Metrics Breakout Group

3rd WGNE Workshop on
Systematic Errors in Climate and NWP Models

San Francisco, CA
12-16 February 2007

Some questions concerning metrics

- What value is there in encouraging a more routine application of metrics to climate models?
- Should we be wary of metrics? What are the dangers?
- What are the outstanding challenges?
  - What is the relationship between skill in simulating observed phenomenon and (unobserved) future climate?
  - For a given application, is there some minimum set of metrics that can be objectively justified for gauging climate model reliability?
  - Is it useful and justifiable to construct a single metric to gauge model performance or weight individual model predictions?
  - Others?

Some questions concerning metrics (cont.)

- Technical issues:
  - Would it be useful to develop a much more comprehensive suite of metrics?
  - What approaches might reduce the number of metrics considered without reducing the information content?
  - Are there ideas on how to prevent metrics from being "played"?
- What can be done to foster/facilitate progress in this area?
  - Coordinating groups (e.g., WCRP, GEWEX)
  - Funding agencies
  - Institutions (e.g., PCMDI)
  - Ad hoc groups (e.g., this break out group)

Diagnostics vs. Metrics

- A metric is a measure of some model characteristic (usually some aspect of model fidelity), which is expressible as a scalar

- It may alert us to some model shortcoming, but
  - Won't indicate *why* something is wrong.
  - Can't suggest how to cure it.

Metrics confusion

- Flavo(u)rs of metrics
  - Assess performance (requires observations)
  - Quantify some model characteristic
- Performance assessment metrics have a long history in NWP:
  - Grade forecasters
  - Monitor changes in performance
  - Gauge relative skill of forecasting systems
- Increasing interest in climate model metrics
- Potential uses of performance metrics
  - Assess model fidelity in simulating present and past climate
  - Determine reliability of future projections (weight individual models?)

Some ideas that seemed to resonate

- Good to provide a "basket" of metrics assessing a wide range of
  - Variables
  - Processes
  - Phenomena
  - Time-scales
  - Regions/space-scales
- Let users select which metrics are most relevant to their particular needs.
- Refrain from computing a single overall model skill index
  - Don't know how to compute (most certainly depends on application)
  - Invites misuse/abuse.

Some ideas that seemed to resonate (cont.)

- Metrics that focus on model fidelity in representing specific processes would be highly useful
  - Might involve characterizing lagged covariance relationships among interacting variables.

Why should we work toward a standard, reasonably comprehensive set of metrics?

- Guard against a tendency for individuals to focus on only the phenomenon/time-scales/space-scales/variables of interest to them.
- Provide information for scientists interested in selecting a model for a specific application.
- Facilitate monitoring and documenting of changes in model performance.
- Promote healthy competition among modeling centers.

Where to start

- Model developers traditionally have tried to get the current climate state right, so a minimum set of metrics should characterize fidelity in this regard.
- Augment this base set with a wide variety of additional metrics.
- Metrics quantifying ability to represent various processes accurately would be be valuable in assessing whether a good simulation is obtained for the right reasons.


PCMDI

Discussions among a subset of the WGNE climate metrics panel

- Focus on more than one field (perhaps ~ 10), but consider only the atmosphere for now.
- Start with climatological annual cycle of the global pattern of these fields.
- Rule out metrics that are sensitive to different observational datasets.
- Avoid metrics that are too difficult to calculate (or too difficult to understand by program managers and non-experts)
- Desirable to quantify uncertainty due to observational error and sampling errors.

Next steps

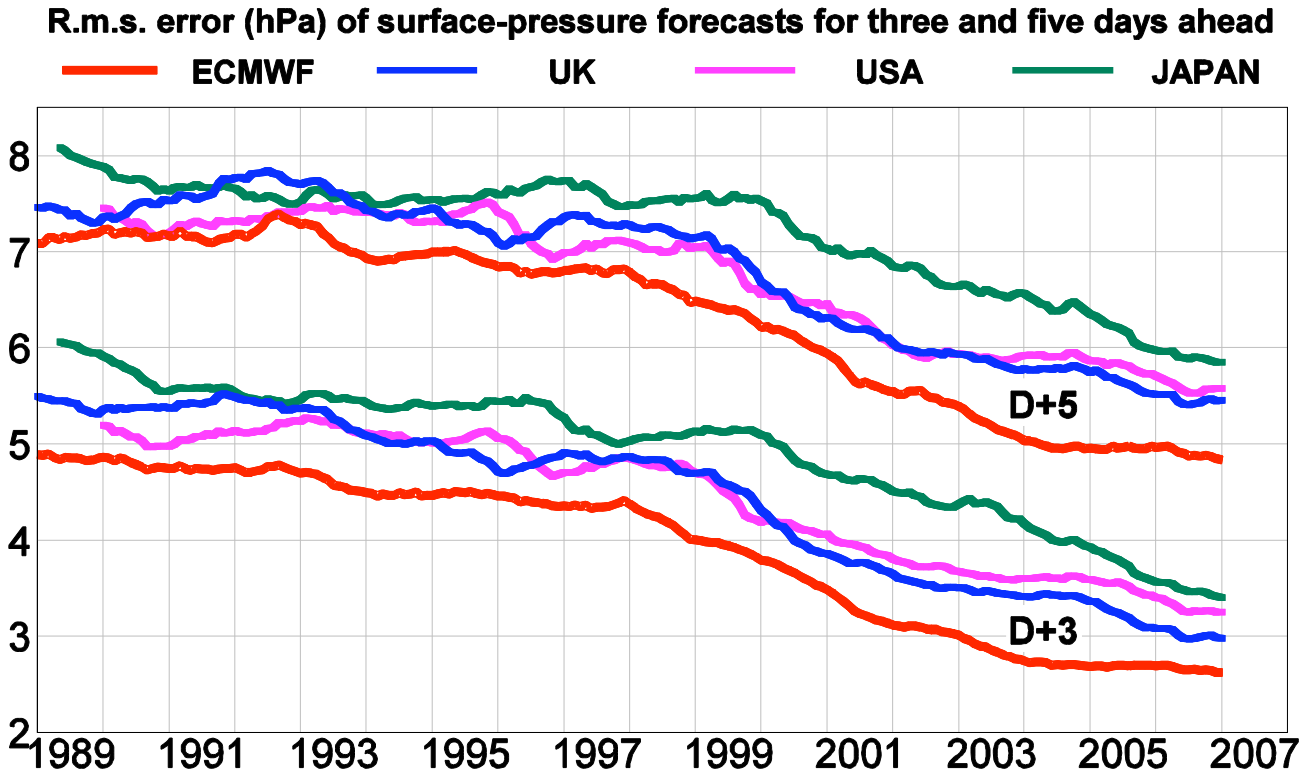
- Propose an initial set of standard (global) climate metrics for atmospheric models.
  - Collect metrics being developed by various researchers
  - Evaluate them against the criteria we've discussed
- Encourage development of metrics for other component models and for specific phenomena
  - Ocean, biogeochemistry, land surface, sea ice …
  - Cloud processes, monsoon, MJO …
- Look to ongoing research to provide rigorous justification for
  - Selecting a minimum set of metrics that need to be considered
  - Applying a metric-based index to weight climate change simulations by models based on their simulation of present climate

The GCSS is interested in metrics useful for assessing skill in simulating clouds and precipitation processes

- Robert Pincus has taken the lead on this.
- Focusing on LW & SW radiation at top of atmosphere and precipitation.
- Other groups are showing interests in developing metrics for assessing other aspects of model simulations.

Monitoring evolution of model performance: An example from operational weather forecast systems

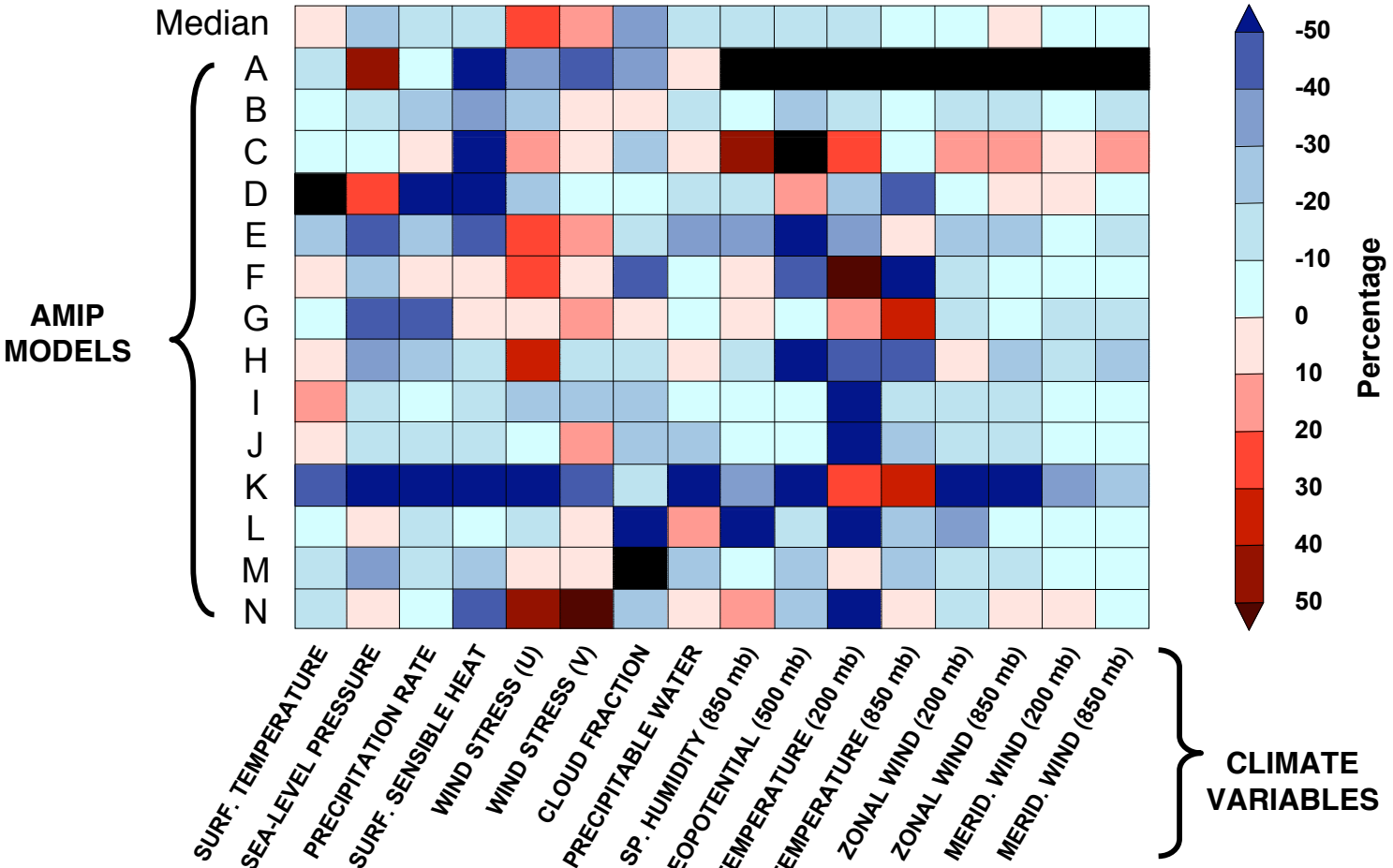
- WGNE routinely reviews skill of daily forecasts
- Indicates improvements and deficiencies in individual forecast systems



Courtesy of M. Miller

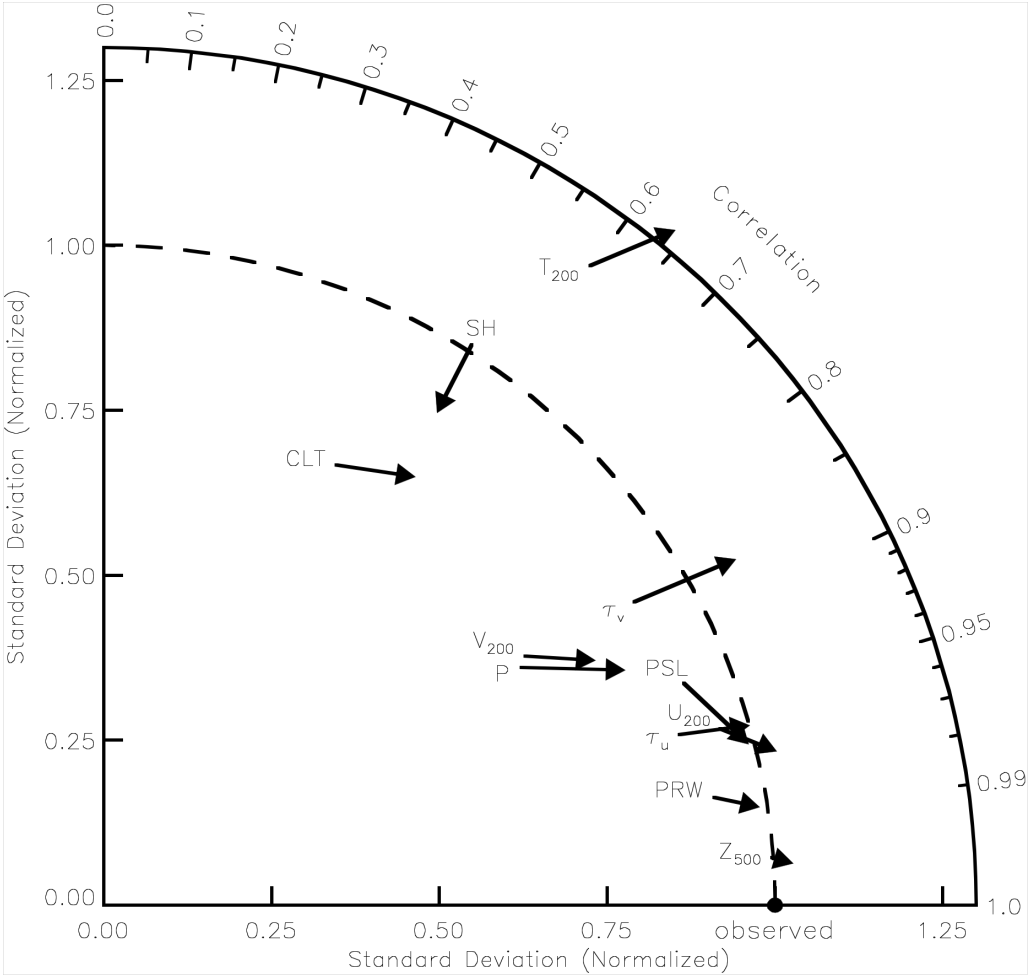
AMIP models showed improvement during the '90s

Annual cycle of global patterns: Percentage change in total error: $100 \times \frac{E_{\text{AMIP2}} - E_{\text{AMIP1}}}{E_{\text{AMIP2}}}$



Multiple statistics for provide a more comprehensive picture of changes in AMIP median model performance

Change from early to late 1990's

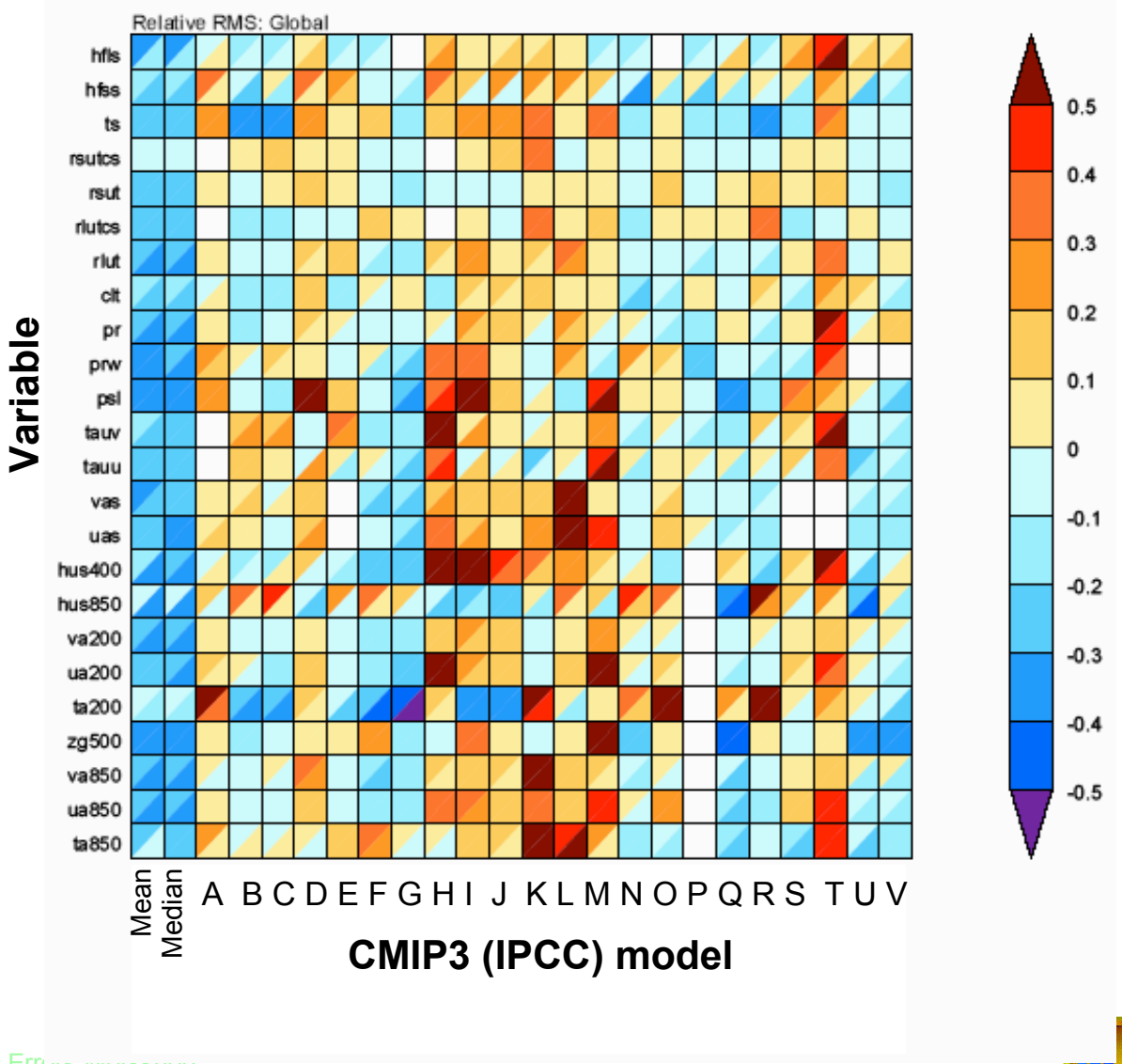


Example: Quantitative assessment of relative skill (*S*) of large collections of models

E_{vm} = RMS error in simulating the spatial pattern of the climatological annual cycle of variable v by model m

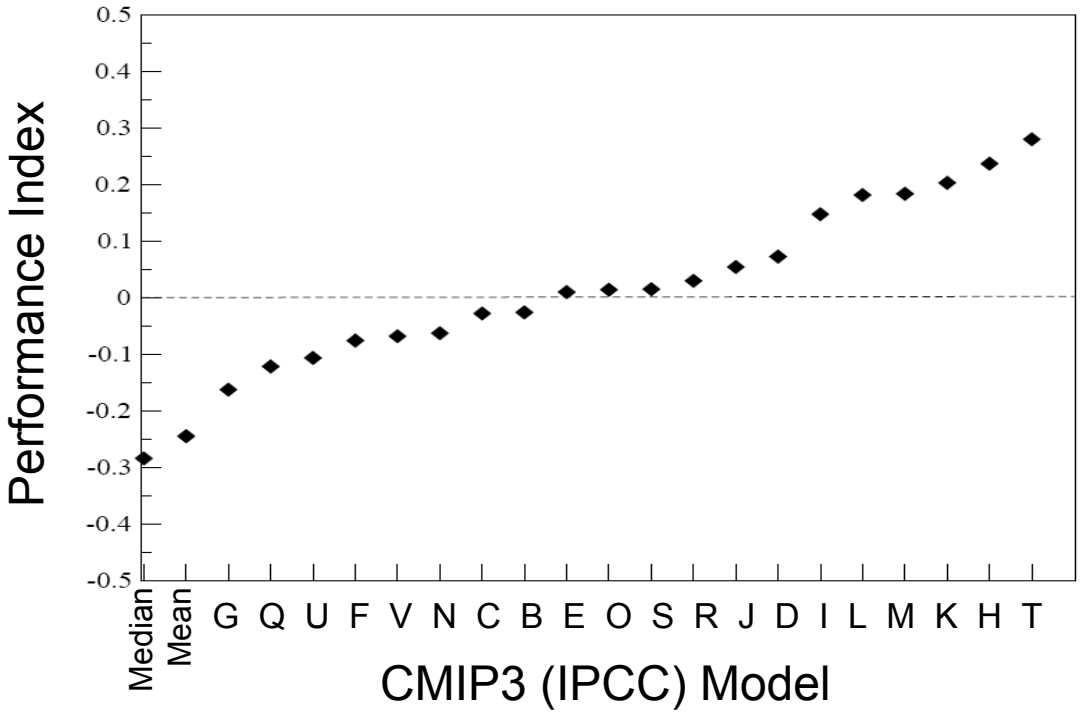
$$S_{vm} = \frac{E_{vm} - \hat{E}_v}{\hat{E}_v}$$

where $\hat{E}_v$ is the median of the individual error measures, E_{vm}



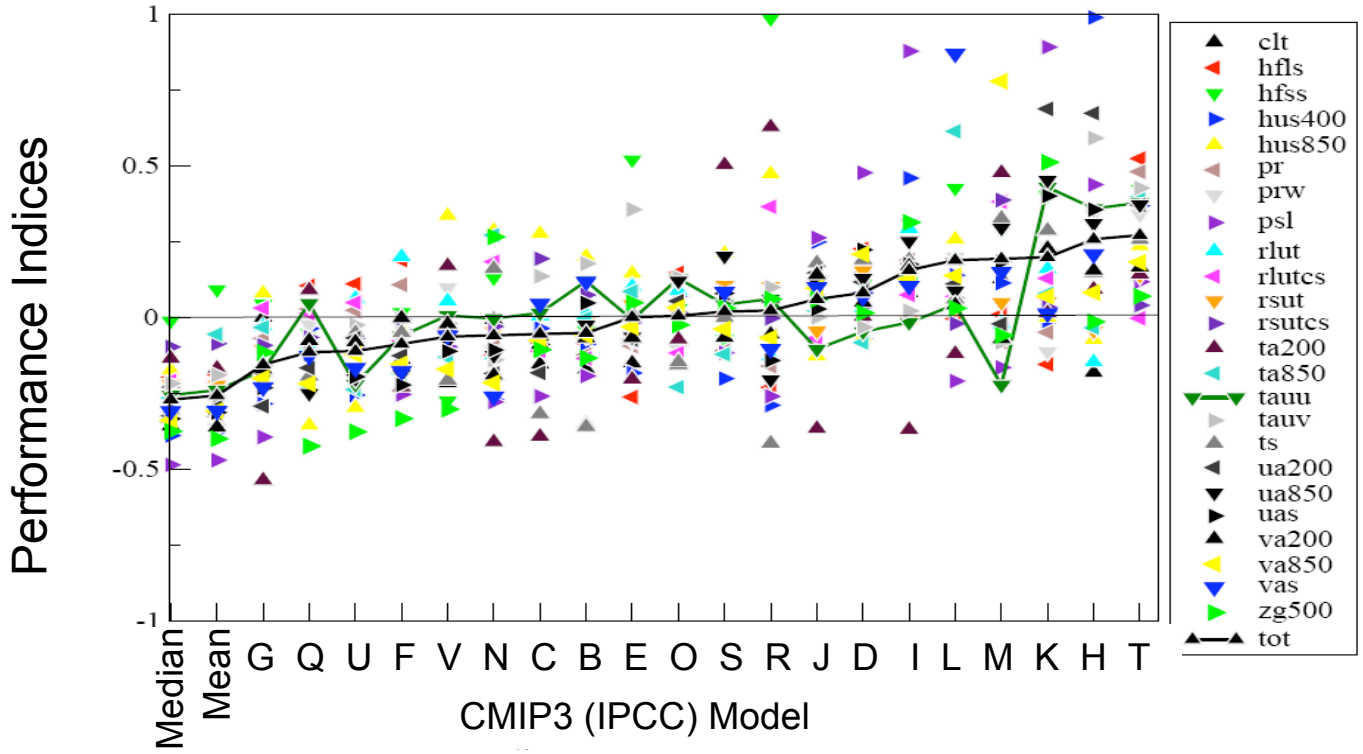
Construction of a "simulation quality" index:

- From performance portrait recall: $S_{vm} = \dfrac{E_{vm} - \hat{E}_v}{\hat{E}_v}$
- Let the performance index $\overline{S}_m$ be the mean of S_{vm} over all the variables.



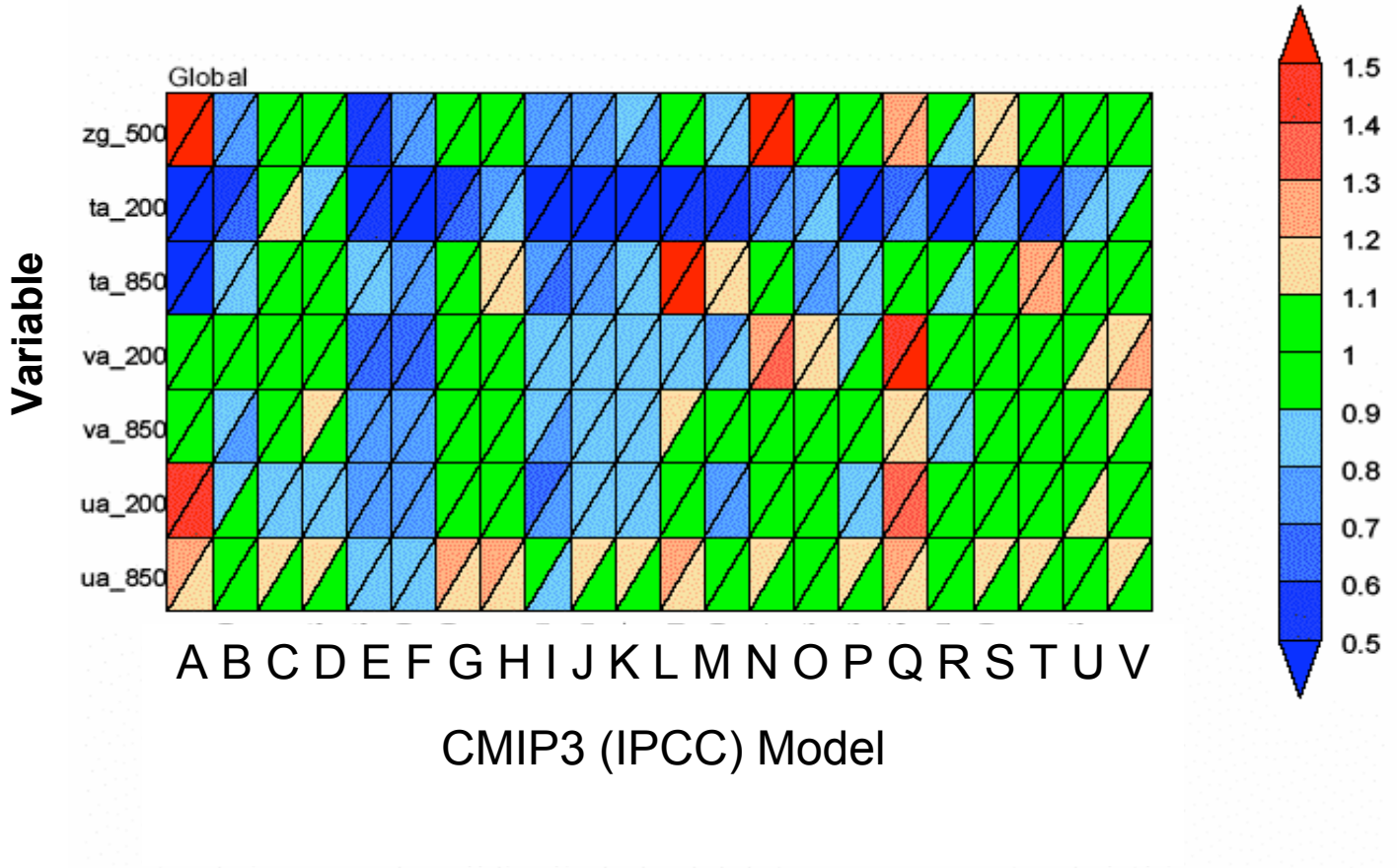
Is the performance index useful?

- Answer is unknown, but it almost certainly depends on the application.
- Does it make sense to rank models based on an index for which even the "best" model simulates some fields with errors larger than those found in most other models?



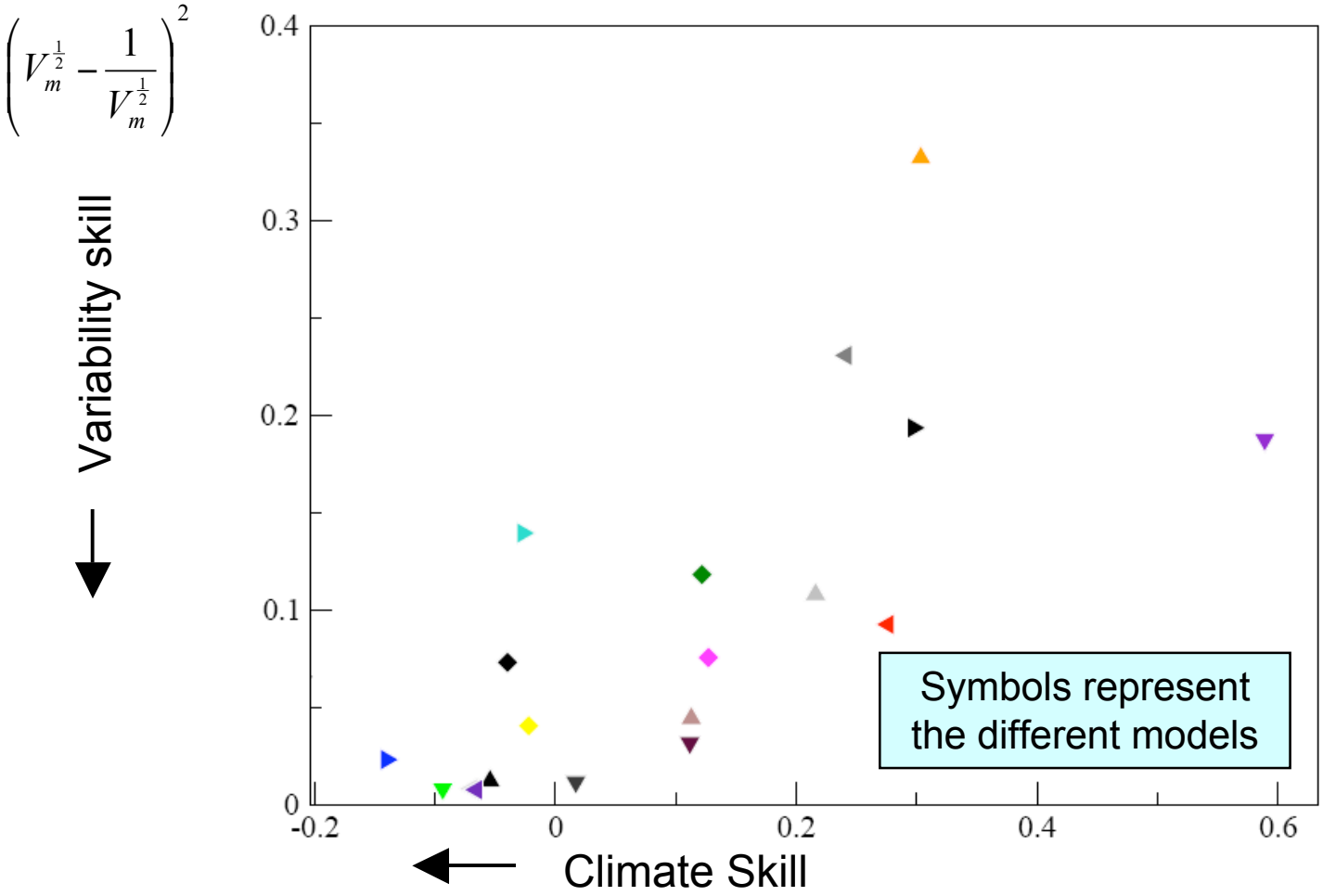
What if we focus on the variability of monthly anomalies in the free-atmosphere fields?

- Plot $V_{vm} = \dfrac{\sigma_{vm}^{\,2}}{\sigma_{v,\mathrm{obs}}^{\,2}}$

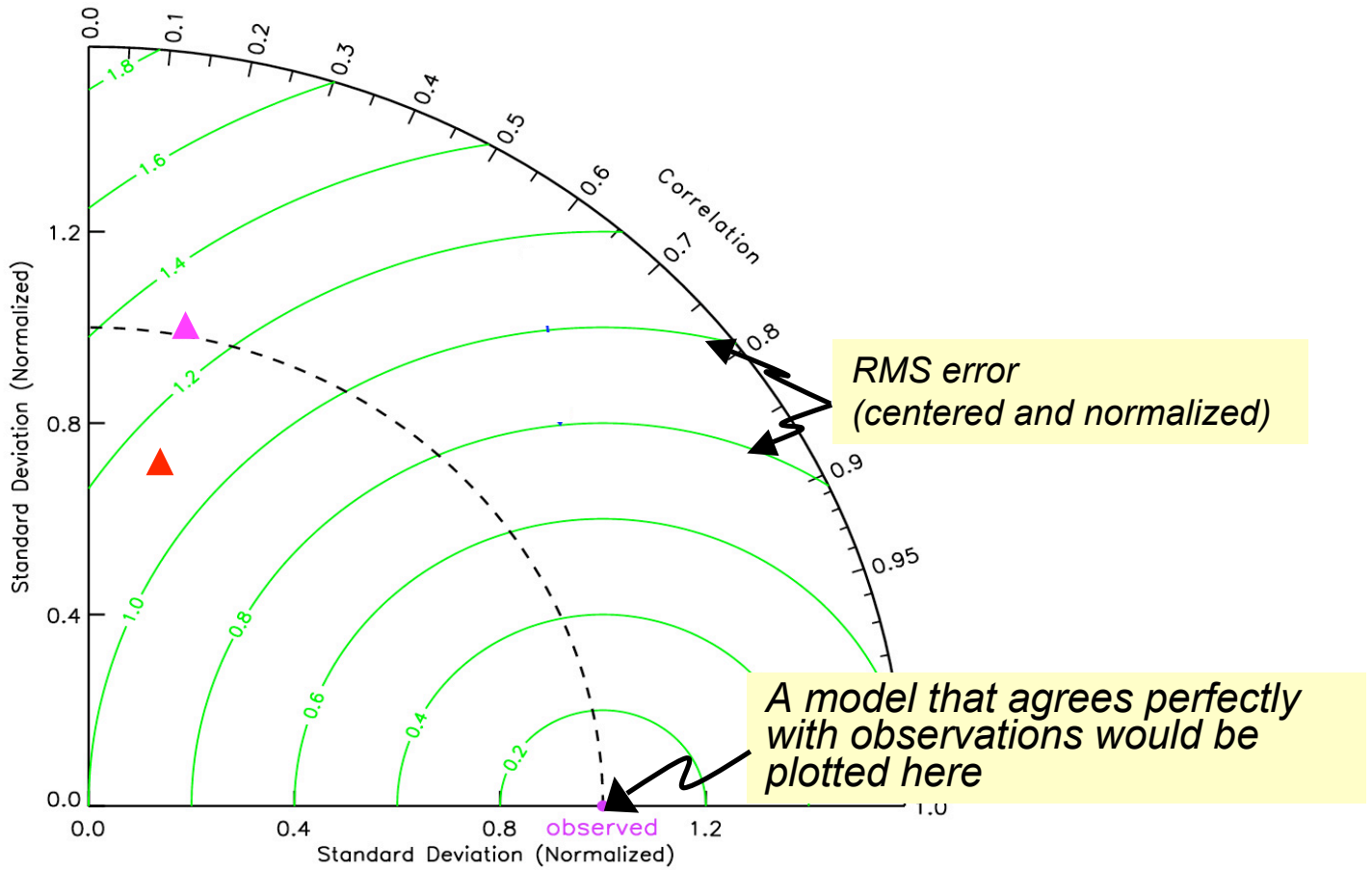


Is skill in simulating the variance of monthly anomalies related to skill in simulating climatology?

Reliance on a single index may be misleading.

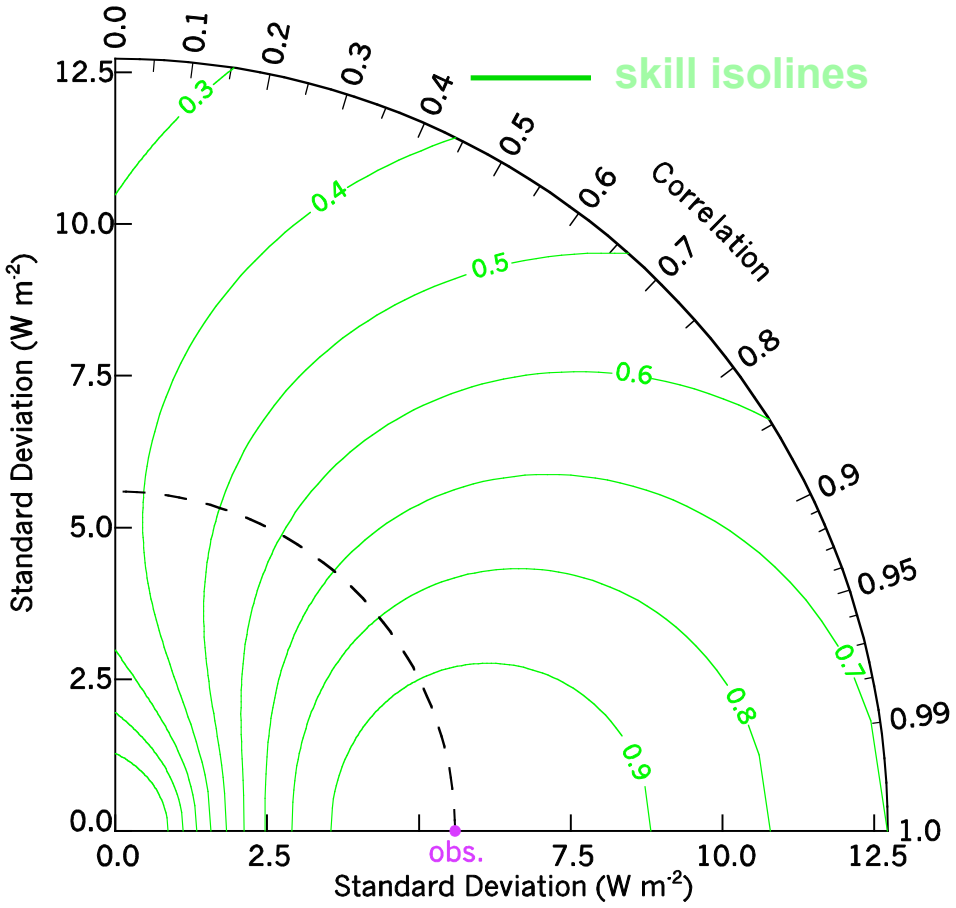


The RMS error can be misleading, especially for poorly simulated fields.



Taylor, *J. Geophys. Res.* (2001)

Prevent "cheating": devise skill scores that penalize filtering



Define "centered" skill score:

$$S' = \exp\left(\frac{-E'^2}{2\sigma_r\sigma_f}\right)$$

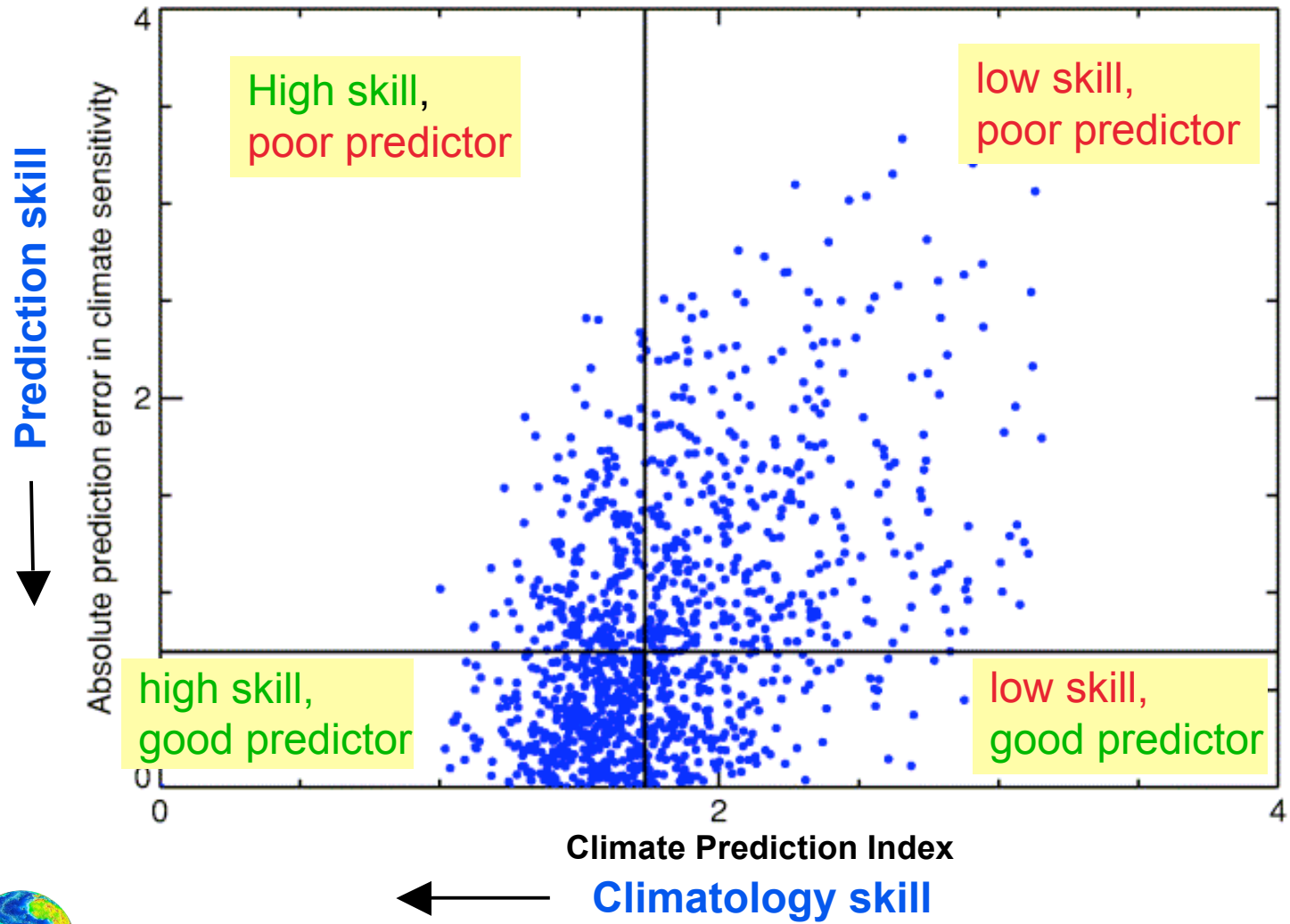
where *E'* is the centered RMS error

This skill score:

- Ranges from 0 to 1
- Decreases with increasing RMS error
- For a given variance, decreases with decreasing correlation
- For a given correlation, decreases as variance strays from correct variance
- Independent of which field is considered the "reference"

Is the climate prediction index relevant to climate change prediction?

Perfect model test



Courtesy of J. Murphy

Summary

- For climate models, we have traditionally summarized model performance with a collection of metrics, mostly focusing on large-scale climatology.
- The scientific community, funding agencies, and policy makers are interested in "which model is best?"
  - This question is not specific enough.
  - Although single "performance indices" can be proposed, there is currently little rigorous scientific justification for paying much attention to them.
- There is value in relying on multi-model ensembles to provide the "best simulation" and to help gauge uncertainty.
- Little work has been done to relate climate model performance (in terms of present day simulation) to quality of climate prediction.
- Metrics can be used to identify model errors, but rarely reveal what's to blame.

The suite of "present climate" metrics should be augmented by statistics characterizing

- Variability on a range of time-scales (from diurnal to long-term trends)
- Regional performance in key areas
- Representation of key physical processes and phenomenon (e.g., Cloud processes, monsoon, MJO …)
- Other components of the climate system (oceans, land-surface, carbon cycle)

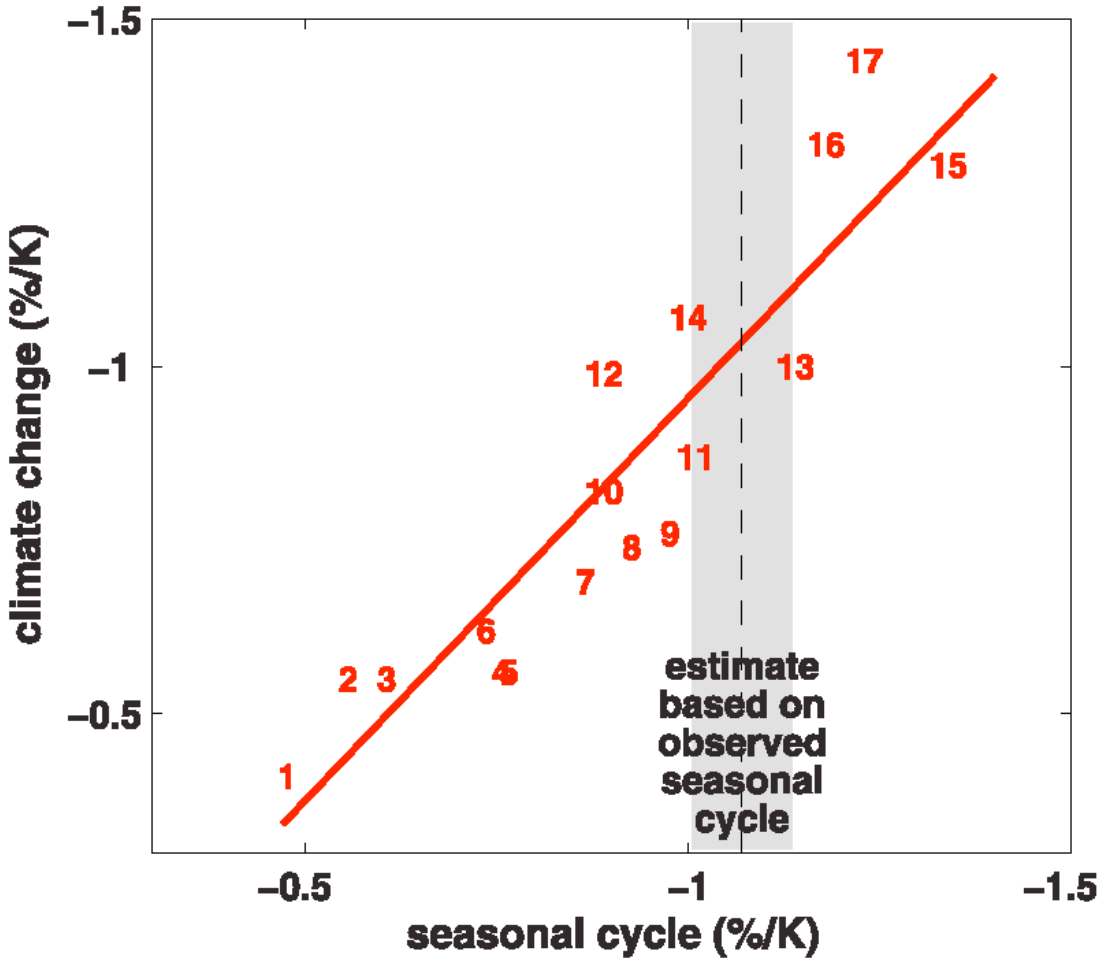
Research and community involvement needed

- PCMDI is working to produce a comprehensive set of metrics.
  - We welcome collaborators!
- PCMDI plans to continue support of "benchmark" experiments (e.g., AMIP, CMIP 20th Century) which
  - Make it possible to track model improvement
  - Can facilitate development of new useful metrics
- With interest from WGNE, GEWEX, and other groups, we should work to establish a set of standard metrics for climate models (following the NWP community).

Fundamental research questions

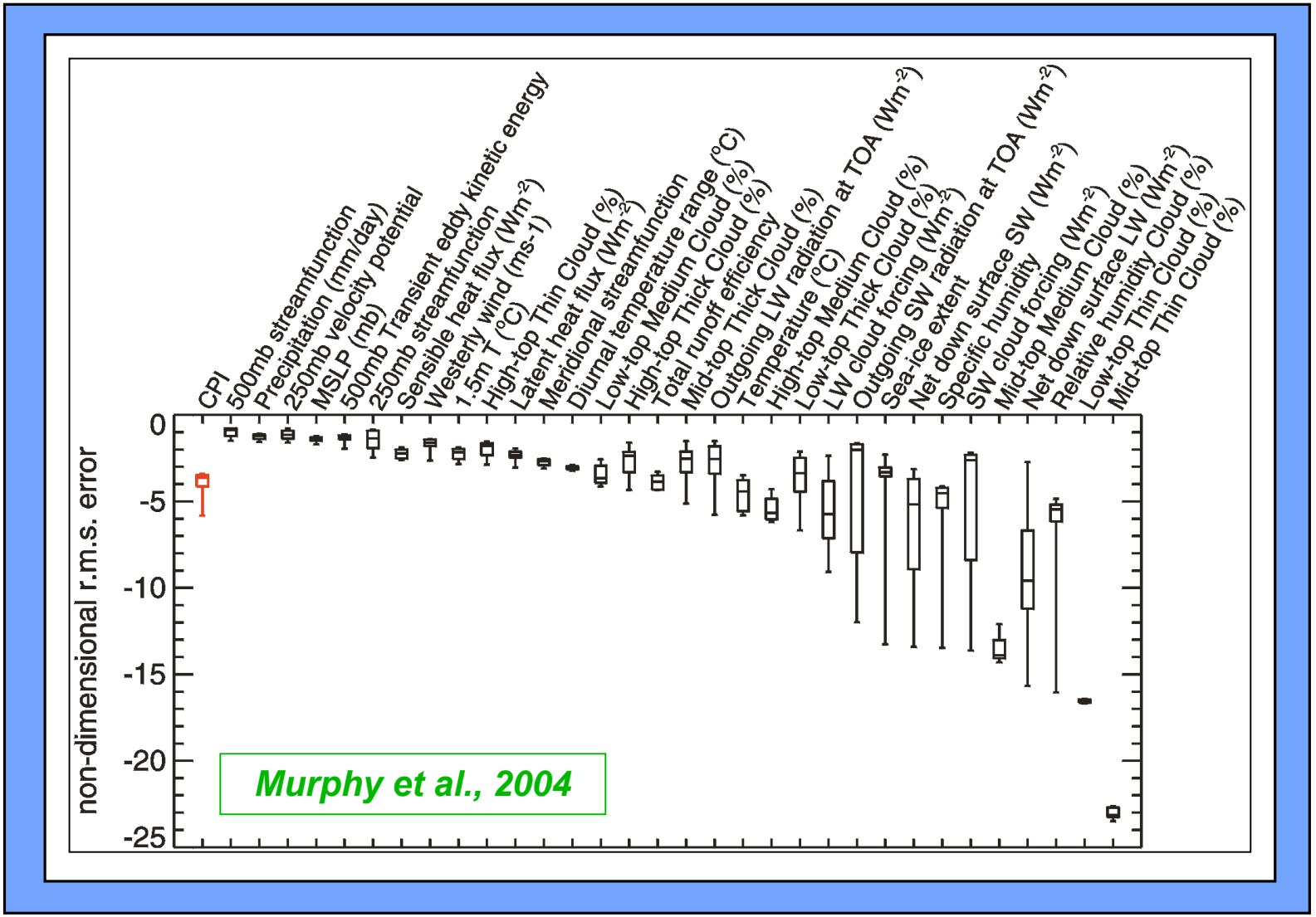
- What is the relationship between skill in simulating observed phenomenon and (unobserved) future climate?
  - "Perfect model" experiments
  - Identification of processes critical to future climate change that can be thoroughly validated on shorter time-scales
- For a given application, is there some minimum set of metrics that can be objectively justified for gauging climate model reliability?
- Can we justifiably construct a single metric
  - To gauge reliability of individual model predictions?
  - To produce an optimally-weighted consensus prediction?

Response of snow cover to global warming in models is related to their snow response to spring warming

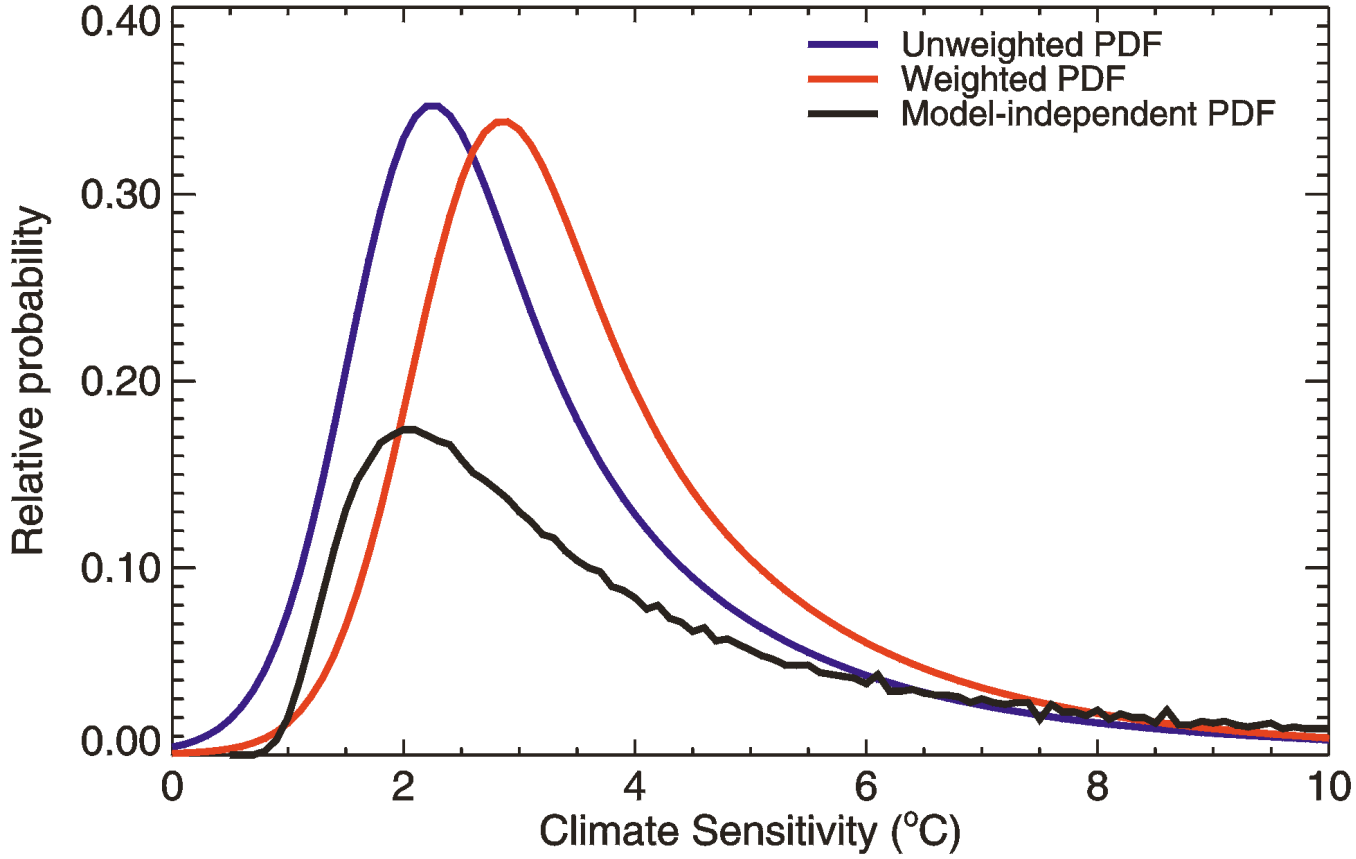


Hall & Xu, 2006

A "climate prediction index" was proposed, based on 32 different fields.

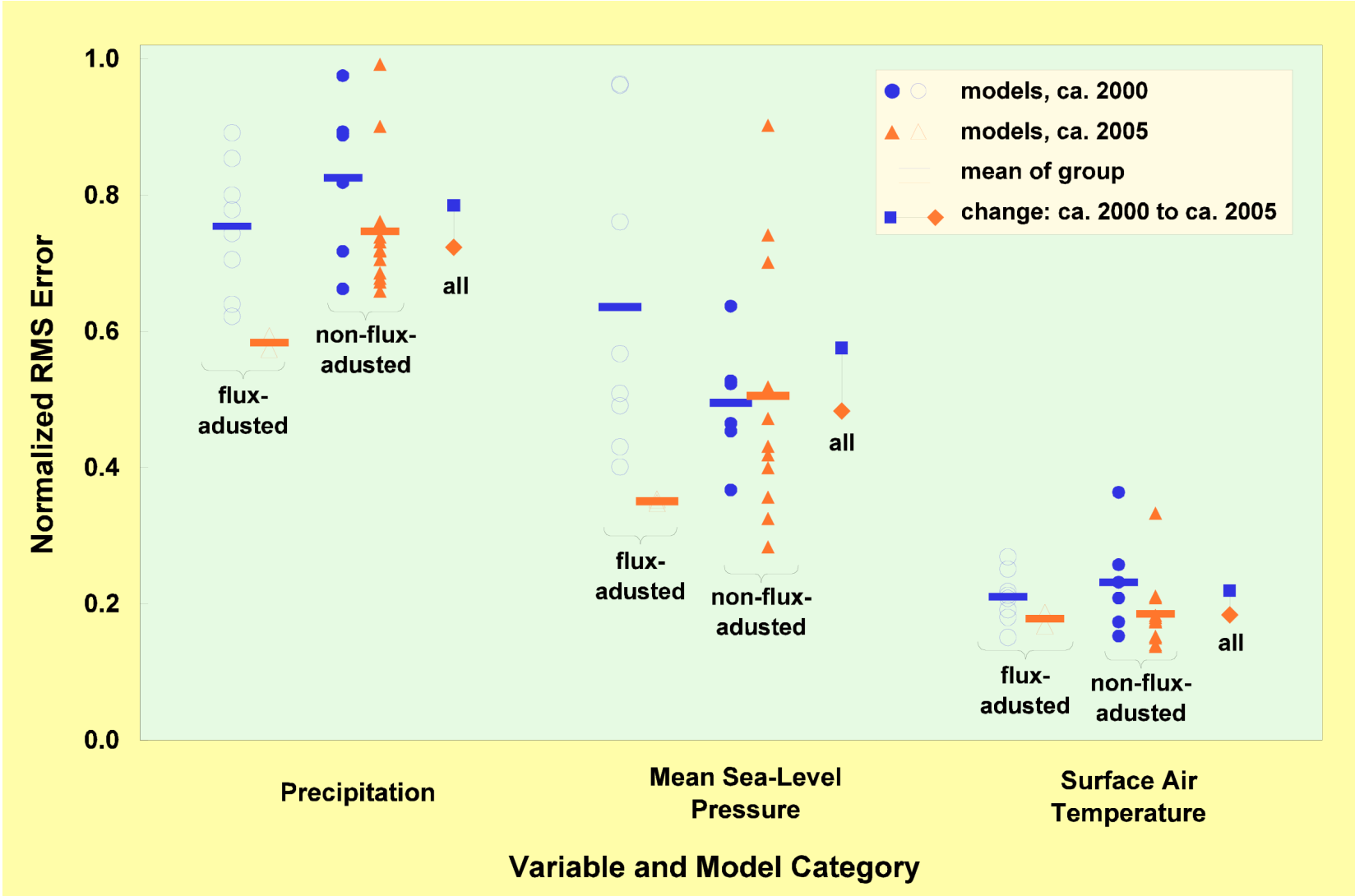


The "climate prediction index" was used to weight results in producing a PDF for climate sensitivity.

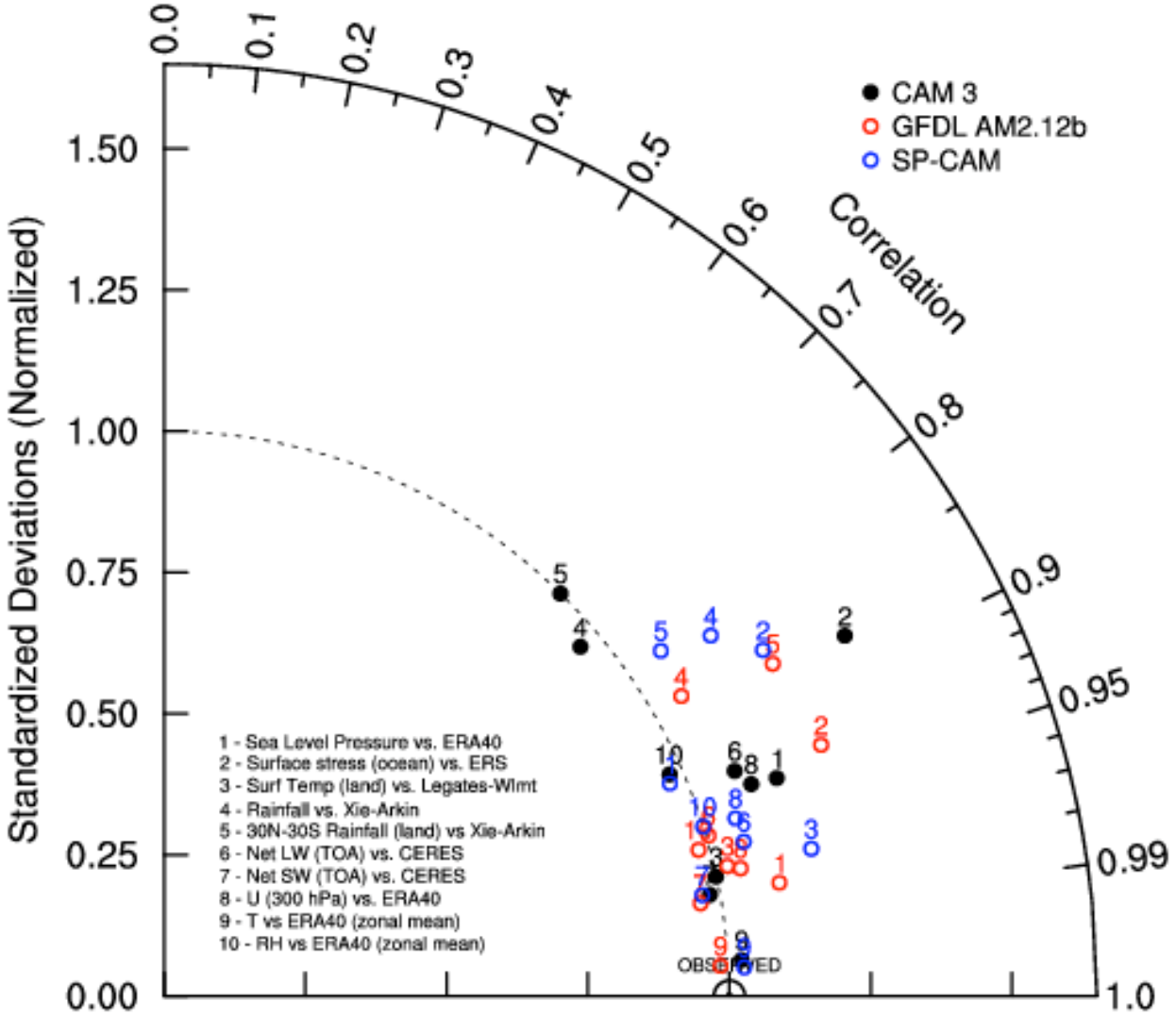


Murphy, Sexton, Barnett, Jones, Webb, 2004

Coupled model improvement in simulating three variables: ca. 2000 to ca. 2005



C. Bretherton has proposed metrics to help in selecting an atmospheric model suitable for coupling to an ocean:



Courtesy of Bretherton & Wyant

Combine error metrics to form a "climate bias index"

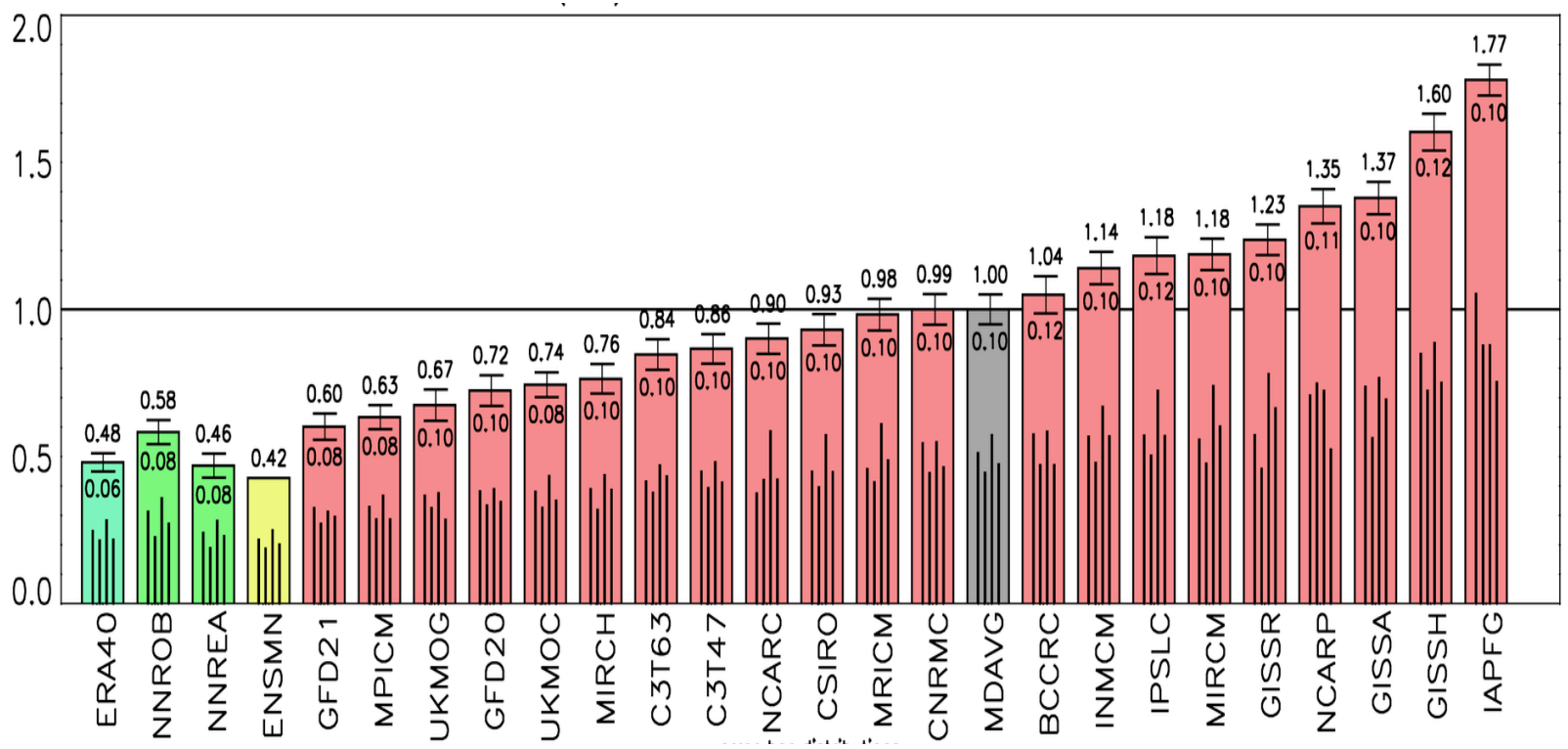
- Climate bias index:

Courtesy of Bretherton & Wyant

$$CBI = \frac{1}{10}\sum_{k=1}^{10}\frac{\varepsilon_k}{\varepsilon_k^{ref}}$$

- Scores calculated for 3 models (with CAM3.0 as reference):
  - CAM3 (T42): 1.00 (AMIP: 0.96; FV2x2.5-AMIP: 0.97)
  - SP-CAM: 0.92
  - AM2.12b: 0.76

An index, based on 35 individual metrics, has been used to rank CMIP3 (IPCC) models.



Courtesy of Reichler & Kim

Apparent relationship between skill in simulating annual cycle + interannual variability and climate sensitivity