JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 113, D06104, doi:10.1029/2007JD008972, 2008

Performance metrics for climate models

P. J. Gleckler,[1] K. E. Taylor,[1] and C. Doutriaux[1]

Received 15 May 2007; revised 3 August 2007; accepted 21 November 2007; published 20 March 2008.

[1] Objective measures of climate model performance are proposed and used to assess simulations of the 20th century, which are available from the Coupled Model Intercomparison Project (CMIP3) archive. The primary focus of this analysis is on the climatology of atmospheric fields. For each variable considered, the models are ranked according to a measure of relative error. Based on an average of the relative errors over all fields considered, some models appear to perform substantially better than others. Forming a single index of model performance, however, can be misleading in that it hides a more complex picture of the relative merits of different models. This is demonstrated by examining individual variables and showing that the relative ranking of models varies considerably from one variable to the next. A remarkable exception to this finding is that the so-called ''mean model'' consistently outperforms all other models in nearly every respect. The usefulness, limitations and robustness of the metrics defined here are evaluated 1) by examining whether the information provided by each metric is correlated in any way with the others, and 2) by determining how sensitive the metrics are to such factors as observational uncertainty, spatial scale, and the domain considered (e.g., tropics versus extra-tropics). An index that gauges the fidelity of model variability on interannual time-scales is found to be only weakly correlated with an index of the mean climate performance. This illustrates the importance of evaluating a broad spectrum of climate processes and phenomena since accurate simulation of one aspect of climate does not guarantee accurate representation of other aspects. Once a broad suite of metrics has been developed to characterize model performance it may become possible to identify optimal subsets for various applications.

Citation: Gleckler, P. J., K. E. Taylor, and C. Doutriaux (2008), Performance metrics for climate models, *J. Geophys. Res.*, *113*, D06104, doi:10.1029/2007JD008972.

1. Introduction

[2] Climate models are routinely subjected to a variety of tests to assess their capabilities. A broad spectrum of diagnostic techniques are relied on in evaluations of this kind, but relatively little effort has been devoted to defining a standard set of measures designed specifically to provide an objective overview or summary of model performance. In this study we explore issues important to the development of climate model performance ''metrics,'' addressing how such measures can be defined and, by means of example, illustrating their potential uses and limitations. We evaluate a recent suite of coupled ocean-atmosphere general circulation models (OAGCMs), focusing primarily on global scales of the simulated mean annual cycle.

[3] Years ago, standard measures of forecast skill were adopted by the Numerical Weather Prediction (NWP) community. Relying on these metrics, the Working Group on Numerical Experimentation (WGNE) routinely reviews the skill of weather forecasts made by the major weather prediction centers [e.g., *WMO*, 1994]. (The WGNE reports to the World Meteorological Organization Commission for Atmospheric Sciences and World Climate Research Programme (WCRP) Joint Scientific Committee.) Monitoring NWP performance in this way has provided quantitative evidence of increases in forecast accuracy over time as well as characterizing the relative skill of individual forecast systems.

[4] Although the value of climate model metrics has been recognized for some time [e.g., *Williamson*, 1995], there are reasons why climate modelers have yet to follow the lead of the NWP community. First, a limited set of observables (e.g., surface pressure anomalies) have proven to be reliable proxies for assessing overall NWP forecast skill, whereas for climate models, examination of a small set of variables may not be sufficient. Because climate models are utilized for such a broad range of research purposes, it seems likely that a more comprehensive evaluation will be required to characterize a host of variables and phenomena on diurnal, intraseasonal, annual, and longer times scales. To date, a succinct set of measures that assess what is important to climate has yet to be identified.

[5] Another reason for the lack of accepted standard measures of climate model performance is that opportunities to test their ability to make predictions is limited. NWP

[1]Program for Climate Model Diagnosis and Intercomparison, Lawrence Livermore National Laboratory, Livermore, California, USA.

0148-0227/08/2007JD008972$09.00

systems are routinely verified against continuously varying weather conditions whereas climate models are repeatedly tested against an observed climatology that is only slowly evolving. Although models are increasingly being tested against field campaign data [e.g., *Phillips et al.*, 2004], and individual processes can sometimes best be evaluated in this way, our emphasis here is on the climatological features at large to global scales.

[6] There are additional factors to consider in developing climate model metrics. A wide variety of variables are of interest, and observational uncertainties are often substantial but poorly estimated. Some aspects of climate model simulations are deterministic, while others are not, making quantitative verification more complex. Lastly, models can to a certain degree be tunable to appear realistic in some respects, but as a result of compensating errors.

[7] In the face of these challenges, initial work in this area has nevertheless been carried out. The WGNE has encouraged development of standard diagnostics (see www-pcmdi.llnl.gov/projects/amip/OUTPUT/WGNEDIAGS) and has established benchmark experiments through model intercomparison projects. Many of the metrics described in this article have been routinely calculated by the Program for Climate Model Diagnosis and Intercomparison (PCMDI) and have been presented at public forums attended by climate modelers. Until now, publication of these results has been withheld because of the real danger that unwarranted importance might be attributed to these measures of model performance without appreciation of their limitations. One of the objectives of the present study is to identify those limitations, as well as to explore the potential uses of metrics.

[8] For the purposes of this study, the term "metric" will be restricted to scalar measures. Many popular statistical measures are consistent with this definition (e.g., mean error, root-mean square error, correlation, and variance). We also limit ourselves to metrics that directly compare model simulations with observationally based reference data, and distinguish them from diagnostics, which are more varied and include maps, time series, distributions, etc. Consistent with this nomenclature, metrics can be derived from diagnostics, generally resulting in a condensation of the original information. In this regard we expect metrics to provide symptoms of problems, but to be less informative than diagnostics for illuminating their causes.

[9] In section 2 we describe both the observationally based reference data and model simulations used in this study. In section 3 we introduce our choice of metrics and apply them to the Coupled Model Intercomparison Project phase 3 (CMIP3) simulations to evaluate the performance of models relative to each other. After demonstrating that our results can be sensitive to various factors (e.g., observational uncertainty), we provide an example of how an overall index of mean climate "skill" can be defined and discuss its limitations. Lastly, in section 4 we summarize our results and discuss how the use of metrics might be expanded and explored further.

2. Model and Reference Data Sets

2.1. The CMIP3 Simulations

[10] The models evaluated in this study come from nearly all the major climate modeling groups and are listed in Table 1. Recently as part of CMIP3, these groups performed an unprecedented suite of coordinated climate simulations. This ambitious undertaking was organized by the WCRP's Working Group on Coupled Modelling (WGCM). The WGCM designed this set of experiments to facilitate the science and model evaluation they felt needed to be addressed in the 4th Assessment Report (AR4) of the Intergovernmental Panel on Climate Change (IPCC). The CMIP3 multimodel data set has been archived by PCMDI and has been made available to the climate research community.

[11] Our analysis focuses on the 20th century simulations, which were "forced" by a variety of externally imposed changes such as increasing greenhouse gas and sulfate aerosol concentrations, changes in solar radiation, and forcing by explosive volcanism. Since considerable uncertainty as to the true forcing remains, the forcing is not the same for all models. Rather, these runs represent each group's best effort to simulate 20th century climate. The models were "spun up" under conditions representative of the pre-industrial climate (nominally 1850). From this point, external time-varying forcing, consistent with the historical period, was introduced, and the simulations were extended through at least the end of the 20th century.

[12] Although the CMIP3 archive includes daily means for a few fields and even some data sampled at 3-hourly intervals, we shall focus here solely on the monthly mean model output. This provides a reasonably comprehensive picture of model performance, but excludes evaluation of some aspects of climate, such as the frequency and intensity of extratropical cyclones and the frequency of some types of extreme events that would only be evident in the daily or 3-hourly data. Moreover, climatically important quadratic quantities such as heat and momentum fluxes are not available in the archive, and will not be considered here. The initial list of fields evaluated here should in future studies be expanded to include these other important aspects of climate.

[13] In this study we focus mostly on climatologies of the last 20 years of the 20th century simulations (1980–1999). During this period, the observational record is most reliable and complete, largely due to the expansion of and advances in space-based remote sensing.

2.2. Reference Data

[14] The climate characteristics of particular interest to us in this study range from large to global scales, so we restrict ourselves to using observationally-based data sets provided on global grids that are readily comparable to climate models. One limitation of most reference data sets is that it is in general difficult to estimate their observational errors. Sources of uncertainty include random and bias errors in the measurements themselves, sampling errors, and analysis error when the observational data are processed through models or otherwise altered. In short, the quality of observational measurements varies considerably from one variable to the next.

[15] Errors in the reference data are in fact usually all but ignored in the evaluation of climate models. It is often argued that this is acceptable as long as these errors remain much smaller than the errors in the models. If models improve to the point where their errors are comparable to

Table 1. Model Identification, Originating Group, and Atmospheric Resolution

IPCC I.D.	Center and Location	Atmosphere Resolution
BCCR-BCM2.0	Bjerknes Centre for Climate Research (Norway)	T63 L31
CGCM3.1(T47)	Canadian Centre for Climate Modelling and Analysis (Canada)	T47 L31
CGCM3.1(T63)		T63 L31
CSIRO-Mk3.0	CSIRO Atmospheric Research (Australia)	T63 L18
CNRM-CM3	Météo-France, Centre National de Recherches Météorologiques (France)	T42 L45
ECHO-G	Meteorological Institute of the University of Bonn, Meteorological Research Institute of KMA, and Model and Data group (Germany and Korea)	T30 L19
GFDL-CM2.0	US Dept. of Commerce, NOAA	N45 L24
GFDL-CM2.1	Geophysical Fluid Dynamics Laboratory (USA)	N45 L24
GISS-AOM	NASA/Goddard Institute for Space Studies (USA)	90 × 60 L12
GISS-EH		72 × 46 L17
GISS-ER		72 × 46 L17
FGOALS-g1.0	LASG/Institute of Atmospheric Physics (China)	128 × 60 L26
INM-CM3.0	Institute for Numerical Mathematics (Russia)	72 × 45 L21
IPSL-CM4	Institut Pierre Simon Laplace (France)	96 × 72 L19
MIROC3.2(medres)	Center for Climate System Research (The University of Tokyo), National Institute for Environmental Studies, and Frontier Research Center for Global Change (JAMSTEC) (Japan)	T42 L20
MIROC3.2(hires)		T106 L56
MRI-CGCM2.3.2	Meteorological Research Institute (Japan)	T42 L30
ECHAM5/MPI-OM	Max Planck Institute for Meteorology (Germany)	T63 L32
CCSM3	National Center for Atmospheric Research (USA)	T85 L26
PCM		T42 L18
UKMO-HadCM3	Hadley Centre for Climate Prediction and Research, Met Office (UK)	96 × 72 L19
UKMO-HadGEM1		N96 L38

observational uncertainty, a more rigorous approach will be required. In a similar vein, some of the satellite data described below are available for only a few years, whereas model climatologies compared with such data are usually computed from twenty or more simulated years. For the metrics considered here, simple sampling tests (performed on model simulations) indicate that the impacts of uncertainties associated with a limited observational record are small when compared to the magnitude of current model errors.

[16] A full quantitative assessment of observational errors is beyond the scope of this study. We shall, however, provide for most fields some indication of the effects of observational uncertainty by comparing model simulations to two different reference data sets. For each of the fields examined here, the reference data used are shown in Table 2. Here, we briefly summarize these reference data sets.

[17] For many fields examined here, the best available reference data come from two 40-year reanalysis efforts, namely the NCEP/NCAR [*Kalnay et al.*, 1996] and ERA40 [*Simmons and Gibson*, 2000]. These products blend many available measurements in a way that ensures their mutual internal consistency as constrained by the physical laws underlying the models. While there are important differences in the analysis systems and models used in these two

Table 2. Observationally-Based Reference Data Sets

Variable I.D.	Description	Reference1/Reference2	Domain
ta	Temperature, °C	ERA40/NCEP-NCAR	200,850 hPa
ua	Zonal wind, m/s	ERA40/NCEP-NCAR	200,850 hPa
va	Meridional wind, m/s	ERA40/NCEP-NCAR	200,850 hPa
zg	Geopotential height, m	ERA40/NCEP-NCAR	500 hPa
hus	Specific humidity, kg/kg	AIRS/ERA40	400, 850 hPa
psl	Sea level pressure, Pa	ERA40/NCEP-NCAR	Ocean-only
uas	Surface (10m) zonal wind speed, m/s	ERA40/NCEP-NCAR	Ocean-only
vas	Surface (10m) meridional wind speed, m/s	ERA40/NCEP-NCAR	Ocean-only
ts	Sea surface temperature, °C	ERSST/HadISST	Ocean-only; equatorward of 50°
tauu	Ocean surface zonal wind stress, Pa	ERA40/NCEP-NCAR	Ocean-only
tauv	Ocean surface meridional wind stress, Pa	ERA40/NCEP-NCAR	Ocean-only
hfls	Ocean surface latent heat flux, W/m^2	SOC/ERA40	Ocean-only
hfss	Ocean surface sensible heat flux, W/m^2	SOC/ERA40	Ocean-only
rlut	Outgoing longwave radiation, W/m^2	ERBE/CERES	
rsut	TOA reflected shortwave radiation, W/m^2	ERBE/CERES	
rlutcs	TOA longwave clear-sky radiation, W/m^2	ERBE/CERES	Equatorward of 60°
rsutcs	TOA shortwave clear-sky radiation, W/m^2	ERBE/CERES	Equatorward of 60°
rlwcrf	Longwave cloud radiative forcing, W/m^2	ERBE/CERES	Equatorward of 60°
rswcrf	Shortwave cloud radiative forcing, W/m^2	ERBE/CERES	Equatorward of 60°
pr	Total precipitation, mm/day	GPCP/CMAP	
clt	Total cloud cover, %	ISCCP-D2/ISCCP-C2	
prw	Precipitable water, g/kg	RSS/NVAP	

products, for the most part they rely on the same observations. Although they are therefore not truly independent, these products do represent the best observationally constrained estimates of the free-atmosphere variables evaluated here. We shall analyze the most commonly studied levels (200 hPa and 850 hpa for ta, ua, and va, 500 hPa for zg, and 850 hPa and 400 hPa for hus — see Table 2 for variable names). Specific humidity (hus) is of a fundamental importantance to climate, but the accuracy of analyzed moisture is much lower than that of other state fields. We use estimates of hus from the Atmospheric Infrared Sounder (AIRS) experiment, which includes a hyperspectral infrared spectrometer and an Advanced Microwave Sounding Unit radiometer, both carried on the Aqua spacecraft in a sun-synchronous orbit [*Aumann et al.*, 2003]. We evaluate hus using AIRS as our primary reference, and ERA40 as our alternate.

[18] Surface air fields (winds at 10 m, and temperature and humidity at 2 m) are also obtained from reanalysis. While these estimates are of reasonable quality, it is important to keep in mind that they are diagnostic quantities in both the reanalyses and the climate models. They are typically computed from complex iterative schemes based on conditions in the lowest model level and at the surface, and are not actually needed in any of the equations that affect the simulation. Thus errors in these fields could reflect, at least in part, poor diagnostics, not fundamental model errors. Similar caveats should be noted in comparisons of 'mean sea level pressure', which is also diagnosed through extrapolation both in reanalyses and in climate models.

[19] We compare model-simulated top-of-the-atmosphere radiation fields to data derived from the Earth Radiation Budget Experiment [ERBE, *Barkstrom*, 1984] and the more recent Clouds and the Earth's Radiant Energy System measurements [CERES, *Wielicki et al.*, 1996]. Estimates of "clear-sky" fluxes at the top-of-the-atmosphere are less accurate than the all-sky fluxes [e.g., *Cess et al.*, 1990], but, nonetheless, useful. Our analysis excludes high latitude clear-sky fluxes which are much less reliable. The ERBE period is from 1985 to 1989, whereas the CERES data are available from 2001 to 2004.

[20] The precipitation data sets we use are from GPCP [*Adler et al.*, 2003] and CMAP [*Huffman et al.*, 1997]. There have been numerous comparisons of these products [e.g., *Yin et al.*, 2004], and while there are important differences between them (particularly over oceans), they have been derived from many of the same data sources. Both are available from 1979 to near present.

[21] For total cloud cover, our primary data set is from the International Satellite Cloud Climatology Project [ISCCP, e.g., *Rossow and Schiffer*, 1999], available for the period 1983 through 2005. As an alternative to the recent ISCCP "D2" data, which is taken as the primary reference, we also use the older "C2" ISCCP data. These two cloud data sets have been compared by *Rossow et al.* [1993].

[22] The two references we have chosen for the sea surface temperature are the NOAA Extended Reconstructed sea surface temperature (SST) data set [ERSST, *Smith and Reynolds*, 2004] and the Hadley Centre Sea Ice and SST data set [HadISST, *Rayner et al.*, 2006]. While these analyses do make use of much of the same data, there are important differences in their approaches. To avoid the influence of sea-ice (both observed and simulated), we exclude regions poleward of 50° from our analysis of SST.

[23] After mid-1997 estimates of precipitable water (or column integrated water content) over the oceans are available from the special sensor microwave imager (SSM/I). We use two products: RSS version 6 [see *Wentz*, 2000] and the NASA water vapor project [NVAP, *Simpson et al.*, 2001]. Although both products derive precipitable water from SSM/I, the independently developed RSS and NVAP algorithms yield substantial differences [e.g., *Trenberth et al.*, 2005].

[24] Several satellite-derived estimates of wind stress over the ocean are available, however directional measurements remain problematic. We use the a European Remote Sensing (ERS) blended product derived from several satellites [*Bourassa et al.*, 1997] as our primary reference and the ERA40 wind stress as an alternate.

[25] Estimates of ocean surface heat fluxes are even more uncertain than wind stress. Significant efforts have been devoted to the improvement of in situ estimates, but uncertainties remain large in all terms, especially in the Southern Hemisphere. For latent and sensible heat fluxes, we contrast the Southampton Oceangraphic Center (SOC) climatology [*Josey et al.*, 1999] with ERA40. It is well known that there are large differences in these [e.g., *Taylor*, 2000], with their patterns in better agreement than the magnitude of their fluxes.

[26] In the following sections these data sets will simply be referred to as "observations", keeping in mind that "observationally-based" or "reference" data is generally more appropriate.

3. Results

3.1. Monthly Mean Statistics

[27] In this section we examine how well the CMIP3 20th century simulations compare with observations during the last two decades of the 20th century (1980–1999), our most complete and accurate 20-year observational record. To portray how the models perform relative to each other, we shall use simple statistical measures to quantify the fidelity of their simulations. Our results will be shown for the following domains: global, tropical (20S–20N), and the extra-tropical zones of the Southern (90S–20S) and Northern (20N–90N) Hemispheres. These regions will respectively be referred to as Global, Tropics, SHEX and NHEX.

[28] One statistical measure of model fidelity is the root-mean square difference (E) between a simulated field F and a corresponding reference data set R. For monthly mean climatological data, the most comprehensive root-mean square (RMS) error statistic accounts for errors in both the spatial pattern and the annual cycle, and it is calculated as follows:

$$E^2 = \frac{1}{W} \sum_i \sum_j \sum_t w_{ijt} \left(F_{ijt} - R_{ijt}\right)^2. \quad (1)$$

[29] The indices i, j, and t correspond to the longitude, latitude and time dimensions, and W is the sum of the weights (w_{ijt}), which for the spatial dimensions are propor-

tional to grid-cell area and for time are proportional to the length of each month. The weights are therefore approximately proportional to the cosine of latitude except when there are missing data in which case $w_{ijk} = 0$ (e.g., over land, when only the ocean portion of the domain is considered). In most cases considered here the sums will be accumulated over all 12 months and over one of the domains of interest (Global, Tropics, SHEX, or NHEX). For these calculations we interpolate each data set (model and reference data) to a T42 grid, which is a resolution comparable to that used in many of the models (Table 1). The sensitivity of our results to the choice of target grid is examined in section 3.3.

[30] We begin by making use of the Taylor diagram [*Taylor*, 2001], which relates the ''centered'' RMS error (calculated as in equation (1), but with the overall time-mean, spatial-mean removed), the pattern correlation and the standard deviation. A reference data set is plotted along the abscissa. Simulated fields are located in the first quadrant if the correlation with the reference data is positive. For both the reference and any model data represented on the plot, the radial distance from the origin is proportional to the standard deviation. The pattern correlation between the simulated field and the reference data is related to the azimuthal angle, and the centered RMS difference between a simulated field and the reference data is proportional to the distance between these two points (i.e., the closer a model is to the observational point, the lower its centered RMS error). The overall model ''bias,'' defined as the difference between the simulated and observed time-mean, spatial-mean fields, is not shown on this diagram.

[31] Three climatological annual-cycle space-time Taylor diagrams are shown in Figure 1: (a) NHEX, (b) TROP and (c) NHEX deviations from the zonal mean. (Here, as in some subsequent figures, space constraints preclude inclusion of statistics for the Southern Hemisphere performance statistics. The observations are more plentiful in the Northern Hemisphere, which is why results are emphasized for this region.) Each field is normalized by the corresponding standard deviation of the reference data, which allows multiple fields (distinguished by color) to be shown in each panel of Figure 1. In this figure, each colored dot represents an individual simulation made with a particular model, whereas each triangle represents the ensemble ''mean model.'' The ''mean model'' statistics are calculated after regridding each model's output to a common (T42) grid, and then computing the multimodel mean value at each grid cell. Note that because the statistics considered here are based on sums of quadratic quantities, the error in a multimodel mean field is not the same as the mean of the error statistics from the individual models. In fact, for each variable considered in Figures 1a–1c, the mean model simulated field matches, in an RMS sense, the reference data more closely than most or all of the individual models. We shall return to this notable feature later.

[32] It is also clear from Figures 1a–1c that the accuracy of a model simulation depends on the field and the domain as well as the model. In Figure 1a (NHEX), some simulated fields have correlations with the reference data of greater than 90% (e.g., ta-850, ua-850, rlut, psl), whereas other fields, have much lower correlations (e.g., clt and pr). In general there is a much larger inter-model spread for fields with poorer correlations. In addition to a range of skill across variables and models (as reflected by the spread of centered RMS or pattern correlation), there are large differences between different domains and between components considered. Some fields with relatively high correlations in the NHEX have a lower skill in the tropics. This might be partly due to the fact that in higher latitudes more of the total variance is forced by the insolation pattern which creates land-sea contrasts and meridional gradients that are relatively easily simulated by models. In Figure 1c, the errors are computed on fields from which the zonal mean has been removed so that the often dominating influence on the statistics of merdional gradients are not considered and the smaller scale features evident in deviations from the zonal mean are emphasized. This aspect of climate involving smaller scale variations that are often strongly affected by internal dynamics are more difficult to simulate accurately, so some of the correlations are noticeable lower. Note that the errors shown in these ''normalized'' diagrams are relative to the magnitude of the variations in the observed field, as quantified by the standard deviation. Thus for a given apparent error shown in the diagram, the actual error will be larger for variables or domains with a larger degree of variation. Using Taylor diagrams, *Pincus et al.* [2008] discuss additional aspects of model performance focusing on clouds, radiation and precipitation in the CMIP3 simulations.

[33] For most fields considered here, the amplitude of spatial and/or temporal variability simulated by models is reasonably close to that observed, as is evident in Figure 1. Thus the centered RMS error and the correlation coefficient are not both needed to summarize model fidelity, and they both omit any overall ''bias'' error. In preference to these metrics, we shall hereafter rely on the full RMS error (i.e., ''uncentered'' RMS error) because it includes the overall bias. Furthermore, unlike the correlation it can be resolved into components which when summed quadratically yield the full mean square error.

[34] As an example that is particularly pertinent to later discussion, Figure 2 shows (for several different variables) how errors in the simulated mean annual cycle for the global domain can be resolved into five orthogonal components: 1) the annual mean, global mean (i.e., the bias), 2) the annual mean zonal mean, with bias removed, 3) the annual mean, deviations from zonal mean, 4) the annual cycle of zonal-mean, with annual mean removed, and 5) the annual cycle of deviations from zonal mean, with annual mean removed. These are expressed as fractions of the total mean square error and are partially accumulated before plotting, so that when the last component is added to the earlier ones, the full error is accounted for. The errors associated with individual components is proportional to the lengths of the shaded segments.

[35] The fractions appearing in Figure 2 are computed by first averaging over all models each of the mean square error components, then accumulating these, and finally dividing by the average over all models of the total mean square error. None of the components dominates, and the error fraction associated with each component depends on the variable. In most cases none of the components is especially small, except in the case of upper air (200 hPa) meridional wind, in which case the error associated with the zonal mean component is negligible. Note that the 500 hPa

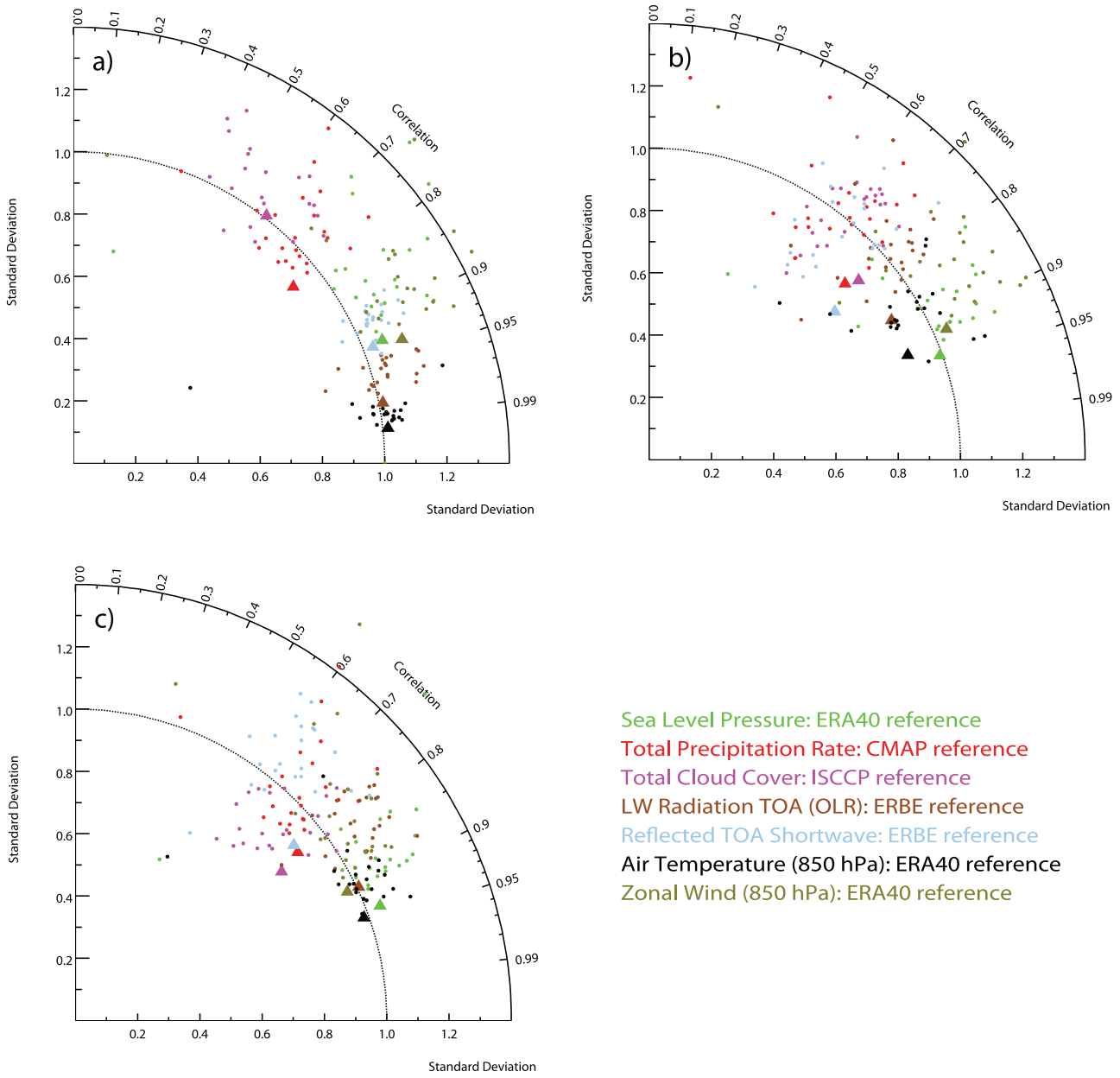

Figure 1. Multivariable Taylor diagrams of the 20th century CMIP3 annual cycle climatology (1980–1999) for (a) NHEX (20N–90N), (b) Tropical (20S–20N), and (c) NHEX as in Figure 1a but with the zonal mean removed. Each colored dot represents an individual simulation made with a particular model, whereas each triangle represents the ensemble "mean model," as defined in the text.

geopotential height bias contributes a noticeably higher fraction of the total error than other fields displayed in Figure 2. It is also clear from Figure 2 that for most of the variables considered, more than a quarter (and in the case of precipitation, more than half) of the mean square error is associated with the annual cycle (with the annual mean removed). Thus it would likely be misleading to judge models solely on their ability to simulate annual mean climate unless the errors in the annual mean were strongly correlated (across models) with errors in the annual cycle.

3.2. Relative Model Performance

[36] Global scale observationally based estimates exist for many more quantities than those shown in Figures 1 and 2. In what follows we make use of all the observationally based data sets given in Table 2, calculating RMS errors for each model using both our primary (R1) and alternate (R2) references. Using both references for each variable, we therefore have twice as many RMS error values as available models (N). For a given field f and reference r, we define a 'typical' model error, $\bar{E}_{fr}$, as the median of our N RMS error calculations. We use the median rather than the mean value to guard against models with unusually large errors (outliers) unduly influencing the results. We then define a

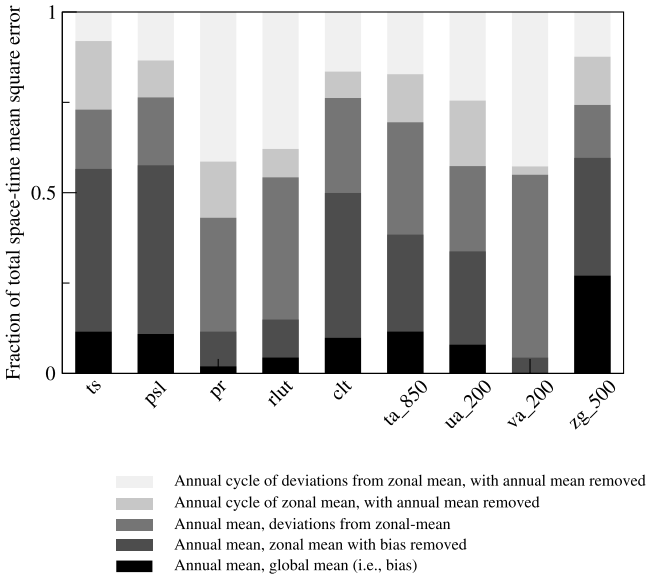

Figure 2. Decomposition of mean square error for the seasonally varying global pattern of simulated fields. The fraction of the total error due to each of the components listed in the legend is represented by the length of the corresponding bar segment.

relative error (E') for a given model m, field f, and reference set r, as:

$$E'_{mfr} = \frac{E_{mfr} - \bar{E}_{fr}}{\bar{E}_{fr}} \quad (2)$$

[37] Normalizing the RMS calculations in this way yields a measure of how well a given model (with respect to a particular reference data set) compares with the typical model error. For example, if the relative error has a value of −0.2, then the model's RMS error (E_{mfr}) is 20% smaller than the typical model. Conversely, if $E'_{mfr} = 0.2$, then the E_{mfr} is 20% greater than $\bar{E}_{fr}$. (The reader should keep in mind the distinction between the typical or median error within the distribution of model errors and the error in the multimodel ensemble mean or median field.)

[38] In Figures 3a–3f we provide a summary of the models' relative errors using "portrait" diagrams in which different colors indicate the size of errors. The portraits are arranged such that the rows are labeled by the variable name (see Table 2) and the columns by the name of the model (see Table 1). Each grid square is split by a diagonal in order to show the relative error with respect to both the primary (upper left triangle) and the alternate (lower right triangle) reference data sets. In each panel the two columns on the far left represent results for the multimodel ensemble mean and median fields. Variables (rows) are loosely organized as surface fields, clouds and radiation, and upper-air fields. Shades of blue indicate cases where a model fares better than the typical model with respect to the reference data, and shades of red the contrary. White squares indicate the unavailability of model data. For the surface fluxes, our results are based only on the centered RMS error (i.e., the overall mean bias is excluded) because the 'observed' large scale patterns of these fields are thought to be more accurate than their absolute magnitudes.

[39] Relative error calculations for the global domain are shown in Figure 3 for (a) the annual cycle of the full spatial pattern, (b) the annual cycle of the zonal means, and (c) the annual cycle of the deviations from the zonal means. The relative errors shown in Figure 3a for the global domain are shown for subglobal domains in Figure 3d (NHEX, 20N–90N), 3e (Tropics, 20S–20N), and 3f (SHEX, 90S–20N).

[40] An obvious feature in all panels of Figure 3 is that for virtually all variables, both the mean and median model fields are in better agreement with observations than the typical model. In most cases the mean and median models score best. This has been noted in other model comparisons [e.g., *Lambert and Boer*, 2001; *Taylor and Gleckler*, 2002], but stands out rather strikingly across the broader spectrum of fields and regions examined here. This favorable characteristic of the multimodel ensemble may in part be due to smoothing of higher frequency and smaller scale features, but this is almost certainly not the full explanation.

[41] One exception to the usual multimodel superiority apparent in Figure 3a is temperature at 200 hPa, where several models (notably GFDL-CM2.1) have noticeably smaller RMS errors than both the mean and the median models. Most atmospheric models continue to have cold biases at high latitudes in the summertime upper troposphere and throughout the stratosphere. Because this error is prevalent (though not universal), it is reflected in both the mean and median models. In the newer generation GFDL model this persistent and common error has been reduced.

[42] Errors in the annual cycle climatology (shown in Figure 3a) can be resolved into two components: 1) the annual cycle of the zonal means (Figure 3b), and 2) the annual cycle of deviations from the zonal mean (Figure 3c). The relative error differences among models is larger for the zonal means than for the deviations from the zonal mean, but the two error portraits are similar in the sense that models that accurately simulate the zonal mean (in a relative sense) also accurately simulate the deviations from the zonal mean. These figures also show that the relative errors in the mean and median fields are again smaller than the errors in the individual model fields, and this is perhaps even more evident in the deviations from the zonal mean. Note that the cold bias in the upper level temperature field affects the zonal mean component in the multimodel mean and median fields, but not the deviations from the zonal mean.

[43] Some models clearly fare better than others in Figure 3, although none scores above average or below average in all respects. For example, in the extra-tropics the UKMO-HadGEM1 relative errors are negative for most fields (indicating better than average performance) in the extratropics, while this model does not appear to excel in the tropics. The ECHO-G model, on the other hand, scores higher in the tropics, but does not stand out elsewhere. With respect to the upper air fields, the relative performance of the GFDL-CM2.1 is exemplary in the Southern Hemisphere (Figure 3f), whereas elsewhere it is closer to average (Figures 3d and 3e). The skill of many of the other models appears to depend more on the variable than the region.

[44] The sensitivity of the relative error (E') to the choice of reference data set can be inferred from the differences in color within any given square of Figure 3. Recall that for

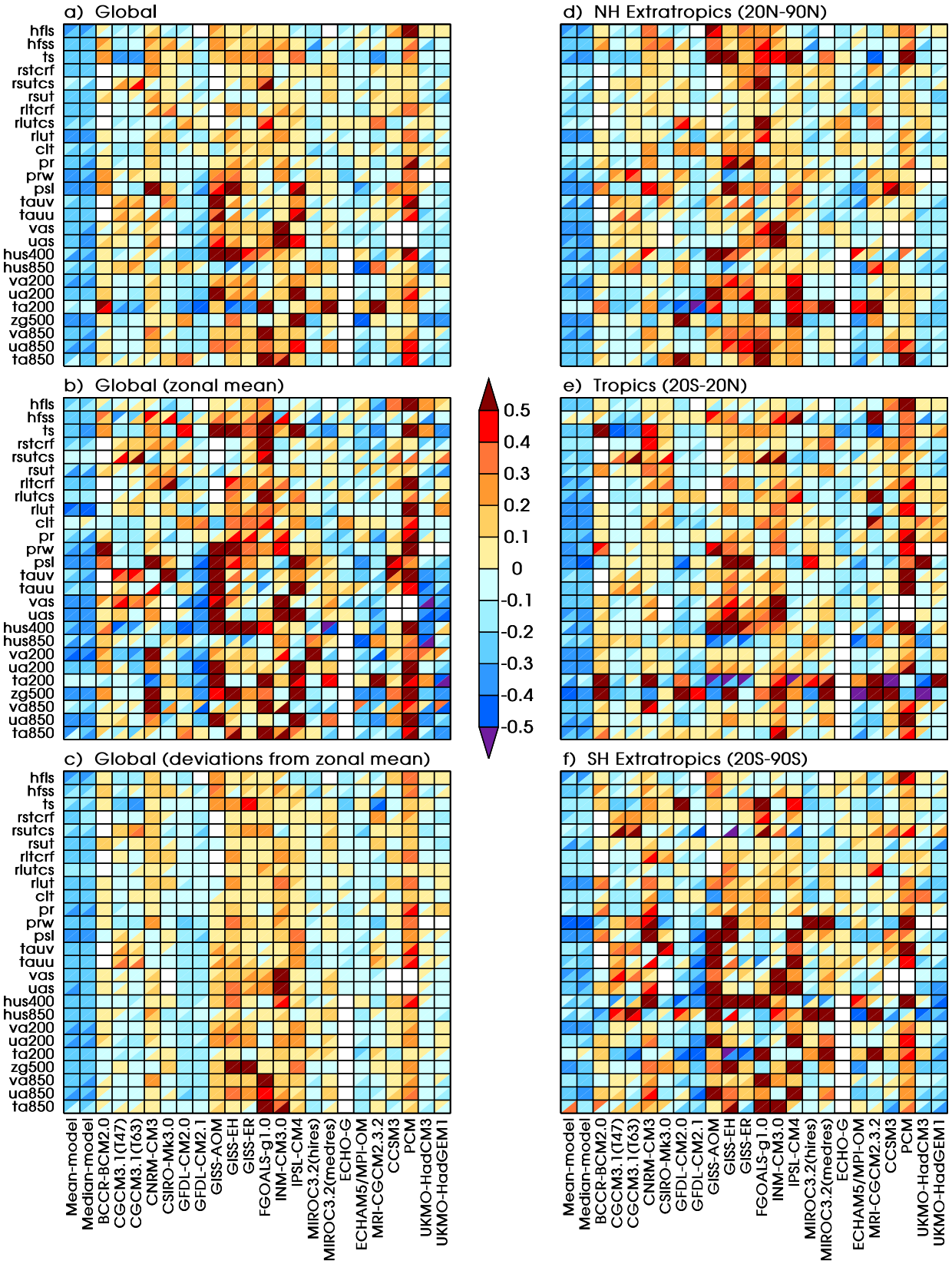

Figure 3. Portrait diagram display of relative error metrics for 20th century CMIP3 annual cycle climatology (1980–1999): (a) full field, (b) zonal mean (with bias removed), and (c) deviations from zonal mean, all based on the full global domain, and for the full field only, based on subglobal domains of (d) NHEX, (e) Tropics, and (f) SHEX. Each grid square is split by a diagonal in order to show the relative error with respect to both the primary (upper left triangle) and the alternate (lower right triangle) reference data sets.

some fields there is only a single reference data set (e.g., top-of-atmosphere clear-sky longwave radiation), and in other cases (e.g., cloud cover) there are strong interdependencies between the reference data, so for these we expect at most minor differences. For many of the other fields, however, there are notable differences. For example, in the NHEX and SHEX it is not unusual for a model's relative error in simulating precipitation (pr) to differ by 10%, depending on which reference is used. A more surprising example is the case of temperature at 850 hPa, a relatively well ''observed'' field. Here in both the Tropics and SHEX we often see differences of 20% or more between the two reference data sets (ERA40 and NCEP/NCAR). Note that it is not uncommon for one model to have two different triangle colors while for another model (and the same field) the colors are identical. (Our 10% discrete color interval is only a partial explanation.)

[45] There are mountainous regions where the surface elevation is high enough that the surface pressure is less than 850 hPa. Here modelers either flagged data as ''missing'' or simply extrapolated from a pressure level above the surface to the ''below ground'' standard pressure level (850 hPa). The choice presumably has at least some influence on the error differences among models. Above 500 hPa this problem must certainly be negligible.

[46] Surface air temperature and humidity (2 m above the surface) and winds (10 m above the surface) are not prognostic variables in the models, and they may be sensitive to the method used to diagnose them. These fields are analyzed over the oceans only.

3.3. The Sensitivity of Metrics to Analysis Choices

[47] One of our objectives is to develop metrics that summarize model performance in a way that is both concise, but also reasonably complete. To a certain extent this has been achieved in Figure 3a where the large-scale characteristics of the simulated annual cycle have been distilled down to several hundred statistical quantities. Even so, one wonders whether the portrait contains unnecessary detail and whether in fact some of the information is redundant. One way to determine whether the individual metrics are each providing independent information is to plot the performance metrics for one variable against the performance metrics of another. Two extreme cases are shown in the scatterplots of Figure 4, which shows the relationships between errors in simulating two different pairs of variables. In Figure 4a the relative errors (taken from Figure 3a) are compared for global precipitation and outgoing longwave radiation. There is clearly a strong relationship (R = 0.92) between how well models simulate precipitation and outgoing longwave radiation. A contrasting example is shown in Figure 4b, which indicates that there is almost no relationship between how well a model simulates geopotential height (500 hPa) and mean sea level pressure.

[48] A more comprehensive summary of the relationship between errors in simulating various pairs of variables is provided by Figure 5a. Here we examine how the ranking of models (based on the RMS error) depends on the variable considered. Figure 5a shows for each pair of variables how much the ranking of a model typically changes in an absolute sense depending on which variable the ranking is based. The ''typical'' change is the average across all models of the absolute change in ranking. For the pair of variables shown in Figure 4a (precipitation and outgoing longwave radiation), the average difference in ranking is 3 (out of 24 models), whereas for geopotential height (500 hPa) and mean sea level pressure it is 6. In more than 90% of the cases, the jump in model position is at least 4, depending on which of the two different variables is used to determine the ranking.

[49] Figures 4a and 5a indicate that there is some redundancy in the information depicted in the global relative error metrics in Figure 3a. This suggests that one might be able to portray mean climate model performance with fewer fields than we have chosen. On the other hand, Figures 4b and 5a also indicate that accurate simulation of one variable does not in most cases imply equally accurate simulation of another. For example, errors in 200 hPa temperature and 850 hPa specific humidity are not well correlated with errors in any of the other fields, so these metrics are clearly not redundant with any of the others. Thus consideration of multiple fields provides a more complete characterization of model performance.

[50] Working toward a reduced set of metrics will require careful consideration of several factors. Even when a strong relationship is found between relative skill in simulating two different fields, it might be difficult to decide whether one field should be eliminated or whether, perhaps, the errors of the two should somehow be combined. In the case of outgoing longwave radiation and precipitation (where the redundancy is greater than for most variable pairs), both are of interest to model developers, and it would be hard to argue that one should be removed in favor of the other.

[51] Another question of interest is whether it is really necessary to consider the full seasonal cycle. Would the relative merits of models be evident simply by considering the annual mean spatial distribution of each variable? In Figure 5b we show how the average absolute change in model ranking depends on whether it is based on the global pattern of deviations from the annual mean or the annual mean pattern itself. The results are mixed. In many cases the differences are small, but there are numerous instances where they are very large. In other words, relative model ranking in some cases strongly depends on whether the annual means or deviations from the annual means are considered. Thus it is not always sufficient to merely consider the annual mean climate.

[52] We next explore several additional factors that might affect the metrics used here to characterize model performance: 1) simulation initial conditions, 2) spatial scales considered, and 3) the period of the simulation evaluated.

[53] Multiple realizations, differing only in their initial conditions, are available for some of the models considered here and are critical for many studies, particularly those in which internal variability of the climate system is of interest. Until now we have focused on a single realization from each model. (We have used the first realization in the CMIP3 database at PCMDI.) In what follows we shall make use of all available realizations in the CMIP3 20th century experiment. We also computed 20-year climatologies for two different periods: 1900–1919 and 1980–1999 (on which the results in Figures 1–5 where based). Last, for the first realization of each model and our primary reference

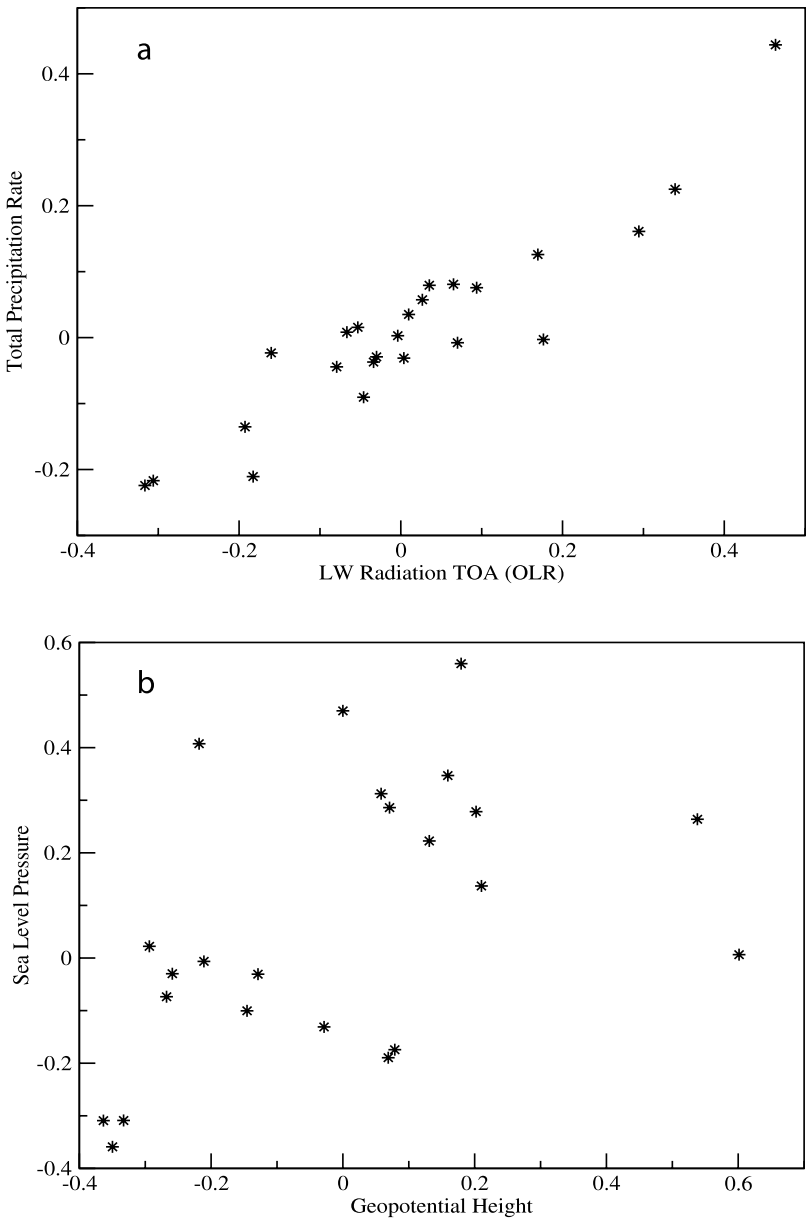

Figure 4. Relative errors for each model (taken from Figure 3a): (a) precipitation versus outgoing longwave radiation, and (b) 500 hPa geopotential height versus mean sea level pressure. Each symbol represents one of the models.

data set, we compute all of our statistics on two additional grids: a coarser grid (45 × 72) of 4° latitude by 5° longitude and a finer, gaussian grid (96 × 192) normally used in spherical harmonic decompositions truncated at T63.

[54] In Figure 6 we show results for precipitation and mean sea level pressure for the NHEX. In each case we have ordered the models, based on the relative errors given in Figure 3d, and denoted this error by an asterisk. We also show errors based on the alternate reference data set (o), other available realizations (−), the 1900–1919 climatology (a) and the lower and higher resolution "target" grids (4 × 5 and T63, +). The results are mixed. In some cases these variations on our analysis choices lead to small differences in a model's relative ranking, whereas in others the differences can be quite large. Rarely, however, would the model rank position change by more than 5 or 6.

[55] The choice of reference data set for precipitation can have a moderate effect on a model's relative ranking in the NHEX, although for several models the difference is small. In the tropics (not shown) the ranking is more systematically sensitive to the choice of precipitation data set. This is perhaps not surprising given the large uncertainties in precipitation estimates, especially over the ocean. The effect of the model initial conditions and the climatology averaging period is not very large. One exception is the GFDL CM2.1 which later (in Figure 8) is shown to have excessive

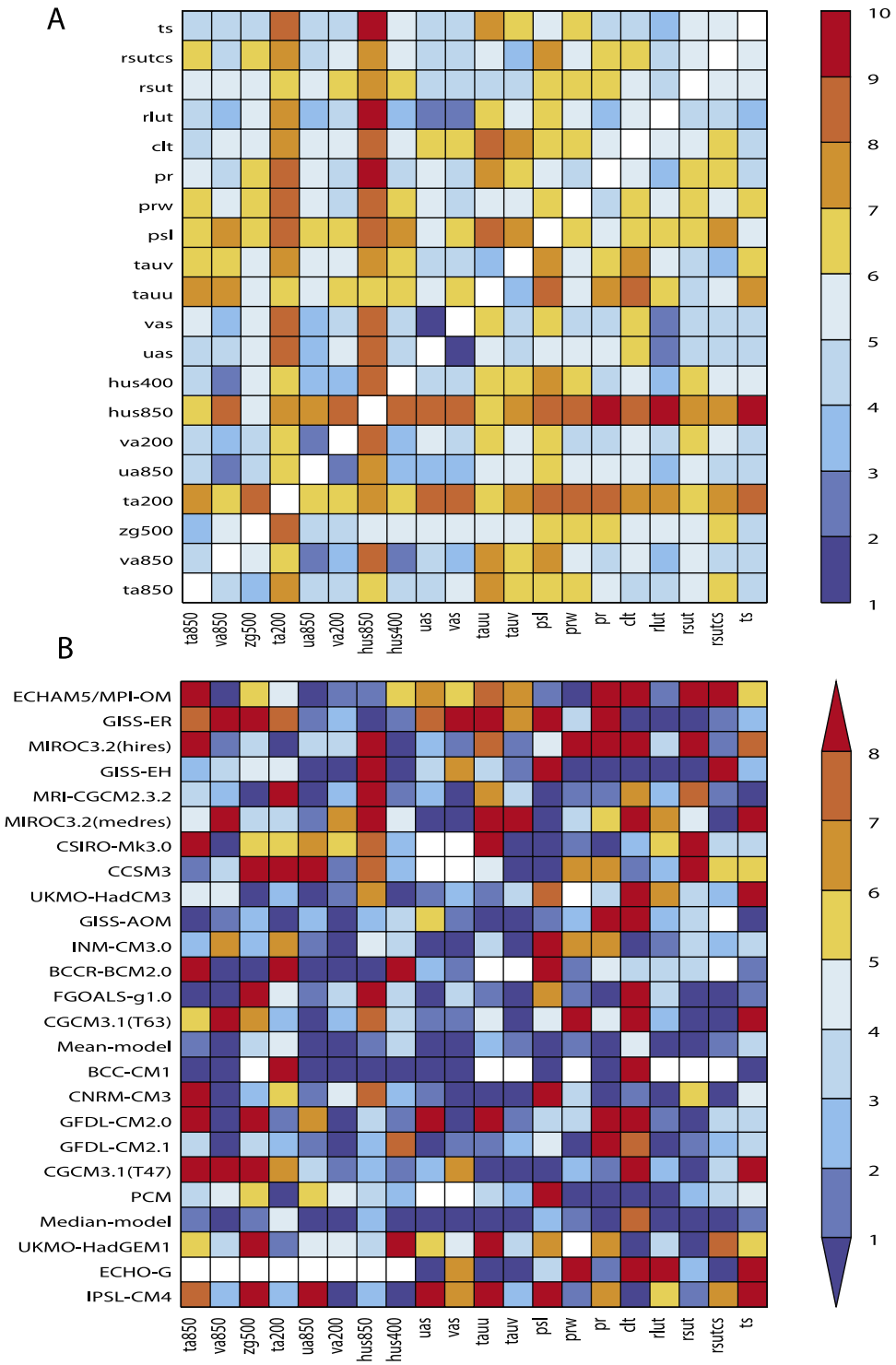

Figure 5. Average absolute change in model ranking: (a) variable-by-variable (taken from Figure 3a), and (b) the annual cycle versus annual mean.

variability, making climatologies more sensitive to the 20-year averaging period. Finally, considering the target grid resolution, we find, generally speaking, the coarser the grid, the better the model agreement with observations of precipitation. The ranking sensitivity of mean sea level pressure is similar in most respects to precipitation, but is clearly less sensitive to the target grid resolution.

3.4. An Overall Model Performance Index?

[56] There would be considerable value in deriving a single index of model reliability. If this were possible and could be justified, the index could be used, for example, to weight individual model results to form more accurate multimodel projections of climate change [*Murphy et al.*, 2004; *Stainforth et al.*, 2005]. Projections by individual

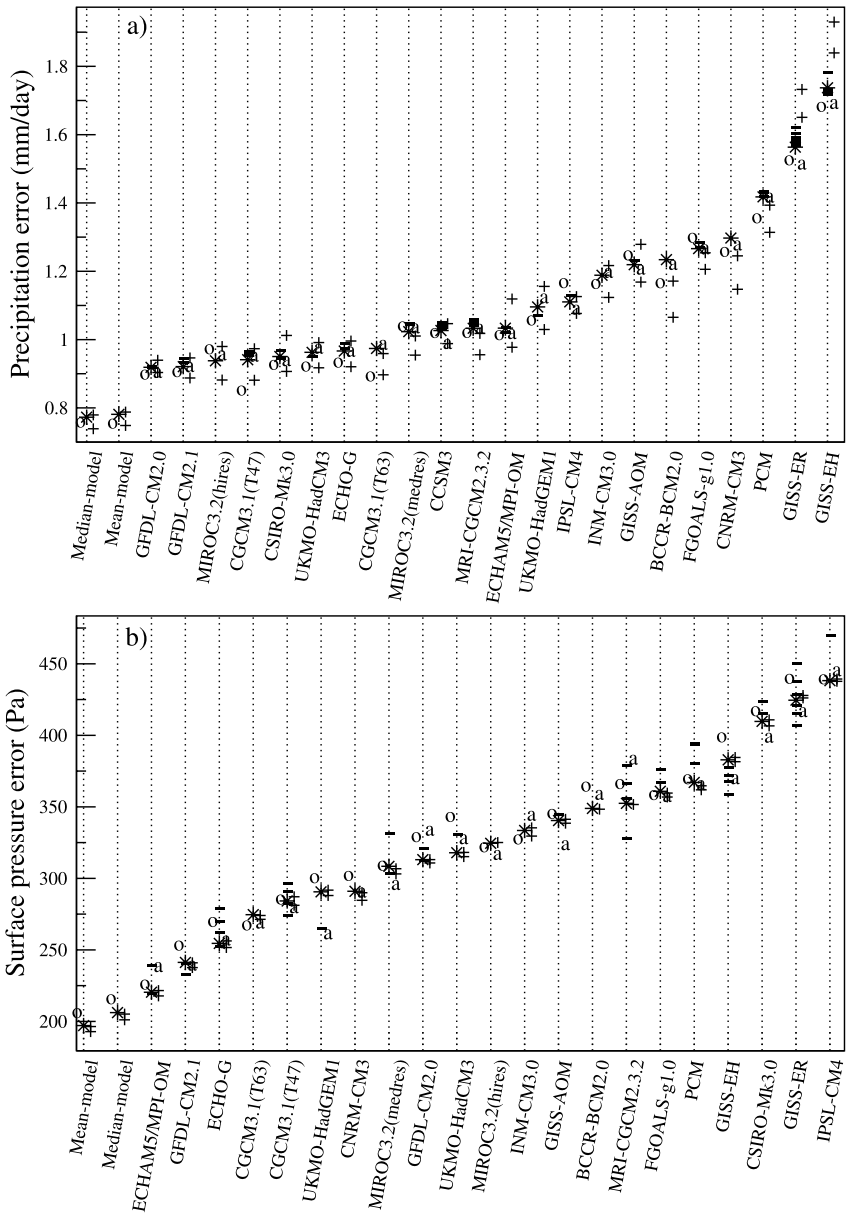

Figure 6. Model RMS errors in the Northern Hemisphere extra-tropics (20N–90N) for (a) precipitation and (b) mean sea level pressure. The sensitivity to different analysis choices are shown by use of different symbols: standard choice (*), alternate reference data set (o), different climatological averaging period (a), target grid at different resolutions (+), and alternate ensemble members (–). The models are ordered according to the errors calculated with the standard analysis procedure.

models judged to be more reliable would be weighted more heavily than the other models in forming the consensus prediction.

[57] Although defining an optimal index of this kind is beyond the scope of this study, for exploratory purposes only a "Model Climate Performance Index" (MCPI) has been constructed. This is done rather arbitrarily by simply averaging each model's relative errors across all of the fields appearing in Figure 3. Results are shown in Figure 7a (NHEX) and Figure 7b (Tropics), with models sorted by their MCPI (black line) and with the zero line indicating the mean result across all models. For each model the relative error for each variable contributing to the index is also shown (symbols). Not surprisingly, the MCPI's based on the multimodel mean and median fields are lower than the MCPI's of individual models. The models that fare well by this measure also stand out in Figure 3, as do the outliers with the larger errors. However, even the "better" models have a large spread in their variable-by-variable performance, particularly in the tropics. Also note that the ranking of models is somewhat different between the NHEX and Tropics. Thus combining these two indices of performance would result in a noticeable loss of information concerning model error.

[58] The average relative error of each model (i.e., MCPI) is a residual of a rather large spread of variable-specific

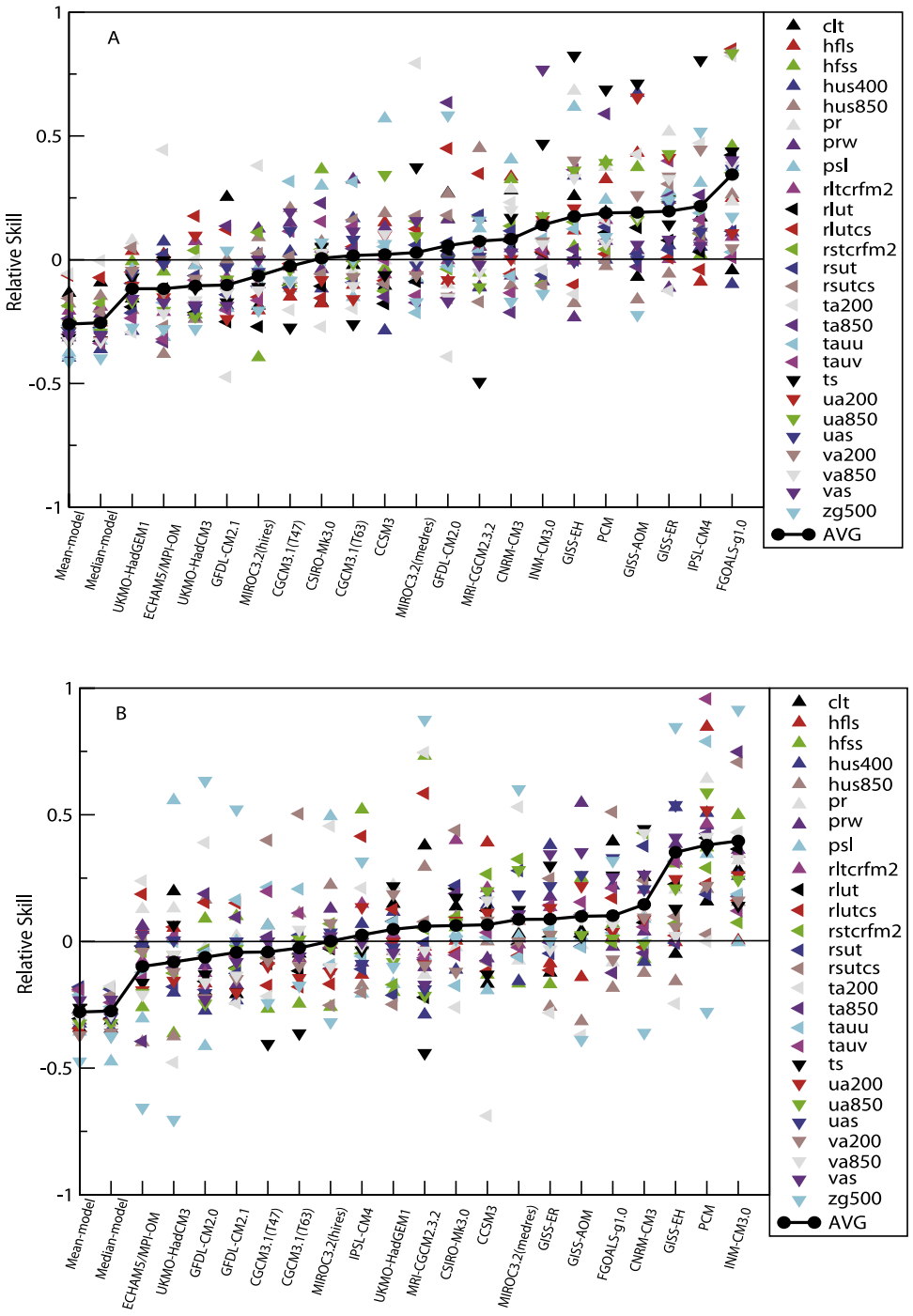

Figure 7. Relative errors, with models ordered by the "Model Climate Performance Index," for (a) NHEX (20N–90N) taken from Figure 3d, and (b) Tropics (20S–20N) taken from Figure 3e. The indices are connected by the solid line, and the colored symbols indicate the relative error for each of the variables that contribute to the index.

model performance, which means that the MCPI can hide substantial model errors. Although the MCPI seems to reinforce the general impression of model performance conveyed by Figures 3d and 3e, it is our view that the complexity of the models and the characteristics of their simulated fields cannot be adequately captured by a single measure of performance. Furthermore, our decision to weight all fields uniformly is highly subjective and arbitrary, and therefore counter to the goal of gauging model performance by measures that are objective. At this point the utility of this or any similarly defined MCPI is therefore unclear. It is likely that depending on the application, it would be appropriate to weight different aspects of model error by different amounts.

[59] The primary appeal of a single index is its simplicity, but a single index could lead some naive individuals to draw

unwarranted conclusions concerning the relative value of different models. This is especially true if the index is based solely on the climatology of global scale fields, completely omitting any evaluation of the wide variety of modes of variability that might indicate whether models have really captured the physics of the climate system. There is understandable concern among model developers that a single index used to rank models could prematurely discourage development of new modeling approaches, which might at first appear inferior, but which might more realistically represent the physics of the system and, after further work, could eventually produce a better model.

3.5. Beyond the Mean Climate

[60] The focus of this study has thus far been on the simulation of the mean annual cycle. While this may be a reasonable starting point, it provides only a limited perspective of overall climate model performance. Here we take a preliminary look at simulated inter-annual variability by examining variances of monthly mean anomalies, computed relative to the monthly climatology. The RMS error is not an appropriate metric for characterizing this aspect of model performance because there is no reason to expect models and observations to agree on the phasing of internal (unforced) variations (e.g., the timing of El Niño events). On the other hand, correctly matching the observed variance does not guarantee correct representation of the modes of variability responsible for the variance.

[61] In order to obtain a robust statistical estimate of observed variability, we need data available for a period of a decade or more. This constraint eliminates some of the reference data sets used in our earlier comparisons of the simulated mean climate state. As a first step in our assessment of model simulation of climate variability, we focus on the upper air fields available from the ERA40 and NCEP-NCAR reanalyses. After removing the mean annual cycle from each reanalysis data set, monthly anomalies are computed for the period 1980–1999. A similar procedure is followed for each of the 20th century simulations, where for each model we use the same realization included in Figures 1–5.

[62] Figure 8 depicts the ratio of simulated to observed variances for the NHEX, Tropics and SHEX (with reference to each of the reanalyses). Ratios close to unity indicate that the variance of simulated monthly anomalies compare well with the reanalyses, whereas lower ratios suggest there is too little simulated variability and higher ratios imply too much. The statistics for the multimodel mean field are omitted in this figure because the phasing of the monthly anomaly variations cannot be expected to be the same from one model to the next because they are not externally forced. Thus the temporal variance found in "mean model" and "median model" anomaly fields is unrealistically small and would in fact become smaller still if additional models were added to the ensemble.

[63] In the NHEX, the variance in both reanalyses are comparable, as evidenced by the preponderance of single color boxes. In the Tropics and in the SHEX the two triangles in each box are more frequently different, indicating larger discrepancies between the two reanalyses. This is not surprising, since the reanalyses are less well constrained by observations outside the Northern Hemisphere midlatitudes. In areas of relatively sparse data, the reference data sets are more heavily influenced by the reanalysis model and therefore the observational estimates of variability are likely to be less accurate.

[64] Most models have too little extra-tropical variability in temperature at 200 hPa. This may be because the high-latitude winter tropopause is well below 200 hPa, and most models have relatively crude representation of the stratosphere. Several models stand out consistently across all fields with too little variability in tropics, whereas others appear to have too much.

[65] We can create an exploratory Model Variability Index (MVI), much as we did for the mean climate by arbitrarily weighting each of the variables in Figure 8 uniformly. To avoid cancellation between excessive and deficient variability, we define for a given model (m) and reference data set (r) the MVI as follows:

$$MVI_{mr} = \sum_{f=1}^{F} \left[\beta_{mrf} - \frac{1}{\beta_{mrf}} \right]^2 \tag{3}$$

where β^2 is the ratio of simulated to observed variance and F is the total number of variables (rows in each of the panels of Figure 8). Defined in this way, the MVI is positive definite, with smaller values indicating better agreement with the reference data.

[66] Is there any relationship between how well a model simulates the mean climate and its ability to capture large-scale characteristics of inter-annual variability? Using our simple indices of performance, we compare each model's climate and variability skill for both the NHEX and Tropics (Figures 9a–9b). For both our MCPI and MVI, the smaller the value, the better the skill, so the "better" models would be found in the lower left corner of Figure 9. In the NHEX there appears to be some relationship between the two performance measures, although there is substantial scatter. In the tropics, on the other hand, there seems to be hardly any relationship between a model's relative skill in simulating the mean climate and its inter-annual variability performance. Note that for consistency, the MCPI values shown in Figure 9 are based on the set of fields used for the MVI. For most models this leads to only modest changes in the MCPI shown in Figure 7, but it is interesting to note that in the tropics there are substantial changes for all three GISS models.

[67] There is a possibility that the results in Figure 9b might be sensitive to the occurrence of relatively rare events such as El Niño. We can check whether our results are sensitive to sampling a particular realization of climate "noise" of this kind by analyzing the multiple realizations available for some models (not shown). We find little change in the results of Figure 9, which suggests that for our global scale measures, a sample size of 20 years yields a robust measure of each model's relative performance in simulating variability.

[68] It can be argued that a realistic simulation of the annual cycle might be required for a model to have reasonable characteristics of inter-annual variability (e.g., to account for the influence of the Asian monsoon on ENSO). This is not necessarily inconsistent with the results displayed in Figure 9, but improving a model's simulation

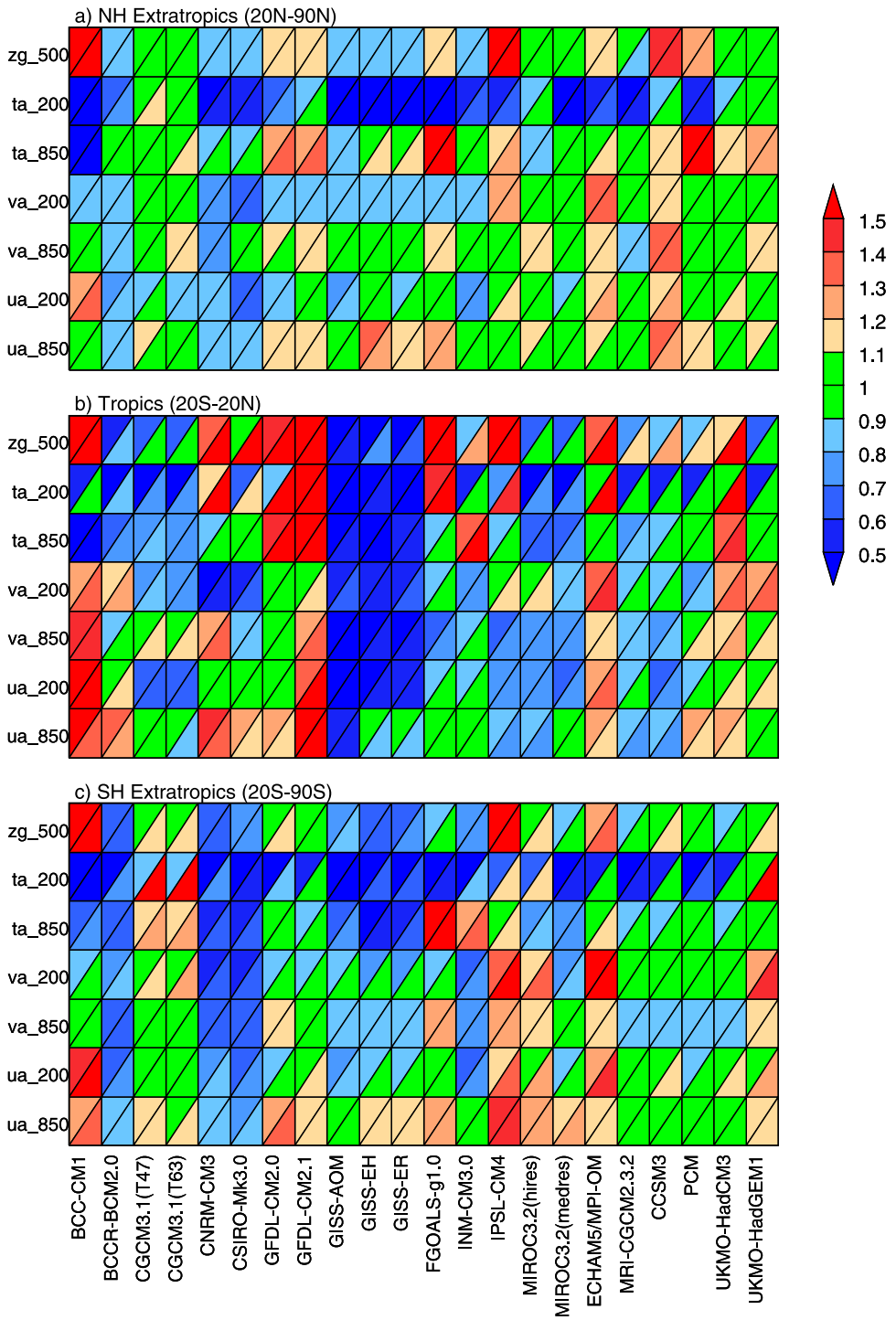

Figure 8. Variance ratios (CMIP3/reanalysis) for 1980–1999 monthly anomalies in the (a) NHEX (20N–90N), (b) Tropics (20S–20N), and (c) SHEX (90S–20N). In each rectangle the upper triangle is based on ERA40 and the lower on NCEP-NCAR.

of the mean climate is no guarantee that its variability will also improve.

[69] These results underscore the limitations of mean climate metrics; they may be woefully inadequate for assessing the multiple facets of model performance. At this time it therefore appears that the climate research community is better served by further work to develop a comprehensive hierarchy of model metrics, which can be used to assess the spectrum of processes and phenomenon considered important for the simulation of climate.

4. Summary and Discussion

[70] Climate model "metrics," as described here, are scalar quantities designed to gauge model performance. Defined for this purpose, metrics can be contrasted with

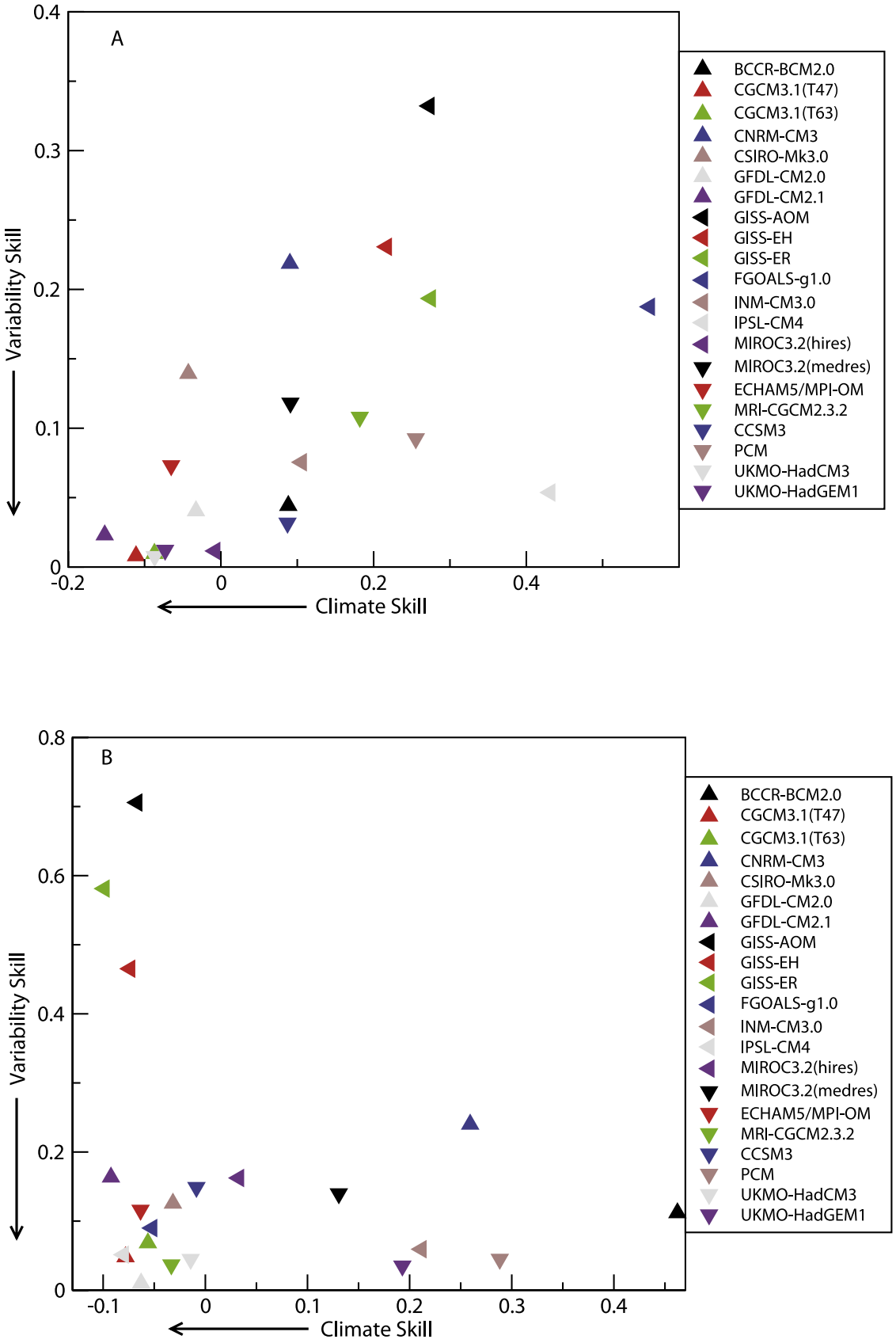

Figure 9. Model Climate Performance Index (MCPI) versus Model Variability Index (MVI) for (a) NHEX (20N–90N) and (b) Tropics (20S–20N). For consistency, the MCPI values shown here are based on the same set of fields used for the MVI shown in Figure 8.

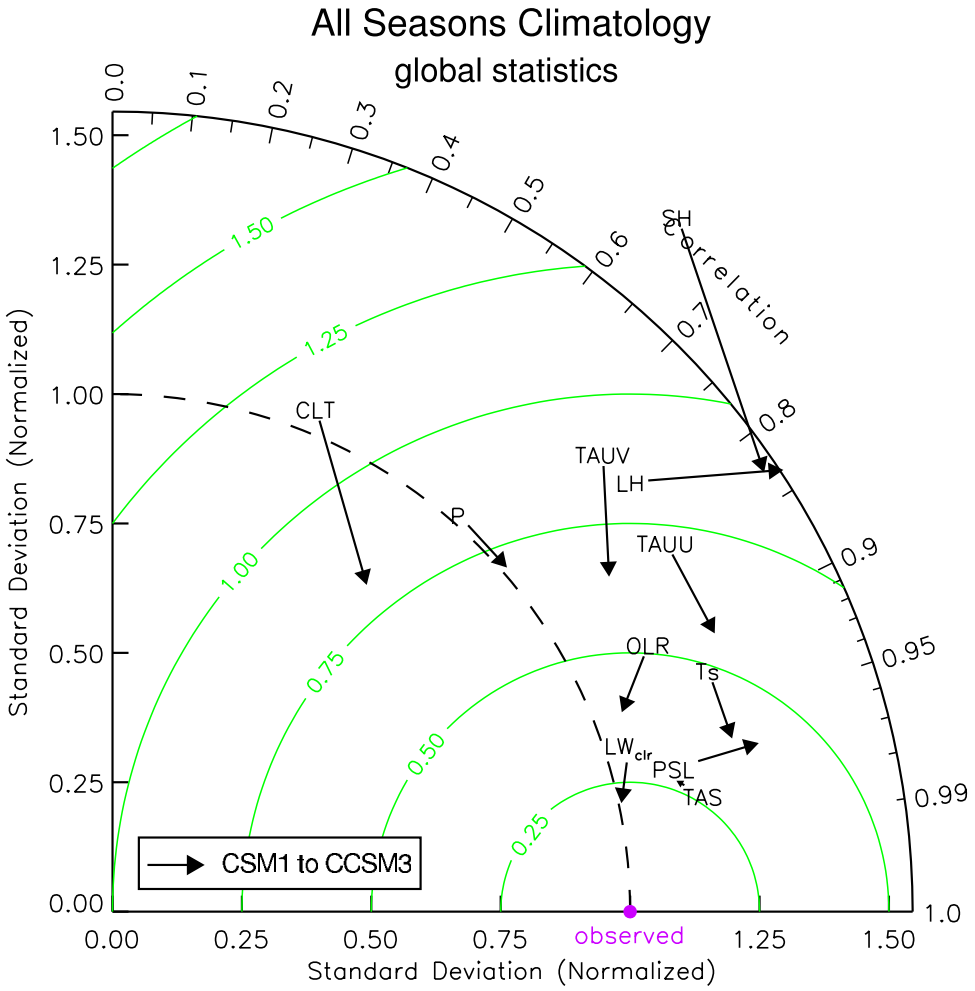

Figure 10. Example of a Taylor diagram used in the evaluation of two different versions of an NCAR coupled climate model. The following equivalence between acronyms applies: OLR = rlut, LH = hfls, SH = hfss, P = pr, LW_{clr} = rlutcs, and TAS = surface air temperature.

"diagnostics," which may take many forms (e.g., maps, time series, power spectra) and may often reveal more about the causes of model errors and the processes responsible for those errors. There is, for a variety of reasons, growing interest within the climate research community in establishing a standard suite of metrics that characterize model performance. Metrics are usually designed to quantify how simulations differ from observations, and they are generally used to characterize how well models compare with each other. Unlike numerical weather prediction, there is currently no widely accepted suite of metrics for evaluating climate model performance. The greatest challenge in selecting metrics for measuring climate model performance is determining what phenomena are important to simulate accurately, and therefore what the metrics need to measure. It remains largely unknown what aspects of observed climate must be correctly simulated in order to make reliable predictions of climate change.

[71] The metrics used here indicate that models are not all equally skillful in simulating the annual cycle climatology and the variance of monthly anomalies. The information provided by our metrics makes it possible for anyone to draw inferences about the relative performance of different models, but here we point out some obvious generalizations. First, the "mean model" and "median model" exhibit clear superiority in simulating the annual cycle climatology. This conclusion is robust across variables, regions (tropics and extra-tropics) and the component considered (e.g., deviations from the annual mean, annual mean, deviations from the zonal mean). Second there are some models that in many respects stand out as superior. In the extra-tropics, for example, we find the UKMO-HadCM3, UKMO-HadGEM1, GFDL-CM2.1, MICROC3.2 (hires), and MPI-ECHAM5 errors are smaller than those found in the "typical" model by more than 10%. Relative to the most poorly performing models, these errors are lower by up to 30%–40%. In the tropics, the UKMO-HadCM3, MPI-ECHAM5, CCCMA-CGCM-1 (at both resolutions), and both GFDL model versions, each have overall errors on the order of 5% less than the typical error, and on the order of 30% lower than the most poorly performing models. While quantitative, these conclusions are drawn by looking collectively at a host of variables having a wide range of observational uncertainty. Moreover, even for these "better" performing models we must reiterate the fact that the range of performance across variables is substantial, and at

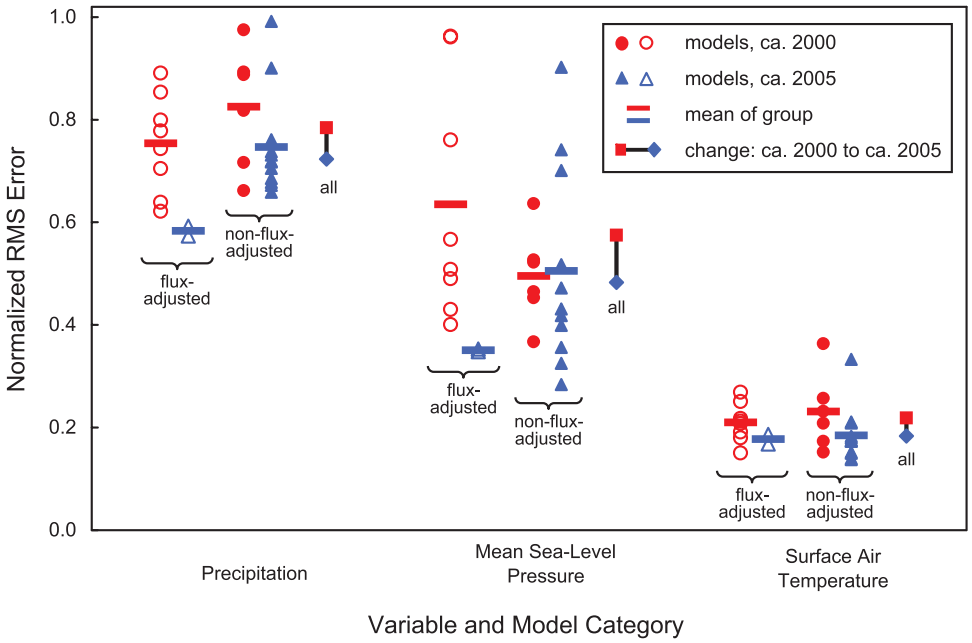

Figure 11. CMIP2 and CMIP3 model errors in simulating precipitation, mean sea level pressure and surface air temperature (prepared for *IPCC*, 2007). All fields were mapped to a 4x5 degree latitude-longitude grid before computing the errors. Results from the earlier generation of models (CMIP2) are based on the output from control runs (specifically, the first 30 years, in the case of temperature, and the first 20 years for the other fields), whereas results from the recent model versions (CMIP3) are based on the 20th century simulations.

least in the tropics there is little indication that relative mean climate performance translates to how well models simulated basic characteristics of variability.

[72] This study has shown that the relative performance of models reflected by metrics proposed here can be sensitive to the choice of reference data, internal variability (e.g., exhibited via multiple realizations of the same model), and even the resolution of the regridded data. While none of these factors leads to a complete rearrangement of the relative ranking of models (the "above average" models typically remain so), their impact can move a model up or down in the ranking by several slots.

[73] There are a number of ways one can arbitrarily construct an "index of climate skill," and we have illustrated one obvious way to do this. While the results are generally consistent with our impressions gleaned from a large suite of metrics (as to which models do relatively well, and which do not), we have not demonstrated that in fact this metric has any specific value in determining which model might be more reliable in predicting climate change. We note that even the "better" models score below average in the simulation of some fields, while the "poorer" models score above average in some respects (especially in the tropics). Thus an overall performance index of this kind is a relatively small residual resulting from the large range of scores on which it is based. Because the component scores are to a considerable degree offsetting, the reasonable, but somewhat arbitrary, decision to weight each variable equally is a critically important one. A different choice could lead to a rather different ranking of the models. It is likely that the optimal weighting of individual metrics contributing to a performance index will depend on the application (e.g., prediction of climate change versus study of ENSO). Additional research is needed to determine which aspects of a model simulation that can be verified against observations are most critical for predictive reliability. Until then, the use of a single index to gauge model performance is unwarranted and scientifically unjustified.

[74] Further research should lead to more comprehensive and useful means of summarizing mean climate performance. Initially it might be fruitful to explore a wide range of metrics, rather than striving for a single index of overall skill, and then to use some objective method to reduce redundant information (e.g., Single Value Decomposition techniques). This might lead us to a more robust measure of skill. In this study our aim has been to set the ground-work for future efforts to define standard metrics by exploring the sensitivity of metrics to various analysis choices. At this stage, however, it would be premature to suggest that this (or any other) set of metrics be adopted as a standard for climate modeling. Eventually it should be possible to establish standard performance metrics that could be rigorously justified as providing a useful guide to their predictive capability, but this will require a better understanding of the relationship between a model's ability to simulate observed phenomena and its ability to simulate climate changes.

[75] Metrics are currently being used by some groups to aid in the development of new model versions by quantifying how model changes affect performance. This practical application of metrics can help modelers distill a range of diagnostic information in ways that can enable them to quickly review chosen performance measures of many model versions. Figure 10 is an example of how the Taylor diagram has been exploited to track performance changes in

an atmospheric model. Knowledgeable model development teams are unlikely to over-emphasize the value of metrics because they will be aware of a much broader suite of diagnostics routinely considered during model development. Any subjectivity associated with the definition of performance indices will be recognized within these groups and taken into account.

[76] Another important use of model metrics is to monitor how climate models improve over time. This is greatly facilitated by established "benchmark" experiments that are expected to be performed whenever a new model version is developed. The protocol for the Atmospheric Model Intercomparison Project [AMIP, *Gates et al.*, 1999] and the control experiments of the Coupled Model Intercomparison Project [CMIP, *Meehl et al.*, 2000] are prime examples of standard experiments of this kind. Figure 11 provides a summary of the ability of OAGCMs to simulate the seasonally varying climate state. The RMS error normalized by the amplitude of the pattern (i.e., the standard deviation) is given for precipitation, sea level pressure and surface temperature. The subset of the climate research centers given in Table 1 who contributed model output to the CMIP database from both an earlier and more recent version of their model are included in the plot. The models in Figure 11 are identified by open or filled symbols, depending on whether or not flux adjustments were applied. Only two of the 8 groups who originally used flux adjustment continue that practice. The figure shows that flux adjusted models on average have smaller errors than those without (in both generations), but considering all the models, the smallest errors in simulating sea level pressure and surface temperature are found in those without flux adjustment. Also, despite the elimination of flux adjustment in all but two of the recent models, the mean error obtained from the recent suite of 14 models is smaller than errors found in the corresponding earlier suite of models. Moreover, both flux adjusted models, as a group, and their non-flux-adjusted counterparts have with one exception improved.

[77] A third use of model metrics is to rely on them to make quantitative judgments on how to use information from a collection of models for a particular application [e.g., *Annamalai and Hamilton*, 2007]. One way this can be done is to establish what is important, and then to weight individual simulations accordingly. A variant of this is to use metrics to eliminate some models from a multimodel ensemble (i.e., assigning them a weight of 0).

[78] Finally, in spite of the increasing use of metrics in the evaluation of models, it is not yet possible to answer the question often posed to climate modelers: "What is the best model?" The answer almost certainly will depend on the intended application. Conceivably, a set of metrics could be developed for a specific application that would accurately quantify the relative merits of different models. It is unlikely, however, that the same set of metrics would be optimally suited for all applications. Therefore, a prudent strategy would be to encourage the development of performance metrics for a wide range of processes and phenomenon of known importance for climate. Among these metrics would be peformance measures of the simulated atmosphere, the ocean [e.g., *McClean et al.*, 2006], and the land and cryosphere. With a reasonably comprehensive set of model metrics identified, the climate research community will be better positioned to evaluate overall climate model performance and determine which metrics are particularly relevant to any given application.

[79] **Acknowledgments.** We would like to thank the members of the CAS/JSC WGNE for their persistent encouragement to pursue this work. We acknowledge the modeling groups, the PCMDI, and the WCRP's WGCM for their roles in making available the WCRP CMIP3 multimodel data set. Support of this data set is provided by the Office of Science, U.S. Department of Energy. The same office, through its Climate Change Prediction Program, supported this research, which was carried out at the University of California Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

References

Adler, R. F., G. J. Huffman, A. Chang et al. (2003), The Version-2 Global Precipitation Climatology Project (GPCP) monthly precipitation analysis (1979–present), *J. Hydrometeorol.*, *4*, 1147–1167.

Annamalai, H., and K. Hamilton (2007), The South Asian summer monsoon and its relationship with ENSO in the IPCC AR4 simulations, *J. Clim.*, *20*, doi:10.1175/jcli4035.1.

Aumann, H. H., et al. (2003), AIRS/AMSU/HSB on the Aqua mission: Design, science objectives, data products, and processing systems, *IEEE Trans. Geosci. Remote Sens.*, *41*, 253–264.

Barkstrom, B. (1984), The earth radiation budget experiment, *Bull. Am. Meteorol. Soc.*, *65*, 1170–1185.

Bourassa, M. A., M. H. Freilich, D. M. Legler, W. T. Liu, and J. J. O'Brien (1997), Wind observations from new satellite and research vessels agree, *EOS Trans. Am. Geophys. Union*, 597–602.

Cess, R. D., et al. (1990), Intercomparison and interpretation of climate feedback processes in 19 atmospheric general circulation models, *J. Geophys. Res.*, *95*(16), 601.

Gates, W. L., et al. (1999), An overview of results from the Atmospheric Model Intercomparison Project (AMIP1), *Bull. Am. Meteorol. Soc.*, *80*, 29–55.

Huffman, G. J., et al. (1997), The Global Precipitation Climatology Project (GPCP) combined data set, *Bull. Am. Meteorol. Soc.*, *78*, 5–20.

IPCC (2007), Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor and H. L. Miller (eds.), Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 996 pp.

Josey, S. A., E. C. Kent, and P. K. Taylor (1999), New insights into the ocean heat budget closure problem from analysis of the SOC air-sea flux climatology, *J. Clim.*, *12*, 2856–2880.

Kalnay, E. M., et al. (1996), The NCAR/NCAR 40-year reanalysis project, *Bull. Am. Meteorol. Soc.*, *77*, 437–471.

Lambert, S. J., and G. J. Boer (2001), CMIP1 evaluation and intercomparison of coupled climate models, *Clim. Dyn.*, *17*, 83–106.

McClean, J. L., M. E. Maltrud, and F. O. Bryan (2006), Measures of the fidelity of eddy resolving ocean models, *Oceanography*, *19*, 104–117.

Meehl, G. A., G. J. Boer, C. Covey, M. Latif, and R. J. Stouffer (2000), The Coupled Model Intercomparison Project (CMIP), *Bull. Am. Meteorol. Soc.*, *81*, 313–318.

Murphy, J. M., et al. (2004), Quantification of uncertainties in large ensembles of climate change prediction, *Nature*, *430*, 768–772, doi:10.1038/nature02771.

Phillips, et al. (2004), Evaluating parameterizations in general circulation models: Climate simulation meets weather prediction, *Bull. Am. Meteorol. Soc.*, *85*, 1903–1915.

Pincus, R., C. P. Batstone, R. J. P. Hofmann, K. E. Taylor, and P. J. Glecker (2008), Evaluating the present-day simulation of clouds, precipitation, and radiation in climate models, *J. Geophys. Res.*, doi:10.1029/2007JD009334, in press.

Rayner, N. A., P. Brohan, D. E. Parker, C. K. Folland, J. J. Kennedy, M. Vanicek, T. Ansell, and S. F. B. Tett (2006), *J. Clim.*, *19*, 446–469.

Rossow, W. B., and R. A. Schiffer (1999), Advances in understanding clouds from ISCCP, *Bull. Am. Meteorol. Soc.*, *80*, 2261–2287.

Rossow, W. B., A. W. Walker, and L. C. Garder (1993), Comparison of ISCCP and other cloud amounts, *J. Clim.*, *6*, 2394–2418.

Simmons, A. J., and J. K. Gibson (2000), *The ERA-40 project plan, ERA-40 Project Report Series No. 1*, 62 pp., Reading, U.K.

Simpson, J. J., et al. (2001), The NVAP global water vapor dataset: Independent cross-comparison and multiyear variability, *Remote Sens. Environ.*, *76*, 112–129.

Smith, T. M., and R. W. Reynolds (2004), *J. Clim.*, *17*, 2466–2477.

Stainforth, D. A., et al. (2005), Uncertainty in predictions of the climate response to rising levels of greenhouse gases, *Nature*, *433*, 403–405.
Taylor, K. E. (2001), Summarizing multiple aspects of model performance in a single diagram, *J. Geophys. Res.*, *106*, 7183–7192.
Taylor, P. (Ed.) (2000), Final report of the Joint WCRP/SCOR Working Group on Air-Sea Fluxes, WCRP-112.
Taylor, K. E., and P. J. Gleckler (2002), The Second Phase of the Atmospheric Model Intecomparison Project, *Lawrence Livermore National Laboratory Report Series, UCRL-PROC-209115*.
Trenberth, K. E., J. Fasullo, and L. Smith (2005), Trends and variability in column-integrated atmospheric water vapor, *Clim. Dyn.*, *24*, 741–758.
Wentz, F. J. (2000), A well-calibrated ocean algorithm for SSM/I, *J. Geophys. Res.*, *102*, 8703–8718.
Wielicki, B. A., et al. (1996), Clouds and the Earth's Radiant Energy System (CERES): An Earth observing system experiment, *Bull. Am. Meteorol. Soc.*, *77*, doi:10.1175/1520-0477, 853–868.
Williamson, D. (1995), Skill scores from the AMIP simulations, *WMO report series, WMO/TD-No. 732*.
WMO (1994), Report on the ninth session of the CAS/JSC Working Group on Numerical Experimentation, CAS/JSC WGNE Report No. 9, WMO/TD-No. 607, 38pp.
Yin, X., A. Gruber, and P. Arkin (2004), Comparison of the GPCP and CMAP merged gauge satellite monthly precipitation products for the period 1979–2001, *J. Hydrometeorol.*, *5*, 1207–1222.

C. Doutriaux, P. J. Gleckler, and K. E. Taylor, Program for Climate Model Diagnosis and Intercomparison, Lawrence Livermore National Laboratory, Livermore, CA 94550, USA. (pgleckler@llnl.gov)