

An international land surface model evaluation  
protocol

Gab Abramowitz  
Climate Change Research Centre, UNSW

# Evaluation should include model standards

Model has technical documentation	Model has no technical documentation
Technical documentation matches what is in the model code	Technical documentation related to what was in the code 5 years ago
Model is open source, community oriented and has hundreds of users	Model is only used by a few people in one organisation
All development of the model is contained in a version control system	Individuals maintain and manage multiple versions in home directories/desktop
Model has a clear user interface and user guide	Model has no user guide and no specific interface
Code is clearly commented, and logically structured	Code is not commented at all and structure is ad hoc
Variable names are consistent throughout the code and relate to their function	Variable names change in each subroutine call and are meaningless
Model changes meet prescribed	Changes are accepted purely on

# Why do we need coordinated model evaluation in LS?

- Evaluation procedures are often limited, ad-hoc, and seen as a matter of personal preference.
- Acceptable standards vary as a function of individuals, workloads etc.
- We can actually do better than “Matches observations well” and “better than the previous model”.
- Comparisons of models are limited to the set of tests included in “intercomparison” experiments.
- Many groups duplicate efforts to develop very similar evaluation programs
- By using a common framework we can consider a wider range of metrics

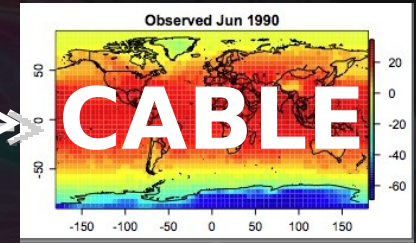
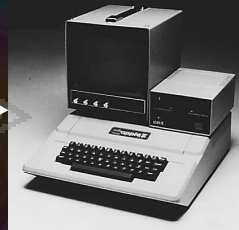
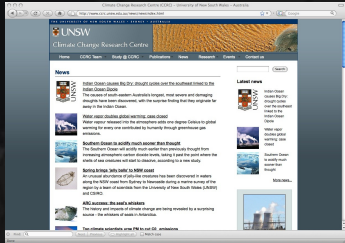
# What the evaluation protocol is and aims to achieve

A web-based server is being built which provides:

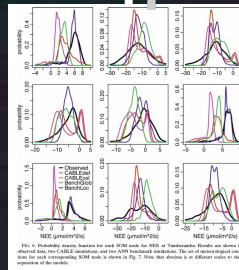
- A broad set of standardised performance measures
- Benchmark levels of performance in these measures
- Standardised, maintained, version controlled observational/synthetic data sets for evaluation.
- An ongoing model comparison experiment using a very wide range of performance measures
- A fast, detailed and free evaluation procedure for model developers
- A quantitative measure of independence between participating LSMs – in which circumstances to LSMs misbehave in the same way?
- Model uncertainty assessment based on the accumulation of submissions
- Raise evaluation standards – any publication using a LSM

# The process for a modeller

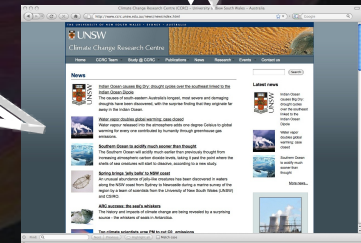
standard  
experiment  
setups, driving  
datasets etc



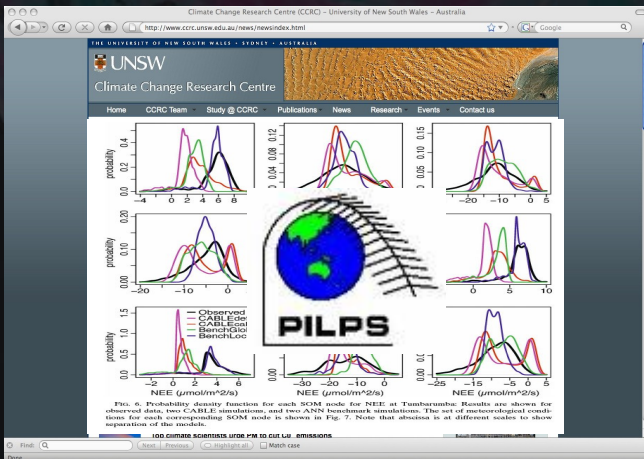
results  
returned  
to user



upload  
model  
simulations



ongoing model  
comparison  
experiment

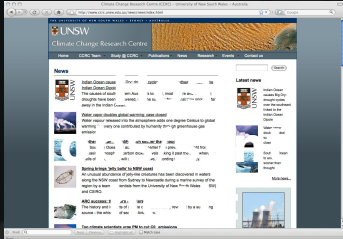




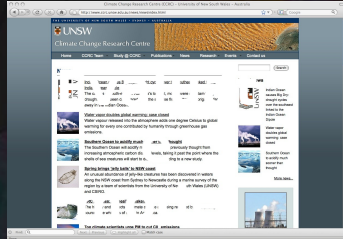
# How it will work

- Two broad categories of users – model users and data providers
- Model users:
  - can submit model runs as often as they like
  - must tag model output to refer to experiment/driving data version info
  - will be encouraged to provide model code that produced run (enforced?)
  - must nominate one or more of their submissions as ‘public’ (to be downloaded by any relevant data providers) and for model comparison
  - must provide meta data on submission: parameter choice, initialisations, structural choices to aid interpretation by others
- Data providers:
  - can update their data sets online
  - must provide meta data: processing, gap-filling, interpolation etc
  - are free to download ‘public’ model runs forced by their data

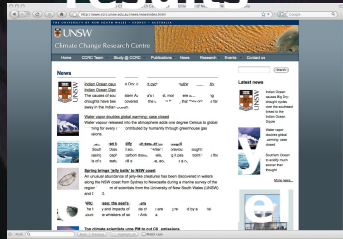
# Roles of the benchmark/protocol server



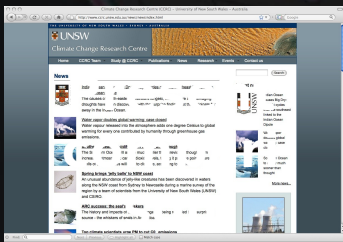
**data**



**results**



**data**



- Data providers able to maintain data sets, version control enforced.
- Format conversion likely needed to protocol netcdf standard.
- Run evaluation scripts; return plots
- Prescribe benchmarks/ produce benchmark empirical model runs
- Probabilistic evaluation techniques can consider collected submissions as an ensemble simulation
- Nominated submissions can be used to show model improvements over time
- Protocol website will host comparison of all participating models in each evaluation measure

# Who's interested?

- The GEWEX Global Land Atmosphere System Study (GLASS) panel
- NCAR – CLM through Gordon Bonan
- UKMO – JULES/MOSES through Martin Best and Mat Williams
- LSCE – ORCHIDEE through Soenke Zaelle
- CAWCR – CABLE through Yingping Wang and Bernard Pak
- Bart van den Hurk (KNMI)
- Dennis Baldocchi (UC Berkely)
- Markus Reichstein (Max-Planck)
- Steve Running (U Montana)
- Dario Papale (U Tuscia)
- Sonia Seneviratne (ETH Zurich)
- Andrew Richardson (Harvard U)

# Where to from here?

- GLASS panel benchmarking meeting, Exeter, June.
- First step likely to be Fluxnet single site data for process evaluation
- Submission protocols need to be developed
- Programmer started building last week

# Questions

- How can we motivate people to participate?
- Any suggestions about its implementation?
- Any difficulties we're likely to encounter?