Met Office

Evaluation and benchmarking of land surface models

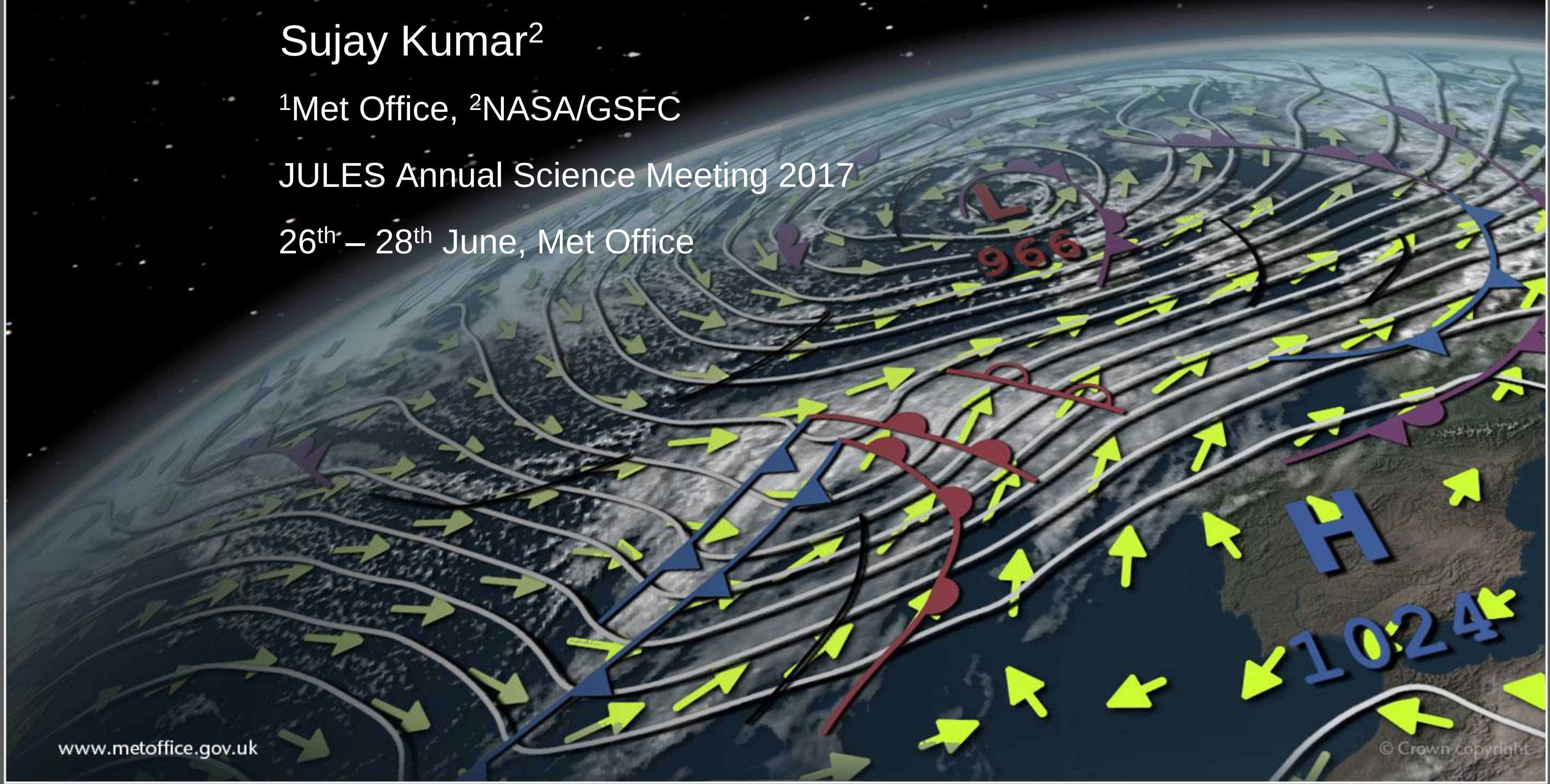
Heather Rumbold[1], Graham Weedon[1], Martin Best[1],

Sujay Kumar[2]

[1]Met Office, [2]NASA/GSFC

JULES Annual Science Meeting 2017

26th – 28th June, Met Office



Outline

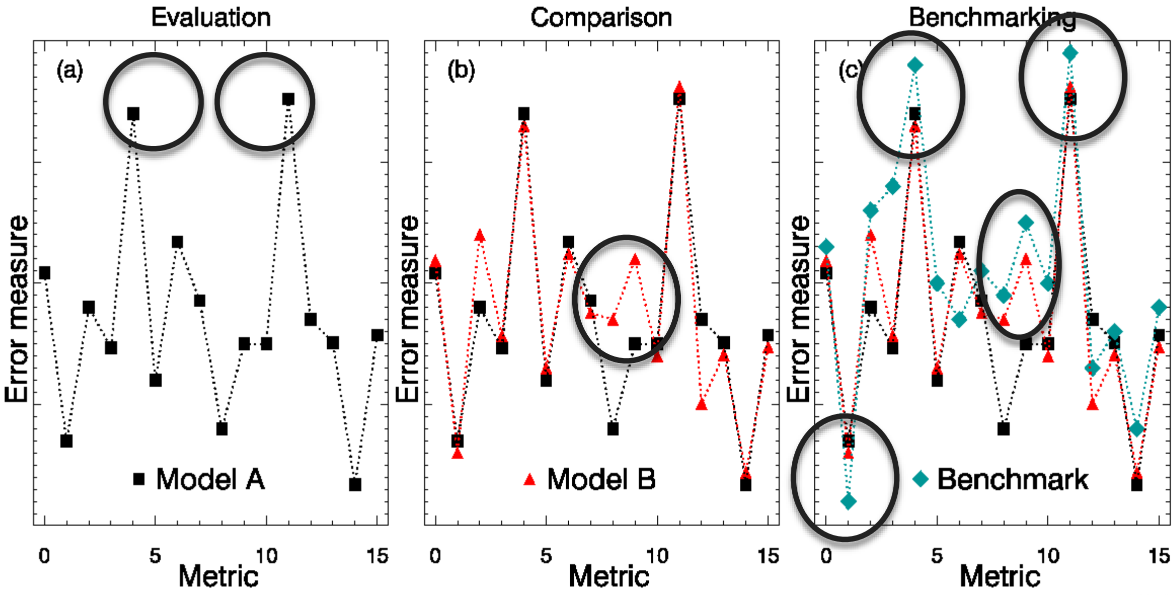
- Evaluation, comparison and benchmarking
- Defining benchmarks
- Existing JULES benchmarks
- Land Validation Toolkit (LVT)
- Examples
- Future plans

Evaluation, Comparisons & Benchmarking

- Evaluation - model outputs are compared to observations to derive an error measure

- Comparison - model is not just compared to observations, but also to other models.

- Benchmarking - performance expectation is defined a priori



Best et al (2015)

Defining benchmarking

There are several ways performance expectations might be defined before running a model:

1. Is it better than another model?

   e.g. set the results from a previous model version as the performance benchmark.

2. Is it fit for a particular application?

   e.g. Can the LSM capture specific impacts

3. Can it effectively utilise available information?

   e.g. If a LSM is given information about vegetation and soil at a location in addition to time varying meteorology it should be expected to perform better than one that is not

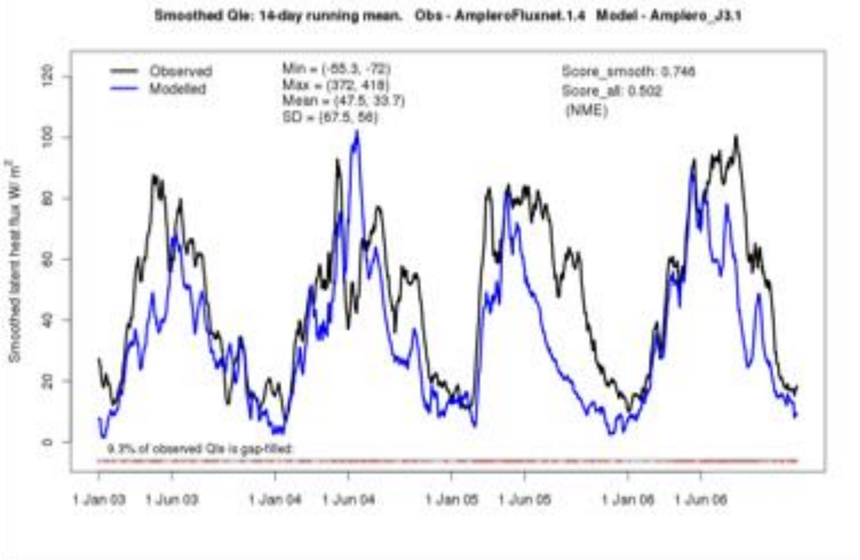
Best et al (2015)

Benchmarking

- Simply comparing models and observations – i.e. “evaluation” – can’t tell us whether any of the models are doing a good job

- Example...

Latent Heat Flux at Amplero



We would typically accept this as a good simulation (good correlation visually)

However, benchmarking will reveal that this is in fact a poor simulation!

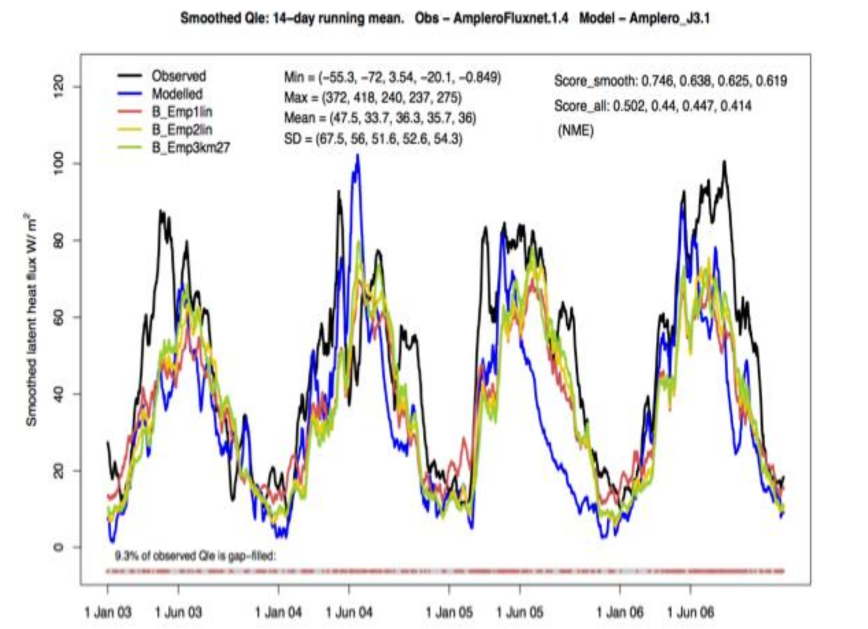
(G. Abramowitz)

Benchmarking example...

- How well should we expect a LSM to predict latent heat (Qle) flux at Amplero site?
  - Take several (19) flux tower sites other than Amplero
  - Train a linear regression between downward shortwave radiation and Qle
  - Use regression parameters to predict Qle at Amplero using site meteorology

This will tell us:

- The extent to which Qle is predictable from SWdown alone.
- How predictable Qle is at Amplero site - is it unusually difficult?



Even the 1-variable regression beats the model!

(G. Abramowitz)

Met Office

Benchmarking for JULES

What is needed?

1. Tests with new developments turned **off**

- Need to check science changes do not break existing code

- JULES Rose stem tests

2. Tests with new developments turned **on**

- Need to check science is performing against previous code

- New benchmarks are required to test model performance

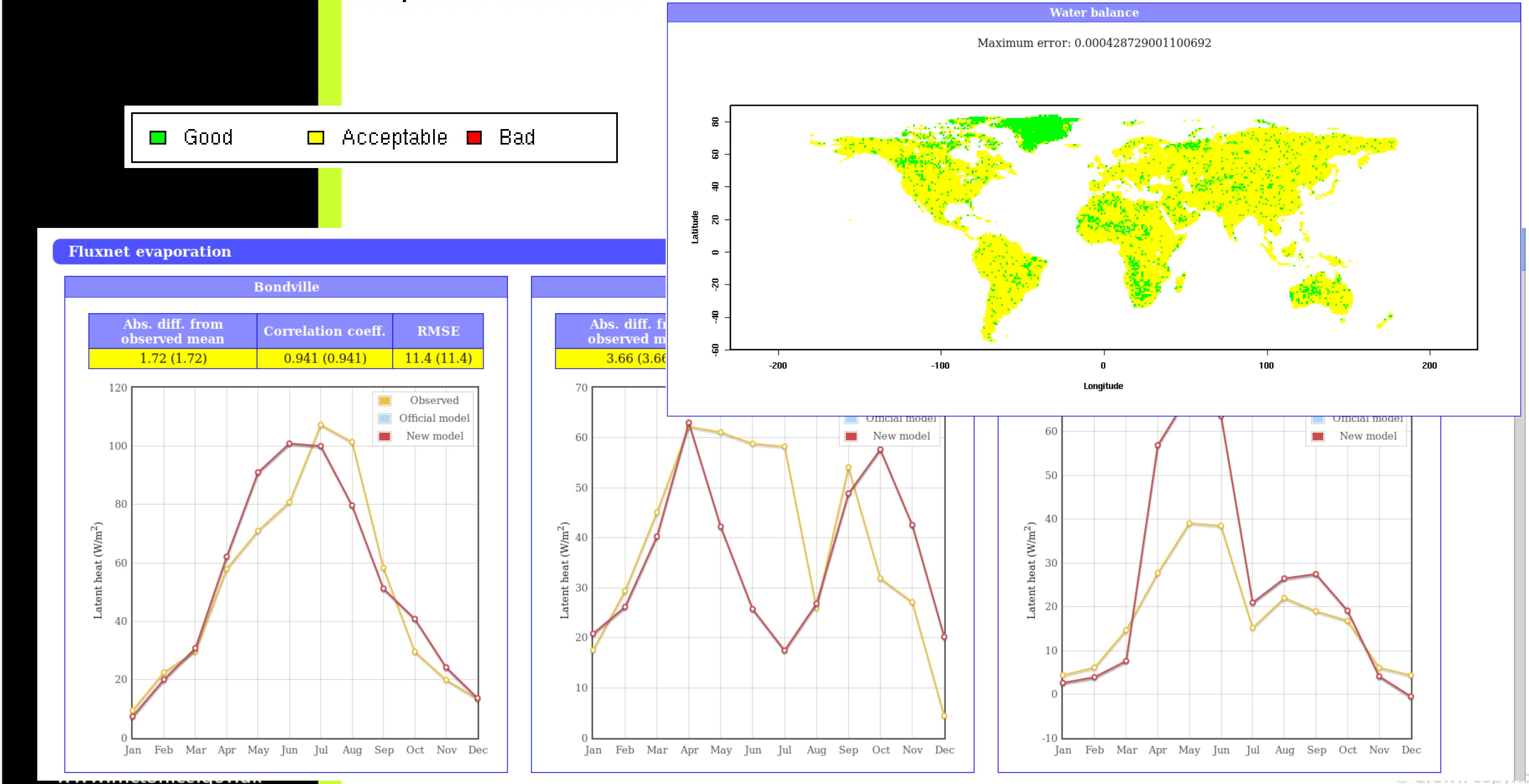
"Ultimate" benchmark – model to be within the 1 observational error of observations!

Existing benchmarking system

- Assessed performance at 10 FLUXNET sites and globally using GSWP2 gridded data.

- Limitations: Only used 10 sites, 1 year for each, didn't check all science aspects of JULES

🟩 Good 🟨 Acceptable 🟥 Bad



Water balance

Maximum error: 0.000428729001100692

Fluxnet evaporation

Bondville

Abs. diff. from observed mean	Correlation coeff.	RMSE
1.72 (1.72)	0.941 (0.941)	11.4 (11.4)

Abs. diff. f observed m
3.66 (3.66

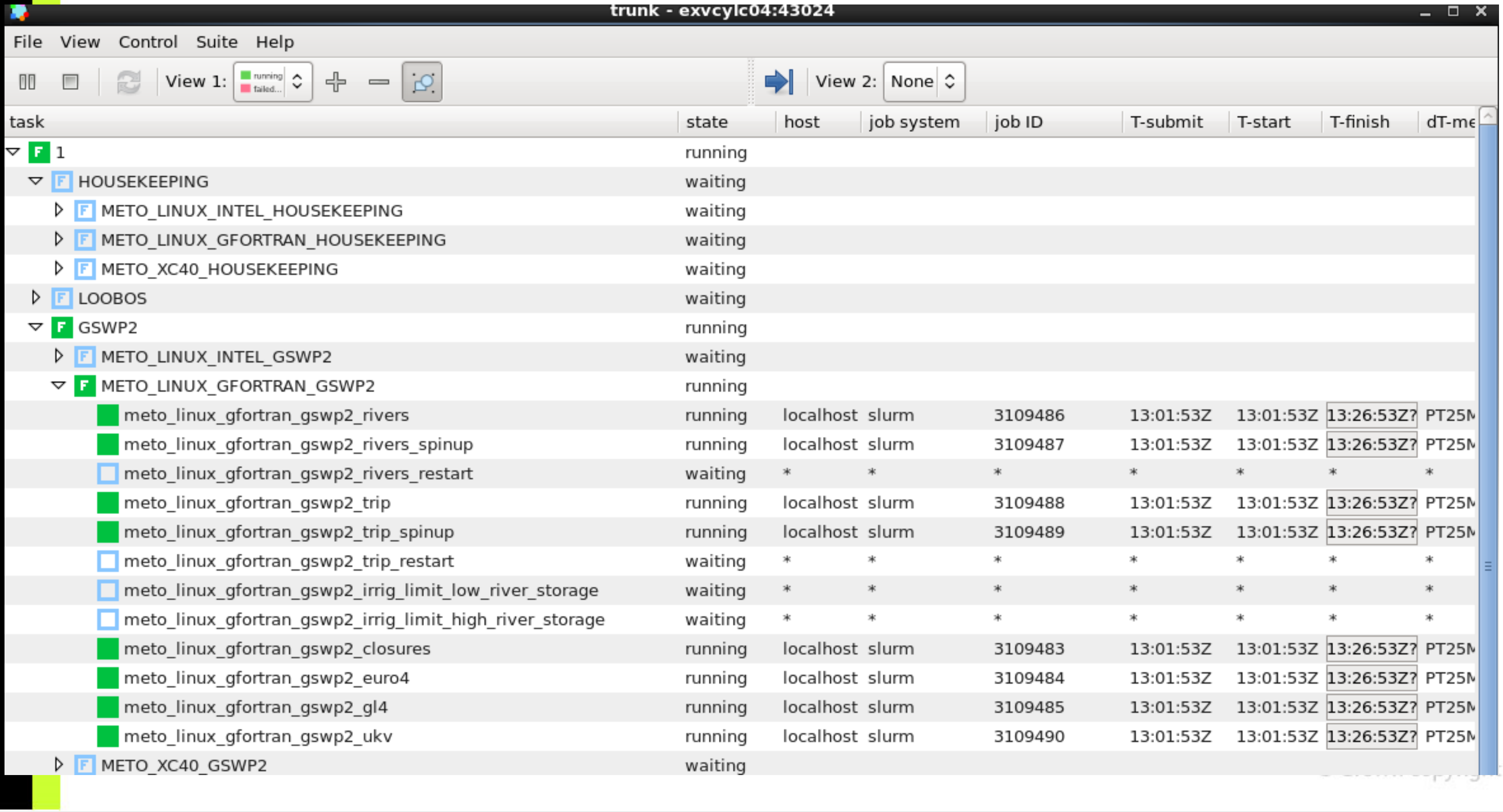
Met Office

Rose-Stem tests

- Makes sure that any code changes do not break (i.e. compromise) any existing science that has a test.

- More tests are being added to provide robustness to the system.

- Rose stem is part of the JULES code and can be run by anyone that has a copy of the code and is running on the Virtual Machine (VM), JASMIN, MONSooN or any other supported site.



task	state	host	job system	job ID	T-submit	T-start	T-finish	dT-me
F 1	running							
F HOUSEKEEPING	waiting							
F METO_LINUX_INTEL_HOUSEKEEPING	waiting							
F METO_LINUX_GFORTRAN_HOUSEKEEPING	waiting							
F METO_XC40_HOUSEKEEPING	waiting							
F LOOBOS	waiting							
F GSWP2	running							
F METO_LINUX_INTEL_GSWP2	waiting							
F METO_LINUX_GFORTRAN_GSWP2	running							
meto_linux_gfortran_gswp2_rivers	running	localhost	slurm	3109486	13:01:53Z	13:01:53Z	13:26:53Z?	PT25M
meto_linux_gfortran_gswp2_rivers_spinup	running	localhost	slurm	3109487	13:01:53Z	13:01:53Z	13:26:53Z?	PT25M
meto_linux_gfortran_gswp2_rivers_restart	waiting	*	*	*	*	*	*	*
meto_linux_gfortran_gswp2_trip	running	localhost	slurm	3109488	13:01:53Z	13:01:53Z	13:26:53Z?	PT25M
meto_linux_gfortran_gswp2_trip_spinup	running	localhost	slurm	3109489	13:01:53Z	13:01:53Z	13:26:53Z?	PT25M
meto_linux_gfortran_gswp2_trip_restart	waiting	*	*	*	*	*	*	*
meto_linux_gfortran_gswp2_irrig_limit_low_river_storage	waiting	*	*	*	*	*	*	*
meto_linux_gfortran_gswp2_irrig_limit_high_river_storage	waiting	*	*	*	*	*	*	*
meto_linux_gfortran_gswp2_closures	running	localhost	slurm	3109483	13:01:53Z	13:01:53Z	13:26:53Z?	PT25M
meto_linux_gfortran_gswp2_euro4	running	localhost	slurm	3109484	13:01:53Z	13:01:53Z	13:26:53Z?	PT25M
meto_linux_gfortran_gswp2_gl4	running	localhost	slurm	3109485	13:01:53Z	13:01:53Z	13:26:53Z?	PT25M
meto_linux_gfortran_gswp2_ukv	running	localhost	slurm	3109490	13:01:53Z	13:01:53Z	13:26:53Z?	PT25M
F METO_XC40_GSWP2	waiting							

Met Office

Some LSM evaluation & benchmarking 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 0.5º x 0.5º 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

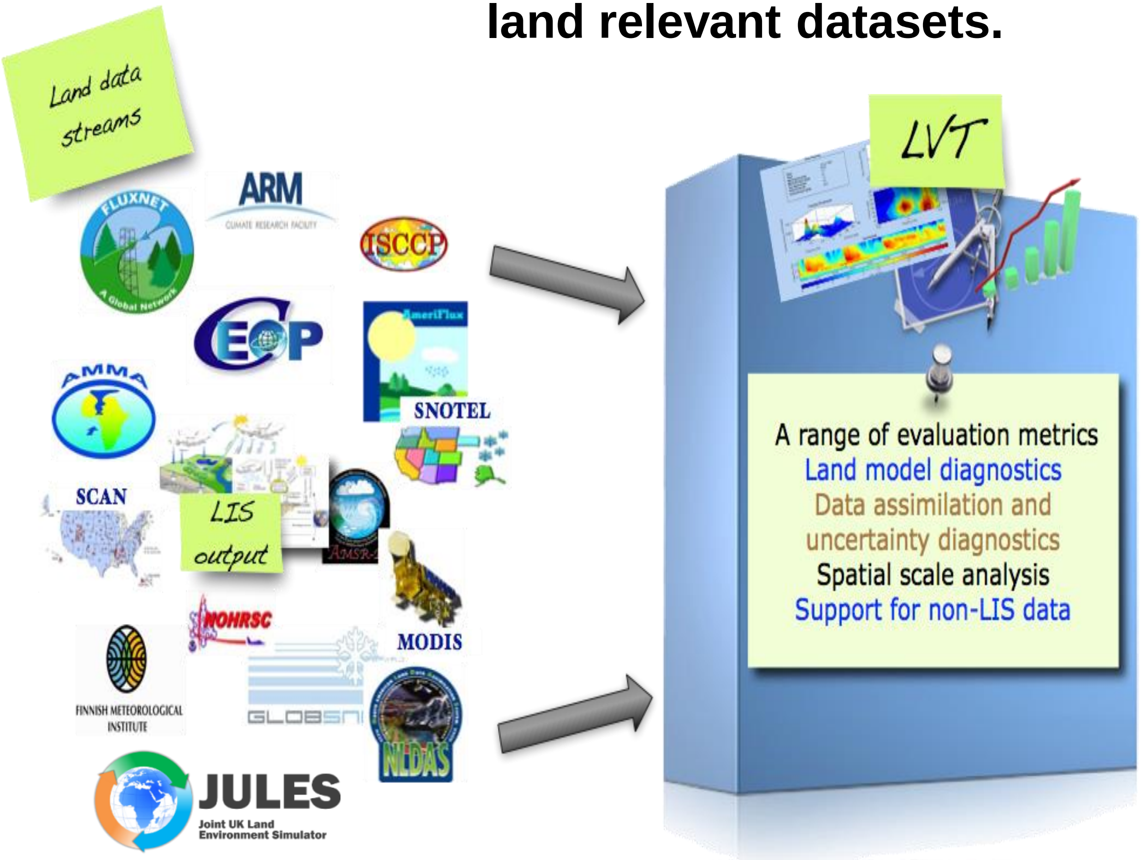
LVT = Land surface Verification Toolkit

Part of NASA LIS (Land Information System). Site or gridded data, any time step, allows for missing data & screening by Quality flag, full range of statistical metrics including 95% confidence intervals.
Kumar et al., 2012, *GMD*, doi: 10.5194/gmd-5-869-2012

G. Weedon (2016) Technical Report. Assessment of available systems for future JULES evaluation and benchmarking

The Land Validation Toolkit (LVT)

- Designed to handle **any two land relevant datasets.**



A range of evaluation metrics
Land model diagnostics
Data assimilation and uncertainty diagnostics
Spatial scale analysis
Support for non-LIS data

- Large range of supported datasets + capability to add bespoke readers for new datasets.
- Completely flexible selection of metrics + capability to add new metrics.
- The supported datasets in LVT can be used to develop benchmarks using simple (regression) to more complex methods.

- Flexibility to carry out analysis at **single sites, regionally and globally** with observations at a wide range of **spatial and temporal scales** as chosen by the user.

Standalone JULES-LVT Rose Suite has been developed

Kumar et al (2012)

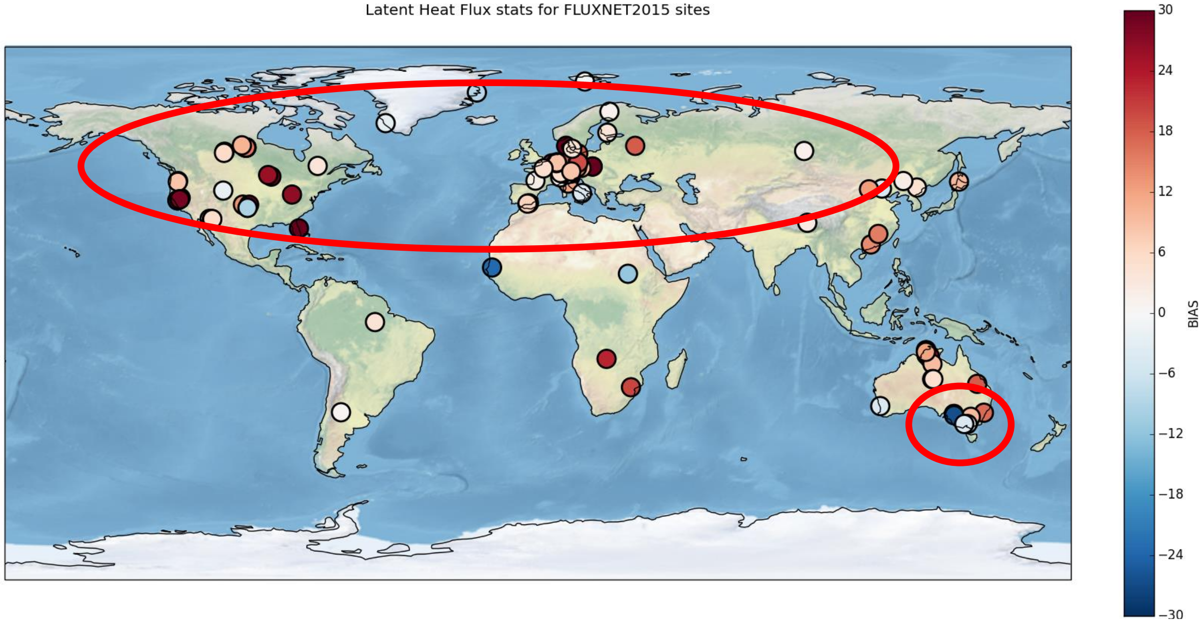
Met Office

Summary Statistics – bias (model minus obs)

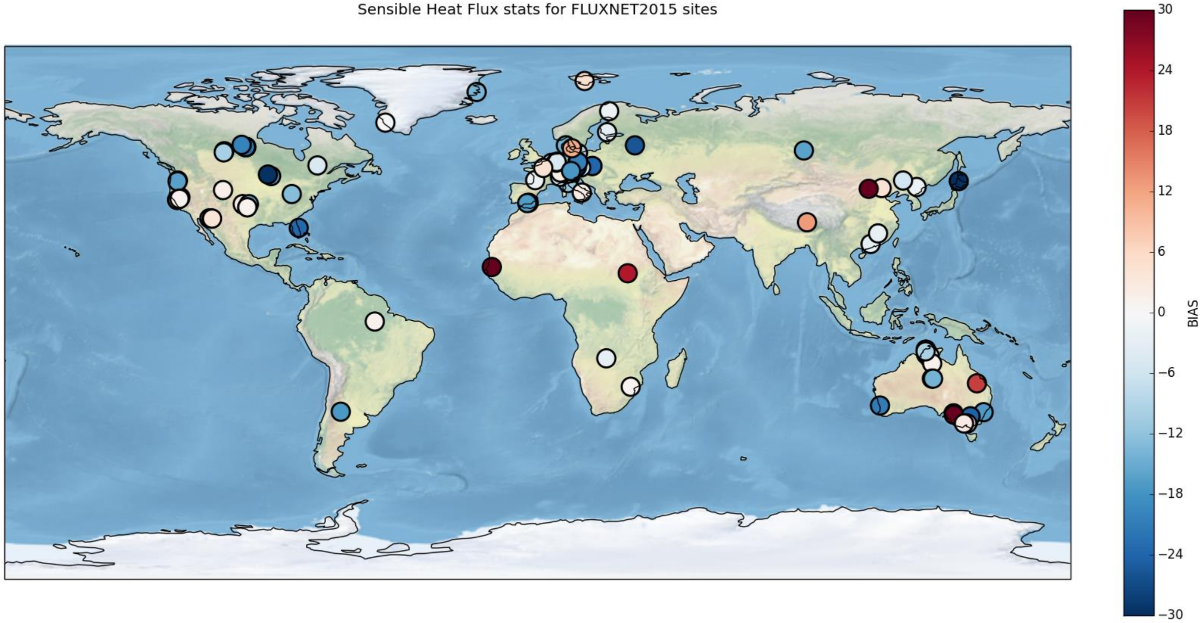
JULES vn4.8, driven with WFDEI, out of the box configuration

JULES vs. FLUXNET2015

Qle


Latent Heat Flux stats for FLUXNET2015 sites

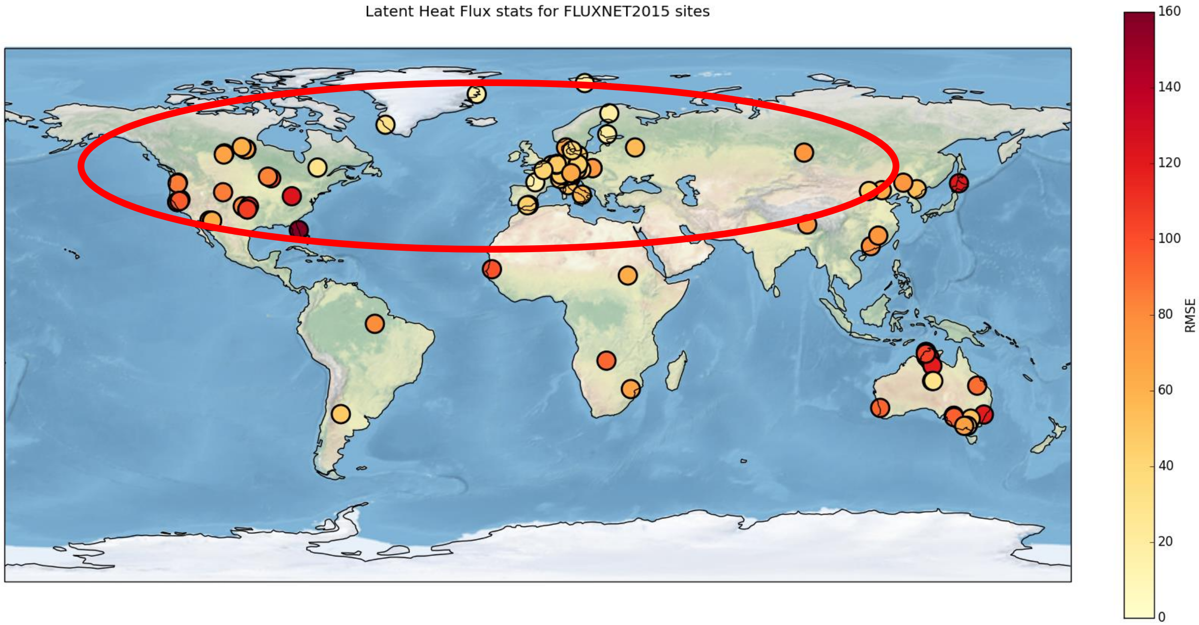
Qh


Sensible Heat Flux stats for FLUXNET2015 sites

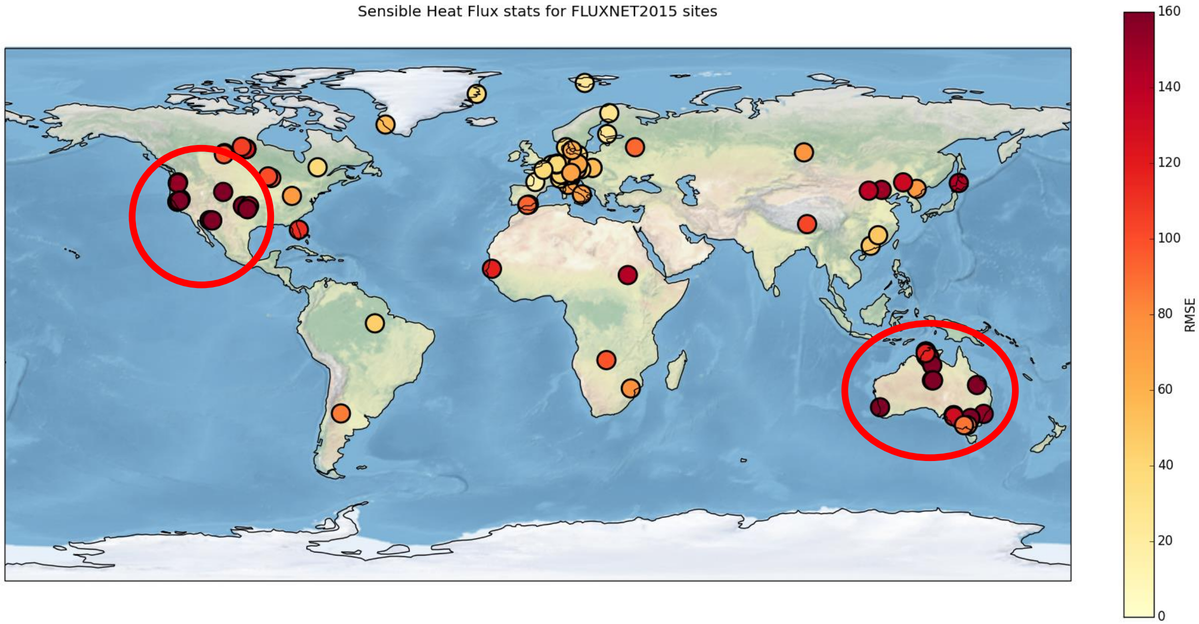
JULES vs. FLUXNET2015

Summary Statistics – RMSE

Qle



Qh



Future Plans

- **Aim** - Develop a fully comprehensive benchmarking suite
- Complete analysis for all four fluxes:
  - Energy, water, carbon and momentum
- Capability to extend to other variables:
  - Soil moisture, LST's, albedo, LAI/NDVI
- Utilise a wider range of observation data including:
  - NRFA stream flows, GRACE, point scale groundwater
  - + ....?
- Enable community contributions

Conclusions

- Evaluation is still a valuable tool for identifying model development needs.

- However, the wider use of benchmarking is likely to identify the more serious challenges in land surface models and accelerate our improvements in the science.

- We are developing a comprehensive benchmarking suite for JULES using NASA's Land Validation Toolkit

- Hoped that the community will adopt this approach in the future, to be used in combination with existing evaluation and comparison tools.

Met Office

Any questions?