Evaluating and benchmarking land surface models

Graham Weedon[1], Heather Ashton[1], Martin Best[1]

[1] = Met Office

WGNE Workshop on Systematic Errors in Weather and Climate Models

Key points about LSM evaluations:

Evaluations = assessment against observations

Comparisons = assessment against other models

Benchmarking = assessment against a pre-defined reference

Definitions: Best et al. 2015, JHM, doi: 10.1175/JHM-D-14-0158.1.

Offline LSM runs, for testing process representation, are forced using:

a) locally observed meteorology or

b) bias-corrected reanalyses

Severe lack of long, high quality observations of LSM output variables – especially soil moisture and soil temperature

Land Surface Models and IPCC AR6

LS3MIP = Land Surface, Snow and Soil moisture Model Intercomparison Project

LMIP = Offline simulations (multiple meteorological forcings)

LFMIP = Atmosphere-Land feedbacks (soil moisture / snow)

Van den Hurk et al., 2016, *GMD* doi:10.5194/gmd-9-2809-2016

LUMIP = Land Use Model Intercomparison Project

Focus on land-use and land cover change (LULCC) & climate

Lawrence et al., 2016, *GMD* doi:10.5194/gmd-9-2973-2016

Some LSM evaluation tools

PALS = Protocol for the Analysis of Land Surface Models

Primarily uses site (FLUXNET) 30min – 1hr observations + R-based standard metrics

Abramowitz, 2012, *GMD*, doi: 10.5194/gmd-5-819-2012

ILAMB = International Land Model Benchmarking

ILAMBv2.0: monthly, gridded 0.5^{o} x 0.5^{o} surface and EO data with a focus on carbon-related processes and bespoke metrics

Luo et al., 2012, *Biogeosciences*, doi: 10.5194/bg-9-3857-2012

ESMValTool = Earth System Model Evaluation Tool

ESM evaluation protocol for CMIP6. Metrics based on climatological means and annual cycles

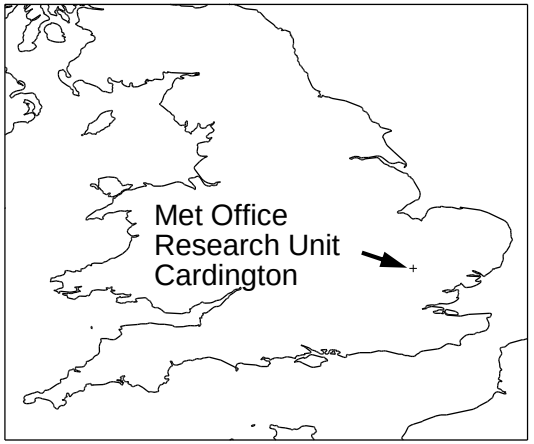
For LSMs near-surface Air Temp.; Evapotransp. v LandFlux-EVAL; Runoff for 12 large catchments

Eyring et al., 2015, *GMD*, doi: 10.5194/gmd-9-1747-2016

LVT = Land surface Verification Toolkit

Part of NASA LIS (Land Information System). Site or gridded data, any time step, allows for missing data & screening obs using quality flags, many statistical metrics including 95% confidence intervals

Kumar et al., 2012, *GMD*, doi: 10.5194/gmd-5-869-2012

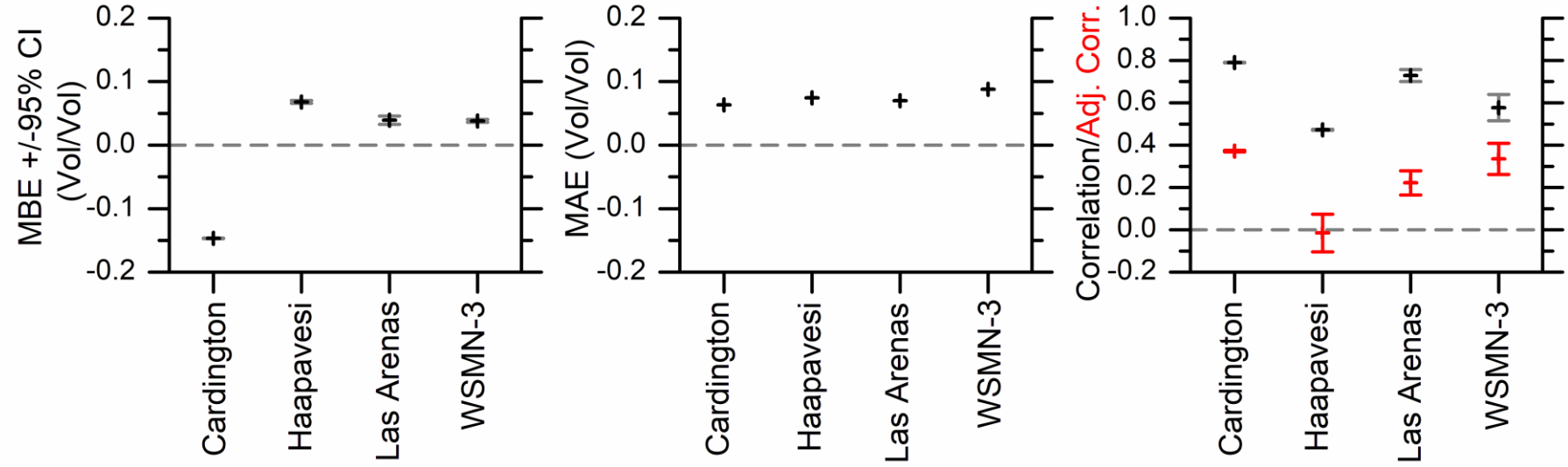


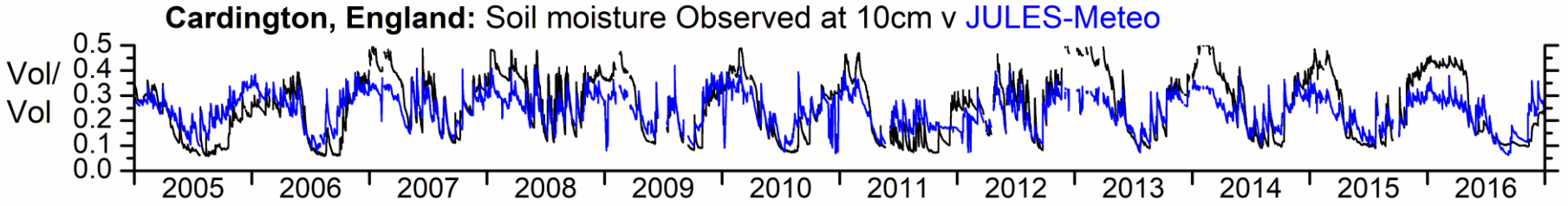
Met Office research unit (MRU) Cardington
30 min surface observations 2005-2016:
(*N* = about 210,000 time steps)

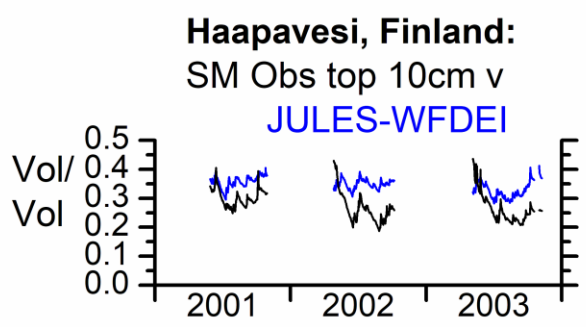
Tower measurements at 1.2m, 10m, 25m, 50m:
Air temperature; humidity; air pressure; wind speed

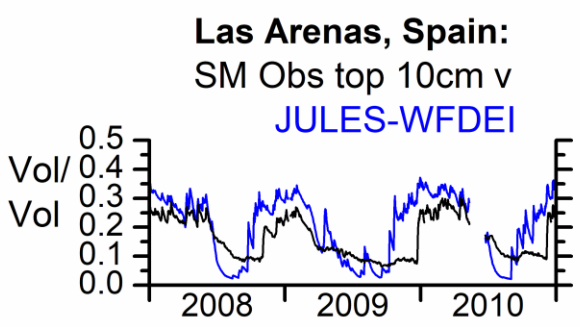
Other LSM-relevant measurements:
Air pressure; rainfall rate; downwards shortwave; downwards longwave; momentum flux; latent heat; sensible heat; soil saturation; soil temperature; albedo; COSMOS-UK soil moisture

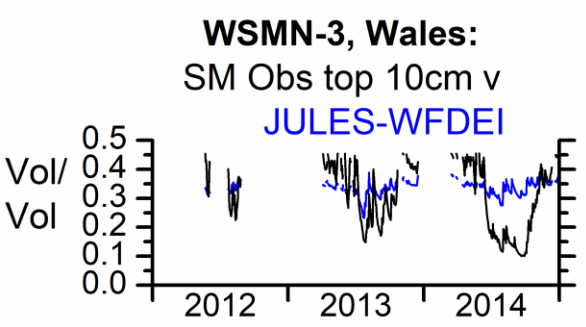
Standard metrics and soil moisture



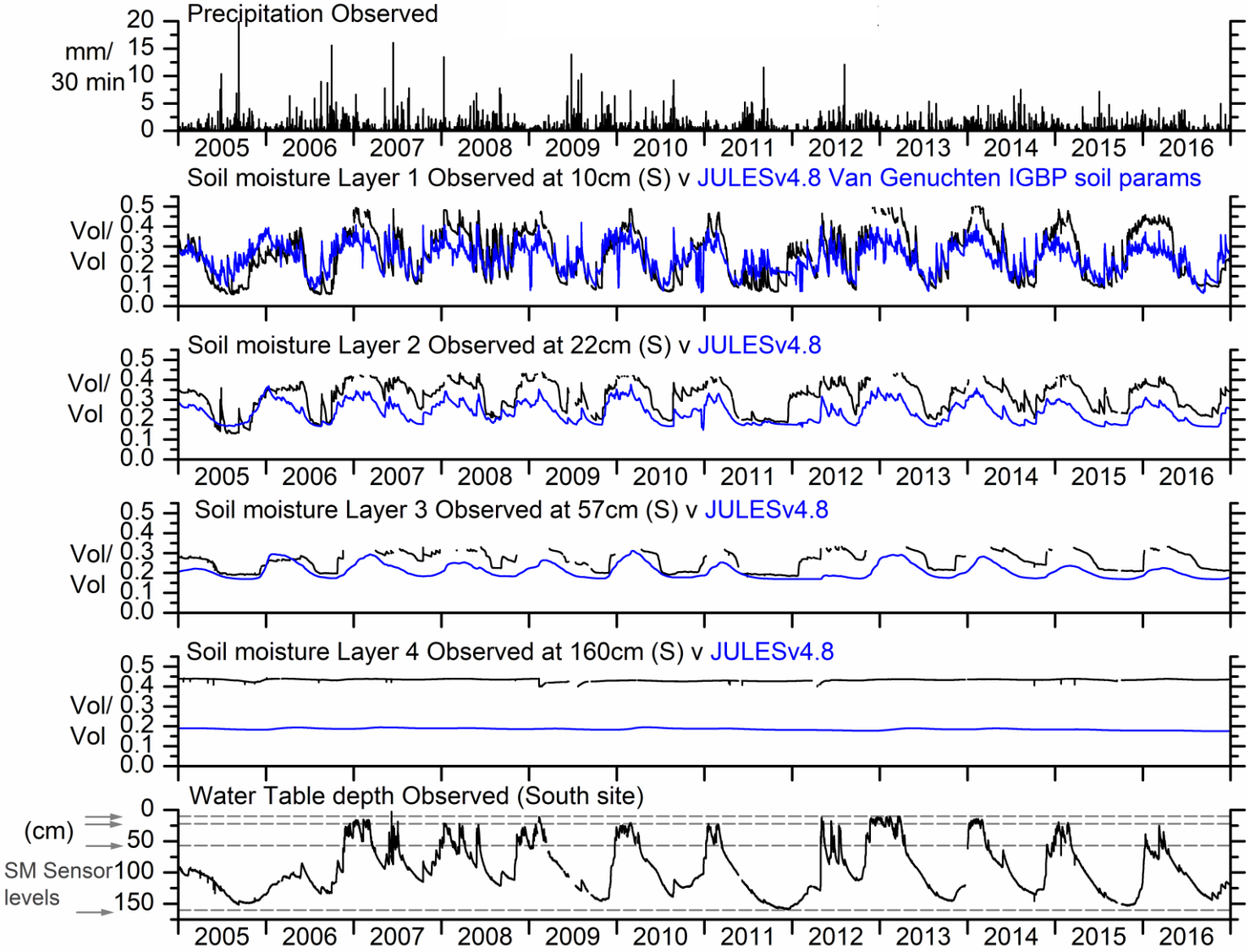




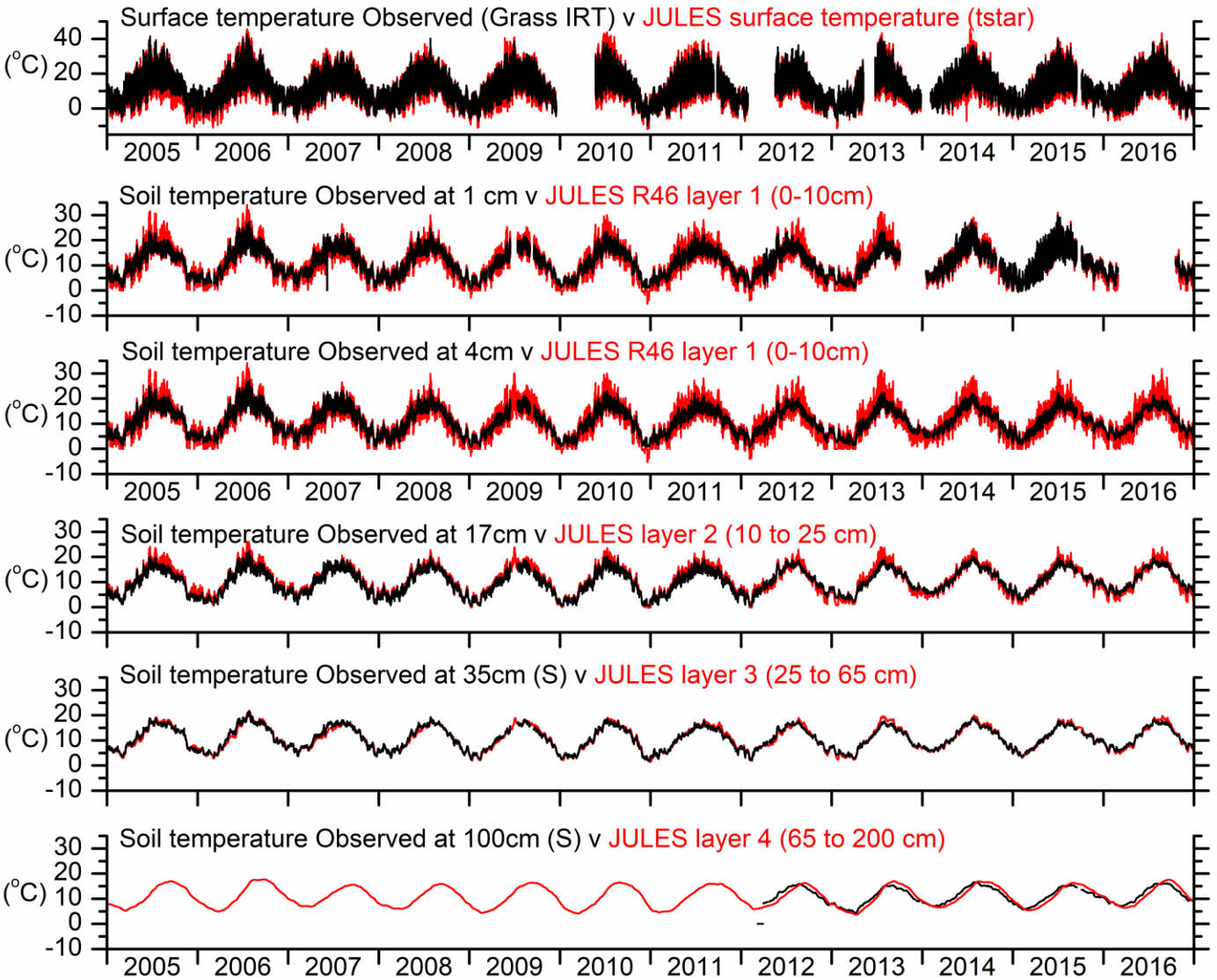




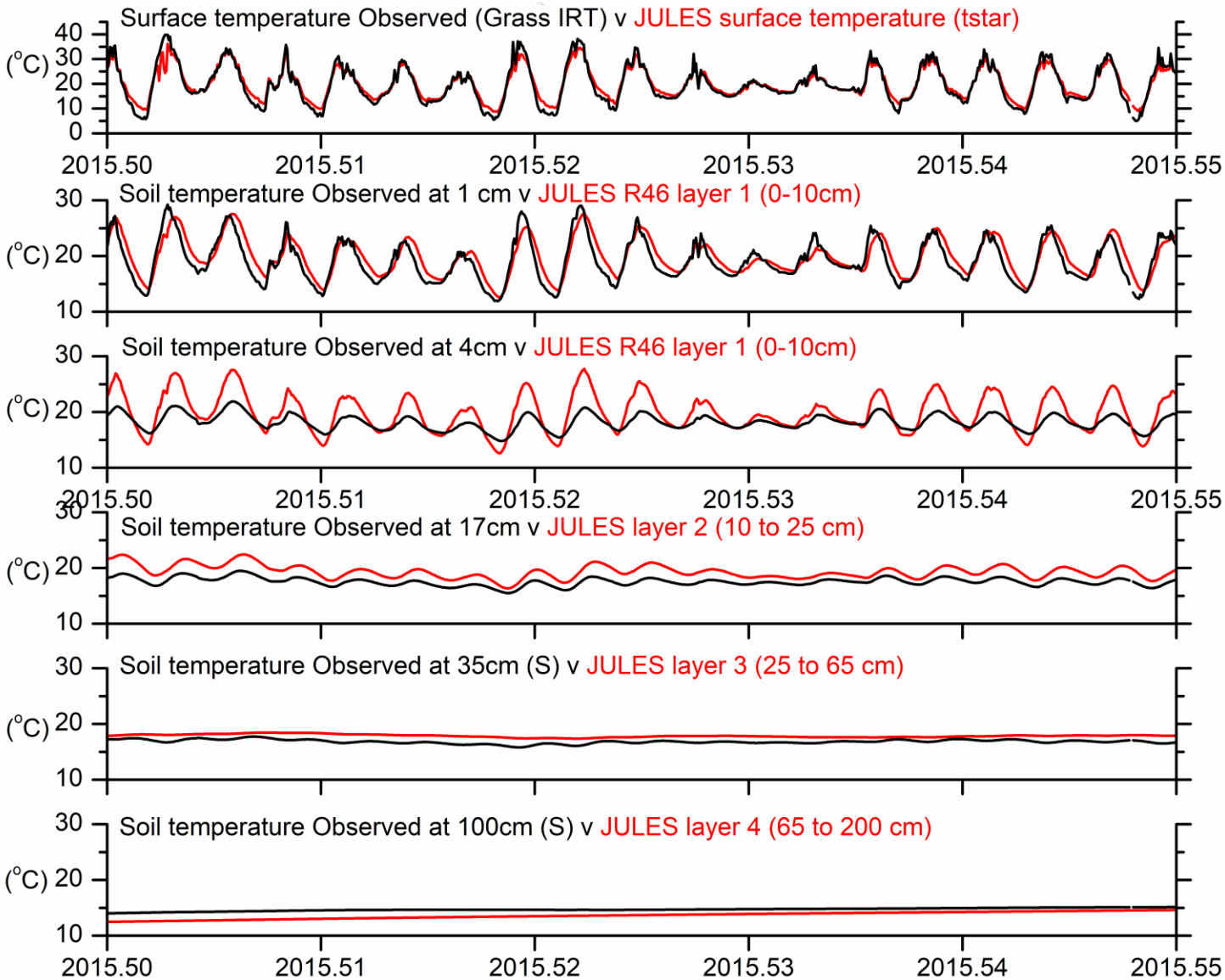
MRU Cardington soil moisture data and water table depths



MRU Cardington soil temperature



MRU Cardington soil temperature June 2015



Cross spectral evaluation of model time series against observations, e.g.: Soil temperature

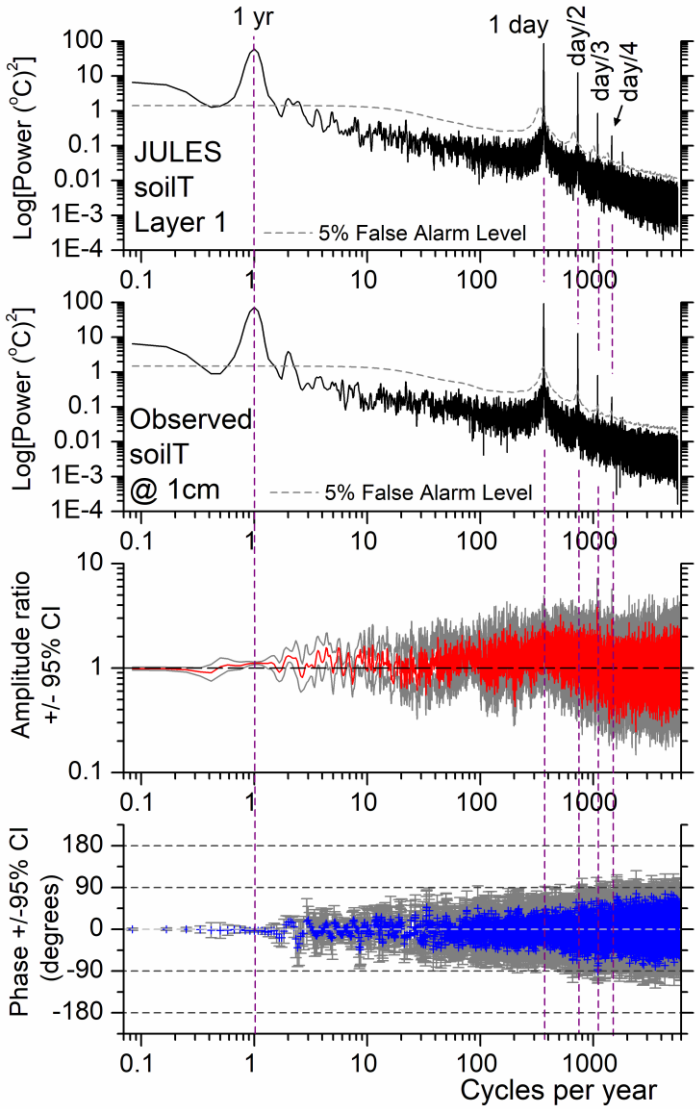
Methodology: Weedon et al., 2015, *JHM*, doi: 10.1175/JHM-D-14-0021.1

Variance or power spectrum

Variance or power spectrum

Amplitude ratio spectrum

Phase spectrum



Cross spectral evaluation of model time series against observations, e.g.: Soil temperature

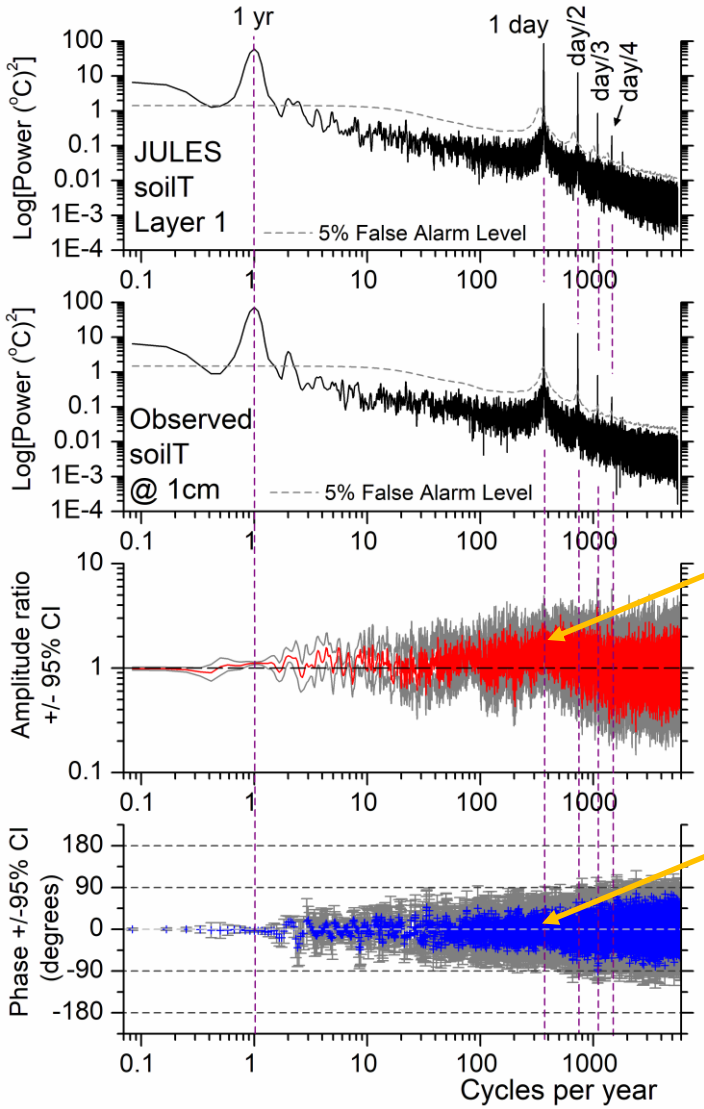
Methodology: Weedon et al., 2015, *JHM*, doi: 10.1175/JHM-D-14-0021.1

Variance or power spectrum

Variance or power spectrum

Amplitude ratio spectrum

Phase spectrum

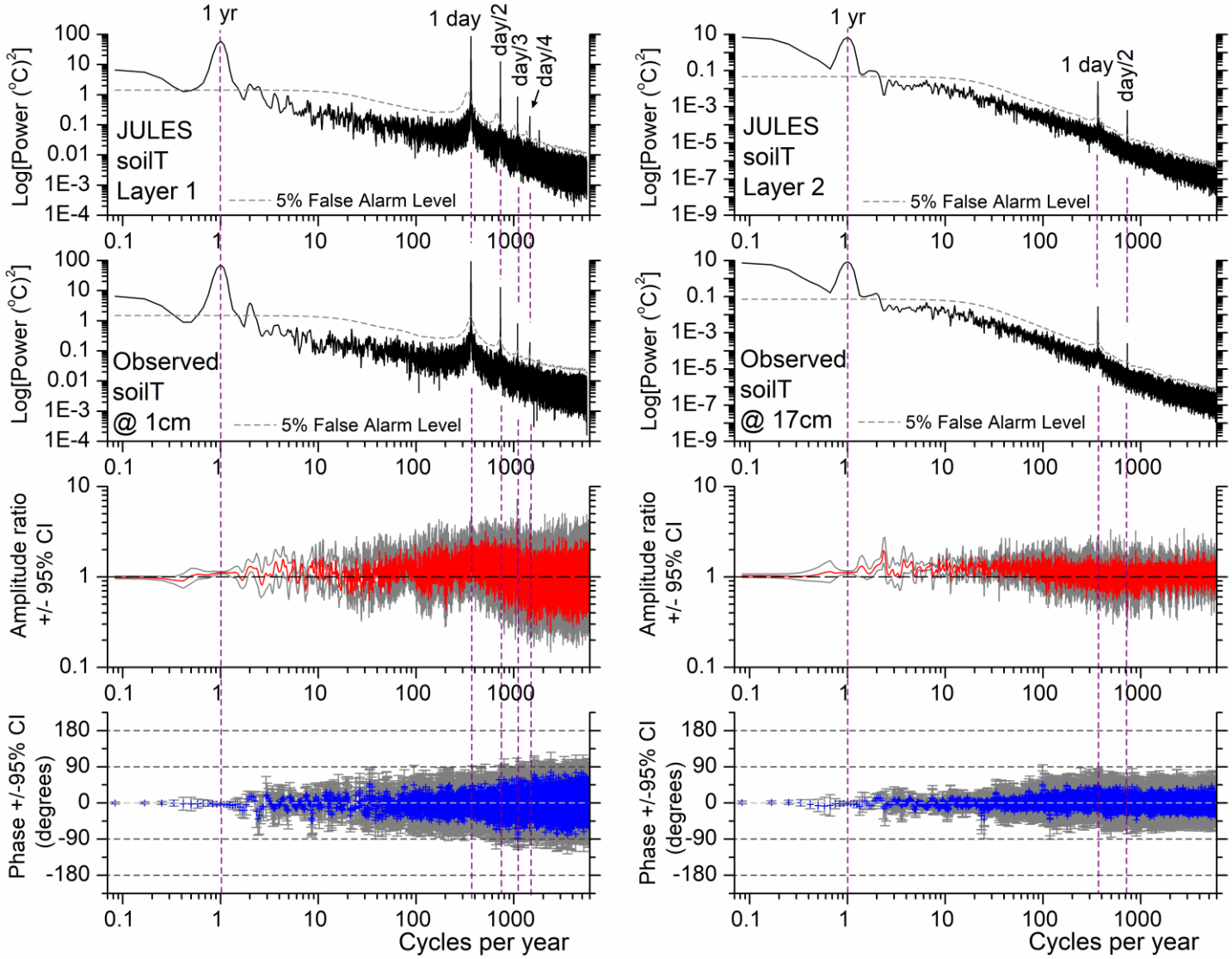


Daily amplitude ratio = 1.70
(1.30 to 2.22)
> Too variable

Daily phase difference = -11.8°
(-36.4° to +13.0°)
> In phase

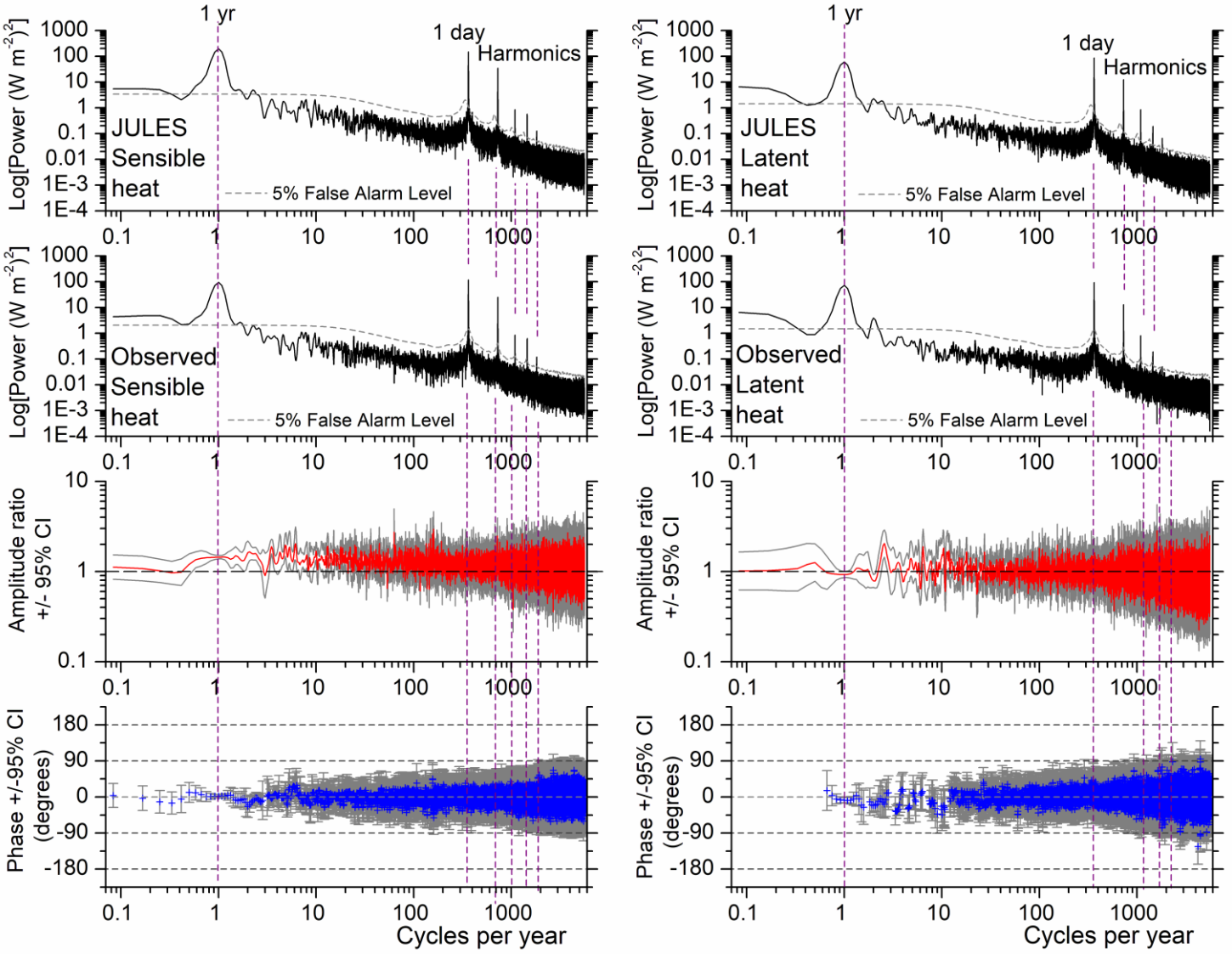
Cross spectral evaluation of model time series against observations, e.g.: Soil temperature

Methodology: Weedon et al., 2015, *JHM*, doi: 10.1175/JHM-D-14-0021.1



Cross spectral evaluation of model time series against observations, e.g.: Sensible & Latent heat

Methodology: Weedon et al., 2015, *JHM*, doi: 10.1175/JHM-D-14-0021.1



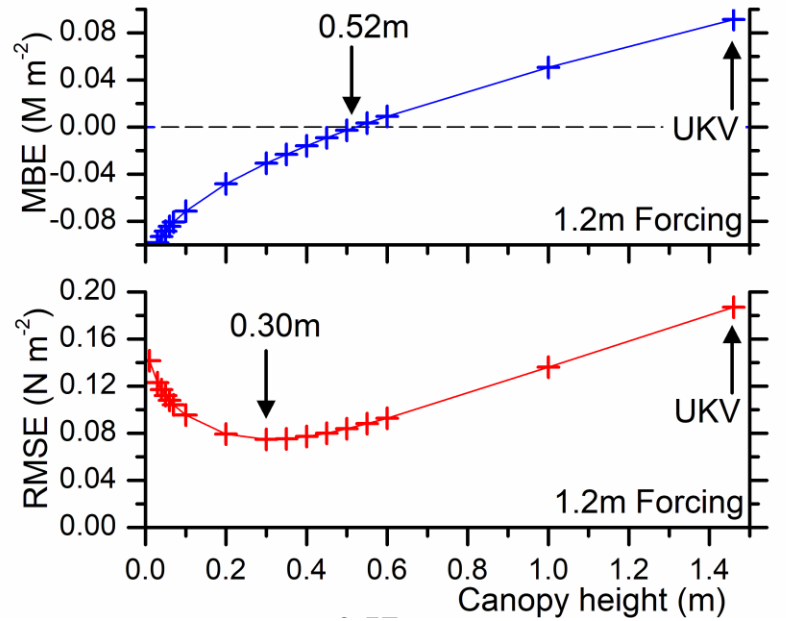
Constraining roughness length via canopy height using observations of momentum flux

1) The JULES offline runs illustrated are based on the UKV configuration – as used in Met Office weather forecasts.

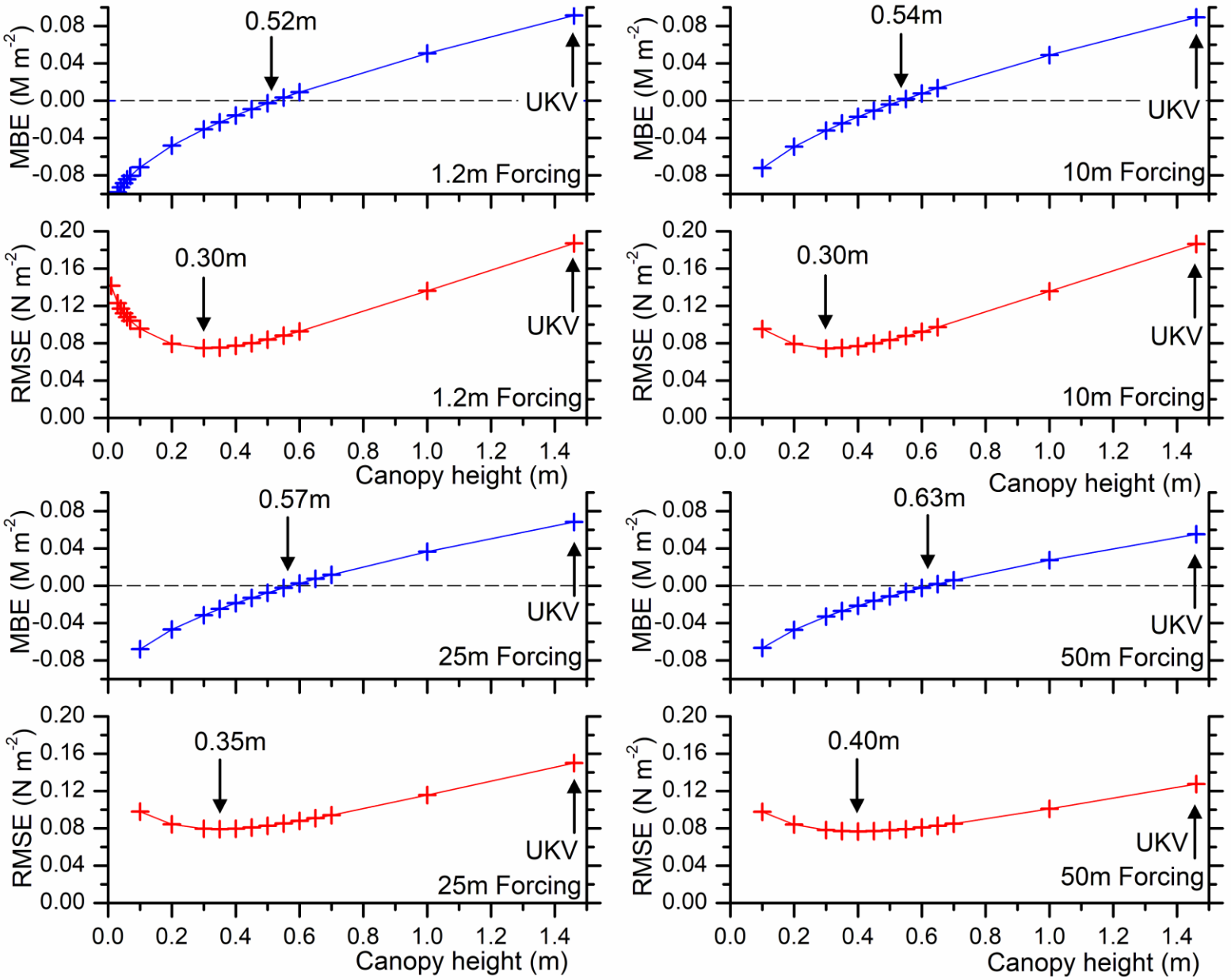
2) In UKV the canopy height (linked to roughness length) is defined as 1.46 m for C3 grass land cover.

3) The Cardington site is C3 grass, so can we use offline runs of JULES plus observations of momentum flux to estimate the optimum canopy height?

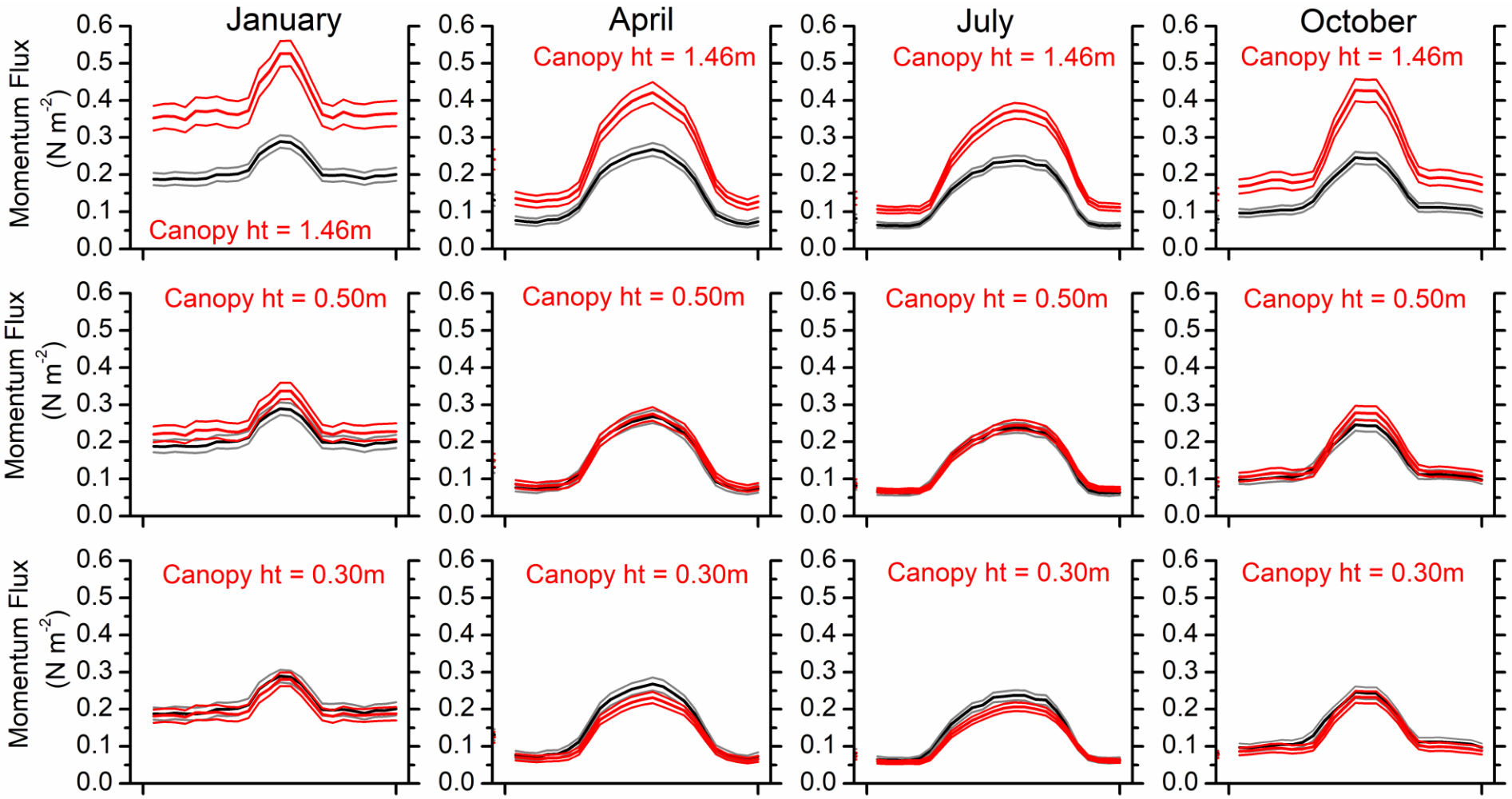
Optimum JULES-UKV **offline** configuration “canopy height” using modelled v. observed momentum flux for different forcing levels



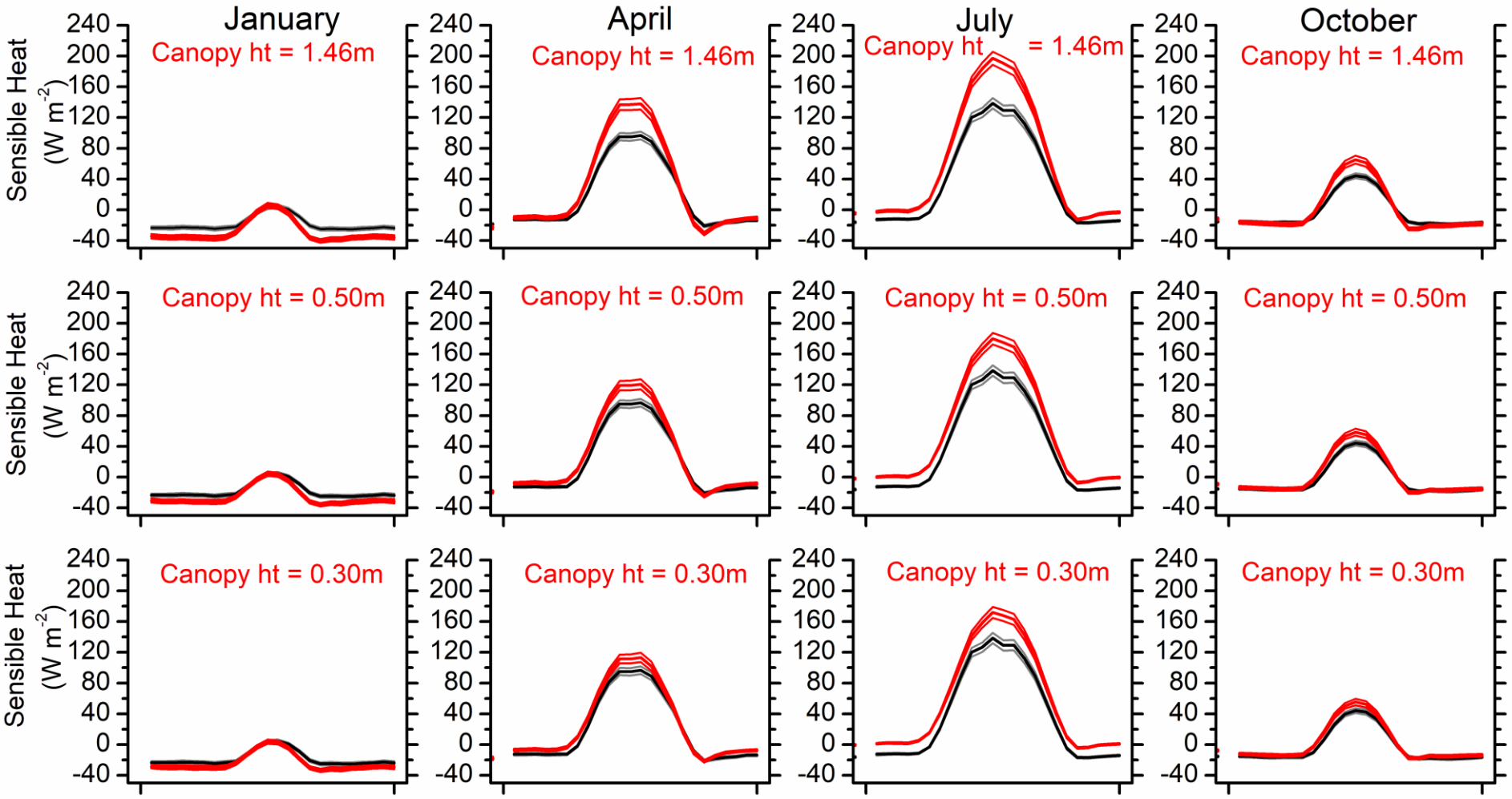
Optimum JULES-UKV **offline** configuration "canopy height" using modelled v. observed momentum flux for different forcing levels



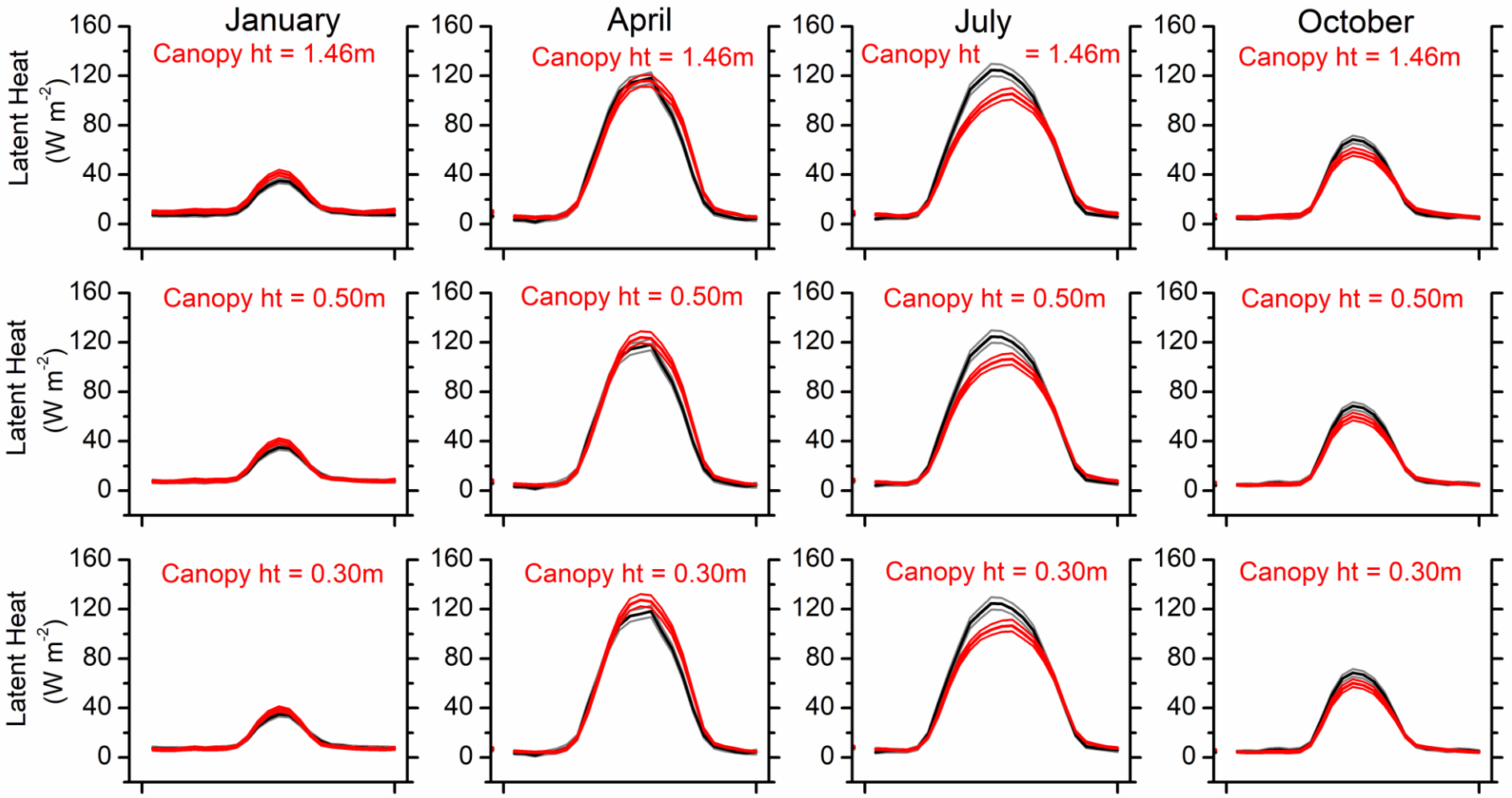
MRU Cardington 2005-2016
Average diurnal cycles in Momentum flux +/- 95% CIs:
1.2 m forcing JULESv4.8 VG IGBP soil params v. Obs.



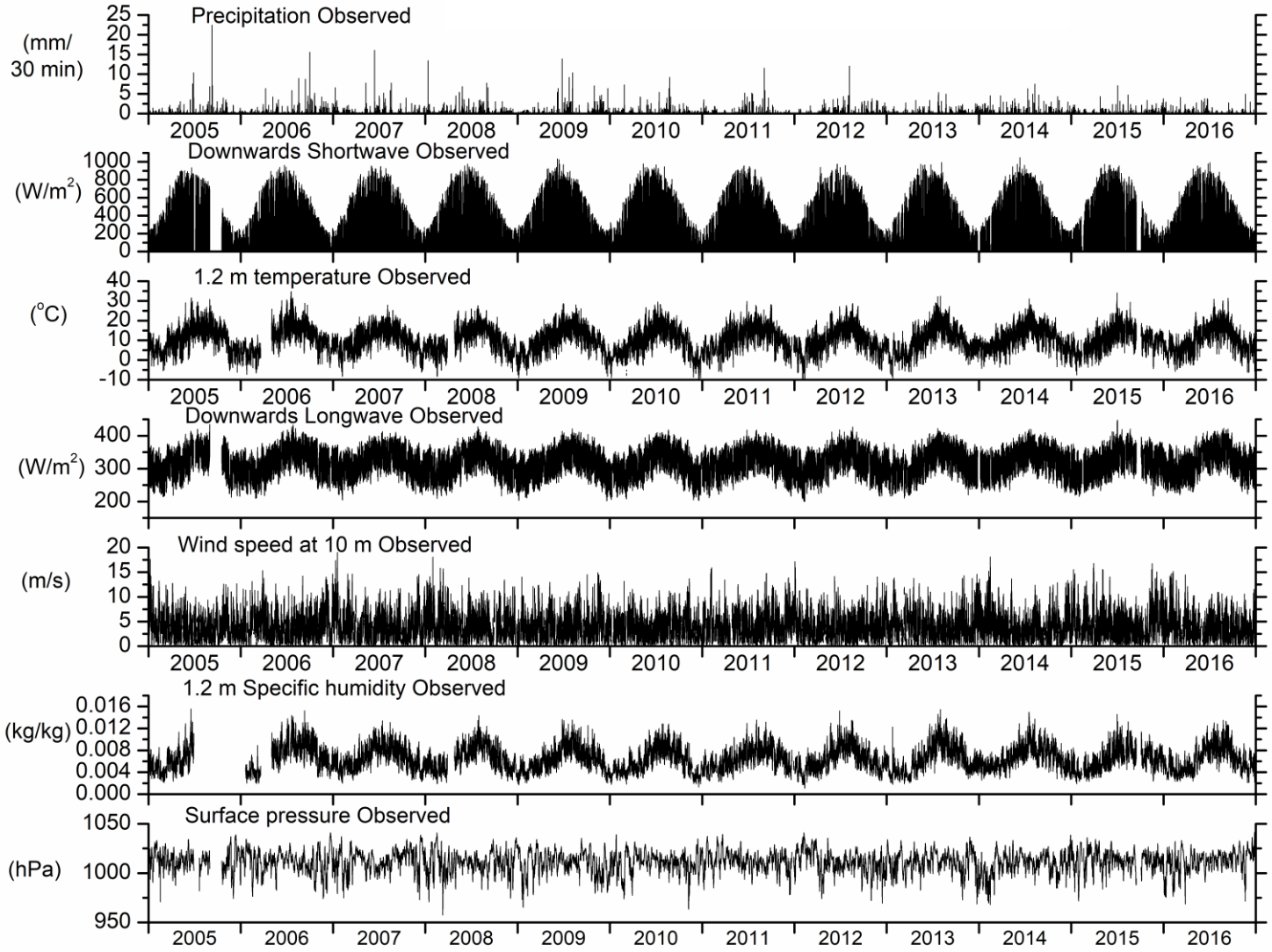
MRU Cardington 2005-2016
Average diurnal cycles in Sensible Heat +/- 95% CIs:
1.2 m forcing JULESv4.8 VG IGBP soil params v. Obs.



MRU Cardington 2005-2016
Average diurnal cycles in Latent Heat +/- 95% CIs:
1.2 m forcing JULESv4.8 VG IGBP soil params v. Obs.

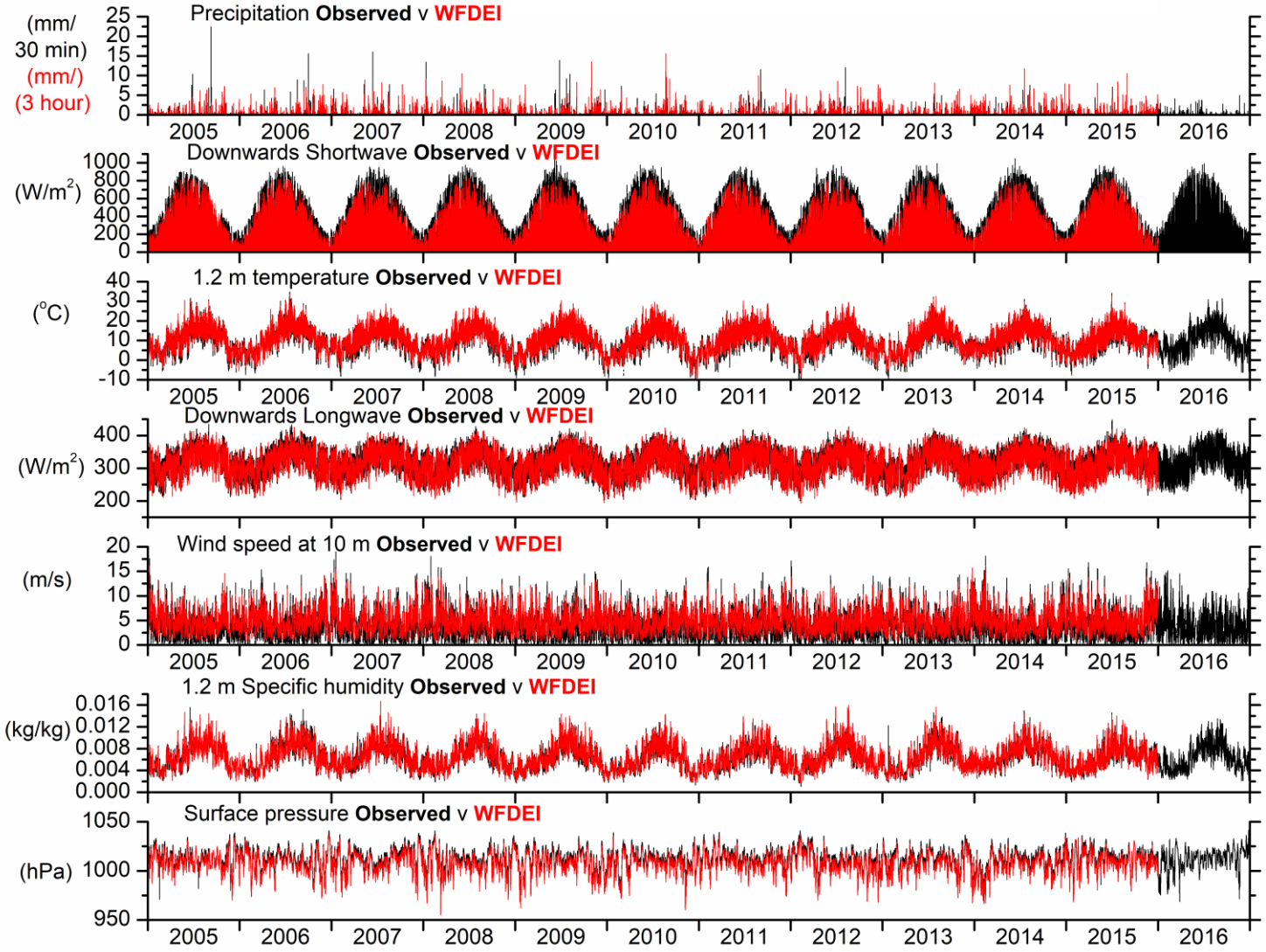


Observed meteorological variables at MRU Cardington

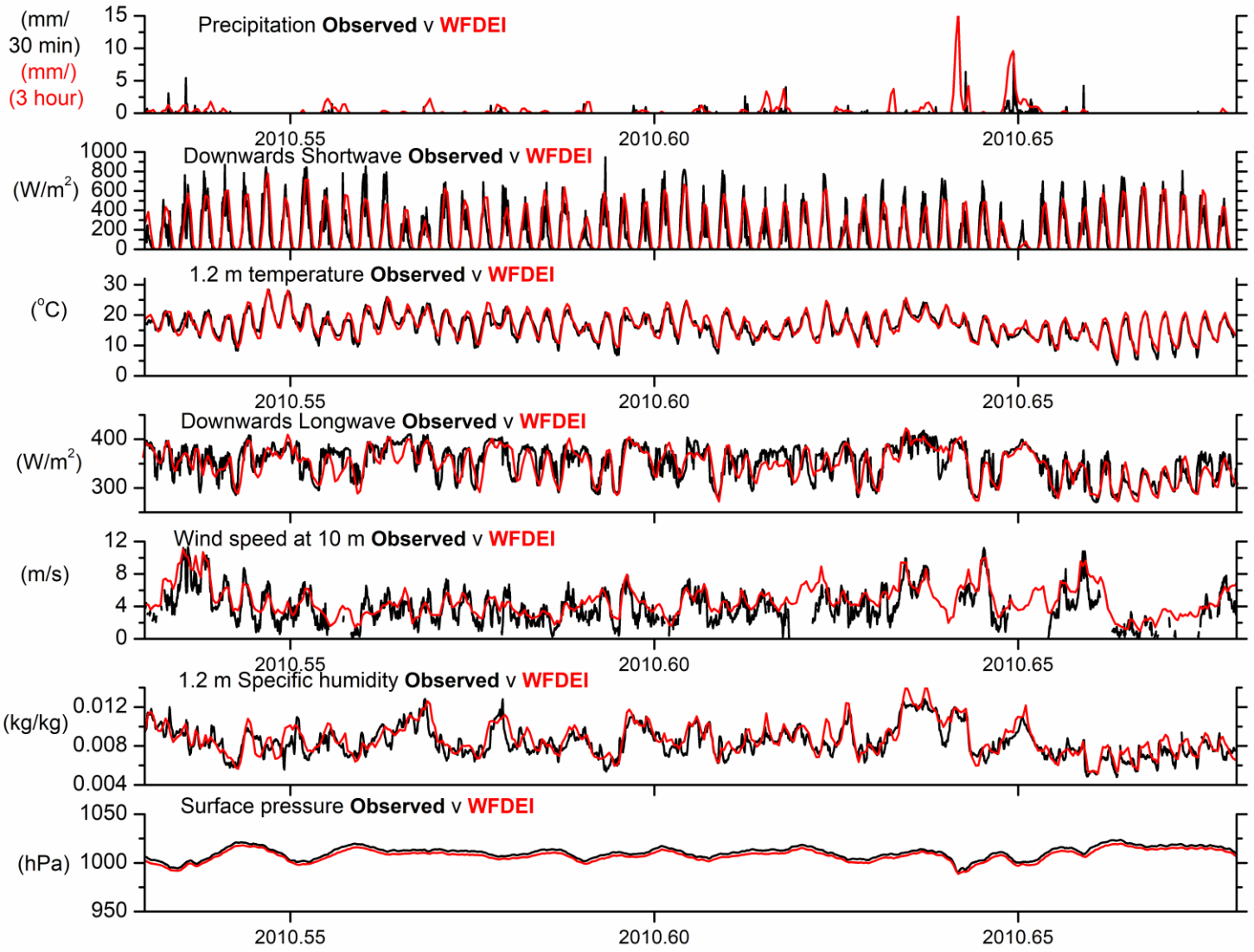


Observed meteorological variables at MRU Cardington versus WFDEI

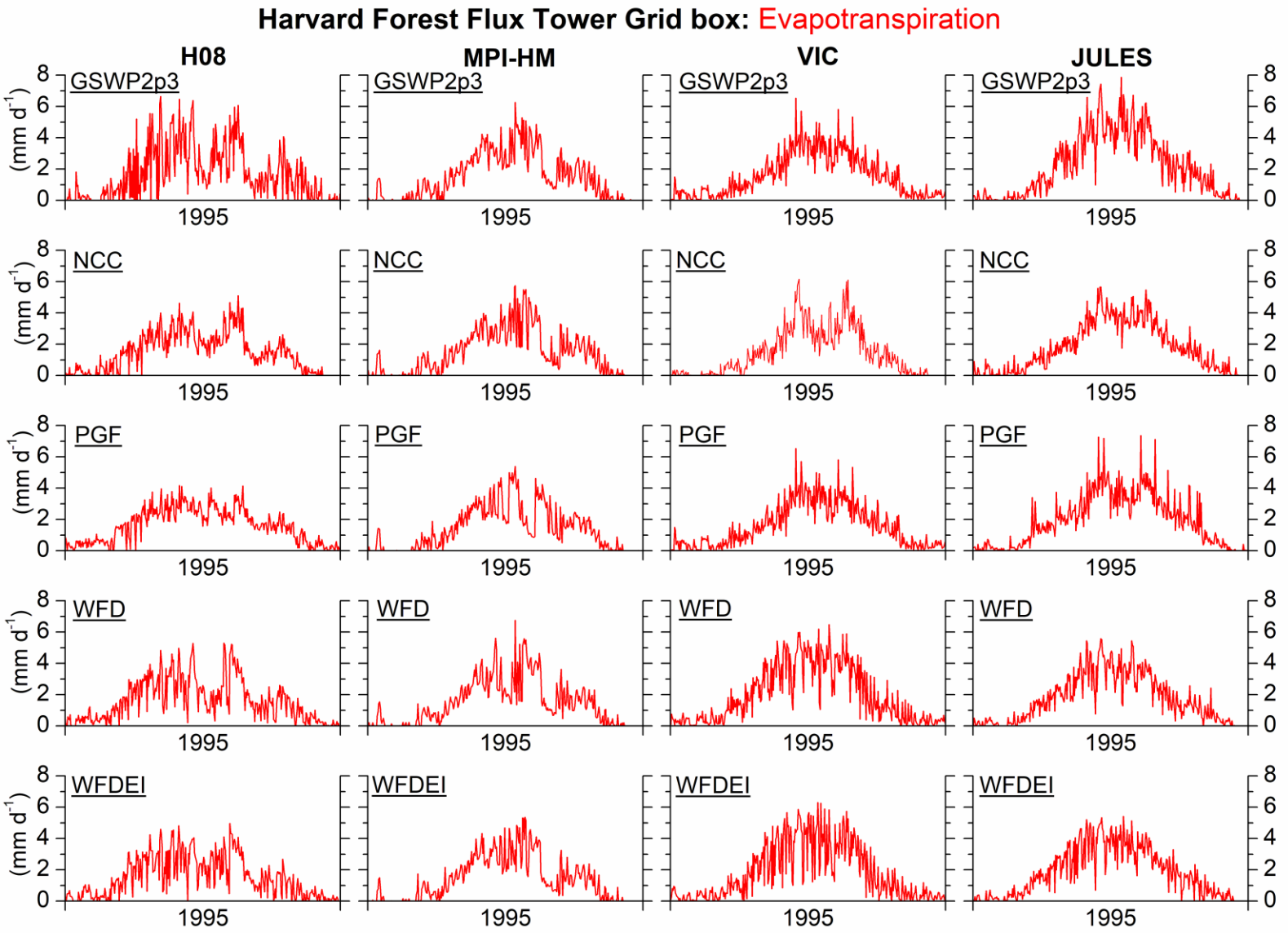
WFDEI description: Weedon, Balsamo et al. 2014, *WRR*, doi: 10.1002/2014WR015638



Observed meteorology (30 min) v WFDEI (3 hourly) for June-July 2010

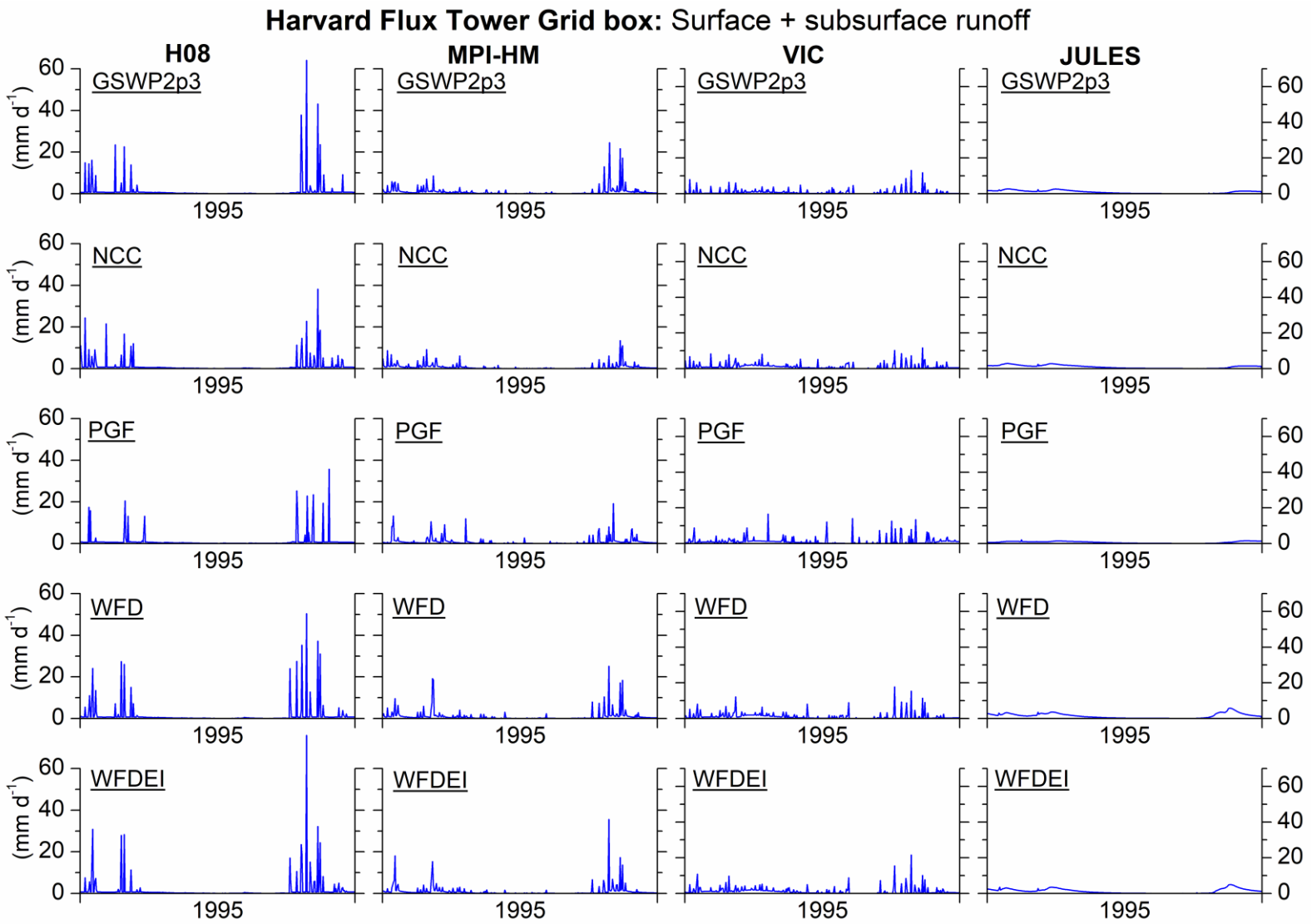


Spread in simulated Latent heat related to meteorological forcing data and model



Weedon, Ellis, Haddeland, Hansaki, Stacke, Sheffield, Best (in prep.)

Spread in simulated runoff related to Meteorological forcing data and model



Weedon, Ellis, Haddeland, Hansaki, Stacke, Sheffield, Best (in prep.)

Sensitivity ratio to measure spread due to forcing *cf.* spread due to model (e.g. evapotranspiration, mm/d)

Forcing dataset	Model H08	Model MPI-HM	Model VIC	Model JULES	Forcing means
GSWP2	1.850	1.622	1.935	2.023	**1.858**
NCC	1.787	1.610	1.409	1.809	**1.654**
PGF	1.836	1.537	1.569	1.857	**1.700**
WFD	1.869	1.592	1.921	1.699	**1.770**
WFDEI	1.791	1.603	1.810	1.887	**1.773**
Model means	**1.827**	**1.593**	**1.729**	**1.855**	-

Weedon, Ellis, Haddeland, Hansaki, Stacke, Sheffield, Best (in prep.)

Stdev (Model means) = 0.118

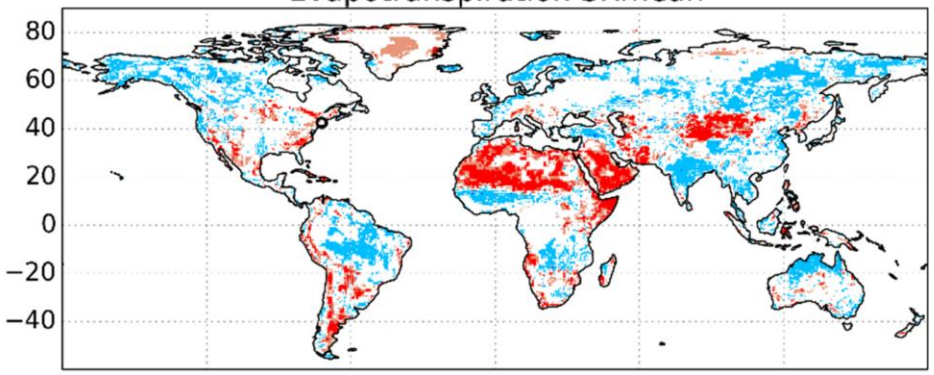
Stdev(Forcing means) = 0.078

Sensitivity ratio (means) = SRmean = 0.078/0.118 = **0.658**

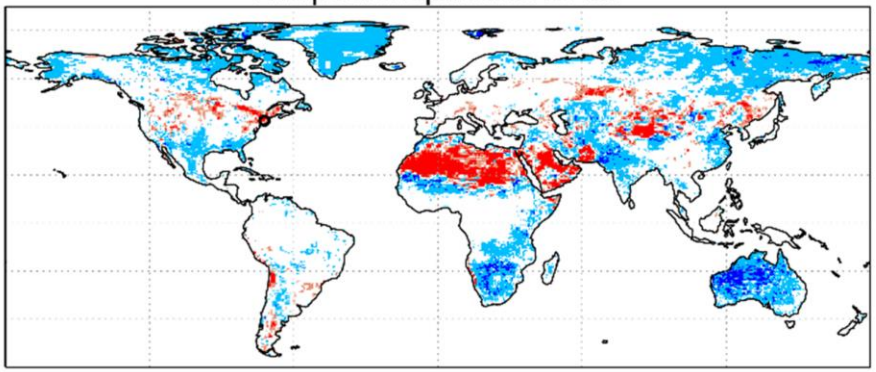
$$\text{SRmean} = \frac{\text{Stdev(Means across Forcing)}}{\text{Stdev(Means across Models)}}$$

$$\text{SRstdev} = \frac{\text{Stdev(Stdev across Forcing)}}{\text{Stdev(Stdev across Models)}}$$

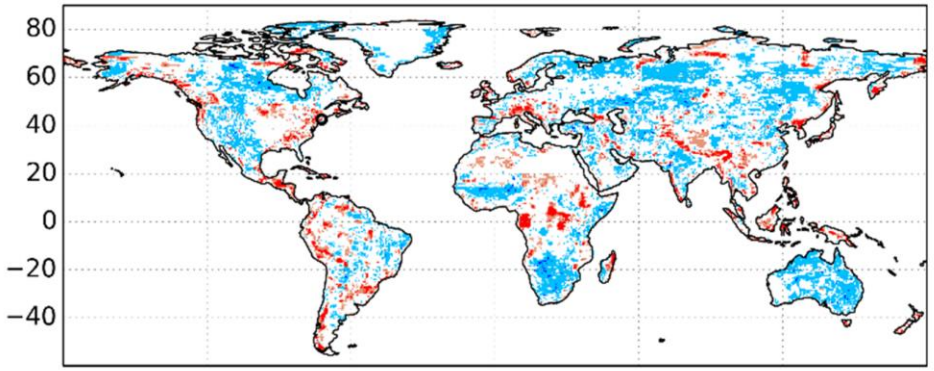
Evapotranspiration SRmean



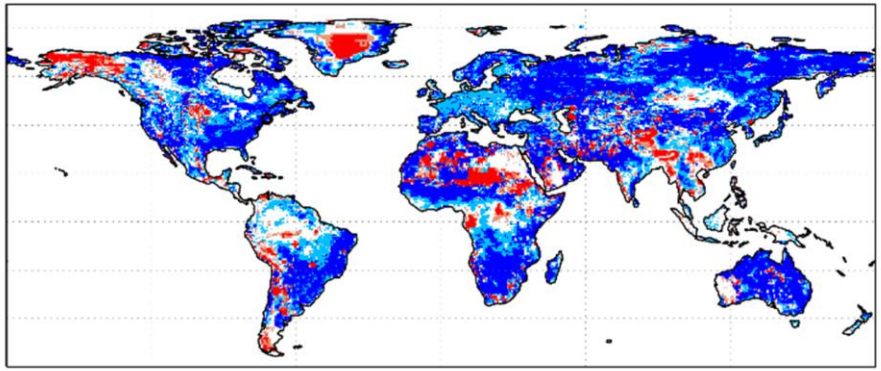
Evapotranspiration SRsd



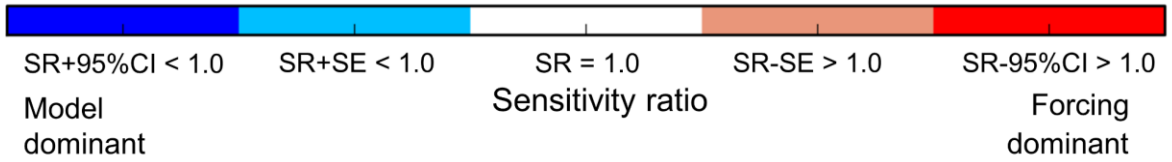
Surface + subsurface runoff SRmean



Surface + subsurface runoff SRsd



Weedon et al. (in prep.)



Conclusions

1) Multiple systems exist for LSM evaluation using standard metrics. Currently LVT seems to be the most flexible in terms of type of data processed, time steps, missing data handling, data quality flag screening and provision of uncertainty values (95% CIs) for evaluation metrics.

2) Cross-spectral analysis, using the Lomb-Scargle Transform for coping with missing time steps, can be useful for investigating mis-matches between LSM simulations at specific time scales in terms of amplitude and phase (and their uncertainties). It is applicable to any model simulations (not just LSMs) given observations for the same time steps. Cross-spectral analysis will be added to LVT.

3) There is currently no widely-adopted specific benchmark (model performance & metrics) that has been defined *a priori* for evaluating LSM performance.

4) Optimizing LSM parameters using offline runs does not guarantee improvements when the model is coupled to the atmosphere.

5) As well as model specifics, LSM performance is influenced, in different ways according to the simulation variable, by the specific re-analysis-based forcing dataset used in offline runs.