Reliability of the CMIP3 ensemble

J. D. Annan[1] and J. C. Hargreaves[1]

Received 1 December 2009; revised 22 December 2009; accepted 24 December 2009; published 20 January 2010.

[1] We consider paradigms for interpretation and analysis of the CMIP3 ensemble of climate model simulations. The dominant paradigm in climate science, of an ensemble sampled from a distribution centred on the truth, is contrasted with the paradigm of a statistically indistinguishable ensemble, which has been more commonly adopted in other fields. This latter interpretation (which gives rise to a natural probabilistic interpretation of ensemble output) leads to new insights about the evaluation of ensemble performance. Using the well-known rank histogram method of analysis, we find that the CMIP3 ensemble generally provides a rather good sample under the statistically indistinguishable paradigm, although it appears marginally over-dispersive and exhibits some modest biases. These results contrast strongly with the incompatibility of the ensemble with the truth-centred paradigm. Thus, our analysis provides for the first time a sound theoretical foundation, with empirical support, for the probabilistic use of multi-model ensembles in climate research. **Citation:** Annan, J. D., and J. C. Hargreaves (2010), Reliability of the CMIP3 ensemble, *Geophys. Res. Lett.*, *37*, L02703, doi:10.1029/2009GL041994.

1. Introduction

[2] The World Climate Research Programme's Coupled Model Intercomparison Project phase 3 (CMIP3) multi-model dataset of more than 20 global climate models developed around the world has proved to be a valuable resource which has motivated and enabled much research. Since the ensemble was not generated though a coordinated attempt to sample a specific distribution but is instead an "ensemble of opportunity" (that is, a collection of model outputs based solely on availability), there has been extensive discussion on how best to interpret its construction and analyse its outputs, in order to make predictions of future climate change [e.g., *Tebaldi and Knutti*, 2007, and references therein].

[3] One paradigm which has formed the basis for many analyses is to consider that the models are "random samples from a distribution of possible models centered around the true climate" [e.g., *Jun et al.*, 2008; *Tebaldi and Knutti*, 2007]. This truth-centred paradigm appears to have arisen as a post-hoc interpretation of the ad-hoc weighting procedure known as "Reliability Ensemble Averaging" or REA [*Giorgi and Mearns*, 2002; *Nychka and Tebaldi*, 2003], and it has since been widely adopted [e.g., *Tebaldi et al.*, 2005; *Smith et al.*, 2009]. When the outputs from the CMIP3 ensemble are analysed, however, they are found to not possess the statistical properties that would be expected of such a sampling distribution. For example, if the models are indeed sampled from a distribution centered on the truth, then the biases of different models should have, on average, near-zero pairwise correlations. In practice, however, the correlations are for the most part strongly positive [*Jun et al.*, 2008]. An immediate consequence of this is that, as models are added to an ensemble, the multi-model mean does not converge to observations as rapidly as would be expected for independent biases [*Knutti et al.*, 2010]. These disquieting results have lead to claims that the model spread is "likely too narrow" [*Knutti et al.*, 2010] and limit the confidence that we can have in analyses and results that are based on this underlying interpretation. Various adaptations and corrections have been proposed to adjust for these problems [e.g., *Jun et al.*, 2008; S. Jewson and E. Hawkins, CMIP3 ensemble spread, model similarity, and climate prediction uncertainty, 2009, available at http://arxiv1.library.cornell.edu/abs/0909.1890], but there is no clear consensus on the way forward.

[4] In this paper, we reconsider paradigms for ensemble generation and interpretation. In the following section, we contrast the truth-centred paradigm described above, with an alternative interpretation which is well-established in numerical weather prediction and other fields: that of an exchangeable or *statistically indistinguishable* ensemble, that is, one where the truth is drawn from the same distribution as the ensemble members, and thus no statistical test can reliably distinguish one from the other. A simple method of evaluating ensemble performance under this interpretation is presented. We contrast the two paradigms through the analysis of some idealised examples in section 3. In section 4 we evaluate outputs from the CMIP3 database in the context of the statistically indistinguishable paradigm. We discuss the implications of our findings and conclude with some suggestions for future research directions.

2. Ensembles and Reliability

[5] The paradigm of a statistical indistinguishable ensemble has been widely adopted in numerical weather prediction [e.g., *Toth et al.*, 2003] but less commonly in climate change research [*Räisänen and Palmer*, 2001]. A fundamental distinction can easily be made between the truth-centred approach described above, and the statistically indistinguishable interpretation; in the latter case the mathematical expectation of the truth is still given by the mean of the sampling distribution, but we no longer expect the truth to be at, or even close to, this location. For example, if the ensemble members and the truth are all drawn independently from the 40-dimensional standard Normal distribution $(Z_1, \ldots, Z_{40})$, then even though the expected value of the truth is $(0, \ldots, 0)$, the probability of it being found in the

[1]Research Institute for Global Change, Yokohama, Japan.

0094-8276/10/2009GL041994$05.00

bounded box $[-2,2]^{40}$ is less than 16%, and the probability of it lying in $[-0.5,0.5]^{40}$ is 2×10^{-17}.

[6] There is a natural probabilistic interpretation of such ensembles based on simple counting arguments. When a proportion p of the ensemble members have a particular property Q (such as the temperature at a particular time and place exceeding a specified threshold) then this is interpreted as assigning probability p to the event Q. Although such a single-event probability cannot be directly validated against the observation, it is, in principle, straightforward to check over a large number of similar forecasts whether events that are predicted to occur with probability p actually do occur on a proportion p of the occasions. A system is called "reliable" if the forecast and observed frequencies do in fact agree [*Murphy*, 1973; *Toth et al.*, 2003].

[7] The rank histogram or Talagrand diagram [*Anderson*, 1996; *Talagrand et al.*, 1997] is a common method for evaluating the reliability of ensemble forecasts. It is based on the histogram of the rank of each of the observations in the ordered set formed by the union of the n predictions of the individual ensemble members together with the one observed value. If the truth and ensemble members are drawn from the same distribution, then the ranks of the observations should be uniformly distributed in $\{1, \ldots, n+1\}$. If the ensemble spread is too low, then observations will frequently lie close to or outside the edges of the ensemble, resulting in a u-shaped rank histogram. Conversely, if the ensemble spread is too broad, then the rank histogram will have a central dome. Computing the histogram of the ranks of the observations, and checking for consistency with uniformity, therefore provides a necessary condition for an ensemble prediction system to be reliable.

3. Idealised Analysis

[8] We now examine the properties of statistically indistinguishable ensembles in more detail, in particular investigating how they perform when analysed using methods developed for the truth-centred paradigm.

[9] We perform three experiments, in each of which we draw 24 ensemble members independently from the 40-dimensional standard Normal $(Z_1, \ldots, Z_{40})$. The truth for the three experiments is sampled from the three distributions $0.5 \times (Z_1, \ldots, Z_{40})$, $(Z_1, \ldots, Z_{40})$ and $2 \times (Z_1, \ldots, Z_{40})$ respectively. Thus, the second ensemble is statistically indistinguishable from the truth, and the first and third are sampled from broader, and narrower, distributions, respectively.

[10] A simple but intuitively appealing investigation into the correlation of errors across the CMIP3 ensemble has been presented by *Knutti et al.* [2010, Figure 3]. We replicate this method of analysis in Figures 1 (top) and 1 (middle), and obtain strikingly similar results. That is, (i) the errors of ensemble members are generally positively correlated, (ii) as the sample size increases, the RMSE of the ensemble mean converges to a value that is substantially greater than zero, and (iii) a few "good" ensemble members can be found, the mean of which outperforms the average of the larger set. It therefore appears that these properties are not directly informative regarding the reliability of the ensemble (in the technical sense introduced above). In fact, in the case of the statistically indistinguishable ensemble, the RMSE ensemble mean will typically converge to a value which is only a factor $1/\sqrt{2}$ smaller than the average RMSE of the individual models. However, the precision with which this ratio is attained will depend on the sampling variability not only of the ensemble, but also of the true value itself, which might by chance lie somewhat closer to or further from the ensemble mean, compared to its expected distance. Figure 1 (middle) provides some indication of these effects, in the discrepancy between the observed (solid red) and theoretically predicted (dashed black) lines. The latter was, in a minor divergence from *Knutti et al.*'s [2010] analysis, not fitted to the results of the specific samples but directly calculated from the underlying sampling distributions.

[11] In contrast to these analyses, the rank histograms contained in Figure 1 (bottom) exhibits (in order from left to right) the characteristic domed, flat, and u shapes that we should expect given the sampling distributions, and thus enable us to correctly classify the experiments. The χ^2 statistic based on the contents of the 24 bins is a common statistical test of uniformity, but it provides a rather poor test of reliability since it is insensitive to order and thus does not directly consider issues such as bias or the overall spread of the ensemble. We therefore decompose the χ^2 statistic according to the method proposed by *Jolliffe and Primo* [2008], using components to evaluate both a bias (linear trend across the histogram) and spread (approximated by a v-shape). The contributions of these two components to the total χ^2 statistic are also presented in Figure 1 (bottom). If the rank histogram was generated from a uniform distribution, both of these statistics should be distributed according to the χ^2 distribution with one degree of freedom. Therefore, we correctly find no signs of any significant biases in these experiments, but the spreads of the first and third experiments are both found to be significantly non-uniform at the $p < 1\%$ level ($\chi^2 > 6.64$).

4. Analysis of CMIP3 Models

[12] We now investigate the reliability of the CMIP3 ensemble using the rank histogram approach. We follow the approach of previous authors in evaluating the mean climatic state against modern observational data. We analyse fields of three climatic variables: surface air temperature, with data obtained from *Brohan et al.* [2006], precipitation, using the data of *Adler et al.* [2003], and sea level pressure against the data of *Allan and Ansell* [2006]. All model and observational data sets are firstly averaged onto 5 degree global grids and over the years 1961–1990 (temperature and sea level pressure) or 1979–1999 (precipitation). We only present results from annual mean values here, but using seasonal averages (DJF, JJA) or the magnitude of the seasonal cycle (JJA minus DJF) give broadly similar results.

[13] Rank histograms of the three sets of observations are shown in Figure 2. These are calculated on an area-weighted basis, with the totals normalised to 40. Since neighbouring grid points are highly correlated, the number of effective degrees of freedom of the data (which determines the precision with which the empirical rank histograms should match the uniform distribution) is not entirely clear. Semi-variograms of the inter-model (and model minus data) differences suggest a decorrelation distance of around

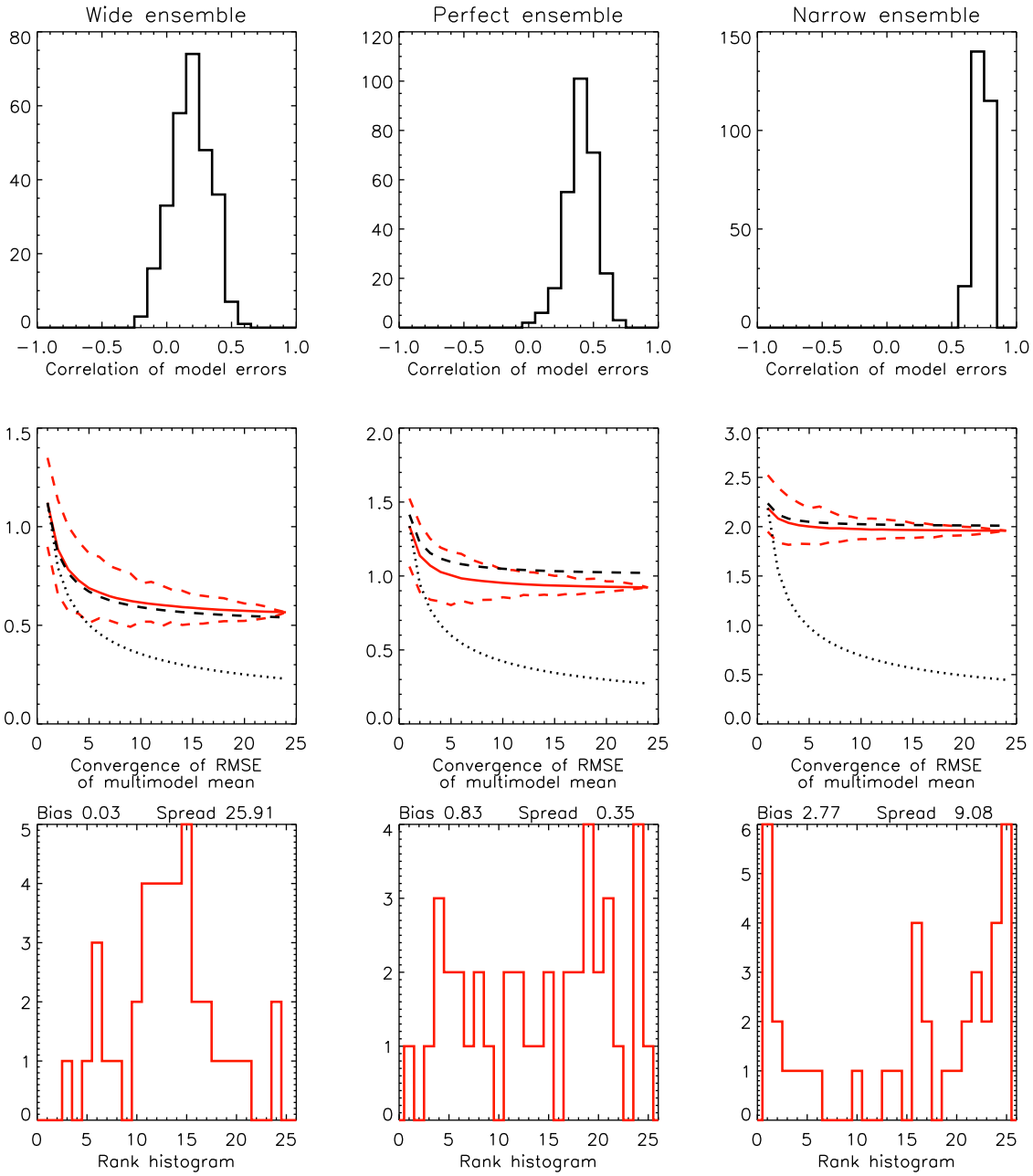

Figure 1. Analysis of three idealised ensembles. (top) Histograms of the pairwise correlation coefficients of the biases of the ensemble members. (middle) RMSE of the mean of random subsets of the ensemble members, plotted as a function of subset size. Solid red lines show the mean over repeated subsampling, and the dashed red lines give the upper and lower ranges obtained. Black dashed lines gives the theoretically-expected result of $\sqrt{1/n + \sigma^2}$ where $\sigma = 0.5, 1, 2$ is the standard deviation of the true variables, and the dotted lines show the $\sqrt{1/n}$ convergence that would be expected for a truth-centred ensemble. (bottom) Rank histograms for the 40 true values of the variables in each experiment, with contributions of bias and spread to the χ^2 statistic (see text for details).

1000–2000km. This would imply at least 40 degrees of freedom for each data field, which motivates our choice of this value both here and for the idealised tests presented in section 3. *Jolliffe and Primo* [2008] suggest using 25 degrees of freedom for a single hemisphere, which is similar to our choice. Changing the number of degrees of freedom alters the statistical significance of our results, but not their qualitative nature.

[14] The total χ^2 statistics of the rank histograms are all insignificant, but the decomposition into bias and spread components does reveal some problems, in that the modelled temperatures are biased low and the spreads of temperature and sea level pressure appear too large, relative to the observations (albeit the error in the spread of temperature does not quite reach the $p < 5\%$ significance threshold of $\chi^2 > 3.84$). We should note, however, that these errors are in fact relatively small compared to the ensemble ranges themselves. The surface temperature histogram can be effectively flattened by both subtracting a mean bias of 0.5C, and adding random noise of magnitude 1C to the data

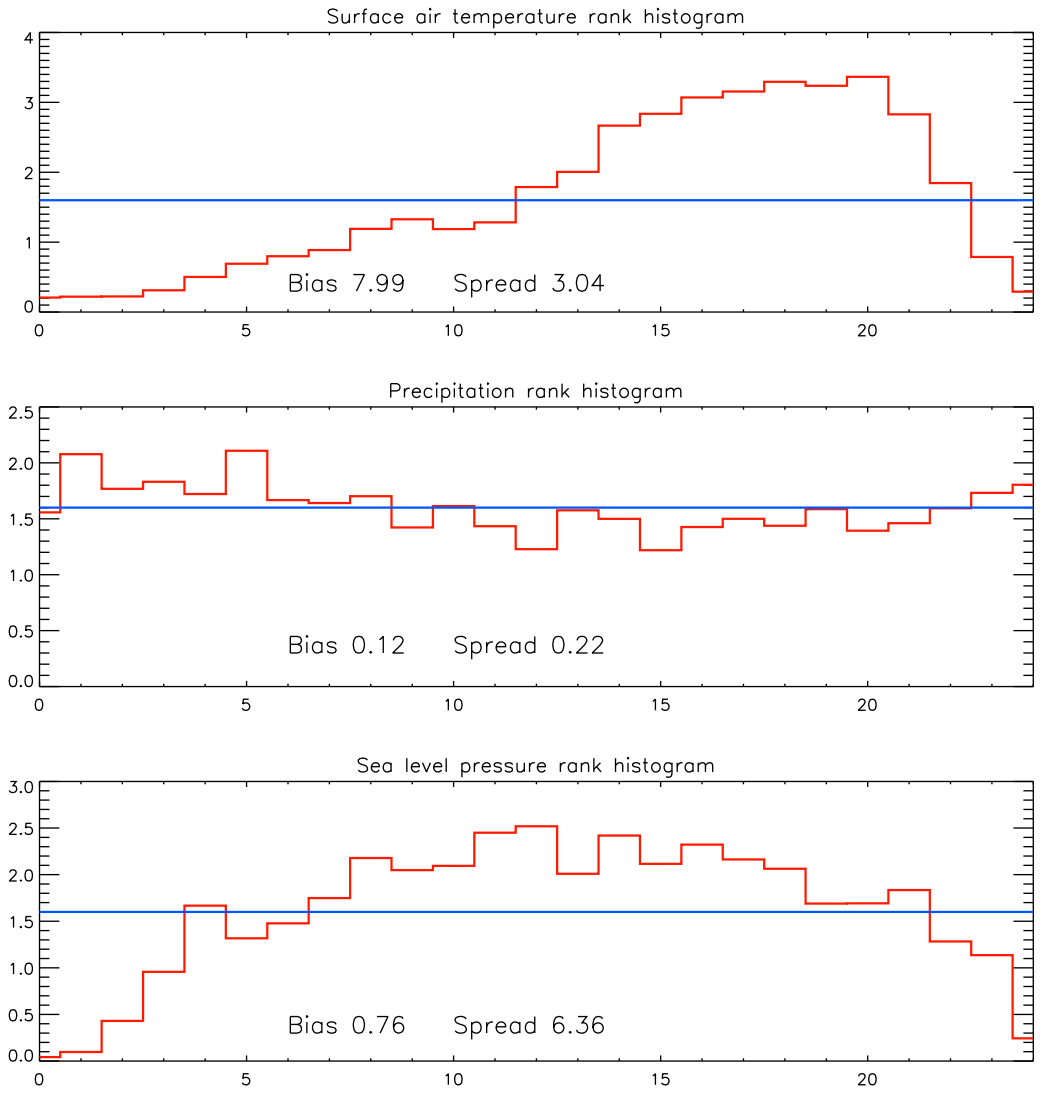

Figure 2. Rank histogram analysis of outputs of CMIP3 models versus observational data. Sea surface temperature, precipitation and sea level pressure are shown from top to bottom. χ^2 statistics for bias and spread are also presented in each subplot (see text for details).

to increase their overall spread. These figures only amount to 7% and 13% respectively of the typical ensemble range of 7.7C at each gridpoint. The sea level pressure histogram can be flattened by adding random noise of magnitude 1hPa to the data, which is less than 8% of the average ensemble spread at each gridpoint. It is also worth noting that this analysis finds the ensemble width to be generally too wide, rather than too narrow as has been previously claimed. This suggests that the direct probabilistic interpretation of the ensemble will err on the side of caution, rather than having a high probability of excluding the truth.

[15] We have also performed a similar analysis for each model in turn, generating its rank histogram among the remaining 23 models. The χ^2 tests for bias and spread frequently fail (eg in 30% of cases at the $p < 5\%$ level), with a similar degree of non-uniformity as found for the observational data. If we assume 5 degrees of freedom rather than 40, the failure rate is reduced to around 5% at the 5% level, and the divergence from uniformity noted for the rank histogram of observational data is no longer statistically significant. Thus, the models appear to be about as reliable at predicting reality as they are at predicting each other, further supporting the hypothesis of exchangeability between the models and true climate system.

5. Discussion

[16] The CMIP3 ensemble has arisen through a process of large numbers of researchers making numerous diverse decisions according to their beliefs about the climate system. It should hardly be surprising that these beliefs are biased in their mean, and that the resulting ensemble of models is not centred on the truth. So long as the range of choices made is commensurate with the errors that exist, however, this in no way precludes the ensemble members from forming a sample which is statistically indistinguishable from the truth.

[17] Given the various contingencies relating to the creation of the CMIP3 ensemble of opportunity, it would be unreasonably optimistic to expect it to be perfectly

reliable in all respects. However, our analysis here suggests, at least for the data considered here, that this assumption is in fact not far from the truth, although there are weak indications that the model spread may be a little too broad.

[18] An important issue, that we do not address here, is the relationship between past and future performance [e.g., *Whetton et al.*, 2007; *Abe et al.*, 2009]. Arguably, the CMIP3 climate models have already been tuned to some extent to the recent climate data that we have used here. It is not immediately clear that this should affect the reliability of the ensemble, as this would require not just that the biases on individual models are reduced, but that the biases change sign, and do so preferentially in one direction. However, this issue is worthy of further investigation. Additionally, it would be interesting to test further the reliability of the ensemble in other ways, for example considering simulations of other epochs, or other climatic observations that are less widely used during model construction and tuning.

[19] The analysis presented here implicitly assigns equal weight to each ensemble member, which would be appropriate if we believe them to be equally good models of the climate system. Such an approach is normal for single-model ensembles generated by perturbing initial conditions, but may not be so sensible when samples arise heterogeneously, as is the case with the CMIP3 ensemble. Therefore, the quality of probabilistic predictions in terms of both reliability and resolution may be improved by some non-uniform weighting. A variety of approaches are in common use, including both heuristic re-weighting methods [*Zhu et al.*, 1996; *Krishnamurti et al.*, 2000] and formal Bayesian methods such as Bayesian Model Averaging [*Raftery et al.*, 2005]. The paradigm of a statistically-indistinguishable ensemble provides an appropriate theoretical foundation for the exploration of these ideas. There is also a wide range of established analysis techniques which may be applicable for evaluating ensemble performance [*Toth et al.*, 2003].

6. Conclusions

[20] An ensemble which is statistically indistinguishable from the truth will appear to be biased and non-convergent when analysed under the assumption that it is a truth-centred ensemble. We have shown that the CMIP3 ensemble appears fairly reliable when tested against recent observations, and if anything tends towards being over-broad, in contrast to recent claims. Thus, our analysis supports the direct probabilistic interpretation of the ensemble, although we expect its reliability (and resolution) could be further improved by non-uniform weighting. We suggest that in place of the truth-centred approach, future research into the use of the CMIP3 and other multi-model ensembles of opportunity should be based on the paradigm of a statistically indistinguishable ensemble, as this is both intuitively plausible and reasonably compatible with observational evidence.

[21] **Acknowledgments.** We are grateful to two reviewers for helpful comments. This work was supported by the S-5-1 project of the MoE, Japan and by the Kakushin Program of MEXT, Japan. We acknowledge the modeling groups, the Program for Climate Model Diagnosis and Intercomparison (PCMDI) and the WCRP's Working Group on Coupled Modelling (WGCM) for their roles in making available the WCRP CMIP3 multi-model dataset. Support of this dataset is provided by the Office of Science, U.S. Department of Energy.

References

Abe, M., H. Shiogama, J. Hargreaves, J. Annan, T. Nozawa, and S. Emori (2009), Correlation between inter-model similarities in spatial pattern for present and projected future mean climate, *SOLA*, *5*, 133–136.

Adler, R., et al. (2003), The version-2 Global Precipitation Climatology Project (GPCP) monthly precipitation analysis (1979–present), *J. Hydrometeorol.*, *4*(6), 1147–1167.

Allan, R., and T. Ansell (2006), A new globally complete monthly historical gridded mean sea level pressure dataset (HadSLP2): 1850–2004, *J. Clim.*, *19*(22), 5816–5842.

Anderson, J. (1996), A method for producing and evaluating probabilistic forecasts from ensemble model integrations, *J. Clim.*, *9*(7), 1518–1530.

Brohan, P., J. J. Kennedy, I. Harris, S. F. B. Tett, and P. D. Jones (2006), Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850, *J. Geophys. Res.*, *111*, D12106, doi:10.1029/2005JD006548.

Giorgi, F., and L. Mearns (2002), Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the "reliability ensemble averaging" (REA) method, *J. Clim.*, *15*(10), 1141–1158.

Jolliffe, I., and C. Primo (2008), Evaluating rank histograms using decompositions of the chi-square test statistic, *Mon. Weather Rev.*, *136*(6), 2133–2139.

Jun, M., R. Knutti, and D. Nychka (2008), Spatial analysis to quantify numerical model bias and dependence: How many climate models are there?, *J. Am. Stat. Assoc.*, *103*(483), 934–947.

Knutti, R., R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl (2010), Challenges in combining projections from multiple climate models, *J. Clim.*, in press.

Krishnamurti, T., C. Kishtawal, Z. Zhang, T. LaRow, D. Bachiochi, E. Williford, S. Gadgil, and S. Surendran (2000), Multimodel ensemble forecasts for weather and seasonal climate, *J. Clim.*, *13*(23), 4196–4216.

Murphy, A. (1973), A new vector partition of the probability score, *J. Appl. Meteorol.*, *12*(4), 595–600.

Nychka, D., and C. Tebaldi (2003), Comments on "Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the 'reliability ensemble averaging' (REA) method", *J. Clim.*, *16*(5), 883–884.

Raftery, A., T. Gneiting, F. Balabdaoui, and M. Polakowski (2005), Using Bayesian model averaging to calibrate forecast ensembles, *Mon. Weather Rev.*, *133*(5), 1155–1174.

Räisänen, J., and T. Palmer (2001), A probability and decision-model analysis of a multimodel ensemble of climate change simulations, *J. Clim.*, *14*(15), 3212–3226.

Smith, R., C. Tebaldi, D. Nychka, and L. Mearns (2009), Bayesian modeling of uncertainty in ensembles of climate models, *J. Am. Stat. Assoc.*, *104*(485), 97–116.

Talagrand, O., R. Vautard, and B. Strauss (1997), Evaluation of probabilistic prediction systems, paper presented at the Workshop on Predictability, Eur. Cent. for Medium-Range Weather Forecasts, Reading, U. K.

Tebaldi, C., and R. Knutti (2007), The use of the multi-model ensemble in probabilistic climate projections, *Philos. Trans. R. Soc., Ser. A*, *365*(1857), 2053.

Tebaldi, C., R. Smith, D. Nychka, and L. Mearns (2005), Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multimodel ensembles, *J. Clim.*, *18*(10), 1524–1540.

Toth, Z., O. Talagrand, G. Candille, and Y. Zhu (2003), Probability and ensemble forecasts, in *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, edited by I. T. Jolliffe and D. B. Stephenson, pp. 137–163, John Wiley, Chichester, U. K.

Whetton, P., I. Macadam, J. Bathols, and J. O'Grady (2007), Assessment of the use of current climate patterns to evaluate regional enhanced greenhouse response patterns of climate models, *Geophys. Res. Lett.*, *34*, L14701, doi:10.1029/2007GL030025.

Zhu, Y., G. Iyengar, Z. Toth, M. Tracton, and T. Marchok (1996), Objective evaluation of the NCEP global ensemble forecasting system, in *Proceedings of the 15th Conference on Weather Analysis and Forecasting*, vol. 15, pp. 79–82, Am. Meteorol. Soc., Boston, Mass.

J. D. Annan and J. C. Hargreaves, Research Institute for Global Change, 3073-25 Showamachi, Yokohama, Kanagawa 236-0001, Japan. (jdannan@jamstec.go.jp; jules@jamstec.go.jp)