Evaluation of CMIP6 Models Toward Dynamical Downscaling Over Eight CORDEX Domains

Meng-Zhuo Zhang
Nanjing University
Zhongfeng Xu (xuzhf@tea.ac.cn)
Institute of Atmospheric Physics Chinese Academy of Sciences https://orcid.org/0000-0002-1274-6438
Ying Han
Institute of Atmospheric Physics Chinese Academy of Sciences
Weidong Guo
Nanjing University

Research Article

Keywords: Multivariable integrated evaluation, Model performance, Model interdependency, CMIP6, CORDEX

Posted Date: September 29th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-911581/v1

Evaluation of CMIP6 models toward dynamical downscaling over eight CORDEX domains

Meng-Zhuo Zhang[1], Zhongfeng Xu[2], Ying Han[2], Weidong Guo[1]

[1] School of Atmospheric Sciences, Nanjing University, Nanjing, China

[2] CAS Key Laboratory of Regional Climate and Environment for Temperate East Asia, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029, China

Correspondence to: Zhongfeng Xu (xuzhf@tea.ac.cn)

Abstract

Both reliability and independence of global climate model (GCM) simulation are essential for model selection to generate a reasonable uncertainty range of dynamical downscaling simulations. In this study, we evaluate the performance and interdependency of 37 GCMs from the Coupled Model Intercomparison Project Phase 6 (CMIP6) in terms of seven key large-scale driving fields over eight CORDEX domains. A multivariable integrated evaluation method is used to evaluate and rank the models' ability to simulate multiple variables in terms of their climatological mean and interannual variability. The results suggest that the model performance varies considerably with seasons, domains, and variables evaluated, and no model outperforms in all aspects. However, the multi-model ensemble mean performs much better than any individual model. Among 37 CMIP6 models, the MPI-ESM1-2-HR, FIO-ESM-2-0, and MPI-ESM1-2-LR rank top three due to their overall good performance across all domains. To measure the model interdependency in terms of multiple fields, we define the similarity of multivariate error fields between pairwise models. Our results indicate that the dependence exists between most of the CMIP6 models, and the models sharing the same idea or/and concept generally show less independence. Furthermore, we hierarchically cluster the top 15 models based on the similarity of multivariate error fields to facilitate the model selection. Our evaluation can provide useful guidance on the selection of CMIP6 models based on their

performance and relative independence, which helps to generate a more reliable ensemble of dynamical downscaling simulations with reasonable inter-model spread.

Keywords Multivariable integrated evaluation · Model performance · Model interdependency · CMIP6 · CORDEX

Declarations

Funding: The study was supported jointly by the National Key Research and Development Program of China (2017YFA0603803) and the National Science Foundation of China (41675105, 42075170, 42075152). This work was also supported by the Jiangsu Collaborative Innovation Center for Climate Change.

Conflicts of interest/Competing interests: The authors declare no conflicts of interest.

Availability of data and material: The data used in this study is available for open access.

Code availability: Not applicable.

Authors' contributions: Meng-Zhuo Zhang, Zhongfeng Xu, and Ying Han designed the study. Meng-Zhuo Zhang performed the analysis and wrote the paper. All authors read and approved the final manuscript.

1 Introduction

Dynamical downscaling is one of the important approaches to generate the regional climate projections using the global climate model (GCM) outputs as initial and lateral boundary conditions (Giorgi et al. 2009; Buontempo et al. 2015; Dai et al. 2020). To advance and coordinate the science and application of the regional climate downscaling, the COordinated Regional Downscaling EXperiment (CORDEX) was implemented under the framework of World Climate Research Program. Using regional climate model (RCM) or empirical statistical downscaling driven by the Coupled Model Intercomparison Project (CMIP) models, the CORDEX can provide more detailed and accurate representation of regional climate, especially for extreme events (Gutowski et al. 2016). Regional climate projections are fundamental for the studies of vulnerability and impacts assessment, and even the development of suitable adaptation and mitigation strategies at the national level (Giorgi et al. 2006, 2016; Ruane et al. 2016; Mishra et al. 2018).

Over past decades, efforts have been made to evaluate the model performance in terms of the large-scale driving variables on the regional scale (e.g., Brands et al. 2013; Elguindi et al. 2014; July et al. 2015). However, most of the studies paid more attention to the model performance in simulating each of the evaluated variables rather than the overall performance of all variables. For example, Elguindi et al. (2014) explored the reliability of CMIP Phase 5 (CMIP5) models in simulating the climatology of both temperature and precipitation over a subset of the CORDEX

88 domains. These studies found that most models exhibit varying performance on the
89 regions, seasons, and variables, and few models show perfect performance in all
90 aspects. In addition, previous studies also showed that the GCMs generally suffer
91 from the sizable bias in simulating driving variables over some regions, such as the
92 air temperature and humidity over the Tibetan Plateau (Xu et al. 2017; Zhu and Yang
93 2020), the sea surface temperature over the tropical central and eastern Pacific (Tian
94 and Dong 2020), and the lower level vector wind over the Asian-Australian monsoon
95 region (Huang et al. 2019). In general, GCMs still have difficulties in reproducing
96 the features of large-scale driving fields.

97 The bias in the large-scale driving fields can propagate into RCM through the
98 underling and lateral boundary of the RCM. Consequently, the downscaling
99 simulations on regional climate strongly depend on the quality of the large-scale
100 driving fields provided by GCMs (Wu et al. 2005; Plavcová and Kyselý 2012; Xu
101 and Yang 2012, 2015; Dosio et al. 2015; McSweeney et al. 2015; Kebe et al. 2017;
102 Rocheta et al. 2020). Therefore, the reliability of GCM simulation is an essential
103 factor in model selection for generating reliable downscaling simulations. So far,
104 more than 100 models are available in the CMIP Phase 6 (CMIP6) (Eyring et al.
105 2016). Each model has its own biases and uncertainties. Although skillful simulation
106 of the historical period might not guarantee skillful projections of future climate for
107 the same GCM, the lack of skill in historical simulation might translate to the lack of
108 skill in future simulation (Knutti 2008). Thus, the historical GCM simulations still

warrant in-depth evaluation to select GCMs and generate more reliable future climate projections.

In addition to model performance, model independence is also of great importance for generating a reasonable uncertainty range of future climate projections. Owing to 'sharing relationship' in terms of code and concept, neither different models in one generation nor different versions of one model are independent of each other (Knutti et al. 2010a, 2010b, 2012; Sanderson et al. 2015a, b). The model interdependency in an ensemble would reduce its effective degree of freedom, consequently leading to an underestimation of the uncertainty range of climate projections. Therefore, evaluating model interdependency and skillfully selecting GCMs can help to minimize the model interdependency in an ensemble. Independent large-scale forcing ensures that its downscaling simulations can sufficiently represent the future climate uncertainty (Bishop and Abramowitz 2013; Mendlik and Gobiet 2016; Herger et al. 2018).

Moreover, which variables are taken into evaluation needs cautious consideration as well, because the evaluation results are variable-dependent. Previous studies usually chose several key variables from the large-scale driving fields, i.e., temperature, sea level pressure, and wind fields (e.g., Elguindi et al. 2014; Jury et al. 2015). Jury et al. (2015) suggested that the evaluation should widely cover the driving variables because the evaluation of a few variables is inadequate as a reference for selecting suitable driving GCMs toward downscaling. On the other hand, the large-scale driving variables, such as sea level pressure, geopotential height,

and wind fields, may depend on each other to a certain extent. Consequently, evaluating all of the large-scale variables would contain redundant information. Thus, the selection of relatively independent variables is also worthy of investigation.

This paper aims to present a comprehensive evaluation on the performance of 37 CMIP6 models in simulating multiple large-scale driving variables over eight CORDEX domains. We also assess the model interdependency in terms of the climatological mean and interannual variability of multiple variables. This evaluation would help to select CMIP6 models towards dynamical downscaling over CORDEX domains.

In Section 2, we briefly introduce datasets used in this study. Section 3 describes the statistics of model evaluation. Section 4 illustrates the selection of evaluated variables. Sections 5 and 6 present the evaluation results of model performance and model interdependency, respectively. Conclusion and discussion are given in Section 7.

2 Data

In the study, we use monthly mean data of the first ensemble run in historical experiments from 37 CMIP6 models during the period of 1979–2013 (Table 1). The variables used include two surface variables (i.e., surface temperature and sea level pressure) and four atmosphere variables (i.e., air temperature, wind fields, specific humidity, and geopotential height) at four different pressure levels (Table 2).

151 Especially, wind fields consisting of zonal and meridional wind components are
152 regarded as a vector variable.

153 As shown in Fig. 1, the evaluation is carried out in eight CORDEX domains
154 including Central America (C-AM), North America (N-AM), Europe (EURO), South
155 Asia (S-AS), East Asia (E-AS), Australasia (AUS), Middle East North Africa
156 (MENA), and South East Asia (SEA). The low-level atmospheric variables in some
157 of the CMIP6 models contain missing values over the high topographic regions
158 within the CORDEX domains. To make model simulations comparable with each
159 other, a common mask across 37 CMIP6 models (the grey shades in Fig. 1) is
160 generated for the atmosphere variables in each domain. Note that these masks are not
161 applied to the surface variables. In addition, eight CORDEX domains are defined on
162 the rotated coordinates, and the horizontal grid spacing of the SEA domain is $0.22 \times$
163 0.22 while that of the other seven domains is 0.44×0.44.

164 Owing to the high resolution of the CORDEX domains, we use the high-resolution
165 reanalysis dataset—the fifth generation of the European Centre for Medium-Range
166 Weather Forecast atmosphere reanalysis (ERA5) as the reference in the evaluation.
167 In addition, the Japan Meteorological Agency and the Central Research Institute of
168 Electric Power Industry Reanalysis-55 (JRA55) and the multi-model ensemble mean
169 (MME) of the 37 CMIP6 models are evaluated at the same time. To facilitate the
170 inter-comparison between models, all model and reanalysis data have been bilinearly
171 interpolated into the CORDEX coordinates before evaluation.

3 Statistical methods

3.1 Model skill scores of the individual variable and correlation analysis

In the analysis of the dependence between model performance in simulating different variables (Sect. 4), we use two skill scores representing the models' ability to simulate individual variables. Therein, the model skill scores S_{v1} and S_{v2} proposed by Xu et al. (2016) are used for a vector variable:

$$S_{v1} = \frac{4(1+R_v)}{(\frac{L_A}{L_O}+\frac{L_O}{L_A})^2(1+R_0)} \quad (1)$$

$$S_{v2} = \frac{4(1+R_v)^4}{(\frac{L_A}{L_O}+\frac{L_O}{L_A})^2(1+R_0)^4} \quad (2)$$

where L_A (L_O) is the root mean square length (RMSL) of the vector field for the model (reference):

$$L_A = \sqrt{\frac{\sum_{i=1}^{N}|\boldsymbol{A}_i|^2}{N}} \text{ and } L_O = \sqrt{\frac{\sum_{i=1}^{N}|\boldsymbol{O}_i|^2}{N}} \quad (3)$$

R_v is the vector similarity coefficient between the model and the reference:

$$R_v = \frac{\sum_{i=1}^{N}\boldsymbol{A}_i\cdot\boldsymbol{O}_i}{\sqrt{\sum_{i=1}^{N}|\boldsymbol{A}_i|^2}\cdot\sqrt{\sum_{i=1}^{N}|\boldsymbol{O}_i|^2}} \quad (4)$$

Here, $\boldsymbol{A}$ and $\boldsymbol{O}$ represent two vector fields derived from the model and the reference, respectively. Each vector field consists of N discrete vectors in time and/or space. RMSL measures the magnitude the vector field, and R_v describes the pattern similarity of two vector fields, ranging from -1 to 1. Thus, both S_{v1} and S_{v2} take RMSL and R_v into account, but S_{v1} puts more emphasis on amplitude simulation while S_{v2} gives more attention to the simulation of pattern similarity. R_0 is the maximum R_v attainable, and we set it to 1 in the study. S_{v1} and S_{v2} increase

192 monotonically with the model performance, and the closer to 1, the better
193 performance they indicate.

194 In addition, when $\boldsymbol{A}$ and $\boldsymbol{O}$ degrade to the scalar field, RMSL and R_v become the
195 root mean square (rms) and the uncentered correlation coefficient (R), respectively.
196 In this case, S_{v1} and S_{v2} can be applied to the evaluation of a scalar variable, termed
197 S_1 and S_2:

198 $$S_1 = \frac{4(1+R)}{(\frac{rms_A}{rms_O}+\frac{rms_O}{rms_A})^2(1+R_0)} \quad (5)$$

199 $$S_2 = \frac{4(1+R)^4}{(\frac{rms_A}{rms_O}+\frac{rms_O}{rms_A})^2(1+R_0)^4} \quad (6)$$

200 S_1 and S_2 are equivalent to the skill scores proposed by Taylor (2001), except that the
201 original scalar field is used here instead of the anomalous scalar field in Taylor's
202 definition.

203 Spearman rank correlation is used to measure the dependence of model
204 performances between different variables. We calculate the cross-correlation
205 between the skill scores derived from two different variables across the 37 CMIP6
206 models. A larger rank correlation coefficient indicates a closer model skill in the
207 simulation of two variables.

208 **3.2 Multivariable integrated skill score**

209 We use the multivariable integrated skill score (MISS) proposed by Zhang et al.
210 (2021) to measure the overall performance of a climate model in simulating multiple
211 fields. MISS is defined as:

$$MISS = \left\{F + 1 - \left[\frac{1}{M}\sum_{m=1}^{M}(R_m^* - 1)^2 + 2 \cdot (1 - R_v)\right]\right\}/(F + 1)\ ,$$

$$where\ R_m^* = \begin{cases} \frac{L_{Am}}{L_{Om}}, & \frac{L_{Am}}{L_{Om}} \le 1 \\ \frac{L_{Om}}{L_{Am}}, & \frac{L_{Am}}{L_{Om}} > 1 \end{cases} \quad (7)$$

where M is the number of individual variables. L_{Am} (L_{Om}) is the RMSL of the m-th variable for the model (reference). R_v measures the pattern similarity between the model and the reference in terms of multiple fields. Here, R_v is computed with Eq. (4), in which the vector field is composed of multiple normalized variables (Xu et al. 2017).

MISS takes the models' ability to simulate the amplitude and the pattern similarity of various variables into consideration simultaneously. F is a weight factor to adjust the relative importance of the amplitude and the pattern similarity in MISS, which is set to 2 in our study. MISS varies monotonically with the overall model performance in simulating multiple fields, usually ranging from 0 to 1. MISS is equal to 1 when the model simulations are exactly the same as the reference.

3.3 Statistics for model interdependency and cluster analysis

Error correlation is one of the statistics used to measure the model interdependency (e.g., Jun et al. 2008; Collins et al. 2011; Bishop and Abramowitz 2013). However, this statistic is only valid for the scalar field. To measure the model interdependency in terms of multiple fields, we define a new statistic—the Multivariable Error Similarity coefficient (MvES) inspired by the multivariable integrated evaluation (MVIE) method (Xu et al. 2017).

Assume that there are three datasets derived from model A, model B, and reference O, respectively. Each dataset includes *M* fields consisting of either scalar or vector fields. Following the idea of MVIE, each scalar (vector) field is normalized by the rms (RMSL) value of the corresponding observed variable. These normalized *M* fields are grouped into a multi-dimensional vector field for model $\boldsymbol{A}$, model $\boldsymbol{B}$, and reference $\boldsymbol{O}$, respectively. Then, the multivariate error fields are defined by the difference between the model and the reference:

$$\boldsymbol{A_j^{err}} = \boldsymbol{A_j} - \boldsymbol{O_j} = (a_{1j}^{err}, a_{2j}^{err}, \ldots, a_{Dj}^{err}); \quad j = 1,2,\ldots,N$$

$$\boldsymbol{B_j^{err}} = \boldsymbol{B_j} - \boldsymbol{O_j} = (b_{1j}^{err}, b_{2j}^{err}, \ldots, b_{Dj}^{err}); \quad j = 1,2,\ldots,N$$

$\boldsymbol{A^{err}}$ and $\boldsymbol{B^{err}}$ are composed of *N* discrete vectors in time and/or space. *D* is the dimension of the multi-dimensional vector field, which sums the dimensions of *M* fields. MvES between model A and model B can be written as:

$$\text{MvES} = \frac{\sum_{i=1}^{D}\sum_{j=1}^{N} a_{ij}^{err} \cdot b_{ij}^{err}}{\sqrt{\sum_{i=1}^{D}\sum_{j=1}^{N}(a_{ij}^{err})^2} \cdot \sqrt{\sum_{i=1}^{D}\sum_{j=1}^{N}(b_{ij}^{err})^2}} \tag{8}$$

MvES represents the similarity between two sets of multivariate error fields derived from pairwise models. It ranges from 1 to -1. Larger MvES means that the errors of multiple variables derived from model A are more dependent on those derived from model B and vice versa.

The hierarchical clustering analysis (HCA) is further carried out based on MvES to separate various CMIP6 models into groups. The HCA constructs a hierarchy of sets of groups, each level of which is formed by merging one pair from the collection of previously defined groups (Wilks 2011). As the basis for merging, the distance

between two models is measured by MvES in our study. Then, group-to-group distance is defined using the average of the distance between all possible pairs of models in two groups being compared:

$$d_{G_1,G_2} = \frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \mathrm{MvES}_{i,j}}{n_1 \cdot n_2} \quad (9)$$

where G_1 and G_2 represent two individual groups with n_1 and n_2 models, respectively. Thus, d_{G_1,G_2} measures the average dependence of models between two groups, and larger d_{G_1,G_2} indicates that two groups are closer to each other.

Assuming there are n models, each one is regarded as a group at the beginning. The first step of the HCA is to find two groups with the highest MvES among the n groups and to combine them into a new group. The distance between this new group and the other groups is computed with Eq. (9) consequently. Then, two groups with the highest average MvES among the $n-1$ groups will be merged and form the second group, and group-to-group distance is updated accordingly. This process continues until n models have been aggregated into a single group. We use a dendrogram to illustrate the progress and intermediate results of the HCA.

4 Selection of evaluated variables

Figures 2 and 3 show the relationship of model performance in simulating various key variables. The filling color indicates the number of domains over which the correlation coefficient of model skill scores between two variables reaches the significance level of 0.05. As shown in Figs. 2 and 3, the number of significant

correlations varies with different pairs of variables, seasons, statistical characteristics of historical climate (i.e., climatological mean and interannual variability), and skill scores. However, model abilities are generally closely related to each other in terms of the simulation of some variables, e.g., Q700–Q500, T850–T700–T500–UV200, SLP–UV850–UV700–UV500–UV200–UV700, and SLP–Z850–Z700–Z500–Z200. Among these pairwise variables, there are generally 5–8 domains where model abilities show significant correlation with each other for both the climatological mean and interannual variability (Figs. 2, 3). Note that the models' ability to simulate the specific humidity is relatively independent of the other variables.

Based on the analysis above, we can select some variables that are relatively independent of each other. Considering that the water vapor is mostly concentrated in the lower troposphere together with the significant correlation between Q500 and Q700, we select and evaluate Q700 and Q850 to measure the model performance in simulating specific humidity. Given the importance of the wind in dynamical downscaling simulations (Rocheta et al. 2020), we select UV200 and UV850 although the wind fields at different levels are generally correlated with each other. Besides, we also select T200 and T500 according to the independence of air temperatures at different levels. Geopotential height is excluded because it is closely correlated with the other variables, e.g., T850–Z850/Z700 in the climatological mean field (Fig. 2), UV850–Z200/Z500 in the interannual variability field (Fig. 3). Sea level pressure is excluded from the evaluation because of the close correlation between SLP and UV850. Instead, surface temperature is included due to its

292 importance in regional climate simulation and its independence from other variables.
293 Therefore, the model evaluations in the following sections focus on seven selected
294 variables (shown in the red font in Figs. 2, 3), i.e., surface temperature, 850-hPa and
295 700-hPa specific humidity, 850-hPa and 200-hPa vector wind fields as well as
296 500-hPa and 200-hPa air temperatures. These variables can reasonably represent the
297 overall performance of a CMIP6 model in simulating multiple large-scale driving
298 fields of dynamical downscaling simulations.

299 # 5 Intercomparison of CMIP6 model performance

300 ## 5.1 Climatological mean

301 In this section, we use MISS as the indicator to evaluate the model performance in
302 terms of multiple fields. Model performance in simulating the climatological mean of
303 seven key variables selected in Sect. 4 is evaluated in different CORDEX domains
304 and seasons against the ERA5 data (Fig. 4). JRA55 and MME are also included in
305 the evaluation in addition to the 37 CMIP6 models. We also rank model performance
306 in simulating climatological mean based on MISS. JRA55 ranks first in all domains
307 and all seasons. MME generally shows much better performance than the individual
308 CMIP6 model, which is likely due to the compensation of multi-model errors. The
309 performances of CMIP6 models depend on seasons and domains. No model performs
310 best in all domains and all seasons. For example, CIESM ranks first of 37 CMIP6

311 models in the C-AM domain during winter and spring, while it ranks out of the top
312 15 in the S-AS domain throughout the year.

313 The rightmost column in Fig. 4 further summarizes the overall model rank based
314 on the sum of 32 ranks (8 CORDEX domains × 4 seasons) of each model, termed the
315 total rank. The results show that EC-Earth3-Veg, AWI-CM-1-1-MR, FIO-ESM-2-0,
316 and EC-Earth3 rank top four out of 37 CMIP6 models, indicating that these models
317 show an overall better climatology than the other models in eight CORDEX domains
318 throughout the year. In addition, some models from the same institution show close
319 performance with each other, e.g., BCC-CSM2-MR vs. BCC-ESM1 in the MENA
320 domain, EC-Earth3 vs. EC-Earth3-Veg in the EURO domain, GISS-E2-1-G vs.
321 GISS-E2-1-H in the E-AS domain, and INM-CM4-8 vs. INM-CM5-0 in the N-AM
322 domain.

323 **5.2 Interannual variability**

324 The interannual variability is measured by the standard deviation of the
325 year-to-year time series. Thereafter, MISS is computed based on the spatial field of
326 the interannual standard deviation of multiple variables (Fig. 5). Similar to the
327 evaluation results of the climatological mean, the ranks of the 37 CMIP6 models are
328 still season- and domain-dependent, and no model performs best in all cases.
329 According to the total ranking, the top four models are KACE-1-0-G, FIO-ESM-2-0,
330 MPI-ESM1-2-HR, and MRI-ESM2-0, with overall better performance in simulating
331 interannual variability of multiple variables in eight CORDEX domains throughout

the year. Notably, JRA55 shows more consistent interannual variability with ERA5 than the CMIP6 models across almost all domains and seasons except for the MENA domain during spring and autumn. Such a difference between JRA55 and ERA5 is mainly induced by the inconsistency of 850-hP specific humidity in the MENA domain, where JRA55 shows a stronger interannual variability than ERA5. Besides, MME still performs much better than most models and is even more consistent with ERA5 than JRA55 in the MENA domain during spring and autumn.

Comparison of the same model in Figs. 4 and 5 indicates that a model with the good ability to simulate climatological mean does not guarantee a good ability to simulate interannual variability. For example, EC-Earth3 ranks top four out of 37 CMIP6 models in reproducing the climatological mean (Fig. 4). However, it shows poor ability to simulate interannual variability (ranking 29th) (Fig. 5). Thus, model evaluation should take into account both climatological mean and interannual variability simultaneously.

5.3 Inter-model spread

To further investigate the inter-model spread of model performance in different domains and seasons, a box plot is shown in Fig. 6 based on the MISSs of 37 CMIP6 models. Note that the MISSs for the climatological mean are mostly higher than those for interannual variability (Fig. 6c). Such a difference is especially apparent in summer compared with the other three seasons over all domains except for the AUS domain (Fig. 6c). Meanwhile, the MISSs of interannual variability also exhibit

353 greater uncertainty ranges compared with those of climatological mean in almost all
354 domains (Fig. 6a, b). These suggest that the CMIP6 models have more difficulties to
355 simulate interannual variability than the climatological mean.

356 In terms of the climatological mean, CMIP6 models show better ability in the
357 AUS domain than the other seven domains characterized by greater MISS on average
358 and smaller inter-model spreads (Fig. 6a). CMIP6 models tend to show large
359 uncertainty with great MME bias over tropical east Pacific Ocean, tropical Indian
360 Ocean, tropical Atlantic Ocean, and the regions with complex terrains, such as the
361 vicinities of Tibetan Plateau and eastern Africa (figure not shown). In contrast, in the
362 AUS domain with flat terrain, models generally show small errors and small
363 inter-model spread. As for the interannual variability, CMIP6 models generally show
364 better performance in the domains of middle and high latitudes, i.e., N-AM and
365 EURO (Fig. 6b). Conversely, CMIP6 models show relatively poor performance in
366 the tropical domains (e.g., SEA and C-AM). In addition, the models' ability to
367 simulate interannual variability also appears to be season-dependent. For example,
368 CMIP6 models generally show poor performance (smaller MISSs) with greater
369 inter-modal spread in summer relative to the other seasons (Figs. 5, 6b).

370 **5.4 Overall performance**

371 We further evaluate and rank the overall model performance in terms of multiple
372 fields combining both climatological mean and interannual variability (Fig. 7). Note
373 that the ranking includes 37 CMIP6 models, MME, and JRA55. JRA55 is more

consistent with ERA5 than any CMIP6 models including MME, and MME outperforms all CMIP6 models. Among 37 CMIP6 models, MPI-ESM1-2-HR, FIO-ESM-2-0, and MPI-ESM1-2-LR are the top three models with a relatively better overall performance characterized by higher MISS in the CORDEX domains. Therein, MPI-ESM1-2-HR even ranks top three in six domains (i.e., C-AM, N-AM, EURO, S-AS, E-AS, and MENA). Similarly, FIO-ESM-2-0 performs well in four domains (i.e., E-AS, AUS, MENA, and SEA). Interestingly, MPI-ESM1-2-HR and FIO-ESM-2-0 show complementary performance to a certain extent. Specifically, FIO-ESM-2-0 shows good performance in the AUS and SEA domains where MPI-ESM1-2-HR shows slightly inadequate ability. Conversely, FIO-ESM-2-0 ranks outside the top three in the C-AM, N-AM, EURO, and S-AS domains, while MPI-ESM1-2-HR ranks in the top three. In addition, some models perform very well in the certain domains but have limited ability in the other domains, e.g., AWI-CM-1-1-MR, KACE-1-0-G, GFDL-ESM4, ACCESS-CM2, CAMS-CSM1-0, and MRI-ESM2-0. These models are also good candidates for driving RCM in the corresponding CORDEX domains.

6 Model interdependency and clustering

In this section, we evaluate the model interdependency in terms of multiple fields with MvES (Fig. 8). The computation of MvES takes the errors of both the climatological mean and interannual variability of seven variables in four seasons into account. Here, we only investigate the model interdependency of the top 15

CMIP6 models in each domain derived from Fig. 7. Thus, we are able to identify some independent models with relatively good performance in each domain. Generally, the dependence between the top 15 CMIP6 models is domain-dependent. For example, the top 15 CMIP6 models in the tropical domains (i.e., SEA and C-AM) tend to be more dependent on each other than those of the other domains. In the SEA and C-AM domains, the close interdependency of the top 15 models primarily results from their similar error fields in reproducing interannual variability of 500-hPa/200-hPa air temperature and surface temperature during summer and autumn (figure not shown).

We can identify some models that appear to be less similar to the others characterized by smaller MvES, such as EC-Earth3-Veg in the N-AM domain (Fig. 8b), INM-CM5-0 in the MENA domain (Fig. 8g). Whereas, the pairwise models with MvES greater than 0.5 reach 43% of the total pairs in eight domains, showing relatively high similarity in multivariate error fields. The models developed by the same institution are more similar to each other characterized by MvES greater than 0.6, e.g., E3SM-1-0 vs. E3SM-1-1, MPI-ESM1-2-LR vs. MPI-ESM1-2-HR, EC-Earth3 vs. EC-Earth3-Veg, and CESM2 vs. CESM2-WACCM. Therein, MPI-ESM1-2-LR and MPI-ESM1-2-HR are generally based on the same model configuration but submitted at a different resolution (Gutjahr et al. 2019). In other model pairs, one model is developed from the other by making some improvements and/or including additional components. For example, EC-Earth3-Veg interactively couples the dynamic global vegetation model based on EC-Earth3 (Döscher et al.

2021); E3SM-1-1 improves the simulation in terms of the carbon cycle and additionally includes the active biogeochemical process, compared with E3SM-1-0 (Burrows et al. 2020). Meanwhile, some models maintained by different institutions also show relatively large MvES (up to 0.8), e.g., MPI series models (i.e., MPI-ESM1-2-HR, MPI-ESM1-2-LR, and MPI-ESM-1-2-HAM) vs. AWI-CM-1-1-MR, ACCESS-CM2 vs. KACE-1-0-G, and CESM2 series models (i.e., CESM, CESM2-FV2, CESM2-WACCM, and CESM2-WACCM-FV2) vs. CIESM/FIO-ESM-2-0/NorESM2-MM/SAM0-UNICON. Sharing code and/or concepts may also account for this similarity. For example, both the MPI series models and AWI-CM-1-1-MR incorporated ECHAM6.3 atmosphere model and JSBACH3.20 land surface model. ACCESS-CM2 and KACE-1-0-G share the same atmosphere (MetUM-HadGEM3-GA7.1) and aerosol (UKCA-GLOMAP-mode) model. Originating from the same ancestor model (CESM1), CIESM and the CESM2 series models are developed on the common code (Danabasoglu et al. 2020; Lin et al. 2020). Thus, sharing the same components could potentially degrade the independence of the climate model simulation. Note that the dependence between the same pair of models also varies with domains. For example, MvES value between ACCESS-CM2 and AWI-CM-1-1-MR is 0.26 in the N-AM domain (Fig. 8b) against 0.59 in the AUS domain (Fig. 8f). In addition to similar parameterizations, such as in the ocean models, ACCESS-CM2 and AWI-CM-1-1-MR differ in their atmospheric, land, and ocean models (Bi et al. 2020; Semmler et al. 2020). Sharing common parameterization schemes in the ocean model may account for the close dependence

between ACCESS-CM2 and AWI-CM-1-1-MR in the domains with large coverage of the ocean, e.g., the SEA domain (0.62 MvES) and the C-AM domain (0.48 MvES) (figure not shown).

To facilitate model selection, we show the clustering of the MvES between the top 15 CMIP6 models in each domain (Fig. 9). After each clustering, models from different groups are relatively more independent of each other compared with those in the same group. The background color of a clustering step represents the number of remaining groups after clustering. Thus, we can easily select some of the models with better performance and independence. For example, if four models to select in the E-AS domain, we can find the clustering position where there are four groups left and select the model with the highest rank (best performance) in each group, i.e., KACE-1-0-G, GFDL-ESM4, CAMS-CSM1-0, and FIO-ESM-2-0 (Fig. 9e). The interdependency of these four models is less than most of the other model combinations. Moreover, they also have excellent overall performance, which rank fifth, fourth, third and first among 37 CMIP6 models, respectively. Similarly, we can select some models for other domains as well. In addition, there are some models with outstanding independence than others, which can be included in any level of model independence selections, i.e., MRI-ESM2-0 in the C-AM domain (Fig. 9a), CanESM5 in the N-AM domain (Fig. 9b), ACCESS-CM2 in the EURO domain (Fig. 9c), INM-CM5-0 in the S-AS domain (Fig. 9d), FGOALS-f3-L in the AUS domain (Fig. 9f), and KACE-1-0-G/INM-CM5-0 in the MENA domain (Fig. 9g).

7 Conclusion and discussion

This paper evaluates both the performance and interdependency of 37 CMIP6 models in simulating seven large-scale driving fields over eight CORDEX domains. Seven evaluated variables (i.e., Ts, Q850, Q700, T500, T200, UV850, and UV200) have been demonstrated as a proxy for the CMIP6 models' quality as driving fields for dynamical downscaling (Figs. 2, 3). In our evaluation, we treat multiple variables as a whole with the support of the multivariable integrated evaluation method, which distinguishes from most previous studies that focused on the model performance and/or interdependency in terms of the individual variable. Therefore, our study is expected to provide a more comprehensive and reasonable evaluation on the performance and independence of CMIP6 models from the perspective of dynamical downscaling.

Our evaluation results indicate that the model performance in terms of seven variables varies considerably with seasons, domains, and climate statistical characteristics. None of the 37 CMIP6 models performs well in all cases. However, EC-Earth3-Veg, AWI-CM-1-1-MR, FIO-ESM-2-0, and EC-Earth3 show good performance in reproducing climatological mean across most of the eight domains (Fig. 4), while in terms of interannual variability, MPI-ESM1-2-HR, MRI-ESM2-0, KACE-1-0-G, and ACCESS-CM2 perform better than the others (Fig. 5). Regarding the overall model performance that considers both climatological mean and interannual variability in four seasons and eight CORDEX domains,

MPI-ESM1-2-HR, FIO-ESM-1-0, and MPI-ESM1-2-LR rank top three out of the 37 CMIP6 models.

Previous studies suggested that an increase in model horizontal resolution helps to improve the model performance (e.g., Reichler and Kim 2008; Ranjha et al. 2016; Han et al. 2021). Therefore, we examine the relationship between the horizontal resolution and the model performance in simulating individual variables evaluated in our study (Fig. 10). The results demonstrate that the models' ability to simulate various variables is significantly correlated with their horizontal resolutions, over many CORDEX domains. In terms of the simulations of climatological mean and interannual variability, approximately 50% and 28% of variables show significant correlation on average, respectively. Among seven variables, the models' ability to simulate 850-hPa wind field is more closely related to the horizontal resolution, with significant correlations in 94% of cases. Previous studies also indicated that the higher horizontal resolution tends to generate improved simulations in the wind field and precipitation (Huang et al. 2019, 2020). Higher horizontal resolution can better resolve topography and the finer-scale dynamic and thermal process, resulting in the improved simulations for various variables.

To measure the model interdependency in terms of multiple fields, we define a new statistic that is the similarity of multivariable error fields between pairwise models (MvES). Among the 37 CMIP6 models, 39% of pairwise models show close dependence on each other with MvES greater than 0.5 (figure not shown). Therefore, the model interdependency should be taken into account in the selection of CMIP6

503 models towards downscaling. Shared code or/and concepts between pairwise models
504 would lead to the interdependency in their simulations (Fig. 8). Some model pairs
505 tend to show robust similarity across all of the eight domains, e.g., E3SM-1-0 vs.
506 E3SM-1-1, MPI series models vs. AWI-CM-1-1-MR, and ACCESS-CM2 vs.
507 KACE-1-0-G. We further hierarchically cluster the top 15 CMIP6 models in each of
508 the eight domains (Fig. 9). In each level of clustering, the interdependency of models
509 from different groups is relatively smaller than that in the same group. In this way,
510 one can easily select models from different groups to effectively reduce the
511 interdependency of the selected models.
512 The results presented in this study can provide useful guidance for model selection
513 toward dynamical downscaling simulations over eight CORDEX domains, as both
514 the reliability and independence of driving fields are essential factors to generate
515 reliable and representative downscaling simulations. In addition, our results could
516 also help to weight climate models to generate more reliable projections of future
517 climate in the CORDEX domains (Kuntti et al. 2017; Brunner et al. 2020).
518

Acknowledgments

We thank the climate modeling groups involved in CMIP6 project for producing and making their model outputs available. The ERA5 Reanalysis data was provided from the website at https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5. Japanese 55-year reanalysis projects were carried out by the Japan Meteorological Agency. The study was supported jointly by the National Key Research and Development Program of China (2017YFA0603803) and the National Science Foundation of China (41675105, 42075170, 42075152). This work was also supported by the Jiangsu Collaborative Innovation Center for Climate Change.

Reference

Bands S, Herrera S, Fernandez J, Gutierrez JM (2013) How well do CMIP5 Earth System Models simulate present climate conditions in Europe and Africa? Clim Dyn 41:803–817. https://doi.org/10.1007/s00382-013-1742-8

Bi D, Dix M, Marsland S, O'Farrell S, Sullivan A, Bodman R et al (2020) Configuration and spin-up of ACCESS-CM2, the new generation Australian community climate and earth system simulator coupled model. J South Hemisph Earth Syst Sci 70(1):225–251. https://doi.org/10.1071/ES19040

Bishop CH, Abramowitz G (2013) Climate model dependence and the replicate Earth paradigm. Clim Dyn 41:885–900. https://doi.org/10.1007/s00382-012-1610-y

Brunner L, Pendergrass AG, Lehner F, Merrifield AL, Lorenz R, Knutti R (2020) Reduced global warming from CMIP6 projections when weighting models by performance and independence. Earth Syst Dynam 11:995–1012. https://doi.org/10.5194/esd-11-995-2020

Buontempo C, Mathison C, Jones R, Willias K, Wang C, cSweeney C (2015) An ensemble climate projection for Africa. Clim Dyn 44:2097–2118. https://doi.org/10.1007/s00382-014-2286-2

Burrows SM, Maltrud M, Yang X, Zhu Q, Jeffery N, Shi X et al (2020) The DOE E3SM v1.1 biogeochemistry configuration: Description and simulated ecosystem-climate responses to historical changes in forcing. J Adv Model Earth Syst 12: e2019MS001766. https://doi.org/10.1029/2019MS001766

Collins M, Booth BBB, Bhaskaran B, Harris GR, Murphy JM, Sexton DMH, Webb MJ (2011) Climate model errors, feedbacks and forcings: a comparison of perturbed physics and multi-model ensemble. Clim Dyn 36:1737–1766. https://doi.org/10.1007/s00382-010-0808-0

Dai A, Rasmussen RM, Ikeda K, Liu C (2020) A new approach to construct representative future forcing data for dynamic downscaling. Clim Dyn 55:315–323. https://doi.org/10.1007/s00382-017-3708-8

Danabasoglu G, Lamarque J-F, Bacmeister J, Bailey DA, DuVivier AK, Edwards J et al (2020) The Community Earth System Model Version 2 (CESM2). J Adv Model Earth Syst 12: e2019MS001916. https://doi.org/10.1029/2019MS001916

Döscher R, Acosta M, Alessandri A, Anthoni P, Arneth A, Arsouze T et al (2021) The EC-Earth3 Earth System Model for the Climate Model Intercomparison Project 6. Geosci Model Dev Discuss Preprint. https://doi.org/10.5194/gmd-2020-446.

Dosio A, Panitz H-J, Schubert-Frisius M, Lüthi D (2015) Dynamical downscaling of CMIP5 global circulation models over CORDEX-Africa with COSMO-CLM: evaluation over the present climate and analysis of the added value. Clim Dyn 44:2637–2661. https://doi.org/10.1007/s00382-014-2262-x

Elguindi N, Giorgi F, Turuncoglu U (2014) Assessment of CMIP5 global model simulations over the subset of CORDEX domains used in the Phase I CREMA. Clim Change 125:7–21. https://doi.org/10.1007/s10584-013-0935-9

Eyring V, Bony S, Meehl GA, Senior CA, Stevens B, Stouffer RJ, Taylor KE (2016) Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. Geosci Model Dev 9: 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016

Giorgi F (2006) Regional climate modeling: status and perspectives. J Phys IV France 139:101–118. https://doi.org/10.1051/jp4:2006139008

Giorgi F, Gutowski WJ (2016) Coordinated Experiments for Projections of Regional Climate Change. Curr Clim Change Rep 2:202–210. https://doi.org/10.1007/s40641-016-0046-6

Giorgi F, Jones C, Asrar GR (2009) Addressing climate information needs at the regional level: the CORDEX framework. Bull World Meteorol Organ 58(3):175–183.

Gutjahr O, Putrasahan D, Lohmann K, Jungclaus JH, Storch J-S, Brüggemann N, Haak H, Stössel A (2019) Max Planck Institute Earth System Model (MPI-ESM1.2) for the High-Resolution Model

576 Intercomparison Project (HighResMIP). Geosci Model Dev 12:3241–3281.
577 https://doi.org/10.5194/gmd-12-3241-2019

578 Gutowski WJ, Giorgi F, Timbal B, Frigon A, Jacob D, Kang H-S, Raghavan K, Lee B, Lennard C,
579 Nikulin G, O'Rourke E, Rixen M, Solman S, Stephenson T, Tangang F (2016) WCRP
580 COordinated Regional Downscaling EXperiment (CORDEX): a diagnostic MIP for CMIP6.
581 Geosci Model Dev 9:4087–4095. https://doi.org/10.5194/gmd-9-4087-2016

582 Han Y, Zhang M-Z, Xu Z, Guo W (2021) Assessing the performance of 33 CMIP6 models in
583 simulating the large-scale environmental fields of tropical cyclones. Clim Dyn Preprint.
584 https://doi.org/10.21203/rs.3.rs-339002/v1

585 Herger N, Abramowitz G, Knutti R, Angélil O, Lehmann K, Sanderson BM (2018) Selecting a
586 climate model subset to optimise key ensemble properties. Earth Syst Dynam 9:135–151.
587 https://doi.org/10.5194/esd-9-135-2018

588 Huang F, Xu Z, Guo W (2019) Evaluating vector winds in the Asian-Australian monsoon region
589 simulated by 37 CMIP5 models. Clim Dyn 53: 491–507.
590 https://doi.org/10.1007/s00382-018-4599-z

591 Huang F, Xu Z, Guo W (2020) The linkage between CMIP5 climate models' abilities to simulate
592 precipitation and vector winds. Clim Dyn 54:4953–4970.
593 https://doi.org/10.1007/s00382-020-05259-6

594 Jun M, Knutti R, Nychka D (2008) Spatial analysis to quantify numerical model bias and dependence:
595 how many climate models are there? J Am Stat Assoc 103:934–947.
596 https://doi.org/10.1198/016214507000001265

597 Jury MW, Prein AF, Truhetz H, Gobiet A (2015) Evaluation of CMIP5 Models in the Context of
598 Dynamical Downscaling over Europe. J Climate 28:5575–5582.
599 https://doi.org/10.1175/JCLI-D-14-00430.1

600 Kebe I, Sylla MB, Omotosho JA, Nikiema PM, Gibba P, Giorgi F (2017) Impact of GCM boundary
601 forcing on regional climate modeling of West African summer monsoon precipitation and
602 circulation features. Clim Dyn 48:1503–1516. https://doi.org/10.1007/s00382-016-3156-x

603 Knutti R (2008) Why are climate models reproducing the observed global surface warming so well?
604 Geophys Res Lett 40:1194–1199. https://doi.org/10.1002/grl.50256

Knutti R (2010a) The end of model democracy?: An editorial comment. Clim Change 102: 395–404. https://doi.org/10.1007/s10584-010-9800-2

Knutti R, Furrer R, Tebaldi C, Cermak J, Meehl GA (2010b) Challenges in combining projections from multiple climate models. J Clim 23:2739–2758. https://doi.org/10.1175/2009JCLI3361.1

Knutti R, Sedláček J (2012) Robustness and uncertainties in the new CMIP5 climate model projections. Nature Clim Change 3:369–373. https://doi.org/10.1038/nclimate1716

Knutti R, Sedláček J, Sanderson BM, Lorenz R, Fischer EM, Eyring V (2017) A climate model projection weighting scheme accounting for performance and interdependence. Geophys Res Lett 44:1909–1918. https://doi.org/10.1002/2016GL072012

Lin Y, Huang X, Liang Y, Qin Y, Xu S, Huang W et al (2020) Community Integrated Earth System Model (CIESM): Description and evaluation. J Adv Model Earth Syst 12: e2019MS002036. https://doi.org/10.1029/2019MS002036

McSweeney CF, Jones RG, Lee RW, Rowell DP (2015) Selecting CMIP5 GCMs for downscaling over multiple regions. Clim Dyn 44:3237–3260. https://doi.org/10.1007/s00382-014-2418-8

Mendlik T, Gobiet A (2016) Selecting climate simulations for impact studies based on multivariate patterns of climate change. Clim Change 135:381–393. https://doi.org/10.1007/s10584-015-1582-0

Mishra SK, Sahany S, Salunke P (2018) CMIP5 vs. CORDEX over the Indian region: how much do we benefit from dynamical downscaling? Theor Appl Climatol 133:1133–1141. https://doi.org/10.1007/s00704-017-2237-z

Plavcová E, Kyselý J (2012) Atmospheric circulation in regional climate models over Central Europe: links to surface air temperature and the influence of driving data. Clim Dyn 39:1681–1695. https://doi.org/10.1007/s00382-011-1278-8

Ranjha R, Tjernstrom M, Svensson G, Semedo A (2016) Modelling coastal low-level wind-jets: does horizontal resolution matter? Meteorol Atmos Phys 128:263–278. https://doi.org/10.1007/s00703-015-0413-1

Reichler T, Kim J (2008) How Well Do Coupled Models Simulate Today's Climate? Bull Am Meteorol Soc 89(3):303-311. https://doi.org/10. 1175/BAMS-89-3-303

633 Rocheta E, Evans JP, Sharma A (2020) Correcting lateral boundary biases in regional climate
634 modeling: the effect of the relaxation zone. Clim Dyn 55:2511–2521.
635 https://doi.org/10.1007/s00382-020-05393-1

636 Ruane AC, Teichmann C, Arnell NW, Carter TR, Ebi KL, Frieler K, Goodess CM, Hewitson B,
637 Horton R, Kovats RS, Lotze HK, Mearns LO, Navarra A, Ojima DS, Riahi K, Rosenzweig C,
638 Themessl M, Vincent K (2016) The vulnerability, impacts, adaptation and climate services
639 advisory board (VIACS AB V1.0) contribution to CMIP6. Geosci Model Dev 9:3493–3515.
640 https://doi.org/10.5194/gmd-9-3493-2016

641 Sanderson BM, Knutti R, Caldwell P (2015a) A Representative Democracy to Reduce
642 Interdependency in a Multimodel Ensemble. J Clim 28:5171–5194.
643 https://doi.org/10.1175/JCLI-D-14-00362.1

644 Sanderson BM, Knutti R, Caldwell P (2015b) Addressing Interdependency in a Multimodel Ensemble
645 by Interpolation of Model Properties. J Clim 28:5150–5170.
646 https://doi.org/10.1175/JCLI-D-14-00361.1

647 Semmler T, Danilov S, Gierz P, Goessling HF, Hegewald J, Hinrichs C et al (2020) Simulations for
648 CMIP6 with the AWI climate model AWI-CM-1-1. J Adv Model Earth Syst 12:
649 e2019MS002009. https://doi.org/10.1029/2019MS002009

650 Taylor KE (2001) Summarizing multiple aspects of model performance in a single diagram. J
651 Geophys Res 106:7183–7192. https://doi.org/10.1029/2000JD900719

652 Tian B, Dong X (2020). The double-ITCZ Bias in CMIP3, CMIP5 and CMIP6 models based on
653 annual mean precipitation. Geophys Res Lett 47:e2020GL087232.
654 https://doi.org/10.1029/2020GL087232

655 Wilks D (2011) Statistical methods in the atmospheric sciences, 3rd edn. Academic Press, USA, pp
656 721–723

657 Wu W, Lynch AH, Rivers A (2005) Estimating the Uncertainty in a Regional Climate Model Related
658 to Initial and Lateral Boundary Conditions. J Climate 18:917–933.
659 https://doi.org/10.1175/JCLI-3293.1

660 Xu J, Gao Y, Chen D, Xiao L, Ou T (2017) Evaluation of global climate models for downscaling
661 applications centred over the Tibetan Plateau. Int J Climatol 37:657–671.
662 https://doi.org/10.1002/joc.4731

663 Xu Z, Han Y, Fu C (2017) Multivariable integrated evaluation of model performance with the vector
664 field evaluation diagram. Geosci Model Dev 10: 3805–3820.
665 https://doi.org/10.5194/gmd-10-3805-2017

666 Xu Z, Hou Z, Han Y, Guo W (2016) A diagram for evaluating multiple aspects of model performance
667 in simulating vector fields. Geosci Model Dev 9:4365–4380. https://doi.org/10.5194/
668 gmd-9-4365-2016

669 Xu Z, Yang Z-L (2012) An Improved Dynamical Downscaling Method with GCM Bias Corrections
670 and Its Validation with 30 Years of Climate Simulations. J Climate 25(18):6271–6286.
671 https://doi.org/10.1175/JCLI-D-12-00005.1

672 Xu Z, Yang Z-L (2015) A new dynamical downscaling approach with GCM bias corrections and
673 spectral nudging. J Geophys Res Atmos 120:3036–3084. https://doi.org/10.1002/2014JD022958

674 Zhang M-Z, Xu Z, Han Y, Guo W (2021) An improved multivariable integrated evaluation method
675 and tool (MVIETool) v1.0 for multimodel intercomparison. Geosci Model Dev 14: 3079–3094.
676 https://doi.org/10.5194/gmd-14-3079-2021

677 Zhu Y-Y, Yang S (2020) Evaluation of CMIP6 for historical temperature and precipitation over the
678 Tibetan Plateau and its comparison with CMIP5. Adv Clim Change Res 11(3):239-251.
679 https://doi.org/10.1016/j.accre.2020.08.001.
680

Figures

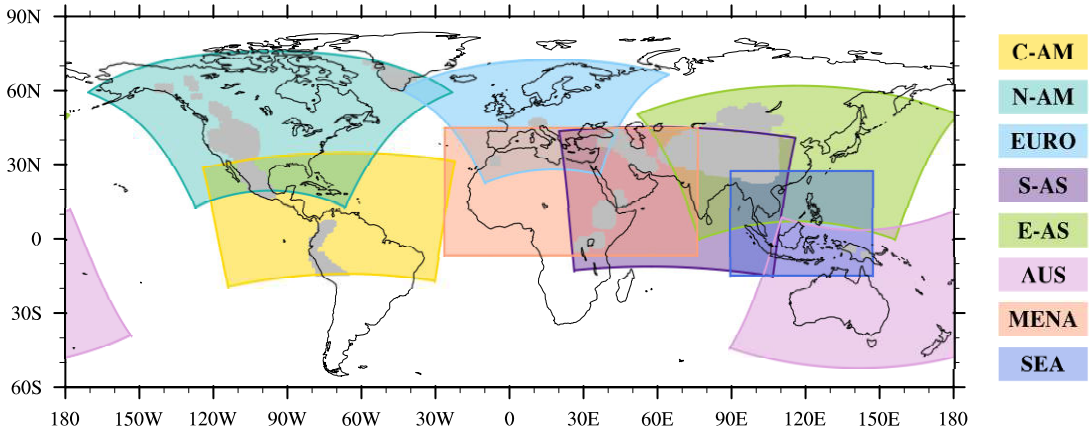
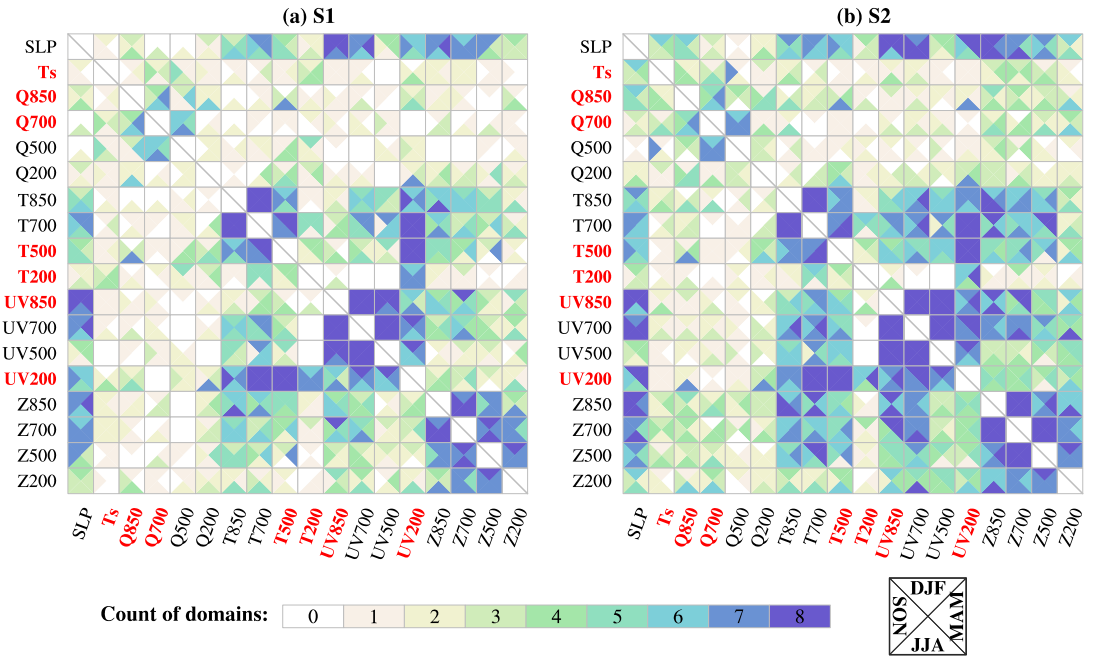


Fig. 1 Eight CORDEX domains for evaluation, i.e., Central America (C-AM), North America (N-AM), Europe (EURO), South Asia (S-AS), East Asia (E-AS), Australasia (AUS), Middle East North Africa (MENA), and South East Asia (SEA). The grey shading within each domain is the mask for the atmosphere variables



688

689 **Fig. 2** Spearman rank correlation of model performance between pairwise variables across 37
690 CMIP6 models. The filling color indicates the number of CORDEX domains over which the
691 correlation coefficient of **a** S_1 or **b** S_2 between two variables reaches the significance level of 0.05.
692 Variables include sea level pressure (SLP), surface temperature (Ts), as well as specific humidity
693 (Q), air temperature (T), vector wind field (UV), and geopotential height (Z) at 850-hP, 700-hPa,
694 500-hPa, and 200-hPa, respectively. Both S_1 and S_2 are used to quantify the model performance
695 in simulating the climatological mean of the individual scalar variable. For the vector wind field,
696 S_{v1} and S_{v2} are used instead. Each square in the table is divided into four triangles, representing
697 four seasons, i.e., December-January-February (DJF), March-April-May (MAM),
698 June-July-August (JJA), and September-October-November (SON). Variables shown in the red
699 font are selected ones to evaluate in our study

700

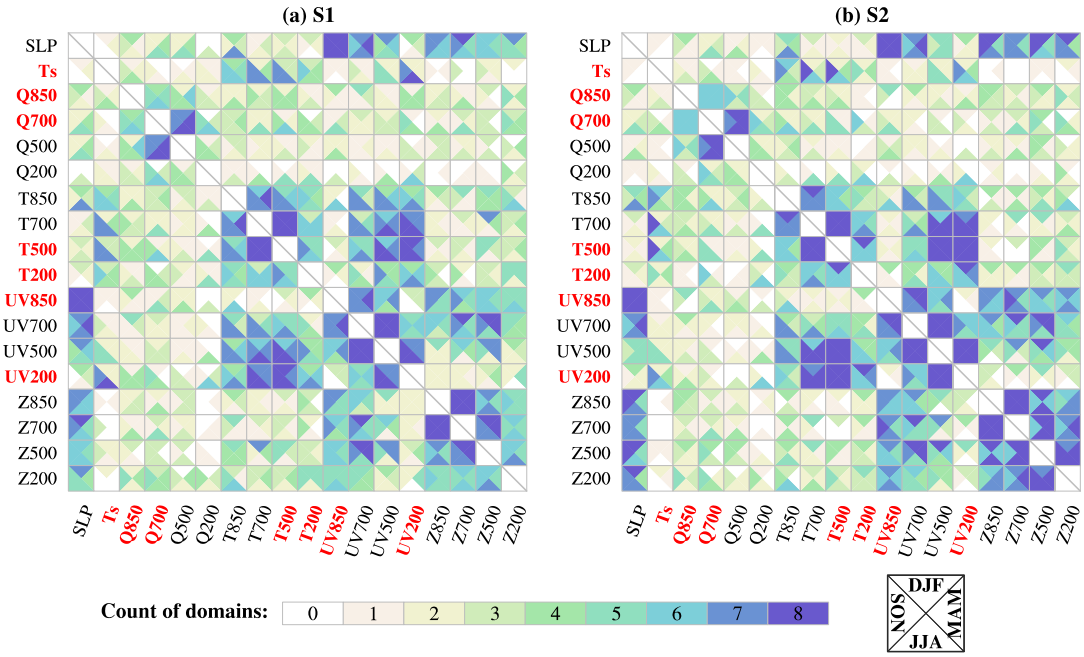


Fig. 3 Same as Fig. 2, except that S_1 and S_2 are computed with the interannual standard deviation fields

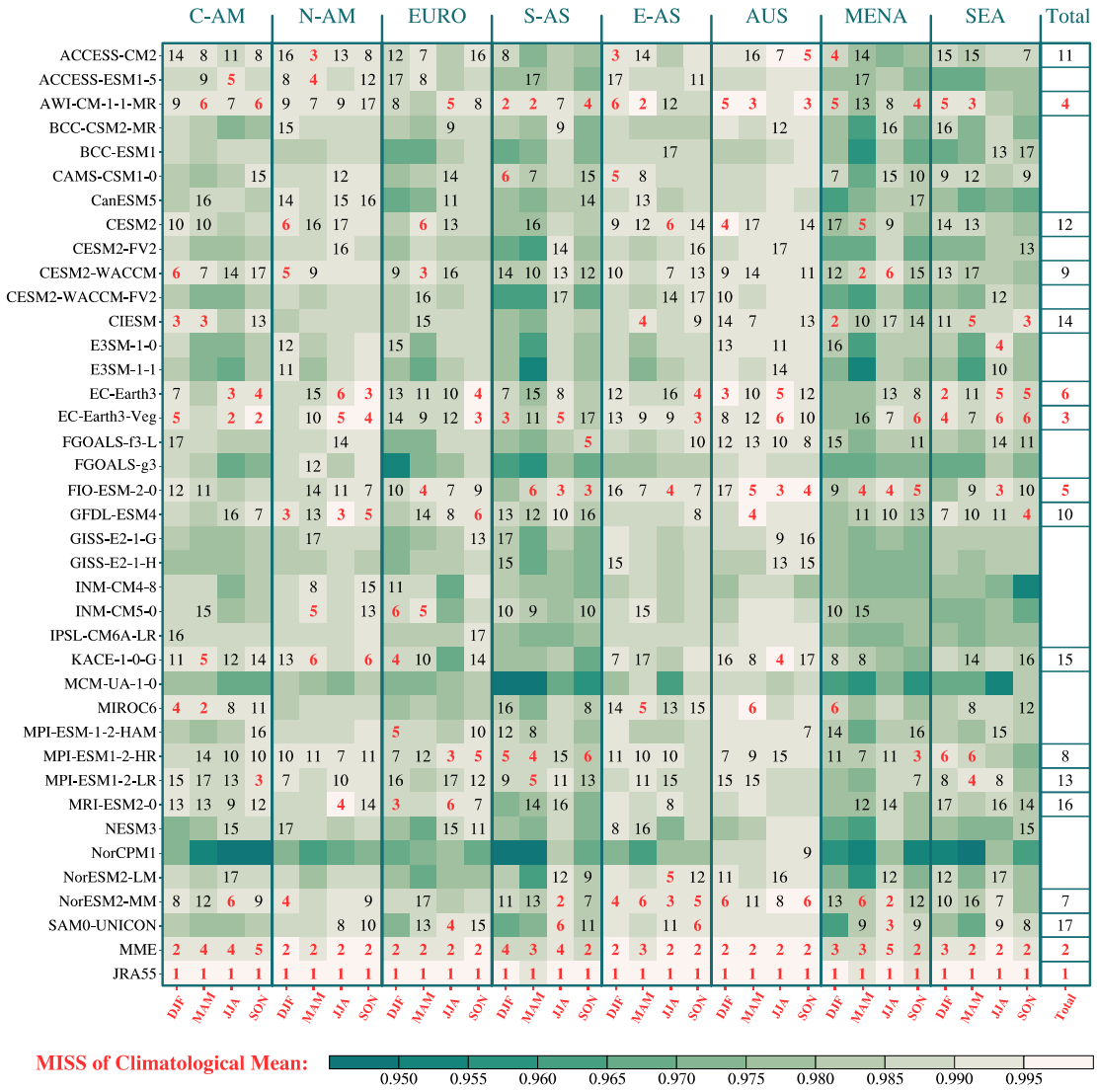


Fig. 4 Rank of CMIP6 models in terms of the model performance in simulating climatological mean of seven variables over eight CORDEX domains in four seasons (i.e., DJF, MAM, JJA, and SON). For comparison, JRA55 and multi-model ensemble mean (MME) are also shown in the figure. All models and JRA55 reanalysis are compared against the ERA5. The rank is determined by MISS, and a greater MISS indicates a better model performance. The rightmost column shows the total ranking of each model by taking all domains and seasons into account. Only the top 15 models among 37 CMIP6 models as well as the JRA55 and MME are numbered, where the top six ones are even numbered with red font

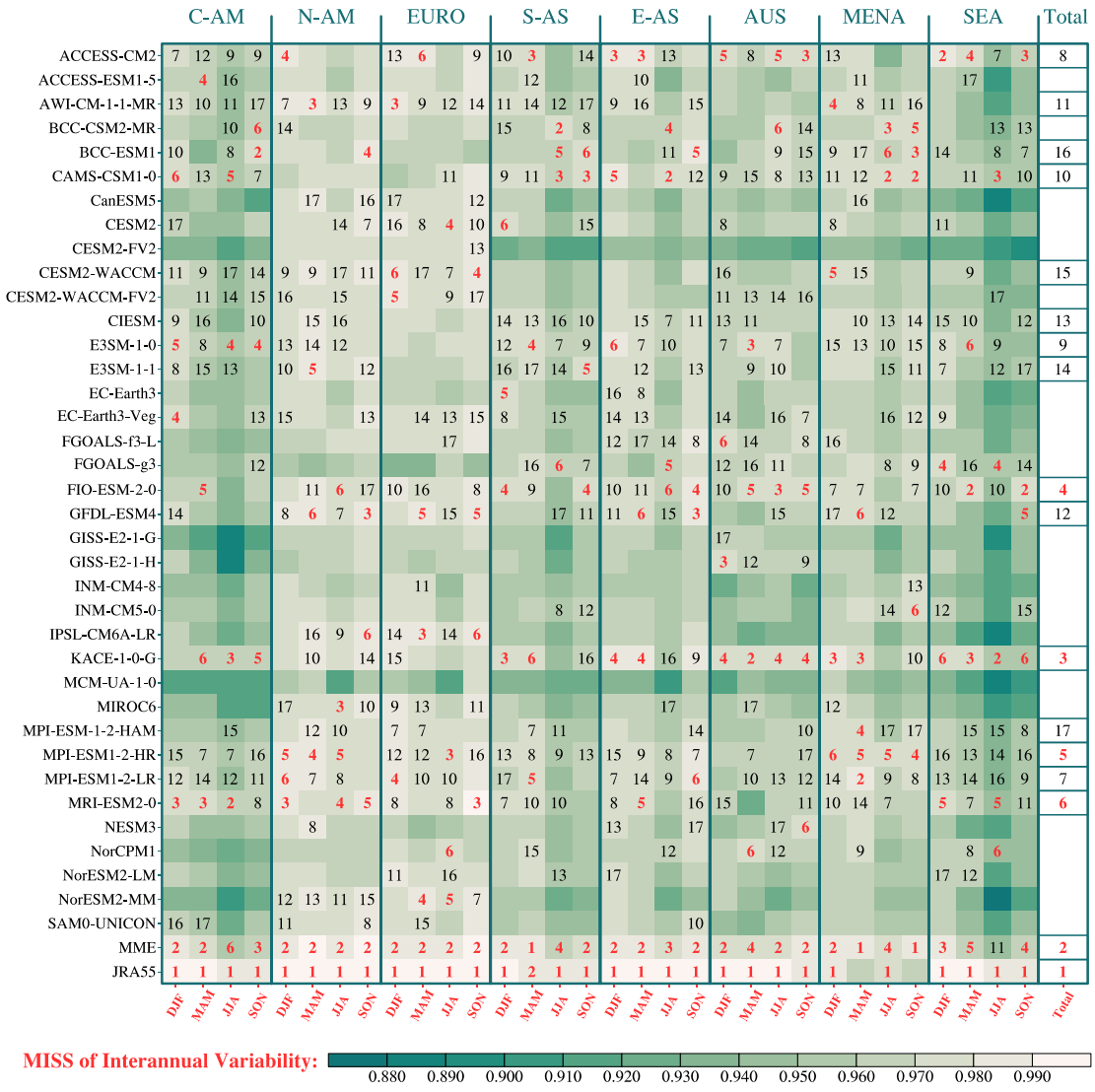
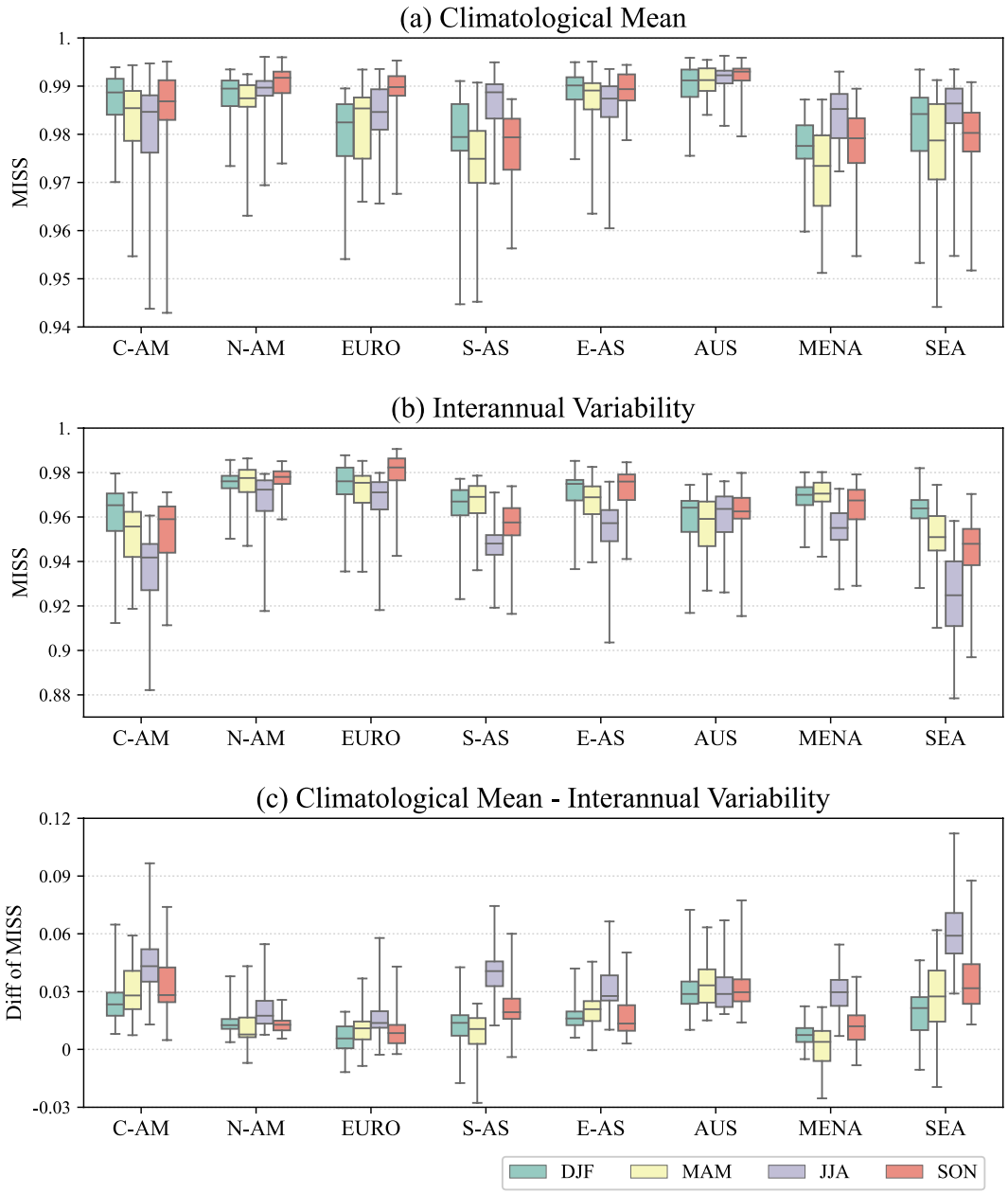


Fig. 5 Same as Fig. 4, except that MISSs are computed with the interannual variability fields of multiple variables. Interannual variability is measured by the standard deviation of year-to-year time series in terms of one individual variable



720

721 **Fig. 6** Box plot of multi-model MISSs in simulating **a** climatological mean and **b** interannual

722 variability of seven variables over eight CORDEX domains. **c** Difference of MISSs between

723 climatological mean and interannual variability. The box plot shows the minimum, 25th quantile,

724 medium, 75th quantile, and maximum MISSs derived from 37 CMIP6 models

725

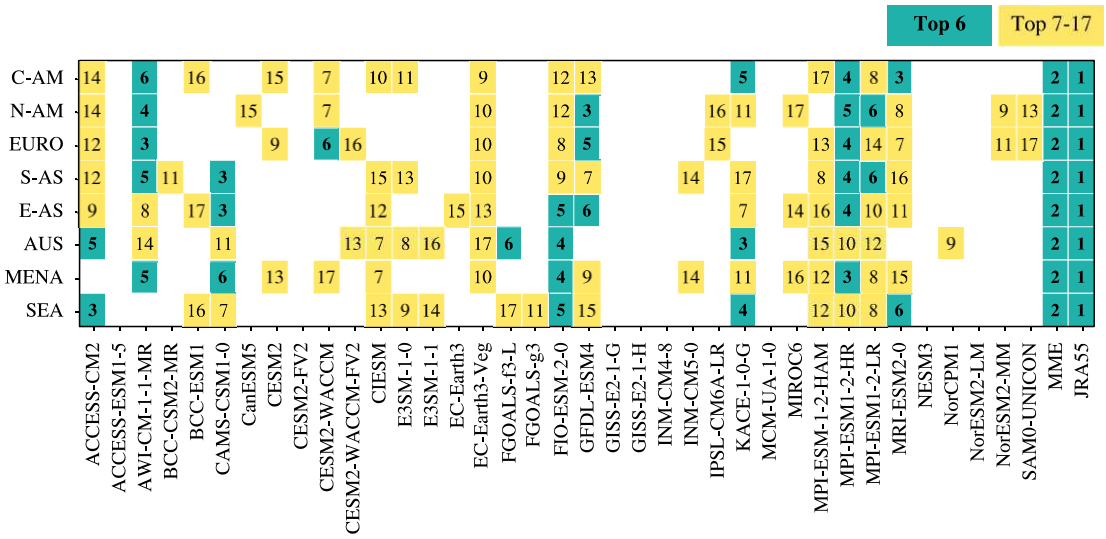


Fig. 7 Rank of 37 CMIP6 models in terms of overall model performance over eight CORDEX domains. The rank is determined by MISS of multiple fields that contain both climatological mean and interannual variability of four seasons (2 climate statistical characteristics × 4 seasons × 7 variables). Only the top 15 models among CMIP6 models as well as the JRA55 and MME, are ranked in each domain

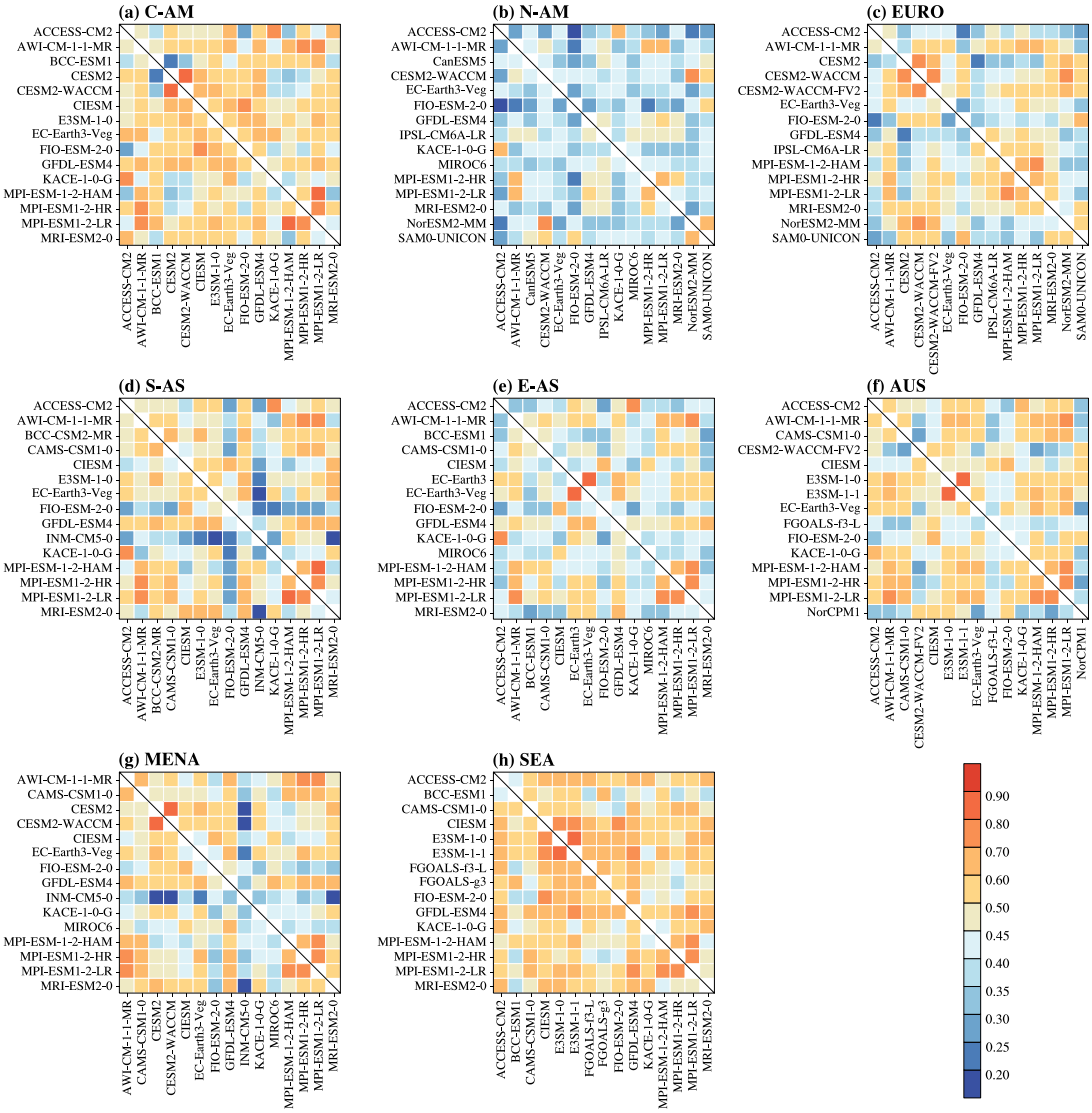


Fig. 8 Model interdependency measured by the similarity of multivariable error fields (MvES) in each of the eight CORDEX domains. Only model interdependency of the top 15 CMIP6 models in terms of the overall performance in simulating multiple fields is shown here. The color filling in each square represents the magnitude of MvES between its corresponding two models in the X-axis and Y-axis, respectively. Larger MvES indicates higher similarity in multivariate error fields of pairwise models and vice versa

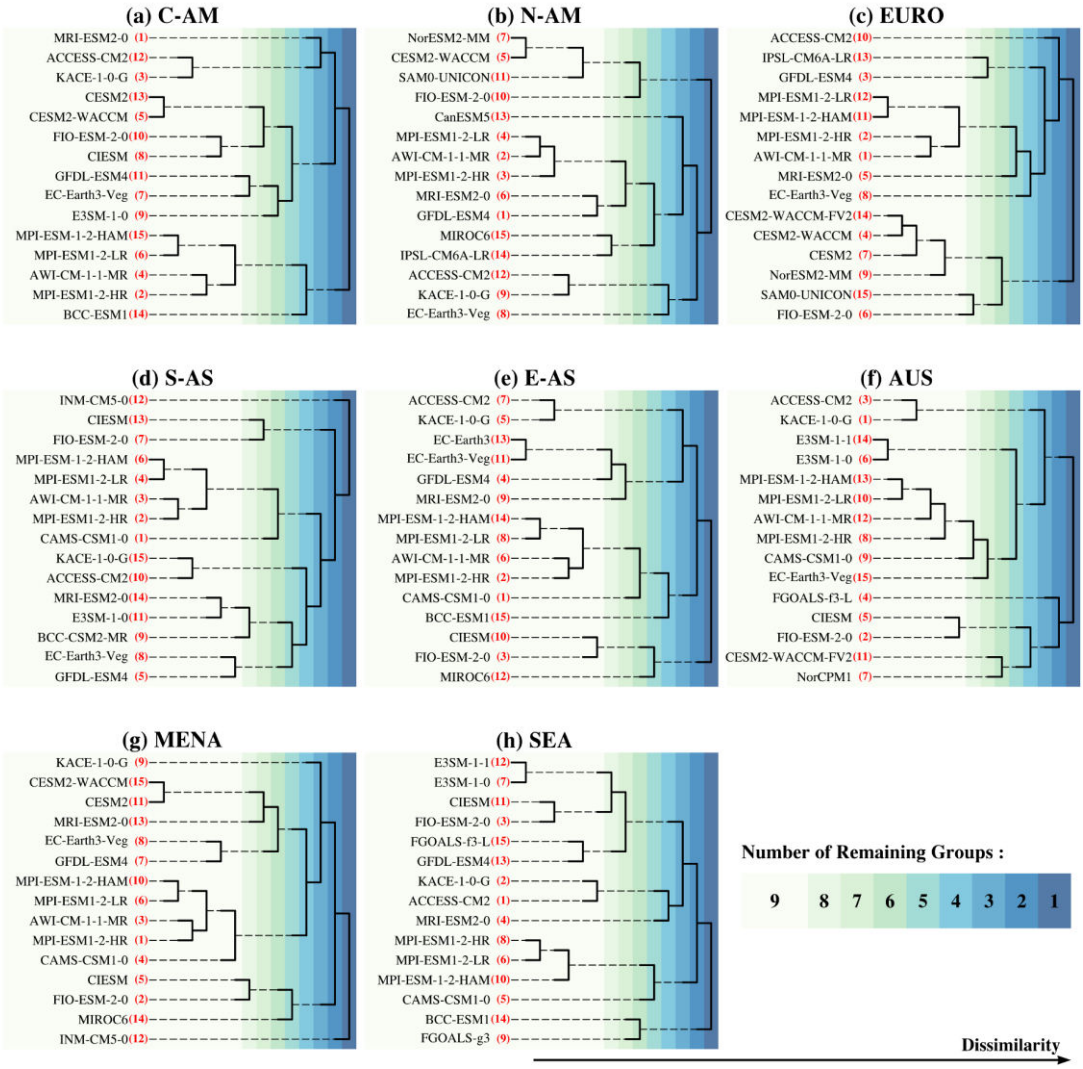
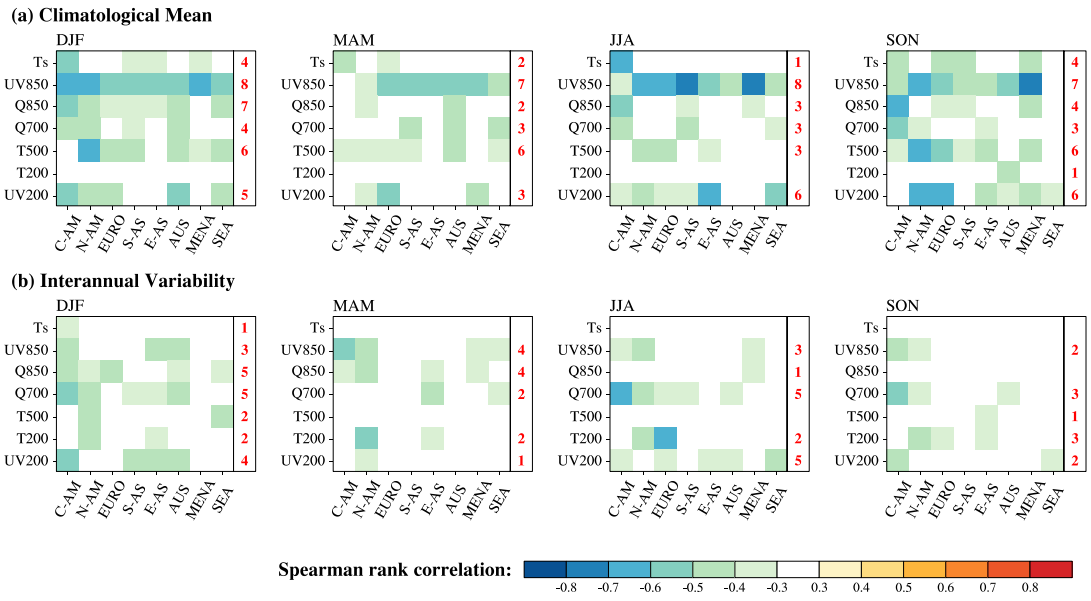


Fig. 9 The dendrogram reflecting the clustering of the top 15 CMIP6 models in each of the eight CORDEX domains. Each model is regarded as an individual group at the beginning. MvES is used as the distance measure between two models, and inter-group distance is the average of MvES between all pairs of models in two groups. In each clustering step, two groups closest to each other are merged into a new group. Inter-group similarity decreases with the clustering progressing. The background color of a clustering represents the number of remaining groups after clustering, which also indicates the number of relatively independent models at the current clustering step. The red number in the bracket indicates the rank of the model determined by the MISSs among 37 CMIP6 models



752

753 **Fig. 10** Spearman rank correlation coefficient table between the models' horizontal resolutions
754 and their ability in simulating **a** climatological mean and **b** interannual variability of seven
755 individual variables over eight CORDEX domains. The model horizontal resolution is defined by
756 the average grid area, which is the product of the zonal and meridional grid spacing. S_2 (S_{v2}) is
757 used to measure the models' ability to simulate the individual scalar (vector) variable. The shaded
758 grid box indicates that the correlation reaches the significance level of 0.05, and the depth of
759 shading donates the value of the correlation. The number in the right column of each season
760 represents the number of domains with a significant correlation between the horizontal
761 resolutions and the model ability

762

Tables

Table 1 37 CMIP6 models used for evaluation together with their institutions and horizontal resolutions (longitude × latitude)

	Model	Horizontal Resolution	Institution
1	ACCESS-CM2	1.875°×1.25°	Commonwealth Scientific and Industrial Research Organization (Australia)
2	ACCESS-ESM1-5	1.875°×1.25°	
3	AWI-CM-1-1-MR	0.94°×0.94°	Alfred Wegener Institute (Germany)
4	BCC-CSM2-MR	1.125°×1.125°	Beijing Climate Center (China)
5	BCC-ESM1	2.81°×2.81°	
6	CAMS-CSM1-0	1.125°×1.125°	Chinese Academy of Meteorological Sciences (China)
7	CanESM5	2.81°×2.81°	Canadian Centre for Climate Modelling and Analysis (Canada)
8	CESM2	1.25°×0.94°	National Center for Atmospheric Research (USA)
9	CESM2-FV2	2.5°×1.89°	
10	CESM2-WACCM	1.25°×0.94°	
11	CESM2-WACCM-FV2	2.5°×1.89°	
12	CIESM	1.25°×0.94°	Department of Earth System Science (China)
13	E3SM-1-0	1°×1°	Lawrence Livermore National Laboratory (USA)
14	E3SM-1-1	1°×1°	
15	EC-Earth3	0.70°×0.70°	European Center–Earth Consortium (Europe)
16	EC-Earth3-Veg	0.70°×0.70°	
17	FGOALS-f3-L	1.25°×1°	Chinese Academy of Sciences (China)
18	FGOALS-g3	2°×2.25°	
19	FIO-ESM-2-0	1.25°×0.94°	First Institute of Oceanograph (China)
20	GFDL-ESM4	1°×1°	National Oceanic and Atmospheric Administration (USA)
21	GISS-E2-1-G	2.48°×2°	Goddard Institute for Space Studies (USA)
22	GISS-E2-1-H	2.48°×2°	
23	INM-CM4-8	2°×1.5°	Institute for Numerical Mathematics (Russia)
24	INM-CM5-0	2°×1.5°	
25	IPSL-CM6A-LR	2.5°×1.27°	Institut Pierre Simon Laplace (France)
26	KACE-1-0-G	1.87°×1.25°	National Institute of Meteorological Sciences/Korea Meteorological Administration (Korea)
27	MCM-UA-1-0	3.75°×2.25°	Department of Geosciences, University of Arizona (USA)
28	MIROC6	1.41°×1.41°	Japan Agency for Marine-Earth Science and Technology (Japan)
29	MPI-ESM-1-2-HAM	1.875°×1.875°	ETH Zurich (Switzerland)
30	MPI-ESM1-2-HR	0.94°×0.94°	Max Planck Institute for Meteorology (Germany)
31	MPI-ESM1-2-LR	1.875°×1.875°	

32	MRI-ESM2-0	1.125°×1.125°	Meteorological Research Institute (Japan)
33	NESM3	1.875°×1.875°	Nanjing University of Information Science and Technology (China)
34	NorCPM1	2.5°×1.89°	Center for International Climate and Environmental Research (Norway)
35	NorESM2-LM	2.5°×1.89°	
36	NorESM2-MM	1.25°×0.94°	
37	SAM0-UNICON	1.25°×0.94°	Seoul National University (Korea)

766

767 **Table 2** Eighteen variables in the initial consideration for evaluation

	Acronyms	Description	Unit
Surface variables	Ts	Surface temperature	°C
	SLP	Sea level pressure	hPa
Atmosphere variables at 850-hPa, 700-hPa, 500-hPa, and 200-hPa	T	Atmosphere temperature	°C
	UV	Zonal wind and meridional wind	m/s
	Q	Specific humidity	grams/kg
	Z	Geopotential height	m^2/s^2

768