How Well Do Coupled Models Simulate Today's Climate?

BY THOMAS REICHLER AND JUNSU KIM

Coupled climate models are sophisticated tools designed to simulate the Earth climate system and the complex interactions between its components. Currently, more than a dozen centers around the world develop climate models to enhance our understanding of climate and climate change and to support the activities of the Intergovernmental Panel on Climate Change (IPCC). However, climate models are not perfect. Our theoretical understanding of climate is still incomplete, and certain simplifying assumptions are unavoidable when building these models. This introduces biases into their simulations, which sometimes are surprisingly difficult to correct. Model imperfections have attracted criticism, with some arguing that model-based projections of climate are too unreliable to serve as a basis for public policy. In particular, early attempts at coupled modeling in the 1980s resulted in relatively crude representations of climate. Since then, however, we have refined our theoretical understanding of climate, improved the physical basis for climate modeling, increased the number and quality of observations, and multiplied our computational capabilities. Against the background of these developments, one may ask how much climate models have improved and how much we can trust the latest coupled model generation.

The goal of this study is to objectively quantify the agreement between model and observations using a single quantity derived from a broad group of variables, which is then applied to gauge several generations of coupled climate models. This approach is new, since previous model intercomparison studies either focused on specific processes, avoided making quantitative performance statements, or considered a rather narrow range of models.

Several important issues complicate the model validation process. First, identifying model errors is difficult because of the complex and sometimes poorly understood nature of climate itself, making it difficult to decide which of the many aspects of climate are important for a good simulation. Second, climate models must be compared against present (e.g., 1979–99) or past climate, since verifying observations for future climate are unavailable. Present climate, however, is not an independent dataset since it has already been used for the model development. On the other hand, information about past climate carries large inherent uncertainties, complicating the validation process of past climate simulations. Third, there is a lack of reliable and consistent observations for present climate, and some climate processes occur at temporal or spatial scales that are either unobservable or unresolvable. Finally, good model performance evaluated from the present climate does not necessarily guarantee reliable predictions of future climate. Despite these difficulties and limitations, model agreement with observations of today's climate is the only way to assign model confidence, with the underlying assumption that a model that accurately describes present climate will make a better projection of the future.

Considering the above complications, it is clear that there is no single "ideal way" to characterize and compare model performances. Most previous model validation studies used conventional statistics to measure the similarity between observed and modeled data. For example, studies by Taylor et al. (2001) and Boer and Lambert (2001) characterized model performance from correlation, root-mean-square (RMS) error, and variance ratio. Both studies found similar ways to combine these three statistics

AFFILIATIONS: REICHLER AND KIM—Department of Meteorology, University of Utah, Salt Lake City, Utah
CORRESPONDING AUTHOR: Thomas Reichler, Department of Meteorology, University of Utah, 135 S 1460 E, Rm 819 (WBB), Salt Lake City, UT 84112-0110
E-mail: thomas.reichler@utah.edu

DOI:10.1175/BAMS-89-3-303

©2008 American Meteorological Society

in a single diagram, resulting in nice graphical visualizations of model performance. This approach, however, is only practical for a small number of models and/or climate quantities. In addition, Taylor's widely used approach requires centered RMS errors with the mean bias removed. We, however, consider the mean bias as an important component of model error. In a 2004 article, Murphy et al. introduced a Climate Prediction Index (CPI), which measures the reliability of a model based on the composite mean-square errors of a broad range of climate variables. More recently, Min and Hense (2006) introduced a Bayesian approach into model evaluation, where skill is measured in terms of a likelihood ratio of a model with respect to some reference.

THREE GENERATIONS OF MODEL DATA. This study includes model output from three different climate model intercomparison projects (CMIP): CMIP-1 (Meehl et al. 2000), the first project of its kind organized in the mid-1990s; the follow-up project CMIP-2 (Covey et al. 2003, Meehl et al. 2005); and CMIP-3 (PCMDI 2007) (aka, IPCC-AR4), representing today's state of the art in climate modeling. The CMIP-3 data were taken from the "climate of the twentieth century" (20C3M) (hereafter simply "present-day") and the "preindustrial control" (PICNTRL) (hereafter simply "preindustrial") experiments. These simulations were driven by a rather realistic set of external forcings, which included the known or estimated history of a range of natural and anthropogenic sources, such as variations in solar output, volcanic activity, trace gases, and sulfate aerosols. The exact formulation of these forcings varied from model to model, with potential implications for model performance. In contrast, the CMIP-1 and CMIP-2 model output was derived from long control runs, in which the forcings were held constant in time. These forcings were only approximately representative for present climate.

MEASURE OF MODEL PERFORMANCE. As outlined before, there are many different ways to measure and depict model performance. Given the extra challenge of this study to evaluate and depict a large number of models and climate variables, we decided to design our own measure. Our strategy was to calculate a single performance index, which can be easily depicted, and which consists of the aggregated errors in simulating the observed climatological mean states of many different climate variables. We focused on validating the time-mean state of climate, since this is the most fundamental and best-observed aspect of climate, and because of restrictions imposed by available model data in calculating higher moments of climate (most CMIP-1 fields are archived as climatological means, prohibiting the derivation of temporal variability). This concept is somewhat similar to the CPI performance measure introduced by Murphy et al. (2004), but in contrast to the present study, Murphy et al. calculated the CPI from a range of rather closely related models.

Our choice of climate variables, which is shown in Table 1, was dictated by the data available from the models. In most cases, we were able to validate the model data against true observation-based data, but for a few variables of the free atmosphere, the usage of reanalyses as validation data was unavoidable. In terms of the specific uncertainties associated with each of those validating datasets, separate analysis showed that the data can be considered as good approximations to the real state of present climate for the purpose of model validation.

We obtained the model performance index by first calculating multiyear annual mean climatologies from global gridded fields of models and validating data. The base period for the observations was 1979–99, covering most of the well-observed post-1979 satellite period. For some observations, fewer years were used if data over the entire period were not available. For the CMIP-1 models, long-term climatologies of the control run for Northern Hemisphere winter (December, January, February) and summer (June, July, August) conditions were downloaded from the archives and averaged to annual mean climatologies. The CMIP-2 climatologies were calculated by averaging the annual mean data of the control run over the years 61–80. The CMIP-3 present-day climatologies were formed using the same base period as for the observations, and the preindustrial climatologies were taken from the last 20 simulation years of the corresponding control run. For any given model, only one member integration was included. In the rare case that a climate variable was not provided by a specific model, we replaced the unknown error by the mean error over the remaining models of the corresponding model generation. One model (BCC-CM1 from CMIP-3) was excluded because it only provided a small subset of variables needed for this study.

In determining the model performance index, we first calculated for each model and variable a normalized error variance e^2 by squaring the grid-point differences between simulated (interpolated to the observational grid) and observed climate, normalizing

Table 1. Climate variables and corresponding validation data. Variables listed as "zonal mean" are latitude–height distributions of zonal averages on twelve atmospheric pressure levels between 1000 and 100 hPa. Those listed as "ocean," "land," or "global" are single-level fields over the respective regions. The variable "net surface heat flux" represents the sum of six quantities: incoming and outgoing shortwave radiation, incoming and outgoing longwave radiation, and latent and sensible heat fluxes. Period indicates years used to calculate observational climatologies.

Variable	Domain	Validation data	Period
Sea level pressure	ocean	ICOADS (Woodruff et al. 1987)	1979–99
Air temperature	zonal mean	ERA-40 (Simmons and Gibson 2000)	1979–99
Zonal wind stress	ocean	ICOADS (Woodruff et al. 1987)	1979–99
Meridional wind stress	ocean	ICOADS (Woodruff et al. 1987)	1979–99
2-m air temperature	global	CRU (Jones et al. 1999)	1979–99
Zonal wind	zonal mean	ERA-40 (Simmons and Gibson 2000)	1979–99
Meridional wind	zonal mean	ERA-40 (Simmons and Gibson 2000)	1979–99
Net surface heat flux	ocean	ISCCP (Zhang et al. 2004), OAFLUX (Yu et al. 2004)	1984 (1981) –99
Precipitation	global	CMAP (Xie and Arkin 1998)	1979–99
Specific humidity	zonal mean	ERA-40 (Simmons and Gibson 2000)	1979–99
Snow fraction	land	NSIDC (Armstrong et al. 2005)	1979–99
Sea surface temperature	ocean	GISST (Parker et al. 1995)	1979–99
Sea ice fraction	ocean	GISST (Parker et al. 1995)	1979–99
Sea surface salinity	ocean	NODC (Levitus et al. 1998)	variable

on a grid-point basis with the observed interannual variance, and averaging globally. In mathematical terms this can be written as

$$e^2_{vm} = \sum_n \left(w_n \left(\overline{s}_{vmn} - \overline{o}_{vn} \right)^2 \Big/ \sigma^2_{vn} \right), \qquad (1)$$

where $\overline{s}_{vmn}$ is the simulated climatology for climate variable (v), model (m), and grid point (n); $\overline{o}_{vn}$ is the corresponding observed climatology; w_n are proper weights needed for area and mass averaging; and σ^2_{vn} is the interannual variance from the validating observations. The normalization with the interannual variance helped to homogenize errors from different regions and variables. In order to ensure that different climate variables received similar weights when combining their errors, we next scaled e^2 by the average error found in a reference ensemble of models—that is,

$$I^2_{vm} = e^2_{vm} \Big/ \overline{e^2_{vm}}^{m=20C3M}, \qquad (2)$$

where the overbar indicates averaging. The reference ensemble was the present-day CMIP-3 experiment. The final model performance index was formed by taking the mean over all climate variables (Table 1) and one model using equal weights,

$$I^2_m = \overline{I^2_{vm}}^{v}. \qquad (3)$$

The final step combines the errors from different climate variables into one index. We justify this step by normalizing the individual error components prior to taking averages [Eqs. (1) and (2)]. This guarantees that each component varies evenly around one and has roughly the same variance. In this sense, the individual I^2_{vm} values can be understood as rankings with respect to individual climate variables, and the final index is the mean over all ranks. Note that a very similar approach has been taken by Murphy et al. (2004).

RESULTS. The outcome of the comparison of the 57 models in terms of the performance index I^2 is illustrated in the top three rows of Fig. 1. The I^2 index varies around one, with values greater than one for underperforming models and values less than one

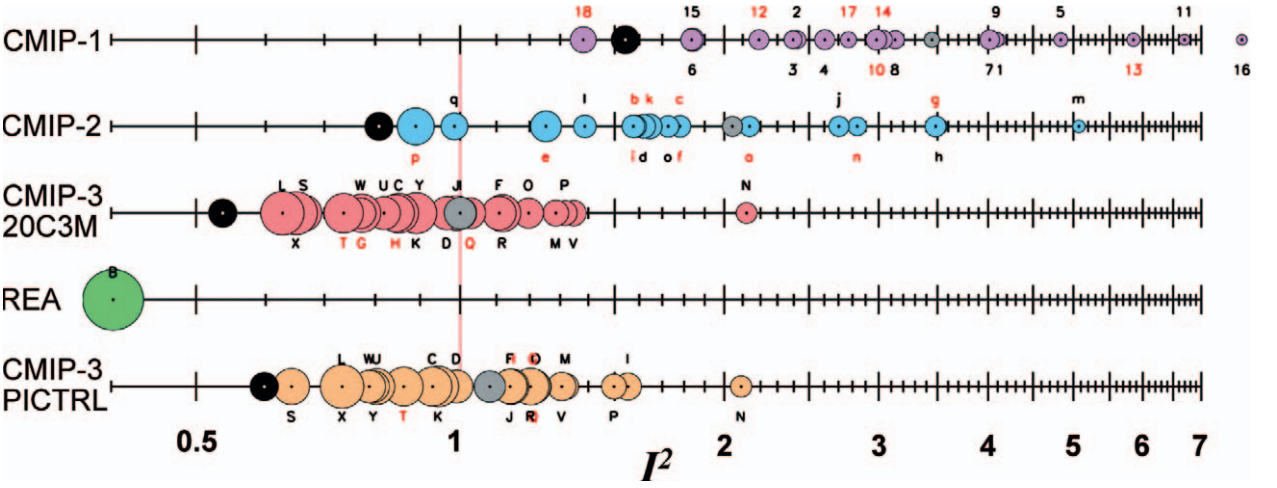

Fig. 1. Performance index I^2 for individual models (circles) and model generations (rows). Best performing models have low I^2 values and are located toward the left. Circle sizes indicate the length of the 95% confidence intervals. Letters and numbers identify individual models (see supplemental online material at doi:10.1175/BAMS-89-3-Reichler); flux-corrected models are labeled in red. Grey circles show the average I^2 of all models within one model group. Black circles indicate the I^2 of the multimodel mean taken over one model group. The green circle (REA) corresponds to the I^2 of the NCEP/NCAR reanalyses. Last row (PICTRL) shows I^2 for the preindustrial control experiment of the CMIP-3 project.

for more accurate models. Since I^2 is an indicator of model performance relative to the mean over the present-day CMIP-3 ensemble, we used a logarithmic scale to display the index. The results indicate large differences from model to model in terms of their ability to match the observations of today's climate. Further, the results clearly demonstrate a continuous improvement in model performance from the early CMIP-1 to the latest CMIP-3 generation. To our knowledge, this is the first systematic attempt to compare the performance of entire generations of climate models by exploring their ability to simulate present climate. Figure 1 also shows that the realism of the best models approaches that of atmospheric reanalysis (indicated by the green circle), but the models achieve this without being constrained by real observations.

We also obtained quantitative estimates of the robustness of the I^2 values by validating the models against a large synthetic ensemble of observational climatologies and by calculating the range of I^2 values encompassed by the 5th and 95th percentiles. The synthetic ensemble was produced by selecting the years included in each climatology using bootstrapping (i.e., random selection with replacement). To the extent that the circles in Fig. 1 overlap, it is not possible to distinguish the performance of the corresponding models in a way that is statistically significant.

ROLE OF FORCINGS. Given the more realistic forcing used for the present-day CMIP-3 simulations, the superior outcome of the corresponding models is perhaps not too surprising. One might ask how important realistic forcing was in producing such good simulations. To this end, we included the preindustrial CMIP-3 simulations in our comparison. Both the present-day and the preindustrial simulations were conducted with identical models. The only difference was the forcing used to drive the simulations, which was similar to preindustrial conditions for the preindustrial experiments and similar to present-day conditions for the present-day experiments.

The outcome of validating the preindustrial experiment against current climate is shown in the bottom row of Fig. 1. As expected, the I^2 values are now larger than for the present-day simulations, indicating poorer performance. However, the mean difference between the two CMIP-3 simulations, which was due only to different forcings, is much smaller than that between CMIP-3 and the previous two model generations. The latter difference was due to different models and forcings combined. We conclude that the superior performance of the CMIP-3 models is mostly related to drastic model improvements, and that the forcings used to drive these models play a more subtle role.

Two developments—more realistic parameterizations and finer resolutions—are likely to be most

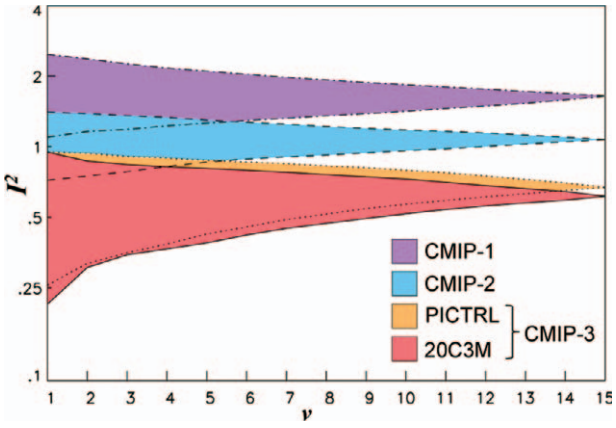

Fig. 2. Spread of I^2 values (lowest to highest) for an increasing number of randomly chosen variables v. Shown are index values averaged individually over the four model groups (corresponding to the grey circles in Fig. 1). In order to avoid nonunity results for 20C3M, all values were normalized by the mean I^2 over all three model generations, and not by the mean of the 20C3M group alone (as in Fig. 1, see Eq. 2).

responsible for the good performance seen in the latest model generation. For example, there has been a constant refinement over the years in how subgrid-scale processes are parameterized in models. Current models also tend to have higher vertical and horizontal resolution than their predecessors. Higher resolution reduces the dependency of models on parameterizations, eliminating problems since parameterizations are not always entirely physical. The fact that increased resolution improves model performance has been shown in various previous studies.

SENSITIVITY OF THE INDEX. We now address the question of how sensitive our results are with respect to our particular choice of variables. We used bootstrapping to investigate how I^2—averaged individually over the four model groups—varies with an increasing number v of variables. For any given v, we calculated I^2 many times, every time using different randomly chosen variable combinations taken from Table 1. As shown in Fig. 2, the spread of outcomes decreases with increasing number of variables. When six or more variables are used to calculate I^2, the average performances of the three model generations are well separated from each other—independent from the exact choice of variables. Only the two CMIP-3 experiments cannot be distinguished from each other, even for a very large number of variables. Also note that CMIP-3 always performs better than CMIP-1, and almost always better than CMIP-2, even when only one variable is included. These results indicate that I^2, when used to compare entire model generations, is robust with respect to the number and choice of selected variables.

VALUE OF THE MULTIMODEL MEAN. We also investigated the performance of the multimodel means (black circles in Fig. 1), which are formed by averaging across the simulations of all models of one model generation and using equal weights. Notably, the multimodel mean usually outperforms any single model, and the CMIP-3 multimodel mean performs nearly as well as the reanalysis. Such performance improvement is consistent with earlier findings by Lambert and Boer (2001), Taylor et al. (2004), and Randall et al. (2007) regarding CMIP-1, AMIP-2, and CMIP-3 model output, respectively.

The use of multimodel ensembles is common practice in weather and short-term climate forecasting, and it is starting to become important for long-term climate change predictions. For example, many climate change estimates of the recently released global warming report of the IPCC are based on the multimodel simulations from the CMIP-3 ensemble. The report dealt with the problem of inconsistent predictions, resulting from the use of different models, by simply taking the average of all models as the best estimate for future climate change. Our results indicate that multimodel ensembles are a legitimate and effective means to improve the outcome of climate simulations. As yet, it is not exactly clear why the multimodel mean is better than any individual

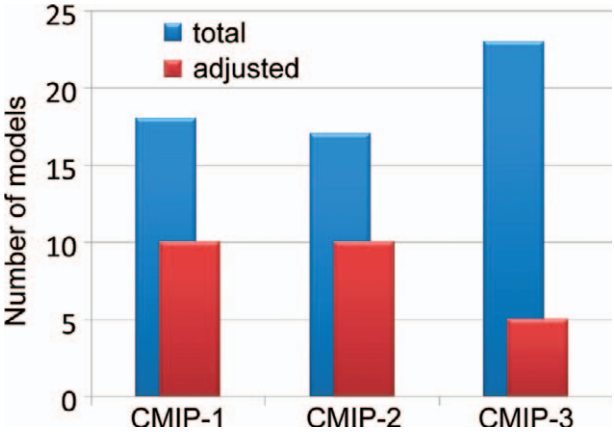

Fig. 3. Fraction of flux-adjusted models among the three model generations.

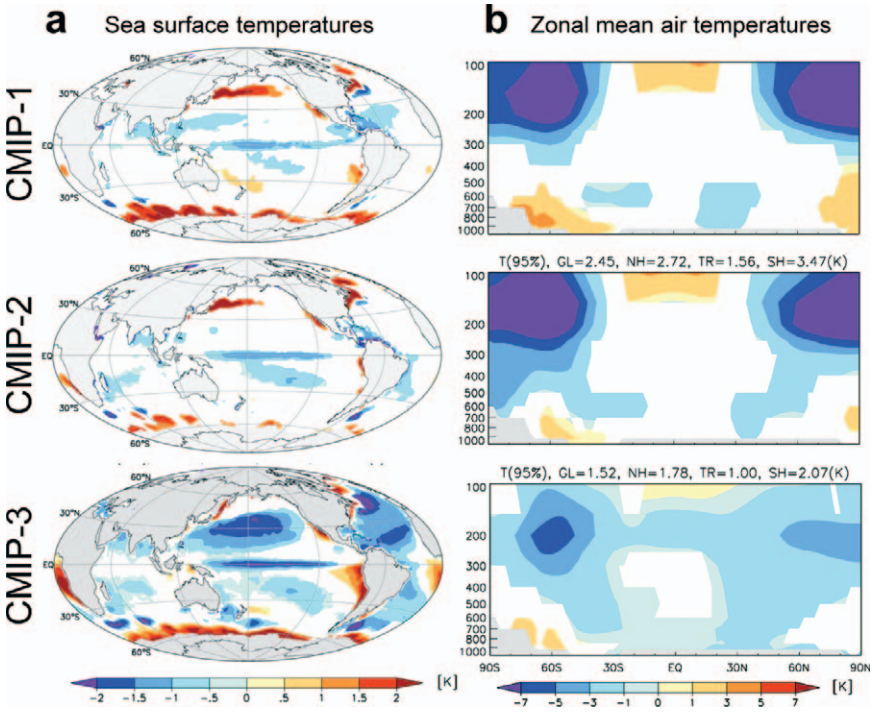

FIG. 4. Systematic biases for the three model generations. (a) Biases in annual mean climatological mean sea surface temperatures (K); (b) Biases in zonal mean air temperatures (K). Statistically significant biases that pass a Student's t-test at the 95% level are shown in color; other values are suppressed and shown in white. Gray areas denote no or insufficient data.

model. One possible explanation is that the model solutions scatter more or less evenly about the truth (unless the errors are systematic), and the errors behave like random noise that can be efficiently removed by averaging. Such noise arises from internal climate variability, and probably to a much larger extent from uncertainties in the formulation of models.

ROLE OF FLUX CORRECTION. When discussing coupled model performances, one must take into account that earlier models are generally flux corrected, whereas most modern models do not require such corrections (Fig. 3). Flux correction, or adding artificial terms of heat, momentum, and freshwater at the air–sea interface, prevents models from drifting to unrealistic climate states when integrating over long periods of time. The drift, which occurs even under unforced conditions, is the result of small flux imbalances between ocean and atmosphere. The effects of these imbalances accumulate over time and tend to modify the mean temperature and/or salinity structure of the ocean. The technique of flux correction attracts concern because of its inherently nonphysical nature. The artificial corrections make simulations at the ocean surface more realistic, but only for artificial reasons. This is demonstrated by the increase in systematic biases (defined as the multimodel mean minus the observations) in sea surface temperatures from the mostly flux-corrected CMIP-1 models to the generally uncorrected CMIP-3 models (Fig. 4a). Because sea surface temperatures exert an important control on the exchange of properties across the air–sea interface, corresponding errors readily propagate to other climate fields. This can be seen in Fig. 4b, which shows that biases in ocean temperatures tend to be accompanied by same-signed temperature biases in the free troposphere. On the other hand, the reduction of strong lower stratospheric cold biases in the CMIP-3 models indicates considerable model improvements. These cold biases are likely related to the low vertical and horizontal resolution of former model generations and to the lack of parameterizations for small-scale gravity waves, which break, deposit momentum, and warm the middle atmosphere over the high latitudes. Modern models use appropriate parameterizations to replace the missing momentum deposition.

CONCLUSION. Using a composite measure of model performance, we objectively determined the ability of three generations of models to simulate present-day mean climate. Current models are certainly not perfect, but we found that they are much more realistic than their predecessors. This is mostly related to the enormous progress in model development that took place over the last decade, which is partly due to more sophisticated model parameterizations, but also to the general increase in computational resources, which allows for more thorough model testing and higher model resolu-

tion. Most of the current models not only perform better, they are also no longer flux corrected. Both improved performance and more physical formulation suggest that an increasing level of confidence can be placed in model-based predictions of climate. This, however, is only true to the extent that the performance of a model in simulating present mean climate is related to the ability to make reliable forecasts of long-term trends. It is hoped that these advancements will enhance the public credibility of model predictions and help to justify the development of even better models.

Given the many issues that complicate model validation, it is perhaps not too surprising that the present study has some limitations. First, we note the caveat that we were only concerned with the time-mean state of climate. Higher moments of climate, such as temporal variability, are probably equally as important for model performance, but we were unable to investigate these. Another critical point is the calculation of the performance index. For example, it is unclear how important climate variability is compared to the mean climate, exactly which is the optimum selection of climate variables, and how accurate the used validation data are. Another complicating issue is that error information contained in the selected climate variables is partly redundant. Clearly, more work is required to answer the above questions, and it is hoped that the present study will stimulate further research in the design of more robust metrics. For example, a future improved version of the index should consider possible redundancies and assign appropriate weights to errors from different climate variables. However, we do not think that our specific choices in this study affect our overall conclusion that there has been a measurable and impressive improvement in climate model performance over the past decade.

ACKNOWLEDGMENTS. We thank Anand Gnanadesikan, Karl Taylor, Peter Gleckler, Tim Garrett, and Jim Steenburgh for useful discussions and comments, Dan Tyndall for help with the figures, and Curt Covey and Steve Lambert for providing the CMIP-1 and CMIP-2 data. The comments of three anonymous reviewers, which helped to improve and clarify the paper, are also appreciated. We acknowledge the modeling groups for providing the CMIP-3 data for analysis, the Program for Climate Model Diagnosis and Intercomparison for collecting and archiving the model output, and the JSC/CLIVAR Working Group on Coupled Modeling for organizing the model data analysis activity. The multimodel data archive is supported by the Office of Science, U.S. Department of Energy. This work was supported by NSF grant ATM0532280 and by NOAA grant NA06OAR4310148.

FOR FURTHER READING

AchutaRao, K., and K. R. Sperber, 2006: ENSO simulation in coupled ocean–atmosphere models: Are the current models better? *Climate Dyn.,* **27,** 1–15.

Armstrong, R. L., M. J. Brodzik, K. Knowles, and M. Savoie, 2005: Global monthly EASE-Grid snow water equivalent climatology. National Snow and Ice Data Center. [Available online at www.nsidc.org/data/nsidc-0271.html.]

Bader, D., Ed., 2004: *An Appraisal of Coupled Climate Model Simulations.* Lawrence Livermore National Laboratory, 183 pp.

Barnett, T. P., and Coauthors, 1994: Forecasting global ENSO-related climate anomalies. *Tellus,* **46A,** 381–397.

Barnston, A. G., S. J. Mason, L. Goddard, D. G. Dewitt, and S. E. Zebiak, 2003: Multimodel ensembling in seasonal climate forecasting at IRI. *Bull. Amer. Meteor. Soc.,* **84,** 1783–1796.

Boer, G. J., and S. J. Lambert, 2001: Second order space–time climate difference statistics. *Climate Dyn.,* **17,** 213–218.

Bony, S., and J.-L. Dufresne, 2005: Marine boundary layer clouds at the heart of tropical cloud feedback uncertainties in climate models. *Geophys. Res. Lett.,* **32,** doi:10.1029/2005GL023851.

Covey, C., K. M. AchutaRao, U. Cubasch, P. Jones, S. J. Lambert, M. E. Mann, T. J. Phillips, and K. E. Taylor, 2003: An overview of results from the Coupled Model Intercomparison Project (CMIP). *Global Planet. Change,* **37,** 103–133.

Gates, W., U. Cubasch, G. Meehl, J. Mitchell, and R. Stouffer, 1993: An intercomparison of selected features of the control climates simulated by coupled ocean–atmosphere general circulation models. World Climate Research Programme WCRP-82 WMO/TD No. 574, World Meteorological Organization, 46 pp.

Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer, 2005: The rationale behind the success of multimodel ensembles in seasonal forecasting. I. Basic concept. *Tellus,* **57A,** 219–233, doi:10.1111/j.1600-0870.2005.00103.x.

Hewitt, C. D., 2005: The ENSEMBLES Project: Providing ensemble-based predictions of climate changes

and their impacts. *EGGS Newsletter,* **13,** 22–25.

IPCC, 2007: Climate Change 2007: The Physical Science Basis—Summary for Policymakers. 21 pp.

Jones, P. D., M. New, D. E. Parker, S. Martin, and I. G. Rigor, 1999: Surface air temperature and its changes over the past 150 years. *Rev. Geophys.,* **37,** 173–199.

Jones, R., 2005: Senate hearing demonstrates wide disagreement about climate change. FYI Number 142, American Institute of Physics. [Available online at www.aip.org/fyi/2005/142.html.]

Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.,* **77,** 437–471.

Krishnamurti, T. N., A. Chakraborty, R. Krishnamurti, W. K. Dewar, and C. A. Clayson, 2006: Seasonal prediction of sea surface temperature anomalies using a suite of 13 coupled atmosphere–ocean models. *J. Climate,* **19,** 6069–6088.

Lahsen, M., 2005: Seductive simulations? Uncertainty distribution around climate models. *Social Stud. Sci.,* 895–922.

Lambert, S. J., and G. J. Boer, 2001: CMIP1 evaluation and intercomparison of coupled climate models. *Climate Dyn.,* **17,** 83–116.

Levitus, S., T. P. Boyer, M. E. Conkright, J. A. T. O' Brien, L. S. C. Stephens, D. Johnson, and R. Gelfeld, 1998: *NOAA Atlas NESDIS 18, World Ocean Database 1998.* Vol. 1: Introduction, U.S. Government Printing Office, 346 pp.

Lin, J.-L., and Coauthors, 2006: Tropical intraseasonal variability in 14 IPCC AR4 climate models. Part I: Convective signals. *J. Climate,* **19,** 2665–2690.

Lindzen, R., 2006: Climate of fear: Global-warming alarmists intimidate dissenting scientists into silence. *Wall Street Journal.* 12 April. [Available online at www.opinionjournal.com/extra/?id=110008220.]

McAvaney, B. J., and Coauthors, 2001: Model evaluation. *Climate Change 2001: The Scientific Basis,* J. T. Houghton et al., Eds., Cambridge Univ. Press, 471–523.

Mechoso, C. R., and Coauthors, 1995: The seasonal cycle over the tropical pacific in coupled ocean–atmosphere general circulation models. *Mon. Wea. Rev.,* **123,** 2825–2838.

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

——, C. Covey, B. McAvaney, M. Latif, and R. J. Stouffer, 2005: Overview of the coupled model intercomparison project. *Bull. Amer. Meteor. Soc.,* **86,** 89–93.

Min, S.-K., and A. Hense, 2006: A Bayesian approach to climate model evaluation and multi-model averaging with an application to global mean surface temperatures from IPCC AR4 coupled climate models. *Geophys. Res. Lett.,* **33,** doi:10.1029/2006GL025779.

Mo, K. C., J.-K. Schemm, H. M. H. Juang, R. W. Higgins, and Y. Song, 2005: Impact of model resolution on the prediction of summer precipitation over the United States and Mexico. *J. Climate,* **18,** 3910–3927.

Mullen, S. L., and R. Buizza, 2002: The impact of horizontal resolution and ensemble size on probabilistic forecasts of precipitation by the ECMWF ensemble prediction system. *Wea. Forecasting,* **17,** 173–191.

Murphy, J. M., D. M. H. Sexton, D. N. Barnett, G. S. Jones, M. J. Webb, M. Collins, and D. A. Stainforth, 2004: Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature,* **430,** 768–772.

Palmer, T. N., and Coauthors, 2004: Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER). *Bull. Amer. Meteor. Soc.,* **85,** 853–872.

Parker, D. E., C. K. Folland, A. Bevan, M. N. Ward, M. Jackson, and K. Maskell, 1995: Marine surface data for analysis of climate fluctuations on interannual to century timescale. *Natural Climate Variability on Decade-to-Century Time Scales,* Climate Research Committee and National Research Council. National Academies Press, 241–250.

PCMDI, 2007: IPCC Model Ouput. [Available online at www-pcmdi.llnl.gov/ipcc/about_ipcc.php.

Randall, D. A., and Coauthors, 2007: Climate models and their evaluation. *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 et al., Eds., Cambridge University Press, 589–662.

Reichler, T., and J. Kim, 2008: Uncertainties in the climate mean state of global observations, reanalyses, and the GFDL climate model. *J. Geophys. Res.,* in press.

Roeckner, E., and Coauthors, 2006: Sensitivity of simulated climate to horizontal and vertical resolution in the ECHAM5 Atmosphere Model. *J. Climate,* **19,** 3771–3791.

Schmidt, G. A., D. T. Shindell, R. L. Miller, M. E. Mann, and D. Rind, 2004: General circulation modelling of Holocene climate variability. *Quaternary Sci. Rev.,* **23,** 2167–2181.

Simmons, A. J., and J. K. Gibson, 2000: *The ERA-40 Project Plan.* ERA-40 Project Rep., Series No. 1, 62 pp.

Singer, S. F., 1999: Human contribution to climate change remains questionable. *Eos Trans. AGU,* **80(16),** 183–187.

Stainforth, D. A., and Coauthors, 2005: Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature,* **433,** 403–406.

Stenchikov, G., A. Robock, V. Ramaswamy, M. D. Schwarzkopf, K. Hamilton, and S. Ramachandran, 2002: Arctic Oscillation response to the 1991 Mount Pinatubo eruption: Effects of volcanic aerosols and ozone depletion. *J. Geophys. Res.,* **107** (D24), doi:10.1029/2002JD002090.

Sun, D.-Z., and Coauthors, 2006: Radiative and dynamical feedbacks over the equatorial cold tongue: Results from nine atmospheric GCMs. *J. Climate,* **19,** 4059–4074.

Taylor, K. E., 2001: Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.,* **106** (D7), 7183–7192, doi:10.1029/2000JD900719.

——, P. J. Gleckler, and C. Doutriaux, 2004: Tracking changes in the performance of AMIP models. *Proc. AMIP2 Workshop,* Toulouse, France, Meteo-France, 5–8.

van Oldenborgh, G. J., S. Y. Philip, and M. Collins, 2005: El Niño in a changing climate: A multi-model study. *Ocean Sci.,* **1,** 81–95.

Williamson, D. L., 1995: Skill scores from the AMIP simulations. *First Int. AMIP Scientific Conf.,* Monterey, CA, World Meteorological Organization, 253–256.

Woodruff, S. D., R. J. Slutz, R. L. Jenne, and P. M. Steurer, 1987: A comprehensive ocean–atmosphere data set. *Bull. Amer. Meteor. Soc.,* **68,** 1239–1250.

Xie, P. P., and P. A. Arkin, 1998: Global monthly precipitation estimates from satellite-observed outgoing longwave radiation. *J. Climate,* **11,** 137–164.

Yu, L., R. A. Weller, and B. Sun, 2004: Improving latent and sensible heat flux estimates for the Atlantic Ocean (1988–1999) by a synthesis approach. *J. Climate,* **17,** 373–393.

Zhang, Y., W. B. Rossow, A. A. Lacis, V. Oinas, and M. I. Mishchenko, 2004: Calculation of radiative fluxes from the surface to top of atmosphere based on ISCCP and other global data sets: Refinements of the radiative transfer model and the input data. *J. Geophys. Res.,* **109,** D19105, doi:10.1029/2003JD004457.