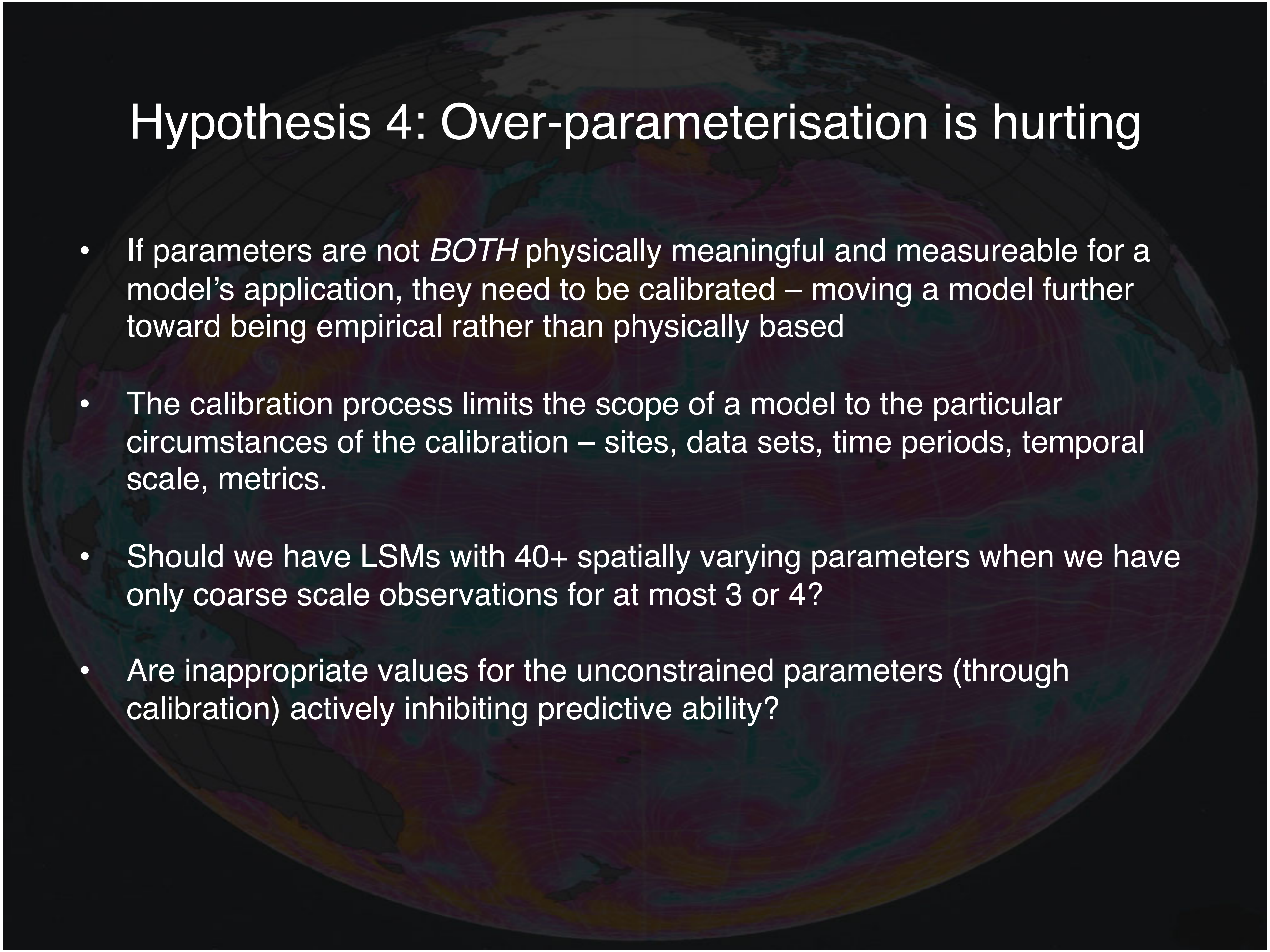
NEE	Too high	Too low	neither
JULES	1	2	17
CABLE	3	3	16


Average monthly NEE: Obs - SylvaniaFluxnet.1.4 Model - Sylvania_J3.1

Hypothesis 3: LSMs are conceptual models only

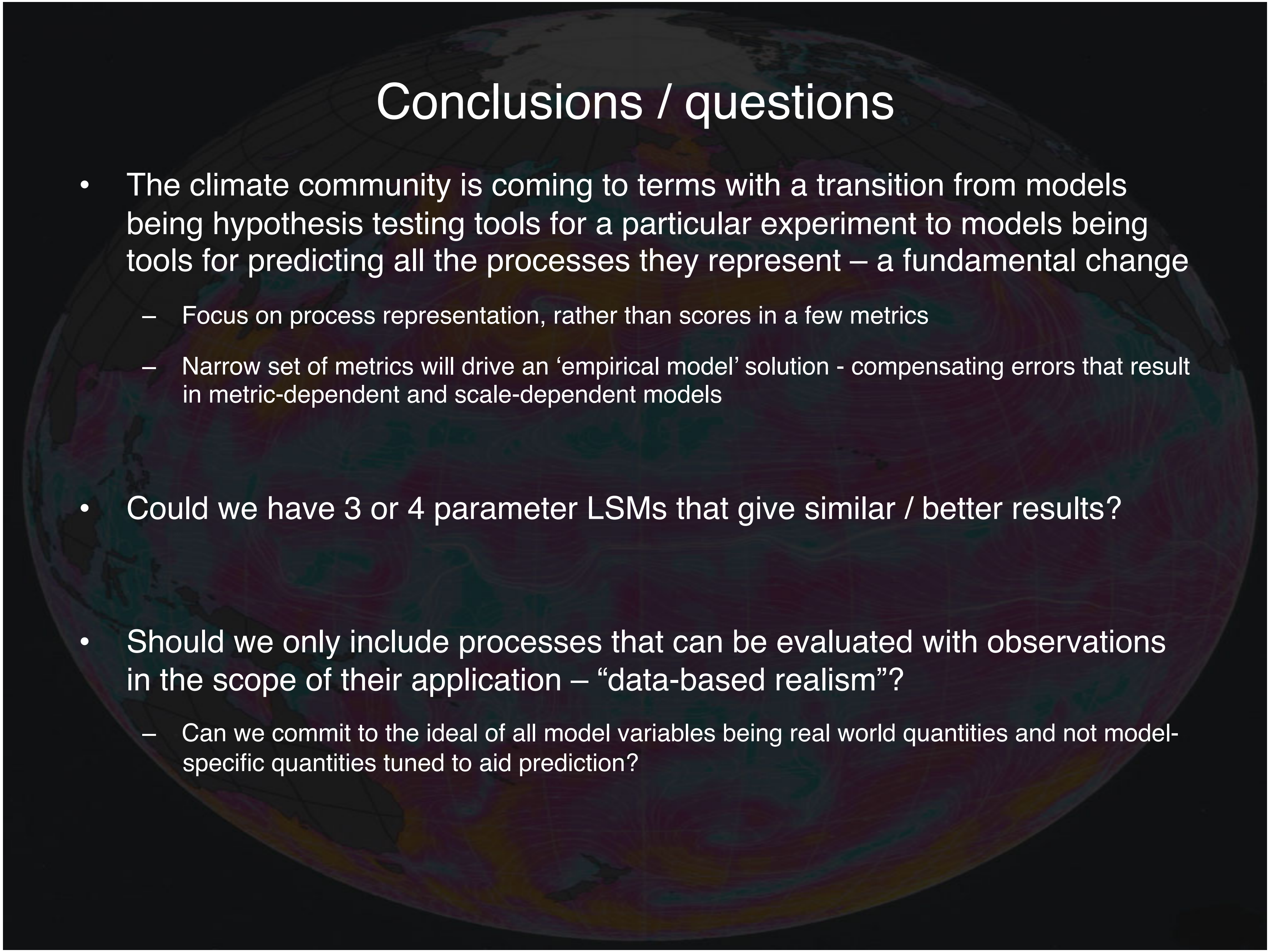


- Most core process representations were developed using very little observational data – few sites, few seasons, few times of day – and were rather based on conceptual models – disagree?
- What does it mean to say we have “physically-based” model of a natural system when we don’t have enough data to construct 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?



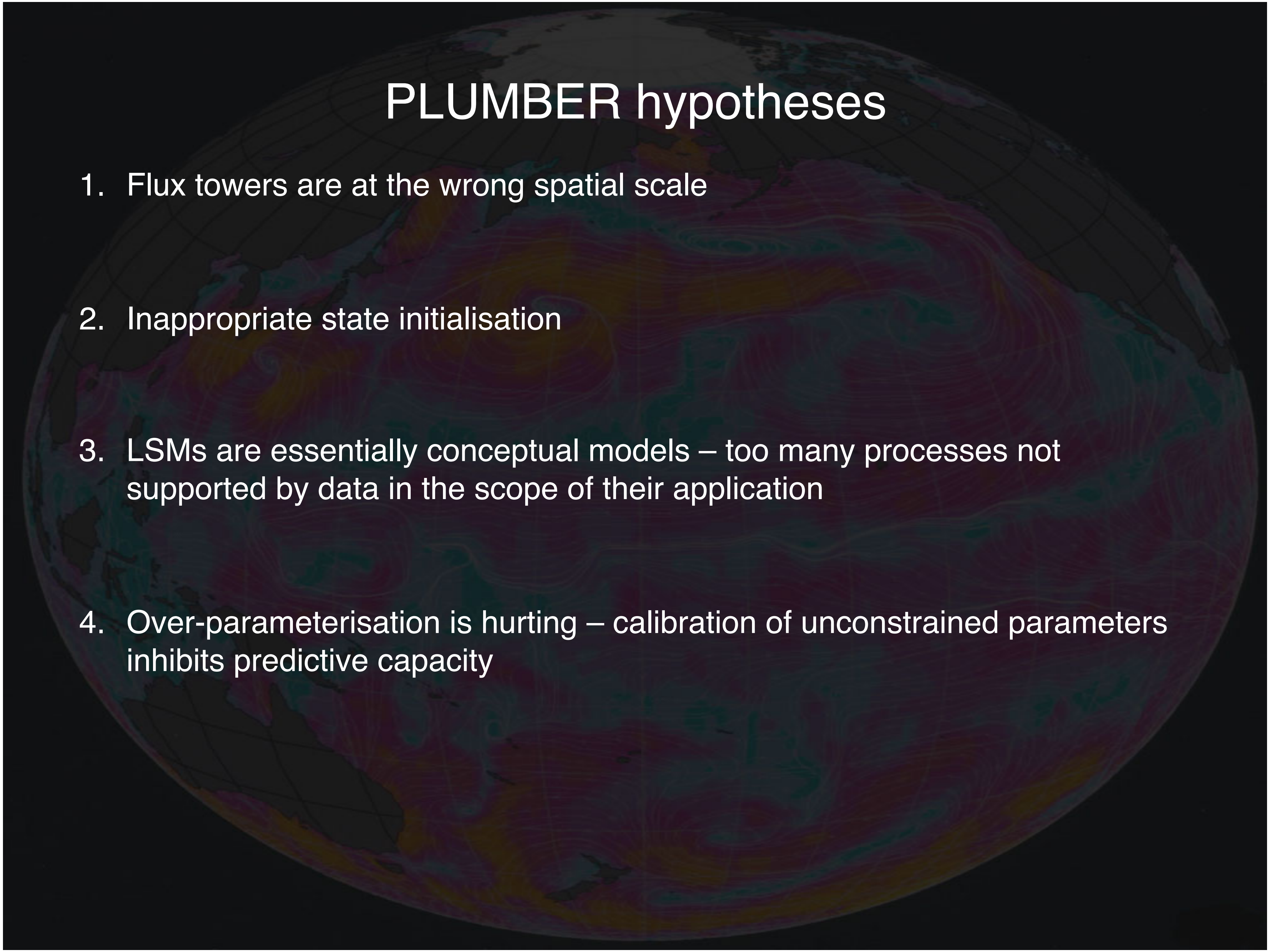
Hypothesis 4: Over-parameterisation is hurting

- If parameters are not *BOTH* physically meaningful and measureable for a model's application, they need to be calibrated – moving a model further toward being empirical rather than physically based
- The calibration process limits the scope of a model to the particular circumstances of the calibration – sites, data sets, time periods, temporal scale, metrics.
- Should we have LSMs with 40+ spatially varying parameters when we have only coarse scale observations for at most 3 or 4?
- Are inappropriate values for the unconstrained parameters (through calibration) actively inhibiting predictive ability?



Conclusions / questions

- The climate community is coming to terms with a transition from models being hypothesis testing tools for a particular experiment to models being tools for predicting all the processes they represent – a fundamental change
  - Focus on process representation, rather than scores in a few metrics
  - Narrow set of metrics will drive an ‘empirical model’ solution - compensating errors that result in metric-dependent and scale-dependent models
- Could we have 3 or 4 parameter LSMs that give similar / better results?
- Should we only include processes that can be evaluated with observations in the scope of their application – “data-based realism”?
  - Can we commit to the ideal of all model variables being real world quantities and not model-specific quantities tuned to aid prediction?



PLUMBER hypotheses

1. Flux towers are at the wrong spatial scale

2. Inappropriate state initialisation

3. LSMs are essentially conceptual models – too many processes not supported by data in the scope of their application

4. Over-parameterisation is hurting – calibration of unconstrained parameters inhibits predictive capacity