Uses of Metrics in the Evaluation and Application of Climate Models

Karl E. Taylor and Peter J. Gleckler

Program for Climate Model Diagnosis and Intercomparison
Lawrence Livermore National Laboratory

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

San Francisco, CA

12-16 February 2007

Themes and Focus

- The usefulness of any particular set of metrics depends on the application.
- The relationship between a climate model's skill in simulating present conditions and its predictive reliability is largely unknown.
- At this time, there is little justification for reliance on any single metric to gauge simulation quality.
- Roadmap
  - → Background, orientation, definitions
  - → Uses of metrics
  - → What's next?

What is meant by metric (as discussed here)?

- Objective measure of some characteristic of model behavior
  - → Objective measure of model performance – observations required.
  - → Other model characteristics – e.g., global climate sensitivity
- Usually a scalar statistical measure, e.g.,
  - → RMS error
  - → Correlation measures
  - → "skill score" based on one or several error measures
- Not usually "targeted" enough to diagnose reasons for model errors
  - → Provides symptoms (manifestations) of problems, but not particularly helpful in diagnosing causes

Performance metrics

- Absolute measure relative to observations – e.g., "The RMS error in simulating temperature relative to ERA40 is 1.3 K."
- Size of error relative to other models – e.g., "Model X's error is 10% larger than model Y's."
- Error relative to "unskilled" model
- Error relative to limits to agreement expected, given
  - → observational uncertainty
  - → sampling issues
  - → unforced variability
  - → the chaotic nature of some aspects of weather and climate

A popular question: "Which model is best?"

- Depends on application (weather? Climate change? ENSO?)
- Depends on computer resources available
- A better question: "Given X computing resources, which (weighted?) combination of models should be used for the purpose of…."

What can we do with metrics?

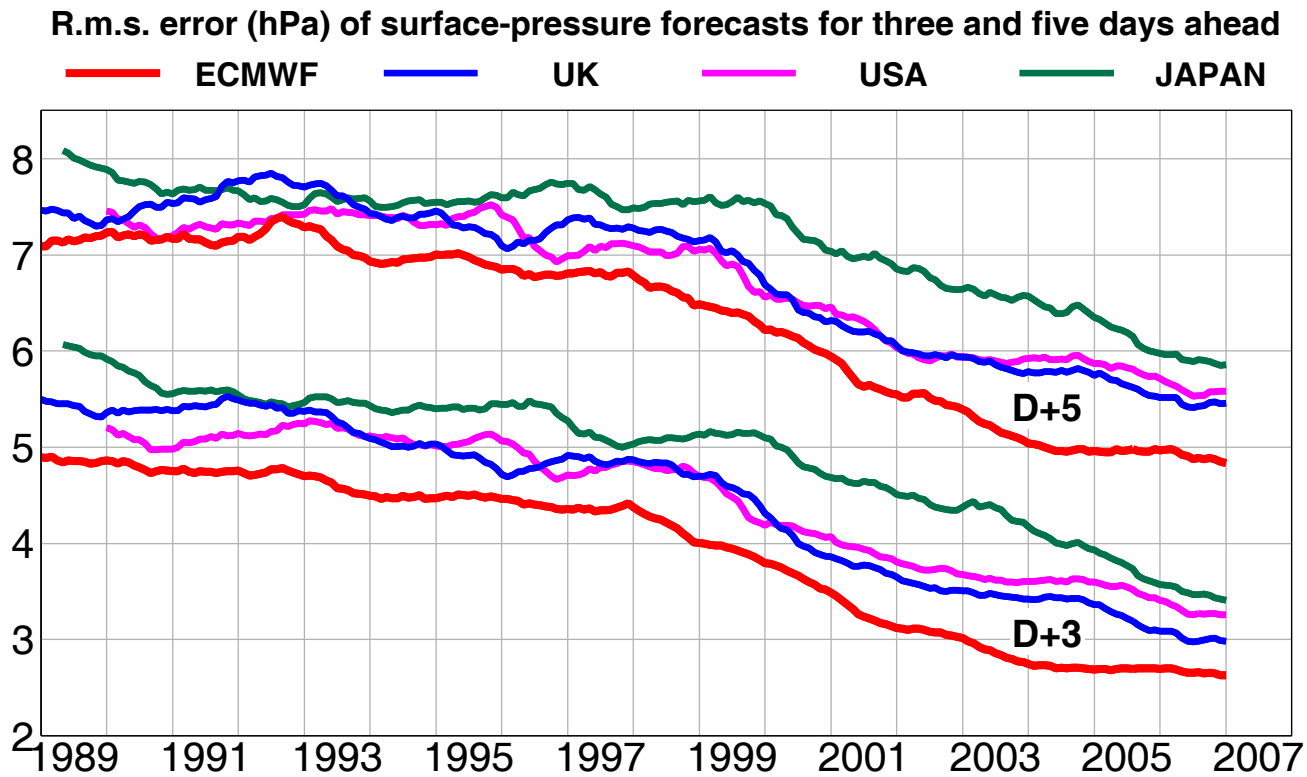
- Monitor changes in performance as models evolve
- Enable quantitative comparisons of model skill
  - → Quantify the relative merits of different models
  - → Aid model development and objective selection of a new model version
- Construct simulation skill indices, but recognize their limitations

What can we do with metrics?

- Monitor changes in performance as models evolve
- Enable quantitative comparisons of model skill
  - → Quantify the relative merits of different models
  - → Aid model development and objective selection of a new model version
- Construct simulation skill indices, but recognize their limitations

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

Challenges faced in evaluating climate models

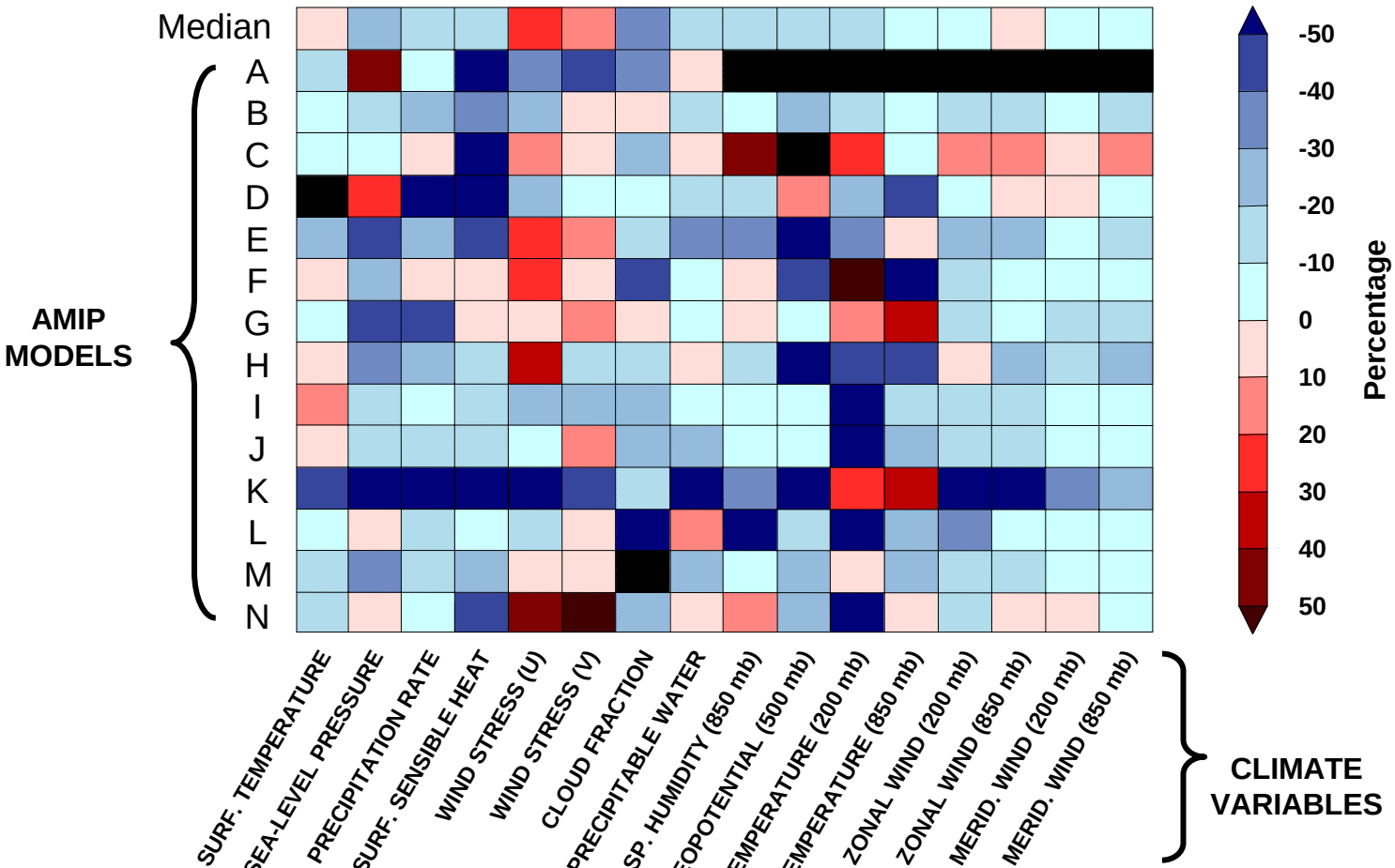
- Variety of fields to consider
- Multitude of phenomena / processes of interest
- Range of time and space scales
- A choice of error measures (e.g., RMS error, correlation)
- Limited opportunities for verification of "forecasts"
- Must account for observational errors/uncertainty
- Some aspects of simulations are not deterministic

Overall model performance (first presented ~1995)

- Initially limited evaluation to:
  - → ~ dozen "well-observed" atmospheric variables
  - → Space-scale: global domain, coarse model grid (~T42)
  - → Time-scale: annual cycle
  - → Phenomenon: large-scale characteristics of mean-climate state

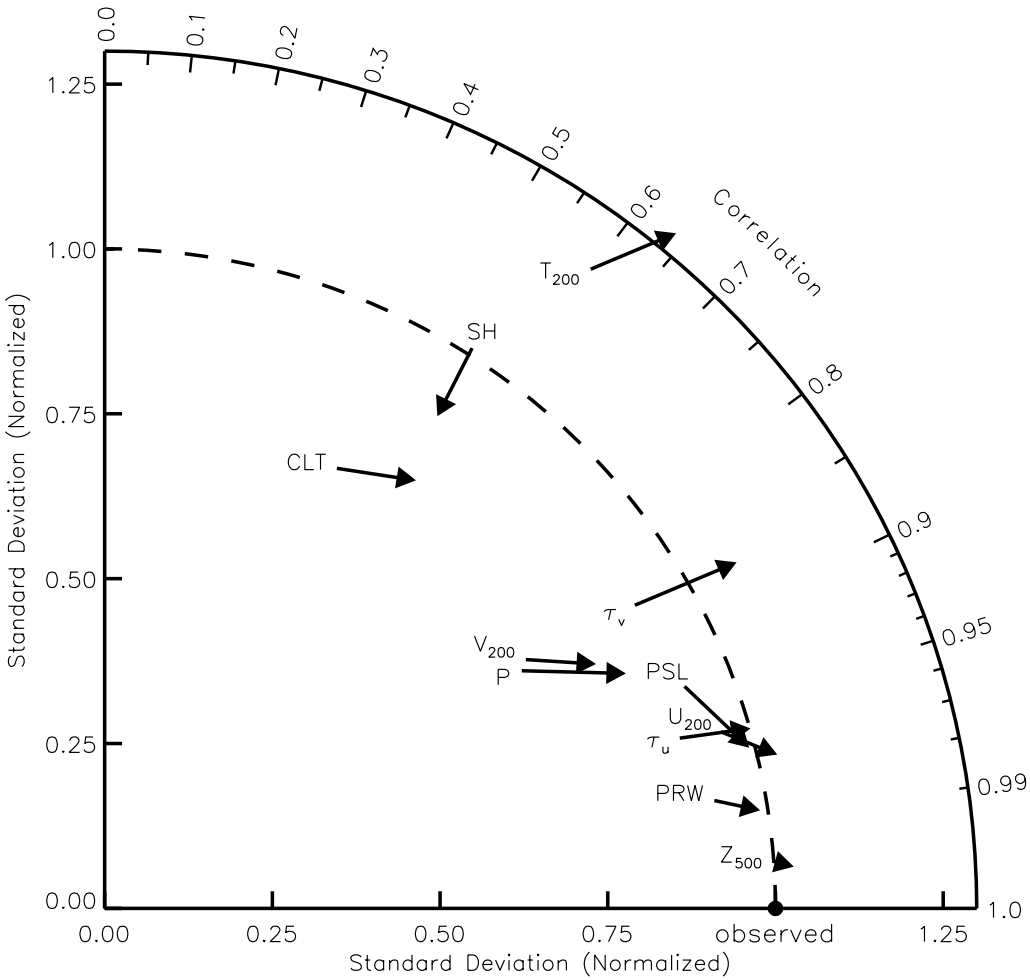
AMIP models showed improvement during the '90s

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



What can we do with metrics?

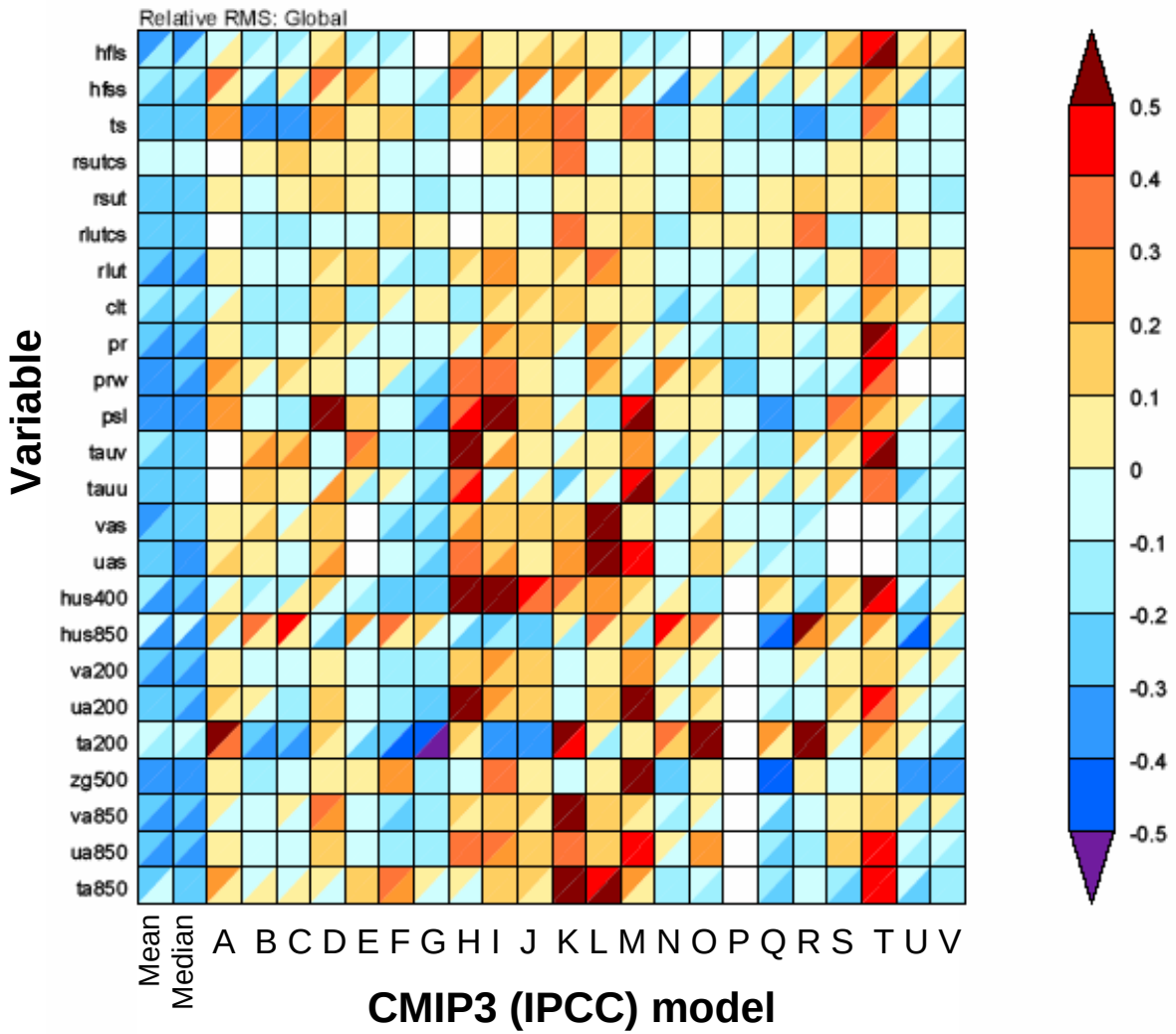
- Monitor changes in performance as models evolve
- Enable quantitative comparisons of model skill
  - → Quantify the relative merits of different models
  - → Aid model development and objective selection of a new model version
- Construct simulation skill indices, but recognize their limitations

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}



What's missed in above performance portrait?

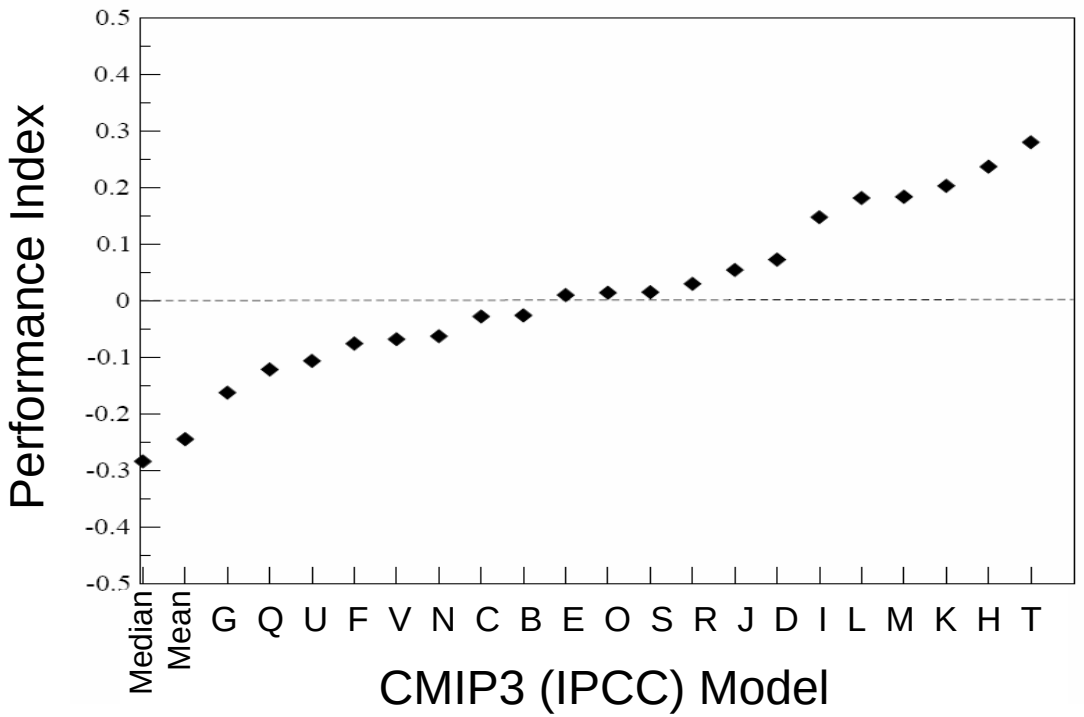
- Skill in simulating specific phenomenon (e.g., weather, convection, boundary layer processes, ENSO, MJO, NAO)
- Regional performance
- Skill in simulating other time scales (e.g, diurnal cycle, long-term trends in climate)
- Skill in simulating unusual events (e.g., heat waves)
- Should we *expand the suite of metrics?*

What can we do with metrics?

- Monitor changes in performance as models evolve
- Enable quantitative comparisons of model skill
  - → Quantify the relative merits of different models
  - → Aid model development and objective selection of a new model version
- Construct simulation skill indices, but recognize their limitations

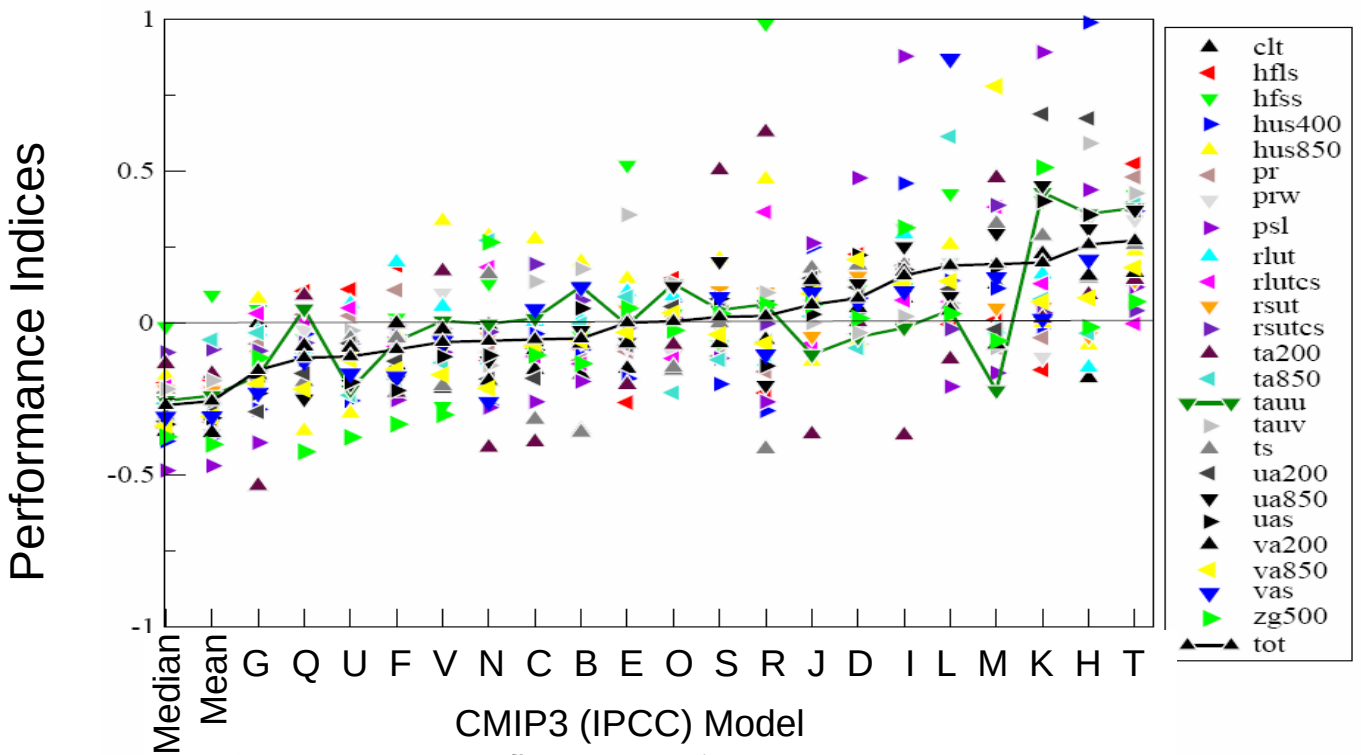
Construction of a "simulation quality" index:

- 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.



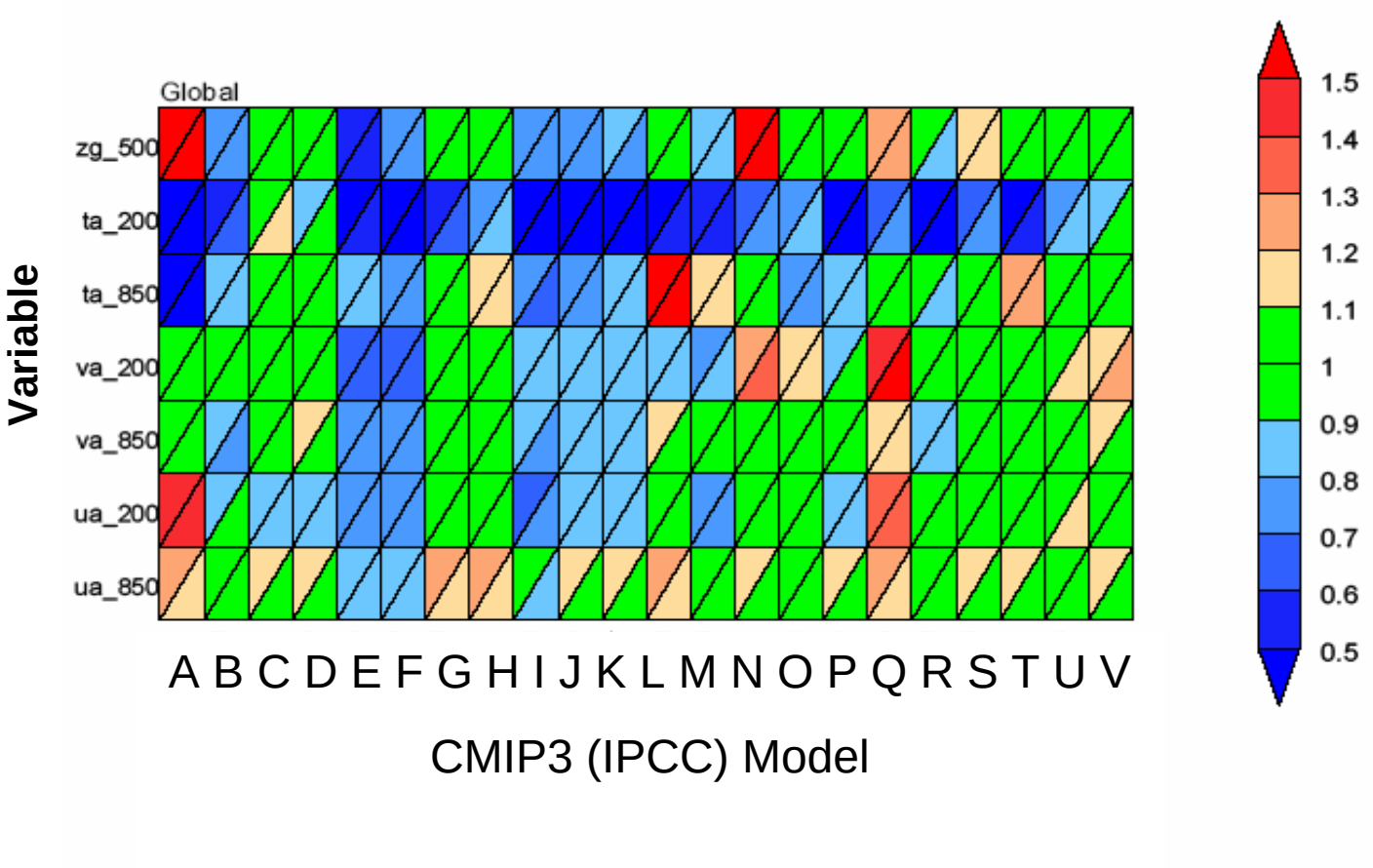
Is the performance index useful?

- 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?

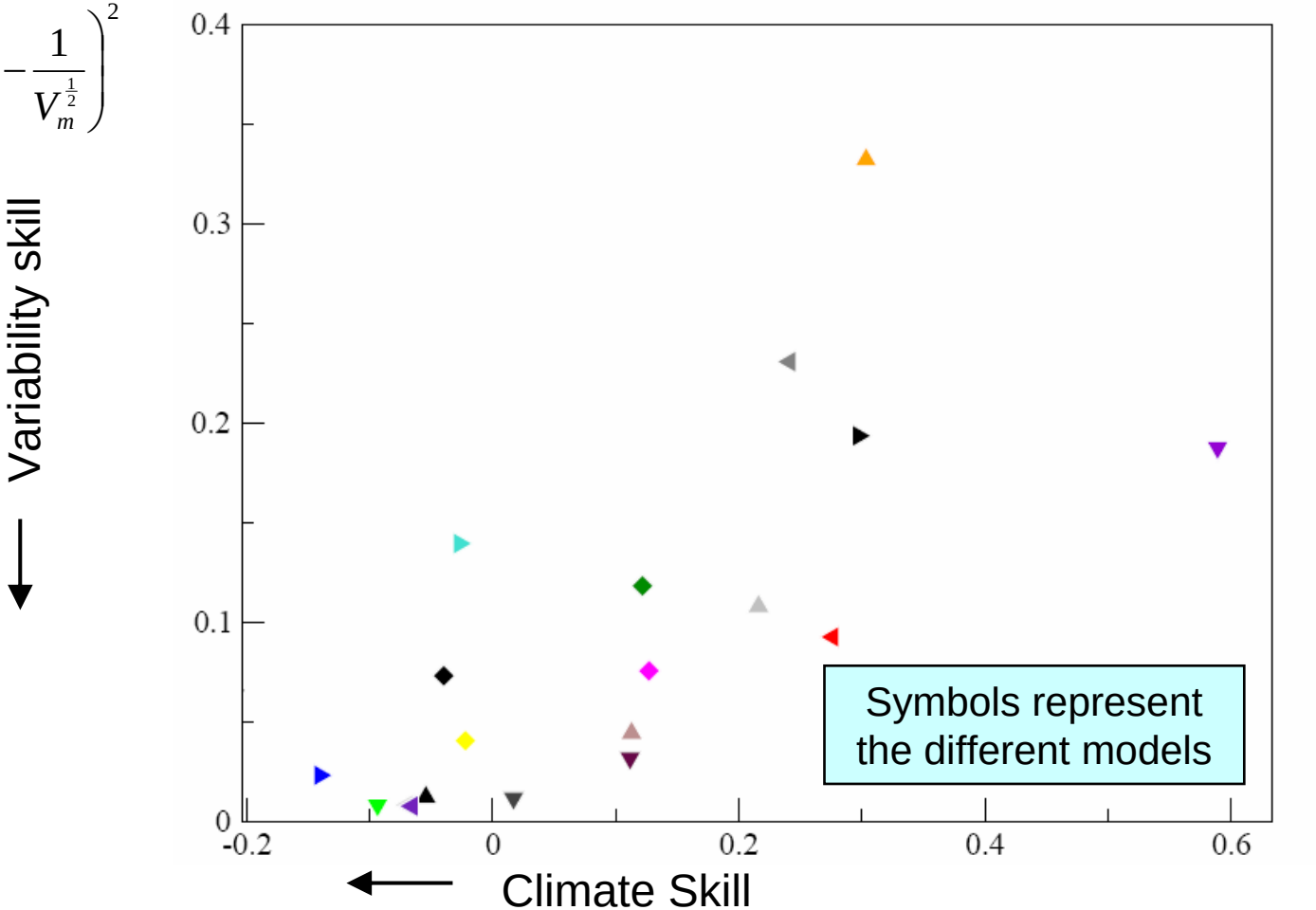
- Plot $V_{vm} = \dfrac{\sigma_{vm}^{\,2}}{\sigma_{v,\mathrm{obs}}^{\,2}}$



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

Reliance on a single index may be misleading.

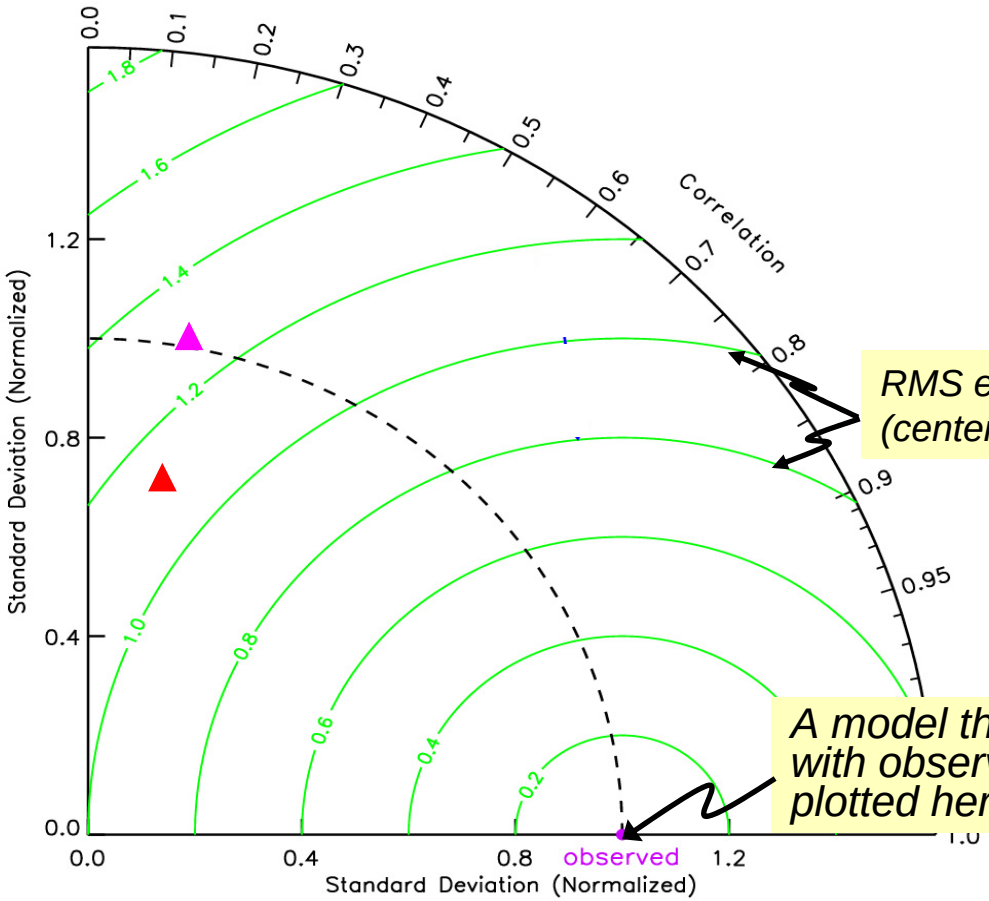
$$\left(V_m^{\frac{1}{2}} - \frac{1}{V_m^{\frac{1}{2}}} \right)^2$$



Limitations of metrics

- Can sometimes identify problems, but rarely lead us to solutions.
- Little work so far in characterizing whether the complex interactions among different fields are correctly represented.
- Metrics (skill scores) can sometimes be "played," hence may be misleading:
  - → Filtering is a specific example of this.
- Little justification for using any particular "performance index" to determine the relative reliability of models.

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

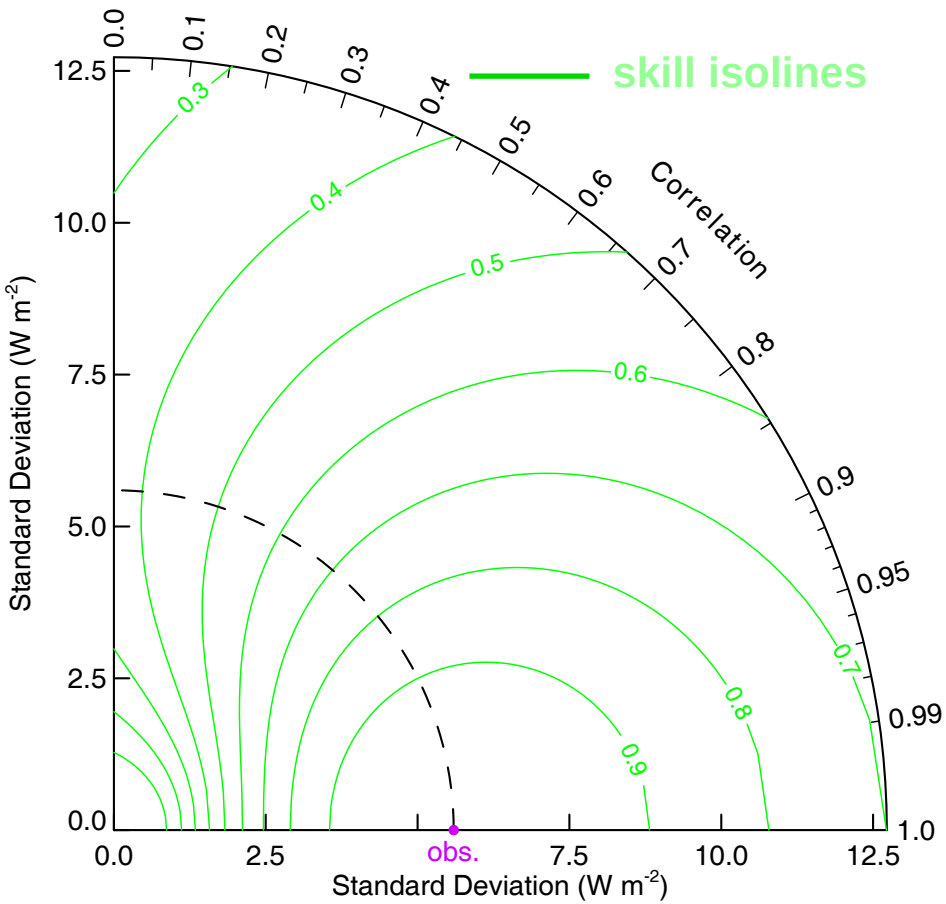


RMS error (centered and normalized)

A model that agrees perfectly with observations would be plotted here

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"

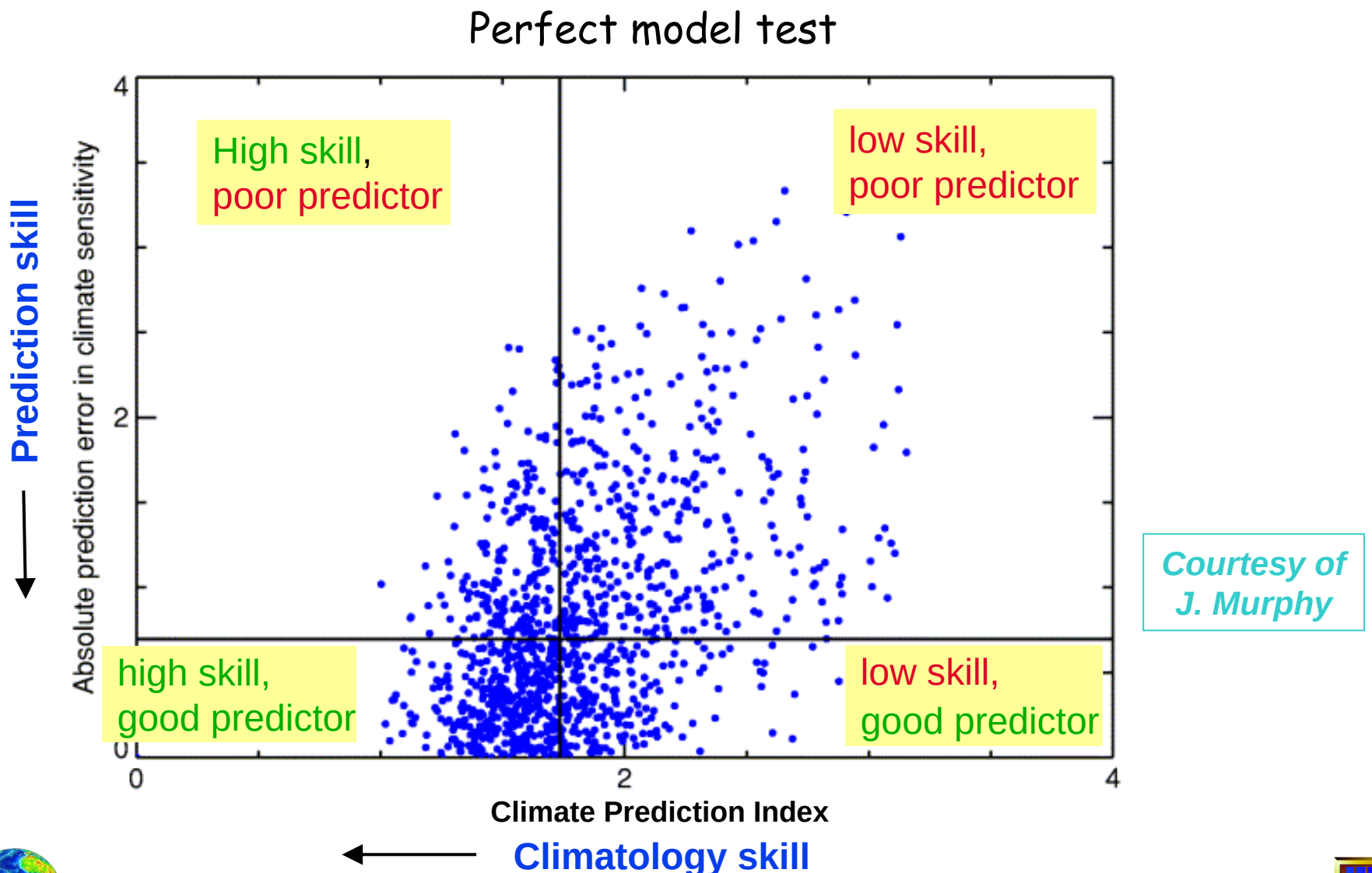
Limitations of metrics

- Can sometimes identify problems, but rarely lead us to solutions.
- Little work so far in characterizing whether the complex interactions among different fields are correctly represented.
- Metrics (skill scores) can be played:
  - → Danger in altering a model in non-physical ways to achieve a higher score.
  - → Filtering is a specific example of this.
- Little justification for using any particular "performance index" to determine the relative reliability of models for any specific purpose.

"Perfect model" experiments can provide evidence of the relevance of performance indices to predictive reliability

- A designated "reference" model's simulation is taken as a substitute for observations.
- Other models are evaluated in terms of their ability to simulate the reference model's
  - → Climatology
  - → Prediction of future climate change

Is the climate prediction index relevant to climate change prediction?



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).

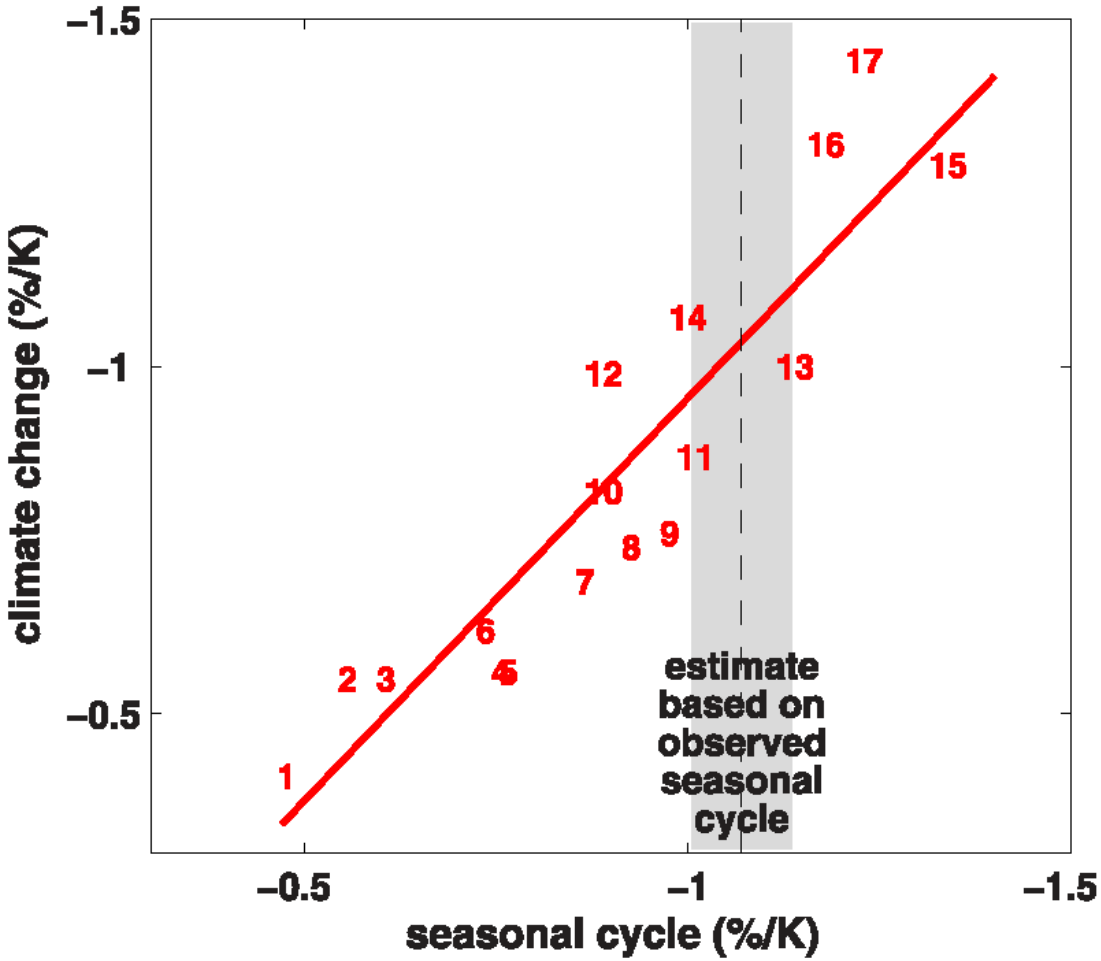
Fundamental research questions

- 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?

Advertisement

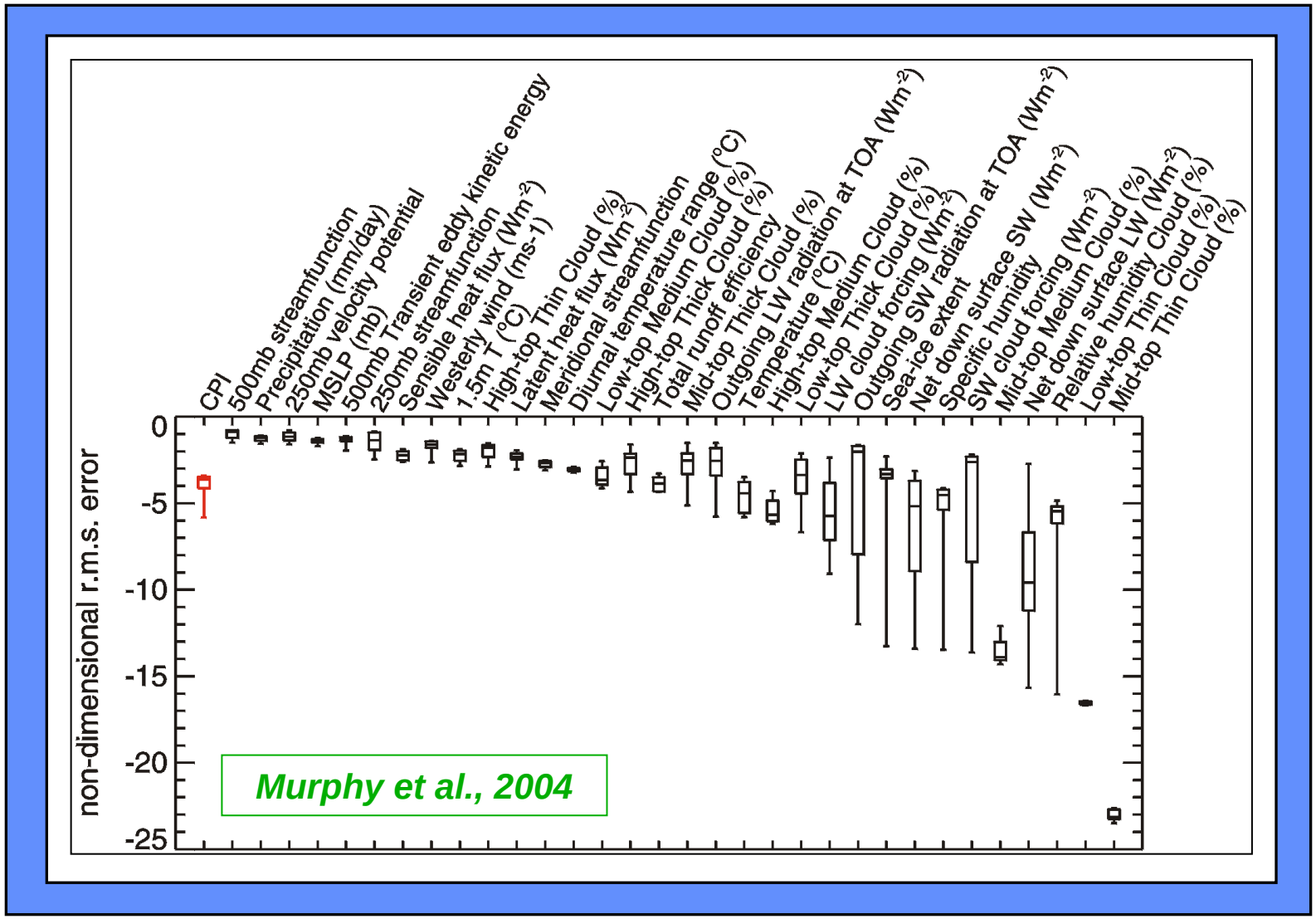
- Break-out group tomorrow afternoon on metrics
- Several posters of interest

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

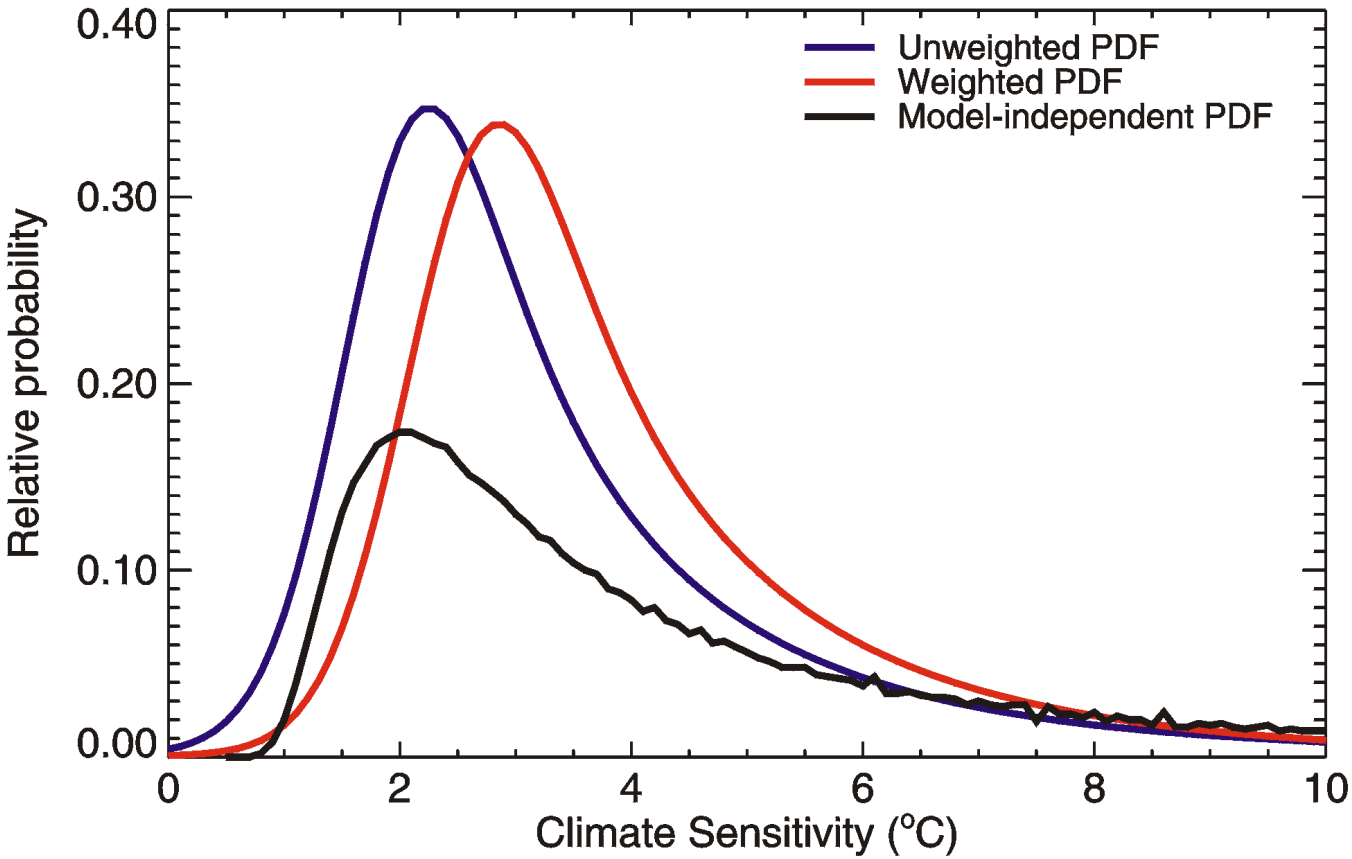


Hall & Xu, 2006

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