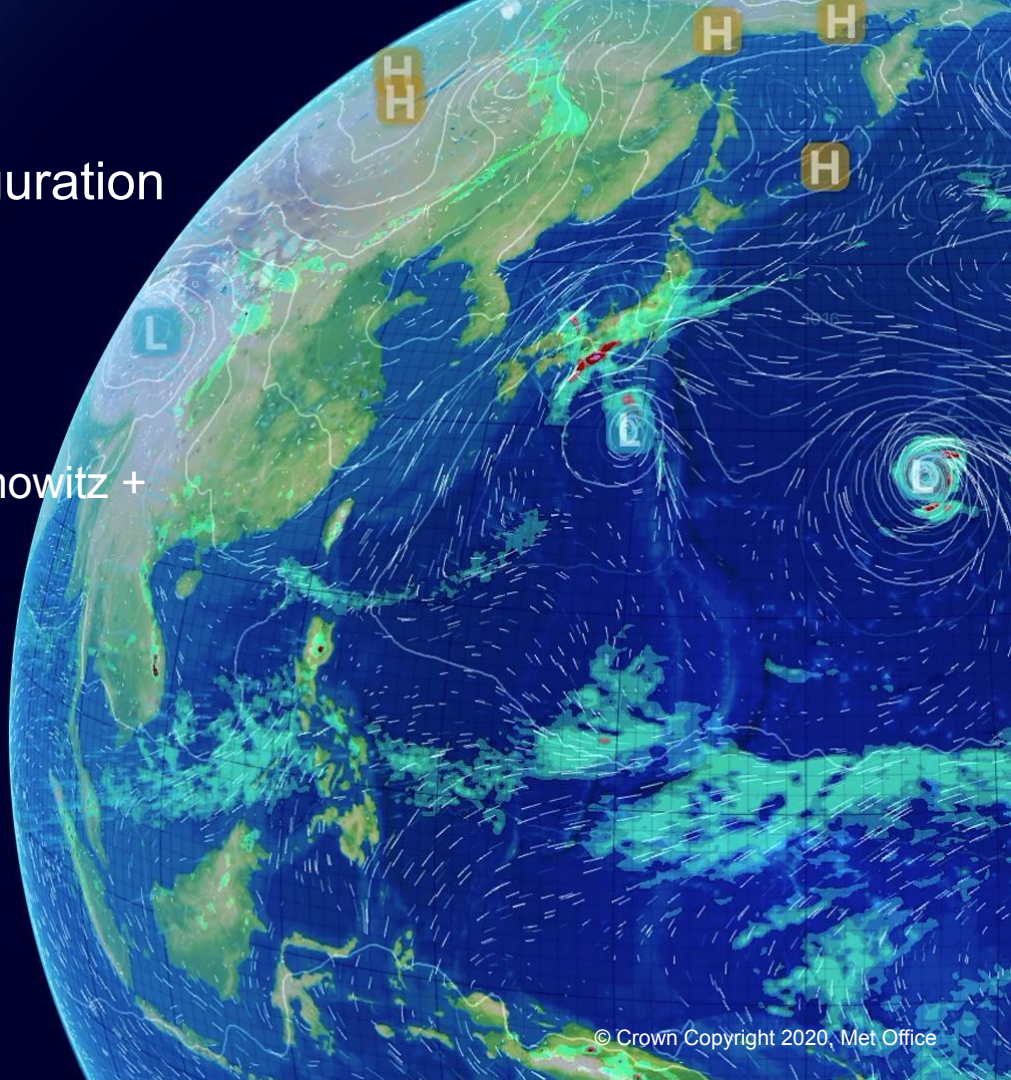
Developing the next standard configuration for standalone JULES using a benchmarking system based on ModelEvaluation.org.

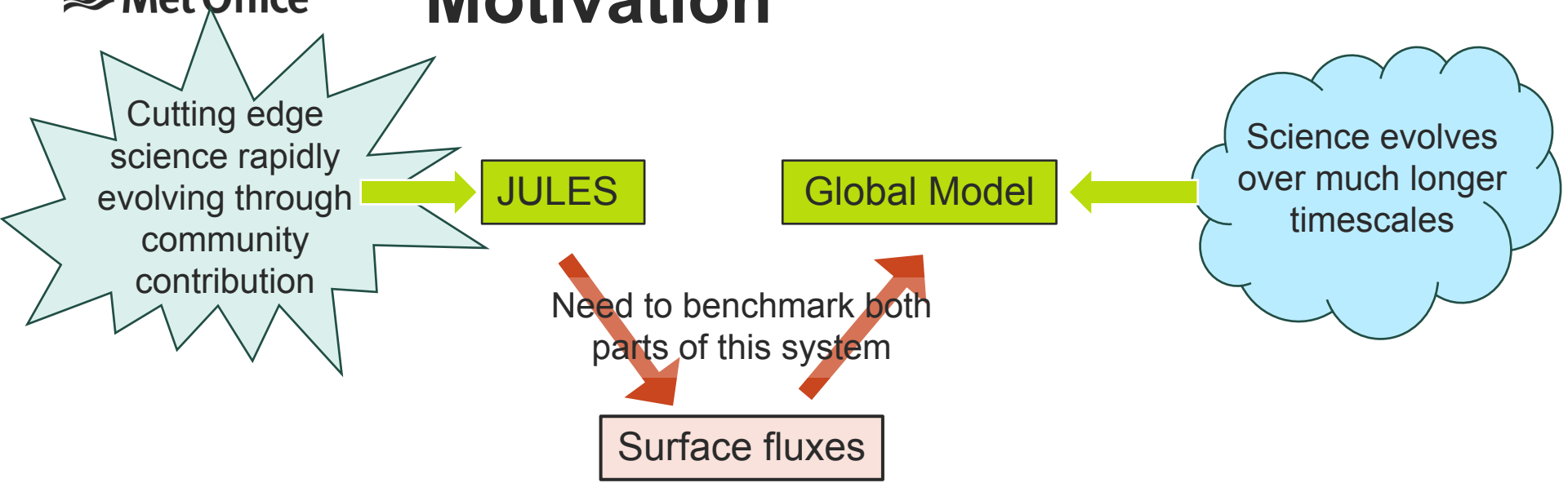
Heather Rumbold, Martin Best, Gab Abramowitz + JULES Community

Seamless Global Modelling Workshop

Thursday June 5th, 2025

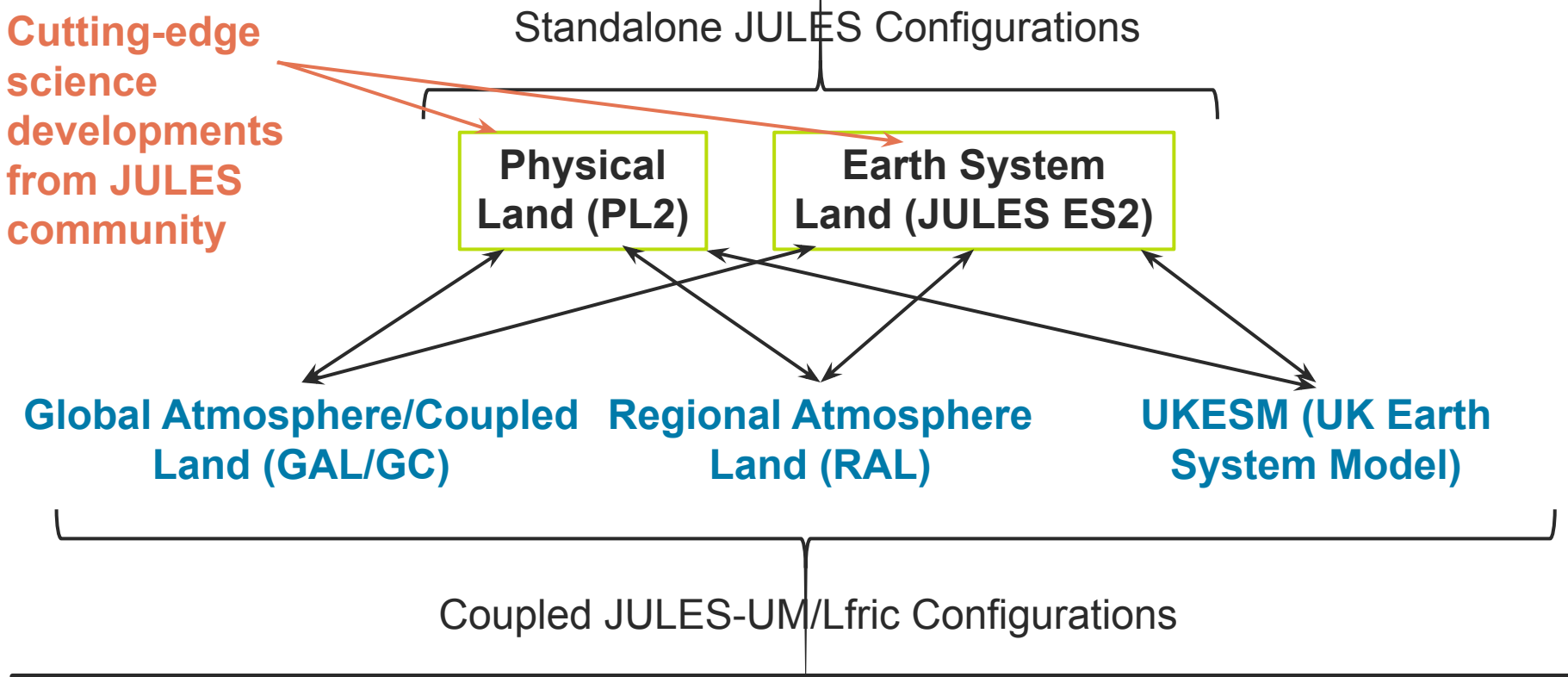


Motivation



- Standalone JULES configurations are evaluated and benchmarked by observations.
- This allows us to isolate the scheme and develop the best configurations for the land.
- However, compromises must be made when science is pulled through to global model configurations.
- Can we use a standalone benchmarking system to identify errors to improve the atmospheric model?

Configurations required for Standalone JULES



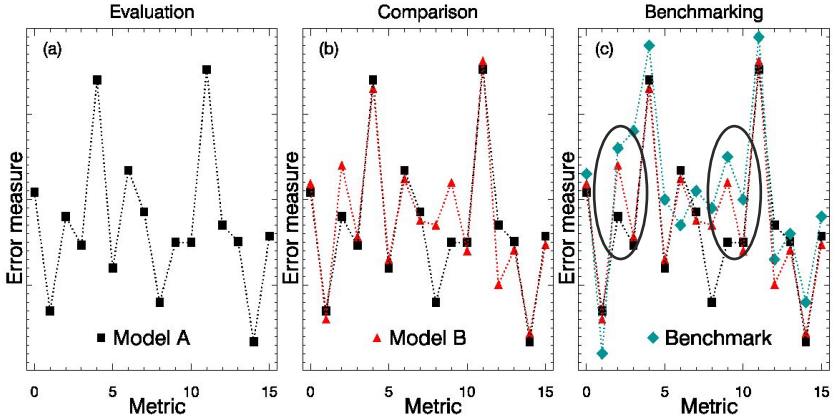
Aim: To create **standard configurations** that give the **best simulation** of the **physical and earth system environments**, demonstrated through **predefined benchmarks.**

What is benchmarking?

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,[c] 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]

(2015) Journal of Hydrometeorology, 16, 1425-1442.



Model outputs are compared to a predefined benchmark.

3 types of benchmark:

1. Is it better than another model?
2. Is it fit for a particular application?
3. Can it effectively utilise available information?

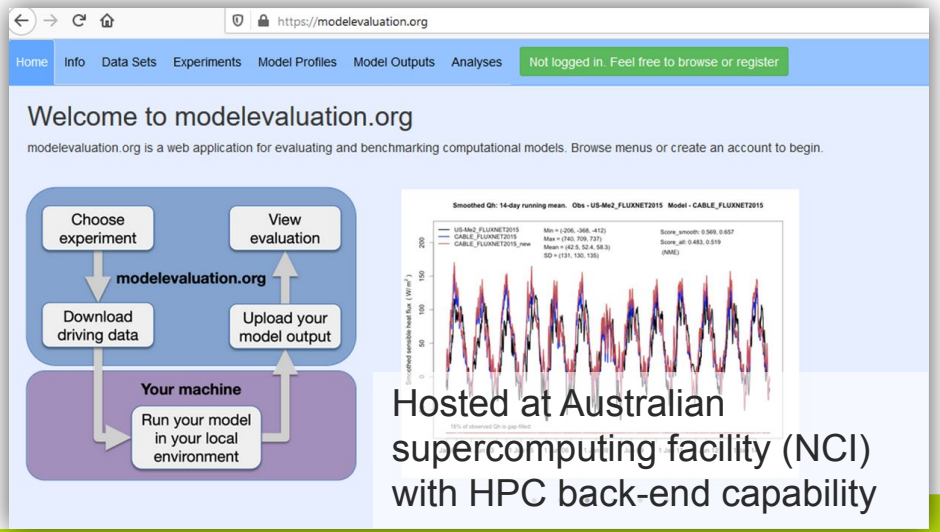
e.g. previous model configuration
Does adding new science code improve JULES compared to the previous configuration?

ModelEvaluation.org

- **Community web-based environment** for model evaluation and benchmarking.
- Users run their **land models offline** in single-site mode, forced by **locally observed sub-daily meteorology** (downloaded from ME.org), for a **wide variety of sites** across the globe.
- Model output is then **uploaded to the web application** and **analysed** using scripts that compare output with evaluation **data products, other models and empirical or ML benchmarks**.

Advantages:

- System shared by many international research groups, providing opportunities to benchmark against many other land surface models.
- Able to utilise standard testing experiments such as linear regression and machine learning models.
- Can utilise an extensive range of metrics and statistics.
- Use of consistent, reliable and robust, quality-controlled data for many sites.



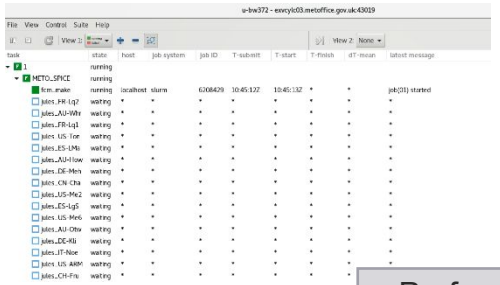
PLUMBER2 Benchmarking Workflow

PLUMBER2 dataset
170 sites from FLUXNET2015, FLUXNET La Thuile & OzFlux
+
canopy height, LAI reference height & IGBP vegetation
+
HWSD soils

Python script → convert jules input variables into json file


```
{
  "AR-SLu": {
    "data_start": "2010-01-01 00:00:00",
    "data_end": "2011-01-01 00:00:00",
    "data_period": 1800,
    "drive_file": "/data/users/hashton/PLUMBER2/met_f
    "latitude": -33.4648,
    "longitude": -66.4598,
    "spinup_start": "2010-01-01 00:00:00",
    "spinup_end": "2011-01-01 00:00:00",
    "main_run_start": "2010-01-01 00:00:00",
    "main_run_end": "2011-01-01 00:00:00",
    "timestep_len": 1800,
    "z1_tq_in": 11.0,
    "z1_uv_in": 11.0
  },
```


Rose workflow
Run JULES for all sites in json file

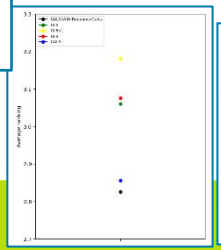


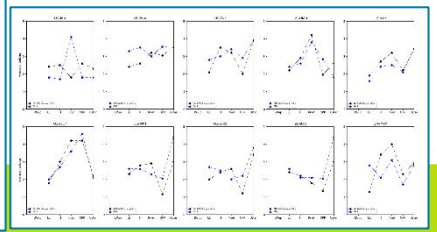
Calculate output metrics*:
JULES vs. Obs for every site, for LWup, LE, H, Rnet, GPP & Ustar, for each configuration

Configuration/ benchmark are ranked for each metric, var, site

Perform the average ranking of metrics for every site and variable

Plot output and document ticket





Benchmark configurations:
Standalone versions of GL8, GL8.1, GL9 & GAL9

* Metrics: RMSE, MBE, NME, StdDevdiff, Correlation, 5thdiff, 95thdiff, PDFoverlap, Skewness, kurtosis

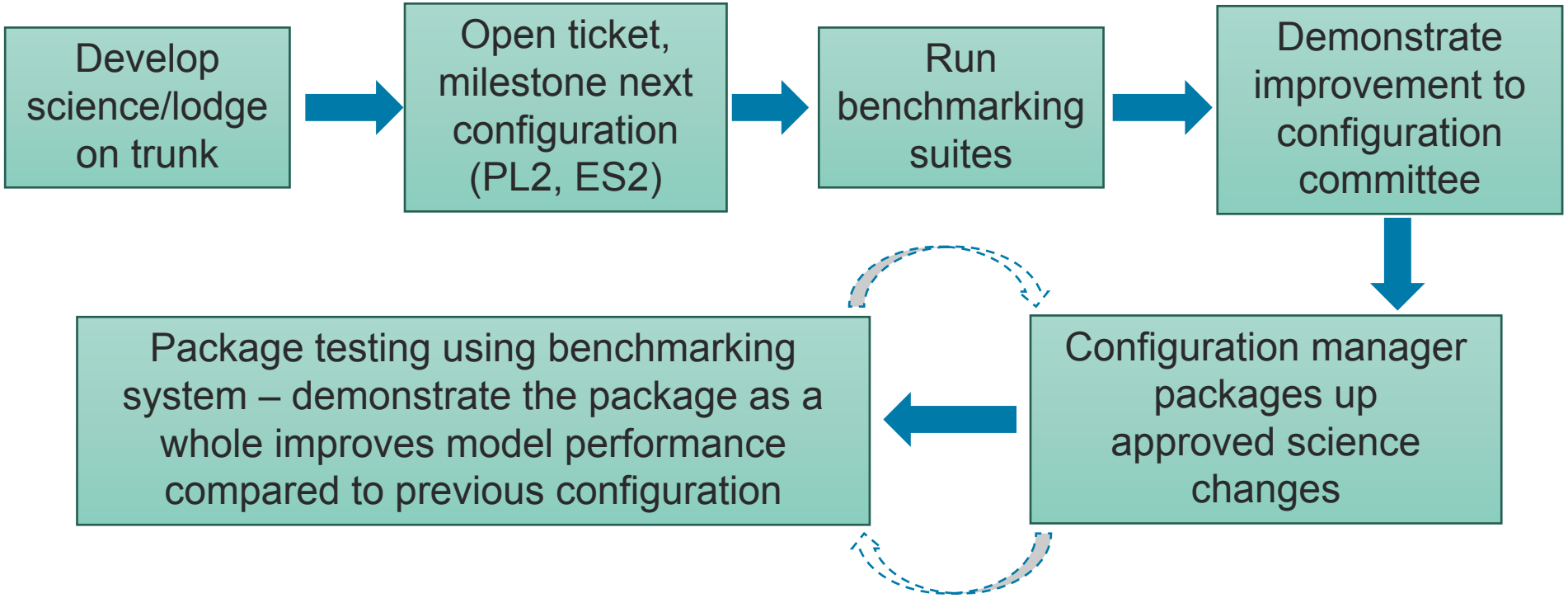
Analysis

- Compares JULES vs Obs for every **site**, **variable**, **model configuration** and **model benchmarks.**
- 10 Statistical metrics are calculated for each instance.

Sites	170 FLUXNET sites, global coverage, with site metadata AND evaluation data
Metrics	RMSE, MBE, NME, StdDevdiff, Correlation, 5thdiff, 95thdiff, PDFoverlap, Skewness, Kurtosis
Variables	LE, H, LWup, Rnet, GPP and momentum flux
New Model Configuration	PL2 package test
Benchmarks	GL8, GL8.1, GL9 & GAL9

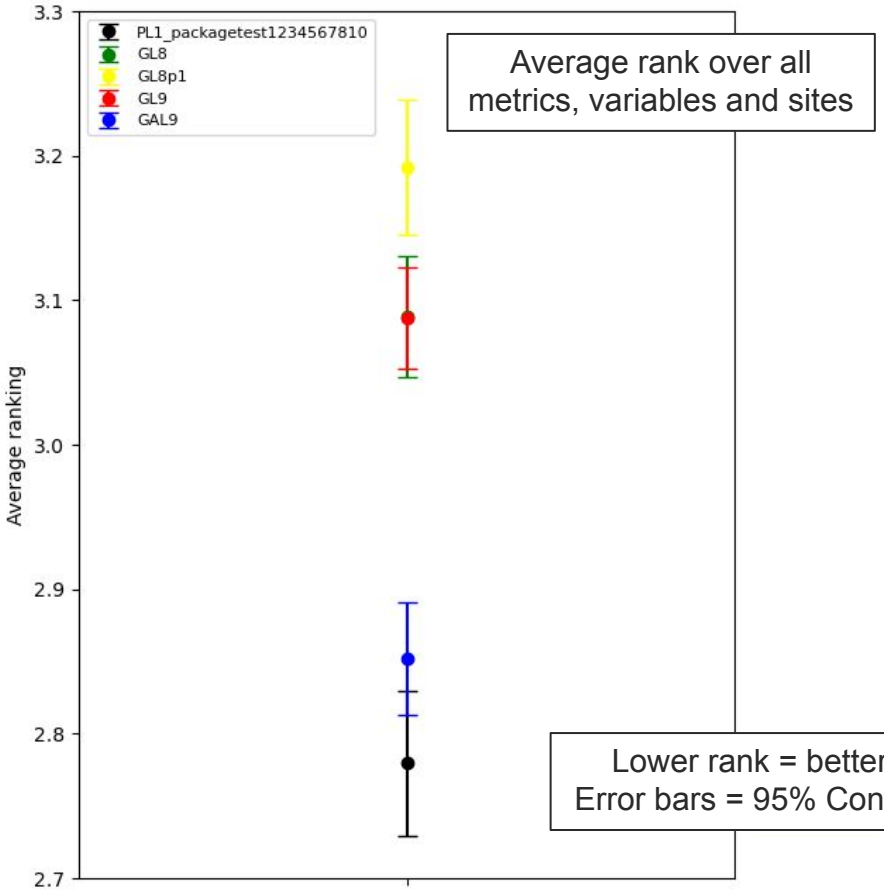
- Metrics values are ranked from 1 to 5 (1 being the best result)
- These rankings are averaged over all statistics and all sites to give an average ranking for all the variables separately.
- A final averaging is performed over all variables to give an overall ranking.

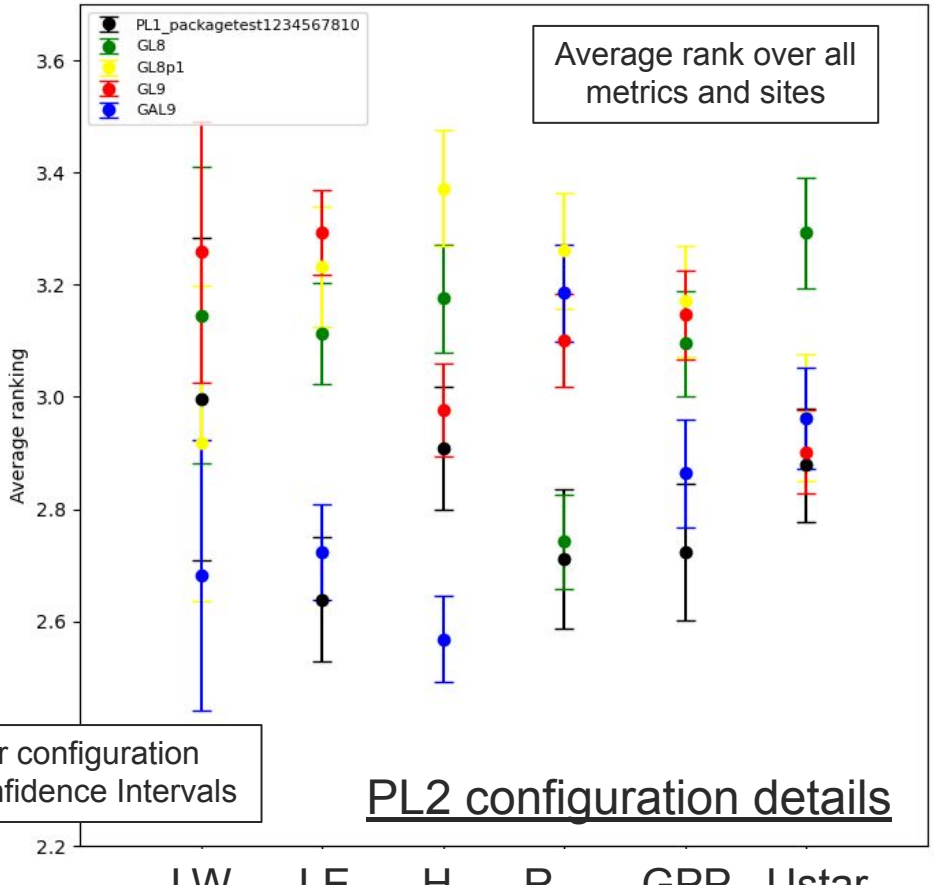
Development Process for standalone JULES configurations



- Same process, as used by regional and global modelling, for the JULES community
- First cycle has been completed… configurations are now ready to release

Physical Land Package: PLUMBER2 benchmarking output

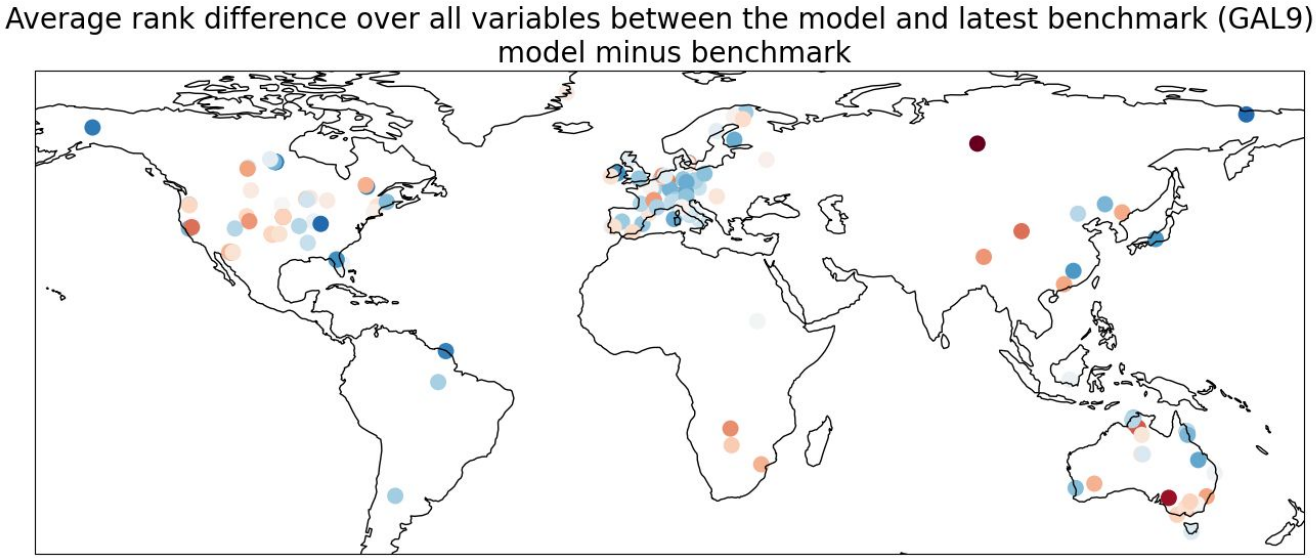




Lower rank = better configuration
Error bars = 95% Confidence Intervals

PL2 configuration details

Average rank over all metrics and variables

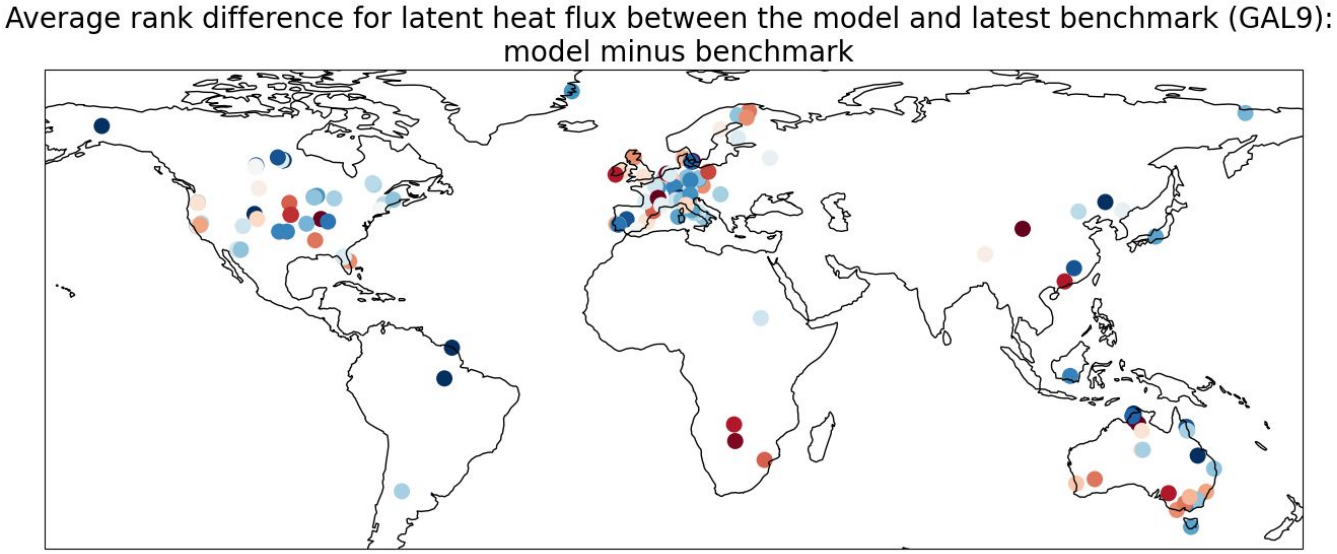


Model: JULES PL2. Benchmark: GAL9

If model rank is > benchmark rank, then benchmark is doing better than model

Blue = better performing model i.e. PL2 configuration

Average rank over all metrics only: Latent heat flux

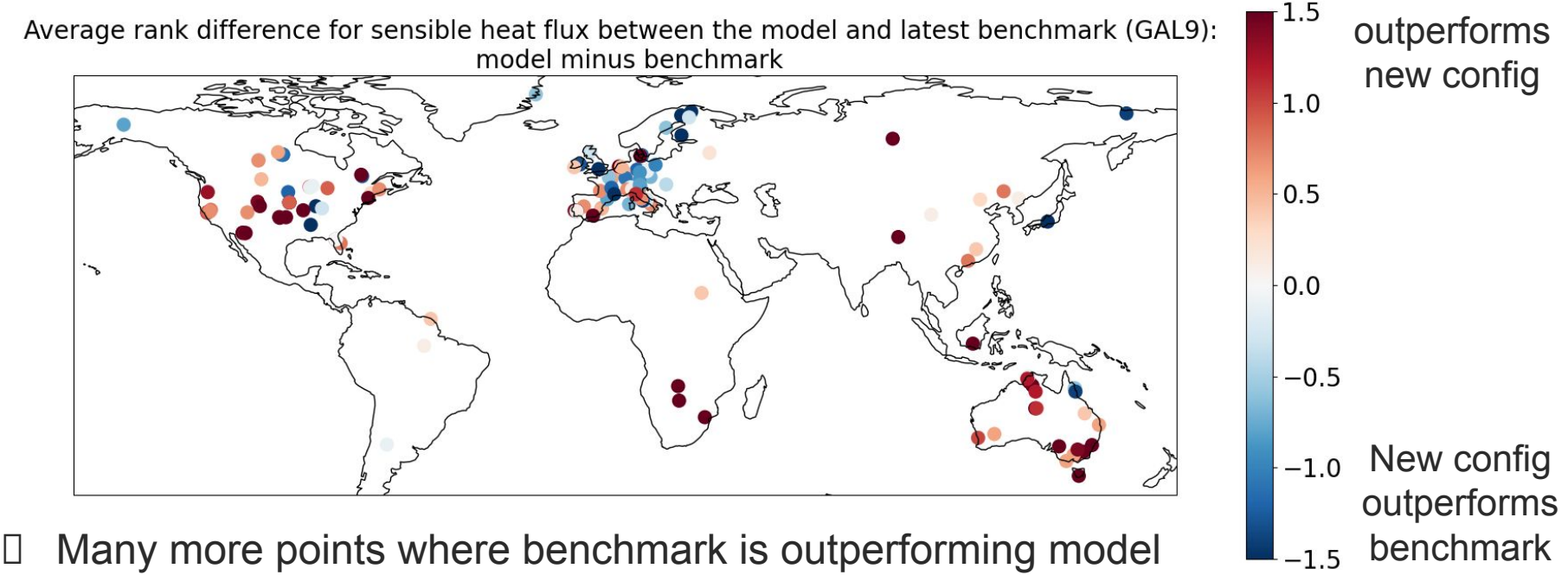


Benchmark outperforms new config

New config outperforms benchmark

- More points where benchmark is outperforming model

Average rank over all metrics only: Sensible heat flux



- Many more points where benchmark is outperforming model
- Highlights the need for a more rigorous system which explores the data space in more detail…

Met Office

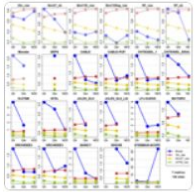
Uses Machine Learning models as our benchmarks - i.e. use ML to predict fluxes using meteorological variables as predictors

See seminar by Gab Abramowitz: “Machine Learning for Benchmarking Land Models”, 21st May, here

https://doi.org/10.5194/bg-21-5517-2024

Research article | 12 Dec 2024

On the predictability of turbulent fluxes from land: PLUMBER2 MIP experimental description and preliminary results



Gab Abramowitz, Anna Ukkola, Sanaa Hobeichi, Jon Cranko Page, Mathew Lipson, Martin G. De Kauwe, Samuel Green, Claire Brenner, Jonathan Frame, Grey Nearing, Martyn Clark, Martin Best, Peter Anthoni, Gabriele Arduini, Souhail Boussetta, Silvia Caldararu, Kyeungwoo Cho, Matthias Cuntz, David Fairbairn, Craig R. Ferguson, Hyungjun Kim, Yeonjoo Kim, Jürgen Knauer, David Lawrence, Xiangzhong Luo, Sergey Malyshev, Tomoko Nitta, Jerome Ogee, Keith Oleson, Catherine Ottlé, Phillipe Peylin, Patricia de Rosnay, Heather Rumbold, Bob Su, Nicolas Vuichard, Anthony P. Walker, Xiaoni Wang-Faivre, Yunfei Wang, and Yijian Zeng

- Simulations of latent, sensible heat, carbon fluxes at 154 flux tower sites.
- 20 different international land models.
- Linear regression, cluster and regression, LSTM, random forest as benchmarks – all out of sample
- Land model provided with observed vegetation type, reference and canopy height, LAI and 30-60min meteorological forcing. No calibration.
- How predictable are the surface fluxes? How good are existing mechanistic models at prediction? Under which conditions/environments do they excel, or we know they can improve?
- Results are broadly consistent: **Simple out-of-sample empirical models, including linear regression, comfortably outperforming mechanistic land models.**

Conclusions

- The **land component** can now be **formally assessed in isolation**, so we have a clearer **understanding** of the science processes.

- We now have a **benchmarking system** that runs very quickly giving us the ability to **assess sites and variables** individually for separate **individual science changes**

- **Pull through** of new science in the global model configurations will allow us to understand the **impacts of land science changes on the global model outputs**.
  - **Compromises** will still be needed to account for biases in the atmospheric model.
  - However, we are now able to repeat local assessment of the impacts of such compromises.

- **PLUMBER2 MIP** has provided lots more information which can be used to explore the **where**, **how** and **why** our land models aren't performing well and identify where it can be improved.
  - Provide the **most accurate fluxes** back to the global model.

- **Benchmarking** is critical to **target challenges in land modelling** and therefore **accelerate improvements** in the science across all scales.

Met Office

Thank you for listening

Any Questions?

heather.rumbold@metoffice.gov.uk