PLUMBER2 – motivation for discussion

Gab Abramowitz

The Plumbing of Land Surface Models: Benchmarking Model Performance

M. J. BEST,[a] G. ABRAMOWITZ,[b] H. R. JOHNSON,[a] A. J. PITMAN,[b] G. BALSAMO,[c] A. BOONE,[d] M. CUNTZ,[e] B. DECHARME,[d] P. A. DIRMEYER,[f] J. DONG,[g] M. EK,[g] Z. GUO,[f] V. HAVERD,[h] B. J. J. VAN DEN HURK,[i] G. S. NEARING,[j] B. PAK,[k] C. PETERS-LIDARD,[j] J. A. SANTANELLO JR.,[j] L. STEVENS,[k] AND N. VUICHARD[l]

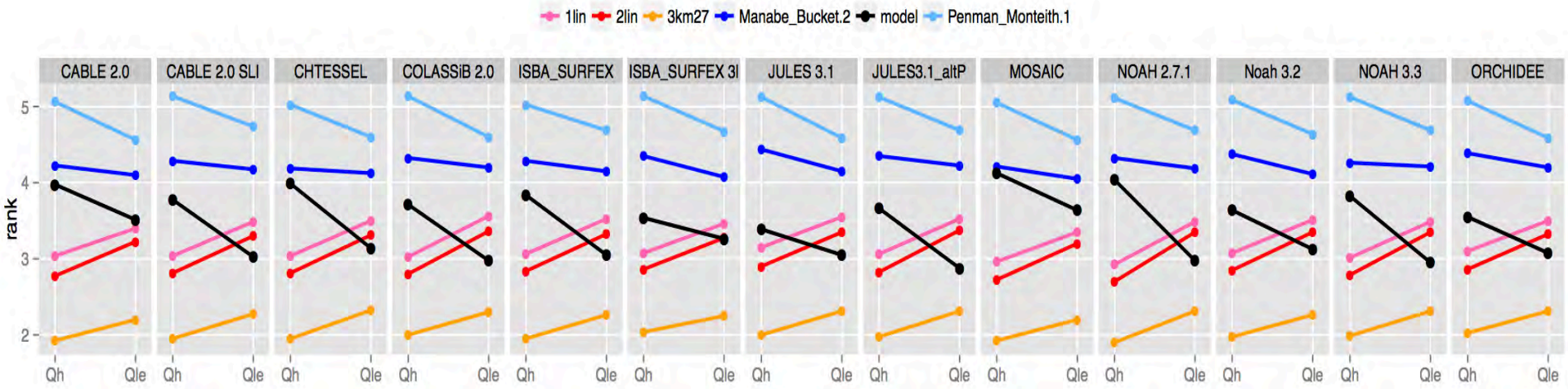
[a] *Met Office, Exeter, United Kingdom*
[b] *ARC Centre of Excellence for Climate System Science, University of New South Wales, Sydney, New South Wales, Australia*
[c] *ECMWF, Reading, United Kingdom*
[d] *CNRM-GAME, Météo-France, Toulouse, France*
[e] *Helmholtz Centre for Environmental Research–UFZ, Leipzig, Germany*
[f] *Center for Ocean–Land–Atmosphere Studies, George Mason University, Fairfax, Virginia*
[g] *NOAA/NCEP/EMC, College Park, Maryland*
[h] *Oceans and Atmosphere Flagship, CSIRO, Canberra, Australian Capital Territory, Australia*
[i] *KNMI, De Bilt, Netherlands*
[j] *Hydrological Sciences Laboratory, NASA GSFC, Greenbelt, Maryland*
[k] *Oceans and Atmosphere Flagship, CSIRO, Aspendale, Victoria, Australia*
[l] *Laboratoire des Sciences du Climat et de l'Environnement, UMR 8212, IPSL-LSCE, CEA-CNRS-UVSQ, Gif-sur-Yvette, France*

(Manuscript received 27 August 2014, in final form 19 December 2014)

ABSTRACT

The Protocol for the Analysis of Land Surface Models (PALS) Land Surface Model Benchmarking Evaluation Project (PLUMBER) was designed to be a land surface model (LSM) benchmarking intercomparison. Unlike the traditional methods of LSM evaluation or comparison, benchmarking uses a fundamentally different

PLUMBER results



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

- 20 Flux tower sites – 9 IGBP vegetation types;
- 4 metrics: bias, correlation, SD, normalised mean error

- 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

The Plumbing of Land Surface Models: Is Poor Performance a Result of Methodology or Data Quality?

NED HAUGHTON,[a] GAB ABRAMOWITZ,[a] ANDY J. PITMAN,[a] DANI OR,[b] MARTIN J. BEST,[c] HELEN R. JOHNSON,[c] GIANPAOLO BALSAMO,[d] AARON BOONE,[e] MATTHIAS CUNTZ,[f] BERTRAND DECHARME,[e] PAUL A. DIRMEYER,[g] JAIRUI DONG,[h] MICHAEL EK,[h] ZICHANG GUO,[g] VANESSA HAVERD,[i] BART J. J. VAN DEN HURK,[j] GREY S. NEARING,[k] BERNARD PAK,[l] JOE A. SANTANELLO JR.,[k] LAUREN E. STEVENS,[l] AND NICOLAS VUICHARD[m]

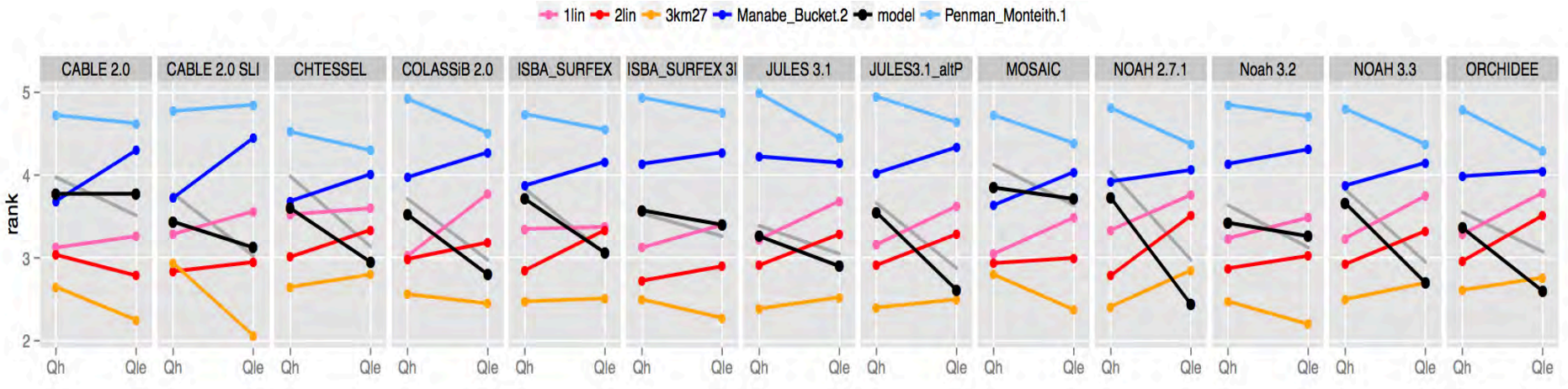
[a] *ARC Centre of Excellence for Climate Systems Science, Sydney, New South Wales, Australia*
[b] *Department of Environmental Systems Science, ETH Zurich, Zurich, Switzerland*
[c] *Met Office, Exeter, United Kingdom*
[d] *ECMWF, Reading, United Kingdom*
[e] *CNRM-GAME, Météo-France, Toulouse, France*
[f] *Helmholtz Centre for Environmental Research (UFZ), Leipzig, Germany*
[g] *Center for Ocean–Land–Atmosphere Studies, George Mason University, Fairfax, Virginia*
[h] *NOAA/NCEP/EMC, College Park, Maryland*
[i] *Oceans and Atmosphere, CSIRO, Canberra, Australian Capital Territory, Australia*
[j] *Royal Netherlands Meteorological Institute (KNMI), De Bilt, Netherlands*
[k] *Hydrological Sciences Laboratory, NASA GSFC, Greenbelt, Maryland*
[l] *Oceans and Atmosphere, CSIRO, Aspendale, Victoria, Australia*
[m] *Laboratoire des Sciences du Climat et de l'Environnement, UMR 8212, IPSL-LSCE, CEA-CNRS-UVSQ, Gif-sur-Yvette, France*

(Manuscript received 16 September 2015, in final form 27 January 2016)

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?

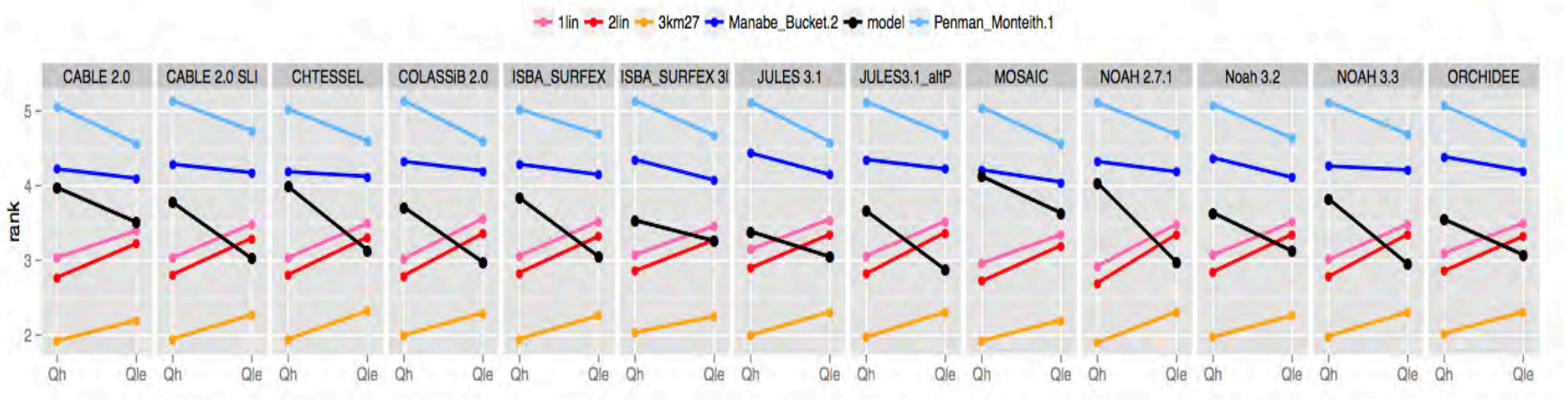
PLUMBER results – why? Not energy conservation.

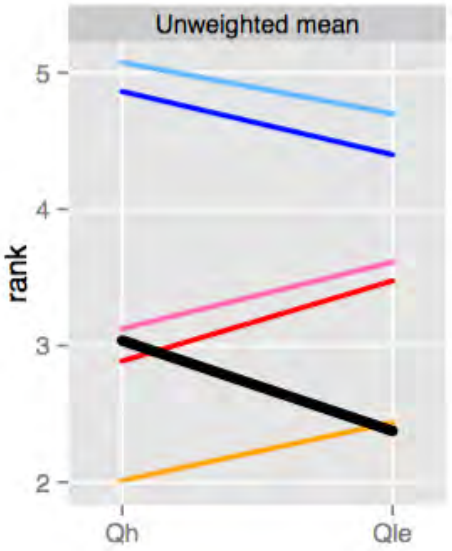


- Constrain each empirical model to have the same sum of (latent + sensible) heat flux as the LSM at every time step
  - Each empirical model then effectively has the same Rnet and ground heat flux as the LSM it's being compared to – and conserves energy.
- Results are mixed but the regression against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

Haughton et al, 2016

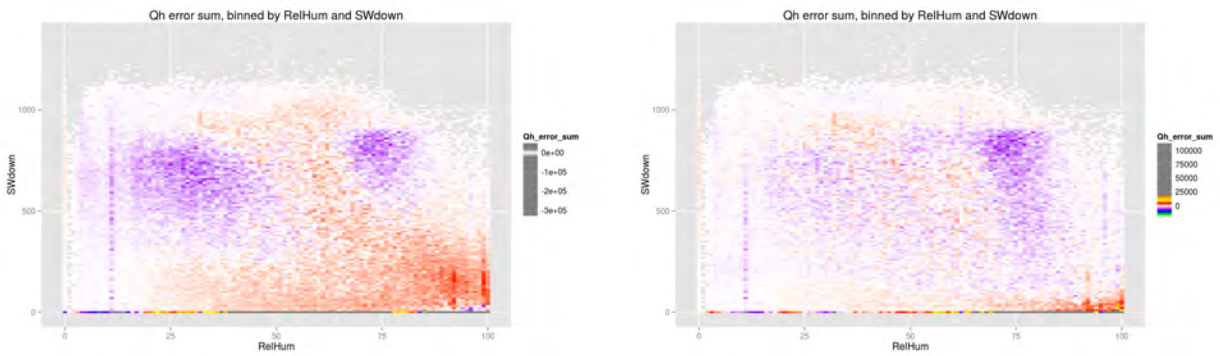
PLUMBER results – shared model issues?



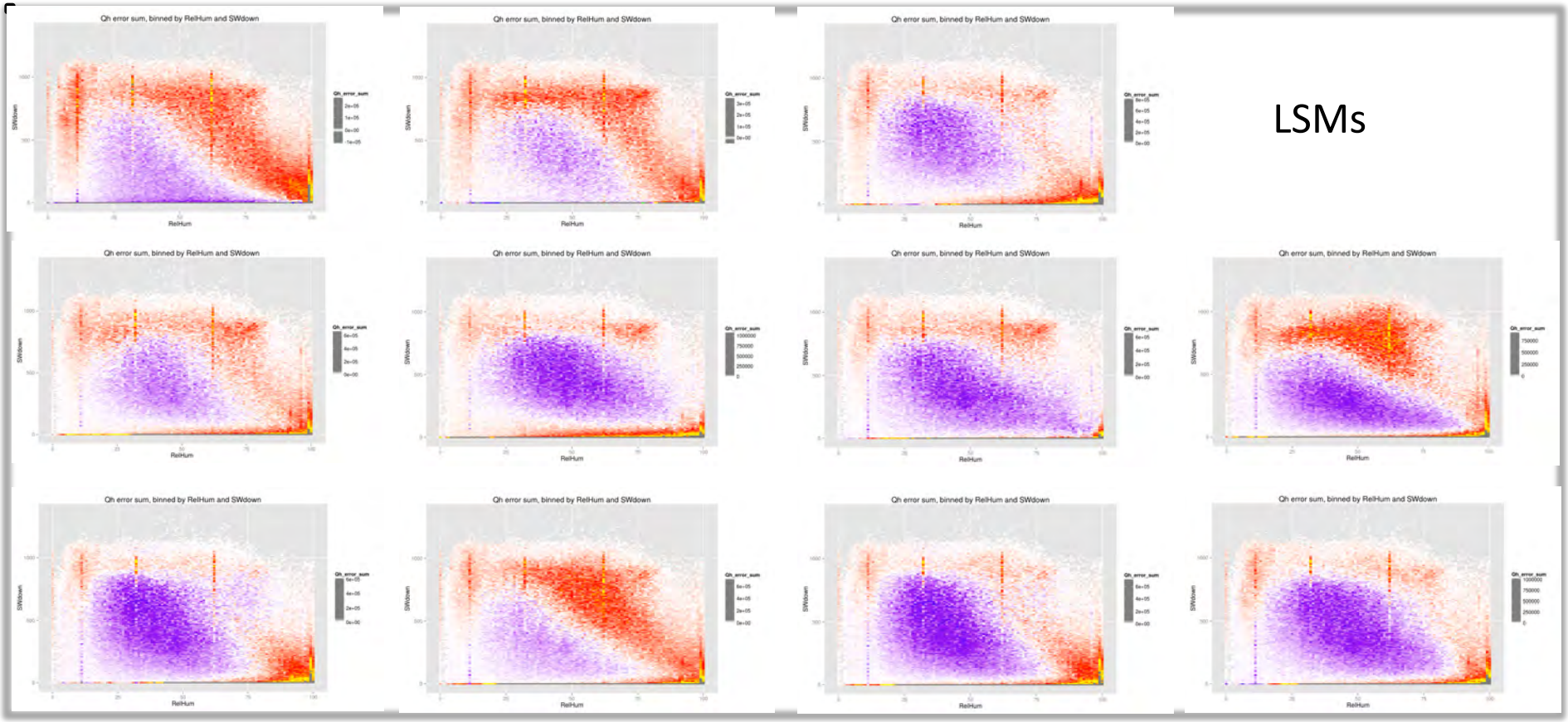


Haughton et al, 2016

Ned Haughton: PLUMBER results – shared model issues?

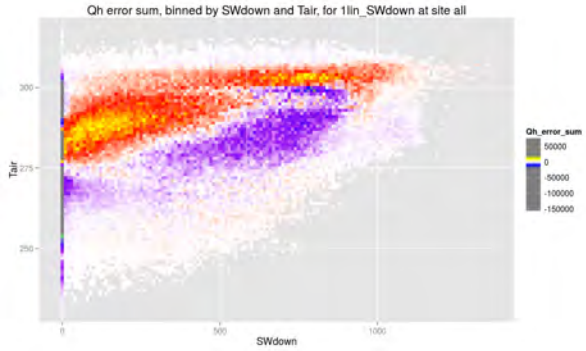


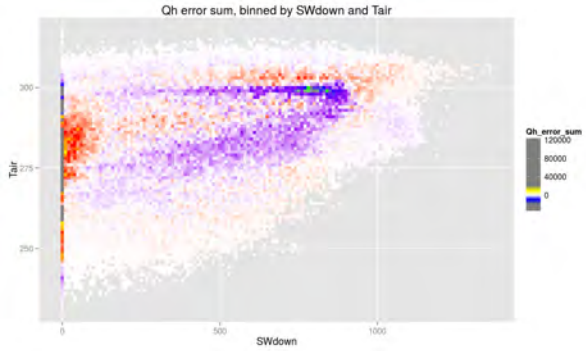
Qh error, binned by (RelHum, SWdown,)



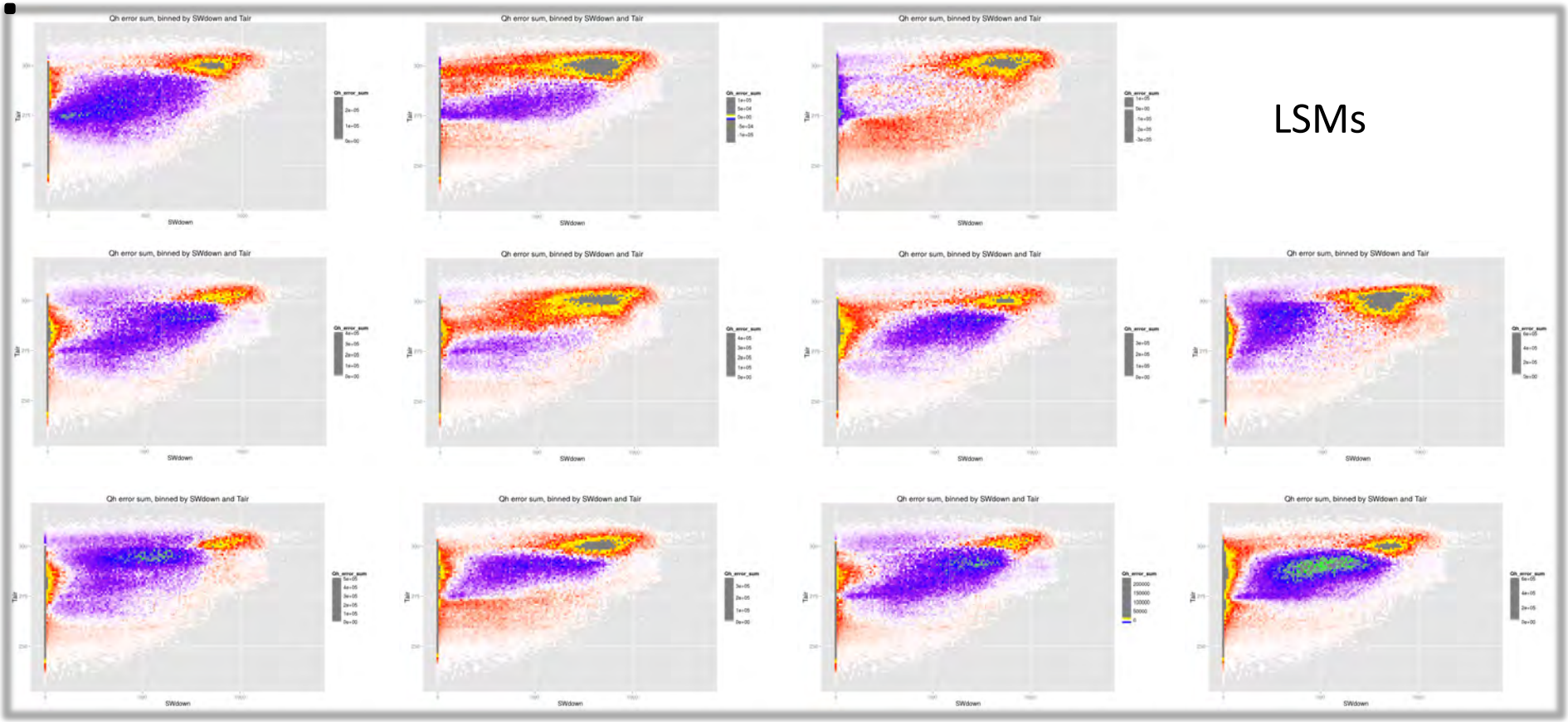
LSMs

Ned Haughton: PLUMBER results – shared model issues?





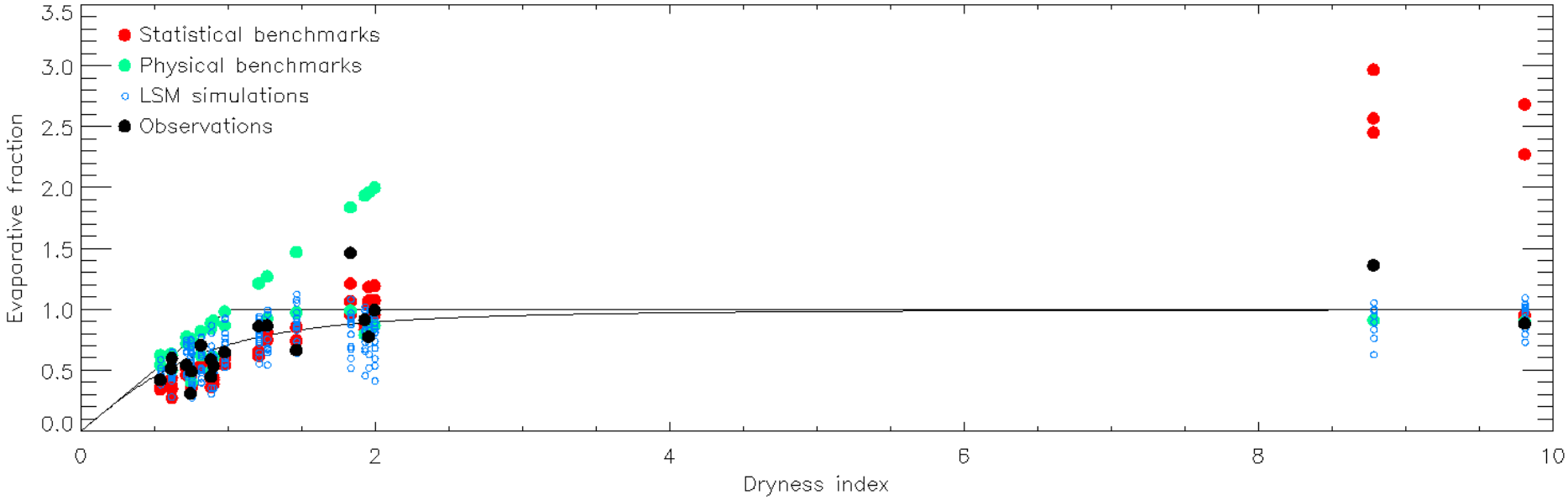
Qh error, binned by (Swdown, Tair)



LSMs

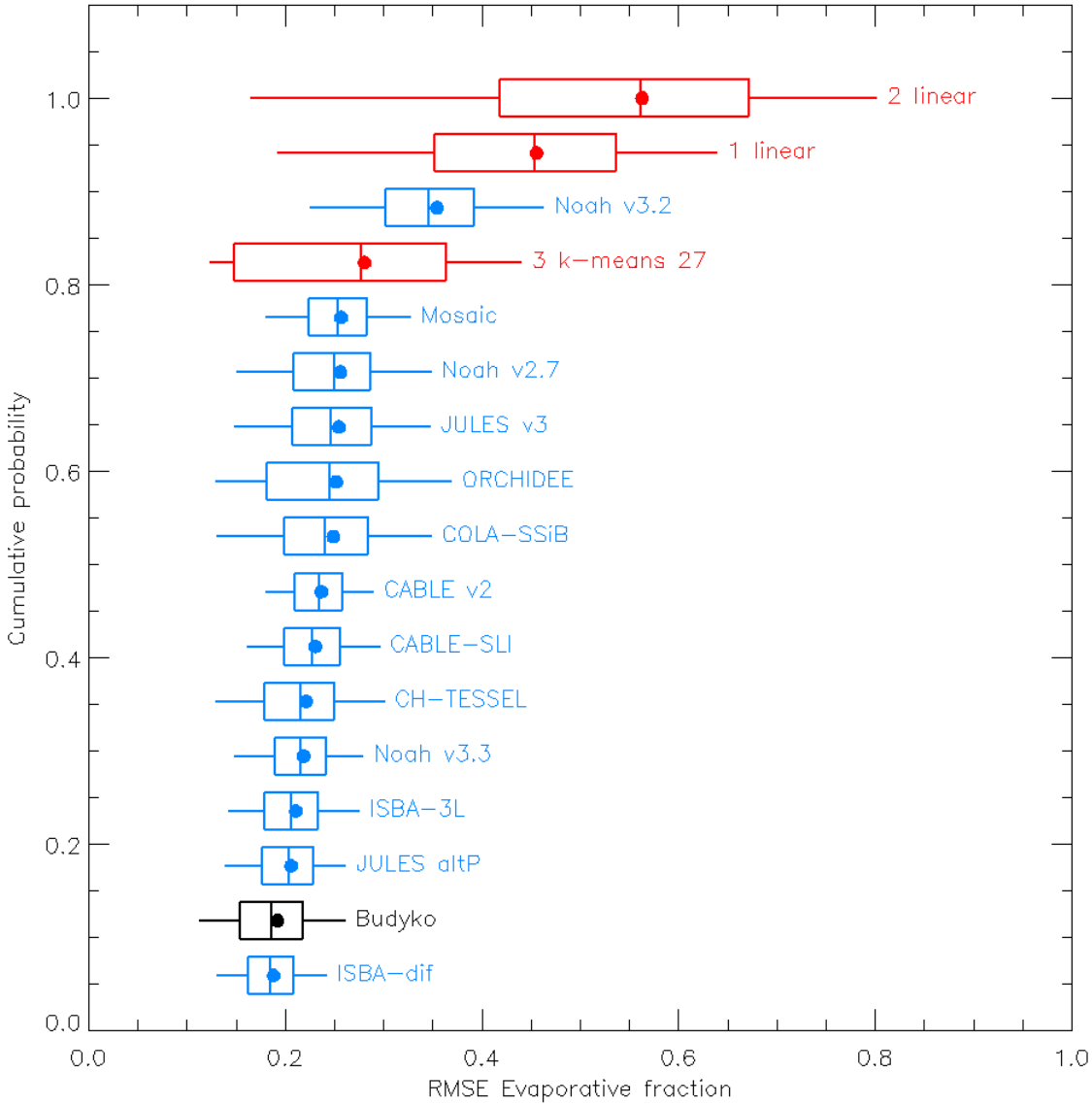
Martyn Clark:
PLUMBER models within a Budyko framework

- The Budyko framework examines how the dryness index (PET/P) affects the evaporative fraction (ET/P).
- The statistical models tend to be lower than the Budyko curve for the wetter sites and higher than the Buyko curves for the drier sites.
- At drier sites the statistical models can have ET greater than P (i.e., an evaporative fraction greater than 1).

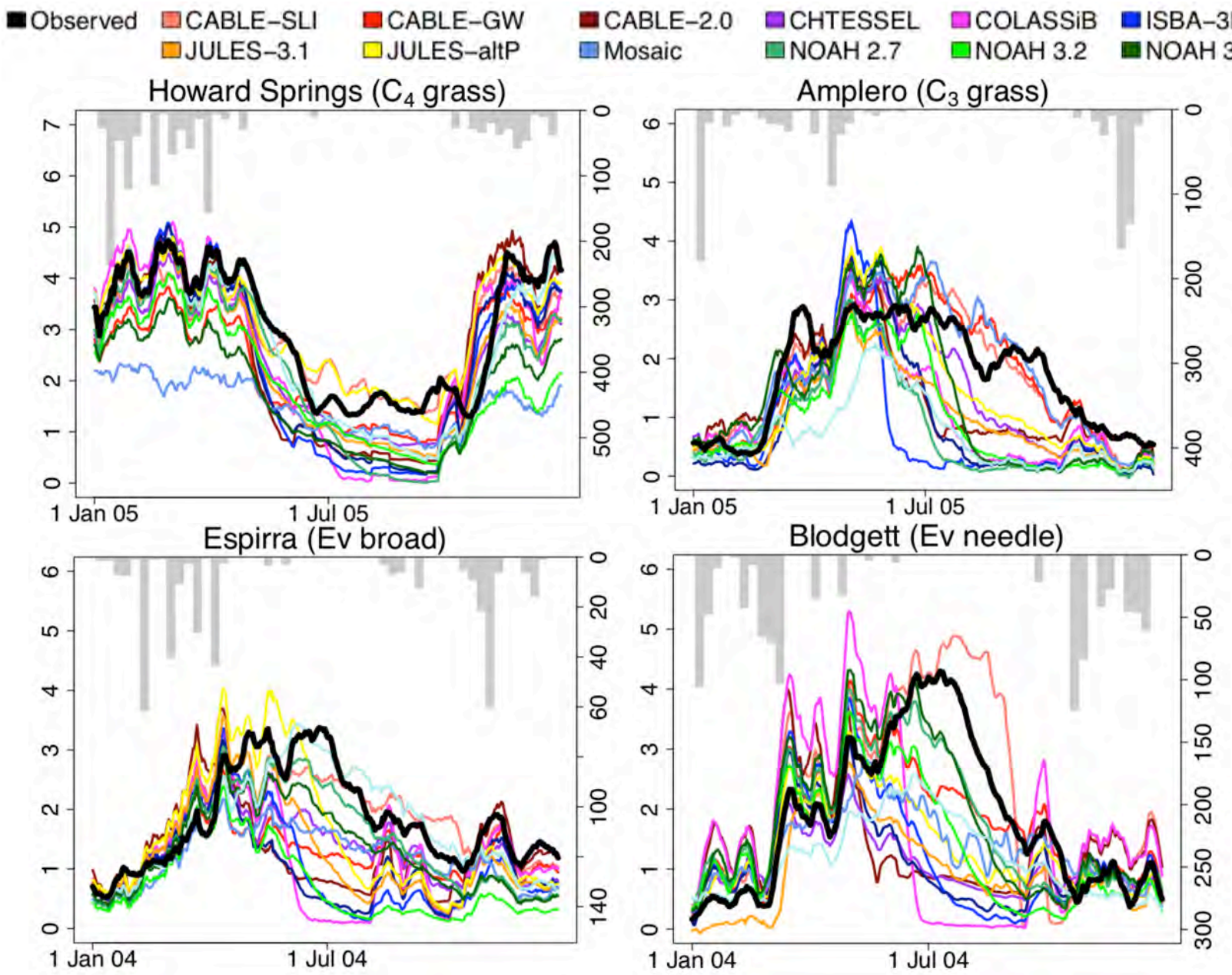


Martyn Clark:
PLUMBER models within a Budyko framework

- Approach
  - RMSE across the 20 fluxnet sites
  - Impact of the small sample size is characterized by resampling the sites (with replacement) 1000 times
- Results
  - Most of the land models actually outperform the statistical models.
  - The Budyko curve provides better predictions than most of the land models, suggesting that the land models are incapable of predicting departures from the Budyko curve.
- The conclusions of PLUMBER still hold, with a simple model (Budyko) outperforming most land models.



Dry-down events at PLUMBER sites (Anna Ukkola)



Ukkola et al, 2016, ERL

Why do it again – what could we improve?

- More, better quality controlled sites
- Energy-balance corrected site data
- Improved hierarchy of empirical model to benchmark against
  - Energy and mass conservation in empirical models
- Report more variables so process representation differences in models can be explored
- Look at sites that have some boundary layer data and run with SCMs? (i.e. compensating biases could be the cause)
- …..more?

Ned Haughton: a hierarchy of better empirical models

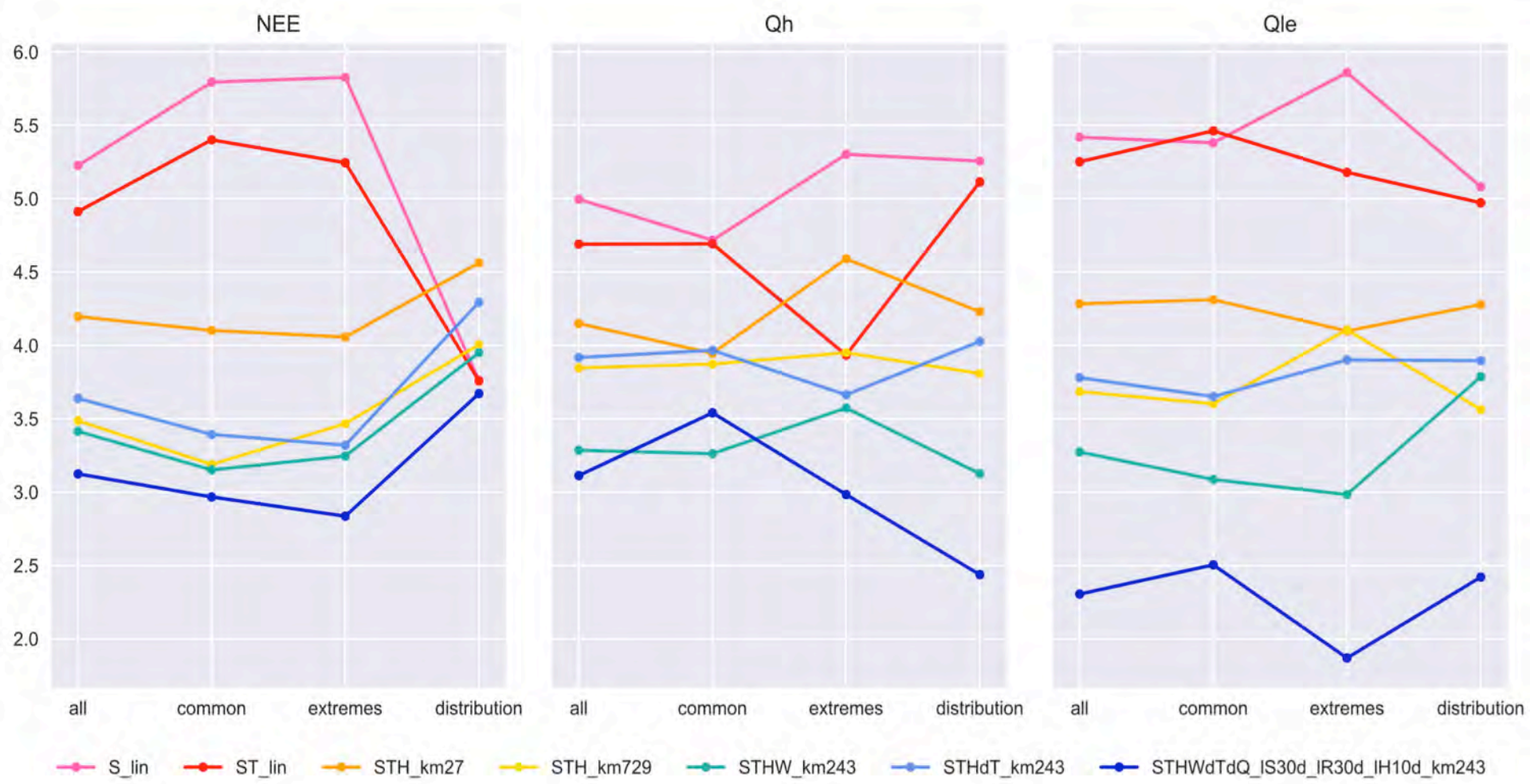


Figure 5: Rank-average plot of the 7 models in the final ensemble.

Options for experimental protocol

- Site selection:
  - FLUXNET2015: ~150 already with QC for PALS release
  - maybe including some with boundary layer data for SCM comparison?
- How much to prescribe, versus leave as LSM default?
  - Prescribe: vegetation type, reference height
  - Soil type, veg height, (+schemes for types – mapping to internal parameters)
- Initialisation? Carbon?
- LAI – progostic vs prescribed? Where do values come from?
- I/O protocol: Hyungjun's ALMA update
  - Can we add structural assumptions and perhaps parameter values to this?
- Can we run this all through PALS?
- Extend methodology to UrbanMIP?