Selecting CMIP6 global climate models (GCMs) for Coordinated Regional Climate Downscaling Experiment (CORDEX) dynamical downscaling over Southeast Asia using a standardised benchmarking framework

Phuong Loan Nguyen[1,2]**, Lisa V. Alexander**[1,2]**, Marcus J. Thatcher**[3]**, Son C. H. Truong**[3]**, Rachael N. Isphording**[1,2]**, and John L. McGregor**[3]

[1]ARC Centre of Excellence of Climate Extremes and Climate Change Research Centre,
UNSW Sydney, Sydney, NSW 2052, Australia
[2]ARC Centre of Excellence for the Weather of the 21st Century, UNSW Sydney, Sydney, NSW 2052, Australia
[3]CSIRO Environment, Commonwealth Scientific and Industrial Research Organisation, Aspendale, Victoria, Australia

Correspondence: Phuong Loan Nguyen (phuongloan.nguyen@unsw.edu.au)

Abstract. Downscaling global climate models (GCMs) provides crucial high-resolution data needed for informed decision-making at regional scales. However, there is no uniform approach to select the most suitable GCMs. Over Southeast Asia (SEA), observations are sparse and have large uncertainties, complicating GCM selection especially for rainfall. To guide this selection, we apply a standardised benchmarking framework to select CMIP6 GCMs for dynamical downscaling over SEA, addressing current observational limitations. This framework identifies fit-for-purpose models through a two-step process: (a) selecting models that meet minimum performance requirements in simulating the fundamental characteristics of rainfall (e.g. bias, spatial pattern, annual cycle and trend) and (b) selecting models from (a) to further assess whether key precipitation drivers (monsoon) and teleconnections from modes of variability are captured, i.e. the El Niño–Southern Oscillation (ENSO) and Indian Ocean Dipole (IOD). GCMs generally exhibit wet biases, particularly over the complex terrain of the Maritime Continent. Evaluations from the first step identify 19 out of 32 GCMs that meet our minimum performance expectations in simulating rainfall. These models also consistently capture atmospheric circulations and teleconnections with modes of variability over the region but overestimate their strength. Ultimately, we identify eight GCMs meeting our performance expectations. There are obvious, high-performing GCMs from allied modelling groups, highlighting the dependency of the subset of models identified from the framework. Therefore, further tests of model independence, data availability and future climate change spread are conducted, resulting in a final subset of two independent models that align with our a priori expectations for downscaling over the Coordinated Regional Climate Downscaling Experiment –Southeast Asia (CORDEX-SEA).

1 Introduction

The Sixth Assessment Report (AR6) of the Intergovernmental Panel on Climate Change (IPCC) underscores with high confidence the escalating water-related risks, losses and damage associated with each increment of global warming (IPCC, 2023). The report specifically notes a projected increase in the frequency and intensity of heavy rainfall, leading to an increased risk of rain-generated localised flooding, particularly in coastal and low-lying cities and regions (Sect. 3, IPCC, 2023). Therefore, climate projections at regional scales are required to inform climate change adaptation strategies and enhance resilience efforts.

Different types of models have been developed and have become fundamental tools for assessing future regional cli-

mate changes, including state-of-the-art global climate models (GCMs) and regional climate models (RCMs). GCMs are generally used to explore climate interactions and underpin climate projections through the Coupled Model Intercomparison Project (CMIP; Meehl et al., 2000), an initiative of the World Climate Research Programme (WCRP). However, with a typical horizontal resolution of 50–250 km, GCMs have limited ability to simulate sub-grid weather (e.g. local variance, persistence, topography, etc.) and therefore cannot accurately define local-scale processes and feedbacks (e.g. deep convection, land–atmosphere interactions, etc.). This limits GCMs' ability to simulate aspects of the present-day water cycle and to determine robust future changes for local and regional applications (Maraun and Widmann, 2018; Douville et al., 2021). RCMs dynamically downscale GCM outputs to create higher spatial resolutions of ~ 2–50 km, providing richer regional spatial information (e.g. small-scale processes and extreme events) for climate assessments and for impact and adaptation studies (Diaconescu and Laprise, 2013; Giorgi and Gao, 2018). However, such experiments are computationally expensive, so it is not practical to choose all GCMs for dynamical downscaling. Thus, a subset of GCMs has to be selected.

The WCRP's Coordinated Regional Climate Downscaling Experiment (CORDEX) initiative delivers dynamically downscaled simulations of various GCMs (Giorgi and Gao, 2018) over 14 regions worldwide. This includes Phase I using CMIP5 (Giorgi et al., 2008) and Phase II, Coordinated Output for Regional Evaluations (CORDEX-CORE; Giorgi et al., 2021), as well as on-going experiments (CMIP6). However, there is no agreed upon approach to selecting which GCMs would be most suitable for dynamical downscaling, either in the recent WCRP guidelines for CMIP6 CORDEX experiments (CORDEX, 2021) or across different CORDEX domains (Di Virgilio et al., 2022; Grose et al., 2023; Sobolowski et al., 2023). In the earliest initiatives, GCMs were eliminated based on their skill at reproducing the current climate for the region of interest, given the fact that the bias in the GCMs can propagate into the RCM through the underlying and lateral boundary conditions (i.e. driven by initial and time-dependent meteorological variables from GCMs; Mote et al., 2011; Overland et al., 2011; McSweeney et al., 2012, 2015). In addition, the selection of GCMs considers the need to generate a reasonable uncertainty range for future climate projections (Mote et al., 2011; Overland et al., 2011). Given the shared physical components of the design of CMIP6 GCMs, there are inherent biases in statistical properties like the multi-model mean or standard deviation of the full ensemble (Boé, 2018; Brands, 2022; Sobolowski et al., 2023). To address this problem, model dependency is also considered. These considerations and methodologies have been integrated into the most recent CMIP6 CORDEX experimental design for specific regions, such as Europe (Sobolowski et al., 2023) or Australia (Di Vir-

gilio et al., 2022) and are recommended for widespread application across other CORDEX domains.

Model evaluation is an essential part of CMIP6 model selection since simulating past performance well is a necessary (but insufficient) condition to have more confidence in future performance. Different metrics are employed to quantify model skill in simulating various climate variables at either global (Kim et al., 2020; Ridder et al., 2021; Wang et al., 2021b; Donat et al., 2023) or regional scales, e.g. in Australia (Deng et al., 2021; Di Virgilio et al., 2022), Europe (Ossó et al., 2023; Palmer et al., 2023), South America (Díaz et al., 2021), Asia (Dong and Dong, 2021) and Southeast Asia (Desmet and Ngo-Duc, 2022; Pimonsree et al., 2023). However, the lack of consistency in the list of metrics used makes it difficult to perform one-to-one comparisons between studies or to track model performance across various regions.

Recently, Isphording et al. (2024) introduced a standardised benchmarking framework (BMF) underpinned by the work of the US DOE (2020), which included a set of baseline performance metrics for assessing model performance in simulating different characteristics of rainfall. The BMF is different from traditional model evaluation in that it defines performance expectations a priori (Abramowitz, 2005, 2012; Best, 2015; Nearing et al., 2018). Under the BMF, a model will not be considered fit-for-purpose if it fails any performance metric. The BMF consists of two tiers of metrics: the first tier includes minimum standard performance metrics related to fundamental characteristics of rainfall, and the second tier allows users to define metrics that help to answer specific scientific research questions. The BMF was initially designed for rainfall but can be widely applied to other climate variables (e.g. surface temperature) depending on the user's purpose (Isphording et al., 2024).

IPCC highlights Southeast Asia (SEA) as a region facing considerable climate change risks from extreme events (e.g. floods, extreme heat, changing precipitation and extremes; IPCC, 2022). However, available regional climate simulations for SEA, particularly from CMIP5 CORDEX-SEA experiments are limited to 13 simulations (Tangang et al., 2020) compared to EURO-CORDEX with 68 simulations (Jacob et al., 2020) or CORDEX-Australasia with 20 simulations (Evans et al., 2021). Consequently, future projections come with a higher degree of uncertainty, especially for rainfall (Tangang et al., 2020; Nguyen et al., 2023). This motivated the CORDEX-SEA community to update their regional climate model simulations with the latest CMIP6 models. Note that over SEA, observations are sparse with large uncertainties, particularly for rainfall (Nguyen et al., 2020), making GCM evaluations more complicated (Nguyen et al., 2022; Nguyen et al., 2023). To date, the performance of various CMIP6 GCMs has been evaluated and ranked over the whole region of SEA (Desmet and Ngo-Duc, 2022; Pimonsree et al., 2023) and its sub-regions, e.g. the Philippines (Ignacio-Reardon and Luo, 2023), Thailand (Kamworapan et al.,

2021) and Vietnam (Nguyen-Duy et al., 2023). Although there are groups of GCMs that consistently perform well (e.g. EC-Earth3, EC-Earth3-Veg, GFDL-ESM4, MPI-ESM1-2-HR, E3SM1-0 and CESM2) and poorly (e.g. FGOALS-g3, CanESM, NESM3 and IPSL-CM6A-LR) across the available literature, their ranking varies given inconsistencies in evaluation metrics and observational reference datasets. This creates challenges in conducting direct intercomparisons across the abovementioned studies. In addition, it is crucial to consider other important aspects discussed above (e.g. observational uncertainty, model dependency and future climate change spread) in identifying the list of reliable models over SEA.

In this research, we aim to apply the lessons learnt from CMIP6 selection over different CORDEX domains for SEA by assessing different aspects of models: model performance, model independence, data availability and future climate change spread. We apply the BMF to provide a consistent set of metrics for holistically evaluating model performance and to deal with large observational uncertainties over the region. Focusing on precipitation, where future projections are much more uncertain, the objectives of this research are twofold.

– We aim to evaluate the performance of CMIP6 GCMs in simulating the fundamental characteristics of precipitation, its drivers and teleconnection with modes of variability over SEA using a standardised benchmark framework and to identify a subset of models that meet our performance expectations.

– We aim to retain models that are relatively independent and are representative of the full range of possible projected change for finalising a subset of CMIP6 GCMs for dynamical downscaling over SEA using model independence tests and assessment of climate change response patterns.

The structure of the paper is as follows: Sect. 2 introduces the data and the benchmarking framework employed in this study. The results are presented in three subsections. Section 3.1 focuses on model assessment using the benchmarking framework, Sect. 3.2 examines the spread of future climate change among models and Sect. 3.3 assesses model dependence through cluster analysis. Finally, we conclude with a discussion of our results in Sect. 4 and a summary of the main conclusions in Sect. 5.

2 Methods

2.1 Data

2.1.1 CMIP6 GCM data

We use the historical daily data of precipitation, near-surface temperature, 850 hPa wind speed and both monthly and daily

sea-surface temperature data from the 32 CMIP6 models listed in Table 1. We consider only models that have a horizontal grid spacing finer than $2° \times 2°$, which are likely to be more suitable for dynamical downscaling. One simulation (typically the first member r1i1f1p1) is utilised in the benchmarking process to enable a fair comparison. At the time of this analysis, the first member of some models (e.g. the CNRM-family models, UKESM1-0-LL and HadGEM3-GC31-MM) was not available, so another member was utilised.

2.1.2 Observations and reanalyses

Given the large observational uncertainty in precipitation over the region (Nguyen et al., 2022), we use multiple daily observed datasets from different in situ and satellite sources to quantify model skill (Table 2). These datasets have been chosen given their high consistency in representing daily precipitation (Nguyen et al., 2022) and extremes (Alexander et al., 2020; Nguyen et al., 2020) over SEA.

The ERA5 reanalysis (~ 31 km grid resolution; Hersbach et al., 2020) was used to benchmark model performance in representing the climatology of atmospheric circulation (e.g. metrics related to horizontal wind at 850 hPa level are described in Sect. 2.2).

We acknowledge that different observational sea-surface temperatures (SSTs) have different abilities to capture signals of the modes of variability. Therefore, we utilise multiple SST products (Table 2) to take account of the observational uncertainties in simulating the teleconnection between rainfall and the main modes of variability, including the El Niño–Southern Oscillation (ENSO) and Indian Ocean Dipole (IOD) as described in Sect. 2.2.

2.2 Benchmarking CMIP6 GCMs over Southeast Asia

Given the large uncertainties and model inconsistency in rainfall projections, our main aim is to identify a subset of CMIP6 GCMs that meet our a priori expectations. That is, as a minimum requirement, a model should simulate past rainfall statistics over SEA reasonably well using consistent criteria. Figure 1 illustrates the GCM selection process applied in this research based on a standardised benchmarking framework (Isphording et al., 2024). A subset of CMIP6 GCMs that meet our model performance expectations are identified through a two-step process: (a) selecting models that meet minimum performance requirements in simulating the fundamental characteristics of rainfall (Fig. 1) and (b) selecting models from (a) to further assess performance in simulating precipitation drivers (e.g. monsoon) and teleconnections with modes of variability (Fig. 1).

2.2.1 Minimum standard metrics

The BMF introduces a set of minimum-standard metrics (MSMs): (1) mean absolute percentage error (MAPE),

Table 1. Information on model components from the CMIP6 GCMs used in this study.

No	Model	Run	Atmosphere long/lat	Reference	Atmospheric component	Land component	Sea ice component	Ocean component
1	ACCESS-CM2*	r1i1p1f1	$1.2° \times 1.8°$	Bi et al. (2020); Ziehn et al. (2020)	UKMO UM v10.6	CABLE 2.5	LANL CICE5.1	MOM5
2	ACCESS-ESM1-5*	r1i1p1f1	$1.2° \times 1.8°$		UKMO UM V7.3	CABLE2.4	LANL CICE4.1	GFDL MOM5
3	BCC-CSM2-MR*	r1i1p1f1	$1.1° \times 1.1°$	Wu et al. (2019)	BCC-AGCM3	BCC-AVIM2	SIS4	MOM4-L40
4	CESM2*	r1i1p1f1	$0.95° \times 1.25°$	Danabasoglu et al. (2020)	CAM6/WACCM6	CLM5.0	CICE5	POP2
5	CMCC-CM2-HR4	r1i1p1f1	$0.95° \times 1.25°$	Cherchi et al. (2019)	CAM v5	CLM4.5	CICE4	NEMO v3.6
6	CMCC-CM2-SR5*	r1i1p1f1	$0.9° \times 0.9°$					
7	CMCC-ESM2*	r1i1p1f1	$0.95° \times 1.25°$					
8	CNRM-CM6-1*	r1i1p1f2	$1.4° \times 1.4°$	Voldoire et al. (2019)	ARPPE-Climat v6.3	Flake	OASIS-MCT	NEMO
9	CNRM-CM6-1-HR	r1i1p1f2	$0.5° \times 0.5°$					
10	CNRM-ESM2-1*	r1i1p1f2	$1.4° \times 1.4°$					
11	E3SM-1-0*	r1i1p1f1	$1° \times 1°$	Zheng et al. (2022)	EAM (CAM 5.3)	MPAS-Ocean	MPAS-Seaice	ELMv0 (CLM4.5)
12	EC-Earth3-AerChem	r1i1p1f1	$0.7° \times 0.9°$	Döscher et al. (2022)	ECMWF IFS	LPJ-GUESS TSI	LIM3	NEMO v3.6
13	EC-Earth3-CC	r1i1p1f1	$0.7° \times 0.9°$					
14	EC-Earth3*	r1i1p1f1	$0.7° \times 0.7°$					
15	EC-Earth3-Veg*	r1i1p1f1	$0.7° \times 0.7°$					
16	EC-Earth3-Veg-LR	r1i1p1f1	$1.125° \times 1.125°$					
17	GFDL-CM4	r1i1p1f1	$1.0° \times 1.3°$	Held et al. (2019); Dunne et al. (2020)	AM4	LM4	SIS 2	OM4 MOM6
18	GFDL-ESM4*	r1i1p1f1	$1.0° \times 1.3°$					
19	HadGEM3-GC31-MM	r1i1p1f3	$0.9° \times 0.9°$	Andrews et al. (2020)	GA7/GL7		GSI8.1 (CICE5.1)	GO6 (NEMO)
20	INM-CM4-8	r1i1p1f1	$1.5° \times 2.0°$	Volodin et al. (2017)	INM-AM4-8/5.0	INM-LND1	INM-ICE1	INM-OM5
21	INM-CM5-0	r1i1p1f1	$1.5° \times 2.0°$					
22	IPSL-CM6A-LR*	r1i1p1f1	$1.3° \times 2.5°$	Boucher et al. (2020)	LMDZ 6A-LR	ORCHIDE2.0	NEMO-LIM3.6	NEMO 3.6
23	IPSL-CM6A-LR-INCA*	r1i1p1f1	$1.27° \times 2.5°$					
24	MIROC6*	r1i1p1f1	$1.4° \times 1.4°$	Tatebe et al. (2019)	MIROC 3.2	MATSIRO	MIROC 3.2	COCO 4.5
25	MPI-ESM1-2-HR*	r1i1p1f1	$0.94° \times 0.94°$	Mauritsen et al. (2019)	ECHAM6.3	JSBACH		MPIOM
26	MPI-ESM1-2-LR*	r1i1p1f1	$1.875° \times 1.875°$					
27	MRI-ESM2-0*	r1i1p1f1	$1.1° \times 1.1°$	Yukimoto et al. (2019)	MRI-AGCM3.5			MRI.COMv4
28	NESM3	r1i1p1f1	$1.9° \times 1.9°$	Cao et al. (2018)	ECHAM6.3	JSBACH	CICE4	NEMO v3.4
29	NorESM2-MM*	r1i1p1f1	$0.9° \times 9°$	Seland et al. (2020)	CAM4-Oslo	CLM4	CICE4	MICOM
30	SAM0-UNICON*	r1i1p1f1	$0.9° \times 1.3°$	Park et al. (2019)	CAM5.3 with UNICON	CLM4	CICE4.0	POP2
31	TaiESM1*	r1i1p1f1	$0.9° \times 0.9°$	Wang et al. (2021a)	Tai AM1	CLM4.0	CICE4	POP2
32	UKESM1-0-LL*	r1i1p1f2	$1.3° \times 1.9°$	Sellar et al. (2019)	MetUM-HadGEM3-GA7.1	JULES-ES-1.0	CICE-HadGEM3-GSI8	NEMO-HadGEM3-GO6.0

* The model offers atmospheric variables available in three dimensions every 6 h for dynamical downscaling at the time of analyses.

(2) spatial correlation (Scor), (3) seasonal cycle (Scyc) and (4) significant changes (SigT) (Isphording et al., 2024) to assess the skill of climate models in simulating fundamental characteristics of precipitation (e.g. magnitude of biases, spatial distributions, annual cycles and temporal variability). Before exploring complex processes, a model should meet performance expectations for these MSMs. Therefore, we initially calculate the MSMs for precipitation. In addition, we acknowledge that models should produce adequate present-day simulations of other fundamental climate variables like near-surface temperature. Hence, we also apply the MSMs for near-surface temperature in the Supplement. Given the strong seasonality of precipitation in the region (Juneng et al., 2016), the analyses related to precipitation

Table 2. The main characteristics of observational datasets used in this study.

Type of dataset	Product short name	Dataset name	Temporal coverage	Spatial resolution	Data source	Reference
Precipitation dataset	APHRODITE	APHRODITE V1101 and V1101XR	1950–2015	$0.5° \times 0.5°$	In situ	Yatagai et al. (2012)
	CHIRPv2	CHIRPSv2	1981–2016	$0.25° \times 0.25°$	In situ and satellite	Funk et al. (2015)
	REGEN_ALL	REGEN Allstns V1 2019	1950–2019	$1° \times 1°$	In situ	Contractor et al. (2020)
	GPCC_v2018	GPCC FDD v2018	1982–2019	$1° \times 1°$	In situ	Schamm et al. (2014)
Sea-surface temperature dataset	HadISST	HadISST1 v1	1870–2021	$1° \times 1°$	In situ and satellite	Rayner et al. (2003)
	OISST	OISST v2.0	1981–2020	$0.25° \times 0.25°$	In situ and satellite	Huang et al. (2021)
	ERSST	ERSST v5	1854–2024	$2° \times 2°$	In situ	Huang et al. (2017)
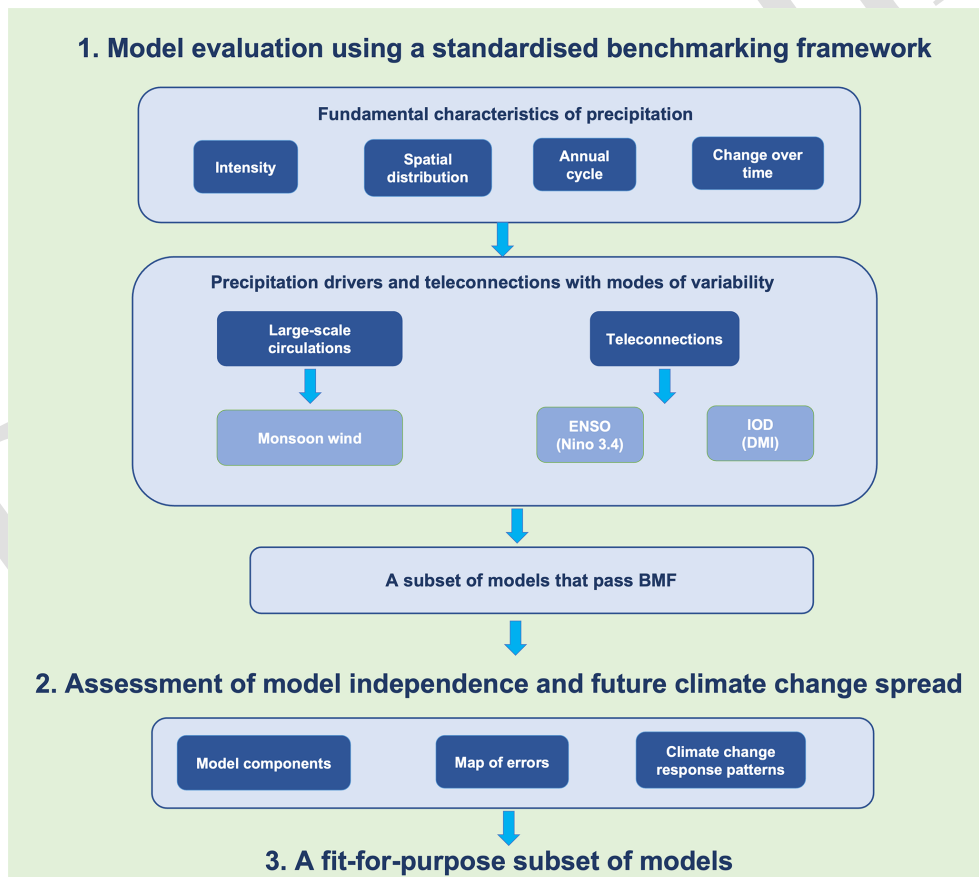
Figure 1. A schematic of the CMIP6 GCM selection process, including (1) model evaluation using a standardised benchmarking framework (BMF) and (2) assessment of model independence and future climate change spread. The BMF includes two steps: minimum standard metrics (MSMs) that assess very basic characteristics of rainfall and second-tier metrics (e.g. versatility metrics) that quantify the model skill of the models that pass the MSMs in simulating precipitation drivers (monsoon) and teleconnections with modes of variability, i.e. the El Niño–Southern Oscillation (ENSO) and Indian Ocean Dipole (IOD).

are conducted at a seasonal scale (e.g. the dry season from November–April – NDJFMA – and the wet season from May–October – MJJASO). Meanwhile, temperature analyses are conducted at the annual scale.

Note that in this research, we focus only on precipitation over land given the lack of in situ references over the ocean. Some satellite-derived products provide oceanic precipitation data, but most of their temporal coverage is not sufficiently long to use as a reference. In addition, the observational uncertainties among satellite clusters in estimating oceanic precipitation over SEA are quite substantial, with discrepancies reaching up to $4\,\mathrm{mm\,d^{-1}}$ (Fig. S1 in the Supplement).

2.2.2 Versatility metrics

The MSMs provide statistical measurements that are not always correlated with future projections (Knutti et al., 2010), given that some models may simulate historical precipitation well for the wrong reasons. A further recommendation is therefore to also assess model performance based on key physical processes (Doe, 2020; Nguyen et al., 2023). This approach offers additional insights into the relative roles of model biases at simulating large-scale environments versus the limitations of model parameterisations in generating precipitation biases. Therefore, we define second-tier versatility metrics to assess the GCMs selected from Sect. 2.2.1 in simulating the complex precipitation-related processes, including drivers and teleconnections with modes of variability.

Monsoon circulation

SEA is situated within the Asian monsoon regime, where atmospheric circulation is modulated by two primary monsoon systems: the Indian monsoon characterised by westerlies from the Bay of Bengal into northern parts of SEA, including the mainland and northern Philippines (along $10°\,\mathrm{N}$) during the boreal summer (JJAS) and reversed in direction during the boreal winter (DJF), and the Australian monsoon, e.g. easterlies from Australia to the Maritime Continent (MC) and Papua New Guinea (Chang et al., 2005). These monsoon systems drive regional rainfall seasonality. Therefore, we focus on assessing model skill in simulating the intensity and direction of monsoon wind (e.g. 850 hPa wind) for JJAS and DJF. While wind speed is evaluated using the MAPE and Scor metrics similar to the MSMs for precipitation and temperature, wind direction is quantified using an equation from Desmet and Ngo-Duc (2022):

$$\mathrm{MD} = \frac{\sum_i u_i \times |\theta_i - \theta_{i,\mathrm{ref}}|_{[0,180]}}{\sum_i u_i},$$

where u_i refers to the simulated wind speed at the grid i and θ_i and $\theta_{i,\mathrm{ref}}$ are the wind direction at grid i in the simulated and reference data, respectively. $|\theta_i - \theta_{i,\mathrm{ref}}|_{[0,180]}$ is the absolute value of the difference at the ith grid between directions

of simulated and reference wind speed (e.g. ERA5). The MD metric allows us to quantify the agreement in wind direction between two datasets in which the impact of high wind speed is taken into account.

ENSO, IOD and teleconnections

Various parts of SEA are also affected by two prominent modes of variability: the El Niño–Southern Oscillation (ENSO; Haylock and McBride, 2001; Chang et al., 2005; Juneng and Tangang, 2005; Qian et al., 2013) and the Indian Ocean Dipole (IOD; Xu et al., 2021) via atmospheric teleconnections. In this research, the teleconnection is defined by the temporal correlation between precipitation anomalies at each grid point and the ENSO–IOD indices.

To track ENSO variability, the Niño3.4 index ($5°\,\mathrm{S}$–$5°\,\mathrm{N}$ and $160°\,\mathrm{E}$–$120°\,\mathrm{W}$) (Trenberth and Hoar, 1997; Shukla et al., 2011) derived for the 1951–2014 period as area-mean monthly SST anomalies with respect to a 1961–1990 climatology is used. For IOD, we use the dipole mode index (DMI; Saji et al., 1999; Meyers et al., 2007). DMI measures differences in monthly SST anomalies between those in the west equatorial Indian Ocean ($10°\,\mathrm{S}$–$10°\,\mathrm{N}$, 50–$70°\,\mathrm{E}$) and those in the east (90–$110°\,\mathrm{S}$, $10°\,\mathrm{S}$–$0°\,\mathrm{N}$).

We use a 5-monthly average Niño3.4 and IOD index to remove seasonal cycles. The resulting monthly time series are detrended using a fourth-order polynomial fit to remove the possible influence of a long-term trend and to better preserve high-amplitude (<10 years) variability (Braganza et al., 2003).

Since ENSO typically matures toward the end of the calendar year (Rasmusson and Wallace, 1983 TS2), we consider ENSO developing years as year (0) and use the DJF means to identify ENSO events. Over SEA, ENSO interacts with the monsoon cycle, and due to the varying monsoon onsets between the northern and southern parts of the region, its seasonal evolution differs across regions (Fig. S2). In particular, there is a lagged negative correlation between rainfall and ENSO over the Maritime Continent (MC) and the Philippines, which develops from May to June, strengthens during July–August and reaches its highest correlation during September–October of the developing year (year 0). On the other hand, this negative correlation becomes prominent over the northern parts during the subsequent boreal spring (from March to May of the year $+1$; Wang et al., 2020; Chen et al., 2023). The negative correlation indicates dry anomalies during El Niño and/or wet anomalies during La Niña. Therefore, in the context of this research, we examine the lead–lag Pearson correlation of the DJF Niño3.4 index in the developing year (year 0) with two different seasonal rainfall patterns: May–October (MJJASO) of the developing year (year 0) and March–May (MAM) of the following year (year $+1$).

Furthermore, considering the stronger influence of the IOD and its associated teleconnection during SON compared to other seasons (McKenna et al., 2020), we calculated the in-

phase Pearson correlation coefficient between the detrended precipitation anomaly and DMI for the SON season. The statistical significance of the correlation coefficient is tested using the Student t test (alpha $= 0.05$). Note that IOD could exist as part of ENSO (Allan et al., 2001; Baquero-Bernal et al., 2002), and their coexistence could have strong impacts on rainfall variability over many parts of SEA (D'Arrigo and Wilson, 2008; Amirudin et al., 2020), which is not investigated in this study.

The previous literature has often focused on assessing the robustness of rainfall teleconnections (e.g. spatial patterns and amplitudes) across CMIP model ensembles. These assessments typically involve examining agreement in the sign of teleconnections such as through rainfall anomaly composites (Langenbrunner and Neelin, 2013) and regional average teleconnection strength over land (Perry et al., 2020) or a combination of both (Power and Delage, 2018) rather than evaluating the skill of an individual model. However, since rainfall teleconnections across SEA exhibit spatial and seasonal variability, the above metrics may be substantially influenced by internal variability. For high-level qualification, we employ spatial correlation and simplified metrics to assess whether there are significant correlations between teleconnections, as recommended by Liu et al. (2024). We assess the similarity in the number of grid points detecting significant signals between observed and modelled teleconnections using a set of three metrics: hit rate (HR), miss rate (MR) and false-alarm rate (FAR) as follows:

$$HR = \frac{\text{Area with correct sign of significant correlation}}{\text{Area with significant correlation in OBS}} \quad (1)$$

$$MR = \frac{\begin{array}{c}\text{Area with significant correlation in OBS but}\\ \text{with no significant correlation in model}\end{array}}{\text{Area with significant correlation in OBS}} \quad (2)$$

$$FAR = \frac{\begin{array}{c}\text{Area with no significant correlation in OBS}\\ \text{but with significant correlation in model}\end{array}}{\text{Area with no significant correlation in OBS}}. \quad (3)$$

These metrics allow us to make sure that the model adequately simulates significant signals across the entire region. While HR ranges from 0 to 1, MR and FAR vary. A desirable model outcome includes a high HR value coupled with low MR and FAR values, indicating the model's ability to adequately capture the significance of the correct signal in the right region (on grid scales) of teleconnections between ENSO and IOD and rainfall pattern.

2.3 GCM independence assessment and future climate change spread

Model independence could be assessed based on model components (e.g. shared atmospheric, land and/or ocean models) and/or model output patterns. In this study, we employ both methods for testing GCM independence. Table 1 provides information on the principal components of the models used in this study. Note that model independence based on this criterion could depend on the model version (e.g. the same model with different levels of complexity). In addition, we acknowledge that the spatial pattern of error maps and future change maps seems to correlate well with model dependency (Knutti et al., 2010; Knutti and Sedláček, 2013; Brunner et al., 2020; Brands, 2022). Therefore, we determine the independence of GCMs simply by calculating the correlation coefficient of historical biases and future projections between models and then apply a hierarchical clustering approach (Rousseeuw, 1987) from this correlation matrix to group models. This cluster analysis has been employed in the previous literature for multiple purposes, e.g. to assess model dependency (Brunner et al., 2020; Masson and Knutti, 2011), spatial patterns of climatology and trends in climate extremes (Gibson et al., 2017), or spatial patterns of precipitation change signals (Gibson et al., 2024).

Note that historical biases are calculated by comparing the climatology of total rainfall over the land area of SEA for the 1951–2014 period to corresponding data from an observed reference. Meanwhile, for future signals, we focus on the relative change (in percentage) between the far future (2070–2099) and the baseline (1961–1990), as suggested by the World Meteorological Organization (WMO). All analyses are conducted for two seasonal periods: the wet MJJASO and dry NDJFMA seasons.

We use the coarsest resolution (i.e. NESM ~ 216 km or $1.9° \times 1.9°$ resolution) among the 32 GCMs as the target resolution for comparison. All data are interpolated into a spatial resolution of $1.9° \times 1.9°$ using a first-order conservative regridding method (Jones, 1999) to better capture the spatial discontinuity of precipitation (Contractor et al., 2018).

Benchmarking CMIP6 GCMs against observations is conducted over land for precipitation and the teleconnections between precipitation and modes of variability, while 850 hPa winds from ERA5 allow the comparison to also be extended over the ocean.

Hereafter, we select APHRODITE as the primary baseline for all the main figures, as it utilises the greatest number of rain gauges of any dataset. We include the results related to all other observational datasets in the Supplement (Figs. S3–S8) and provide a detailed explanation of related results in the main text for intercomparison purposes.

3 Results

3.1 Minimum standard metrics (MSMs)

3.1.1 Mean absolute percentage error (MAPE) and spatial correlation (Scor)

We initially assess the performance of CMIP6 GCMs in reproducing the spatial distribution of precipitation using the first two MSMs: MAPE and Scor. Previous studies have emphasised strong seasonal and regional contrasts in rain-
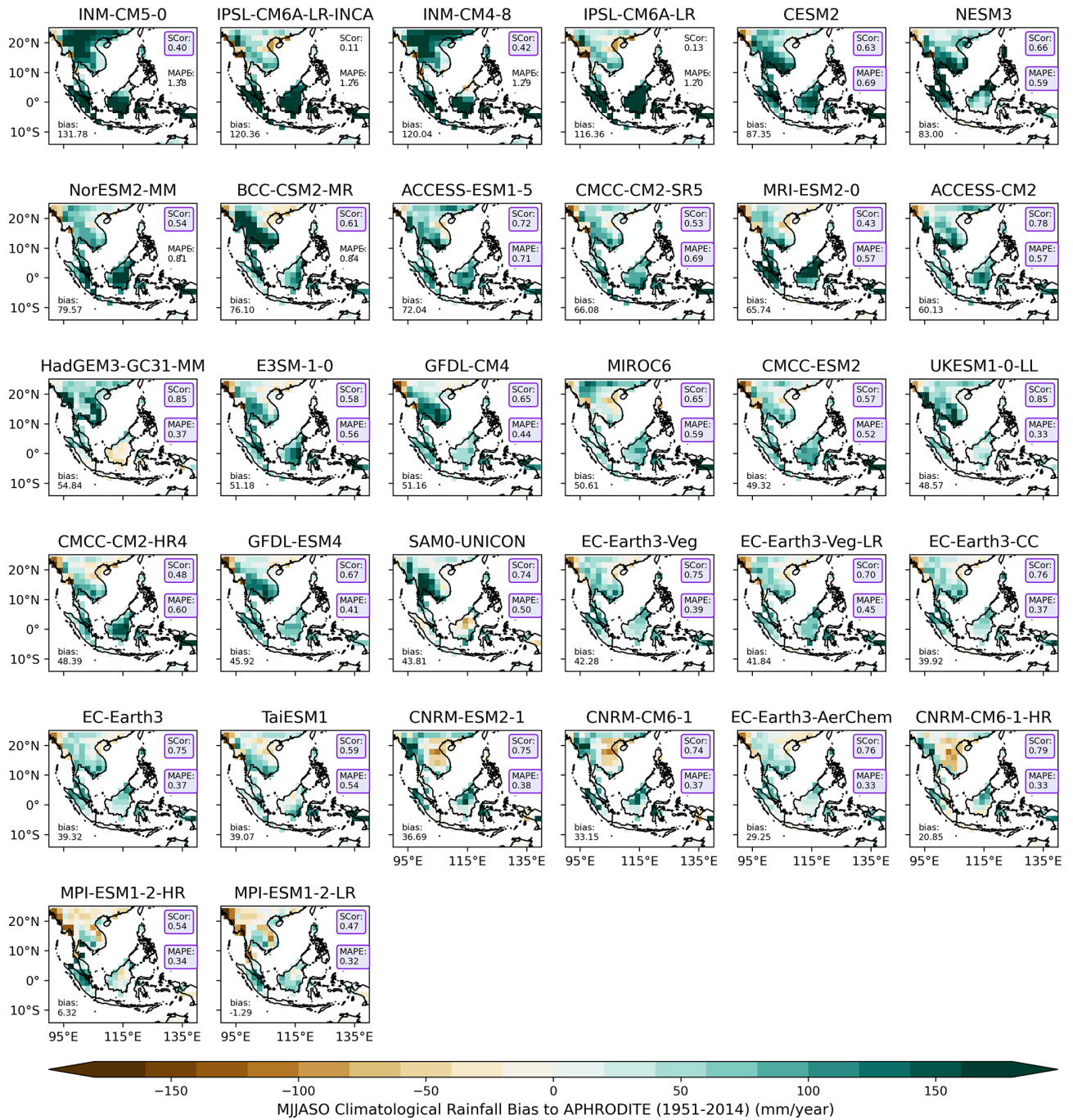
Figure 2. The seasonal climatological (1951–2014) bias (in mm yr^{-1}) for each model against the APHRODITE observational product during the wet season (May–October; MJJASO), ranked from wettest to driest based on regionally averaged bias. The mean absolute percentage errors (MAPEs) and spatial correlations (Scor's CE1) calculated against APHRODITE are shown in the upper-right corner. Values highlighted in purple-coloured boxes indicate values that meet our defined benchmarking thresholds. All analyses are considered at the resolution of the coarsest CMIP6 GCM (i.e. NESM3, ∼ 216 km).
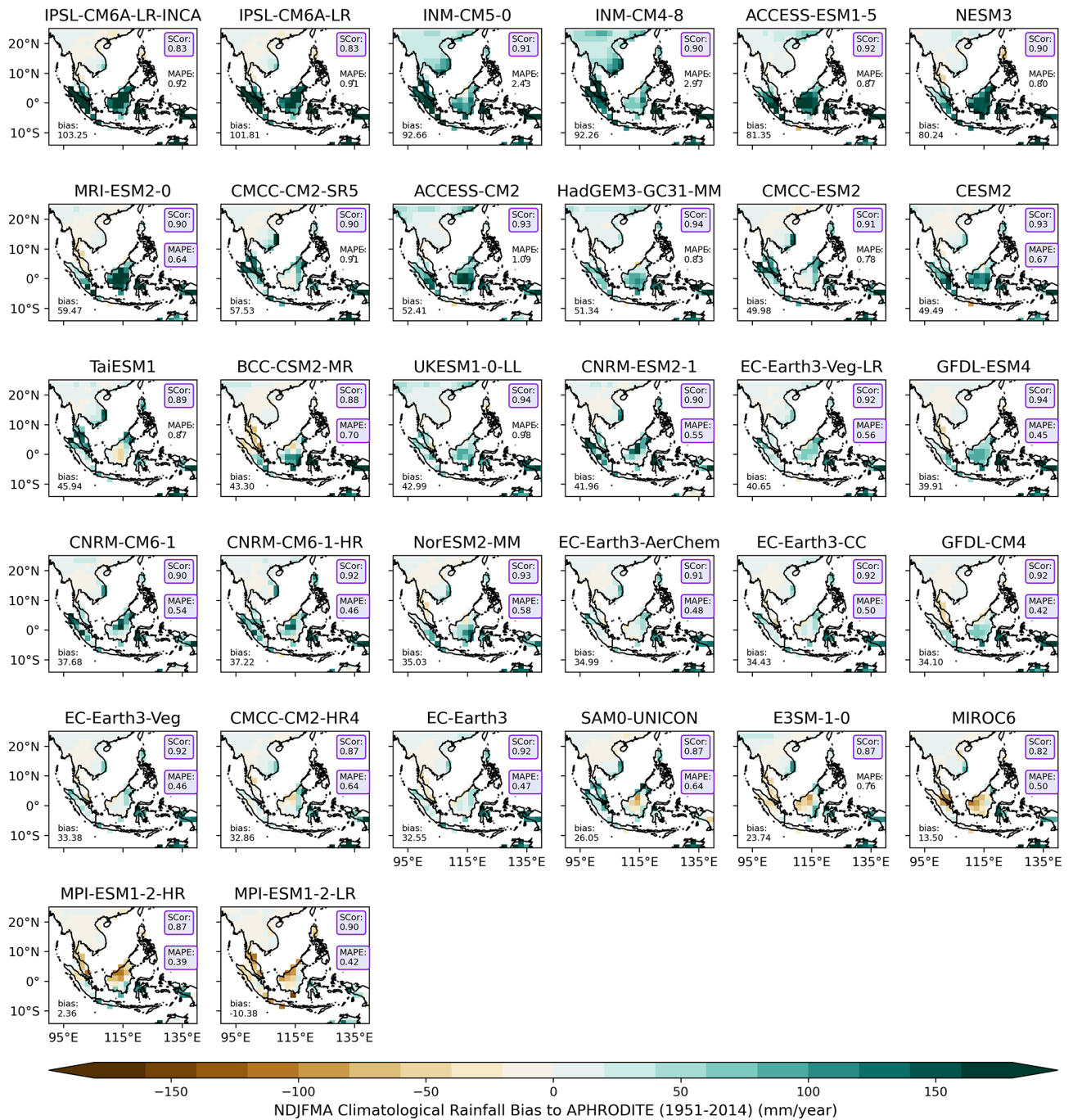
Figure 3. Same as Fig. 2 but for the dry season (November–April; NDJFMA). TS3

fall distribution over Southeast Asia (Nguyen et al., 2023). Therefore, we focus on comparing the seasonal climatology (1951–2014) of total rainfall during wet days (e.g. precipitation $\geq 1\,\text{mm}$) between models and APHRODITE for both the wet MJJASO and dry NDJFMA seasons (Figs. 2 and 3, respectively). For MSMs, our strategy is to retain as many models as possible. We establish benchmarking thresholds based on the requirements of downscaling CMIP6 from CORDEX communities and our understanding of reasonable model performance based on current scientific understanding. In particular, GCMs should adequately reproduce the spatial distribution of rainfall without a strong wet or dry bias. In addition, we also identify observational uncertainties through intercomparison of multiple precipitation datasets.

Considering variations in model performance across seasons, we also set different thresholds for benchmarking models for different seasons. In particular, due to the better ability of the models to capture spatial variability in precipitation during the dry season compared to the wet season (Desmet and Ngo-Duc, 2022), we adopt a more lenient approach by relaxing our expectation for a spatial distribution metric, setting the Scor threshold ≥ 0.4 for the wet season and ≥ 0.75 for the dry season. However, for the MAPE score, we apply a stricter criterion, as we require models to closely simulate observed rainfall intensity over SEA. For both wet and dry seasons, we set the benchmarking threshold for MAPE at ≤ 0.75. With this threshold, our objective is to identify models capable of capturing the spatial variability in rainfall across at least 40 % (Scor ≥ 0.4) or 75 % (Scor ≥ 0.75) of the domain during wet and dry seasons, respectively, with a wet/dry bias of no more than 75 % compared to observations (MAPE ≤ 0.75) for both seasons.

We first discuss key features of the wet season (MJJASO; Fig. 2). Models are ranked from wettest to driest based on their regionally averaged climatologies (i.e. the average of accumulated precipitation over all land grid points inside the domain). Models that meet our benchmarking thresholds for MAPE and Scor (i.e. calculated against APHRODITE) are highlighted by purple-coloured boxes. In general, CMIP6 GCMs demonstrate a wet bias in terms of regional averages ranging from 6.32 to 131.78 mm yr^{-1}, except for MPI-ESM1-2-LR (-1.29 mm yr^{-1}). However, there is spatial variability in the distribution of wet and dry biases. While most of these models consistently show wet biases over the MC, dry biases are observed in different locations on the mainland across models (e.g. along the west coast in the EC-Earth, IPSL and CMCC families or on the east coast in the CNRM family, as well as in some northern regions such as the MPI family). Among the wettest GCMs, including INM, IPSL, NorESM2-MM and the CESM2 family, the largest biases are predominantly over the MC. Interestingly, most CMIP6 GCMs can capture the spatial variability in rainfall (Scor around or greater than 0.5), except for the IPSL-family simulations (Scor's of 0.11 and 0.13). Using the threshold definitions mentioned above, six models fail to meet these benchmarks, exhibiting obvious grouping by GCM group. For example, IPSL-CM6A-LR and IPSL-CM6A-LR-INCA fail due to their low Scor's (0.13 and 0.11, respectively) and high MAPEs (1.20 and 1.26, respectively). While the INM-CM5-0 and INM-CM4-8 models meet our set expectation in relation to spatial variability, they fail to meet the MAPE threshold due to their overestimation of rainfall across the entire region (e.g. MAPEs ranging from 1.29 to 1.38, respectively). All failed models mentioned exhibit high MAPE values ranging from 0.81 to 1.28.

The corresponding results for the dry season reveal some interesting features (Fig. 3). First, there are substantial similarities in the spatial distribution of climatological rainfall biases across models during this season. CMIP6 GCMs consistently show small biases over Indochina and large wet biases over the MC. A better spatial correlation with observations (i.e. Scor > 0.8) is obtained during the dry season, consistent with previous findings, e.g. CORDEX-CMIP5 RCMs (Nguyen et al., 2022) or CMIP6 GCMs (Desmet and Ngo-Duc, 2022), in highlighting the dependence of model performance on the season. With improved performance in capturing the spatial variation in total precipitation intensity compared to the wet season, all models meet our expected performance in spatial variability. However, INM- and IPSL-family models still fail the MAPE criterion since they exhibit much-higher precipitation intensity than APHRODITE, particularly over the MC. Note that over SEA, APHRODITE is drier than other precipitation products, particularly over the MC (Nguyen et al., 2020).

It is important to note that whether a model passes or fails the benchmarking is strongly dependent on the choice of threshold, as emphasised in Isphording et al. (2024). For instance, more simulations would fail this test if we set a higher threshold for Scor, notably for the MJJASO season case.

3.1.2 Seasonal cycle

In this section, we follow the simplified method developed by Isphording et al. (2024) in quantifying the phase and structure of the seasonal cycle. In particular, we rank total monthly precipitation from wettest to driest. We then define the benchmarking threshold such that the 4 wettest and driest observed months must fall within the 6 wettest and driest months simulated by models (Fig. 4).

Overall, most CMIP6 GCMs reproduce the phase well but tend to overestimate precipitation intensity, notably for the observed precipitation peaks during boreal summer (Fig. S3). The INM- and IPSL-family simulations stand out, consistent with the wettest biases observed in spatial patterns (Sect. 3.1.1).

According to the benchmarking threshold definitions, all models meet the benchmark for simulating the 4 wettest observed months. However, six models do not pass the benchmark for simulating the 4 driest observed months, as highlighted in orange in Fig. 4. Specifically, 1 of the 4 driest months according to the APHRODITE dataset (December through March) is ranked as the 6th wettest month (ranked 7th in Fig. 4) by these models.

3.1.3 Significant trend

The final MSM aims to explore how rainfall changes over time (Isphording et al., 2024). In this part, we compare the signal of statistically significant simulated and observed trends using the wet (Fig. 5) and dry (Fig. 6) season accumulated precipitation. A Theil–Sen trend is calculated over a 65-year period (1951–2014) and tested at a 5 % significance level using a Mann–Kendall significant test (Kendall, 1975).
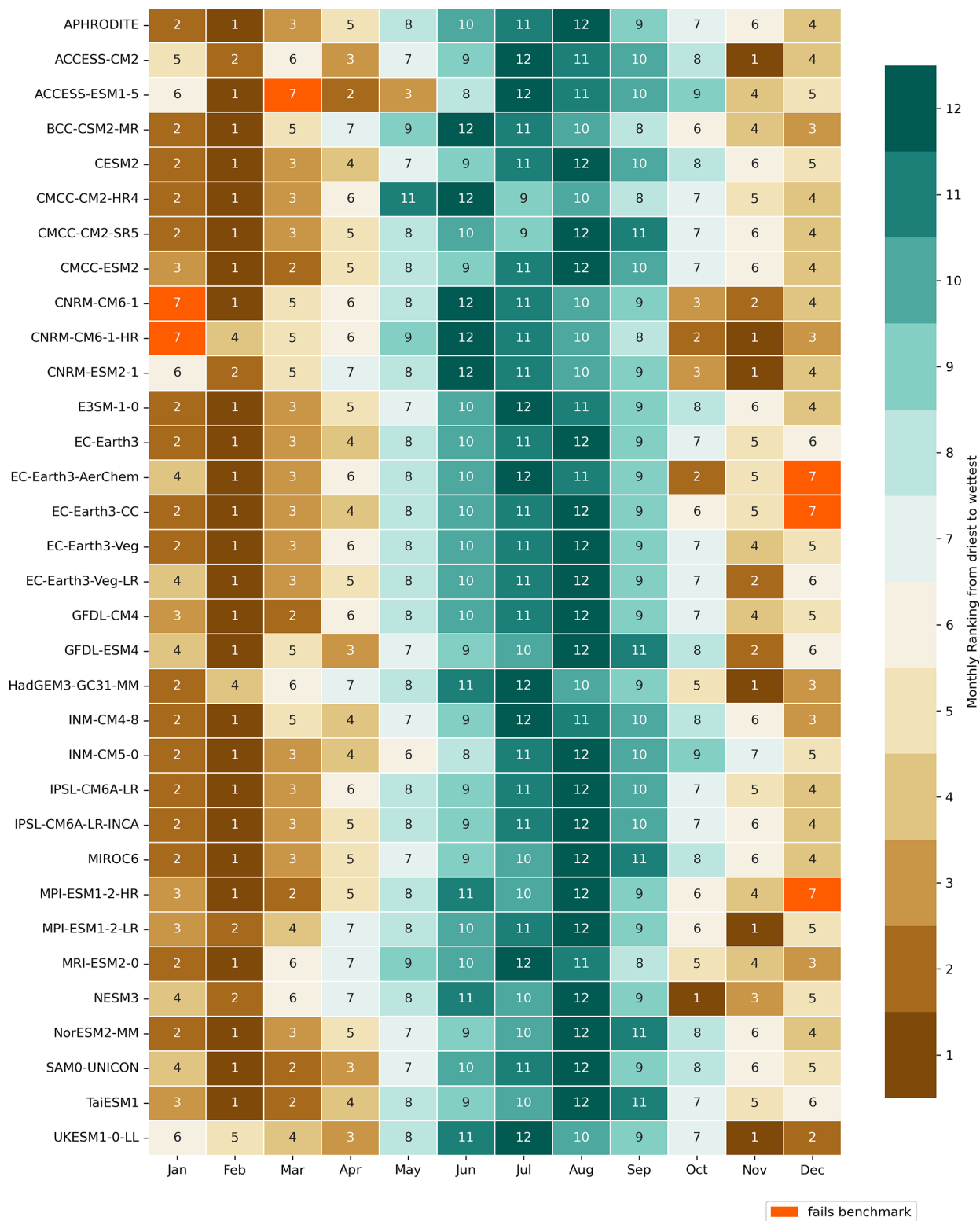
Figure 4. The climatological (1951–2014) average total monthly rainfall over mainland Southeast Asia ranked from driest to wettest for each CMIP6 simulation. Brown shades (1–6) indicate the 6 driest months while teal colours (7–12) indicate the 6 wettest months. The models that failed the benchmarking are highlighted in orange. All analyses are considered at the coarsest CMIP6 GCM resolution (i.e. NESM3, ∼ 216 km).

Figure 5. The observed (top row) and modelled seasonal average total precipitation across Southeast Asia land areas during the wet season (May–October; MJJASO) for the period 1951–2014. The direction of the observed Theil–Sen trend is the benchmark (top row). The Theil–Sen trend line for each of the simulations is plotted in grey if the models fail the benchmark and in purple if they pass. The magnitude of the trend is noted in the top middle, and the results of the Mann–Kendall significance test are noted in the bottom-right corner. Models are sorted based on the magnitude of the spatial average to match the order of Fig. 2. All analyses are considered at the coarsest CMIP6 GCM resolution (i.e. NESM3, ∼ 216 km). All models pass the benchmark.

Figure 6. Same as Fig. 5 but for the boreal dry season (November–April; NDJFMA).

There is a significant decreasing trend in observed total precipitation during the wet season (Fig. 5 – top panel), while the dry season has a significant increasing trend (Fig. 6 – top panel). A model fails this benchmark if it exhibits an opposite significant trend to that of the observations. Using this definition, all models pass this benchmark during the wet season, but MRI-ESM2-0 and MPI-ESM1-2-HR fail during the dry season.

Note that the AR6 (Chap. 8, Douville et al., 2021) stated much more confidence in precipitation trends over the MC after 1980. Therefore, we conducted an additional trend calculation (figures not shown) over the 33-year (1982–2014) period for all considered observational products. Although there are differences in the slope of changes among observational products, their direction (not shown) remains the same as the 1951–2014 period.

Table 3 summarises the MSM benchmarking results for the 32 CMIP6 GCMs tested. There are 19 simulations that pass all MSMs and therefore meet the minimum requirements for the purpose of this study.

While the BMF was designed for precipitation, we can also apply the MSMs to other climate variables such as annual mean near-surface temperature (see Figs. S4–S7 and Table S1 in the Supplement). For temperature, we use the APHRODITE daily temperature datasets (version V1204R1 and V1204XR; Yatagai et al., 2012) that span 1961–2015. In general, CMIP6 GCMs show biases in average temperature, with a greater number of GCMs exhibiting cold biases rather than warm biases (Fig. S4). Almost all models succeed in simulating the observed spatial distribution (e.g. Scor's greater than 0.75), phases (e.g. no model fails the benchmarking for the temperature annual cycle; Figs. S5 and S6) and historical trends (e.g. an increasing trend; Fig. S7) in temperature. Overall, models are better at simulating temperature characteristics (e.g. spatial patterns, annual cycles and trends) than precipitation over SEA. Out of the four models that fail the MSMs for near-surface temperature, two INM-family simulations do not meet the expected spatial distribution benchmark (Scor $\geq$ 0.85), while CNRM-CM6-1-HR and NESM3 show the largest relative errors compared to APHRODITE (MAPE $=$ 0.08). These four models also fail the MSMs for precipitation, as discussed above.

3.2 Versatility metrics – process-oriented metrics

In addition to the MSMs, our aim is to select a subset of GCMs for dynamical downscaling that simulate precipitation mechanisms. Therefore, in the next steps we focus on process-oriented metrics that capture the relationships between precipitation and other variables well.

3.2.1 Monsoon wind

We seek to identify models that adequately depict the low-level circulation over SEA during two prominent seasons: boreal summer (June–September; JJAS) and winter (December–February; DJF) by comparing them to ERA5 (Figs. 7 and 8, respectively). To measure the agreement between simulated and observed wind patterns in terms of intensity and direction, we employ three metrics, Scor, MAPE and MD (see Sect. 2.2.2), and we set the benchmarking threshold for each metric in dealing with limited simulations at this versatility stage. In particular, we define the threshold for wind intensity as MAPE $\leq$ 0.65 to seek models that do not overestimate the amplitude of monsoon wind. In terms of wind structure, we set a stricter benchmarking threshold for Scor of $\geq$ 0.70, aiming to retain models that adequately represent the distribution of wind intensity across the whole region. Recognising that wind magnitude might be the same at a location but different directions could substantially impact rainfall patterns, we consider a threshold for direction MD of $\leq$ 20°. This criterion helps to eliminate models where high-speed wind direction deviates significantly from observed patterns.

During summer, ERA5 shows westerly winds flowing from the Bay of Bengal into Indochina, then deviating northward to the northern Philippines (along 10° N). Concurrently, easterly winds from Australia traverse the MC and Papua New Guinea (see Fig. 7). Conversely, in winter, the wind patterns are largely reversed (Fig. 8). The easterly and northeasterly winds from the north pass through the Philippines, reaching the southern coast of Vietnam and the Malaysian peninsula, while westerly winds predominate between the Indonesian islands towards Papua New Guinea.

Overall, the subset of CMIP6 GCMs capture the circulation structure relatively well (Scor's ranging from 0.72 to 0.92 for DJF and from 0.81 to 0.95 for JJAS) but tend to overestimate the wind intensity relative to ERA5, particularly over high-speed wind areas. For example, the westerly component from the Bay of Bengal during JJAS or the easterly component over the MC during DJF is too strong compared to ERA5. These might link with the wet biases discussed in Sect. 4 TS5. Interestingly, all MSM-selected models for precipitation capture the direction of the main components of JJAS monsoon flow well.

Using the definition of benchmark thresholds mentioned above, all models meet our expectations for wind intensity (MAPE) during the summer season but two fail for the winter season (i.e. MAPEs of 0.79 for CMCC-CM2-HR4 and 0.69 for MIROC6). Interestingly, only one model fails the benchmarking for wind spatial distribution and direction: CNRM-ESM2-1 (MD is 21.67 during DJF; Fig. 8).

3.2.2 Rainfall teleconnections with modes of variability

The rainfall teleconnection for DJF ENSO is examined for two different seasons. We look at the extended summer season of the developing year (MJJASO of year 0) and the boreal spring of the following year (MAM of year +1), while the precipitation–IOD teleconnection is analysed for boreal

Table 3. Summary of model performance against the MSMs for precipitation. Models that pass the benchmarks are highlighted in bold. TS4

Simulations	Wet season		Dry season		Seasonal cycle	Trend		Pass (7 is passing)
	MAPE	Scor	MAPE	Scor		Wet	Dry	
ACCESS-CM2	+	+	+	+	+	+	+	**7**
ACCESS-ESM1-5	+	+	+	+	-	+	+	6
BCC-CSM2-MR	+	+	+	+	+	+	+	**7**
CESM2	+	+	+	+	+	+	+	**7**
CMCC-CM2-HR4	+	+	+	+	+	+	+	**7**
CMCC-CM2-SR5	+	+	+	+	+	+	+	**7**
CMCC-ESM2	+	+	+	+	+	+	+	**7**
CNRM-CM6-1	+	+	+	+	-	+	+	6
CNRM-CM6-1-HR	+	+	+	+	-	+	+	6
CNRM-ESM2-1	+	+	+	+	+	+	+	**7**
E3SM-1-0	+	+	+	+	+	+	+	**7**
EC-Earth3-AerChem	+	+	+	+	-	+	+	6
EC-Earth3-CC	+	+	+	+	-	+	+	6
EC-Earth3	+	+	+	+	+	+	+	**7**
EC-Earth3-Veg	+	+	+	+	+	+	+	**7**
EC-Earth3-Veg-LR	+	+	+	+	+	+	+	**7**
GFDL-CM4	+	+	+	+	+	+	+	**7**
GFDL-ESM4	+	+	+	+	+	+	+	**7**
HadGEM3-GC31-MM	+	+	+	+	+	+	+	**7**
INM-CM4-8	-	+	-	+	+	+	+	5
INM-CM5-0	-	+	-	+	+	+	+	5
IPSL-CM6A-LR	-	-	-	+	+	+	+	4
IPSL-CM6A-LR-INCA	-	-	-	+	+	+	+	4
MIROC6	+	+	+	+	+	+	+	**7**
MPI-ESM1-2-HR	+	+	+	+	-	+	-	5
MPI-ESM1-2-LR	+	+	+	+	+	+	+	**7**
MRI-ESM2-0	+	-	+	+	+	+	+	6
NESM3	+	+	-	+	+	+	+	5
NorESM2-MM	-	+	+	+	+	+	+	6
SAM0-UNICON	+	+	+	+	+	+	+	**7**
TaiESM1	+	+	+	+	+	+	+	**7**

autumn (SON). To benchmark CMIP6 GCMs, three metrics (HR, MR and FAR; see Sect. 2.2.2) are calculated for each GCM considering the thresholds ≥ 0.5 for HR and ≤ 0.65 for MR and FAR, given the limited number of simulations used at this stage.

The results for observations and CMIP6 GCMs selected from MSMs are shown in Figs. 9–11. The observed teleconnections vary widely by region and season. In general, ENSO-induced summer rainfall variability is dominant over the MC (e.g. Sumatra and Java; Fig. 9), while spring variability is dominant over Indochina, northern Borneo and Philippines (Fig. 10), which agrees with the evolution and seasonal circulation migration mentioned in the previous literature (Juneng and Tangang, 2005; Supari et al., 2018; Wang et al., 2020). On the other hand, IOD-induced rainfall variability is more pronounced during the SON season over the MC (Fig. 11).

CMIP6 GCMs (Fig. 10) demonstrate reasonable accuracy in simulating the spatial distribution of the ENSO teleconnection but tend to overestimate its strength, particularly over regions where observed temporal correlation coefficients are non-significant. During MJJASO of the developing year, most models successfully reproduce significant negative signals over the MC (e.g. high HR values ranging from 0.66 to 0.7 and low MR values less than 0.4). During boreal spring of the following year (MAM of year 1), the ENSO signals in CMIP6 GCMs match the observed pattern better than those during MJJASO of the developing year (Fig. 9), particularly over Indochina. Higher values of HR and lower MRs are found in most CMIP6 GCMs. This is consistent with the previous literature that highlights that GCMs tend to overestimate ENSO variability across much of the equatorial Pacific (McKenna et al., 2020) and produce a poor representation of the ENSO life cycle (Taschetto et al., 2014; McKenna et al., 2020) and interaction between ENSO and IOD (McKenna et al., 2020; Planton et al., 2021). Note that certain models consistently perform well across seasons, such as EC-Earth3-Veg, EC-Earth3-CC,
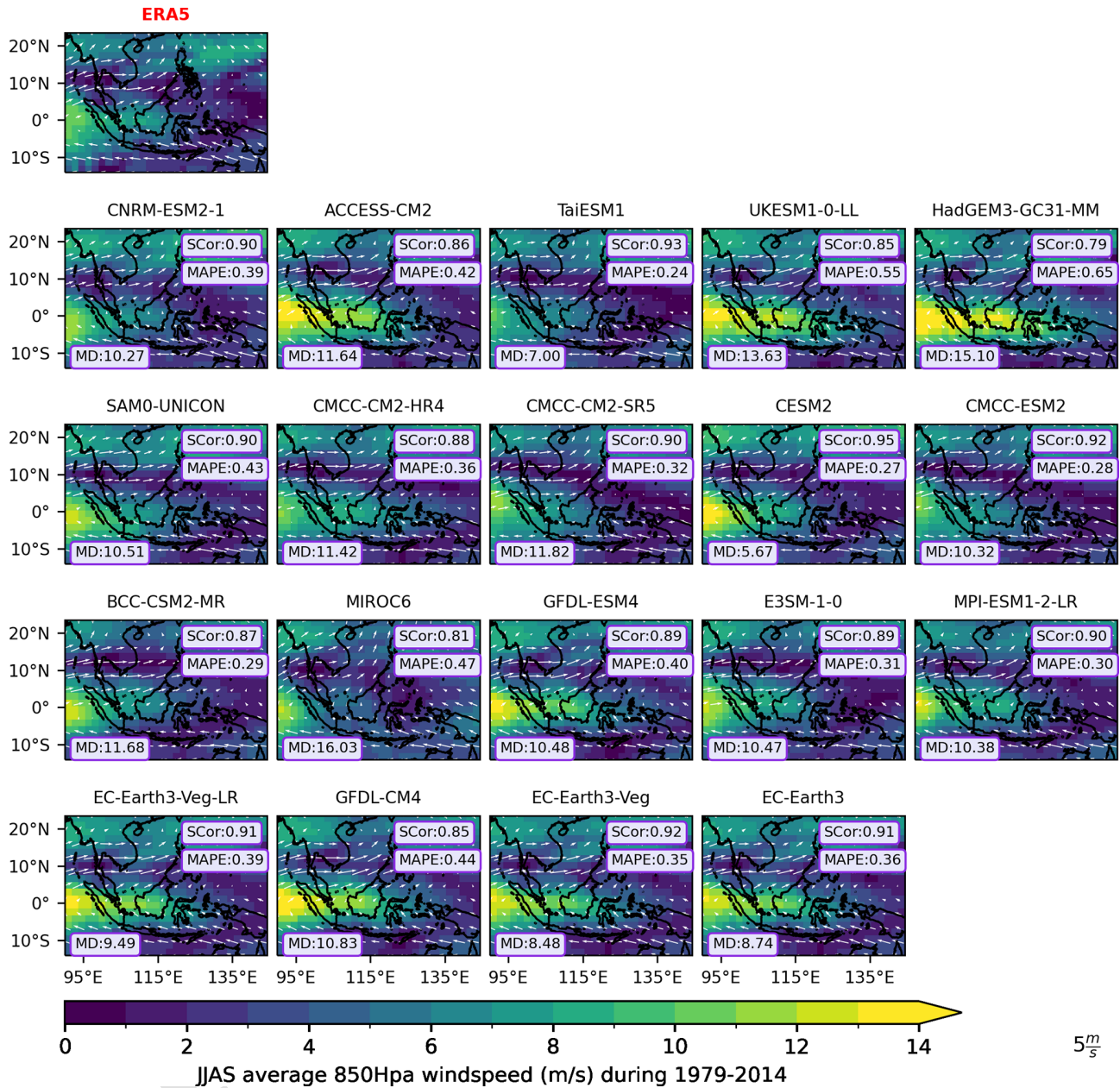
Figure 7. The spatial distribution of the climatology (1979–2014) of low-level wind circulation during the summer (JJAS; vectors) in ERA5 reanalysis (highlighted by red title) and for individual simulations selected using MSMs. All analyses are considered at the coarsest CMIP6 GCM resolution (i.e. NESM3, ∼ 216 km). Shading indicates the magnitude of the wind (in m s^{-1}). The mean absolute percentage errors (MAPEs) and spatial correlations (Scor's) calculated against ERA5 are plotted in the upper-right corners. The mean of difference in wind direction (MD) referenced to ERA5 is shown in the lower-left corner. Values highlighted in purple-coloured boxes indicate that they meet our defined benchmarking thresholds. Models are ranked from highest to lowest values of MD.

GFDL-ESM4 or HadGEM3-GM31-MM, while others, like BCC-CSM2-MR and CESM2, exhibit less-favourable performance in capturing ENSO teleconnections over the region (Figs. 9 and 10). Eight out of the 19 models, including the EC-Earth3 family, ACCESS-CM2, E3SM1-0, GFDL-ESM4, HadGEM3-GCM31-MM, MPI-ESM1-2-LR, SAM0-UNICON and UK-ESM1-0-LL, meet the ENSO teleconnec-

tion benchmark. Among models that did not pass the benchmark, many indicate an overestimation of observed non-significant ENSO signals (FAR) over the mainland during the MJJASO of year 0 (e.g. the FARs of CMCC-CM2-HR, TaiESM1 and GFDL-CM4 are 0.76, 0.75 and 0.72, respectively) or over the MC during MAM of the following year
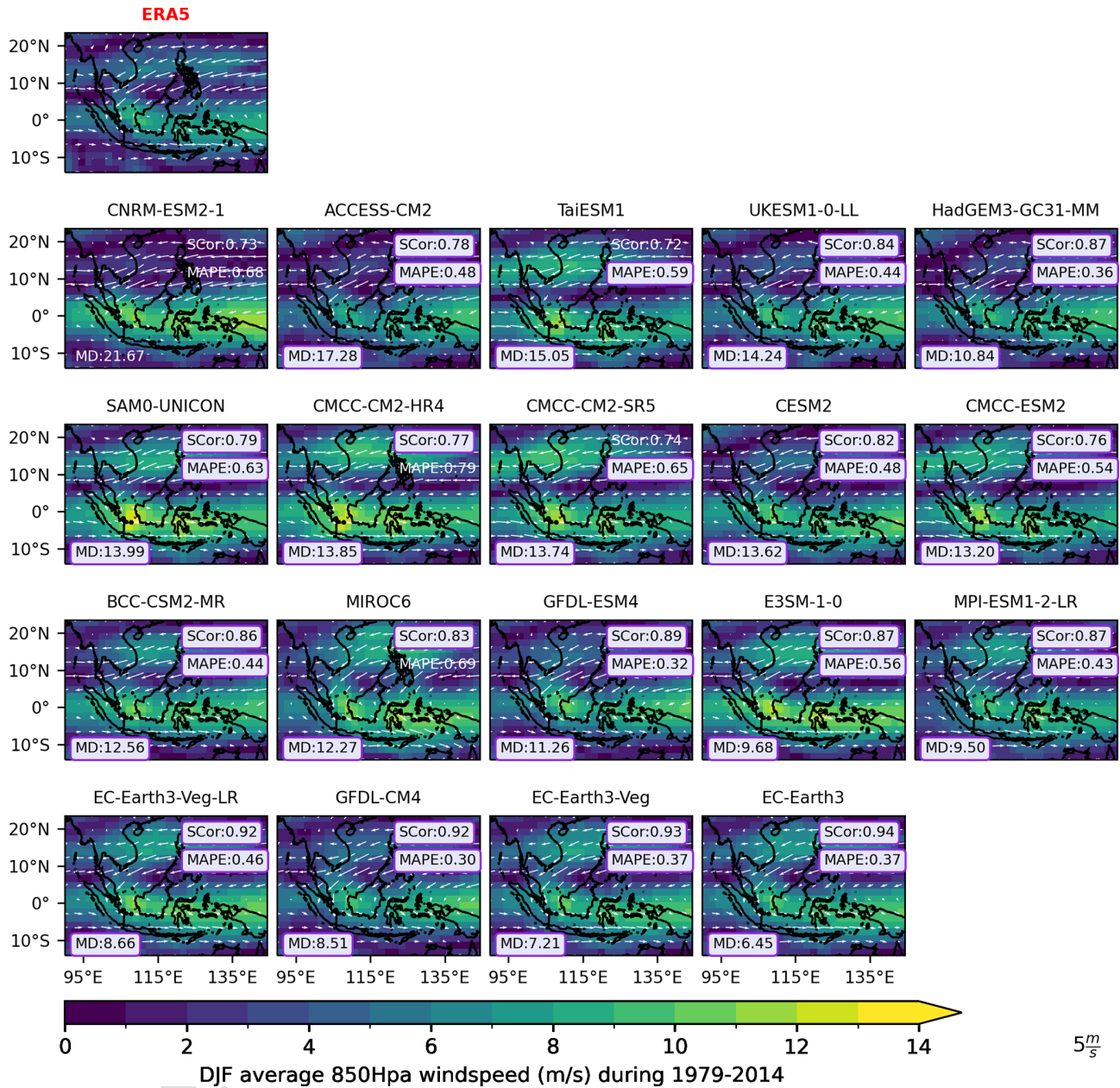
Figure 8. Same as Fig. 7 but for the boreal winter wind (December–February; DJF).

(e.g. the FARs of CMCC-CM2-SR5 CE2, EC-Earth3-Veg-LR and CMCC-ESM2 are 0.84, 0.77 and 0.74, respectively).

Interestingly, the precipitation–IOD teleconnection shows some notable similarities among the 18 CMIP6 GCMs considered at the versatility metrics stage (Fig. 11). Most models capture the significant negative correlation over Java and southern Borneo, resulting in high HR values (ranging from 0.58 to 0.75). An exception is CESM2, which produces non-significant signals over the entire region (Fig. 11). Interestingly, models that demonstrate weak performance in simulating ENSO teleconnections (e.g. BCC-CSM2-MR, CESM2 and CMCC-CM2-HR) also struggle to accurately simulate

the IOD teleconnection. Using the same threshold definitions as established for assessing the ENSO teleconnection, we identify 14 out of 18 models that pass the benchmarking for IOD teleconnection.

Given the large observational uncertainty, particularly in rainfall estimation over the region (Nguyen et al., 2020, 2022), we apply the BMF using different reference datasets while maintaining a consistent benchmarking threshold definition. This evaluation identifies a similar list of models meeting the minimum standards of performance (Table S1). However, exceptions are noted; for instance, MPI-ESM1-2-LR fails to meet the MSMs when compared with
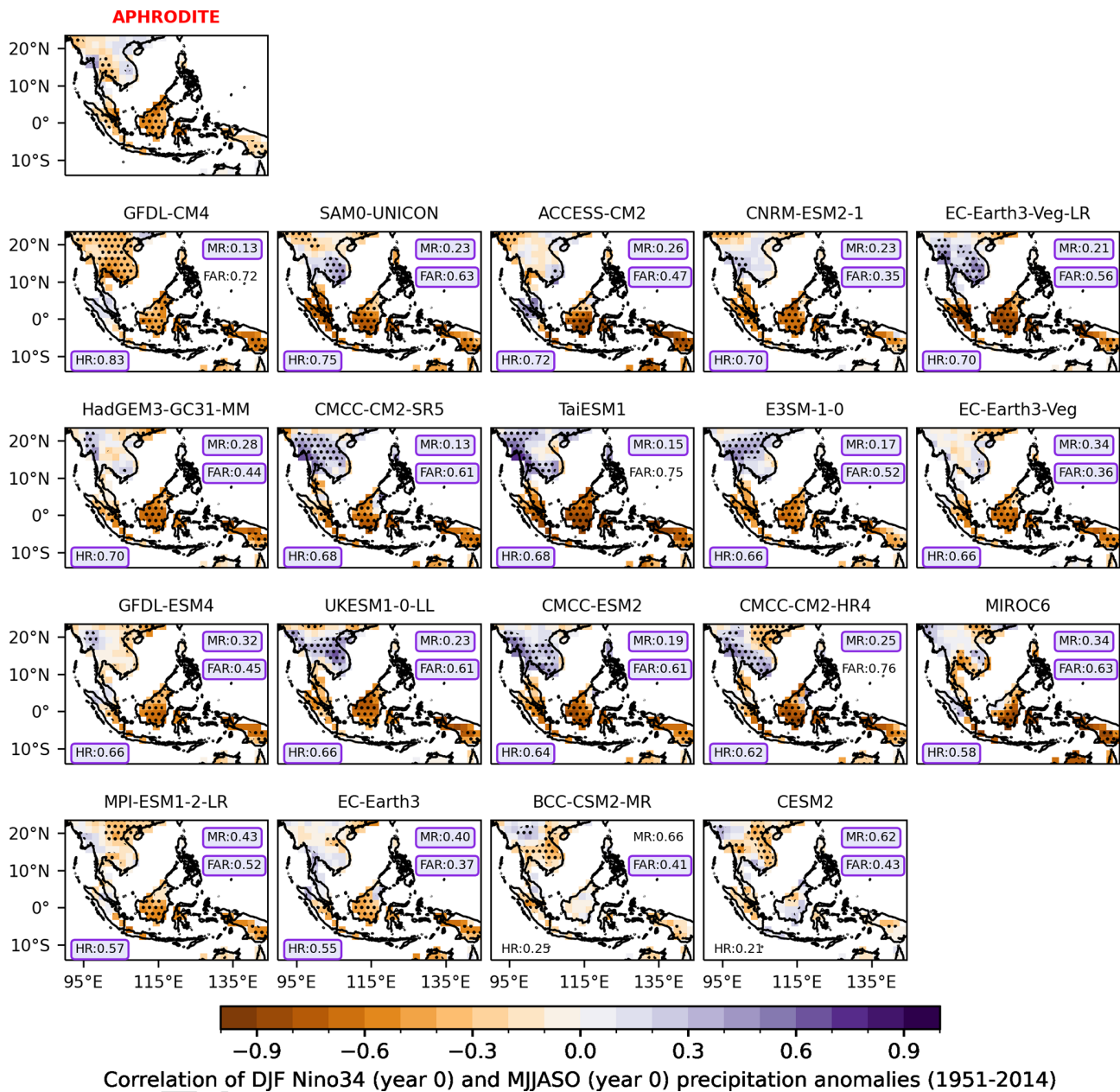
Figure 9. Lead correlation coefficients of the boreal summer (May–October; MJJASO year 0) rainfall with the mature phase of ENSO (December–January–February; DJF year 0 of Niño3.4 indices) for observations from APHRODITE with HadISST: individual CMIP6 GCM models during the period 1951–2014. The stippling indicates the grid points where the correlation coefficient is statistically significant at a 95 % confidence level according to the Student t test. The hit rates (HRs), miss rates (MRs) and false-alarm rates (FARs) calculated against APHRODITE are shown in the bottom-left and upper-right corners. All analyses are considered at the coarsest CMIP6 GCM resolution (i.e. NESM3, ∼ 216 km). Values highlighted in purple-coloured boxes indicate values that meet our defined benchmarking thresholds. Models are ranked from highest to lowest values of HR.

GPDD_FDD CE3 but passes with other references. Similarly, NorESM2-MM exhibits varying performance across different observational products. However, even if these two models are included in the subsequent selection steps, they fail to meet one or more versatility metrics. For instance, MPI-ESM1-2-LR fails the IOD–teleconnection benchmark

(Fig. 11 and Table 4), while NorESM2-MM fails the ENSO–teleconnection benchmark (Fig. S9).

It is acknowledged that different SST products vary in capturing the teleconnection. Figure S8 indicates the notable similarities among SST products in capturing the response of precipitation with modes of variability over SEA, except
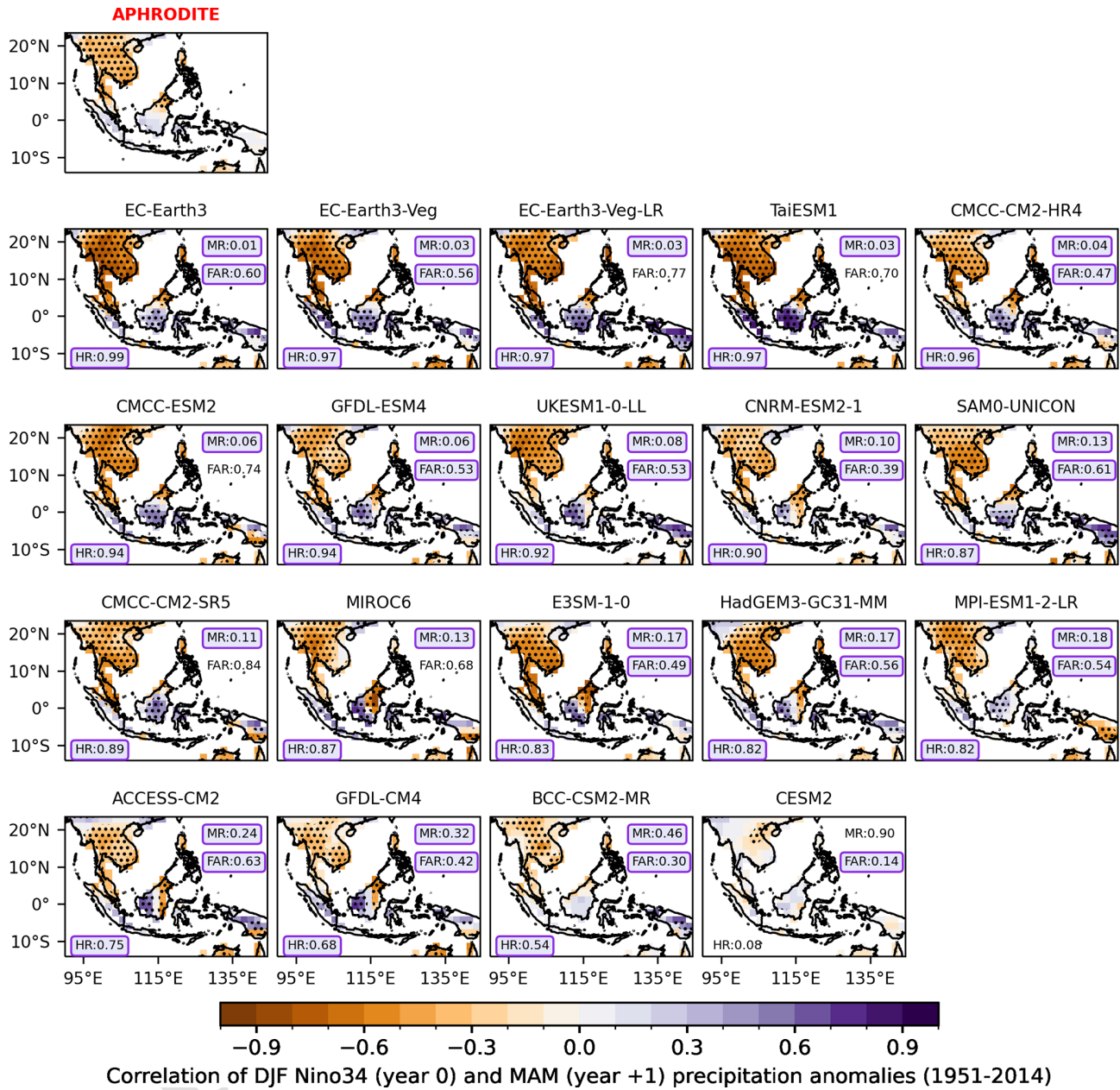
Figure 10. Similar to Fig. 9 but for the lag correlation coefficients of the mature phase of ENSO (December–January–February; DJF year 0 of Niño3.4 indices) with the boreal spring (March–April–May; MAM year +1) rainfall for (top) observations from APHRODITE with HadISST and individual CMIP6 GCM models during the period 1951–2014. Models are ranked from highest to lowest values of HR. All analyses are considered at the coarsest CMIP6 GCM resolution (i.e. NESM3, ∼ 216 km).

for the teleconnection between DJF (year 0) ENSO and MJ-JASO (year 0) precipitation. However, despite the diversity in SST products, the final selection of models passing the BMF remains the same.

Table 4 summarises the results of benchmarking 19 CMIP6 GCMs selected from the MSM for the versatility metrics. At the point of applying the BMF, we find eight models (ACCESS-CM2, E3SM1-0, EC-Earth3, EC-Earth3-Veg, GFDL-CM4, HadGEM3-GC31-MM, SAM0-UNICON and UKESM1-0-LL) that meet our expectations in simulating precipitation drivers and teleconnections with modes of variability. This could be due to the fact that IOD is an ENSO artefact (Dommenget, 2011).
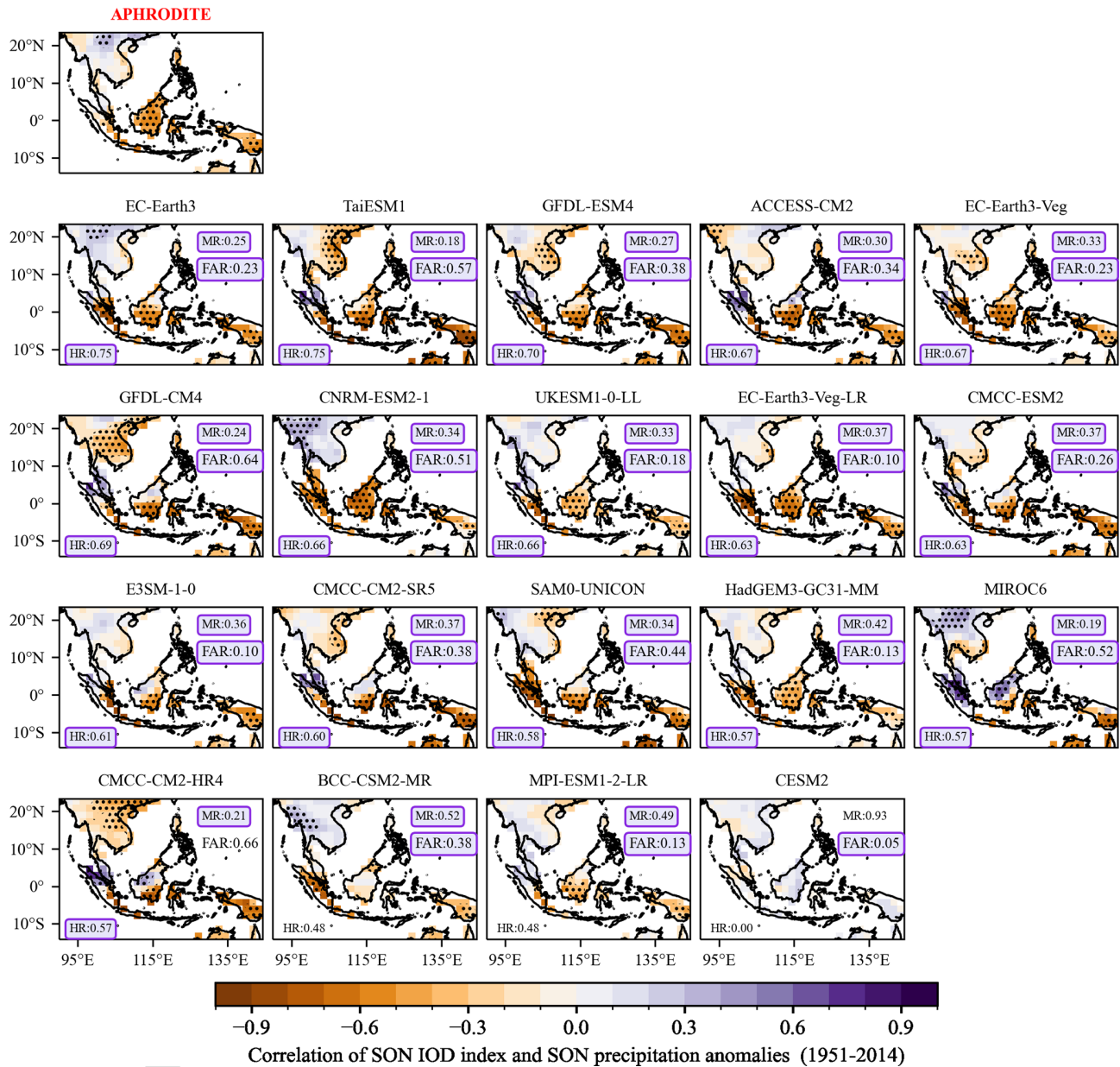
Figure 11. Correlation coefficients of the boreal autumn (September–October–November; SON) rainfall with IOD (DMI) for observations from APHRODITE with HadISST and for individual CMIP6 GCMs during the period 1951–2014. The stippling indicates the grid points where the correlation coefficient is statistically significant at a 95 % confidence level according to the Student t test. The hit rates (HRs), miss rates (MRs) and false-alarm rates (FARs) calculated against APHRODITE are plotted in the bottom-left and upper-right corners. Values highlighted in purple-coloured boxes indicate values that meet our defined benchmarking thresholds. Models are ranked from highest to lowest values of HR. All analyses are considered at the coarsest CMIP6 GCM resolution (i.e. NESM3, $\sim$ 216 km).

3.3 Future climate change signals and model dependence

In this section, we examine the climate change signals from CMIP6 GCMs that provide at least mean temperature and precipitation data for the Shared Socioeconomic Pathway (SSP3-7.0) scenario across two distinct seasons (see Fig. 12). Note that some models, such as CNRM-CM6-1-HR and EC-Earth3-Veg-LR (listed in Table 1), do not offer the sub-daily data (e.g. atmospheric variables in three dimensions at 6 h intervals) required for dynamical downscaling at the time of writing. Nevertheless, we include these models in our analysis to gain insights into the future climate change responses of CMIP6 GCMs. Interestingly, while temperature projections show general agreement regarding an increasing trend (ranging from 2.1 to 5.1 °C), precipitation projections ex-

Table 4. Summary model performance against the versatility metrics that focused on precipitation drivers and modes of variability (ENSO and IOD teleconnections). Models that meet or exceed the benchmarks are highlighted in bold. All analyses are considered at the coarsest CMIP6 GCM resolution (i.e. NESM3, ∼ 216 km).

Simulations	Monsoon circulation						ENSO teleconnection						IOD teleconnection			Pass (15 is
	JJAS			DJF			MJJASO			MAM			SON			passing)
	Scor	MD	MAPE	Scor	MD	MAPE	MR	FAR	HR	MR	FAR	HR	MR	FAR	HR	
ACCESS-CM2	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	**15**
BCC-CSM2-MR	+	+	+	+	+	+	-	+	-	+	+	+	+	+	+	13
CESM2	+	+	+	+	+	+	-	+	+	-	-	+	-	-	+	10
CMCC-CM2-HR4	+	+	+	-	+	+	+	+	-	+	+	+	-	+	-	11
CMCC-CM2-SR5	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	14
CMCC-ESM2	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	14
CNRM-ESM2-1	+	+	+	-	-	-	+	+	+	+	+	+	+	+	+	12
E3SM-1-0	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	**15**
EC-Earth3	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	**15**
EC-Earth3-Veg	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	**15**
EC-Earth3-Veg-LR	+	+	+	+	+	+	+	+	+	+	+	-	+	+	+	14
GFDL-CM4	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	14
GFDL-ESM4	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	**15**
HadGEM3-GC31-MM	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	**15**
MIROC6	+	+	+	-	+	+	+	+	+	+	+	-	+	+	+	13
MPI-ESM1-2-LR	+	+	+	+	+	+	+	+	+	+	+	+	-	+	+	14
SAM0-UNICON	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	**15**
TaiESM1	+	+	+	+	+	+	+	+	-	+	+	-	+	+	+	13
UKESM1-0-LL	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	**15**
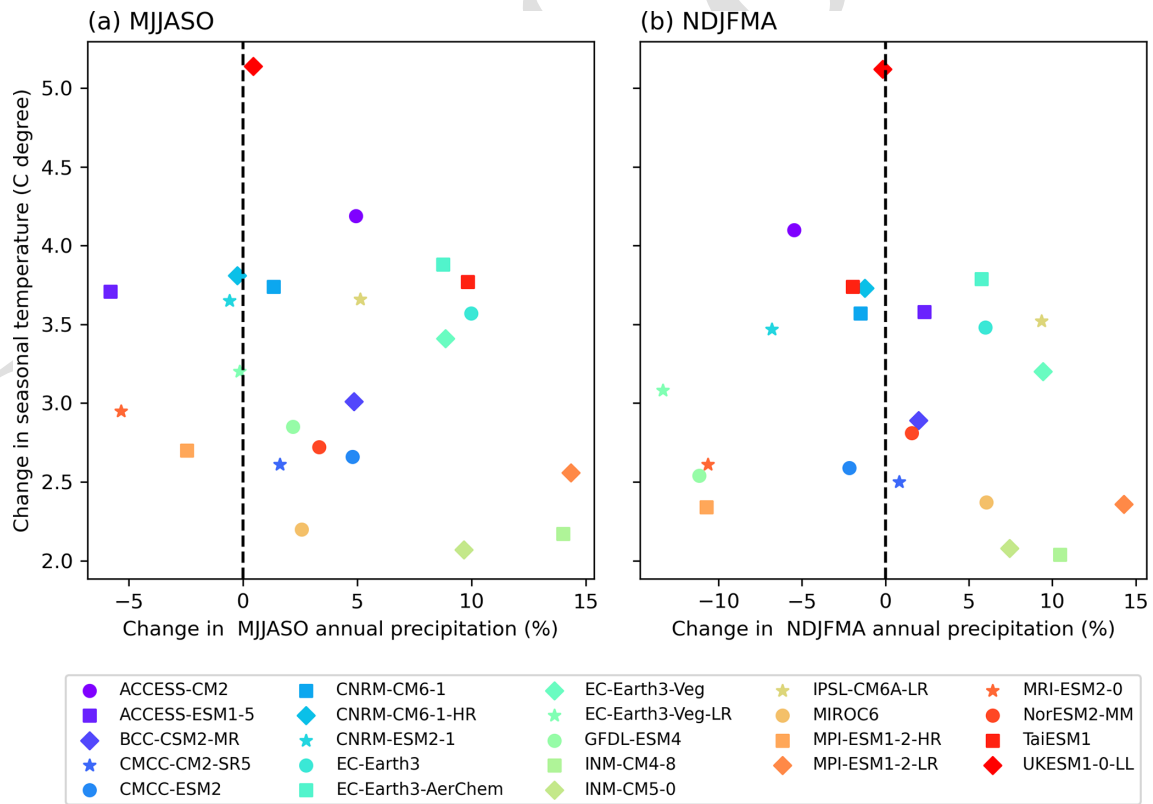
Figure 12. CMIP6 GCM climate change signal (2070–2099 relative to 1961–1990) over mainland Southeast Asia during **(a)** the wet season (MJJASO) and **(b)** the dry season (NDJFMA). The analyses are conducted for the GCMs that simulated at least monthly near-surface air temperature (tas) and precipitation (pr) for the SSP3-7.0 scenario. Note that some models that did not simulate tas or pr for SSP3-7.0 (e.g. E3SM1-0, HadGEM3-GCM31-MM and SAM0-UNICON) are not plotted.

hibit large variation in both signal and magnitude (ranging from −4.3 % to 12.9 %). Therefore, we cannot see the linear relationship between the change in regional total precipitation and temperature. Among the eight models that pass our BMF a priori expectations, there are only five models that provide at least data for monthly near-surface temperature (tas) and precipitation (pre), and they are distributed across the wide range of temperature and precipitation signals over SEA. They include the wettest models in both seasons with mid-range projected temperatures, e.g. EC-Earth3 (10 % and 3.6 °C) and EC-Earth3_Veg (8.9 % and 3.4 °C) for the MJJASO season (Fig. 12a), UKESM1-0-LL as the model with the largest increase in temperature (e.g. 5.1 °C during the MJJASO season), GFDL-ESM4 as the model with larger response in precipitation and lower warming (e.g. −11.2 % and 2.5 °C during the MJJASO season), and ACCESS-CM2 as the model with a high-range temperature and mid-range precipitation response (e.g. 4.9 % and 4.2 °C during the MJJASO season).

The dendrogram and matrix of spatial correlations between CMIP6 GCMs are shown for Southeast Asia for climatological bias (Figs. S10 and S11) and long-term changes (Figs. 13 and 14) in total precipitation. As before, we focus on the wet (MJJASO) and dry (NDJFMA) seasons. Historical correlations highlight notable similarities between models in historical bias maps (mostly significant and greater than 0.5), except UK-ESM1-0-LL, which shows poorer relationships with other models (e.g. correlation coefficients with other models are less than 0.5; Figs. S10 and S11). However, there is higher independence in projection maps compared with that in historical maps. This interesting feature needs further investigation.

Clustering analysis indicates three main spatial change clusters for the MJJASO season, as shown in the dendrogram (Fig. 13). This indicates similarities in the spatial patterns of the climate change response maps (e.g. correlations greater than 0.5) not only among models from the same families, e.g. among the Met Office GCM-based family (i.e. UKESM1-0-LL, ACCESS's family) and in model families that share the same model components (e.g. the UK-ESM1-0-LL and EC-Earth3 families share the same ocean model, NEMO3.6; Table 1) but also in less obvious families like the CNRM and INM families or the EC-Earth-based and GFDL-based simulations. An exception is EC-Earth-Veg-LR, which appears in different main clusters compared with other EC-Earth-based simulations. As indicated in the MJJASO dendrogram, the BMF-passing models that have data available for dynamical downscaling are in two main clusters, including the EC-Earth3–EC-Earth-Veg–GFDL-ESM4 cluster and the UKESM1-0-LL–ACCESS-CM2 cluster.

Figure 14 indicates two main spatial change clusters in the dry season. Interestingly, some models from the same family (e.g. EC-Earth3 and EC-Earth-Veg) still belong to the same main cluster but span different branches of the dendrogram. This might be related to the different role of internal variability in determining the level of uncertainty for precipitation during different seasons and needs further investigation. Interestingly, among models that pass the BMF, EC-Earth3 and EC-Earth-Veg appear on a main cluster, while UKESM1-0-LL, ACCESS-CM2 and GFDL-ESM4 are in the other main cluster for the NDJFMA dendrogram. This highlights the dependence of the clustering analysis on the season.

We acknowledge that a model's good performance in simulating historical climate conditions does not necessarily guarantee similar accuracy in future climate projections, a well-recognised issue in climate modelling (Herger et al., 2019). However, there are no arguments in the literature suggesting that models with weaker skill in simulating historical climatology perform better in future projections. On the contrary, we believe that models demonstrating good performance in both statistical and process-based metrics are more likely to provide credible future projections, given their proven ability to accurately simulate the physical mechanisms responsible for generating rainfall in the region.

In general, based on our evaluation of model performance, model dependence and future climate change spread, we identify two independent groups of models to use for dynamical downscaling over SEA, that is, EC-Earth3–EC-Earth-Veg and ACCESS-CM2–UKESM1-0-LL. Models from these two groups also offer the atmospheric variables in three dimensions at 6h intervals required for dynamical downscaling (Table 1). Given the inconsistency of classification of GFDL-ESM4 during different seasons and metrics, we suggest considering GFDL-ESM4 with caution.

4 Discussion

Our results somewhat differ from traditional model evaluation studies like Desmet and Ngo-Duc (2022), which ranks models by evaluation metrics and identifies a list of the best models including EC-Earth3, EC-Earth3-Veg, CNRM-CM6-1-HR, FGOALS-f3-L, HadGEM3-GC31-MM, GISS-E2-1-G, GFDL-ESM4, CIESM-WACCM and FIO-ESM-2-0. First, rather than ranking models, our aim is to retain models that meet our predefined expectations (e.g. benchmarking thresholds). Second, the list of examined models is different since we especially focus on models with a resolution greater than 2° to avoid the impacts of coarser resolutions in GCMs on dynamical downscaling. Furthermore, while Desmet and Ngo-Duc (2022) combine model performance in simulating surface climates (e.g. precipitation and near-surface temperature) and climate processes (e.g. low-level atmospheric circulation), our focus is solely on precipitation, its drivers and teleconnections with modes of variability.

We acknowledge that the list of models passing the BMF might change depending on how the benchmarking thresholds are defined. Isphording et al. (2024) note that the definition of the benchmarking thresholds for the MSMs and versatility metrics can be subjective, and they should be chosen to
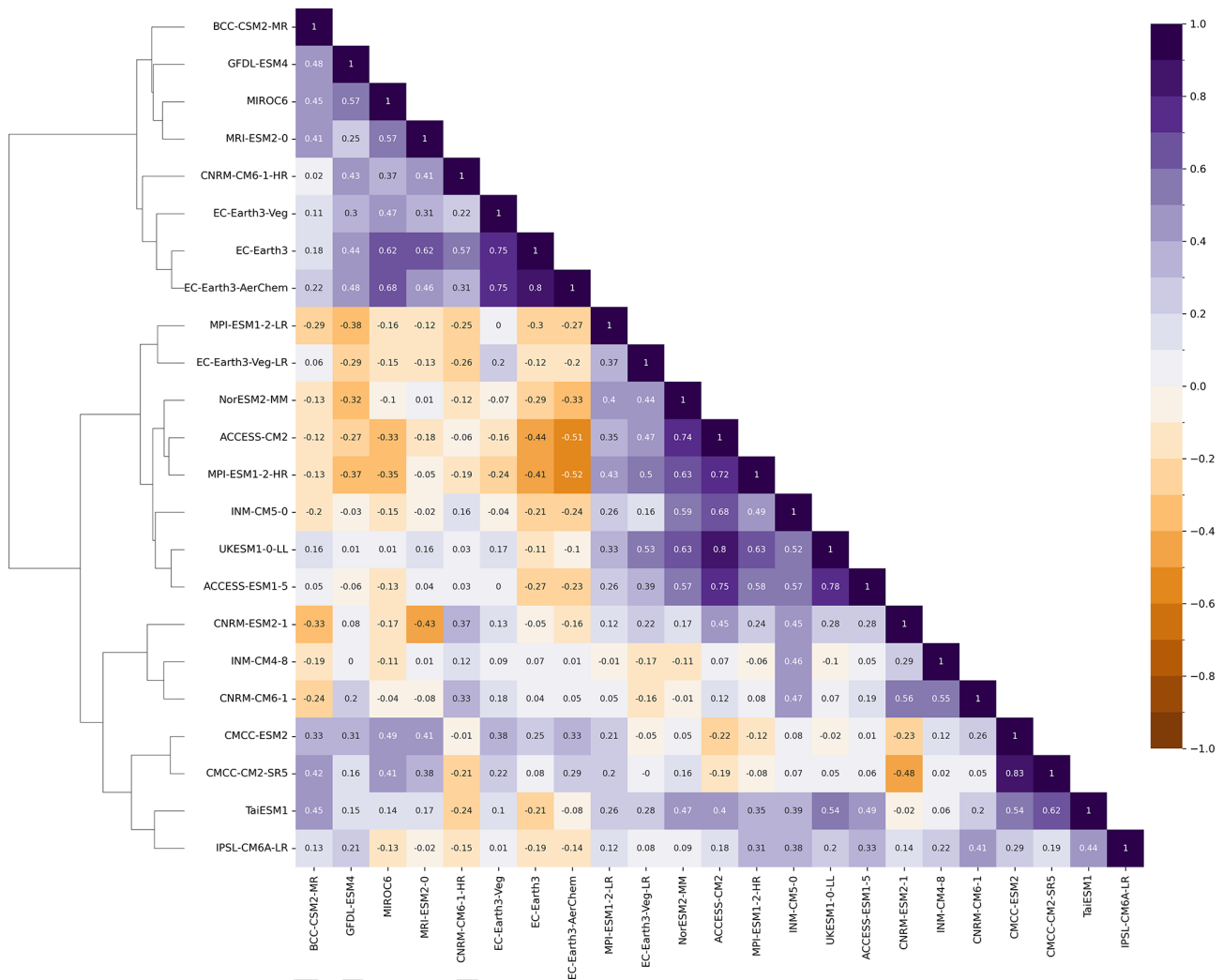
Figure 13. Dendrogram with hierarchical clustering applied to a matrix of spatial correlation coefficients between CMIP6 climate models for the long-term changes (2070–2099 SSP3-7.0 relative to 1961–1990) in total precipitation during the wet season (MJJASO). The matrix is plotted for GCMs that simulated at least monthly near-surface air temperature (tas) and precipitation (pr) for the SSP3-7.0 scenario only. Models are clustered using Ward's linkage criterion.

fit the purpose of the study while incorporating strong scientific reasoning. The strategy employed here involves defining the benchmarking thresholds based on our knowledge of observational uncertainty over the region. In addition, we aim to give each model the benefit of doubt, thus retaining a broad range of plausible future climate change responses. In particular, in the initial step of the BMF framework, we are generous in defining the benchmark threshold for the wet season given the lower model performance compared to the dry season. This approach results in 19 out of 32 models passing the MSMs. Subsequently we employ versatility metrics to cover a more process-based assessment. Given that previous studies have highlighted the overestimation of GCMs in simulating precipitation drivers and their teleconnections and limited possible simulations at this stage, we also set relaxed thresholds for various metrics to maximise the number of models

passing the BMF. We feel this is a pragmatic approach to retain a reasonable sample size and explore plausible futures. However, we acknowledge that dynamical downscaling experiments often require significant computing resources and only a small subset of GCMs should be pre-selected. Therefore, we narrow down our selection of 8 GCMs for further assessment using metrics related to model dependency and future climate change spread.

Previous studies suggest the potential impact of smoothing the extreme values when interpolating to coarser resolutions, which might affect the skill score metrics used to measure percentage errors in a simulation relative to a reference (i.e. MAPE). Although we observe a higher number of failed models for the same skill when conducting the BMF at the GCM original resolutions (Table S4), we identify a similar subset of models meeting all minimum performance require-
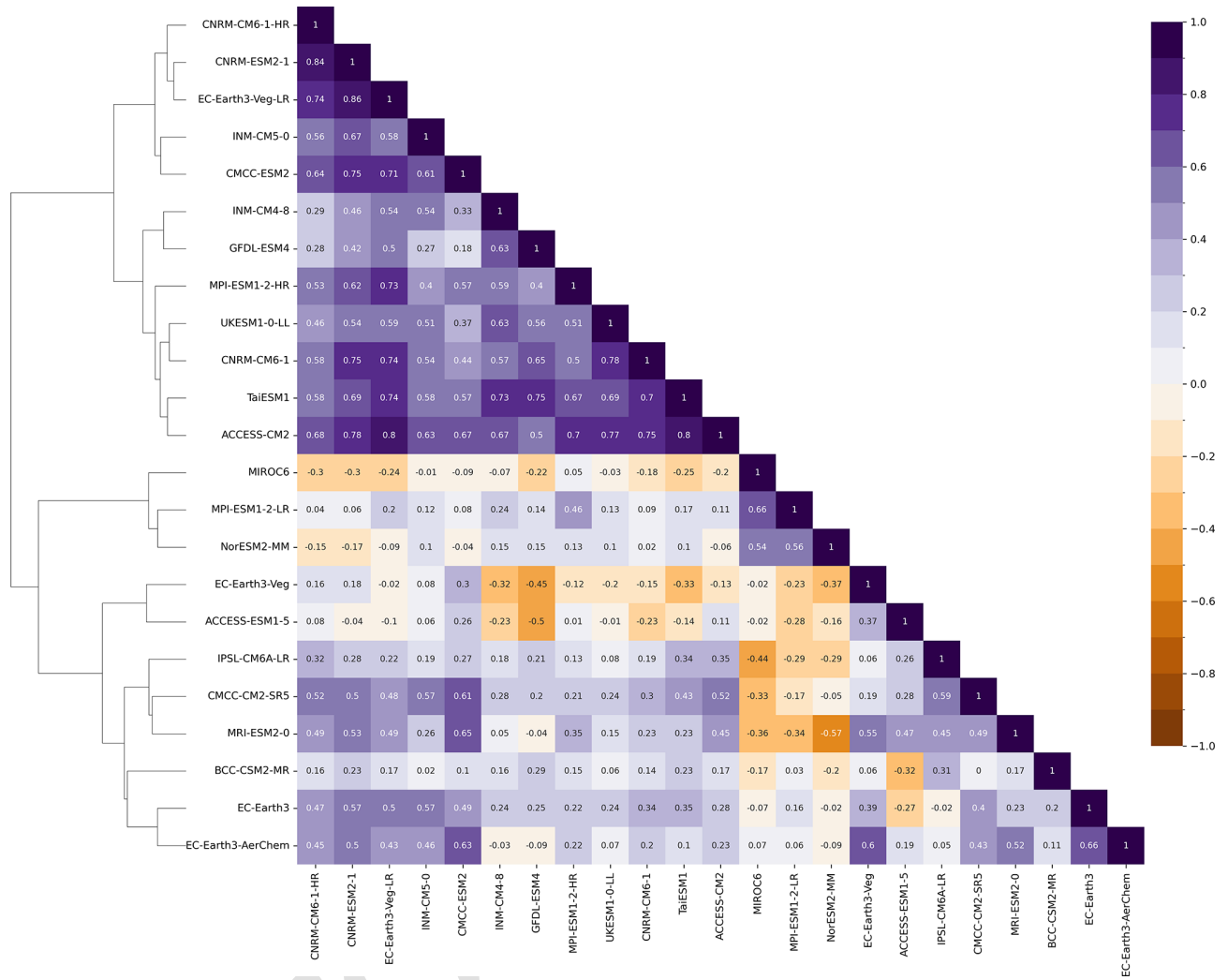
Figure 14. The same as Fig. 13 but for the dry season (November–April; NDJFMA).

ments (Table S4). This suggests that the coarser resolution of ~ 210 km used for benchmarking is not the main reason behind the results of quantifying model skill used in this study. This is in line with Nguyen et al. (2022), where they demonstrate that model components (e.g. configurations in different schemes) are the main reason behind the model biases rather than model resolution.

The relationship between model structures and model biases is investigated in the model dependency section using cluster analysis. We acknowledge that grouping of models might changes for not only the periods and seasons considered (as discussed in Sect. 3.3) but also the metrics considered. Interestingly, using mean percentage changes as distance measure between models, we identify similar main clusters of EC-Earth3–EC-Earth-Veg and ACCESS-CM2–UKESM1-0-LL among models that passed the BMF (Figs. S12 and S13). This subset of models is suitable for dynamical downscaling over Southeast Asia.

The customised BMF implemented in this study offers a consistent framework for model evaluation across the whole CORDEX-SEA domain. The framework can be further developed and applied extensively to sub-regions of interest, in particular within the upcoming Climatic hazard Assessment to enhance Resilience against climate Extremes for Southeast Asian megacities (CARE for SEA megacities) project of CORDEX-SEA. In this project, each megacity can identify their climate priority and the associated metrics to select a fit-for-purpose subset of models. This framework could also be implemented in impact-related projections over SEA for credible future projections for particular sectors: agriculture, forestry, water, etc.

5 Conclusion

In this paper, we apply the insight gained from the CMIP6 selection process for dynamical downscaling across vari-

ous CORDEX domains to Southeast Asia by encompassing several critical factors: model performance, model independence, data availability and the spread of future climate change projections.

Rather than exhaustively evaluating all performance aspects of the models in simulating the Southeast Asian climate, our focus is on selecting models that simulate precipitation well, including its drivers and teleconnections, given the high uncertainty in rainfall projections over the region. In addition, we apply a novel standardised benchmarking framework – a new approach to identify a subset of fit-for-purpose models that align with a user's a priori performance expectations. This framework has two stages of assessment: statistical-based metrics and process/regime-based metrics, conducted for both the wet (MJJASO) and dry (NDJFMA) seasons.

From the first step we identify 19 GCMs that meet our minimum criteria for simulating the fundamental characteristics (e.g. bias, spatial distribution, seasonality and trends) of seasonal rainfall. GCMs generally exhibit wet biases, particularly over the complex terrain of the Maritime Continent. These models then undergo a second evaluation, focusing on their ability to simulate climate processes and teleconnections with modes of variability. While these models consistently capture atmospheric circulation and teleconnections with modes of variability over the region, they exhibit a tendency to overestimate their strength. Ultimately, our framework narrows down the selection to eight GCMs that meet our model performance expectations in simulating fundamental characteristics of precipitation, key drivers and teleconnections over Southeast Asia. There are obvious high-performing GCMs from allied modelling groups, highlighting the dependency of the subset of models identified from the framework. Consequently, additional tests on model independence, data availability for SSP3-7.0 and the spread of future climate change are conducted. These tests lead to the identification of two independent groups of models (e.g. EC-Earth3-Veg/EC-Earth3 and ACCESS-CM2/UKESM1-0-LL) that align with our a priori expectations for dynamical downscaling over CORDEX-SEA. It is recommended that only one model from each group be chosen to avoid models that are too closely related.

Code and data availability. Code for benchmarking the CMIP6 GCM performance (Isphording, 2024) is available from https://doi.org/10.5281/zenodo.8365065 (Isphording, 2023). Data used in this study are available through CMIP6 GCMs at the Earth System Grid Federation (ESGF): https://esgf.nci.org.au/projects/esgf-nci/ (NCI Australia, 2024 TS6). ERA5 (Hersbach et al., 2020) is available at https://doi.org/10.24381/cds.bd0915c6 (Hersbach et al., 2023). OISST version 2.1 (Huang et al., 2021) is available at https://www.psl.noaa.gov/data/gridded/data.noaa.oisst.v2.highres.html (NOAA, 2024 TS7). ERSST version 5 (Huang et al., 2017) is available at https://doi.org/10.7289/V5T72FNM TS8. APHRODITE version V1101R1 and V1101 XR (Yatagai et al., 2012) are available at https://www.chikyu.ac.jp/precip/english/index.html.

Supplement. The supplement related to this article is available online at: https://doi.org/10.5194/gmd-17-1-2024-supplement.

Author contributions. RNI built the BMF used in this research. PLN applied and developed the BMF for the region of interest, performed the analysis and prepared the original paper. LVA, MJT, SCHN and JLM supervised the research and reviewed and edited the paper.

Competing interests. The contact author has declared that none of the authors has any competing interests.

ther geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors.

Acknowledgements. This work was supported by the Australian Research Council (ARC), grant FT210100459. Lisa V. Alexander and Rachael N. Isphording are also supported by the ARC grant CE17010023. The research was undertaken with the assistance of resources and services from the National Computational Infrastructure (NCI), which is supported by the Australian Government. The codes and graphic visualisation for the assessment of CMIP6 GCMs are based on the benchmarking framework suggested in Isphording et al. (2024) and Isphording (2024). CE4

Financial support. This research has been supported by the Australian Research Council (grant nos. FT210100459 and CE17010023). TS9

Review statement. This paper was edited by Stefan Rahimi-Esfarjani and reviewed by three anonymous referees.

References

Abramowitz, G.: Towards a benchmark for land surface models, Geophys. Res. Lett., 32, L22702, https://doi.org/10.1029/2005GL024419, 2005.

Abramowitz, G.: Towards a public, standardized, diagnostic benchmarking system for land surface models, Geosci. Model Dev., 5, 819–827, https://doi.org/10.5194/gmd-5-819-2012, 2012.

Alexander, L. V., Bador, M., Roca, R., Contractor, S., Donat, M. G., and Nguyen, P. L.: Intercomparison of annual precipitation indices and extremes over global land areas from in situ, space-

based and reanalysis products, Environ. Res. Lett., 15, 055002, https://doi.org/10.1088/1748-9326/ab79e2, 2020.

Allan, R., Chambers, D., Drosdowsky, W., Hendon, H., Latif, M., Nicholls, N., Smith, I., Stone, R., and Tourre, Y.: Is there an Indian Ocean dipole and is it independent of the El Niño-Southern Oscillation, CLIVAR exchanges, 21, 18–22, 2001.

Amirudin, A. A., Salimun, E., Tangang, F., Juneng, L., and Zuhairi, M.: Differential influences of teleconnections from the Indian and Pacific Oceans on rainfall variability in Southeast Asia, Atmosphere, 11, 886, https://doi.org/10.3390/atmos11090886 TS10, 2020.

Andrews, M. B., Ridley, J. K., Wood, R. A., Andrews, T., Blockley, E. W., Booth, B., Burke, E., Dittus, A. J., Florek, P., Gray, L. J., Haddad, S., Hardiman, S. C., Hermanson, L., Hodson, D., Hogan, E., Jones, G. S., Knight, J. R., Kuhlbrodt, T., Misios, S., Mizielinski, M. S., Ringer, M. A., Robson, J., and Sutton, R. T.: Historical Simulations With HadGEM3-GC3.1 for CMIP6, J. Adv. Model. Earth Sy., 12, e2019MS001995, https://doi.org/10.1029/2019MS001995, 2020.

Baquero-Bernal, A., Latif, M., and Legutke, S.: On dipolelike variability of sea surface temperature in the tropical Indian Ocean, J. Climate, 15, 1358–1368, 2002.

Best, M. J.: The plumbing of land surface models: benchmarking model performance, J. Hydrometeorol., 16, 1425, https://doi.org/10.1175/JHM-D-14-0158.1, 2015.

Bi, D., Dix, M., Marsland, S., O'Farrell, S., Sullivan, A., Bodman, R., Law, R., Harman, I., Srbinovsky, J., Rashid, H. A., Dobrohotoff, P., Mackallah, C., Yan, H., Hirst, A., Savita, A., Dias, F. B., Woodhouse, M., Fiedler, R., and Heerdegen, A.: Configuration and spin-up of ACCESS-CM2, the new generation Australian Community Climate and Earth System Simulator Coupled Model, Journal of Southern Hemisphere Earth Systems Science, 70, 225–251, https://doi.org/10.1071/ES19040, 2020.

Boé, J.: Interdependency in Multimodel Climate Projections: Component Replication and Result Similarity, Geophys. Res. Lett., 45, 2771–2779, https://doi.org/10.1002/2017GL076829, 2018.

Boucher, O., Servonnat, J., Albright, A. L., Aumont, O., Balkanski, Y., Bastrikov, V., Bekki, S., Bonnet, R., Bony, S., Bopp, L., Braconnot, P., Brockmann, P., Cadule, P., Caubel, A., Cheruy, F., Codron, F., Cozic, A., Cugnet, D., D'Andrea, F., Davini, P., de Lavergne, C., Denvil, S., Deshayes, J., Devilliers, M., Ducharne, A., Dufresne, J.-L., Dupont, E., Éthé, C., Fairhead, L., Falletti, L., Flavoni, S., Foujols, M.-A., Gardoll, S., Gastineau, G., Ghattas, J., Grandpeix, J.-Y., Guenet, B., Guez, L., E., Guilyardi, E., Guimberteau, M., Hauglustaine, D., Hourdin, F., Idelkadi, A., Joussaume, S., Kageyama, M., Khodri, M., Krinner, G., Lebas, N., Levavasseur, G., Lévy, C., Li, L., Lott, F., Lurton, T., Luyssaert, S., Madec, G., Madeleine, J.-B., Maignan, F., Marchand, M., Marti, O., Mellul, L., Meurdesoif, Y., Mignot, J., Musat, I., Ottlé, C., Peylin, P., Planton, Y., Polcher, J., Rio, C., Rochetin, N., Rousset, C., Sepulchre, P., Sima, A., Swingedouw, D., Thiéblemont, R., Traore, A. K., Vancoppenolle, M., Vial, J., Vialard, J., Viovy, N., and Vuichard, N.: Presentation and Evaluation of the IPSL-CM6A-LR Climate Model, J. Adv. Model. Earth Sy., 12, e2019MS002010, https://doi.org/10.1029/2019MS002010, 2020.

Braganza, K., Karoly, D., Hirst, A., Mann, M., Stott, P., Stouffer, R., and Tett, S.: Simple indices of global climate variability and change: Part I – variability and correlation structure, Clim. Dynam., 20, 491–502, https://doi.org/10.1007/s00382-002-0286-0, 2003.

Brands, S.: Common Error Patterns in the Regional Atmospheric Circulation Simulated by the CMIP Multi-Model Ensemble, Geophys. Res. Lett., 49, e2022GL101446, https://doi.org/10.1029/2022GL101446, 2022.

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

Cao, J., Wang, B., Yang, Y.-M., Ma, L., Li, J., Sun, B., Bao, Y., He, J., Zhou, X., and Wu, L.: The NUIST Earth System Model (NESM) version 3: description and preliminary evaluation, Geosci. Model Dev., 11, 2975–2993, https://doi.org/10.5194/gmd-11-2975-2018, 2018.

Chang, C. P., Wang, Z., McBride, J., and Liu, C.-H.: Annual Cycle of Southeast Asia – Maritime Continent Rainfall and the Asymmetric Monsoon Transition, J. Climate, 18, 287–301, https://doi.org/10.1175/JCLI-3257.1, 2005.

Chen, C., Sahany, S., Moise, A. F., Chua, X. R., Hassim, M. E., Lim, G., and Prasanna, V.: ENSO–Rainfall Teleconnection over the Maritime Continent Enhances and Shifts Eastward under Warming, J. Climate, 36, 4635–4663, https://doi.org/10.1175/JCLI-D-23-0036.1, 2023.

Cherchi, A., Fogli, P. G., Lovato, T., Peano, D., Iovino, D., Gualdi, S., Masina, S., Scoccimarro, E., Materia, S., Bellucci, A., and Navarra, A.: Global Mean Climate and Main Patterns of Variability in the CMCC-CM2 Coupled Model, J. Adv. Model. Earth Sy., 11, 185–209, https://doi.org/10.1029/2018MS001369, 2019.

Contractor, S., Donat, M. G., and Alexander, L. V.: Intensification of the daily wet day rainfall distribution across Australia, Geophys. Res. Lett., 45, 8568–8576, 2018.

Contractor, S., Donat, M. G., Alexander, L. V., Ziese, M., Meyer-Christoffer, A., Schneider, U., Rustemeier, E., Becker, A., Durre, I., and Vose, R. S.: Rainfall Estimates on a Gridded Network (REGEN) – a global land-based gridded dataset of daily precipitation from 1950 to 2016, Hydrol. Earth Syst. Sci., 24, 919–943, https://doi.org/10.5194/hess-24-919-2020, 2020.

CORDEX: CORDEX experiment design for dynamical downscaling of CMIP6, https://cordex.org/wp-content/uploads/2021/05/CORDEX-CMIP6_exp_design_RCM.pdf (last access: 29 June 2024), 2021.

D'Arrigo, R. and Wilson, R.: El Nino and Indian Ocean influences on Indonesian drought: implications for forecasting rainfall and crop productivity, Int. J. Climatol., 28, 611–616, 2008.

Danabasoglu, G., Lamarque, J.-F., Bacmeister, J., Bailey, D. A., DuVivier, A. K., Edwards, J., Emmons, L. K., Fasullo, J., Garcia, R., Gettelman, A., Hannay, C., Holland, M. M., Large, W. G., Lauritzen, P. H., Lawrence, D. M., Lenaerts, J. T. M., Lindsay, K., Lipscomb, W. H., Mills, M. J., Neale, R., Oleson, K. W., Otto-Bliesner, B., Phillips, A. S., Sacks, W., Tilmes, S., van Kampenhout, L., Vertenstein, M., Bertini, A., Dennis, J., Deser, C., Fischer, C., Fox-Kemper, B., Kay, J. E., Kinnison, D., Kushner, P. J., Larson, V. E., Long, M. C., Mickelson, S., Moore, J. K., Nienhouse, E., Polvani, L., Rasch, P. J., and Strand, W. G.: The Community Earth System Model Version 2 (CESM2), J. Adv. Model. Earth Sy., 12, e2019MS001916, https://doi.org/10.1029/2019MS001916, 2020.

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

Desmet, Q. and Ngo-Duc, T.: A novel method for ranking CMIP6 global climate models over the southeast Asian region, Int. J. Climatol., 42, 97–117, https://doi.org/10.1002/joc.7234, 2022.

Di Virgilio, G., Ji, F., Tam, E., Nishant, N., Evans, J. P., Thomas, C., Riley, M. L., Beyer, K., Grose, M. R., Narsey, S., and Delage, F.: Selecting CMIP6 GCMs for CORDEX Dynamical Downscaling: Model Performance, Independence, and Climate Change Signals, Earths Future, 10, e2021EF002625, https://doi.org/10.1029/2021EF002625, 2022.

Diaconescu, E. P. and Laprise, R.: Can added value be expected in RCM-simulated large scales?, Clim. Dynam., 41, 1769–1800, https://doi.org/10.1007/s00382-012-1649-9, 2013.

Díaz, L. B., Saurral, R. I., and Vera, C. S.: Assessment of South America summer rainfall climatology and trends in a set of global climate models large ensembles, Int. J. Climatol., 41, E59–E77, https://doi.org/10.1002/joc.6643, 2021.

Dommenget, D.: An objective analysis of the observed spatial structure of the tropical Indian Ocean SST variability, Clim. Dynam., 36, 2129–2145, https://doi.org/10.1007/s00382-010-0787-1, 2011.

Donat, M. G., Delgado-Torres, C., De Luca, P., Mahmood, R., Ortega, P., and Doblas-Reyes, F. J.: How Credibly Do CMIP6 Simulations Capture Historical Mean and Extreme Precipitation Changes?, Geophys. Res. Lett., 50, e2022GL102466, https://doi.org/10.1029/2022GL102466, 2023.

Dong, T. and Dong, W.: Evaluation of extreme precipitation over Asia in CMIP6 models, Clim. Dynam., 57, 1751–1769, https://doi.org/10.1007/s00382-021-05773-1, 2021.

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

Douville, H., Raghavan, K., Renwick, J., Allan, R. P., Arias, P. A., Barlow, M., Cerezo-Mota, R., Cherchi, A., Gan, T. Y., Gergis, J., Jiang, D., Khan, A., Pokam Mba, W., Rosenfeld, D., Tierney, J., and Zolina, O.: Water Cycle Changes, in: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M. I., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J. B. R., Maycock, T. K., Waterfield, T., Yelekçi, O., Yu, R., and Zhou, B., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, https://doi.org/10.1017/9781009157896.010, 1055–1210, 2021.

Dunne, J. P., Horowitz, L. W., Adcroft, A. J., Ginoux, P., Held, I. M., John, J. G., Krasting, J. P., Malyshev, S., Naik, V., Paulot, F., Shevliakova, E., Stock, C. A., Zadeh, N., Balaji, V., Blanton, C., Dunne, K. A., Dupuis, C., Durachta, J., Dussin, R., Gauthier, P. P. G., Griffies, S. M., Guo, H., Hallberg, R. W., Harrison, M., He, J., Hurlin, W., McHugh, C., Menzel, R., Milly, P. C. D., Nikonov, S., Paynter, D. J., Ploshay, J., Radhakrishnan, A., Rand, K., Reichl, B. G., Robinson, T., Schwarzkopf, D. M., Sentman, L. T., Underwood, S., Vahlenkamp, H., Winton, M., Wittenberg, A. T., Wyman, B., Zeng, Y., and Zhao, M.: The GFDL Earth System Model Version 4.1 (GFDL-ESM 4.1): Overall Coupled Model Description and Simulation Characteristics, J. Adv. Model. Earth Sy., 12, e2019MS002015, https://doi.org/10.1029/2019MS002015, 2020.

Evans, J., Virgilio, G., Hirsch, A., Hoffmann, P., Remedio, A. R., Ji, F., Rockel, B., and Coppola, E.: The CORDEX-Australasia ensemble: evaluation and future projections, Clim. Dynam., 57, 1385–1401, https://doi.org/10.1007/s00382-020-05459-0, 2021.

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., and Michaelsen, J.: The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes, Scientific Data, 2, 150066, https://doi.org/10.1038/sdata.2015.66, 2015.

Gibson, P. B., Perkins-Kirkpatrick, S. E., Alexander, L. V., and Fischer, E. M.: Comparing Australian heat waves in the CMIP5 models through cluster analysis, J. Geophys. Res.-Atmos., 122, 3266–3281, https://doi.org/10.1002/2016JD025878, 2017.

Gibson, P. B., Rampal, N., Dean, S. M., and Morgenstern, O.: Storylines for Future Projections of Precipitation Over New Zealand in CMIP6 Models, J. Geophys. Res.-Atmos., 129, e2023JD039664, https://doi.org/10.1029/2023JD039664, 2024.

Giorgi, F. and Gao, X.: Regional earth system modeling: review and future directions, Atmospheric and Oceanic Science Letters, 11, 189–197, https://doi.org/10.1080/16742834.2018.1452520, 2018.

Giorgi, F., Jones, C., and Asrar, G.: Addressing climate information needs at the regional level: The CORDEX framework, WMO Bulletin, 53, 2008. TS11

Giorgi, F., Coppola, E., Teichmann, C., and Jacob, D.: Editorial for the CORDEX-CORE Experiment I Special Issue, Clim. Dynam., 57, 1265–1268, https://doi.org/10.1007/s00382-021-05902-w, 2021.

Grose, M. R., Narsey, S., Trancoso, R., Mackallah, C., Delage, F., Dowdy, A., Di Virgilio, G., Watterson, I., Dobrohotoff, P., Rashid, H. A., Rauniyar, S., Henley, B., Thatcher, M., Syktus, J., Abramowitz, G., Evans, J. P., Su, C.-H., and Takbash, A.: A CMIP6-based multi-model downscaling ensemble to underpin climate change services in Australia, Climate Services, 30, 100368, https://doi.org/10.1016/j.cliser.2023.100368, 2023.

Haylock, M. and McBride, J.: Spatial coherence and predictability of Indonesian wet season rainfall, J. Climate, 14, 3882–3887, https://doi.org/10.1175/1520-0442(2001)014<3882:SCAPOI>2.0.CO;2, 2001.

Held, I. M., Guo, H., Adcroft, A., Dunne, J. P., Horowitz, L. W., Krasting, J., Shevliakova, E., Winton, M., Zhao, M., Bushuk, M., Wittenberg, A. T., Wyman, B., Xiang, B., Zhang, R., Anderson, W., Balaji, V., Donner, L., Dunne, K., Durachta, J., Gauthier, P. P. G., Ginoux, P., Golaz, J. C., Griffies, S. M., Hallberg, R., Harris, L., Harrison, M., Hurlin, W., John, J., Lin, P., Lin, S. J., Malyshev, S., Menzel, R., Milly, P. C. D., Ming, Y., Naik, V., Paynter, D., Paulot, F., Ramaswamy, V., Reichl, B., Robinson, T., Rosati, A., Seman, C., Silvers, L. G., Underwood, S., and Zadeh, N.: Structure and Performance of GFDL's CM4.0 Climate Model, J. Adv. Model. Earth Sy., 11, 3691–3727, https://doi.org/10.1029/2019MS001829, 2019.

Herger, N., Abramowitz, G., Sherwood, S., Knutti, R., Angélil, O., and Sisson, S. A.: Ensemble optimisation, multiple constraints and overconfidence: a case study with future Australian precipitation change, Clim. Dynam., 53, 1581–1596, https://doi.org/10.1007/s00382-019-04690-8, 2019.

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., and Thépaut, J.-N.: The ERA5 global reanalysis, Q. J.Roy. Meteor. Soc., 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., and Thépaut, J.-N.: ERA5 hourly data on pressure levels from 1940 to present, Copernicus Climate Change Service (C3S) Climate Data Store (CDS) [data set], https://doi.org/10.24381/cds.bd0915c6, 2023.

Huang, B., Thorne, P. W., Banzon, V. F., Boyer, T., Chepurin, G., Lawrimore, J. H., Menne, M. J., Smith, T. M., Vose, R. S., and Zhang, H.-M.: Extended Reconstructed Sea Surface Temperature, Version 5 (ERSSTv5): Upgrades, Validations, and Intercomparisons, J. Climate, 30, 8179–8205, https://doi.org/10.1175/JCLI-D-16-0836.1, 2017.

Huang, B., Liu, C., Banzon, V., Freeman, E., Graham, G., Hankins, B., Smith, T., and Zhang, H.-M.: Improvements of the Daily Optimum Interpolation Sea Surface Temperature (DOISST) Version 2.1, J. Climate, 34, 2923–2939, https://doi.org/10.1175/JCLI-D-20-0166.1, 2021.

Huang, B., Thorne, P. W., Banzon, V. F., Boyer, T., Chepurin, G., Lawrimore, J. H., Menne, M. J., Smith, T. S., Vose, R. S., and Zhang, H.-M.: NOAA Extended Reconstructed Sea Surface Temperature (ERSST), Version 5, NOAA National Centers for Environmental Information [data set], https://doi.org/10.7289/V5T72FNM TS12, 2017. TS13

Ignacio-Reardon, S. and Luo, J.-j.: Evaluation of the Performance of CMIP6 Climate Models in Simulating Rainfall over the Philippines, Atmosphere, 14, 1459, https://doi.org/10.3390/atmos14091459, 2023.

IPCC: Summary for Policymakers, in: Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Shukla, P. R., Skea, J.,

Slade, R., Al Khourdajie, A., van Diemen, R., McCollum, D., Pathak, M., Some, S., Vyas, P., Fradera, R., Belkacemi, M., Hasija, A., Lisboa, G., Luz, S., and Malley, J., Cambridge University Press, Cambridge, UK and New York, NY, USA, https://doi.org/10.1017/9781009157926.001, 2022.

IPCC: Summary for Policymakers, in: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, Core Writing Team, Lee, H., and Romero, J., IPCC, Geneva, Switzerland, 1–34, https://doi.org/10.59327/IPCC/AR6-9789291691647.001, 2023.

Isphording, R.: aus_precip_benchmarking: Benchmarking Precipitation in Regional Climate Models: Jupyter Notebooks (Python) (v1.0), Zenodo [code], https://doi.org/10.5281/zenodo.8365065, 2023.

Isphording, R. N.: ni422/aus_precip_benchmarking: Updated – Benchmarking Precipitation in Regional Climate Models: Jupyter Notebooks (Python) for the Minimum Standard Metrics and two simple example applications (v1.1), Zenodo [data set], https://doi.org/10.5281/zenodo.10668767, 2024.

Isphording, R. N., Alexander, L. V., Bador, M., Green, D., Evans, J. P., and Wales, S.: A Standardized Benchmarking Framework to Assess Downscaled Precipitation Simulations, J. Climate, 37, 1089–1110, https://doi.org/10.1175/JCLI-D-23-0317.1, 2024.

Jacob, D., Teichmann, C., Sobolowski, S., Katragkou, E., Anders, I., Belda, M., Benestad, R., Boberg, F., Buonomo, E., Cardoso, R. M., Casanueva, A., Christensen, O. B., Christensen, J. H., Coppola, E., De Cruz, L., Davin, E. L., Dobler, A., Domínguez, M., Fealy, R., Fernandez, J., Gaertner, M. A., García-Díez, M., Giorgi, F., Gobiet, A., Goergen, K., Gómez-Navarro, J. J., Alemán, J. J. G., Gutiérrez, C., Gutiérrez, J. M., Güttler, I., Haensler, A., Halenka, T., Jerez, S., Jiménez-Guerrero, P., Jones, R. G., Keuler, K., Kjellström, E., Knist, S., Kotlarski, S., Maraun, D., van Meijgaard, E., Mercogliano, P., Montávez, J. P., Navarra, A., Nikulin, G., de Noblet-Ducoudré, N., Panitz, H.-J., Pfeifer, S., Piazza, M., Pichelli, E., Pietikäinen, J.-P., Prein, A. F., Preuschmann, S., Rechid, D., Rockel, B., Romera, R., Sánchez, E., Sieck, K., Soares, P. M. M., Somot, S., Srnec, L., Sørland, S. L., Termonia, P., Truhetz, H., Vautard, R., Warrach-Sagi, K., and Wulfmeyer, V.: Regional climate downscaling over Europe: perspectives from the EURO-CORDEX community, Reg. Environ. Change, 20, 51, https://doi.org/10.1007/s10113-020-01606-9, 2020.

Jones, P. W.: First- and Second-Order Conservative Remapping Schemes for Grids in Spherical Coordinates, Mon. Weather Rev., 127, 2204–2210, https://doi.org/10.1175/1520-0493(1999)127<2204:FASOCR>2.0.CO;2, 1999.

Juneng, L. and Tangang, F. T.: Evolution of ENSO-related rainfall anomalies in Southeast Asia region and its relationship with atmosphere–ocean variations in Indo–Pacific sector, Clim. Dynam., 25, 337–350, 2005.

Juneng, L., Tangang, F., Chung, J. X., Ngai, S. T., Teh, T. W., Narisma, G., Cruz, F., Phan-Van, T., Ngo-Duc, T., Santisirisomboon, J., Singhruck, P., Gunawan, D., and Aldrian, E.: Sensitivity of Southeast Asia rainfall simulations to cumulus and air–sea flux parameterizations in RegCM4, Clim. Res., 69, 59–77, https://doi.org/10.3354/cr01386, 2016.

Kamworapan, S., Bich Thao, P. T., Gheewala, S. H., Pimonsree, S., and Prueksakorn, K.: Evaluation of CMIP6

GCMs for simulations of temperature over Thailand and nearby areas in the early 21st century, Heliyon, 7, e08263, https://doi.org/10.1016/j.heliyon.2021.e08263, 2021.

Kendall, M. G.: Rank Correlation Methods, 4th edn., Charles Griffin, London, 278–281, https://doi.org/10.1007/978-0-387-32833-1_211, 1975.

Kim, Y.-H., Min, S.-K., Zhang, X., Sillmann, J., and Sandstad, M.: Evaluation of the CMIP6 multi-model ensemble for climate extreme indices, Weather and Climate Extremes, 29, 100269, https://doi.org/10.1016/j.wace.2020.100269, 2020.

Knutti, R. and Sedláček, J.: Robustness and uncertainties in the new CMIP5 climate model projections, Nat. Clim. Change, 3, 369–373, https://doi.org/10.1038/nclimate1716, 2013.

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., and Meehl, G. A.: Challenges in Combining Projections from Multiple Climate Models, J. Climate, 23, 2739–2758, https://doi.org/10.1175/2009jcli3361.1, 2010.

Langenbrunner, B. and Neelin, J. D.: Analyzing ENSO Teleconnections in CMIP Models as a Measure of Model Fidelity in Simulating Precipitation, J. Climate, 26, 4431–4446, https://doi.org/10.1175/JCLI-D-12-00542.1, 2013.

Liu, Y. L., Alexander, L. V., Evans, J. P., and Thatcher, M.: Sensitivity of Australian rainfall to driving SST datasets in a variable-resolution global atmospheric model, J. Geophys. Res.-Atmos., 129, e2024JD040954, https://doi.org/10.22541/essoar.170913616.69157157/v1, 2024.

Maraun, D. and Widmann, M.: Statistical downscaling and bias correction for climate research, Cambridge University Press, https://doi.org/10.1017/9781107588783, 2018.

Masson, D. and Knutti, R.: Spatial-Scale Dependence of Climate Model Performance in the CMIP3 Ensemble, J. Climate, 24, 2680–2692, https://doi.org/10.1175/2011JCLI3513.1, 2011.

Mauritsen, T., Bader, J., Becker, T., Behrens, J., Bittner, M., Brokopf, R., Brovkin, V., Claussen, M., Crueger, T., Esch, M., Fast, I., Fiedler, S., Fläschner, D., Gayler, V., Giorgetta, M., Goll, D. S., Haak, H., Hagemann, S., Hedemann, C., Hohenegger, C., Ilyina, T., Jahns, T., Jimenéz-de-la-Cuesta, D., Jungclaus, J., Kleinen, T., Kloster, S., Kracher, D., Kinne, S., Kleberg, D., Lasslop, G., Kornblueh, L., Marotzke, J., Matei, D., Meraner, K., Mikolajewicz, U., Modali, K., Möbis, B., Müller, W. A., Nabel, J. E. M. S., Nam, C. C. W., Notz, D., Nyawira, S.-S., Paulsen, H., Peters, K., Pincus, R., Pohlmann, H., Pongratz, J., Popp, M., Raddatz, T. J., Rast, S., Redler, R., Reick, C. H., Rohrschneider, T., Schemann, V., Schmidt, H., Schnur, R., Schulzweida, U., Six, K. D., Stein, L., Stemmler, I., Stevens, B., von Storch, J.-S., Tian, F., Voigt, A., Vrese, P., Wieners, K.-H., Wilkenskjeld, S., Winkler, A., and Roeckner, E.: Developments in the MPI-M Earth System Model version 1.2 (MPI-ESM1.2) and Its Response to Increasing CO2, J. Adv. Model. Earth Sy., 11, 998–1038, https://doi.org/10.1029/2018MS001400, 2019.

McKenna, S., Santoso, A., Gupta, A. S., Taschetto, A. S., and Cai, W.: Indian Ocean Dipole in CMIP5 and CMIP6: characteristics, biases, and links to ENSO, Sci. Rep.-UK, 10, 11500, https://doi.org/10.1038/s41598-020-68268-9, 2020.

McSweeney, C. F., Jones, R. G., and Booth, B. B. B.: Selecting Ensemble Members to Provide Regional Climate Change Information, J. Climate, 25, 7100–7121, https://doi.org/10.1175/JCLI-D-11-00526.1, 2012.

McSweeney, C. F., Jones, R. G., Lee, R. W., and Rowell, D. P.: Selecting CMIP5 GCMs for downscaling over multiple regions, Clim. Dynam., 44, 3237–3260, https://doi.org/10.1007/s00382-014-2418-8, 2015.

Meehl, G. A., Boer, G. J., Covey, C., Latif, M., and Stouffer, R. J.: The Coupled Model Intercomparison Project (CMIP), B. Am. Meteorol. Soc., 81, 313–318, https://doi.org/10.1175/1520-0477(2000)081<0313:Tcmipc>2.3.Co;2, 2000.

Meyers, G., McIntosh, P., Pigot, L., and Pook, M.: The Years of El Niño, La Niña, and Interactions with the Tropical Indian Ocean, J. Climate, 20, 2872–2880, https://doi.org/10.1175/JCLI4152.1, 2007.

Mote, P., Brekke, L., Duffy, P. B., and Maurer, E.: Guidelines for constructing climate scenarios, Eos T. Am. Geophys. Un., 92, 257–258, https://doi.org/10.1029/2011EO310001, 2011.

NCI Australia: NCI ESGF Node, NCI Australia [data set], https://esgf.nci.org.au/projects/esgf-nci/, last access: 2024. TS14

Nearing, G. S., Ruddell, B. L., Clark, M. P., Nijssen, B., and Peters-Lidard, C.: Benchmarking and Process Diagnostics of Land Models, J. Hydrometeorol., 19, 1835–1852, https://doi.org/10.1175/JHM-D-17-0209.1, 2018.

Nguyen, P.-L., Bador, M., Alexander, L. V., Lane, T. P., and Funk, C. C.: On the Robustness of Annual Daily Precipitation Maxima Estimates Over Monsoon Asia, Frontiers in Climate, 2, 578785, https://doi.org/10.3389/fclim.2020.578785, 2020.

Nguyen, P.-L., Bador, M., Alexander, L. V., Lane, T. P., and Ngo-Duc, T.: More intense daily precipitation in CORDEX-SEA regional climate models than their forcing global climate models over Southeast Asia, Int. J. Climatol., 42, 6537–6561, https://doi.org/10.1002/joc.7619, 2022.

Nguyen, P.-L., Bador, M., Alexander, L. V., and Lane, T. P.: Selecting regional climate models based on their skill could give more credible precipitation projections over the complex Southeast Asia region, Clim. Dynam., 61, 3431–3452, https://doi.org/10.1007/s00382-023-06751-5, 2023.

Nguyen-Duy, T., Ngo-Duc, T., and Desmet, Q.: Performance evaluation and ranking of CMIP6 global climate models over Vietnam, J. Water Clim. Change, 14, 1831–1846, https://doi.org/10.2166/wcc.2023.454, 2023.

NOAA: NOAA OI SST V2 High Resolution Dataset, NOAA [data set], https://www.psl.noaa.gov/data/gridded/data.noaa.oisst.v2.highres.html, last access: 2024. TS15

Ossó, A., Craig, P., and Allan, R. P.: An assessment of CMIP6 climate signals and biases in temperature, precipitation and soil moisture over Europe, Int. J. Climatol., 43, 5698–5719, https://doi.org/10.1002/joc.8169, 2023.

Overland, J. E., Wang, M., Bond, N. A., Walsh, J. E., Kattsov, V. M., and Chapman, W. L.: Considerations in the Selection of Global Climate Models for Regional Climate Projections: The Arctic as a Case Study, J. Climate, 24, 1583–1597, https://doi.org/10.1175/2010JCLI3462.1, 2011.

Palmer, T. E., McSweeney, C. F., Booth, B. B. B., Priestley, M. D. K., Davini, P., Brunner, L., Borchert, L., and Menary, M. B.: Performance-based sub-selection of CMIP6 models for impact assessments in Europe, Earth Syst. Dynam., 14, 457–483, https://doi.org/10.5194/esd-14-457-2023, 2023.

Park, S., Shin, J., Kim, S., Oh, E., and Kim, Y.: Global Climate Simulated by the Seoul National University Atmosphere Model Version 0 with a Unified Convection Scheme (SAM0-UNICON),

J. Climate, 32, 2917–2949, https://doi.org/10.1175/JCLI-D-18-0796.1, 2019.

Perry, S. J., McGregor, S., Sen Gupta, A., England, M. H., and Maher, N.: Projected late 21st century changes to the regional impacts of the El Niño-Southern Oscillation, Clim. Dynam., 54, 395–412, https://doi.org/10.1007/s00382-019-05006-6, 2020.

Pimonsree, S., Kamworapan, S., Gheewala, S. H., Thongbhakdi, A., and Prueksakorn, K.: Evaluation of CMIP6 GCMs performance to simulate precipitation over Southeast Asia, Atmos. Res., 282, 106522, https://doi.org/10.1016/j.atmosres.2022.106522, 2023.

Planton, Y. Y., Guilyardi, E., Wittenberg, A. T., Lee, J., Gleckler, P. J., Bayr, T., McGregor, S., McPhaden, M. J., Power, S., Roehrig, R., Vialard, J., and Voldoire, A.: Evaluating Climate Models with the CLIVAR 2020 ENSO Metrics Package, B. Am. Meteorol. Soc., 102, E193–E217, https://doi.org/10.1175/BAMS-D-19-0337.1, 2021.

Power, S. B. and Delage, F. P. D.: El Niño–Southern Oscillation and Associated Climatic Conditions around the World during the Latter Half of the Twenty-First Century, J. Climate, 31, 6189–6207, https://doi.org/10.1175/JCLI-D-18-0138.1, 2018.

Qian, J.-H., Robertson, A. W., and Moron, V.: Diurnal cycle in different weather regimes and rainfall variability over Borneo associated with ENSO, J. Climate, 26, 1772–1790, https://doi.org/10.1175/JCLI-D-12-00178.1, 2013.

Rasmusson, E. M. and Wallace, J. M.: Meterological Aspects of the El Niño/Southern Oscillation, Science, 222, 1195–1202, http://www.jstor.org/stable/1691792 (last access: TS16), 1983.

Rayner, N. A., Parker, D. E., Horton, E. B., Folland, C. K., Alexander, L. V., Rowell, D. P., Kent, E. C., and Kaplan, A.: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century, J. Geophys. Res.-Atmos., 108, https://doi.org/10.1029/2002JD002670, 2003.

Ridder, N. N., Pitman, A. J., and Ukkola, A. M.: Do CMIP6 Climate Models Simulate Global or Regional Compound Events Skillfully?, Geophys. Res. Lett., 48, e2020GL091152, https://doi.org/10.1029/2020GL091152, 2021.

Rousseeuw, P. J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math., 20, 53–65, https://doi.org/10.1016/0377-0427(87)90125-7, 1987.

Saji, N. H., Goswami, B. N., Vinayachandran, P. N., and Yamagata, T.: A dipole mode in the tropical Indian Ocean, Nature, 401, 360–363, https://doi.org/10.1038/43854, 1999.

Schamm, K., Ziese, M., Becker, A., Finger, P., Meyer-Christoffer, A., Schneider, U., Schröder, M., and Stender, P.: Global gridded precipitation over land: a description of the new GPCC First Guess Daily product, Earth Syst. Sci. Data, 6, 49–60, https://doi.org/10.5194/essd-6-49-2014, 2014.

Seland, Ø., Bentsen, M., Olivié, D., Toniazzo, T., Gjermundsen, A., Graff, L. S., Debernard, J. B., Gupta, A. K., He, Y.-C., Kirkevåg, A., Schwinger, J., Tjiputra, J., Aas, K. S., Bethke, I., Fan, Y., Griesfeller, J., Grini, A., Guo, C., Ilicak, M., Karset, I. H. H., Landgren, O., Liakka, J., Moseid, K. O., Nummelin, A., Spensberger, C., Tang, H., Zhang, Z., Heinze, C., Iversen, T., and Schulz, M.: Overview of the Norwegian Earth System Model (NorESM2) and key climate response of CMIP6 DECK, historical, and scenario simulations, Geosci. Model Dev., 13, 6165–6200, https://doi.org/10.5194/gmd-13-6165-2020, 2020.

Sellar, A. A., Jones, C. G., Mulcahy, J. P., Tang, Y., Yool, A., Wiltshire, A., O'Connor, F. M., Stringer, M., Hill, R., Palmieri, J., Woodward, S., de Mora, L., Kuhlbrodt, T., Rumbold, S. T., Kelley, D. I., Ellis, R., Johnson, C. E., Walton, J., Abraham, N. L., Andrews, M. B., Andrews, T., Archibald, A. T., Berthou, S., Burke, E., Blockley, E., Carslaw, K., Dalvi, M., Edwards, J., Folberth, G. A., Gedney, N., Griffiths, P. T., Harper, A. B., Hendry, M. A., Hewitt, A. J., Johnson, B., Jones, A., Jones, C. D., Keeble, J., Liddicoat, S., Morgenstern, O., Parker, R. J., Predoi, V., Robertson, E., Siahaan, A., Smith, R. S., Swaminathan, R., Woodhouse, M. T., Zeng, G., and Zerroukat, M.: UKESM1: Description and Evaluation of the U. K. Earth System Model, J. Adv. Model. Earth Sy., 11, 4513–4558, https://doi.org/10.1029/2019MS001739, 2019.

Shukla, R. P., Tripathi, K. C., Pandey, A. C., and Das, I.: Prediction of Indian summer monsoon rainfall using Niño indices: a neural network approach, Atmos. Res., 102, 99–109, 2011.

Sobolowski, S., Somot, S., Fernandez, J., Evin, G., Maraun, D., Kotlarski, S., Jury, M., Benestad, R. E., Teichmann, C., Christensen, O. B., Bülow, K., Buonomo, E., Katragkou, E., Steger, C., Sørland, S., Nikulin, G., McSweeney, C., Dobler, A., Palmer, T., Wilcke, R., Boé, J., Brunner, L., Ribes, A., Qasmi, S., Nabat, P., Sevault, F., Oudar, T., and Brands, S.: EURO-CORDEX CMIP6 GCM Selection & Ensemble Design: Best Practices and Recommendations, Zenodo, https://doi.org/10.5281/zenodo.7673400, 2023.

Supari, S., Tangang, F., Salimun, E., Aldrian, E., Sopaheluwakan, A., and Juneng, L.: ENSO modulation of seasonal rainfall and extremes in Indonesia, Clim. Dynam., 51, 2559–2580, https://doi.org/10.1007/s00382-017-4028-8, 2018.

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., and Kumar, P.: Projected future changes in rainfall in Southeast Asia based on CORDEX–SEA multi-model simulations, Clim. Dynam., 55, 1247–1267, https://doi.org/10.1007/s00382-020-05322-2, 2020.

Taschetto, A. S., Gupta, A. S., Jourdain, N. C., Santoso, A., Ummenhofer, C. C., and England, M. H.: Cold Tongue and Warm Pool ENSO Events in CMIP5: Mean State and Future Projections, J. Climate, 27, 2861–2885, https://doi.org/10.1175/JCLI-D-13-00437.1, 2014.

Tatebe, H., Ogura, T., Nitta, T., Komuro, Y., Ogochi, K., Takemura, T., Sudo, K., Sekiguchi, M., Abe, M., Saito, F., Chikira, M., Watanabe, S., Mori, M., Hirota, N., Kawatani, Y., Mochizuki, T., Yoshimura, K., Takata, K., O'ishi, R., Yamazaki, D., Suzuki, T., Kurogi, M., Kataoka, T., Watanabe, M., and Kimoto, M.: Description and basic evaluation of simulated mean state, internal variability, and climate sensitivity in MIROC6, Geosci. Model Dev., 12, 2727–2765, https://doi.org/10.5194/gmd-12-2727-2019, 2019.

Trenberth, K. E. and Hoar, T. J.: El Niño and climate change, Geophys. Res. Lett., 24, 3057–3060, https://doi.org/10.1029/97GL03092, 1997.

U.S. DOE: Benchmarking Simulated Precipitation in Earth System Models Workshop Report, DOE/SC-0203, U. S. Department of Energy Office of Science, Biological and Environmental Research (BER) Program, Germantown, Maryland, USA, 2020.

Voldoire, A., Saint-Martin, D., Sénési, S., Decharme, B., Alias, A., Chevallier, M., Colin, J., Guérémy, J.-F., Michou, M., Moine, M.-P., Nabat, P., Roehrig, R., Salas y Mélia, D., Séférian, R., Valcke, S., Beau, I., Belamari, S., Berthet, S., Cassou, C., Cattiaux, J., Deshayes, J., Douville, H., Ethé, C., Franchistéguy, L., Geoffroy, O., Lévy, C., Madec, G., Meurdesoif, Y., Msadek, R., Ribes, A., Sanchez-Gomez, E., Terray, L., and Waldman, R.: Evaluation of CMIP6 DECK Experiments With CNRM-CM6-1, J. Adv. Model. Earth Sy., 11, 2177–2213, https://doi.org/10.1029/2019MS001683, 2019.

Volodin, E. M., Mortikov, E. V., Kostrykin, S. V., Galin, V. Y., Lykosov, V. N., Gritsun, A. S., Diansky, N. A., Gusev, A. V., and Yakovlev, N. G.: Simulation of modern climate with the new version of the INM RAS climate model, Izv. Atmos. Ocean. Phy.+, 53, 142–155, https://doi.org/10.1134/S0001433817020128, 2017.

Wang, B., Luo, X., and Liu, J.: How Robust is the Asian Precipitation–ENSO Relationship during the Industrial Warming Period (1901–2017)?, J. Climate, 33, 2779–2792, https://doi.org/10.1175/JCLI-D-19-0630.1, 2020.

Wang, Y.-C., Hsu, H.-H., Chen, C.-A., Tseng, W.-L., Hsu, P.-C., Lin, C.-W., Chen, Y.-L., Jiang, L.-C., Lee, Y.-C., Liang, H.-C., Chang, W.-M., Lee, W.-L., and Shiu, C.-J.: Performance of the Taiwan Earth System Model in Simulating Climate Variability Compared With Observations and CMIP6 Model Simulations, J. Adv. Model. Earth Sy., 13, e2020MS002353, https://doi.org/10.1029/2020MS002353, 2021a.

Wang, Z., Zhan, C., Ning, L., and Guo, H.: Evaluation of global terrestrial evapotranspiration in CMIP6 models, Theor. Appl. Climatol., 143, 521–531, https://doi.org/10.1007/s00704-020-03437-4, 2021b.

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

Xu, J., Koldunov, N. V., Xu, M., Zhu, X., Fraedrich, K., Jiang, X., Zhu, S., and Zhi, X.: Impacts of Indian Ocean Dipole–Like SST on Rice Yield Anomalies in Jiangsu Province, Front. Earth Sci., 8, 690, https://doi.org/10.3389/feart.2020.568365, 2021.

Yatagai, A., Kamiguchi, K., Arakawa, O., Hamada, A., Yasutomi, N., and Kitoh, A.: APHRODITE: Constructing a Long-Term Daily Gridded Precipitation Dataset for Asia Based on a Dense Network of Rain Gauges, B. Am. Meteorol. Soc., 93, 1401–1415, https://doi.org/10.1175/BAMS-D-11-00122.1, 2012 (data available at: https://www.chikyu.ac.jp/precip/english/index.html, last access: TS17).

Yukimoto, S., Kawai, H., Koshiro, T., Oshima, N., Yoshida, K., Urakawa, S., Tsujino, H., Deushi, M., Tanaka, T., Hosaka, M., Yabu, S., Yoshimura, H., Shindo, E., Mizuta, R., Obata, A., Adachi, Y., and Ishii, M.: The Meteorological Research Institute Earth System Model Version 2.0, MRI-ESM2.0: Description and Basic Evaluation of the Physical Component, J. Meteorol. Soc. Jpn. Ser. II, 97, 931–965, https://doi.org/10.2151/jmsj.2019-051, 2019.

Zheng, X., Li, Q., Zhou, T., Tang, Q., Van Roekel, L. P., Golaz, J.-C., Wang, H., and Cameron-Smith, P.: Description of historical and future projection simulations by the global coupled E3SMv1.0 model as used in CMIP6, Geosci. Model Dev., 15, 3941–3967, https://doi.org/10.5194/gmd-15-3941-2022, 2022.

Ziehn, T., Chamberlain, M. A., Law, R. M., Lenton, A., Bodman, R. W., Dix, M., Stevens, L., Wang, Y.-P., and Srbinovsky, J.: The Australian Earth System Model: ACCESS-ESM1.5, Journal of Southern Hemisphere Earth Systems Science, 70, 193–214, https://doi.org/10.1071/ES19035, 2020.

Remarks from the language copy-editor

CE1   Apostrophes have been added for this abbreviation when it has been used in the plural. This is our standard for plural abbreviations that end in lowercase letters. Sorry for the late change.

CE2   "CMS" was changed to "CM2" here. Please check. Sorry for the late change.

CE3   Should this be GPCC FDD?

CE4   Please note the language edits to this section.

Remarks from the typesetter

TS1   Please confirm.

TS2   Please confirm citation.

TS3   Please give an explanation of why this needs to be changed. We have to ask the handling editor for approval. Thanks.

TS4   Please give an explanation of why the table needs to be changed. We have to ask the handling editor for approval. Thanks.

TS5   Please confirm.

TS6   Please confirm citation.

TS7   Please confirm citation.

TS8   Please confirm that URL (https://www.ncei.noaa. gov/pub/data/cmb/ersst/v5/netcdf/) was changed to DOI.

TS9   Please make sure that all funders and grant numbers are mentioned in the financial support section. If the repeated funding information should be removed in the acknowledgements, please let us know.

TS10   Please check DOI.

TS11   Please provide DOI.

TS12   Please check DOI.

TS13   Please confirm reference list entry.

TS14   Please confirm reference list entry and add last access date.

TS15   Please confirm reference list entry and add last access date.

TS16   Please provide date of last access.

TS17   Please confirm addition and add last access date.