Performance metrics and observations in global climate model evaluation

Veronika Eyring (DLR, Germany)

CMUG Integration-3 meeting

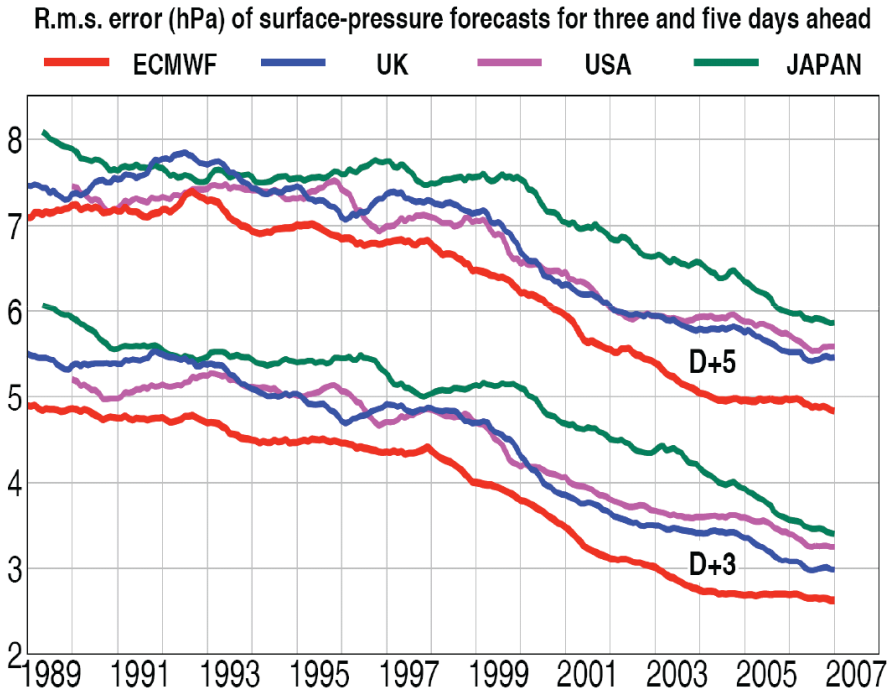
MPI-M, Hamburg, Germany

5 June 2013

DLR

Towards a More Quantitative Evaluation: Performance Metrics

Performance Metric: statistical measure of agreement between a **simulated and observed field** (or co-variability btw. fields) that assigns a quantitative measure of performance to individual models.



Back in 2011:

- The climate community does not yet calculate performance metrics routinely.
- Some pioneering work had been published (e.g. Schmittner et al., 2005; Connolley and Bracegirdle, 2007; Reichler and Kim, 2008; Gleckler et al., 2008; Pincus et al., 2008; Waugh and Eyring, 2008)

In 2013:

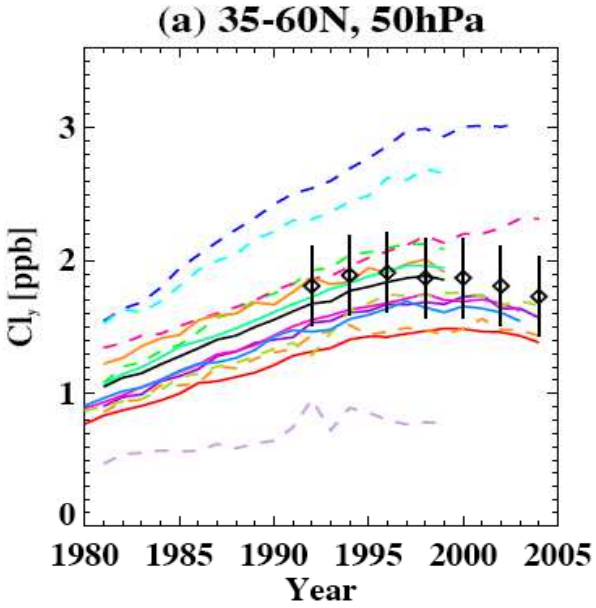
- Much more awareness that performance metrics are useful.
- General acceptance within the climate community has grown.
- Growing literature on performance metrics.

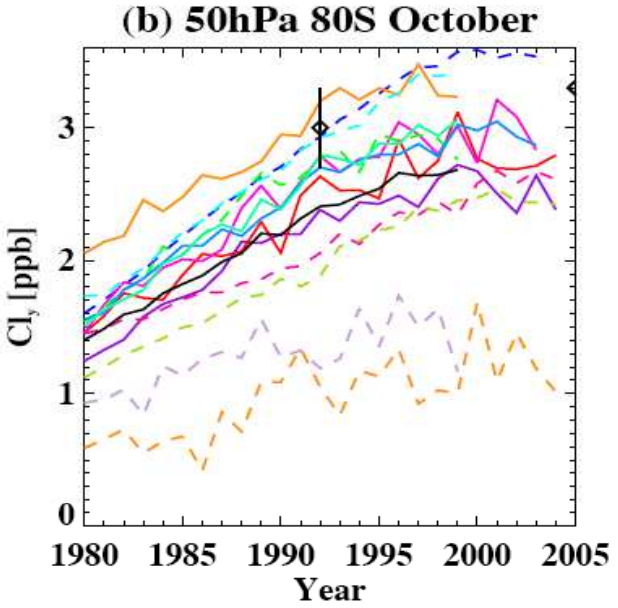
DLR

Performance Metrics: Examples for quantitative evaluation from Chemistry-Climate Model Validation Activity (CCMVal)

Inorganic Chlorine Cl_y

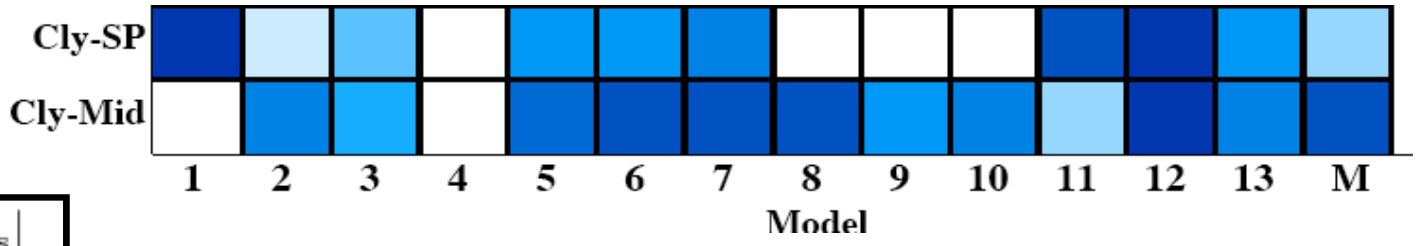
Diagnostics





Performance Metrics

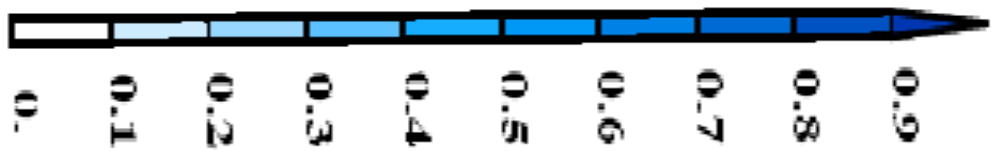
Waugh & Eyring, ACP, 2008; SPARC CCMVal, 2010



Grades

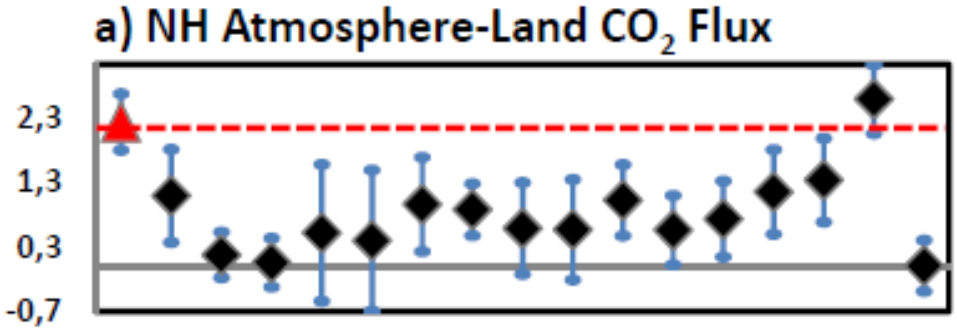
$$g = 1 - \frac{1}{n_g} \frac{|\mu_{\text{model}} - \mu_{\text{obs}}|}{\sigma_{\text{obs}}}$$

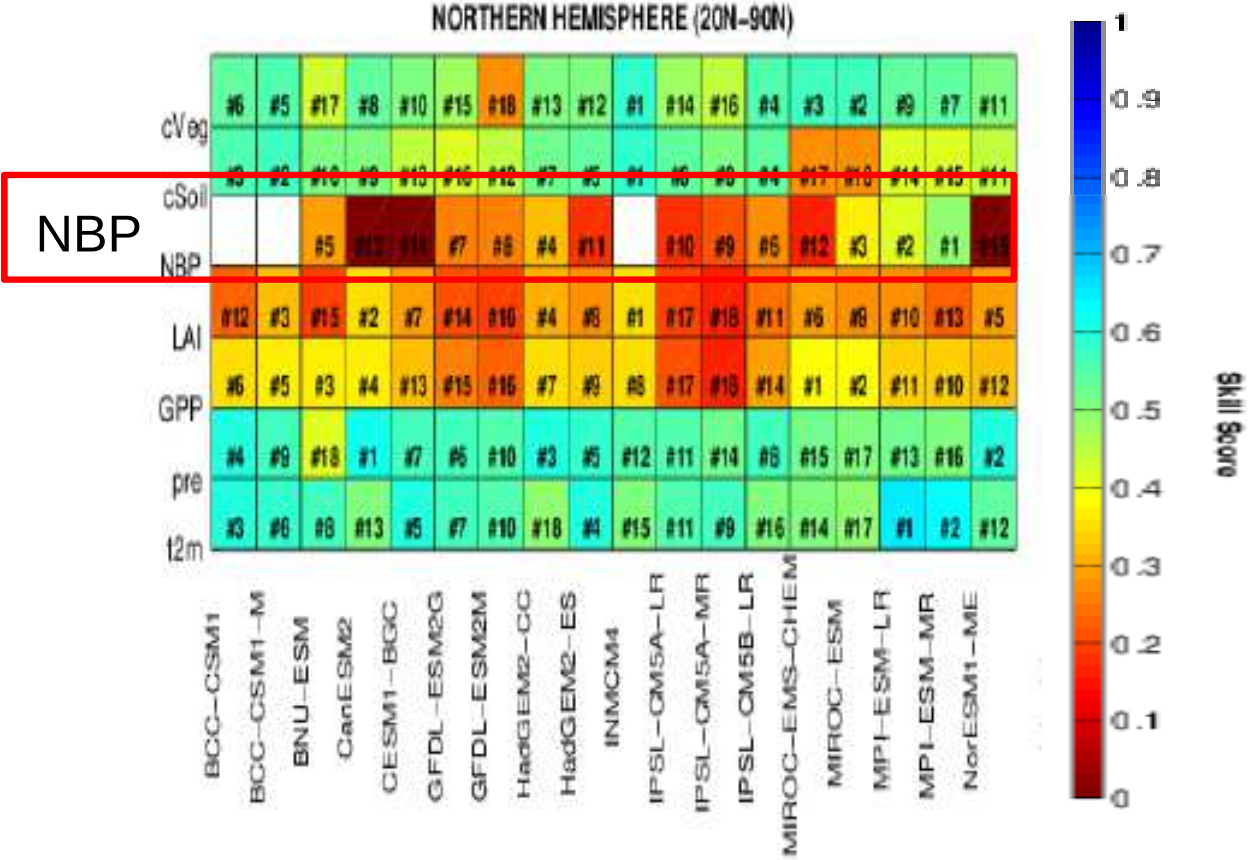
bad *good*



DLR

Model evaluation: Carbon cycle performance metrics





PDF based skill scores (mean state and IAV)

$$s = \sum_{1}^{N} min\left(Z_m, Z_0\right)$$

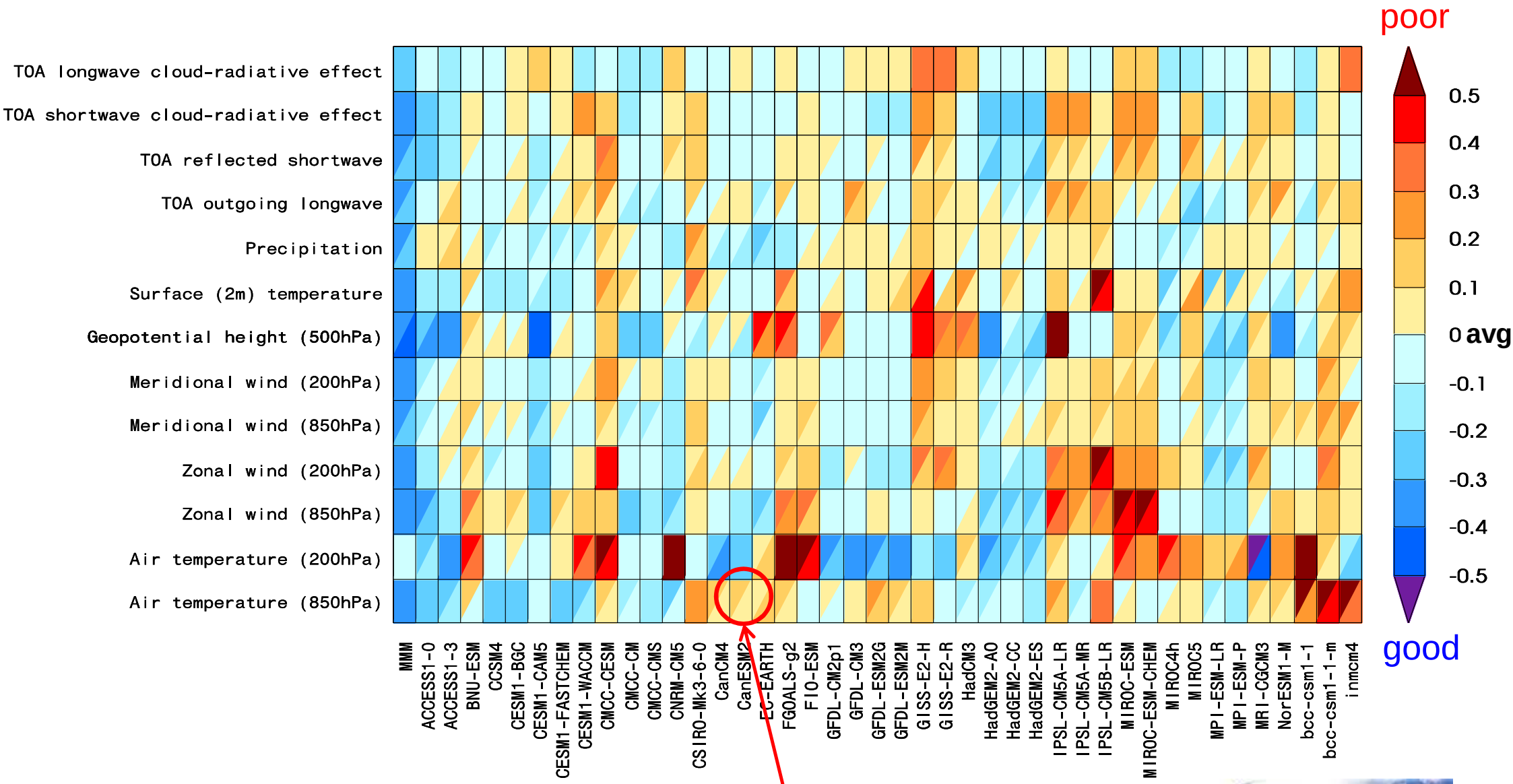
- Surface Temperature and precipitation show general good agreement globally and in the SH and NH, poorer skills in the tropical region.
- In the NH over land almost all models systematically underestimate the sink, reflecting absence of sinks due to nitrogen deposition or poor representation of forest regrowth.
- Good scores for global NBP, but low scores in the NH.

Anav et al., 2013

UNIVERSITY OF EXETER

Model performance physical climate:
revisiting Gleckler et al. (2008) portrait plot with CMIP5

Relative global Root Mean Square Error (RMSE) in climatological annual mean

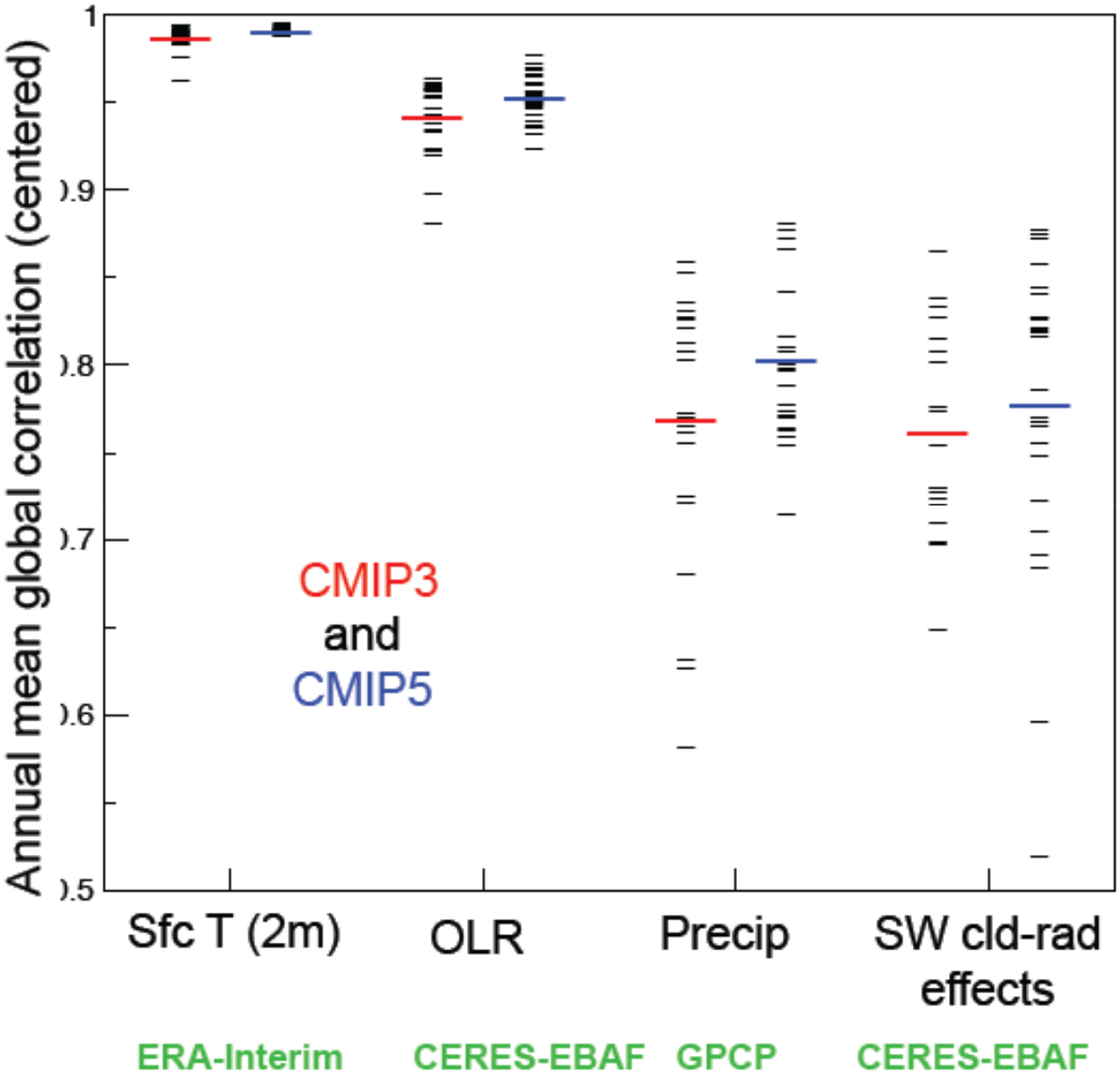


Courtesy of Peter Gleckler

Two different sets of observations

PCMDI

Tracking model performance: Incremental improvement from CMIP3 to CMIP5



Ensemble average results in CMIP5 are incrementally better than CMIP3

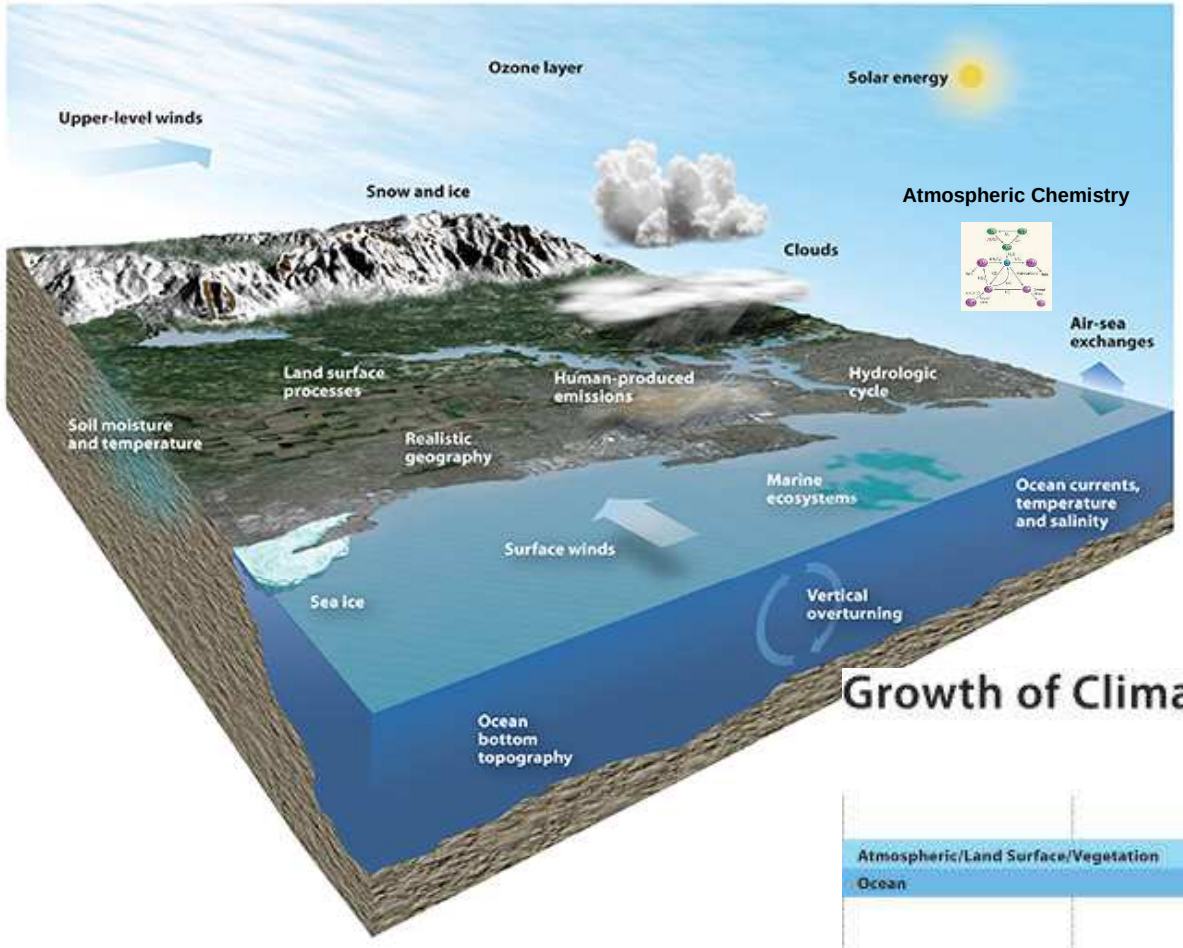
At least in part due to a reduction in spread, with fewer poor outliers

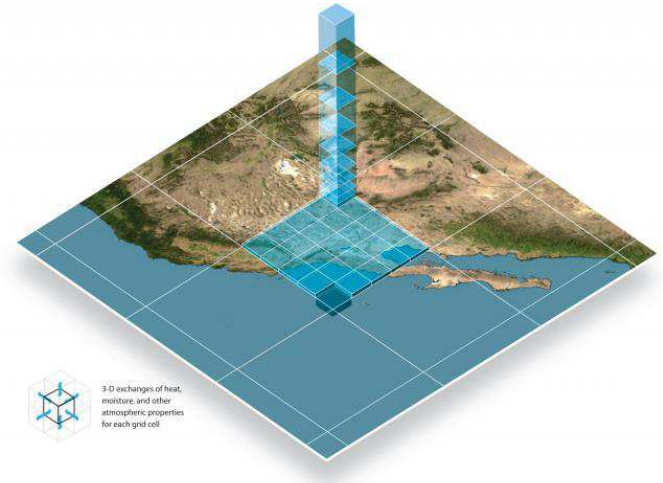
But some evidence of the better models getting better

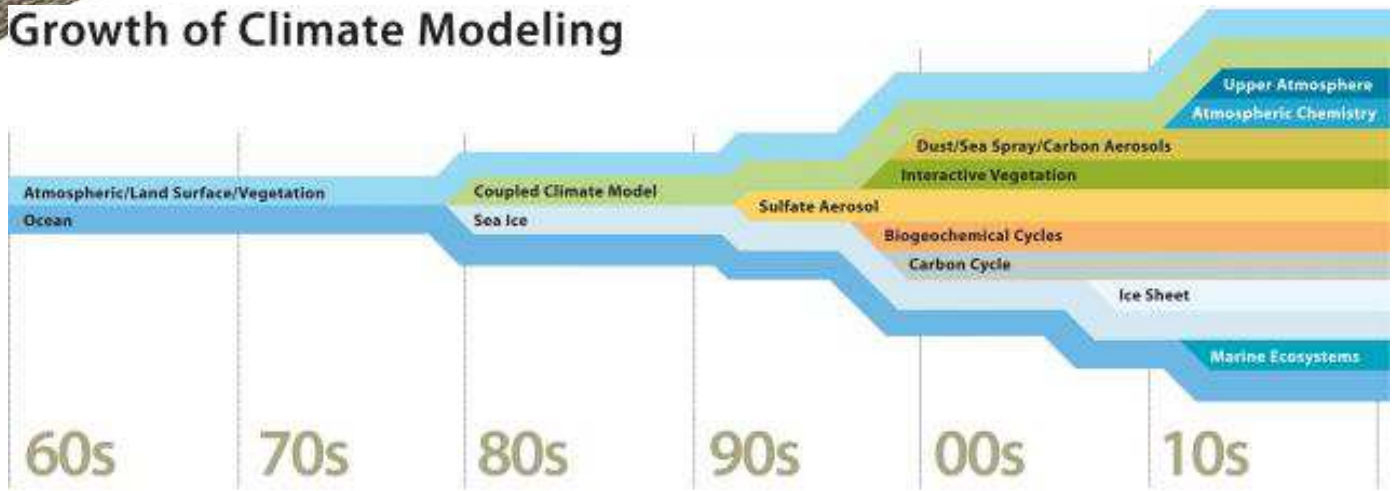
Courtesy of Peter Gleckler

PCMDI

Models are increasing in complexity and resolution
- From AOGCMs to Earth System Models with biogeochemical cycles -







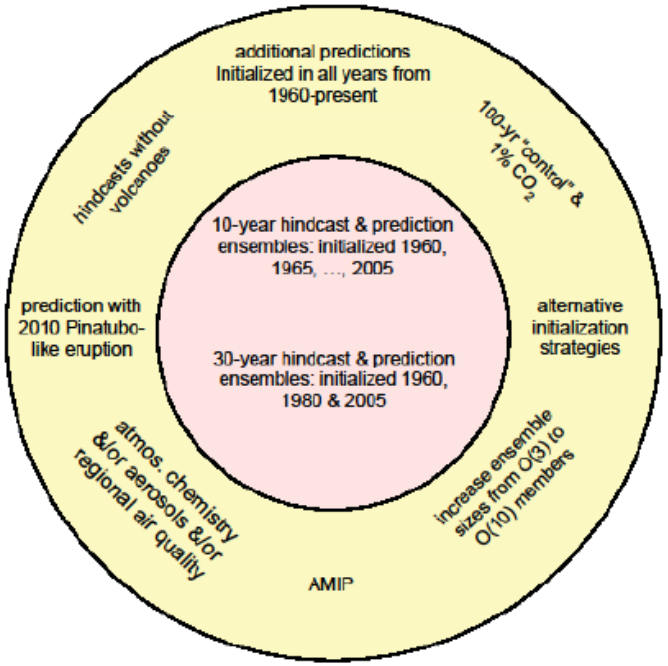
https://www2.ucar.edu/news/understanding-climate-change-multimedia-gallery

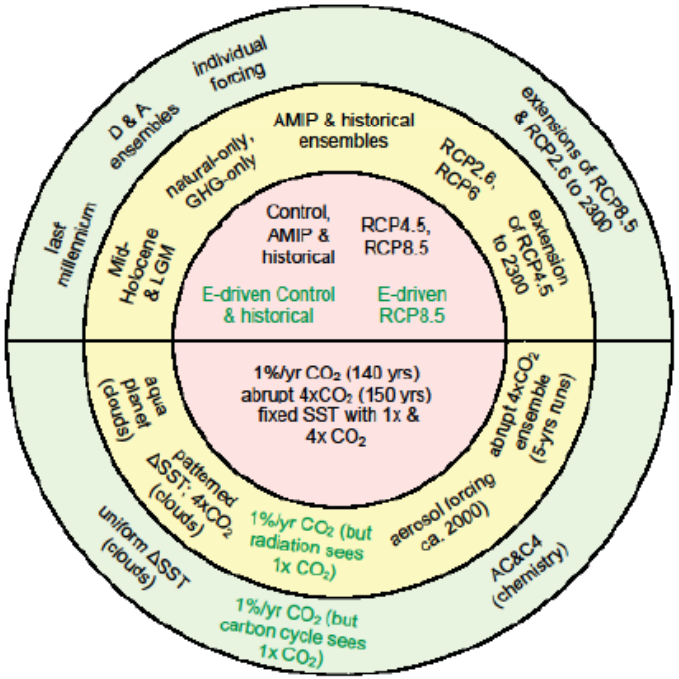
DLR

CMIP5 model simulations: Two classes of models to address two time frames and two sets of science questions (Taylor et al., BAMS, 2012)

Near-Term (2005-2030, decadal predictions) high resolution, no carbon cycle, some chemistry, single scenario; **Science question:** e.g. regional extremes

Longer term (to 2100 and beyond) lower resolution, carbon cycle, benchmark stabilization concentration scenarios; **Science question**: e.g. feedbacks.



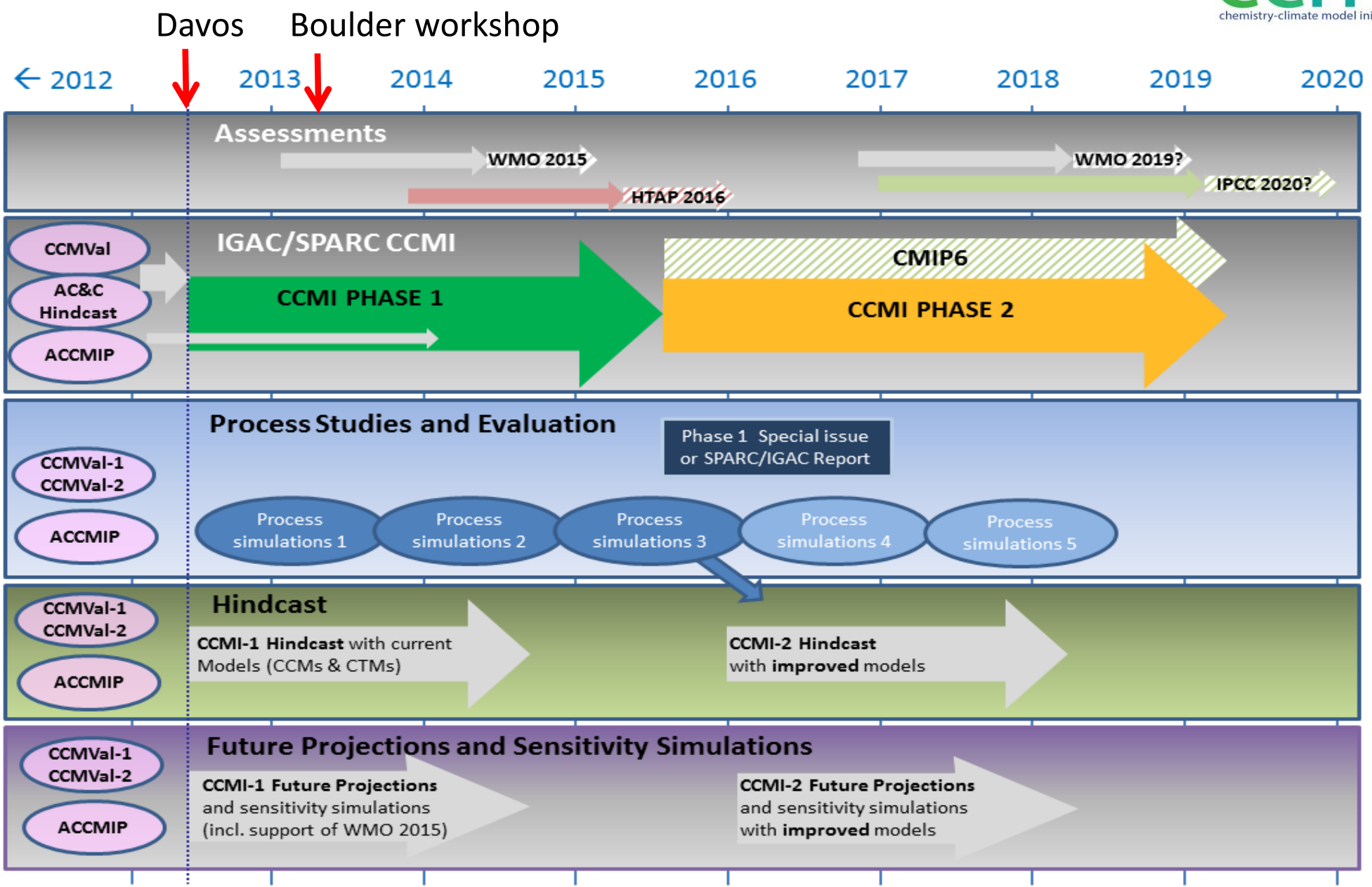


- Different model simulations require different evaluation strategies

IGBP IGAC / WCRP SPARC Chemistry-Climate Model Initiative

Co-Chairs: Veronika Eyring (DLR) & Jean-Francois Lamarque (NCAR)

ccmi
chemistry-climate model initiative



Eyring et al., SPARC Newsletter, 2013

=> Earth System Model (ESM) evaluation is complex!

Models:

- Models are getting more complex, with more processes included
- More models (20+ groups; ~ 40 models) in CMIP5, ~ 22 model groups in CCMI-1
- Different climate model simulation setups (e.g. decadal, long-term, nudged) require different evaluation strategy

Models and observations (Recommendations WCRP Modeling Council Meeting):

- Provide observations with quantified uncertainties.
- Reduce the gaps between modeling and observations communities.
- Create better infrastructure to facilitate access to observations including estimates of uncertainties in datasets.
- Promote the systematic collection of observations in regions, such as polar areas, the upper troposphere / lower stratosphere (UTLS), the deep ocean, where the lack of data is impeding progress..

Requires a community

effort to make it happen…

Related International Efforts

WGNE/WGCM Climate Model Metrics Panel

An effort to advance the routine evaluation of climate models

Beth Ebert (BMRC) – JWGV/WWRP, WMO forecast metrics
Veronika Eyring (DLR Germany) – WGCM/SPARC, CCMI, CMIP, ESMs
Pierre Friedlingstein (U. Exeter) – IGBP, carbon cycle, ESMs
Peter Gleckler (PCMDI), chair – WGNE, atmosphere, ocean
Simon Marsland (CSIRO) – WGOMD, ocean
Robert Pincus (NOAA) – GEWEX/GCSS, clouds/radiation
Karl Taylor (PCMDI) – WGCM, CMIP5, atmosphere
Helene Hewitt (U.K. Met Office) – polar ocean and sea-ice

This panel aims at making results from routine performance metrics more accessible, and in doing so clarify their limitations. It also seeks to gradually facilitate the incorporation a diverse set of more in-depth performance tests.

http://www-metrics-panel.llnl.gov/wiki

-> See poster by Peter Gleckler

DLR

The WGNE/WGCM metrics panel package and code repository

The metrics panel package of routine metrics

- Simple package to offer modeling groups, providing them with the ability to easily benchmark their models against others.
- This package includes carefully selected observational data, a few very simple codes, and a database of results for all CMIP3 and CMIP5 models.
- Distribution to modeling groups expected within 3-4 months, with a survey requesting feedback for improvement.
- Other efforts are underway to develop analysis codes that will be available to the research community (e.g., EMBRACE).

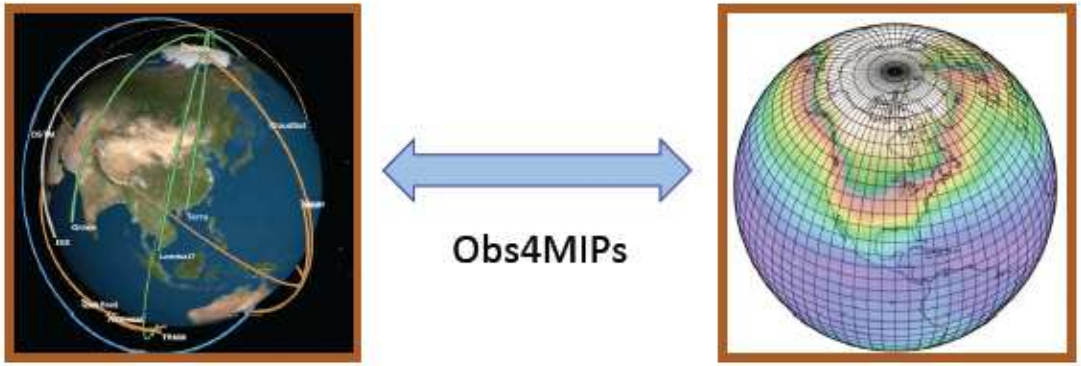
A Community-wide diagnostic & performance metrics code repository?

- At this site scientists involved in climate model evaluation are encouraged to contribute codes that can be used to compute metrics and associated diagnostics.
- One goal of this collection is to facilitate the sharing of analysis packages (ranging from routine calculations to advanced or novel efforts).
- It is hoped that this repository might facilitate an increasing openness to climate model evaluation.

DLR

Recommendation: Contribute data to obs4MIPs

obs4MIPs: An Overview and Update



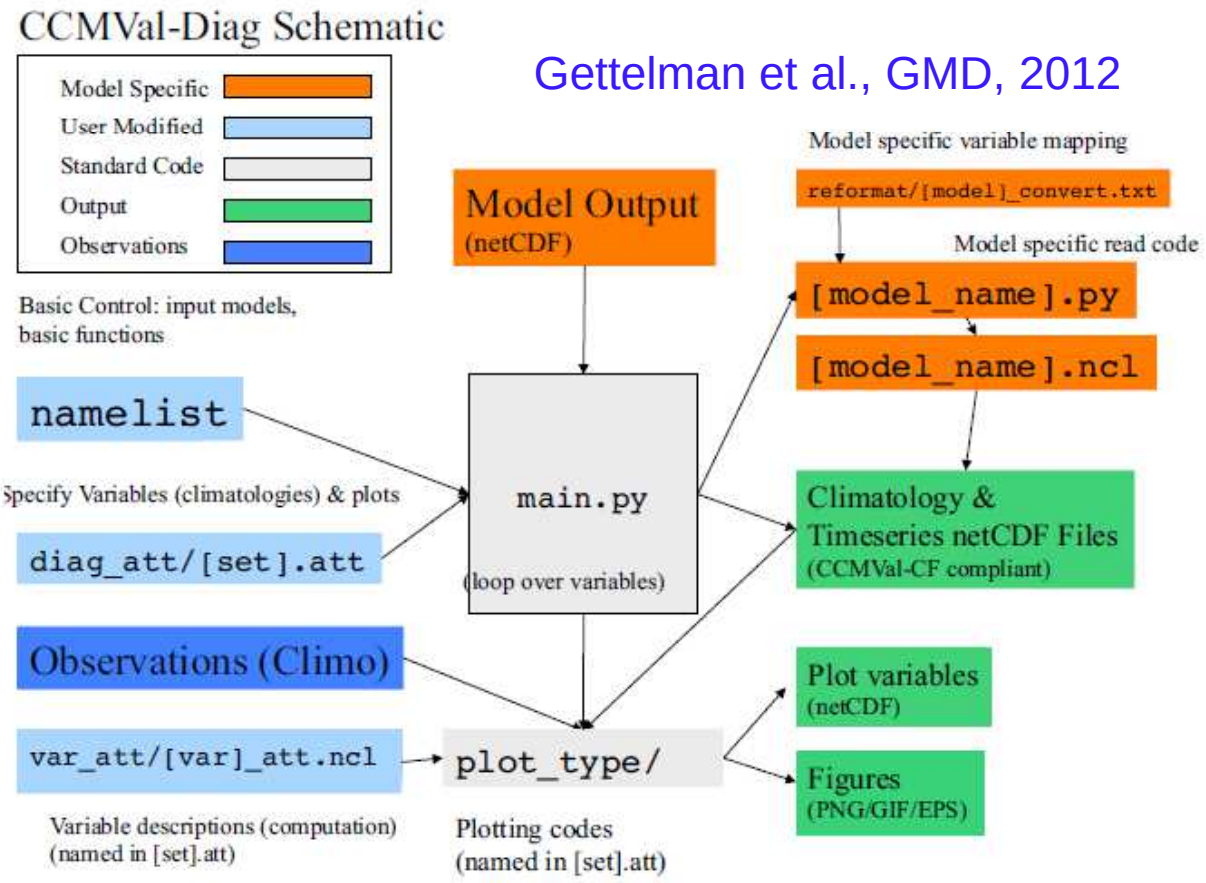
1. Use the CMIP5 simulation protocol (Taylor et al. 2009) as guideline for deciding which observations to select.
   Initial Target was monthly averaged (OMON, AMON) products on 1 x 1 degree grid
2. Convert (Satellite) Observations to CMIP model output format
   CMOR output, NetCDF files, CF Convention Metadata, CMIP standard pressure levels, etc.
   Not a new product. Independent QC check before release.
3. Includes a 6-8 page Technical Note describing strengths/weaknesses, uncertainties, caveats regardingcomparisons with models.
   (at graduate student level)
4. Host side by side on the ESG with CMIP5

Courtesy of Duane Waliser, WGCM Meeting 2012

Development of an Earth System Model Evaluation Tool

Within EMBRACE: DLR, SMHI & EMBRACE partners in collaboration with NCAR, PCMDI, GFDL

- **Open Source:** **Python** Script that calls **NCL (NCAR Command Language)**
- **Input:** CF compliant NetCDF model output (CCMVal, CMIP, etc.)
- **Observations:** Can be easily added
- **Extensible:** easy to (a) read models (b) process output [diagnostic] with observations and (c) use a standard plot type (e.g. lat-lon map)



Gettelman et al., GMD, 2012

Current developments include

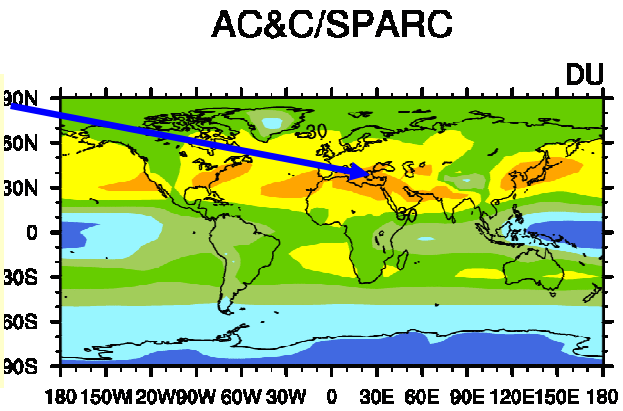
- Essential Climate Variables, e.g.
  - Sea-Ice
  - Temperatures
  - Water Vapor
  - Radiation
  - CO2
  - Ozone
- Tropical variability (incl. Monsoon)
- Southern Ocean
- More Observations (e.g., obs4MIPs, ESA CCI)

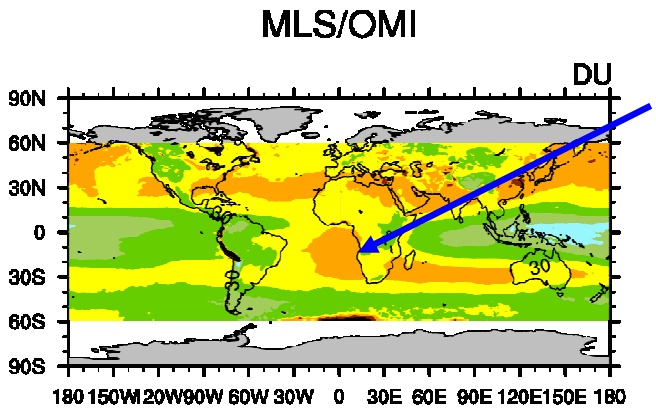
Goal: Standard namelists to reproduce certain reports or papers (e.g., Ch8 AR4, Ch9 AR5)

Extensions within the IGAC/SPARC Chemistry-Climate Model Initiative (CCMI)

Additional tropospheric diagnostics (composition, climate)
Example: Evaluation of IGAC/SPARC Ozone Database

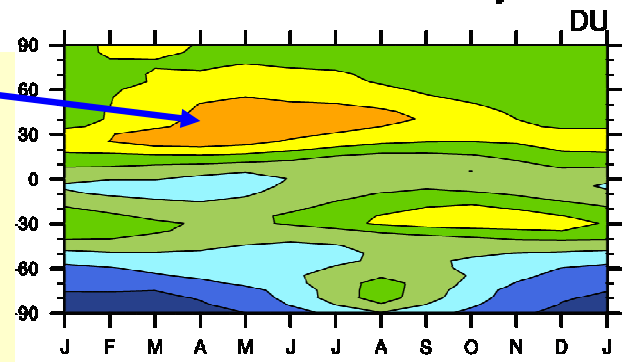
Large values over industrial emission regions (Chandra 2004)

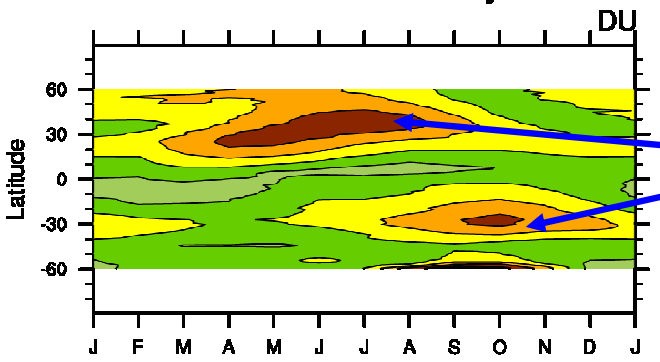


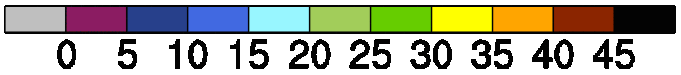


Effect of biomass burning in both Africa and South America (Ziemke et al., 2006)

Seasonal increase of tropospheric ozone is slightly understimate compared to MLS/OMI







Seasonal increase due to the combined effects of STE and photochemistry (de Laat et al., 2005). It is larger in the NH because of the larger emissions of anthropogenic pollution, biogenic VOCs and NOx (Ziemke et al., 2006)

Cionni et al., ACP (2011)

Summary and Conclusions

Advantages of a quantitative multi-model evaluation demonstrated, e.g.

- E.g., quantitative assessment of model improvements for single model and generations of models (e.g. CCMVal-1 vrs CCMVal-2, CMIP3 versus CMIP5).
- Will continue into CMIP6
- Requires accurate observations with quantified uncertainty (beyond ECVs)

Several international activities in place to move ahead ESM evaluation

- WGNE/WGCM metrics panel
- Model Intercomparison Projects (MIPs)
- obs4MIPs ←→ CMIP
- Beginning of sharing of diagnostic code and common tools

ESA CCI great initiative, some recommendations:

- Show the feasibility of ESA CCI data for climate model applications
- Ensure easy and free access to ESA CCI products
- Preferably: contribute data to obs4MIPs along with technical documentation
- Contribute ESA data to the development of community diagnostic tools

=> Will do a long way towards data being used by the climate model community

DLR

Future Emphasis and Needs

Evaluation:

- **Spatial and temporal covariability patterns** between Essential Climate Variables (ECVs) across the atmosphere, ocean and terrestrial domains.
- An in-depth analysis of the underlying controls of the **seasonal cycle, IAV** and **long-term temporal trends** in these variables and **processes** can help to understand the spread in model projections over the coming decades => priority for model development?
- **Development of statistical frameworks** rather than heuristic model weighting.

Model Simulations:

- **Targeted model simulations and output** to improve process-understanding
- **Coupled and uncoupled runs**, better handling of forcings

Observations for Model Evaluation:

- **Need good enough observations to discriminate models (note: models are getting better!)**
- **Fully exploit available observations for model evaluation considering uncertainties.**
- **Identify additional observations to include in obs4MIPs** (broader participation, e.g. ESA CCI, CFMIP observations) with guidance from WCRP (e.g. WDAC, WMAC, WGs & projects).
- **Improve comparability between models and observations** (e.g., CCMI insitu expert group, satellite simulators like COSP).

Diagnostic and Performance Metrics Tools:

- **Develop and share common diagnostic tools** that routinely run on CMIP output and according observations (obs4MIPs) on the Earth System Grid Federation (ESGF).

=> Benchmark for model evaluation that will over time lead to model improvements