A Standardized Benchmarking Framework to Assess Downscaled Precipitation Simulations

Rachael N Isphording, Lisa V Alexander, Margot Bador, Donna Green, Jason P. Evans, Scott Wales

A Standardized Benchmarking Framework to Assess Downscaled Precipitation Simulations

Rachael N. Isphording,[a,b] Lisa V. Alexander,[a,b] Margot Bador,[c] Donna Green,[a,b] Jason P. Evans,[a,b] Scott Wales[d]

[a] *Climate Change Research Centre, University of New South Wales, Sydney, New South Wales, Australia*

[b] *ARC Centre of Excellence for Climate Extremes, University of New South Wales, Sydney, New South Wales, Australia*

[c] *CECI Université de Toulouse, CERFACS/CNRS, Toulouse, France*

[d] *Bureau of Meteorology, Melbourne, Victoria, Australia*

Corresponding author: Rachael Isphording, r.isphording@unsw.edu.au

ABSTRACT

Presently, there is no standardized framework or metrics identified to assess regional climate model precipitation output. Because of this, it can be difficult to make a one-to-one comparison of their performance between regions, studies, or against coarser resolution global climate models. To address this, we introduce the first steps towards establishing a dynamic, yet standardized, benchmarking framework that can be used to assess model skill in simulating various characteristics of rainfall. Benchmarking differs from typical model evaluation in that it requires that performance expectations are set *a priori*. This framework has innumerable applications to underpin scientific studies that assess model performance, inform model development priorities, and aid stakeholder decision-making by providing a structured methodology to identify fit-for-purpose model simulations for climate risk assessments and adaptation strategies. While this framework can be applied to regional climate model simulations at any spatial domain, we demonstrate its effectiveness over Australia using high-resolution, 0.5° x 0.5° simulations from the CORDEX-Australasia ensemble. We provide recommendations for selecting metrics and pragmatic benchmarking thresholds depending on the application of the framework. This includes a top tier of Minimum Standard Metrics to establish a minimum benchmarking standard for ongoing climate model assessment. We present multiple applications of the framework using feedback received from potential user communities and encourage the scientific and user community to build on this framework by tailoring benchmarks and incorporating additional metrics specific to their application.

SIGNIFICANCE STATEMENT

We introduce a standardized benchmarking framework for assessing the skill of regional climate models in simulating precipitation. This framework addresses the lack of a uniform approach in the scientific community and has diverse applications in scientific research, model development, and societal decision-making. We define a set of minimum standard metrics to underpin ongoing climate model assessments that quantify model skill in simulating fundamental characteristics of rainfall. We provide guidance for selecting metrics and defining benchmarking thresholds, demonstrated using multiple case studies over

Australia. This framework has broad applications for numerous user communities and provides a structured methodology for the assessment of model performance.

1. Introduction

The Sixth Assessment Report (AR6) by the Intergovernmental Panel on Climate Change (IPCC) highlights the exacerbation of water-related crises in a changing climate. According to this report, nearly half of the global population is facing annual, severe water shortages, and over 50% of disaster events since 1970 are due to rainfall extremes, including floods and droughts (Caretta et al., 2022). Despite the widespread impact of these water crises and rainfall-related disasters driving international efforts to adapt to changing rainfall patterns, global climate models (GCMs) still struggle to simulate many aspects of rainfall. Most notably attributed to GCM rainfall biases are model parameterizations and coarse model resolution that cannot resolve key thermodynamic and dynamic processes relevant to rainfall simulation (Flato et al. 2013). There has been improvement across generations of the Coupled Model Intercomparison Project (CMIP) (Flato et al. 2013; IPCC, 2021). However, these improvements are heterogeneous across regions, timespans, and rainfall characteristics. Many studies detail sustained problems in how GCMs simulate tropical rainfall (Oueslati and Bellon 2015; Fiedler et al. 2020), rainfall extremes (Sillmann et al. 2013), seasonal rainfall patterns (Dunning et al. 2017), long-term annual precipitation trends (Vicente-Serrano et al. 2022), the diurnal cycle (Covey et al. 2016), and the 'drizzle bias' where models tend to rain too little, too often (Dai 2006; Chen et al. 2021). A lack of consistency in the methods or metrics used to quantify models' skill in simulating different aspects of rainfall makes it difficult to make a one-to-one comparison between studies, or efficiently track progress across CMIP generations. However, recent efforts have prompted standardization in assessing how GCMs simulate rainfall (Eyring et al. 2016; Baker and Taylor 2016; Eyring et al. 2019; Lauer et al. 2020; U.S. DOE 2020; Ahn et al. 2023).

A standardized benchmarking framework to assess simulated precipitation in GCMs across different generations of models was outlined in U.S. DOE (2020). They identify a set of performance metrics that can serve as a baseline to gauge model performance in simulating the spatial distribution, seasonal cycle, temporal variability, observed distributions of intensity and frequency, wet extremes, and drought. This 'benchmarking' framework was primarily established to better gauge progress across CMIP generations. Benchmarking

differs from standard model evaluation in that benchmarking requires performance expectations to be defined *a priori* (Abramowitz 2005, 2012). Since its publication, many additional studies have investigated the diurnal cycle (Tang et al. 2021), temporal variability (Ahn et al. 2022), and daily distributions of rainfall (Martinez-Villalobos et al. 2022) in GCMs, underpinned by the work presented in U.S. DOE (2020).

International efforts to coordinate the production and evaluation of dynamically downscaled models and reanalyses (Giorgi and Gutowski 2015) have allowed for far greater accessibility of high-resolution, regional climate model (RCM) simulations to the scientific community and regional decision-makers. While there has been progress to standardize how GCM simulations of rainfall are assessed, efforts to standardize the assessment of RCMs have been regionally heterogeneous. Presently, there is no standardized framework or metrics identified to assess RCM precipitation output. Previous studies have shown that RCMs tend to differ in magnitude and spatial variability when compared to GCMs. However, these studies are frequently limited in the scope of performance metrics evaluated, and commonly only assess a handful of indices using the ensemble mean instead of individual model performance. There are also regional inconsistencies where RCMs tend to run wetter than their forcing GCM [e.g. over Europe (Boé et al. 2020) and south-east Asia (Nguyen et al. 2022)], and RCMs tend to run drier over Africa (Dosio et al. 2021). However, there is little consistency between the metrics used in these evaluation studies which makes it difficult to make a one-to-one comparison or properly assess RCM performance in simulating precipitation.

To address this inconsistency, we present a standardized benchmarking framework underpinned by the work presented in U.S. DOE (2020), to holistically assess the skill of downscaled precipitation simulations. This framework could be used to guide scientific studies to assess model performance and inform model development priorities, and for stakeholders to identify fit-for-purpose model simulations to underpin climate risk assessments and inform climate adaptation strategies.

This paper is organized as follows: Section 2 presents the benchmarking framework, with Sections 2b and 2c describing tiers of performance metrics and Section 2d describing recommendations for defining *a priori* benchmarking thresholds. Section 3 showcases multiple applications of the benchmarking framework to the CORDEX-Australasia ensemble

prefaced by a description of the data used and pre-processing steps completed. We summarize and discuss key points in Section 4.

2. The Benchmarking Framework

Model evaluation and benchmarking differ in significant ways. Model evaluation gauges how well a model simulates a given variable compared to observations (Flato et al., 2013). Benchmarking seeks to understand how well a model should perform by defining performance expectations *a priori* (Abramowitz 2005, 2012). Benchmarking reframes traditional model evaluation by incorporating predefined performance thresholds. This step is equally beneficial and challenging, as we discuss in Section 2.d. An established benchmarking framework already exists for land surface models (Best et al. 2015), and early work has been completed to benchmark precipitation in GCMs (see Section 1). While precipitation encompasses liquid precipitation (rainfall) and solid precipitation (snow, hail, etc.), we use rainfall synonymously with precipitation in this paper as solid precipitation is negligible for our case study region of Australia.

We have developed this framework to establish a consistent, systematic, foundational methodology to quantify RCM skill in simulating precipitation for various user communities. The Benchmarking Framework (BMF) consists of two tiers of metrics: the first tier defines a set of minimum standard performance metrics, and the second tier encourages user-defined metrics relevant to the study. The BMF can be applied to RCM simulations across any region at any spatial resolution. This framework can be used by stakeholder user communities to distill a subset of fit-for-purpose model simulations or subset model simulations to develop storylines for informed decision-making. Scientific and research user communities can use this framework for innumerable applications to highlight gaps in model performance and guide model development priorities. Model developers can use the BMF as a first step to efficiently assess model performance, broadly quantify biases and uncertainties, and identify the sources of these uncertainties. Model developers and evaluators can also use the BMF to test the impact of higher spatial resolutions (Bador et al. 2020a; Nishant et al. 2022) or bias correction techniques (Casanueva et al. 2016), quantify model progress across generations (Alexander and Arblaster 2009; Flato et al., 2013; Sillmann et al. 2013; Alexander and Arblaster 2017; Fiedler et al. 2020), test different model set-ups and parameterizations (Ji et al. 2014), or assess model performance when different downscaling techniques are used, such

as spectral nudging, statistical downscaling, or machine learning (Hobeichi et al. 2023). Additionally, scientific researchers can use this framework to underpin studies assessing regime- and process-oriented properties of rainfall, such as frontal precipitation (Berry et al. 2011) and teleconnections (Fita et al. 2017), respectively. These more complex assessments of simulated rainfall are essential for better understanding model biases and limitations, improving future simulations of rainfall, and improving the scientific community's physical interpretation of performance metrics.

a. Observational Uncertainty

A common, yet unavoidable, problem in traditional model evaluation to quantify model skill in simulating precipitation is observational uncertainty (Evans et al. 2016; Gibson et al. 2019); this is true for benchmarking as well. It is well-known that there are vast differences in global observations of precipitation (Sun et al. 2018), particularly for extreme precipitation (Herold et al. 2017; Bador et al. 2020b). This tends to be true regionally as well (see Supplemental Figures 1-3; Contractor et al. 2015; Yin et al. 2015), especially as there are significant regional heterogeneities in data quality and spatial and temporal availability (Alexander et al. 2019). Because of these regional differences in observational data quality and coverage, there is not one best way to quantify observational uncertainty. We acknowledge that further research is required in this area, and it is likely better to quantify observational uncertainty differently depending on the spatial/temporal scale and region of study. In the following sections, we discuss using a single observational product to quantify model skill for simplicity, acknowledging that real-world applications of the benchmarking framework should incorporate multiple observational products (see the Supplemental Material) or other methods to quantify observational uncertainty.

b. Minimum Standard Metrics

We first define a set of foundational, minimum-standard metrics (MSMs) that address very fundamental characteristics of rainfall to provide consistency, simplicity, and pragmatism in how RCM skill is measured (Table 1). We define four, equally weighted MSMs that quantify mean-state biases in model performance with respect to the amount of rainfall, the spatial distribution of rainfall, the timing of rainfall, and the temporal variability of rainfall. The MSMs are calculated using area-weighted, average total rainfall, providing a well-rounded synopsis of RCM performance and mean-state biases in simulating rainfall that

accounts for the different sizes of grid cells across latitudes. Before more complex processes or rainfall characteristics are assessed, a model should meet performance expectations (i.e. benchmarks, see Section 2.d.) for all the MSMs.

Fundamental Rainfall Characteristic	Quantifying Metric
How much does it rain?	Mean Absolute Percentage Error (MAPE)
Where does it rain?	Spatial Correlation
When does it rain?	Seasonal Cycle
How does rainfall ***change over time***?	Direction of a Significant Trend

Table 1. The minimum standard metrics (MSMs) quantify very fundamental characteristics of rainfall. These metrics should be calculated based on area-weighted, average total rainfall for the region of interest.

To quantify the mean-state model skill in simulating the amount of rainfall, we recommend the mean absolute percentage error (MAPE) where n is the number of grid cells in the spatial domain. (Eq. 1).

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|\, model_i - obs_i \,|}{obs_i}$$

This provides a metric that is robust against large biases in a small region of the study domain and expresses the relative error of the model simulation compared to observations. Because the MAPE can quickly be converted to a percentage error, it is also easy to interpret by non-research communities. To quantify the mean-state spatial distribution of simulated rainfall, we recommend using the spatial correlation tested against the observational product. The spatial correlation is a standard metric for quantifying the agreement of spatial patterns between two datasets and ranges from 0 to 1. Because both metrics can be thought of as a percentage error of different rainfall characteristics, they are easy to compare. Further, the definition of benchmarking thresholds is very intuitive. For example, if a user wants to identify models that capture the spatial variability across at least 65% of their study domain

with a wet/dry bias of no more than 70% compared to observations, then the user would define a benchmarking threshold for the spatial correlation as >= 0.65 and the threshold for the MAPE as <=0.7 (see Section 2.d for more on defining benchmarking thresholds).

To quantify model skill in simulating the timing of rainfall, we prescribe a simple quantification of the seasonal cycle that emphasizes quantifying model skill in simulating the phase of rainfall. We recommend calculating the climatological total monthly precipitation across the study domain and ranking the months from driest to wettest for the observational product and the RCM simulations. Then, for a unimodal (bimodal) seasonal cycle, we use the three (six) wettest and driest months of the observational product to quantify the phase of the seasonal cycle. 100% of the three (six) wettest observed months must be among the six wettest modelled months, and 100% of the three (six) driest observed months must be among the six driest modelled months. Models where rainfall peaks or troughs slightly out of phase with observations will likely still pass this metric. Again, the MSMs are intended to highlight any fundamental flaws in the simulation of basic characteristics of rainfall. This metric will flag simulations where the seasonal cycle is inverted or largely out of phase with observations. This metric establishes a consistent and simple assessment among studies with flexibility appropriate for the large differences in rainfall seasonality between regions. While this metric does neglect the amplitude of the seasonal cycle, the purpose here is to broadly quantify model skill in simulating the timing of precipitation (Table 1). More detailed assessments of the seasonal cycle, including the amplitude, can be incorporated in further steps as outlined in the Seasonality section of the Versatility Metrics (see Section 2.c.1 and Section 4.b).

For a low-level quantification of the temporal trend, we recommend using the direction of a significant trend, tested using at least at a 10% significance level, in the time series of the reference observational dataset using at least 30 years of data. Ideally, a longer time series will be used if data are available. We recommend using standard, non-parametric statistical methods that do not assume a Gaussian distribution including the Thiel-Sen Trend tested for significance using the Mann-Kendall significance test (Hamed 2008). This metric tests the direction of the simulated trend in precipitation, neglecting the magnitude of the trend. If the model does not have a significant trend in the same direction as the observational dataset, then the model would not meet minimum performance standards. The models that meet

performance expectations for all the MSMs can then be assessed against the more complex metrics in the second tier (Figure 1) based on the need and/or scientific interest of the user.

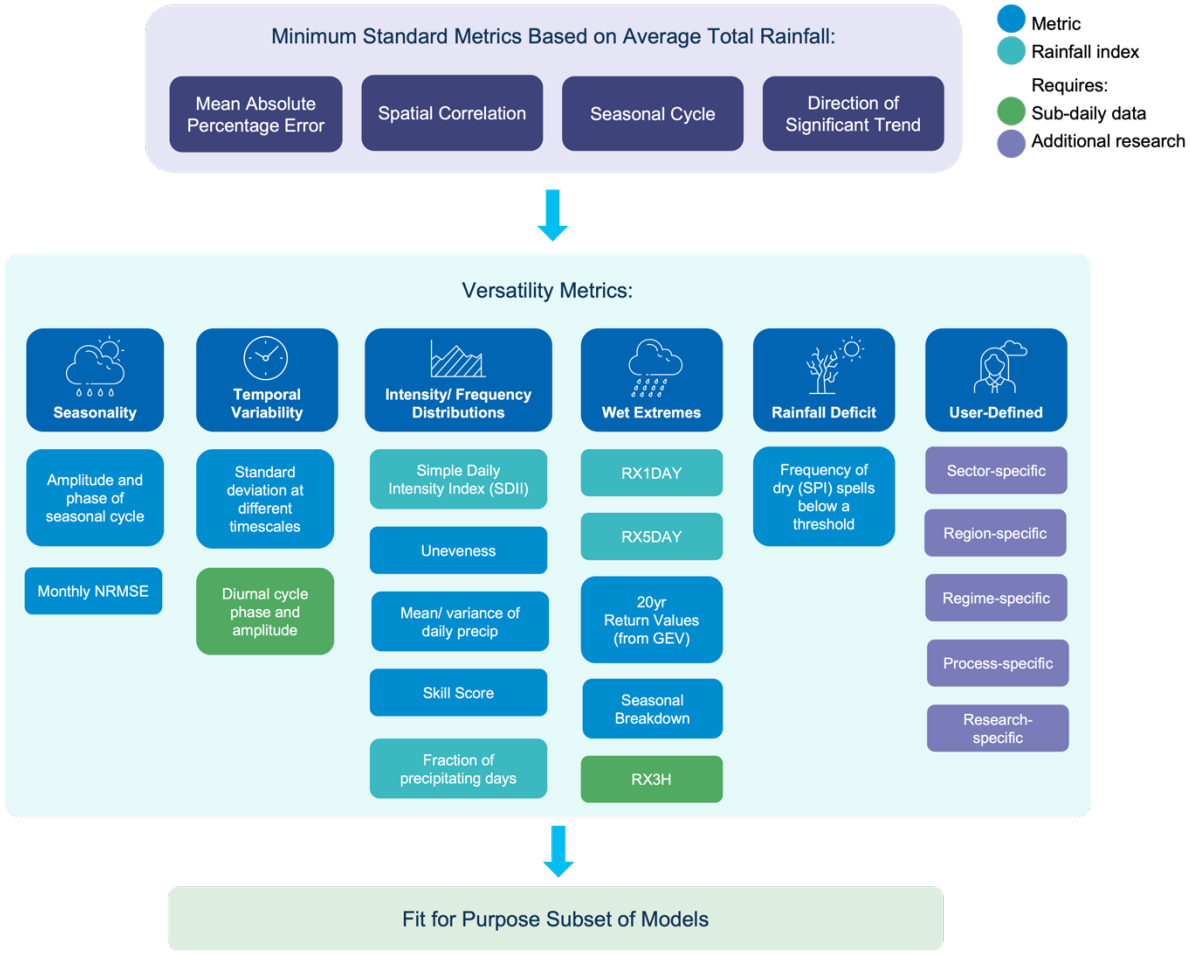

Fig. 1. Schematic for the tiers of metrics for the benchmarking framework, underpinned by U.S. DOE (2020). The Minimum Standard Metrics quantify very basic characteristics of rainfall. The second-tier metrics offer a non-exhaustive list of metrics to further assess additional characteristics of rainfall. These were largely consolidated by a group of international experts specializing in various aspects of modeled and observed rainfall (U.S. DOE, 2020) and have been updated for downscaled rainfall. We encourage users to incorporate additional metrics relevant to their application of the framework.

c. Versatility Metrics

Quantifying model skill in simulating regional rainfall is very complex, regardless of the region of interest, aspect of rainfall, or the spatial or temporal scale. The second tier, also referred to as the Versatility Tier, provides a non-exhaustive list of recommended metrics and indices to quantify model skill across rainfall characteristics (Figure 1). These metrics were largely consolidated from the scientific literature by a group of international experts in U.S. DOE (2020) but were amended here to apply to downscaled data. Primarily, we have added

the User-Defined column to explicitly address the diverse applications of RCMs across stakeholder and research communities. RCMs are commonly used to inform climate adaption planning and risk assessments or research atmospheric phenomena that are not resolved at the coarser spatial resolutions of GCMs. We also explicitly incorporate user-defined benchmarking thresholds (see Section 2.d.) to identify fit-for-purpose simulations wherein DOE (2020) established a framework intended to gauge model performance across CMIP generations. We encourage users to apply other metrics and develop additional techniques to better quantify model skill for any standard characteristic of rainfall (e.g., seasonality, temporal variability, intensity/frequency distributions, wet extremes, or rainfall deficits). Additionally, the User-defined column prompts users to incorporate additional metrics for more complex aspects of rainfall and broader characteristics of the water cycle. For example, users can define and develop metrics based on the region and/or sector of interest, to quantify a specific rainfall regime or process, or incorporate any other metric or technique that is relevant to the research question. It is expected that this collection of recommended metrics will be updated as further research is completed.

1) SEASONALITY

The low-level quantification of the seasonal cycle used in the MSMs will not be sufficient for many applications of the BMF especially in regions largely impacted by inter-annual and/or decadal variability. Many users will require a deeper analysis that better captures the amplitude, phase, onset/cessation, or other characteristics of rainfall seasonality. For instance, sector users in agriculture (Basso et al. 2012), hydroelectric power supply (de Jong et al. 2018), or water resource management (Barua et al. 2013) could be very interested in assessing long-term variability in the seasonal cycle and would benefit from applying more advanced techniques to calculate the onset and cessation of the rainy season(s). This could include calculating the cumulative rainfall anomaly (Dunning et al. 2017, 2016) or setting a fixed threshold for continued rainfall as is frequently used in the agriculture sector (Liebmann et al. 2012).

Additionally, a better scientific understanding of the physical drivers of regional rainfall seasonality–and a quantification of how well models capture the influence of these drivers–would require a more complete breakdown of the seasonal cycle. Therefore, it could be

beneficial to employ Harmonic Analysis (Wang and LinHo 2002) to quantify the amplitude and phase of the seasonal cycle for efficient comparison against observations. Further, due to the vast regional variability in rainfall seasonality, many studies have shown the benefits of tailoring metrics and analysis techniques to the region of interest to quantify rainfall seasonality (Seregina et al. 2019; Dey et al. 2021).

2) Temporal Variability

Quantifying model skill in simulating the temporal variability of rainfall is challenging as rainfall varies at timescales ranging from sub-daily to multi-decadal. The simplest way to quantify temporal variability is to calculate the standard deviation at different timescales, although this provides limited insight into model performance. Advanced methods can yield a more comprehensive assessment of model performance at different timescales. As an example, Covey et al. (2016) propose the "harmonic dial" diagram, created by vector spatial averaging Fourier amplitude and phases across land and ocean separately, to assess the diurnal cycle of rainfall simulations. Using this method, they find that members of the CMIP5 ensemble tend to rain too early in the day. Other methods such as Harmonic Analysis and Principal Component Analysis (EOFs) are frequently used to distill modes of temporal variability such as those from the El Niño-Southern Oscillation (ENSO), the Atlantic Multi-decadal Oscillation (AMO) the Indian Ocean Dipole (IOD), or long-term trends (Cai et al. 2011; Roundy 2015; Xiao et al. 2015; Yang et al. 2015; Chen and Tung 2018; Tippett and L'Heureux 2020). Ahn et al. (2022) also introduce techniques to quantify temporal variability at sub-daily to interannual scales using power spectra analysis and time-averaging, highlighting that these robust methods are not sensitive to differences in observations. These techniques can be incredibly effective ways to investigate different drivers of temporal rainfall variability.

3) Intensity and Frequency Distributions

While the MSMs provide a high-level quantification of how well models simulate fundamental characteristics of rainfall, they do not capture the full distribution of rainfall. Quantifying the intensity and frequency distribution of rainfall provides a deeper insight into the strengths and weaknesses of simulated rainfall. This can also provide insight into the

causes of biases and limitations (i.e. model set-ups and parameterizations). RCM performance in simulating the distribution of rainfall can be quantified using many established techniques. For instance, established skill scores can be used to quantify how well models simulate the distribution of rainfall compared to observations (Perkins et al. 2007; Nguyen et al. 2022). Martinez-Villalobos et al. (2022) outline the strengths and limitations of several metrics that quantify GCM performance in simulating the distribution of daily rainfall that could be applied to RCMs. This study also highlights the necessity of incorporating multiple metrics in studies of modeled rainfall as model performance varies across metrics.

4) Wet Extremes

Rainfall extremes are of high importance for stakeholder decision-making and climate adaptation planning but are not explicitly captured in the MSMs. There are numerous climate indices defined by the World Meteorological Organisation (WMO) and the World Climate Research Program (WCRP) such as those by the former Expert Team on Climate Change Detection and Indices (ETCCDI) to quantify the intensity, severity, and frequency of moderately extreme rainfall (Zhang et al. 2011; Alexander et al. 2019). However, users should be thoughtful in the selection and interpretation of these climate indices. Percentile indices, such as very wet days as defined by rainfall in the 95th percentile of a given time period (R95p) and extremely wet days as defined by rainfall in the 99th percentile (R99p), are particularly subjective to the reference period (Alexander et al. 2019; Bador et al. 2020b). Further, it is pertinent to acknowledge the large variability between rainfall extremes in global observational datasets (Bador et al., 2020b) and the impact of different pre-processing steps used in the creation of gridded observational datasets (Alexander et al., 2019). When using the BMF to benchmark model performance in simulating wet extremes, it is highly recommended to incorporate multiple observational datasets to quantify model performance (see Supplemental Figures 2-3).

5) Rainfall Deficit

While impacts are felt during periods of severe dryness, a deficit of rainfall is not always extreme. It is very dependent on the spatio-temporal scale in which a deficit occurs. There are many established metrics and methods to quantify how well models simulate a lack of

rainfall, extreme or otherwise. In this section, we focus primarily on an extreme deficit of rainfall or conditions that lead to meteorological drought. Metrics, indices, and thresholds used to quantify meteorological drought vary based on the study region. However, one universal index that works well globally (WMO and GWP 2016) is the Standardized Precipitation Index (SPI) (Mckee et al. 1993). The SPI is a measure of how much rainfall deviates from the long-term average and can be calculated at different time spans. Then, users can calculate how often rainfall falls below a given threshold; thresholds are standardized to indicate the severity of drought (see Section 3.d. for an example). Another commonly used index to study meteorological dryness is the Maximum Annual Number of Consecutive Dry Days (CDD) which quantifies the duration of dry spells (Chu et al. 2010; Haylock and Goodess 2004). However, this index is not appropriate for regions with a distinct dry season or general arid climate (see Alexander et al., 2019).

6) USER-DEFINED

All performance metrics discussed previously are limited in scope in that they only require precipitation data for their calculation. The User-Defined column explicitly encourages users to incorporate and develop additional metrics and techniques to evaluate rainfall and broader aspects of the water cycle. This could include established rainfall indices that incorporate other meteorological variables, such as the Standardized Precipitation and Evapotranspiration Index (SPEI) which is calculated using temperature and rainfall data and is commonly used to study drought (Spinoni et al. 2021). This could also include sector- or research-specific metrics or indices. For instance, stakeholder user communities can incorporate user-defined metrics or indices to benchmark aspects of rainfall that are specific to their decision-making process. This can facilitate broader opportunities for co-designed research and help stakeholders optimize the utility of downscaled data and climate information. Additionally, users can incorporate metrics specific to their application of the BMF. For instance, if the BMF were used to underpin added value studies (Choudhary et al. 2019; Torma et al. 2015; De Haan et al. 2015; Di Virgilio et al. 2020; Solman and Blázquez 2019; Rummukainen 2016), then users would ideally incorporate established added value metrics (Kanamitsu and Dehaan 2011; Di Luca et al. 2012; Di Virgilio et al. 2020; Ciarlo` et al. 2020) to quantify RCM performance compared to GCMs. Further, certain methods have been shown to better capture intricate rainfall characteristics in different regions. For

example, Seregina et al. (2019) found that replacing Fourier harmonics (Wang and LinHo, 2002) with a low-pass Lanczos Filter better captured the complex seasonality of rainfall in the Greater Horn of Africa. There are many established methods and metrics that can be incorporated when using the BMF to quantify model performance in simulating rainfall beyond what is explicitly listed in Figure 1.

Additionally, there are many aspects of rainfall that require additional research to determine appropriate benchmarking metrics. For instance, as computing capabilities improve, we can simulate rainfall at higher resolutions. This facilitates the development and application of methods and metrics that are effective in quantifying model performance in simulating complex rainfall regimes, such as frontal systems or mesoscale convective systems, and rainfall processes, such as teleconnections and orographic rainfall. For example, methods that are frequently used in forecast verification can be leveraged for RCM assessment at higher resolutions such as the fraction skill score to assess the distribution of precipitation in convective-permitting models (Prein et al. 2015) or storm tracking methods to identify the source of simulated precipitation (Feng et al. 2021). There are many benefits to further developing metrics to quantify these complexities of rainfall. Outside of improving our scientific understanding of these processes, scientists can identify parameterizations and other model structures that cause biases and other erroneous representations of different rainfall characteristics. We strongly encourage users to incorporate, develop, and test other performance metrics to improve ongoing benchmarking capabilities.

d. Defining a Benchmark

Benchmarking requires that model performance expectations are defined prior to the analysis. Therefore, we must define performance benchmarks–the criteria that will be used to assess model performance–and benchmarking thresholds–how well a model should score against a given metric. There is no one-size-fits-all definition for performance benchmarks. The benchmarking definition, and associated benchmarking thresholds, that define acceptable model performance should be informed by strong scientific reasoning, the scientific research question, the region or sector of interest, and the general purpose for benchmarking model performance.

Benchmarks can be defined more objectively for some metrics and applications than for others. For instance, as is done in standard model evaluation, we can use observational

products (see Section 3) or a range of observational uncertainty from multiple observational products (Martinez-Villalobos et al., 2022; see Supplemental Figures 1-3) as the benchmark and benchmarking thresholds as appropriate in certain cases. However, other performance metrics must be informed more subjectively using strong scientific reasoning based on the application of the framework. This is particularly important when working with stakeholder user communities to identify fit for purpose simulations. Scientific expertise should be used to define *reasonable* model performance expectations that fall within current capabilities of the modelling community. For the MSMs, benchmarking thresholds that are generous in the definition of 'reasonable performance' are encouraged because these metrics are intended to identify models with fundamental shortcomings in simulating precipitation.

Due to the diverse applications of the BMF, the range of effective benchmarking definitions and thresholds is also vast. For instance, model evaluators can use the BMF to gauge model improvement across generations. A reasonable benchmarking definition here could be that models must perform at least as well as the previous generation of models. Since GCM-RCM pairings typically change across generations, the users could define the benchmarking thresholds as the range of performance from the ensemble of the previous generation(s) against a selection of metrics. Likewise, model developers could use the BMF in a similar way to adjust model set-ups and parameterizations in response to performance against the MSMs (Table 1). Further, the BMF can underpin 'added value' studies (Choudhary et al. 2019; Torma et al. 2015; De Haan et al. 2015; Di Virgilio et al. 2020; Solman and Blázquez 2019; Rummukainen 2016) that seek to quantify the benefits of downscaling GCMs. For these studies, the benchmarking definition and thresholds could be that RCMs must perform at least as well as their forcing GCM against a given set of metrics.

We do not seek to prescribe the best definition of a benchmark or the associated benchmarking thresholds. Instead, we provide guidance, emphasizing again that benchmarks should be informed by the purpose of applying the benchmarking framework and should be fit for purpose. As different user communities apply the BMF, this process may become more prescriptive over time. In the next section, we use scientific expertise to translate stakeholder performance needs into reasonable definitions of performance benchmarks using the CORDEX-Australasia ensemble as a case study.

3. Benchmarking the CORDEX-Australasia Ensemble

In this section, we showcase an application of the Benchmarking Framework over terrestrial Australia where we have confidence in our observational record (defined in Section 3a) using 24 simulations from the CORDEX-Australasia ensemble (Table 3).

Using feedback from discussions with potential users in humanitarian aid, water resource management, and the scientific research community, we present two simplified hypothetical applications of the framework. While these stakeholders specifically had very different concerns depending on their location in Australia, in-house scientific resources, and the aspect of their decision-making in question, we distilled their feedback to create a simplified case study to test the BMF. Broadly speaking, these stakeholders wanted models that best captured Australia's highly variable rainfall seasonality (with equal emphasis on the spatial variability, timing, and quantity of rainfall) and the frequency of rainfall deficits. We present these as two different case studies for simplicity. For the MSMs, these stakeholders emphasized the need for models that are skilled in capturing the spatial distribution of rainfall. Specifically, it was important for these stakeholders to know if rain would fall in a particular watershed or catchment area to use in allocating water resources or where areas would not receive rainfall and may need more aid or water conservation actions. It was less important to identify models that have a large wet or dry bias as these stakeholders are accustomed to Australia's characteristically extreme wet and dry periods.

In the following sections, we translate these qualitative stakeholder needs into quantitative model performance expectations. The performance expectations for the MSMs will be the same for both hypotheticals. Then, one application will seek to identify models better at simulating the amplitude and phase of the seasonal cycle, and the other application will seek to identify models better at simulating the frequency of rainfall deficits over Australia. Again, the benchmarks used to test the MSMs are not meant to be too restrictive. At this stage, we only want to remove models that have low-level, systematic biases in simulating fundamental characteristics of rainfall. The benchmarks should reflect the stakeholder needs while also incorporating scientific expertise to inform reasonable model performance expectations. We will use a regionally developed observational product for Australia to quantify model skill, noting again the need to account for observational uncertainty in real applications of the BMF (see Section 2.a.; Supplemental Material).

a. Data and Preprocessing

We use daily precipitation from 24 simulations of the CORDEX-Australasia ensemble that includes 7 RCMs forced by 10 GCMs (Evans et al. 2021) and daily precipitation observations from the Australian Gridded Climate Dataset (Jones et al. 2009) for 1976-2005 (Table 2). By only using the AGCD product instead of global, gridded observational products (i.e. Roca et al. 2019; see Supplemental Material) we can assess RCM performance at a higher resolution. We improve confidence in our observational dataset by creating a quality mask that removes grid points not containing at least one observing station based on the Global Historical Climatology Network daily (GHCN-daily) database (see Figure 2). This removes grid points where rainfall observations are artificially created through the interpolation algorithms used to create the gridded dataset. We also remove grid points that contain more than 50% ocean.

First, all datasets were interpolated to a Cartesian coordinate system with a spatial resolution of 0.5° x 0.5° using first-order conservative interpolation to better capture the spatial discontinuity of precipitation (Jones 1998). This meant interpolating the AGCD data and some of the CORDEX simulations to a coarser resolution so all datasets were on a common grid. Then, we used Climpact, an open-source software package developed under the auspices of the World Meteorological Organization (WMO) (Alexander and Herold, 2015; see https://climpact-sci.org), to calculate a set of 51 climate indices for the AGCD data (Isphording and Liu, 2023) and the CORDEX-Australasia ensemble (Isphording et al. 2023). This order of operations is recommended as it has been shown to be less sensitive to the interpolation methods used in regridding (Avila et al. 2015). It is also recommended that a gridded observational product is used to assess model performance against the MSMs because RCMs provide area averaged values at each grid point; it is therefore pertinent that a fair assessment of model performance is based on a comparison to observed area-averaged values. See the Supplemental Figure S1 for additional guidance in selecting observational products to use for benchmarking.

Institute	RCM	Driving CMIP5 GCM	Available Experiments	Available Time Period
CSIRO	CCAM-1704	ACCESS1-0	historical, RCP 4.5, RCP 8.5	1960-2099
		CNRM-CM5		
		GFDL-ESM2M		

		HadGEM2-CC		
		MIROC5		
		NorESM1-M		
	CCAM-2008	ACCESS1-0	historical, RCP 4.5, RCP 8.5	1960-2099
		CanESM2		
		GFDL-ESM2M		
		MIROC5		1961-2099
		NorESM1-M		1960-2099
CLMcom-HZG	CCLM5-0-15	HadGEM2-ES	historical, RCP 8.5	1950-2099
		MPI-ESM-LR		1950-2100
		NorESM1-M		
ICTP	RegCM4-7	HadGEM2-ES	historical, RCP 8.5	1970-2099
		MPI-ESM-MR		
		NorESM1-M		
GERICS	REMO2015	HadGEM2-ES	historical, RCP 8.5	1970-2100
		MPI-ESM-LR		
		NorESM1-M		
UNSW	WRF360J	ACCESS1-0	historical, RCP 4.5, RCP 8.5	1951-2100
		CanESM2		1951-2099
	WRF360K	ACCESS1-0		1951-2100
		CanESM2		1951-2099

Table 2. Summary of CORDEX-Australasia simulations to be used in this study (Evans et al. 2021).

b. Minimum Standard Metrics

1) MAPE AND SPATIAL CORRELATION

The first two MSMs we use to benchmark the CORDEX-Australasia ensemble are the MAPE and spatial correlation. As these metrics are tested against the AGCD dataset and require users to specifically define a benchmarking threshold, we define the benchmarking thresholds based on scientific reasoning, feedback received from the potential user communities, and the objectives of the two hypothetical applications. Feedback from stakeholder user communities across Australia (i.e. humanitarian aid and water resource management) emphasized the need for RCMs that reasonably capture the spatial distribution of rainfall, while their decision-making allows for a generous amount of wet or dry bias due to Australia's characteristically extreme rainfall variability. Further, during data preprocessing and data exploration, we evaluated several precipitation indices (Zhang et al. 2011) of the CORDEX-Australasia ensemble at a coarser resolution to incorporate additional global gridded observational datasets, with and without the quality mask (see Supplemental Material). We also evaluated gridded observational products against the AGCD product to determine a range of observational uncertainty across different characteristics of rainfall. This preliminary assessment underpinned our understanding of reasonable model performance based on the current scientific capabilities in both regional climate modeling and gridded observations over Australia that was used to define the benchmarking thresholds for the MAPE and the spatial correlation. Further, since both hypothetical user case studies will later distill a subset of models without a strong wet or dry bias, we set the benchmarking threshold for the MAPE as <= 0.75. However, in setting the benchmarking threshold for the spatial correlation we are stricter because we do want models that reasonably capture Australia's highly variable spatial rainfall patterns. We set the benchmarking threshold for the spatial correlation as >= 0.7.

In Figure 2, we show the climatological (1976-2005) rainfall bias for each model against AGCD, ranked from wettest to driest based on the weighted spatial average of the bias. Areas in grey show where the quality mask has been applied. At the bottom of each plot the MAPE and the spatial correlation, calculated against the AGCD data, are shown where values highlighted in purple indicate those that meet the performance benchmarking thresholds. Two models fail these benchmarks. The HadGEM2-ES RegCM4-7 fails due to the rainfall bias being too large, and the CanESM2 WRF360K fails as it does not reasonably capture the mean spatial distribution of rainfall. It's important to note how the definition of the benchmarking thresholds for these two metrics impacts how we assess the performance of the simulations.

For instance, if the MAPE benchmarking threshold had been lower (higher) or the spatial correlation higher (lower), more simulations would fail (pass) this test. We would need to increase the MAPE threshold by over 50% for all models to meet performance expectations, but the CanESM-2 CCAM-2008 wouldn't meet our performance expectations if the MAPE threshold was any lower. If we decreased the spatial correlation by approximately 5% than all models would meet the benchmark for this metric. In this case, our thresholds largely identify outliers within the ensemble. The bias maps in Figure 2 also show the substantial variability among simulations of climatological rainfall across Australia, highlighting the need to have metrics that quantify both spatial variability and biases in routine studies assessing model performance.

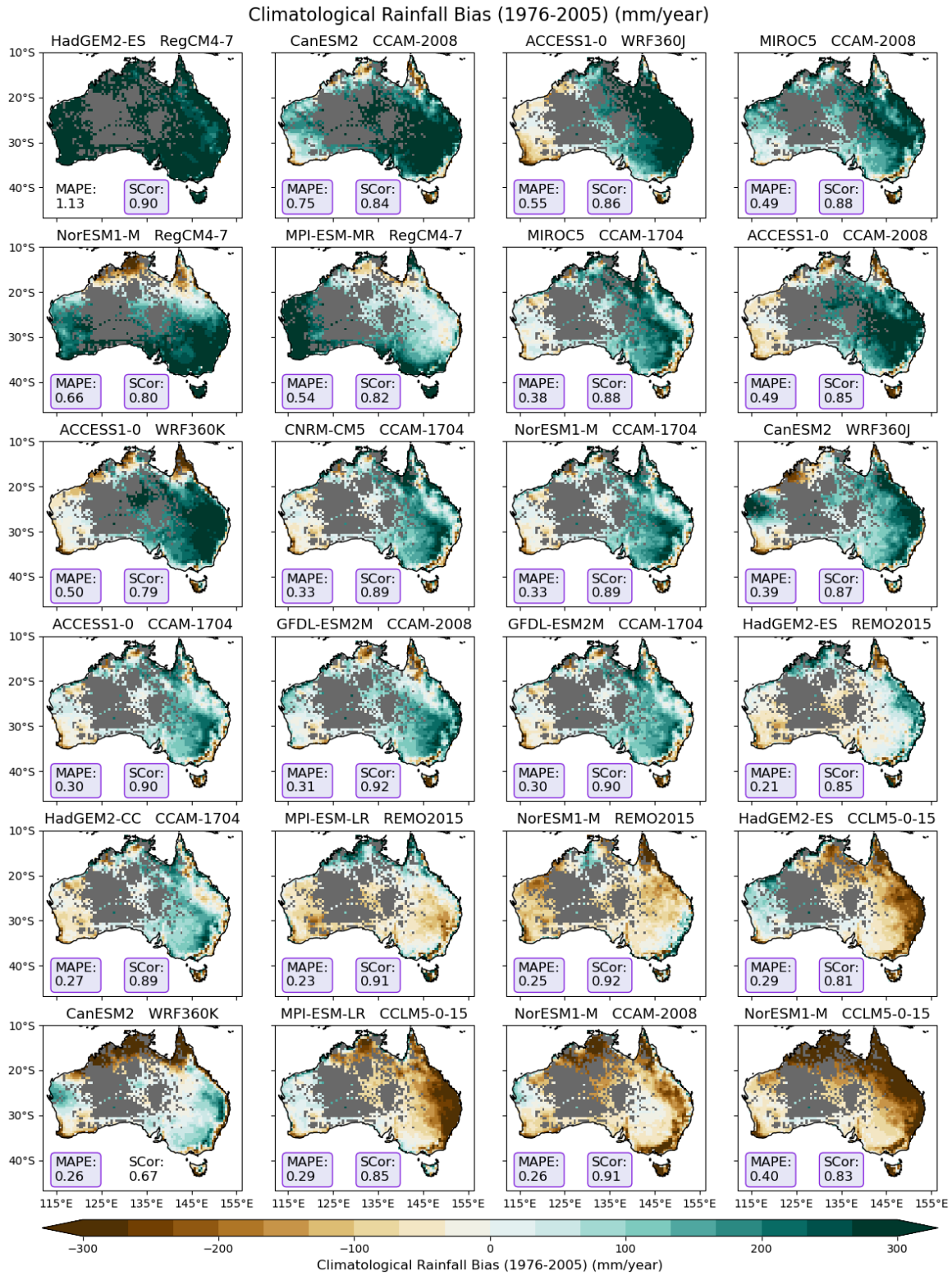

Fig. 2. The climatological (1976-2005) bias for each model against the AGCD observational product, ranked wettest to driest based on the area-weighted spatial average of the bias. Areas in dark grey indicate grid boxes where we don't have at least one observation station within that grid box. In the bottom left corner, we plot the MAPE and the spatial correlation (SCor) calculated against the AGCD data. Values highlighted in purple indicate values that meet our defined benchmarking thresholds. The AGCD climatology for this period is provided as Supplemental Figure S2.

533

534 2) SEASONAL CYCLE

535 The quantification of the Seasonal Cycle for the MSMs differs from the Seasonality
536 column within the Versatility Tier metrics. For our example, there is a unimodal seasonal
537 cycle when averaging rainfall across all of Australia (Figure 3). We assess model
538 performance in simulating the seasonal cycle by ranking the months from wettest to driest
539 and define our benchmarking threshold as the three wettest and driest observed months must
540 be among the six wettest and driest modelled months (Figure 4). This method captures the
541 unimodal structure and the phase of the observed seasonal cycle at a high level. This method
542 also does not restrict how the models simulate the onset and offset of the climatological wet
543 season. Using this definition, two models fail this benchmark as both models have one of the
544 wettest six months falling within AGCD's driest three months (Figure 4). The NorESM1-M
545 CCLM-0-15 and NorESM1-M CCAM-2008 simulations fail as the sixth wettest months
546 (ranked as the seventh driest month in Figure 4) falls within the climatological driest three
547 months of AGCD (Figure 4).

548 While this is an easy way to capture the phase and structure of the seasonal cycle, it
549 provides limited information as to the amplitude of the seasonal cycle. For instance, the
550 CanESM2 WRF360K simulation has a somewhat muted seasonal cycle: the range between
551 the driest month and the wettest month is substantially smaller than that in AGCD (Figure 3).
552 Based on the monthly rankings, this model would pass the benchmark. This is acceptable as
553 the MSMs are meant to be very low-level. If a more precise quantification of model skill in
554 simulating the seasonal cycle is required, then more complex analyses can be completed as
555 recommended in the Versatility Tier section (see Section 3c).

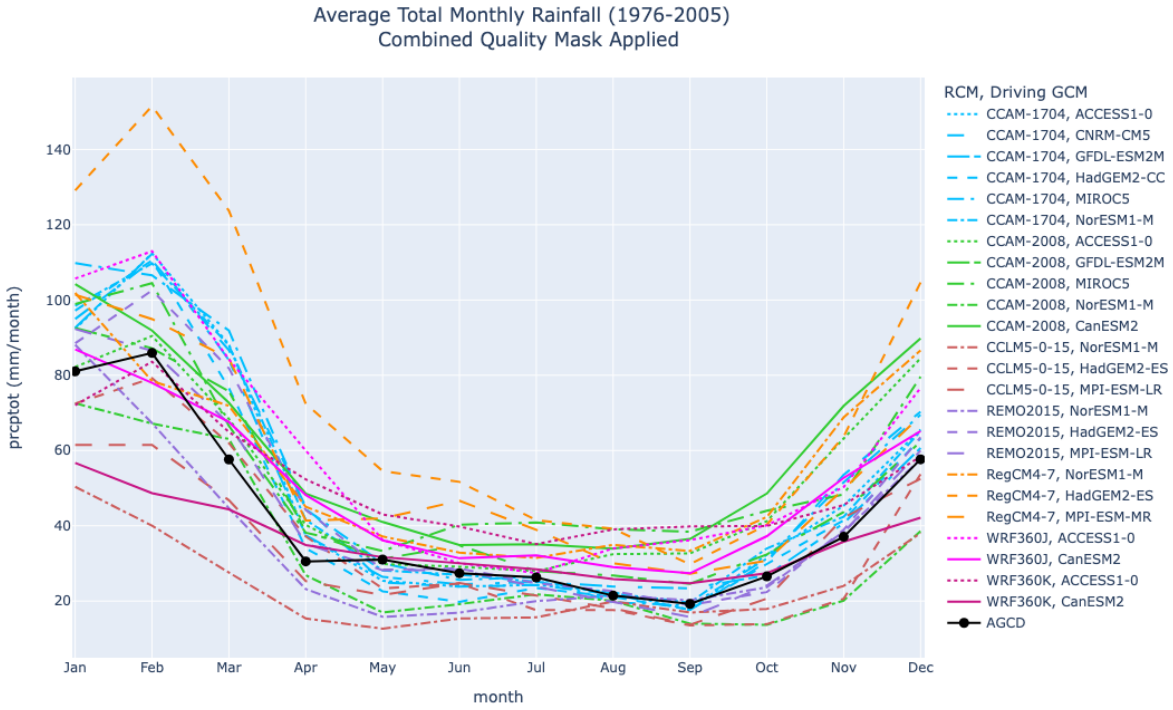

Fig. 3. The climatological (1976-2005), latitudinally weighted, average total monthly rainfall (prcptot) across Australia with the combined quality mask applied. Colors indicate the RCM, and the line styles indicate the forcing GCM. The AGCD data, used as the benchmark, is shown in black.

		Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
AGCD		11	12	9	6	7	5	3	2	1	4	8	10
CCAM-1704	ACCESS1-0	11	12	10	7	5	3	4	2	1	6	8	9
	CNRM-CM5	11	12	10	7	5	3	4	2	1	6	8	9
	GFDL-ESM2M	11	12	10	7	5	3	4	2	1	6	8	9
	HadGEM2-CC	11	12	10	7	4	2	5	3	1	6	8	9
	MIROC5	12	11	10	7	5	3	4	2	1	6	8	9
	NorESM1-M	11	12	10	7	5	4	3	2	1	6	8	9
CCAM-2008	ACCESS1-0	10	12	9	6	3	2	1	4	5	7	8	11
	CanESM2	12	11	9	6	5	2	3	1	4	7	8	10
	GFDL-ESM2M	12	11	10	7	4	6	3	2	1	5	8	9
	MIROC5	11	12	9	2	1	5	6	4	3	7	8	10
	NorESM1-M	12	11	10	8	3	4	7	5	2	1	6	9
CCLM5-0-15	HadGEM2-ES	11	12	10	8	6	7	3	4	1	2	5	9
	MPI-ESM-LR	12	11	9	7	5	6	4	2	1	3	8	10
	NorESM1-M	12	11	9	3	1	2	4	7	5	6	8	10
REMO2015	HadGEM2-ES	11	12	10	8	5	6	4	3	1	2	7	9
	MPI-ESM-LR	12	11	10	7	5	6	3	2	1	4	8	9
	NorESM1-M	12	11	9	6	1	2	3	5	4	7	8	10
RegCM4-7	HadGEM2-ES	11	12	10	8	6	5	4	2	1	3	7	9
	MPI-ESM-MR	12	11	10	5	6	7	4	2	1	3	8	9
	NorESM1-M	12	10	9	7	5	2	1	4	3	6	8	11
WRF360J	ACCESS1-0	11	12	10	8	5	2	1	3	4	6	7	9
	CanESM2	12	11	10	7	5	3	4	2	1	6	8	9
WRF360K	ACCESS1-0	11	12	10	8	6	3	1	2	4	5	7	9
	CanESM2	12	11	10	7	6	5	4	2	1	3	8	9

Fig. 4. The climatological (1976-2005) area-weighted, average total monthly rainfall across Australia with the combined quality mask applied (see Figure 3) are ranked from driest (1) to wettest (12) for each CORDEX simulation, grouped by RCM. Brown shades (1-6) indicate the driest six months and teal colors

(7-12) indicate the wettest six months. The monthly rankings for the AGCD data, used as the benchmark, are on the top row, and the instances where the two simulations fail the benchmark are outlined in red.

3) Direction of a Significant Trend

The final MSM is the direction of a significant observed trend using the annual time series of annual average total precipitation. We use the direction of the significant Thiel-Sen trend of the AGCD product spatially averaged over all of Australia after the quality mask has been applied (Figure 5) as the benchmark. We test the significance of the trend using the Mann-Kendall significance test at a 5% significance level (Hussain et al., 2019). There is no significant positive or negative trend for the AGCD product, so our benchmarking threshold is 'no trend' (Figure 5). We replicate this analysis for each simulation (Figure 5). All models pass this benchmark as no trends are significantly positive or negative, meeting our performance benchmark. Because the RCMs, and their forcing GCMs, are not forced by observational datasets, we do not expect the time series of the simulations to be aligned with that of observations. We are only concerned with the direction of a significant trend—neglecting magnitude and a comprehensive quantification of interannual temporal variability—once again emphasizing the MSMs are low-level performance metrics. If simulations are driven by reanalysis data, then it is expected that users would quantify temporal consistency as appropriate in the Versatility Metrics.

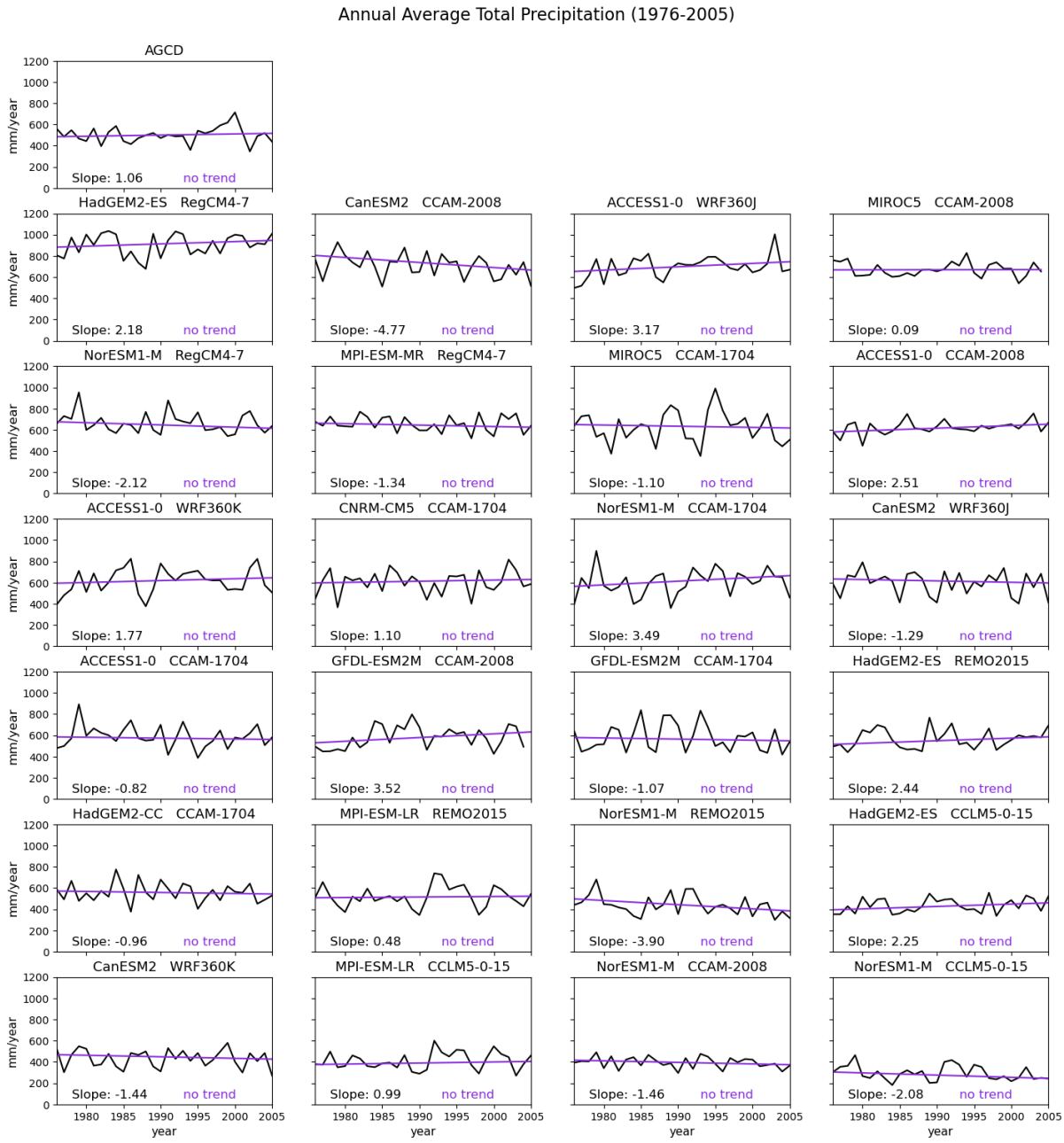

Fig. 5. The observed (top row) and modeled area weighted annual average total precipitation across Australia, with the combined quality mask applied, for 1976-2005. The direction of the observed Thiel-Sen trend is the benchmark (top row). The Thiel-Sen trend line for each of the simulations is plotted in purple. The magnitude of the trend is noted in the bottom left corner and the results of the Mann-Kendall significance test (Hussain et al., 2019) is noted in the bottom right corner. Models are sorted based on the magnitude of the latitudinally weighted spatial average to match the order of Figure 2. All models pass the benchmark.

4) SUBSET OF MODELS

After testing the CORDEX-Australasia ensemble against the four MSMs, 20 simulations out of 24 meet the minimum performance requirements for the hypothetical case studies

(Figure 6). At this point in applying the benchmarking framework, we eliminate the four simulations that failed the minimum performance standards from further analysis. However, nearly all the simulations within the CORDEX-Australasia ensemble assessed here simulate the fundamental characteristics of precipitation quite well over Australia, noting regional biases (Figure 2). Further, we cannot identify any RCM or forcing GCM that is routinely less skillful in simulating these characteristics across all of Australia.

It is important to note that the model subset depends on the performance requirements, i.e., benchmarks and benchmarking thresholds, defined in the earlier section. Because the benchmarking thresholds so strongly influence the quantification of model performance, it's critical to define them in a way that is fit for purpose and incorporates strong scientific reasoning.

RCM	Forcing GCM	Mean Absolute Percentage Error	Spatial Correlation	Seasonal Cycle	Direction of a Significant Trend	TOTAL
CCAM-1704	ACCESS1-0	✓	✓	✓	✓	4
	CNRM-CM5	✓	✓	✓	✓	4
	GFDL-ESM2M	✓	✓	✓	✓	4
	HadGEM2-CC	✓	✓	✓	✓	4
	MIROC5	✓	✓	✓	✓	4
	NorESM1-M	✓	✓	✓	✓	4
CCAM-2008	ACCESS1-0	✓	✓	✓	✓	4
	CanESM2	✓	✓	✓	✓	4
	GFDL-ESM2M	✓	✓	✓	✓	4
	MIROC5	✓	✓	✓	✓	4
	NorESM1-M	✓	✓		✓	3
CCLM5-0-15	HadGEM2-ES	✓	✓	✓	✓	4
	MPI-ESM-LR	✓	✓	✓	✓	4
	NorESM1-M	✓	✓		✓	3
RegCM4-7	HadGEM2-ES		✓	✓	✓	3
	MPI-ESM-MR	✓	✓	✓	✓	4
	NorESM1-M	✓	✓	✓	✓	4
REMO2015	HadGEM2-ES	✓	✓	✓	✓	4
	MPI-ESM-LR	✓	✓	✓	✓	4
	NorESM1-M	✓	✓	✓	✓	4
WRF360J	ACCESS1-0	✓	✓	✓	✓	4
	CanESM2	✓	✓	✓	✓	4
WRF360K	ACCESS1-0	✓	✓	✓	✓	4
	CanESM2	✓		✓	✓	3

Fig. 6. Summary of model performance against the MSMs. 20/24 models pass all the MSMs, highlighted in green in the far-right column.

c. Hypothetical User 1 - Seasonality

Australia's climate is characterized by highly diverse rainfall patterns, which vary significantly across different regions and seasons. For the first hypothetical case study, we seek to identify models that best capture the amplitude and phase of the seasonal cycle across Australia as compared to observations. We will emphasize benchmarking the models against the amplitude as our assessment of the seasonal cycle in the MSMs neglected amplitude. This means that we will be stricter in our definition of the benchmarking threshold for the amplitude than for the phase. To calculate the amplitude and phase of the seasonal cycle, we first calculate the climatological seasonal cycle (Figure 3) at each grid point. We define the amplitude as the difference between the maximum and mean monthly rainfall (Figure 7) and the phase as the month of maximum rainfall (Figure 8). To benchmark the subset of simulations from the CORDEX-Australasia ensemble (Figure 6), we calculate the circular spatial correlation against the AGCD observational product for the phase and the NRMSE (Eq. 1) for the amplitude. For the phase, we assign an integer to each month (1-12) and calculate the circular spatial correlation against the maps of these values using Eq. 2 (Jammalamadaka and SenGupta 2001) where α and β indicate the month value of the observational product and model simulation, respectively, expressed as angles around a circle, and $\bar{\alpha}$ and $\bar{\beta}$ are the circular mean of this angle taken over all grid cells across Australia. We use this metric to account for the circularity of the seasonal cycle.

$$\rho_c(\alpha,\beta) = \frac{\sum_{i=1}^{n} \sin(\alpha_i - \bar{\alpha})\ \sin(\beta_i - \bar{\beta})}{\sqrt{\sum_{i=1}^{n} \sin^2(\alpha_i - \bar{\alpha})\ \sin^2(\beta_i - \bar{\beta})}}$$

Similar to how we defined benchmarking thresholds for the MAPE and spatial correlation of the MSMs, benchmarking thresholds for these seasonality metrics must be defined using scientific reasoning and the purpose for applying the BMF. We set the benchmarking threshold for the amplitude as >= 0.6 to identify models that best simulate the amplitude of the seasonal cycle across Australia with a maximum relative error of 0.6. To set this benchmarking threshold, we explored the skill of each of the simulations in simulating the seasonal cycle at smaller scales (see Supplemental Figure S6) and in simulating the amplitude

641 across our domain. Then, we intuitively set a threshold that is rather strict given our
642 understanding of model performance across Australia (see Supplemental Material) but is also
643 a reasonable performance expectation. As another example, a less subjective threshold could
644 have been to identify the 50% best performing models and not identify a specific
645 benchmarking threshold. While benchmarking does require *a priori* performance
646 expectations, it is very unlikely that benchmarking thresholds can ever truly be informed
647 without any relevant assessment of model performance to establish scientific expertise. Using
648 the benchmarking threshold of 0.6, nine models meet our performance expectations (Figure
649 7). Recognizing that rainfall may peak in the same season but in a different month, we
650 benchmark the phase as a statistically significant, positive circular correlation tested at the
651 5% significance level. We compute the 95% confidence interval (see Supplemental Table S2)
652 by applying bootstrapping methods that randomly resample the rainfall phase data across
653 60% of our domain for the observations and the model simulations. We use identical subsets
654 of the observations and simulations in each of our 5000 iterations to retain the spatial
655 relationship between our datasets. We calculate the circular correlation coefficient on our
656 resampled datasets to create our confidence interval. This definition does not overextend our
657 expectations of reasonable model performance but is strict enough to eliminate models that
658 too often peak early or late in the rainy season across Australia. The AGCD product also
659 captures much finer scale features of the rainfall phase than the models do, leading to
660 consistently low correlation values (Figure 8). If we smoothed the rainfall phase using a low-
661 pass filter or similar techniques, we would expect the simulations to have a higher
662 correlation. Based on this benchmarking definition, all models except the NorESM-1
663 REMO2015 simulation pass our performance expectations (Figure 8). There are eight models
664 that meet performance expectations for both seasonality benchmarks (amplitude and phase)
665 and would therefore be the subset of models that meet all our performance expectations for
666 the MSMs and our first hypothetical case study.

667 These methods to quantify the seasonality of rainfall will likely be too restrictive for most
668 applications of the BMF, especially over a large spatial domain with high seasonal
669 variability. Observations will likely capture finer features of seasonality that are smoothed by
670 models. There are many other ways to quantify rainfall seasonality (see Section 2.c.1), and
671 we emphasize that users should select metrics and benchmarks that are appropriate for their
672 study.

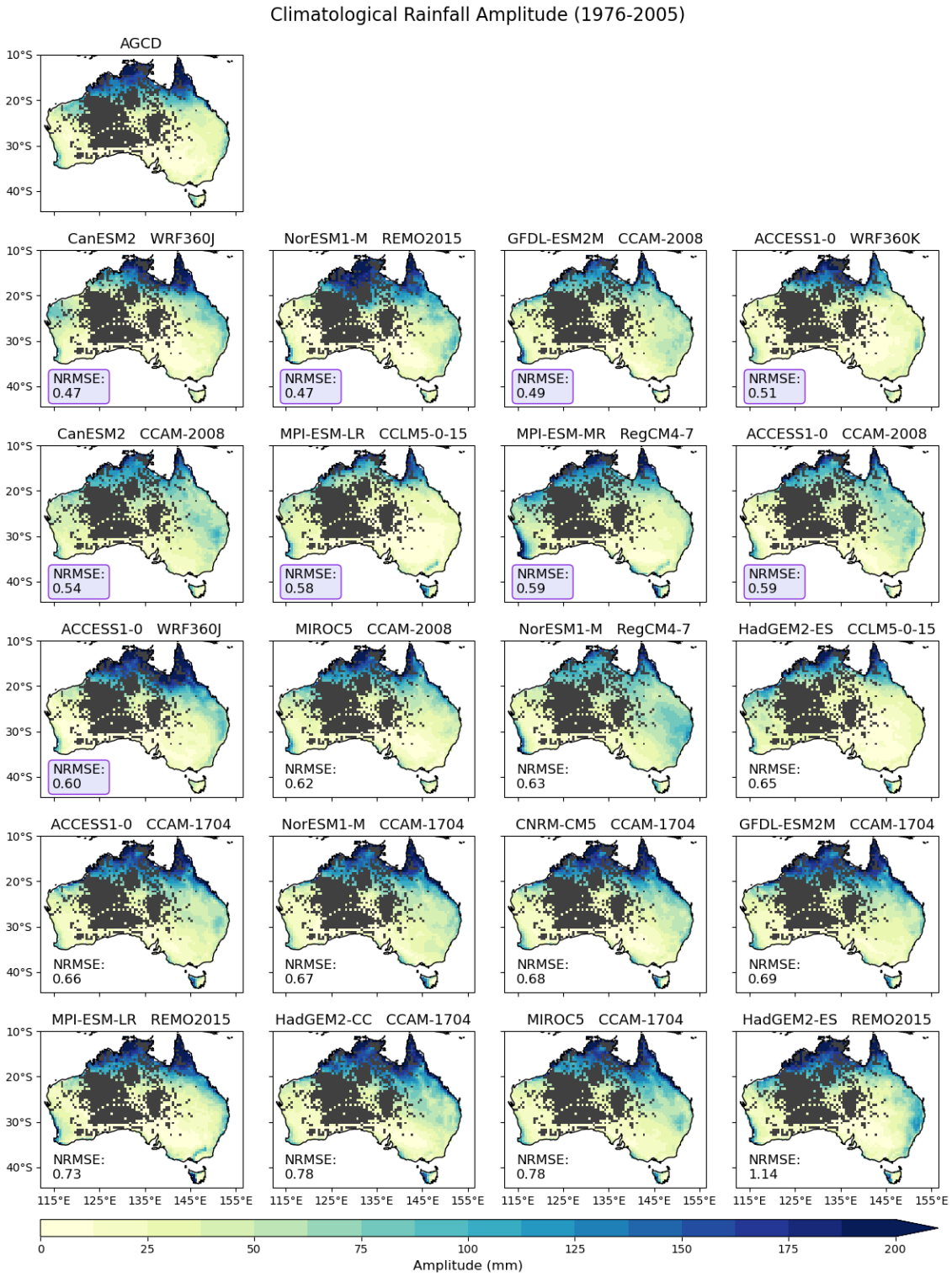

Fig. 7. The climatological (1976-2005) amplitude of rainfall. The AGCD dataset, used as the benchmark, is in the top left panel. Each of the models from Fig 6 follows, and they are sorted by the score of the spatial correlation tested against the AGCD dataset, shown in the bottom left corner of each panel. Simulations that pass the benchmark are highlighted in purple.

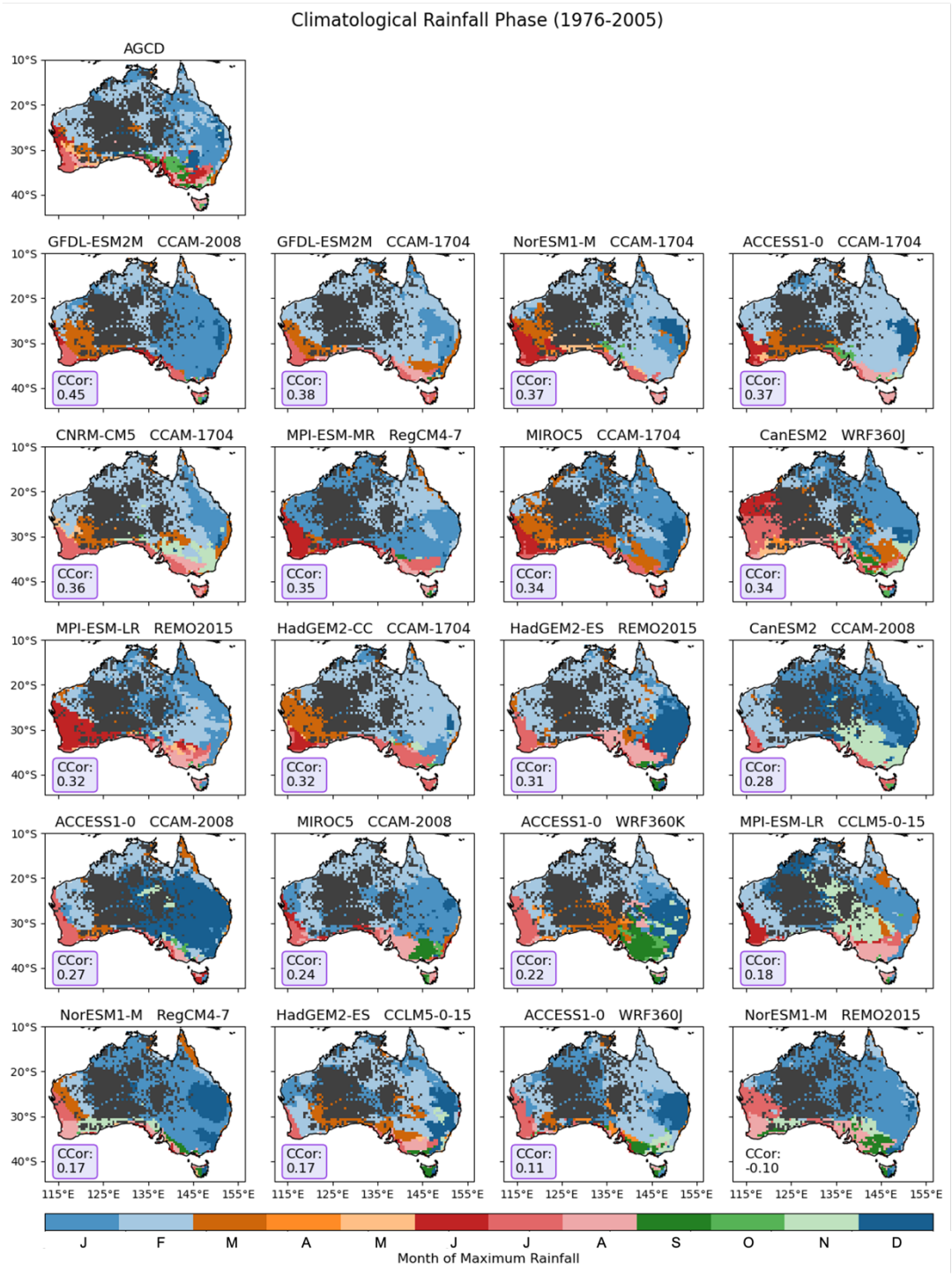

Fig. 8. The climatological (1976-2005) phase of rainfall (month of maximum rainfall) based on monthly rainfall totals. The AGCD dataset, used as the benchmark, is in the top left panel. Each of the models from Figure 6 follows, and they are sorted by the score of the circular spatial correlation (Eq. 2), shown in the bottom left corner of each panel. Simulations that pass the benchmark are highlighted in purple. Colors indicate the month in which rainfall climatologically peaks, and shades of similar colors indicate the season.

d. Hypothetical User 2 - Rainfall Deficit

Australia can be thought of as "always being in drought" broken up by periods of drought-breaking rains. Drought is a very complex hazard, and there are many ways to define drought. Further, there are many metrics and methods that can be used to quantify drought including when a deficit of rainfall is categorized as drought. For this second hypothetical application of the BMF, we seek to identify models that reasonably simulate time spent in meteorological drought over Australia as defined by a deficit of rainfall. We also do not want to include models that underestimate the percentage of time spent in any category of drought during our time period to align with the goals of the hypothetical user. We use the Standardized Precipitation Index (SPI) (McKee et al., 1993 and WMO, 2012) to identify models that reasonably simulate the extent and severity of drought over time. The SPI is a measure of how much rainfall has deviated from the average, based on historical records for a particular location and timespan. The SPI can be calculated across different temporal averaging periods relevant to different usable water resources including soil moisture, groundwater, streamflow, snowpack, and reservoir storage (Mckee et al. 1993). McKee et al. (1993) also define thresholds to identify different categories of drought based on the SPI value that can be used for all the temporal averaging periods. However, these thresholds have been updated by the WMO to re-categorize a 'mild drought' as 'near normal conditions' (WMO, 2012).

For our application, we calculate the SPI at each grid box over Australia at a 12-month averaging period for the AGCD product and the subset of members of the CORDEX-Australasia ensemble (Figure 6) using the Climpact software (Alexander and Herold, 2015). Figure 9 shows the area-averaged 12-month SPI for AGCD in the top-left panel followed by the model simulations sorted alphabetically based on the RCM – GCM name. Colored vertical columns indicate periods of drought where the color indicates the severity of drought as defined by the WMO (2012). To benchmark this metric, we calculate the percentage of the time series spent in each category of drought and define the benchmarking threshold as 0 to 10 percentage points of the AGCD value for each category of drought (Figure 10). Models must meet this benchmark for all categories of drought to meet our performance requirements. We set the benchmarking threshold as such because previous studies have shown that models struggle to capture observed dry periods over Australia (Ukkola et al. 2018; Kirono et al. 2020), and we do not expect the simulated rainfall deficits to be

718 synchronized with observations. Using this definition, 14 models pass this benchmark (Figure
719 10).

720 This is only one example of how to benchmark RCMs to identify models that reasonably
721 simulate drought-level rainfall deficits. Using the same metric, we could rank the
722 performance of models based on tiered benchmarking thresholds. For instance, models that
723 fall within +/- 5% of the observational percentage could be ranked excellent, +/- 10% as
724 good, +/- 15% as adequate, etc. One could also benchmark the SPI using other drought
725 indices, or vice versa. For instance, Joetzjer et al. (2013) use the standardized runoff index, a
726 measure of river discharge, to benchmark several meteorological drought indices. Again, the
727 benchmarking thresholds should be defined based on the application of the benchmarking
728 framework and incorporate sound scientific reasoning.

729

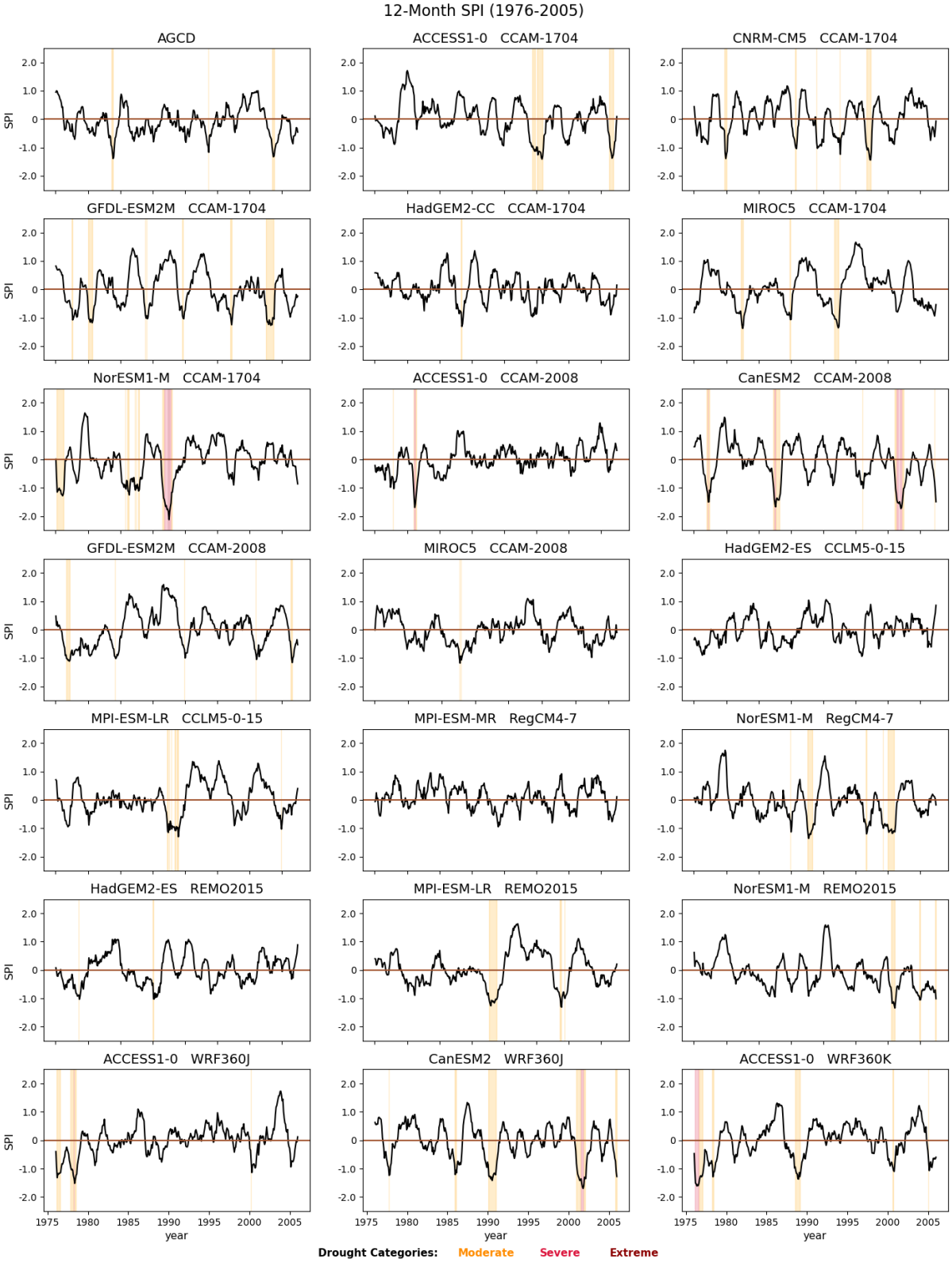

Fig. 9. The area weighted averaged 12-month SPI values across Australia (with the combined quality mask applied) for 1976-2005. Vertical bars indicate the category of drought as defined by the WMO (2012). The top-left figure shows the SPI for AGCD, and the 20 members of the CORDEX-Australasia ensemble follow, sorted alphabetically by RCM-GCM name to match Table 3.

Percentage of Time Series in Each Drought Category				
Dataset name		**Moderate -1.00 to -1.49**	**Severe -1.50 to -1.99**	**Extreme ≤ -2.00**
AGCD		**2.22**	**0.00**	**0.00**
CCAM-1704	ACCESS1-0	5.56	0.00	0.00
	CNRM-CM5	4.17	0.00	0.00
	GFDL-ESM2M	7.78	0.00	0.00
	HadGEM2-CC	0.56	0.00	0.00
	MIROC5	3.61	0.00	0.00
	NorESM1-M	6.67	2.50	0.28
CCAM-2008	ACCESS1-0	1.11	0.56	0.00
	CanESM2	5.56	2.78	0.00
	GFDL-ESM2M	3.33	0.00	0.00
	MIROC5	0.57	0.00	0.00
CCLM5-0-15	HadGEM2-ES	0.00	0.00	0.00
	MPI-ESM-LR	2.78	0.00	0.00
RegCM4-7	MPI-ESM-MR	0.00	0.00	0.00
	NorESM1-M	6.11	0.00	0.00
REMO2015	HadGEM2-ES	0.83	0.00	0.00
	MPI-ESM-LR	4.44	0.00	0.00
	NorESM1-M	2.50	0.00	0.00
WRF360J	ACCESS1-0	4.17	0.28	0.00
	CanESM2	8.06	1.11	0.00
WRF360K	ACCESS1-0	5.83	1.67	0.00

Fig. 10. The percentage of the 12-month SPI time series (see Figure 9) that falls within each drought category as defined by WMO (2012). For moderate droughts the benchmarking range is 2.22 % (AGCD) - 12.22% (+10 percentage points), for severe and extreme droughts 0-10%.14 models pass the benchmarking threshold of 0 to +10% of the AGCD product (top row) for all categories of drought. Instances where models fail the benchmark are highlighted in red.

4. Summary and Conclusions

To date, there is no standardized framework available for the scientific community to quantify RCM skill in simulating various characteristics of rainfall. We have developed this framework primarily to establish a uniform approach for holistically assessing RCM performance in simulating rainfall and, for stakeholder user communities to identify fit-for-purpose model simulations. This framework can underpin future model assessments of existing and new simulations, including studies to compare dynamical or statistical downscaling techniques, added value studies, quantifying model skill across CORDEX

generations and/or regions, or testing machine learning techniques. We introduce a tiered set of performance metrics that establishes a consistent yet versatile framework with wide-ranging applications across research and stakeholder user groups, and we walk users through two example applications of the BMF, summarized in Figure 11.

It is critical that users are thoughtful and transparent in their definition of benchmarking thresholds and their selection of additional Versatility metrics. While the MSMs provide consistency in quantifying model performance, the definition of benchmarking thresholds for the MSMs and additional metrics can be subjective. If users are not clear about their justification for defining benchmarks, this can lead to erroneous conclusions about model performance. For instance, our definition of benchmarking thresholds and small selection of Versatility Metrics yield nine and fourteen simulations in the subsets for the first and second hypothetical user examples, respectively (Figure 11). This is not conclusive or prescriptive for which CORDEX-Australasia simulations are best at representing these rainfall characteristics over Australia. Ideally, users should incorporate multiple metrics when assessing how well models simulate rainfall characteristics that fall within the Versatility Tier as model performance can vary across metrics used to quantify skill for the same aspect of precipitation (Martinez-Villalobos et al., 2022). Further, if possible, it is recommended to benchmark models across regions with a similar climate regime, such as the IPCC regions (see Supplemental Figure S6). This will prevent key regional features from being overshadowed by large-scale features, such as the seasonal cycle in southern Australia compared to the rest of the continent (Figure 8).

We apply the BMF to 24 simulations of the CORDEX-Australasia ensemble. Of the 24 simulations, 20 meet our performance requirements for the MSMs (Figures 6 and 11), showing that across Australia, most members of the ensemble perform reasonably well at simulating fundamental characteristics of rainfall. While there are no obvious groupings of RCMs or GCMs that routinely perform better at simulating these characteristics of rainfall across all of Australia, there are regional patterns of RCM performance. For instance, the CCAM-1704 and CCAM-2008 models tend to run drier (out of the full ensemble) over west-southwest Western Australia and the southeast coastline. Similarly, the WRF360K simulations underestimate mean rainfall in northern Australia and in southwest Western Australia. The CCLM-0-15 simulations are routinely drier across much of Australia, with a consistent dry bias across the eastern half and northern Australia. This is largely true for the

REMO2015 simulations as well, although there is much more spatial variability based on the forcing GCM (Figure 2). These patterns seem to indicate that regionally and sub-regionally, the choice of RCM has more influence in how rainfall is simulated over Australia than the forcing GCM. For the models in the subsets from our two case studies (Figure 11), there are few obvious groupings of RCMs or GCMs that perform especially well at simulating the seasonality and rainfall deficits across Australia. However, all HadGEM2-ES/CC forced simulations underestimate the observed time spent in drought and fail this benchmark (Section 3d). It's expected that this would change if we investigated over a smaller region though (Figure 2). Combined, our subsets include 8/10 GCMs and 7/7 RCMs. No HadGEM-ES/CC forced simulations meet performance standards for either the Seasonality or Rainfall Deficit investigations, but all WRF360J/K simulations within the MSM subset meet performance expectations for both investigations. (Figure 11). It is also important to acknowledge the dependence of the subset of models that are identified through the application of the BMF. It is well known that GCMs and RCMs cannot be considered independent due to shared code and parameterizations among model developers and institutions (Knutti et al. 2013).

The BMF presented here is a significant, first step in establishing consistency in how the scientific community quantifies RCM skill in simulating various characteristics of rainfall and the broader water cycle. The flexibility incorporated in the development of the framework makes it suitable for application across regions and numerous user communities. While the BMF facilitates consistency in the methodical assessment of RCM skill, there is still a pressing need for high quality, high resolution global observational precipitation datasets to fully establish regional consistency and equity in assessing RCM skill. We acknowledge and encourage a broader application of this framework beyond what has been discussed here, and we hope users will build on this framework by customizing benchmarks, developing and incorporating additional metrics, and identifying best-practice standards for benchmarking.

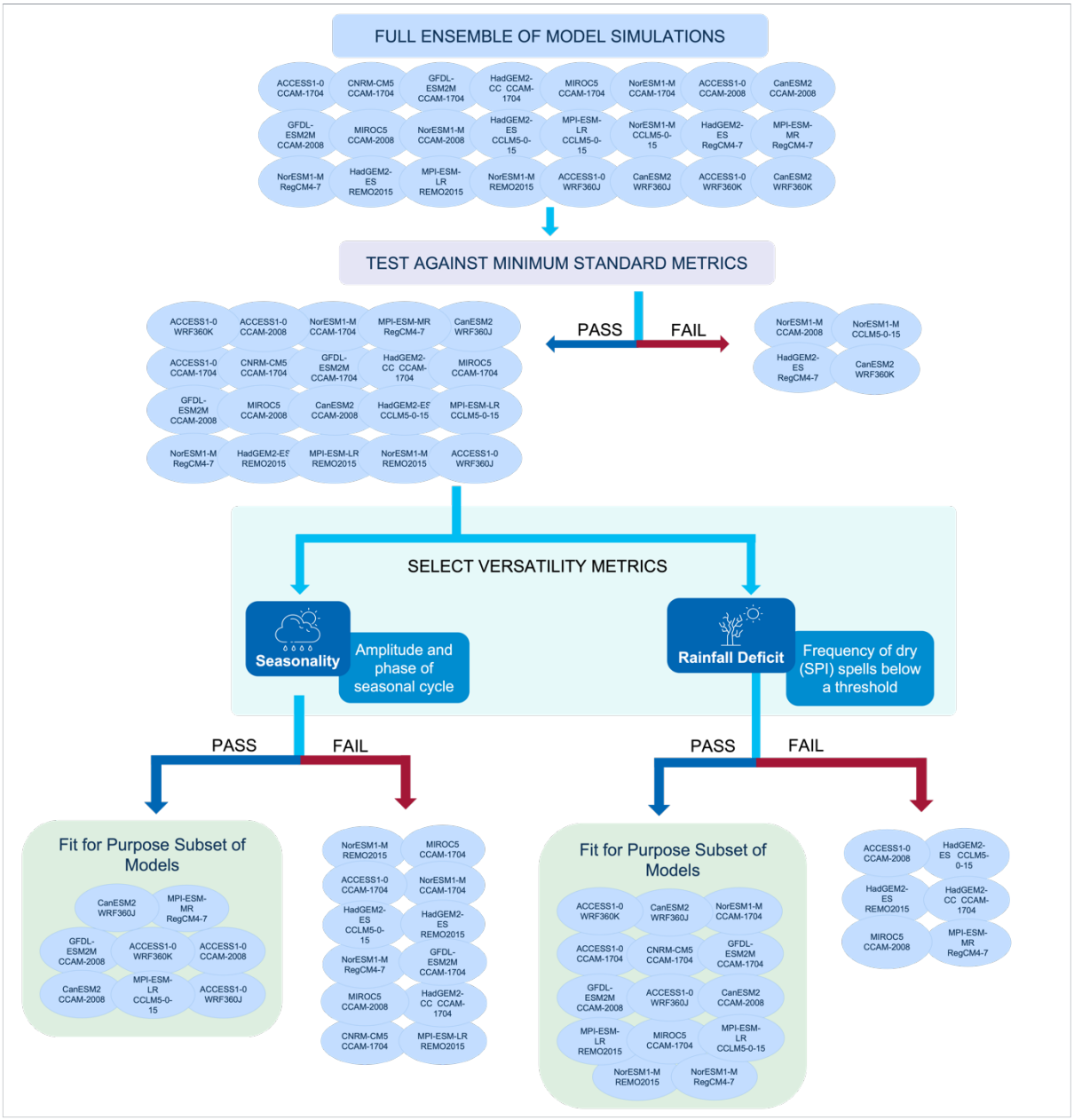

Fig. 11. Schematic flowchart summarizing our example applications of the benchmarking framework.

Acknowledgments.

We thank all members of the Australian Research Council (ARC) Centre of Excellence for Climate Extremes (CLEX) Computational Modelling Systems (CMS) Team for their assistance in data acquisition, pre-processing, and analysis. We also thank Joshua-Brent Amoils for early assistance in data pre-processing and Georgina Harmer for assistance with graphic design. This work received funding from the University of New South Wales (UNSW) and ARC grant FT210100459. This project was supported by CLEX (ARC Grant

No. CE170100023) and received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Grant agreement No 101027577. R.N.I. is also supported by a Scientia PhD scholarship from UNSW (Program code 1476). The analyses and dataset publications completed through this project used resources and services provided by the National Computational Infrastructure which is supported by the Australian Government.

Data Availability Statement.

The datasets used in this study are publicly available. The raw CORDEX-Australasia datasets are available at https://cordex.org/data-access/, and the raw AGCD dataset is available at https://geonetwork.nci.org.au/geonetwork/srv/eng/catalog.search#/metadata/f6475_9317_5747_6204. The full suite of Climpact indices for the CORDEX-Australasia ensemble are available from Isphording et al. (2023). Python scripts used for analysis and figure creation are available from Isphording, R. N. (2023).

REFERENCES

Abramowitz, G., 2005: Towards a benchmark for land surface models. *Geophys Res Lett*, **32**, 1–4, https://doi.org/10.1029/2005GL024419.

Abramowitz, G., 2012: 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.

Ahn, M.-S., P. A. Ullrich, P. J. Gleckler, J. Lee, A. C. Ordonez, and A. G. Pendergrass, 2023: Evaluating precipitation distributions at regional scales: a benchmarking framework and application to CMIP5 and 6 models. *Geosci Model Dev*, **16**, 3927–3951, https://doi.org/10.5194/gmd-16-3927-2023

Ahn, M.-S., P. J. Gleckler, J. Lee, A. G. Pendergrass, and C. Jakob, 2022: Benchmarking Simulated Precipitation Variability Amplitude across Timescales. *J Clim*, **35**, 6773–6796, https://doi.org/10.1175/JCLI-D-21-0542.1.

Alexander, L. V., and J. M. Arblaster, 2009: Assessing trends in observed and modelled climate extremes over Australia in relation to future projections. *Int J Climatol*, Vol. 29 of, 417–435.

——, and N. Herold, 2015: ClimPACTv2 indices and software. WMO, https://github.com/ARCCSS-extremes/climpact2.

——, and J. M. Arblaster, 2017: Historical and projected trends in temperature and precipitation extremes in Australia in observations and CMIP5. *Weather Clim Extrem*, **15**, 34–56, https://doi.org/10.1016/J.WACE.2017.02.001.

——, and Coauthors, 2019: On the use of indices to study extreme precipitation on sub-daily and daily timescales. *Environ Res Lett*, **14**, https://doi.org/10.1088/1748-9326/ab51b6.

——, M. Bador, R. Roca, S. Contractor, M. G. Donat, and P. L. Nguyen, 2020: Intercomparison of annual precipitation indices and extremes over global land areas from in situ, space-based and reanalysis products. *Environ Res Lett*, **15**, https://doi.org/10.1088/1748-9326/ab79e2.

Avila, F. B., Dong, S., Menang, K. P., Rajczak, J., Renom, M., Donat, M. G., & Alexander, L. V., 2015: Systematic investigation of gridding-related scaling effects on annual statistics of daily temperature and precipitation maxima: A case study for south-east Australia. *Weather Clim Extrem*, 9, 6–16. https://doi.org/10.1016/j.wace.2015.06.003

Bador, M., and Coauthors, 2020a: Impact of Higher Spatial Atmospheric Resolution on Precipitation Extremes Over Land in Global Climate Models. *J Geophys Res: Atmospheres*, **125**, https://doi.org/10.1029/2019jd032184.

——, L. V. Alexander, S. Contractor, and R. Roca, 2020b: Diverse estimates of annual maxima daily precipitation in 22 state-of-the-art quasi-global land observation datasets. *Environ Res Lett*, **15**, https://doi.org/10.1088/1748-9326/ab6a22.

Baker, N. C., and P. C. Taylor, 2016: A framework for evaluating climate model performance metrics. *J Clim*, **29**, https://doi.org/10.1175/JCLI-D-15-0114.1.

Barua, S., N. Muttil, A. W. M. Ng, and B. J. C. Perera, 2013: Rainfall trend and its implications for water resource management within the Yarra River catchment, Australia. *Hydrol Process*, **27**, 1727–1738, https://doi.org/10.1002/hyp.9311.

877 Basso, B., C. Fiorentino, D. Cammarano, G. Cafiero, and J. Dardanelli, 2012: Analysis of
878 rainfall distribution on spatial and temporal patterns of wheat yield in Mediterranean
879 environment. *European J Agronomy*, **41**, 52–65,
880 https://doi.org/10.1016/j.eja.2012.03.007.

881 Berry, G., C. Jakob, and M. Reeder, 2011: Recent global trends in atmospheric fronts.
882 *Geophys Res Lett*, **38**, https://doi.org/10.1029/2011GL049481.

883 Best, M. J., and Coauthors, 2015: The plumbing of land surface models: Benchmarking
884 model performance. *J Hydrometeorol*, **16**, 1425–1442, https://doi.org/10.1175/JHM-D-
885 14-0158.1.

886 Boé, J., S. Somot, L. Corre, and P. Nabat, 2020: Large discrepancies in summer climate
887 change over Europe as projected by global and regional climate models: causes and
888 consequences. *Clim Dyn*, **54**, 2981–3002, https://doi.org/10.1007/s00382-020-05153-1.

889 Cai, W., A. Sullivan, and T. Cowan, 2011: Interactions of ENSO, the IOD, and the SAM in
890 CMIP3 models. *J Clim*, **24**, 1688–1704, https://doi.org/10.1175/2010JCLI3744.1.

891 Caretta, M.A., and Coauthors, 2022: Water. *Climate Change 2022: Impacts, Adaptation and*
892 *Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the*
893 *Intergovernmental Panel on Climate Change*, H.-O. Pörtner et al., Eds, Cambridge
894 University Press. pp. 551–712, doi:10.1017/9781009325844.006.

895 Casanueva, A., and Coauthors, 2016: Daily precipitation statistics in a EURO-CORDEX
896 RCM ensemble: added value of raw and bias-corrected high-resolution simulations. *Clim*
897 *Dyn*, **47**, 719–737, https://doi.org/10.1007/s00382-015-2865-x.

898 Chen, D., A. Dai, and A. Hall, 2021: The Convective-To-Total Precipitation Ratio and the
899 "Drizzling" Bias in Climate Models. *J Geophys Res: Atmos*, **126**,
900 https://doi.org/10.1029/2020JD034198.

901 Chen, X., and K. K. Tung, 2018: Global-mean surface temperature variability: space–time
902 perspective from rotated EOFs. *Clim Dyn*, **51**, 1719–1732,
903 https://doi.org/10.1007/s00382-017-3979-0.

904 Choudhary, A., A. P. Dimri, and H. Paeth, 2019: Added value of CORDEX-SA experiments
905 in simulating summer monsoon precipitation over India. *Int J Climatol*, **39**, 2156–2172,
906 https://doi.org/10.1002/joc.5942.

907 Chu, P. S., Y. R. Chen, and T. A. Schroeder, 2010: Changes in precipitation extremes in the
908 Hawaiian Islands in a warming climate. *J Clim*, **23**,
909 https://doi.org/10.1175/2010JCLI3484.1.

910 Ciarlo, J. M., and Coauthors, 2020: A new spatially distributed added value index for
911 regional climate models: the EURO-CORDEX and the CORDEX-CORE highest
912 resolution ensembles. *Clim Dyn*, https://doi.org/10.1007/s00382-020-05400-5.

913 Contractor, S., L. V. Alexander, M. G. Donat, and N. Herold, 2015: How Well Do Gridded
914 Datasets of Observed Daily Precipitation Compare over Australia? *Advances in*
915 *Meteorology*, **2015**, https://doi.org/10.1155/2015/325718.

916 Covey, C., P. J. Gleckler, C. Doutriaux, D. N. Williams, A. Dai, J. Fasullo, K. Trenberth, and
917 A. Berg, 2016: Metrics for the diurnal cycle of precipitation: Toward routine
918 benchmarks for climate models. *J Clim*, **29**, https://doi.org/10.1175/JCLI-D-15-0664.1.

919 Dai, A., 2006: Precipitation Characteristics in Eighteen Coupled Climate Models. *J Clim*,
920 4605–4630.

921 Dey, R., M. Bador, L. V. Alexander, and S. C. Lewis, 2021: The drivers of extreme rainfall
922 event timing in Australia. *Int J Climatol*, **41**, 6654–6673,
923 https://doi.org/10.1002/joc.7218.

924 Dosio, A., and Coauthors, 2021: Projected future daily characteristics of African precipitation
925 based on global (CMIP5, CMIP6) and regional (CORDEX, CORDEX-CORE) climate
926 models. *Clim Dyn*, https://doi.org/10.1007/s00382-021-05859-w.

927 Dunning, C. M., E. C. L. Black, and R. P. Allan, 2016: The onset and cessation of seasonal
928 rainfall over Africa. *J Geophys Res*, **121**, 11405–11424,
929 https://doi.org/10.1002/2016JD025428.

930 ——, R. P. Allan, and E. Black, 2017: Identification of deficiencies in seasonal rainfall
931 simulated by CMIP5 climate models. *Environ Res Lett*, **12**, https://doi.org/10.1088/1748-
932 9326/aa869e.

933 Evans, J. P., K. Bormann, J. Katzfey, S. Dean, and R. Arritt, 2016: Regional climate model
934 projections of the South Pacific Convergence Zone. *Clim Dyn*, **47**, 817–829,
935 https://doi.org/10.1007/s00382-015-2873-x.

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

939 Eyring, V., and Coauthors, 2016: ESMValTool (v1.0)-a community diagnostic and
940 performance metrics tool for routine evaluation of Earth system models in CMIP. *Geosci*
941 *Model Dev*, **9**, 1747–1802, https://doi.org/10.5194/gmd-9-1747-2016.

942 ——, and Coauthors, 2019: Taking climate model evaluation to the next level. *Nat Clim*
943 *Chang*, **9**, 102–110, https://doi.org/10.1038/s41558-018-0355-y.

944 Feng, Z., and Coauthors, 2021: A Global High-Resolution Mesoscale Convective System
945 Database Using Satellite-Derived Cloud Tops, Surface Precipitation, and Tracking. *J*
946 *Geophys Res: Atmos*, **126**, https://doi.org/10.1029/2020JD034202.

947 Fiedler, S., and Coauthors, 2020: Simulated Tropical Precipitation Assessed across Three
948 Major Phases of the Coupled Model Intercomparison Project (CMIP). *Mon Weather Rev*,
949 **148**, 3653–3680, https://doi.org/10.1175/MWR-D-19.

950 Fita, L., J. P. Evans, D. Argüeso, A. King, and Y. Liu, 2017: Evaluation of the regional
951 climate response in Australia to large-scale climate modes in the historical NARCliM
952 simulations. *Clim Dyn*, **49**, 2815–2829, https://doi.org/10.1007/s00382-016-3484-x.

953 Flato, G., and Coauthors, 2013: Evaluation of climate models. *Climate Change 2013: The*
954 *Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report*
955 *of the Intergovernmental Panel on Climate Change,* T.F. Stocker et al., Eds., Cambridge
956 University Press.

957 Gibson, P. B., D. E. Waliser, H. Lee, B. Tian, and E. Massoud, 2019: Climate Model
958 Evaluation in the Presence of Observational Uncertainty: Precipitation Indices over the
959 Contiguous United States. *J Hydrometeorol*, **20**, 1339–1357,
960 https://doi.org/10.1175/JHM-D-18.

961 Giorgi, F., and W. J. Gutowski, 2015: Regional Dynamical Downscaling and the CORDEX
962 Initiative. *Annu Rev Environ Resour*, **40**, https://doi.org/10.1146/annurev-environ-
963 102014-021217.

De Haan, L. L., M. Kanamitsu, F. De Sales, and L. Sun, 2015: An evaluation of the seasonal added value of downscaling over the United States using new verification measures. *Theor Appl Climatol*, **122**, 47–57, https://doi.org/10.1007/S00704-014-1278-9.

Hamed, K. H., 2008: Trend detection in hydrologic data: The Mann-Kendall trend test under the scaling hypothesis. *J Hydrol (Amst)*, **349**, 350–363, https://doi.org/10.1016/j.jhydrol.2007.11.009.

Haylock, M. R., and C. M. Goodess, 2004: Interannual variability of European extreme winter rainfall and links with mean large-scale circulation. *Int J Climatol*, **24**, https://doi.org/10.1002/joc.1033.

Herold, N., A. Behrangi, and L. V. Alexander, 2017: Large uncertainties in observed daily precipitation extremes over land. *J Geophys Res*, **122**, https://doi.org/10.1002/2016JD025842.

Hobeichi, S., N. Nishant, Y. Shao, G. Abramowitz, A. Pitman, S. Sherwood, C. Bishop, and S. Green, 2023: Using Machine Learning to Cut the Cost of Dynamical Downscaling. *Earths Future*, **11**, https://doi.org/10.1029/2022ef003291.

Hussain et al., 2019: pyMannKendall: a python package for non parametric Mann Kendall family of trend tests. *J Open Source Software*, **4**(39), 1556, https://doi.org/10.21105/joss.01556.

IPCC, 2021: Climate Change 2021: *The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* [Masson-Delmotte, V., P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, In press, doi:10.1017/9781009157896.

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

992 Isphording, R. N., Y. L. Liu, J. Amoils, and S. Wales, 2023: Climate Indices for the
993 CORDEX-Australasia (CMIP5) Ensemble, v1.0. National Computational Infrastructure
994 (NCI) Australia, https://dx.doi.org/10.25914/6cw7-fz24.

995 Jammalamadaka, S. R. and A SenGupta, 2001: *Topics in Circular Statistics*. World Scientific
996 Publishing Co. Pte. Ltd., 176-178 pp.

997 Ji, F., M. Ekström, J. P. Evans, and J. Teng, 2014: Evaluating rainfall patterns using physics
998 scheme ensembles from a regional atmospheric model. *Theor Appl Climatol*, **115**, 297–
999 304, https://doi.org/10.1007/s00704-013-0904-2.

1000 Joetzjer, E., H. Douville, C. Delire, P. Ciais, B. Decharme, and S. Tyteca, 2013: Hydrologic
1001 benchmarking of meteorological drought indices at interannual to climate change
1002 timescales: A case study over the Amazon and Mississippi river basins. *Hydrol Earth*
1003 *Syst Sci*, **17**, 4885–4895, https://doi.org/10.5194/hess-17-4885-2013.

1004 Jones, D. A., W. Wang, and R. Fawcett, 2009: High-quality spatial climate data-sets for
1005 Australia. *Australian Meteorological and Oceanographic Journal*, **58**, 233–248.

1006 Jones, P. W., 1998: First-and Second-Order Conservative Remapping Schemes for Grids in
1007 Spherical Coordinates. *Mon Weather Rev*, **127**, 2204–2210.

1008 de Jong, P., C. A. S. Tanajura, A. S. Sánchez, R. Dargaville, A. Kiperstok, and E. A. Torres,
1009 2018: Hydroelectric production from Brazil's São Francisco River could cease due to
1010 climate change and inter-annual variability. *Science of the Total Environ*, **634**, 1540–
1011 1553, https://doi.org/10.1016/j.scitotenv.2018.03.256.

1012 Kanamitsu, M., and L. Dehaan, 2011: The Added Value Index: A new metric to quantify the
1013 added value of regional models. *J Geophys Res Atmospheres*, **116**,
1014 https://doi.org/10.1029/2011JD015597.

1015 Kirono, D. G. C., V. Round, C. Heady, F. H. S. Chiew, and S. Osbrough, 2020: Drought
1016 projections for Australia: Updated results and analysis of model simulations. *Weather*
1017 *Clim Extrem*, **30**, https://doi.org/10.1016/j.wace.2020.100280.

1018 Knutti, R., D. Masson, and A. Gettelman, 2013: Climate model genealogy: Generation
1019 CMIP5 and how we got there. *Geophys Res Lett*, **40**, 1194–1199,
1020 https://doi.org/10.1002/grl.50256.

1021 Lauer, A., and Coauthors, 2020: Earth System Model Evaluation Tool (ESMValTool) v2.0 -
1022 Diagnostics for emergent constraints and future projections from Earth system models in
1023 CMIP. *Geosci Model Dev*, **13**, 4205–4228, https://doi.org/10.5194/gmd-13-4205-2020.

1024 Liebmann, B., I. Bladé, G. N. Kiladis, L. M. V. Carvalho, G. B. Senay, D. Allured, S.
1025 Leroux, and C. Funk, 2012: Seasonality of African precipitation from 1996 to 2009. *J*
1026 *Clim*, **25**, 4304–4322, https://doi.org/10.1175/JCLI-D-11-00157.1.

1027 Di Luca, A., R. de Elía, and R. Laprise, 2012: Potential for added value in precipitation
1028 simulated by high-resolution nested Regional Climate Models and observations. *Clim*
1029 *Dyn*, **38**, 1229–1247, https://doi.org/10.1007/s00382-011-1068-3.

1030 Martinez-Villalobos, C., J. David Neelin, and A. G. Pendergrass, 2022: Metrics for
1031 Evaluating CMIP6 Representation of Daily Precipitation Probability Distributions. *J*
1032 *Clim*, **35**, 5719–5743, https://doi.org/10.1175/JCLI-D-21.

1033 Mckee, T. B., N. J. Doesken, and J. Kleist, 1993: The Relationship of Drought Frequency and
1034 Duration to Time Scales. *Eighth Conference on Applied Climatol*, 17–22.

1035 Nguyen, P. L., M. Bador, L. V. Alexander, T. P. Lane, and T. Ngo-Duc, 2022: More intense
1036 daily precipitation in CORDEX-SEA regional climate models than their forcing global
1037 climate models over Southeast Asia. *Int J Climatol*, **42**, 6537–6561,
1038 https://doi.org/10.1002/joc.7619.

1039 Nishant, N., S. Sherwood, A. Prasad, F. Ji, and A. Singh, 2022: Impact of Higher Spatial
1040 Resolution on Precipitation Properties Over Australia. *Geophys Res Lett*, **49**,
1041 https://doi.org/10.1029/2022GL100717.

1042 Oueslati, B., and G. Bellon, 2015: The double ITCZ bias in CMIP5 models: interaction
1043 between SST, large-scale circulation and precipitation. *Clim Dyn*, **44**, 585–607,
1044 https://doi.org/10.1007/s00382-015-2468-6.

1045 Perkins, S. E., A. J. Pitman, N. J. Holbrook, and J. McAneney, 2007: Evaluation of the AR4
1046 climate models' simulated daily maximum temperature, minimum temperature, and
1047 precipitation over Australia using probability density functions. *J Clim*, **20**, 4356–4376,
1048 https://doi.org/10.1175/JCLI4253.1.

1049 Prein, A. F., A. Gobiet, M. Suklitsch, H. Truhetz, N. K. Awan, K. Keuler, and G.
1050 Georgievski, 2013: Added value of convection permitting seasonal simulations. *Clim*
1051 *Dyn*, **41**, 2655–2677, https://doi.org/10.1007/s00382-013-1744-6.

1052 Roca, R., L. V. Alexander, G. Potter, M. Bador, R. Jucá, S. Contractor, M. G. Bosilovich, and
1053 S. Cloché, 2019: FROGS: A daily 1° × 1° gridded precipitation database of rain gauge,
1054 satellite and reanalysis products. *Earth Syst Sci Data*, **11**, 1017–1035,
1055 https://doi.org/10.5194/essd-11-1017-2019.

1056 Roundy, P. E., 2015: On the interpretation of EOF analysis of ENSO, atmospheric Kelvin
1057 waves, and the MJO. *J Clim*, **28**, 1148–1165, https://doi.org/10.1175/JCLI-D-14-
1058 00398.1.

1059 Rummukainen, M., 2016: Added value in regional climate modeling. *Wiley Interdiscip Rev*
1060 *Clim Change*, **7**, https://doi.org/10.1002/wcc.378.

1061 Seregina, L. S., A. H. Fink, R. van der Linden, N. A. Elagib, and J. G. Pinto, 2019: A new
1062 and flexible rainy season definition: Validation for the Greater Horn of Africa and
1063 application to rainfall trends. *Int J Climatol*, **39**, 989–1012,
1064 https://doi.org/10.1002/joc.5856.

1065 Sillmann, J., V. V. Kharin, X. Zhang, F. W. Zwiers, and D. Bronaugh, 2013: Climate
1066 extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the
1067 present climate. *J Geophys Res Atmos*, **118**, 1716–1733,
1068 https://doi.org/10.1002/jgrd.50203.

1069 Solman, S. A., and J. Blázquez, 2019: Multiscale precipitation variability over South
1070 America: Analysis of the added value of CORDEX RCM simulations. *Clim Dyn*, **53**,
1071 1547–1565, https://doi.org/10.1007/s00382-019-04689-1.

1072 Spinoni, J., and Coauthors, 2021: Global exposure of population and land-use to
1073 meteorological droughts under different warming levels and SSPs: A CORDEX-based
1074 study. *Int J Climatol*, **41**, 6825–6853, https://doi.org/10.1002/joc.7302.

1075 Sun, Q., C. Miao, Q. Duan, H. Ashouri, S. Sorooshian, and K. L. Hsu, 2018: A Review of
1076 Global Precipitation Data Sets: Data Sources, Estimation, and Intercomparisons. *Rev*
1077 *Geophys*, 56, 79–107, https://doi.org/10.1002/2017RG000574.

1078 Tang, S., P. Gleckler, S. Xie, J. Lee, M.-S. Ahn, C. Covey, and C. Zhang, 2021: Evaluating
1079 Diurnal and Semi-Diurnal Cycle of Precipitation in CMIP6 Models Using Satellite- and
1080 Ground-Based Observations. *J Clim*, **34**, 3189–3210, https://doi.org/10.1175/jcli-d-20-
1081 0639.1.

1082 Tippett, M. K., and M. L. L'Heureux, 2020: Low-dimensional representations of Niño 3.4
1083 evolution and the spring persistence barrier. *NPJ Clim Atmos Sci*, **3**,
1084 https://doi.org/10.1038/s41612-020-0128-y.

1085 Torma, C., F. Giorgi, and E. Coppola, 2015: Added value of regional climate modeling over
1086 areas characterized by complex terrain-precipitation over the Alps. *J Geophys Res*, **120**,
1087 https://doi.org/10.1002/2014JD022781.

1088 Ukkola, A. M., A. J. Pitman, M. G. De Kauwe, G. Abramowitz, N. Herger, J. P. Evans, and
1089 M. Decker, 2018: Evaluating CMIP5 model agreement for multiple drought metrics. *J*
1090 *Hydrometeorol*, **19**, https://doi.org/10.1175/JHM-D-17-0099.1.

1091 U.S. DOE, 2020: *Benchmarking Simulated Precipitation in Earth System Models*. DOE/SC-
1092 0203, U.S. Department of Energy Office of Science, Biological and Environmental
1093 Research (BER) Program. Germantown, Maryland, USA.

1094 Vicente-Serrano, S. M., and Coauthors, 2022: Do CMIP models capture long-term observed
1095 annual precipitation trends? *Clim Dyn*, **58**, 2825–2842, https://doi.org/10.1007/s00382-
1096 021-06034-x.

1097 Di Virgilio, G., J. P. Evans, A. Di Luca, M. R. Grose, V. Round, and M. Thatcher, 2020:
1098 Realised added value in dynamical downscaling of Australian climate change. *Clim Dyn*,
1099 **54**, 4675–4692, https://doi.org/10.1007/s00382-020-05250-1.

1100 Wang, B., and LinHo, 2002: Rainy Season of the Asian-Pacific Summer Monsoon *. *J Clim*,
1101 **15**, 386–398.

1102 World Meteorological Organization (WMO), 2012: Standardized Precipitation Index User
1103 Guide (M. Svoboda, M. Hayes and D. Wood). (WMO-No. 1090), Geneva.

1104 World Meteorological Organization (WMO) and Global Water Partnership (GWP), 2016:
1105 Handbook of Drought Indicators and Indices (M. Svoboda and B.A. Fuchs). Integrated
1106 Drought Management Programme (IDMP), Integrated Drought Management Tools and
1107 Guidelines Series 2. Geneva.

1108 Xiao, M., Q. Zhang, and V. P. Singh, 2015: Influences of ENSO, NAO, IOD and PDO on
1109 seasonal precipitation regimes in the Yangtze River basin, China. *Int J Climatol*, **35**,
1110 3556–3567, https://doi.org/10.1002/joc.4228.

1111 Yang, W., R. Seager, M. A. Cane, and B. Lyon, 2015: The annual cycle of East African
1112 precipitation. *J Clim*, **28**, 2385–2404, https://doi.org/10.1175/JCLI-D-14-00484.1.

1113 Yin, H., M. G. Donat, L. V. Alexander, and Y. Sun, 2015: Multi-dataset comparison of
1114 gridded observed temperature and precipitation extremes over China. *Int J Climatol*, **35**,
1115 2809–2827, https://doi.org/10.1002/joc.4174.

1116 Zhang, X., L. Alexander, G. C. Hegerl, P. Jones, A. K. Tank, T. C. Peterson, B. Trewin, and
1117 F. W. Zwiers, 2011: Indices for monitoring changes in extremes based on daily
1118 temperature and precipitation data. *Wiley Interdiscip Rev Clim Change*, **2**, 851–870,
1119 https://doi.org/10.1002/wcc.147.