

Relationship of the Reproducibility of Multiple Variables among Global Climate Models

Kazuaki NISHII^{*}, Takafumi MIYASAKA, Hisashi NAKAMURA

Department of Earth and Planetary Science, University of Tokyo, Tokyo[#]

Yu KOSAKA

International Pacific Research Center, University of Hawaii, Honolulu, Hawaii, USA

Satoru YOKOI, Yukari N. TAKAYABU

Atmosphere and Ocean Research Institute, the University of Tokyo, Kashiwa

Hirokazu ENDO

Meteorological Research Institute, Tsukuba

Hiroki ICHIKAWA

Graduate School of Environmental Studies, Nagoya University, Nagoya

Tomoshige INOUE

Graduate School of Life and Environmental Sciences, University of Tsukuba, Tsukuba

Kazuhiro OSHIMA

Faculty of Environmental Earth Science, Hokkaido University, Sapporo

Naoki SATO

*Tokyo Gakugei University, Tokyo, and Japan Agency for Marine-Earth Science and
Technology, Yokohama*

and

Yoko TSUSHIMA

Met Office Hadley Centre, FitzRoy Road, Exeter EX1 3PB, UK

Submitted March 18, 2011, revised

^{*} Corresponding author: Kazuaki Nishii: Research Center for Advanced Science and
Technology, University of Tokyo, University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo,
153-8904, Japan. Email: nishii@atmos.rcast.u-tokyo.ac.jp.

[#]Current affiliation: Research Center for Advanced Science and Technology, University of Tokyo.

Abstract

Numerous efforts have been made for evaluating the performance of global climate models with such expectation that those models with higher reproducibility of the current climate should provide more reliable projections of climate changes into the future. Attempts have been made to define a single general metric through which the overall performance of a global climate model can be assessed. On the basis of general metrics defined through several techniques of multivariate analysis, the present study compares global climate models from a viewpoint of their reproducibility of climatological-mean fields of multiple variables. The analyses indicate that a reproducibility of a particular variable is not necessarily independent of that of others, which may bring redundant information into a general metric. The model reproducibility in upper and mid-tropospheric temperature and lower-tropospheric humidity, for example, tends to be anti-correlated with that in upper and mid-tropospheric humidity. It is argued that attention has to be paid to this kind of trade-off relationships among some variables and resultant redundancy in synthesizing multiple metrics. A possibility is suggested that an arbitrary selection of variables can yield some redundant information of variables. The redundancy is, however, found to exert no serious influence on the quality of a general metric as long as it is based on the sufficient number of variables. In our attempt to evaluate the climate models by introducing general performance metrics with reduced redundancy of variables, the overall model ranking is found rather insensitive to the specific definition of the metric.

1. Introduction

Quantitative projections of future climate changes depend more or less on numerical climate models. A multi-model ensemble (MME) is known to outperform individual models in reproducing the current climatic state owing to a tendency for their biases to cancel each other (e.g., Knutti et al. 2010). The MME future projection has therefore been believed to be more reliable than the corresponding projection based on a single model, as exemplified in the Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report (AR4; Solomon et al. 2007). In AR4 a simple algebraic average of the outputs from more than 20 global climate models that participated in the World Climate Research Programme's (WCRP's) Coupled Model Intercomparison Project Phase 3 (CMIP3; Meehl et al. 2007) is used as the best guess for the future projection. The cancellation of model biases is, however, not necessarily perfect. For example, a group of the CMIP3 models in which a particular parameterization scheme is commonly adopted, say, for cumulus convection may suffer from a common bias, suggesting that model biases are not necessarily distributed randomly. Even if model biases were distributed randomly, the number of available models would be unlikely sufficient for their perfect cancellation (e.g., Knutti et al. 2010). In fact, the effective number (or degrees of freedom: DOFs) of the CMIP3 models has been estimated to be only between five and ten (Jun et al. 2008a, 2008b; Knutti et al. 2010; Pennell and Reichler 2010). In other words, the amount of information provided as an ensemble of those models may be less than what would be expected under the assumption that all the models were mutually independent¹.

Spatial similarity of biases in such a model variable as climatological-mean surface air temperature (SAT) is often used as a measure of independency among the models. In addition to the insufficient effective number of models as discussed above, the effective

¹ Annan and Hargreaves (2010) showed that in a paradigm of *statistically indistinguishable* ensemble, CMIP3 models are well distributed in a sense that observations can be considered as a member of the CMIP3 ensemble.

number of these measures may also be limited. In fact, Yokoi et al. (2011) have demonstrated that a performance metric for a given variable (hereafter referred to as “variable metric”), which quantifies the similarity of its model-simulated distribution to its observational counterpart, may be correlated with other variable metrics under the constraint, for example, of thermal wind balance that relates circulation and thermal fields.

Efforts have been made to define a single general performance metric (hereafter referred to as “general metric”) into which various aspects of model performance are incorporated (Gleckler et al. 2008; Reichler and Kim 2008). This general metric can be used for determining weights for individual models to synthesize their outputs for defining an optimal MME (e.g., Murphy et al. 2004). Usually in defining a general metric, reproducibility of various variables is estimated separately on the basis of variable metrics before summed up, but what variables to be chosen is rather arbitrary. In fact, Knutti et al. (2010) pointed out “there is virtually an infinite number of metrics that can be defined”. Furthermore, Yokoi et al. (2011) argued that a general performance metric might be marred if seriously biased variables are incorporated into it. Furthermore, adding a new variable metric to a general metric may not necessarily lead to an effective increase in the information included in the metric, if the new variable is linked closely to any of the variables that have already been incorporated into the metric. In this case, the addition will introduce some redundant information, or even some bias, to the new general metric. In any case, “we currently have no basis for assigning unequal weights for any variables” (Sexton and Murphy 2003) in defining a general metric.

This study is motivated by Gleckler et al. (2008), who argued “it might be fruitful to explore a wide range of metrics, rather than striving for a single index of overall skill, and then to use some objective method to reduce redundant information (e.g., SVD)”. We examine linkages among variable metrics for the CMIP3 models by applying several techniques of multivariate analysis. We identify positively-correlated variable metrics in

106 particular variable groups and other metrics showing trade-off reproducibility of variables.
107 We then propose several definitions for general metrics in our attempt to reduce
108 redundancy.

109 The metrics defined in the following sections are based only on the climatological-mean
110 state. It should be pointed out that they do not necessarily capture every aspect of the
111 performance of a climate model, since its reproducibility of the mean state and that of
112 natural variability around it do not necessarily correlate positively (Gleckler et al. 2008;
113 Santer et al. 2009). Another possible defect of our metrics arises from their rather
114 straightforward definition. It has been pointed out that most of such straightforward metrics
115 as area-mean biases and root-mean-square errors for the present day climate do not
116 necessarily be applicable well to future projections (Whetton et al. 2007; Abe et al. 2009;
117 Girogi and Coppola 2010; Knutti et al. 2010). Recently, efforts have been devoted to finding
118 metrics that can connect current climate reproducibility reasonably to future projection (Hall
119 and Qu 2006; Boe et al. 2009; Shiogama et al. 2011), where these metrics are expected to
120 reduce uncertainty in future projections based on ensembles of climate models. In addition,
121 a new paradigm of a *statistically indistinguishable* ensemble has been proposed (Annan
122 and Hargreaves 2010), which differs from the particular paradigm we adopt here that
123 ensemble members are assumed to be distributed around the true climate. Despite the
124 defects mentioned above, we nevertheless use our metrics because our main goal is to
125 explore inter-variable relationships of multiple metrics.

126

127 **2. Data and analysis methods**

128 *2.1 Climate models and observed data*

Table 1

129 The multi-model dataset of the 20th Century Climate in Coupled Models (20C3M)
130 experiment in CMIP3 (Meehl et al. 2007) is utilized in this study. In Table 1, the 22 variables
131 used for our analysis are listed with their abbreviations for reference. For each of the

variables, model output data from 24 climate models are compared with observational data whose source and available periods are also listed in Table 1. Most of the variables are obtained from the Japanese 25-year reanalysis (JRA-25) of the global atmosphere (Onogi et al. 2007). We have verified that the usage of the European Centre Medium-Range Weather Forecast 40-yr Reanalysis (ERA40) data set (Uppala et al. 2005) in place of JRA-25 yields no substantial changes in the results presented below. We define a variable metric for the i -th model ($i = 1, \dots, I$) and the j -th variable ($j = 1, \dots, J$) as

$$C_{ij} = \frac{1}{\sigma_j} \sqrt{\frac{1}{12W} \sum_k^{12} \sum_l^L w_l (m_{ijkl} - o_{ijkl})^2} \quad (1)$$

where σ_j denotes standard deviation of the observed interannual variability of the j -th variable, w_l a local area weighting factor at the l -th grid point ($l = 1, \dots, L$), $W = \sum w_l$, and m_{ijkl} and o_{ijkl} are the simulated and observed climatological means of the j -th variable for the k -th calendar month ($k = 1, \dots, 12$), respectively. $\sum_j C_{ij}^2 / J$ is equivalent to the Climate Prediction Index (CPI; Murphy et al. 2004) for the i -th model. Since available periods for observed OLR and SWTOA are too short for a robust estimation of their interannual variances (Table 1), the estimation was based on the JRA25 data. A shortcoming of such metrics as ours that include mean square errors is that they cannot incorporate the signs of model errors. This may artificially reduce the effective variable number estimated in our analysis.

The inter-model variance in C is not necessarily comparable in magnitude among the variables. For example, standard deviations are large in upper and mid-tropospheric temperature and specific humidity fields (Fig. 1). In section 3, variances in C have been standardized with inter-model standard deviations, to explore relationships among variable metrics. However, no standardization has been applied to C in section 4, where we discuss general performance metrics that have to be related to the model reproducibility of variables and therefore their inter-model variances must be explicitly incorporated.

Fig. 1

2.2 Multivariate analysis techniques

158 In this subsection we briefly introduce three multivariate analysis techniques applied to C
 159 in the present study. One of them is a cluster analysis. As in Yokoi et al. (2011), we apply a
 160 cluster analysis to a set of variable metrics, to identify several groups of variable metrics
 161 that exhibit similar behaviors. We adopt so-called Ward (1967) method, which is based on
 162 the Euclidian distance between any pair of clusters in the phase space.

163 Unlike the cluster analysis, a principal component analysis (PCA), or an empirical
 164 orthogonal function (EOF) analysis, seeks for basis vectors that can be regarded as new
 165 “variable” metrics each of which can represent behaviors of multiple variable metrics.
 166 Before performing a PCA the RMS biases of individual variables within the model ensemble
 167 have been subtracted from the CPI matrix C defined in (1):

$$168 \quad C' = \{C'_{ij}\} = \{C_{ij} - \frac{1}{I} \sum_i C_{ij}\}. \quad (2)$$

169 The resultant matrix C' can be decomposed in PCA into a pair of orthogonal matrices:

$$170 \quad C' = U'V'^T, \text{ or } C'_{ij} = \sum_r^R U'_{ir} V'_{jr}, \quad (3)$$

171 where $U' = \{U'_{ir}\}$, $V' = \{V'_{jr}\}$ and $r = 1, \dots, R$ ($R = \min(I, J)$). In this factorization, the i -th row
 172 vector of C' (a set of variable metrics for the i -th model) is represented by a linear
 173 combination of the R column vectors in V' , called basis vectors or EOFs, with the
 174 corresponding i -th row vector of U' that represents a set of their coefficients that scores
 175 reproducibility of the i -th model.

176 As in the case of PCA, non-negative matrix factorization (NMF; Lee and Seung 1999)
 177 decomposes the CPI matrix C in (1). Unlike PCA, however, NMF decomposes C directly:

$$178 \quad C \sim PQ^T, \quad (4)$$

179 taking advantage of the fact that every element of C is nonnegative. In (4), P and Q are
 180 nonnegative $I \times R$ and $J \times R$ matrices, respectively, but not necessarily orthogonal. Here, a
 181 positive integer R satisfies $R < I J (I + J)^{-1}$.

Fig. 2

182 Figure 2 schematically compares basis vectors obtained through (a) PCA and (b) NMF
 183 applied to a hypothetical two-variable metric data set. The origin of the PCA basis vectors is

184 situated at the center of balance between the two model groups that corresponds to the
185 RMS bias in (1). The leading PCA vector is in the direction of the maximum variability of the
186 metrics, and the second PCA vector must be orthogonal to the leading vector. In contrast,
187 the NMF basis vectors are not orthogonal mutually. In a hypothetical situation where there
188 are only two groups of climate models as in Fig. 2, the two NMF basis vectors are inclined
189 to point those groups. In the particular phase space illustrated in Fig. 2, a model with lower
190 reproducibility of the current climatic state tends to be more distant from the origin². The
191 particular distance can therefore be regarded as a general performance metric, and the
192 projection of the state vector of a given model onto a NMF basis vector can thus be
193 considered as a new variable metric that comprises multiple variables showing similar
194 behaviors. A general performance metric thus defined should be subject to a certain degree
195 of redundancy, which can nevertheless be reduced in synthesizing these projections. This
196 contrasts with the PCA vectors that do not necessarily point the origin of the phase space
197 but may rather represent trade-off reproducibility among the variables.

198 While some suggestions have been made on how many basis vectors should be retained
199 for PCA, no objective criterion has been proposed thus far for determining R in NMF. In fact,
200 Schlink and Thiem (2009), who applied NMF to identify dominant patterns of atmospheric
201 variability, determined R empirically after several trials in varying R . While relative
202 importance of a given set of PCA basis vectors can be assessed with the corresponding
203 eigenvalues, the order of NMF basis vectors cannot be uniquely determined. With this
204 peculiarity of NMF, all the basis vectors should be treated evenly.

205

206 **3. Relationship among multiple variable metrics**

207 **3.1 Cluster analysis**

Fig. 3

208 Figure 3 shows a dendrogram based on our cluster analysis that was applied to a set of

² Here we assume that both internal climate variability and observational errors are much smaller than the model bias, as is likely the case for most of the models.

209 C after standardizing inter-model variances. We adopted a stopping rule of Calinski and
 210 Harabasz (1974). Though not particularly distinct, the maximum of the pseudo- F function in
 211 their definition, which is the ratio of the inter-cluster variance based on the means of the
 212 individual clusters to the mean of the intra-cluster variances, was found to be realized when
 213 the model members were categorized into two main clusters. This result of our cluster
 214 analysis may be attributable to the artifact of RMSE-based metrics where the signs of
 215 biases are neglected. One of the two main clusters consists of upper and mid-tropospheric
 216 temperature (T300, T500, T600, T700) and lower-tropospheric humidity (Q850), whose
 217 combination may be understandable except for humidity. The other main cluster, which
 218 consists of the 17 other variables, comprises several sub-clusters. One of them consists of
 219 lower-tropospheric temperature (T850), SAT and sea surface temperature (SST), whose
 220 close association in the climate models is understandable. However, interpretation of some
 221 of the other sub-clusters is not necessarily straightforward. It seems counterintuitive, for
 222 example, that model biases in surface sensible and latent heat fluxes are not closely
 223 related to those in either SAT or SST. As argued by Yokoi et al. (2011), the mixture
 224 between variables that can yield model biases in their global-mean values (e.g., SLP and
 225 temperature fields) and those that cannot (e.g., meridional wind velocity) may complicate
 226 the interpretation.

227

228 3.2 PCA

Fig. 4

229 We applied PCA to the same set of C as above through the eigenvalue decomposition of
 230 its correlation matrix (Fig. 4). Fractions of the total variance explained by these modes are
 231 36%, 21%, 9%, 7%, 6% and 5%. Thus more than 80% of the total variance is explained by
 232 the six leading modes, which means that most of the information of the 22 variables can be
 233 accounted for only by these six modes. The first mode represents the overall model
 234 performance (Fig. 4a). Models that earn large negative scores of this mode tend to show

high reproducibility in 16 out of the 22 variables but not for upper and mid-tropospheric temperature (T300, T500, T600, T700), lower-tropospheric humidity (Q850) and cloud cover (Fig. 4d). Meanwhile, reproducibility of most of these six variables is measured by the second PCA mode (Fig. 4b), and its large negative score represents high reproducibility of those variables (Fig. 4e). In contrast to these two leading modes, the higher modes represent trade-off relationships in reproducibility among the 22 variables (Figs. 4c and 4f), and therefore none of these modes alone can be used as a measure of the overall performance of a given model. The trade-off relationships found in the analysis by Yokoi et al. (2011) and ours may suggest that one should not focus too much on the model reproducibility only of a particular aspect, in order to avoid its over-tuning at the expense of other aspects. We should keep in mind, however, that the trade-off relationships represented by the higher modes tend to be more or less overemphasized due to an artifact of PCA (Lee and Seung 1999).

248

249 3.3 NMF

Fig. 5

Our cluster analysis implies that the DOFs of the variable metrics of C may be only two, while the six leading modes are retained for our PCA. In recognition of this uncertainty, we repeatedly applied NMF to the standardized C , changing R from two to six. Figure 5 presents the results for $R = 2$ as a typical example. In Fig. 5, a small value in P_{ir} suggests high reproducibility of the i -th model in a particular aspect represented by the r -th column vector of Q . The first NMF mode for $R=2$ measures the reproducibility of upper and mid-tropospheric temperature and lower-tropospheric humidity, whereas that of upper and mid-tropospheric humidity is scored effectively by the second mode. The grouping of the variables into the two NMF modes is overall consistent with the corresponding grouping in our cluster analysis and PCA. The characteristic of the first mode for $R=2$ is fairly robust as it is reproduced in the second mode for $R=3$ (not shown). A positive score of the first NMF

mode with $R=3$ corresponds to lower reproducibility of upper and mid-tropospheric humidity, T850, SAT and SST. The third mode implies better reproducibility of temperature fields in those models with large Q values at the expense of that of other variables.

4. Attempts for synthesizing multiple variable metrics for reduced redundancy

Several methods have been proposed for synthesizing multiple variable metrics, but some of them, including an algebraic mean of the variable metrics, are rather ad hoc. Utilizing the multivariate analyses discussed above, we make several attempts to reduce redundant information in a set of multiple variable metrics in defining a scalar metric as a measure of model's general performance ("general metric"), as in Yokoi et al. (2011). In our attempts, we try to evaluate the overall performance of the i -th model with R ($r = 1, 2, \dots, R$) new variable metrics defined as:

$$\tilde{C}_{ir} = \frac{\sum_j \omega_{jr} C_{ij}}{\sum_j \omega_{jr}}, \quad (5)$$

where ω_{jr} signifies the weighting for the r -th metric that has been defined through one of the analysis methods discussed above. For the cluster-analysis-based CPI, $\omega_{jr} = 1$ if the j -th variable belongs to the r -th variable cluster or $\omega_{jr} = 0$ otherwise. For the NMF-based metrics, $\omega_{jr} = Q_{jr}$. A new general metric for the i -th model with reduced redundancy may thus be given as

$$\hat{D}_i = \frac{\sum_r \tilde{C}_{ir}^2}{R}. \quad (6)$$

Our cluster analysis of the unnormalized C gives us $R = 3$, because the pseudo F reaches its maximum for three main clusters, whereas PCA for the unnormalized C suggests $R = 4$, because the four leading modes explain more than 80% of the total variance represented as the trace of the covariance matrix of C . On the basis of these results $R = 3$ and 4 are tested for our NMF, but their difference is so small that only results for $R = 3$ are discussed in the following.

We also utilize total energy (TE; Talagrand 1981), which has been used as a norm for

287 evaluating forecast errors. In our practice, TE is integrated over the global domain A :

$$288 \quad TE = \frac{1}{2} \iint \{u^2 + v^2 + \frac{C_p}{T_r} T'^2 + R T_r (\frac{p'_s}{p_r})^2 + \frac{L^2}{C_p T_r} q'^2\} dA dp, \quad (7)$$

289 where primes denote deviations from the observations, u westerlies, v southerlies, C_p
 290 specific heat at constant pressure, L latent heat, R gas constant, T temperature, T_r
 291 reference temperature, and q specific humidity. In (7), the vertical integration was
 292 performed between the $p = 200$ and 1000 (hPa) levels. No evaluation was made, however,
 293 for the term that includes surface pressure (p_s), which is not available in some of the CMIP3
 294 model output. Strictly speaking, TE cannot be regarded as a general metric for model
 295 performance, since solar and terrestrial radiations, surface heat fluxes and cloud cover are
 296 all excluded from it. It can nevertheless offer a physically meaningful means for
 297 synthesizing dynamical and thermal variables in defining a metric. As another general
 298 metric, we also adopt the same definition as the Model Climate Performance Index (MCPI;
 299 Gleckler et al. 2008), which is a simple summation of the conventional variable metrics but
 300 with the variable metrics listed in Table 1.

Fig. 6

301 Figure 6 compares the model rankings based on the aforementioned general metrics.
 302 Models that are evaluated at higher rankings based on a particular general metric tend to
 303 be ranked at higher positions based on the other general metrics. Although the TE-based
 304 model ranking tends to deviate slightly from those based on the other metrics, the overall
 305 consistency among the model rankings based on the various general metrics implies that
 306 the reproducibility of the dynamical variables is more or less related to that of the physical
 307 variables.

Fig. 7

308 Figure 7 shows the numbers of variable metrics that are ranked as the top five (squares
 309 with solid line) and bottom five (triangles with dotted line) among the 24 CMIP3 models.
 310 The models are listed in descending order according to the CPI-based general metric. The
 311 figure indicates an overall tendency for models with higher (lower) ranking based on the
 312 CPI-based general metric to exhibit higher (lower) reproducibility with respect to a greater

number of variable metrics. For example, ECHAM5/MPI-OM, the best model based on the general metric, is ranked among the top five of the 24 models with respect to as many as 15 variable metrics, while only a single variable metric ranks this model (ECHAM5/MPI-OM) among the bottom five. In contrast, the three models that earn the lowest scores of the general metric are not ranked among the top five with respect to any of the variable metrics. Meanwhile, such models as GFDL-CM2.1, MRI-CGCM2.3.2, CSIRO-Mk3.5 and GFDL-2.0 earn the top five scores in as many variable metrics as the higher-ranked models based on the general metric do so. Those models exhibit, however, the relatively low reproducibility in air temperature and humidity, whose inter-model variances tend to be large (Fig. 1). This is hinted at in Figs. 5b and 5d, where these models earn high scores in P . Our results suggest that a general metric based on an unnormalized matrix C may likely be influenced substantially by the reproducibility of variables with large inter-model variances.

5. Discussion and conclusions

In this paper, we have compared several multivariate analysis methods that can be used for extracting relationships among variable metrics. While details are dependent of specific analysis methods, there are nevertheless some common features in the resultant grouping of the variable metrics. Some groups of the metrics obtained as the leading PCA or NMF modes are characterized by variable metrics whose inter-model variances are large and can thereby score the overall performance of the models. In contrast, other groups represent trade-off relationships among the variables in their model reproducibility.

We have also proposed several methods to reduce redundancy in variable metrics before defining a general metric that scores the general performance of climate models. Model rankings are, however, rather insensitive to the particular definition of the general performance metric (Fig. 6). These results suggest that (i) a general performance metric

339 that consists of a sufficiently large number of variable metrics is unlikely to be influenced
340 significantly by the redundancy of variables, and (ii) good models tend to show high
341 reproducibility in various aspects, at least based on the metrics used in this study (Fig. 7).

342 Basically our metrics are based on RMSE from the observed climatology³, even in the
343 estimation with the total energy norm. Thus one may consider that this similarity in the
344 definition of the metrics based on CPI, MCPI and TE may lead to the similarity among the
345 model rankings based on those metrics as shown in Fig. 6. We have compared the model
346 ranking based on CPI with those on the pattern correlations and RMSE of global-mean
347 biases (Fig. 8). Although the similarity among those three rankings is weaker if compared to
348 that among the rankings shown in Fig. 6, there is still a tendency that those models with
349 higher rankings based on CPI tend to be ranked also in higher positions based on the
350 pattern correlation and global-mean biases.

Fig. 8

Fig. 9

351 As noted in the introduction, metrics that are related to future projections have been
352 sought (Hall and Qu 2006; Boe et al. 2009; Shiogama et al. 2011). Though it is beyond the
353 scope of the present study, it will be valuable to assess briefly whether the simple metrics
354 defined in this study may have any relevance to future projection. Following Abe et al.
355 (2009), we compared inter-model similarity of present-day climate simulation with that of
356 projected future change. The inter-model similarity is evaluated between possible pairs of
357 the CMIP3 models based on CPI (Fig. 9a) or single variable metrics (Fig. 9b). The future
358 change is based on the difference between the averages for the two periods, one for
359 2070-2099 of the A1B scenario experiment and the other for 1970-1999 of the 20C3M
360 experiment. More specifically, the former average is assigned to m_{ijkl} and the latter to o_{ijkl} in
361 (1). σ_j is based on the current climate. Figure 9b summarizes the correlation in the
362 inter-model similarities between the present-day climate and projected future change

³ Note that RMSE-based metrics provide us with mixture of information on the similarity in model-simulated and observed climatological-mean fields of a given variable from multiple perspectives: the global-mean bias and pattern similarities with respect to spatial distribution and local amplitude.

363 based on the same scatter plot as in Fig. 9a but based on respective variable metrics. The
364 figure indicates fairly high correlation between current climate and future change projection
365 based on single variable metrics, especially in OLR, SWTOA and Prec, except for
366 tropospheric temperatures. The high correlations of variable metrics suggests that a pair of
367 models that simulate similar mean fields for the present-day climate tends to yield similar
368 future projection in the mean field, as long as the similarity is measured by those variables.
369 The correlation lowers if these variables are synthesized in the form of CPI (0.21), while the
370 correlation is improved slightly (0.31) if temperature metrics are excluded. This modest
371 correlation implies that uncertainty that could emerge in the future projection may not be
372 well constrained by using a synthesized metric that consists of multiple aspects, even if
373 each of the metrics shows high correlation between the present-day climate and future
374 projection. In our analysis, high correlations are found in some variable metrics, but the
375 physical reasoning has not been uncovered.

376 Previous studies have pointed out that the CMIP3 models are not mutually independent
377 and their effective number is only between five and ten (Jun et al. 2008a, 2008b; Pennell
378 and Reichler 2010). The estimation of the effective model number by using PCA is
379 equivalent to that of the number of effective metrics or measures of inter-model similarity,
380 since the numbers of nonzero eigenvalues of inter-model and inter-variable covariance
381 matrices of C are identical. As there are infinite ways to define metrics, incorporating
382 additional metrics may increase the effective model number. While precise estimation of
383 the effective numbers of models and variables may be of little worth, it will be worthwhile to
384 deepen our understanding of inter-model and inter-metric relationships. In section 3,
385 linkages were revealed among different variable metrics for the CMIP3 models. Some of
386 them seem to reflect physical relationships among the variables or in parameterization
387 schemes, while others may be mere artifacts of constraints among the variables by a
388 particular analysis method. Further investigation is needed to identify the origins of the

revealed relationships. In section 4, we attempted to reduce redundancy among the variable metrics in quantifying general performance of the CMIP3 models. Still, no attempt has been made for avoiding inter-model dependency that may distort the uncertainty (i.e., PDF) of the future projection in the ensemble of the CMIP3 models.

In the present study, we have focused on the reproducibility of the climatological-mean fields, whereas most of the studies on model reproducibility also focus on time-variability and long-term trends. From a regional viewpoint, however, assessing the model reproducibility of atmospheric phenomena, including tropical and midlatitude cyclones and large-scale teleconnection patterns, is necessary for reliable projection of their future changes. Several studies applied process-oriented performance metrics to the CMIP3 models (e.g. Yokoi and Takayabu 2009; Nishii et al. 2009). Especially, Kosaka and Nakamura (2011) found that models with better reproducibility of the climatological-mean fields tend to show better reproducibility of the most dominant summertime anomaly pattern over the western North Pacific. Exploring the relationships among process-oriented regional metrics and global metrics based on climatological-mean fields and their trends will be valuable in improving global climate models.

405

Acknowledgments

This study is supported by the Global Environment Research Fund (S-5) of the Ministry of Environment and by the Grant-in-Aid for Scientific Research (B) #22340135 by the Ministry of Education, Culture, Sports, Science and Technology (MEXT). The “Data Integration and Analysis System (DIAS)” Fund for the National Key Technology by the MEXT provides us with efficient environment for handling massive data. We acknowledge the modeling groups, the PCMDI and the WCRP’s Working Group on Coupled Modelling (WGCM) for their roles in making available the WCRP CMIP3 multi-model dataset. Support for this dataset is provided by the Office of Science, U.S. Department of Energy.

415

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.
- Annan, J. D., and J. C. Hargreaves, 2010: Reliability of the CMIP3 ensemble, *Geophys. Res. Lett.*, **37**, L02703, doi:10.1029/2009GL041994
- Barkstrom, B., E. Harrison, G. Smith, R. Green, J. Kibler, R. Cess, and the ERBE Science Team, 1989: Earth Radiation Budget Experiment (ERBE) archival and April 1985 results. *Bull. Amer. Meteor. Soc.*, **70**, 1254-1262.
- Boe, J. L., A. Hall, and X. Qu, 2009: September sea-ice cover in the Arctic Ocean projected to vanish by 2100. *Nat. Geosci.*, **2**, 341-343.
- Calinski, R. B., and J. Harabasz, 1974: A dendrite method for cluster analysis. *Communications in Statistics*, **3**, 1-27.
- Giorgi, F., and E. Coppola, 2010: Does the model regional bias affect the projected regional climate change? An analysis of global model projections. *Climatic Change*, **100**, 787-795.
- Gleckler, P. J., K. E. Taylor, and C. Doutriaux, 2008: Performance metrics for climate models. *J. Geophys. Res.*, **113**, D06014, doi:10.1029/2007JD008972.
- Hall, A., and X. Qu, 2006: Using the current seasonal cycle to constrain snow albedo feedback in future climate change. *Geophys. Res. Lett.*, **33**, L03502, doi:10.1029/2005GL025127.
- Jun, M., R. Knutti, and D. W. Nychka, 2008a: Local eigenvalue analysis of CMIP3 climate model errors. *Tellus*, **60A**, 992-1000.
- Jun, M., R. Knutti, and D. W. Nychka, 2008b: Spatial analysis to quantify numerical model bias and dependence: How many climate models are there? *J. Amer. Stat. Assoc.*, **103**, 934947, doi:10.1198/016214507000001265.

442 Lee, D. D., and H. S. Seung, 1999: Learning the parts of objects by non-negative matrix
 443 factorization. *Nature*, **40**, 788-791.

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

446 Kosaka, Y. and H. Nakamura, 2011: Dominant mode of climate variability, intermodel
 447 diversity and projected future changes over the summertime western North Pacific
 448 simulated in the CMIP3 models. *J. Climate*, **24**, 3935-3955.

449 Meehl, G. A., C. Covey, T. Delworth, M. Latif, B. McAvaney, J. F. B. Mitchell, R. J. Stouffer,
 450 and K. E. Taylor, 2007: The WCRP CMIP3 multi-model dataset: A new era in climate
 451 change research. *Bull. Amer. Meteor. Soc.*, **88**, 1383-1394.

452 Murphy, J. M., D. M. H. Sexton, D. N. Barnett, G. S. Jones, M. J. Webb, M. Collins and D. A.
 453 Stainforth, 2004: Quantification of uncertainties in large ensembles of climate change
 454 prediction. *Nature*, **430**, 768–772.

455 Nishii, K., T. Miyasaka, Y. Kosaka, and H. Nakamura, 2009: Reproducibility and future
 456 projection of the midwinter storm-track activity over the Far East in the CMIP3 climate
 457 models in relation to the occurrence of the first spring storm (Haru-Ichiban) over Japan. *J.*
 458 *Meteor. Soc. Japan*, **87**, 581-588.

459 Onogi, K., and co-authors, 2007: The JRA-25 reanalysis. *J. Meteorol. Soc. Japan.*, **85**,
 460 369-432.

461 Pennell, C., and T. Reichler, 2011: On the effective number of climate models. *J. Climate*,
 462 **24**, 2358-2367.

463 Rayner, N. A., P. Brohan, D. E. Parker, C. K. Folland, J. J. Kennedy, M. Vanicek, T. Ansell,
 464 and S. F. B. Tett, 2006: Improved analyses of changes and uncertainties in sea surface
 465 temperature measured in situ since the mid-nineteenth century: The HadSST2 dataset. *J.*
 466 *Climate*, **19**, 4464-69.

467 Reichler, T., and J. Kim, 2008: How well do coupled models simulate today's climate? *Bull.*

468 *Amer. Meteor. Soc.*, **89**, 3033-11.

469 Rossow, W. B., and R. A. Schiffer, 1999: Advances in understanding clouds from ISCCP.

470 *Bull. Amer. Meteor. Soc.*, **80**, 2261-2287.

471 Santer, B. D., and co-authors, 2009: Incorporating model quality information in climate

472 change detection and attribution studies. *Proc. Natl. Acad. Sci. USA*, **106**, 14778–

473 14783.

474 Schlink, U., and A. Thiem, 2009: Non-negative matrix factorization for the identification of

475 patterns of atmospheric pressure and geopotential for the Northern Hemisphere. *Int. J.*

476 *Climatol.*, **30**, 909-925.

477 Shiogama, H., S. Emori, N. Hanasaki, M. Abe, Y. Masutomi, K. Takahashi, and T. Nozawa,

478 2011: Observational constraints indicate risk of drying in the Amazon basin. *Nat.*

479 *Commun.*, 2:253. doi: 10.1038/ncomms1252.

480 Solomon, S., D. Qin, M. Manning, M. Marquis, K. Averyt, M. M. B. Tignor, H. L. Miller Jr.,

481 and Z. Chen, 2007: Climate Change 2007: The Physical Science Basis, Cambridge

482 University Press, 996 pp.

483 Talagrand, O., 1981: A study of the dynamics of four-dimensional data assimilation. *Tellus*,

484 **33**, 4360.

485 Uppala, S. M., and co-authors, 2005: The ERA-40 reanalysis. *Q. J. R. Meteorol. Soc.*, **131**,

486 2961–3012.

487 Ward, J. H. Jr., 1967: Hierarchical grouping to optimize an objective function. *J. Amer. Stat.*

488 *Assoc.*, **58**, 236-244.

489 Whetton, P., I. Macadam, J. Bathols, and J. O’Grady, 2007: Assessment of the use of

490 current climate patterns to evaluate regional enhanced greenhouse response patterns of

491 climate model. *Geophys. Res. Lett.*, **34**, L14701, doi:10.1029/2007GL030025.

492 Xie, P., and P. A. Arkin, 1997: Global precipitation: A 17-year monthly analysis based on

493 gauge observations, satellite estimates, and numerical model outputs. *Bull. Amer.*

Meteor. Soc., **78**, 2539-2558.

Yokoi, S., and Y. N. Takayabu, 2009: Multi-model Projection of Global Warming Impact on Tropical Cyclone Genesis Frequency over the Western North Pacific. *J. Meteor. Soc. Japan*, **87**, 525-538.

Yokoi, S., Y. N. Takayabu, K. Nishii, H. Nakamura, H. Endo, H. Ichikawa, T. Inoue, M. Kimoto, Y. Kosaka, T. Miyasaka, K. Oshima, N. Sato, Y. Tsushima and M. Watanabe, 2011: Application of cluster analysis to climate model performance metrics. *J. Appl. Meteorol. Climatol.*, **50**, 1666-1675.

List of Figures

Fig. 1 Inter-model standard deviations of C_{ij} , defined in (1). See text for details.

Fig. 2 Comparison among basis vectors, represented by arrows **a** and **b**, obtained by (a) PCA and (b) NMF, in a phase space of a hypothetical two-variable coordinate system. Ovals in each panel denote distributions of the points that represent errors (biases) of the individual models.

Fig. 3 Dendrogram of the cluster analysis that is applied to C_{ij} defined in (1).

Fig. 4 (a) Loadings of individual variables (abscissa) for the leading PCA modes, and (b) its scores for individual models (abscissa). (c, e) Same as in (a), but for second and third modes, respectively. (d, f) Same as in (b), but for the second and third modes, respectively.

Fig. 5 (a) First column vector of Q that represents weights of individual variables (abscissa) for measure of the reproducibility of the CMIP3 models (abscissa) as represented by (b) the column vectors of P both for the first mode of NMF with $R=2$. (c, d) As in (a, b), respectively, but for the second mode.

Fig. 6 Ranking (ordinate) of the CMIP3 models (abscissa) determined through general metrics based on the CPI (square), cluster analysis (rhombus), NMF (downward-pointing triangle), MCPI (upward-pointing triangle) and TE (rightward-pointing triangle), as indicated.

521 See text for details.

522 Fig. 7 The number of variable metrics (metrics) that are ranked as the top five (squares
523 with solid line) and bottom five (triangles with dotted line) among the models. Models
524 (abscissa) are listed in descending order according to the rank of CPI.

525 Fig. 8 Comparison of rankings of the CMIP3 models. Small squares, rhombuses and
526 double triangles denote the rankings based on CPI, horizontal pattern correlation and
527 RMSE of global-mean biases, respectively, between simulated and observed climatological
528 fields. In the evaluation of the latter two, the pattern correlations and global-mean biases for
529 single variables are first estimated, and then their rankings among the models are
530 averaged, respectively. Models (abscissa) are listed in descending order according to the
531 rank of CPI. Note that ECHO-G is not listed, whose humidity data were lost due to a
532 computer trouble.

533 Fig. 9 (a) Scatter plot between inter-model similarity of the 20c3m experiment (abscissa)
534 and that of the projected future changes (ordinate) for all possible pairs of the CMIP3
535 models. The similarity is measured by CPI that has been evaluated without specific
536 humidity. The future climate is based on the projection of the projection with the A1B
537 scenario average for 2070-2099. A line represents a regression line. (b) Correlations
538 between the inter-model similarity of the 20c3m experiment and that of the future
539 change, which is based on the same scatter plots as in (a) but for variables used in this
540 study. The last one “CPI w/o T” denotes CPI evaluated without T300, T500, T600, T700,
541 and T850.

542

543

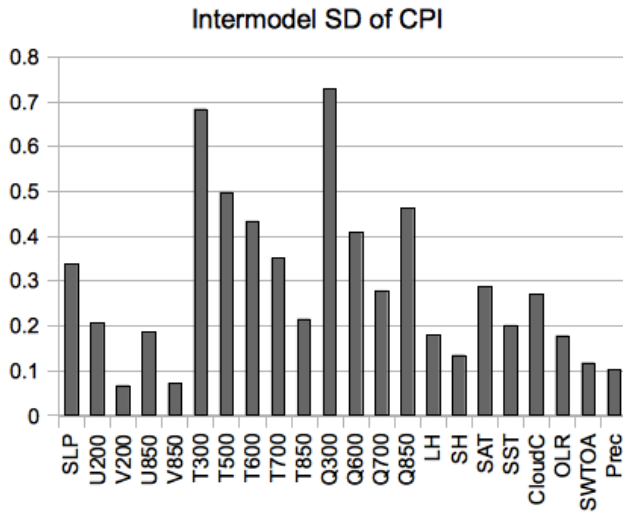


Fig. 1 Inter-model standard deviations of C_{ij} , defined in (1). See text for details.

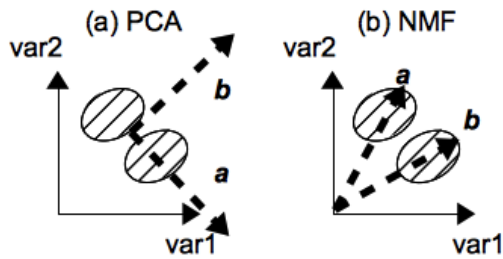


Fig. 2 Comparison among basis vectors, represented by arrows **a** and **b**, obtained by (a) PCA and (b) NMF, in a phase space of a hypothetical two-variable coordinate system. Ovals in each panel denote distributions of the points that represent errors (biases) of the individual models.

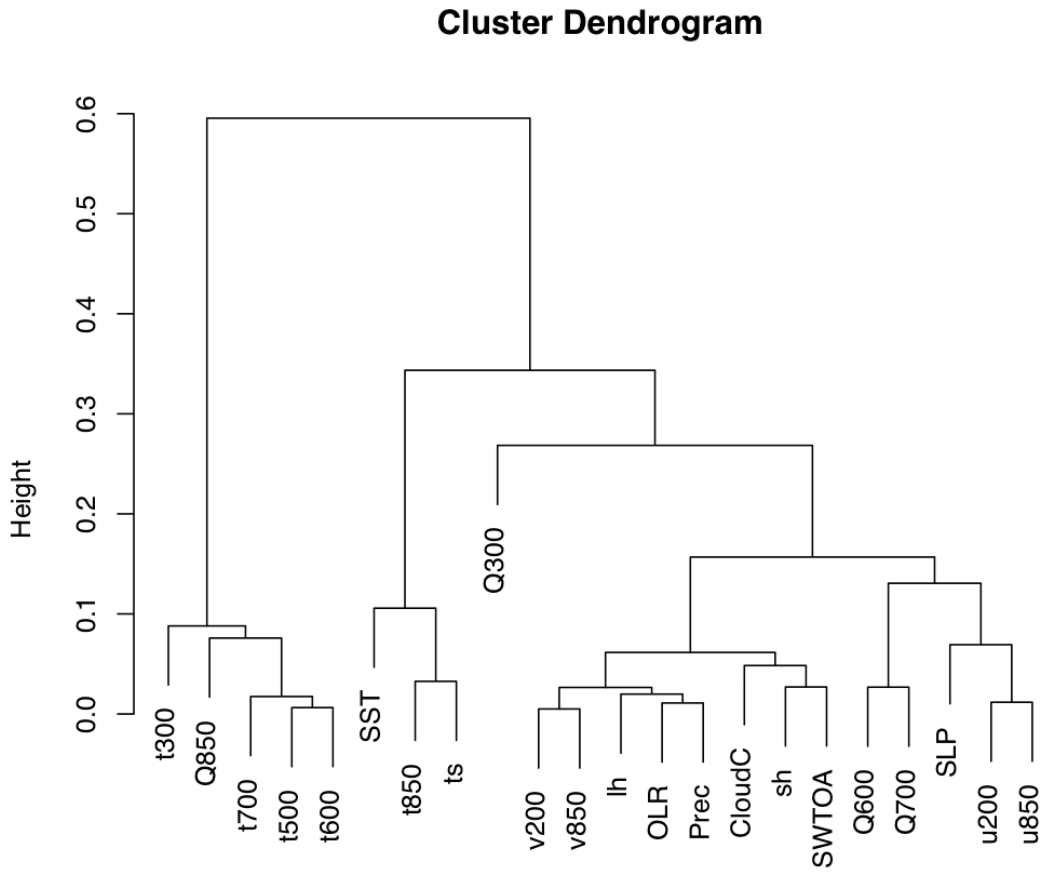


Fig. 3 Dendrogram of the cluster analysis that is applied to C_{ij} defined in (1).

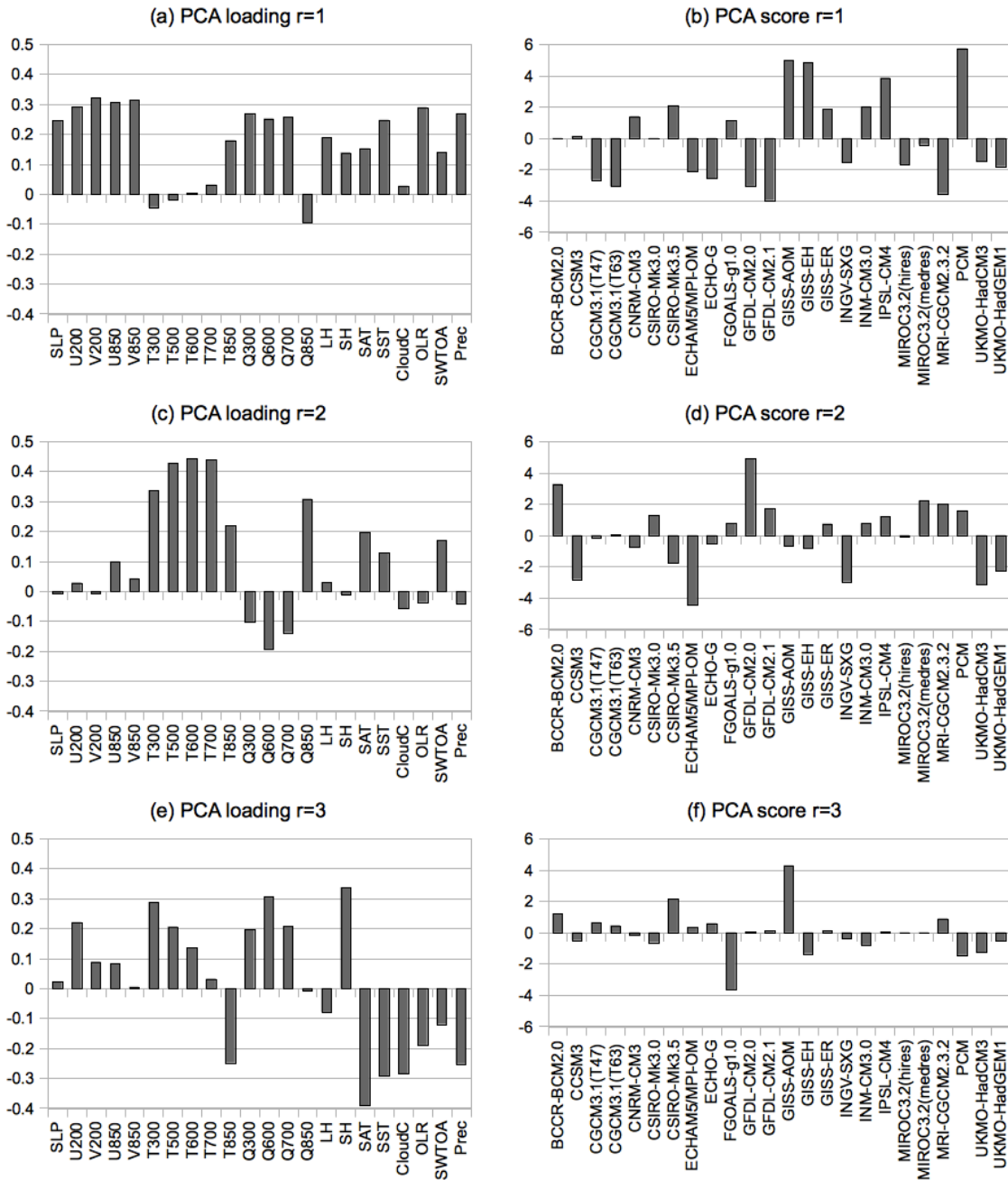


Fig. 4 (a) Loadings of individual variables (abscissa) for the leading PCA modes, and (b) its scores for individual models (abscissa). (c, e) Same as in (a), but for second and third modes, respectively. (d, f) Same as in (b), but for the second and third modes, respectively.

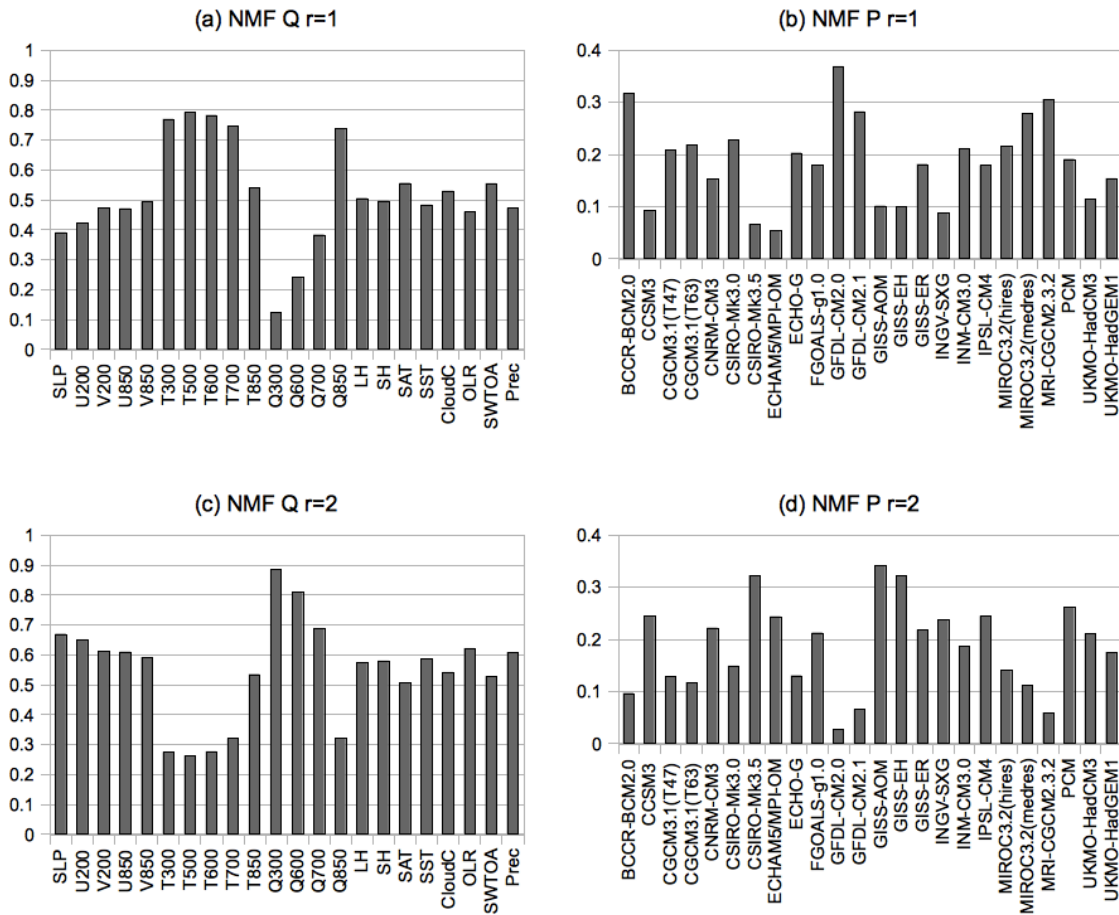


Fig. 5 (a) First column vector of Q that represents weights of individual variables (abscissa) for measure of the reproducibility of the CMIP3 models (abscissa) as represented by (b) the column vectors of P both for the first mode of NMF with $R=2$. (c, d) As in (a, b), respectively, but for the second mode.

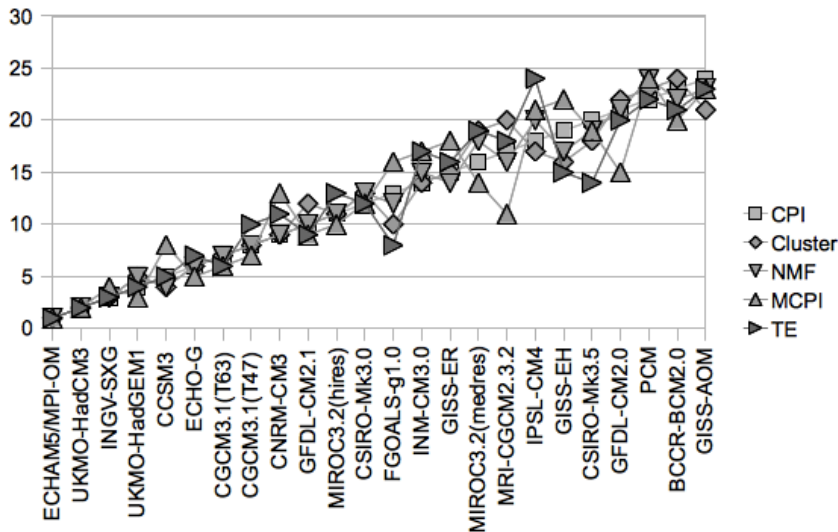


Fig. 6 Ranking (ordinate) of the CMIP3 models (abscissa) determined through general metrics based on the CPI (square), cluster analysis (rhombus), NMF (downward-pointing triangle), MCPI (upward-pointing triangle) and TE (rightward-pointing triangle), as indicated. See text for details.

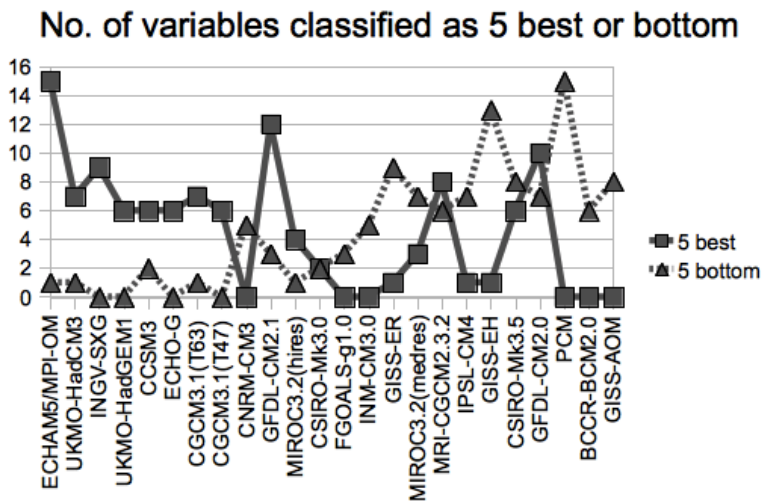


Fig. 7 The number of variable metrics (metrics) that are ranked as the top five (squares with solid line) and bottom five (triangles with dotted line) among the models. Models (abscissa) are listed in descending order according to the rank of CPI.

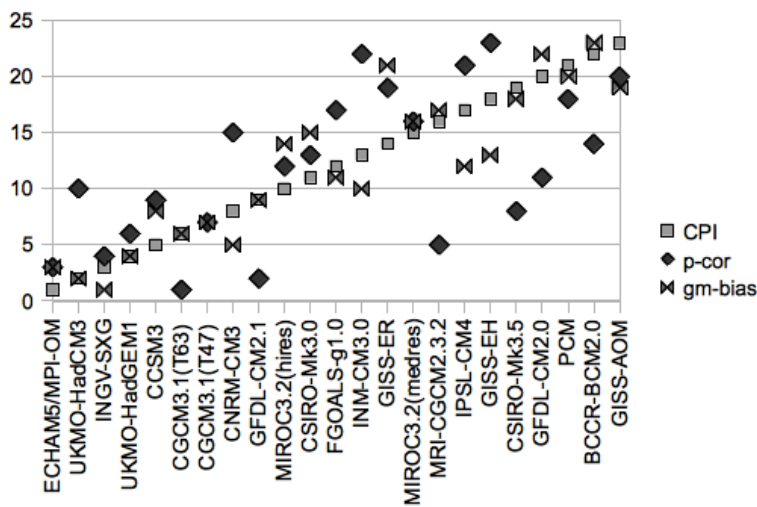


Fig. 8 Comparison of rankings of the CMIP3 models. Small squares, rhombuses and double triangles denote the rankings based on CPI, horizontal pattern correlation and RMSE of global-mean biases, respectively, between simulated and observed climatological fields. In the evaluation of the latter two, the pattern correlations and global-mean biases for single variables are first estimated, and then their rankings among the models are averaged, respectively. Models (abscissa) are listed in descending order according to the rank of CPI. Note that ECHO-G is not listed, whose humidity data were lost due to a computer trouble.

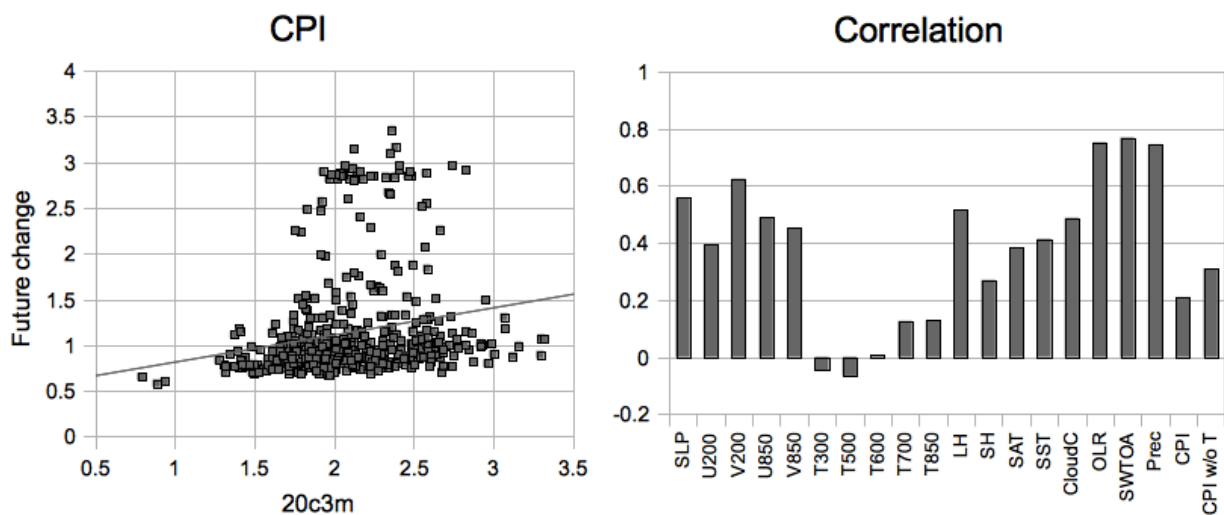


Fig. 9 (a) Scatter plot between inter-model similarity of the 20c3m experiment (abscissa) and that of the projected future changes (ordinate) for all possible pairs of the CMIP3 models. The similarity is measured by CPI that has been evaluated without specific humidity. The future change is based on the difference between the averages for the two periods, one for 2070-2099 of the A1B scenario experiment and the other for 1970-1999 of the 20C3M experiment. A line represents a regression line. (b) Correlations between the inter-model similarity of the 20c3m experiment and that of the future change, which is based on the same scatter plots as in (a) but for variables used in this study. The last one “CPI w/o T” denotes CPI evaluated without T300, T500, T600, T700, and T850.

List of Tables

Table 1 List of used variables and reference dataset. JRA25 is for Japan Re-Analysis (Onogi et al. 2007). HadSST2 is for the Second Hadley Centre Sea Surface Temperature dataset (Rayner et al. 2006). ISCCP is for the International Satellite Cloud Climatology Project (Rossow and Schiffer 1999). ERBE is for Earth Radiation Budget Experiment (Barkstrom et al. 1989). CMAP is for the CPC Merged Analysis of Precipitation (Xie and Arkin 1997).

677 Table 1 List of used variables and reference dataset. JRA25 is for Japan Re-Analysis
678 (Onogi et al. 2007). HadSST2 is for the Second Hadley Centre Sea Surface Temperature
679 dataset (Rayner et al. 2006). ISCCP is for the International Satellite Cloud Climatology
680 Project (Rossow and Schiffer 1999). ERBE is for Earth Radiation Budget Experiment
681 (Barkstrom et al. 1989). CMAP is for the CPC Merged Analysis of Precipitation (Xie and
682 Arkin 1997).

Variable	Description	Reference	Period
SLP	Sea level pressure	JRA25	1979-1999
U200	200-hPa zonal wind	JRA25	1979-1999
U850	850-hPa zonal wind	JRA25	1979-1999
V200	200-hPa meridional wind	JRA25	1979-1999
V850	850-hPa meridional wind	JRA25	1979-1999
T300	300-hPa air temperature	JRA25	1979-1999
T500	500-hPa air temperature	JRA25	1979-1999
T600	600-hPa air temperature	JRA25	1979-1999
T700	700-hPa air temperature	JRA25	1979-1999
T850	850-hPa air temperature	JRA25	1979-1999
Q300	300-hPa Specific humidity	JRA25	1979-1999
Q600	600-hPa Specific humidity	JRA25	1979-1999
Q700	700-hPa Specific humidity	JRA25	1979-1999
Q850	850-hPa Specific humidity	JRA25	1979-1999
LH	Surface latent heat flux	JRA25	1979-1999
SH	Surface sensible heat flux	JRA25	1979-1999
SAT	Surface (2m) air temperature	JRA25	1979-1999
SST	Sea surface temperature	HadSST2	1979-1999
CloudC	Cloud cover	ISCCP-D2	1984-1999
OLR	Outgoing longwave radiation	ERBE	Feb. 1985 - Feb. 1990
SWTOA	Reflected shortwave radiation	ERBE	Feb. 1985 - Feb. 1990
Prec	Total precipitation	CMAP	1979-1999

683

684

685