https://doi.org/10.5194/egusphere-2026-5469
Preprint. Discussion started: 16 November 2026
© Author(s) 2026. CC BY 4.0 License.

Advancing climate model evaluation and intercomparison: an error decomposition-based index to quantify statistical sources of climate model errors

Meng-Zhuo Zhang[1], Zhongfeng Xu[2], Ying Han[3], Weidong Guo[1]

[1]School of Atmospheric Sciences, Nanjing University, Nanjing, China
[2]State Key Laboratory of Earth System Numerical Modelling and Application, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, 100029, China
[3]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. Root Mean Square Error (RMSE) and Kling-Gupta Efficiency (KGE) are widely used performance metrics for quantifying the mismatch between simulated and observed data. They can be expressed in terms of three components: mean error, anomaly amplitude error, and anomaly similarity error. This decomposition makes it possible to identify the primary error source affecting overall model performance. However, the RMSE does not vary monotonically with model performance, which may mislead the evaluation results, while KGE is insensitive to anomaly similarity error, especially when the correlation coefficient value is high. To address these shortcomings, we proposed a novel error-decomposition index, termed the Decomposable Error Index (DEI), for measuring overall model error. The DEI is built from individual penalty terms that target mean, anomaly amplitude, and anomaly similarity errors. When the other two errors are held constant, the DEI varies monotonically with each of these three errors, thereby avoiding the non-monotonic limitation of RMSE. Moreover, the DEI imposes stronger penalties on the anomaly similarity error than KGE does. The DEI error decomposition statistically identifies the primary error source of the overall model error and pinpoints the primary driver for the inter-model performance difference. Beyond the evaluation of a scalar field, we also extended DEI and its decomposition into vector or multivariate fields. Applying DEI and its decomposition to evaluate CMIP models over East Asia, we identified the anomaly similarity error as the primary source of the overall performance for precipitation and multi-level winds across 26 CMIP6 models in both winter and summer, explaining 42–99% of DEI^2. It underscores the priority of improving pattern similarity for their simulations in future model development. Specifically, the pattern similarity improvement in wind and summer precipitation can significantly benefit from enhancing the horizontal resolution. In evaluating generational progress of 13 CMIP5-6 model pairs, the DEI decomposition pinpoints the reduced anomaly similarity error (mean error) as the key driver responsible for the improved overall performance of multi-level wind simulations in winter and summer (multi-level air temperature simulation in summer) over this region, with over 53% contribution.

https://doi.org/10.5194/egusphere-2026-5469
Preprint. Discussion started: 16 November 2026
© Author(s) 2026. CC BY 4.0 License.

1 Introduction

State-of-the-art climate model simulations are pivotal in understanding past, present, and future climate and routinely serve as the scientific basis for future climate change projections (Meehl et al., 2014; Eyring et al., 2016; Simpkins, 2017). With the growing number of models participating in the International Coupled Model Intercomparison Project (CMIP), quantitative evaluation and intercomparison of climate models emerge as the basis for downstream application of model outputs and are essential for advancing their further development (Knutti et al., 2013; Eyring et al., 2016, 2019; Chen et al., 2021). To this end, there is a pressing need for an overall model performance indicator that can be decomposed into separate error components, thereby enabling the identification of the primary statistical source of overall model error, investigation of their potential causes, and guidance for next-generation model development.

In particular, it is acknowledged that mean error, anomaly amplitude error, and anomaly similarity error between the model and the observation are key aspects of model performance and have been given great attention in previous model evaluation studies (Taylor, 2001; Flato et al., 2013; Yin et al., 2013; Zhu et al., 2020). Some summarized metrics indicating overall model performance, such as Root Mean Square Error (RMSE) and Kling-Gupta Efficiency (KGE), simultaneously integrate these three aspects of model performance, indicating that they can be mathematically decomposed into error components related to mean, anomaly amplitude, and anomaly similarity errors (Murphy, 1988; Gupta et al., 2009; Xu and Han, 2019). An emerging question of great importance appears: which error component dominates overall model performance? Addressing this question is critical for prioritizing model modifications and guiding model development toward greater accuracy and capability, as different errors enable causal inference of model deficiencies.

It is well-established that current model errors primarily stem from inadequate representations of realistic physical processes (Li et al., 2023; Pan et al., 2024). Mean error is often linked to systematic errors, and modifications to model modules affecting all regions could efficiently alleviate this bias. For example, overestimated mean precipitation is likely linked to homogeneous biases caused by unreasonable parameterizations, such as the trigger function of convection activities, entrainment and detrainment rates, parameters related to convection memory effect, and the calculation of cloud cover in cloud schemes (Wu, 2012; Wu et al., 2019; Kawai et al., 2019; Hwong et al., 2023). In contrast, the anomaly amplitude and similarity measure model capability in reproducing the heterogeneity of the anomaly field. Therein, anomaly amplitude quantifies the model's ability to reproduce the magnitude of spatial and/or temporal variability. Improving this aspect requires refining model modules that influence simulated anomaly variation, such as detailed categories of land use and cover, modification to vegetation cover, and accounting for the transition of land use and land cover change in the land model for better simulating the spatial and temporal heterogeneity of surface temperature, planetary boundary layer, etc (Mauritsen et al., 2019; Wu et al., 2019). Similarly, anomaly similarity enables the causal inference of poor model performance from the perspective of affecting the anomalous distribution characteristics of the simulation. For example, increasing the horizontal resolution of climate models may enhance the pattern similarity in near-surface anomalous wind fields (Roberts et al., 2015; Huang et al., 2019a; Han et al., 2022; Zhang et al., 2022). Likewise, improving vertical resolution contributes to more accurate simulations of cloud

https://doi.org/10.5194/egusphere-2026-5469
Preprint. Discussion started: 16 November 2026
© Author(s) 2026. CC BY 4.0 License.

and precipitation anomalies by better capturing their distribution characteristics (Hagemann et al., 2006; Roeckner et al., 2006; Volosciuk et al., 2015). Therefore, quantifying the contributions of mean error, anomaly amplitude error, and anomaly similarity error to the overall error helps to identify the dominant error source of the model.

Besides, quantifying the contributions of the three errors to the overall error can also enable valuable insight into what dominates the overall performance of a model to be better or worse compared with others. Performance disparities between models from different generations have gathered widespread attention (Kim et al., 2020; Deng et al., 2021; Khadka et al., 2021; Guo et al., 2022). Pinpointing which error critically influences overall model performance changes across generations fosters a deep understanding of intergenerational differences in CMIP models and offers guidance for future model development.

Among the current summarized metrics, widely used RMSE and KGE are explicit functions of mean error, anomaly amplitude error, and anomaly similarity error between the model and the observation for a scalar field. However, for any given mean and anomaly similarity errors, RMSE does not decrease monotonically as the modelled anomaly amplitude approaches the observed one (Taylor, 2001; Gupta et al., 2009). Given the same mean value and anomaly similarity measured by correlation coefficient, a model with a more accurate amplitude may yield a larger RMSE than a model with a less accurate amplitude. Thus, RMSE could lead to misleading overall model performance estimates if focusing solely on the three errors in the model evaluation (detailed illustration shown in Section 2.3). In addition, KGE integrates mean, anomaly amplitude, and anomaly similarity errors by the Euclidean distance of model performance to the ideal one, and can be individually decomposed into the three error components (Gupta et al., 2009). However, the penalty term for the anomaly similarity error may result in insufficient emphasis on this non-systematic error in KGE (detailed illustration shown in Section 2.3). Therefore, a more reasonable metric summarizing overall model performance is warranted, which can reasonably integrate mean, anomaly amplitude, and anomaly similarity errors and avoid these limitations of current metrics for the purpose of statistical error decomposition and attribution.

Thus, our study proposes a new overall performance indicator termed the Decomposable Error Index (DEI) for climate model evaluation. DEI comprises three individual penalty terms targeting mean, anomaly amplitude, and anomaly similarity errors and has advantages compared with RMSE and KGE. In addition to the scalar field, we extend DEI to measure overall model performance in simulating multiple fields for multivariable integrated evaluation. With the aid of DEI, we can quantify the contributions of the three source errors to the overall performance of a single model and inter-model performance differences. Thus, applying DEI and its decomposition in model evaluation pinpoints the primary error driving the overall performance of individual models and performance differences among models, which previous studies have not explored. In this study, we use DEI to evaluate CMIP model performance in simulating various variables, including precipitation, temperature, and wind fields over the East Asian region. Through these evaluations, we reveal the added diagnostic information of overall model performance, thereby advancing model inter-comparison and development.

The paper is organized as follows. Section 2 defines the DEI, introduces its decomposition in terms of mean, anomaly amplitude, and anomaly similarity errors, and compares DEI with RMSE and KGE. Section 3 applies this method to evaluate

https://doi.org/10.5194/egusphere-2026-5469
Preprint. Discussion started: 16 November 2026
© Author(s) 2026. CC BY 4.0 License.

the overall model performance of various CMIP Phase 5 (CMIP5) and CMIP Phase 6 (CMIP6) models in terms of six variables over East Asia, and illustrates the evaluation results. Section 4 presents the summary and discussion.

2 Decomposable Error Index (DEI)

2.1 Defining DEI for model performance of a scalar field

The DEI is built on the decomposition of RMSE, and we start with its decomposition into mean, anomalous amplitude, and anomalous similarity errors. Following Murphy (1988), this relationship can be written as:

$$RMSE^2 = ME^2 + SD_A^2 + SD_O^2 - 2 \cdot SD_A \cdot SD_O \cdot R \quad (1)$$

where ME (R) represents the mean error (correlation coefficient) of a scalar field between the model and the observation. SD_A and SD_O denote modelled and observed standard deviations, respectively, and their ratio measures the anomaly amplitude error for a scalar field in the model.

DEI for a scalar field also comprises three independent penalty terms related to mean error, anomaly amplitude error, and anomaly similarity error, respectively. By setting the model performance in simulating anomaly amplitude and similarity to be perfect in Eq.(1), we define the penalty term related to mean error as:

$$E_{Mean} = ME^2 \quad (2)$$

Similarly, the penalty terms related to anomaly amplitude and anomaly similarity errors are defined as:

$$E_{AnoA} = (SD_A - SD_O)^2 \quad (3)$$

$$E_{AnoS} = 2 \cdot {SD_O}^2 \cdot (1 - R) \quad (4)$$

Then, the DEI measuring the overall model error of a scalar field is expressed as:

$$DEI = \sqrt{E_{Mean} + E_{AnoA} + E_{AnoS}} \quad (5)$$

where a smaller value of DEI indicates a better model's overall performance, and vice versa. When two of the three errors are held constant, the DEI varies monotonically with the remaining error, avoiding the limitation of RMSE.

We also give a geometric interpretation of DEI and its decomposed error terms, illustrated using the Taylor diagram (Fig. 1). Points A and R represent the model and observation, respectively. We first focus on the error terms related to the anomaly field (left panel of Fig. 1). The line segment AO (RO and BO) represents modelled (observed) standard deviation, and thereby line segment AB refers to $\sqrt{E_{AnoA}}$. Assuming that the model performance in simulating amplitude reaches its best by moving point A to B, and then the line segment BR denotes $\sqrt{E_{AnoS}}$. Then, place point B as the center and use the length of AB as the radius to construct a circle, and consequently, the distance between point B and any point along this arc also equals $\sqrt{E_{AnoA}}$.

https://doi.org/10.5194/egusphere-2026-5469
Preprint. Discussion started: 16 November 2026
© Author(s) 2026. CC BY 4.0 License.

Add the point C on the arc to make the line segment perpendicular to BR, and CR indicates $\sqrt{E_{AnoA} + E_{AnoS}}$. Further, add line segment DA vertical with aspect to the Taylor diagram with the length of $\sqrt{E_{Mean}}$, and line segment DR indicates DEI (right panel of Fig. 1b). Clearly, BC, BR, and DC are all orthogonal to each other. In geometry, BC, BR, DC, and DR are the length, width, height, and diagonal line of a cube, respectively. Thus, the DEI can be decomposed into three independent error components, which represent the anomaly amplitude error, anomaly similarity error, and mean error, respectively.

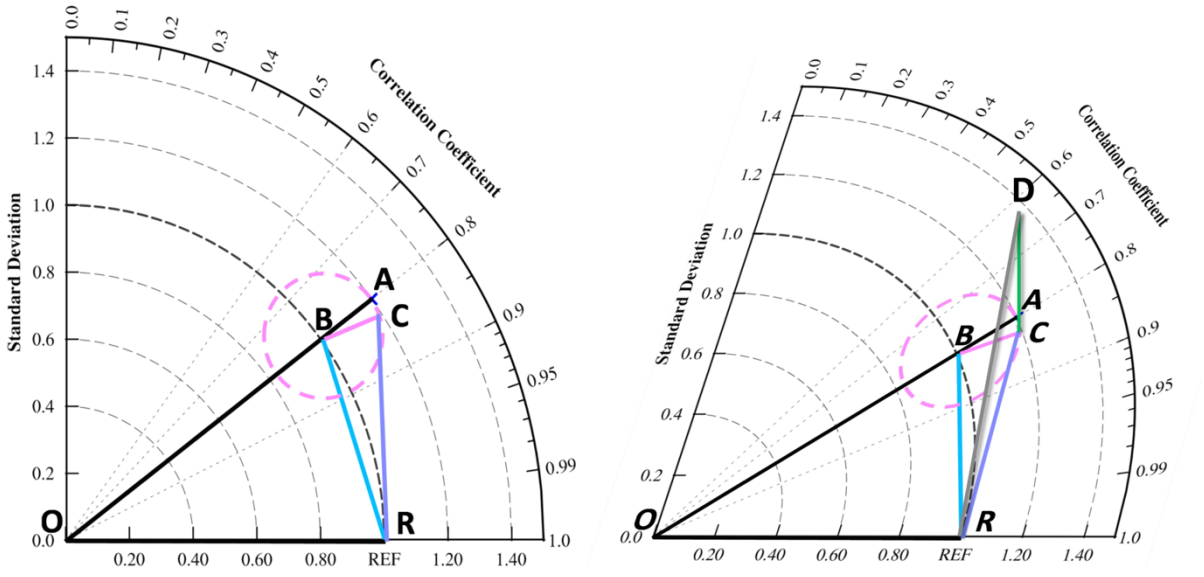


Figure 1: Schematic diagram displaying the relationship between $\sqrt{E_{Mean}}$, $\sqrt{E_{AnoA}}$, $\sqrt{E_{AnoS}}$, and DEI based on the Taylor diagram. Point A denotes the model, and R represents the observation. $\sqrt{E_{Mean}}$, $\sqrt{E_{AnoA}}$, and $\sqrt{E_{AnoS}}$ are the roots of penalty terms targeting mean error (line DC), anomaly amplitude error (line BC or BA), and anomaly similarity error (line BR) in DEI (line DR), respectively. The left (right) panel is the diagram in two (three) dimensions. The line segment DA is vertical for the Taylor diagram in the right panel.

Based on Eqs. (2–5), the relative contributions of mean error (RC_{Mean}), anomaly amplitude error (RC_{AnoA}), and anomaly similarity error (RC_{AnoS}) to the over model error (DEI) can be computed as:

$$RC_{Mean} = \frac{E_{Mean}}{DEI^2} \times 100\% = \frac{ME^2}{ME^2+(SD_A-SD_O)^2+2\cdot SD_O{}^2\cdot(1-R)} \times 100\% \quad (6)$$

$$RC_{AnoA} = \frac{E_{AnoA}}{DEI^2} \times 100\% = \frac{(SD_A-SD_O)^2}{ME^2+(SD_A-SD_O)^2+2\cdot SD_O{}^2\cdot(1-R)} \times 100\% \quad (7)$$

$$RC_{AnoS} = \frac{E_{AnoS}}{DEI^2} \times 100\% = \frac{2\cdot SD_O{}^2\cdot(1-CORR)}{ME^2+(SD_A-SD_O)^2+2\cdot SD_O{}^2\cdot(1-R)} \times 100\% \quad (8)$$

with the aid of RC_{Mean}, RC_{AnoA}, and RC_{AnoS}, we can identify which aspect of model performance dominates the overall model error.

Moreover, we define the normalized DEI (nDEI) by dividing the observed SD to enable fair comparison of overall model performance across variables and regions:

https://doi.org/10.5194/egusphere-2026-5469
Preprint. Discussion started: 16 November 2026
© Author(s) 2026. CC BY 4.0 License.

$$nDEI = \sqrt{E_{Mean}^{Nor} + E_{AnoA}^{Nor} + E_{AnoS}^{Nor}} = \sqrt{\frac{ME^2}{SD_O^2} + (\frac{SD_M}{SD_O} - 1)^2 + 2 \cdot (1 - R)} \quad (9)$$

2.2 Extending DEI to model evaluation of multiple fields

Following the framework of the multivariable integrated evaluation method, multiple fields can be composed of various individual variables for evaluation, including the scalar variable (e.g., temperature and precipitation) and the vector variable (e.g., vector wind field) (Xu et al., 2017; Zhang et al., 2021). After normalizing each of the scalar (vector) variables with the respective root mean square value (root mean square length) of the observation, these variables can be grouped into a multi-dimensional vector field for evaluation.

Regarding vector fields or multiple fields, vector mean error (VME) measures the modelled mean error against the observation, centered root-mean-square vector length (cRMSL) measures the anomaly amplitude, and centered vector similarity coefficient (cVSC) measures the similarity of anomalies between model and observation (Xu et al., 2017; Zhang et al., 2021). The calculation formulas of VME, cRMSL, and cVSC are provided in Text S1 in the Supporting Information. Thus, by replacing ME, SD, and CORR with corresponding metrics for multiple fields—VME, cRMSL, and cVSC in Eqs. (2–5), we obtain the metric for measuring overall model errors in simulating multiple fields. We termed this metric the Decomposable Vector Error Index (DVEI), and DVEI and its three error penalty terms are expressed as:

$$DVEI = \sqrt{E_{Mean} + E_{AnoA} + E_{AnoS}} \quad (10)$$

with:

$$E_{Mean} = VME^2$$

$$E_{AnoA} = (cL_A - cL_O)^2$$

$$E_{AnoS} = 2 \cdot {cL_O}^2 \cdot (1 - cR_v)$$

where cL_A and cL_O denote modelled and observed cRMSL, respectively, and cR_v represents the cVSC of multiple fields.

Analogous to Eqs. (6–8), the relative contributions of mean, anomaly amplitude, and anomaly similarity errors to the overall model error for multiple fields are computed as:

$$RC_{Mean} = \frac{VME^2}{VME^2 + (cL_A - cL_O)^2 + 2 \cdot {cL_O}^2 \cdot (1 - cR_v)} \times 100\% \quad (11)$$

$$RC_{AnoA} = \frac{(cL_A - cL_O)^2}{VME^2 + (cL_A - cL_O)^2 + 2 \cdot {cL_O}^2 \cdot (1 - cR_v)} \times 100\% \quad (12)$$

$$RC_{AnoS} = \frac{2 \cdot {cL_O}^2 \cdot (1 - cR_v)}{VME^2 + (cL_A - cL_O)^2 + 2 \cdot {cL_O}^2 \cdot (1 - cR_v)} \times 100\% \quad (13)$$

Following the same normalization strategy as in Eq. (9), we define the normalized DVEI (nDVEI) as follows:

https://doi.org/10.5194/egusphere-2026-5469
Preprint. Discussion started: 16 November 2026
© Author(s) 2026. CC BY 4.0 License.

$$nDVEI = \sqrt{\frac{VME^2}{cL_O^2} + (\frac{cL_M}{cL_O} - 1)^2 + 2 \cdot (1 - cR_v)} \quad (14)$$

2.3 Comparing DEI with Root Mean Square Error and Kling-Gupta Efficiency

By comparing the DEI with other summarized metrics, we aimed to illustrate the advantage and suitability of DEI against RMSE and KGE in model evaluation. Given that KGE is a normalized index, we focus on the normalized versions of RMSE and DEI here.

Similar to nDEI (Eq. 9), the normalized RMSE (nRMSE) is also obtained by dividing by the observed SD, which can be written as:

$$n\text{RMSE} = \sqrt{\frac{ME^2}{SD_O^2} + (\frac{SD_M}{SD_O} - 1)^2 + 2 \cdot (1 - R) + 2 \cdot (\frac{SD_M}{SD_O} - 1) \cdot (1 - R)} \quad (15)$$

Here, we decomposed $nRMSE^2$ into four terms, and the first three terms are identical to the three independent error components in $nDEI^2$ (Eq. 9). The fourth term involves the non-linear interaction between anomaly amplitude and anomaly similarity errors of the model. This term tends to increase nRMSE when the modelled SD is overestimated, and R is positive, but decreases nRMSE when the modelled SD is underestimated. Consequently, $RMSE^2$ cannot be decomposed into the independent penalty components that separately target mean error, anomaly amplitude error, and anomaly similarity error. As a measure of the accuracy of the modelled field relative to the observed one, a smaller value of nRMSE usually indicates better performance.

The widely used KGE proposed by Gupta et al. (2009) is an expression of Euclidean distance away from perfect model performance by considering the mean, anomaly amplitude, and anomaly similarity errors simultaneously. However, the penalty term related to the mean error in traditional KGE is the ratio of the modelled mean value to the observed one, which may lead to an unrealistic estimation of mean error (Santos et al., 2018; Clark et al., 2021; Tang et al., 2021). To address this issue, a modified version of KGE (KGE") developed by Tang et al. (2021) is considered here:

$$KGE'' = 1 - \sqrt{\frac{ME^2}{SD_O^2} + (\frac{SD_M}{SD_O} - 1)^2 + (R - 1)^2} \quad (16)$$

The three terms under the square root penalize the mean error, anomaly amplitude error, and anomaly similarity error, respectively. For the mean and anomaly amplitude errors, KGE" adopts the same penalty terms as nDEI (Eq. 9). As for the anomaly similarity error, $(1+KGE")^2$ penalizes the square deviation of the correlation coefficient from its ideal value, whereas $nDEI^2$ penalizes twice that distance. As CORR ranges from -1 to 1, the penalty imposed by nDEI on the anomaly similarity error is always greater than that imposed by KGE". By adopting a stronger penalty on this component, nDEI places more emphasis on the anomaly similarity error compared with KGE". Normally, the mean and anomaly amplitude errors can be regarded as systematic errors of the climate model outputs, which can be adjusted conveniently by adding or multiplying by a scaling factor. In contrast, errors in the anomaly similarity are not easily remedied by such bias correction techniques. Hence,

https://doi.org/10.5194/egusphere-2026-5469
Preprint. Discussion started: 16 November 2026
© Author(s) 2026. CC BY 4.0 License.

greater attention should be given to a model's ability to reproduce anomaly similarity than to mean and anomaly amplitude errors. For this reason, nDEI offers a more valuable evaluation of model performance than KGE".

We further compared the three metrics of the overall model performance by exploring their relationship with CORR and SD_M (modelled SD), assuming the ME is zero, and the observed SD is equal to 1 (Fig. 2). For a given CORR, the optimal SD_M that minimizes nRMSE lies on the 1:1 line (left panel of Fig. 2a). In this case, an underestimated SD_M produces a smaller nRMSE than an SD_M equal to the observed SD. Clearly, nRMSE inherently favors models with underestimated SD. As a consequence, nRMSE may give misleading evaluations of model performance when the overall model performance evaluation focuses on the mean, anomaly amplitude, and anomaly similarity errors. This issue is particularly acute when the correlation coefficient is low (left panel of Fig. 2a). The limitation arises from the non-linear term in the nRMSE formula (Eq. 15) that couples the anomaly amplitude and similarity errors. For a given correlation coefficient, this non-linear term causes an underestimated SD_M to yield a smaller nRMSE than an overestimated SD_M with the same absolute deviation from SD_O (right panel of Fig. 2a). In contrast, both nDEI and KGE" avoid this pitfall. As shown in the left panels of Figs. 2b and c, nDEI decreases (KGE" increases) monotonically as either anomaly amplitude or similarity errors decrease. Furthermore, when the underestimated and overestimated SD_M have equal absolute deviation from SD_O, nDEI or KGE" exhibits the identical functional curve with varying CORR (the right panels of Figs. 2b and c).

Due to the distinct penalty ways for anomaly similarity error, the sensitivity of the KGE" value to CORR decreases when CORR is large compared with nDEI, and this issue is more evident with increased anomaly amplitude error (right panels of Figs. 2b and c, Fig. S1 in the Supporting Information). This implies that when CORR is relatively high, the improvement in CORR is less noticeably reflected in the KGE" value than in nDEI. When the model achieves high CORR but still suffers from sustainable mean and anomaly amplitude errors, KGE" assigns less importance to the anomaly similarity than nDEI, thereby shifting the model ranking criterion toward mean and anomaly amplitude errors. It suggests that KGE" is more suitable for model evaluation where mean error and/or anomaly amplitude error are the primary concern, or when CORR has already reached a satisfactory level.

However, in most of the climate model evaluations, the anomaly similarity is usually more important than the mean value and the anomaly amplitude, because without temporal and/or pattern similarity, the accuracy of mean and anomaly amplitude simulation is often less meaningful. In this context, the penalty design of nDEI makes it more suitable for most model evaluation applications compared with KGE". Moreover, nDEI becomes increasingly sensitive to CORR as CORR rises (Fig. S1), which constitutes another advantage of this index. This merit is particularly relevant because improving model performance in terms of anomaly correlation becomes progressively harder when the CORR is already high; for example, an increase in CORR from 0.80 to 0.85 may be as significant and difficult to achieve as an increase from 0.30 to 0.50. Thus, nDEI can reasonably reflect the influence of CORR changes, assigning due importance to both small improvements at high levels and large gains at low levels. In contrast, the KGE" exhibits a decreasing sensitivity as CORR increases, which runs counter to the practical requirement that even modest enhancements at high correlation are as valuable as substantial ones at low correlation.

https://doi.org/10.5194/egusphere-2026-5469
Preprint. Discussion started: 16 November 2026
© Author(s) 2026. CC BY 4.0 License.

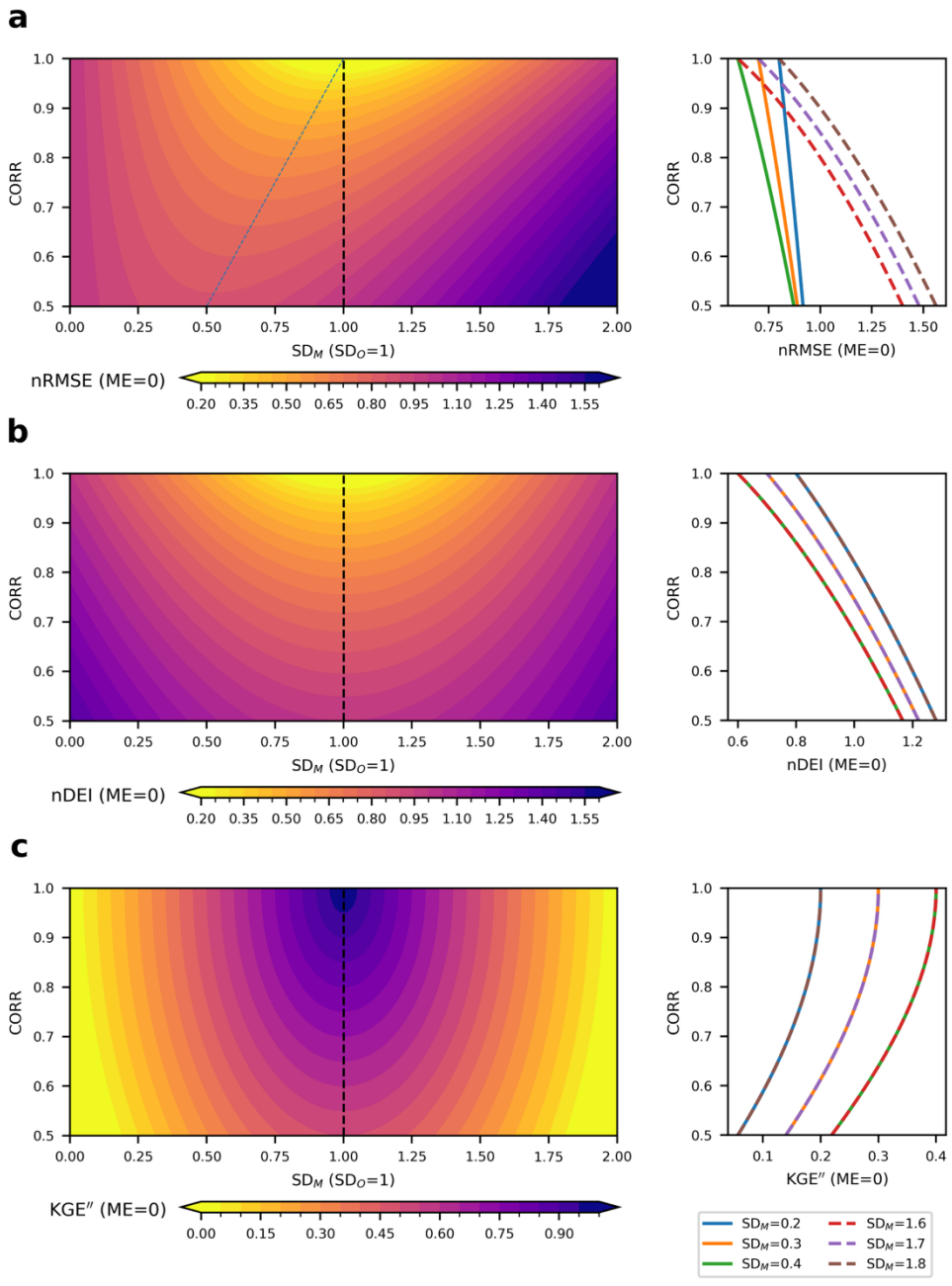


Figure 2: Relationship of (a) normalized Root Mean Square Error (nRMSE), (b) normalized Decomposable Error Index (nDEI), (c) modified Kling-Gupta Efficiency (KGE″) with CORR and SD_M, assuming that ME is zero and SD_O is 1. CORR (ME) is the correlation coefficient (mean error) between the model and the observation. SD_M (SD_O) indicates the modelled (observed) standard deviation.

https://doi.org/10.5194/egusphere-2026-5469
Preprint. Discussion started: 16 November 2026
© Author(s) 2026. CC BY 4.0 License.

3 Applying DEI in climate model evaluation

DEI indicates the overall performance of climate models, thereby favoring model ranking and selection. Quantifying the relative contribution of mean, anomaly amplitude, and anomaly similarity errors to DEI can help identify the dominant source of the model's overall error. Meanwhile, the three error penalty terms for the same counterparts enable investigation of the primary driver of differences in overall performance across multiple models, especially changes in performance from generation to generation.

With the aid of DEI and its decomposition, we evaluated CMIP model performance in simulating various variables over East Asia, referring to CORDEX-EA (0.2°N–50.5°N, 51.6°E–181.5°E). Here, we used monthly outputs from 26 CMIP6 models (Eyring et al., 2016; Table S1 in the Supporting Information) derived from historical experiments from 1979 to 2005. The evaluated variables include precipitation, near-surface temperature, as well as air temperature, geopotential height, specific humidity, and vector wind at 850 hPa, 500 hPa, and 200 hPa (Table S2 in the SI). Therein, we treated the meridional and zonal winds as the vector variable. The same air variables at different pressure levels are grouped into multiple fields for evaluation. The climatological means of the surface and air variables during December–January–February (DJF) and June–July–August (JJA) are evaluated in this section. In addition to 26 CMIP6 model simulations, we also used 13 CMIP5 model outputs, each paired with a corresponding CMIP6 model (Taylor et al., 2012; Table S3 in SI), to investigate generational differences in overall model performance.

For precipitation, we used the ensemble mean of three products as the reference, derived from the Multi-Source Weighted-Ensemble Precipitation (MSWEP; Beck et al., 2019), the Global Precipitation Climatology Center (GPCC; Schneider et al., 2011), and the Climatic Research Unit Time-Series (CRU TS; Harris et al., 2020). Regarding the near-surface temperature, the ensemble mean of three observations was also considered, including the Berkeley Earth Temperature (Rohde et al., 2020), CRU TS, and GHCN_CAMS Gridded 2m Temperature (Fan and Dool, 2008). European Centre for Medium-Range Weather Forecasts Reanalysis 5 (ERA5) was used as the reference for evaluating air variables simulated by climate models (Hersbach et al., 2023). Further details of the reference datasets are provided in Table S4 in the SI.

Prior to evaluation, all GCM simulations and reference datasets, except for precipitation, were regridded to 1.25º × 1.25º resolution by the bilinear interpolation method. For precipitation, we applied the conservative interpolation method to regrid both model simulations and observations to the same 1.25º × 1.25º grid. Note that all performance metrics take the area weighting into account in this section.

3.1 Model ranking by DEI

Figs. 3a and b present the overall performance measured by nDEI of 26 CMIP6 models and their multi-model ensemble mean (MME) in simulating the climatological mean of various variables during JJA and DJF over East Asia. Based on nDEI, we further rank these models. Generally, the MME ranks among the top six among 27 simulations (26 individual models plus MME) for all six variables during two seasons, demonstrating substantially better performance than any single CMIP6 model

https://doi.org/10.5194/egusphere-2026-5469
Preprint. Discussion started: 16 November 2026
© Author(s) 2026. CC BY 4.0 License.

(Figs. 3a and b). Although the overall performance of CMIP6 models varies with the variable and season, several models consistently exhibit better overall performance across all surface or air variables. In terms of two surface variables, EC-Earth and NorESM-MM perform better during JJA, and CESM2-WACCM, CESM2, and NorESM-MM stand out during DJF, all ranking within the top 25% of 26 CMIP6 models. As for the simulations of the four air variables, NorESM-MM outperforms the other models during JJA.

We also evaluated the overall performance using nRMSE and the modified KGE (KGE") for the same 26 CMIP6 models and their MME, and derived model rankings based on these metrics (Figs. S2 and S3 in the SI). The consistency between model rankings based on different summarized performance metrics varies with season, variables, and the specific metric pairs. In general, the model rankings by nDEI and nRMSE show closer agreement than those by nDEI and KGE" (Figs. 3a, 3b, S2, and S3). To quantify the agreement between two performance metrics in ranking models, we calculated the proportion of concordant pairs among all model pairs (Table S5 in the SI). For any two models, we examined whether two metrics agreed on which model performed better. Across possible pairs of models (here are 325 pairs), we counted the number of pairs where the two metrics ranked the two models in the same relative order, and the proportion of concordant pairs was then obtained by dividing this count by the total number of model pairs. Results show that over 90% of model pairs are ranked consistently by nDEI and nRMSE, while nDEI and KGE" yield consistent ranks for 71–92% of model pairs. These differences in model rankings arise from the distinct penalty forms applied to the mean, anomaly amplitude, and anomaly similarity errors and their trade-off in these metrics. Note that we did not attempt to judge the reasonableness of the three metrics here, because each metric has its suitable applications in the model evaluation, as discussed in Sect. 2.3.

When the focus is on the mean, anomaly amplitude, and anomaly similarity errors of climate models, both nDEI and KGE" are more reasonable indicators for model evaluation than nRMSE, as they vary monotonically with each of the errors when keeping the other two constant. Their property of error decomposition also allows us to examine how the penalty terms of each error influence the relative overall model performance indicated by these two metrics. We illustrated this using the performance of FGOALS-f3-L and CanESM5 in simulating the multi-level winds during JJA as an example (Figs. 3a and S3a). Based on nDEI (KGE"), FGOALS-f3-L ranks 8 (19), and CanESM5 ranks 22 (17). As shown in Fig. 3c, the mean error is the same for the two models, whereas FGOALS-f3-L (CanESM5) reproduces the anomaly similarity (amplitude) better than CanESM5 (FGOALS-f3-L). Note that the model ranks of the two models based on nDEI and KGE" are opposite: nDEI indicates that FGOALS-f3-L performs better, while KGE" indicates that CanESM5 performs better (Figs. 3a and S3a). For the sake of comparison with nDEI, the decomposition of $(1+\mathrm{KGE}")^2$ is presented here (formulas see Text S2 in the SI), with a smaller value indicating better overall performance. Through statistical error decomposition of the two metrics, the penalty terms targeting the mean error and the anomaly amplitude error are the same for each model in nDEI^2 and $(1+\mathrm{KGE}")^2$ (Fig. 3c). However, the penalty terms for the anomaly similarity error differ between the two metrics, with nDEI assigning greater relative importance to this component than KGE". In this case, the inter-model difference in penalty terms targeting the anomaly similarity error dominates the nDEI-based model rankings in this example, while those by KGE" depend more on the anomaly

https://doi.org/10.5194/egusphere-2026-5469
Preprint. Discussion started: 16 November 2026
© Author(s) 2026. CC BY 4.0 License.

300 amplitude error (Fig. 3a). Thus, different penalty forms of anomaly similarity errors affect the trade-off among the three error components in the nDEI and KGE", and consequently lead to the difference in model ranking based on these metrics.

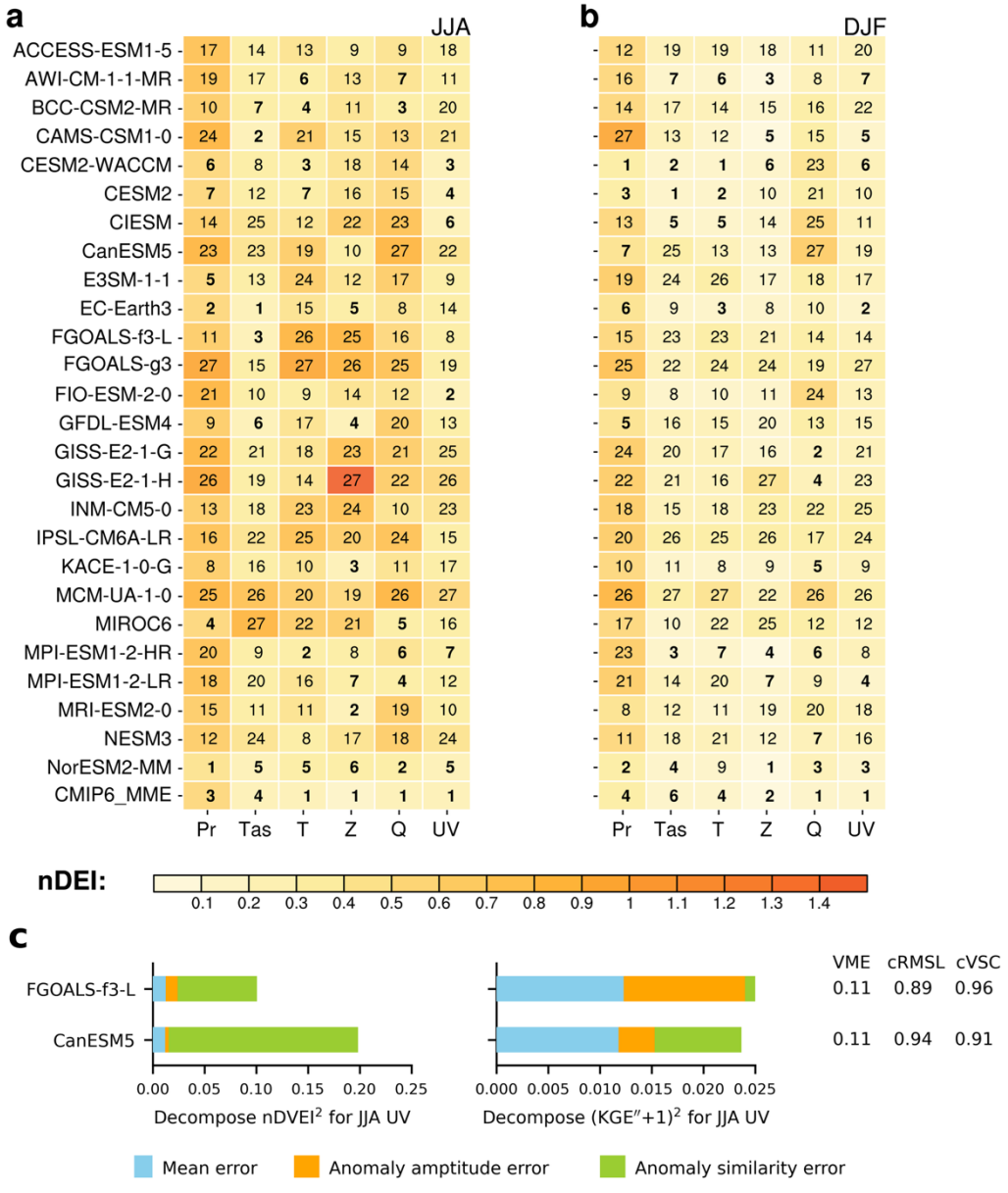


Figure 3: Overall performance by the normalized Decomposable Error Index (nDEI) of 26 CMIP6 models and their multi-model ensemble mean (MME) in simulating climatological mean fields of six variables during (a) June–July–August (JJA) and (b) 305 **December–January–February (DJF) throughout 1979–2005 over the CORDEX-EA region.** The color filled in each cell indicates the nDEI, and the value represents the model rank. Evaluated variables include precipitation (Pr), near-surface temperature (Tas), multi-level air temperatures (T), multi-level geopotential heights (Z), multi-level specific humidities (Q), and multi-level vector winds (UV). As T, Z, Q, and UV are multiple fields, the nDEI here is the normalized Decomposable Vector Error Index (nDVEI). (c) Decompose nDVEI^2 (left panel) and (KGE"+1)^2 (right panel) to the three error penalty terms targeting mean error, anomaly amplitude error, and anomaly similarity

https://doi.org/10.5194/egusphere-2026-5469
Preprint. Discussion started: 16 November 2026
© Author(s) 2026. CC BY 4.0 License.

error for FGOALS-f3-L and CanESM5 in simulating JJA UV fields. KGE" is the modified Kling-Gupta Efficiency. The values provided on the right of the bar chart are performance metrics of two models in terms of vector mean bias (VME), the ratio of modelled centered root-mean-square vector length to the reference one (cRMSL), and centered vector similarity correlation (cVSC).

3.2 Identifying sources of model overall bias through relative contributions

Fig. 4a demonstrates the relative contributions of mean, anomaly amplitude, and anomaly similarity errors to nDEI^2 of 26 CMIP6 models in simulating the climatological mean of precipitation during JJA throughout 1979–2005 over the East Asia region. Both the mean and anomaly amplitude errors separately explain 0–18% of DEI^2, while the anomaly similarity error accounts for 70–99% of DEI^2. Hence, the anomaly similarity error plays a dominant role in the overall performance of JJA precipitation across all 26 models.

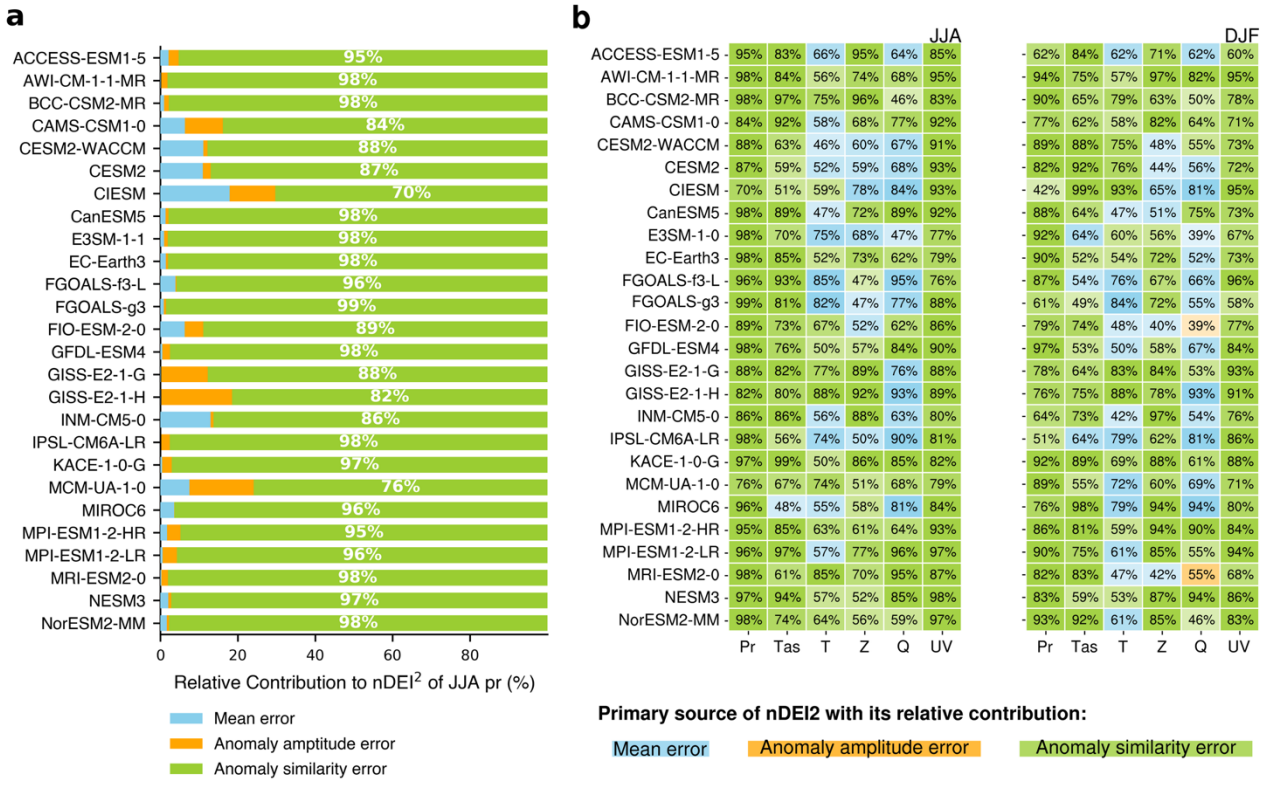


Figure 4: (a) Relative contributions of mean, anomaly amplitude, and anomaly similarity errors to the nDEI^2 for 26 CMIP6 models in simulating JJA precipitation in terms of the climatological mean over the CORDEX-EA region. (b) Primary source error (i.e., mean, anomaly amplitude, and anomaly similarity errors) for nDEI^2 in simulating the climatological mean of six variables during JJA and DJF for the same models and region. The value marked in each cell implies the relative contribution of the primary source error to nDEI^2 of a model. Different colors filled in the cell denote the primary source error for nDEI^2, with the color shade indicating its relative contribution to nDEI^2. Evaluated variables include precipitation (Pr), near-surface temperature (Tas), multi-level air temperatures (T), multi-level geopotential heights (Z), multi-level specific humidities (Q), and multi-level vector winds (UV). As T, Z, Q, and UV are multiple fields, the nDEI here is the nDVEI.

We then expanded this model evaluation to all six variables across two seasons and displayed the primary error source of nDEI^2 with its relative contribution in Fig. 4b. Overall, the primary source of the overall error varies considerably across

https://doi.org/10.5194/egusphere-2026-5469
Preprint. Discussion started: 16 November 2026
© Author(s) 2026. CC BY 4.0 License.

variables and seasons. For most models, their overall performance primarily depends on their ability to simulate anomaly similarity or mean values. For example, the anomaly similarity error mainly drives the overall performance of CESM2 in simulating the precipitation, near-surface temperature, and multi-level vector winds during JJA, while the performance for multi-level air temperatures, geopotential heights, and specific humidity is mainly affected by the mean error (Fig. 4b). Notably, regarding precipitation, near-surface temperature, and multi-level vector winds, the anomaly similarity error mainly determines the overall performance of over 88% models, where this source error explains over 42% of $nDEI^2$ (Fig. 4b). For precipitation and multi-level vector winds, it emerges as the primary error source for all 26 models. Therefore, alleviating models' overall bias in these variables should involve efforts to improve model performance in simulating pattern similarity. For precipitation, effective model modification may include increasing the horizontal and vertical resolutions for better representation of complex topography and cloud cover, improving the representation of aerosol, and improving the transport of cloud water (Kawai et al., 2019; Kim et al., 2019; Mauritsen et al., 2019; Swart et al., 2019a; Song et al., 2024; Zheng et al., 2024). Enhancing the spatial similarity of the near-surface temperature would require modifications such as increasing the vertical resolution to promote a more precise pattern of clouds and refining surface energy exchange by revising land use and land cover (Wilson et al., 2008; Guo et al., 2015; Mauritsen et al., 2019; Wu et al., 2019). Besides, the anomaly amplitude error plays a dominant role in the overall performance of multi-level specific humidities for FIO-ESM-2-0 and MRI-ESM2-0 during DJF (Fig. 4b).

3.3 Investigating inter-model differences in the overall error through decomposed error penalty terms

The three error penalty terms in DEI targeting mean, anomaly similarity, and anomaly amplitude errors provide a basis for attribution of inter-model differences in the overall performance. To demonstrate this application, we first compared the overall performance among climate models from the same generation, using the overall performance of 26 CMIP6 models in simulating JJA near-surface temperature throughout 1979–2005 as an example. For each pair of models, we quantify the relative contribution of each error to the inter-model $nDEI^2$ difference (see algorithm in Text S3 in the Supporting Information), thereby identifying the primary source error responsible for the inter-model difference in overall error. Fig. 5 indicates that the primary driver of inter-model differences in overall performance varies with different model pairs. Generally, anomaly similarity error emerges as the primary driver for 63% of all analysed model pairs, explaining over 51% of inter-model $nDEI^2$ differences. By contrast, the mean error underlies the overall error difference for 29% of CMIP6 model pairs, such as the relatively poor overall performance of CIESM and MIROC6 compared to the other models (Fig. 5). There are also some model pairs with the anomaly amplitude error predominantly leading to their $nDEI^2$ difference, including MRI-ESM2-0 .vs. NorESM2-MM, GISS-E2-1-H .vs. BCC-CSM2-MR, and CanESM5 .vs. MPI-ESM1-2-LR. Note that the relatively poor performance of MIROC6 compared to the other 25 models stems from its greater mean error. It indicates that MIROC6 suffers from systematic mean error among 26 CMIP6 models (Table S6).

https://doi.org/10.5194/egusphere-2026-5469
Preprint. Discussion started: 16 November 2026
© Author(s) 2026. CC BY 4.0 License.

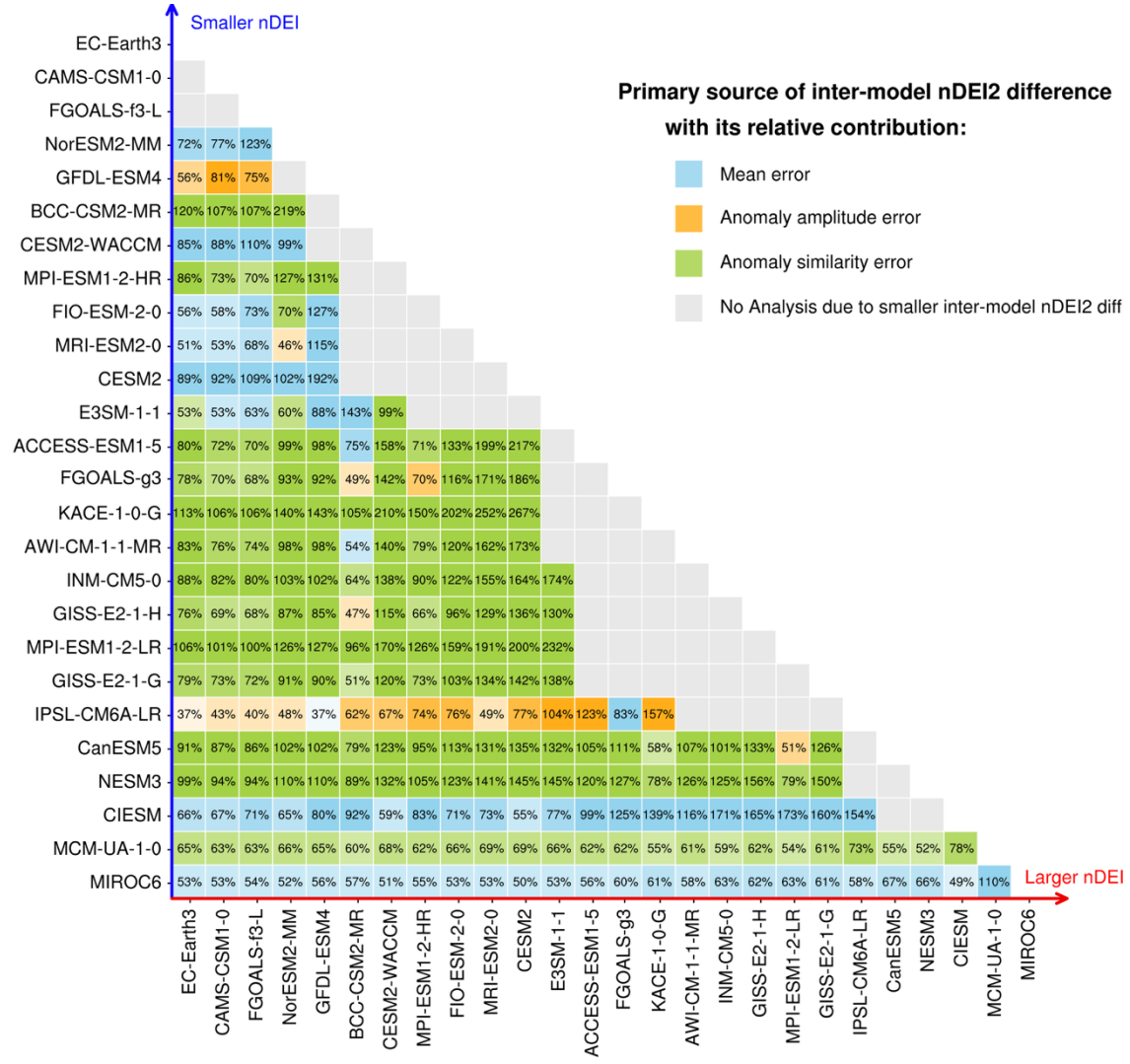


Figure 5: The primary source error (i.e., mean error, anomaly amplitude error, and anomaly similarity error) driving the inter-model difference in nDEI^2 among 26 CMIP6 models in simulating the climatological mean of JJA near-surface temperature throughout 1979–2005 over the CORDEX-EA region. The color filled in the cell indicates the different statistical errors. The value denotes the relative contribution of the primary error source to the inter-model nDEI^2 difference. Note that when the inter-model difference is small, with the absolute difference between the two models' nDEI smaller than 2% of the average of the two models' nDEI, the analysis for the primary error source of inter-model performance difference is not conducted.

This application also gives valuable insight into the changes in overall model performance across generations. Using 13 pairs of CMIP5 and CMIP6 models (Table S4 in the SI), we evaluated their overall performance in simulating six variables during JJA and DJF. For each pair, we quantified the change in DEI from CMIP5 to CMIP6 and attributed this change to the three error penalty terms. Fig. 6a displays the nDEI change for precipitation and near-surface temperature during JJA, along with the changes in the three error terms of nDEI. For the climatological mean, the CMIP6 model outperforms its CMIP5 predecessor for 10 model pairs in precipitation and 11 model pairs in near-surface temperature, with nDEI reductions of 2–

https://doi.org/10.5194/egusphere-2026-5469
Preprint. Discussion started: 16 November 2026
© Author(s) 2026. CC BY 4.0 License.

48%. In terms of precipitation, these improvements are driven by better representation of the spatial similarity across all model pairs (left panel of Fig. 6a). Regarding near-surface temperature, better spatial similarity also serves as the main contributor to the 9 pairs of CMIP5-CMIP models, with over 66% contributions for $nDEI^2$ change (right panel of Fig. 6a). Moreover, there are two model pairs,, i.e., INM and CanESM, the reduced mean error primarily promoting improvements from CMIP5 to CMIP6 models.

Fig. 6b extends the analysis to model performance of all six variables across two seasons. The direction of performance change from CMIP5 to CMIP6 varies considerably by variable, season, and model pair. No variable shows a consistent improvement across all 13 model pairs. Overall, CMIP6 outperforms CMIP5 in 7–11 model pairs depending on the variable. Notably, the BCC-CSM, EC-Earth, and NorESM pairs show superior CMIP6 performance across six variables, with average improvements of 27%, 43%, and 40%, respectively (Fig. 6b). Attribution of the $nDEI^2$ differences between CMIP5 and CMIP6 additionally reveals that for multi-level vector winds during two seasons, the generational improvement mainly stems from better anomaly spatial similarity for most model pairs, with relative contributions of more than 53% to the decreased $nDEI^2$. This issue may be related to the increase in horizontal resolution, implementation of shallow convection parameterization, enforcement of angular conservation, etc. (Tatebe et al., 2019; Semmlar et al., 2020; Döscher et al., 2022). In terms of multi-level air temperatures, the generational reduction in $nDEI^2$ is primarily due to the decreased mean error for most model pairs, with over 53% contribution. Moreover, there are some model pairs with the CMIP6 model displaying degraded overall performance against its CMIP5 predecessor, e.g., the CanESM and IPSL-CM pairs for multi-level specific humidity, and the driver for this degradation varies with different model pairs (Fig. 6b).

https://doi.org/10.5194/egusphere-2026-5469
Preprint. Discussion started: 16 November 2026
© Author(s) 2026. CC BY 4.0 License.

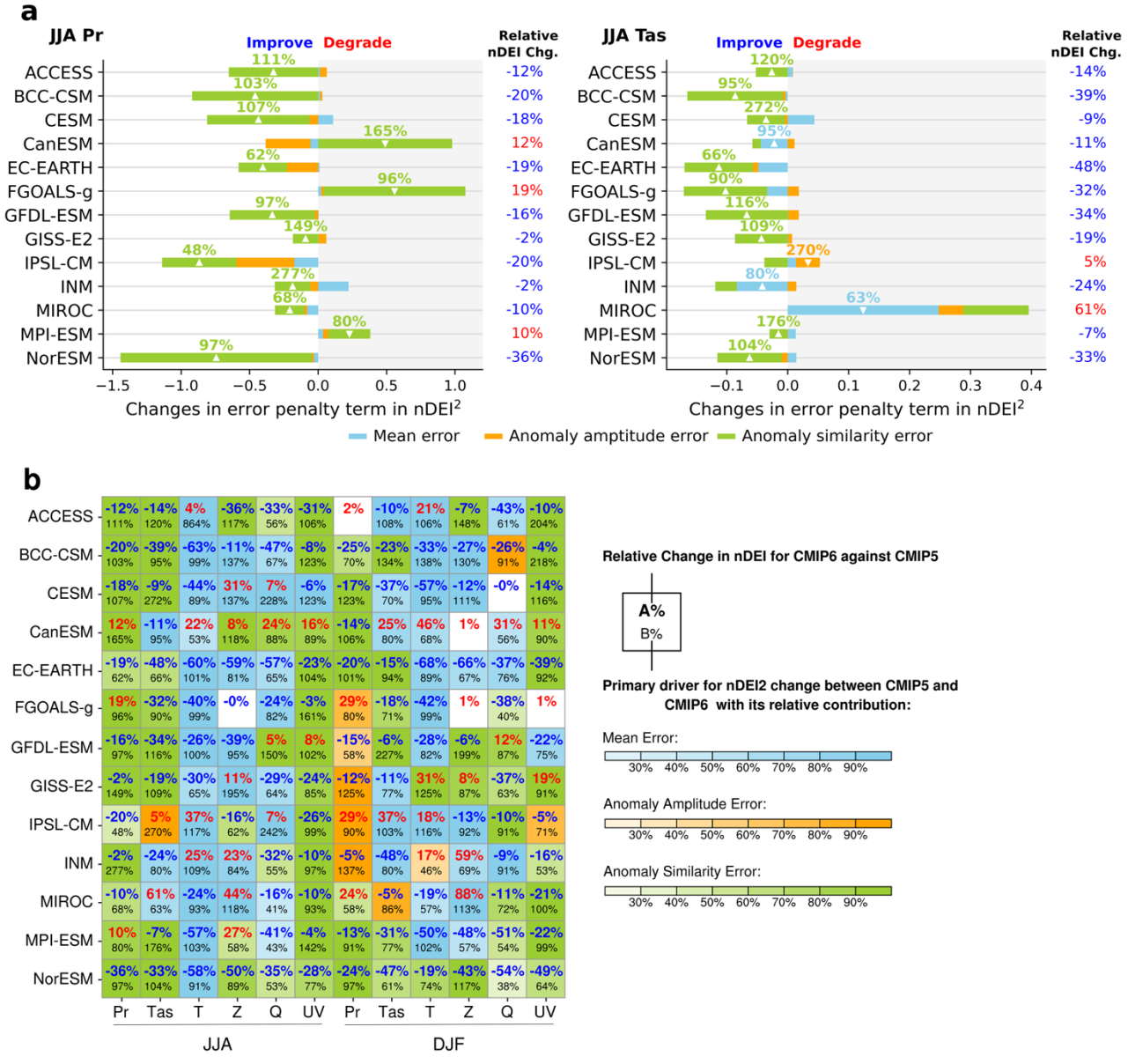


Figure 6: (a) Decompose nDEI² change from CMIP5 to CMIP6 into the changes in three penalty terms targeting mean, anomaly amplitude, and anomaly similarity errors for 13 model pairs' simulation of climatological mean precipitation (Pr) and near-surface temperature (Tas) during JJA throughout 1979–2005. Relative nDEI change in CMIP5 against CMIP6 models in blue (red) denotes an improved (degraded) overall performance for the CMIP6 model against its CMIP5 predecessor. The white triangle marked on the bar indicates the primary driver for model performance change, with the percent value for its contribution to the nDEI² change between the CMIP5 and CMIP6 models. **(b) The primary source error driving the CMIP5-CMIP6 difference in nDEI² for the six variables.** The color filled in the cell indicates the different statistical errors. As multi-level air temperature (T), multi-level geopotential height (Z), multi-level humidity (Q), and multi-level wind fields (UV) are multiple fields, the nDEI here is the nDVEI. Note that when the relative change in nDEI from CMIP5 to CMIP6 is smaller than 2%, the analysis for the primary source error of generational performance difference is not conducted.

https://doi.org/10.5194/egusphere-2026-5469
Preprint. Discussion started: 16 November 2026
© Author(s) 2026. CC BY 4.0 License.

4 Summary and discussion

This study developed a novel error-decomposition metric termed DEI indicating the overall model performance, providing valuable insight into how different statistical errors contribute to the overall bias. DEI comprises independent penalty terms targeting three source errors, i.e., mean error, anomaly similarity error, and anomaly amplitude error of a scalar field (Eqs. 2–5) based on the decomposition of RMSE. Each error penalty term keeps positive values, and DEI varies monotonically as each of the three source errors changes, avoiding the limitation of RMSE (Figs. 2a and b). The relative contributions of the three source errors to DEI^2 (Eqs. 6–8) are also defined to address how much the overall model performance stems from the mean error, how much is attributable to the difference in anomaly amplitude, and how much is due to poor similarity of anomaly fields between the model and observation. In addition, we proposed the normalized version of DEI (nDEI, Eq. 9) to meet evaluation needs across variables and regions.

Considering that the scalar field is a special case of vector/multiple fields, we further expanded the DEI to the multivariable integrated evaluation termed DVEI, along with the quantification of the relative contributions of three source errors to it. Related calculation formulas for DEI and DVEI are summarized in Table S7. Under the framework of the MVIE method, the multiple fields comprise various individual variables, either scalar or vector. Investigating the relationship of the error penalty terms of DVEI/DEI between the multiple fields and individual variables can help understand how the source errors of the individual variables influence those for multiple fields (Text S4 in the SI). Eq. (S17) indicates that the penalty term of mean error to DVEI in multiple fields is the linear sum of those for individual variables. By contrast, the penalty term of anomaly amplitude or similarity error to DVEI does not satisfy this relationship, and they involve additional terms accounting for the modelled and observed amplitude of the anomaly fields across all individual variables (Eqs. S18 and S19).

In this study, we demonstrated the differences between nDEI and other widely used metrics for the overall performance, nRMSE (normalized version of RMSE) and KGE'' (modified KGE), in terms of their evaluation context. Under the focus of climate model evaluation on the three source errors, despite being a straightforward measure of the overall model-observation deviation, nRMSE exhibits the limitation that it does not vary monotonically with the decrease in the anomaly amplitude error when keeping the mean and anomaly similarity errors held constant. In contrast, both nDEI and KGE'' satisfy the monotonicity requirement and can be decomposed into three independent error components that target mean, anomaly amplitude, and anomaly similarity errors. Building on different evaluation philosophies, nDEI and KGE'' differ in the penalization of the anomaly similarity error, with nDEI imposing stronger penalties on the anomaly similarity error in the design of the metric (Eqs. 2–5 and 16). This feature of nDEI provides a more practical basis for model evaluation and application because the anomaly similarity error is difficult to resolve through the correction process relative to the mean and anomaly amplitude errors. Thus, nDEI is applicable to most model evaluations of climate models. Besides, users can assign different weights to the three error components in nDEI according to evaluation needs.

As a summarized indicator for the overall model performance, DEI can guide the model ranking. As shown in Sect. 3.1, we evaluated 26 CMIP6 models in simulating climatological means of various variables over the East Asia region during

https://doi.org/10.5194/egusphere-2026-5469
Preprint. Discussion started: 16 November 2026
© Author(s) 2026. CC BY 4.0 License.

1979–2005, including precipitation, near-surface temperature, as well as multi-level air temperature, geopotential height, specific humidity, and vector winds. nDEI-based model ranking indicates that the overall performance of CMIP6 models varies with different variables and seasons, and no model excels in simulating all variables (Figs. 3a and b). Consistent with the related studies, nDEI-based model rankings indicate that CMIP6 MME generally outperforms individual models, ranking the top 6 among 26 CMIP6 models and MME (Guo et al., 2022; Zhang et al., 2022; Wei et al., 2023). Moreover, we also compared the model rankings based on nDEI with those based on nRMSE and KGE". Due to distinct penalty forms applied to the mean, anomaly amplitude, and anomaly similarity errors and their trade-off in these metrics, the ranking consistency between nDEI and nRMSE exceeds 90% of model pairs across six variables and two seasons, while ranging from 71% to 92% between nDEI and KGE" (Figs. 3a, S1, S2, and Table S5).

Moreover, the decomposed property of DEI provides valuable insights into the deficiencies of climate models from the perspective of the three statistical errors, informing model development to a certain extent. By analysing the relative contributions of mean, anomaly amplitude, and anomaly similarity errors to DEI, we can identify the source of the overall model error. Through its application to model evaluation, we found that the overall performance of 26 CMIP6 models is dominated by their ability to simulate anomaly similarity or mean values across six variables (Fig. 4). Therein, the anomaly similarity error primarily drives the overall performance of the precipitation and multi-level vector winds across all CMIP6 models, explaining over 42–99% of $nDEI^2$. This underscores the priority of improving pattern similarity for the simulations of these variables (Fig. 4b).

Previous studies have demonstrated that the horizontal resolution of climate models plays a vital role in the precipitation and wind simulations (Roberts et al., 2015; Huang et al., 2019a; Han et al., 2022; Zhang et al., 2022). In our study, we found significant positive correlations between horizontal resolution and the penalty term of anomaly similarity error in the DEI (E_{AnoS}) across 26 CMIP6 models for wind fields during both DJF and JJA, suggesting that finer resolution substantially improves the anomaly spatial pattern (Fig. 7). Similarly, the pattern similarity of JJA precipitation simulated by CMIP6 models is significantly affected by the horizontal resolution. This can be attributed to the fact that summer precipitation over East Asia is largely governed by mesoscale convective systems, representation of which benefits considerably from enhanced horizontal resolution in GCMs (Gao et al., 2006; Park et al., 2013; Guan et al., 2020). By contrast, no significant correlation is observed between model resolution and E_{AnoS} of DJF precipitation simulation. As winter precipitation in this region is largely associated with synoptic-scale systems, simply increasing resolution cannot adequately resolve the anomaly similarity error of winter precipitation in most models (Park et al., 2013). It is widely recognized that extratropical cyclones provide most of the precipitation during winter over the midlatitudes, modulated by ENSO, the Arctic Oscillation, and other factors (Catto et al., 2012; Hawcroft et al., 2012; Hu et al., 2018; He et al., 2022; Song & Zhou, 2022). Simultaneously, CMIP6 models still have common deficiencies in the simulation of ENSO teleconnection and Arctic sea ice (Watts et al., 2021; Jiang et al., 2022). Improving the performance of CMIP6 models in simulating winter precipitation over East Asia may need to start from these aspects.

https://doi.org/10.5194/egusphere-2026-5469
Preprint. Discussion started: 16 November 2026
© Author(s) 2026. CC BY 4.0 License.

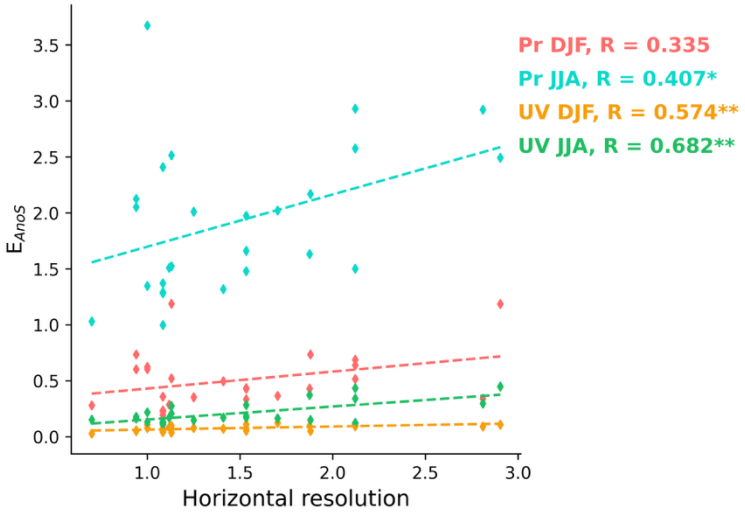


Figure 7: Relationship between horizontal resolution and the penalty term of anomaly similarity error in the DEI (E_{AnoS}) among 26 CMIP6 models in simulating precipitation (Pr) and multi-level wind fields (UV) during DJF and JJA. Horizontal resolution is measured by the square of the product of latitude and longitude resolutions (Table S2), and a lower value denotes a higher resolution. The diamond indicates a CMIP6 model. R represents the correlation coefficient between horizontal resolution and E_{AnoS}. The ** (*) behind the value of the correlation coefficient indicates the correlation passes the 0.01 (0.05) significance level by the R test.

Meanwhile, the absolute contributions of the three source errors help pinpoint the primary driver of the differences in overall performance among multiple models, such as the inter-model comparison of overall performance in simulating the climatological mean of near-surface temperature during JJA over East Asia between CMIP6 models (Fig. 5). For 63% of all analysed model pairs, anomaly similarity error dominates the inter-model overall performance difference, explaining over 51% of inter-model $nDEI^2$ differences. We also applied the absolute contributions to pinpoints the primary driver for overall performance changes in simulating of six variables between 13 CMIP5 and CMIP6 model pairs over the same region (Fig. 6). We found that, the reduced anomaly similarity error largely emerges as the primary driver for the generational improvement of JJA and DJF multi-level winds simulations as well as JJA precipitation, with relative contributions of more than 53% to the decreased $nDEI^2$ (Fig. 6b). In contrast, improved mean value mainly drives the better overall performance in multi-level air temperatures simulated by most CMIP6 models during JJA, with over 53% contribution. (Fig. 6b). Insights derived from the absolute contribution analysis may enhance the understanding of the inter-model differences in configurations and parameterization schemes. Note that compared with a single variable, the decomposition of the overall model performance regarding various variables gives more comprehensive information about model deficiencies and differences.

Furthermore, the error-decomposed DEI may assist in tuning the key properties of the global climate model (e.g., the radiation balance at the top of the atmosphere, cloud, sea surface temperature, and wind fields) to best reproduce the observed climate. Combining with the insightful understanding of the climate model, this can be achieved by adjusting parameters in cloud microphysics and convection schemes, climate sensitivity, etc. (Mauritsen et al., 2012, 2019; Kawai et al., 2019; Park et al., 2019; Wu et al., 2019). Since different aspects of model performance are influenced by different model parameters, the

https://doi.org/10.5194/egusphere-2026-5469
Preprint. Discussion started: 16 November 2026
© Author(s) 2026. CC BY 4.0 License.

DEI and its decomposition have the potential to improve the efficiency of model tuning to a certain extent. For example, applying the DEI to the precipitation simulated by a climate model allows for the identification of its primary error source, such as the mean error. Then, model parameters related to the precipitation mean error can be adjusted in priority order, which provides a targeted way to improve the accuracy of the precipitation simulation. Beyond the climate model evaluation, quantifying the overall discrepancy between two sets of samples across various disciplines, DEI can also serve as a summarized performance metric when optimizing parameters to train machine learning models. By applying this index, one can examine the influence of different hyperparameters and their varying values on the three aspects of model performance to a certain extent, ultimately providing additional insight into the hyperparameter-tuning process of the machine learning model.

https://doi.org/10.5194/egusphere-2026-5469
Preprint. Discussion started: 16 November 2026
© Author(s) 2026. CC BY 4.0 License.

Code and data availability

Code for generating the results illustrated in this study is available at https://doi.org/10.5281/zenodo.23681351.

CMIP6 simulations used in the study are derived from:

-ACCESS-ESM1-5, https://doi.org/10.22033/ESGF/CMIP6.4272
- AWI-CM-1-1-MR, https://doi.org/10.22033/ESGF/CMIP6.2686
- BCC-CSM2-MR, https://doi.org/10.22033/ESGF/CMIP6.2948
- CAMS-CSM1-0, https://doi.org/10.22033/ESGF/CMIP6.9754
- CESM2-WACCM, https://doi.org/10.22033/ESGF/CMIP6.10071
- CESM2, https://doi.org/10.22033/ESGF/CMIP6.7627
- CIESM, https://doi.org/10.22033/ESGF/CMIP6.8843
- CanESM5, https://doi.org/10.22033/ESGF/CMIP6.3610
- E3SM-1-1, https://doi.org/10.22033/ESGF/CMIP6.11485
- EC-Earth3, https://doi.org/10.22033/ESGF/CMIP6.4700
- FGOALS-f3-L, https://doi.org/10.22033/ESGF/CMIP6.3355
- FGOALS-g3, https://doi.org/10.22033/ESGF/CMIP6.3356
- FIO-ESM-2-0, https://doi.org/10.22033/ESGF/CMIP6.9199
- GFDL-ESM4, https://doi.org/10.22033/ESGF/CMIP6.8597
- GISS-E2-1-G, https://doi.org/10.22033/ESGF/CMIP6.7127
- GISS-E2-1-H, https://doi.org/10.22033/ESGF/CMIP6.7128
- INM-CM5-0, https://doi.org/10.22033/ESGF/CMIP6.5070
- IPSL-CM6A-LR, https://doi.org/10.22033/ESGF/CMIP6.5195
- KACE-1-0-G, https://doi.org/10.22033/ESGF/CMIP6.8378
- MCM-UA-1-0, https://doi.org/10.22033/ESGF/CMIP6.8888
- MIROC6, https://doi.org/10.22033/ESGF/CMIP6.5603
- MPI-ESM1-2-HR, https://doi.org/10.22033/ESGF/CMIP6.6594
- MPI-ESM1-2-LR, https://doi.org/10.22033/ESGF/CMIP6.6595
- MRI-ESM2-0, https://doi.org/10.22033/ESGF/CMIP6.6842
- NESM3, https://doi.org/10.22033/ESGF/CMIP6.8769
- NorESM2-MM, https://doi.org/10.22033/ESGF/CMIP6.8040

CMIP5 simulations used in the study are derived from:

- CSIRO-Mk3-6-0, https://doi.org/10.1594/wdcc/cmip5.cqmkhi
- BCC-CSM1-1, https://doi.org/10.1594/wdcc/cmip5.bcb1hi

https://doi.org/10.5194/egusphere-2026-5469
Preprint. Discussion started: 16 November 2026
© Author(s) 2026. CC BY 4.0 License.

- CNRM-CM5, https://doi.org/10.1594/wdcc/cmip5.cec5hi
- CanESM2, https://doi.org/10.22033/wdcc/cmip5.cce2hi
- EC-Earth, https://doi.org/10.5281/zenodo.24050128
- FGOALS-g2, https://doi.org/https://doi.org/10.1594/wdcc/cmip5.lsf2hi
- GFDL-ESM2M, https://doi.org/10.1594/wdcc/cmip5.ngemhi
- GISS-E2-R, https://doi.org/10.1594/wdcc/cmip5.gigrhi
- IPSL-CM5A-LR, https://doi.org/10.1594/wdcc/cmip5.ipilhi
- inmcm4, https://doi.org/10.1594/wdcc/cmip5.inc4hi
- MIRCO5, https://doi.org/10.1594/WDCC/CMIP5.MIM5hi
- MPI-ESM-MR, https://doi.org/10.1594/wdcc/cmip5.mxmrhi
- NorESM1-M, https://doi.org/10.1594/wdcc/cmip5.nccnmhi

The ERA5 reanalysis dataset can be accessed from https://doi.org/10.5281/zenodo.24158403, https://doi.org/10.5281/zenodo.24158650, https://doi.org/10.5281/zenodo.24158679, https://doi.org/10.5281/zenodo.24158665, and https://doi.org/10.5281/zenodo.24158606. Monthly precipitation data derived from MSWEP v2 used to construct our evaluation input can be obtained from https://doi.org/10.5281/zenodo.24051544. GPCC precipitation data is available at https://doi.org/10.5281/zenodo.24159880. CRU TS v4.09 precipitation and near-surface temperature data can be accessed at https://doi.org/10.5281/zenodo.24051071. Berkeley Earth near-surface temperature data is available at https://doi.org/10.5281/zenodo.24051305. CPC GHCN_CAMS Gridded 2m Temperature is available at https://doi.org/10.5281/zenodo.24050885.

Supplement link

The link to the supplement will be included by Copernicus, if applicable.

Author contributions

Meng-Zhuo Zhang: Conceptualization, Data curation, Formal analysis, Methodology, Visualization, Writing, Reviewing, Editing; Ying Han, Zhongfeng Xu: Conceptualization, Methodology, Supervision, Reviewing, Editing; Weidong Guo: Supervision, Reviewing.

https://doi.org/10.5194/egusphere-2026-5469
Preprint. Discussion started: 16 November 2026
© Author(s) 2026. CC BY 4.0 License.

Competing interests

All authors declare that they have no conflict of interest.

Disclaimer

Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

Acknowledgements

We thank the climate modelling groups involved in the CMIP project for producing and making available their model outputs. The ERA5 reanalysis data set was provided by the European Centre for Medium-Range Weather Forecasts.

Financial support

This research has been supported by the National Key Research and Development Program of China (2022YFF0802503), the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM907), and the National Natural Science Foundation of China (42405178). This work was also supported by the Jiangsu Collaborative Innovation Center for Climate Change and the National Key Scientific and Technological Infrastructure project "Earth System Numerical Simulation Facility" (EarthLab).

References

Beck, H. E., Wood, E. F., Pan, M., et al.: MSWEP v2 Global 3-hourly 0.1° precipitation: Methodology and quantitative assessment, Bulletin of the American Meteorological Society, 100(3), 473–500, 2019.

Catto, J. L., Jakob, C., Berry, G., and Nicholls, N.: Relating global precipitation to atmospheric fronts, Geophysical Research Letters, 39, 2012GL051736, https://doi.org/10.1029/2012GL051736, 2012.

Chen, D., M. Rojas, B.H. Samset, K. Cobb, et al.: Framing, Context, and Methods. In Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, pp. 147–286, 2021.

https://doi.org/10.5194/egusphere-2026-5469
Preprint. Discussion started: 16 November 2026
© Author(s) 2026. CC BY 4.0 License.

Clark, M. P., Vogel, R. M., Lamontagne, J. R., Mizukami, N., Knoben, W. J. M., Tang, G., Gharari, S., Freer, J. E., Whitfield, P. H., Shook, K. R., and Papalexiou, S. M.: The Abuse of Popular Performance Metrics in Hydrologic Modeling, Water Resources Research, 57, e2020WR029001, https://doi.org/10.1029/2020WR029001, 2021.

Deng, X., Perkins-Kirkpatrick, S. E., Lewis, S. C., and Ritchie, E. A.: Evaluation of Extreme Temperatures Over Australia in the Historical Simulations of CMIP5 and CMIP6 Models, Earth's Future, 9, e2020EF001902, https://doi.org/10.1029/2020EF001902, 2021.

Döscher, R., Acosta, M., Alessandri, A., Anthoni, P., Arsouze, T., Bergman, T., Bernardello, R., Boussetta, S., Caron, L.-P., Carver, G., Castrillo, M., Catalano, F., Cvijanovic, I., Davini, P., Dekker, E., Doblas-Reyes, F. J., Docquier, D., Echevarria, P., Fladrich, U., Fuentes-Franco, R., Gröger, M., V. Hardenberg, J., Hieronymus, J., Karami, M. P., Keskinen, J.-P., Koenigk, T., Makkonen, R., Massonnet, F., Ménégoz, M., Miller, P. A., Moreno-Chamarro, E., Nieradzik, L., Van Noije, T., Nolan, P., O'Donnell, D., Ollinaho, P., Van Den Oord, G., Ortega, P., Prims, O. T., Ramos, A., Reerink, T., Rousset, C., Ruprich-Robert, Y., Le Sager, P., Schmith, T., Schrödner, R., Serva, F., Sicardi, V., Sloth Madsen, M., Smith, B., Tian, T., Tourigny, E., Uotila, P., Vancoppenolle, M., Wang, S., Wårlind, D., Willén, U., Wyser, K., Yang, S., Yepes-Arbós, X., and Zhang, Q.: The EC-Earth3 Earth system model for the Coupled Model Intercomparison Project 6, Geosci. Model Dev., 15, 2973–3020, https://doi.org/10.5194/gmd-15-2973-2022, 2022.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: 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, 2016.

Eyring, V., Cox, P. M., Flato, G. M., Gleckler, P. J., Abramowitz, G., Caldwell, P., Collins, W. D., Gier, B. K., Hall, A. D., Hoffman, F. M., Hurtt, G. C., Jahn, A., Jones, C. D., Klein, S. A., Krasting, J. P., Kwiatkowski, L., Lorenz, R., Maloney, E., Meehl, G. A., Pendergrass, A. G., Pincus, R., Ruane, A. C., Russell, J. L., Sanderson, B. M., Santer, B. D., Sherwood, S. C., Simpson, I. R., Stouffer, R. J., and Williamson, M. S.: Taking climate model evaluation to the next level, Nature Clim Change, 9, 102–110, https://doi.org/10.1038/s41558-018-0355-y, 2019.

Fan, Y., and Dool, H. van den: A global monthly land surface air temperature analysis for 1948-present, J. Geophys. Res., 113, D01103, https://doi.org/10.1029/2007JD008470, 2008.

Flato, G., Marotzke, J., Abiodun, B., et al.: Evaluation of Climate Models, in: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, UK, New York, NY, USA, 746– 748, 2013.

Gao, X. J., Xu, Y., and Zhao, Z.: Impacts of horizontal resolution and topography on the numerical simulation of East Asia precipitation, Chinese Journal of Atmospheric Sciences, 30, 185-192, 2006.

Guan, P., Chen, G., Zeng, W., and Liu, Q.: Corridors of Mei-Yu-Season Rainfall over Eastern China, Journal of Climate, 33, 2603–2626, https://doi.org/10.1175/JCLI-D-19-0649.1, 2020.

Guo, H., Zhan, C., Ning, L., Li, Z., and Hu, S.: Evaluation and comparison of CMIP6 and CMIP5 model performance in simulating the runoff, Theor Appl Climatol, 149, 1451–1470, https://doi.org/10.1007/s00704-022-04118-0, 2022.

https://doi.org/10.5194/egusphere-2026-5469
Preprint. Discussion started: 16 November 2026
© Author(s) 2026. CC BY 4.0 License.

Guo, Z., Wang, M., Qian, Y., Larson, V. E., Ghan, S., Ovchinnikov, M., Bogenschutz, P. A., Zhao, C., Lin, G., and Zhou, T.: A sensitivity analysis of cloud properties to CLUBB parameters in the single-column Community Atmosphere Model (SCAM5), J Adv Model Earth Syst, 6, 829–858, https://doi.org/10.1002/2014MS000315, 2014.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, Journal of Hydrology, 377, 80–91, https://doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

Hagemann, S., Arpe, K., and Roeckner, E.: Evaluation of the Hydrological Cycle in the ECHAM5 Model, Journal of Climate, 19, 3810–3827, https://doi.org/10.1175/JCLI3831.1, 2006.

Han, Y., Zhang, M.-Z., Xu, Z., and Guo, W.: Assessing the performance of 33 CMIP6 models in simulating the large-scale environmental fields of tropical cyclones, Clim Dyn, 58, 1683–1698, https://doi.org/10.1007/s00382-021-05986-4, 2022.

Harris, I., Osborn, T.J., Jones, P. et al.: Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset, Sci Data 7, 109, https://doi.org/10.1038/s41597-020-0453-3, 2020.

Hawcroft, M. K., Shaffrey, L. C., Hodges, K. I., and Dacre, H. F.: How much Northern Hemisphere precipitation is associated with extratropical cyclones?, Geophysical Research Letters, 39, 2012GL053866, https://doi.org/10.1029/2012GL053866, 2012.

He, S., Gao, Y., Li, F., Wang, H., and He, Y.: Impact of Arctic Oscillation on the East Asian climate: A review, Earth-Science Reviews, 164, 48–62, https://doi.org/10.1016/j.earscirev.2016.10.014, 2017.

Hersbach, H. B. B., Berrisford, P., Biavati, G., et al.: ERA5 monthly averaged data on pressure levels from 1940 to present, Copernicus Climate Change Service (C3S) Climate Data Store (CDS), https://doi.org/10.24381/cds.6860a573 (Accessed on 01-01-2025), 2023.

Hu, K., Huang, G., Wu, R., and Wang, L.: Structure and dynamics of a wave train along the wintertime Asian jet and its impact on East Asian climate, Clim Dyn, 51, 4123–4137, https://doi.org/10.1007/s00382-017-3674-1, 2018.

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

Hwong, Y. -L., Colin, M., Aglas-Leitner, P., Muller, C. J., and Sherwood, S. C.: Assessing Memory in Convection Schemes Using Idealized Tests, J Adv Model Earth Syst, 15, e2023MS003726, https://doi.org/10.1029/2023MS003726, 2023.

Jiang, W., Gong, H., Huang, P., Wang, L., Huang, G., and Hu, L.: Biases and improvements of the ENSO-East Asian winter monsoon teleconnection in CMIP5 and CMIP6 models, Clim Dyn, 59, 2467–2480, https://doi.org/10.1007/s00382-022-06220-5, 2022.

Kawai, H., Yukimoto, S., Koshiro, T., Oshima, N., Tanaka, T., Yoshimura, H., and Nagasawa, R.: Significant improvement of cloud representation in the global climate model MRI-ESM2, Geosci. Model Dev., 12, 2875–2897, https://doi.org/10.5194/gmd-12-2875-2019, 2019.

https://doi.org/10.5194/egusphere-2026-5469
Preprint. Discussion started: 16 November 2026
© Author(s) 2026. CC BY 4.0 License.

Khadka, D., Babel, M. S., Abatan, A. A., and Collins, M.: An evaluation of CMIP5 and CMIP6 climate models in simulating summer rainfall in the Southeast Asian monsoon domain, Intl Journal of Climatology, 42, 1181–1202, https://doi.org/10.1002/joc.7296, 2022.

Kim, I.-W., Oh, J., Woo, S., and Kripalani, R. H.: Evaluation of precipitation extremes over the Asian domain: observation and modelling studies, Clim Dyn, 52, 1317–1342, https://doi.org/10.1007/s00382-018-4193-4, 2019.

Kim, M., Yu, D., Oh, J., Byun, Y., Boo, K., Chung, I., Park, J., Park, D. R., Min, S., and Sung, H. M.: Performance Evaluation of CMIP5 and CMIP6 Models on Heatwaves in Korea and Associated Teleconnection Patterns, JGR Atmospheres, 125, e2020JD032583, https://doi.org/10.1029/2020JD032583, 2020.

Knutti, R., Masson, D., and Gettelman, A.: Climate model genealogy: Generation CMIP5 and how we got there, n.d.

Li, X., Feng, M., Ran, Y., Su, Y., Liu, F., Huang, C., Shen, H., Xiao, Q., Su, J., Yuan, S., and Guo, H.: Big Data in Earth system science and progress towards a digital twin, Nat Rev Earth Environ, 4, 319–332, https://doi.org/10.1038/s43017-023-00409-w, 2023.

Mauritsen, T., Bader, J., Becker, T., Behrens, J., Bittner, M., Brokopf, R., Brovkin, V., Claussen, M., Crueger, T., Esch, M., Fast, I., Fiedler, S., Fläschner, D., Gayler, V., Giorgetta, M., Goll, D. S., Haak, H., Hagemann, S., Hedemann, C., Hohenegger, C., Ilyina, T., Jahns, T., Jimenéz-de-la-Cuesta, D., Jungclaus, J., Kleinen, T., Kloster, S., Kracher, D., Kinne, S., Kleberg, D., Lasslop, G., Kornblueh, L., Marotzke, J., Matei, D., Meraner, K., Mikolajewicz, U., Modali, K., Möbis, B., Müller, W. A., Nabel, J. E. M. S., Nam, C. C. W., Notz, D., Nyawira, S., Paulsen, H., Peters, K., Pincus, R., Pohlmann, H., Pongratz, J., Popp, M., Raddatz, T. J., Rast, S., Redler, R., Reick, C. H., Rohrschneider, T., Schemann, V., Schmidt, H., Schnur, R., Schulzweida, U., Six, K. D., Stein, L., Stemmler, I., Stevens, B., Von Storch, J., Tian, F., Voigt, A., Vrese, P., Wieners, K., Wilkenskjeld, S., Winkler, A., and Roeckner, E.: Developments in the MPI-M Earth System Model version 1.2 (MPI-ESM1.2) and Its Response to Increasing CO_2, J Adv Model Earth Syst, 11, 998–1038, https://doi.org/10.1029/2018MS001400, 2019.

Mauritsen, T., Stevens, B., Roeckner, E., Crueger, T., Esch, M., Giorgetta, M., Haak, H., Jungclaus, J., Klocke, D., Matei, D., Mikolajewicz, U., Notz, D., Pincus, R., Schmidt, H., and Tomassini, L.: Tuning the climate of a global model, J Adv Model Earth Syst, 4, 2012MS000154, https://doi.org/10.1029/2012MS000154, 2012.

Meehl, G. A., Moss, R., Taylor, K. E., Eyring, V., Stouffer, R. J., Bony, S., and Stevens, B.: Climate Model Intercomparisons: Preparing for the Next Phase, EoS Transactions, 95, 77–78, https://doi.org/10.1002/2014EO090001, 2014.

Murphy, A.: Skill scores based on the mean square error and their relationships to the correlation coefficient, Monthly Weather Review, 116, 2417–2424, 1988.

Pan, X., Chen, D., Pan, B., Huang, X., Yang, K., Piao, S., Zhou, T., Dai, Y., Chen, F., and Li, X.: Evolution and prospects of Earth system models: Challenges and opportunities, Earth-Science Reviews, 260, 104986, https://doi.org/10.1016/j.earscirev.2024.104986, 2025.

Park, J., Oh, S., and Suh, M.: Impacts of boundary conditions on the precipitation simulation of RegCM4 in the CORDEX East Asia domain, JGR Atmospheres, 118, 1652–1667, https://doi.org/10.1002/jgrd.50159, 2013.

https://doi.org/10.5194/egusphere-2026-5469
Preprint. Discussion started: 16 November 2026
© Author(s) 2026. CC BY 4.0 License.

Park, S., Shin, J., Kim, S., Oh, E., and Kim, Y.: Global Climate Simulated by the Seoul National University Atmosphere Model Version 0 with a Unified Convection Scheme (SAM0-UNICON), Journal of Climate, 32, 2917–2949, https://doi.org/10.1175/JCLI-D-18-0796.1, 2019.

Roberts, M. J., Vidale, P. L., Mizielinski, M. S., Demory, M.-E., Schiemann, R., Strachan, J., Hodges, K., Bell, R., and Camp, J.: Tropical Cyclones in the UPSCALE Ensemble of High-Resolution Global Climate Models*, Journal of Climate, 28, 574–596, https://doi.org/10.1175/JCLI-D-14-00131.1, 2015.

Roeckner, E., Brokopf, R., Esch, M., Giorgetta, M., Hagemann, S., Kornblueh, L., Manzini, E., Schlese, U., and Schulzweida, U.: Sensitivity of Simulated Climate to Horizontal and Vertical Resolution in the ECHAM5 Atmosphere Model, Journal of Climate, 19, 3771–3791, https://doi.org/10.1175/JCLI3824.1, 2006.

Rohde, R. A. and Hausfather, Z.: The Berkeley Earth Land/Ocean Temperature Record, Earth Syst. Sci. Data, 12, 3469–3479, https://doi.org/10.5194/essd-12-3469-2020, 2020.

Santos, L., Thirel, G., and Perrin, C.: Technical note: Pitfalls in using log-transformed flows within the KGE criterion, Hydrol. Earth Syst. Sci., 22, 4583–4591, https://doi.org/10.5194/hess-22-4583-2018, 2018.

Schneider, U., Becker, A., Finger, P. et al.: GPCC Full Data Reanalysis Version 6.0 at 1.0°: Monthly Land-Surface Precipitation from Rain-Gauges built on GTS-based and Historic Data, https://doi.org/10.5676/DWD_GPCC/FD_M_V7_100, 2011.

Semmler, T., Danilov, S., Gierz, P., Goessling, H. F., Hegewald, J., Hinrichs, C., Koldunov, N., Khosravi, N., Mu, L., Rackow, T., Sein, D. V., Sidorenko, D., Wang, Q., and Jung, T.: Simulations for CMIP6 With the AWI Climate Model AWI-CM-1-1, J Adv Model Earth Syst, 12, e2019MS002009, https://doi.org/10.1029/2019MS002009, 2020.

Simpkins, G.: Progress in climate modelling, Nature Clim Change, 7, 684–685, https://doi.org/10.1038/nclimate3398, 2017.

Song, X., Xu, M., Kang, S., Wang, R., and Wu, H.: Evaluation and projection of changes in temperature and precipitation over Northwest China based on CMIP6 models, Intl Journal of Climatology, 44, 5039–5056, https://doi.org/10.1002/joc.8622, 2024.

Song, Z. and Zhou, B.: CMIP6 projected response of the East Asian winter climate to the sea ice-free Arctic, Intl Journal of Climatology, 43, 1529–1542, https://doi.org/10.1002/joc.7931, 2023.

Swart, N. C., Cole, J. N. S., Kharin, V. V., Lazare, M., Scinocca, J. F., Gillett, N. P., Anstey, J., Arora, V., Christian, J. R., Hanna, S., Jiao, Y., Lee, W. G., Majaess, F., Saenko, O. A., Seiler, C., Seinen, C., Shao, A., Sigmond, M., Solheim, L., Von Salzen, K., Yang, D., and Winter, B.: The Canadian Earth System Model version 5 (CanESM5.0.3), Geosci. Model Dev., 12, 4823–4873, https://doi.org/10.5194/gmd-12-4823-2019, 2019a.

Tang, G., Clark, M. P., and Papalexiou, S. M.: The use of serially complete station data to improve the temporal continuity of gridded precipitation and temperature estimates, Journal of Hydrometeorology, https://doi.org/10.1175/JHM-D-20-0313.1, 2021.

Tatebe, H., Ogura, T., Nitta, T., Komuro, Y., Ogochi, K., Takemura, T., Sudo, K., Sekiguchi, M., Abe, M., Saito, F., Chikira, M., Watanabe, S., Mori, M., Hirota, N., Kawatani, Y., Mochizuki, T., Yoshimura, K., Takata, K., O'ishi, R., Yamazaki, D., Suzuki, T., Kurogi, M., Kataoka, T., Watanabe, M., and Kimoto, M.: Description and basic evaluation of simulated mean state,

https://doi.org/10.5194/egusphere-2026-5469
Preprint. Discussion started: 16 November 2026
© Author(s) 2026. CC BY 4.0 License.

internal variability, and climate sensitivity in MIROC6, Geosci. Model Dev., 12, 2727–2765, https://doi.org/10.5194/gmd-12-2727-2019, 2019.

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

Taylor, K.E., R.J. Stouffer, G.A. Meehl: An Overview of CMIP5 and the experiment design, Bull. Amer. Meteor. Soc., **93**, 485-498, doi:10.1175/BAMS-D-11-00094.1, 2012.

Volosciuk, C., Maraun, D., Semenov, V. A., and Park, W.: Extreme Precipitation in an Atmosphere General Circulation Model: Impact of Horizontal and Vertical Model Resolutions, Journal of Climate, 28, 1184–1205, https://doi.org/10.1175/JCLI-D-14-00337.1, 2015.

Watts, M., Maslowski, W., Lee, Y. J., Kinney, J. C., and Osinski, R.: A Spatial Evaluation of Arctic Sea Ice and Regional Limitations in CMIP6 Historical Simulations, Journal of Climate, 34, 6399–6420, https://doi.org/10.1175/JCLI-D-20-0491.1, 2021.

Wei, L., Xin, X., Li, Q., Wu, Y., Tang, H., Li, Y., and Yang, B.: Simulation and projection of climate extremes in China by multiple Coupled Model Intercomparison Project Phase 6 models, Intl Journal of Climatology, 43, 219–239, https://doi.org/10.1002/joc.7751, 2023.

Wilson, D. R., Bushell, Andrew. C., Kerr-Munslow, A. M., Price, J. D., Morcrette, C. J., and Bodas-Salcedo, A.: PC2: A prognostic cloud fraction and condensation scheme. II: Climate model simulations, Quart J Royal Meteoro Soc, 134, 2109–2125, https://doi.org/10.1002/qj.332, 2008.

Wu, T., Lu, Y., Fang, Y., Xin, X., Li, L., Li, W., Jie, W., Zhang, J., Liu, Y., Zhang, L., Zhang, F., Zhang, Y., Wu, F., Li, J., Chu, M., Wang, Z., Shi, X., Liu, X., Wei, M., Huang, A., Zhang, Y., and Liu, X.: The Beijing Climate Center Climate System Model (BCC-CSM): the main progress from CMIP5 to CMIP6, Geosci. Model Dev., 12, 1573–1600, https://doi.org/10.5194/gmd-12-1573-2019, 2019.

Wu, T.: A mass-flux cumulus parameterization scheme for large-scale models: description and test with observations, Clim Dyn, 38, 725–744, https://doi.org/10.1007/s00382-011-0995-3, 2012.

Xu, Z. and Han, Y.: Short communication comments on 'DISO: A rethink of Taylor diagram,' Intl Journal of Climatology, 40, 2506–2510, https://doi.org/10.1002/joc.6359, 2020.

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

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

Yin, L., Fu, R., Shevliakova, E., and Dickinson, R. E.: How well can CMIP5 simulate precipitation and its controlling processes over tropical South America?, Clim Dyn, 41, 3127–3143, https://doi.org/10.1007/s00382-012-1582-y, 2013.

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

https://doi.org/10.5194/egusphere-2026-5469
Preprint. Discussion started: 16 November 2026
© Author(s) 2026. CC BY 4.0 License.

Zhang, M.-Z., Xu, Z., Han, Y., and Guo, W.: Evaluation of CMIP6 models toward dynamical downscaling over 14 CORDEX domains, Clim Dyn, 62, 4475–4489, https://doi.org/10.1007/s00382-022-06355-5, 2024.

Zheng, M., Wu, T., Xin, X., Liu, X., Lu, Y., Jie, W., Xie, C., and Zhou, Y.: Simulation of MJO with improved deep convection scheme in different resolutions of BCC-CSM2 models, Clim Dyn, 62, 2161–2185, https://doi.org/10.1007/s00382-023-07015-y, 2024.

Zhu, H., Jiang, Z., Li, J., Li, W., Sun, C., and Li, L.: Does CMIP6 Inspire More Confidence in Simulating Climate Extremes over China?, Adv. Atmos. Sci., 37, 1119–1132, https://doi.org/10.1007/s00376-020-9289-1, 2020.