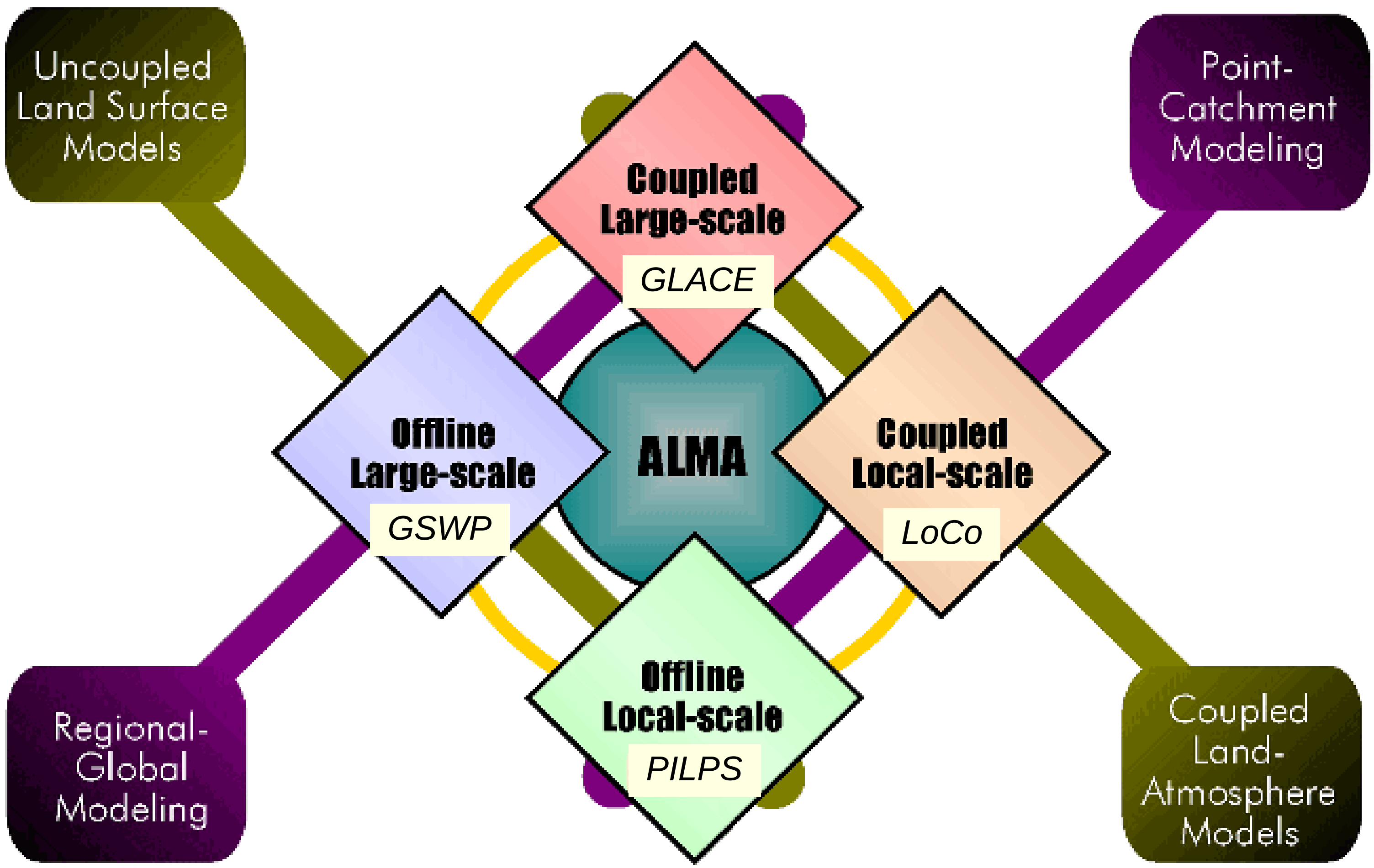
Land surface models benchmarking: offline validation and verification in NWP

ECMWF/GLASS workshop

M.J. Best, November 2009

Previous GLASS structure



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” or “better than the previous version of the 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

Courtesy of Gab Abramowitz

Not just science - model standards are still an issue:

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 performance/realism/functionality checks	Changes are accepted purely on the basis of personal preference.

Courtesy of Gab Abramowitz

C-LAMP – The Carbon-Land Model Intercomparison Project

Forrest Hoffman, James Randerson,

Perter Thornton, Natalie Mahowald,

Keith Lindsey, Yen-Huei Lee,

Cynthia Nevison, Scott Doney,

Gordon Bonan, Reto Stockli, Curtis Covey,

Steven Running, Inez Fung ….

Met Office
Hadley Centre

C-LAMP experiments

1. Examine the influence of climate variability, prescribed atmospheric CO_2 and land cover on terrestrial carbon fluxes during the 20^{th} Century

2. Examine the effect of a coupled biosphere-atmosphere for carbon fluxes and climate during the 20^{th} Century

http://www.climatemodeling.org/c-lamp/

ILAMB – International Land Atmosphere Model Benchmarking

Eleanor Blyth, Colin Prentice, Pru Foster,

Pierre Friedlingstein, Stephen Sitch,

Josh Fisher, Martin Best, Natalie de Noblet,

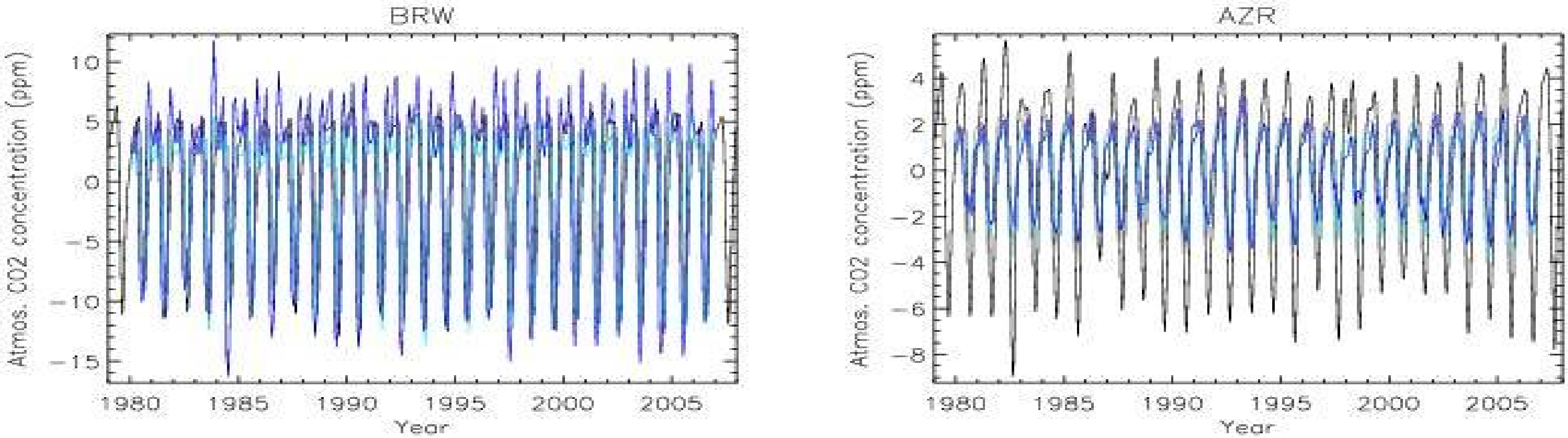
Dieter Gerten, Thmoas Hickler,

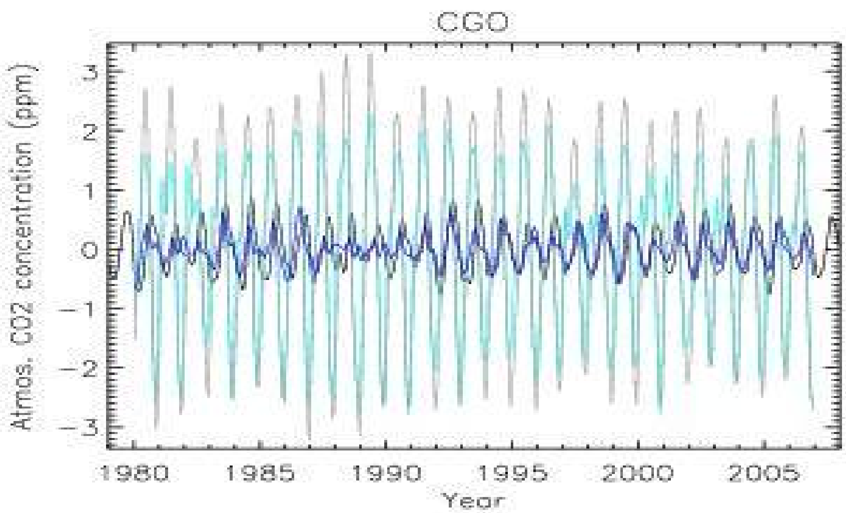
Marko Schulze, Angela Gallego-Sala,

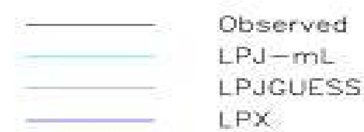
Steve Murray, Richard Betts,

Andy Wiltshire….

Atmospheric CO_2

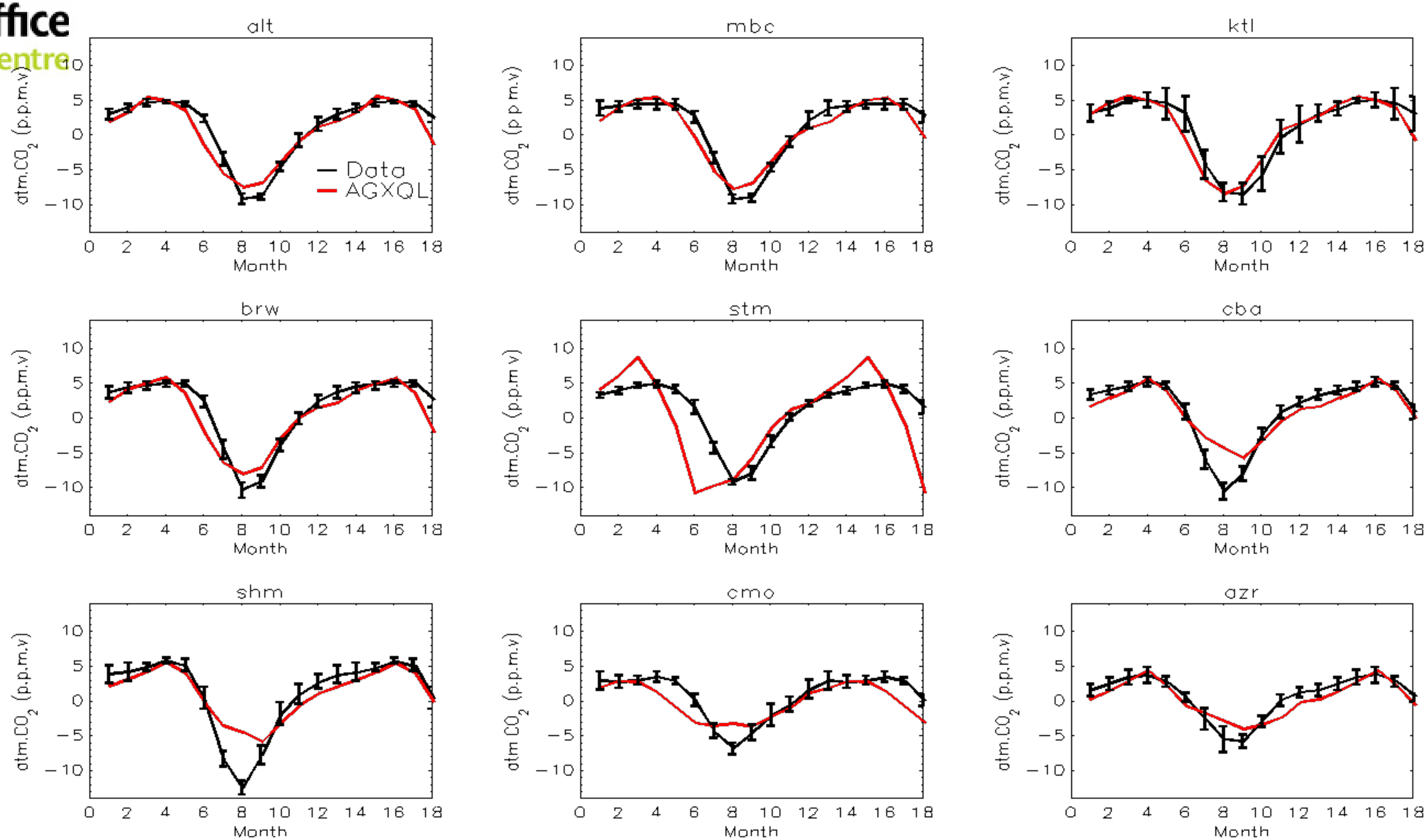






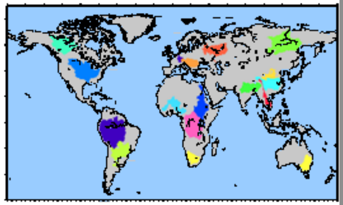
De-trended seasonal cycle of atmospheric CO2, observed and modelled by 3 models

Atmospheric CO_2

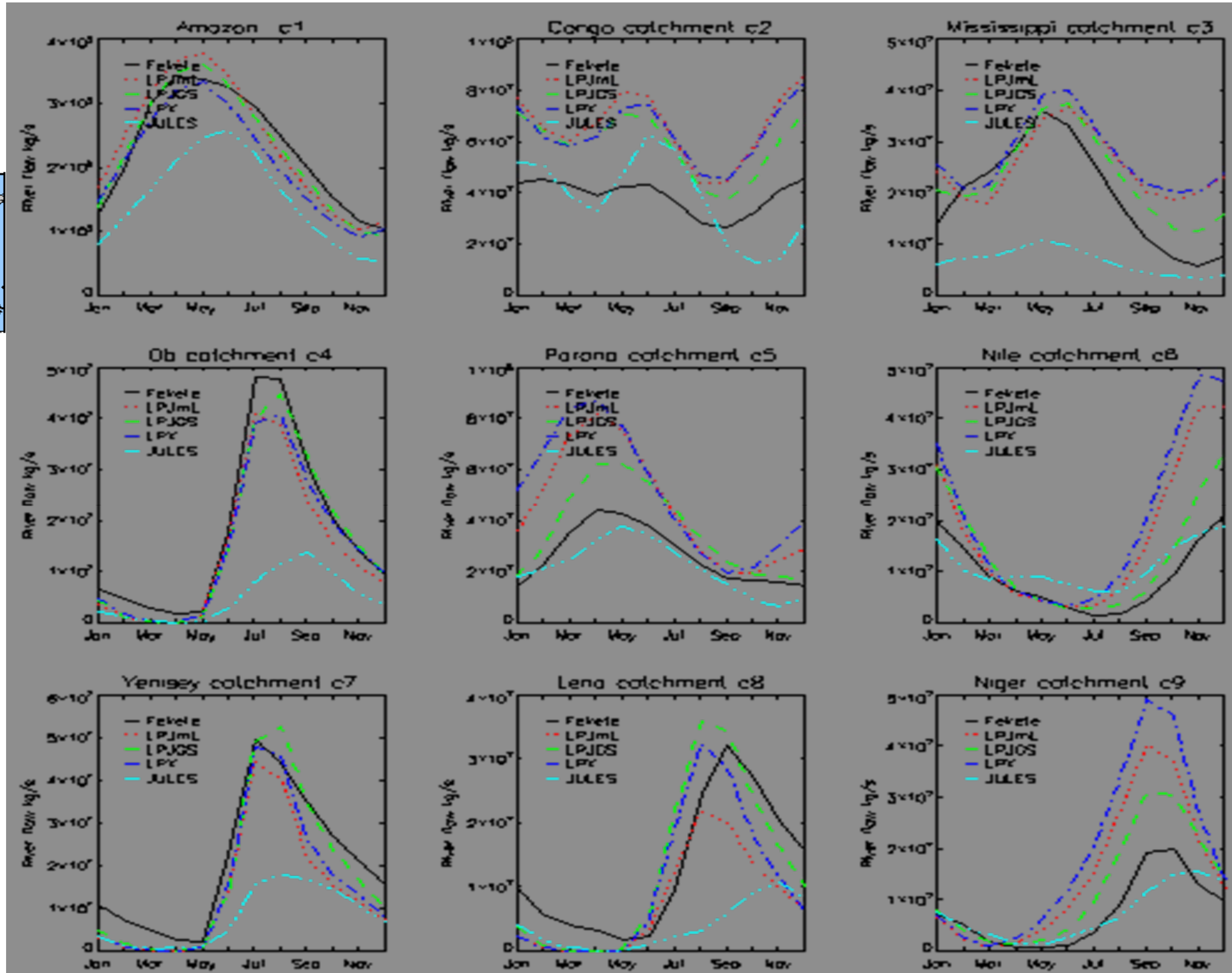


Example mean seasonal cycle of atmospheric CO2 from JULES (in coupled mode).

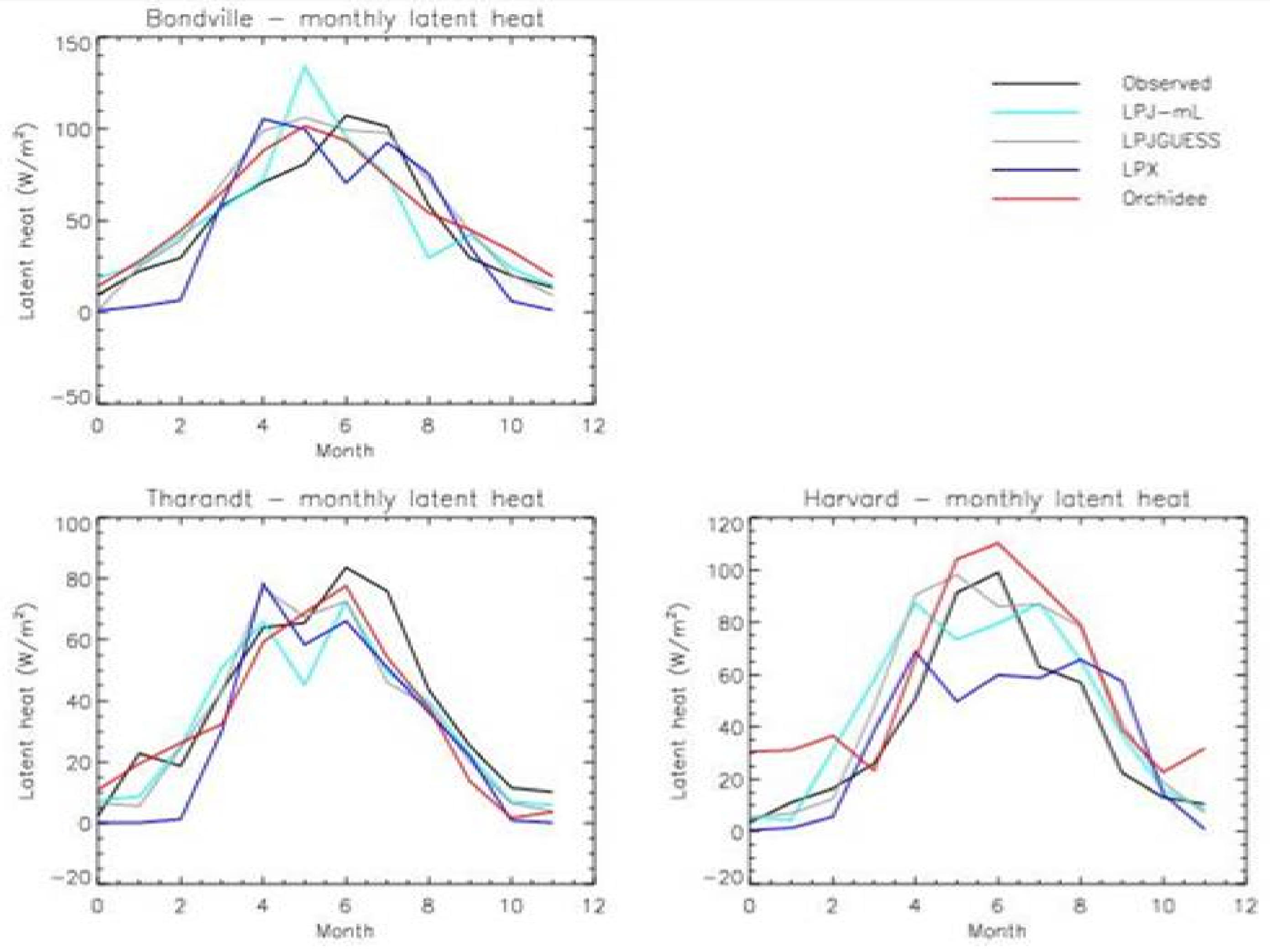
River flow



Mean seasonal riverflow for 4 models



Fluxnet data (Evap only so far)



PALS – Protocol for the Analysis of Land Surface models

Gab Abramowitz ….

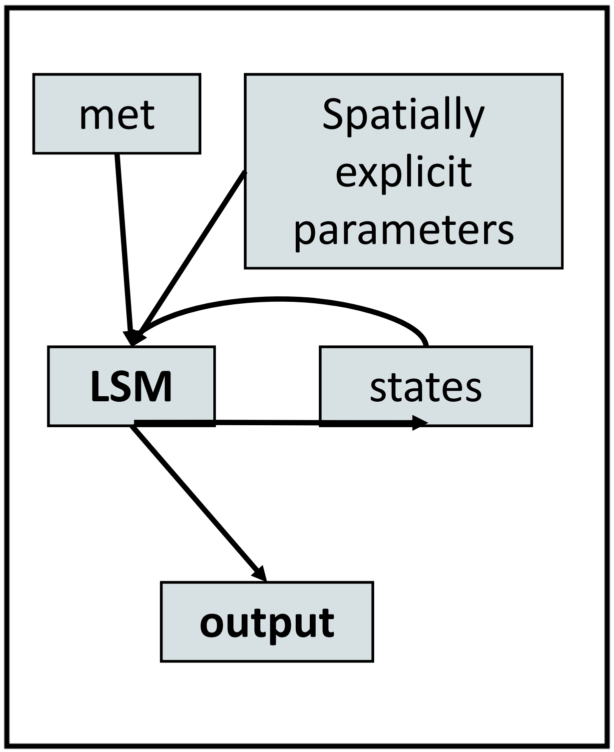
What the evaluation protocol is and aims to achieve

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 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
- Improved evaluation standards – any publication using a LSM could refer to its performance in particular standard tests on this site.

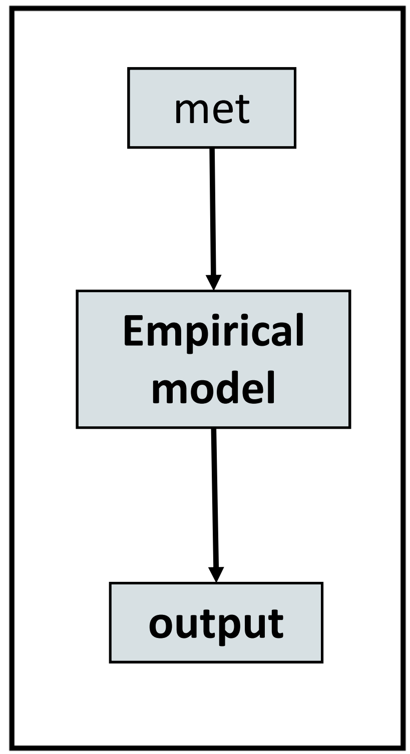
Statistical benchmark

Normal LSM:



Physically based

Empirical model:



Statistically based

COMPARE

NEE CO_2, latent heat, sensible heat

- Multiple linear regression (MLR)
- Neural Network
- SOLO
- other machine learning…

By manipulating the relationship between training and testing data sets we can test how well a LSM utilises the information available to it…

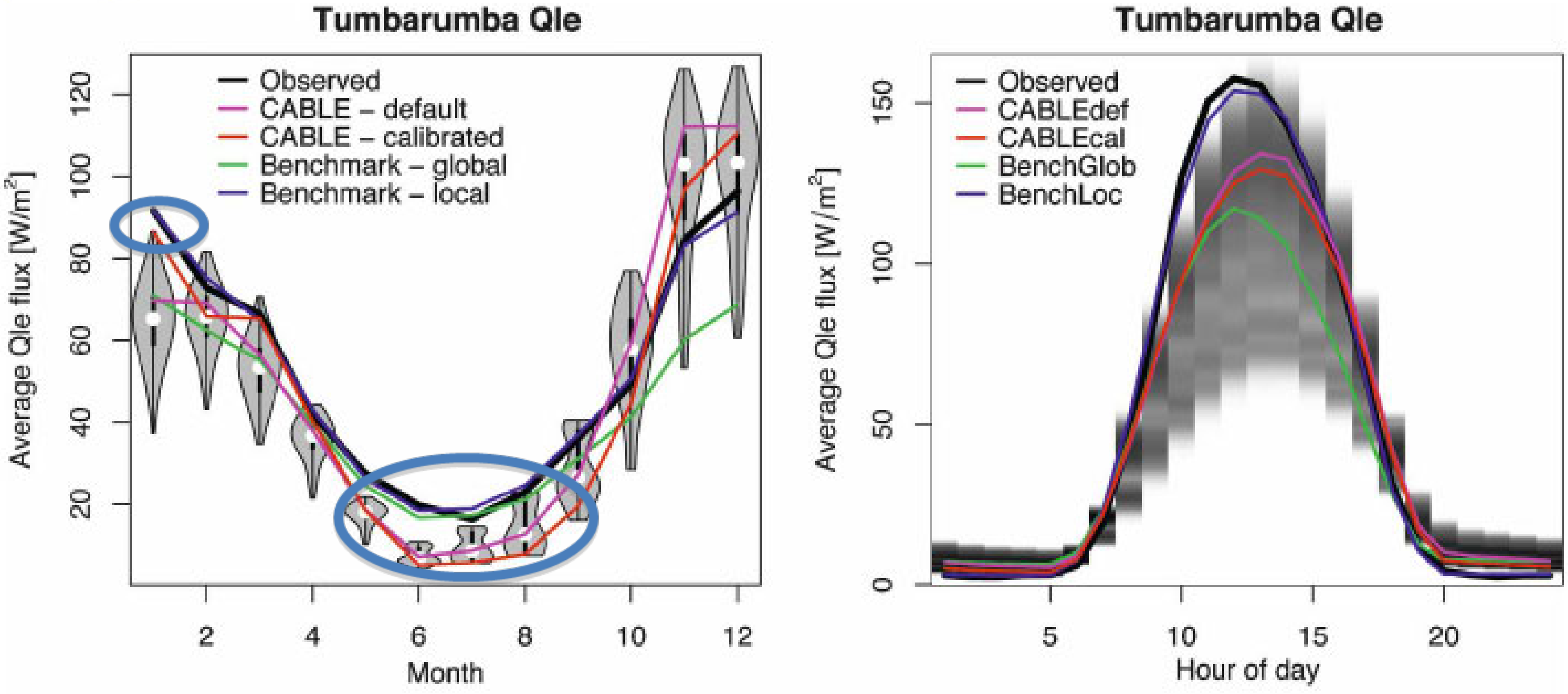
		NEE of CO_2 ($\mu mol/m^2/s$)					Latent heat flux (W/m^2)					Sensible heat flux (W/m^2)				
		mean (obs)	rmse	grad	int	rsq	mean (obs)	rmse	grad	int	rsq	mean (obs)	rmse	grad	int	rsq
Tumbarumba	CAB	-2.3 (-0.1)	5.48	0.55	-2.2	0.52	49 (52)	55.9	0.71	12.0	0.61	38 (54)	58.3	0.79	-5.0	0.82
	ORC	-0.5 (-0.1)	5.16	0.47	-0.4	0.50	46 (52)	58.1	0.73	8.1	0.59	14 (54)	118.	0.16	6.0	0.78
	CLM	-1.2 (-0.1)	6.13	0.24	-1.6	0.35	26 (52)	72.9	0.38	6.5	0.40	56 (54)	84.5	0.83	10.9	0.64
	MLR	-1.5 (-0.1)	4.40	0.55	-1.4	0.70	45 (52)	47.6	0.59	14.5	0.74	43 (54)	54.9	0.69	5.2	0.87
	ANN	-1.3 (-0.1)	4.65	0.52	-1.2	0.64	42 (52)	49.9	0.54	14.1	0.74	41 (54)	51.4	0.81	-3.0	0.86
Bondvill	CAB	-1.7 (-0.9)	6.13	0.61	-1.2	0.49	43 (47)	64.5	0.89	1.8	0.55	-0.1 (26)	71.8	0.62	-16.1	0.32
	ORC	-0.4 (-0.9)	9.07	0.68	0.2	0.29	30 (47)	48.0	0.84	-9.4	0.70	3.2 (26)	52.2	0.69	-14.5	0.56
	CLM	-1.2 (-0.9)	7.92	0.12	-1.1	0.08	33 (47)	50.5	0.63	3.0	0.62	30 (26)	57.9	1.01	4.3	0.59
	MLR	-0.7 (-0.9)	7.38	0.21	-0.5	0.20	38 (47)	45.8	0.53	13.6	0.73	26 (26)	47.7	1.08	-2.0	0.71
	ANN	0.0 (-0.9)	7.69	0.20	0.2	0.15	39 (47)	45.7	0.53	14.4	0.72	27 (26)	49.0	1.14	-2.2	0.73
Tharand	CAB	-1.2 (-1.7)	4.18	0.50	-0.3	0.77	46 (38)	47.4	0.98	8.0	0.61	-34 (26)	84.3	0.62	-50.4	0.56
	ORC	-0.5 (-1.7)	4.08	0.65	0.6	0.72	32 (38)	41.8	0.68	5.7	0.55	-28 (26)	76.8	0.64	-45.0	0.62
	CLM	-1.2 (-1.7)	5.28	0.31	-0.7	0.70	28 (38)	41.1	0.68	2.1	0.58	14 (26)	56.3	0.92	-9.9	0.69
	MLR	0.7 (-1.7)	5.81	0.32	1.3	0.62	28 (38)	37.3	0.58	5.4	0.65	16 (26)	43.3	0.69	-2.1	0.79
	ANN	0.1 (-1.7)	5.28	0.36	0.7	0.70	31 (38)	36.6	0.57	9.5	0.65	10 (26)	44.5	0.75	-9.6	0.78
Washita	CAB	-0.7 (0.6)	4.15	0.34	-0.9	0.15	46 (39)	64.7	0.95	9.6	0.46	31 (38)	77.9	0.67	5.4	0.41
	ORC	-0.1 (0.6)	4.13	0.52	-0.4	0.23	33 (39)	37.7	0.86	-0.2	0.69	30 (38)	56.0	0.78	0.5	0.64
	CLM	-0.9 (0.6)	4.00	0.29	-1.1	0.16	42 (39)	48.6	0.87	8.5	0.57	53 (38)	56.8	0.98	15.3	0.72
	MLR	-2.5 (0.6)	6.29	0.60	-2.8	0.16	61 (39)	43.7	0.96	24	0.72	31 (38)	40.4	1.00	-7.0	0.83
	ANN	-2.5 (0.6)	8.51	1.31	-3.3	0.29	65 (39)	58.4	1.33	14.0	0.75	29 (38)	46.9	0.73	0.9	0.73
Weiden	CAB	-1.0 (0.2)	2.90	0.66	-1.1	0.67	26 (24)	40.0	1.21	-2.9	0.7	-23 (-5)	62.7	0.51	-20.1	0.67
	ORC	-0.5 (0.2)	2.70	0.89	-0.7	0.73	11 (24)	38.0	0.89	-10	0.61	-19 (-5)	62.4	0.51	-16.2	0.65
	CLM	-1.1 (0.2)	3.37	0.42	-1.2	0.62	18 (24)	35.5	0.79	-0.2	0.59	15 (-5)	57.4	0.73	18.6	0.70
	MLR	-0.5 (0.2)	3.01	0.53	-0.6	0.62	29 (24)	28.3	0.70	12.1	0.69	20 (-5)	62.3	0.54	22.8	0.70
	ANN	-0.4 (0.2)	2.45	0.81	-0.6	0.75	30 (24)	27.0	0.73	12.5	0.72	19 (-5)	60.0	0.59	21.7	0.71
Metoliu	CAB	-0.8 (-0.7)	3.02	0.50	-0.4	0.50	29 (40)	56.7	0.51	8.8	0.26	22 (34)	61.8	0.97	-11.2	0.71
	ORC	-1.0 (-0.7)	3.27	0.93	-0.3	0.60	34 (40)	53.6	0.71	5.0	0.38	9 (34)	63.9	0.68	-14.1	0.64
	CLM	-0.7 (-0.7)	3.54	0.26	-0.5	0.32	21 (40)	51.5	0.45	2.9	0.32	40 (34)	57.5	0.83	11.7	0.69
	MLR	-1.5 (-0.7)	3.36	0.71	-1.0	0.50	37 (40)	48.7	0.72	7.9	0.43	42 (34)	53.7	0.82	13.6	0.72
	ANN	-0.8 (-0.7)	2.88	0.81	-0.2	0.61	36 (40)	47.7	0.69	8.6	0.43	46 (34)	55.3	0.88	15.7	0.73

Which model wins?

- CABLE : 12
- CLM : 15
- ORCHIDEE : 17
- simple neural network : 21
- linear regression : 25
- Neural net and linear regression are NOT trained at the sites at which they're tested and have no mechanism to distinguish between vegetation types
- Enough information in SWdown, Tair and humidity, globally, to outperform LSMs with many spatially explicit parameters.

Abramowitz et al, *J Climate*, 2008

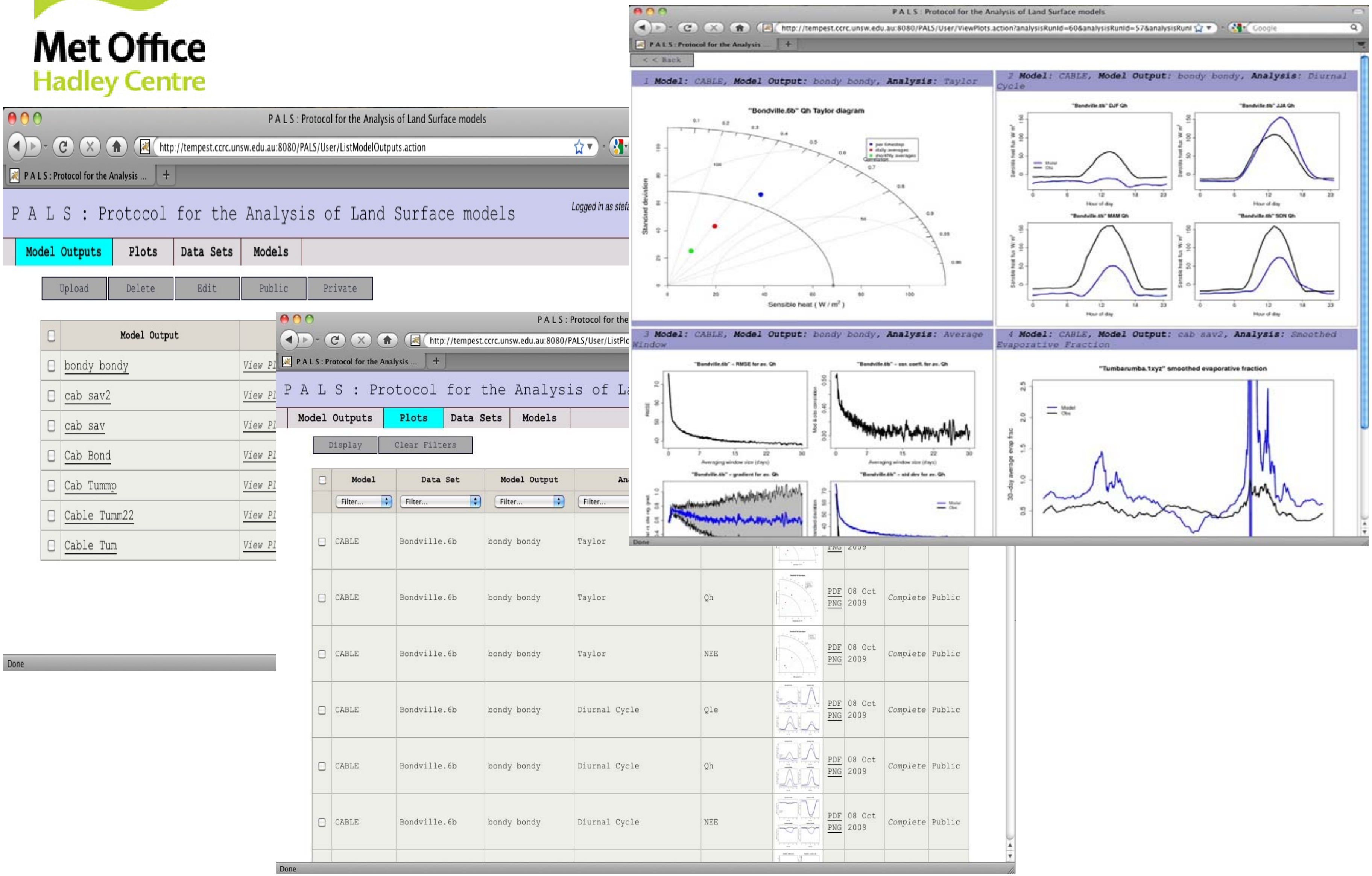
What calibration can and cannot achieve



Abramowitz et al, *J Climate*, 2008

- PDFs represent perturbed parameter ensemble (5 parameters, ~10k runs)
- Parameters and their ranges chosen in consultation with model builders

Prototype of web system



Summary of benchmarking activities

- QUEST/GLASS benchmarking meeting June 2009

Generic Categories

- Numerical Weather Prediction (NWP)
- Climate Systems
- Impacts on humans and ecosystems
- Process Studies

Minimum Benchmark

- NWP – Persistence
- Climate Systems – ??? (*Climatology for current climate? What about future climate?)*

- Next Steps
  - ➢Model data exchange for C-LAMP and ILAMB
  - ➢Development of PALS web-based system
  - ➢Follow on meeting planned for 2010
- And beyond …..
  - ➢Merger of systems into an internationally agreed benchmarking tool for land surface models with agreed metrics

CEOP

Coordinated Energy and Water Cycle Observations Project

To understand and predict continental to local-scale hydroclimates for hydrologic applications.

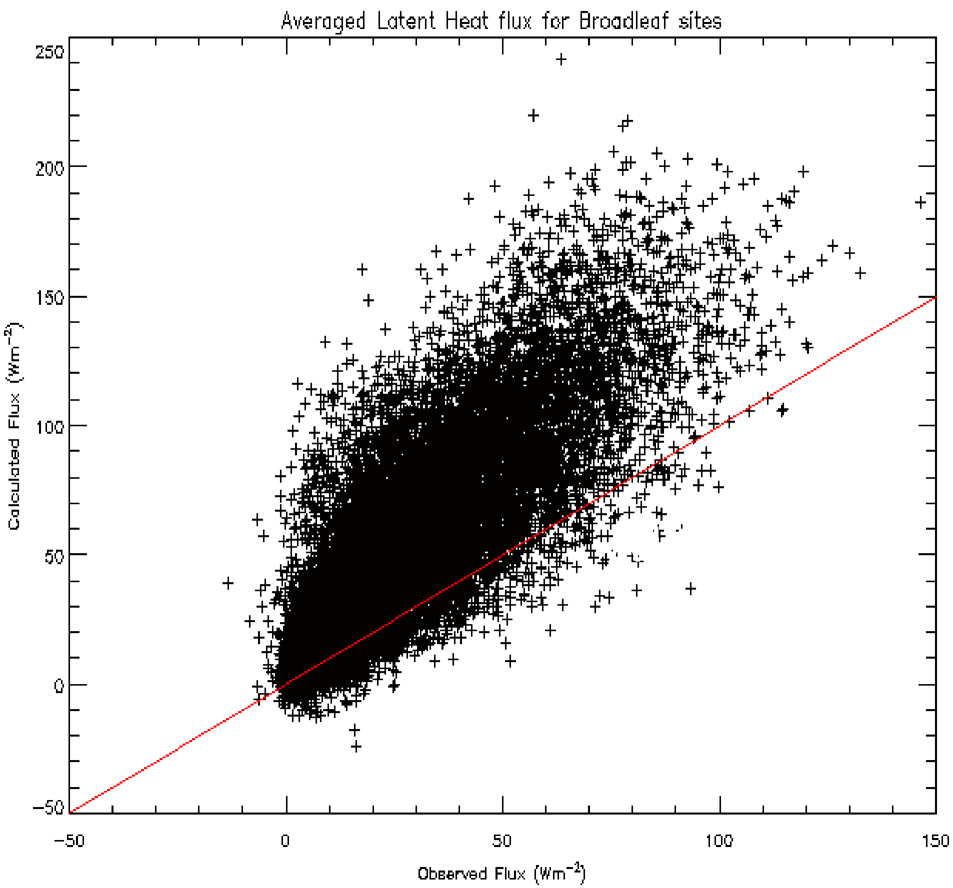
- **Objective 1**: Produce data sets of the Earth's energy budget and water cycle for climate system analysis and **model development and evaluation**.
- **Objective 2**: Enhance understanding of energy/water cycle processes & climate feedbacks.
- **Objective 3**: **Improve the predictive capability** for key water and energy cycle variables and feedbacks through **improved parameterizations**.
- **Objective 4**: Undertake joint activities with operational hydrometeorological services assessing the consequences of climate predictions and global change for water resources.

- **CEOP Phase 1** : 2001-2004 with main extended observation periods EOP3 (Oct 2002-Sep 2003) and EOP4 (Oct 2003-Dec 2004).
- **CEOP Phase 2** : Jan 2007- Dec 2010. Ten year "synthesis" dataset (2001-2010) planned.

Met Office Strategy for CEOP

- Evaluate Operational Global and Regional (NAE) NWP models against CEOP observations for short-range (12-36 hour) forecasts (2007-2010 CEOP phase 2).
- Testbed for evaluating parametrizations over land – land surface, BL, cloud, aersols & radiative forcing.
- Use JULES run offline and forced by observational data at CEOP sites to explore errors in land surface parametrizations.
- Evaluate model errors in longer timescale predictions – THORPEX, seasonal, Decadal to explore feedbacks between land surface errors and atmospheric circulation (e.g. Monsoons)

JULES Example
Excess evaporation



Courtesy of Sean Milton

FLUXNET sites

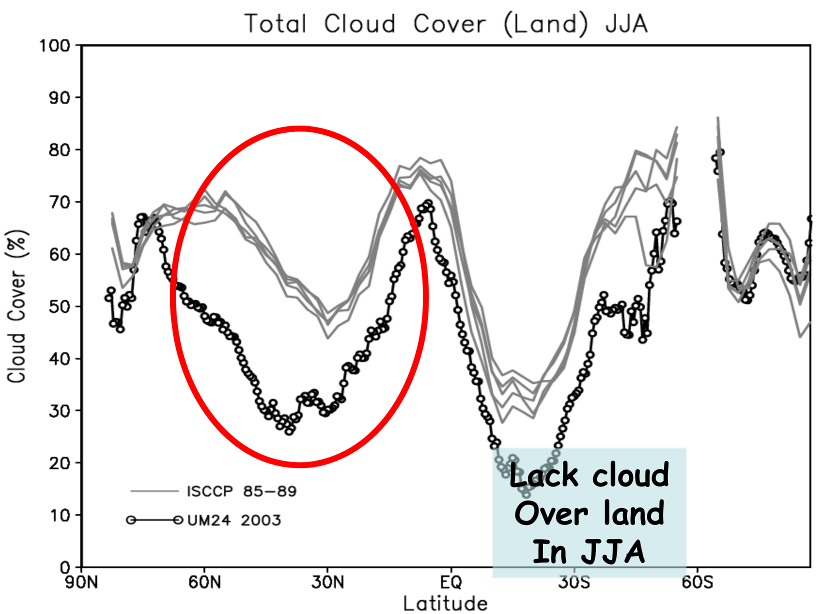
Location	Error in latent heat flux	Error in sensible heat flux
Aberfeldy	11.27	21.99
Bondville	14.32	-3.21
Bordeaux	19.52	6.26
Brasschaat	23.16	7.11
Castelporziano	12.92	-12.56
Flakaliden	5.87	3.09
Gunnarsholt	21.35	26.04
Harvard	27.47	-2.95
Hesse	22.48	3.54
Hyytiala	11.22	-10.21
Little	39.63	6.85
Loobos	-8.14	18.26
Metolius	-0.78	69.82
Sky Oak Old	-2.07	-7.08
Soroe	22.96	-10.38
Tharand	15.03	-11.67
Upad	2.46	1.12
Vielsalm	15.14	-2.59
Walker Branch	39.09	12.51
Weidenbrunnen	18.71	16.63

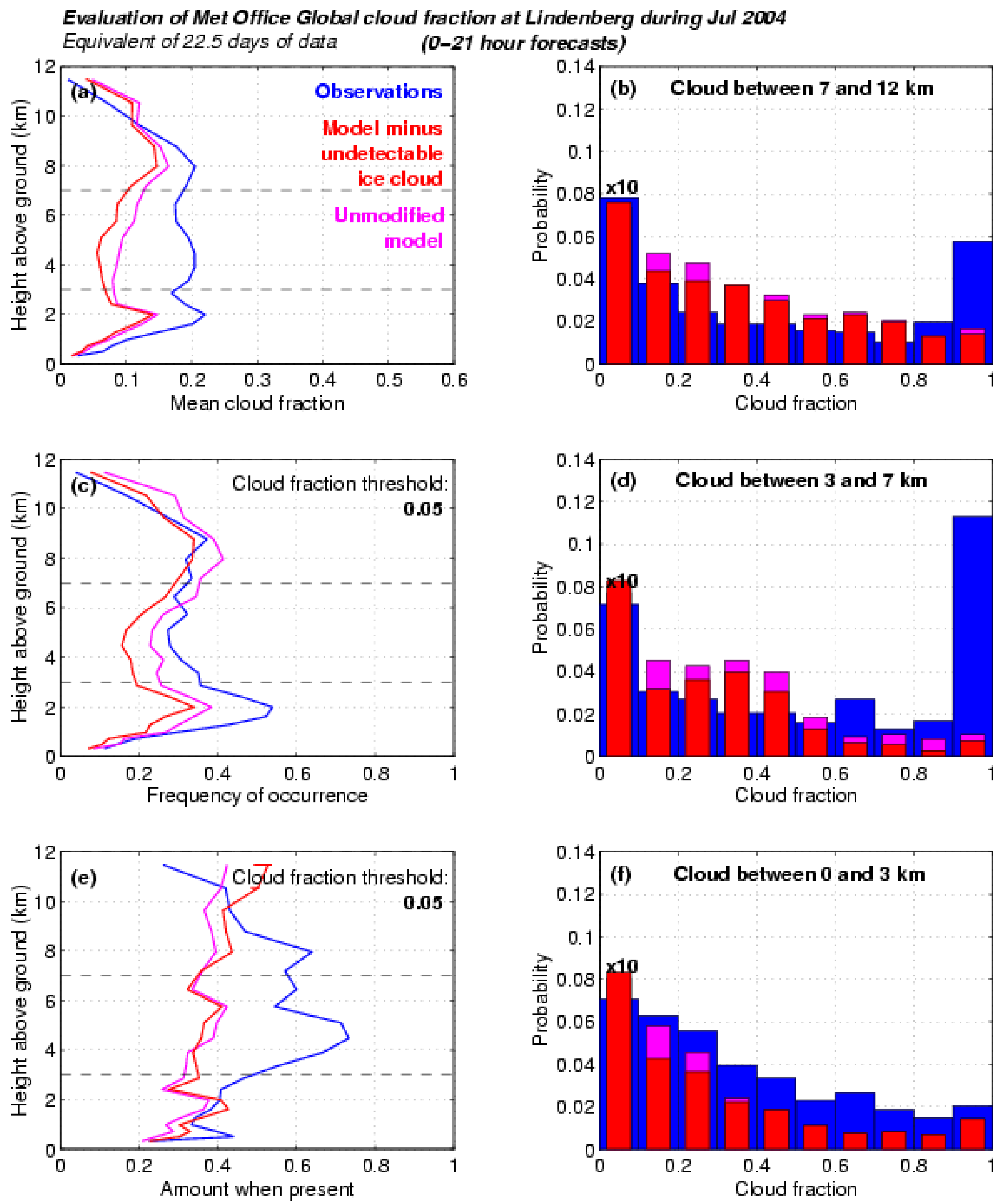
Cloudnet Comparisons.

Lindenberg

July 2004

Underestimation of cloud fraction
Error in amount when present rather than frequency of occurrence.



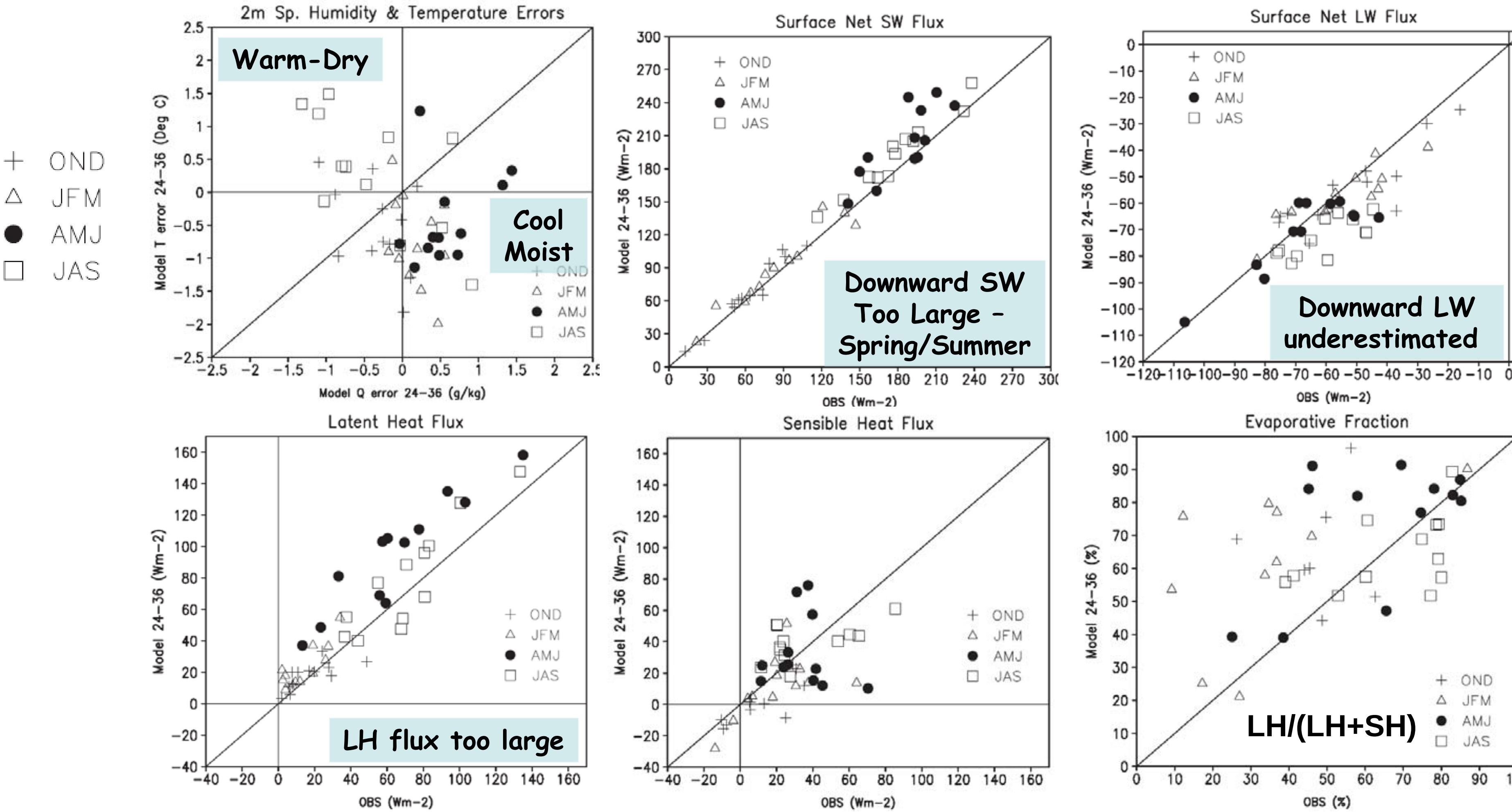


Courtesy of Sean Milton

Surface Energy Balance

24-36 hr forecasts Mid-Latitude sites

ARM-SGP, Tongyu, Cabauw, Bondville



Courtesy of Sean Milton

Questions and answers