The PALS Land sUrface Model Benchmarking Evaluation pRoject

(PLUMBER)

Summary and ongoing work

Gab Abramowitz

UNSW Sydney & ARCCSS

Martin Best, Ned Haughton, Martyn Clark, Anna Ukkola, Andy Pitman, Dani Or

+ other PLUMBER coauthors: 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 AUSTRALIA

ARC CENTRE OF EXCELLENCE FOR CLIMATE SYSTEM SCIENCE

TERN Terrestrial Ecosystem Research Network

GLOBAL LAND ATMOSPHERE SYSTEM STUDY GLASS

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*

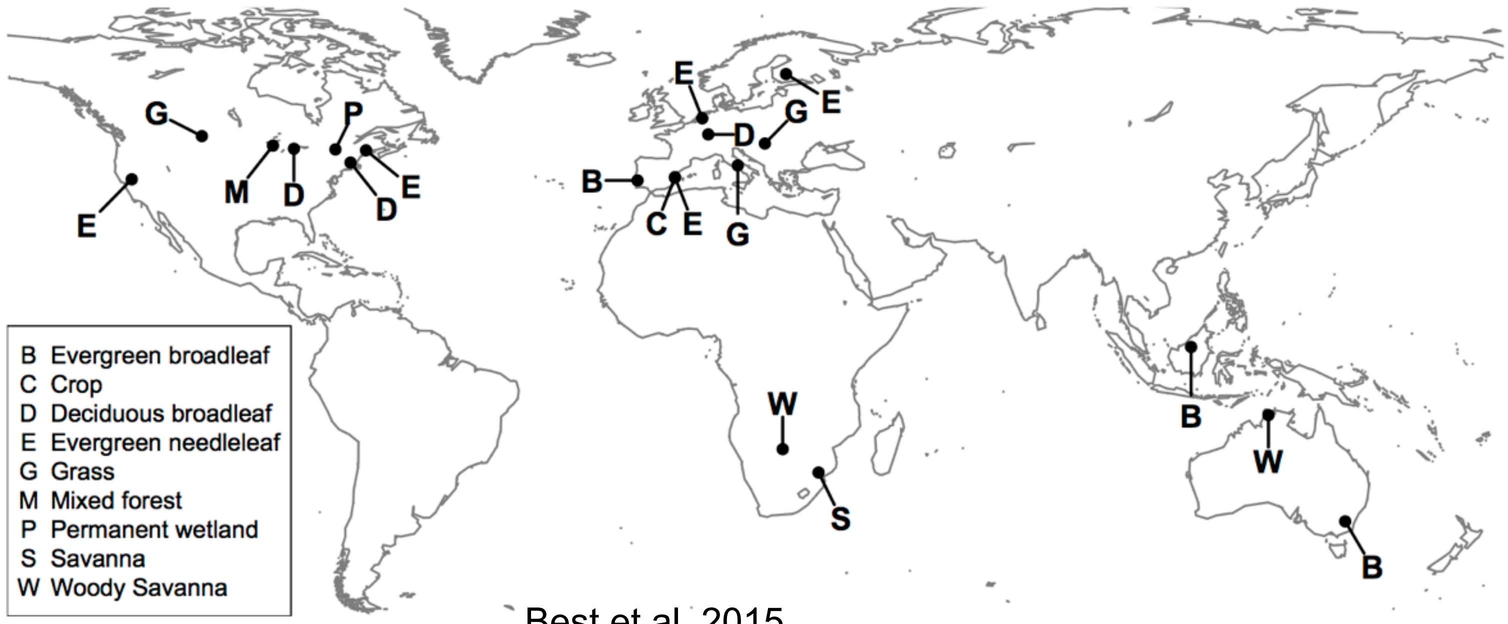
(Manuscript received 27 August 2014, in final form 19 December 2014)

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

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
- Benchmarks: two ‘physical’ – PM and Manabe bucket; 3 empirical



Best et al, 2015

The three empirical benchmarks in PLUMBER

- All 3 empirical models relate met forcing and a flux

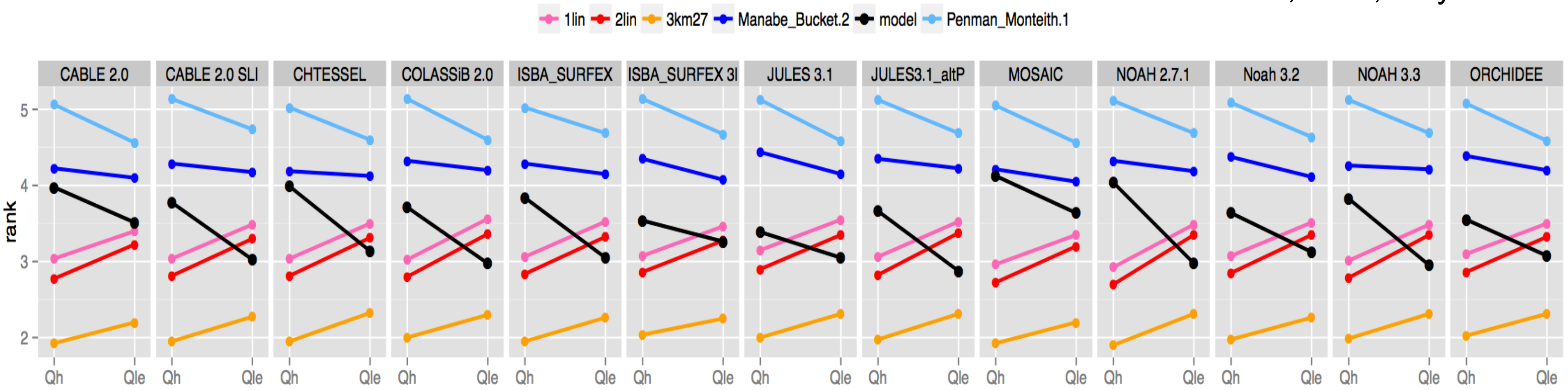
- Trained with data from sites other than the testing site (i.e. out of sample)

- 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.

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 – 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]

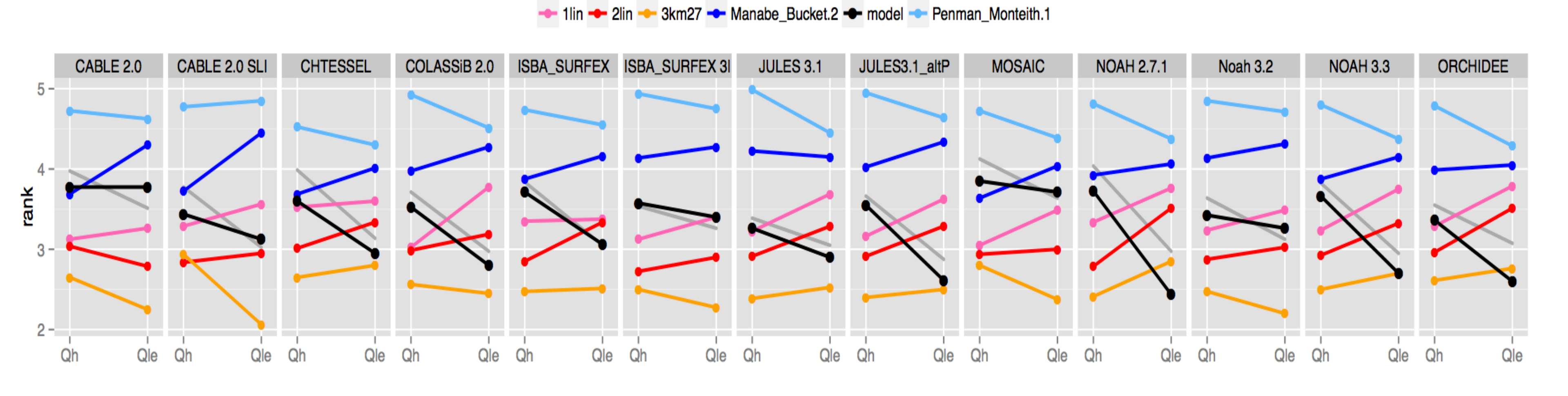
[a] *ARC Centre of Excellence for Climate Systems Science, Sydney, New South Wales, Australia*
[b] *Department of Environmental Systems Science, ETH Zurich, Zurich, Switzerland*
[c] *Met Office, Exeter, United Kingdom*
[d] *ECMWF, Reading, United Kingdom*
[e] *CNRM-GAME, Météo-France, Toulouse, France*
[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?

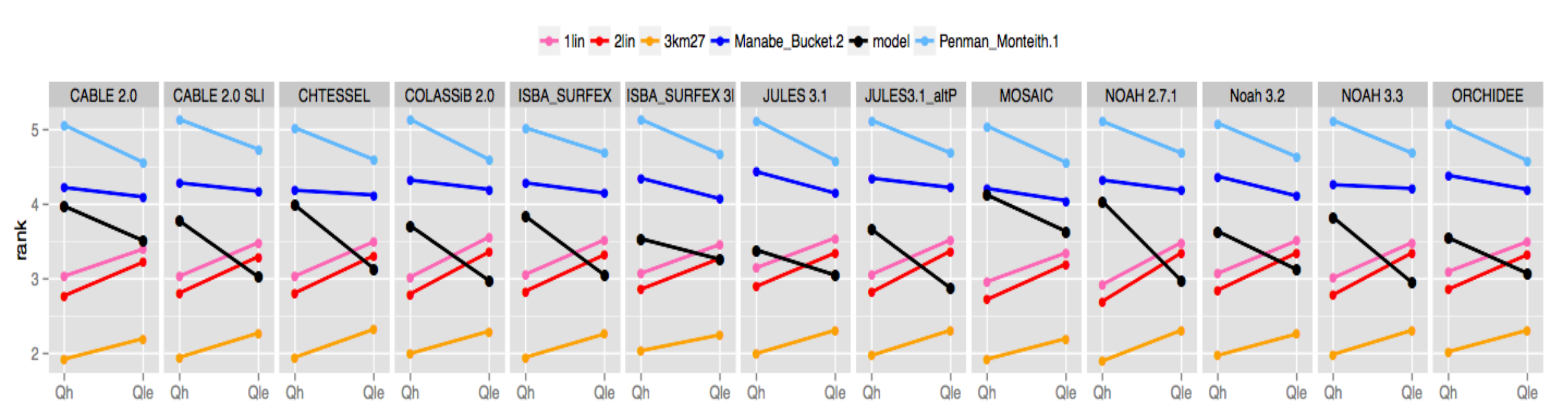
PLUMBER results – why? Not energy conservation.

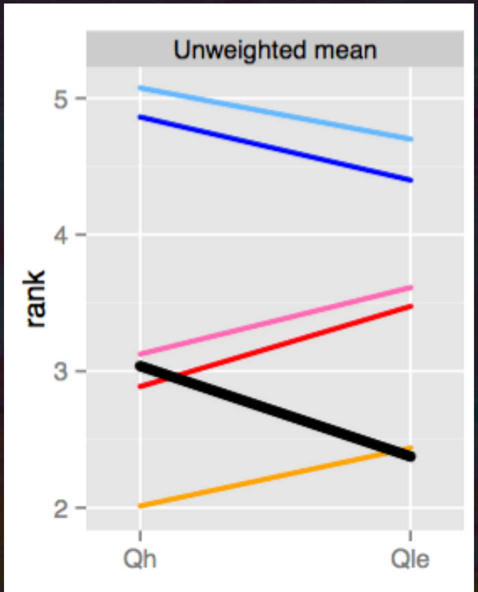


- Constrain each empirical model to have the same sum of (latent + sensible) heat flux as the LSM at every time step
  - Each empirical model then effectively has the same Rnet and ground heat flux as the LSM it's being compared to – and conserves energy.
- Results are mixed but the regression against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

Haughton et al, 2016

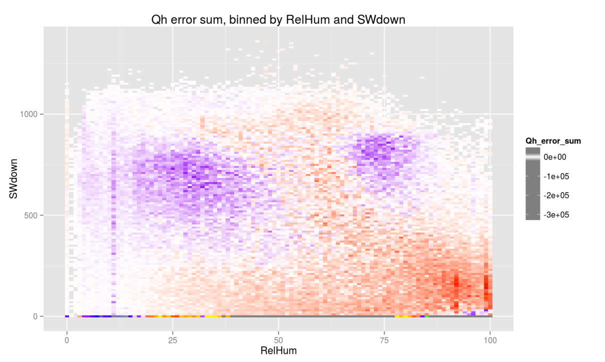
PLUMBER results – shared model issues?

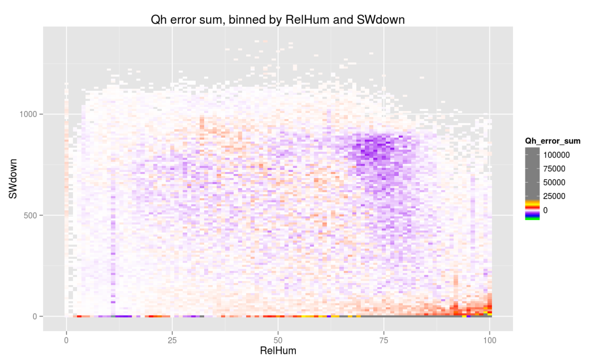




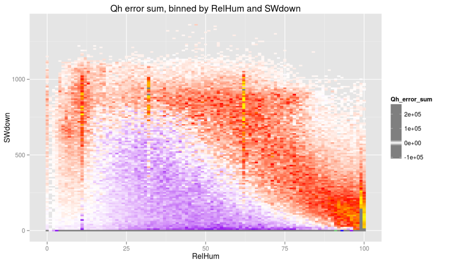
Haughton et al, 2016

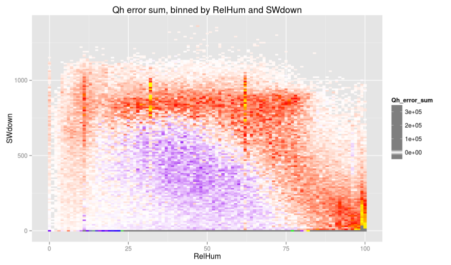
PLUMBER results – shared model issues?

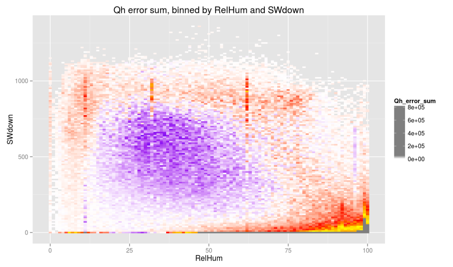




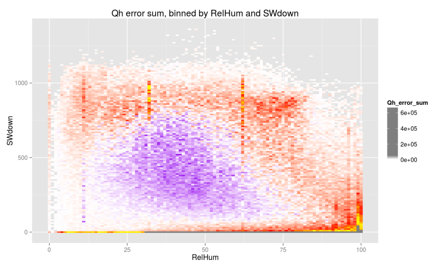
Qh error, binned by (RelHum, SWdown,)

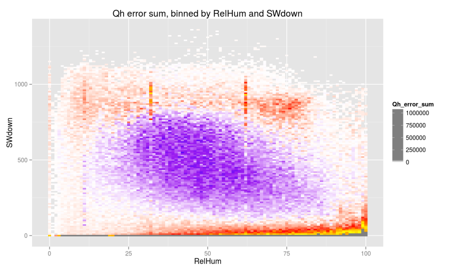


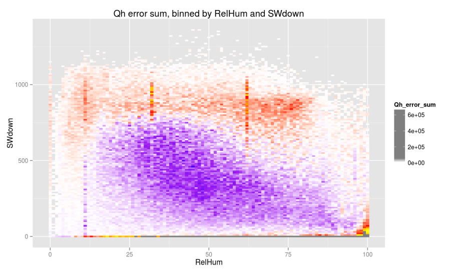


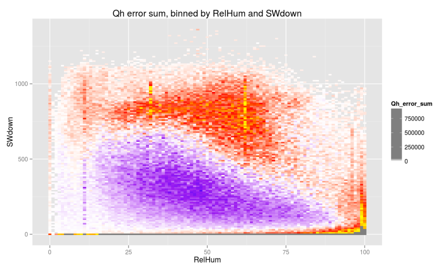


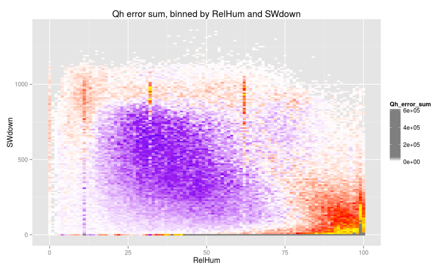
LSMs

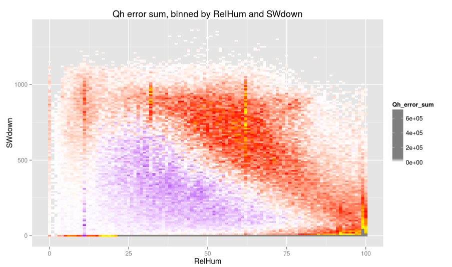


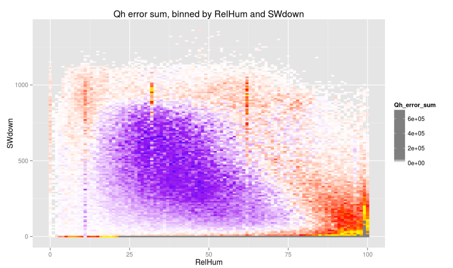


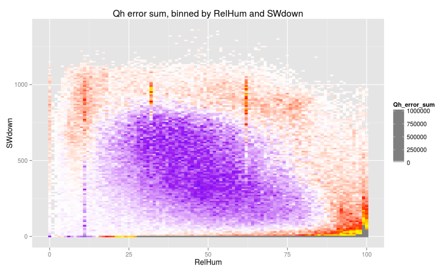




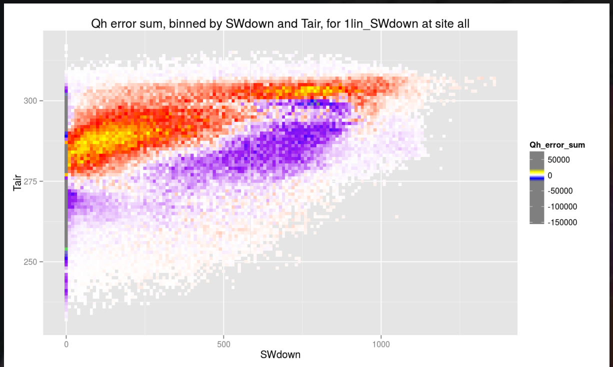


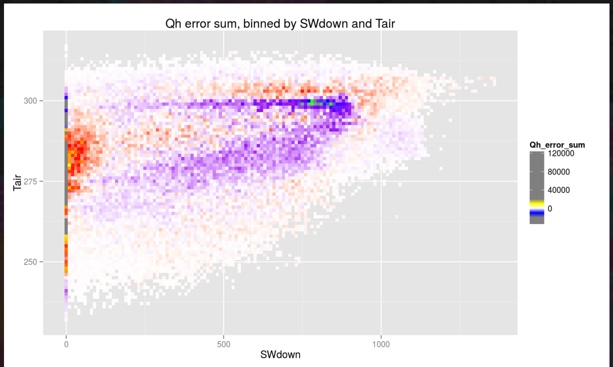




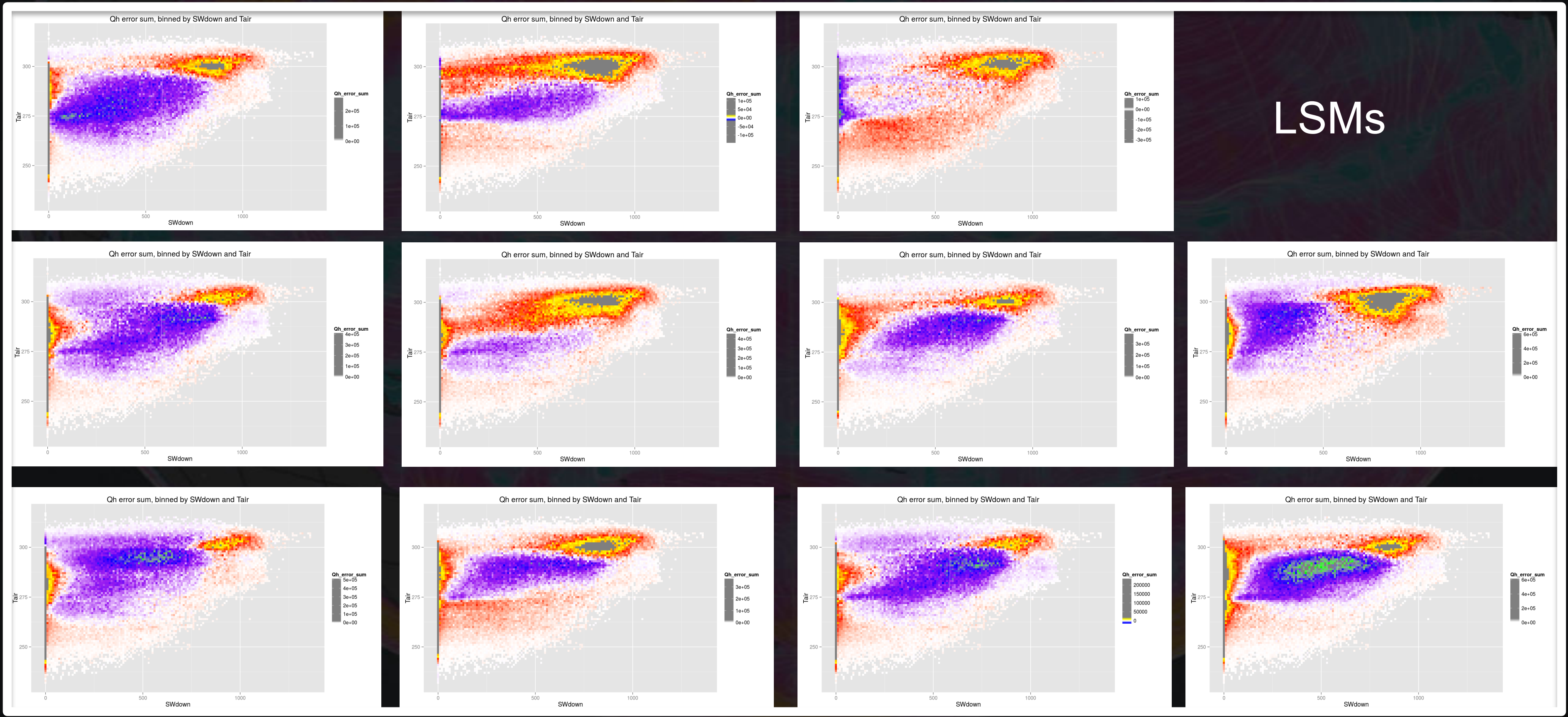


PLUMBER results – shared model issues?


Qh error sum, binned by SWdown and Tair, for 1lin_SWdown at site all


Qh error sum, binned by SWdown and Tair

Qh error, binned by (Swdown, Tair)


LSMs

Qh error sum, binned by SWdown and Tair

Qh error sum, binned by SWdown and Tair

Qh error sum, binned by SWdown and Tair

Qh error sum, binned by SWdown and Tair

Qh error sum, binned by SWdown and Tair

Qh error sum, binned by SWdown and Tair

Qh error sum, binned by SWdown and Tair

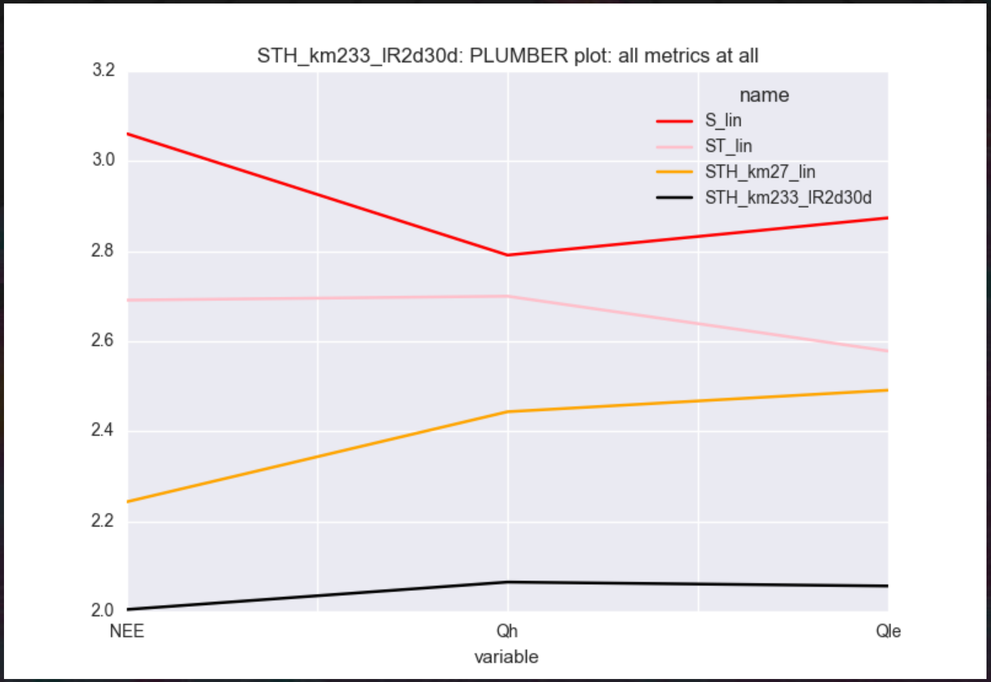
Qh error sum, binned by SWdown and Tair

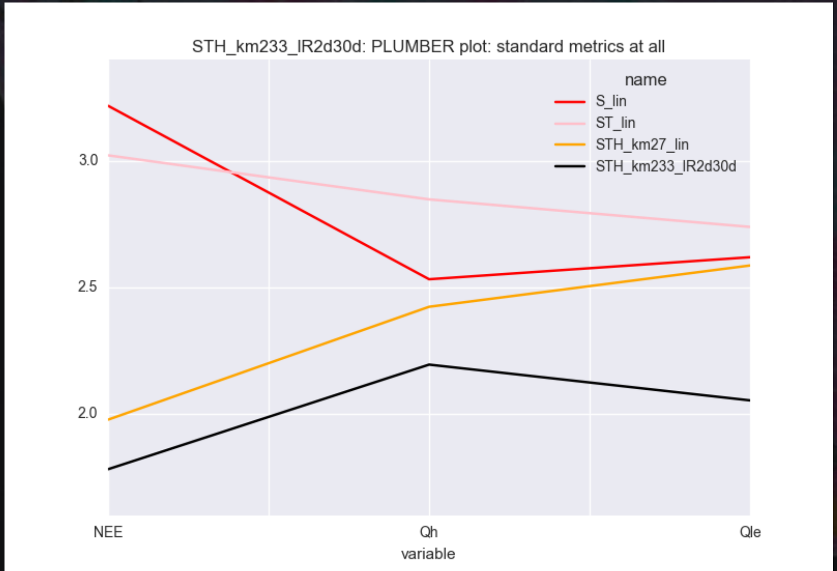
Qh error sum, binned by SWdown and Tair

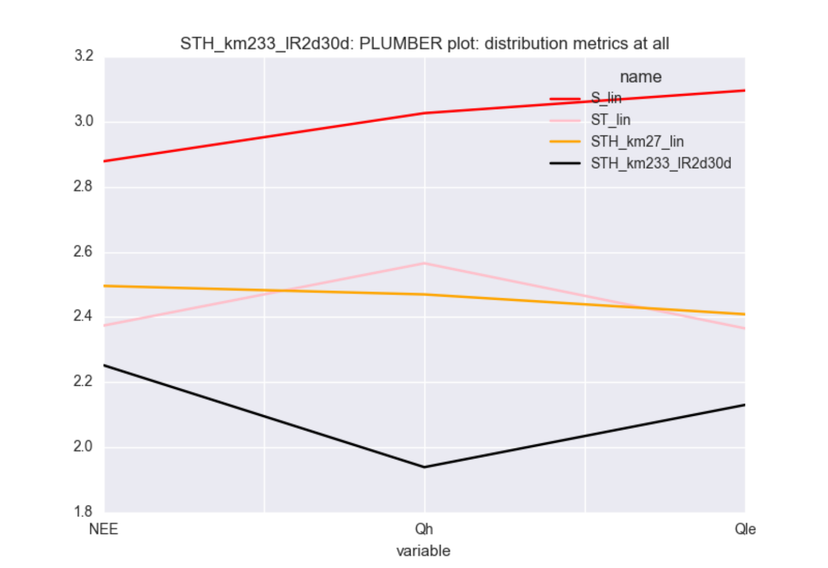
Qh error sum, binned by SWdown and Tair

Qh error sum, binned by SWdown and Tair

Can we build a better empirical model?





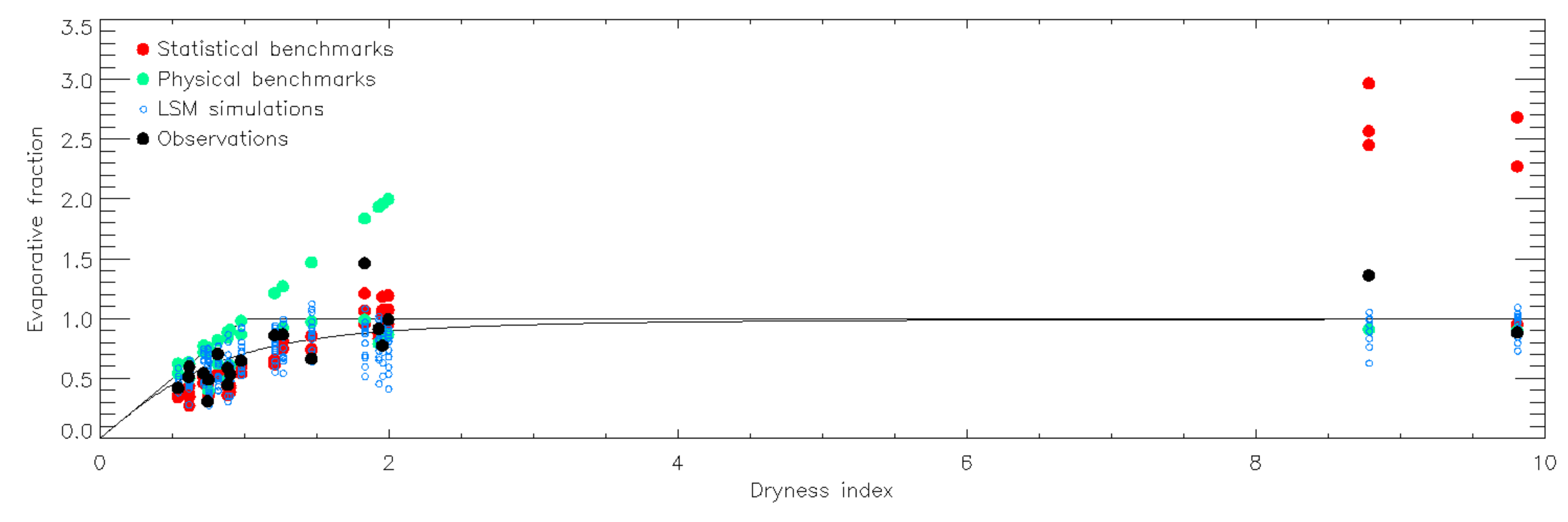


Martyn Clark: PLUMBER models within a Budyko framework

- The Budyko framework examines how the dryness index (PET/P) affects the evaporative fraction (ET/P).

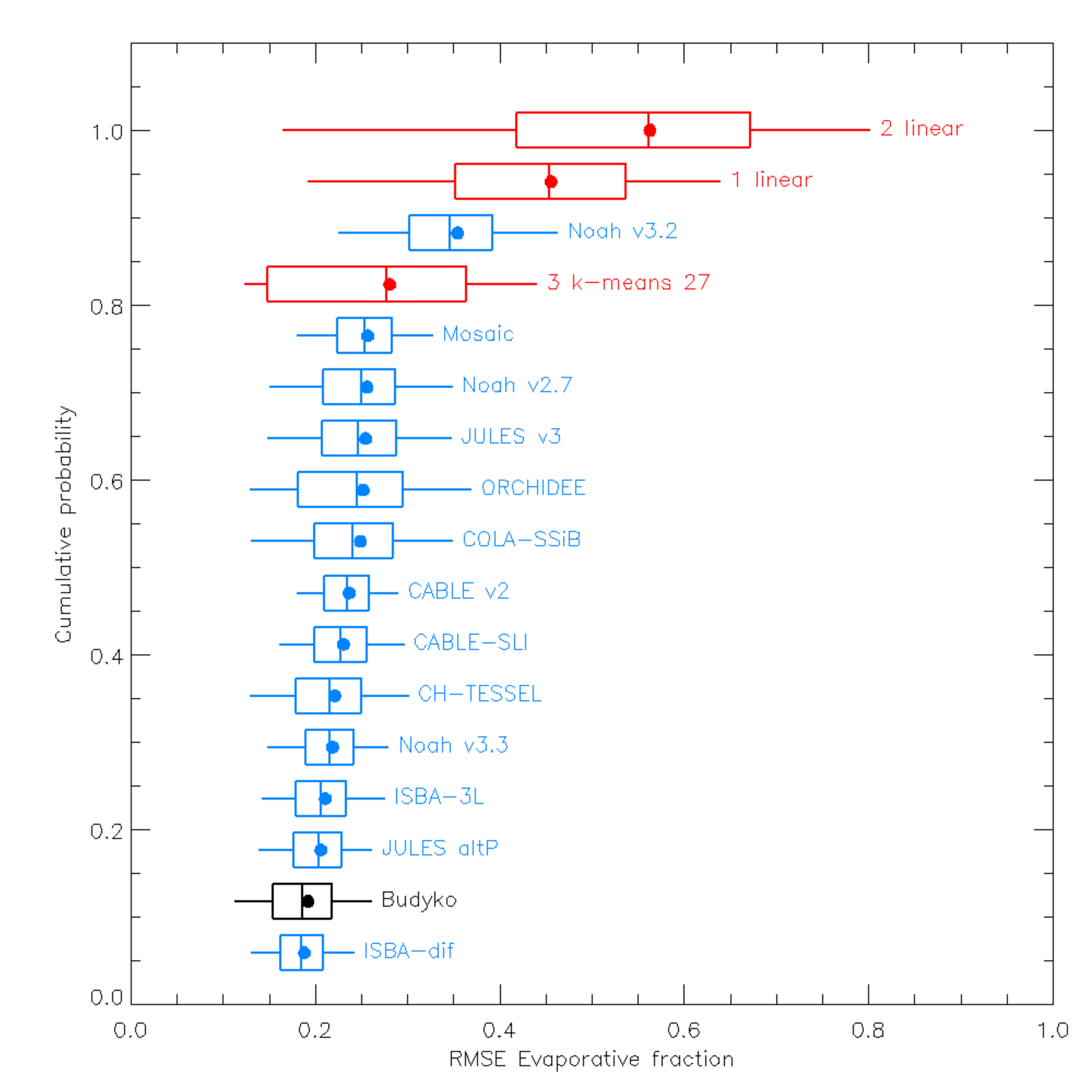
- The statistical models tend to be lower than the Budyko curve for the wetter sites and higher than the Buyko curves for the drier sites.

- At drier sites the statistical models can have ET greater than P (i.e., an evaporative fraction greater than 1).

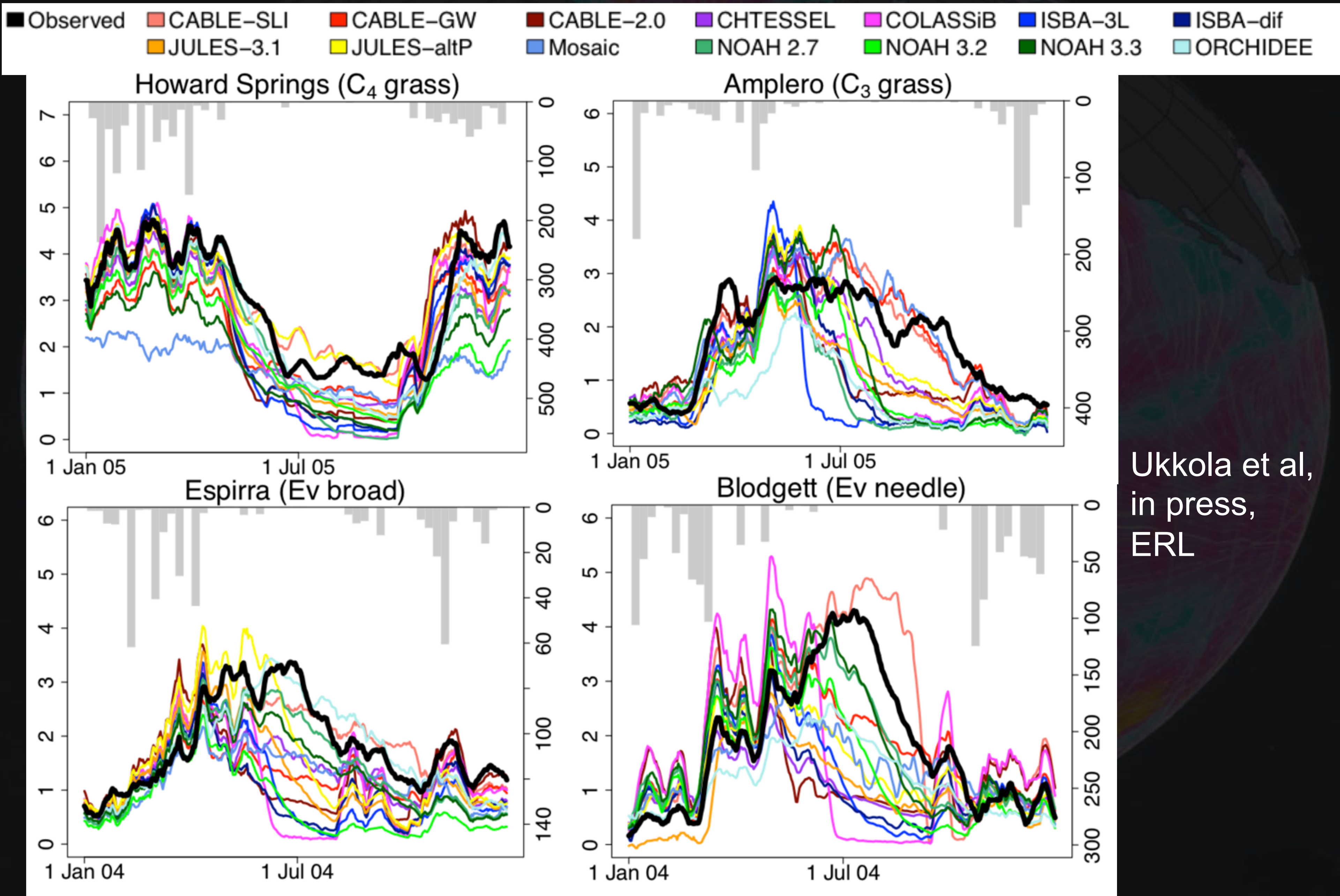


Martyn Clark:
PLUMBER models within a Budyko framework

- 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.

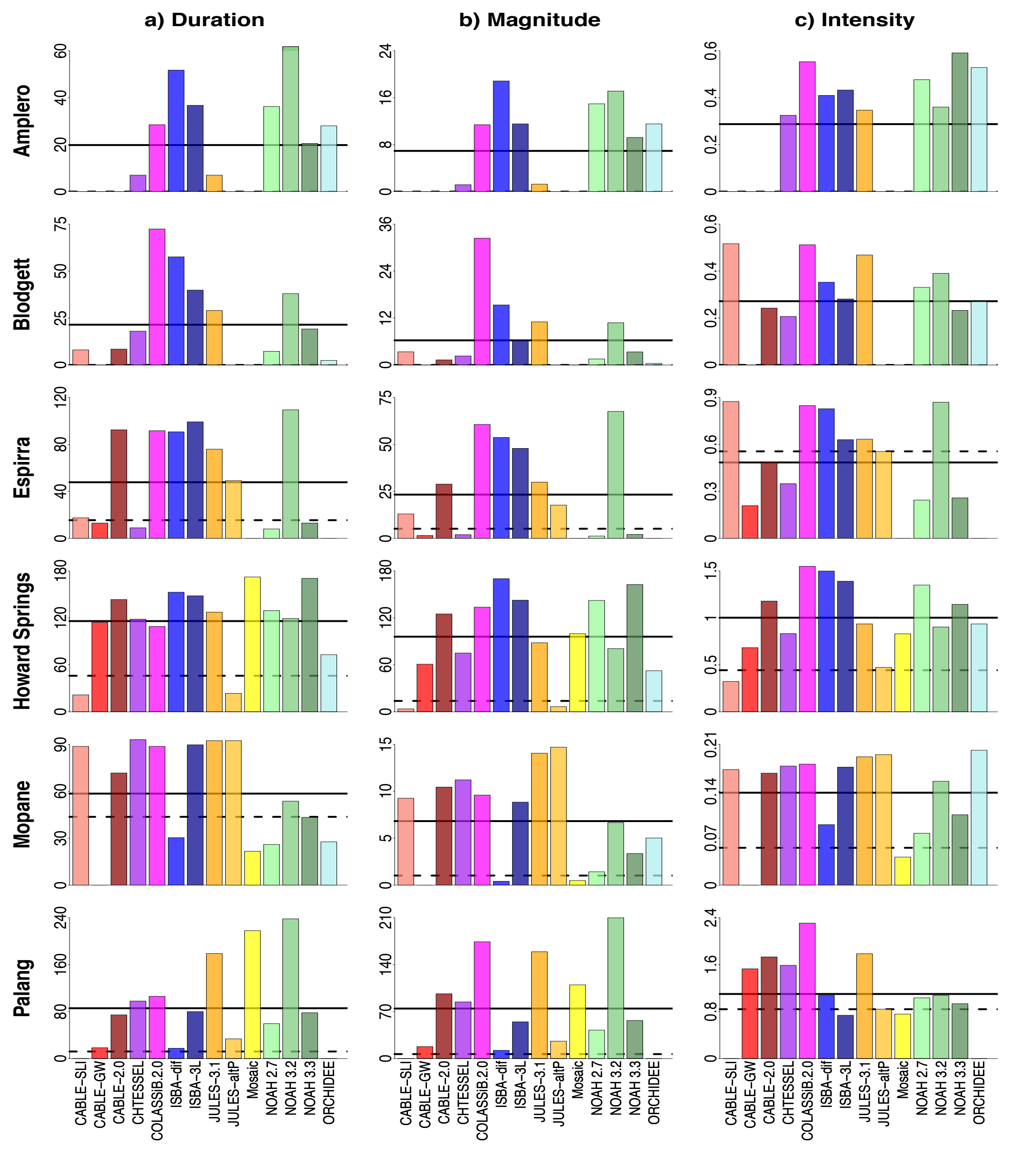


Dry-down events at PLUMBER sites (Anna Ukkola)



Ukkola et al, in press, ERL

Evaporative drought at PLUMBER sites



Ukkola et al, in press, ERL

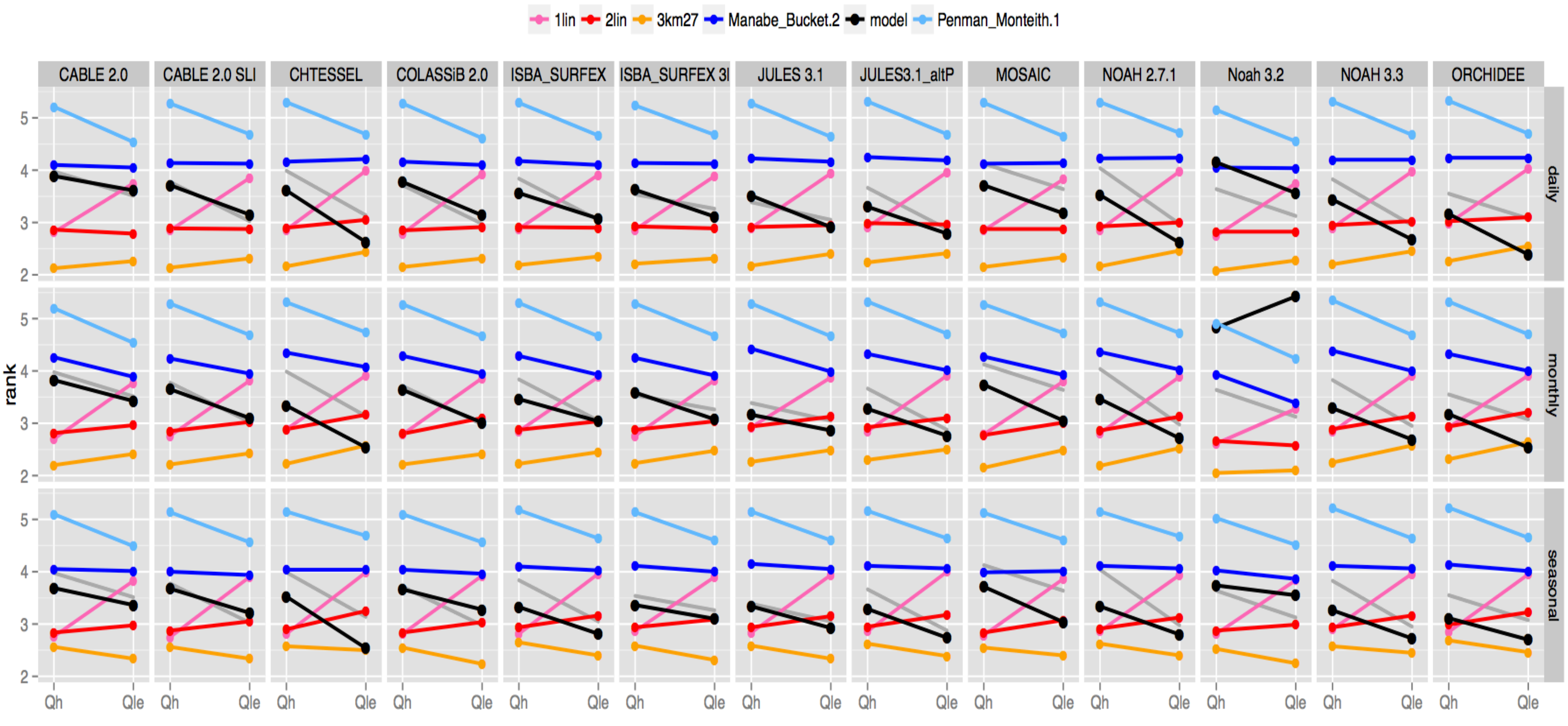
Ongoing work around PLUMBER

- Ned Haughton (UNSW Sydney, + Gab) – how good can empirical models be?
  - Continuing to look for an uber-model using all met variables, flux history, markov chain approach etc
  - Build in conservation?

- Martyn Clark (UCAR) – investigating PLUMBER with SUMMA modelling architecture

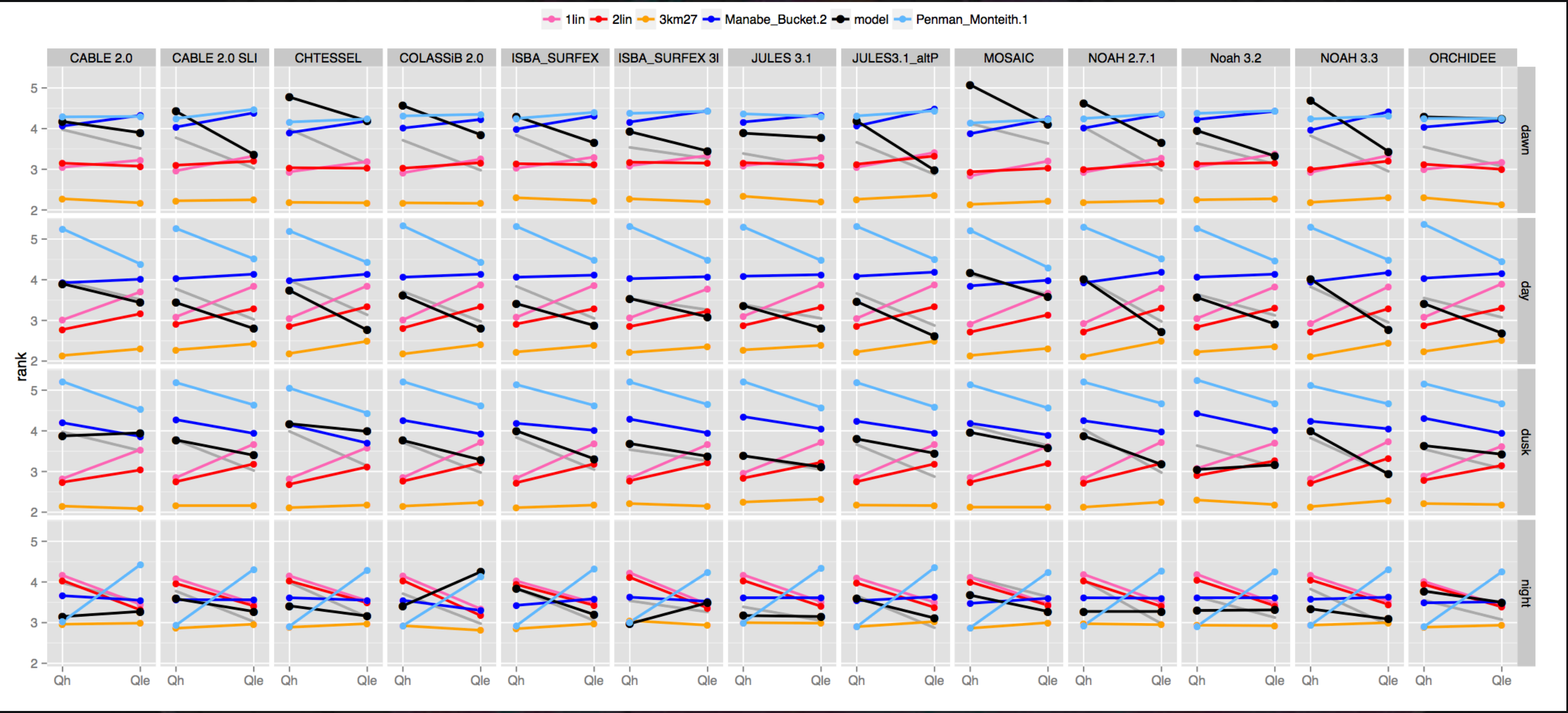
- Martin Best has mentioned wanting to use PLUMBER data for some kind of frequency domain analysis

PLUMBER results – timescale?



Haughton et al, 2016

PLUMBER results – time of day?



Haughton et al, 2016

PLUMBER results – initialisation?

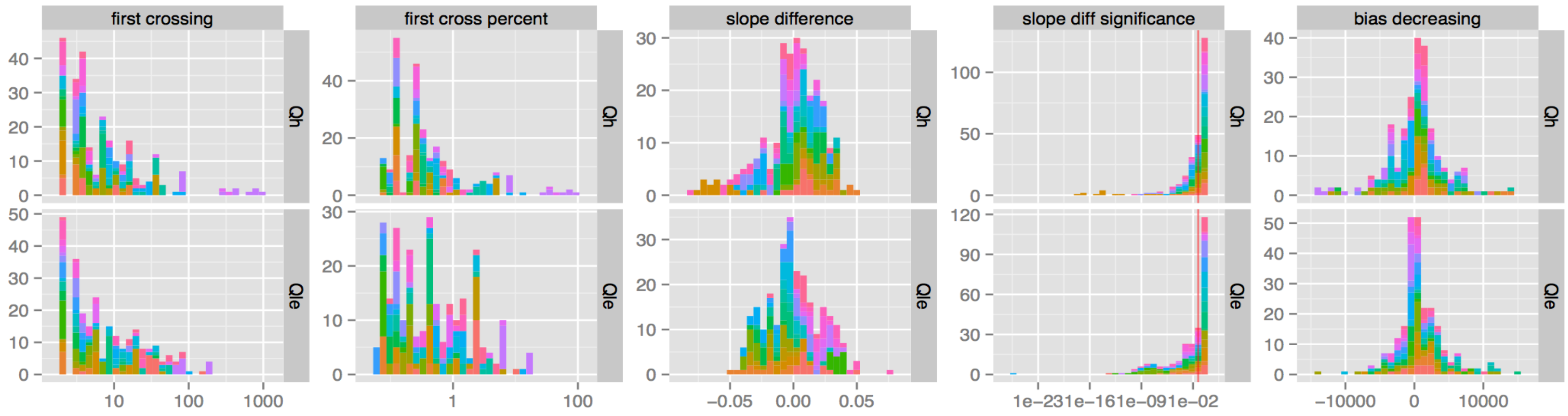
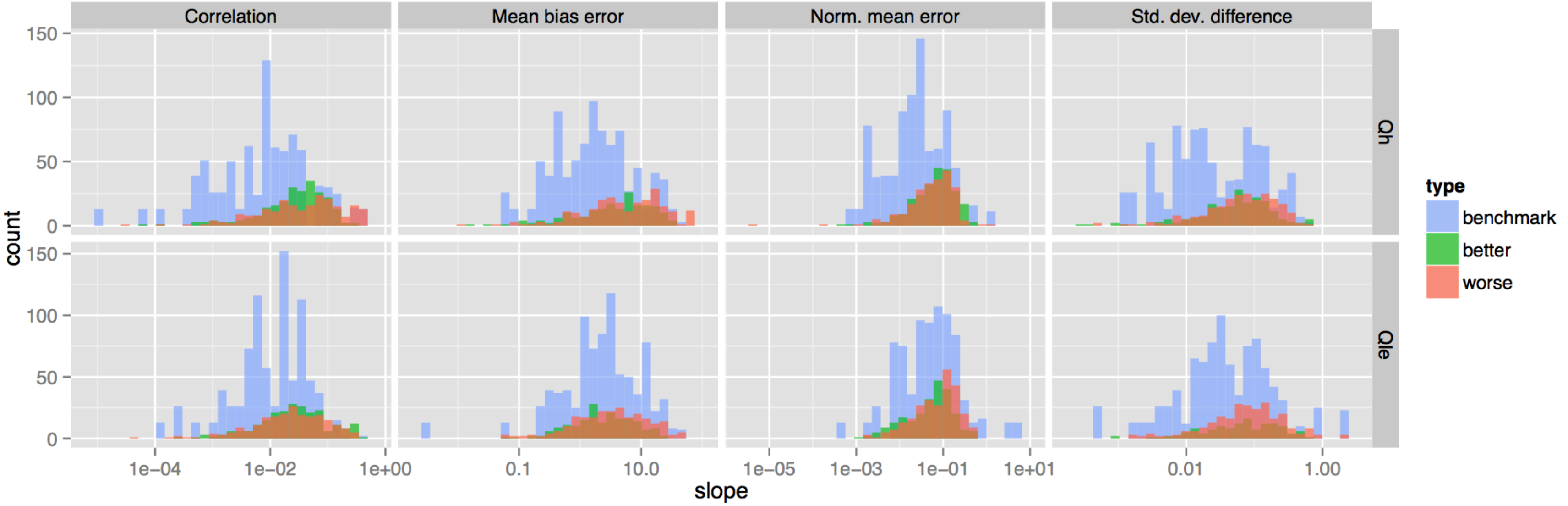


FIG. 7. Model spin-up metrics, based on daily averages, from all LSMs at all sites. From left to right: 1) day at which the simulated series crosses the observed series; 2) as previous, but as a percentage of the time series; 3) difference in the slopes of linear regressions of simulated and observed 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, 2016

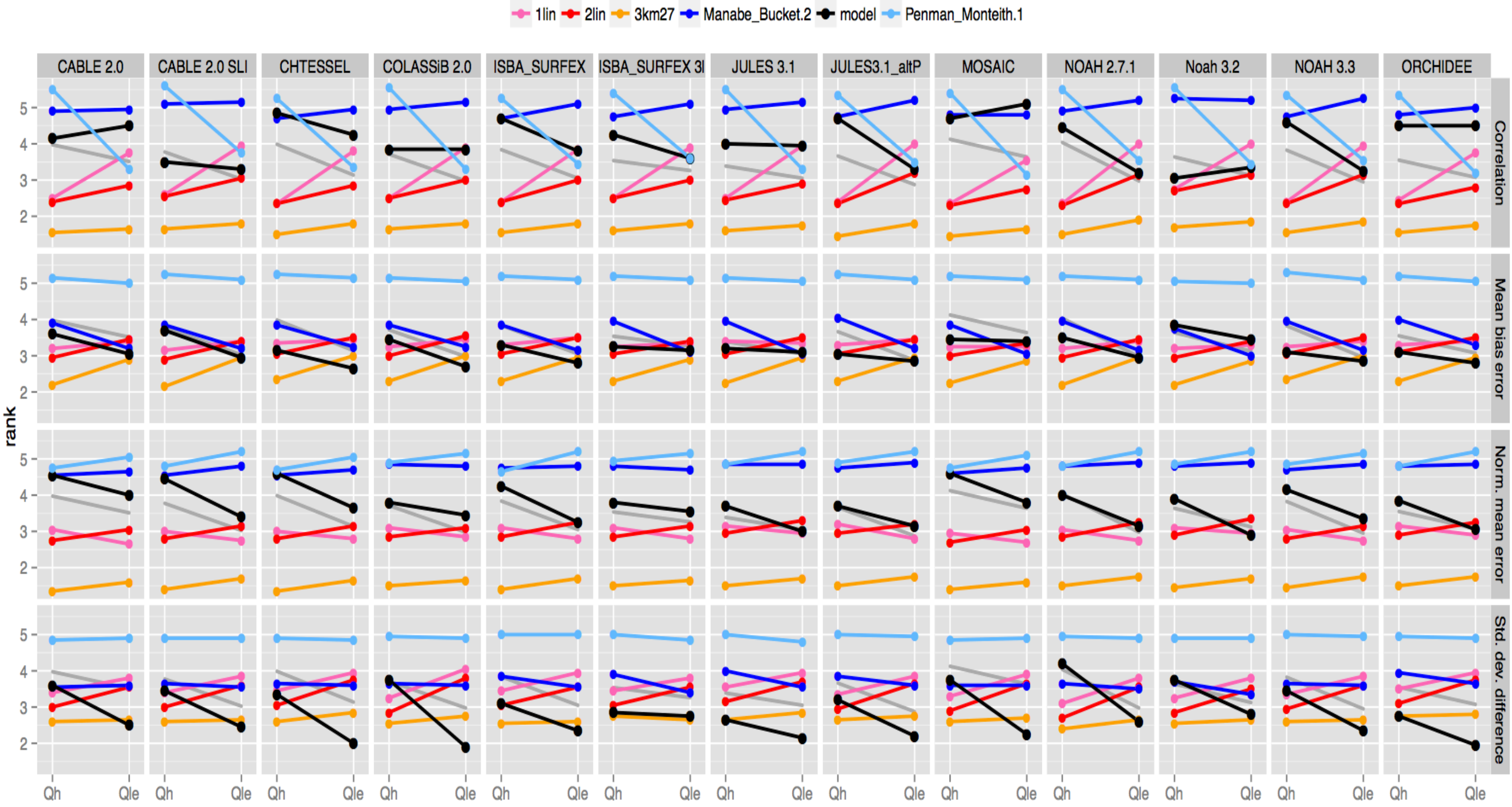
PLUMBER results – ranks vs metric values?



Haughton et al, J Hydromet, in review

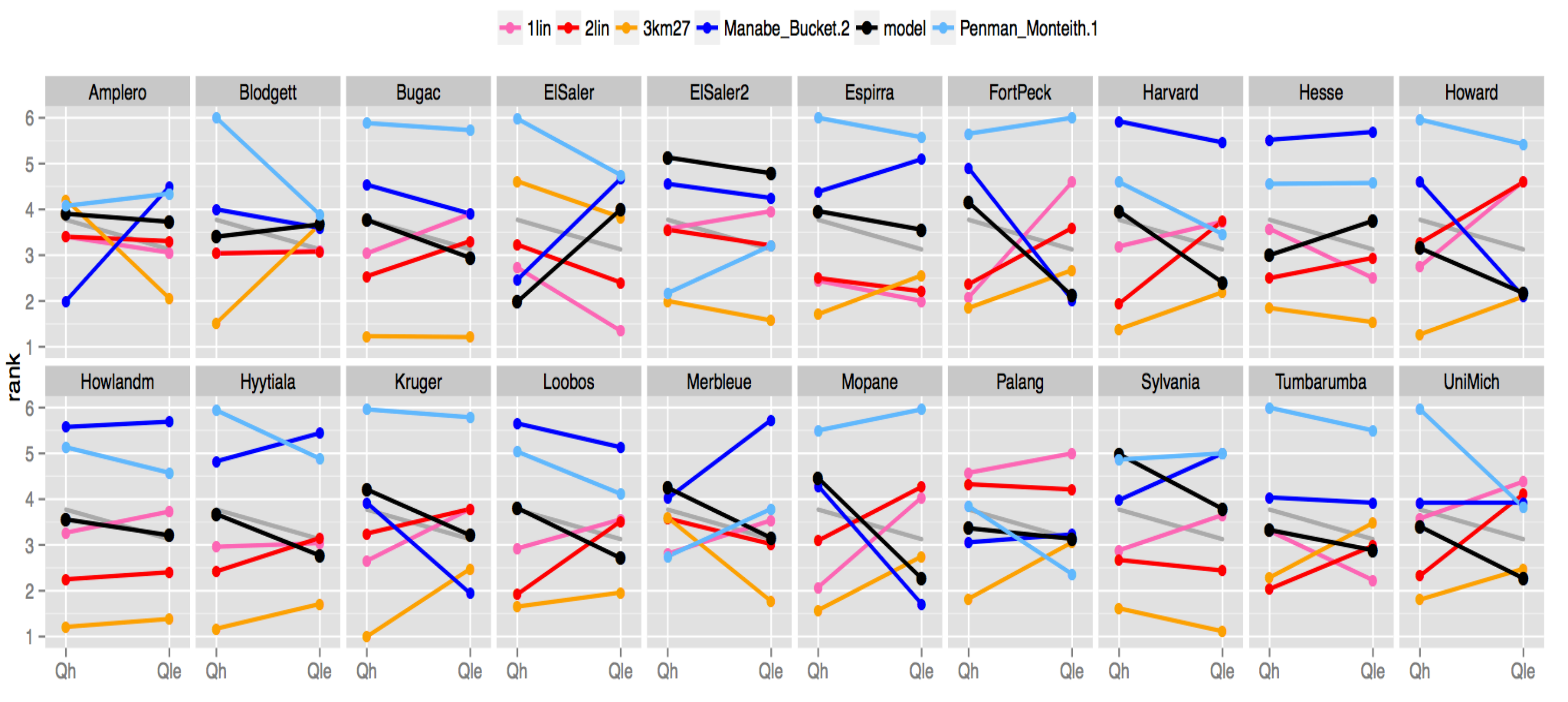
FIG. 3. Histograms of differences between metric values for benchmarks and models with neighbouring ranks, for all models at all sites. Values are calculated by taking the difference of the metric value for each model from the model ranked next-worst in for each LSM, Fluxnet site, metric, and variable. 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, 2016

PLUMBER results – sites?



Haughton et al, J Hydromet, in review