Developing the next benchmarking system for JULES based on ModelEvaluation.org

Heather Rumbold, Martin Best, Gab Abramowitz, Adrian Lock
Configuration Manager for the Global Land

My role…

• Maintain the standalone physical land model configuration versions on both the Met Office and NERC systems.
  • GL9 standalone is being finalised on Jasmin, ready for use shortly

• Build and maintain the comprehensive benchmarking system that will be used to assess new components for future configurations.
  • Generating a new benchmarking tool using ModelEvaluation.org for use along side existing tools
JULES Standard Configuration

https://code.metoffice.gov.uk/trac/jules/wiki/JulesConfigurations

- Set of model and ancillary generation switches and parameter values
- GL/RL/ES are all required for the coupled system.
- Standalone only requires configurations for:
  1. Physical Land (weather/climate)
  2. Earth System
JULES Standard Configuration

A configuration is not:
- Driving data
- The resolution
- The ancillary files (but can include the data sources)
- Application specific

Best combination of settings to give the best description of the physical environment

How do we know when we have a better description of the environment?
What is benchmarking?

➢ 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?

“Ultimate” benchmark – model to be within the observational error

The Plumbing of Land Surface Models: Benchmarking Model Performance


What will benchmarking do for JULES?

• JLMP – Require a single configuration which generates the best simulation of JULES as a whole system
  – Is the new JULES configuration better the previous model configuration? (i.e. no 1)
  – E.g. Does adding X piece of new science code improve JULES compared to the previous configuration?
  – Old configuration version will become the benchmark

• JULES community are aiming for 2 or 3 – Best science for a specific area
  – E.g. Can the new configuration capture specific impacts (e.g. the river flow or snow depth) better than the old configuration?
  – E.g. If supplied with better inputs (e.g. high resolution veg ancillaries) it should be expected to perform better than a configuration without this.
A new benchmarking suite

PLUMBER2 data
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

Rose suite Run JULES for all sites in json file

Upload data to modelevaluation.org

Benchmarking output

Perform analysis

Coming soon… upload automated within the suite
Welcome to modelevaluation.org

modelevaluation.org is a web application for evaluating and benchmarking computational models. Browse menus or create an account to begin.

Choose experiment

View evaluation

Download driving data

Upload your model output

Your machine

Run your model in your local environment
Single site vs Observations
Multi site analysis

- Each model is ranked according to every metric, variable and site
- Ranks are averaged to give a single value....
- 3 models: JULES GL8, 1 var linear regression, 2 var linear regression

0 = Perfect model
1 = Worst model

JULES is better than the linear regression models!
Multi site analysis

- Variable breakdown...
- JULES LE beats the linear regression models
- H does not, however:

quantile value =
(highest rank – JULES model)/
(lowest rank – highest rank)

- H isn’t as bad as it looks!
- Overall JULES is as good as or better than the benchmarks
Multi site analysis

Breakdown by site

Overall there are more sites where we are doing well compared to the empirical benchmarks
Development Process for standalone Physical Land configurations (work in progress)

1. Develop science/lodge on trunk
2. Open ticket, milestone next configuration (JCX?)
3. Run benchmarking suite
4. Demonstrate improvement to configuration committee (who?)

Package testing using benchmarking system – demonstrate the package as whole improves from previous configuration

Configuration manager packages up

Benchmarking is an important part of this process!!
How does this fit in with other tools going forward?

Is the new JULES configuration better than the previous model configuration?