Agromet 38 (1): 19-35, 2024

Agromet
journal.ipb.ac.id/index.php/agromet
ISSN : 0126 - 3633

A Comparison of the Performance of the Weighted Ensembles Means in CORDEX-SEA Precipitation Simulations

Tugiyo Aminoto[1,2], Akhmad Faqih[2], Perdinan[2], Yonny Koesmaryono[2], Bambang Dwi Dasanto[2]

[1]Department of Physics, University of Jambi, Jambi, 36129, Indonesia
[2]Department of Geophysics and Meteorology, IPB University, Bogor, 16680, Indonesia

ARTICLE INFO

Received
03 November 2023

Revised
30 November 2023

Accepted for Publication
20 February 2024

Published
19 March 2024

doi: 10.29244/j.agromet.38.1.19-35

Correspondence:
Akhmad Faqih
Department of Geophysics and Meteorology, IPB University, Bogor, 16680, Indonesia
Email: akhmadfa@apps.ipb.ac.id

ABSTRACT

Numerous studies stated that the performance of ensemble mean derived from multiple climate models generally surpassed the individual member model, and applying weighting factors potentially increase the ensemble mean of performance. This study aims to assess the performance of unweighted and weighted ensemble means of 9-modelled precipitation datasets in the CORDEX-SEA multi-model simulations for 1981-2005. The 9 datasets included: CNRM_a, ECE_b, GFDL_b, IPSL_b, HadGEM2_a, HadGEM2_c, HadGEM2_d, MPI_c, and NorESM1_d. The weighting factors were derived from the models' skill scores measured using five statistical-based metrics, namely Taylor, Pierce (SS), Tian skill score (Tian), Climate prediction index (CPI), and Performance and Independence (PI). The ERA5 and GPCP precipitation datasets were used as the references for comparison. Then, reliable metrics will be used to determine the weighting factor. The results found that three metrics namely Taylor, SS, and Tian were more reliable than the other two metrics (CPI and PI). Spatially, the weighted ensemble mean based on a random method was superior to other ensemble mean methods and individual models. We found that the CNRM_a and GFDL_b models were spatially performed best. In contrast, most the ensemble means was temporally less performed compared to the individual model. Our findings suggested that by removal of low performance models will significantly influence on the overall ensemble model performance. Further, the research may provide valuable considerations of climate models selection for climate projection assessments, especially in the Southeast Asia region.

KEYWORDS
climate models, model skill score, performance evaluation, statistical-based metrics, weighting factors

INTRODUCTION

The numerous climate models with their respective strengths and weaknesses undoubtedly require evaluation, selection, or combination to obtain more accurate climate projection results (Lutz et al., 2016; Pierce et al., 2009). As it is essential to evaluate the output of climate models (Kim et al., 2014), it needs to be done more efficiently and consistently (Eyring et al., 2016). However, previous assessments of climate models that combine many different factors aren't always helpful in certain situations. A model that works well for a specific variable, time scale and area, might not work well for other variable, time scale and area (Schaller et al., 2011). This is mainly due to different formulations and parameterizations within each model (Rummukainen, 2016). Thus, evaluating the proficiency of climate models remains a complex task (Siew et al.,

2014). Evaluation of the CORDEX-SEA dataset, focusing on the ensemble mean, reveals a higher fit between the ensemble mean and the observational data. A study by Tangang et al., (2020) compared the ensemble mean of RCMs and GCMs with the result that RCMs which have a higher resolution can produce a better performance even in areas with complex topography. The ensemble mean has a higher correlation value and a lower RMSE than its individual component models (Lee and Wang, 2014).

Several studies stated that the ensemble mean shows a better agreement to observational data than the single model (Schaller et al., 2011). It might be caused by the possibility that wet and dry biases from individual models cancel each other out (Nguyen et al., 2022). In addition, using the ensemble mean in climate model analysis aims to reduce uncertainty (Doblas-Reyes, 2021; Wang et al., 2019). Studies in CORDEX-SEA ensemble mean calculations often use an averaging method without considering weight factors for constituent model performance (Tangang et al., 2020). While investigations suggest that unequally weighted multi-model combinations may have varying success (Delsole and Tippett, 2012; Christensen et al., 2010). Some studies show improved results with weight factors applied (Brunner et al., 2019; Wang et al., 2019; Knutti et al., 2017b).

Weighting methods are based on model performance and have the potential to increase the reliability of climate model performance (Casanova and Ahrens, 2009; Weiland et al., 2021). Hence, it is necessary to explore the potential of incorporating weighting factors in the ensemble mean of the CORDEX-SEA multi-model simulations. As there is no standardized approach for determining weighting factors in forming the weighted ensemble mean, the general knowledge is the potential source of these weights lies in the models' performance, evaluated using the relevant metrics (Flato et al., 2013). Moreover, as the performance of the models are commonly assessed by a metric that quantifiably measure the similarity of a model to the reference (Ningrum et al., 2023; Reed et al., 2022) different types of metrics frequently produce varying results. Hence, assessing the robustness of metrics being used to measure the models' performance is crucial.

The aim of this study is to assess the performance of the weighted ensemble means from nine CORDEX-SEA output models in simulating precipitation during the historical period from 1981 to 2005. Given that the accuracy of the weighting factors used relies on the resilience of the chosen metric, this study also investigates the robustness of several metrics to consistently measure the models' performance. The weighting factors used were determined based on the scores obtained from the selected metrics and from random weights (Chen et al., 2017). The findings of this study are expected to serve as additional important considerations when selecting the appropriate ensemble mean for climate projection analysis in the Southeast Asian region.

RESEARCH METHODS

Data

The datasets utilized in this study comprised of nine models extracted from the CORDEX-SEA output simulations, encompassing the Southeast Asia region

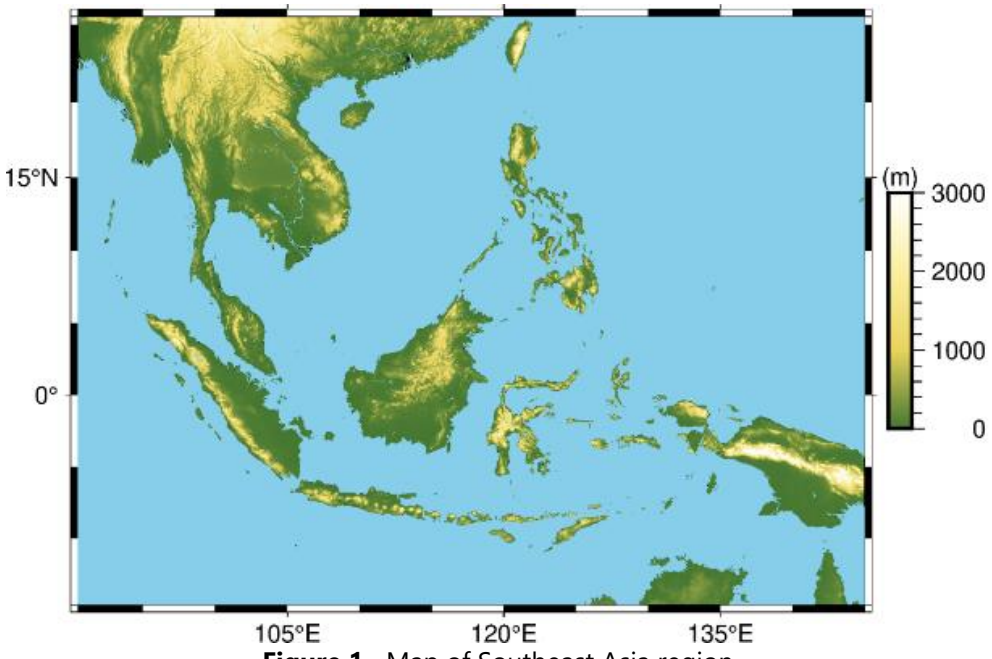

Figure 1. Map of Southeast Asia region.

Table 1. Climate models' overview. The lowercase letters represent the model ID in respect to their RCM models.

No	GCM (Driving model)		RCM (Dynamical downscaling)		Grid size (lat, lon)	Model ID
	Institute	Model	Institute	Model		
1	CNRM (France)	CNRM-CM5	SMHI (Sweden)	RCA4	194, 264	CNRM_a
2	EC-Earth (Europe)	EC-EARTH	ICTP (Italy)	RegCM4-3	191, 253	ECE_b
3	NOAA (USA)	GFDL-ESM2M	ICTP	RegCM4-3	191, 253	GFDL_b
4	IPSL (France)	IPSL-CM5A-LR	ICTP	RegCM4-3	191, 253	IPSL_b
5	Hadley Centre (UK)	HadGEM2-ES	SMHI (Sweden)	RCA4	194, 264	HadGEM2_a
6	Hadley Centre (UK)	HadGEM2-ES	ICTP	RegCM4-7	189, 335	HadGEM2_c
7	Hadley Centre (UK)	HadGEM2-ES	GERICS (Germany)	REMO2015	201, 273	HadGEM2_d
8	MPI (Germany)	MPI-ESM-MR	ICTP	RegCM4-7	189, 335	MPI_c
9	NCC (Norway)	NorESM1-M	GERICS	REMO2015	201, 273	NorESM1_d

geographically positioned between 90.5°E to 145°E and 14.5°S to 25.5°N (Figure 1). The utilized climate models comprise monthly historical precipitation data spanning the period from 1981 to 2005, with a resolution of 0.22° (approximately 25 km). The choice of the evaluation time period was determined by data availability. Specifically, the observational dataset (GPCP) covers the period from 1979 to 2017, whereas the model historical datasets are available only until 2005. The description of the models is presented in the Table 1.

Data Processing

Considering that the ensemble mean's superiority may stem from the potential cancellation of wet and dry biases from individual models (Nguyen et al., 2022), it suggests that this superiority may depend on the size of the domain where, bigger is better. Therefore, this study covers both land and ocean regions. Acquiring reference datasets that encompass both land and ocean areas, while maintaining a resolution comparable to the models under assessment, presents a challenge.

Hence, ERA5 (Hersbach et al., 2020) and GPCP (Adler et al., 2018) were selected as the reference datasets as both of them encompass land and ocean. Several studies have shown the advantages of ERA5 as a reference dataset, not only because of its higher resolution but also its comparable quality (Vanella et al., 2022; Jiao et al., 2021; Tarek et al., 2020). The ERA5 and GPCP datasets exhibit a high level of quality when compared to other observational datasets used for comparison, as shown in Figure A1, with their similarity suggesting potential interchangeability (Tangang et al., 2020) as reference. The description of the references data that used is presented in the Table A1.

This study employed RCMES (Regional Climate Model Evaluation System), an open-source Python-based program (Lee et al., 2018), as the primary tool for processing, analyzing, and visualizing the data. Adjustments were made to match it to the specific characteristics of the data and the study domain. The evaluation stage began with checking the minimum and maximum values in the data to ensure the absence of outlier data. Next, a selection was conducted among the considered multi-metrics to identify the most robust metrics that would be used in determining the weighting factors. The considered metrics are Taylor's skill score (S) (Taylor, 2001), Climate Prediction Index (CPI) (Murphy et al., 2004), Pierce's skill score (SS) (Pierce et al., 2009), Tian's skill score (Tian) (Tian et al., 2017), and Performance and Independence weighting (PI) (Sanderson et al., 2017; Knutti et al., 2017a; Brunner et al., 2019), and orderly listed in the Equation (1-5).

$$S = \frac{4(1+r)}{(\sigma_r+1/\sigma_r)^2\,(1+r_0)} \quad (1)$$

$$CPI = \exp\left[-0.5\,\frac{(s-o)^2}{\sigma^2}\right] \quad (2)$$

$$SS = r_{m,o}^2 - \left[r_{m,o} - \left(\frac{S_m}{S_o}\right)\right]^2 - \left[\frac{(\bar{m}-\bar{o})^2}{S_o}\right]^2 \quad (3)$$

$$Tian = \frac{1+R}{2}\left[1 - MSE/(Bias^2 + \sigma_m^2 + \sigma_o^2)\right] \quad (4)$$

$$PI_i = \frac{e^{-\frac{D_i}{\sigma_D}}}{1+\sum_{j\neq i}^{M} e^{-\frac{S_{ij}}{\sigma_S}}} \quad (5)$$

In these equations, r represents the correlation coefficient for the observed and simulated data, r_0 denotes the maximum correlation (set to 1 in this study), σ_r is the standard deviation, s is the mean of simulated data, o is the mean of observed data, σ is variance, S_m is standard deviation of modeled data, S_o is standard deviation of observed data, and the bar symbol on top of m and o indicate the average value of modeled and observed data respectively.

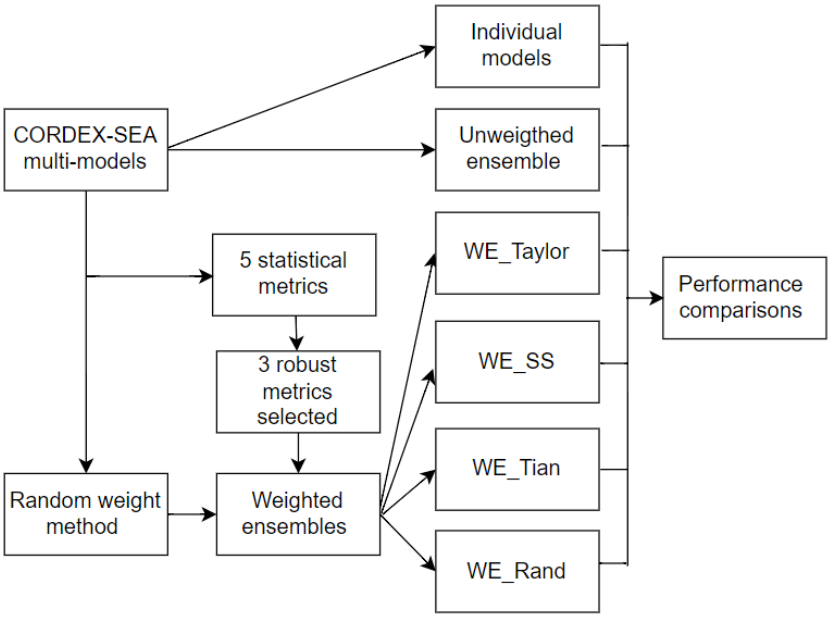

Figure 2. Schematic diagram of the research methods.

Ensemble Means Performance

This study assessed the performance of two types of ensemble means. The first is the unweighted or equally weighted ensemble mean (Herger et al., 2017), commonly referred to as the multimodel ensemble mean (MME) (Shin et al., 2020). In this type, each model is assigned the same weighting factor (usually 1) in the averaging process regardless of its performance score. The second type is the unequally weighted ensemble mean, where each model is assigned different weighting factors based on its respective performance scores. In this study, we formed three types of weighted ensemble means (WE_Taylor, WE_SS, WE_Tian) based on the three most robust metrics used to assign weight factors according to their performance scores. The model performance scores were obtained by calculating the temporal average, standard deviation, correlation, and other statistical parameters required by each metric (Equation 1-5). These calculations are based on the annual rainfall time series data.

Additionally, we created the fourth type, one (WE_Rand) using the random weight method by generating weights randomly 100 times and selecting the weight configuration that yielded a standard deviation ratio and correlation closest to 1. These steps are illustrated in Figure 2. Subsequently, this study assessed the performance of the models and the ensemble means in terms of various aspects, including spatial mean, zonal mean, and seasonal-to-inter-annual variability. Furthermore, as the SEA region experiences strong seasonal contrasts in precipitation distribution (Nguyen et al., 2022), to accurately capture these seasonal contrasts, the evaluations were mostly conducted in summer (JJA) and winter (DJF) seasons. Finally, the performance results of the models are summarized in the last section.

RESULTS AND DISCUSSION

In this study, to validate the data used, the initial phase of model performance evaluation involves screening the model's maximum and minimum values. This step is crucial for identify any extreme values in the error category caused by the presence of relaxing zones in the RCMs (Giorgi, 2019). If these relaxing zones persist in the model's dataset, they need to be excluded, as the objective is to evaluate valid datasets. Among the 9 model datasets examined, it was discovered that IPSL_b and GFDL_b models exhibited unreasonable maximum values (resulting from relaxing zone effects) at the lateral boundaries domain. To mitigate the impact of such errors on subsequent stages of evaluation, the study domain was shifted to latitude and longitude boundaries that were free from these errors. Screening the minimum values plays a crucial role in identifying invalid rainfall amounts below zero. Some models exhibited negative values, but these values have very small magnitudes (on the order of < E-10), rendering them insignificant.

Robust Metrics and Weighting Factors

The annual rainfall from each models is shown in Figure 3. Comparison of the results from five metrics is illustrated in Figure 4a revealing considerable variability in performance scores across each metric. In Figure 4a, two main patterns are clearly visible. Firstly, most metrics provide higher scores for the observation dataset (GPCP). Secondly, most metrics also demonstrate a high agreement with the ensemble

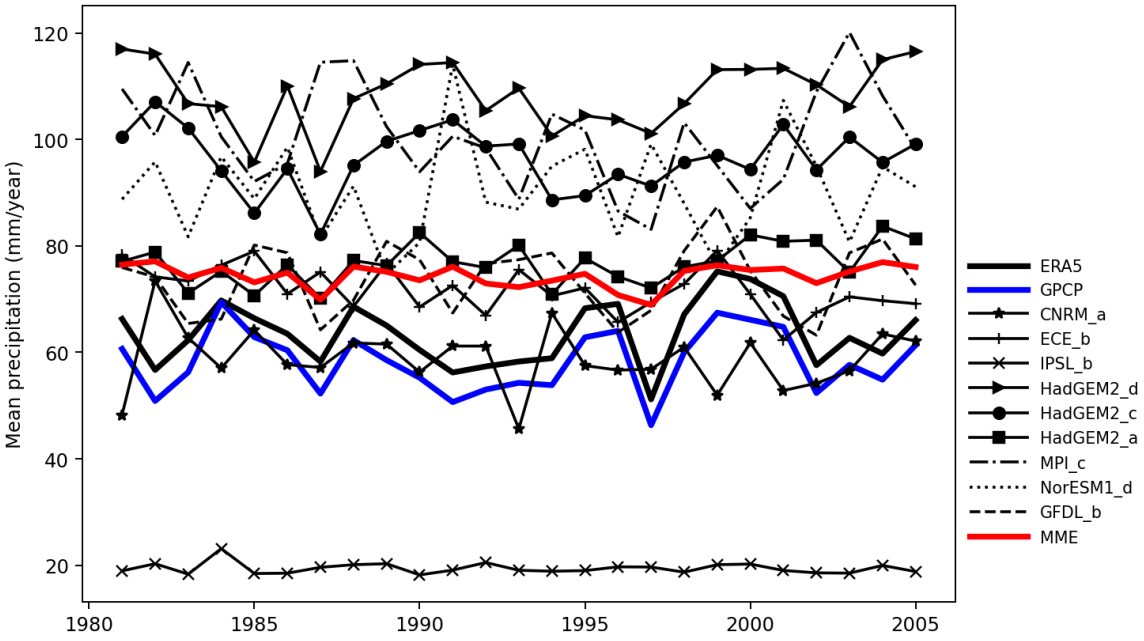

Figure 3. Time series of the annual mean precipitation over SEA region.

mean (MME). This result is consistent with typical model evaluations, where the MME generally outperforms individual models in most cases (Tangang et al., 2020; Lee and Wang, 2014; Schaller et al., 2011). In contrast, for the models, the metrics produce varied results. In this stage, to identify robust metrics, we examined how the metrics scored the models that were clearly close or far from the reference dataset. For instance, we observed how the metric scores CNRM_a, which is very close to the reference. In this regard, it should receive a higher score as it is very close to the reference dataset.

However, the PI metric gives a very lower score, indicating that in this case, PI is not a robust metric. This is also supported by the fact that this metric also gives lower scores to ECE_b, HadGEM2_a, NorESM1_d and GFDL_b, while most metrics give them higher scores. Therefore, this analysis indicates that the PI metric exhibits the lowest of robustness compared to other metrics. Additionally, Gleckler et al., (2008) highlighted the importance of assessing the associations between the data provided by each metric to determine their robustness. Based on this, Figure 4b presents the cross-correlation results among the metrics, indicating that the CPI metric yields the lowest result. This is because CPI demonstrates a weak correlation with two metrics, unlike other metrics that exhibit a weak correlation with no more than one metric. Taking this into account, the CPI metric should also not be deemed a robust metric.

Although no single evaluation technique or performance measure is deemed superior (Flato et al., 2013), and a better approach is to use a combination of several metrics or average the results from various metrics (Reichler and Kim, 2008), in this study we chose to minimize the number of metrics used by excluding some based on the above robustness considerations. Upon excluding the PI and CPI metric, comparing the remaining three metrics is challenging, as each possesses its own slight strengths and weaknesses. This is evident from how they score the IPSL_b and ECE_b models.

In this stage, Tian metric is preferable because it assigns a lower score to IPSL_b (which has a significant underestimation) while the other metrics (Taylor and SS) give higher scores. However, when scoring ECE_b (which has a significant overestimation), the situation is reversed (Taylor and SS are better). Considering this, the study selected Taylor, SS, and Tian as the most robust metrics for quantitatively measuring the skill scores of the models. These skill scores including scores from random weighting method, in Figure 4c were subsequently employed as weighting factors in the creation of four weighted ensemble means, namely WE_Taylor, WE_SS, WE_Tian, and WE_Rand.

Spatial-Based Evaluation

This evaluation was conducted on the average of climatological rainfall over 25 years in the winter (DJF) and summer (JJA) seasons. The SEA region shows significant seasonal variations in precipitation distribution (Nguyen et al., 2022). Hence, to accurately capture these variations, separate evaluations were performed for the summer (JJA) and winter (DJF) seasons separately. It can be seen from Figure 5 that among the ensemble mean types, the highest result is obtained by the weighted ensemble mean of WE_Rand

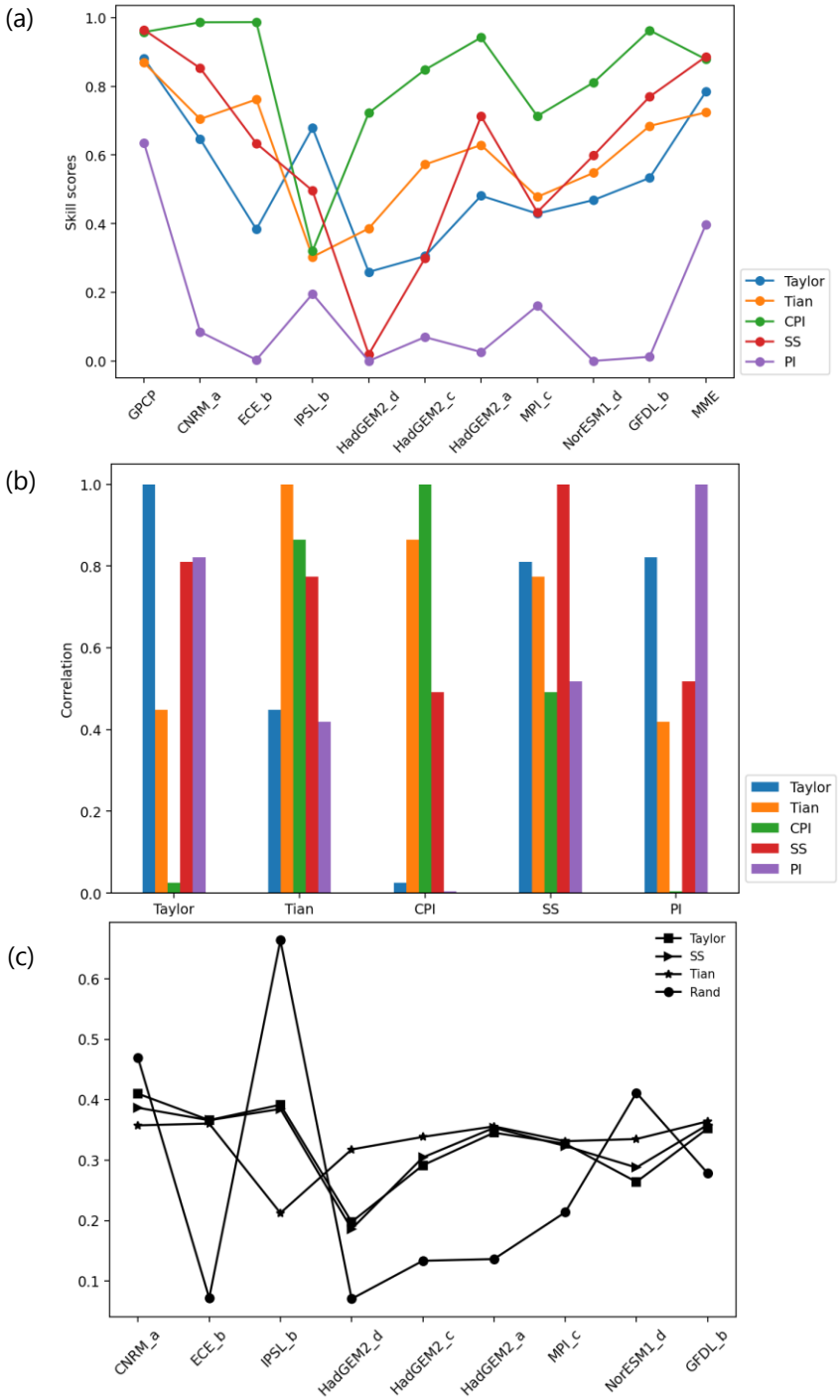

Figure 4. Plot of (a) skills scores of the models based on each metric (Taylor's, Climate Prediction Index (CPI), skill score (SS), Tian's skill score (Tian), and Performance and Independence weighting (PI)) (b) the cross-correlation among those five metric scores, and (c) four weighting factors obtained from three performance metrics and random (rand) method.

(0.93), followed by WE_SS (0.91), and MME (0.90), WE_Tian (0.89) and WE_Taylor (0.88). This indicates that ensemble means with weighting factors, slightly produce better results. This supports the findings that weighting methods have the potential to increase the climate model performance (Casanova and Ahrens, 2009; Weiland et al., 2021) and it might be only suitable for a specific case (Delsole and Tippett, 2012; Christensen et al., 2010). In addition, GPCP, exhibits remarkably similar pattern (the highest similarity score of 0.94), indicating its significant proximity to ERA5. Moreover, the models that are most similar to the reference datasets are CNRM_a with the similarity score of 0.83 followed by ECE_b (0.73) and HadGEM2_a (0.73). In JJA season (Figure A2), the results are slightly lower than DJF. Atmospheric circulation in Southeast Asian region is largely modulated by the Asian-Australian monsoon, where the migration of the inter-tropical

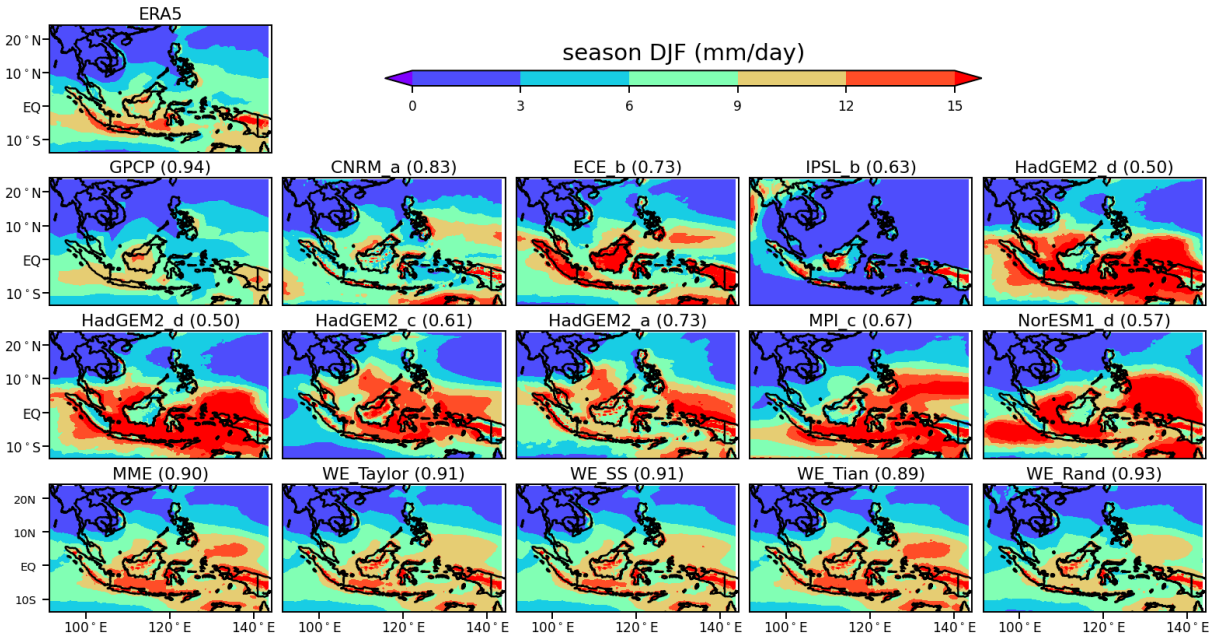

Figure 5. Climatological mean rainfall from 1981 to 2005 in the SEA region in winter (DJF) season. The value inside the brackets represents the pattern similarity scores with the references (ERA5). MME (and four last figures) represents ensemble mean without (with) weighting factors.

convergence zone (ITCZ) and monsoon (Robertson et al., 2011), trough affects the distribution of precipitation in the region. Therefore, the models must be able to simulate this regional circulation to capture the distribution of precipitation correctly (Tangang et al., 2020). Figure 6 shows the distribution of average precipitation around the equator obtained from the climatological and zonal mean (with respect to longitude). The GPCP and ERA5 datasets have an identical pattern, namely the letter 'A'. The temporal-spatial distribution pattern of rainfall (letter A) over the SEA results from the year-round ITCZ migration pattern around the equator (Tangang et al., 2020).

It is interesting to note that all ensemble means outperform all individual models, and the highest score is for the weighted ensemble mean WE_Rand (0.97), followed by WE_SS (0.96), WE_Taylor (0.96), and MME (0.94). The individual model that has the highest performance is CNRM_a (0.85) and the lowest is IPSL_b (0.44). Since the comparison of model performances is

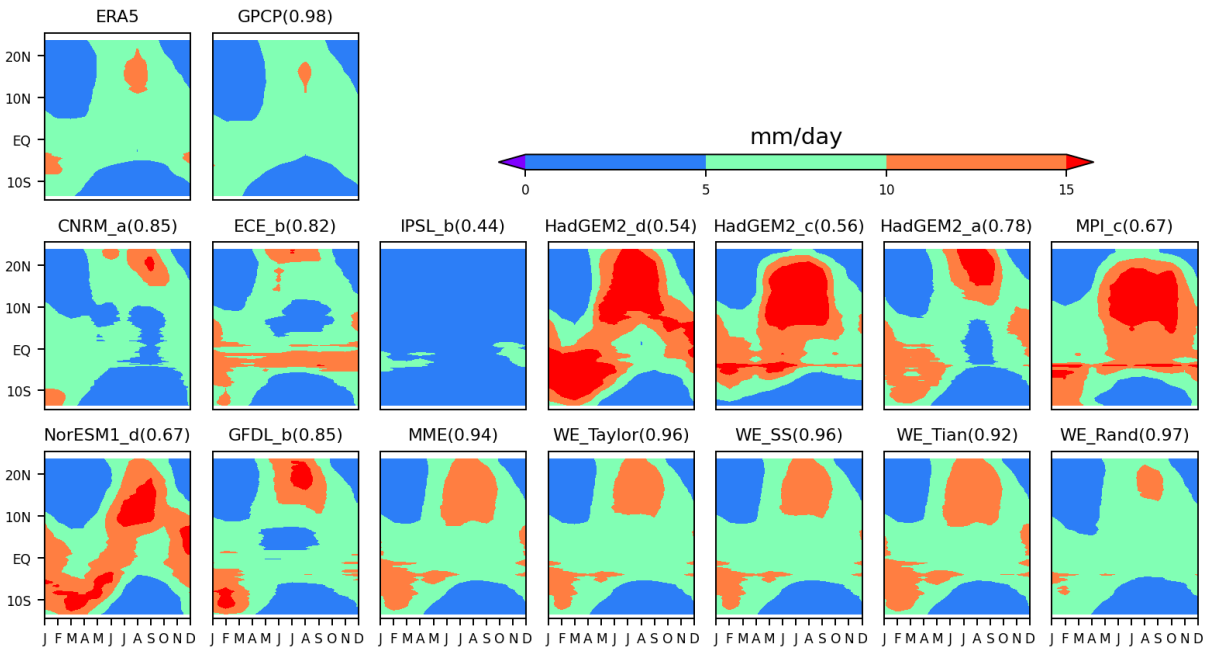

Figure 6. Zonally averaged annual cycle of precipitation over SEA region in the CORDEX-SEA models and five types of their ensemble means. The values in the brackets represent the similarity of the model to the reference (ERA5).

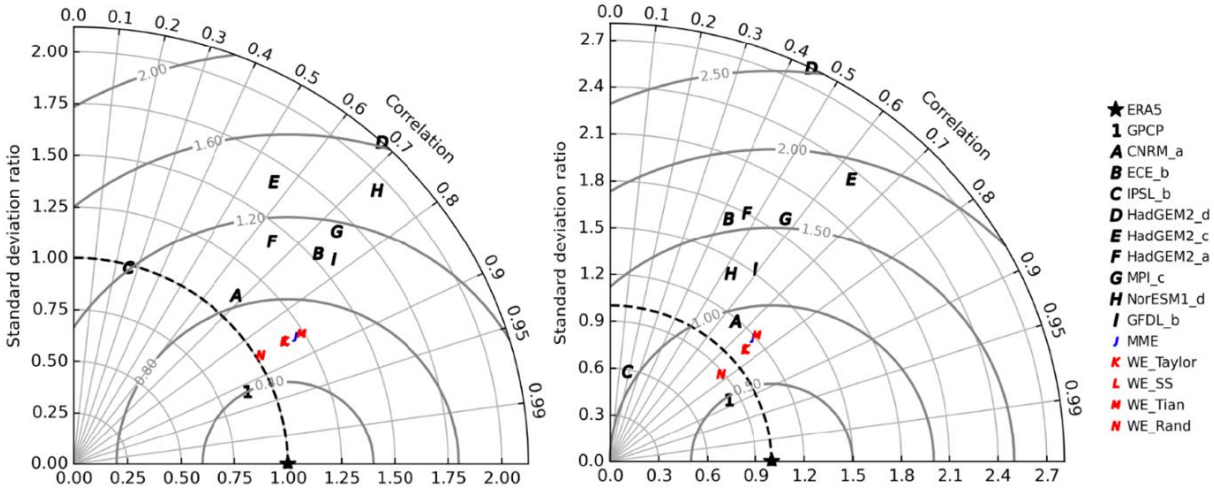

Figure 7. Taylor diagram of DJF (left) and JJA (right) seasons showing the performance of 9 models (A-I), ensemble mean (J), and four weighted ensemble means: Taylor (K), SS (L), Tian (M), Random (N).

typically presented using a Taylor diagram, this section provides an additional plot for comparison. The calculation of the standard deviation ratio and pattern correlation utilized in the Taylor diagram is performed on a spatial basis. Figure 7 clearly shows that each model has varied performance, and the weighted ensemble mean of WE_Rand (N) is in the leading position, followed by WE_SS, WE_Taylor, and the unweighted ensemble mean MME (J). It is apparent that the expected improvement in performance from the weighted ensemble mean over the unweighted ensemble mean occur slightly in this case, as mentioned in several studies (Sanderson et al. 2017; Knutti et al. 2017; Brunner et al. 2019).

It is also evident that most models and ensemble means exhibit higher performance during the winter (DJF) compared to the summer (JJA) which aligns with Tangang et al., (2020) findings. However, this is not a case in the study by Tuyet et al., (2019), which also assessed the same models, but they presented a Taylor diagram based on an annual time scale that may not be reliable due to the strong seasonal contrasts in precipitation distribution in the SEA region. To accurately capture these seasonal contrasts, the evaluation should be conducted for specific seasons (Nguyen et al., 2022).

Temporal-Based Evaluation

This section examines the seasonal to inter-annual variability of modeled and observed rainfall by employing a wavelet analysis (Jiang et al., 2013; Chao et al., 2014). This method is superior to the power

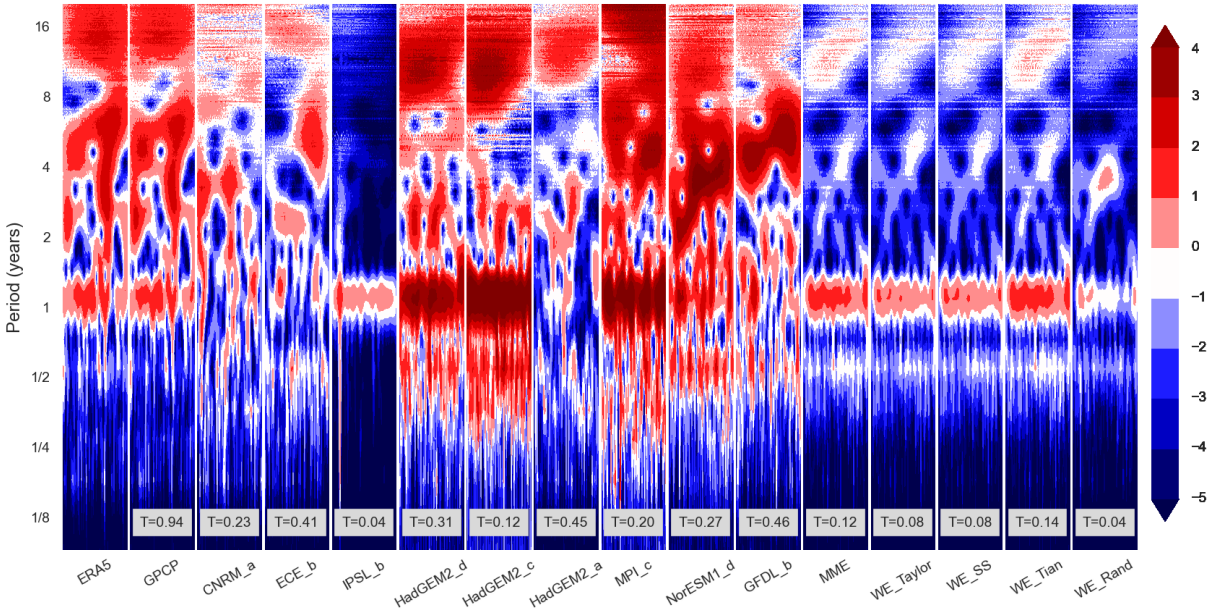

Figure 8. Wavelet profile of the SEA mean precipitation variability. The T value represents the similarity pattern to the reference (ERA5).

Table 2. Summary of models and ensemble means with the highest skill scores.

Aspect of performance evaluation	Model	Score	Ensemble mean	
			Type	Score
Spatial mean in DJF season	CNRM_a	0.83	Rand	0.93
			SS	0.91
Spatial mean in JJA season	CNRM_a	0.80	Rand	0.87
			SS	0.87
Zonal mean	CNRM_a	0.85	Rand	0.97
	GFDL_b	0.85	SS	0.96
Taylor diagram in DJF season	GFDL_b	*	Rand	*
			SS	*
Taylor diagram in JJA season	CNRM_a	*	Rand	*
			SS	*
Wavelet	GFDL_b	0.46	Tian	0.14
FFT	GFDL_b	0.80	MME	0.85
			Tian	0.85

Note: Rand = Random; SS = Pierce; Tian = Tian' skill score; MME = Multimodel Ensemble Mean

spectral was used by Siew et al., (2014). The modes are related to monsoonal ENSO and IOD teleconnections, and pacific decadal oscillation (Mantua and Hare, 2002). In this context, the wavelet were constructed from the time series of mean rainfall data in the model datasets. To quantitatively measures to the models' resemblance to reference dataset, Taylor's skill score was utilized, and the outcomes are presented on each wavelet graph to facilitate comparison. It is evident in Figure 8 that most variability modes are concentrated within three dominant patterns with periodicities of approximately 1, 2-8, and around 16 years. The ENSO periodicity of 2-8 years is slightly different from Siew et al., (2014) (2-5 years), where they used GPCP (version v2.2) as the reference and power spectrum as the method.

It is also clear that GPCP and ERA5 exhibit an almost identical pattern, with a similarity score of 0.94. However, for the ensemble mean types, all of them have a low similarity to the reference datasets. This finding requires further investigation as the ensemble mean generally outweighs all the individual model in most cases through cancelling out errors (Hagedorn et al., 2005; Weigel et al., 2014) or wet and dry bias (Nguyen et al., 2022) among models that occur during averaging operation. In addition, among the models, GFDL_b is the highest (0.46), followed by HadGEM2_a (0.45) and ECE_b (0.41). Conversely, the IPSL_b model performs poorly (0.04).

The CNRM_a model, which previously exhibited superior performance, in this wavelet plot has a relatively low score (0.23). This discrepancy could be attributed to the fact that IPSL_b, being one of the members of the ensemble mean constituent, exhibited inferior performance across most evaluation stages. Due to the underwhelming performance of the IPSL_b model, it was excluded from the evaluation section that utilized new ensemble means without IPSL_b. Excluding a model of lower quality from ensemble mean formation is justifiable (Knutti et al., 2017). The outcomes of this new configuration (Figure A3 to A5) suggest that this revised configuration yielded significant alterations. This discovery affirms that the subpar skill observed across all ensemble means in the wavelet analysis is partly due to the subpar performance of the IPSL_b model. Consequently, further investigation is warranted to uncover the primary underlying causes.

To compare the wavelet plot to other methods, we also present a Fast Fourier Transform (FFT) plot based on the period (Figure A6). These FFT plots demonstrate higher scores in comparison to the results obtained from wavelet analysis. All types of ensemble means (MME, WE_Taylor, WE_SS, and WE_Tian) exhibit higher similarity scores ranging from 0.82 to 0.85, surpassing all individual models that range from 0.28 to 0.80. The GFDL_b is the model that has highest similarity score (0.80), followed by NorESM1_d (0.67). A summary of the varied abilities of the individual models, unweighted (MME) and the weighted ensemble means (Taylor, SS, Tian) is presented in Table 2. The model that has slightly lower score than the best model is included in the summary as the additional comparison. For the Taylor diagram, it has no explicit score as it visually represent the models' performance.

This study revealed that no model consistently exhibits significant performance across all evaluation aspects in the Southeast Asia region. While on the temporal basis, they did not show better performance, on the spatial basis, the weighted ensemble means, especially WE_Ran, demonstrated better performance

than other models. This is somewhat aligned with (Bukovsky et al., 2019), who mentioned that the metrics and resulting weights do not significantly differentiate between the simulations. Among the individual models, the CNRM_a model demonstrated better performance, and this aligns with other studies (Tuyet et al., 2019; Kamworapan and Surussavadee, 2017; Siew et al., 2014).

CONCLUSIONS

This study has evaluated the performance of the equally and unequally weighted ensemble mean of nine modeled precipitations in the CORDEX-SEA output models in historical period from 1981 to 2005. The weighting factors used were derived from the models' skill scores. The findings indicate that, when it comes to spatial performance, the weighted ensemble mean (WE_Rand) outperforms all other models, with CNRM_a showing the highest performance among individual models, followed by GFDL_b. When examining temporal aspects, all types of ensemble mean produces lower results than most individual models, while the individual model GFDL achieves the highest score, followed by NorESM1_d. Furthermore, eliminating the lowest-performing model has a significant impact on the ensemble mean's performance.

These research findings are anticipated to provide valuable additional insights when selecting a more accurate ensemble mean for climate projection assessments, particularly in Southeast Asia. Here are some practical strategies highlighting the advantages of using a weighted ensemble mean over an unweighted one or individual models, firstly, weighting models based on their historical performance or skill scores allows the ensemble mean to give more emphasis to better-performing models. This approach tends to improve overall climate projection accuracy and reliability compared to an unweighted ensemble or individual models. Secondly, models might have inherent biases or systematic errors. Assigning weights based on the ability of models to represent certain climatic features or historical accuracy can help correct biases in the ensemble mean, providing more realistic climate projections.

ACKNOWLEDGEMENT

We thank LPDP-Indonesia, which provided funding to the first author through the Doctoral Scholarship Research Grant scheme. Also, we thank to the anonymous reviewers for their valuable comments and feedbacks during review process.

REFERENCES

Adler, R.F., Sapiano, M.R.P., Huffman, G.J., Wang, J.J., Gu, G., Bolvin, D., Chiu, L., Schneider, U., Becker, A., Nelkin, E., Xie, P., Ferraro, R., Shin, D. Bin, 2018. The Global Precipitation Climatology Project (GPCP) monthly analysis (New Version 2.3) and a review of 2017 global precipitation. Atmosphere 9. https://doi.org/10.3390/atmos9040138.

Brunner, L., Lorenz, R., Zumwald, M., Knutti, R., 2019. Quantifying uncertainty in European climate projections using combined performance-independence weighting. Environ. Res. Lett. 14 14. https://doi.org/10.1088/1748-9326/ab492f.

Bukovsky, M.S., Thompson, J.A., Mearns, L.O., 2019. Weighting a regional climate model ensemble: Does it make a difference? Can it make a difference? Climate Research 77, 23–43. https://doi.org/10.3354/cr01541.

Casanova, S., Ahrens, B., 2009. On the Weighting of Multimodel Ensembles in Seasonal and Short-Range Weather Forecasting. American Meteorological Society 3811–3822. https://doi.org/10.1175/2009MWR2893.1.

Chao, B.F., Chung, W., Shih, Z., Hsieh, Y., 2014. Earth's rotation variations: A wavelet analysis. Terra Nova 26, 260–264. https://doi.org/10.1111/ter.12094.

Chen, J., Brissette, F.P., Lucas-Picher, P., Caya, D., 2017. Impacts of weighting climate models for hydro-meteorological climate change studies. Journal of Hydrology 549, 534–546. https://doi.org/10.1016/j.jhydrol.2017.04.025.

Christensen, J.H., Kjellström, E., Giorgi, F., Lenderink, G., Rummukainen, M., 2010. Weight assignment in regional climate models. Climate Research 44, 179–194. https://doi.org/10.3354/cr00916.

Delsole, T., Tippett, M.K., 2012. Is unequal weighting significantly better than equal weighting for multi-model forecasting ? Q. J. R. Meteorol. Soc. https://doi.org/10.1002/qj.1961.

Doblas-Reyes, 2021. SPM 10: Linking Global to Regional Climate Change. https://doi.org/10.1017/9781009157896.012.1364.

Eyring, V., Gleckler, P.J., Heinze, C., Stouffer, R.J., Taylor, K.E., 2016. Towards improved and more routine Earth system model evaluation in CMIP. Earth Syst. Dynam. 813–830. https://doi.org/10.5194/esd-7-813-2016.

Flato, G., Marotzke, J., B. Abiodun, Braconnot, P., 2013. Evaluation of climate models. Climate Change 2013 the Physical Science Basis: Working Group I Contribution to the Fifth Assessment

Report of the Intergovernmental Panel on Climate Change 9781107057, 741–866. https://doi.org/10.1017/CBO9781107415324.020.

Giorgi, F., 2019. Thirty Years of Regional Climate Modeling : Where Are We and Where Are We Going next ? Journal of Geophysical Research : Atmospheres. Journal of Geophysical Research: Atmospheres 124, 5696–5723. https://doi.org/10.1029/2018JD030094.

Gleckler, P.J., Taylor, K.E., Doutriaux, C., 2008. Performance metrics for climate models. Journal of Geophysical Research Atmospheres 113, 1–20. https://doi.org/10.1029/2007JD008972.

Hagedorn, R., Doblas-Reyes, F.J., Palmer, T.N., 2005. The rationale behind the success of multi-model ensembles in seasonal forecasting – I. Basic concept. Tellus A: Dynamic Meteorology and Oceanography 57, 219. https://doi.org/10.3402/tellusa.v57i3.14657.

Herger, N., Abramowitz, G., Knutti, R., Angélil, O., Lehmann, K., Sanderson, M., 2017. Selecting a climate model subset to optimise key ensemble properties. Earth Syst. Dynam. Discuss. 1–24. https://doi.org/10.5194/esd-2017-28.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R.J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., Thépaut, J.N., 2020. The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society 146, 1999–2049. https://doi.org/10.1002/qj.3803.

Huang, B., 2015. Assessment of precipitation climatology in an ensemble of CORDEX-East Asia regional climate simulations. Climate Research 64, 141–158. https://doi.org/10.3354/cr01302.

Jiang, P., Gautam, M.R., Zhu, J., Yu, Z., 2013. How well do the GCMs/RCMs capture the multi-scale temporal variability of precipitation in the Southwestern United States? Journal of Hydrology 479, 75–85. https://doi.org/10.1016/j.jhydrol.2012.11.041.

Jiao, D., Xu, N., Yang, F., Xu, K., 2021. Evaluation of spatial-temporal variation performance of ERA5 precipitation data in China. Scientific Reports 11, 1–13. https://doi.org/10.1038/s41598-021-97432-y.

Kamworapan, S., Surussavadee, C., 2017. Performance of CMIP5 global climate models for climate simulation in Southeast Asia. IEEE Region 10 Annual International Conference, Proceedings/TENCON 2017-Decem, 718–722. https://doi.org/10.1109/TENCON.2017.8227954.

Kim, J., Waliser, D.E., Mattmann, C.A., Goodale, C.E., Hart, A.F., Zimdars, P.A., Crichton, D.J., Jones, C., Nikulin, G., Hewitson, B., Jack, C., Lennard, C., Favre, A., 2014. Evaluation of the CORDEX-Africa multi-RCM hindcast: Systematic model errors. Climate Dynamics 42, 1189–1202. https://doi.org/10.1007/s00382-013-1751-7

Knutti, R., Sedl, J., Sanderson, B.M., Lorenz, R., Fischer, E., Eyring, V., 2017a. A climate model projection weighting scheme accounting for performance and interdependence. https://doi.org/10.1002/2016GL072012.

Knutti, R., Sedláček, J., Sanderson, B.M., Lorenz, R., Fischer, E.M., Eyring, V., 2017b. A climate model projection weighting scheme accounting for performance and interdependence. Geophysical Research Letters 44, 1909–1918. https://doi.org/10.1002/2016GL072012.

Lee, H., Goodman, A., McGibbney, L., Waliser, D.E., Kim, J., Loikith, P.C., Gibson, P.B., Massoud, E.C., 2018. Regional climate model evaluation system powered by Apache Open Climate Workbench v1.3.0: An enabling tool for facilitating regional climate studies. Geoscientific Model Development 11, 4435–4449. https://doi.org/10.5194/gmd-11-4435-2018.

Lee, J., Wang, B., 2014. Future change of global monsoon in the CMIP5. Climate Dynamics 42, 101–119. https://doi.org/10.1007/s00382-012-1564-0.

Lutz, A.F., ter Maat, H.W., Biemans, H., Shrestha, A.B., Wester, P., Immerzeel, W.W., 2016. Selecting representative climate models for climate change impact studies: an advanced envelope-based selection approach. International Journal of Climatology 36, 3988–4005. https://doi.org/10.1002/joc.4608.

Mantua, N.J., Hare, S.R., 2002. The Pacific Decadal Oscillation. Journal of Oceanography 58, 35–44. https://doi.org/10.1023/A:101582061638 .

Murphy, J.M., Sexton, D.M.H., Barnett, D.N., Jones, G.S., 2004. Quantification of modelling uncertainties in a large ensemble of climate change simulations James. Nature 396, 1–5. https://doi.org/10.1038/nature02770.1

Ningrum, W., Boer, R., Apip, 2023. Statistical Assessment of High-Resolution Climate Model Rainfall Data in the Ciliwung Watershed, Indonesia. Agromet 37, 21–33. https://doi.org/10.29244/j.agromet.37.1.21-33.

Nguyen, P.-L., Bador, M., Alexander, L. V, Lane, T.P., Ngo-Duc, T., 2022. More intense daily precipitation in CORDEX-SEA regional climate models than their forcing global climate models over Southeast Asia. Int. J. Climatol 6537–6561. https://doi.org/10.1002/joc.7619.

Pierce, D.W., Barnett, T.P., Santer, B.D., Gleckler, P.J., 2009. Selecting global climate models for regional climate change studies. Proceedings of the National Academy of Sciences of the United States of America 106, 8441–8446. https://doi.org/10.1073/pnas.0900094106.

Reed, K.A., Goldenson, N., Grotjahn, R., Gutowski, W.J., Jagannathan, K., Jones, A.D., Leung, L.R., McGinnis, S.A., Pryor, S.C., Srivastava, A.K., Ullrich, P.A., Zarzycki, C.M., 2022. Metrics as tools for bridging climate science and applications. Wiley Interdisciplinary Reviews: Climate Change 13. https://doi.org/10.1002/wcc.799.

Robertson, A.W., Moron, V., Qian, J.H., Chang, C.P., Tangang, F., Aldrian, E., Koh, T.Y., Juneng, L., 2011. The Maritime Continent Monsoon. The Global Monsoon System: Research and Forecast 85–98. https://doi.org/10.1142/9789814343411_0006.

Rummukainen, M., 2016. Added value in regional climate modeling. Wiley Interdisciplinary Reviews: Climate Change 7, 145–159. https://doi.org/10.1002/wcc.378.

Sanderson, B.M., Wehner, M., Knutti, R., Sanderson, B.M., 2017. Skill and independence weighting for multi-model assessments. Geosci. Model Dev. 2379–2395. https://doi.org/10.5194/gmd-10-2379-2017.

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 Atmospheres 116, 1–14. https://doi.org/10.1029/2010JD014963.

Shin, Y., Lee, Y., Park, J.-S., 2020. A Weighting Scheme in A Multi-Model Ensemble for Bias-Corrected Climate Simulation. Atmosphere 11, 7–10. https://doi.org/10.3390/atmos11080775.

Siew, J.H., Tangang, F.T., Juneng, L., 2014. Evaluation of CMIP5 coupled atmosphere-ocean general circulation models and projection of the Southeast Asian winter monsoon in the 21st century. International Journal of Climatology 34, 2872–2884. https://doi.org/10.1002/joc.3880.

Tangang, F., Chung, J.X., Juneng, L., Supari, Salimun, E., Ngai, S.T., Jamaluddin, A.F., Mohd, M.S.F., Cruz, F., Narisma, G., Santisirisomboon, J., Ngo-Duc, T., Van Tan, P., Singhruck, P., Gunawan, D., Aldrian, E., Sopaheluwakan, A., Grigory, N., Remedio, A.R.C., Sein, D. V., Hein-Griggs, D., McGregor, J.L., Yang, H., Sasaki, H., Kumar, P., 2020. Projected future changes in rainfall in Southeast Asia based on CORDEX–SEA multi-model simulations. Climate Dynamics 55, 1247–1267. https://doi.org/10.1007/s00382-020-05322-2.

Tarek, M., Brissette, F.P., Arsenault, R., 2020. Evaluation of the ERA5 reanalysis as a potential reference dataset for hydrological modelling over North America. Hydrology and Earth System Sciences 24, 2527–2544. https://doi.org/10.5194/hess-24-2527-2020.

Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a Single Diagram. Journal of Geophysical Research Atmospheres 106, 7183–7192.

Tian, B., Lee, H., Waliser, D.E., Ferraro, R., Kim, J., Case, J., Iguchi, T., Kemp, E., Wu, D., Putman, W., Wang, W., 2017. Development of a model performance metric and its application to assess summer precipitation over the U.S. great plains in downscaled climate simulations. Journal of Hydrometeorology 18, 2781–2799. https://doi.org/10.1175/JHM-D-17-0045.1.

Tuyet, N.T., Thanh, N.D., Tan, P.-V., 2019. Performance of SEACLID/CORDEX-SEA multi-model experiments in simulating temperature and rainfall in Vietnam. Vietnam Journal of Earth Sciences 374–387. https://doi.org/DOI: 10.15625/0866-7187/41/4/14259.

Vanella, D., Longo-Minnolo, G., Belfiore, O.R., Ramírez-Cuesta, J.M., Pappalardo, S., Consoli, S., D'Urso, G., Chirico, G.B., Coppola, A., Comegna, A., Toscano, A., Quarta, R., Provenzano, G., Ippolito, M., Castagna, A., Gandolfi, C., 2022. Comparing the use of ERA5 reanalysis dataset and ground-based agrometeorological data under different climates and topography in

Italy. Journal of Hydrology: Regional Studies 42, 101182. https://doi.org/10.1016/j.ejrh.2022.101182.

Wang, H., Chen, J., Xu, C., Chen, H., Guo, S., Xie, P., Li, X., 2019. Does the weighting of climate simulations result in a better quantification of hydrological impacts ? Hydrol. Earth Syst. Sci. 4033–4050. https://doi.org/10.5194/hess-23-4033-2019.

Weigel, A.P., Liniger, M.A., Appenzeller, C., 2014. Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? J. R. Meteorol. Soc. 1227, 496. https://doi.org/10.1002/qj.

Weiland, FCS., Visser, R.D., Greve, P., Bisselink, B., Brunner, L., Weerts, A.H., 2021. Estimating Regionalized Hydrological Impacts of Climate Change Over Europe by Performance-Based Weighting of CORDEX Projections. Frontiers in Water 3. https://doi.org/10.3389/frwa.2021.713537.

ANNEX

Table A1. Reference datasets.

Dataset	Product	Type*	Spatial covered*	Resolution, Time Covered	Reference
ERA5 (https://cds.climate.copernicus.eu)	Monthly averaged reanalysis	R	All	0.25° Monthly 1950-present	(Hersbach et al., 2020)
GPCP (https://esgf-node.llnl.gov/search/obs4mips)	v7-7A	S, G	All	2.5° Monthly 1979-2017	(Adler et al., 2018)

*G = Gauge; S = Satellite; R = Reanalysis; All = Land and Ocean

Figure A1. Taylor diagram illustrating the proximity of ERA5 (2) and GPCP (3) performance to GPCC (Global Precipitation Climatology Centre) and SAOBS (Southeast Asia Observation) dataset during the DJF (left) and JJA (right) seasons.

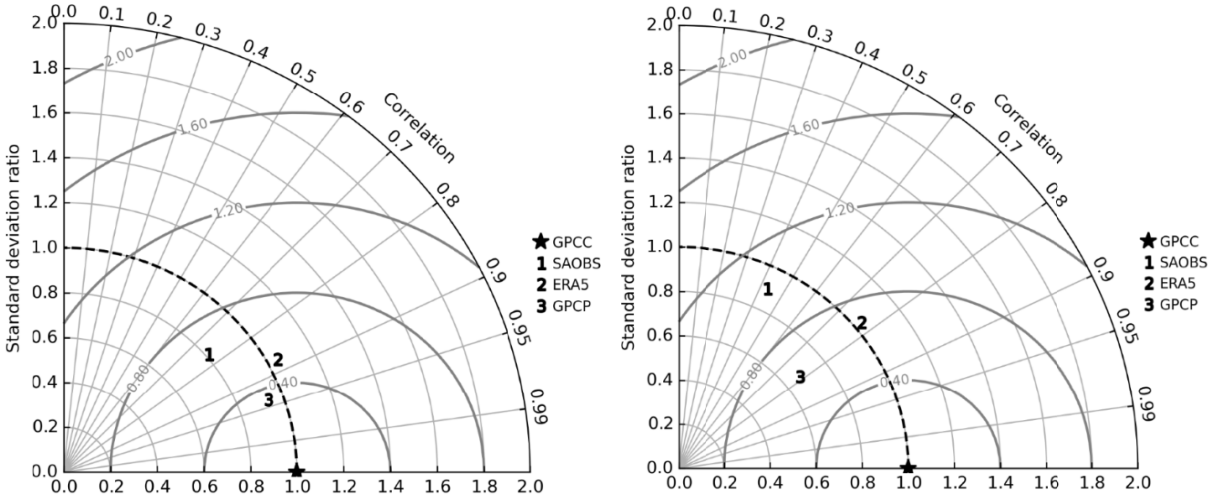

Figure A2. Climatological mean rainfall from 1981 to 2005 in the SEA region in summer (JJA) season. The value inside the brackets represents the pattern similarity scores with the references (ERA5). MME (and four last figures) represents ensemble mean without (with) weighting factors.

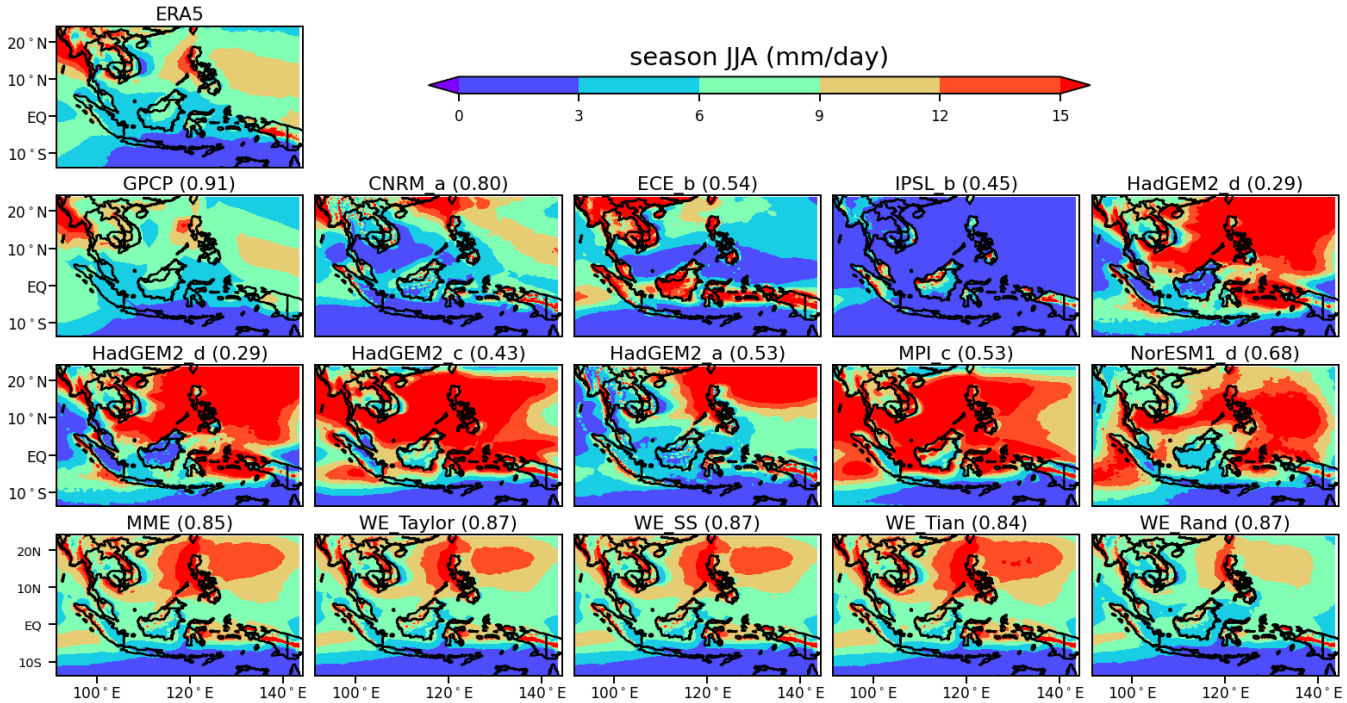

Figure A3. Zonally averaged annual cycle of precipitation (mm/month) in the SEA region for the new configuration (omitting IPSL). The value inside the bracket represents the pattern similarity of the model to the reference.

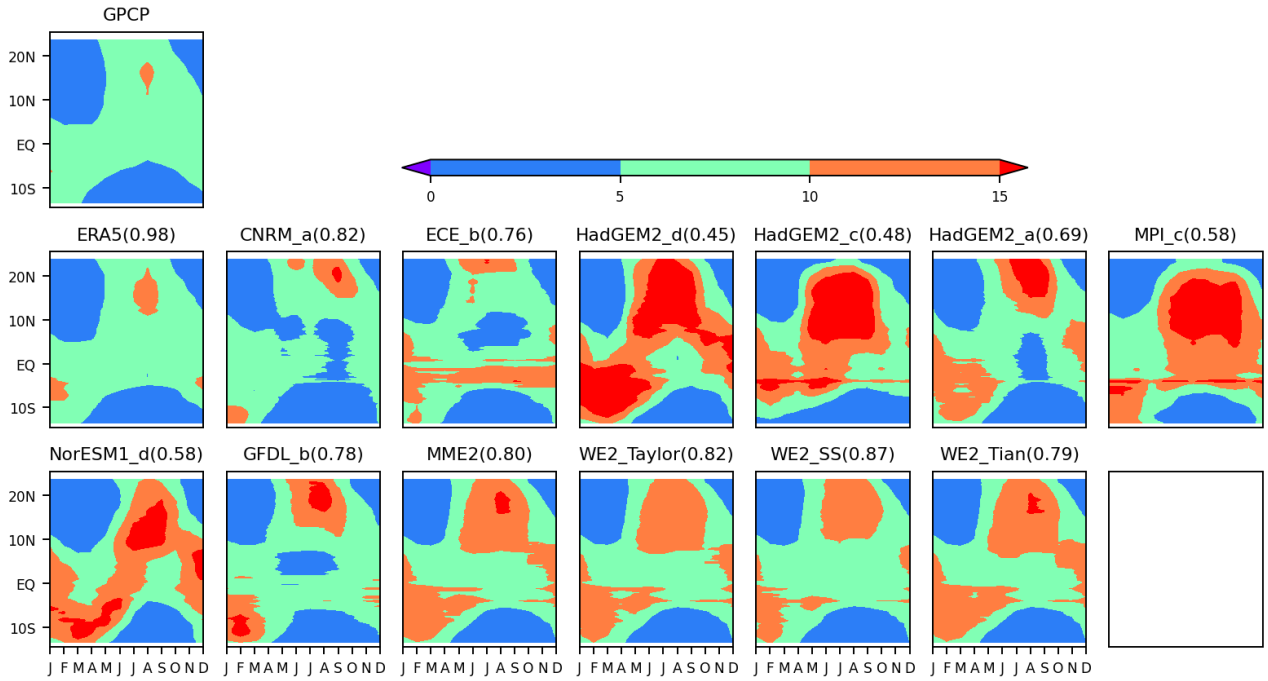

Figure A4. Taylor diagram of DJF (left) and JJA (right) season for new configuration (omitting IPSL), showing the performance of 8 models (A-H), ensemble mean (I), and weighted ensemble mean: Taylor (J), SS (K), and Tian (L).

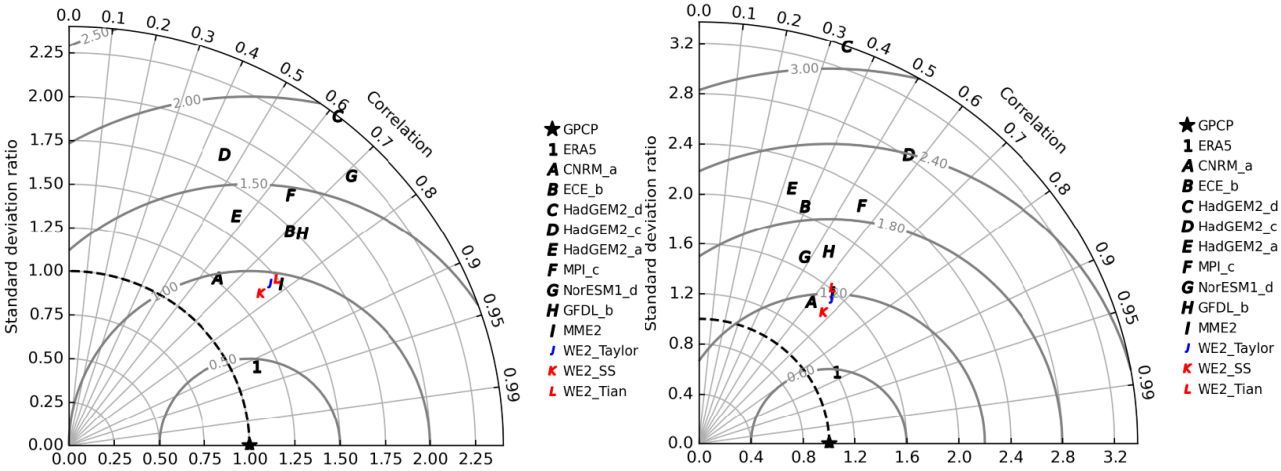

Figure A5. Wavelet profile of the SEA precipitation variability for new configuration (IPSL_b and PI weighting factor not included). The S value represents the pattern similarity with the reference (ERA5).

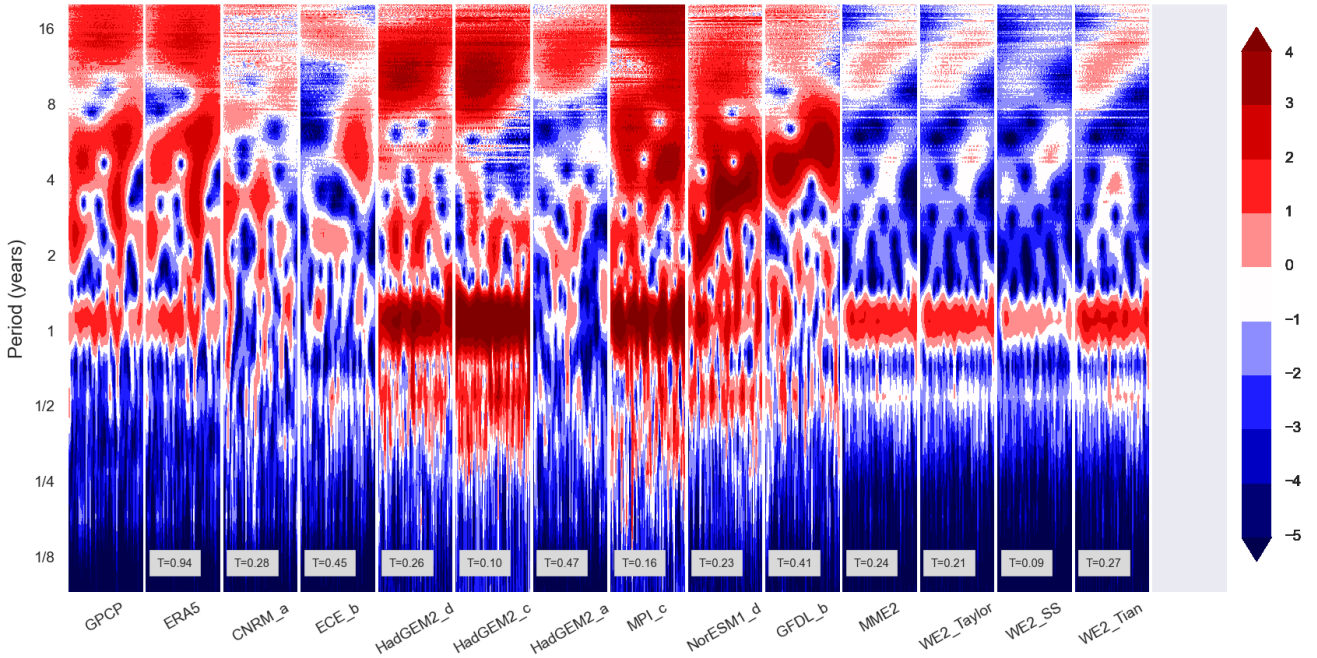

Figure A6. Plot of FFT for models (red) vs. reference (black) of the SEA precipitation variability. The value inside the bracket represents the pattern similarity with the reference (ERA5).