Application of Cluster Analysis to Climate Model Performance Metrics

SATORU YOKOI,* YUKARI N. TAKAYABU,* KAZUAKI NISHII,[+] HISASHI NAKAMURA,[+,#] HIROKAZU ENDO,[@] HIROKI ICHIKAWA,[&] TOMOSHIGE INOUE,[**] MASAHIDE KIMOTO,* YU KOSAKA,[+] TAKAFUMI MIYASAKA,[+] KAZUHIRO OSHIMA,[++] NAOKI SATO,[##,#] YOKO TSUSHIMA,[#] AND MASAHIRO WATANABE*

[*] *Atmosphere and Ocean Research Institute, The University of Tokyo, Kashiwa, Japan*
[+] *Graduate School of Science, The University of Tokyo, Tokyo, Japan*
[#] *Research Institute for Global Change, Japan Agency for Marine-Earth Science and Technology, Yokohama, Japan*
[@] *Meteorological Research Institute, Tsukuba, Japan*
[&] *Graduate School of Environmental Studies, Nagoya University, Nagoya, Japan*
[**] *Graduate School of Life and Environmental Sciences, University of Tsukuba, Tsukuba, Japan*
[++] *Faculty of Environmental Earth Science, Hokkaido University, Sapporo, Japan*
[##] *Tokyo Gakugei University, Tokyo, Japan*

(Manuscript received 31 August 2010, in final form 8 March 2011)

ABSTRACT

The overall performance of general circulation models is often investigated on the basis of the synthesis of a number of scalar performance metrics of individual models that measure the reproducibility of diverse aspects of the climate. Because of physical and dynamic constraints governing the climate, a model's performance in simulating a certain aspect of the climate is sometimes related closely to that in simulating another aspect, which results in significant intermodel correlation between performance metrics. Numerous metrics and intermodel correlations may cause a problem in understanding the evaluation and synthesizing the metrics. One possible way to alleviate this problem is to group the correlated metrics beforehand. This study attempts to use simple cluster analysis to group 43 performance metrics. Two clustering methods, the *K*-means and the Ward methods, yield considerably similar clustering results, and several aspects of the results are found to be physically and dynamically reasonable. Furthermore, the intermodel correlation between the cluster averages is considerably lower than that between the metrics. These results suggest that the cluster analysis is helpful in obtaining the appropriate grouping. Applications of the clustering results are also discussed.

1. Introduction

In recent years, the importance of general circulation models (GCMs) for climate science and future projection has been increasingly acknowledged. For instance, more than 20 GCMs participated in phase 3 of the Climate Model Intercomparison Project (CMIP3; Meehl et al. 2007a) to contribute to the Intergovernmental Panel for Climate Change (IPCC) Fourth Assessment Report. The archived output data from these GCMs are recognized as being critically important for multimodel-based analyses of the current state of the climate and its future projection (Meehl et al. 2007b).

One approach to synthesize outputs of various GCMs is a democratic ''one model–one vote'' framework, whereas another approach is weighting or screening GCMs depending on their performances (e.g., Santer et al. 2009; Knutti et al. 2010). In the latter, determination of the weights or screening criteria is a major issue. One way of determination is based on the evaluation of the overall performance of GCMs in simulating diverse aspects of the climate relevant to the purpose of the weighting or screening. Several studies have examined the overall performance (Murphy et al. 2004; Gleckler et al. 2008; Pincus et al. 2008; Reichler and Kim 2008). Murphy et al. (2004) introduced a climate prediction index on the basis of the reproducibility of the climatological mean fields of 32 components of surface and atmospheric variables for use as a weighting function of GCMs for climate prediction. Gleckler et al. (2008) evaluated the reproducibility of the global

Corresponding author address: Satoru Yokoi, Atmosphere and Ocean Research Institute, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8568, Japan.
E-mail: yokoi@aori.u-tokyo.ac.jp

DOI: 10.1175/2011JAMC2643.1

distribution of 22 variables, including atmospheric temperature, wind, sea surface temperature, and ocean surface heat and momentum fluxes, by using root-mean-square errors (RMSE) between the simulations and the observations as the metrics to evaluate the model performance. Pincus et al. (2008) evaluated the model reproducibility of atmospheric variables related to clouds, precipitation, and radiation.

Because of the complexity of the climate, numerous performance metrics that measure individual aspects of the climate can be defined. Because even GCMs that exhibit superior overall performance have some weaknesses in reproducing certain aspects of the climate (Gleckler et al. 2008), we tend to end up with handling numerous performance metrics to evaluate the overall performance. It is complicated to understand the details of the GCM performance when examining such a wide range of performance metrics.

The overall performance metric defined for the weighting or screening of GCMs may be a weighted average of the individual metrics. For instance, Gleckler et al. (2008) experimentally presented an overall metric that is a simple average of the individual metrics they examined. For these types of overall metric, we have to determine the weights for numerous metrics, which is also a complicated task.

Incorporating numerous metrics may cause another problem. If more metrics are incorporated in the evaluation, it is more likely that some will exhibit significant positive intermodel correlation reflecting physical and dynamic constraints, as briefly mentioned in Gleckler et al. (2008). For instance, the geostrophic balance tends to relate errors between horizontal wind and geopotential height; GCMs with smaller horizontal wind errors tend to exhibit smaller geopotential height errors.

Such strong intermodel relationships may further complicate the determination of the weights. Suppose we examine three metrics, and two of them exhibit a significant positive correlation and apparently represent the same aspect of the climate, whereas the other represents a different aspect. If we want to incorporate the performances in simulating the two aspects in an overall metric with equal weight, the weight of the last metric should be twice as large as the former. Unlike this simple example, because we should examine numerous metrics and aspects, it is much more complicated to completely consider the relationships among the metrics for determining of the weights.

Sometimes statistical approaches are used to synthesize the performance metrics. The strong relationships may make it difficult to apply statistical approaches, however, because some require statistical independence among the metrics. For instance, we may perform a multiple linear regression analysis between trends of a certain climatological aspect projected by GCMs and performance metrics to obtain reliable projection. This method assumes the explaining variables (performance metrics in this case) to be independent of each other, however, so it will be difficult to appropriately determine regression coefficients, which is the so-called multicollinearity problem.

We consider grouping such correlated performance metrics beforehand to be one solution for addressing these problems. By calculating averages of the metrics in individual groups and using them as a new set of metrics instead of the original ones, the number of metrics to be considered will be reduced and the intermodel correlation among the metrics will be generally less significant. This will simplify the model evaluation and the understanding of the model performance. In this study, we attempt to use simple cluster analysis to demonstrate its applicability in obtaining the appropriate grouping.

The rest of the paper is organized as follows: section 2 explains the GCMs, variables, performance metrics, and cluster analysis considered in this study. Section 3 demonstrates successful grouping of the performance metrics by cluster analysis. Section 4 discusses the application of the clustering results. Section 5 briefly summarizes this study.

2. Data and performance metrics

a. Data

To demonstrate the significant intermodel correlations between the performance metrics and the applicability of the cluster analysis, we examine a set of performance metrics for 3-month mean climatological fields in the boreal summer (June–August) of 14 variables (Table 1). We evaluate the performance of 22 GCMs that participated in CMIP3. The GCMs include the Bjerknes Centre for Climate Research Bergen Climate Model, version 2.0 (BCCR-BCM2.0); National Center for Atmospheric Research Community Climate System Model, version 3 (CCSM3); Canadian Centre for Climate Modelling and Analysis (CCCma) Coupled General Circulation Model, version 3.1, with spectral triangular truncation at wave number 47 [CGCM3.1(T47)]; CGCM3.1(T63); Centre National de Recherches Météorologiques Coupled Global Climate Model, version 3 (CNRM-CM3); Commonwealth Scientific and Industrial Research Organisation, Mark version 3.0 (CSIRO Mk3.0); CSIRO Mk3.5; ECHAM5/Max Planck Institute Ocean Model (MPI-OM); Flexible Global Ocean–Atmosphere–Land System Model, gridpoint version 1.0 (FGOALS-g1.0); Geophysical Fluid Dynamics Laboratory Climate Model, version 2.0 (GFDL CM2.0); GFDL CM2.1; Goddard Institute for Space Studies Atmosphere–Ocean Model

TABLE 1. Variables used for evaluation and reference datasets.

Variable	Description	Reference	Year
U85, U50, U20	850-, 500-, 200-hPa zonal wind	ERA-40 (Uppala et al. 2005)	1979–99
V85, V50, V20	850-, 500-, 200-hPa meridional wind	ERA-40 (Uppala et al. 2005)	1979–99
T85, T50, T20	850-, 500-, 200-hPa temperature	ERA-40 (Uppala et al. 2005)	1979–99
Z85, Z50, Z20	850-, 500-, 200-hPa geopotential height	ERA-40 (Uppala et al. 2005)	1979–99
Q85, Q50, Q30	850-, 500-, 300-hPa specific humidity	ERA-40 (Uppala et al. 2005)	1979–99
Tsf	Surface temperature	ERA-40 (Uppala et al. 2005)	1979–99
SLP	Sea level pressure	ERA-40 (Uppala et al. 2005)	1979–99
OLR	Outgoing longwave radiation	Clouds and the Earth's Radiant Energy System (CERES; Wielicki et al. 1996)	2001–04
OSR	Reflected shortwave radiation at the top of the atmosphere	CERES (Wielicki et al. 1996)	2001–04
CLD	Cloud fraction	International Satellite Cloud Climatology Project (ISCCP)-D2 (Rossow and Schiffer 1999)	1984–99
LHF	Surface latent heat flux	Southampton Oceanography Centre (SOC; Josey et al. 1999)	1980–93
SHF	Surface sensible heat flux	SOC (Josey et al. 1999)	1980–93
PRC	Precipitation	Global Precipitation Climatology Project (GPCP; Adler et al. 2003)	1979–99
SST	Sea surface temperature	Hadley Centre Sea Ice and Sea Surface Temperature dataset (HadISST; Rayner et al. 2006)	1979–99

(GISS-AOM); GISS Model E-H (GISS-EH); GISS Model E-R (GISS-ER); Istituto Nazionale di Geofisica e Vulcanologia global climate model, version SXG (INGV-SXG); Institute of Numerical Mathematics Coupled Model, version 3.0 (INM-CM3.0); L'Institut Pierre-Simon Laplace Coupled Model, version 4 (IPSL CM4); Model for Interdisciplinary Research on Climate 3, high-resolution version (MIROC3hi); MIROC 3, medium-resolution version (MIROC3med); Meteorological Research Institute Coupled General Circulation Model, version 2.3.2 (MRI CGCM2.3.2); Parallel Climate Model (PCM); and the third climate configuration of the Met Office Unified Model (UKMO HadCM3). The simulated climatological fields of the variables are calculated from monthly-mean outputs of twentieth-century climate simulations (Meehl et al. 2007a) for the period of 1979–99, archived in the Program for Climate Model Diagnosis and Intercomparison database.

The simulated fields are compared with the reference datasets, which are also listed in Table 1. Among 14 variables, seven variables of zonal (U) and meridional (V) winds, temperature T, geopotential height Z, specific humidity Q, surface temperature Tsf, and sea level pressure (SLP) fields are obtained from the 40-yr European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis (ERA-40) dataset. The other seven variables are obtained from products retrieved from satellite datasets. Note that U, V, T, Z, and Q are evaluated separately in the lower (850-hPa level), middle (500-hPa level), and upper (300-hPa level for Q and 200-hPa level for the others) troposphere. These pressure levels are symbolized in a manner such that, for example, ''20'' stands for the 200-hPa level. Thus, 200-hPa zonal wind is symbolized as ''U20.''

Because of the data availability, reference climatological fields of several variables are calculated for different periods from the model evaluation period 1979–99. In particular, outgoing longwave radiation (OLR) and reflected shortwave radiation (OSR) at the top of the atmosphere have a period of only four years (2001–04) that does not overlap the model evaluation period. Thus, it seems inappropriate to regard the 4-year average as the climatological field because of the possible existence of natural variability with interannual to decadal time scales. When we use averages over 4-yr periods (such as 1979–82 and 1983–86) as simulated climatological fields to examine the sensitivity of the performance metrics of OLR and OSR to a selection of target periods, however, the differences in the metrics among the different periods for each model are considerably smaller than those among the GCMs (figure not shown). This suggests that simulated natural variability has little impact on the intermodel comparison. Although characteristics in simulated natural variability may be different from those in the real atmosphere, this result implies that the inconsistency and the shortness of the period do not significantly affect our main results.

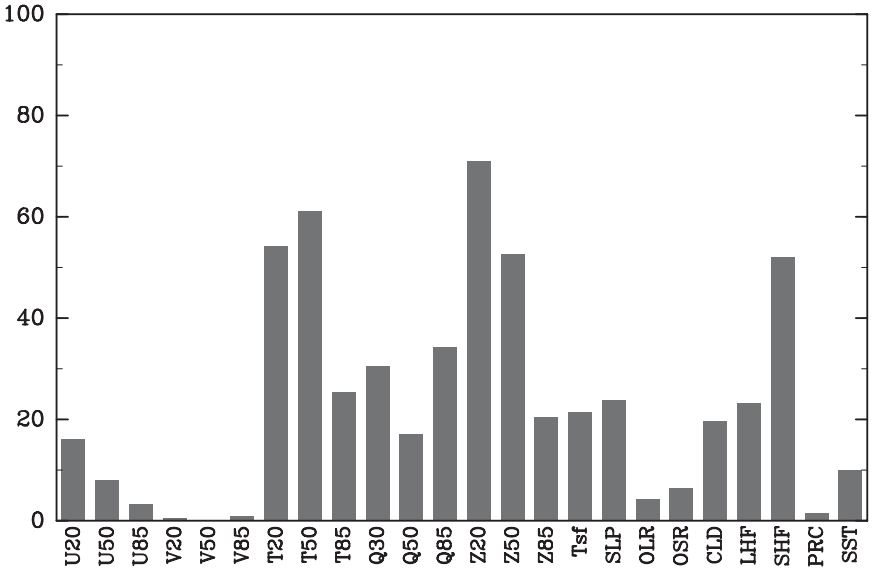

FIG. 1. Intermodel average of the ratio of the squared mean bias to MSE for each variable.

For the polar regions (south of 60°S or north of 70°N), we decide to use only the satellite datasets for evaluation and not to use the reanalysis dataset because of the lack of sufficient observational data assimilated in the reanalysis procedure. The evaluation of SLP is further limited to areas for which the surface altitude is below 1000 m. In addition, latent heat flux (LHF), sensible heat flux (SHF), and sea surface temperature (SST) data exist only over the ocean.

Because the horizontal resolutions of the GCM outputs and the reference data are different, we first interpolated all data onto 2.5° × 2.5° grids and then compared the simulated and reference data for each grid.

b. Performance metrics

Among various kinds of statistical measures that represent similarity between simulated and reference fields, we use two measures as the performance metrics: the magnitude of the global mean bias and the centered (unbiased) RMSE. The global mean bias b_m is defined as

$$b_m = \frac{1}{IJ}\sum_i \sum_j [X_m(i,j) - X_o(i,j)],$$

where X_m is the simulated climatological field of a certain variable X of the mth model; X_o is the corresponding reference field; i and j indicate longitude and latitude, respectively; and I and J indicate grid numbers in longitude and latitude, respectively. The first metric is simply the absolute value of b_m ($|b_m|$). The second metric c_m is the RMSE between the simulated and reference fields after eliminating the corresponding global mean. Its mathematical form is

$$c_m = \left\{\frac{1}{IJ}\sum_i \sum_j [X_m(i,j) - X_o(i,j) - b_m]^2\right\}^{1/2}.$$

This measure represents the similarity in the spatial pattern between the two fields. For both metrics, smaller values represent better performance. For brevity, the mean-bias metric and the centered-RMSE metric are hereinafter denoted with suffixes "b" and "c," respectively. For instance, the mean-bias metric of U20 (200-hPa zonal wind) is abbreviated as U20b.

The square sum of these two metrics is equal to the mean-square error (MSE), which is given as

$$|b_m|^2 + c_m^2 = \frac{1}{IJ}\sum_i \sum_j [X_m(i,j) - X_o(i,j)]^2 = \mathrm{MSE}_m.$$

Relative magnitudes of the mean-bias metric with respect to the centered-RMSE metric are considered to depend on the variables (Gleckler et al. 2008). If one of the two metrics explains most of the MSE, the other metric is considered to be unimportant; thus, it will be better to exclude the latter from the cluster analysis. Figure 1 shows the ratio of squared mean bias ($|b_m|^2$) to MSE averaged over all models. The values of the squared mean biases of temperature and geopotential height in the middle and upper troposphere are higher than 50% of their MSEs. On the other hand, horizontal winds, radiation at the top of the atmosphere (OLR and OSR), and precipitation have small relative magnitudes of the squared mean biases. In

particular, V20b, V50b, and V85b squared are less than 1% of their respective MSEs. Therefore, we exclude these three metrics from the analysis.

In addition, because the magnitudes of the observed surface turbulent fluxes (LHF and SHF) tend to be considerably less accurate than their horizontal patterns (Gleckler et al. 2008), we consider only the centered RMSE for these variables (LHFc and SHFc). By considering these factors, we decided to retain 43 metrics (19 mean-bias metrics and 24 centered-RMSE metrics) for our cluster analysis.

At this point, b_m and c_m have dimensions that are the same as the respective variables. Therefore, we next introduced the normalized mean-bias metric b_m^* and the normalized centered-RMSE metric c_m^* such that

$$b_m^* = (b_m - a_b)/s_b \quad \text{and}$$
$$c_m^* = (c_m - a_c)/s_c.$$

Here, a_b and a_c indicate the intermodel mean of b_m and c_m, respectively, and s_b and s_c indicate the respective intermodel standard deviation.

As discussed later, 17% of the pairs of the 43 metrics exhibit a statistically significant positive correlation. We will use the cluster analysis to group such correlated metrics.

c. Clustering algorithms

In this study, the aforementioned performance metrics are regarded as vectors with 22 dimensions, which correspond to the number of GCMs used. The similarity among the metrics used for the cluster analysis is defined as the Euclidean distance in the 22-dimensional phase space.

We apply two hard clustering methods, in which each metric belongs to only one cluster, in contrast to soft (fuzzy) clustering methods in which each metric can belong to more than one cluster. One method is a hierarchical approach of the Ward method (Ward 1963). It begins with 43 one-member clusters and stepwise merges two clusters until all the metrics belong to a single cluster. The other is a nonhierarchical approach of the *K*-means method (Forgy 1965). In this method, we first define the number of clusters and the initial grouping of the metrics, and then the method iteratively modifies the grouping to find the appropriate grouping. Note that the clustering approaches have recently been used in climate studies. For instance, Williams and Tselioudis (2007) and Williams and Webb (2009) applied the *K*-means method to joint cloud optical depth–cloud-top pressure histograms and determined cloud regimes at individual grid points.

For cluster analysis, one important issue is the determination of the number of clusters. Among the various statistical approaches proposed for assessing the relevance of the cluster number, we adopted the pseudo-*F* and pseudo-t^2 statistics (Fovell and Fovell 1993) for the Ward method and the pseudo-*F* statistic for the *K*-means method. The pseudo-*F* statistic is based on the ratio of the variance among the cluster means to that among the members within individual clusters; the statistical significance levels at which the former variance is greater than the latter are evaluated. The pseudo-t^2 statistic represents an error increment by cluster merger, and its abrupt increase suggests the number of clusters to be retained. Note that, as Fovell and Fovell (1993) stated, these statistical measures are "pseudo tests" because they violate several assumptions that underlie statistic theories. Therefore, we should use these measures with caution, although they provide useful information on the number of clusters to be retained.

For the *K*-means method, the clustering result is subject to the initial grouping. Therefore, we perform the method 100 000 times with a different initial grouping to find the most appropriate result with the maximum significance level of the pseudo-*F* statistic.

3. Results

First, we determine the number of clusters from a statistical viewpoint. Figure 2a shows the statistical significance levels of the pseudo-*F* statistic. Because the significance levels are calculated using the standard *F* statistics theory, it is inappropriate to focus on their magnitudes. Instead, we attempted to compare the significance levels among the different numbers of clusters. In both clustering methods, significance levels do not vary considerably if the total number of clusters is more than seven, whereas it falls sharply if the number is less than seven. This suggests that retaining seven clusters is statistically appropriate. The pseudo-t^2 statistic for the Ward method (Fig. 2b) supports this finding. We can find three relatively large stepwise error increases at cluster numbers 11, 7, and 5. Based on these results, we decide to retain seven clusters for both methods.

The clustering result with seven clusters based on the *K*-means method is shown in Table 2. The number of members included in a single cluster ranges from 3 to 17. The result from the Ward method is very consistent with that from the *K*-means method; the only difference is that T50c belongs to cluster B in the Ward method instead of to cluster A.

We can point out several aspects of the clustering results that are reasonable from the physical and dynamic viewpoints, although it is difficult to interpret the entire aspects. For instance, most of the centered-RMSE metrics of the horizontal winds and the mid- and

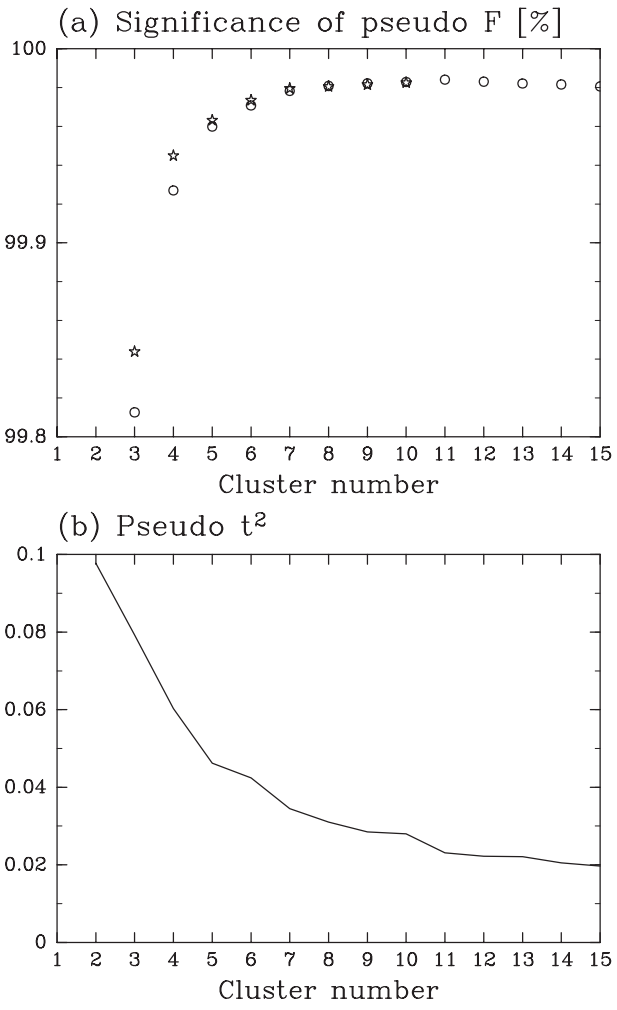

FIG. 2. (a) Statistical significance of pseudo-F statistic for the Ward (circles) and K-means (stars) methods, and (b) pseudo-t^2 statistic for the Ward method as a function of cluster number.

lower-tropospheric thermodynamic variables are included in cluster A. This feature is probably related to the geostrophic balance, mass continuity, and hydrostatic balance that link large-scale wind, pressure, and temperature fields. Centered-RMSE metrics of OLR, cloudiness, and precipitation are also grouped in cluster A, possibly representing a link between the horizontal distribution in the convection and circulation fields. In cluster C, the mean-bias metrics of temperature in the middle and lower troposphere and geopotential height in the middle and upper troposphere are grouped, which is associated with the radiative–convective equilibrium and the hydrostatic balance. In cluster E, OLRb and OSRb are grouped together with PRCb, which is related to a bias in condensation and radiative heating. The consistency between the two clustering results and the conceivable physical explanation for the results suggest that the resultant clusters are not merely a statistical artifact.

TABLE 2. Members of the seven clusters for the K-means clustering. The mean-bias metrics ($|b_m|$) and the centered-RMSE metrics (c_m) are indicated by italic and boldface type, respectively.

Cluster	Metrics
A	**U20c**, **U50c**, **U85c**, **V20c**, **V50c**, **V85c**, **T50c**, **T85c**, **Z50c**, **Z85c**, **Q50c**, **Q85c**, **Tsfc**, **SLPc**, **OLRc**, **CLDc**, **PRCc**
B	*U20b*, *U50b*, **Z20c**, *Q30b*, **Q30c**, *Q50b*, **SHFc**
C	*T50b*, *T85b*, *Z20b*, *Z50b*, *Tsfb*
D	*T20b*, **T20c**, **OSRc**, *CLDb*, **LHFc**
E	*OLRb*, *OSRb*, *PRCb*
F	*U85b*, *Z85b*, *SLPb*
G	*Q85b*, *SSTb*, **SSTc**

We also performed the K-means method for 3-month-mean fields for three other seasons (September–November, December–February, and March–May) and for the annual-mean field. For each of the four periods, retaining six or seven clusters is statistically appropriate. There are several features in the clustering results that are consistent with those for the June–August season. For instance, a combination of U20c, U50c, U85c, V20c, V50c, V85c, T50c, Z50c, Z85c, and SLPc and a combination of T50b, T85b, Tsfb, Z20b, and Z50b are found in all periods. These similarities support the argument that the clustering results reflect underlying physical and dynamic constraints independent of the seasons. On the other hand, there are some differences in the clustering results among seasons, which suggest that it is necessary to perform the cluster analysis to obtain the most appropriate results for a target season of individual studies.

Because we arbitrarily selected the 43 metrics, the appropriateness of the metric selection and the robustness of the clustering result against the metric selection may be questioned. To address this, we examined the stability of the clustering result against the removal of some of the 43 metrics from the Ward clustering. We removed several (from one to five) metrics, applied the Ward method to the remaining metrics, and compared the results with the reference result obtained by analyzing all 43 metrics. We examined all combinations of the removals for the one-, two-, three-, and four-metric-removal tests, and we sampled randomly 100 000 combinations for the five-metric-removal test. Figure 3 shows the ratios of the experiments in which clustering results are completely consistent with the reference result and those in which only one, two, and three metrics are classified into different clusters. We consider the transfer of up to three metrics as moderately consistent. For the one-metric-removal test, most of the experiments exhibit such moderately consistent results, whereas the ratio of experiments with moderately consistent results decreases with an increasing number of removals. For the

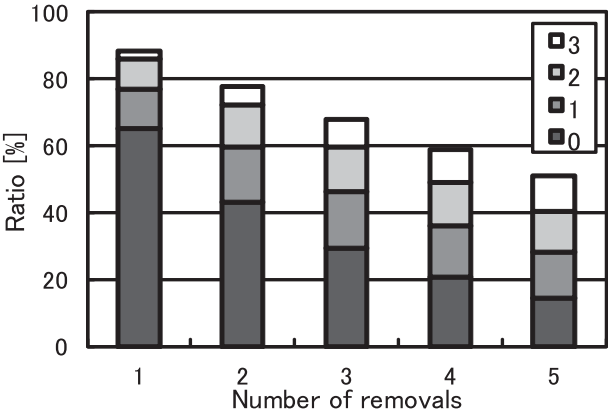

FIG. 3. Results of the robustness test, showing ratios of experiments by which clustering results are completely consistent with the reference result (darkest), and those in which only one (second darkest), two (second lightest), and three (the lightest) metrics are classified into different clusters.

five-metric-removal test, only approximately one-half of the experiments exhibit moderately consistent results. Note that more than 65% of the experiments of the one-metric-removal test exhibit a clustering result that is completely consistent. These results imply that the clustering result is not sensitive to the artificial removal of one or a few metrics from the cluster analysis, although removal of five or more metrics sometimes considerably influences the clustering result.

It is sometimes difficult to obtain robust results by using cluster analysis if it is applied to data with large dimensions. This is because the Euclidean distance between a given pair of data points in phase space tends to increase with data dimension. This may seriously affect the analysis in this study performed within the phase space having as many as 22 dimensions. Our analysis is surprisingly successful in defining clusters that are reasonable from the physical and dynamic viewpoints and that are not particularly sensitive to the artificial removal of one or a few metrics, however.

Cluster means, which are averages over the normalized performance metrics in individual clusters, can be used as new metrics instead of the original performance metrics. Note that, as with the original performance metrics, smaller values of the cluster means represent higher performance of the models in simulating the aspects of the climate relevant to the cluster.

Figure 4 compares intermodel significant relationships among the performance metrics with those among the cluster means. Statistical significance of the coefficients is evaluated based on the Student's t test with two different assumptions about the degrees of freedom (DOF). The first assumption is that the 22 models are independent, which means that DOF = 20. This assumption may be inappropriate, because five pairs of the GCMs are considered to be only different version of the same model, such as MIROC3hi and MIROC3med. Therefore, we also adopt a more reasonable assumption: the number of independent GCMs is considered to be 17 (DOF = 15). Results of the significance tests with the first and second assumption are presented with color tones and symbols, respectively, in Fig. 4. Intermodel correlation coefficients among the 43 metrics (Fig. 4a) show that, in general, a pair of metrics that exhibits a statistically significant correlation tends to be grouped in a single cluster—in particular, clusters A, C, and G. Furthermore, the correlation between two metrics in different clusters is generally insignificant. On the other hand, intermodel correlation coefficients among the cluster means for the K-means method (Fig. 4b) show that there are only three pairs that exhibit moderate positive correlation at the 95% confidence level (less than at the 99% level) for the test with DOF = 20, and only one pair for that with DOF = 15.

Table 3 summarizes the percentage of metric pairs and cluster-mean pairs with significant intermodel correlation for the test with DOF = 15. Whereas 17% of the metric pairs (156 out of 903 pairs) exhibit statistically significant positive correlation above the 95% levels, only 5% of the cluster-mean pairs (1 out of 21 pairs) exhibit such correlation. Furthermore, whereas 9% of the metric pairs exhibit correlation at a higher significance level (99% and 99.9%), such high correlation is not observed for the cluster-mean pairs. These results suggest that the problem of significant intermodel relationship among the metrics can be considerably alleviated by using the cluster means instead of the original metrics.

4. Discussion

One advantage of the clustering method is that we can simplify the evaluation studies of GCM performance by reducing the number of metrics considered and by reducing the significant intermodel correlation between the metrics. Using the clustering result, we can roughly grasp the similarities and the differences in reproducibility between the different versions of the GCM, which is useful information during model improvement activities. As an example, we apply the K-means clustering result to the comparison in model performance among three versions of our Japanese GCM: MIROC3med, MIROC3hi (K-1 Model Developers 2004), and MIROC5 (Watanabe et al. 2010). These GCMs were developed by the Atmosphere and Ocean Research Institute (AORI) of the University of Tokyo, the National Institute for Environmental Studies (NIES), and the Japan Agency for Marine-Earth Science and Technology (JAMSTEC). The former two versions participated in CMIP3. The major difference between the

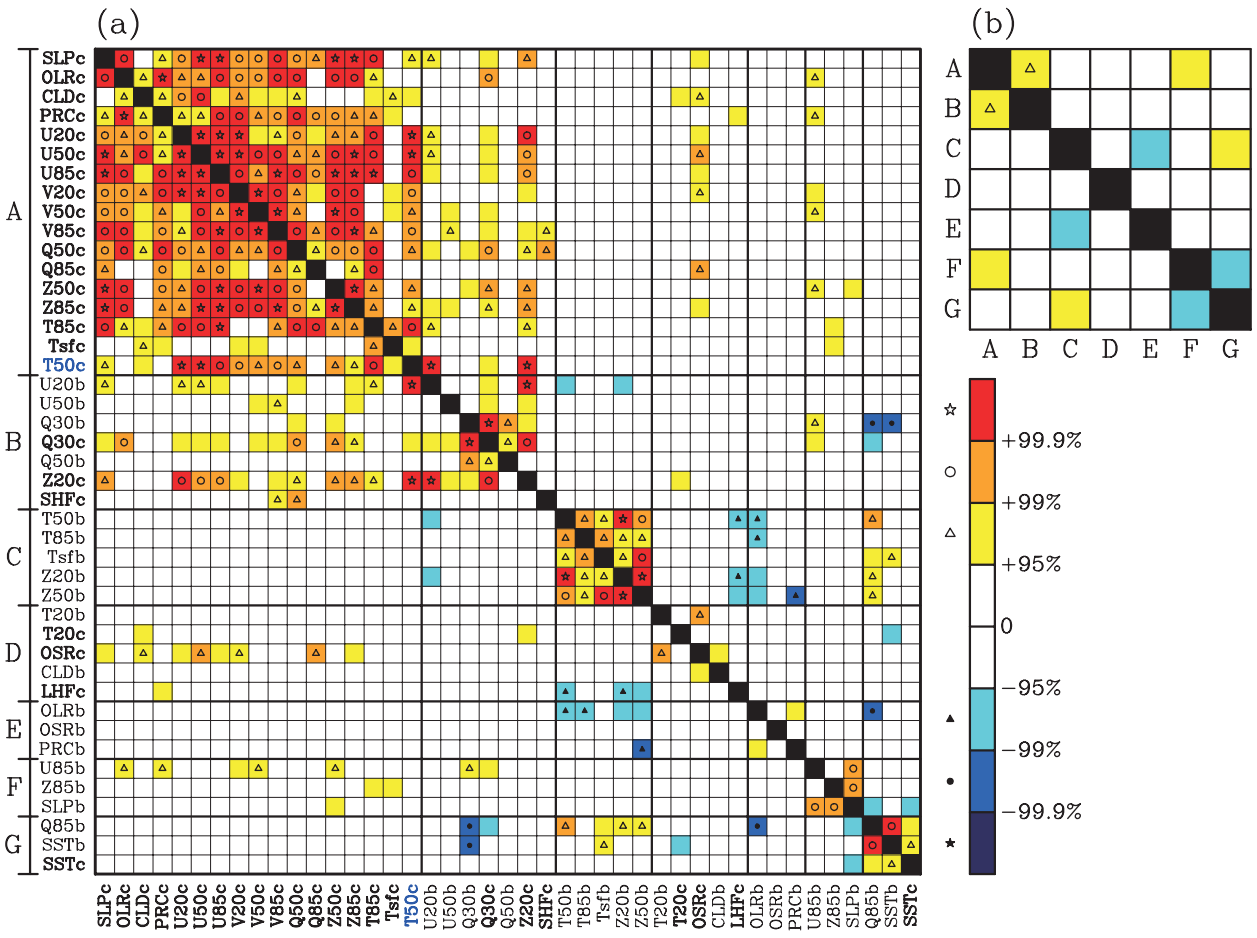

FIG. 4. (a) Statistical significance of correlation coefficients between performance metrics based on the Student's *t* test with DOF = 20 (color tones) and DOF = 15 (symbols). The clustering result of the *K*-means methods is shown at the left of the figure. Metric identifications in boldface type represent the centered-RMSE metric, and those without boldface represent the mean-bias metric. Note that T50c, shown by blue type, belongs to cluster B in the Ward clustering method. (b) As in (a), except showing statistical significance of correlation coefficients between cluster means of the *K*-means method.

two versions is the horizontal and vertical resolution of the atmosphere and ocean components; the horizontal resolution of the atmosphere is T42 for MIROC3med and is T106 for MIROC3hi. The third model, MIROC5, is a newly developed version designed for the next IPCC report. It has numerous improvements in the parameterization schemes. In particular, MIROC5 adopts new schemes for cumulus convection, large-scale condensation, cloud microphysics, and turbulence parameterization (Watanabe et al. 2010).

For comparing the three MIROC versions, we calculate cluster means of the normalized performance metrics by using the clustering result in the last section (shown in Table 2). Because the intermodel mean and the standard deviation used for the normalization are obtained from the 22 CMIP3 models, negative (positive) values represent a higher (lower) performance than the average performance over the 22 CMIP3 models. Figure 5 shows such cluster means for the three MIROC versions. On comparing MIROC3med and MIROC3hi, it is revealed that five out of seven cluster means of MIROC3hi are smaller than those of MIROC3med. The other two cluster means are comparable. These results suggest an advantage of higher resolution.

On the other hand, the performance of MIROC5 is, despite its coarser resolution, superior to that of MIROC3hi in five out of seven clusters, owing to the result of substantial improvements in the model. As a result of the

TABLE 3. Ratio of the metric pairs and the cluster-mean pairs that exhibit statistically significant positive intermodel correlation coefficients above 99.9%, 99%, and 95% significance levels with DOF = 15.

Significance levels	≥99.9%	≥99%	≥95%
Metric pairs	3%	9%	17%
Cluster-mean pairs	0%	0%	5%

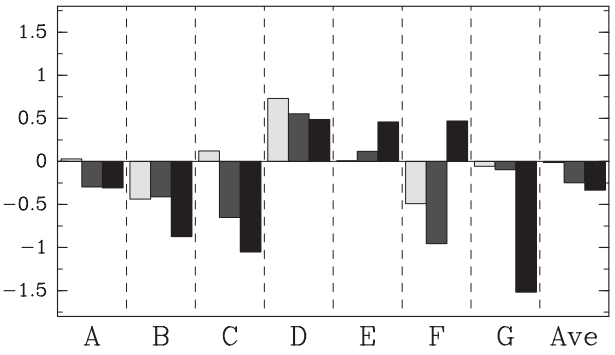

FIG. 5. Cluster means of the normalized performance metrics of MIROC3med (light gray), MIROC3hi (gray), and MIROC5 (dark gray). Smaller values indicate better performance, and negative (positive) values indicate higher (lower) performance than the average performance over the 22 CMIP3 models. The clustering results of the *K*-means method are used. Simple averages over the seven cluster means are shown at the farthest right (Ave).

clustering method, we can roughly discuss the aspects of the climate that are better simulated by the new model. The most notable improvement can be found in cluster G, wherein the SST distribution has become more realistic than that in the former versions, as pointed out by Watanabe et al. (2010). The simulations of the upper-tropospheric jet and moisture field represented by cluster B and the mean temperature profile represented by cluster C are also improved. If we experimentally define an overall performance metric as a simple average of the seven cluster means as proposed by Gleckler et al. (2008), MIROC5 exhibits slightly higher overall performance than MIROC3hi. Note, however, that this evaluation is no doubt dependent on the definition of the overall metric, which should be determined in the context of the purpose of the overall metric.

The clustering approaches will also be applicable to various kinds of GCM studies and future projections. For instance, they can help to select GCMs whose simulations are used as initial and boundary conditions for dynamic downscaling (e.g., Kawase et al. 2009) and impact assessment models (e.g., Iizumi et al. 2010). They can also be applied to multimodel detection and attribution studies (e.g., Santer et al. 2009) and multimodel-based future assessments (e.g., Murphy et al. 2004).

In such studies, we may have to define a synthesized overall performance metric that measures overall reproducibility of diverse aspects of the climate. If we have to define it on the basis of a weighted sum of original metrics, determination of specific weighting will be complicated because of the numerous metrics and the significant intermodel correlations between them. On the other hand, the use of cluster means as a new set of metrics will alleviate this difficulty because of the reduced number of metrics and generally lower intermodel correlations. In addition, when we want to consider all the conceivable variables for which appropriate reference data can be obtained, the clustering approach will allow it without introducing concerns about increases in the number of performance metrics.

Note that the selection of the performance metrics depends on the purpose of the evaluation. In addition to the global field, it may be important to evaluate the regional aspects for dynamic downscaling and impact assessment studies. As seen in the last section, the clustering results also depend on the analyzed season. Furthermore, aspects of the climate are not limited to climatological mean fields. For instance, interannual variability is probably important for assessing future projection, and spatial distribution of the variance and the phase of diurnal variability and seasonal cycle may be important for impact assessments on agriculture.

5. Summary

The evaluation of overall GCM performance in simulating climatological mean fields should be based on a synthesis of various kinds of atmospheric and oceanic variables (Gleckler et al. 2008). We have to consider numerous performance metrics, however, some of which are sometimes significantly correlated because of the underlying physical and dynamic constraints. If we reduce the number of the metrics and the significant relationship, we may be able to simplify the evaluation studies of the overall GCM performance. For this purpose, we attempt to use the cluster analysis to group the performance metrics that are mutually linked. We apply two clustering methods to the 43 performance metrics that measure the reproducibility of the boreal summer climatological fields of 14 variables. The two methods suggest that retaining seven clusters is statistically the optimal choice in this case, yielding consistent clustering results between the two methods. Several aspects of the clustering results are agreeable from the physical and dynamic viewpoints and/or are common among the four seasons. Furthermore, the clustering results are not particularly sensitive to the removal of one or a few metrics from the cluster analysis. Examining the intermodel correlation analysis among the metrics and among the cluster means, we find that the ratio of the cluster-mean pairs that exhibit statistically significant positive correlation is considerably less than that of the metric pairs. These results suggest that the cluster analysis is useful in obtaining appropriate grouping. In addition, the problem of a significant intermodel relationship between the metrics will be alleviated by using the cluster means as new metrics instead of using the original metrics.

We believe that the cluster analysis will be helpful in selecting GCMs for the dynamic downscaling approach and for impact assessment and in synthesizing outputs of GCMs for multimodel detection and attribution studies and multimodel future projection. In addition, cluster analysis will help in summarizing details of model improvement, which may be useful information for model development activities.

It is pointed out that, in addition to the two clustering methods adopted in this study, there are several approaches that may be potentially useful for alleviating a significant intermodel relationship problem. These approaches include soft (fuzzy) clustering, empirical orthogonal function analysis, and single-value decomposition analysis. Comparisons of these approaches are useful for identifying their advantages if these approaches are applied to the evaluation of model performance, which remains our future topic of research.

Acknowledgments. The authors acknowledge the modeling groups, the Program for Climate Model Diagnosis and Intercomparison, and the World Climate Research Programme's Working Group on Coupled Modelling for their roles in providing the CMIP3 multimodel dataset. The Data Integration and Analysis System Fund for the National Key Technology from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan, supported the authors in obtaining and analyzing the dataset. The development and simulation of MIROC5 were conducted with the support of the Innovative Program of Climate Change Projection for the 21st Century (''Kakushin'' program) from MEXT, Japan. The twentieth-century climate simulation of MIROC5 was performed by Dr. Tokuta Yokohata of JAMSTEC. This study is supported by the Global Environmental Research Fund (S-5-2) of the Ministry of Environment, Japan.

REFERENCES

Adler, R. F., and Coauthors, 2003: The version-2 Global Precipitation Climatology Project (GPCP) monthly precipitation analysis (1979–present). *J. Hydrometeor.,* **4,** 1147–1167.

Forgy, E. W., 1965: Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications. *Biometrics,* **21,** 768–769.

Fovell, R. G., and M.-Y. C. Fovell, 1993: Climate zones of the conterminous United States defined using cluster analysis. *J. Climate,* **6,** 2103–2135.

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

Iizumi, T., M. Nishimori, and M. Yokozawa, 2010: Diagnostics of climate model biases in summer temperature and warm-season insolation for the simulation of regional paddy rice yield in Japan. *J. Appl. Meteor. Climatol.,* **49,** 574–591.

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. Climate,* **12,** 2856–2880.

K-1 Model Developers, 2004: K-1 coupled model (MIROC) description. K-1 Tech. Rep., H. Hasumi and S. Emori, Eds., 34 pp. [Available online at http://www.ccsr.u-tokyo.ac.jp/~agcmadm/.]

Kawase, H., T. Yoshikane, M. Hara, F. Kimura, T. Yasunari, B. Ailikun, H. Ueda, and T. Inoue, 2009: Intermodel variability of future changes in the Baiu rainband estimated by the pseudo global warming downscaling method. *J. Geophys. Res.,* **114,** D24110, doi:10.1029/2009JD011803.

Knutti, R., R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl, 2010: Challenges in combining projections from multiple climate models. *J. Climate,* **23,** 2739–2758.

Meehl, G. A., C. Covey, T. Delworth, M. Latif, B. McAvaney, J. F. B. Mitchell, R. J. Stoufeer, and K. E. Taylor, 2007a: The WCRP CMIP3 multimodel dataset. *Bull. Amer. Meteor. Soc.,* **88,** 1383–1394.

——, and Coauthors, 2007b: Global climate projections. *Climate Change 2007: The Physical Science Basis,* S. Solomon et al., Eds., Cambridge University Press, 747–846.

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 modeling uncertainties in a large ensemble of climate change simulations. *Nature,* **430,** 768–772.

Pincus, R., C. P. Batstone, R. J. P. Hofmann, K. E. Taylor, and P. J. Gleckler, 2008: Evaluating the present-day simulation of clouds, precipitation, and radiation in climate models. *J. Geophys. Res.,* **113,** D14209, doi:10.1029/2007JD009334.

Rayner, N. A., P. Brohan, D. E. Parker, C. K. Folland, J. J. Kennedy, M. Vanicek, T. Ansell, and S. F. B. Tett, 2006: Improved analyses of changes and uncertainties in sea surface temperature measured in situ since the mid-nineteenth century: The HadSST2 dataset. *J. Climate,* **19,** 446–469.

Reichler, T., and J. Kim, 2008: How well do coupled models simulate today's climate? *Bull. Amer. Meteor. Soc.,* **89,** 303–311.

Rossow, W. B., and R. A. Schiffer, 1999: Advances in understanding clouds from ISCCP. *Bull. Amer. Meteor. Soc.,* **80,** 2261–2287.

Santer, B. D., and Coauthors, 2009: Incorporating model quality information in climate change detection and attribution studies. *Proc. Natl. Acad. Sci. USA,* **106,** 14 778–14 783.

Uppala, S. M., and Coauthors, 2005: The ERA-40 Re-Analysis. *Quart. J. Roy. Meteor. Soc.,* **131,** 2961–3012.

Ward, J. H., Jr., 1963: Hierarchical grouping to optimize an objective function. *J. Amer. Stat. Assoc.,* **58,** 236–244.

Watanabe, M., and Coauthors, 2010: Improved climate simulation by MIROC5: Mean states, variability, and climate sensitivity. *J. Climate,* **23,** 6312–6335.

Wielicki, B. A., B. R. Barkstrom, E. F. Harrison, R. B. Lee III, G. L. Smith , and J. E. Cooper, 1996: Clouds and the Earth's Radiant Energy System (CERES): An Earth Observing System experiment. *Bull. Amer. Meteor. Soc.,* **77,** 853–868.

Williams, K. D., and G. Tselioudis, 2007: GCM intercomparison of global cloud regimes: Present-day evaluation and climate change response. *Climate Dyn.,* **29,** 231–250.

——, and M. J. Webb, 2009: A quantitative performance assessment of cloud regimes in climate models. *Climate Dyn.,* **33,** 141–157.