Exploratory precipitation metrics: spatiotemporal characteristics, process-oriented, and phenomena-based

L. Ruby Leung[1*], William R. Boos[2], Jennifer L. Catto[3], Charlotte DeMott[4], Gill M. Martin[5], J. David Neelin[6], Travis A. O'Brien[7,8], Shaocheng Xie[9], Zhe Feng[1], Nicholas P. Klingaman[10,11], Yi-Hung Kuo[6], Robert W. Lee[10,11], Cristian Martinez-Villalobos[12], S. Vishnu[2], Matthew Priestley[3], Cheng Tao[9], and Yang Zhou[8]

[1] Pacific Northwest National Laboratory, Richland, WA, USA
[2] University of California at Berkeley, CA, USA
[3] University of Exeter, Exeter, United Kingdom
[4] Colorado State University, Fort Collins, CO, USA
[5] Met Office, Exeter, United Kingdom
[6] University of California at Los Angeles, CA, USA
[7] Indiana University, Bloomington, IN, USA
[8] Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[9] Lawrence Livermore National Laboratory, Livermore, CA, USA
[10] National Centre for Atmospheric Science, Reading, United Kingdom
[11] University of Reading, Reading, United Kingdom
[12] Universidad Adolfo Ibáñez, Peñalolén, Santiago, Chile

* Corresponding author: L. Ruby Leung (ruby.leung@pnnl.gov)

Abstract

Precipitation sustains life and supports human activities, making its prediction one of the most societally relevant challenges in weather and climate modeling. Limitations in modeling precipitation underscore the need for diagnostics and metrics to evaluate precipitation in simulations and predictions. While routine use of basic metrics is important for documenting model skill, more sophisticated diagnostics and metrics aimed at connecting model biases to their sources and revealing precipitation characteristics relevant to how model precipitation is used are critical for improving models and their uses. This paper illustrates examples of exploratory diagnostics and metrics including: (1) spatiotemporal characteristics such as diurnal variability, probability of extremes, duration of dry spells, spectral characteristics, and spatiotemporal coherence of precipitation; (2) process-oriented metrics based on the rainfall-moisture

Early Online Release: This preliminary version has been accepted for publication in *Journal of Climate*, may be fully cited, and has been assigned DOI 10.1175/JCLI-D-21-0590.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

coupling and temperature-water vapor environments of precipitation; and (3) phenomena-based metrics focusing on precipitation associated with weather phenomena including low pressure systems, mesoscale convective systems, frontal systems, and atmospheric rivers. Together, these diagnostics and metrics delineate the multifaceted and multiscale nature of precipitation, its relations with the environments, and its generation mechanisms. The metrics are applied to historical simulations from the Coupled Model Intercomparison Project Phase 5 and Phase 6. Models exhibit diverse skill as measured by the suite of metrics, with very few models consistently ranked as top or bottom performers compared to other models in multiple metrics. Analysis of model skill across metrics and models suggests possible relationships among subsets of metrics, motivating the need for more systematic analysis to understand model biases for informing model development.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

1. Introduction

Precipitation is a key component of the water cycle connecting processes across the atmosphere, land, ocean, and cryosphere (Trenberth et al. 2007). Through decades of development, the current generation of climate models uses increasingly sophisticated, physically-based subgrid parameterizations of convection and cloud microphysics to simulate precipitation, although their horizontal resolutions are still typically much coarser than needed to explicitly resolve precipitation formation processes. When, where, how often, and how much precipitation falls have significant implications for the energy, water, and biogeochemical cycles of the Earth system. For example, biases in soil moisture can often be linked to biases in precipitation amount, frequency, and intensity, which influence the partitioning of precipitation into evapotranspiration, runoff, and soil moisture storage, with subsequent impact on surface temperature through evaporative cooling (Qian et al. 2006). Relatedly, biases in modeling the surface water and energy balance due to precipitation biases can influence clouds, convection, and precipitation through energetic constraints and land-atmosphere feedbacks. Because of the myriad Earth system interactions and feedbacks mediated by precipitation, skillful modeling of precipitation and understanding and attribution of precipitation biases are scientifically challenging (Dai 2006; Covey et al. 2016; Chen et al. 2021). As precipitation biases are among the most consequential in limiting the use of climate models for decision support, there is an urgent need to improve precipitation modeling across a wide range of spatial and temporal scales (Tapiador et al. 2019).

Quantifying and understanding model precipitation biases is an important step towards improving the overall quality of climate simulations and predictions. Metrics are objective measures for benchmarking model performance against observations and facilitating model intercomparison. Common metrics of precipitation have focused on aspects such as the spatial distribution of annual and seasonal mean precipitation, daily precipitation amount, frequency, and intensity, and the probability density function of precipitation rate (Deser et al. 2012; Chen and Dai 2018, 2019). Increasingly, metrics related to extremes such as annual maximum daily precipitation and consecutive dry days have also been used to evaluate precipitation characteristics connected more closely to societal impacts. These metrics have revealed multiple longstanding precipitation biases in climate models. For example, climate models tend to produce too frequent light daily precipitation, but not enough high intensity daily precipitation compared to observations (Dai 2006; Stephens et al. 2010; Chen et al. 2021), while sub-daily intensities can vary considerably between models (e.g., Klingaman et al., 2017). Most global climate models simulate a spurious inter-tropical convergence zone (ITCZ) in the southeastern Pacific and South Atlantic, resulting in a double-ITCZ bias that is most prominent during boreal winter (Mechoso et al. 1995; Lin 2007;

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

Mapes and Neale 2011; Hwang and Frierson 2013; Oueslati and Bellon 2013; Hirota et al. 2014; Tian 2015; Tian and Dong 2020). Erroneous diurnal timing of precipitation over land is another common bias, which is most noticeable during boreal summer in regions such as the central U.S. featuring nocturnal peaks in precipitation (Dai et al. 1999; Tang et al. 2021). Precipitation biases have also been identified in regions with complex terrain such as the western U.S. (Mejia et al. 2018) and Europe (Mehran et al. 2014), in Amazonia (Yin et al. 2013), and in monsoon regions such as Asia (Sperber et al. 2013).

Although precipitation diagnostics and metrics have been incorporated in model evaluation and diagnostic packages such as ESMValTool (Eyring et al. 2020) and PCMDI Metrics Package (PMP; Glecker et al. 2016) used by climate modeling centers and the climate science community, they focus on limited aspects of precipitation for benchmarking global climate simulations. At the same time, over the past few years new precipitation diagnostics and metrics have been developed to deconvolve and better understand model precipitation biases. For example, Ma et al. (2013) proposed a set of metrics and diagnostics to evaluate and diagnose tropical precipitation biases and associated moist processes in climate models. Their proposed diagnostics include stratiform fraction of precipitation, probability density function of daily precipitation intensity, composites of column water vapor, column relative humidity, temperature, and specific humidity profiles as a function of precipitation intensity, and composites of stratiform rainfall fraction as a function of column relative humidity. Klingaman et al. (2017) developed a set of diagnostics and metrics for analyzing precipitation intensity and coherence on a range of time and space scales.

This study represents a collaborative effort as an outgrowth of a workshop on "Benchmarking Simulated Precipitation in Earth System Models" (Pendergrass et al. 2020) to develop more advanced precipitation metrics and demonstrate their use in benchmarking diverse aspects of precipitation from climate simulations. Three types of precipitation diagnostics and metrics are presented: (1) spatiotemporal characteristics, such as diurnal variability, probability of extremes, duration of dry spells, spectral characteristics, and spatiotemporal coherence of precipitation; (2) process-oriented, based on the rainfall-moisture coupling and temperature-water vapor environments of precipitation; and (3) phenomena-based, focusing on precipitation associated with weather phenomena such as low pressure systems, mesoscale convective systems, frontal systems, and atmospheric rivers. These diagnostics and metrics take advantage of analysis building on advances in understanding the thermodynamic environments of precipitation (e.g., Bretherton et al. 2004; Neelin et al. 2009; Kuo et al. 2018; Chen et al. 2020) and their role in modes of variability (e.g., Wolding et al. 2020), and in tracking weather features such as atmospheric rivers (e.g., Shields et al. 2018).

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

While examples of the above metrics have been reported in recent literature (e.g., Klingaman et al. 2017; Ahmed and Neelin 2020; Feng et al. 2021a), they are deemed exploratory partly because they have not been widely used or implemented in standard metrics and diagnostics packages and partly because they allow deeper exploration of precipitation characteristics and associated processes. Some of these diagnostics and metrics require variables besides precipitation to evaluate relationships with environmental conditions, or to track weather features, so their data requirements go beyond the baseline precipitation metrics already implemented in widely used metrics and diagnostics packages (Pendergrass et al. 2020). Furthermore, additional research may be needed on interpretations of results from use of these metrics, to standardize their use, or to address technical or computational issues. Here, we apply the exploratory metrics to a common set of climate simulations from the Coupled Model Intercomparison Project Phase 5 (Taylor et al. 2012) and Phase 6 (CMIP6) (Eyring et al. 2016). While our aim is not to provide an exhaustive study of the ability of these models to represent precipitation, we illustrate how such diagnostics and metrics may be used to evaluate broader aspects of precipitation in climate simulations and to explore insights that may be gained through comparative analysis of multiple metrics. With increasing model resolutions to better resolve weather and large-scale environments (e.g., Haarsma et al. 2016), the exploratory diagnostics and metrics may be even more relevant not only for benchmarking models but also for understanding the causes of model precipitation biases. They also provide useful information to support the growing and more diverse uses of precipitation from climate models and improve communications of climate model performance by connecting precipitation to commonly understood weather phenomena. A collection of such exploratory diagnostics and metrics is a valuable addition to the existing precipitation diagnostics and metrics packages that are used in the community.

We briefly summarize the observational data, climate model output, and the feature tracking methods in Section 2. Key results are presented in Sections 3-5 for the spatiotemporal characteristics metrics, process-oriented metrics, and phenomena-based metrics, respectively. Each area is presented as a module describing the diagnostics and metrics and the results of applying them to climate model outputs summarized in a multi-panel figure. We conclude with discussion and summary in Section 6.

2. Data and Feature Tracking Methods

2.1 Observational data and climate model outputs

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

Several observational precipitation data products are used for benchmarking precipitation from climate simulations. These include: (1) Tropical Rainfall Measurement Mission (TRMM) Multi-satellite Precipitation Analysis (TMPA-RT) (3B42; Huffman et al. 2007); (2) Remote Sensing Systems TRMM Microwave Imager (TMI) Daily Environmental Suite on 0.25 deg grid, Version 7.1 (Wentz et al. 2015); (3) TRMM Precipitation Radar (PR) Rainfall Rate and Profile L2 1.5 hours V7 (2A25; TRMM 2011); (4) Monthly and daily Global Precipitation Climatology Project (GPCP) V3 combined precipitation data set (Huffman et al. 2020); (5) CMORPH bias-corrected integrated satellite precipitation estimates (Joyce et al. 2011); (6) Precipitation Estimation from Remotely Sensed Information (PERSIANN) (Ashouri et al. 2015); and (7) Global Precipitation Measurement (GPM) Multi-satellitE Retrievals (IMERG) precipitation data V06B (Tan et al. 2019). They represent a diverse set of precipitation data derived from satellite and ground-based remote sensing retrievals. In addition, ground-based precipitation observations at the DOE Atmospheric Radiation Measurement (ARM) program's Southern Great Plains (SGP) and Manacapuru (MAO) sites are also used. The ARM data used in this study are from the ARM best estimate (ARMBE, Xie et al. 2010) data products and the ARM long-term continuous variational analysis (VARANAL, Xie et al. 2004). At these ARM sites, the available surface rain gauge measurements and/or radar retrievals provide additional information to validate satellite-based precipitation products. Table 1 summarizes the spatial and temporal resolution and domain coverage of these datasets. While the highest spatial resolution available for the data set is given in Table 1, coarse-graining of the data for comparison to models is described with each metric. As different exploratory diagnostics and metrics have different requirements for precipitation data, we do not standardize the use of observational precipitation data in calculating the metrics, but recognize the need to address uncertainty in observed precipitation products in use and interpretation of metrics.

Besides precipitation data, several global reanalysis products are used to provide gridded data of the atmospheric environments needed for calculation of some process-oriented metrics and identification and tracking of weather features for the phenomena-based metrics. They include: (1) ERA-Interim (Dee et al. 2011); (2) ERA5 (Hersbach et al. 2020; Hoffman et al. 2019); (3) MERRA-2 (Gelaro et al. 2017); and (4) CFSR (Saha et al. 2010). Lastly, the NASA Global Merged IR V1 infrared brightness temperature (Tb) data (Janowiak et al. 2017) are also used to track mesoscale convective systems (MCSs). The spatial and temporal resolutions of the reanalysis products and T_b data are also summarized in Table 1.

The exploratory metrics are applied to benchmark precipitation from the Coupled Model Intercomparison Project Phase 5 (Taylor et al. 2012) and Phase 6 (Eyring et al. 2016), with typical horizontal resolution of ~1 degree. Two of the metrics on low pressure systems and mesoscale convective systems are applied to

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

precipitation from several high-resolution simulations from HighResMIP (Haarsma et al. 2016) as these weather features are better defined and more reasonably resolved at higher resolution. In HighResMIP, high-resolution simulations have nominal resolutions ranging from 0.25 to 0.5 degrees, with their low-resolution counterparts ranging from 1.0 to 1.4 degrees. Table 2 summarizes the variables and their temporal frequency used to calculate the various metrics.

Table 1. Observational and reanalysis data for benchmarking models. P, Q, U, V, T, CVW, and IR T_b are precipitation, specific humidity, zonal wind, meridional wind, temperature, column water vapor, and infrared brightness temperature, respectively.

	Variables	**Temporal resolution**	**Max. spatial resolution**	**Period of coverage**	**Domain of coverage**
GPCP	P	Monthly	0.25º x0.25º	1979-2020	Global
GPCP 1DD	P	Daily	1º	1996-present	Global
CMORPH	P	30 min	8 km	1998-2017	60ºS – 60ºN
PERSIANN-CDR	P	Monthly	0.25º x0.25º	1983-2017	60ºS – 60ºN
TRMM 3B42	P	3-hourly	0.25º x0.25º	1998-2019	50ºS – 50ºN
TRMM-TMI	CWV	Twice-daily snapshot	0.25º x0.25º	2002-2014	40ºS – 40ºN
TRMM PR 2A25	P	Twice daily snapshot	~5 km	2002-2014	40ºS – 40ºN
GPM-IMERG	P	Hourly	0.1º x0.1º	2001-2020	60ºS – 60ºN
ARMBE	P	Hourly	Single point	SGP: 1993-2018 MAO: 2015-2015	Single point
VARANAL	P	SGP: hourly MAO: 3-hourly	0.5º x0.5º	SGP: 2004-2018 MAO: 2014-2015	SGP: 3º x3º MAO: 2º x2º
NASA Global Merged IR V1	IR T_b	Hourly	Raw data at 4 km but coarsened to 0.1º x0.1º	2000-2019	60ºS – 60ºN
ERAI	Q, U, V, T	3-hourly	80 km	1979-2019	Global
ERA5	Q, U, V, T	1-hourly	30 km	1979-2019	Global
MERRA-2	Q, U, V, T	3-hourly	50 km	1980-2019	Global
CFSR	Q	6-hourly	38 km	1979-2019	Global

Table 2. Variables and their temporal frequency used to calculate various precipitation metrics and the objectives of the metrics.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

Metrics	Variables and temporal frequency	Objectives
Diurnal cycle of precipitation	3-hourly precipitation	Intercompare a large number of models with observations and with each other on the diurnal cycle of precipitation over different climate regimes
Extremes of daily precipitation and duration of dry spells	Daily precipitation	Use characteristic scales governing probabilities in the large-event regime for dry and wet precipitation extremes to capture the performance of models
Spectral analysis of precipitation	3-hourly and daily precipitation sampled over ~20 years of data, by season and annual.	Examine the ability of models to represent the range of precipitation intensities typically occurring at any location, on 3-hourly and daily timescales
Coherence analysis of precipitation	3-hourly and daily precipitation	Measure and compare the spatial and temporal scales of precipitation across observations and models
MJO East/West power ratio and Maritime Continent propagation	Daily precipitation	Evaluate the relationship between precipitation spatial coherence and MJO propagation across the Maritime Continent in models
Rainfall-moisture coupling	Daily precipitation and vertically integrated water vapor and saturation water vapor	Evaluate the coupling of tropical rainfall and moisture in models and how this coupling affects MJO simulation
Temperature-water vapor environment	3-hourly/hourly vertically integrated saturation humidity and snapshots of column water vapor (CWV) and precipitation	Quantify the thermodynamic environment that produces most precipitation at sub-daily timescales
Low pressure systems	6-hourly 850 hPa values of zonal wind, meridional wind, temperature, and specific humidity; 6-hourly precipitation	Track LPS in observations and simulations and compare their depiction of number, structure, and rainfall
Mesoscale convective systems	Hourly outgoing longwave radiation and precipitation	Track MCSs in observations and simulations and compare their depiction of MCS number and

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

		rainfall.
Frontal precipitation	6-hourly 850-hPa zonal and meridional wind components, specific humidity, temperature, and daily precipitation	Use fronts as a precipitation regime to decompose precipitation errors into frontal and non-frontal, and to quantify the representation of the dynamical impact of fronts on precipitation intensity.
Atmospheric rivers	3-hourly or 6-hourly zonal and meridional wind components, specific humidity, surface pressure, and precipitation	Assess whether models simulate AR-related precipitation in the correct locations and with enough contrast between regions with high AR precipitation and low AR precipitation.

2.2 Feature Identification and tracking methods

The phenomena-based metrics require identification and tracking of weather features in observations and simulated precipitation. A brief description of methods used to track low pressure systems (LPS), mesoscale convective systems (MCS), front systems (FRT), and atmospheric rivers (AR) are provided below while more detailed descriptions are provided in the cited references.

Low pressure systems (LPS): The TempestExtremes feature tracking algorithm (Ullrich and Zarzycki, 2017) is used to track tropical low pressure systems (LPS) by identifying extrema in candidate tracking variables. A systematic assessment of multiple candidate variables, hundreds of quantitative tracking criteria, and several vertical levels led to selection of the streamfunction of the 850 hPa horizontal wind (Vishnu et al. 2020) as the optimal tracking variable. Streamfunction minima were used to identify lower-tropospheric cyclonic vortices within 35º of the equator in the ERA5 reanalysis, 4 HighResMIP models, and the 0.25º-resolution E3SM model (Caldwell et al. 2019). The streamfunction was calculated from the horizontal wind for each dataset, with any wind velocities that were extrapolated below Earth's surface (e.g., in ERA5) set to zero before solving the Poisson problem for the streamfunction (Vishnu et al. 2020). The resulting track dataset for ERA5, together with tracks for 4 other reanalyses, are available in a Zenodo repository (doi:10.5281/zenodo.3890646).

Mesoscale convective systems (MCS): The FLEXTRKR algorithm is used to track MCSs in observations and model simulations. An MCS is defined as a convective system with: (1) cold cloud shield (CCS) >

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

$4x10^4$ km^2 containing a precipitation feature (PF) with major axis length > 100 km, (2) PF area, mean rain rate, rain rate skewness and heavy rain volume ratio larger than corresponding lifetime dependent thresholds, and (3) both (1) and (2) last continuously for longer than 4 hours. As in Feng et al. (2021b), CCS is tracked using geostationary satellite T_b data and defined using a threshold of T_b < 241 K. For model simulations, T_b is derived based on simulated outgoing longwave radiation following the empirical formulation provided by Yang and Slingo (2001). PF is tracked using the IMERG hourly precipitation data and PFs are defined as contiguous areas within the CCS with hourly rain rate > 2 mm h^{-1}.

Frontal precipitation (FRT): Fronts are identified using an automated method applied to 6-hourly gridded data at 2.5 degree resolution (Berry et al 2011, Catto et al 2015). This method calculates a thermal front parameter (TFP) as function of a thermal parameter:

$$TFP(\theta_W) = -\nabla|\nabla\theta_W|.\left(\frac{\nabla\theta_W}{|\nabla\theta_W|}\right)$$

While many variables can be used to calculate the thermal front parameter (TFP) (Thomas and Schultz 2019), we have used the wet bulb potential temperature (θ_W) as in Hewson (1998). After calculating the TFP, the field is masked where this is above a fixed negative threshold. Frontal points are then defined as the locations where the gradient of the TFP is equal to zero. These points are joined into contiguous lines and regridded as binary objects with an area of influence of plus and minus one grid box. Fronts can be separated into warm, cold, and quasi-stationary fronts, but here we have maintained simplicity by considering all fronts together.

Atmospheric rivers (AR): With few exceptions, previous studies have utilized only a single AR detection tool (ARDT) in each study, whereas over 30 ARDTs currently exist (Shields et al 2018; Rutz et al 2019). Recent results from the Atmospheric River Tracking Method Intercomparison Project (ARTMIP) project have demonstrated that different ARDTs can produce different scientific results, which suggests that multiple ARDTs may need to be used when evaluating climate models in order to gain a complete picture of model skill in simulating ARs (O'Brien et al 2020b). ARs are detected globally using six independently-developed ARDTs, which we refer to by the following code names: ARCONNECT v2 (Shearer et al. 2020), GuanWaliser v2 (Guan et al. 2018), Lora v2 (Skinner et al. 2020), Mundhenk v3 (Mundhenk et al. 2016), TECA BARD v1.0 (O'Brien et al. 2020), and Tempest LR (McClenny et al. (2020)) (O'Brien et al. 2021). These ARDTs were run on output from the MERRA-2 reanalysis as part of the ARTMIP Tier 1 experiment (Shields et al 2018) and on output from the CMIP5 and CMIP6 multimodel ensembles as part of the ARTMIP Tier 2 CMIP5/6 experiment (O'Brien et al 2021). The

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

methods use a variety of heuristic rules to objectively identify atmospheric rivers from integrated vapor transport (IVT; the vertical integral of horizontal moisture transport) and/or integrated water content. For example, the widely-used GuanWaliser v2 algorithm identifies ARs as continuous regions of integrated vapor transport exceeding the climatological 85th percentile, if the continuous regions meet specific geometric thresholds indicative of long and narrow regions of intense poleward moisture transport. We employ multiple ARDTs because recent literature indicates that different ARDTs may, in some instances, lead to qualitatively different answers to the same question (O'Brien et al 2020a; O'Brien et al 2020b; Zhou et al 2021).

3. Spatiotemporal characteristics metrics

Precipitation variability at different spatial and temporal scales is associated with specific processes such as convection driven by diurnal solar heating at the land surface, seasonal moisture convergence related to monsoon systems, disturbances related to convectively coupled equatorial waves, and large-scale atmosphere-ocean interactions. Diagnostics and metrics of spatiotemporal precipitation characteristics are therefore useful for relating model biases to specific mechanisms of precipitation generation at relevant ranges of spatiotemporal scales. These metrics are also useful for informing use of climate model precipitation data at appropriate spatiotemporal scales. Four metrics to benchmark the diurnal cycle of precipitation, daily precipitation and duration of dry spells, fractional contribution to the total mean rainfall from different intensities, and spatial and temporal coherence of precipitation are discussed in this section.

Diurnal cycle of precipitation

The Fourier analysis has been widely applied to quantifying the diurnal cycle of precipitation in both observations and GCMs. However, the model simulated rainfall is often quite noisy and therefore is poorly fit by low-order Fourier harmonics at single grid points. With these, Covey et al. (2016) proposed a summary metric which illustrates the model simulated Fourier amplitude and phase, averaged separately over all land and all ocean areas, in a single two-dimensional map. This metric enables intercomparison of climate models with observations and with each other over different climate regimes, but it becomes problematic when the size of models increases. Here we extend the procedure of Covey et al. (2016) and propose a metric that clearly displays the Fourier amplitude and phase of each individual model from a large number of groups in one bar plot (Tao et al. 2021).

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

Figure 1a shows an example of the composite diurnal harmonic amplitude and phase (in LST) of summertime precipitation from 24 CMIP6 models vs observations over land. To generate the figure, we first produce a composite diurnal time series of precipitation, averaged over many years, for each grid point. We then apply Fourier analysis on the composite diurnal cycle of precipitation and focus on the first harmonic component, following Dai (2001). Here, the diurnal harmonic amplitude and phase are averaged over all land points between 50°S-50°N using a vector averaging method, which automatically down weights the areas with a weak diurnal cycle (Covey and Cleckler, 2014; Covey et al. 2016). Model precipitation is evaluated for the period of 1996-2005. Previous studies (e.g., Dai et al. 2007) have indicated that a stable diurnal cycle can be obtained with just a few years of data. As shown, the two satellite-based observations (TRMM 3B42 v7 and GPM-IMERG) agree quite well with each other in terms of both diurnal amplitude and phase. Over land, the major deficiency of the models is the too early diurnal precipitation peak, consistent with previous studies (e.g., Dai, 2006; Xie et al. 2019). The majority of the models show a diurnal harmonic phase peaking between 12-15 LST instead of early evening from the observations. The observed early morning diurnal harmonic phase over the ocean is generally captured by most of the CMIP6 models (Fig. 1b) while the corresponding diurnal harmonic amplitude is somewhat underestimated in all 24 CMIP6 models. To highlight the models with the best performance, Fig. 1c and 1d shows the scatter plot of absolute model bias to TRMM observations in diurnal harmonic phase vs. amplitude over land and over ocean, respectively. Over ocean, interestingly, models that perform better in the diurnal cycle phase tend to perform worse in amplitude (Fig. 1d). The relationship between model biases in precipitation diurnal phase and amplitude over land is less significant than that over ocean but there is a tendency for models with smaller bias in phase to have correspondingly smaller bias in amplitude (Fig. 1c). Particularly, EC-Earth3, EC-Earth3-Veg and EC-Earth3-Veg-LR compare the best to the observations over land in terms of both diurnal amplitude and phase. Similar results are found by interpolating the data to a common grid (not shown). Generally, the impact of model resolution on the simulated diurnal cycle of precipitation is minimal.

The metric diagram can also be easily computed for smaller scales and at different locations where rich ground-based high-frequency observations are available. Figures 1e-1f compare the simulated diurnal harmonic amplitude and phase to observations at the two ARM sites (SGP and MAO) where precipitation shows distinct diurnal variability with SGP featuring a nocturnal peak while MAO having an afternoon peak. Despite some discrepancies, the satellite-based products agree fairly well with the ground-based rain gauge and/or radar measurements in general. As is shown in Fig. 1e, there is a large model spread in both diurnal amplitude and phase at SGP, with most of the models (except for two) failing to capture the observed nocturnal peak around mid-night in which half of the models actually show a diurnal precipitation peak in

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

the afternoon. CNRM-CM6-1-HR and FGOALS-g3 simulate the diurnal phase much closer to that observed but both significantly underestimate the diurnal harmonic amplitude. The majority of models show a diurnal precipitation peak around noon at MAO, a few hours earlier than the observed (Fig. 1f). In general, the CMIP6 models show a diverse result in simulating the diurnal amplitude with some overestimating the observed value and some underestimating it, but they often show consistent biases in simulating the diurnal phase. Almost all the models peak too early during the day and miss the nocturnal diurnal peak at certain regions.

To summarize, the metric developed here provides quick comparison with observations and among models, and reasonably summarizes the systematic model errors in reproducing the diurnal cycle of precipitation over both large areas and single point locations. Particularly, by displaying the diurnal harmonic amplitude and phase from the Fourier analysis in one bar plot, this metric enables the evaluations with a focus on individual model performance from a large number of models.

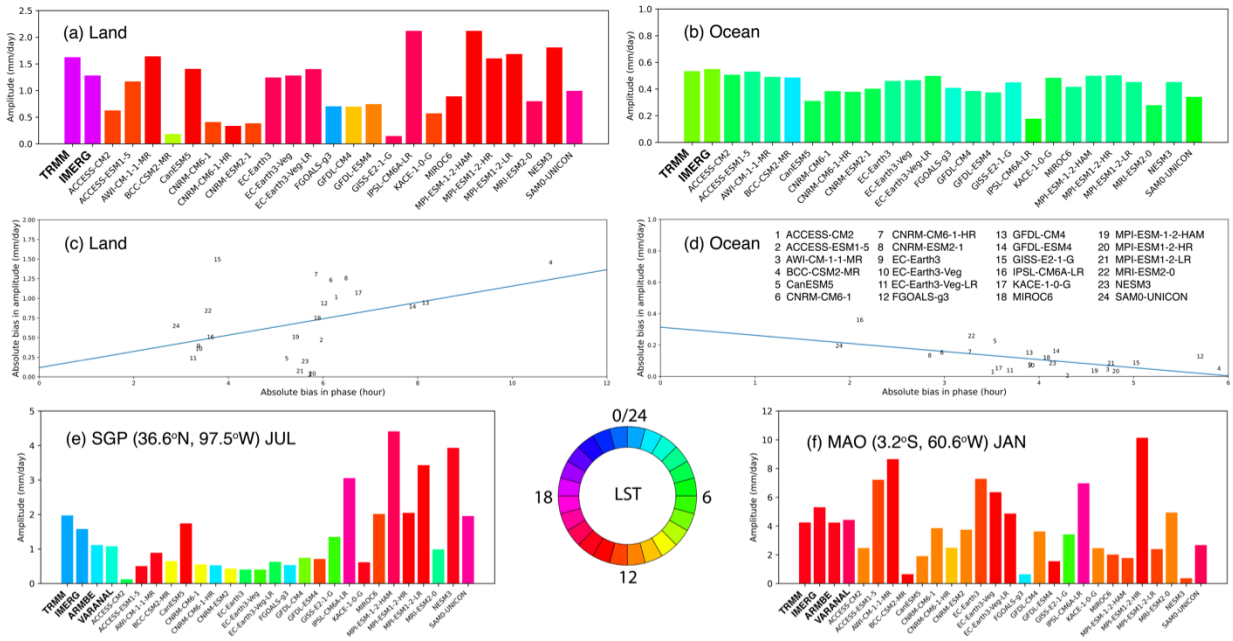

Figure 1. (a) Bar plot of the composite mean diurnal harmonic amplitude (y-axis) and phase in LST (color) of summertime precipitation averaged over land, (b) Same as (a) but over ocean, (c) Scatter plot of absolute bias in diurnal harmonic phase vs. amplitude over land, (d) Same as (c) but over ocean, (e)-(f) Same as (a), but for ARM SGP and MAO sites, respectively. Here, summertime precipitation refers to July for the Northern Hemisphere and January for the Southern Hemisphere. Model precipitation for 24 CMIP6 historical simulations are examined for the years 1996-2005.

Extremes: daily precipitation and duration of dry spells

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

While seemingly contrasting variables, daily precipitation and the duration of dry spells share many features in the shape of their probability distributions. The probability density functions (PDFs) of both quantities are characterized by a power law range, where the probability decreases slowly with each order of magnitude increase in precipitation rate or duration of dry spells, up to a cutoff-scale (denoted P_L for daily precipitation and t_L for the duration of dry spells; see Figs. 2a,b) where the probability decreases roughly exponentially (Figs. 2a,b), ultimately controlling the size of extreme percentiles (Martinez-Villalobos and Neelin 2018,2021; Chang et al., 2020). These quantities have connections with the moisture budget, with P_L (and hence also extreme percentiles) scaling with the amplitude of moisture convergence fluctuations within precipitating events (Neelin et al., 2017; Martinez-Villalobos and Neelin 2019), and t_L scaling with the balance between moisture convergence fluctuations at dry times and the mean moisture source tendency (Pierrehumbert et al., 2007; Stechmann and Neelin 2014).

Recently, Martinez-Villalobos and Neelin (2021) shows that the shape of the large daily precipitation probability tail and the spatial pattern of the cutoff-scale are well simulated by GCMs but there is a bias in the magnitude of P_L compared to observational datasets (see also Fig. 2a). This suggests that two metrics can succinctly summarize the general model behavior of daily precipitation and dry-spell duration extremes. The first one is the spatial correlation coefficient over 50S-50N (the spatial extent of TRMM-3B42; see Table 1) between model simulated P_L and t_L patterns (see Figs. 2c,d for their CMIP6 multi-model mean) and their observational counterparts (TRMM-3B42 in this case). The second metric is the scaling factor, defined as the model area weighted mean P_L or t_L divided by the TRMM-3B42 observational estimate of the same quantity. The first metric tests whether extremes are well simulated spatially regardless of magnitude (values can range between -1 and 1, with 1 denoting a model that simulates the spatial pattern of TRMM-3B42), and the second tests the overall magnitude of the pattern (values can range between 0 and infinity, with 1 being the best). To gauge model behavior, we also calculate the same metrics comparing GPCP vs TRMM-3B42 as a measure of observational uncertainty. The differences between observational precipitation products can be large, thus, model results may be sensitive to the choice of target observational product. This sensitivity is discussed in section 4 and Fig. S2 of the Supplementary Material. We note the caveat that part of the differences between models and observational products noted below may be the result of sampling different internal variability realizations (Deser et al., 2012) due to the relatively short span in which precipitation observational products have been available. However, different realizations from models of the same family (e.g., GFDL models, CNRM models) tend to perform similarly, which suggest that sampling variability has only a minor effect on the results. More details on these metrics and methodology are given in the Supplementary Material.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

Figures 2e and 2f show the results for 35 CMIP6 models and for the multi-model mean ensemble (MME) for P_L and t_L respectively. We first find that there is a substantial observational uncertainty for P_L. The overall magnitude of P_L in GPCP is about 70% (scaling factor of 0.68) of TRMM-3B42 magnitude and the correlation coefficient of the patterns is 0.81. There are several models that are closer to TRMM-3B42 than the observational uncertainty. Among these we highlight HADGEM3-GC31-MM as the model with the closest P_L spatial pattern (r=0.89) and GFDL-ESM4 as the model with the closest overall magnitude to TRMM-3B42 (scaling factor=0.98). The MME benefits from the good performance of the best models in the spatial structure and averages the overall magnitude of P_L in the different models. This results in a multi-model mean that is closer than GPCP to TRMM-3B42 in both P_L spatial pattern and magnitude.

The model performance on the duration of dry spells is similarly encouraging. While all individual models and the MME simulate longer duration of dry spells than both TRMM-3B42 and GPCP (even after the models wet-day biases are greatly reduced. See Supplementary Information), the t_L pattern correlation in almost all models is comparable, although reduced, with the pattern correlation between TRMM-3B42 and GPCP. Even though the magnitude of P_L and t_L (hence also extreme percentiles) differs from TRMM-3B42 in almost all models, the fact that the patterns are well correlated helps boost confidence in model projections of relative (i.e., percent) changes of daily precipitation and dry-spell duration extremes.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

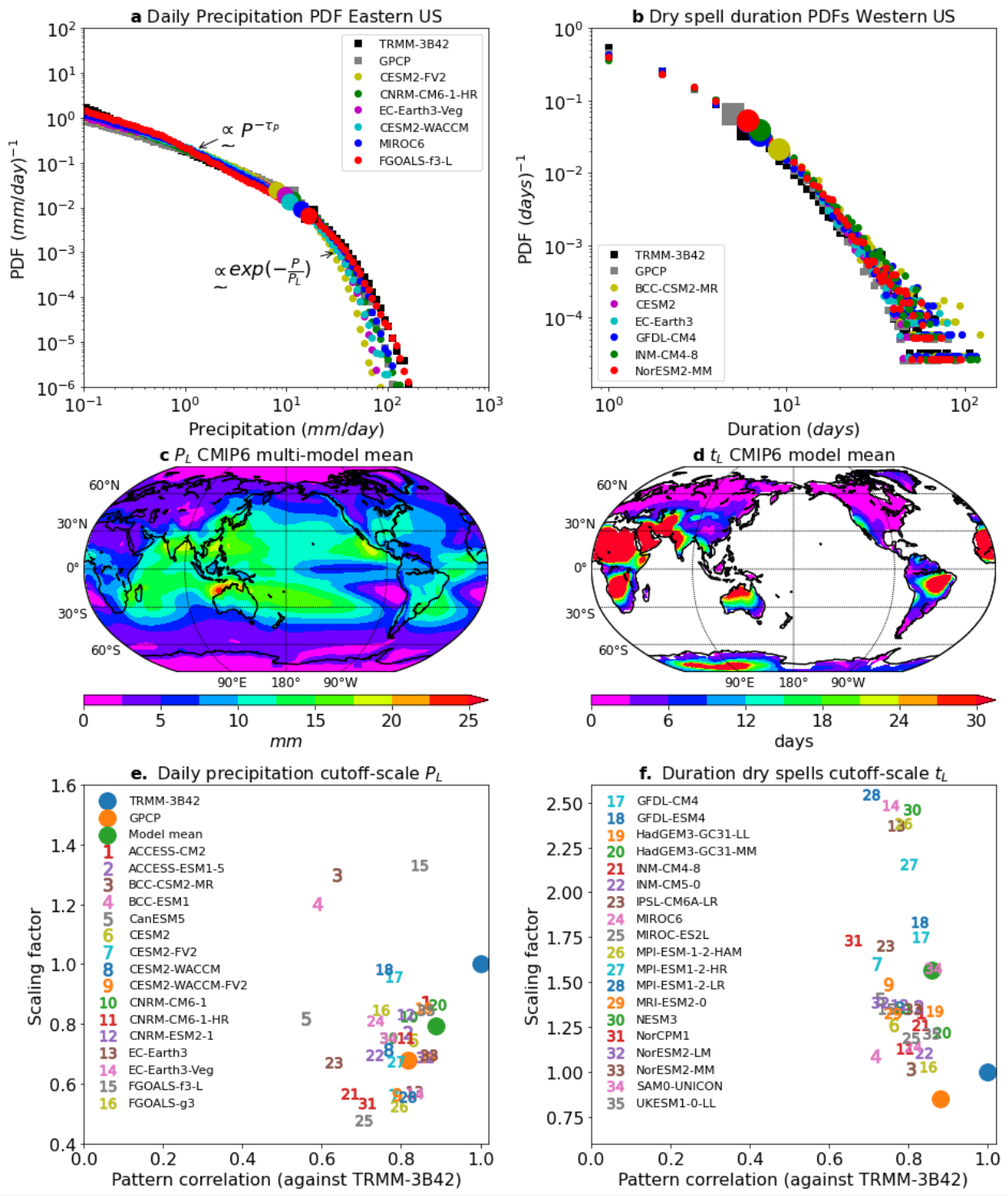

Figure 2. Observational (GPCP and TRMM-3B42) and selected models: a. daily precipitation PDFs in the Eastern United States (25N-48N,257E-294E) and b. dry spell durations PDFs in the Western United States (30N-48N,236E-257E). In a,b the cutoff-scales are shown by a large circle (for models) or large squares (for observational datasets). Note that the larger or longer the cutoff-scale, the more extreme is the large event tail. c. Multi-model mean (out of 35 models) of the daily precipitation cutoff-scale P_L pattern. d. Multi-model mean of the dry spell duration cutoff-scale t_L pattern (with model-dependent dry-day precipitation threshold). e. Scatter plot of the P_L scaling factor and pattern correlation coefficient against TRMM-3B42 for individual models (numbers; legend across panels e-f gives

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

corresponding acronyms), multi-model mean (green dot), GPCP (orange dot), and TRMM-3B42 (blue dot. (1,1) by definition). f. As in e, but for t_L scaling factor and pattern correlation coefficient.

Spectral analysis

Following the method of Klingaman et al. (2017) implemented in Analyzing Scales of Precipitation (ASoP) version 1.0, we calculate the fractional contribution to the total mean rainfall from different intensities, at 3hr and daily timescales, sorted into 100 bins of varying width ranging from 0.005 to 2,360 mm day^{-1}. This reveals the relative importance of precipitation events in a given intensity bin to the total precipitation. The calculation is performed at each grid box, using a horizontal resolution that is sufficiently coarse for at least some spatial averaging to be carried out for all of the models and the observations. In order to avoid removing important spatial detail, we limit this resolution to 2° x 2°, thereby requiring us to omit models whose resolution is similar to or coarser than this. Calculations are performed for the whole year (ANN) and for each season, over 25 years (1990-2014) of CMIP6 historical simulations.

In order to evaluate the models, we use a similarity index (Perkins et al., 2007) to compare the fractional histograms from each model with those obtained from 19 years of GPM-IMERG observations (2001-2019) at each grid point between 60°S and 60°N. This measures the overlap between the model and observed histograms, with values closer to 1.0 indicating that the histograms match better and 0.0 if they are entirely separated. Metrics are the spatial root mean square of these indices over selected regions. Any region could be chosen for metric evaluation; here we have used 6 regions: Global (60S to 60N); Tropics (15S to 15N); Land-only (30S to 30N); Sea-only (30S to 30N); Northern hemisphere (NH) mid-latitudes (30-60N); Southern hemisphere (SH) mid-latitudes (30-60S).

Figure 3(a) shows an example map of the indices from 3hr rainfall data from HadGEM3-GC31-MM vs GPM-IMERG. This suggests that performance is better over land than ocean, and over mid-latitudes than the Tropics. Figure 3(b) shows the overall metric summary information for the 3hr timescale. This confirms that the pattern seen for HadGEM3-GC31-MM is similar for the other CMIP6 models and is consistent through the seasons. The stars indicate comparison of GPM-IMERG with other observation datasets, providing a measure of uncertainty. The metrics from the models nearly all lie outside this uncertainty range. Figure 3(c) provides additional information about the model-observation differences: the models are generally biased towards smaller rainfall accumulations, although there are a few models for which there is a greater than observed contribution from the largest rainfall accumulations. We find similar results for daily accumulations.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

The metrics are a useful guide to the overall model performance, but the fact that the histograms are analyzed at each grid point, and that the calculations can be performed on any temporal or spatial scale, means there is much more information available from these diagnostics to users and model developers that could be used to understand model errors on a range of timescales (see, for example, Martin et al., 2017). There is also the potential for sub-sampling of rainfall associated with organized systems or phenomena (such as tropical cyclones, fronts, MJO) prior to the histogram analysis, which could increase our understanding of these systems as well as providing information on model errors. The metrics could also be used to examine the influence of model resolution, ocean-atmosphere coupling and the inclusion of Earth-System processes on the spread of rainfall intensities.

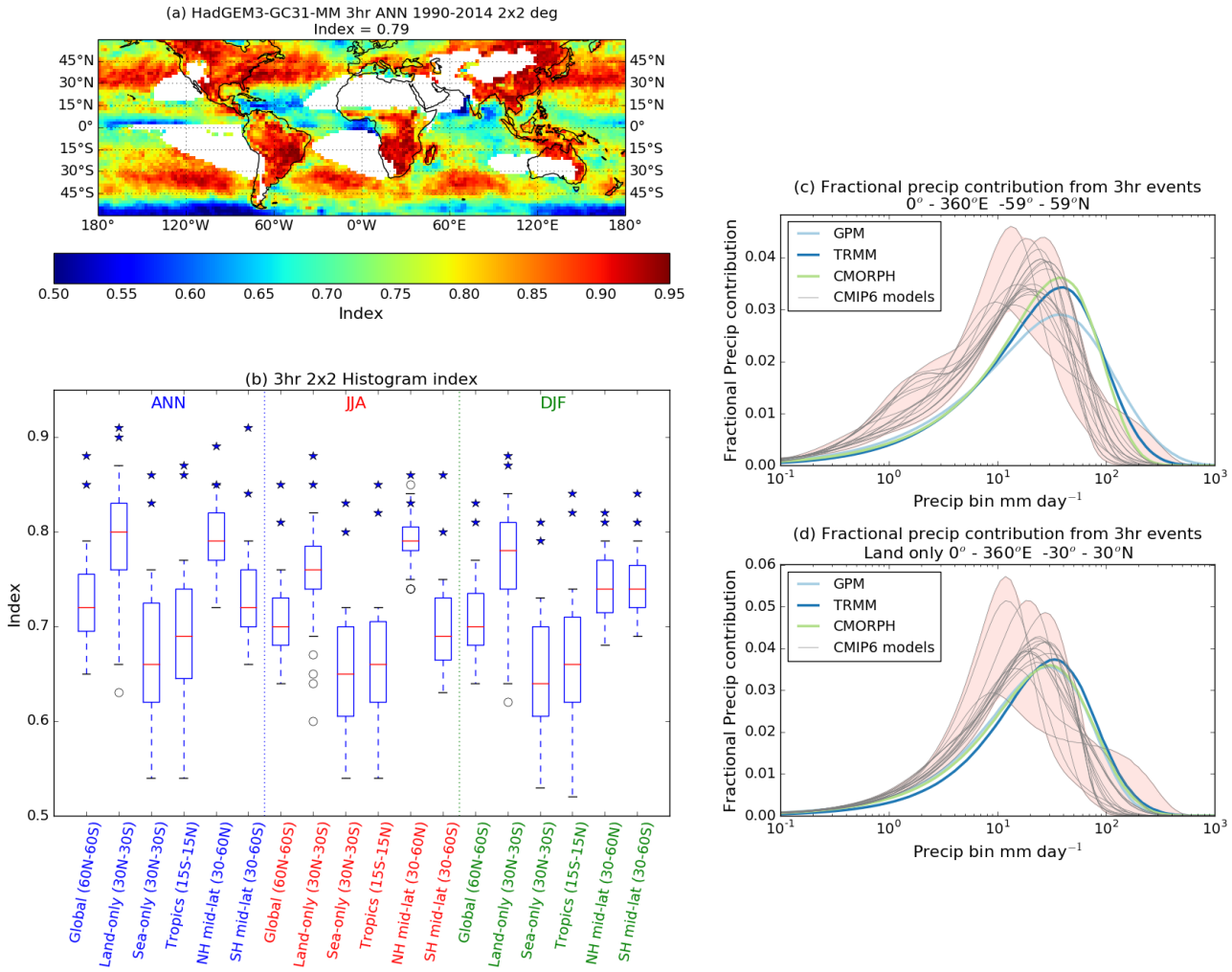

Figure 3. (a) Example map of index from 3hr rainfall data from HadGEM3-GC31-MM vs GPM; (b) Summary metrics for different regions from timeseries of 3hr rainfall data from 23 CMIP6 models compared with GPM-

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

IMERG observations. Boxes show inter-quartile range while whiskers indicate the full range of model indices. Red line shows the median. Filled stars indicate other observational datasets (TRMM and CMORPH). (c) Histograms of 3hr rainfall data from 23 CMIP6 models and 3 observational datasets. (top) Global (60S-60N); (bottom) Land-only (30S-30N). All model and observation data were averaged to a 2° x 2° grid, using conservative area-weighted averaging, before analysis.

Coherence analysis

The "Analyzing Scales of Precipitation" (ASoP) diagnostics (Klingaman et al., 2017) can measure, and compare, the spatial and temporal scales of precipitation across observations and GCMs. The "ASoP Coherence" package was designed to produce a single diagnostic or metric for a chosen region. Here, we extend the package to operate on gridded data. We measure spatial and temporal coherence in 3-hourly and daily precipitation in GPM-IMERG observations and CMIP6 historical simulations. We perform these calculations on a common 2° x 2° grid, a horizontal resolution that is sufficiently coarse for at least some spatial averaging to be carried out for all of the models and the observations while also avoiding removing important spatial detail. This requires us to omit models whose resolution is similar to, or coarser than this. The calculations are performed between 60°S-60°N, neglecting any point with annual-mean rainfall < 1 mm/day and, in the remaining points, any months in the dry season, defined as months that contribute, in the mean, less than $1/24^{th}$ of the annual precipitation.

Figures 4(a-c) use 3-hourly data to show the temporal scale, defined as the first lag at which the temporal auto-correlation is < 0.2, for the CMIP6 historical multi-model mean (Fig. 4a), GPM-IMERG (Fig. 4b), and the multi-model mean bias (Fig. 4c). Throughout much of the tropical and subtropical oceanic regions, the CMIP6 multi-model mean precipitation is too persistent, highlighting an area for model improvement. Figures 4(d-f) use daily-mean data to show the spatial scale, which is computed from the temporal correlation of the precipitation between each gridpoint and its surrounding grid points, using intervals of radii given in the colorbar beneath the panel. The scale is defined as the first search radius at which the spatial correlation is < 0.2. Daily precipitation spatial scales are larger in the CMIP6 multi-model mean (Fig. 4d) than in GPM (Fig. 4e), particularly in the eastern equatorial Pacific and Atlantic ocean, and in near-equatorial regions of the Indian ocean, as well as much of the subtropical oceans (Fig. 4f). Combined with the temporal scale results above, this suggests that CMIP6 models produce precipitation features that are too large and that last too long, particularly in the tropical oceans.

Klingaman et al. (2017) also defines spatial and temporal coherence metrics. The spatial metric is derived from the likelihood of coincidence of upper-quartile and lower-quartile precipitation at neighbouring grid points; the temporal metric is derived from the likelihood of consecutive time steps of upper-quartile and lower-quartile at the same grid point. Quartiles are computed for each gridpoint and each month of the

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

seasonal cycle. For the temporal coherence metric, we show the aggregated grid point metrics (computed 60°S–60°N) as Taylor diagrams for global land (Fig. 4g), ocean (Fig. 4h) and all points (Fig. 4i). The CMIP6 models show higher centered RMS difference and lower correlations, against GPM-IMERG, for land points than for ocean points, indicating that persistence of land precipitation is another area for model improvement. The spatial standard deviation values of the coherence metrics shown in the Taylor diagrams can provide further insights for model improvements: models that have a smaller standard deviation than GPM-IMERG are typically too persistent across all grid points, as the mean bias is positive for nearly all models (not shown), while models that have a greater standard deviation are typically too persistent in some regions and too intermittent in others. These standard deviations show stronger negative biases over land than over ocean, indicating that models show little spatial variability in temporal coherence over land and hence cannot distinguish regions dominated by longer-lived rain-bearing systems from regions dominated by shorter-lived systems.

Next, we demonstrate the ability to compare the spatial scale of precipitation (now restricted to the tropical Indian Ocean - 10°S–10°N; 60°E–90°E; using daily data; determined as correlations at a distance of 800 km) with two metrics of the MJO in CMIP6 models, two satellite observation datasets, and ERA5 (Fig. 4j). The satellite observations and ERA5 have an average precipitation spatial coherence of 0.06-0.09, and the CMIP6 models cover the range −0.03-+0.26. CMIP6 models have a relatively close relationship between precipitation spatial coherence and the MJO Maritime Continent propagation metric (Ahn et al. 2020; R^2=0.489). This suggests that those climate models with a higher spatial coherence of daily precipitation propagate the MJO more robustly east over the Maritime Continent. The relationship is weaker (R^2=0.114) between precipitation spatial coherence and the MJO East/West power ratio, which measures MJO spatio-temporal structure (e.g., Sperber and Kim, 2012; Ahn et al., 2017). There is no relationship between precipitation temporal scale and either MJO metric. Comparisons between spatiotemporal characteristics metrics and process- or phenomena-based metrics may be able to lead to greater insights and understanding of the origins of biases and model errors.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

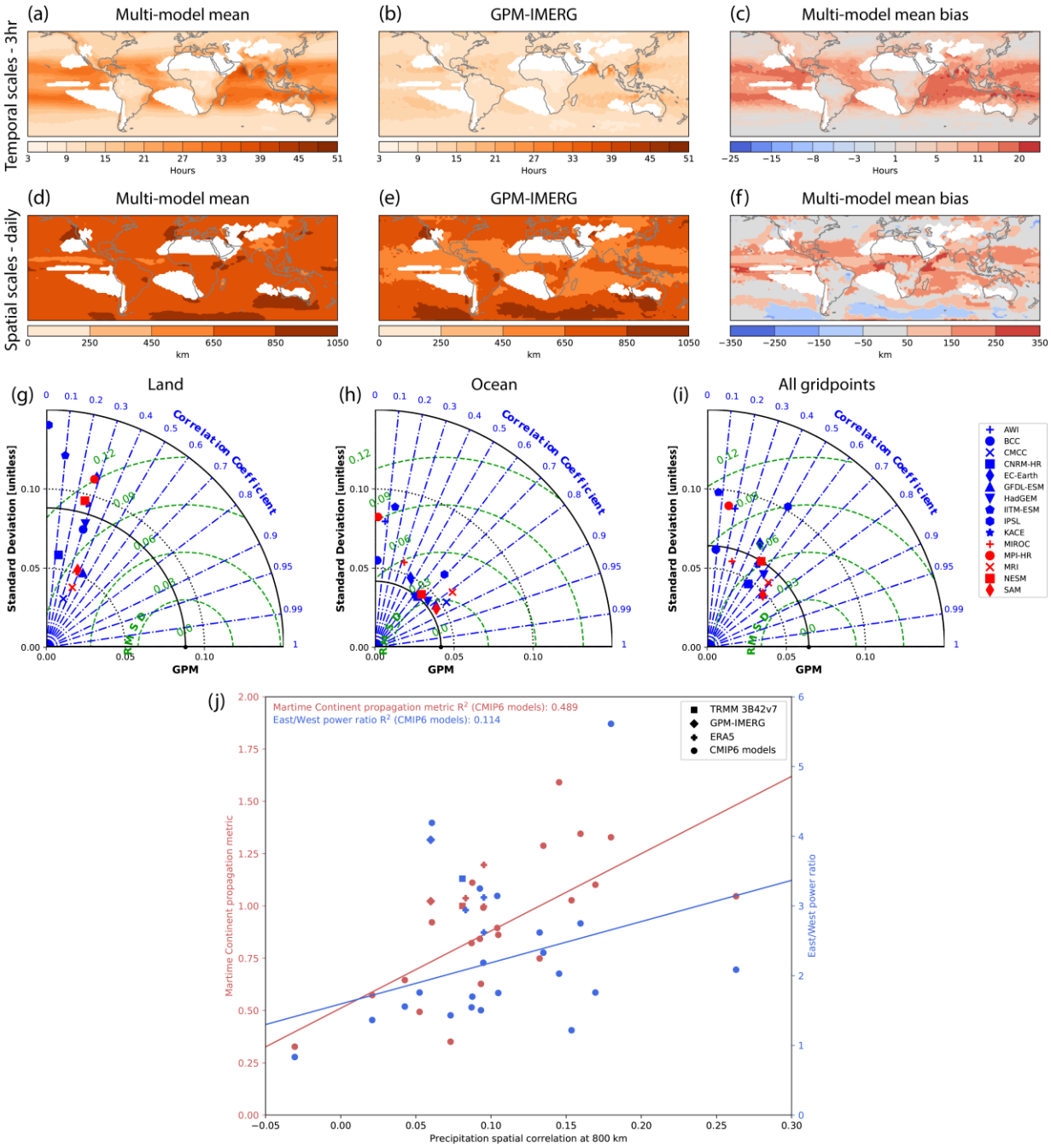

Figure 4. (top row) Temporal scale (hours; defined as the first lag at which the autocorrelation of 3hr precipitation is < 0.2; within 60°S–60°N) a) the CMIP6 historical multi-model mean, b) GPM-IMERG, and c) the multi-model mean bias. (second row) Spatial scale (km; defined as the first distance at which the correlation between a gridpoint and neighbouring points within a distance bin is < 0.2, with bin edges given as divisions of the colorbar) for d) the CMIP6 historical multi-model mean, e) GPM-IMERG, and f) the multi-model mean bias. In (a-f), white shading denotes grid points with annual-mean precipitation < 1 mm/day, which are not included in the analysis. Summary Taylor diagrams (third row) of temporal coherence, using 3-hourly data, over: g) land-only, h) ocean-only, and i) all grid points, within 60°S–60°N, for the CMIP6 models vs GPM-IMERG. Bottom row, j) Tropical Indian Ocean (10°S–10°N; 60°E–90°E) precipitation spatial correlation at 800 km (4× the grid scale) vs MJO Maritime Continent propagation metric (Ahn et al., 2020; left-side axis - red) and MJO East/West power ratio (e.g., Sperber and Kim,

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

2012, Ahn et al., 2017; right-side axis - blue) for two satellite observational datasets (TRMM and GPM-IMERG), ERA5 over three different periods, and the CMIP6 models. Quoted R^2 values and lines of best fit are for CMIP6 models only.

4. Process-oriented metrics

Although metrics of spatiotemporal characteristics are suggestive of the processes contributing to precipitation biases at different spatial and temporal scales, they are not by themselves representing processes related to precipitation. Here, process-oriented metrics are used to reveal relationships between precipitation and the thermodynamic environments, which provide important information on the ability of models to reproduce the observed relationships and the potential contributions of large-scale biases in the atmospheric environments to the precipitation biases. Here, we discuss two metrics highlighting the coupling of precipitation with the thermodynamic environments.

Rainfall-moisture coupling

Latent heating from tropical rainfall formation forces large-scale circulation anomalies that affect weather patterns globally through the tropical-extratropical teleconnection response (Stan et al. 2017). The onset of tropical heavy rainfall is critically dependent upon the relative saturation of the atmosphere (Bretherton et al. 2004; Neelin et al. 2009), while the teleconnection response is sensitive to the spatial and temporal scale of the heating anomaly (Yadav and Straus, 2017; Wang et al. 2020). The MJO is a prominent example of a large-scale tropical disturbance that is strongly governed by column moisture (Adames and Kim 2016) and is also a major driver of tropical-extratropical teleconnections (e.g., Henderson et al. 2017). With thissection, we aim to understand how tropical rainfall and moisture are coupled and how this coupling affects MJO simulation in CMIP6 models.

Following Wolding et al. (2020), daily tendencies of precipitation (P) and column saturation fraction (CSF; i.e., vertically integrated column water vapor divided by vertically integrated saturation column water vapor) over the Indo-Pacific Warm Pool are averaged within conditionally sampled CSF and P bins. All data are first remapped onto a common 2.5°×2.5° degree grid. In Fig. 5a-c, joint CSF-P tendencies are shown with vectors, which indicate if CSF-P departures above or below the mean CSF-P line leads to column moistening or drying. In observations and in most CMIP6 models, the vectors rotate clockwise about the mode (red circles in Fig. 5a-f) that corresponds to the quasi-equilibrium state (Neelin et al. 2008; Wolding et al. 2020). This clockwise rotation indicates that anomalously high precipitation for a given CSF is associated with column moistening, while anomalously low precipitation is associated with

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

column drying. The strength of this rotation in each CSF-P bin can be diagnosed using a vorticity-like metric based on non-dimensionalized CSF and P tendencies where positive values denote clockwise rotation (Fig. 5b). A scalar rotation metric, R, is then computed as the frequency-weighted rotation in CSF-P space.

For models with R>0, positive moistening and rainfall tendencies are largest during the dry-to-moist transition when P is much greater than its mean value for a given CSF (solid red line in Fig. 5a-c). Analysis of radar data collected over the tropical Indian Ocean indicate that this state is associated with a transition from trade wind cumulus to cumulus congestus (Wolding et al. 2020). Negative moistening and rainfall tendencies are largest when CSF is greater than its average value for a given P (red dashed line in Fig. 5a-c), a state associated with widespread stratiform rainfall with embedded convection. For models with R<0, higher-than-average rainfall at intermediate CSF is associated with strong drying; positive P tendencies are only observed at high CSF. Rainfall-moisture coupling in R<0 models suggests that exaggerated depletion of column water vapor by rainfall leads to excessive drying at intermediate CSF, thus reducing the likelihood of subsequent heavy precipitation. Heavy precipitation in these models is only observed at high CSF, where the environment cannot be rapidly dried by rainfall.

Correlations between the R-metric and several MJO propagation "pattern correlation" metrics for a subset of CMIP6 models suggest that tropical rainfall-moisture coupling plays an important role in regulating MJO periodicity. Various MJO pattern correlation metrics have been used to assess MJO propagation in models by correlating simulated and observed rainfall lagged-regressions over the Warm Pool. Jiang et al. (2015) computed pattern correlations of regression coefficient using the composite propagation plotted in Fig. 5h (i.e., the "full" metric). Wang et al. (2017) and DeMott et al. (2019) reduced the influence of MJO period on the pattern correlation by masking coefficients within ±15° longitude of the rainfall basepoint (the "masked" metric), while Ahn et al. (2020) completely removed periodicity effects by only considering positive coefficients in a small portion of the domain east the Maritime Continent (the "MC-crossing metric"). Correlations between the R-metric and the full, masked, and MC-crossing propagation metrics are 0.47, 0.23, and 0.11, respectively. The correlation is only statistically significant for the full pattern correlation metric, which measures the combined effects of MJO propagation and period.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

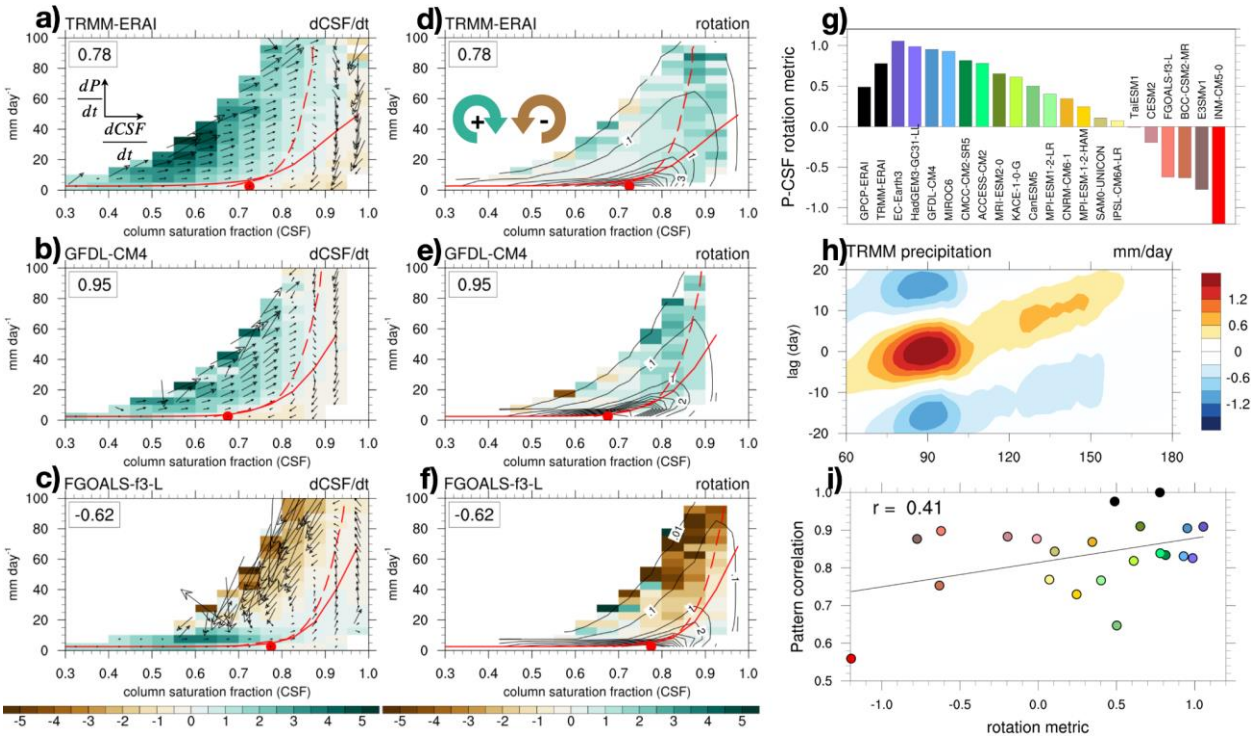

Figure 5. Left column: Daily mean column saturation fraction (CSF) tendency (dCSF/dt in % per day; shading) and the daily mean joint CSF-rainfall rate tendencies (vectors) as a function of CSF rainfall rate (P) for the Indo-Pacific Warm Pool (ocean-only grid points from 20S-20N and 30E-180E) from ERA-Interim reanalysis and TRMM 3B42 (top) the GFDL-CM4 (middle; the median high-performing model) and the FGOALS-f3-L (bottom; the median low-performing model). Red filled circle is mode of observations; red solid and dashed lines are mean rainfall rate and CSF, respectively, for a given CSF or rainfall rate bin. Center column: Non-dimensional rotation (d[dCSF/dt]/dCSF – d[dP/dt]/dP; shading with clockwise rotation shaded green; value shown in upper left of panel) and CSF-P probability distribution function (contours). Right column: frequency-weighted mean rotation for ERAI-TRMM and CMIP6 model, i.e., the "rotation metric" (top); the lagged regression of TRMM 3B42 tropical precipitation (10S-10N averaged) anomalies onto the 20-100 day filtered eastern Indian Ocean area-averaged (5S-5N; 85E-95E) rainfall (middle); scatter plot of individual model convection-moisture rotation metric against the Jiang et al. (2015) MJO propagation metric (see text for details), where colors of dots match bar colors above (bottom). The correlation of the two metrics is r=0.49.

Temperature-water vapor environment

The aim of this module is to create metrics that capture the typical range of moisture and temperature over which precipitation is produced by condensing information from prior diagnostics (Kuo et al. 2018, 2020; which also provides information on sensitivity to sampling and resolution). Here we use a thermodynamic space in which temperature is measured by the vertically integrated saturation humidity, q_{sat}, and moisture is measured by column relative humidity, CRH=CWV/qsat, where CWV is column water vapor, for each q_{sat}. Figure 6a shows, for q_{sat} = 65.5 mm over tropical oceans, the conditional mean precipitation rate (circles) and precipitation contribution (lines) from observations and one model instance. For

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

observations, we use precipitation from the TRMM Precipitation Radar (PR), column water vapor from the TRMM Microwave Imager (TMI), and ERA-5 temperature for computing q_{sat} (for an alternative combination of observations, we use MERRA-2 temperature in Figs. 6c-d). The PR is coarse-grained to 0.25^o x 0.25^o, compatible with the CWV resolution; results are insensitive to resolution up to 1.5° (Kuo et al. 2018). The observed precipitation rate sharply picks up as CRH increases above a certain threshold. The precipitation contribution peaks near this value because the system spends less time at the high precipitation values and the many occurrences of low CRH contribute little to precipitation. The MIROC-E2SL model exhibits qualitatively similar behavior, although the precipitation pickup is too weak and begins at lower CRH than observed, as seen more clearly in the peak of the precipitation contribution. To characterize the moisture range over which precipitation is produced, we identify the CRH values associated with the 25th and 75th percentile of precipitation contribution for each q_{sat}. These CRH values for q_{sat} (tropospheric temperature environment) between the 25th and 75th percentile of q_{sat} (blue lines) are shown in Fig. 6b, together with the precipitation contribution as a function of CRH and q_{sat} (color contours). A notable feature is that the CRH values associated with the 25th and 75th percentile as well as peak of precipitation contribution decrease as q_{sat} increases, i.e., precipitation is produced at lower CRH in a warmer environment.

The values associated with these percentiles provide a good summary of the observed thermodynamic range associated with precipitation, shown by the blue trapezoid in Fig. 6b. We choose a visual reference range (grey box) and repeat it in Figs. 6c-d. Figure 6c presents typical thermodynamic ranges associated with precipitation from a subset of CMIP6 historical simulations and two observational combinations. Deviations of the trapezoids from the observed along the q_{sat} axis indicate cold/warm biases in the simulation, and deviations along the CRH axis indicate that models tend to produce precipitation outside the observed CRH range. Figure 6d exhibits the thermodynamic ranges as in Fig. 6c, but for the 17 available CMIP6 models, ranked by the precipitation contribution error defined as the L^2-difference between the observed and model-simulated precipitation contribution (i.e., the mean square of the dotted area in Fig. 6a), averaged over the four most probable q_{sat} bins. This scalar metric focuses on relative humidity rather than temperature bias. It is encouraging to see that some of the models can produce most of their precipitation in a thermodynamic environment close to the observed range both by the scalar metric and rhomboid location. Other models fare poorly by these measures. Most models capture the decrease in the CRH for the 75th percentile precipitation contribution with increasing temperature, but only about half capture this feature for the 25th percentile.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

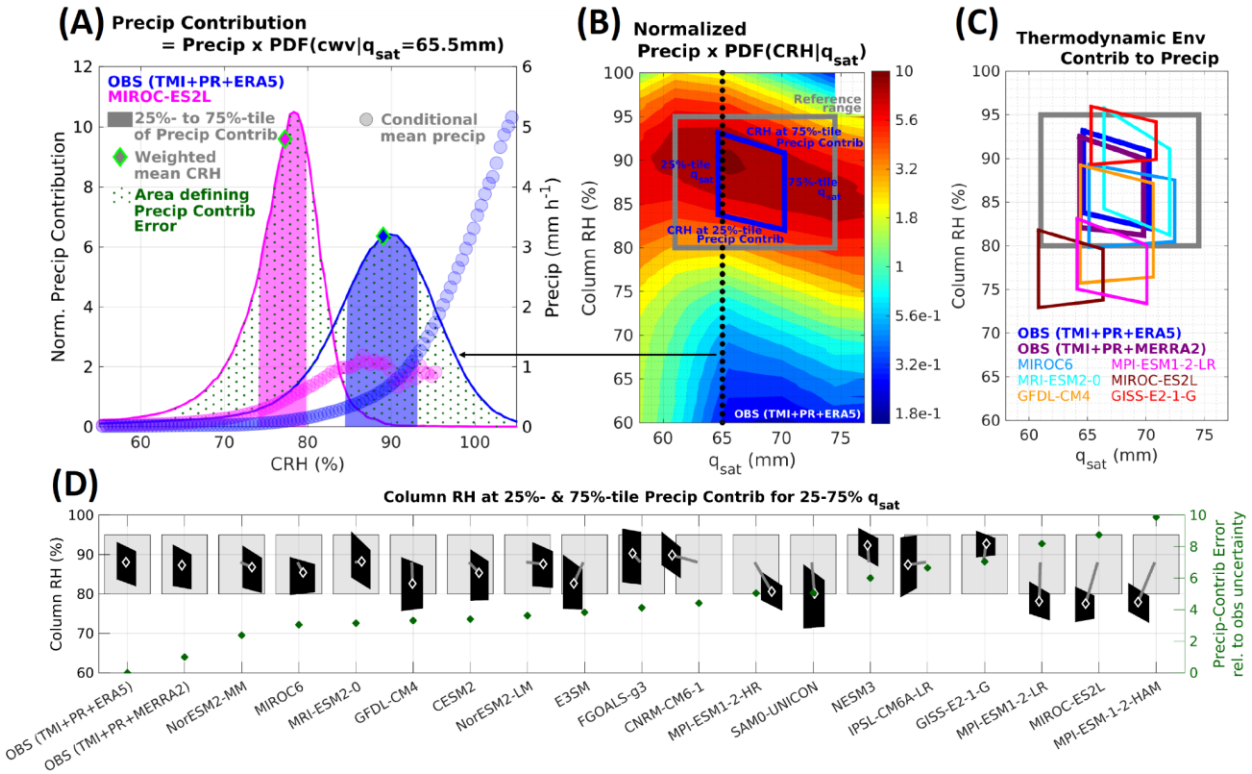

Figure 6. (A) Observed (blue; TMI+PR+ERA5, see text) and an example model (magenta) conditional mean precipitation rate (circles) and precipitation contribution (lines) as a function of column relative humidity (CRH) for column-integrated saturation humidity q_{sat} = 65.5 mm (bin-width 4.5 mm) for tropical oceans within 20°S-20°N. Note that precipitation contributions here are normalized so the area under each curve is one. The 25th to 75th percentile of precipitation contributions are indicated by shaded areas, and the mean CRH weighted by 25%-75% precipitation contribution by diamonds. The precipitation contribution error [in (D)] is defined by considering the L^2-difference between the observed and model-simulated precipitation contributions, i.e., the mean square of the dotted area. (B) Color contours: Precipitation contributions normalized for each q_{sat}. Blue trapezoid: the CRH at the 25th and 75th percentile of the precipitation contribution between the 25th to 75th percentile of qsat. The grey box indicates a visual reference range which remains invariant in (C) and (D). (C) Trapezoids as in (B), but from a set of CMIP6 historical simulations; the observed trapezoid from (B) is repeated (blue) and an additional observational combination (TMI+PR+MERRA2, see text) shown in purple. (D) Black trapezoids and gray boxes as in (C). The white diamonds indicate the 50th percentile q_{sat} and the mean CRH weighted by the precipitation contribution within the rhomboid range. The precipitation-contribution error (dark green) is defined as the L^2-difference between the observed and model-simulated precipitation contribution, averaged over the four most probable q_{sat} bins. Here the difference between the two observational combinations provides a simple measure of observational uncertainty and is used to normalize the precipitation contribution error.

5. Phenomena-based metrics

Phenomena-based metrics emphasize weather features such as synoptic systems and different types of storms that generate precipitation. While synoptic systems such as fronts may be broadly resolved by GCMs at typical 1-degree resolution, storms such as tropical cyclones, LPS, and MCS require higher resolution modeling. Models' ability to simulate these storms is critical as they are key contributors to

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

extreme precipitation in many regions. Feature tracking (briefly summarized in Section 2.2) is used to identify and track the weather features, allowing precipitation associated with these features be isolated and evaluated using different metrics that measure model-observation differences. Here, four examples of weather features and associated precipitation are discussed.

Low pressure systems

A wide variety of synoptic-scale disturbances that consist of balanced flow around a pressure minimum produce precipitation in Earth's tropics and extratropics. Classic examples are midlatitude baroclinic waves, which often produce intense precipitation through semi-geostrophic uplift in their frontal zones, and tropical cyclones, which produce precipitation through the radial, frictionally balanced component of their circulation. Understanding the mechanisms by which such systems amplify and generate precipitation requires tracking the systems from initial genesis; this can be a difficult task, requiring data of sufficiently fine resolution and algorithms of adequate robustness to unambiguously represent the weak and sometimes horizontally small low pressure center. Here we illustrate how a strategic choice of variables allows for improved tracking of low pressure systems (LPS) in the South Asian monsoon, which produce a large fraction of that region's annual mean rainfall as well as many extreme precipitation events. This tracking exercise allows the relationship of circulation with precipitation to be characterized in observations and model ensembles.

Tropical LPS are most commonly identified and tracked using lower-tropospheric vorticity or sea level pressure. Even for strong tropical cyclones, ambiguities in the criteria used in the tracking algorithm can lead to large uncertainties in the number of storms identified in observationally constrained gridded data (e.g., Murakami 2014). This issue is even more problematic for weak LPS, where the noisiness of the vorticity field produces irregular, broken tracks for systems that seem to move smoothly when tracked subjectively using a standard suite of meteorological data (Fig. 7a). Sea level pressure, which is less noisy, is sometimes used to track LPS instead but is ill suited for South Asian LPS, which typically have winds that peak around 3 km above the surface; geopotential height near the level of maximum wind also does not capture the full rotational flow given the low latitude and high Rossby number of these storms. Physical reasoning, as well as systematic assessment of multiple candidate variables with hundreds of combinations of quantitative tracking criteria, showed that the streamfunction of the horizontal 850 hPa wind is an optimal variable to use for tracking these LPS (Fig. 7b, and Vishnu et al. 2020). This streamfunction represents the full non-divergent wind, even when geostrophic balance does not hold, yet

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

retains the smoothness of the geopotential or sea level pressure fields; it was inverted using a method to avoid contamination by any wind data extrapolated below Earth's surface (Vishnu et al. 2020).

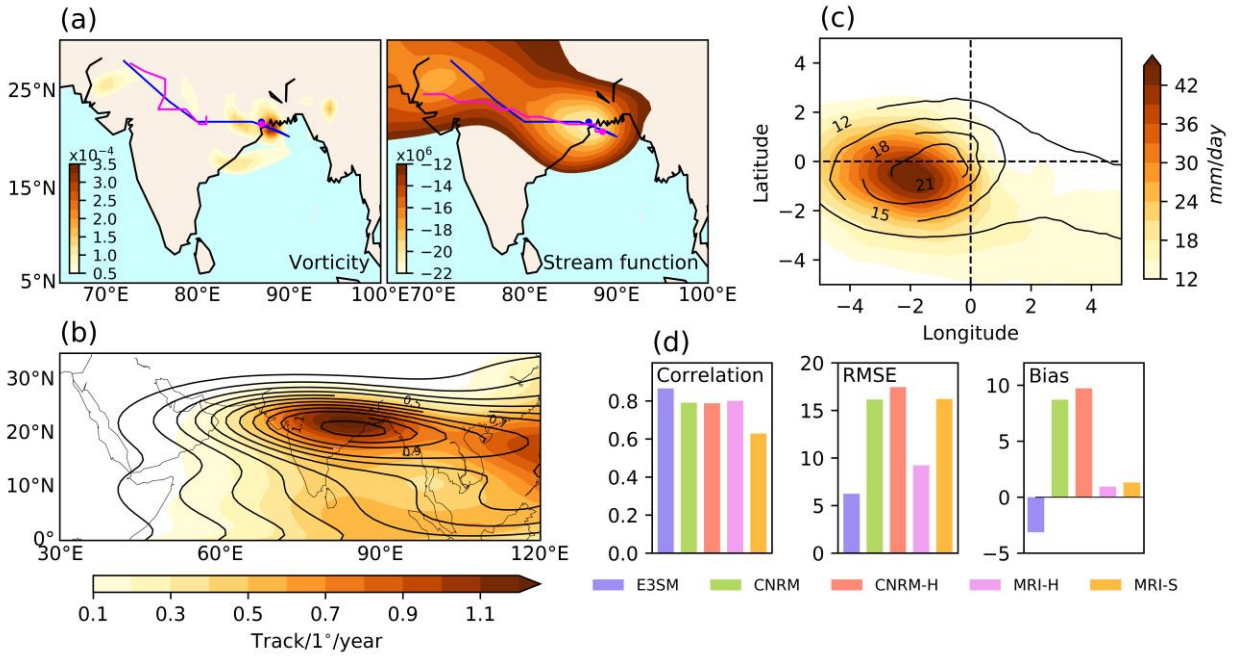

Figure 7. (a) Example of the influence of variable choice on tracking skill: compared to the 850 hPa relative vorticity, which is commonly used to track tropical disturbances, a more continuous track that better matches the subjectively analyzed reference track is obtained using the streamfunction of the 850 hPa horizontal wind (magenta lines show tracks obtained from an automated algorithm applied to ERA5, while blue lines show the reference track). (b, c) Comparison of model (black contours, for E3SM) and observed (shading) representations of climatological mean track density (b) and vortex-centered composite rain rate (c) for South Asian monsoon low pressure systems. E3SM simulates a reasonable track density but produces disturbances that rain too little with peak rainfall biased slightly toward the vortex center. Observed tracks are from ERA5; observed precipitation from TRMM. (d) Metrics showing skill of E3SM and four HighResMIP models in simulating the spatial structure of rainfall in South Asian low pressure systems. For vortex-centered composite rain rates (as in (c)), we show the correlation coefficient, root-mean-squared error (in mm/day), and horizontal mean bias (in mm/day, averaged over a 10^{o}x10^{o} box around the composite vortex center) compared to TRMM. Note that E3SM has the highest correlation and the lowest RMSE but a larger magnitude bias in horizontal-mean precipitation than the MRI models; the MRI model skill degrades at finer resolution (MRI-S is finer resolution than MRI-H), while CNRM model skill has little sensitivity to resolution.

Precipitation in South Asian monsoon LPS is known to fall southwest of the storm center, where the interaction of the storm's rotational flow with the background vertical shear produces quasi-geostrophic uplift (Rao and Rajamani 1970, Sanders 1984). This placement of peak precipitation is well-captured when compositing TRMM precipitation relative to ERA5 LPS tracks (Fig. 7c). The ERA5 reanalysis also accurately represents the well-known distribution of track density, with storm frequency peaking strongly over the northwest Bay of Bengal (Fig. 7b). Recent work has shown that LPS frequency likely peaks in that small region because the large-scale, low-level monsoon winds are barotropically unstable there

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

(Diaz and Boos 2019) and vapor pressures are large with strong horizontal gradients (Ditchek et al. 2016, Adames and Ming 2018). Wind-enhanced evaporation from the Bay of Bengal may also enhance LPS intensity there (Murthy and Boos 2020, Fujinami et al. 2020, Diaz and Boos 2021).

By tracking LPS in ensembles of GCMs, we can create composites that allow model precipitation bias to be assessed in a phenomenon-based system rather than in a space- or time-based system that averages many types of atmospheric disturbances. One high-resolution GCM (E3SM integrated at 0.25-degree resolution) represents the track density of South Asian monsoon LPS well, in addition to the spatial structure of precipitation relative to the vortex center (Fig. 7b, c). This is notable given the poor ability of some coarse-resolution GCMs to simulate these LPS (Praveen et al. 2015). However, the E3SM model simulates monsoon LPS rainfall that is too weak, with the peak storm-centered composite precipitation being about half that observed (Fig. 7c). Other models exhibit a variety of biases in their representation of monsoon LPS precipitation with differing sensitivities to model resolution. Storm-centered composites in the CNRM models have overly strong precipitation with little sensitivity to model resolution, while the MRI models produce roughly the right amount of precipitation over the entire storm but with a spatial pattern that, unexpectedly, degrades at finer model resolution (Fig 7d). These biases are large for some models, exceeding 50% of the system-averaged TRMM rain rate of 15 mm day^{-1}; interannual variations in LPS activity and storm-centric rain rates is substantially more modest (e.g. Sikka 1980, Krishnamurthy and Ajayamohan 2010, Vishnu et al. 2020).

Such assessment of model skill in representing the synoptic systems that produce extreme rainfall, such as monsoon LPS, is an important step in producing reliable projections of future extreme rainfall. The LPS dataset used here, which is available for 5 modern reanalysis products, provides LPS tracks throughout the global tropics that can be used to better understand a variety of synoptic-scale phenomena, including the weak progenitors of tropical cyclones.

Mesoscale convective systems

Mesoscale convective systems (MCSs) are ubiquitous over the tropics year-round and in the mid-latitudes during the warm season. Besides contributing to over 50% of the annual precipitation in most regions of the tropics and selected regions in the midlatitudes (Nesbitt et al. 2006; Feng et al. 2021b), MCSs are also key contributors to extreme precipitation, partly because of their larger size and longer lifetime compared to individual convective storms (Stevenson and Schumacher 2014). Because of the distinctive nocturnal timing of MCS, erroneous diurnal timing of summer precipitation produced by models has been used to

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

infer their failure in simulating MCSs. Recent efforts in developing algorithms to identify and track MCSs in observations (Feng et al., 2018) and model simulations (Feng et al. 2021a) have provided unprecedented opportunities to directly evaluate MCSs and their characteristics in weather and climate models using MCS-specific metrics.

Using FLEXTRKR, an algorithm developed to track MCSs using both infrared brightness temperature (T_b) and precipitation feature (PF) (Feng et al. 2018; 2019), a global (60°S-60°N) MCS tracking database has been developed at ~10 km and hourly resolution (Feng et al. 2021b). Combining the track locations and precipitation, this database can be used to derive information of the MCS number, MCS precipitation and its fractional contribution to the total precipitation, MCS maximum precipitation rate, MCS lifetime, and MCS translation speed and direction. As MCSs are not well defined at coarser spatial resolution, we develop MCS metrics mainly for use in evaluating high-resolution weather and climate simulations with grid spacing < 50 km. Instead of coarse graining the observations and model outputs, which correspond to a range of grid spacing, to a common resolution, we use specific PF criteria derived for a given resolution for MCS tracking to facilitate comparison across datasets of different resolutions (Feng et al., 2021).

Figs. 8a&b compare the MCS number tracked using two algorithms, a more commonly used method that tracks MCSs using T_b only vs. FLEXTRKR that tracks MCSs using both T_b and PF. These two methods produce similar observed total MCS number and spatial distribution in the tropics, but larger differences are noticeable in the mid-latitudes. Including PF in MCS tracking noticeably reduces the number of MCSs in the mid-latitudes by disqualifying large cold cloud systems (e.g., synoptically forced) with small area and/or low rainfall intensity PF as MCSs. Using only IR T_b, the model (E3SM) simulates too many MCSs (blue contours) except in a few locations. In contrast, using both IR T_b and PF, E3SM simulates two few MCSs (magenta contours) except in a few locations. These results show that large cold cloud systems are produced by the model too frequently but many of them fail to meet the PF thresholds. This is supported by the composited MCS rain rates shown in Fig. 8c for northeast moving MCSs in the central U.S. during spring (MAM) and summer (JJA). The simulated and observed rain rate composites have similar size, but the model produces much lower peak rain rates. A higher fraction (65%) of MCSs in the model have a northeast propagation than observed (44%).

Fig. 8d summarizes the MCS precipitation metrics for four models in HighResMIP. The pattern correlation, root-mean-square error (RMSE), and bias are calculated based on comparison of the observed and simulated composited MCS rain-rates over the central U.S. Since hourly T_b is not available from the HighResMIP models except E3SM, MCSs are tracked using an algorithm that depends only on PF,

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

trained using MCSs tracked using both T_b and PF (Feng et al. 2016). Note that E3SM is a free-running fully coupled simulation with constant 1950 forcing while other simulations are atmosphere-only simulations driven by observed sea surface temperature and sea ice distribution. The models exhibit a range of biases from larger negative (E3SM) to larger positive (NICAM) and the skills are generally lower during summer than spring. The seasonal difference is particularly large for NICAM. Unlike the other models that parameterized deep convection, no deep convection scheme was used in NICAM at 56 km grid spacing. Lastly, it is worth noting that metrics based on composited MCS precipitation can only reveal differences in PF qualified as MCS. All models evaluated here display significant dry bias in the summer, consistent with the ubiquitous warm-dry bias noted in CMIP5 (Lin et al. 2017), as the models simulate much lower numbers of MCSs compared to observations. Therefore, we emphasize the importance of using multiple metrics for comprehensive evaluation of precipitation in models.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

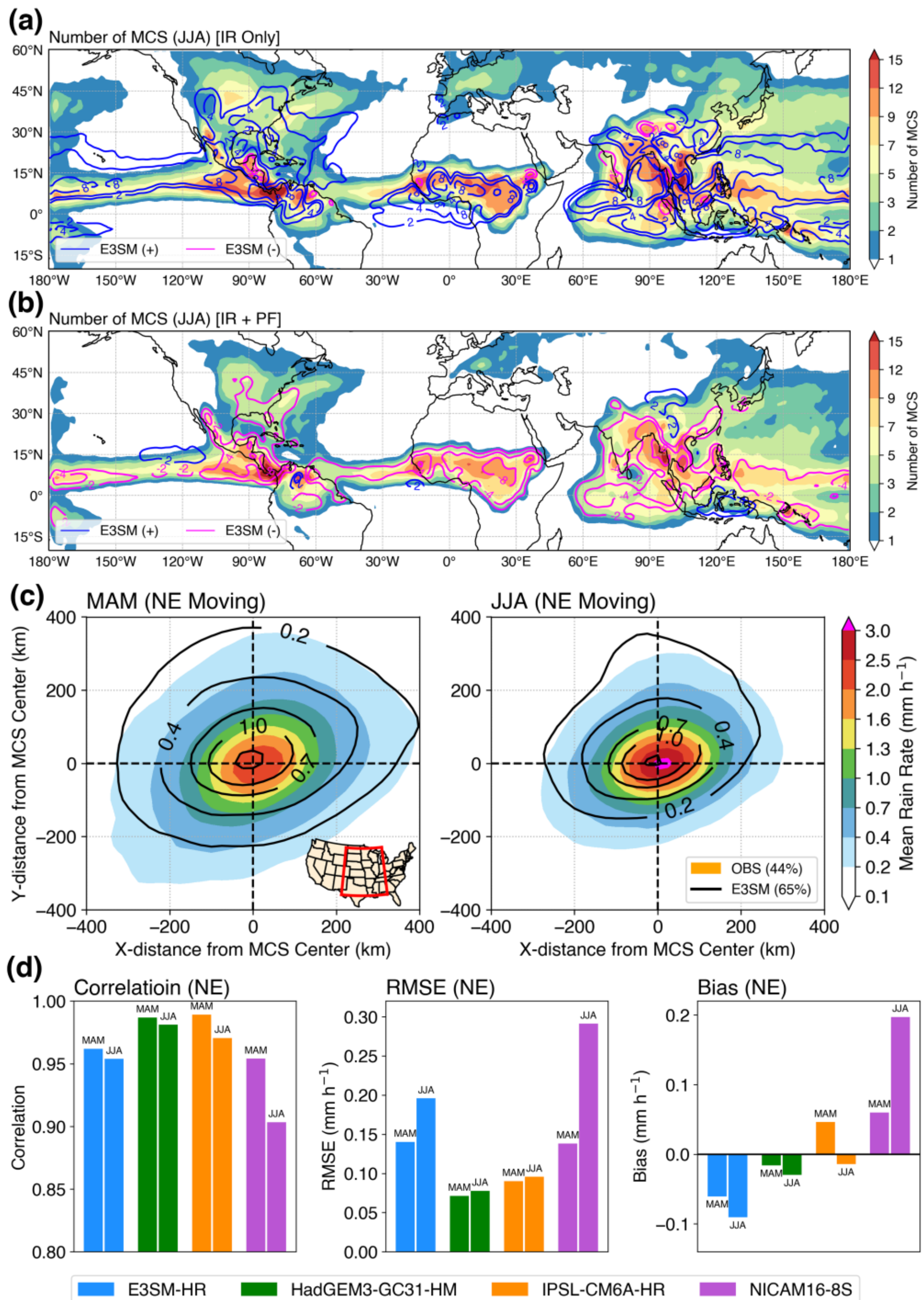

Figure 8. Influence of variable choice on MCS tracking. MCSs are tracked using only infrared brightness temperature (IR T_b) (a) and using both IR T_b and precipitation feature (PF) (b). The observed number of MCS is shown in color shading and the model bias is shown in color contours (blue/magenta for positive/negative bias). (c) Comparison of simulated (black contours, for E3SM) and observed (colored shading) MCS rain rates (mm h^{-1}) composited with a center co-located with the geometric centroid of the MCS PF. Composites are shown for spring (left, March-April-May) and summer (right, June-July-August) for northeast moving MCSs inside the central U.S. (red region) shown in the left panel inset. Trained on the MCS statistics tracked using both IR T_b and PF, the MCSs used in these composites are tracked using only PF to facilitate comparison with other models for which hourly precipitation but not hourly outgoing longwave radiation is available. (d) Metrics showing skill of E3SM and three

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

HighResMIP models in simulating the spatial structure of MCS rainfall in the central U.S. Based on the rain rate composites (as shown in (c) for E3SM), three metrics - correlation coefficient, root-mean-square error (mm h^{-1}), and mean bias (mm h^{-1}) - are used to evaluate different aspects of the model MCS rainfall.

Frontal precipitation

Fronts have been identified using the method described in Section 2, applied to ERA-Interim and five CMIP6 models, giving gridded front objects on a 2.5° grid. The fronts are linked to daily precipitation, using GPCP 1dd as an observational precipitation estimate. The precipitation data are regridded to the same resolution as the fronts in order to make the linking simpler. We consider precipitation only if it is above a threshold of 1mm, which is the minimum 24-hour precipitation a gauge can measure, and this eliminates some of the "drizzle problem" that models tend to have (Stephens et al. 2010). The precipitation is associated with a front if it lies within the front area of influence (which is equivalent to being in the same grid box or the surrounding eight grid boxes) during any of the four 6-hourly reanalysis times in the 24-hour precipitation period. From this association of fronts and precipitation, we can produce the diagnostics of frontal (and non-frontal) precipitation frequency (F_f, F_{nf}), frontal (and non-frontal) precipitation intensity (I_f, I_{nf}), frontal amplification factor ($A_f = I_f/I_{nf}$), and fraction of total precipitation from fronts (P_f) (See Catto et al., 2013, 2015 for full details). Comparing the model diagnostics to the observational estimates from ERA-Interim and GPCP, we can produce a number of metrics, including the correlation, RMSE and bias of these values.

Since precipitation biases (E_p) in the models depend on the frequency of fronts, the frequency of precipitation, and the intensity of the precipitation, we can also decompose the bias of each model into components associated with these characteristics as follows:

$$E_p = \Delta F_f I_{f,o} + F_{f,o}\Delta I_f + \Delta F_f \Delta I_f + \Delta F_n f I_{nf,o} + F_{nf,o}\Delta I_{nf} + \Delta F_{nf}\Delta I_{nf}$$

where subscript *o* represents the observational estimate, and Δ represents the difference between model and observational estimate. The cross terms (3 and 6) are generally very small and are not shown.

Maps (Fig. 9a) of the error decomposition for term 1 (contribution from frequency of frontal precipitation) show that there are large regions of positive bias contribution. Errors are largely confined to the regions of maximum storm track activity and in the NH the largest positive bias contributions can be seen over the Kuroshio Current, over western Europe and parts of the North Atlantic, and at the end of the Pacific storm track into North America. In the SH the largest positive contributions are in a band between 30 and 40° South, particularly around the south coast of Australia. Term 2 errors (contribution from

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

intensity of frontal precipitation) are generally largest in the same regions and indicate negative contributions to the total bias, with this being particularly notable over the North Atlantic region. The maps indicate a compensation of biases between terms 1 and 2, which is confirmed for each of the models in Fig. 9b, and consistent with the CMIP5 models (Catto et al 2015). In the midlatitudes the contribution to the total precipitation error from the non-frontal precipitation terms is small (Fig. 9b), as expected due to the high frequency of fronts.

The models all overestimate A_f due to larger negative biases in the non-frontal precipitation intensity than the negative biases in frontal precipitation intensity (not shown). These biases are large compared to the GPCP A_f of 1.28 in NH DJF and 1.35 in SH JJA and are strongly correlated with the model biases in the intensity of the frontal precipitation (not shown). The spatial correlation is between 0.4 and 0.6 in the NH and between 0.3 and 0.4 in the SH, indicating a better representation in the NH.

The proportion of total precipitation associated with fronts in the winter seasons is 0.50 in the NH and 0.54 in the SH for GPCP and ERA-Interim. The biases in this quantity range between 0.02 and 0.27 (Fig. 9d), with most models showing a better representation in the SH. The models that perform better for the proportion do not necessarily show better performance in the A_f metric, indicating the utility of looking at more than one metric.

Analyzing the ranks of the models using the various calculated metrics, we can see that some models that perform well in metrics that quantify magnitude differences (e.g. the decomposition terms and biases), also perform poorly in their spatial correlation, (e.g. IPSL-CM6A-LR). Again, this points to the importance of considering a number of different metrics to investigate the model performance.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

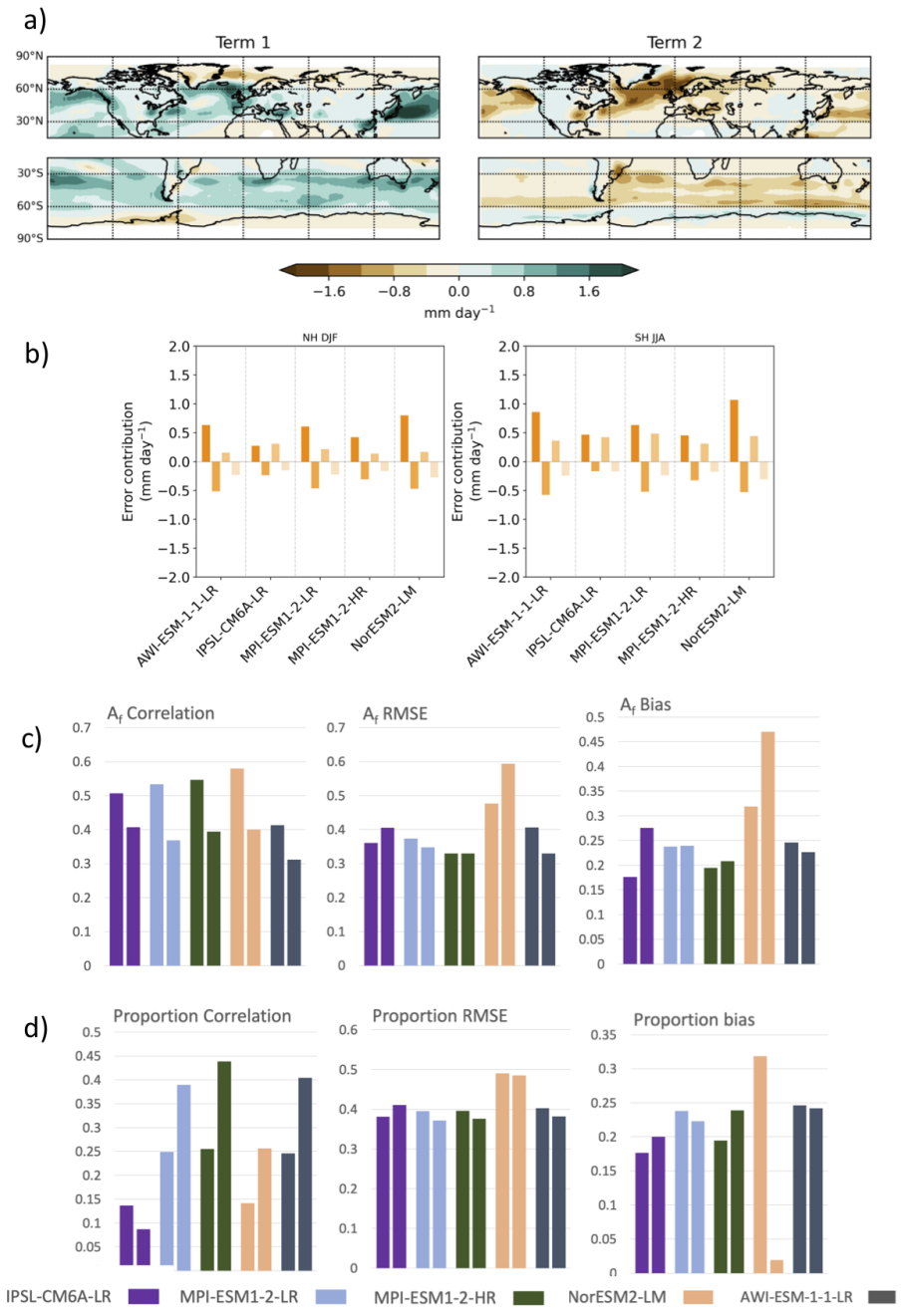

Figure 9. Representation of frontal precipitation in 5 CMIP6 models (1980-2014) compared to ERAI fronts with GPCP daily precipitation (1997-2017) for winter (DJF in the NH and JJA in the SH). (a) Multi-model mean of the first and second terms of the decomposition in mm/day. (b) Area averaged decomposition terms (term 1, 2, 4, 5) for each of the models in the NH and SH extratropics (15-90). (c) The correlation, RMSE and bias for the frontal amplification factor Af=Pf/Pnf. For each model the left bar is the NH extratropics in DJF and the right bar is the SH extratropics in JJA. (d) The correlation, RMSE, and bias for the proportion of precipitation associated with fronts.

Atmospheric rivers

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

Atmospheric rivers (ARs) are long narrow bands of poleward vapor transport often associated with the warm sector in advance of midlatitude cyclone cold fronts (Ralph et al 2018). They account for a large fraction of wet-season precipitation in a number of regions (Dettinger et al 2011; Rutz et al 2014; Guan and Waliser 2015), and they account for a majority of the poleward moisture transport (Gimeno et al 2014). Previous studies examining ARs in climate model simulations have assessed the ability of models to adequately simulate relevant characteristics of ARs, including: global and landfalling frequency, intensity, precipitation, duration, lifecycle, etc. (Dettinger 2011; Payne and Magnusdottir 2015; Shields and Kiehl 2016; Goldenson et al 2018). In this module, we present two metrics aimed at answering the questions: (1) Do models simulate AR-related precipitation in the correct locations? (2) Do models simulate enough contrast between regions with high AR precipitation and low AR precipitation? (3) Does the diversity of AR detection and tracking (ARDTs) affect the above conclusions?

We utilize output from six global ARDTs that participated in the ARTMIP Tier 1 experiment and Tier 2 CMIP5/6 experiment (see Section 2.2); these ARDTs identified ARs in the MERRA-2 reanalysis and in historical simulations from nine members of the CMIP5 and CMIP6 multi-model ensembles. We quantitatively define 'AR-precipitation', for each ARDT, as precipitation occurring when AR conditions are identified by a given ARDT. We calculate AR-precipitation for MERRA-2 (using the precipitation field from MERRA-2) and for the CMIP5 and CMIP6 simulations. We calculate 30-year averages of these quantities and regrid all to a common 2°x2° grid to facilitate direct comparison of the fields between the simulations and the reanalysis. Additionally, we calculate AR-precipitation for the ERA 20C reanalysis (1900-2010) to provide a combined estimate of observational uncertainty and natural variability (since we use a different time period than with MERRA-2). Figures 10a,b show the bias in AR-precipitation between two CMIP6 models, with one model's bias field indicating some regional biases in AR-precipitation (Fig. 10a) and another model's bias field indicating systematically too little AR-precipitation (Fig. 10b).

The spatial correlation coefficient of AR-precipitation between each model simulation and MERRA-2 is used to answer question (1) above, and the ratio of the spatial standard deviation of AR-precipitation between each model and MERRA-2 is used to assess question (2). These quantities are calculated for all available model-ARDT pairs in order to assess question (3). Figure 10c shows a Taylor diagram constructed by plotting the spatial correlations on the azimuthal axis and the ratio of the standard deviations on the radial axis.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

It appears that models generally produce AR-precipitation in the correct regions, but they do not have enough spatial variability in AR-precipitation. The models have relatively high spatial correlation coefficients--regardless of which ARDT is used--with most models having coefficients between 0.8 and 0.95. It is notable, however, that the value of the spatial correlation coefficient can depend strongly on which ARDT is used. Consider results from the CMIP5 CCSM4 simulation (navy blue markers), which range from about 0.7 when evaluated with the GuanWaliser v2 ARDT to over 0.9 with the ARCONNECT v2, Lora v2, and TECA BARD v1.0 ARDTs. In contrast to the spatial correlation, all models have less variability than the MERRA-2 simulation, and models exhibit a wide range of skill in this metric.

Across the ARDTs used, some models form distinct clusters in the Taylor diagram, with the CMIP6 MRI-ESM2-0 and CMIP5 CCSM4 simulations having systematically low Taylor skill values and the CMIP5 CanESM2 simulation having systematically high Taylor skill values. These distinct clusters indicate consensus among the ARDTs about the model skill. In contrast, some models span the Taylor diagram; e.g., the skill of the CMIP5 IPSL-CM5A-LR simulation depends strongly on which ARDT is used, with the TECA-BARD v1.0 giving a Taylor skill score of approximately 0.87 and ARCONNECT v2 giving a skill score of only about 0.32. Comparing between generations, the CMIP6-CM6A-LR simulation has systematically higher Taylor skill scores than either of the CMIP5 IPSL simulations. Further, the CMIP6-CM6A-LR simulation forms a distinct cluster in the Taylor diagram, suggesting a consensus among ARDTs that the CMIP6 version of the IPSL model is superior to the CMIP5 versions.

The ARDTs exhibit distinctive differences in model evaluation. Metrics calculated with the TECA-BARD v1.0 ARDT (star markers in Fig. 10c) are systematically higher than any other ARDT, and most models evaluated by TECA-BARD v1.0 appear skillful at simulating AR-precipitation. The notable exceptions are the CMIP6 MRI-ESM2-0 and CMIP5 CCSM4 simulations, which--as noted previously--have low metric scores no matter which ARDT is used, which is due to a systematic low bias in AR-precipitation in the simulations (e.g., Fig. 10b). Other ARDTs, such as ARCONNECT v2, have a wide spread in the AR-precipitation metrics.

These differences among ARDTs are partly related to their designs. ARCONNECT v2 utilizes an absolute threshold in IVT when identifying ARs, which would make the ARDT much more sensitive to biases in model humidity and/or winds. If a simulation has a systematic low bias in IVT, for example, then the ARCONNECT v2 ARDT will detect systematically fewer ARs in that simulation. Other ARDTs,

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

such as Lora v2 and TECA-BARD v1.0, utilize relative IVT thresholds, which may be less sensitive to model bias.

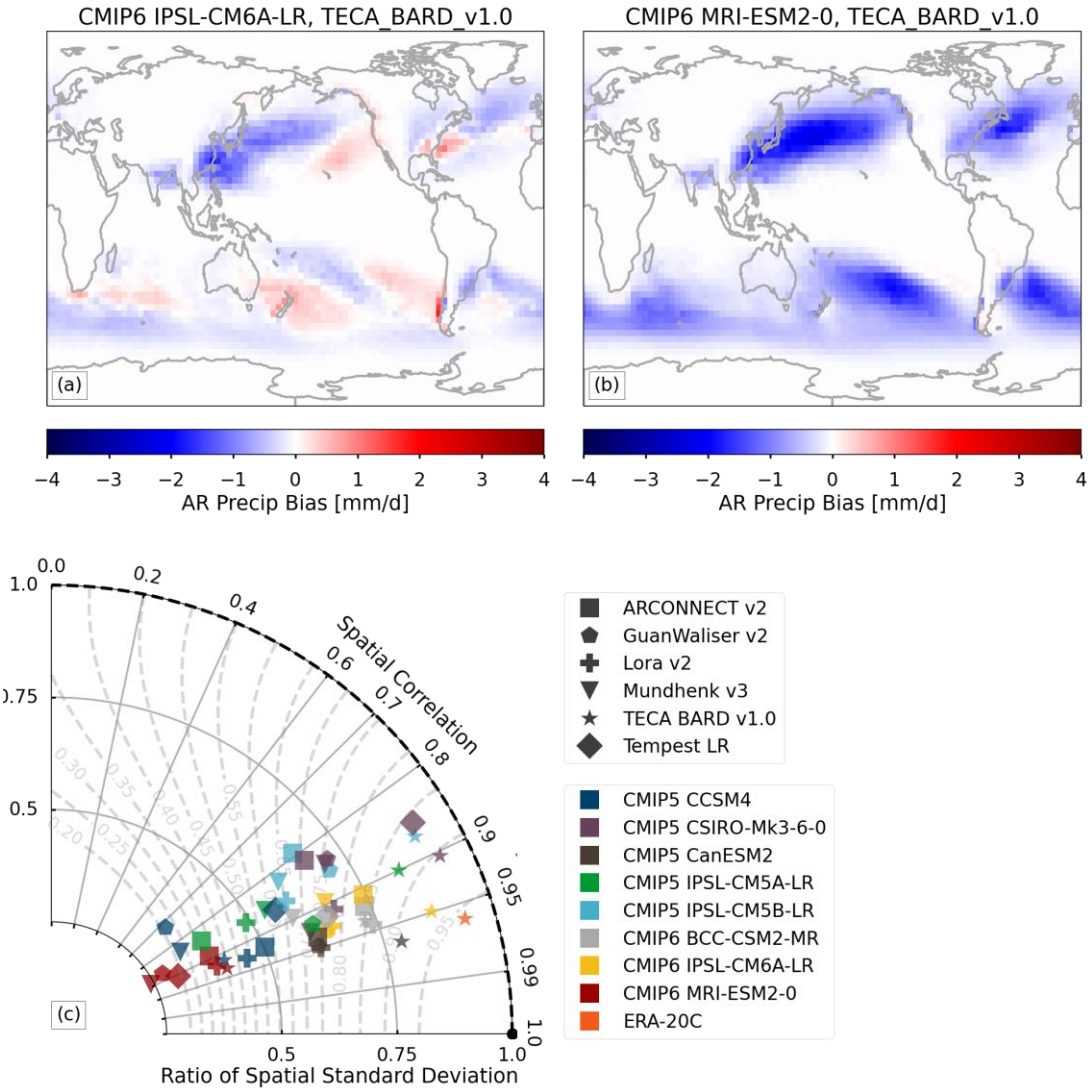

Figure 10. AR precipitation metrics considering AR detection diversity. (a, b) Bias in mean annual precipitation (mm/d) associated with ARs detected using the TECA BARD v1.0 ARDT (a) the CMIP6 historical simulation from the IPSL-CM6A-LR model (1950-1986) and MERRA-2 (1980-2016); and (b) the CMIP6 historical simulation from the MRI-ESM2-0 model (1950-1986) and MERRA-2 (1980-2016). (c) A Taylor diagram comparing the spatial correlations and spatial standard deviations of AR-precipitation between simulations and MERRA-2, using multiple ARDTs. Colors are associated with models, and markers are associated with ARDTs. The dashed gray curves in (e) show contours of constant Taylor skill metric.

6. Discussion and summary

With a primary goal of introducing a suite of exploratory precipitation metrics and demonstrating their use in evaluating precipitation in climate models, we minimized the hurdle by allowing different groups

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

to apply their diagnostics and metrics to readily available model outputs using their preferred or readily available benchmark datasets. Although most of the metrics were applied to CMIP6 simulations including HighResMIP, the number of models evaluated ranges between 4 and 35. Because feature tracking generally requires more variables and higher temporal frequency data, the LPS, MCS, FRT, and AR metrics were demonstrated using only 4-9 simulations. Although all other metrics were applied to a much larger number of CMIP6 simulations (17-35), differences in the specific simulations used and whether a single or multiple members of a model family were used make comparison across models and metrics difficult.

Despite the difficulty in drawing broad conclusions, some general observations can be made for each metric and by comparing across models and metrics. For precipitation diurnal cycle, models generally perform much better over ocean than over land, as models have a tendency to produce peak precipitation in the afternoon over land while the observed peak precipitation occurs in the late afternoon/early evening. There is a relatively strong negative inter-model correlation between biases in the diurnal amplitude and phase over ocean but such correlation is positive and weaker over land. Almost all the examined models fail to capture the nocturnal peak observed at the ARM SGP site. For precipitation and dry spells, models perform well in simulating the spatial pattern of both daily precipitation and duration of dry-spells cutoff-scales, which means that models would also do well in simulating the spatial distribution of extremes. However, there is a larger spread in terms of scaling factor (i.e., the overall magnitude of the patterns), with the daily precipitation cutoff-scale closer to observations than the dry spell duration cutoff-scale. Pattern correlation and scaling factor are largely independent metrics as their inter-model correlations are relatively low. In contrast with the precipitation diurnal cycle, spectral analysis shows that models perform better over land than ocean (between 30S-30N) and better over the NH mid-latitudes (30-60N) than the Tropics (15S-15N). The majority of the models analyzed have their spectra overlapping with observations by more than 60% in all of the regions and seasons, but the metrics from the models nearly all lie outside the spread of the observation datasets used. The temporal and spatial coherence analysis highlights that the CMIP6 models generally produce precipitation features that are too large and that last too long, particularly in the tropical oceans. Despite these general tendencies, models have a wide range of abilities, with some producing good spatial and temporal variability while others perform poorly at both. There are stronger negative biases over land than over ocean, indicating that models show little spatial variability in temporal coherence over land and hence cannot distinguish regions dominated by longer-lived rain-bearing systems from regions dominated by shorter-lived systems. In the tropical Indian Ocean there are some relationships between the precipitation coherence and MJO metrics (Maritime Continent propagation).

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

For the process-oriented metrics, coupling of rainfall tendencies and CSF tendencies over the Indo-Pacific Warm Pool (Figure 5) is well-simulated in five of the 20 models analyzed for that metric, and poorly simulated in eight models; the remaining models with neutral skill may either overestimate or underestimate the rainfall-moisture "rotation" metric derived from this diagnostic. While the rotation metric is modestly correlated with the MJO pattern correlation metric (r=0.41), several models may perform well in one metric, but poorly in another, indicating that rainfall-moisture coupling alone is not a good predictor of a model's ability to simulate the MJO. For the temperature-water vapor environment, almost half of the models produce most of their precipitation over tropical oceans in a temperature-moisture environment that is reasonably close to the observed range (using twice the distance between the two observational estimates as the reference range). This reflects that the deep-convective parameterizations in these models have included a substantial dependence of convective updrafts on lower free-tropospheric humidity (Kuo et al. 2017). Such a precipitation-temperature-water vapor relationship, however, is not perfectly aligned with other metrics related to precipitation and atmospheric moisture, as will be discussed further below.

In the category of phenomena-based metrics, all HighResMIP models examined here simulated synoptic-scale vortices (i.e., LPS) over South Asia with the qualitatively correct spatial structure of rainfall, with no improvement in model skill at finer horizontal resolution in the two models for which low- and high-resolution versions were examined. This contrasts with prior studies that found LPS were simulated more accurately at finer resolutions; that different result may be due to use of a range of coarser resolutions than examined here (Praveen et al. 2015) or the use of only one model (Sabin et al. 2013). In contrast to the general skill in simulating the spatial structure of precipitation within LPS, models exhibited a wide range of biases in representing the amplitude of LPS precipitation, with the three main models examined showing large negative bias, large positive bias, and low bias, with the bias magnitude changing little or, unexpectedly, even degrading at finer resolution. For MCS metrics, the four HighResMIP models evaluated show varying skill in reproducing the observed composited MCS rainfall in Central US, with model ability to simulate intense convective precipitation a distinguishing factor. Skill scores are worse in summer than spring in all models, consistent with the more dominating frontal large-scale environments of MCS in spring, which are more skillfully simulated by global models (Song et al., 2019). The precipitation error decomposition into frontal precipitation frequency and intensity indicates that all the models evaluated have compensating biases. They produce frontal precipitation too frequently, with intensity that is too low. This is consistent with the results from CMIP5 in Catto et al 2015, although the CMIP6 models so far seem to have smaller errors. The total precipitation coming from fronts is well-

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

represented in the models, including the spatial patterns, indicating good representation of fronts themselves. For the AR precipitation metric, the ERA-20C reanalysis has a Taylor skill score of 0.96 relative to MERRA-2 when assessed using the TECA_BARD_ARDT AR tracking method, which provides a measure of observational uncertainty in the metric. Considering the inter-ARDT spread in the Taylor skill score, no models perform well in simulating AR precipitation as none is within one standard deviation of the ERA 20C reanalysis score.

As our diagnostic analysis has been summarized succinctly using scalar metrics, meta-analysis of model skill can be facilitated by developing a matrix of skill scores for models vs. metrics to reveal possible relationships among metrics and models. Comparing across metrics and models, it is clear that model skill varies substantially. To help reveal potential relationships among metrics and models, we identified the top-5 and bottom-5 simulations evaluated by each category of metrics (e.g., diurnal precipitation) and its sub-categories (e.g., amplitude and phase of diurnal precipitation). The results of this relative model ranking are not shown, as we focus on insights that can be gained from the comparative analysis rather than highlighting the performance of specific models. Consistent with the diverse model skill exhibited across metrics and models, only two model families are in the top-5 group for more than three different categories of metrics and are not in the bottom-5 group in any metrics. Similarly, only one model family is in the bottom-5 group for more than three categories of metrics and is not in the top-5 group in any metrics. Many models perform well in some metrics but poorly in other metrics. There is a general tendency for simulations produced by the same model family but using different resolutions, model versions, or model configurations, to perform similarly, although some exceptions can also be found.

Focusing on the actual model skill for each metric, we also identified the good and poor performing models in an absolute sense to determine how well models perform for each metric, and subsequently ranked the metrics according to those in which most models performed well or poorly. This absolute skill and ranking was determined by the developers of each metric based on their own judgement, which generally involved comparing model skill relative to some uncertainty related to observation data, and for ARs, uncertainty in tracking methods is also considered. A few metrics that stand out with more models performing well and poorly are highlighted here. Notably, more than 50% of the simulations evaluated based on the diurnal amplitude and phase of precipitation over ocean are considered skillful, while the same is true for the evaluation of spectral characteristics over land and the NH mid-latitudes, and for the scaling factor of daily precipitation and the pattern correlation of the cutoff scale between the simulated and observed duration of dry spells. In contrast, two metrics stand out as more challenging for models with more than 50% of the simulations considered to be performing poorly. These are correlation

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

coefficients of the Taylor skill score for spatial coherence over both land and ocean and the AR precipitation Taylor skill score. Lastly, more than 50% of the simulations are considered neutral (neither skillful nor poor) with respect to several metrics including: diurnal amplitude and phase over land, spectral analysis over ocean, Tropics and SH mid-latitudes, and MJO pattern correlation. For other metrics, models are more mixed in how well they represent the specific precipitation characteristics evaluated.

Based on the relative and absolute ranking, additional insights can be gained on the potential relationships among the metrics by calculating the correlation coefficients between the model ranking based on different metrics for the overlapping models, although not all metrics should be connected (e.g., due to geographical differences). For illustrative purposes, we calculated the correlation coefficients between the model ranking based on the temperature-water vapor environment and the model ranking based on other metrics for the overlapping models. We found relatively strong correlations ($r > 0.5$) of model skill in temperature-water vapor environment with model skill in precipitation cutoff-scale (both pattern correlation and scaling factor), spectral analysis, temporal and spatial coherence, and MJO propagation based on the relative ranking. On the other hand, model skill in temperature-water vapor environment has very low ($r < 0.2$) or negative correlations with model skill in diurnal precipitation over land (both amplitude and phase), diurnal precipitation over ocean (amplitude only), and dry spell cutoff-scale (pattern correlation). Notably, correlations with the rotation metric and MJO east/west power ratio metric are also rather low (<0.3).

The above analysis is suggestive of some predictive power of the model skill in temperature-water vapor environment on the model skill in several other precipitation characteristics. This motivates future work to understand these relationships by performing additional diagnostic analysis, and also to apply the exploratory metrics more systematically to the same set of model simulations using comparable benchmark datasets in order to support quantitative analysis of skill across models and metrics. This may reveal less obvious relationships among metrics and models, reflecting relationships among processes and/or weather phenomena highlighted by the metrics, or relationships among models due to commonality such as parameterization schemes. Such information is useful for guiding model development and model tuning. Machine learning approaches may be used to develop predictive models of the relationships among the different metrics presented here, or between those metrics and others such as metrics for the modes of climate variability (e.g., MJO, ENSO), circulation indices (e.g., monsoon), sea surface temperature pattern, etc. Such mapping of model skill scores across metrics may help focus

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

efforts on improving model prediction skill given the important role of modes of variability in predicting precipitation at various timescales.

Going beyond baseline metrics that evaluate basic precipitation features, data requirements are an obstacle for systematic application of the metric suite because high temporal frequency data and certain variables (e.g., outgoing longwave radiation) are not commonly available from the CMIP data archive. Table 2 is a good starting point for expansion in the future when more exploratory metrics are added. Communicating the data requirements to community efforts such as CMIP and demonstrating the usefulness of the exploratory metrics are both important for increasing awareness of, and advocacy for, the data needs of model evaluation and diagnostics to support the broad use of climate model output.

While the metrics described in this study are useful individually, combining or connecting them may potentially provide more powerful metrics to benchmark models as well as revealing the underlying reasons or sources of the model biases. At the same time, decomposition of the metrics into independent components is useful for attributing model biases to multiple factors. Future work to standardize the metrics, addressing uncertainties in observation data and tracking methods, and improving interpretations of the metrics, may facilitate more robust use of the exploratory metrics. There may also be a need to reconcile features attributed to different phenomena simultaneously. For example, precipitation from MCSs embedded within frontal systems could potentially be attributed to both MCS and frontal precipitation (e.g., Dowdy and Catto 2017; Catto and Dowdy 2021). In regions such as the Bay of Bengal where LPS and MCS are both prominent, it is not clear if certain precipitation events could be attributed simultaneously to LPS and MCS and what implications this may have on metrics built upon these phenomena. Coding and software aspects may also require some attention in the future to facilitate implementation of the exploratory metrics in community packages for broader adoption and use.

Through this study, we have developed methodologies and analysis codes to calculate metrics and track weather phenomena. Applying them to the CMIP6 output and observation data has generated intermediate quantities and datasets such as tracks of LPS, MCS, fronts, and AR and associated precipitation and large-scale environments. These datasets are useful not only for model evaluation but also for scientific investigations. For example, datasets derived from the historical simulations could be combined with similar datasets for simulations of the future climate to investigate the response of various precipitation metrics to radiative forcing. Different metrics may also be combined to understand the connections between different weather phenomena and storm types and their connections to the temperature-water vapor environments and modes of variability. Among the metrics described in this study, the spectral and

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

coherence metrics have already been included in ASoP, and some atmospheric river tracking algorithms are available from Coordinated Model Evaluation Capabilities (CMEC). Efforts are ongoing to coordinate the development and implementation of metrics to be incorporated in community diagnostic packages to facilitate broader use to improve quantification and understanding of precipitation biases in weather and climate models.

Acknowledgements:

This study represents a collaborative effort as an outgrowth of a workshop on "Benchmarking Simulated Precipitation in Earth System Models", sponsored by the Office of Science of the U.S. Department of Energy (DOE) Biological and Environmental Research through the Regional and Global Model Analysis (RGMA) program area. RGMA also supported Leung and Feng under the WACCEM scientific focus area, O'Brien and Zhou under the CASCADE scientific focus area, Boos and Vishnu under Award DE-SC0019367, DeMott under Award DE- SC0020092, and Klingaman and Lee under Award DE-SC0020324. O'Brien's efforts were also partially supported by the Environmental Resilience Institute, funded by Indiana University's Prepared for Environmental Change Grand Challenge initiative. Neelin, Kuo and Martinez-Villalobos efforts were supported by National Science Foundation grant AGS-1936810 and National Oceanic and Atmospheric Administration grant NA18OAR4310280. Martinez-Villalobos was also supported by Proyecto Corfo Ingeniería 2030 código 14ENI2-26865. Work at LLNL was supported by the DOE Office of Science Biological and Environmental Research through the Earth System Model Development program area and the Atmospheric Radiation Measurement program, and performed under the auspices of the U.S. DOE by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. Pacific Northwest National Laboratory is operated for the Department of Energy by Battelle Memorial Institute under contract DE-AC05-76RL01830. Martin was supported by the UK-China Research & Innovation Partnership Fund through the Met Office Climate Science for Service Partnership (CSSP) China, as part of the Newton Fund, and by the Weather and Climate Science for Service Partnership (WCSSP) India, a collaborative initiative between the Met Office, supported by the UK Government's Newton Fund, and the Indian Ministry of Earth Sciences (MoES). This research used resources of the National Energy Research Scientific Computing Center (NERSC), also supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. We acknowledge the World Climate Research Programme, which, through its Working Group on Coupled Modelling, coordinated and promoted CMIP6. We thank the climate modeling groups for producing and making available their model output, the Earth System Grid Federation (ESGF) for archiving the data and providing access, and the multiple funding agencies who support CMIP6 and

ESGF. We thank DOE's RGMA program area, the Data Management program, and NERSC for making this coordinated CMIP6 analysis activity possible.

References:

Adames, Á. F., & Kim, D. (2016). The MJO as a dispersive, convectively coupled moisture wave: Theory and observations. *Journal of the Atmospheric Sciences*, *73*(3), 913-941.

Ahn, MS., Kim, D., Sperber, K.R., Kang, I.-S., Maloney, E., Waliser, D., Hendon, H., on behalf of WGNE MJO Task Force. (2017). MJO simulation in CMIP5 climate models: MJO skill metrics and process-oriented diagnosis. Clim Dyn 49, 4023–4045. https://doi.org/10.1007/s00382-017-3558-4

Ahn, M.- S., Kim, D., Kang, D., Lee, J., Sperber, K. R., Gleckler, P. J., Jiang, X., Yoo-Geun, H., Kim, H. (2020). MJO propagation across the Maritime Continent: Are CMIP6 models better than CMIP5 models? Geophysical Research Letters, 47, e2020GL087250. https://doi.org/10.1029/2020GL087250

Ahn, M. S., Kim, D., Kang, D., Lee, J., Sperber, K. R., Gleckler, P. J., ... & Kim, H. (2020). MJO propagation across the Maritime Continent: Are CMIP6 models better than CMIP5 models?. *Geophysical Research Letters*, *47*(11), e2020GL087250.

Ashouri, H., Hsu, K. L., Sorooshian, S., Braithwaite, D. K., Knapp, K. R., Cecil, L. D., ... & Prat, O. P. (2015). PERSIANN-CDR: Daily precipitation climate data record from multisatellite observations for hydrological and climate studies. *Bulletin of the American Meteorological Society*, *96*(1), 69-83.

Berry, G., Reeder, M. J., & Jakob, C. (2011). A global climatology of atmospheric fronts. *Geophysical Research Letters*, *38*(4).

Bretherton, C. S., Peters, M. E., & Back, L. E. (2004). Relationships between water vapor path and precipitation over the tropical oceans. *Journal of climate*, *17*(7), 1517-1528.

Caldwell, P. M., Mametjanov, A., Tang, Q., Van Roekel, L. P., Golaz, J.-C., Lin, W. et al. (2019). The DOE E3SM coupled model version 1: Description and results at high resolution. *Journal of Advances in Modeling Earth Systems*, 11. https://doi.org/10.1029/2019MS001870.

Catto, J. L., & Pfahl, S. (2013). The importance of fronts for extreme precipitation. *Journal of Geophysical Research: Atmospheres*, *118*(19), 10-791.

Catto, J. L., Jakob, C., & Nicholls, N. (2015). Can the CMIP5 models represent winter frontal precipitation?. *Geophysical Research Letters*, *42*(20), 8596-8604.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

Chang, M., Liu, B., Martinez-Villalobos, C., Ren, G., Li, S., & Zhou, T. (2020). Changes in extreme precipitation accumulations during the warm season over continental China, *Journal of Climate*, 33(24), 10799-10811.

Catto, J. L., and A. J. Dowdy (2021) Understanding compound hazards from a weather system perspective, Weather and Climate Extremes, 32, 100313, https://doi.org/10.1016/j.wace.2021.100313.3

Chen, D., and A. Dai, 2018: Dependence of estimated precipitation frequency and intensity on data resolution. Climate Dynamics, 50, 3625–3647. DOI: 10.1007/s00382-017-3830-7.

Chen, D. and A. Dai, 2019: Precipitation characteristics in the Community Atmosphere Model and their dependence on model physics and resolution. J. Adv. Model. Earth Syst., 11, 2352-2374. https://doi.org/10.1029/2018MS001536.

Chen, D., A. Dai and A. Hall, 2021, Precipitation partitioning and the "drizzling" bias in CMIP5 models. J. Geophys. Res., 126, e2020JD034198. https://doi.org/10.1029/2020JD034198.

Chen, J., A. Dai, and Y. Zhang, 2020: Linkage between projected precipitation and atmospheric thermodynamic changes. J. Climate, 33, 7155-7178, https://doi.org/10.1175/JCLI-D-19-0785.1.

Covey, C., and P. Gleckler, 2014: Standard diagnostics for the diurnal cycle of precipitation. Lawrence Livermore National Laboratory Tech. Rep. LLNL-TR-659685, 11 pp. [Available online at https://e-reports-ext.llnl.gov/pdf/780868.pdf.]

Covey, C., and Coauthors, 2016: Metrics for the Diurnal Cycle of Precipitation: Toward Routine Benchmarks for Climate Models. *J. Climate*, 29, 4461-4471.

Dai, A., F. Giorgi, and K. E. Trenberth, 1999: Observed and model simulated diurnal cycles of precipitation over the contiguous United States. J. Geophys. Res., 104, 6377-6402.

Dai, A., 2001: Global precipitation and thunderstorm frequencies. Part II: Diurnal variations. *J. Climate*, 14, 1112-1128.

Dai, A., 2006: Precipitation characteristics in eighteen coupled climate models. *J. Climate*, 19, 4605-4630.

Dai, A., X. Lin, and K.-L. Hsu, 2007: The frequency, intensity, and diurnal cycle of precipitation in surface and satellite observations over low- and mid-latitudes. *Clim. Dyn.*, 29, 727-744.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., ... & Vitart, F. (2011). The ERA- Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the royal meteorological society*, *137*(656), 553-597.

DeMott, C. A., Klingaman, N. P., Tseng, W. L., Burt, M. A., Gao, Y., & Randall, D. A. (2019). The convection connection: How ocean feedbacks affect tropical mean moisture and MJO propagation. *Journal of Geophysical Research: Atmospheres*, *124*(22), 11910-11931.

Deser, C., Phillips, A., Bourdette, V., & Teng, H. (2012). Uncertainty in climate change projections: the role of internal variability. *Climate dynamics*, 38(3), 527-546.

Dettinger, M., 2011: Climate change, atmospheric rivers, and floods in California - a multimodel analysis of storm frequency and magnitude changes. *J. Am. Water Resour. Assoc*., 47, 514–523, https://doi.org/10.1111/j.1752-1688.2011.00546.x.

Dowdy, A., and J. L. Catto (2017), Extreme weather caused by concurrent cyclone, front and thunderstorm occurrences, *Scientific Reports*, 7:40359, DOI: 10.1038/srep40359

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937-1958.

Eyring, V., Righi, M., Lauer, A., Evaldsson, M., Wenzel, S., Jones, C., ... & Williams, K. D. (2016). ESMValTool (v1. 0)–a community diagnostic and performance metrics tool for routine evaluation of Earth system models in CMIP. *Geoscientific Model Development*, *9*(5), 1747-1802.

Feng, Z., F. Song, K. Sakaguchi, and L.R. Leung. 2021a. "Evaluation of Mesoscale Convective Systems in Climate Simulations: Methodological Development and Results from MPAS-CAM over the U.S." *J. Clim*., 34(7), 2611-2633, doi:10.1175/JCLI-D-20-0136.1.

Feng, Z., L.R. Leung, N. Liu, J. Wang, R.A. Houze Jr., J. Li, J.C. Hardin, and J. Guo. 2021b. "A Global High-Resolution Mesoscale Convective System Database Using Satellite-derived Cloud Tops, Surface Precipitation, and Tracking." *J. Geophys. Res*., 126, doi:10.1029/2020JD034202.

Feng, Z., R.A. Houze, Jr., L.R. Leung, F. Song, J. Hardin, J. Wang, W. Gustafson, Jr., and C. Homeyer. 2019. "Spatiotemporal Characteristics and Large-scale Environment of Mesoscale Convective Systems East of the Rocky Mountains." *J. Clim*., 32, 7303-7328, doi:10.1175/JCLI-D-19-0137.1.

Feng, Z., L.R. Leung, R.A. Houze, Jr., S. Hagos, J. Hardin, Q. Yang. B. Han, and J. Fan. 2018. "Structure and Evolution of Mesoscale Convective Systems: Sensitivity to Cloud Microphysics in Convection-Permitting Simulations Over the U.S." *J. Adv. Model. Earth Syst*., 10, doi: 10.1029/2018MS001305.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

Feng, Z., L.R. Leung, S. Hagos, R.a. Houze, Hr., C.D. Burleyson, and K. Balaguru. (2016). "More frequent intense and long-lived storms dominate the trend in central U.S. rainfall." *Nature Commun*., 7, 13429, doi:10.1038/ncomms13429.

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., ... & Zhao, B. (2017). The modern-era retrospective analysis for research and applications, version 2 (MERRA-2). *Journal of climate*, *30*(14), 5419-5454.

Gimeno, L., R. Nieto, M. Vázquez, and D. A. Lavers, 2014: Atmospheric rivers: a mini-review. Front. Earth Sci., 2, 1–6, https://doi.org/10.3389/feart.2014.00002.

Gleckler, P. J., C. Doutriaux, P. J. Durack, K. E. Taylor, Y. Zhang, D. N. Williams, E. Mason, and J. Servonnat (2016), A more powerful reality test for climate models, Eos, 97, doi:10.1029/2016EO051663. Published on 3 May 2016.

Guan, B., and D. E. Waliser, 2015: Detection of atmospheric rivers: Evaluation and application of an algorithm for global studies. *J. Geophys. Res. Atmos*., 120, 12514–12535, https://doi.org/10.1002/2015JD024257.

Guan, B., Waliser, D. E., & Ralph, F. M. (2018). An Intercomparison between Reanalysis and Dropsonde Observations of the Total Water Vapor Transport in Individual Atmospheric Rivers, *Journal of Hydrometeorology*, 19(2), 321-337. Retrieved Jul 23, 2021, from https://journals.ametsoc.org/view/journals/hydr/19/2/jhm-d-17-0114_1.xml

Haarsma, R.J., M. Roberts, P.L. Vidale, C.A. Senior, A. Bellucci, Q. Bao, P. Chang, S. Corti, N.S. Fučkar, V. Guemas, J. von Hardenberg, W. Hazeleger, C. Kodama, T. Koenigk, L.R. Leung, J. Lu, J.-J. Luo, J. Mao, M. Mizielinski, R. Mizuta, P. Nobre, M. Satoh, E. Scoccimarro, T. Semmler, J. Small, and J.-S. von Storch. 2016. "High Resolution Model Intercomparison Project (HighResMIP v1.0) for CMIP6." *Geosci. Mod. Dev.*, 9, 4185-4208, doi:10.5194/gmd-9-4185-2016.

Henderson, S. A., Maloney, E. D., & Son, S. W. (2017). Madden–Julian oscillation Pacific teleconnections: The impact of the basic state and MJO representation in general circulation models. *Journal of Climate*, *30*(12), 4567-4587.

Hersbach, H., and coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Royal Meteorol. Soc*., 146, 1999-2049.

Hirota, H., Y. N. Takayabu, M. Watanabe, M. Kimoto, and M. Chikira, 2014: Role of convective entrainment in spatial distributions of and temporal variations in precipitation over tropical oceans. J. Climate, 27, 8707–8723, https://doi.org/10.1175/JCLI-D-13-00701.1.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

Hoffmann, L., and Coauthors, 2019: From ERA-Interim to ERA5: The Considerable Impact of Ecmwf's Next-Generation Reanalysis on Lagrangian Transport Simulations. *Atmos. Chem. Phys.*, **19,** 3097-3124.

Huffman, G. J., Adler, R. F., Bolvin, D. T., & Gu, G. (2009). Improving the global precipitation record: GPCP version 2.1. *Geophysical Research Letters*, *36*(17).

Huffman, G. J., Bolvin, D. T., Nelkin, E. J., Wolff, D. B., Adler, R. F., Gu, G., ... & Stocker, E. F. (2007). The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. *Journal of hydrometeorology*, *8*(1), 38-55.

Huffman, G. J., Adler, R. F., Behrangi, A., Bolvin, D. T., Nelkin, E., Song, Y., 2020: Algorithm Theoretical Basis Document (ATBD) for Global Precipitation Climatology Project Version 3.1 Precipitation Data. (https://docserver.gesdisc.eosdis.nasa.gov/public/project/MEaSUREs/GPCP/GPCP_ATBD_V3.1.pdf)

Hwang, Y.-T., and D. M. W. Frierson, 2013: Link between the double-intertropical convergence zone problem and cloud biases over the Southern Ocean. Proc. Natl. Acad. Sci. USA, 110, 4935–4940, https://doi.org/10.1073/pnas.1213302110.

Janowiak, J., Joyce, B., & Xie, P. (2017). NCEP/CPC L3 half hourly 4km global (60S - 60N) merged IR V1. Retrieved from https://doi.org/10.5067/P4HZB9N27EKU.

Jiang, X., et al. (2015). Vertical structure and physical processes of the Madden-Julian oscillation: Exploring key model physics in climate simulations. *Journal of Geophysical Research Atmospheres*, 120, 4718-4748, doi:10.1002/2014JD022375.

Joyce, R. J., & Xie, P. (2011). Kalman filter–based CMORPH. *Journal of Hydrometeorology*, *12*(6), 1547-1563.

Klingaman, N.P., G.M. Martin and A.F. Moise (2017): ASoP (v1.0): A set of methods for analyzing scales of precipitation in general circulation models. *Geosci. Model Dev.*, 10, 57-83, doi:10.5194/gmd-10-57-2017, 2017.

Klingaman, N. P., Jiang, X., Xavier, P. K., Petch, J., Waliser, D., & Woolnough, S. J. (2015). Vertical structure and physical processes of the Madden- Julian oscillation: Synthesis and summary. *Journal of Geophysical Research: Atmospheres*, *120*(10), 4671-4689.

Krishnamurthy, V., & Ajayamohan, R. S. (2010). Composite structure of monsoon low pressure systems and its relation to Indian rainfall. Journal of Climate, 23, 4285-4305.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

Kuo, Y.-H., J. D. Neelin and C. R. Mechoso, 2017: Tropical convective transition statistics and causality in the water vapor-precipitation relation. *J. Atmos. Sci*., 74(3), 915-931.

Kuo, Y.-H., K. A. Schiro and J. D. Neelin, 2018: Convective transition statistics over tropical oceans for climate model diagnostics: Observational baseline. *J. Atmos. Sci.*, 75(5), 1553-1570.

Kuo, Y.-H., and Coauthors, 2020: Convective transition statistics over tropical oceans for climate model diagnostics: GCM evaluation. *J. Atmos. Sci*., 77(1), 379-403.

Lin, J.-L. (2007). The double-ITCZ problem in IPCC AR4 coupled GCMs: ocean-atmosphere feedback analysis. *J. Clim*., 20, 4497-4525. https://doi.org/10.1175/JCLI4272.1.

Lin, Y., W. Dong, M. Zhang, Y. Xie, W. Xue, J. Huang, and Y. Luo. (2017). Causes of model dry and warm bias over central U.S. and impact on climate projections. *Nature Commun*., 8, 881, doi:10.1038/s41467-017-01040-2.

Ma, H.-Y., Xie, S., Boyle, J. S., Klein, S. A., and Zhang, Y. (2013). Metrics and diagnostics for precipitation-related processes in climate model short-range hindcasts. J. Clim., 26, 1516-1534, doi:10.1175/JCLI-D-12-00235.1.

Mechoso, C. R., Robertson A. W., Barth, N., Davey, M. K., Delecluse, P. et al. (1995). The seasonal cycle over the tropical Pacific in coupled ocean-atmosphere general circulation models. Mon. Wea. Rev., 123, 2825-2838. https://doi.org/10.1175/1520-0493(1995)123<2825:TSCOTT>2.0.CO;2.

McClenny, E. E., Ullrich, P. A., & Grotjahn, R. (2020). Sensitivity of atmospheric river vapor transport and precipitation to uniform sea-surface temperature increases. n/a(n/a), e2020JD033421. doi:10.1029/2020jd033421

Mapes, B., and R. Neale, 2011: Parameterizing convective organization to escape the entrainment dilemma. J. Adv. Model. Earth Syst., 3, M06004, https://doi.org/10.1029/2011MS000042.

Martin, G.M., N.P. Klingaman and A.F. Moise (2017): Connecting spatial and temporal scales of tropical precipitation in observations and the MetUM-GA6. *Geosci. Model Dev*., 10, 105-126 doi:10.5194/gmd-10-105-2017, 2017

Martinez-Villalobos, C., & Neelin, J.D. (2019). Why do precipitation intensities tend to follow Gamma distributions? *Journal of the Atmospheric Sciences*, 76(11), 3611-3631

Martinez- Villalobos, C., & Neelin, J. D. (2018). Shifts in precipitation accumulation extremes during the warm season over the United States. *Geophysical Research Letters*, 45, 8586– 8595.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

Martinez-Villalobos, C., and J. D. Neelin, 2021: Climate models capture key features of extreme precipitation probabilities across regions. *Environ. Res. Lett.* 16 024017

Mehran, A., AghaKouchak, A., & Phillips, T. J. (2014). Evaluation of CMIP5 continental precipitation simulations relative to satellite- based gauge- adjusted observations. *Journal of Geophysical Research: Atmospheres*, *119*(4), 1695-1707.

Mejia, J. F., Koračin, D., & Wilcox, E. M. (2018). Effect of coupled global climate models sea surface temperature biases on simulated climate of the western United States. *International Journal of Climatology*, *38*(14), 5386-5404.

Mundhenk, B. D., Barnes, E. A., & Maloney, E. D. (2016). All-Season Climatology and Variability of Atmospheric River Frequencies over the North Pacific. *Journal of Climate*, 29(13), 4885-4903. doi:10.1175/Jcli-D-15-0655.1

Neelin, J. D., Peters, O., Lin, J. W. B., Hales, K., & Holloway, C. E. (2008). Rethinking convective quasi-equilibrium: Observational constraints for stochastic convective schemes in climate models. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *366*(1875), 2579-2602.

Neelin, J. D., S. Sahany, S. N. Stechmann, D. N. Bernstein (2017). Global warming precipitation accumulation increases above the current-climate cutoff scale. *Proceedings of the National Academy of Sciences,* 114(6) 1258-1263

O'Brien, T. A., and Coauthors, 2020a: Detection Uncertainty Matters for Understanding Atmospheric Rivers. Bull. Am. Meteorol. Soc., 101, E790–E796, https://doi.org/10.1175/BAMS-D-19-0348.1.

O'Brien, T. A., and Coauthors, 2020b: Detection of atmospheric rivers with inline uncertainty quantification: TECA-BARD v1.0.1. *Geosci. Model Dev*., 13, 6131–6148, https://doi.org/10.5194/gmd-13-6131-2020.

O'Brien, T. A., and Coauthors, 2021: Increases in Future AR Count and Size: Overview of the ARTMIP Tier 2 CMIP5/6 Experiment. *Geophys. Res. Lett*.., In Revision, https://doi.org/10.1002/essoar.10504170.1.

Oueslati, B., and G. Bellon, 2013: Convective entrainment and large-scale organization of tropical precipitation: Sensitivity of the CNRM-CM5 hierarchy of models. J. Climate, 26, 2931–2946, https://doi.org/10.1175/JCLI-D-12-00314.1.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

Payne, A. E., and G. Magnusdottir, 2015: An evaluation of atmospheric rivers over the North Pacific in CMIP5 and their response to warming under RCP 8.5. *J. Geophys. Res. Atmos*., 120, 11,173-11,190, https://doi.org/10.1002/2015JD023586.

Pendergrass, A. G., Gleckler, P. J., Leung, L. R., & Jakob, C. (2020). Benchmarking Simulated Precipitation in Earth System Models. *Bulletin of the American Meteorological Society*, *101*(6), E814-E816.

Perkins, S. E. et al., 2007: Evaluation of the AR4 Climate Models' Simulated Daily Maximum Temperature, Minimum Temperature, and Precipitation over Australia Using Probability Density Functions. *J. Climate*, 20, 4356–4376, https://doi.org/10.1175/JCLI4253.1.

Pierrehumbert, R. T., , H. Brogniez, , and R. Roca, 2007: On the relative humidity of the earth's atmosphere. *The Global Circulation of the Atmosphere: Phenomena, Theory, Challenges*, T. Schneider and A. H. Sobel, Eds., Princeton University Press, 143–185.

Qian, T., A. Dai, K. E. Trenberth, and K. W. Oleson, 2006: Simulation of global land surface conditions from 1948-2004. Part I: Forcing data and evaluation. J. Hydrometeorology, 7, 953-975.

Ralph, F. M., M. D. Dettinger, M. M. Cairns, T. J. Galarneau, and J. Eylander, 2018: Defining "Atmospheric River": How the Glossary of Meteorology Helped Resolve a Debate. *Bull. Am. Meteorol. Soc*., 99, 837–839, https://doi.org/10.1175/BAMS-D-17-0157.1.

Rutz, J. J., W. James Steenburgh, and F. Martin Ralph, 2014: Climatological characteristics of atmospheric rivers and their inland penetration over the western United States. *Mon. Weather Rev*., 142, 905–921, https://doi.org/10.1175/MWR-D-13-00168.1.

Rutz, J. J., and Coauthors, 2019: The Atmospheric River Tracking Method Intercomparison Project (ARTMIP): Quantifying Uncertainties in Atmospheric River Climatology. *J. Geophys. Res. Atmos*., 124, 13777–13802, https://doi.org/10.1029/2019JD030936.

Saha, S., Moorthi, S., Pan, H. L., Wu, X., Wang, J., Nadiga, S., ... & Goldberg, M. (2010). The NCEP climate forecast system reanalysis. *Bulletin of the American Meteorological Society*, *91*(8), 1015-1058.

Shearer, E. J., Nguyen, P., Sellars, S. L., Analui, B., Kawzenuk, B., Hsu, K. L., & Sorooshian, S. (2020). The Atmospheric River-CONNected objeECT (AR-CONNECT) algorithm applied to the National Aeronautics and Space Administration (NASA) Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA V2) - 1983 to 2016. UC San Diego Library Digital Collections. https://doi.org/10.6075/J0D21W00

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

Shields, C. A., and J. T. Kiehl, 2016: Simulating the Pineapple Express in the half degree Community Climate System Model, CCSM4. *Geophys. Res. Lett*., 43, 7767–7773, https://doi.org/10.1002/2016GL069476.

Shields, C. A., Rutz, J. J., Leung, L. Y., Ralph, F. M., Wehner, M., Kawzenuk, B., ... & Nguyen, P. (2018). Atmospheric river tracking method intercomparison project (ARTMIP): project goals and experimental design. *Geoscientific Model Development*, *11*(6), 2455-2474.

Sikka, D.R., (1980). Some aspects of the large scale fluctuations of summer monsoon rainfall over India in relation to fluctuations in the planetary and regional scale circulation parameters. Proceedings of the Indian Academy of Sciences-Earth and Planetary Sciences, 89, 179-195.

Skinner, C. B., Lora, J. M., Payne, A. E., & Poulsen, C. J. (2020). Atmospheric river changes shaped mid-latitude hydroclimate since the mid-Holocene. *Earth and Planetary Science Letters*, 541, 116293. doi:https://doi.org/10.1016/j.epsl.2020.116293

Sperber, K.R., Kim, D. (2012), Simplified metrics for the identification of the Madden–Julian oscillation in models. *Atmosph. Sci. Lett*., 13: 187-193. https://doi.org/10.1002/asl.378

Sperber, K. R., Annamalai, H., Kang, I. S., Kitoh, A., Moise, A., Turner, A., ... & Zhou, T. (2013). The Asian summer monsoon: an intercomparison of CMIP5 vs. CMIP3 simulations of the late 20th century. *Climate dynamics*, *41*(9-10), 2711-2744.

Stan, C., Straus, D. M., Frederiksen, J. S., Lin, H., Maloney, E. D., & Schumacher, C. (2017). Review of tropical- extratropical teleconnections on intraseasonal time scales. *Reviews of Geophysics*, *55*(4), 902-937.

Stechmann , S.N., & Neelin, J.D. (2014). First-passage-time prototypes for precipitation statistics. *Journal of the Atmospheric Sciences*, 71(9), 3269-3291

Stephens, G. L., L'Ecuyer, T., Forbes, R., Gettelmen, A., Golaz, J. C., Bodas- Salcedo, A., ... & Haynes, J. (2010). Dreary state of precipitation in global models. *Journal of Geophysical Research: Atmospheres*, *115*(D24).

Tan, J., Huffman, G. J., Bolvin, D. T., & Nelkin, E. J. (2019a). Diurnal cycle of IMERG V06 precipitation. *Geophysical Research Letters*, 46(22), 13584–13592. https://doi.org/10.1029/2019GL085395

Tang, S., Gleckler, P., Xie, S., Lee, J., Ahn, M. S., Covey, C., & Zhang, C. (2021). Evaluating the Diurnal and Semidiurnal Cycle of Precipitation in CMIP6 Models Using Satellite-and Ground-Based Observations. *Journal of Climate*, *34*(8), 3189-3210.

Tao C., S. Xie, et al. 2021: Diurnal cycle of precipitation over monsoon regimes simulated in CMIP6. In preparation.

Tapiador, F. J., Roca, R., Del Genio, A., Dewitte, B., Petersen, W. and Zhang, F. (2019). Is precipitation a good metric for model performance? Bulletin of American Meteorological Society, 223-233, doi:10.1175/BAMS-D-17-0218.1.

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. *Bulletin of the American meteorological Society*, *93*(4), 485-498.

Thomas, C. M., & Schultz, D. M. (2019). What are the best thermodynamic quantity and function to define a front in gridded model output? *Bulletin of the American Meteorological Society*, *100*(5), 873-895.

Tian, B. (2015). Spread of model climate sensitivity linked to double-Intertropical Convergence Zone bias, *Geophys. Res. Lett*., 42, 4133–4141, doi:10.1002/2015GL064119.

Tian, B., & Dong, X. (2020). The double- ITCZ bias in CMIP3, CMIP5, and CMIP6 models based on annual mean precipitation. *Geophysical Research Letters*, *47*(8), e2020GL087232.

Trenberth, K. E., L. Smith, T. Qian, A. Dai, and J. Fasullo, 2007: Estimates of the global water budget and its annual cycle using observational and model data. J. Hydrometeor., 8, 758–769, https://doi.org/10.1175/JHM600.1.

TRMM, 2011: TRMM Precipitation Radar rainfall rate and profile L2 1.5 hours V7. Goddard Earth Sciences Data and Information Services Center, accessed 19 August 2016, https://disc.gsfc.nasa.gov/datacollection/TRMM_2A25_7.html.

Ullrich, P. A., & Zarzycki, C. M. (2017). TempestExtremes: A framework for scale-insensitive pointwise feature tracking on unstructured grids. *Geoscientific Model Development*, *10*(3), 1069-1090.

Vishnu, S., Boos, W. R., Ullrich, P. A., & O'Brien, T. A. (2020). Assessing historical variability of South Asian monsoon lows and depressions with an optimized tracking algorithm. *Journal of Geophysical Research: Atmospheres*, *125*(15), e2020JD032977.

Wang, B., Lee, S. S., Waliser, D. E., Zhang, C., Sobel, A., Maloney, E., ... & Ha, K. J. (2018). Dynamics-oriented diagnostics for the Madden–Julian oscillation. *Journal of Climate*, *31*(8), 3117-3135.

Wang, J., Kim, H., Kim, D., Henderson, S. A., Stan, C., & Maloney, E. D. (2020). MJO Teleconnections over the PNA Region in Climate Models. Part II: Impacts of the MJO and Basic State. *Journal of Climate*, *33*(12), 5081-5101.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.

Wentz, F. J., C. Gentemann, and K. A. Hilburn, 2015: Remote Sensing Systems TRMM TMI Daily Environmental Suite on 0.25 deg grid, version 7.1. Remote Sensing Systems, accessed 8 July 2016, www.remss.com/missions/tmi.

Wolding, B., Dias, J., Kiladis, G., Ahmed, F., Powell, S. W., Maloney, E., & Branson, M. (2020). Interactions between moisture and tropical convection. Part I: The coevolution of moisture and convection. *Journal of the Atmospheric Sciences*, *77*(5), 1783-1799.

Xie, S., R. T. Cederwall, and M. Zhang, 2004: Developing long-term single-column model/cloud system–resolving model forcing data using numerical weather prediction products constrained by surface and top of the atmosphere observations. *J. Geophys. Res*., 109, D01104, https://doi.org/10.1029/2003JD004045.

Xie, S., and Coauthors, 2010: Clouds and more: ARM climate modeling best estimate data. *Bull. Amer. Meteor. Soc*., 91, 13– 20, https://doi.org/10.1175/2009BAMS2891.1.

Xie, S., and Coauthors, 2019: Improved diurnal cycle of precipitation in E3SM with a revised convective triggering function. *J. Adv. Model. Earth Syst*., 11, 2290-2310.

Yadav, P., & Straus, D. M. (2017). Circulation response to fast and slow MJO episodes. *Monthly Weather Review*, *145*(5), 1577-1596.

Yang, G. Y., and Slingo, J. (2001). The diurnal cycle in the tropics. *Monthly Weather Review*, 129(4), 784–801. https://doi.org/10.1175/-0493(2001)129%3C0784:Tdcitt%3E2.0.Co;2.

Yin, L., Fu, R., Shevliakova, E., & Dickinson, R. E. (2013). How well can CMIP5 simulate precipitation and its controlling processes over tropical South America?. *Climate Dynamics*, *41*(11-12), 3127-3143.

Zhou, Y., O'Brien, T. A., Ullrich, P. A., Collins, W. D., Patricola, C. M., & Rhoades, A. M. (2021). Uncertainties in atmospheric river lifecycles by detection algorithms: Climatology and variability. *Journal of Geophysical Research: Atmospheres*, 126, e2020JD033711. https://doi.org/10.1029/2020JD033711.

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0590.1.