

Evaluation of Climate Models Using Bergen Metrics as a Multi-Metrics Comparison over Indonesia

Mas Agus MARDYANTO^{1,*}, Asyam Mulayyan DARY¹, R Irwan Bagyo SANTOSO¹, Ervin NURHAYATI¹, Hilmi Putra PRADANA¹, Achmad MUZAKKY¹

¹ Department of Environmental Engineering, Sepuluh Nopember Institute of Technology, Surabaya, Indonesia.

* corresponding author: mardyanto@its.ac.id

Date of Submission: 23 January 2026
Revision Date: 23 February 2026
Date of Acceptance: 24 February 2026



Civil and Environmental Engineering
Journal of the Faculty of Civil Engineering | University of Žilina

Abstract

Recent climate events have become increasingly unpredictable, emphasizing the need for climate model projections with reliable performance. Future climate projections derived from the Shared Socioeconomic Pathways of Coupled Model Intercomparison Project Phase 6 require careful evaluation against historical observations to ensure their applicability. This study evaluates six CMIP6 historical simulations namely ACCESS-CM2, CNRM-CM6-HR, EC-Earth3, EC-Earth3-Veg, HadGEM3-GC31-MM, and MPI-ESM1-2-HR using cases in Indonesia. Model performance was assessed using the SA-OBS dataset with a horizontal resolution of 0.25 degrees. Five evaluation metrics were applied namely Taylor Skill Score, Interannual Variability Score, normalized Mean Absolute Error, Kling Gupta Efficiency, and Willmott Index. These metrics were synthesized using the Bergen Metrics framework to provide an integrated measure of model performance. In addition, a rank based weighting approach was employed as a comparative assessment. Results indicate that all models show limited skill in simulating precipitation, while minimum, mean, and maximum temperatures are reproduced more accurately. Across all variables, EC-Earth3-Veg consistently achieved the best overall performance, followed closely by EC-Earth3. The remaining models exhibited comparable performance but ranked lower. These findings are consistent with previous regional evaluations over Southeast Asia. The relatively strong performance of EC-Earth3-based models is likely associated with their finer spatial resolution. Overall, this study demonstrates that Bergen Metrics provide a robust and flexible framework for climate model evaluation, while rank based weighting offers complementary insight into relative model consistency.

Keywords

Bergen metrics; Climate model evaluation; Climate projection validation; Metrics evaluation; Indonesia.

1. Introduction

Recent changes in global climate have occurred more rapidly than previously anticipated, with increasing surface temperatures and enhanced climate variability observed across many regions (Beaugrand et al., 2025; Hansen et al., 2025). While global mean surface temperatures have risen by approximately 0.8°C to 0.9°C since the 1980s, recent observational data indicates that the warming rate has accelerated from 0.18°C per decade (1970–2010) to nearly 0.30°C per decade in the post-2010 period (Forster et al., 2025). This rapid escalation is primarily driven by the sustained accumulation of anthropogenic greenhouse gases, further compounded by a significant reduction in the cooling effect of sulphate aerosols

(Hansen et al., 2025). In addition to temperature changes, global precipitation patterns have also shifted, with increasing trends observed in many tropical regions and decreasing trends in parts of the subtropics and mid latitudes (Gu & Adler, 2018; Nazeri Tahroudi, 2025). These changes have intensified the occurrence of extreme events such as floods (A. Ahmed et al., 2022; Fofana et al., 2022) and droughts in recent decades (Ha et al., 2022; Tripathy et al., 2023). Such developments highlight the urgent need for effective climate mitigation and adaptation strategies supported by reliable climate projections (Ouyang et al., 2025).

Climate projections are commonly derived from the Coupled Model Intercomparison Project, a coordinated framework that facilitates the comparison of global climate models developed by research institutions worldwide (Eyring et al., 2016). The most recent phase, CMIP6, provides simulations under the Shared Socioeconomic Pathways that integrate greenhouse gas emission trajectories with socioeconomic development scenarios (Riahi et al., 2017). These simulations have been widely used in climate change impact assessments at global and regional scales (Gebrechorkos et al., 2023; Gutowski Jr. et al., 2016; Tebaldi et al., 2021). However, the performance of CMIP6 models varies substantially across regions and climate variables, particularly in regions with complex atmospheric and oceanic interactions such as Southeast Asia (Khadka et al., 2022; Ngo-Duc et al., 2024). Therefore, systematic evaluation of model performance against historical observations is essential before applying these datasets to impact studies (Wagener et al., 2022).

The importance of accurate climate model evaluation is further underscored by its direct implications for hazard risk assessment and adaptation planning. Studies in Southeast Asia have demonstrated that precipitation projections from downscaled climate models serve as the primary input for projecting future flood and drought occurrence at the basin scale, where errors in simulated precipitation propagate directly into hydrological risk estimates (Voon et al., 2022). In Indonesia specifically, precipitation intensity is also closely linked to slope instability and landslide hazard, particularly in mountainous regions where high rainfall rates trigger slope failure (Nofrizal et al., 2025). These downstream applications make it critical that the GCMs used as forcing data adequately represent the spatial and temporal characteristics of precipitation, reinforcing the need for rigorous multi-metric evaluation prior to any regional application.

Evaluating climate model performance is inherently challenging because models must reproduce mean climate states, spatial patterns, temporal variability, and error magnitude across multiple scales (Eyring et al., 2019; Karpasitis et al., 2025). Numerous statistical metrics have been developed to quantify different aspects of model skill, including correlation based (Taylor, 2001), variability based (Gupta et al., 2009), and error-based measures (Legates & McCabe, 1999). Reliance on a single metric may lead to misleading conclusions, as a model may perform well according to one criterion while performing poorly according to another. Consequently, multi metric evaluation frameworks have gained increasing attention as a means to provide a more comprehensive assessment of model performance (Eyring et al., 2019).

Despite the growing number of CMIP6 evaluations, relatively few studies have applied integrative frameworks that combine multiple performance metrics into a single diagnostic measure (Akinsanola et al., 2021; Desmet & Ngo-Duc, 2022). Many previous studies focus on individual metrics or specific climate variables, often overlooking the integration between spatial, temporal, and statistical dimensions of model performance. To address this limitation, this study applies the Bergen Metrics framework, which integrates multiple evaluation metrics using Euclidean distance (Samantaray et al., 2024). This approach offers the advantage of simultaneously accounting for spatial, temporal, and magnitude-based performance in a transparent and reproducible manner. However, it carries an inherent limitation in that equal weighting is assigned to all constituent metrics by default, which may not reflect the relative importance of each metric for specific applications, and the Euclidean distance measure is sensitive to the scaling of individual metrics.

The objective of this study is to evaluate the performance of six CMIP6 models over Indonesia using the SA-OBS observational dataset. Model skills are assessed for precipitation, minimum temperature, mean temperature, and maximum temperature using five complementary metrics. The results are synthesized using Bergen Metrics and compared with a rank based weighting approach to identify robust model performance patterns. This study aims to support informed model selection for regional climate analysis, downscaling, and impact assessment.

2. Methodology

2.1. Domain of Research

Indonesia is situated in the maritime continent, spanning approximately 95°E to 141°E longitude and 6°N to 11°S latitude, making it one of the most climatically complex regions in the world. The country consists of over 17,000 islands encompassing a total land area of roughly 1.9 million km², with the five major islands of Sumatra, Java, Kalimantan, Sulawesi, and Papua accounting for most of its landmass. Its equatorial position subjects the entire archipelago to a tropical climate characterized by high temperatures, high humidity, and abundant rainfall throughout the year. Mean annual temperatures generally range between 25°C and 28°C at low elevations, with limited seasonal variation in temperature but pronounced seasonality in precipitation driven by the Asian-Australian monsoon system. The northwest monsoon (November to March) brings wet conditions across much of the western archipelago, while the southeast monsoon (May to September) introduces drier conditions, particularly over Java, Bali, and the Lesser Sunda Islands (Aldrian & Dwi Susanto, 2003). This contrasting monsoon regime creates a strong west–east gradient in rainfall seasonality across the archipelago.

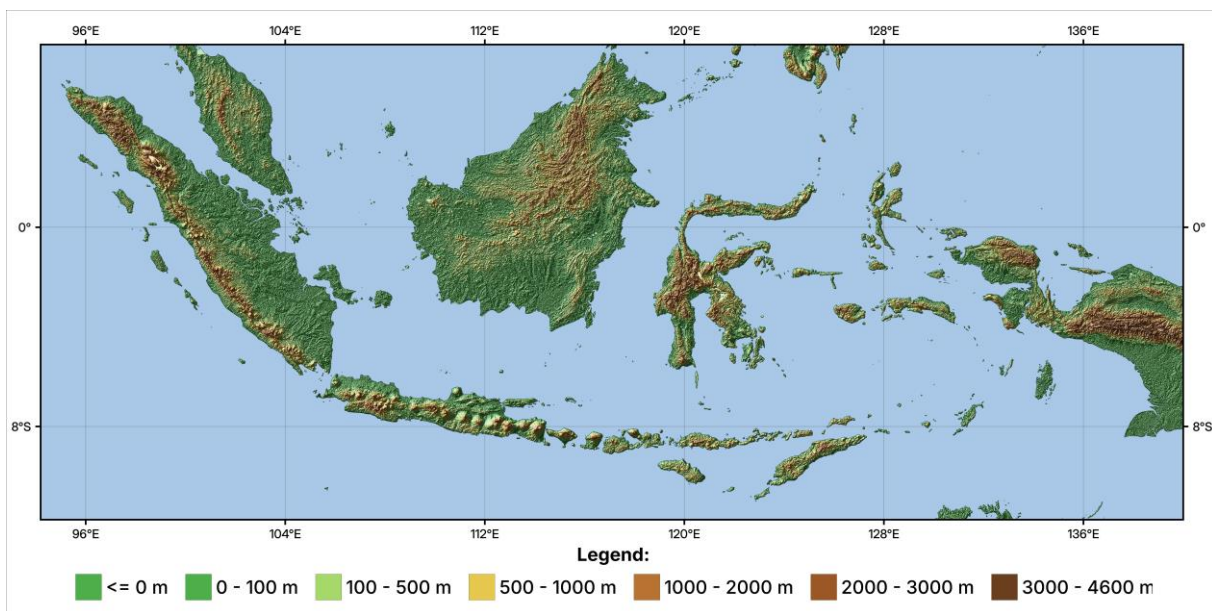


Figure 1: Topographic map of Indonesia and surrounding countries

Topography plays a critical role in shaping Indonesia's climate shown on Figure 1, particularly for precipitation and temperature distributions. The terrain is highly heterogeneous, with several mountain ranges exceeding 4,000 meters above sea level, most notably in Papua where Puncak Jaya reaches approximately 4600 m. The orographic effect of these mountain ranges enhances convective precipitation on windward slopes and creates rain shadows on leeward sides, resulting in substantial spatial variability in rainfall at sub-regional scales. The complex land–sea configuration, extensive coastlines, and the presence of numerous narrow straits further amplify mesoscale convective systems and sea breeze circulations, which contribute significantly to local precipitation patterns. These physiographic characteristics pose inherent challenges for global climate models, whose coarse horizontal resolutions are often insufficient to resolve the fine-scale topographic and coastal features that govern Indonesia's climate variability. Understanding this geographical and climatic context is therefore essential for interpreting the model evaluation results presented in this study.

2.2. Data Collection

The CMIP6 models were gathered from Earth System Grid Federation (<https://esgf.nci.org.au/search>), from 1981 to 2014, listed on Table 1. The selected CMIP6 models were chosen based on several evaluation studies over Southeast Asia (Desmet & Ngo-Duc, 2022; Moise et al., 2024; Nguyen et al., 2024). Daily dataset was used for precipitation (pr), average near-surface air temperature (tg), maximum near-surface air temperature (tx), and minimum near-surface air temperature (tn). These variables were chosen due to the observation of past study over Indonesia across three decades, which revealed changes in extreme temperature and precipitation and as indicator for tracking climate change in the region. All of the datasets then remapped using Climate Data Operators (CDO) (Uwe, 2023), *remapcon* for precipitation and *remapbil* for temperatures, into 0.25° horizontal resolution. The methods were based on how the variable behave: temperature is mostly continuous, so bilinear interpolation is the most suitable and as precipitation is representing amount over an area, so conservative remapping is suitable to preserve the statistics (Diaconescu et al., 2015). The same tool was also used to calculate each metric, using the default function and customized equation. The models then were compared with the observation dataset from SA-OBS (van den Besselaar et al., 2017), which provide gridded observational dataset based on several meteorological stations across the Southeast Asia. The SA-OBS has finer horizontal resolution at 0.25°, compared to CRU-TS at 0.5° and limited temporal resolution at monthly basis (Harris et al., 2020).

Table 1: List of global climate model dataset

No.	Institution	Model	Horizontal Resolution
1	Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia (Dix et al., 2019)	ACCESS-CM2	~1.875° × 1.25° (192 x 144 grid)
2	Centre National de Recherches Meteorologiques (CNRM), France (Voldoire, 2019)	CNRM-CM6-HR	~0.5° (720 x 360 grid)
3	EC-Earth Consortium ((EC-Earth), 2019a)	EC-Earth3	~0.7° (512 × 256 grid)
4	EC-Earth Consortium ((EC-Earth), 2019b)	EC-Earth3-Veg	~0.7° (512 × 256 grid)
5	Met Office Hadley Centre, United Kingdom (Ridley et al., 2019)	HadGEM3-GC31-MM	~0.83° (432 x 324 grid)
6	Max Planck Institute for Meteorology, Germany (Jungclaus et al., 2019)	MPI-ESM1-2-HR	~0.94° (384 x 192 grid)

2.3. Data Analysis

Taylor Skill Score (TSS) is used to measure the overall performance of the model in producing spatial pattern (Taylor, 2001). The TSS is calculated as:

$$TSS = \frac{4(1 + R_m)^2}{\left(\frac{\sigma_m}{\sigma_o} + \frac{\sigma_o}{\sigma_m}\right)^2 (1 + R_o)^2} \quad (1)$$

Where:

R_m - the spatial correlation between simulation and observation,

R_o - the maximum correlation between simulation and observation,

σ_m and σ_o - the spatial standard deviation of simulation and observation respectively.

The ideal score is equal to 1, when there is a perfect spatial match between simulation and observation, and score equal to 0 means inverse model performance. TSS were calculated spatially and then averaged by temporal, to obtain a single score for each model.

Interannual Variability Score (IVS) is used to evaluate the temporal variability of the simulation against the observed (Scherrer, 2011). The IVS is calculated as:

$$IVS = \left(\frac{STD_m}{STD_o} - \frac{STD_o}{STD_m} \right)^2 \quad (2)$$

Where:

STD_m and STD_o – the temporal standard deviation of simulation and observation, respectively.

IVS were calculated temporally for each grid and then spatially averaged to get a single score for each model.

Mean Absolute Error was used in this study to evaluate how far the error of the simulations against observation. MAE was chosen because it is better than Root Mean Squared Error in visualizing the error (Willmott & Matsuura, 2005). The MAE was normalized using observation's standard deviation of each month of the entire timeline, using *ymonstd* in CDO.

$$MAE = \frac{n^{-1} \sum_{i=1}^n |e_i|}{\sigma_o} \quad (3)$$

MAE was calculated by averaging the difference or the error (e_i) between simulation and observation and divided by the standard deviation (σ_o) of the observation. The decision to normalize the MAE was due to the difference in error between precipitation and temperatures. Thus, it can be compared to the performance between variables without such skewness. The ideal score is 0, which indicates no skew between simulations and observation, and extends to ∞ . The MAE was calculated temporally for each grid and then averaged by its spatial.

Kling-Gupta Efficiency (KGE) is mostly used as a goodness indicator of a hydrological model, in this case is precipitation or rainfall data (Kling et al., 2012). Although it is a hydrological indicator, a study also used KGE to evaluate models in producing temperature data (Mohammadlou et al., 2022). The KGE used in this study was the revised version of the original KGE, and calculated as follows:

$$KGE = 1 - \sqrt{(r-1)^2 + \left(\frac{\mu_m}{\mu_o} - 1 \right)^2 + \left(\frac{\sigma_m/\mu_m}{\sigma_o/\mu_o} - 1 \right)^2} \quad (4)$$

Where:

r – the temporal correlation between simulation and observation,

μ_m and μ_o - temporal averages of simulation and observation respectively,

σ_m and σ_o are temporal standard deviation of simulation and observation respectively.

According to Knoben et al. (2019) the value of KGE is not symmetrical, negative value does not necessarily mean bad performance. They argued that $-0.41 < KGE \leq 1$ is considered good performance, if the mean flow $KGE = 1 - \sqrt{2}$ is used. The single score of KGE for each model was calculated the same as the MAE and IVS.

Willmott Index (WI) or Index of Agreement, measures how well the model's predictions match the observed data (Willmott et al., 2012). Several studies used Willmott Index to measure climate model's performance in generating meteorological variables (K. Ahmed et al., 2019; Vicente-Serrano et al., 2025). The Willmott Index used in this study was the refined version, which is shown below:

$$WI = \begin{cases} 1 - \frac{\sum_{i=1}^n |M_i - O_i|}{c \sum_{i=1}^n |O_i - \bar{O}|}, & \text{when } \sum_{i=1}^n |M_i - O_i| \leq c \sum_{i=1}^n |O_i - \bar{O}| \\ \frac{c \sum_{i=1}^n |O_i - \bar{O}|}{\sum_{i=1}^n |M_i - O_i|} - 1, & \text{when } \sum_{i=1}^n |M_i - O_i| > c \sum_{i=1}^n |O_i - \bar{O}| \end{cases} \quad (5)$$

As stated in their study, the Willmott Index was calculated by temporal to better represent the score. M_i and O_i are the i^{th} value of simulation and observation respectively. $\bar{O}$ is the average of the observation, which we calculated using *ymonmean* in CDO as opposed to averaging it by seasonal as their recommendation. By using monthly-year average, we could rectify the sensitivity of precipitation. Since Indonesia is a tropical country with only two seasons (dry and wet season), the observed variability of precipitation would be low as the yearly mean is high relative to the daily data. The suppression is also applied to the temperatures. Temperature variability in Indonesia is relatively stable, compared to the temperate regions. Thus, errors are highly sensitive in Willmott Index, as the observed variability would be low. As for the value of c , we used the recommended value $c = 2$. The same calculation of the single score of each model was used in this metric, same as KGE, MAE, and IVS.

To summarize the scoring system, we utilized Bergen Metrics (Samantaray et al., 2024). The calculations are entirely based on the ideal score for each metric, which measures the distance between ideal score and the actual score of said metrics. The equation is as follows:

$$BM = \sqrt[p]{(1 - TSS)^p + (IVS)^p + (nMAE)^p + (1 - KGE)^p + (1 - WI)^p} \quad (6)$$

The Bergen Metrics were calculated per variable, five metrics for each variable: (1) precipitation, (2) minimum temperature, (3) average temperature, and (4) maximum temperature. The p value used in this study is $p = 2$. The calculation follows this regiment: $(ideal\ score - actual\ score)$. With exception if the ideal score is 0, then the calculation is $(metrics)^p$. As for the final score, the Bergen Metrics for each variable then also calculated using Bergen Metrics.

In addition to the Bergen Metrics scoring, a rank-based weighting (RANK) approach was applied to generate model weights based on performance across all five-evaluation metrics. This method follows the framework of Chen et al. (2011) **Chyba! Nenašiel sa žiaden zdroj odkazov.** and was generalized here for TSS, IVS, nMAE, KGE, and WI. For each metric, all models were ranked from best to worst. Metrics where higher values indicate performance (TSS, KGE, WI) were ranked in descending order, whereas the other (IVS, nMAE) were ranked in ascending order. For each model i , the Total Rank Score (S_i) is then calculated as the simple sum of its rank positions across all M metrics for a given variable:

$$S_i = \sum_{m=1}^M r_{i,k} \quad (7)$$

The Total Rank Score (S_i) is an error proxy, meaning a lower value indicates better performance. To transform this into a performance indicator where a higher value reflects better model skill, the score is calculated as the inverse of S_i , normalized by the sum of all model rank scores. The resulting performance indicator, R_i , for a single variable is defined as:

$$R_i = \frac{\sum_{j=1}^N S_j}{S_i} \quad (8)$$

Where:

N – the total number of models in the ensemble.

A higher value of R_i indicates superior model skill. To obtain a single, robust weight reflecting the model's overall performance across all analyzed climate variables (V), the following steps were followed: (1) the indicator R_i was averaged across all V climate variables to yield the model's mean performance score (R_M), and then (2) the final weight (W_i) was derived by normalizing the mean performance score R_M , ensuring the sum of all model weights equals unity.

$$W_i = \frac{R_i}{\sum_{j=1}^N R_j} \quad (6)$$

The final weight W_i represents the fractional contribution of model i to the final weighted multi-model ensemble projection. A model assigned to a higher W_i value is deemed to have the highest overall skill in simulating historical climate.

3. Results and Discussion

3.1. Precipitation

All evaluated models demonstrate limited skill in simulating daily precipitation over Indonesia, as shown in **Table 2**. Taylor Skill Score values below 0.25 indicate weak spatial correspondence with observed rainfall patterns. Willmott Index values are positive for all models, indicating some agreement in precipitation magnitude, although the agreement remains weak. In contrast, Kling Gupta Efficiency values are negative across all models, indicating poor combined representation of mean state, variability, and temporal correlation.

Table 2: Model's performance on precipitation

Model	TSS	IVS	nMAE	KGE	WI
ACCESS-CM2	0.212	0.716	1.231	-0.564	0.168
CNRM-CM6-1-HR	0.210	0.855	1.185	-0.588	0.213
EC-Earth3	0.231	0.632	1.096	-0.494	0.241
EC-Earth3-Veg	0.230	0.631	1.099	-0.494	0.241
HADGEM-GC31-MM	0.184	1.092	1.191	-0.449	0.225
MPI-ESM1-2-HR	0.235	1.261	0.972	-0.331	0.303

Normalized Mean Absolute Error further highlights persistent absolute biases in precipitation simulations, even after normalization by monthly observational variability. MPI-ESM1-2-HR exhibits the lowest nMAE among the models, indicating relatively smaller average errors, followed by EC Earth3 and EC-Earth3-Veg. Nevertheless, performance in the precipitation remains weak overall.

The poor precipitation skill reflects the inability of global climate models to resolve mesoscale convection, complex coastlines, and strong land sea interactions that dominate rainfall processes over the Indonesian archipelago (Aslam, 2025; Love et al., 2011; Wei et al., 2020). Differences in convective parameterization schemes further contribute to inter model variability (Baba, 2020; HARA et al., 2009; Tost et al., 2006). Similar limitations in CMIP6 precipitation simulations over the Maritime Continent have been widely documented (Agiel et al., 2024; Hariadi et al., 2023; Moise et al., 2024).

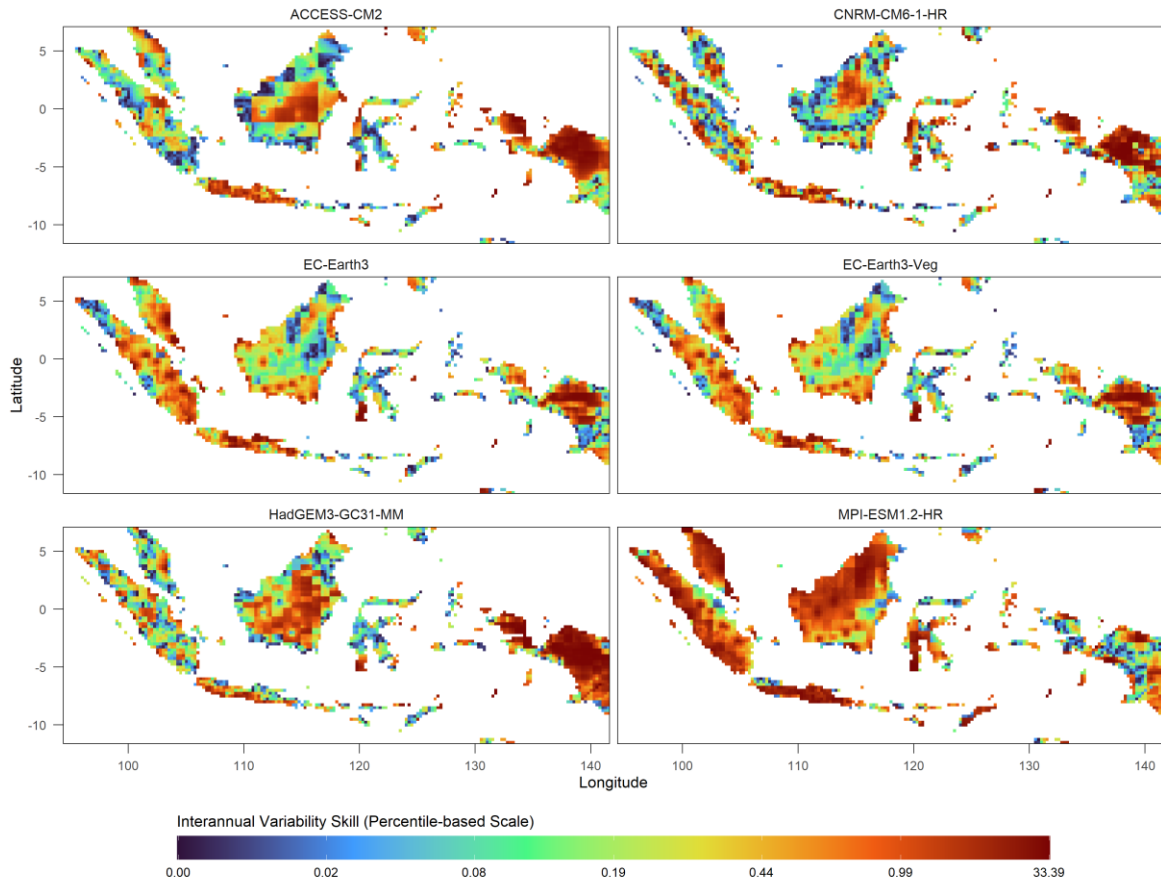


Figure 2: Spatial mapping of IVS performance in precipitation

Based on Figure 2, the spatial distribution of the Interannual Variability Skill (IVS) is consistent with the previously discussed summary scores. All models show pronounced difficulty in reproducing interannual precipitation variability over Papua, where IVS values are generally high, indicating poor temporal skill. In contrast, Sulawesi exhibits the most favorable performance across models, followed by Borneo and Sumatra. Java also shows relatively weak temporal performance, although its IVS is consistently lower than that over Papua, suggesting moderately better representation of interannual variability.

A comparable spatial pattern is evident in Figure 3, which presents the normalized mean absolute error (nMAE). All models display large biases over Papua, reinforcing the indication of poor model performance in this region. However, Java shows the lowest bias among the major islands, which contrasts with its relatively weak IVS performance. This discrepancy suggests that, over Java, models can reproduce mean precipitation magnitudes more accurately, while still struggling to capture year to year variability. Across the remaining islands, model performance is broadly similar, although ACCESS-CM2 and CNRM-CM6-HR exhibit systematically larger biases than the other models.

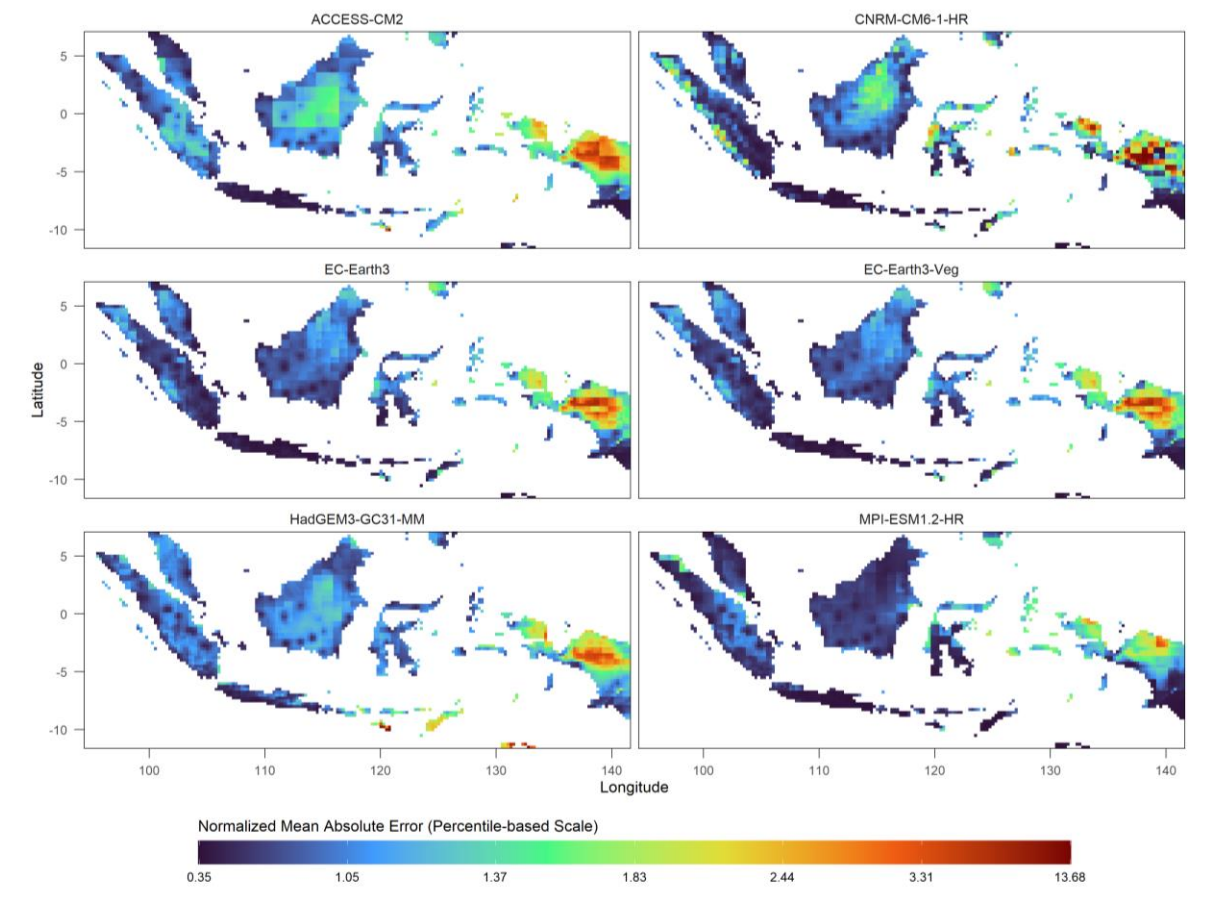


Figure 3: Spatial mapping of nMAE performance in precipitation

3.2. Minimum Temperature

Model performance improves substantially for minimum temperature relative to precipitation on Table 3. Moderate Taylor Skill Score values indicate that spatial patterns of minimum temperature are reasonably well captured. Interannual Variability Score values further indicate that year to year variability is reproduced with moderate skill.

Table 3: Model’s performance at minimum temperature

Model	TSS	IVS	nMAE	KGE	WI
ACCESS-CM2	0.392	0.558	3.748	0.001	-0.392
CNRM-CM6-1-HR	0.681	0.587	2.497	0.005	-0.211
EC-Earth3	0.541	0.499	2.419	0.037	-0.193
EC-Earth3-Veg	0.540	0.519	2.365	0.039	-0.176
HADGEM-GC31-MM	0.560	0.548	2.615	-0.084	-0.224
MPI-ESM1-2-HR	0.431	0.586	2.648	-0.010	-0.197

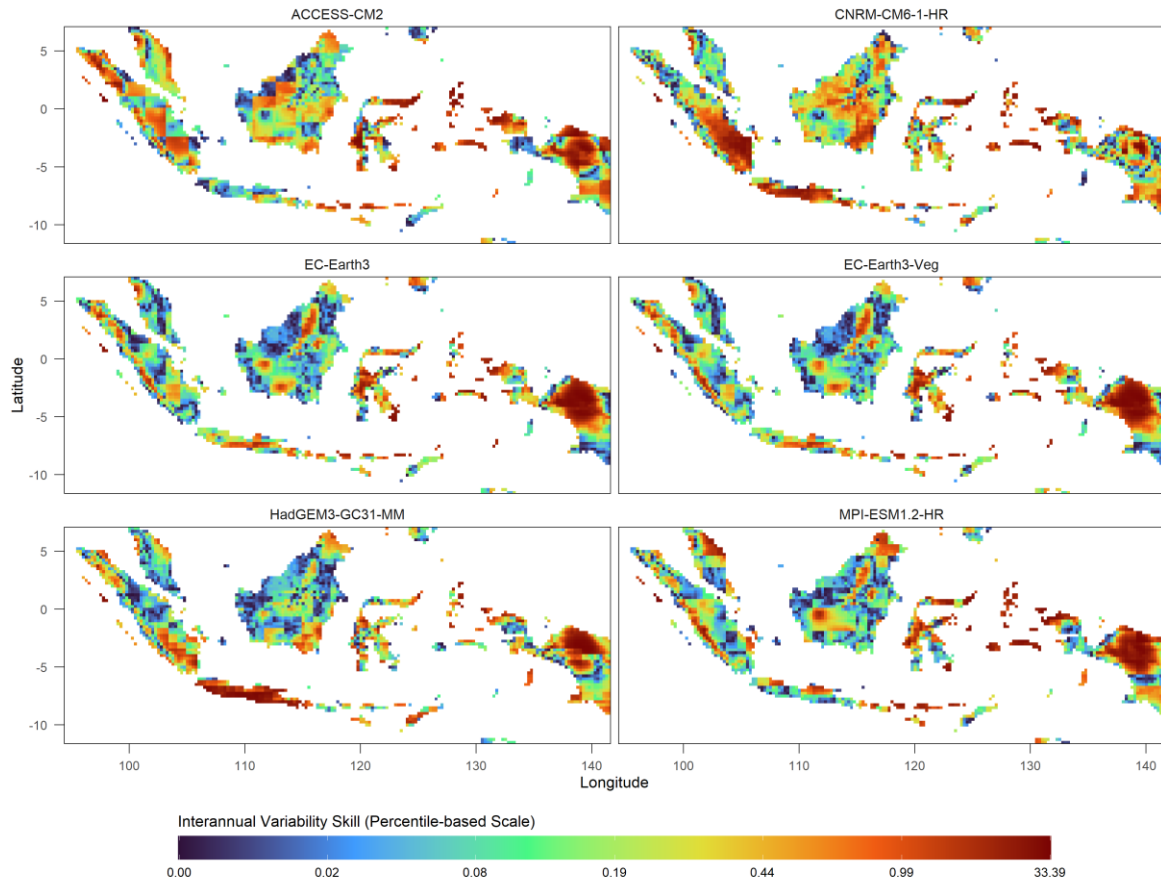


Figure 4: Spatial mapping of IVS performance at minimum temperature

Despite this, Willmott Index values remain negative but close to zero, indicating limited agreement in magnitude. Kling Gupta Efficiency values range from slightly negative to near zero, suggesting small but persistent systematic biases. These results indicate that while models successfully capture the spatial distribution and temporal variability of minimum temperature, they struggle to reproduce the absolute magnitude accurately.

EC-Earth3-Veg and CNRM-CM6-HR achieve the lowest normalized Mean Absolute Error values and the best Bergen Metrics scores for minimum temperature, reflecting balanced spatial and temporal performance. Remaining biases are most evident in coastal and mountainous regions, where local surface processes influence nocturnal cooling (Fan et al., 2022; Nguyen-Duy et al., 2023; Zhang et al., 2024).

Based on Figure 4, the spatial distribution of the Interannual Variability Skill (IVS) for minimum temperature indicates substantially better temporal performance compared to precipitation across most regions. All models exhibit relatively low IVS values over Sumatra, Borneo, Sulawesi, and Java, suggesting that interannual variability in minimum temperature is generally well captured. In contrast to precipitation, poor IVS values are more spatially confined and appear primarily over Papua and parts of eastern Indonesia, where models still struggle to reproduce temporal variability.

Differences among models are more apparent in localized regions rather than at the island scale. EC-Earth3 and EC-Earth3-Veg show comparatively lower IVS over large portions of western Indonesia, whereas ACCESS-CM2 and CNRM-CM6-HR exhibit more spatial heterogeneity, with pockets of degraded skill, particularly over mountainous and coastal regions.

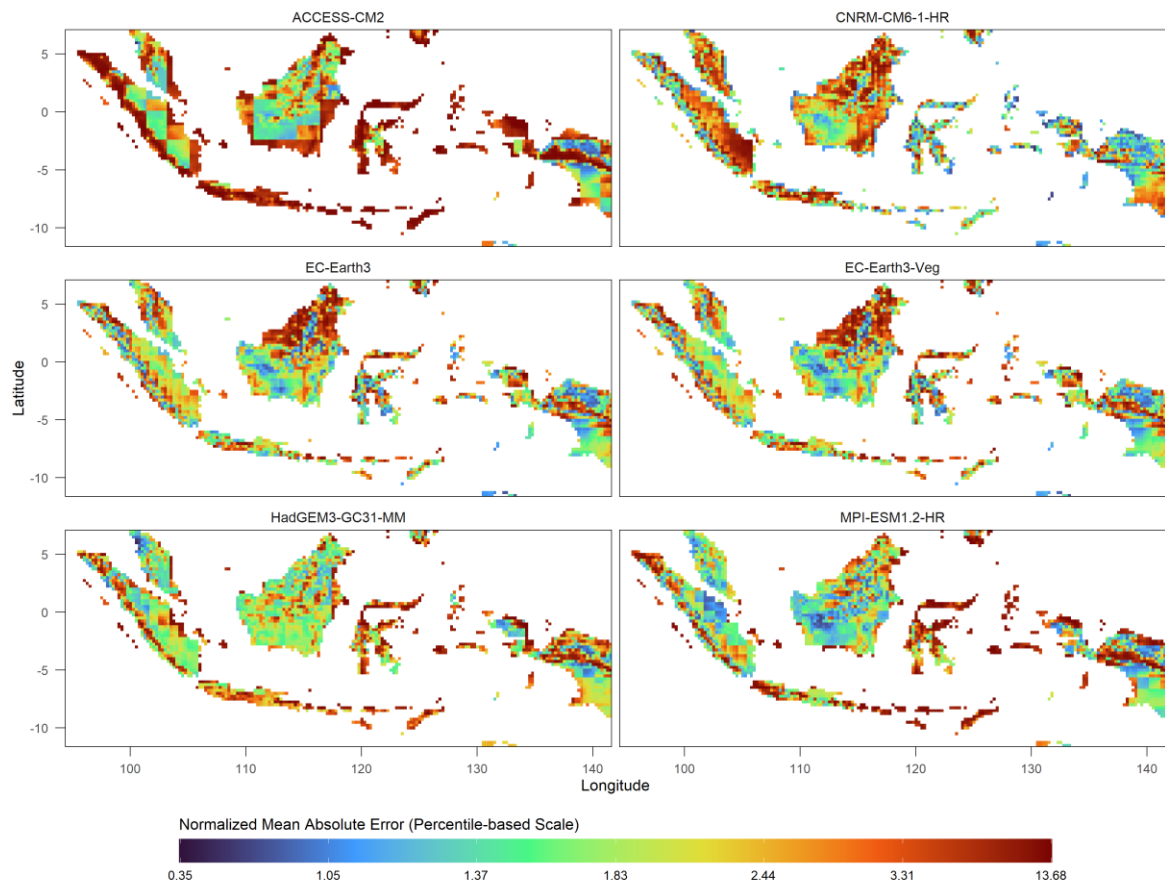


Figure 5: Spatial mapping of nMAE performance at minimum temperature

Figure 5 presents the normalized mean absolute error (nMAE) for minimum temperature and reveals a markedly different spatial behavior from IVS. While temporal variability is generally well represented, systematic biases remain widespread. All models display elevated nMAE over Papua and eastern Indonesia, indicating persistent magnitude errors in minimum temperature. Java and Sumatra show relatively lower biases, although nMAE values remain non-negligible, suggesting that accurate temporal variability does not necessarily imply accurate absolute temperature values.

Overall, the comparison between IVS and nMAE highlights that minimum temperature is more robustly simulated in terms of interannual variability than precipitation, but spatial biases in magnitude persist, particularly in regions with complex topography and land surface heterogeneity.

3.3. Average Temperature

Mean temperature simulations exhibit consistent skill across all evaluated models, shown on Table 4. Taylor Skill Score values exceeding 0.5 indicate reasonable spatial agreement with observations, while Interannual Variability Score values show that year to year temperature fluctuations is well represented.

Table 4: Model’s performance on average temperature

Model	TSS	IVS	nMAE	KGE	WI
ACCESS-CM2	0.376	0.363	2.618	0.127	-0.150
CNRM-CM6-1-HR	0.766	0.343	3.032	0.119	-0.352
EC-Earth3	0.577	0.643	2.635	0.078	-0.259

Model	TSS	IVS	nMAE	KGE	WI
EC-Earth3-Veg	0.574	0.645	2.554	0.083	-0.241
HADGEM-GC31-MM	0.594	0.338	2.118	-0.038	-0.094
MPI-ESM1-2-HR	0.489	0.831	2.642	0.066	-0.266

Kling Gupta Efficiency values are positive for all models, indicating satisfactory representation of mean state and variability. However, Willmott Index values remain negative, highlighting persistent magnitude bias. This combination indicates that models reproduce temporal structure and variability well but exhibit systematic offsets in absolute temperature values.

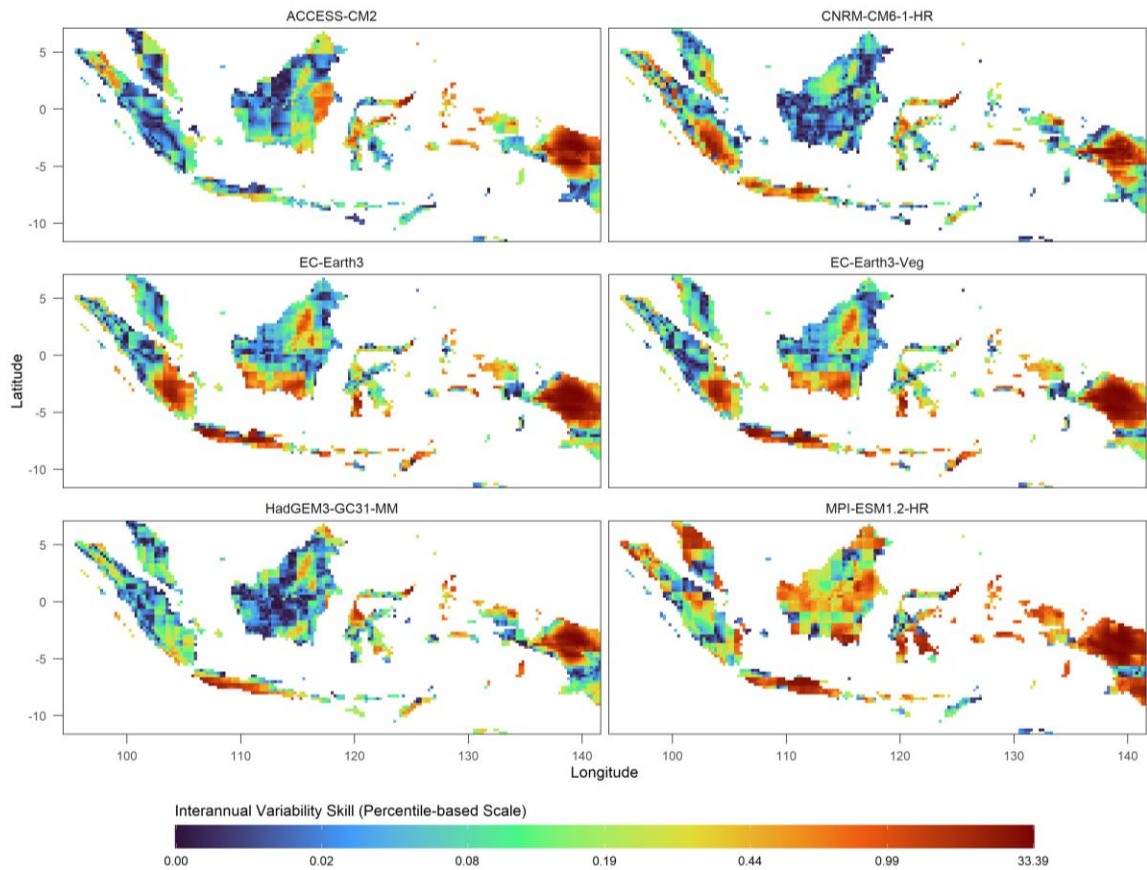


Figure 6: Spatial mapping of IVS performance in average temperature

The relatively small inter model spread reflects the dominance of large-scale radiative forcing in controlling mean temperature, making it easier to simulate than precipitation (Andrews et al., 2010; Yang et al., 2021). Systematic magnitude biases suggest that simple bias correction methods could substantially improve model performance for temperature-based applications.

Figure 6 shows the spatial distribution of the Interannual Variability Skill (IVS) for average temperature. Compared to precipitation and minimum temperature, average temperature exhibits the most spatially coherent and overall strongest temporal performance across all models. Low IVS values dominate over Sumatra, Borneo, Sulawesi, and Java, indicating that interannual variability in mean temperature is consistently well captured. Regions of degraded IVS are largely confined to Papua and parts of eastern Indonesia, where complex terrain and land–atmosphere interactions remain challenging for all models.

Model differences are relatively subtle for average temperature. EC-Earth3 and EC-Earth3-Veg tend to exhibit slightly lower IVS over western Indonesia, while ACCESS-CM2 and CNRM-CM6-HR show more localized variability, particularly over coastal and mountainous regions. Nevertheless, the inter-model spread in IVS is smaller than that observed for precipitation, suggesting greater robustness of average temperature simulations.

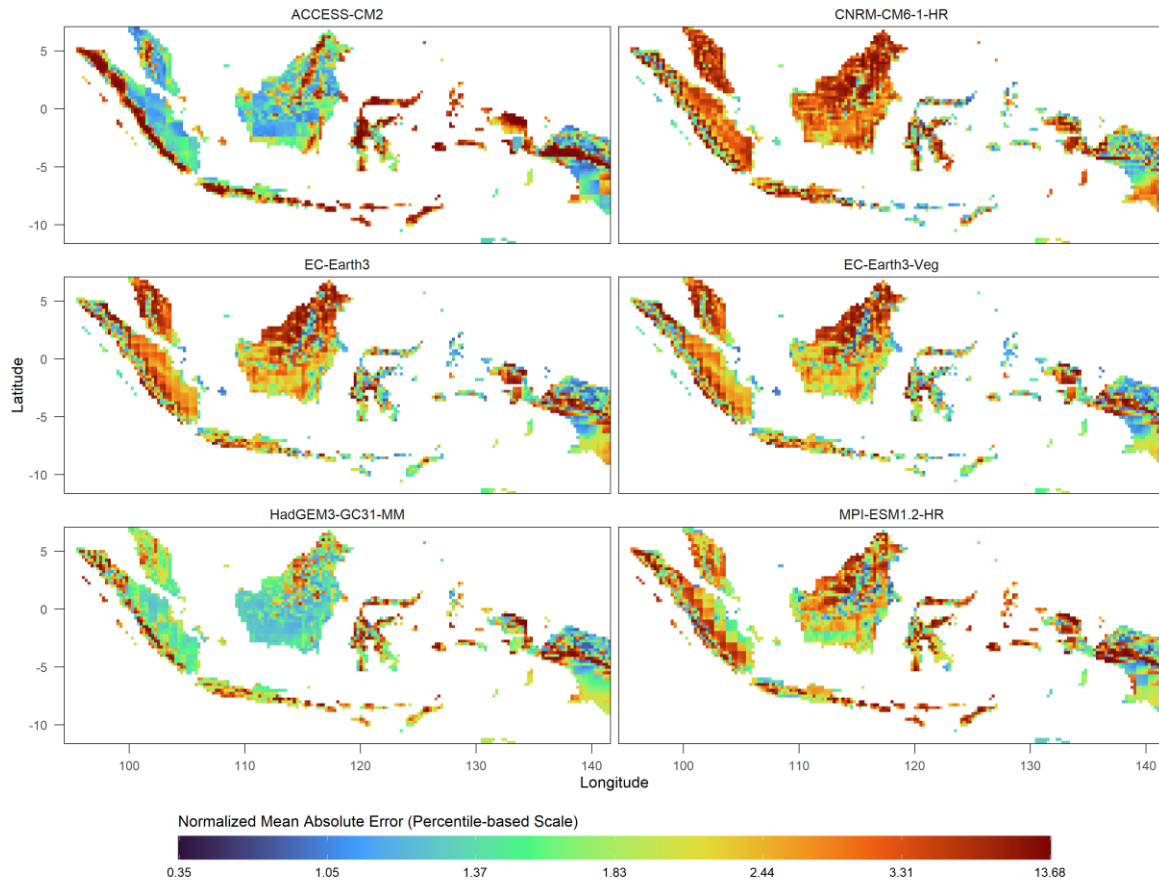


Figure 7: Spatial mapping of nMAE performance in average temperature

The normalized mean absolute error (nMAE) patterns shown in Figure 7 reveal that, despite strong temporal variability skill, substantial magnitude biases persist. All models display elevated nMAE over Papua, while Java and Sumatra consistently show lower biases. Like minimum temperature, this contrast indicates that accurate representation of interannual variability in average temperature does not necessarily translate into accurate absolute temperature values. MPI-ESM1.2-HR exhibits larger biases over several regions, particularly in eastern Indonesia, whereas EC-Earth3-based models generally show more spatially homogeneous errors.

Overall, average temperature emerges as the best simulated variable in terms of temporal variability, yet persistent spatial biases highlight the continued challenge of accurately representing mean temperature fields in regions with complex geography.

3.4. Maximum Temperature

Maximum temperature is simulated with moderate skill across all models (Table 5). Taylor Skill Score values indicate reasonable spatial consistency and normalized Mean Absolute Error values remain relatively low. Kling Gupta Efficiency values suggest moderate efficiency with some model-dependent biases.

Table 5: Model’s performance at maximum temperature

Model	TSS	IVS	nMAE	KGE	WI
ACCESS-CM2	0.329	0.259	1.850	0.119	-0.060
CNRM-CM6-1-HR	0.673	0.395	3.030	0.034	-0.420
EC-Earth3	0.499	0.436	2.466	0.094	-0.273
EC-Earth3-Veg	0.495	0.442	2.388	0.101	-0.252
HADGEM-GC31-MM	0.512	0.367	2.184	-0.064	-0.194
MPI-ESM1-2-HR	0.457	0.672	2.589	0.048	-0.307

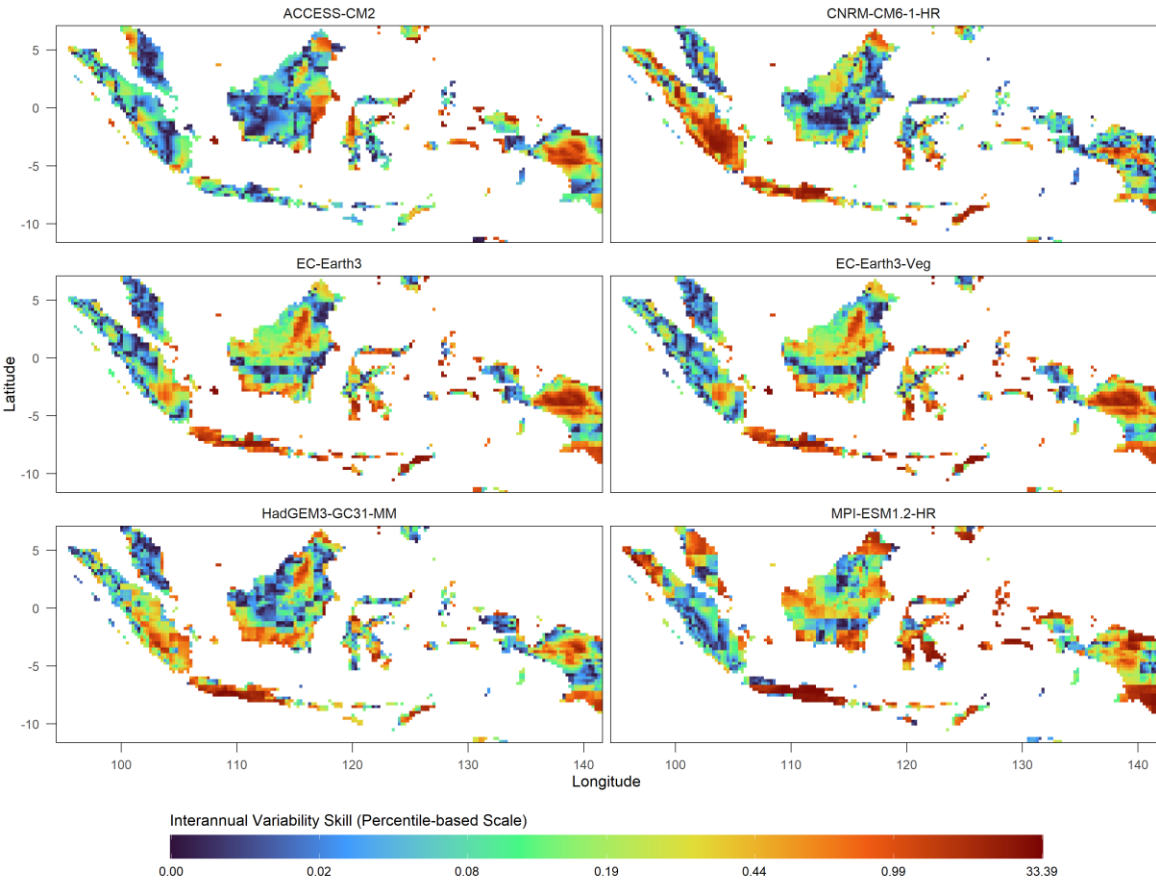


Figure 8: Spatial mapping of IVS performance at maximum temperature

Willmott Index values are negative for all models, although their magnitudes are small, indicating modest but consistent disagreement in magnitude. This pattern suggests that daytime heating processes are captured in relative terms, but systematic biases remain.

Differences in maximum temperature performance are likely related to the representation of surface radiation, cloud cover, and land surface energy balance processes (Lauer et al., 2023; Wang et al., 2023, 2025). Despite these limitations, temperature simulations remain substantially more reliable than precipitation simulations over Indonesia.

Figure 8 presents the spatial distribution of the Interannual Variability Skill (IVS) for maximum temperature. Overall, TX exhibits good temporal performance across much of western Indonesia, with low IVS values over Sumatra, Java, and large parts of Borneo and Sulawesi. This indicates that interannual variability in maximum temperature is generally well captured

by all models. However, compared to average temperature, TX shows slightly higher spatial heterogeneity, particularly over coastal regions and areas influenced by strong land–sea thermal contrasts.

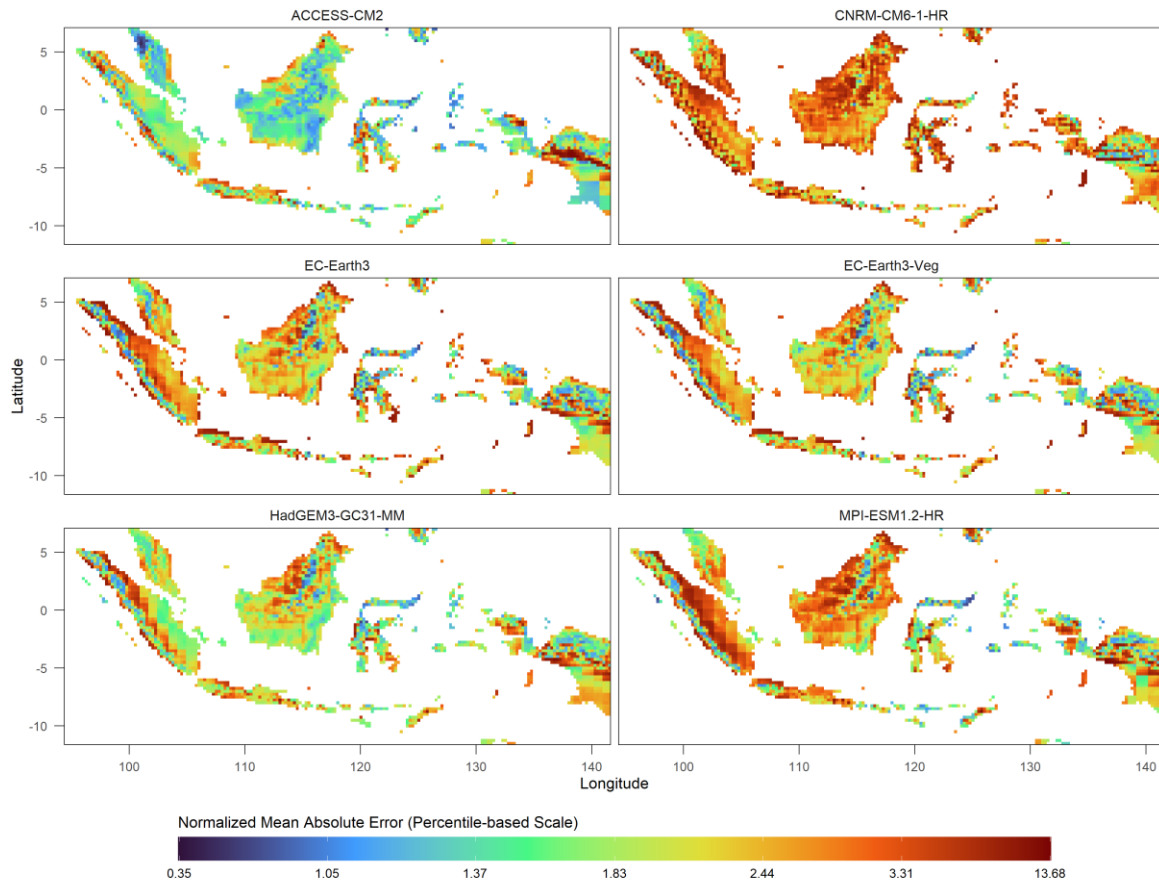


Figure 9: Spatial mapping of nMAE performance at maximum temperature

As with the other variables, Papua stands out as the region with the weakest IVS, where elevated values indicate persistent difficulty in reproducing year to year variability. This behavior is consistent across all models and mirrors the patterns observed for minimum and average temperature, underscoring the challenges associated with complex topography and land surface processes in eastern Indonesia. Among the models, EC-Earth3 and EC-Earth3-Veg display comparatively smoother IVS patterns, while ACCESS-CM2 and CNRM-CM6-HR show more localized regions of degraded skill. The normalized mean absolute error (nMAE) maps in Figure 9 reveal widespread and pronounced magnitude biases for maximum temperature. Compared to minimum and average temperature, TX exhibits the largest nMAE values across most regions, particularly over Sumatra, Borneo, and Papua. This indicates that, although temporal variability is reasonably well represented, models systematically misrepresent the magnitude of daily maximum temperatures. CNRM-CM6-HR and MPI-ESM1.2-HR show especially high biases over large areas, whereas EC-Earth3-based models exhibit relatively lower but still substantial errors.

In summary, maximum temperature demonstrates strong interannual variability skill comparable to other temperature variables, but it is the most biased in terms of absolute magnitude. This contrast highlights the sensitivity of TX to surface energy balance, cloud cover, and land–atmosphere coupling processes, which remain challenging to simulate accurately.

3.5. Overall Performance

Integrated Bergen Metrics' results, in

Table 6, show that EC-Earth3-Veg achieves the lowest overall distance, followed closely by HadGEM3-GC31-MM and EC-Earth3. Models such as ACCESS-CM2 and CNRM-CM6-HR exhibit larger distances, indicating lower overall performance across variables.

Rank based weighting produces broadly consistent conclusions, with EC-Earth3-Veg and EC-Earth3 receiving the highest weights, displayed on Table 7. While Bergen Metrics accounts for metric magnitude through Euclidean distance, rank based weighting emphasizes relative consistency across metrics. The agreement between both approaches strengthens confidence in the identification of robust top performing models.

Table 6: Bergen metrics on overall model's performance

Model	Precipitation	Minimum Temperature	Average Temperature	Maximum Temperature	BM
ACCESS-CM2	2.41	4.20	3.08	2.42	6.23
CNRM-CM6-1-HR	2.43	3.02	3.46	3.52	6.28
EC-Earth3	2.24	2.94	3.16	2.99	5.71
EC-Earth3-Veg	2.24	2.90	3.08	2.92	5.61
HADGEM-GC31-MM	2.44	3.16	2.65	2.78	5.54
MPI-ESM1-2-HR	2.32	3.18	3.23	3.17	6.00

Table 7: Rank-based weighting on overall model's performance

Model	Precipitation	Minimum temperature	Average temperature	Maximum temperature	RANK
ACCESS-CM2	4.38	4.04	7.00	10.50	0.161
CNRM-CM6-1-HR	4.38	6.18	6.18	5.00	0.135
EC-Earth3	9.55	10.50	5.53	5.83	0.195
EC-Earth3-Veg	7.50	11.67	6.18	6.18	0.196
HADGEM-GC31-MM	4.77	5.25	9.55	7.50	0.168
MPI-ESM1-2-HR	10.50	4.57	4.04	4.20	0.145

3.6. Implications and Limitations

The results of this study demonstrate that the choice of evaluation metrics strongly influences the interpretation of climate model performance. The combined use of spatial, temporal, and magnitude-based metrics reveals that CMIP6 models may reproduce spatial patterns and interannual variability well while still exhibiting systematic biases in absolute values. This is particularly evident for temperature variables, where relatively high Taylor Skill Score and Interannual Variability Score values coexist with negative Willmott Index values. Such behavior indicates that models correctly capture spatial gradients and temporal variability but suffer from consistent magnitude offsets, a characteristic commonly reported in global climate model evaluations (Bock et al., 2020; Krinner & Flanner, 2018; Schaller et al., 2011).

In this study, it was revealed that all the models demonstrated low performance of producing precipitation. All models scored low on spatial performance (TSS) and goodness-of-fit (KGE). Since both TSS and IVS used standard deviations, we could estimate the problem. IVS between 0.6 and 1.25 indicates that the ratio of standard deviations (simulation divided by observation) were either 0.585–0.68 or 1.47–1.7, which require the model to had at most spatial correlation of 0.1 to produce score 0.18 and 0.23 of spatial performance. Only MPI-ESM1-2-HR has KGE higher than the -

0.41 threshold. This also translates the performance across TSS, nMAE, and WI that was the highest among all models, although it performed poorly on IVS. The poor performance of KGE might be due to the low spatial correlation of at most 0.1 and the nature of the variability in precipitation. A slight error in producing precipitation could heavily skew the ratio of standard deviation and ratio of average precipitation, which were used in KGE. As for the Willmott Index, the error was suppressed by the variability of the observation.

From all models and variables, the performance was elevation dependent, as shown on the spatial distribution of nMAE. For example, on maximum temperature, both EC-Earth3 and EC-Earth3-Veg performed better over mountainous areas of Sumatra, Java, Borneo, Sulawesi, and Papua. Similar results could also be observed for minimum and average temperatures. As for precipitation, all models overestimated Papua's mountain range depicted in **Figure 1**, with pronounced wet biases along the central cordillera where orographic enhancement of rainfall is strongest and local convective processes are poorly represented at the native CMIP6 resolution (Lalande et al., 2021). This elevation-dependent overestimation implies that model skill assessed at the national or island-mean scale can conceal substantial regional errors, particularly over highlands that are critical for water resources and downstream hydrological impacts.

A further limitation of Bergen Metrics relates to its sensitivity to inter-model spread. The framework measures the Euclidean distance of each model from the ideal score, meaning that when all models perform similarly and their scores cluster closely together, the resulting distances may be numerically small and difficult to distinguish. Under such conditions, small differences in any single metric can disproportionately shift the overall ranking, reducing the reliability of the composite score as a discriminator. Samantaray et al. (2024) acknowledged that Bergen Metrics results should be interpreted alongside individual metric behavior rather than the composite score alone, precisely because the Euclidean distance is sensitive to how individual metrics resolve differences between models. This behavior contrasts with the rank-based weighting approach, which is inherently ordinal and always differentiates between models regardless of the absolute magnitude of performance differences, making it more robust when inter-model differences are small. This sensitivity to the choice of evaluation framework is broadly consistent with findings from other regions: Graff et al. (2025) demonstrated for Arctic climate storylines that no single performance score is universally applicable across all metrics and regions, and that model rankings shift depending on which variables are prioritized. A parallel pattern is evident in neighboring Peninsular Malaysia, where Sa'adi et al. (2026) found that the top-performing GCMs identified through an optimized multi-metric framework differed substantially from those selected using a single metric, with Pour et al. (2022) arriving at an entirely different set of recommended models when relying on KGE alone, underscoring that the choice of evaluation approach is itself a determinant of which models are deemed suitable for regional climate projections. Similarly, a multi-method comparison in southern Italy found that different ranking methods identified different best-performing models depending on the timescale of analysis, reinforcing that composite metrics should not be applied in isolation (Pellicone & Caloiero, 2025). **Chyba! Záložka nie je definovaná..** Taken together, Bergen Metrics is most informative when the evaluated models show substantial spread in performance, and its results in this study should be interpreted with particular caution for temperature variables, where inter-model differences are small.

The application of Bergen Metrics in this study also introduces several limitations that are important to acknowledge when interpreting the integrated performance scores. First, the Euclidean distance formulation implicitly assumes that all standardized metrics contribute equally to the final score, so large systematic precipitation errors over key regions such as the Papua highlands can be partially compensated for by better performance in temperature or in other islands, leading to an overall distance that appears moderate despite serious regional deficiencies. Second, Bergen Metrics aggregates performance over space and variables without explicit regional weighting, which means that complex-terrain hotspots with strong elevation-dependent biases receive the same influence as relatively well-simulated lowland areas, even though they may be more relevant for impact-oriented applications such as flood risk or water resources planning. Third, the method is sensitive to the choice and scaling of the underlying metrics, so different combinations or normalizations of TSS, IVS, nMAE, KGE, and WI could yield different distances, adding a layer of subjectivity that is not always transparent to end users. In contrast, rank-based weighting emphasizes the relative ordering of models rather than the absolute magnitude of metric

differences, which tends to preserve the identification of consistently poor precipitation performance—even when aggregated across variables – and can therefore offer a more robust perspective for model selection in regions with strong spatial heterogeneity such as the Papua mountain range.

4. Conclusion

This study evaluated the performance of six CMIP6 global climate models over Indonesia using a multi-metric framework synthesized through Bergen Metrics and complemented by a rank-based weighting approach. Overall, the results show that model skill is strongly variable across both variables and regions, with temperature fields being simulated more reliably than precipitation. Minimum, mean, and maximum temperatures exhibit reasonable spatial and temporal performance, although systematic magnitude biases persist, particularly over coastal and mountainous regions. In contrast, all models demonstrate limited skill in simulating daily precipitation, with low spatial correspondence, weak temporal variability, and large absolute errors. These deficiencies are most pronounced over the Papua mountain range, where complex topography and mesoscale convective processes lead to strong elevation-dependent overestimation that is not resolved at the native CMIP6 resolution.

The results reveal a consistent contrast in model suitability across climate variables. All evaluated models demonstrate limited skill in reproducing daily precipitation over Indonesia, with weak spatial correspondence and poor goodness-of-fit scores observed across the archipelago. This limitation is most pronounced over high-elevation regions such as Papua, where the complex terrain amplifies the inability of coarse-resolution global models to capture mesoscale convective processes and orographic rainfall. In contrast, temperature variables are simulated with considerably greater skill across all models, with average temperature exhibiting the strongest interannual variability performance and the most spatially coherent patterns. Minimum and maximum temperatures fall between these two extremes, showing reasonable spatial skill but persistent magnitude biases, particularly over elevated terrain. EC-Earth3-Veg emerges as the best overall performing model across all variables, followed by HadGEM3-GC31-MM and EC-Earth3, and is therefore most suitable for applications requiring a single model selection over Indonesia. The comparison between Bergen Metrics and rank-based weighting highlights that the choice of evaluation framework can substantially influence the interpretation of overall model performance. Bergen Metrics provides a compact and physically interpretable summary of multi-metric behavior, but its Euclidean aggregation can partially mask regional hotspots of poor performance when models compensate with better skill in other variables or regions. Rank-based weighting, on the other hand, is more sensitive to the relative ordering of models and tends to better preserve consistent weaknesses in precipitation, especially over complex terrain such as Papua. Taken together, the results emphasize that integrated model evaluation over Indonesia should not rely on a single metric or a single aggregation method. Instead, robust model selection for climate impact studies in the region should combine multi-metric synthesis, rank-based diagnostics, and region-specific analyses, with particular attention to high-risk areas where biases are elevation dependent and precipitation remains highly uncertain.

Acknowledgements

This work was supported by Sepuluh Nopember Institute of Technology (ITS) under the university research grant. The authors would like to thank the Department of Environmental Engineering ITS for providing the resources and facilities necessary to conduct this research.

Author Contributions

M.A.M. conceived and designed the study. **A.M.D.** conducted the experiments, processed the CMIP6 and observational datasets, and performed the formal analysis. **R.I.B.S.** contributed to the development of the evaluation methodology and statistical analysis. **E.N.** contributed to data curation, interpretation of the results, and drafting of the manuscript. **H.P.P.** assisted with visualization and preparation of figures and tables. **A.M.** contributed

to resources, supervision, and critical revision of the manuscript. All authors critically reviewed and approved the final version of the manuscript and agreed to be accountable for all aspects of the work.

Disclosure of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability Statement

All of the CMIP6 dataset were obtained from the ESGF (<https://esgf.nci.org.au/search>) and the SA-OBS dataset were obtained from the SACAD website (<https://sacad.bmkg.go.id/>).

References

- Ahmed, A., Al-Said, T., Madhusoodhanan, R., Naqvi, S. W. A., Sarkar, A., Fernandes, L., Thuslim, F., Al-Zakri, W., & Al-Yamani, F. (2022). Environmental impact of a series of flash flood events on a hypersaline subtropical system in the Northwestern Arabian Gulf. *Marine Pollution Bulletin*, 175, 113394. <https://doi.org/10.1016/J.MARPOLBUL.2022.113394>.
- Ahmed, K., Shahid, S., Wang, X., Nawaz, N., & Khan, N. (2019). Evaluation of Gridded Precipitation Datasets over Arid Regions of Pakistan. *Water*, 11(2), 210. <https://doi.org/10.3390/w11020210>.
- Akinsanola, A. A., Ongoma, V., & Kooperman, G. J. (2021). Evaluation of CMIP6 models in simulating the statistics of extreme precipitation over Eastern Africa. *Atmospheric Research*, 254, 105509. <https://doi.org/10.1016/j.atmosres.2021.105509>.
- Aldrian, E., & Dwi Susanto, R. (2003). Identification of three dominant rainfall regions within Indonesia and their relationship to sea surface temperature. *International Journal of Climatology*, 23(12), 1435–1452. <https://doi.org/10.1002/joc.950>.
- Andrews, T., Forster, P. M., Boucher, O., Bellouin, N., & Jones, A. (2010). Precipitation, radiative forcing and global temperature change. *Geophysical Research Letters*, 37(14). <https://doi.org/10.1029/2010GL043991>.
- Aslam, A. A. (2025). Rainfall over the Maritime Continent: key processes, scale interactions and model representation. *Weather*, 80(6), 176–185. <https://doi.org/10.1002/wea.7731>.
- Baba, Y. (2020). Diurnal cycle of precipitation over the Maritime Continent simulated by a spectral cumulus parameterization. *Dynamics of Atmospheres and Oceans*, 91, 101160. <https://doi.org/10.1016/j.dynatmoce.2020.101160>.
- Beaugrand, G., Reid, P. C., & Conversi, A. (2025). Acceleration of changes in global surface temperature after the Earth system has crossed a tipping point. *Climate Research*, 95, 1–17. <https://doi.org/10.3354/cr01756>.
- Bock, L., Lauer, A., Schlund, M., Barreiro, M., Bellouin, N., Jones, C., Meehl, G. A., Predoi, V., Roberts, M. J., & Eyring, V. (2020). Quantifying Progress Across Different CMIP Phases With the ESMValTool. *Journal of Geophysical Research: Atmospheres*, 125(21). <https://doi.org/10.1029/2019JD032321>.
- Chen, W., Jiang, Z., & Li, L. (2011). Probabilistic projections of climate change over China under the SRES A1B scenario using 28 AOGCMs. *Journal of Climate*, 24(17), 4741–4756. <https://doi.org/10.1175/2011JCLI4102.1>.
- Desmet, Q., & Ngo-Duc, T. (2022). A novel method for ranking CMIP6 global climate models over the southeast Asian region. *International Journal of Climatology*, 42(1), 97–117. <https://doi.org/10.1002/joc.7234>.
- Diaconescu, E. P., Gachon, P., & Laprise, R. (2015). On the Remapping Procedure of Daily Precipitation Statistics and Indices Used in Regional Climate Model Evaluation. *Journal of Hydrometeorology*, 16(6), 2301–2310. <https://doi.org/10.1175/JHM-D-15-0025.1>.
- Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, 9(5), 1937–1958. <https://doi.org/10.5194/gmd-9-1937-2016>.
- 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., ... Williamson, M. S. (2019). Taking climate model evaluation to the next level. *Nature Climate Change*, 9(2), 102–110. <https://doi.org/10.1038/s41558-018-0355-y>.

- Fan, X., Duan, Q., Shen, C., Wu, Y., & Xing, C. (2022). Evaluation of historical CMIP6 model simulations and future projections of temperature over the Pan-Third Pole region. *Environmental Science and Pollution Research*, 29(18), 26214–26229. <https://doi.org/10.1007/s11356-021-17474-7>.
- Fofana, M., Adoukpe, J., Larbi, I., Hounkpe, J., Djan'na Koubodana, H., Toure, A., Bokar, H., Dotse, S. Q., & Limantol, A. M. (2022). Urban flash flood and extreme rainfall events trend analysis in Bamako, Mali. *Environmental Challenges*, 6, 100449. <https://doi.org/10.1016/J.ENVC.2022.100449>.
- Forster, P. M., Smith, C., Walsh, T., Lamb, W. F., Lamboll, R., Cassou, C., Hauser, M., Hausfather, Z., Lee, J.-Y., Palmer, M. D., von Schuckmann, K., Slangen, A. B. A., Szopa, S., Trewin, B., Yun, J., Gillett, N. P., Jenkins, S., Matthews, H. D., Raghavan, K., ... Zhai, P. (2025). Indicators of Global Climate Change 2024: annual update of key indicators of the state of the climate system and human influence. *Earth System Science Data*, 17(6), 2641–2680. <https://doi.org/10.5194/essd-17-2641-2025>.
- Gebrechorkos, S., Leyland, J., Slater, L., Wortmann, M., Ashworth, P. J., Bennett, G. L., Boothroyd, R., Cloke, H., Delorme, P., Griffith, H., Hardy, R., Hawker, L., McLelland, S., Neal, J., Nicholas, A., Tatem, A. J., Vahidi, E., Parsons, D. R., & Darby, S. E. (2023). A high-resolution daily global dataset of statistically downscaled CMIP6 models for climate impact analyses. *Scientific Data*, 10(1), 611. <https://doi.org/10.1038/s41597-023-02528-x>.
- Graff, L. S., Landgren, O. A., Parding, K. M., Levine, X., Williams, R. S., & Mooney, P. A. (2025). Selecting CMIP6 Models for Future Arctic Storylines Using a Novel Performance Score. *Tellus*, 78(1), 1–21. <https://doi.org/10.16993/tellusa.4111>
- Gu, G., & Adler, R. F. (2018). Precipitation Intensity Changes in the Tropics from Observations and Models. *Journal of Climate*, 31(12), 4775–4790. <https://doi.org/10.1175/JCLI-D-17-0550.1>.
- Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, 377(1–2), 80–91. <https://doi.org/10.1016/j.jhydrol.2009.08.003>.
- Gutowski Jr., W. J., Giorgi, F., Timbal, B., Frigon, A., Jacob, D., Kang, H.-S., Raghavan, K., Lee, B., Lennard, C., Nikulin, G., O'Rourke, E., Rixen, M., Solman, S., Stephenson, T., & Tangang, F. (2016). WCRP COordinated Regional Downscaling EXperiment (CORDEX): a diagnostic MIP for CMIP6. *Geoscientific Model Development*, 9(11), 4087–4095. <https://doi.org/10.5194/gmd-9-4087-2016>.
- Ha, T. V., Huth, J., Bachofer, F., & Kuenzer, C. (2022). A Review of Earth Observation-Based Drought Studies in Southeast Asia. *Remote Sensing*, 14(15), 3763. <https://doi.org/10.3390/rs14153763>.
- Hansen, J. E., Kharecha, P., Sato, M., Tselioudis, G., Kelly, J., Bauer, S. E., Ruedy, R., Jeong, E., Jin, Q., Rignot, E., Velicogna, I., Schoeberl, M. R., von Schuckmann, K., Amponsem, J., Cao, J., Keskinen, A., Li, J., & Pokela, A. (2025). Global Warming Has Accelerated: Are the United Nations and the Public Well-Informed? *Environment: Science and Policy for Sustainable Development*, 67(1), 6–44. <https://doi.org/10.1080/00139157.2025.2434494>.
- HARA, M., YOSHIKANE, T., TAKAHASHI, H. G., KIMURA, F., NODA, A., & TOKIOKA, T. (2009). Assessment of the Diurnal Cycle of Precipitation over the Maritime Continent Simulated by a 20 km Mesh GCM Using TRMM PR Data. *Journal of the Meteorological Society of Japan. Ser. II*, 87A, 413–424. <https://doi.org/10.2151/jmsj.87A.413>.
- Hariadi, M. H., van der Schrier, G., Steeneveld, G., Ratri, D. N., Sopaheluwakan, A., Tank, A. K., Aldrian, E., Gunawan, D., Moine, M., Bellucci, A., Senan, R., Tourigny, E., Putrasahan, D. A., & Linarka, U. A. (2023). Evaluation of extreme precipitation over Southeast Asia in the Coupled Model Intercomparison Project Phase 5 regional climate model results and HighResMIP global climate models. *International Journal of Climatology*, 43(3), 1639–1659. <https://doi.org/10.1002/joc.7938>.
- Harris, I., Osborn, T. J., Jones, P., & Lister, D. (2020). Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. *Scientific Data*, 7(1), 109. <https://doi.org/10.1038/s41597-020-0453-3>.
- Khadka, D., Babel, M. S., Abatan, A. A., & Collins, M. (2022). An evaluation of CMIP5 and CMIP6 climate models in simulating summer rainfall in the Southeast Asian monsoon domain. *International Journal of Climatology*, 42(2), 1181–1202. <https://doi.org/10.1002/joc.7296>.
- Kling, H., Fuchs, M., & Paulin, M. (2012). Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *Journal of Hydrology*, 424–425, 264–277. <https://doi.org/10.1016/j.jhydrol.2012.01.011>.
- Knoben, W. J. M., Freer, J. E., & Woods, R. A. (2019). Technical note: Inherent benchmark or not? Comparing Nash-Sutcliffe and Kling-Gupta efficiency scores. *Hydrology and Earth System Sciences*, 23(10), 4323–4331. <https://doi.org/10.5194/hess-23-4323-2019>.
- Krinner, G., & Flanner, M. G. (2018). Striking stationarity of large-scale climate model bias patterns under strong climate change. *Proceedings of the National Academy of Sciences*, 115(38), 9462–9466. <https://doi.org/10.1073/pnas.1807912115>.

- Lalande, M., Ménégoz, M., Krinner, G., Naegeli, K., & Wunderle, S. (2021). Climate change in the High Mountain Asia in CMIP6. *Earth System Dynamics*, 12(4), 1061–1098. <https://doi.org/10.5194/esd-12-1061-2021>.
- Lauer, A., Bock, L., Hassler, B., Schröder, M., & Stengel, M. (2023). Cloud Climatologies from Global Climate Models—A Comparison of CMIP5 and CMIP6 Models with Satellite Data. *Journal of Climate*, 36(2), 281–311. <https://doi.org/10.1175/JCLI-D-22-0181.1>.
- Legates, D. R., & McCabe, G. J. (1999). Evaluating the use of "goodness-of-fit" Measures in hydrologic and hydroclimatic model validation. *Water Resources Research*, 35(1), 233–241. <https://doi.org/10.1029/1998WR900018>.
- Love, B. S., Matthews, A. J., & Lister, G. M. S. (2011). The diurnal cycle of precipitation over the Maritime Continent in a high-resolution atmospheric model. *Quarterly Journal of the Royal Meteorological Society*, 137(657), 934–947. <https://doi.org/10.1002/qj.809>.
- Mohammadlou, M., Bahremand, A., Princz, D., Kinar, N., Haghnegahdar, A., & Razavi, S. (2022). Objective evaluation of the Global Environmental Multiscale Model (GEM) with precipitation and temperature for Iran. *Natural Resource Modeling*, 35(3). <https://doi.org/10.1111/nrm.12343>.
- Nazeri Tahroudi, M. (2025). Comprehensive global assessment of precipitation trend and pattern variability considering their distribution dynamics. *Scientific Reports*, 15(1), 22458. <https://doi.org/10.1038/s41598-025-06050-5>.
- Ngo-Duc, T., Nguyen-Duy, T., Desmet, Q., Trinh-Tuan, L., Ramu, L., Cruz, F., Dado, J. M., Chung, J. X., Phan-Van, T., Pham-Thanh, H., Truong-Ba, K., Tangang, F. T., Juneng, L., Santisirisomboon, J., Srisawadwong, R., Permana, D., Linarka, U. A., & Gunawan, D. (2024). Performance ranking of multiple CORDEX-SEA sensitivity experiments: towards an optimum choice of physical schemes for RegCM over Southeast Asia. *Climate Dynamics*. <https://doi.org/10.1007/s00382-024-07353-5>.
- Nguyen, P. L., Alexander, L. V., Thatcher, M. J., Truong, S. C. H., Isphording, R. N., & McGregor, J. L. (2024). Selecting CMIP6 global climate models (GCMs) for Coordinated Regional Climate Downscaling Experiment (CORDEX) dynamical downscaling over Southeast Asia using a standardised benchmarking framework. *Geoscientific Model Development*, 17(19), 7285–7315. <https://doi.org/10.5194/gmd-17-7285-2024>.
- Nguyen-Duy, T., Ngo-Duc, T., & Desmet, Q. (2023). Performance evaluation and ranking of CMIP6 global climate models over Vietnam. *Journal of Water and Climate Change*, 14(6), 1831–1846. <https://doi.org/10.2166/wcc.2023.454>.
- Nofrizal, Fauzan, Hakam, A., Istijono, B., & Aprisal. (2025). Analysis of the Effect of Rainfall Intensity and Slope Steepness on Landslide Disaster. *Civil and Environmental Engineering*, 21(2), 959–973. <https://doi.org/10.2478/cee-2025-0073>.
- Ouyang, H., Baklanov, A., Tang, X., Wang, P., & Zhang, R. (2025). Urgency and importance of local-scale modeling tools to support climate adaptation and sustainable development. *Frontiers of Environmental Science & Engineering*, 19(12), 171. <https://doi.org/10.1007/s11783-025-2091-7>.
- Pellicone, G., & Caloiero, T. (2025). Assessment and ranking of CMIP6-global climate models over the Calabria region (southern Italy). *Journal of Hydrology*, 661, 133655. <https://doi.org/10.1016/j.jhydrol.2025.133655>.
- Pour, S. H., Shahid, S., & Mainuddin, M. (2022). Relative performance of CMIP5 and CMIP6 models in simulating rainfall in Peninsular Malaysia. *Theoretical and Applied Climatology*, 149(1–2), 709–725. <https://doi.org/10.1007/s00704-022-04076-7>.
- Riahi, K., van Vuuren, D. P., Kriegler, E., Edmonds, J., O'Neill, B. C., Fujimori, S., Bauer, N., Calvin, K., Dellink, R., Fricko, O., Lutz, W., Popp, A., Cuaresma, J. C., KC, S., Leimbach, M., Jiang, L., Kram, T., Rao, S., Emmerling, J., ... Tavoni, M. (2017). The Shared Socioeconomic Pathways and their energy, land use, and greenhouse gas emissions implications: An overview. *Global Environmental Change*, 42, 153–168. <https://doi.org/10.1016/j.gloenvcha.2016.05.009>.
- Sa'adi, Z., Awhari, D. P., Kemarau, R. A., Zainon Noor, Z., Taining, Z., Ahmad, M. F., Salem, A., Muniandy, G., Mohammed Dafalla, M. F., Nursiwan, W. A., Maha, M. R. A., Astuti, A. D., & Al-Husban, S. A. (2026). Optimizing category-based statistical metrics for selecting global climate models in rainfall projections for Peninsular Malaysia. *Theoretical and Applied Climatology*, 157(2), 67. <https://doi.org/10.1007/s00704-025-05952-8>.
- Samantaray, A. K., Mooney, P. A., & Vivacqua, C. A. (2024). Bergen metrics: Composite error metrics for assessing performance of climate models using EURO-CORDEX simulations. *Geoscientific Model Development*, 17(8), 3321–3339. <https://doi.org/10.5194/gmd-17-3321-2024>.
- Schaller, N., Mahlstein, I., Cermak, J., & Knutti, R. (2011). Analyzing precipitation projections: A comparison of different approaches to climate model evaluation. *Journal of Geophysical Research*, 116(D10), D10118. <https://doi.org/10.1029/2010JD014963>.
- Scherrer, S. C. (2011). Present-day interannual variability of surface climate in CMIP3 models and its relation to future warming. *International Journal of Climatology*, 31(10), 1518–1529. <https://doi.org/10.1002/joc.2170>.
- Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research Atmospheres*, 106(D7), 7183–7192. <https://doi.org/10.1029/2000JD900719>.

- Tebaldi, C., Debeire, K., Eyring, V., Fischer, E., Fyfe, J., Friedlingstein, P., Knutti, R., Lowe, J., O'Neill, B., Sanderson, B., van Vuuren, D., Riahi, K., Meinshausen, M., Nicholls, Z., Tokarska, K. B., Hurtt, G., Kriegler, E., Lamarque, J.-F., Meehl, G.,... & Ziehn, T. (2021). Climate model projections from the Scenario Model Intercomparison Project (ScenarioMIP) of CMIP6. *Earth System Dynamics*, 12(1), 253–293. <https://doi.org/10.5194/esd-12-253-2021>.
- Tost, H., Jöckel, P., & Lelieveld, J. (2006). Influence of different convection parameterisations in a GCM. *Atmospheric Chemistry and Physics*, 6(12), 5475–5493. <https://doi.org/10.5194/acp-6-5475-2006>.
- Tripathy, K. P., Mukherjee, S., Mishra, A. K., Mann, M. E., & Williams, A. P. (2023). Climate change will accelerate the high-end risk of compound drought and heatwave events. *Proceedings of the National Academy of Sciences*, 120(28). <https://doi.org/10.1073/pnas.2219825120>.
- van den Besselaar, E. J. M., van der Schrier, G., Cornes, R. C., Iqbal, A. S., & Klein Tank, A. M. G. (2017). SA-OBS: A Daily Gridded Surface Temperature and Precipitation Dataset for Southeast Asia. *Journal of Climate*, 30(14), 5151–5165. <https://doi.org/10.1175/JCLI-D-16-0575.1>.
- Vicente-Serrano, S. M., Trambly, Y., Reig, F., González-Hidalgo, J. C., Beguería, S., Brunetti, M., Kalin, K. C., Patalen, L., Kržiž, A., Lionello, P., Lima, M. M., Trigo, R. M., El-Kenawy, A. M., Eddenjal, A., Türkes, M., Koutroulis, A., Manara, V., Maugeri, M., Badi, W., ... Potopová, V. (2025). High temporal variability not trend dominates Mediterranean precipitation. *Nature*, 639(8055), 658–666. <https://doi.org/10.1038/s41586-024-08576-6>.
- Voon, K. L., Tan, K. W., & Chin, K. S. (2022). Assessment of the Flood and Drought Occurrence Using Statistically Downscaled Local Climate Models: A Case Study in Langat River Basin, Malaysia. *Civil and Environmental Engineering*, 18(1), 221–233. <https://doi.org/10.2478/cee-2022-0021>.
- Wagener, T., Reinecke, R., & Pianosi, F. (2022). On the evaluation of climate change impact models. *WIREs Climate Change*, 13(3). <https://doi.org/10.1002/wcc.772>.
- Wang, L., Liu, Z., & Jiang, D. (2025). Overestimation of Interannual Temperature Variability Over China by CMIP6 Simulations: The Role of Snow Cover and Cloud Representation. *Journal of Geophysical Research: Atmospheres*, 130(12). <https://doi.org/10.1029/2025JD043551>.
- Wang, L., Wang, L., Li, Y., & Wang, J. (2023). A century-long analysis of global warming and earth temperature using a random walk with drift approach. *Decision Analytics Journal*, 7, 100237. <https://doi.org/10.1016/J.DAJOUR.2023.100237>.
- Wei, Y., Pu, Z., & Zhang, C. (2020). Diurnal Cycle of Precipitation Over the Maritime Continent Under Modulation of MJO: Perspectives From Cloud-Permitting Scale Simulations. *Journal of Geophysical Research: Atmospheres*, 125(13). <https://doi.org/10.1029/2020JD032529>.
- Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30, 79–82. www.int-res.com.
- Willmott, C. J., Robeson, S. M., & Matsuura, K. (2012). A refined index of model performance. *International Journal of Climatology*, 32(13), 2088–2094. <https://doi.org/10.1002/joc.2419>.
- Yang, X., Zhou, B., Xu, Y., & Han, Z. (2021). CMIP6 Evaluation and Projection of Temperature and Precipitation over China. *Advances in Atmospheric Sciences*, 38(5), 817–830. <https://doi.org/10.1007/s00376-021-0351-4>.
- Zhang, S., Hu, Q., Meng, X., Lü, Y., & Yang, X. (2024). Spatiotemporal Evaluation and Future Projection of Diurnal Temperature Range over the Tibetan Plateau in CMIP6 Models. *Advances in Atmospheric Sciences*, 41(11), 2245–2258. <https://doi.org/10.1007/s00376-024-3346-0>.
- Agíel, H. M., Pratama, A., Mareta, L., & Birastrí, W. (2024). Uncertainty of Coupled Model Intercomparison Projects 6 (CMIP6) in Indonesia's maritime continental region for the historical period. *IOP Conference Series: Earth and Environmental Science*, 1314(1), 012020. <https://doi.org/10.1088/1755-1315/1314/1/012020>.
- Dix, M., Bi, D., Dobrohotoff, P., Fiedler, R., Harman, I., Law, R., Mackallah, C., Marsland, S., O'Farrell, S., Rashid, H., Srbinovsky, J., Sullivan, A., Trenham, C., Vohralik, P., Watterson, I., Williams, G., Woodhouse, M., Bodman, R., Dias, F. B.,... & Yang, R. (2019). CSIRO-ARCCSS ACCESS-CM2 model output prepared for CMIP6 CMIP historical. *Earth System Grid Federation*. <https://doi.org/10.22033/ESGF/CMIP6.4271>.
- (EC-Earth), E.-E. C. (2019a). EC-Earth-Consortium EC-Earth3 model output prepared for CMIP6 CMIP historical. *Earth System Grid Federation*. <https://doi.org/10.22033/ESGF/CMIP6.4700>.
- (EC-Earth), E.-E. C. (2019b). EC-Earth-Consortium EC-Earth3-Veg model output prepared for CMIP6 CMIP historical. *Earth System Grid Federation*. <https://doi.org/10.22033/ESGF/CMIP6.4706>.
- Jungclaus, J., Bittner, M., Wieners, K.-H., Wachsmann, F., Schupfner, M., Legutke, S., Giorgetta, M., Reick, C., Gayler, V., Haak, H., de Vrese, P., Raddatz, T., Esch, M., Mauritsen, T., von Storch, J.-S., Behrens, J., Brovkin, V., Claussen, M., Crueger, T., ... Roeckner, E. (2019). MPI-M MPI-ESM1.2-HR model output prepared for CMIP6 CMIP historical. *Earth System Grid Federation*. <https://doi.org/10.22033/ESGF/CMIP6.6594>.

- Karpatitis, A., Hadjinicolaou, P., & Zittis, G. (2025). A new efficiency metric for the spatial evaluation and inter-comparison of climate and geoscientific model output. <https://doi.org/10.5194/egusphere-2025-1471>.
- Ridley, J., Menary, M., Kuhlbrodt, T., Andrews, M., & Andrews, T. (2019). MOHC HadGEM3-GC31-MM model output prepared for CMIP6 CMIP historical. Earth System Grid Federation. <https://doi.org/10.22033/ESGF/CMIP6.6112>.
- Uwe, S. (2023). CDO User Guide (2.3.0). Zenodo. <https://doi.org/10.5281/zenodo.10020800>.
- Voldoire, A. (2019). CNRM-CERFACS CNRM-CM6-1-HR model output prepared for CMIP6 CMIP historical. Earth System Grid Federation. <https://doi.org/10.22033/ESGF/CMIP6.4067>.

How to Cite This Article

Mardyanto, M. A., Dary, A. M., Santoso, R. I. B., Nurhayati, E., Pradana, H. P., & Muzakky, A, (2026). Evaluation of Climate Models Using Bergen Metrics as a Multi-Metrics Comparison over Indonesia. Civil and Environmental Engineering, 0 (0). <https://doi.org/10.2478/cee-2026-0107>
