Systematic and Objective Evaluation of Earth System Models: PCMDI Metrics Package (PMP) version 3

Jiwoo Lee[1], Peter J. Gleckler[1], Min-Seop Ahn[2,3], Ana Ordonez[1], Paul A. Ullrich[1,4], Kenneth R. Sperber[1,a], Karl E. Taylor[1], Yann Y. Planton[5,6], Eric Guilyardi[7,8], Paul Durack[1], Celine Bonfils[1], Mark D. Zelinka[1], Li-Wei Chao[1], Bo Dong[1], Charles Doutriaux[1], Chengzhu Zhang[1], Tom Vo[1], Jason Boutte[1], Michael F. Wehner[9], Angeline G. Pendergrass[10,11], Daehyun Kim[12], Zeyu Xue[13], Andrew T. Wittenberg[14], and John Krasting[14]

[1] Lawrence Livermore National Laboratory, Livermore, California, USA
[2] NASA Goddard Space Flight Center, Greenbelt, MD, USA
[3] ESSIC, University of Maryland, College Park, MD, USA
[4] University of California, Davis, Davis, California, USA
[5] NOAA Pacific Marine Environmental Laboratory, Seattle, Washington, USA
[6] Monash University, Clayton, Australia
[7] LOCEAN-IPSL, CNRS-IRD-MNHN-Sorbonne Université, Paris, France
[8] National Centre for Atmospheric Science-Climate, University of Reading, Reading, UK
[9] Lawrence Berkeley National Laboratory, Berkeley, California, USA
[10] Department of Earth and Atmospheric Science, Cornell University, Ithaca, New York, USA
[11] National Center for Atmospheric Research, Boulder, Colorado, USA
[12] School of Earth and Environmental Sciences, Seoul National University, Seoul, South Korea
[13] Pacific Northwest National Laboratory, Richland, WA, USA
[14] NOAA Geophysical Fluid Dynamics Laboratory, Princeton, NJ, USA
[a] Retired

Submitted to Geoscientific Model Development (GMD) in November 2023
Revised in December 2023

Corresponding to: Jiwoo Lee (lee1043@llnl.gov)
7000 East Ave, Livermore, California 94550, USA

Abstract

Systematic, routine, and comprehensive evaluation of Earth System Models (ESMs) facilitates benchmarking improvement across model generations and identifying the strengths and weaknesses of different model configurations. By gauging the consistency between models and observations, this endeavor is becoming increasingly necessary to objectively synthesize thousands of simulations contributed to the Coupled Model Intercomparison Project (CMIP) to date. The Program for Climate Model Diagnosis and Intercomparison (PCMDI) Metrics Package (PMP) is an open-source Python software package that provides "quick-look" objective comparisons of ESMs with one another and with observations. The comparisons include metrics of large- to global-scale climatologies, tropical inter-annual and intra-seasonal variability modes such as El Niño-Southern Oscillation (ENSO) and Madden-Julian Oscillation (MJO), extratropical modes of variability, regional monsoons, cloud radiative feedbacks, and high-frequency characteristics of simulated precipitation, including extremes. The PMP results are produced in the context of all model simulations contributed to CMIP6 and earlier CMIP phases. An important priority of the PMP is to document performance of ESMs participating in the recent phases of CMIP, together with providing version-controlled information for all data sets, software packages and analysis codes being used in the evaluation processes. Among other purposes, this also enables modeling groups to assess performance changes during the ESM development cycle in the context of the error distribution of the multi-model ensemble. Quantitative model evaluation provided by the PMP can assist modelers in their development priorities. In this paper, we provide an overview of the PMP including its latest capabilities, and discuss its future direction.

1 Introduction

Earth System Models (ESMs) are key tools for projecting climate change and conducting research to enhance our understanding of the Earth system. With the advancements in computing power and the increasing importance of climate projections, there has been an exponential growth of data size and diversity of ESM simulations. During the 1990's, the Atmospheric Model Intercomparison Project (AMIP; Gates, 1992; Gates et al., 1999) was a centralizing activity within the modeling community, which led to the creation of the Coupled Model Intercomparison Project (CMIP; Meehl et al., 1997, 2000, 2007; Covey et al., 2003; Taylor et al., 2012). Since 1989, the Program for Climate Model Diagnosis and Intercomparison (PCMDI) has worked closely with the World Climate Research Programme's (WCRP) Working Group on Coupled Models (WGCM) and Working Group on Numerical Experimentation (WGNE) to design and implement these projects (Potter et al., 2011). The most recent phase of CMIP (CMIP6; Eyring et al., 2016) provides a set of well-defined experiments that most climate modeling centers perform, and subsequently makes results available for a large and diverse community to analyze.

Evaluating ESMs is a complex endeavor, given the vast range of climate characteristics across space and time scales. A necessary step involves quantifying the consistency between ESMs with available observations. Climate model performance metrics have been widely used to objectively and quantitatively gauge the agreement between observations and simulations to summarize model behavior with a wide range of climate characteristics. Simple examples include either the model bias or the pattern similarity (correlation) between an observed and simulated field (e.g., Taylor, 2001). With the rapid growth in the number, scale, and complexity of simulations, the metrics have been used more routinely as exemplified by the Intergovernmental Panel on Climate Change (IPCC) Assessment Reports (e.g., Gates et al., 1995; McAvaney et al., 2001; Randall et al., 2007; Flato et al., 2014; Eyring et al., 2021). A few studies have been exclusively devoted to objective model performance assessment using summary statistics. Lambert and Boer (2001) evaluated the first set of CMIP models from CMIP1 using statistics for the large-scale mean climate. Gleckler et al. (2008) identified a variety of factors relevant to model metrics and demonstrated techniques to quantify the relative strengths and weaknesses of the simulated mean climate. Reichler and Kim (2008) attempted to gauge model improvements across the early phases of CMIP. The scope of objective model evaluation has greatly broadened beyond the mean state in recent years (e.g., Gleckler et al., 2016; Eyring et al., 2019), including attempts to establish performance metrics for a wide range of climate variability (e.g., Kim et al., 2009; Sperber et al., 2013; Ahn et al., 2017; Fasullo et al., 2020; Lee et al., 2021b; Planton et al., 2021) and extremes (e.g., Sillmann et al., 2013; Srivastava et al., 2020; Wehner et al., 2020, 2021). Guilyardi et al. (2009) and Reed et al. (2022) emphasized that metrics should be concise, interpretable, informative, and intuitive.

With the growth of data size and diversity of ESM simulations, there has been a pressing need for the research community to become more efficient and systematic in evaluating ESMs and documenting their performances. To respond to the need, PCMDI developed the PCMDI Metrics Package (PMP) and released its first version in 2015 (see Code and Data Availability section for all versions). A centralizing goal of the PMP then and now is to quantitatively synthesize results from the archive of CMIP simulations via performance metrics that help characterize the overall agreement between models and observations (Gleckler et al., 2016). For our purposes, "performance metrics" are typically (but not exclusively) well-established statistical measures that quantify the consistency between observed

and simulated characteristics. Common examples include a domain average bias, a root-mean-square error (RMSE), a spatial pattern correlation, or others, typically selected depending on the application. Another goal of the PMP is to further diversify the suite of high-level performance tests that help characterize the simulated climate. The results provided by the PMP are frequently used to address two overarching and recurring questions: 1) What are the relative strengths and weaknesses between different models? and 2) How are models improving with further development? Addressing the second question is often referred to as "benchmarking" and this motivates an important emphasis of the effort described in this paper—striving to advance the documentation of all data and results of the PMP in an open and ultimately reproducible manner.

In parallel, the current progress towards systematic model evaluation remains dynamic, with evolving approaches and many independent paths being pursued. This has resulted in the development of diversified model evaluation software packages. Examples in addition to the PMP include the ESMValTool (Eyring et al., 2016, 2019, 2020; Righi et al., 2020), the Model Diagnostics Task Force (MDTF) Diagnostics package (Maloney et al., 2019; Neelin et al., 2023), the International Land Model Benchmarking (ILAMB) Software System (Collier et al., 2018) that focuses on land surface and carbon cycle metrics, and the International Ocean Model Benchmarking (IOMB) Software System (Fu et al., 2022) that focuses on surface and upper ocean biogeochemical variables. Some tools have been developed with a more targeted focus on a specific subject area, such as the Climate Variability Diagnostics Package (CVDP) that diagnoses climate variability modes (Phillips et al., 2014; Fasullo et al., 2020), and the Analyzing Scales of Precipitation (ASoP) that focuses on analyzing precipitation scales across space and time (Klingaman et al., 2017; Martin et al., 2017; Ordonez et al., 2021). The regional climate community also has actively developed metrics packages such as the Regional Climate Model Evaluation System (RCMES; Lee et al., 2018a; Whitehall et al. 2012). Separately, a few climate modeling centers have developed their own model evaluation packages to assist in their in-house ESM development, e.g., the E3SM Diags (Zhang et al., 2022). There also have been other efforts to enhance the usability of in-situ and field campaign observations in ESM evaluations, such as Atmospheric Radiation Measurement (ARM) GCM Diag (Zhang et al., 2018, 2020) and Earth System Model Aerosol–Cloud Diagnostics (ESMAC Diags; Tang et al., 2022, 2023). While they all have their own scientific priorities and technical approaches, the uniqueness of the PMP is its focus on the objective characterization of the physical climate system as simulated by community models. An important prioritization of the PMP is to advance all aspects of its workflow, in an open, transparent, and reproducible manner, which is critical for benchmarking. The PMP summary statistics characterizing CMIP simulations are version-controlled and made publicly available as a resource to the community.

In this paper, we describe the latest update of the PMP and its focus on providing a diverse suite of summary statistics that can be used to construct "quick-look" summaries of ESM performance from simulations made publicly available to the research community, notably CMIP. The rest of the paper is organized as follows. In section 2, we provide a technical description of the PMP and its accompanying reference datasets. In section 3, we describe various sets of simulation metrics that provide an increasingly comprehensive portrayal of physical processes across time scales ranging from hours to centurial. In section 4, we introduce the usage of PMP for model benchmarking. We discuss the future direction and the remaining challenges in section 5 and conclude with a summary in section 6. To assist the reader, the table in Appendix A summarizes the acronyms used in this paper.

2 Software package and data description

The PMP is a Python-based open-source software framework (https://github.com/PCMDI/pcmdi_metrics) designed to objectively gauge the consistency between ESMs and available observations via well-established statistics such as those discussed in Section 3. The PMP has been mainly used for the evaluation of CMIP-participating models. A subset of CMIP experiments, those conducted using the observation forcings such as "Historical" and "AMIP" (Eyring et al., 2016), is particularly well suited for comparing models with observations. The AMIP experiment protocol constrains the simulation with prescribed sea surface temperature (SST), and the "Historical" experiment is conducted using coupled model simulations driven by observed varying natural and anthropogenic forcings. Some of the metrics applicable to these experiments may also be relevant to others (e.g., multi-century coupled control runs called "PiControl" and idealized "4xCO2" simulations that are designed for estimating climate sensitivity).

The PMP has been applied to multiple generations of CMIP models in a quasi-operational fashion as new simulations are made available, new analysis methods are incorporated, or new observational data become accessible (e.g., Gleckler et al. 2016; Planton et al., 2021; Lee et al., 2021b; Ahn et al. 2022). Shortly after simulations from the most recent phase of the CMIP (i.e., CMIP6) became accessible, PMP quick-look summaries were provided on the PCMDI's website (https://pcmdi.llnl.gov/metrics/), offering a resource to scientists involved in CMIP or others interested in the evaluation of ESMs. To facilitate this, at PCMDI the PMP is technically linked to the Earth System Grid Federation (ESGF) that is the CMIP data delivery infrastructure (Williams et al., 2016).

The primary deliverable of the PMP is a collection of summary statistics. We strive to make the baseline results (raw statistics) publicly available and well-documented, and continue to make advances with this priority. For our purposes, we are referring to model performance "summary statistics" and "metrics" interchangeably, although in some situations we consider there to be an important distinction. For us, a genuine performance metric constitutes a well-defined and established statistic that has been used in a very specific way (e.g., a particular variable, analysis, and domain) for long-term benchmarking (see Section 4). The distinction between summary statistics and metrics is application-dependent and evolving as the community advances efforts to establish quasi-operational capabilities to gauge ESM performance. Some visualization capabilities described in Section 3 are made available through the PMP. Users can also further explore the model data comparisons using their preferred visualization methods or incorporate the results into their own studies from the summary statistics from the PMP. Noting the above, the scope of the PMP is fairly targeted. It is not intended to be "all-purpose", e.g. by incorporating the vast range of diagnostics used in model evaluation.

The PMP is designed to readily work with model output that has been processed using the Climate Model Output Rewriter (CMOR; https://cmor.llnl.gov/), which is a software library developed to prepare model output following the CF Metadata Conventions (Hassell et al., 2017; Eaton et al., 2022, http://cfconventions.org/) in Network Common Data Form (NetCDF) format. The CMOR is used by most modeling groups contributing to CMIP, ensuring all model output adheres to the CMIP data structures that themselves are based on the CF conventions. It is possible to use the PMP on model output that has not been prepared by CMOR, but this usually requires additional work, e.g., mapping the data to meet the community standards.

219   For reference datasets, the PMP uses observational products processed to be compliant with the Observations
220   for Model Intercomparison Projects (obs4MIPs; https://pcmdi.github.io/obs4MIPs/). The obs4MIPs effort was
221   initiated circa 2010 (Gleckler et al., 2011) to advance the use of the observations in model evaluation and research.
222   Substantial progress has been made in establishing obs4MIPs data standards that technically align with CMIP model
223   output (e.g., Teixeira et al., 2014; Ferraro et al., 2015), with the data products published on the ESGF (Waliser et al.,
224   2020). Obs4MIPs-compliant data were prepared with CMOR, and the data directly available via obs4MIPs are used
225   as PMP reference datasets.

226   The PMP leverages other Python-based open-source tools and libraries, such as *xarray* (Hoyer and Hamman,
227   2017), *eofs* (Dawson, 2016), and many others. One of the primary fundamental tools used in the latest PMP version
228   is the Python package, Xarray Climate Data Analysis Tools (xCDAT; Vo et al., 2023; https://xcdat.readthedocs.io).
229   The xCDAT is developed to provide a more efficient, robust, and streamlined user experience in climate data analysis
230   when using xarray (https://docs.xarray.dev/). Portions of the PMP rely on the precursor of the xCDAT, a Python
231   library called Community Data Analysis Tools (CDAT, Williams et al., 2009; Williams, 2014; Doutriaux et al., 2019),
232   which has been fundamental since the early development stages of the PMP. The xarray software provides much of
233   the functionality of CDAT (e.g., I/O, indexing, and subsetting). However, it lacks some key climate domain features
234   that have been frequently used by scientists and exploited by the PMP (e.g., regridding, utilization of spatial/temporal
235   bounds for computational operations) which motivated the development of the xCDAT. Completing the transition
236   from CDAT to xCDAT is a technical priority for the next version of PMP.

237   To help advance open and reproducible science, the PMP has been maintained with an open-source policy
238   with accompanying metadata for data reproducibility and reusability. The PMP code is distributed and released with
239   version control. The installation process of PMP is streamlined and user-friendly, leveraging the *Anaconda* distribution
240   and the *conda-forge* channel. By employing *conda* and *conda-forge*, users benefit from a simplified and efficient
241   installation experience, ensuring seamless integration of PMP's functionality with minimal dependencies. This
242   approach not only facilitates a straightforward deployment of the package but also enhances reproducibility and
243   compatibility across different computing environments, thereby facilitating the accessibility and widespread adoption
244   of PMP within the scientific community. The pointer to the installation instructions can be found in the Code and Data
245   Availability section. The PMP's online documentation (http://pcmdi.github.io/pcmdi_metrics/) also includes
246   installation instructions and user demo Jupyter Notebooks. We also release a database of pre-calculated PMP statistics
247   for all AMIP and Historical simulations in the CMIP archive are also available online. The archive of these statistics
248   stored as JSON files (Crockford, 2006; Crockford and Morningstar, 2017) includes versioning details for all codes,
249   and dependencies and data that were used for the calculations. These files provide the baseline results of the PMP (See
250   the Code and Data Availability section for details). Advancements in model evaluation along with the number of
251   models and complexity of simulations motivate more systematic documentation of performance summaries. With
252   PMP workflow provenance information being recorded and the model and observational data standards maintained
253   by PCMDI and colleagues, PMP strives to make all its results reproducible.
254

3 Current PMP capabilities

The capabilities of the PMP have been expanded beyond its traditional large-scale performance summaries of the mean climate (Gleckler et al., 2008; Taylor, 2001). Various evaluation metrics have been implemented to the PMP for climate variability such as El Niño-Southern Oscillation (ENSO) (Planton et al., 2021; Lee et al., 2021a), extratropical modes of variability (Lee et al., 2019, 2021b), intra-seasonal oscillation (Ahn et al., 2017), monsoons (Sperber and Annamalai, 2014), cloud feedback (Zelinka et al., 2022), and the characteristics of simulated precipitation (Pendergrass et al., 2020; Ahn et al., 2022, 2023) and extremes (Wehner et al., 2020, 2021). These PMP capabilities were built upon model performance tests that have resulted from research by PCMDI scientists and their collaborators. This section will provide an overview of each category of the current PMP evaluation metrics with their usage demonstrations.

3.1 Climatology

Mean state metrics quantify how well models simulate observed climatological fields at a large scale, gauged by a suite of well-established statistics such as RMSE, mean absolute error (MAE), and pattern correlation that have been used in climate research for decades. The focus is on the coupled "Historical" and atmospheric-only AMIP (Gates et al., 1999) simulations which are well-suited for comparison with observations. The PMP extracts seasonally and annually averaged fields of multiple variables from large-scale observationally based datasets and results from model simulations. Different obs4MIPs-compliant reference datasets are used depending on the variable examined. When multiple reference datasets are available, one of them is considered as a "default" (e.g., see Table 1) while others are identified as "alternatives". The default datasets are typically state-of-the-art products, but in general, we lack definitive measures as to which is the most accurate, so the PMP metrics are routinely calculated with multiple products so that it can be determined what difference the selection of alternative observations makes to judgment made about model fidelity. The suite of mean climate metrics (all area weighted) includes spatial and spatiotemporal RMSE, centered spatial RMSE, spatial-mean bias, spatial standard deviation, spatial pattern correlation, and spatial and spatiotemporal MAE of the annual or seasonal climatological time-mean (Gleckler et al., 2008). Often, a space-time statistic is used that gauges both the consistency of the observed and simulated climatological pattern as well as its seasonal evolution (see Eq. 1 from Gleckler et al., 2008). By default, results are available for selected large-scale domains, including: "Global", "Northern Hemisphere (NH) Extratropics" (30ºN-90ºN), "Tropics" (30ºS-30ºN), and "Southern Hemisphere (SH) Extratropics" (30ºS-90ºS). For each domain, results can also be computed for the land and ocean, land only, or ocean only. These commonly used domains highlight the application of the PMP mean climate statistics at large to global scales, but we note that PMP allows users to define their own domains of interest, including at regional scales. Detailed instructions can be found on the PMP's online documentation (http://pcmdi.github.io/pcmdi_metrics).

Although the primary deliverable of the PMP is the metrics, these PMP results can be visualized in various ways. For individual fields, we often first plot Taylor Diagrams, a polar plot leveraging the relationship between the centered RMSE, the pattern correlation, and the observed and simulated standard deviation (Taylor, 2001). The Taylor Diagram has become a standard plot in the model evaluation workflow across modeling centers and research

329 communities (see Section 5). To interpret results across CMIP models for many variables, we routinely construct
330 normalized Portrait Plots or Gleckler Plots (Gleckler et al., 2008) that provide a quick-look examination of the
331 strengths and weaknesses of different models. For example, in Figure 1, the PMP results display quantitative
332 information of simulated seasonal climatologies of various meteorological model variables via a normalized global
333 spatial RMSE (Gleckler et al., 2008). Variants of this plot have been widely used for presenting model evaluation
334 results, for example, in the IPCC Fifth (Flato et al., 2014, Figures 9.7, 9.12, and 9.37) and Sixth Assessment Reports

335 (Eyring et al., 2021, Chapter 3, Figure 3.42). Because the error distribution across models is variable dependent, the
336 statistics are often normalized to help reveal differences, in this case via the median RMSE across all models (see
337 Gleckler et al. 2008 for more details). This normalization enables a common color scale to be used for all statistics on
338 the Portrait Plot, highlighting the relative strengths and weaknesses of different models. In this example (Fig. 1), an
339 error of -0.5 indicates that a model's error is 50% smaller than the typical (median) error across all models, whereas
340 an error of 0.5 is 50% larger than the typical error in the multi-model ensemble. In many cases, the horizontal bands
341 in the Gleckler plots show that simulations from a given modeling center have similar error structures relative to the
342 multi-model ensemble.
343   The Parallel Coordinate Plot (Inselberg, 1997, 2008, 2016; Johansson and Forsell, 2016) that retains the
344 absolute value of the error statistics is used to complement the Portrait plot. Some previous studies have utilized
345 Parallel Coordinate Plots for analyzing climate model simulations (e.g., Steed et al., 2012; Wong et al., 2014; Wang
346 et al., 2017), but to date, only a few studies have applied it to collective multi-ESM evaluations (e.g., see Fig. 7 of
347 Boucher et al., 2020). In the PMP, we generally construct Parallel Coordinate Plots using the same data as in a portrait
348 plot. However, a fundamental difference is that metrics values can be more easily scaled to highlight absolute values
349 rather than the normalized relative results of the portrait plot. In this way, the Portrait and Parallel Coordinate plots
350 complement one another, and in some applications, it can be instructive to display both. Figure 2 shows the
351 spatiotemporal RMSE, defined as the temporal average of spatial RMSE calculated in each month of the annual cycle,
352 of CMIP5 and CMIP6 models in the format of Parallel Coordinate Plot. Each vertical axis represents a different scalar
353 measure gauging a distinct aspect of model fidelity. While polylines are frequently used to connect data points from
354 the same source (i.e., metric values from the same model, in our case) in Parallel Coordinate Plots, we display results
355 from each model using an identification symbol to reduce visual clutter on the plot and help identify outlier models.
356 In the example of Fig. 2, each vertical axis is aligned with the median value midway through its max/min range scale.
357 Thus, for each axis, the models in the lower half of the plot perform better than the CMIP5-CMIP6 multi-model
358 median, while in the upper half, the opposite is true. For each vertical axis that is for a different model variable, we
359 have added violin plots (Hintze and Nelson, 1998) to show probability density functions representing the distributions
360 of model performance obtained from CMIP5 (shaded in blue, left side of the axis) and CMIP6 (shaded in orange, right
361 side of the axis). Medians of each CMIP5 and CMIP6 group are highlighted using polylines, which indicates that the
362 RMSE is reduced in CMIP6 relative to CMIP5 in general for the majority of the subset of model variables.
363

3.2 El Niño-Southern Oscillation

366

367   The El Niño-Southern Oscillation (ENSO) is Earth's dominant interannual mode of climate variability, which
368 impacts global climate via both regional oceanic effects and far-reaching atmospheric teleconnections (McPhaden et
369 al., 2006, 2020). In response to increasing interest in a community approach to ENSO evaluation in models (Bellenger
370 et al., 2014), the International Climate and Ocean Variability, Predictability and Change (CLIVAR) Research Focus
371 on ENSO in a Changing Climate, together with the CLIVAR Pacific Region Panel, developed the CLIVAR ENSO
372 Metrics Package (Planton et al., 2021) which is now utilized within the PMP. The ENSO metrics used to
373 assess/evaluate the models are grouped into three categories: *Performance* (i.e., background climatology and basic
374 ENSO characteristics), *Teleconnections* (ENSO's worldwide teleconnections), and *Processes* (ENSO's internal
375 processes and feedback). Planton et al. (2021) found that CMIP6 models generally outperform CMIP5 models in
376 several ENSO metrics in particular for those related to tropical Pacific seasonal cycles and ENSO teleconnections.
377 This effort is discussed in more detail in Planton et al. (2021), and detailed descriptions of each metric in the package
378 are available in the ENSO Package online open-source code repository on its GitHub Wiki pages (see
379 https://github.com/CLIVAR-PRP/ENSO_metrics/wiki).

380   Figure 3 demonstrates the application of the ENSO metrics to CMIP6, showing the magnitudes of inter-
381 model and inter-ensemble spreads, along with observational uncertainty varying across metrics. For a majority of the
382 ENSO Performance metrics model error and inter-model spread are substantially larger than observational uncertainty
383 (Figs. 3a-n). This highlights the systematic biases like the double intertropical convergence zone (ITCZ) (Fig. 3a) that
384 are persisting through CMIP phases (Tian and Dong, 2020). Similarly, ENSO Processes metrics (Figs. 3t-w) indicate
385 large errors in the feedback loops generating SST anomalies, indicating a different balance of processes in the model
386 and in the reference and possibly compensating errors (Bayr et al., 2019, Guilyardi et al. 2020). In contrast, for ENSO
387 Teleconnection metrics, the observational uncertainty is substantially larger, thus challenging validation of model
388 error (Figs. 3o-r). For some metrics, such as the ENSO duration (Fig. 3f), the ENSO Asymmetry metric (Fig. 3i), and
389 the Ocean driven SST metric (Fig. 3s), there are larger inter-ensemble spreads than the inter-model spreads. From
390 such results, Lee et al. (2021a) examined the inter-model and inter-member spread of these metrics from the large
391 ensembles available from CMIP6 and the US CLIVAR Large Ensemble Working Group. They argued that to robustly
392 characterize baseline ENSO characteristics and physical processes, larger ensemble sizes are needed, compared to
393 existing state-of-the-art ensemble projects. By applying the ENSO metrics to historical and piControl simulations of
394 CMIP6 via the PMP, Planton et al. (2023) developed equations based on statistical theory to estimate the required
395 ensemble size for a user-defined uncertainty range.

396

3.3 Extratropical Modes of Variability

397

398   The PMP includes objective measures of the pattern and amplitude of extratropical modes of variability from
399 PCMDI's research, which has expanded beyond its traditional large-scale performance summaries to include
400 interannual variability, considering increasing interest in setting an objective approach for the collective evaluation of
401 multiple modes. Extratropical modes of variability (ETMoV) metrics in the PMP were developed by Lee et al. (2019a)
402 that stem from earlier works (e.g., Stoner et al., 2009; Phillips et al., 2014). Lee et al. (2019a) illustrated a challenge

405 when evaluating modes of variability using the traditional empirical orthogonal functions (EOF). In particular, when
406 a higher-order EOF of a model more closely corresponds to a lower-order observationally based EOF (or vice versa),
407 it can significantly affect conclusions drawn about model performance. To circumvent this issue in evaluating the
408 interannual variability modes, Lee et al. (2019a) used the Common Basis Function (CBF) approach that projects the
409 observed EOF pattern onto model anomalies. This approach has been previously applied for the evaluation of
410 intraseasonal variability modes (Sperber, 2004; Sperber et al., 2005). In the PMP, the CBF approach is taken as a
411 default method, and the traditional EOF approach is also enabled as an option for the ETMoV metrics calculations.

412 The ETMoV metrics in the PMP measure simulated patterns and amplitudes of ETMoV, and quantify their
413 agreement with observations (e.g., Lee et al., 2019a, 2021b). The PMP's ETMoV metrics evaluate 5 atmospheric
414 modes – the Northern Annular Mode (NAM), North Atlantic Oscillation (NAO), Pacific North America pattern
415 (PNA), North Pacific Oscillation (NPO), and Southern Annular Mode (SAM), and 3 ocean modes diagnosed by the
416 variance of sea-surface temperature – Pacific Decadal Oscillation (PDO), North Pacific Gyre Oscillation (NPGO),
417 and Atlantic Multi-decadal Oscillation (AMO). The AMO is included for experimental purposes, considering the
418 significant uncertainty in detecting the AMO (Deser and Philips 2021; Zhao et al., 2022). The amplitude metric,
419 defined as the ratio of standard deviations of the model and observed principal components, has been used to examine
420 the evolution of the performance of models across different CMIP generations (Fig. 4). Green shading predominates,
421 indicating where the simulated amplitude of variability is similar to observations. In some cases, such as for
422 SAM_SON, the models overestimate the observed amplitude.

423 The PMP's ETMoV metrics have been used in several model evaluation studies. For example, Orbe et al.
424 (2020) analyzed models from U.S. climate modeling groups including the U.S. Department of Energy (DOE), National
425 Aeronautics and Space Administration (NASA), National Center for Atmospheric Research (NCAR), and National
426 Oceanic and Atmospheric Administration (NOAA), where they found that the improvement in the ETMoV
427 performance is highly dependent on mode and season, when comparing across different generations of those models.
428 Sung et al. (2021) examined the performance of models run at the Korea Meteorological Administration (K-ACE and
429 UKESM1) in reproducing ETMoVs from their Historical simulations, and concluded that these models reasonably
430 capture most ETMoVs. Lee et al. (2021b) collectively evaluated ~130 models from CMIP3, 5, and 6 archive databases
431 using their ~850 Historical and ~300 AMIP simulations, where they found the spatial pattern skill improved in CMIP6
432 compared to CMIP5 or CMIP3 for most modes and seasons, while the improvement in amplitude skill is not clear.
433 Arcodia et al. (2023) used the PMP to derive PDO and AMO to investigate their role in decadal variability of
434 subseasonal predictability of precipitation over the western coast of North America and concluded that no significant
435 relationship was found.

436

437 *3.4 Intraseasonal Oscillation*

438 The PMP has implemented metrics for the Madden-Julian Oscillation (MJO; Madden and Julian, 1971, 1972,
439 1994). The MJO is the dominant mode of tropical intraseasonal variability, characterized by a pronounced eastward
440 propagation of large-scale atmospheric circulation coupled with convection with a typical periodicity of 30-60 days.

Moved down [10]: Kim et al., 2020

Selected metrics from the MJO diagnostics package, developed by the CLIVAR MJO Working Group (Waliser et al., 2009), have been implemented in the PMP following Ahn et al. (2017).

We have particularly focused on metrics for the MJO propagation: East/West power Ratio (EWR) and East power normalized by Observation (EOR). The EWR is proposed by Zhang and Hendon (1997) which is defined as the ratio of the total spectral power over the MJO band (eastward propagating, wavenumber 1-3 and period of 30-60 days) to that of its westward propagating counterpart in the wavenumber-frequency power spectra. The EWR metric has been widely used in the community, to examine the robustness of the eastward propagating feature of the MJO (e.g., Hendon et al., 1999; Lin et al., 2006; Kim et al., 2009; Ahn et al., 2017). The EOR is formulated by normalizing a model's spectral power within the MJO band by the corresponding observed value. Ahn et al. (2017) showed EWRs and EORs of the CMIP5 models. Using daily precipitation, the PMP calculates EWR and EOR separately for boreal winter (November to April) and boreal summer (March to October). We apply the frequency-wavenumber decomposition method to precipitation from observations (GPCP-based; 1997-2010) and the CMIP5 and CMIP6 Historical simulations for 1985-2004. For disturbances with wavenumbers 1-3 and frequencies corresponding to 30-60 days, it is clear in observations that the eastward propagating signal dominates over its westward propagating counterpart with an EWR value of about 2.49 (Fig. 5a). Figure 5b shows the wavenumber-frequency power spectrum from CMIP5 IPSL-CM5B-LR as an example, which has an EWR value that is comparable to the observed value.

Figure 6 shows the EWR from individual models' multiple ensemble members and their average. The average EWR of the CMIP6 model simulations is more realistic than that of the CMIP5 models. Interestingly, a substantial spread exists across models and also among ensemble members of a single model. For example, while the average EWR value for the CESM2 ensemble is 2.47 (close to 2.49 from the GPCP observations), the EWR values of the individual ensemble members range from 1.87 to 3.23. Kang et al. (2020) suggested that the ensemble spread in the propagation characteristics of the MJO can be attributed to the differences in the moisture mean state, especially its meridional moisture gradient. A cautionary note should be given to the fact that the MJO frequency and wavenumber windows are chosen to capture the spectral peak in observations. Thus, while the EWR provides an initial evaluation of the propagation characteristics of the observed and simulated MJO, it is instructive to look at the frequency-wavenumber spectra, as in some cases the dominant periodicity and wavenumber in a model may be different than in observations. It is worthwhile to note that the PMP can be used to obtain EWR and EOR of other daily variables for MJO analysis, such as outgoing longwave radiation (OLR) or zonal wind at 850 hPa (U-850) or 250 hPa (U-250), as shown in Ahn et al. (2017).

3.5 Monsoons

Based on the work of Sperber and Annamalai (2014), skill metrics in the PMP quantify how well models represent the onset, decay, and duration of regional monsoons. From observations and Historical simulations, the climatological pentads of precipitation are area-averaged for six monsoon-related domains: All-India Rainfall, Sahel, Gulf of Guinea, North American Monsoon, South American Monsoon, and Northern Australia, as seen in Fig. 7. For the domains in the Northern Hemisphere, the 73 climatological pentads run from January to December, while for the domains in the Southern Hemisphere, the pentads run from July to June. For each domain, the precipitation is

495 accumulated at each subsequent pentad and then divided by the total precipitation to give the fractional accumulation
496 of precipitation as a function of pentad. Thus, the annual cycle behavior is evaluated irrespective of whether a model
497 has a dry or wet bias. Except for the Gulf of Guinea, the onset and decay of monsoon occur for a fractional
498 accumulation of 0.2 and 0.8, respectively. Between these fractional accumulations, the accumulation of precipitation
499 is nearly linear as the monsoon season progresses. Comparison of the simulated and observed onset, duration, and
500 decay are presented in terms of the difference in the pentad index obtained from the model and observations (i.e.,
501 model minus observations). Therefore, negative values indicate that the onset or decay in the model occurs earlier
502 than in observations, while positive values indicate the opposite. For duration, negative values indicate that for the
503 model it takes fewer pentads to progress from onset to decay compared to observations (i.e., the simulated monsoon
504 period is too short), while positive values indicate the opposite.

505 For CMIP5, we find systematic errors in the phase of the annual cycle of rainfall. The models are delayed in
506 the onset of summer rainfall over India, the Gulf of Guinea, and the South American Monsoon, with early onset
507 prevalent for the Sahel and the North American Monsoon. The lack of consistency in the phase error across all domains
508 suggests that a ''global'' approach to the study of monsoons may not be sufficient to rectify the regional differences.
509 Rather, regional process studies are necessary for diagnosing the underlying causes of the regionally specific
510 systematic model biases over the different monsoon domains. Assessment of the monsoon fidelity in CMIP6 models
511 using the PMP is in progress.

512

3.6 Cloud feedback and mean-state

514 Uncertainties in cloud feedback are the primary driver of model-to-model differences in climate sensitivity
515 – the global temperature response to a doubling of atmospheric CO_2. Recently, an expert synthesis of several lines of
516 evidence spanning theory, high-resolution models, and observations was conducted to establish quantitative
517 benchmark values (and uncertainty ranges) for several key cloud feedback mechanisms. The assessed feedbacks are
518 those due to changes in high-cloud altitude, tropical marine low-cloud amount, tropical anvil cloud area, land cloud
519 amount, middle latitude marine low-cloud amount, and high latitude low-cloud optical depth. The sum of these six
520 components yields the total assessed cloud feedback, which is part of the overall radiative feedback that fed into the
521 Bayesian calculation of climate sensitivity in Sherwood et al. (2020). Zelinka et al. (2022) estimated these same
522 feedback components in climate models and evaluated them against the expert-judgment values determined in
523 Sherwood et al. (2020), ultimately deriving a root mean square error metric that quantifies the overall match between
524 each model's cloud feedback and those determined through expert judgment.

525 Figure 8 shows the model-simulated values for each individual feedback computed in *amip-p4K* simulations
526 as part of CMIP5 and CMIP6 alongside the expert judgment values. Each model is color-coded by its equilibrium
527 climate sensitivity (determined using *abrupt-4xCO2* simulations as described in Zelinka et al., 2020), and the values
528 from an illustrative model (GFDL-CM4) are highlighted. Among the key results apparent from this figure is that
529 models typically underestimate the strength of both positive tropical marine low-cloud feedback and the negative anvil
530 cloud feedback relative to the central expert assessed value. The sum of all six assessed feedback components is

533 positive in all but two models, with a multimodel mean value that is close to the expert-assessed value, but exhibits
534 substantial intermodel spread.

535 In addition to evaluating the ability of models to match the assessed cloud feedback components, Zelinka et
536 al. (2022) investigated whether models with less erroneous mean-state clouds tend to have smaller errors in their
537 overall cloud feedback RMSE. This involved computing the mean-state cloud property error metric developed by
538 Klein et al. (2013). This error metric quantifies the spatiotemporal error in climatological cloud properties for clouds
539 with optical depths greater than 3.6, weighted by their net top-of-atmosphere (TOA) radiative impact. The
540 observational baseline against which the models are compared comes from the International Satellite Cloud
541 Climatology Project H-series Gridded Global (ISCCP HGG) dataset (Young et al., 2018). Zelinka et al. (2022) showed
542 that models with smaller mean-state cloud errors tend to have stronger but not necessarily better (less erroneous) cloud
543 feedback, which suggests that improving mean-state cloud properties does not guarantee improvement in the cloud
544 response to warming. However, the models with the smallest errors in cloud feedback tend also to have less erroneous
545 mean-state cloud properties, and no models with poor mean-state cloud properties have feedback in good agreement
546 with expert judgment.

547 The PMP implementation of this code computes cloud feedback by differencing fields from *amip-p4K* and
548 *amip* experiments and normalizing by the corresponding global mean surface temperature change rather than from
549 differencing *abrupt-4xCO2* and *piControl* experiments and computing feedback via regression (as was done in Zelinka
550 et al., 2022). This choice is made to reduce the computational burden and also because cloud feedbacks derived from
551 these simpler atmosphere-only simulations have been shown to closely match those derived from fully coupled
552 quadrupled CO_2 simulations (Qin et al., 2022). The code produces figures in which the user-specified model results
553 are highlighted and placed in the context of the CMIP5 and CMIP6 multi-model results (e.g., Figure 8).

554

555 *3.7 Precipitation*

556 Recognizing the importance of accurately simulating precipitation in ESMs and a lack of objective and
557 systematic benchmarking for it, and motivated by discussions with WGNE and WGCM working groups of WCRP,
558 the DOE has initiated an effort to establish a pathway to help modelers gauge improvement (U.S. DOE, 2020). The
559 2019 DOE workshop "Benchmarking Simulated Precipitation in Earth System Models" generated two sets of
560 precipitation metrics: *baseline* and *exploratory* metrics (Pendergrass et al., 2020). In the PMP, we have focused on
561 implementing the *baseline* metrics for benchmarking simulated precipitation. In parallel, a set of *exploratory* metrics
562 that could be added to metrics suites including PMP in the future was illustrated by Leung et al. (2022) to extend the
563 evaluation scope to include process-oriented and phenomena-based diagnostics and metrics.

564 The *baseline* metrics gauge the consistency between ESMs and observations, focusing on the holistic set of
565 observed rainfall characteristics (Fig. 9). For example, the spatial distribution of mean state precipitation and seasonal
566 cycle are outcomes of the PMP's Climatology metrics (described in Section 3.1), which provides collective evaluation
567 statistics such as RMSE, standard deviation, and pattern correlation over various domains (e.g., global, NH and SH
568 extratropics, and Tropics, with each domain as a whole, and over land and ocean, in separate). Evaluation of
569 precipitation variability across many timescales with PMP is documented in Ahn et al. (2022); we summarize some

of the findings here. The precipitation variability metric measures forced (diurnal and annual cycles) and internal variability across timescales (subdaily, synoptic, subseasonal, seasonal, and interannual) in a framework based on power spectra of 3-hourly total and anomaly precipitation. Overall, CMIP5 and CMIP6 models underestimate the internal variability, which is more pronounced in the higher frequency variability, while they overestimate the forced variability (Fig. 10). For the diurnal cycle, PMP includes metrics from Covey et al. (2016). Additionally, the intensity and distribution of precipitation are assessed following Ahn et al. (2023). Extreme daily precipitation indices and their 20-year return values are calculated using a non-stationary Generalized Extreme Value statistical method. From the CMIP5 and CMIP6 historical simulations we evaluate model performance of these indices and their return values in comparison with gridded land-based daily observations. Using this approach, Wehner et al. (2020) found that at models' standard resolutions, no meaningful differences were found between the two generations of CMIP models. Wehner et al. (2021) extended the evaluations of simulated extreme precipitation to seasonal 3-hourly precipitation extremes produced by available HighResMIP models and concluded that the improvement is minimal with the models' increased spatial resolutions. They also noted that the order of operations of regridding and calculating extremes affects the ability of models to reproduce observations. Drought metrics developed by Xue and Ullrich (2021) are not implemented in PMP directly, but are wrapped by the Coordinated Model Evaluation Capabilities (CMEC; Ordonez et al. 2021), which is a parallel framework for supporting community-developed evaluation packages. Together, these metrics provide a streamlined workflow for running the entire baseline metrics via the PMP and CMEC that is ready for use by operational centers and in the CMIP7.

3.8 Relating metrics to underlying diagnostics

Considering the extensive collection of information generated from the PMP, efforts have supported improved visualizations of metrics using interactive graphic user interfaces. These capabilities can facilitate the interpretation and synthesis of vast amounts of information associated with the diverse metrics and the underlying diagnostics from which they were derived. Via the interactive navigation interface, we can explore the underlying diagnostics behind the PMP's summary plots. On the PCMDI website, we provide interactive graphical interfaces to enable navigating the supporting plots to the underlying diagnostics of each model's ensemble members and their average. For example, on the interactive mean climate plots (https://pcmdi.llnl.gov/metrics/mean_clim/), hovering the mouse cursor over a square or triangle in the Portrait Plot, or over the markers or lines in the Parallel Coordinate Plot, reveals the diagnostic plot from which the metrics were generated. It allows the user to toggle between several metrics (e.g., RMSE, bias, and correlation) and regions (e.g., global, Northern/Southern Hemisphere, and Tropics), along with relevant provenance information. Users can click on the interactive plots to get dive-down diagnostics information for the model of interest which provides detailed analysis to better understand how the metric was calculated. As with the PMP's mean climate metrics output, we currently provide interactive summary graphics for ENSO (https://pcmdi.llnl.gov/metrics/enso/), extratropical modes of variability (https://pcmdi.llnl.gov/metrics/variability_modes/), monsoon (https://pcmdi.llnl.gov/metrics/monsoon/), MJO (https://pcmdi.llnl.gov/metrics/mjo/), and precipitation benchmarking (https://pcmdi.llnl.gov/metrics/precip/). We plan to expand this capability to other metrics in the PMP, such as the cloud feedback analysis. The majority of the

608 PMP's interactive plots have been developed using Bokeh (https://bokeh.org/), a Python data visualization library that
609 enables the creation of interactive plots and applications for web browsers.
610

611 **4 Model Benchmarking**
612   While the PMP originally focused on evaluating multiple models (e.g., Gleckler et al., 2008), in parallel there
613 has been increasing interest from model developers and modeling centers to leverage the PMP to track performance
614 evolution in the model development cycle, as discussed in Gleckler et al. (2016). For example, metrics from the PMP
615 have been used to document performance of ESMs developed in the U.S. DOE Exascale Earth System Model (E3SM;
616 Caldwell et al., 2019; Golaz et al., 2019; Rasch et al., 2019; Hannah et al., 2021; Tang et al., 2021), NOAA
617 Geophysical Fluid Dynamics Laboratory (GFDL; Zhao et al., 2018), Institut Pierre-Simon Laplace (IPSL; Boucher et
618 al., 2020; Planton et al., 2021), National Institute of Meteorological Sciences-Korea Meteorological Administration
619 (NIMS-KMA; Sung et al., 2021), University of California, Los Angeles (Lee et al., 2019b), and the Community
620 Integrated Earth System Model (CIESM) project (Lin et al., 2020).
621   To make the PMP more accessible and useful for modeling groups, efforts are underway to broaden workflow
622 options. Currently, a typical application involves computing a particular class of performance metrics (e.g., mean
623 climate) for all CMIP simulations available via ESGF. To facilitate the ability of modeling groups to routinely use the
624 PMP during their development process, we are working to provide a customized workflow option to run all the PMP
625 metrics more seamlessly on a single model, and to compare these results with a database of PMP results obtained from
626 CMIP simulations (see Code and Data Availability section). Via the PMP-documented and pre-calculated metrics
627 from simulations in the CMIP archive, it is possible to readily incorporate CMIP results into the assessment of new
628 simulations, without retrieving all CMIP simulations and recomputing the results. The resulting quick-look feedback
629 can highlight model improvement (or deterioration) and can assist in determining development priorities or in the
630 selection of a new model version.
631   As an example, here, we show PMP results obtained from GFDL-CM3 from CMIP5 and GFDL-CM4 from
632 CMIP6, for a demonstration of using the Taylor Diagram to compare versions of a given model (Fig. 11). One
633 advantage of the Taylor Diagram is that it collectively represents three statistics (i.e., centered RMSE, standard
634 deviation, and correlation) in a single plot (Taylor, 2001), which synthesizes the performance intercomparison of
635 multiple models (or different versions of a model). In this example, four variables were selected to summarize
636 performance evolution (shown by arrows) in multiple seasons. Except for boreal winter, both model versions are
637 nearly identical in terms of net TOA radiation, however in all seasons the longwave cloud radiative effect is clearly
638 improved in the newer model version. The TOA flux improvements likely contributed to the precipitation
639 improvements, by improving the balances of radiative cooling and latent heating. The improvement in the newer
640 model version is consistent with that documented by Held et al., (2019) and evident via the arrow directions pointing
641 to the observational reference point.
642   Parallel Coordinate Plots can also be used to summarize the comparison of two simulations for their
643 performance. In Fig 12, we demonstrate the comparison of selected metrics: the mean climate (see Section 3.1), ENSO
644 (Section 3.2), and ETMoV (Section 3.3). To facilitate comparison of a subset of models, a few models can be selected

and highlighted as connected lines across individual vertical axes on the plot. A proposed application of it from PMP is to select two models or two versions of a model to contrast their performance (solid lines) against the backdrop of results from other models, shown as violin plots for the distribution of statistics from other models on each vertical axis. In this example, we contrast the performance of two GFDL models: GFDL-CM3 and GFDL-CM4. Fig 12a is a modified version of Figure 2 that is designed to highlight the difference in performance more efficiently. Each vertical axis indicates performance for each metric defined for climatology of variables (i.e., temporally averaged spatial RMSE of annual cycle climatology patterns, Fig. 12a), ENSO characteristics (Fig. 12b), or interannual variability mode obtained from seasonal or monthly averaged time series (Fig. 12c). It is shown that GFDL-CM4 is superior to GFDL-CM3 for most cases across selected metrics (downward arrows in green) while inferior for a few cases (upward arrows in red), which is consistent with previous findings (Held et al., 2019; Planton et al., 2021; Chen et al., 2021). Such applications of the Parallel Coordinate Plot can enable quick overall assessment and tracking of the ESM performance evolution during its development cycle. More examples showing other models are available in the Supplementary material (Figs. S1 to S3).

It is worth noting that there have been efforts to coalesce objective model evaluation concepts used in the research community (e.g., Knutti et al., 2010). However, the field continues to evolve rapidly with definitions still being debated and finessed. Via the PMP, we produce hundreds of summary statistics, enabling a broad net to be cast in the objective characterization of a simulation, at times helping modelers identify previously unknown deficiencies. For benchmarking, efforts are underway to establish a more targeted path which likely involves a consolidated set of carefully selected metrics.

5 Discussion

Efforts are underway to include new metrics into the PMP to advance the systematic objective evaluation of ESMs. For example, in coordination with the World Meteorological Organization (WMO)'s WGNE MJO Task Force, additional candidate MJO metrics for PMP inclusion have been identified to facilitate more comprehensive assessments of the MJO. Implementation of metrics for MJO amplitude, periodicity, and structure into the PMP is planned. An ongoing collaboration with NCAR aims to incorporate metrics related to the upper atmosphere, specifically the Quasi-Biennial Oscillation (QBO) and QBO-MJO metrics (e.g. Kim et al., 2020). We also have plans to grow the scope of PMP beyond its traditional atmospheric realm, for example including the ocean and polar regions through collaboration with the U.S. DOE's project entitled High Latitude Application and Testing of ESMs (HiLAT, https://www.hilat.org/). In addition, the PMP framework is also well poised to contribute to high-resolution climate modeling activities, such as the High-Resolution Model Intercomparison Project (HighResMIP; Haarsma et al., 2016) and the DYnamics of the Atmospheric general circulation Modeled On Non-hydrostatic Domains (DYAMOND; Stevens et al., 2019). This motivates the development of specialized metrics for high-resolution models, targeting the simulation features enabled by high-resolution models. Another potential avenue for the PMP involves leveraging Machine Learning (ML) techniques, and other state-of-the-art data science techniques being used for process-oriented ESM evaluation works (e.g., Nowack et al., 2020; Labe and Barnes, 2022; Dalelane et al., 2023). Applications of ML detection, such as for storms using TempestExtremes (Ullrich and Zarzycki 2017; Ullrich et al., 2021) and fronts (e.g,

Biard and Kunkel, 2019), can enable additional specialized storm metrics for high-resolution simulations. For convection-permitting models, yet more storm metrics can be applied such as Mesoscale convective systems. Atmospheric blocking metrics and atmospheric river evaluation metrics using the ML pattern detection capabilities in the latest TempestExtremes (Ullrich et al., 2021) are currently under development to be implemented into the PMP. These example enhancements of the PMP are indicative of an increasing priority to target regional simulation characteristics. With a deliberate emphasis on processes intrinsic to specific regions, this may lead to enabling potential applications of the PMP within the regional climate modeling activities such as the Coordinated Regional Downscaling Experiment (CORDEX; Gutowski Jr. et al., 2016).

The comprehensive database of PMP results offers a resource for exploring the range of structural errors in CMIP class models and their interrelationships. For example, examination of cross-metric relationships between mean-state and variability biases can shed additional light on the propagation of errors (e.g., Kang et al., 2020; Lee et al., 2021b). There continues to be interest in ranking models for specific applications (e.g., Ashfaq et al., 2022; Goldenson et al., 2023; Longmate et al., 2023; Papalexiou et al., 2020; Singh and AchutaRao, 2020) or to "move beyond one model one vote" in multi-model analysis to reduce uncertainties in the spread of multi-model projections (e.g., Knutti, 2010; Knutti et al., 2017; Sanderson et al., 2017; Herger et al., 2018; Hausfather et al., 2022; Merrifield et al., 2023). While we acknowledge potential interests in using the results of the PMP or equivalent to rank models or identify performance outliers (e.g., Sanderson and Wehner, 2017), we believe the many challenges associated with model weighting are application dependent, and thus leave it up to users of the PMP to make those judgments.

In addition to the scientific challenges associated with diversifying objective summaries of model performance, there is potential to leverage rapidly evolving technologies, including new open-source tools and methods available to scientists. We expect that the ongoing PMP code modernization effort to fully adapt the xCDAT and xarray will facilitate greater community involvement. As the PMP evolves with these technologies we will continue to maintain rigor in the calculation of statistics for the PMP metrics, for example by incorporating the latest advancements in the field. A prominent example in the objective comparison of models and observations involves the methodology of horizontal interpolation, and in future versions of the PMP we are planning a more stringent conservation method (Taylor, 2023). To improve the clarity of key messages from multivariate PMP metrics data, we will consider implementing the advances in high-dimensional data visualization, e.g., the circular plot discussed in Lee et al. (2018b) and variations of Parallel Coordinate Plots proposed in this paper and by Hassan et al. (2019) and Lu et al. (2020).

Current progress towards systematic model evaluation is exemplified by the diversity of tools being developed (e.g., the PMP, ESMValTool, MDTF, ILAMB, IOMB, and other packages). Each of these tools has its own scientific priorities and technical approaches. We believe that this diversity has made, and will continue to make, the model evaluation process even more comprehensive and successful. The fact that there is some overlap in a few cases is advantageous because it enables the cross-verification of results, which is particularly useful in more complex analyses. Despite possible advantages, having no single best or widely accepted approach for the community to follow, does introduce complexity to the coordination of model evaluation. To facilitate the collective usage of individual evaluation tools, the CMEC has initiated the development of a unified code base that technically coordinates the

Moved (insertion) [15]

Moved (insertion) [16]

Moved (insertion) [17]

Moved (insertion) [18]

Moved (insertion) [19]

Moved up [3]: progress towards systematic model evaluation remains dynamic, with evolving approaches and many independent paths being pursued. This has resulted in the development of diversified model evaluation software packages.

Moved up [4]: ESMValTool (Eyring et al., 2016, 2019, 2020; Righi et al.,

Moved up [5]: (Maloney et al., 2019; Neelin et al., 2023

Moved up [6]: 2014; Fasullo et al.,

Moved up [7]: Klingaman et al., 2017; Martin et al., 2017; Ordonez et al.,

Moved up [8]: regional climate community also has actively developed metrics packages such as the Regional Climate Model Evaluation System (RCMES; Lee et al., 2018a; Whitehall et al. 2012). Separately, a few climate modeling centers have developed their own model evaluation packages to assist in their in-house ESM development, e.g., the E3SM Diags (Zhang et al., 2022). There also have been other efforts to enhance the usability of in-situ and field campaign observations in ESM evaluations, such as Atmospheric Radiation Measurement (ARM) GCM Diag (Zhang et al., 2018, 2020) and Earth System Model Aerosol–Cloud Diagnostics (ESMAC Diags; Tang et al., 2022, 2023).

operation of distinct but complementary tools (Ordonez et al. 2021). Currently, the PMP, ILAMB, MDTF, and ASoP have become CMEC-compliant by adopting common interface standards that define how evaluation tools interact with observational data and climate model output. We expect that CMEC can also help the model evaluation community to establish standards for archiving the metrics output, much as the community did for the conventions to describe climate model data (e.g., CMIP application of CF Metadata Conventions (http://cfconventions.org/); Hassell et al., 2017; Eaton et al., 2022).

6 Summary and Conclusion

The PCMDI has actively developed the PMP with support from the U.S. DOE to improve the understanding of ESMs and to provide systematic and objective ESM evaluation capabilities. With its focus on physical climate, the current evaluation categories enabled in the PMP include seasonal and annual climatology of multiple variables, ENSO, various variability modes in the climate system, MJO, monsoon, cloud feedback and mean state, and simulated precipitation characteristics. The PMP provides quasi-operational ESM evaluation capabilities that can be rapidly deployed to objectively summarize a diverse suite of model behavior with results made publicly available. This can be of value in the assessment of community intercomparisons like CMIP, the evaluation of large ensembles, or the model development process. By documenting objective performance summaries produced by the PMP and making them available via detailed version control, additional research is made possible beyond the baseline model evaluation, model intercomparison, and benchmarking. The outcomes of PMP's calculations applied to the CMIP archive culminate in the PCMDI Simulation Summary (https://pcmdi.llnl.gov/metrics/) that has served as a comprehensive data portal for objective model-to-observation comparisons and model-to-model benchmarking and intercomparisons. Special attention is dedicated to the most recent ensemble of models contributing to CMIP6. By offering a diverse and comprehensive suite of evaluation capabilities, the PMP framework equips model developers with quantifiable benchmarks to validate and enhance model performance.

We expect that the PMP will continue to play a crucial role in benchmarking ESMs. Improvements in the PMP, along with progress in interconnected MIP community projects, will greatly contribute to advancing the evaluation of ESMs including in connection to the community efforts (e.g., the CMIP Benchmarking Task Team). Enhancements in version control and transparency within obs4MIPs are set to enhance the provenance and reproducibility of PMP results, thereby strengthening the foundation for rigorous and repeatable performance benchmarking. The PMP's collaboration with the CMIP Forcing Task Team, through the Input4MIPs (Durack et al., 2018) and the CMIP6Plus projects, will further expand the utility of performance metrics in identifying problems associated with the forcing dataset and their application and use in reproducing the observed record of historical climate. Furthermore, as ESMs advance towards more operationalized configurations to meet the demands of decision-making processes (Jakob et al., 2023), the PMP holds significant potential to provide interoperable ESM evaluation and benchmarking capabilities to the community.

992 **Appendix A: Table of acronyms**

993

Acronym	Description
AMIP	Atmospheric Model Intercomparison Project
AMO	Atlantic Multi-decadal Oscillation
ARM	Atmospheric Radiation Measurement
ASoP	Analyzing Scales of Precipitation
CBF	Common Basis Function
CDAT	Community Data Analysis Tools
CIESM	Community Integrated Earth System Model
CLIVAR	Climate and Ocean Variability, Predictability and Change
CMEC	Coordinated Model Evaluation Capabilities
CMIP	Coupled Model Intercomparison Project
CMOR	Climate Model Output Rewriter
CVDP	Climate Variability Diagnostics Package
DOE	U.S. Department of Energy
ENSO	El Niño-Southern Oscillation
EOF	Empirical Orthogonal Functions
EOR	East power normalized by Observation

ESGF	Earth System Grid Federation
ESM	Earth System Model
ESMAC Diags	Earth System Model Aerosol–Cloud Diagnostics
ETMoV	Extratropical modes of variability
EWR	East/West power Ratio
GFDL	Geophysical Fluid Dynamics Laboratory
ILAMB	International Land Model Benchmarking
IOMB	International Ocean Model Benchmarking
IPCC	Intergovernmental Panel on Climate Change
IPSL	Institut Pierre-Simon Laplace
ISCCP HGG	International Satellite Cloud Climatology Project H-series Gridded Global
ITCZ	Intertropical Convergence Zone
JSON	JavaScript Object Notation
MAE	Mean Absolute Error
MDTF	Model Diagnostics Task Force
MIPs	Model Intercomparison Projects
MJO	Madden-Julian Oscillation
NAM	Northern Annular Mode

NAO	North Atlantic Oscillation
NASA	National Aeronautics and Space Administration
NCAR	National Center for Atmospheric Research
NetCDF	Network Common Data Form
NH	Northern Hemisphere
NIMS-KMA	National Institute of Meteorological Sciences-Korea Meteorological Administration
NOAA	National Oceanic and Atmospheric Administration
NPGO	North Pacific Gyre Oscillation
NPO	North Pacific Oscillation
PCMDI	Program for Climate Model Diagnosis and Intercomparison
PDO	Pacific Decadal Oscillation
PMP	PCMDI Metrics Package
PNA	Pacific North America pattern
RCMES	Regional Climate Model Evaluation System
RMSE	Root-Mean-Square Error
SAM	Southern Annular Mode
SH	Southern Hemisphere
SST	Sea Surface Temperature

TOA	Top of Atmosphere
WCRP	World Climate Research Programme
WGCM	Working Group on Coupled Models
WGNE	Working Group on Numerical Experimentation
xCDAT	Xarray Climate Data Analysis Tools

994

Code and Data Availability

The source code of the PMP (Lee et al., 2023b) is available as an open-source Python package: https://github.com/PCMDI/pcmdi_metrics (last access: 21 February 2024) with all released versions archived on Zenodo DOI: https://doi.org/10.5281/zenodo.592790 (last access: 21 February 2024). The online documentation is available at http://pcmdi.github.io/pcmdi_metrics (last access: 21 February 2024). The PMP results database (Lee et al., 2023a) that includes calculated metrics is available on the GitHub repository at https://github.com/PCMDI/pcmdi_metrics_results_archive (last access: 21 February 2024) with versions archived on Zenodo DOI: https://doi.org/10.5281/zenodo.10181201. PMP's installation process is streamlined using the *Anaconda* distribution and the *conda-forge* channel (https://anaconda.org/conda-forge/pcmdi_metrics, last access: 21 February 2024). The installation instructions are available at http://pcmdi.github.io/pcmdi_metrics/install.html (last access: 21 February 2024). The interactive visualizations of the PMP results are available on the PCMDI website at https://pcmdi.llnl.gov/metrics (last access: 21 November 2023). The CMIP5 and CMIP6 model outputs and obs4MIPs datasets used in this paper are available via the Earth System Grid Federation at https://esgf-node.llnl.gov/ (last access: 21 February 2024).

Author Contributions

All authors contributed to the design and implementation of the research, analysis of the results, and to writing of the manuscript. All authors contributed to the development of codes/metrics in the PMP, its ecosystem tools, and/or the establishment of use cases. JL and PJG led and coordinated the paper with input from all authors.

Competing interests

At least one of the coauthors is a member of the editorial board of *Geoscientific Model Development*. The peer-review process was guided by an independent editor, and the authors also have no other competing interests to declare.

Acknowledgment

We acknowledge the World Climate Research Programme, which, through its Working Group on Coupled Modeling, 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 that support CMIP6 and ESGF. This work is performed under the auspices of the U.S. DOE by Lawrence Livermore National Laboratory (LLNL) under Contract No. DE-AC52-07NA27344. Efforts of JL, PJG, MA, AO, PU, KET, PD, CB, MDZ, LC, and BD were supported by the Regional and Global Model Analysis (RGMA) program of the U.S. Department of Energy (DOE) Office of Science (OS), Biological and Environmental Research (BER) program. MFW was supported by the Director, OS, BER of the U.S. DOE through the RGMA program under Contract No. DE340AC02-05CH11231. AGP was supported by U.S. DOE through BER RGMA through Award Number DE-SC0022070 and via National Science Foundation (NSF) IA 1947282, and by National Center for Atmospheric Research (NCAR), which is a major facility sponsored by the NSF under Cooperative Agreement No. 1852977. YYP and EG were supported by the Agence Nationale de la Recherche ARISE project, under Grant ANR-

18-CE01-0012, and the Belmont project GOTHAM, under Grant ANR-15-JCLI-0004-01, the European Commission's H2020 Programme "Infrastructure for the European Network for Earth System Modelling Phase 3 (IS-ENES3)" project under Grant Agreement 824084. DK was supported by the New Faculty Startup Fund from Seoul National University and the KMA R&D program (KMI2022-01313). The authors thank Program Manager Renu Joseph of the U.S. DOE for the support and advocacy for the Program for Climate Model Diagnosis and Intercomparison (PCMDI) project and the PMP. We thank Stephen Klein for his leadership for the PCMDI project from 2019 to 2022. We acknowledge contributions from our LLNL colleagues, Lina Muryanto and Zeshawn Shaheen (Now at Google LLC) during the early stage of the PMP, and Sasha Ames, Jeff Painter, Chris Mauzey, and Stephen Po-Chedley for the PCMDI's CMIP database management. The authors also thank Liping Zhang for her comments during GFDL's internal review process.

References

Adler, R.F., Sapiano, M. R., Huffman, G.J., Wang, J.J., Gu, G., Bolvin, D., Chiu, L., Schneider, U., Becker, A., Nelkin, E., Xie, P., Ferraro, R., Shin, D.-B.: The Global Precipitation Climatology Project (GPCP) monthly analysis (new version 2.3) and a review of 2017 global precipitation. Atmosphere, 9, 138, https://doi.org/10.3390/atmos9040138, 2018.

Ahn, M.-S., Kim, D. H., Sperber, K. R., Kang, I.-S., Maloney, E. D., Waliser, D. E., and Hendon, H. H.: MJO simulation in CMIP5 climate models: MJO skill metrics and process-oriented diagnosis, Climate Dynamics, 49, 4023–4045, https://doi.org/10.1007/s00382-017-3558-4, 2017.

Ahn, M.-S., Gleckler, P. J., Lee, J., Pendergrass, A. G., and Jakob, C.: Benchmarking Simulated Precipitation Variability Amplitude across Time Scales, Journal of Climate, 35, 3173–3196, https://doi.org/10.1175/jcli-d-21-0542.1, 2022.

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

Arcodia, M., Barnes, E. A., Mayer, K., Lee, J., Ordonez, A., and Ahn, M.-S.: Assessing decadal variability of subseasonal forecasts of opportunity using explainable AI, Environmental Research, https://doi.org/10.1088/2752-5295/aced60, 2023.

Ashfaq, M., Rastogi, D., Kitson, J., Abid, M. A., and Kao, S.-C.: Evaluation of CMIP6 GCMs over the CONUS for downscaling studies, Journal of Geophysical Research: Atmospheres, 127, e2022JD036659. https://doi.org/10.1029/2022JD036659, 2022.

Bayr, T., Wengel, C., Latif, M., Dommenget, D., Lübbecke, J., and Park, W.: Error compensation of ENSO atmospheric feedbacks in climate models and its influence on simulated ENSO dynamics, Climate Dynamics, 53, 155–172, https://doi.org/10.1007/s00382-018-4575-7, 2019.

Biard, J. C. and Kunkel, K. E.: Automated detection of weather fronts using a deep learning neural network, Adv. Stat. Clim. Meteorol. Oceanogr., 5, 147–160, https://doi.org/10.5194/ascmo-5-147-2019, 2019.

1083   Bellenger, H., Guilyardi, E., Leloup, J., Lengaigne, M., and Vialard, J.: ENSO representation in climate models: from
1084   CMIP3 to CMIP5, Climate Dynamics, 42, 1999–2018, https://doi.org/10.1007/s00382-013-1783-z, 2013.

1085   Boucher, O., Servonnat, J., Albright, A. L., Aumont, O., Balkanski, Y., Bastrikov, V., Bekki, S., Bonnet, R., Bony,
1086   S., Bopp, L., Braconnot, P., Brockmann, P., Cadule, P., Caubel, A., Cheruy, F., Codron, F., Cozic, A., Cugnet,
1087   D., D'Andrea, F., Davini, P., De Lavergne, C., Denvil, S., Deshayes, J., Devilliers, M., Ducharne, A.,
1088   Dufresne, J. L., Dupont, E., Ethé, C., Fairhead, L., Falletti, L., Flavoni, S., Foujols, M. A., Gardoll, S.,
1089   Gastineau, G., Ghattas, J., Grandpeix, J. Y., Guenet, B., Lionel, E. G., Guilyardi, E., Guimberteau, M.,
1090   Hauglustaine, D., Hourdin, F., Idelkadi, A., Joussaume, S., Kageyama, M., Khodri, M., Krinner, G., Lebas,
1091   N., Levavasseur, G., Lévy, C., Li, L., Lott, F., Lurton, T., Luyssaert, S., Madec, G., Madeleine, J.-B.,
1092   Maignan, F., Marchand, M., Marti, O., Mellul, L., Meurdesoif, Y., Mignot, J., Musat, I., Ottlé, C., Peylin, P.,
1093   Planton, Y., Polcher, J., Rio, C., Rochetin, N., Rousset, C., Sepulchre, P., Sima, A., Swingedouw, D.,
1094   Thiéblemont, R., Traore, A. K., Vancoppenolle, M., Vial, J., Vialard, J., Viovy, N., and Vuichard, N.:
1095   Presentation and evaluation of the IPSL-CM6A-LR Climate Model, Journal of Advances in Modeling Earth
1096   Systems, 12, https://doi.org/10.1029/2019ms002010, 2020.

1097   Caldwell, P., Mametjanov, A., Tang, Q., Van Roekel, L., Golaz, J.-C., Lin, W., Bader, D. C., Keen, N. D., Feng, Y.,
1098   Jacob, R., Maltrud, M., Roberts, A., Taylor, M. A., Veneziani, M., Wang, H., Wolfe, J. D., Balaguru, K.,
1099   Cameron-Smith, P. J., Dong, L., Klein, S. A., Leung, L. R., Li, H., Li, Q., Liu, X., Neale, R., Pinheiro, M.
1100   C., Qian, Y., Ullrich, P. A., Xie, S., Yang, Y., Zhang, Y., Zhang, K., and Zhou, T.: The DOE E3SM Coupled
1101   Model Version 1: description and results at high resolution, Journal of Advances in Modeling Earth Systems,
1102   11, 4095–4146, https://doi.org/10.1029/2019ms001870, 2019.

1103   Chen, H.-C., Jin, F.-F., Zhao, S., Wittenberg, A. T., and Xie, S.: ENSO dynamics in the E3SM-1-0, CESM2, and
1104   GFDL-CM4 climate models, Journal of Climate, 34, 9365-9384, https://doi.org/10.1175/JCLI-D-21-0355.1,
1105   2021.

1106   Collier, N., Hoffman, F. M., Lawrence, D. M., Keppel-Aleks, G., Koven, C. D., Riley, W. J., Mu, M., and Randerson,
1107   J. T.: The International Land Model Benchmarking (ILAMB) System: Design, theory, and implementation,
1108   Journal of Advances in Modeling Earth Systems, 10, 2731–2754, https://doi.org/10.1029/2018ms001354,
1109   2018.

1110   Covey, C., AchutaRao, K., Cubasch, U., Jones, P., Lambert, S. J., Mann, M., Phillips, T. J., and Taylor, K. E.: An
1111   overview of results from the Coupled Model Intercomparison Project, Global and Planetary Change, 37, 103–
1112   133, https://doi.org/10.1016/s0921-8181(02)00193-5, 2003.

1113   Covey, C., Gleckler, P. J., Doutriaux, C., Williams, D. N., Dai, A., Fasullo, J. T., Trenberth, K. E., and Berg, A.:
1114   Metrics for the diurnal cycle of precipitation: toward routine benchmarks for climate models, Journal of
1115   Climate, 29, 4461–4471, https://doi.org/10.1175/jcli-d-15-0664.1, 2016.

1116   Crockford, D.: The application/json media type for javascript object notation (json) (No. rfc4627), https://www.rfc-
1117   editor.org/rfc/pdfrfc/rfc4627.txt.pdf (last access: 5 March 2024), 2006.

1118   Crockford, D. and Morningstar, C.: The JSON Data Interchange Syntax, ECMA-404, ECMA International, 2017.

Dalelane, C., Winderlich, K., and Walter, A.: Evaluation of global teleconnections in CMIP6 climate projections using complex networks, Earth Syst. Dynam., 14, 17–37, https://doi.org/10.5194/esd-14-17-2023, 2023.

Dawson, A.: eofs: A Library for EOF Analysis of Meteorological, Oceanographic, and Climate Data, Journal of Open Research Software (JORS), 4, e14, https://doi.org/10.5334/jors.122, 2016.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N., and Vitart, F.: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, Quarterly Journal of the Royal Meteorological Society, 137, 553–597, https://doi.org/10.1002/qj.828, 2011

Deser, C. and Phillips, A. S.: Defining the internal component of Atlantic multidecadal variability in a changing climate, Geophysical Research Letters, 48, https://doi.org/10.1029/2021gl095023, 2021.

Doutriaux, C., Nadeau, D., Wittenburg, S., Lipsa, D., Muryanto, L., Chaudhary, A., and Williams, D. N.: CDAT/cdat: CDAT 8.1, Zenodo [Code], https://doi.org/10.5281/zenodo.2586088, 2019.

Durack, P. J., Taylor, K. E., Eyring, V., Ames, S., Hoang, T., Nadeau, D., Doutriaux, C., Stockhause, M., and Gleckler, P. J.: Toward standardized data sets for climate model experimentation, Eos, Transactions American Geophysical Union, 99, https://doi.org/10.1029/2018eo101751, 2018.

Eaton, B., Gregory, J., Drach, B., Taylor, K., Hankin, S., Blower, J., Caron, J., Signell, R., Bentley, P., Rappa, G., Höck, H., Pamment, A., Juckes, M., Raspaud, M., Horne, R., Whiteaker, T., Blodgett, D., Zender, C., Lee, D., Hassell, D., Snow, A. D., Kölling, T., Allured, D., Jelenak, A., Soerensen, A. M., Gaultier, L., Herlédan, S.: NetCDF Climate and Forecast (CF) Meta-data Conventions V1.10, available at: http://cfconventions.org/Data/cf-conventions/cf-conventions-1.10/cf-conventions.html (last access: 6 November 2023), 2022.

Eyring, V., Righi, M., Lauer, A., Evaldsson, M., Wenzel, S., Jones, C., Anav, A., Andrews, O., Cionni, I., Davin, E. L., Deser, C., Ehbrecht, C., Friedlingstein, P., Gleckler, P. J., Gottschaldt, K.-D., Hagemann, S., Juckes, M., Kindermann, S., Krasting, J. P., Kunert, D., Levine, R. C., Loew, A., Mäkelä, J., Martin, G., Mason, E., Phillips, A. S., Read, S., Rio, C., Roehrig, R., Senftleben, D., Sterl, A., Van Ulft, L. H., Walton, J., Wang, S., and Williams, K. D.: ESMValTool (v1.0) – a community diagnostic and performance metrics tool for routine evaluation of Earth system models in CMIP, Geoscientific Model Development, 9, 1747–1802, https://doi.org/10.5194/gmd-9-1747-2016, 2016a.

Eyring, V., Bony, S., Meehl, G. A., A, C., Senior, Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geoscientific Model Development, 9, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016, 2016b.

Eyring, V., Cox, P. M., Flato, G. M., Gleckler, P. J., Abramowitz, G., Caldwell, P., Collins, W. D., Gier, B. K., Hall, A., Hoffman, F. M., Hurtt, G. C., Jahn, A., Jones, C. D., Klein, S. A., Krasting, J. P., Kwiatkowski, L., Lorenz, R., Maloney, E. D., Meehl, G. A., Pendergrass, A. G., Pincus, R., Ruane, A. C., Russell, J. L.,

1167   Sanderson, B. M., Santer, B. D., Sherwood, S. C., Simpson, I. R., Stouffer, R. J., and Williamson, M. S.:
1168   Taking climate model evaluation to the next level, Nature Climate Change, 9, 102–110,
1169   https://doi.org/10.1038/s41558-018-0355-y, 2019.

1170 Eyring, V., Bock, L., Lauer, A., Righi, M., Schlund, M., Andela, B., Arnone, E., Bellprat, O., Brötz, B., Caron, L.-P.,
1171   Carvalhais, N., Cionni, I., Cortesi, N., Crezee, B., Davin, E. L., Davini, P., Debeire, K., De Mora, L., Deser,
1172   C., Docquier, D., Earnshaw, P., Ehbrecht, C., Gier, B. K., Gonzalez-Reviriego, N., Goodman, P. J.,
1173   Hagemann, S., Hardiman, S. C., Hassler, B., Hunter, A., Kadow, C., Kindermann, S., Koirala, S., Koldunov,
1174   N., Lejeune, Q., Lembo, V., Lovato, T., Lucarini, V., Massonnet, F., Müller, B., Pandde, A., Pérez-Zanón,
1175   N., Phillips, A. S., Predoi, V., Russell, J. L., Sellar, A., Serva, F., Stacke, T., Swaminathan, R., Torralba, V.,
1176   Vegas-Regidor, J., Von Hardenberg, J., Weigel, K., and Zimmermann, K.: Earth System Model Evaluation
1177   Tool (ESMValTool) v2.0 – an extended set of large-scale diagnostics for quasi-operational and
1178   comprehensive evaluation of Earth system models in CMIP, Geoscientific Model Development, 13, 3383–
1179   3438, https://doi.org/10.5194/gmd-13-3383-2020, 2020.

1180 Eyring, V., Gillett, N.P., Achuta Rao, K.M., Barimalala, R., Barreiro Parrillo, M., Bellouin, N., Cassou, C., Durack,
1181   P.J., Kosaka, Y., McGregor, S. and Min, S., Morgenstern, O., and Sun, Y.: Human Influence on the Climate
1182   System. In Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth
1183   Assessment Report of the Intergovernmental Panel on Climate Change. 105, 423-552,
1184   https://doi.org/10.1017/9781009157896.005, 2021.

1185 Fasullo, J. T.: Evaluating simulated climate patterns from the CMIP archives using satellite and reanalysis datasets
1186   using the Climate Model Assessment Tool (CMATv1), Geoscientific Model Development, 13, 3627–3642,
1187   https://doi.org/10.5194/gmd-13-3627-2020, 2020.

1188 Fasullo, J. T., Phillips, A. S., and Deser, C.: Evaluation of leading modes of climate variability in the CMIP archives,
1189   Journal of Climate, 33, 5527–5545, https://doi.org/10.1175/jcli-d-19-1024.1, 2020.

1190 Ferraro, R., Waliser, D. E., Gleckler, P. J., Taylor, K. E., and Eyring, V.: Evolving OBS4MIPS to support Phase 6 of
1191   the Coupled Model Intercomparison Project (CMIP6), Bulletin of the American Meteorological Society,
1192   https://doi.org/10.1175/bams-d-14-00216.1, 2015.

1193 Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S.C., Collins, W., Cox, P., Driouech, F., Emori, S., Eyring,
1194   V. and Forest, C.: Evaluation of climate models. In *Climate change 2013: the physical science basis.*
1195   *Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate*
1196   *Change* (pp. 741-866). Cambridge University Press. 2014.

1197 Fu, W., Moore, J. K., Primeau, F., Collier, N., Ogunro, O. O., Hoffman, F. M. and Randerson, J. T.: Evaluation of
1198   ocean biogeochemistry and carbon cycling in CMIP earth system models with the international ocean model
1199   benchmarking (IOMB) software System. Journal of Geophysical Research: Oceans, 127, e2022JC018965,
1200   https://doi.org/10.1029/2022JC018965, 2022.

1201 Gates, W.L.: AN AMS continuing series: Global CHANGE–AMIP: The Atmospheric Model Intercomparison Project,
1202   Bulletin of the American Meteorological Society, 73, 1962-1970, 1992.

1204 Gates, W.L., Henderson-Sellers, A., Boer, G.J., Folland, C.K., Kitoh, A., McAvaney, B.J., Semazzi, F., Smith, N.,
1205   Weaver, A.J. and Zeng, Q.C.: Climate models—evaluation. *Climate change* 1: 229-284, 1995.
1206 Gates, W.L., Boyle, J.S., Covey, C., Dease, C.G., Doutriaux, C.M., Drach, R.S., Fiorino, M., Gleckler, P.J., Hnilo,
1207   J.J., Marlais, S.M. and Phillips, T.J.: An overview of the results of the Atmospheric Model Intercomparison
1208   Project (AMIP I). Bulletin of the American Meteorological Society, 80, 29-56, 1999.
1209 Gleckler, P. J., Taylor, K. E., and Doutriaux, C.: Performance metrics for climate models, Journal of Geophysical
1210   Research, 113, https://doi.org/10.1029/2007jd008972, 2008.
1211 Gleckler, P. J., Ferraro, R., and Waliser, D. E.: Improving use of satellite data in evaluating climate models, Eos,
1212   Transactions American Geophysical Union, 92, 172, https://doi.org/10.1029/2011eo200005, 2011.
1213 Gleckler, P. J., Doutriaux, C., Durack, P. J., Taylor, K. E., Zhang, Y., Williams, D. N., Mason, E., and Servonnat, J.:
1214   A more powerful reality test for climate models, Eos, Transactions American Geophysical Union, 97,
1215   https://doi.org/10.1029/2016eo051663, 2016.
1216 Golaz, J.-C., Caldwell, P., Van Roekel, L., Petersen, M. R., Tang, Q., Wolfe, J. D., Abeshu, G. W., Anantharaj, V.,
1217   Asay-Davis, X., Bader, D. C., Baldwin, S., Bisht, G., Bogenschutz, P., Branstetter, M. L., Brunke, M. A.,
1218   Brus, S., Burrows, S. M., Cameron-Smith, P. J., Donahue, A. S., Deakin, M., Easter, R. C., Evans, K. J.,
1219   Feng, Y., Flanner, M., Foucar, J. G., Fyke, J., Griffin, B. M., Hannay, C., Harrop, B. E., Hoffman, M. J.,
1220   Hunke, E., Jacob, R., Jacobsen, D. W., Jeffery, N., Jones, P. W., Keen, N. D., Klein, S. A., Larson, V. E.,
1221   Leung, L. R., Li, H. Y., Lin, W., Lipscomb, W. H., Lun, P., Mahajan, S., Maltrud, M., Mametjanov, A.,
1222   McClean, J. L., McCoy, R., Neale, R., Price, S., Qian, Y., Rasch, P. J., Eyre, J. E. J. R., Riley, W. J., Ringler,
1223   T. D., Roberts, A., Roesler, E. L., Salinger, A. G., Shaheen, Z., Shi, X., Singh, B., Tang, J., Taylor, M. A.,
1224   Thornton, P. E., Turner, A. K., Veneziani, M., Wan, H., Wang, H., Wang, S., Williams, D. N., Wolfram, P.
1225   J., Worley, P. H., Xie, S., Yang, Y., Yoon, J., Zelinka, M. D., Zender, C. S., Zeng, X., Zhang, C., Zhang, K.,
1226   Zhang, Y., Zheng, X., Zhou, T., and Zhu, Q.: The DOE E3SM Coupled Model Version 1: Overview and
1227   evaluation at standard resolution, Journal of Advances in Modeling Earth Systems, 11, 2089–2129,
1228   https://doi.org/10.1029/2018ms001603, 2019.
1229 Goldenson, N., Leung, L. R., Mearns, L. O., Pierce, D. W., Reed, K. A., Simpson, I. R., Ullrich, P., Krantz, W., Hall,
1230   A., Jones, A. and Rahimi, S.: Use-Inspired, Process-Oriented GCM Selection: Prioritizing Models for
1231   Regional Dynamical Downscaling, Bulletin of the American Meteorological Society, E1619–E1629,
1232   https://doi.org/10.1175/BAMS-D-23-0100.1, 2023.
1233 Guilyardi, E., Wittenberg, A., Fedorov, A., Collins, M., Wang, C., Capotondi, A., Van Oldenborgh, G.J. and
1234   Stockdale, T.: Understanding El Niño in ocean–atmosphere general circulation models: Progress and
1235   challenges,   Bulletin   of   the   American   Meteorological   Society,   90,   325-340,
1236   https://doi.org/10.1175/2008BAMS2387.1, 2009.
1237 Guilyardi E., Capotondi, A., Lengaigne, M., Thual, S., Wittenberg, A. T.: ENSO modelling: history, progress and
1238   challenges, in: El Niño in a changing climate, edited by: McPhaden, M. J., Santoso, A., Cai, W., AGU
1239   monograph, ISBN: 9781119548164, https://doi.org/10.1002/9781119548164.ch9, 2020.

Gutowski Jr., W. J., Giorgi, F., Timbal, B., Frigon, A., Jacob, D., Kang, H.-S., Raghavan, K., Lee, B., Lennard, C., Nikulin, G., O'Rourke, E., Rixen, M., Solman, S., Stephenson, T., and Tangang, F.: WCRP Coordinated Regional Downscaling EXperiment (CORDEX): a diagnostic MIP for CMIP6, Geosci. Model Dev., 9, 4087–4095, https://doi.org/10.5194/gmd-9-4087-2016, 2016.

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

Hannah, W. M., Bradley, A. M., Guba, O., Tang, Q., Golaz, J.-C., and Wolfe, J. D.: Separating physics and dynamics grids for improved computational efficiency in spectral element Earth system models, Journal of Advances in Modeling Earth Systems, 13, https://doi.org/10.1029/2020ms002419, 2021.

Hassan, K. A., Rönnberg, N., Forsell, C., Cooper, M. and Johansson, J.: A study on 2D and 3D parallel coordinates for pattern identification in temporal multivariate data, in: 2019 23rd International Conference Information Visualisation (IV), 145-150, https://doi.org/10.1109/IV.2019.00033, 2019.

Hassell, D., Gregory, J. M., Blower, J., Lawrence, B., and Taylor, K. E.: A data model of the Climate and Forecast metadata conventions (CF-1.6) with a software implementation (cf-python v2.1), Geoscientific Model Development, 10, 4619–4646, https://doi.org/10.5194/gmd-10-4619-2017, 2017.

Hausfather, Z., Marvel, K., Schmidt, G. A., Nielsen-Gammon, J. W. and Zelinka, M.: Climate simulations: Recognize the 'hot model' problem, Nature, 605, 26-29, https://doi.org/10.1038/d41586-022-01192-2, 2022.

Held, I. M., Guo, H., Adcroft, A., Dunne, J. P., Horowitz, L. W., Krasting, J., Shevliakova, E., Winton, M., Zhao, M., Bushuk, M., Wittenberg, A. T., and coauthors: Structure and performance of GFDL's CM4. 0 climate model, Journal of Advances in Modeling Earth Systems, 11, 3691-3727, https://doi.org/10.1029/2019MS001829, 2019.

Hendon, H. H., Zhang, C., and Glick, J. D.: Interannual Variation of the Madden–Julian Oscillation during Austral Summer, Journal of Climate, 12, 2538–2550, 1999.

Herger, N., Abramowitz, G., Knutti, R., Angélil, O., Lehmann, K., and Sanderson, B. M.: Selecting a climate model subset to optimise key ensemble properties, Earth System Dynamics Discussions, 9, 135–151, https://doi.org/10.5194/esd-9-135-2018, 2018.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D. and coauthors: The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society, 146, 1999-2049, https://doi.org/10.1002/qj.3803, 2020.

Hintze, J. L., and Nelson, R. D.: Violin plots: A box plot-density trace synergism, The American Statistician, 52, 181–184, https://doi.org/10.1080/00031305.1998.10480559, 1998.

Hoyer, S. and Hamman, J.: xarray: N-D labeled Arrays and Datasets in Python, Journal of Open Research Software, 5, 10. https://doi.org/10.5334/jors.148, 2017.

1280 Huffman, G. J., Adler, R. F., Morrissey, M. M., Bolvin, D. T., Curtis, S., Joyce, R., McGavock, B. and Susskind, J.:
1281   Global precipitation at one-degree daily resolution from multisatellite observations, Journal of
1282   hydrometeorology, 2, 36-50, 2001.

1283 Huffman, G. J., Bolvin, D. T., Braithwaite, D., Hsu, K., Joyce, R., Kidd, C., Nelkin, E. J., Sorooshian, S., Tan, J., and
1284   Xie, P.: NASA global precipitation measurement (GPM) integrated multi-satellite retrievals for GPM
1285   (IMERG). Algorithm theoretical basis document (ATBD) version, 4, p.30., 2015.

1286 Inselberg, A.: Multidimensional detective, in: Proceedings of IEEE Symposium on Information Visualization, 100–
1287   107, https://doi.org/10.1109/INFVIS.1997.636793, 1997.

1288 Inselberg, A.: Parallel Coordinates: Visualization, Exploration and Classification of High-Dimensional Data, in:
1289   Handbook of Data Visualization, edited by Chen, C., Härdle, W., and Unwin, A., Springer, Berlin,
1290   Heidelberg, Germany, 643-680, https://doi.org/10.1007/978-3-540-33037-0_25, 2008.

1291 Inselberg, A.: Parallel Coordinates, in: Encyclopedia of Database Systems. Springer, edited by Liu, L., and Özsu, M.
1292   T., Springer, New York, NY, U.S.A., https://doi.org/10.1007/978-1-4899-7993-3_262-2, 2016.

1293 Johansson, J. and Forsell, C.: Evaluation of parallel coordinates: Overview, categorization and guidelines for future
1294   research, IEEE Transactions on Visualization and Computer Graphics, 22, 579-588,
1295   https://doi.org/10.1109/TVCG.2015.2466992, 2016.

1296 Jakob, C., Gettelman, A. and Pitman, A.: The need to operationalize climate modelling, Nat. Clim. Chang. 13, 1158–
1297   1160, https://doi.org/10.1038/s41558-023-01849-4, 2023.

1298 Joyce, R. J., Janowiak, J. E., Arkin, P. A. and Xie, P.: CMORPH: A method that produces global precipitation
1299   estimates from passive microwave and infrared data at high spatial and temporal resolution, Journal of
1300   hydrometeorology, 5, 487-503, 2004.

1301 Kang, D., Kim, D. H., Ahn, M.-S., Neale, R., Lee, J., and Gleckler, P. J.: The role of the mean state on MJO simulation
1302   in CESM2 ensemble simulation, Geophysical Research Letters, 47, https://doi.org/10.1029/2020gl089824,
1303   2020.

1304 Kim, D., Sperber, K. R., Stern, W., Waliser, D. E., Kang, I. S., Maloney, E. D., Wang, W., Weickmann, K. M.,
1305   Benedict, J. J., Khairoutdinov, M., Lee, M.-I., Neale, R., Suarez, M. J., Thayer-Calder, K., and Zhang, G.:
1306   Application of MJO simulation diagnostics to climate models, Journal of Climate, 22, 6413–6436,
1307   https://doi.org/10.1175/2009jcli3063.1, 2009.

1308 Kim, H., Caron, J. M., Richter, J. H. and Simpson, I. R.: The lack of QBO-MJO connection in CMIP6 models,
1309   Geophysical Research Letters, 47, e2020GL087295, https://doi.org/10.1029/2020GL087295, 2020.

1310 Klein, S. A., Zhang, Y., Zelinka, M. D., Pincus, R., Boyle, J., and Gleckler, P. J.: Are climate model simulations of
1311   clouds improving? An evaluation using the ISCCP simulator, Journal of Geophysical Research:
1312   Atmospheres, 118, 1329–1342, https://doi.org/10.1002/jgrd.50141, 2013.

1313 Klingaman, N. P., Martin, G., and Moise, A.: ASoP (v1.0): a set of methods for analyzing scales of precipitation in
1314   general circulation models, Geoscientific Model Development, 10, 57–83, https://doi.org/10.5194/gmd-10-
1315   57-2017, 2017.

1326 Knutti, R.: The end of model democracy? Climatic Change, 102, 395–404, https://doi.org/10.1007/s10584-010-9800-
1327   2, 2010.

1328 Knutti, R., Sedláček, J., Sanderson, B. M., Lorenz, R., Fischer, E. M., and Eyring, V.: A climate model projection
1329   weighting scheme accounting for performance and interdependence, Geophysical Research Letters,
1330   https://doi.org/10.1002/2016gl072012, 2017.

1331 Knutti, R., Abramowitz, G., Collins, M., Eyring, V., Gleckler, P. J., Hewitson, B., and Mearns, L.: Good Practice
1332   Guidance Paper on Assessing and Combining Multi Model Climate Projections, in: Meeting Report of the
1333   Intergovernmental Panel on Climate Change Expert Meeting on Assessing and Combining Multi Model
1334   Climate Projections, edited by Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M., and Midgley, P. M., IPCC
1335   Working Group I Technical Support Unit, University of Bern, Bern, Switzerland, 2010.

1336 Labe, Z. M. and Barnes, E. A.: Comparison of Climate Model Large Ensembles With Observations in the Arctic Using
1337   Simple Neural Networks, Earth and Space Science, e2022EA002348,
1338   https://doi.org/10.1029/2022EA002348, 2022.

1339 Lambert, S. J. and Boer, G. J.: CMIP1 evaluation and intercomparison of coupled climate models, Climate Dynamics,
1340   17, 83-106, https://doi.org/10.1007/PL00013736, 2001.

1341 Lee, H., Goodman, A., McGibbney, L. J., Waliser, D. E., Kim, J., Loikith, P. C., Gibson, P. B., and Massoud, E.:
1342   Regional Climate Model Evaluation System powered by Apache Open Climate Workbench v1.3.0: an
1343   enabling tool for facilitating regional climate studies, Geoscientific Model Development, 11, 4435–4449,
1344   https://doi.org/10.5194/gmd-11-4435-2018, 2018a.

1345 Lee, J., Ahn, M.-S., Ordonez, A., Gleckler, P., and Ullrich, P.: PCMDI/pcmdi_metrics_results_archive, Zenodo [data],
1346   https://doi.org/10.5281/zenodo.10181201, 2023a.

1347 Lee, J., Gleckler, P., Ordonez, A., Ahn, M.-S., Ullrich, P., Tom, V., Jason, B., Charles, D., Durack, P., Shaheen, Z.,
1348   Muryanto, L., Painter, J., and Krasting, J.: PCMDI/pcmdi_metrics: PMP Version 3.1.1, Zenodo [code],
1349   https://doi.org/10.5281/zenodo.592790, 2023b.

1350 Lee, J., Gleckler, P., Sperber, K., Doutriaux C., and Williams, D.: High-dimensional Data Visualization for Climate
1351   Model Intercomparison: Application of the Circular Plot, in: Proceedings of the 8th International Workshop
1352   on Climate Informatics: CI 2018. NCAR Technical Note NCAR/TN-550+PROC, 12-14,
1353   http://dx.doi.org/10.5065/D6BZ64XQ, 2018b.

1354 Lee, J., Planton, Y., Gleckler, P. J., Sperber, K. R., Guilyardi, E., Wittenberg, A. T., McPhaden, M. J., and Pallotta,
1355   G.: Robust evaluation of ENSO in climate models: How many ensemble members are needed?, Geophysical
1356   Research Letters, 48, https://doi.org/10.1029/2021gl095041, 2021a.

1357 Lee, J., Sperber, K. R., Gleckler, P. J., Bonfils, C., and Taylor, K. E.: Quantifying the agreement between observed
1358   and simulated extratropical modes of interannual variability, Climate Dynamics, 52, 4057–4089,
1359   https://doi.org/10.1007/s00382-018-4355-4, 2019a.

1360 Lee, J., Sperber, K. R., Gleckler, P. J., Taylor, K. E., and Bonfils, C.: Benchmarking performance changes in the
1361   simulation of extratropical modes of variability across CMIP generations, Journal of Climate, 1–70,
1362   https://doi.org/10.1175/jcli-d-20-0832.1, 2021b.

1363   Lee, J., Xue, Y., De Sales, F., Diallo, I., Marx, L., Ek, M., Sperber, K. R., and Gleckler, P. J.: Evaluation of multi-
1364   decadal UCLA-CFSv2 simulation and impact of interactive atmospheric-ocean feedback on global and
1365   regional variability, Climate Dynamics, 52, 3683–3707, https://doi.org/10.1007/s00382-018-4351-8, 2019b.

1366   Leung, L. R., Boos, W. R., Catto, J. L., DeMott, C. A., Martin, G. M., Neelin, J. D., O'Brien, T. A., Xie, S., Feng, Z.,
1367   Klingaman, N. P. Kuo, Y.-H., Lee, R. W., Martinez-Villalobos, C., Vishnu S., Priestley, M. D. K., Tao, C.,
1368   and Zhou, Y.: Exploratory precipitation metrics: Spatiotemporal characteristics, process-oriented, and
1369   phenomena-based, Journal of Climate, 35, https://doi.org/10.1175/JCLI-D-21-0590.1, 3659-3686, 2022.

1370   Lin, J.-P., Kiladis, G. N., Mapes, B. E., Weickmann, K. M., Sperber, K. R., Lin, W., Wheeler, M. C., Schubert, S. D.,
1371   Del Genio, A. D., Donner, L. J., Emori, S., Guérémy, J.-F., Hourdin, F., Rasch, P. J., Roeckner, E., and
1372   Scinocca, J.: Tropical intraseasonal variability in 14 IPCC AR4 climate Models. Part I: Convective Signals,
1373   Journal of Climate, 19, 2665–2690, https://doi.org/10.1175/jcli3735.1, 2006.

1374   Lin, Y., Huang, X., Liang, Y., Qin, Y., Xu, S., Huang, W., Xu, F., Liu, L., Wang, Y., Peng, Y. and Wang, L.:
1375   Community integrated earth system model (CIESM): Description and evaluation, Journal of Advances in
1376   Modeling Earth Systems, 12, https://doi.org/10.1029/2019ms002036, 2020.

1377   Loeb, N. G., Doelling, D. R., Wang, H., Su, W., Nguyen, C., Corbett, J. G., Liang, L., Mitrescu, C., Rose, F. G., and
1378   Seiji, K.: Clouds and the Earth's Radiant Energy System (CERES) Energy Balanced and Filled (EBAF) top-
1379   of-atmosphere (TOA) Edition-4.0 data product, International Journal of Climatology, 31, 895–918,
1380   https://doi.org/10.1175/JCLI-D-17-0208.1, 2018.

1381   Longmate, J. M., Risser, M. D. and Feldman, D. R.: Prioritizing the selection of CMIP6 model ensemble members for
1382   downscaling projections of CONUS temperature and precipitation, Clim Dyn 61, 5171–5197,
1383   https://doi.org/10.1007/s00382-023-06846-z, 2023.

1384   Lu, L., Wang, W. and Tan, Z.: Double-arc parallel coordinates and its axes re-ordering methods, Mobile Networks
1385   and Applications, 25, 1376-1391, https://doi.org/10.1007/s11036-019-01455-9, 2020.

1386   Madden, R. A. and Julian, P.: Detection of a 40–50 day oscillation in the zonal wind in the Tropical Pacific, Journal
1387   of the Atmospheric Sciences, 28, 702–708, https://doi.org/10.1175/1520-0469(1971)028, 1971.

1388   Madden, R. A. and Julian, P.: Description of Global-Scale Circulation Cells in the Tropics with a 40–50 Day Period,
1389   Journal of the Atmospheric Sciences, 29, 1109–1123, https://doi.org/10.1175/1520-0469(1972)029, 1972.

1390   Madden, R. A. and Julian, P.: Observations of the 40–50-Day Tropical Oscillation—A Review, Monthly Weather
1391   Review, 122, 814–837, https://doi.org/10.1175/1520-0493(1994)122, 1994.

1392   Martin, G. M., Klingaman, N. P., and Moise, A. F.: Connecting spatial and temporal scales of tropical precipitation in
1393   observations and the MetUM-GA6, Geoscientific Model Development, 10, 105–126,
1394   https://doi.org/10.5194/gmd-10-105-2017, 2017.

1395   Maloney, E. D., Gettelman, A., Ming, Y., Neelin, J. D., Barrie, D., Mariotti, A., Chen, C., Coleman, D., Kuo, Y. H.,
1396   Singh, B., Annamalai, H., Berg, A., Booth, J. F., Camargo, S. J., Dai, A., Gonzalez, A., Hafner, J., Jiang, X.,
1397   Jing, X., Kim, D. H., Kumar, A., Moon, Y., Naud, C. M., Sobel, A. H., Suzuki, K., Wang, F., Wang, J., Wing,
1398   A. A., Xu, X., and Zhao, M.: Process-Oriented evaluation of climate and weather forecasting models, Bulletin
1399   of the American Meteorological Society, 100, 1665–1686, https://doi.org/10.1175/bams-d-18-0042.1, 2019.

1400 McAvaney, B.J., Covey, C., Joussaume, S., Kattsov, V., Kitoh, A., Ogana, W., Pitman, A.J., Weaver, A.J., Wood,
1401   R.A. and Zhao, Z.C.: Model evaluation. In Climate Change 2001: The scientific basis. Contribution of WG1
1402   to the Third Assessment Report of the IPCC (TAR) 471-523, Cambridge University Press, 2001.

1403 McPhaden, M. J., Zebiak, S. E., and Glantz, M. H.: ENSO as an integrating concept in Earth Science, Science, 314,
1404   1740–1745, https://doi.org/10.1126/science.1132588, 2006.

1405 McPhaden, M. J., Santoso, A., Cai, W. (Eds.): El Niño Southern oscillation in a changing climate, American
1406   Geophysical Union, USA, 528 pp., ISBN:9781119548126, https://doi.org/10.1002/9781119548164, 2020.

1407 Mears, C. A., Smith, D. K., Ricciardulli, L., Wang, J., Huelsing, H., & Wentz, F. J.: Construction and uncertainty
1408   estimation of a satellite-derived total precipitable water data record over the world's oceans, Earth and Space
1409   Science, 5, 197–210, https://doi.org/10.1002/2018EA000363, 2018.

1410 Meehl, G. A., Boer, G. J., Covey, C., Latif, M., and Stouffer, R. J.: The Coupled Model Intercomparison Project
1411   (CMIP), Bulletin of the American Meteorological Society, 81, 313–318, 2000.

1412 Meehl, G. A., Boer, G. J., Covey, C., Latif, M., and Stouffer, R. J.: Intercomparison makes for a better climate model,
1413   Eos, Transactions American Geophysical Union, 78, 445, https://doi.org/10.1029/97eo00276, 1997.

1414 Meehl, G. A., Covey, C., Delworth, T. L., Latif, M., McAvaney, B. J., Mitchell, J. F. B., Stouffer, R. J., and Taylor,
1415   K. E.: THE WCRP CMIP3 Multimodel Dataset: A new era in climate change research, Bulletin of the
1416   American Meteorological Society, 88, 1383–1394, https://doi.org/10.1175/bams-88-9-1383, 2007.

1417 Merrifield, A., Brunner, L., Lorenz, R., Humphrey, V., and Knutti, R.: Climate model Selection by Independence,
1418   Performance, and Spread (ClimSIPS v1.0.1) for regional applications, Geoscientific Model Development,
1419   16, 4715–4747, https://doi.org/10.5194/gmd-16-4715-2023, 2023.

1420 Neelin, J. D., Krasting, J. P., Radhakrishnan, A., Liptak, J., Jackson, T. J., Ming, Y., Dong, W., Gettelman, A.,
1421   Coleman, D., Maloney, E. D., Wing, A. A., Kuo, Y. H., Ahmed, F., Ullrich, P. A., Bitz, C. M., Neale, R.,
1422   Ordonez, A., and Maroon, E.: Process-oriented diagnostics: principles, practice, community development
1423   and common standards, Bulletin of the American Meteorological Society, https://doi.org/10.1175/bams-d-
1424   21-0268.1, 2023.

1425 Nowack, P., Runge, J., Eyring, V., and Haigh, J. D.: Causal networks for climate model evaluation and constrained
1426   projections, Nature Communications, 11, https://doi.org/10.1038/s41467-020-15195-y, 2020.

1427 Orbe, C., Van Roekel, L., Adames, Á. F., Dezfuli, A., Fasullo, J. T., Gleckler, P. J., Lee, J., Li, W., Nazarenko, L.,
1428   Schmidt, G. A., Sperber, K. R., and Zhao, M.: Representation of modes of variability in six U.S. climate
1429   models, Journal of Climate, 33, 7591–7617, https://doi.org/10.1175/jcli-d-19-0956.1, 2020.

1430 Ordonez, A. C., Klingaman, N. P., and Martin, G.: Analysing scales of precipitation, OSTI OAI (U.S. Department of
1431   Energy Office of Scientific and Technical Information), https://doi.org/10.11578/dc.20211029.5, 2021.

1432 Papalexiou, S. M., Rajulapati, C. R., Clark, M. P., and Lehner, F.: Robustness of CMIP6 historical global mean
1433   temperature simulations: Trends, long-term persistence, autocorrelation, and distributional shape, Earth's
1434   Future, 8, e2020EF001667, https://doi.org/10.1029/2020EF001667, 2020.

Pascoe, C., Lawrence, B. N., Guilyardi, E., Juckes, M., and Taylor, K. E.: Documenting numerical experiments in support of the Coupled Model Intercomparison Project Phase 6 (CMIP6), Geosci. Model Dev., 13, 2149–2167, https://doi.org/10.5194/gmd-13-2149-2020, 2020.

Pendergrass, A. G., Gleckler, P. J., Leung, L. R., and Jakob, C.: Benchmarking simulated precipitation in earth system models, Bulletin of the American Meteorological Society, 101, E814–E816, https://doi.org/10.1175/bams-d-19-0318.1, 2020.

Phillips, A. S., Deser, C., and Fasullo, J. T.: Evaluating modes of variability in climate models, Eos, Transactions American Geophysical Union, 95, 453–455, https://doi.org/10.1002/2014eo490002, 2014.

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

Planton, Y. Y., Lee, J., Wittenberg, A. T. Gleckler, P. J., Guilyardi, E., McGregor, S., and McPhaden, M. J.: Estimating uncertainty in simulated ENSO statistics, Journal of Advances in Modeling Earth Systems (under review), ESS Open Archive, https://doi.org10.22541/essoar.170196744.48068128/v1, 2023.

Potter, G. L., Bader, D. C., Riches, M., Bamzai, A. and Joseph, R.: Celebrating two decades of the Program for Climate Model Diagnosis and Intercomparison. Bulletin of the American Meteorological Society, 92, 629-631, https://doi.org/10.1175/2011BAMS3018.1, 2011.

Qin, Y., Zelinka, M. D., and Klein, S. A.: On the Correspondence Between Atmosphere-Only and Coupled Simulations for Radiative Feedbacks and Forcing From CO_2, Journal of Geophysical Research: Atmospheres, 127, https://doi.org/10.1029/2021jd035460, 2022.

Randall, D.A., Wood, R.A., Bony, S., Colman, R., Fichefet, T., Fyfe, J., Kattsov, V., Pitman, A., Shukla, J., Srinivasan, J. and Stouffer, R.J.: Climate models and their evaluation. In Climate change 2007: The physical science basis. Contribution of Working Group I to the Fourth Assessment Report of the IPCC (FAR), 589-662, Cambridge University Press, 2007.

Randerson, J. T., Hoffman, F. M., Thornton, P. E., Mahowald, N. M., Lindsay, K., Lee, Y.-H., Nevison, C. D., Doney, S. C., Bonan, G. B., Stöckli, R., Covey, C., Running, S. W., and Fung, I.: Systematic assessment of terrestrial biogeochemistry in coupled climate-carbon models, Global Change Biology, 15, 2462–2484, https://doi.org/10.1111/j.1365-2486.2009.01912.x, 2009.

Rasch, P. J., Xie, S., Ma, P.-L., Lin, W., Wang, H., Tang, Q., Burrows, S. M., Caldwell, P., Zhang, K., Easter, R. C., Cameron-Smith, P. J., Singh, B., Wan, H., Golaz, J.-C., Harrop, B. E., Roesler, E. L., Bacmeister, J. T., Larson, V. E., Evans, K. J., Qian, Y., Taylor, M. A., Leung, L. R., Zhang, Y., Brent, L., Branstetter, M. L., Hannay, C., Mahajan, S., Mametjanov, A., Neale, R., Richter, J. H., Yoon, J.-H., Zender, C. S., Bader, D. C., Flanner, M., Foucar, J. G., Jacob, R., Keen, N. D., Klein, S. A., Liu, X., Salinger, A. G., Shrivastava, M., and Yang, Y.: An overview of the atmospheric component of the Energy Exascale Earth System model, Journal of Advances in Modeling Earth Systems, 11, 2377–2411, https://doi.org/10.1029/2019ms001629, 2019.

1471 Reichler, T. and Kim, J.: How well do coupled models simulate today's climate?, Bulletin of the American
1472   Meteorological Society, 89, 303–312, https://doi.org/10.1175/bams-89-3-303, 2008.

1473 Righi, M., Andela, B., Eyring, V., Lauer, A., Predoi, V., Schlund, M., Vegas-Regidor, J., Bock, L., Brötz, B., De
1474   Mora, L., Diblen, F., Dreyer, L., Drost, N., Earnshaw, P., Hassler, B., Koldunov, N., Little, B., Tomas, S. L.,
1475   and Zimmermann, K.: Earth System Model Evaluation Tool (ESMValTool) v2.0 – technical overview,
1476   Geoscientific Model Development, 13, 1179–1199, https://doi.org/10.5194/gmd-13-1179-2020, 2020.

1477 Sanderson, B. M. and Wehner, M. F.: Weighting strategy for the Fourth National Climate Assessment, in: Climate
1478   Science Special Report: Fourth National Climate Assessment, Volume I, edited by Wuebbles, D. J., Fahey,
1479   D. W., Hibbard, K. A., Dokken, D. J., Stewart, B. C., and Maycock, T.K., U.S. Global Change Research
1480   Program, Washington, DC, USA, 436-442, https://doi.org/10.7930/J06T0JS3, 2017.

1481 Sanderson, B. M., Wehner, M., and Knutti, R.: Skill and independence weighting for multi-model assessments,
1482   Geoscientific Model Development, 10, 2379–2395, https://doi.org/10.5194/gmd-10-2379-2017, 2017.

1483 Sherwood, S. C., Webb, M. J., Annan, J. D., Armour, K. C., Forster, P. M., Hargreaves, J. C., Hegerl, G. C., Klein, S.
1484   A., Marvel, K., Rohling, E. J., Watanabe, M., Andrews, T., Braconnot, P., Bretherton, C. S., Foster, G. L.,
1485   Hausfather, Z., Von Der Heydt, A. S., Knutti, R., Mauritsen, T., Norris, J. R., Proistosescu, C., Rugenstein,
1486   M., Schmidt, G. A., Tokarska, K., and Zelinka, M. D.: An assessment of Earth's climate sensitivity using
1487   multiple lines of evidence, Reviews of Geophysics, 58, https://doi.org/10.1029/2019rg000678, 2020.

1488 Singh, R., and AchutaRao, K.: Sensitivity of future climate change and uncertainty over India to
1489   performance-based model weighting. Clim. Change, https://doi.org/10.1007/s10584-019-02643-y, 2020.

1490 Sillmann, J., Kharin, V. V., Zhang, X., Zwiers, F. W., and Bronaugh, D.: Climate extremes indices in the CMIP5
1491   multimodel ensemble: Part 1. Model evaluation in the present climate, Journal of Geophysical Research:
1492   Atmospheres, 118, 1716–1733, https://doi.org/10.1002/jgrd.50203, 2013.

1493 Sperber, K. R.: Madden-Julian variability in NCAR CAM2.0 and CCSM2.0, Clim Dyn 23, 259–278,
1494   https://doi.org/10.1007/s00382-004-0447-4, 2004.

1495 Sperber, K. R., Annamalai, H., Kang, I.-S., Kitoh, A., Moise, A., Turner, A., Wang, B., and Zhou, T.: The Asian
1496   summer monsoon: an intercomparison of CMIP5 vs. CMIP3 simulation of the late 20th century. Clim Dyn
1497   41, 2711-2744, https://doi.org/10.1007/s00382-012-1607-6, 2013.

1498 Sperber K. R., Gualdi, S., Legutke, S., Gayler, V.: The Madden–Julian oscillation in ECHAM4 coupled and uncoupled
1499   general circulation models, Clim Dyn 25, 117–140, https://doi.org/10.1007/s00382-005-0026-3, 2005.

1500 Srivastava, A., Grotjahn, R., and Ullrich, P. A.: Evaluation of historical CMIP6 model simulations of extreme
1501   precipitation over contiguous US regions, Weather and Climate Extremes, 29, 100268,
1502   https://doi.org/10.1016/j.wace.2020.100268, 2020.

1503 Steed, C. A., Shipman, G., Thornton, P., Ricciuto, D., Erickson, D. and Branstetter, M.: Practical application of parallel
1504   coordinates for climate model analysis, Procedia Computer Science, 9, 877-886,
1505   https://doi.org/10.1016/j.procs.2012.04.094, 2012.

1506 Stevens, B., Satoh, M., Auger, L., Biercamp, J., Bretherton, C. S., Chen, X., Düben, P., Judt, F., Khairoutdinov, M.,
1507   Klocke, D., Kodama, C., Kornblueh, L., Lin, S.-J., Neumann, P., Putman, W. M., Röber, N., Shibuya, R.,

1508   Vanniere, B., Vidale, P. L., Wedi, N., and Zhou, L.: DYAMOND: the DYnamics of the Atmospheric general
1509   circulation Modeled On Non-hydrostatic Domains, Progress in Earth and Planetary Science, 6,
1510   https://doi.org/10.1186/s40645-019-0304-z, 2019.
1511   Stoner, A. M. K., Hayhoe, K., and Wuebbles, D. J.: Assessing general circulation model simulations of atmospheric
1512   teleconnection patterns, Journal of Climate, 22, 4348–4372, https://doi.org/10.1175/2009jcli2577.1, 2009.
1513   Sung, H. M., Kim, J., Shim, S., Seo, J., Kwon, S.-H., Sun, M.-A., Moon, H.-J., Lee, J., Lim, Y. C., Boo, K.-O., Kim,
1514   Y., Lee, J., Lee, J., Kim, J.-S., Marzin, C., and Byun, Y.-H.: Climate change projection in the Twenty-First
1515   Century simulated by NIMS-KMA CMIP6 model based on new GHGs concentration pathways, Asia-pacific
1516   Journal of Atmospheric Sciences, 57, 851–862, https://doi.org/10.1007/s13143-021-00225-6, 2021.
1517   Tang, Q., Prather, M. J., Hsu, J., Ruiz, D. J., Cameron-Smith, P. J., Xie, S., and Golaz, J.-C.: Evaluation of the
1518   interactive stratospheric ozone (O3v2) module in the E3SM version 1 Earth system model, Geoscientific
1519   Model Development, 14, 1219–1236, https://doi.org/10.5194/gmd-14-1219-2021, 2021.
1520   Tang, S., Fast, J. D., Zhang, K., Hardin, J. C., Varble, A. C., Shilling, J. E., Mei, F., Zawadowicz, M. A., and Ma, P.-
1521   L.: Earth System Model Aerosol–Cloud Diagnostics (ESMAC Diags) package, version 1: assessing E3SM
1522   aerosol predictions using aircraft, ship, and surface measurements, Geosci. Model Dev., 15, 4055–4076,
1523   https://doi.org/10.5194/gmd-15-4055-2022, 2022.
1524   Tang, S., Varble, A. C., Fast, J. D., Zhang, K., Wu, P., Dong, X., Mei, F., Pekour, M., Hardin, J. C., and Ma, P.-L.:
1525   Earth System Model Aerosol–Cloud Diagnostics (ESMAC Diags) package, version 2: assessing aerosols,
1526   clouds, and aerosol–cloud interactions via field campaign and long-term observations, Geosci. Model Dev.,
1527   16, 6355–6376, https://doi.org/10.5194/gmd-16-6355-2023, 2023.
1528   Taylor, K. E.: Truly Conserving with Conservative Remapping Methods, Geosci. Model Dev. Discuss. [preprint],
1529   https://doi.org/10.5194/gmd-2023-177, in review, 2023.
1530   Taylor, K. E.: Summarizing multiple aspects of model performance in a single diagram, Journal of Geophysical
1531   Research, 106, 7183–7192, https://doi.org/10.1029/2000jd900719, 2001.
1532   Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An overview of CMIP5 and the experiment design, Bulletin of the
1533   American Meteorological Society, 93, 485–498, https://doi.org/10.1175/bams-d-11-00094.1, 2012.
1534   Teixeira, J., Waliser, D. E., Ferraro, R., Gleckler, P. J., Lee, T., and Potter, G. L.: Satellite observations for CMIP5:
1535   The Genesis of OBS4MIPs, Bulletin of the American Meteorological Society, 95, 1329–1334,
1536   https://doi.org/10.1175/bams-d-12-00204.1, 2014.
1537   Tian, B., and Dong, X.: The Double-ITCZ Bias in CMIP3, CMIP5, and CMIP6 Models Based on Annual Mean
1538   Precipitation, Geophysical Research Letters, 47, e2020GL087232, https://doi.org/10.1029/2020GL087232,
1539   2020
1540   Ullrich, P. A. and Zarzycki, C. M.: TempestExtremes: a framework for scale-insensitive pointwise feature tracking on
1541   unstructured grids, Geoscientific Model Development, 10, 1069–1090, https://doi.org/10.5194/gmd-10-
1542   1069-2017, 2017.

1543 Ullrich, P. A., Zarzycki, C. M., McClenny, E., Pinheiro, M. C., Stansfield, A. M., and Reed, K. A.: TempestExtremes
1544   v2.1: a community framework for feature detection, tracking, and analysis in large datasets, Geoscientific
1545   Model Development, 14, 5023–5048, https://doi.org/10.5194/gmd-14-5023-2021, 2021.

1546 U.S. Department of Energy (DOE): Benchmarking Simulated Precipitation in Earth System Models Workshop Report,
1547   DOE/SC-0203, U.S. Department of Energy Office of Science, Biological and Environmental Research (BER)
1548   Program. Germantown, Maryland, USA. 2020.

1549 Vo, T., Po-Chedley, P., Boutte, J., Zhang, C., Lee, J., Gleckler, P., Durack, P., Taylor, K., and Golaz, J.-C.: Xarray
1550   Climate Data Analysis Tools (xCDAT): A Python Package for Simple and Robust Analysis of Climate Data,
1551   The 103rd AMS Annual Meeting, Abstract, 2023.

1552 Waliser, D. E., Gleckler, P. J., Ferraro, R., Taylor, K. E., Ames, S., Biard, J., Bosilovich, M. G., Brown, O. B., Chepfer,
1553   H., Cinquini, L., Durack, P. J., Eyring, V., Mathieu, P.-P., Lee, T., Pinnock, S., Potter, G. L., Rixen, M.,
1554   Saunders, R., Schulz, J. B., Thépaut, J.-N., and Tuma, M.: Observations for Model Intercomparison Project
1555   (Obs4MIPs): status for CMIP6, Geoscientific Model Development, 13, 2945–2958,
1556   https://doi.org/10.5194/gmd-13-2945-2020, 2020.

1557 Waliser, D. E., Sperber, K. R., Hendon, H. H., Kim, D., Maloney, E. D., Wheeler, M. C., Weickmann, K. M., Zhang,
1558   C., Donner, L. J., Gottschalck, J., Higgins, W., Kang, I. S., Legler, D. M., Moncrieff, M. W., Schubert, S. D.,
1559   Stern, W., Vitart, F., Wang, B., Wang, W., and Woolnough, S. J.: MJO Simulation Diagnostics, Journal of
1560   Climate, 22, 3006–3030, https://doi.org/10.1175/2008jcli2731.1, 2009.

1561 Wang, J., Liu, X., Shen, H. W. and Lin, G.: Multi-resolution climate ensemble parameter analysis with nested parallel
1562   coordinates plots, IEEE Transactions on Visualization and Computer Graphics, 23, 81-90,
1563   https://doi.org/10.1109/TVCG.2016.2598830, 2017.

1564 Wehner, M., Gleckler, P. J., and Lee, J.: Characterization of long period return values of extreme daily temperature
1565   and precipitation in the CMIP6 models: Part 1, model evaluation, Weather and Climate Extremes, 30,
1566   100283, https://doi.org/10.1016/j.wace.2020.100283, 2020.

1567 Wehner, M., Lee, J., Risser, M. D., Ullrich, P. A., Gleckler, P. J., and Collins, W. D.: Evaluation of extreme sub-daily
1568   precipitation in high-resolution global climate model simulations, Philosophical Transactions of the Royal
1569   Society A, 379, 20190545, https://doi.org/10.1098/rsta.2019.0545, 2021.

1570 Whitehall, K., Mattmann, C., Waliser, D., Kim, J., Goodale, C., Hart, A., Ramirez, P., Zimdars, P., Crichton, D.,
1571   Jenkins, G., Jones, C., Asrar, G., and Hewitson, B.: Building Model Evaluation and Decision Support
1572   Capacity for CORDEX, WMO Bulletin, 61, available at:
1573   https://public.wmo.int/en/resources/bulletin/building-model-evaluation-and-decision-support-capacity-
1574   cordex (last access date: 14 September 2023), 2012.

1575 Williams, D. N.: Visualization and analysis tools for ultrascale climate data, Eos, Transactions American Geophysical
1576   Union, 95, 377–378, https://doi.org/10.1002/2014eo420002, 2014.

1577 Williams, D. N., Balaji, V., Cinquini, L., Denvil, S., Duffy, D. Q., Evans, B., Ferraro, R., Hansen, R., Lautenschlager,
1578   M., and Trenham, C.: A global repository for Planet-Sized experiments and observations, Bulletin of the
1579   American Meteorological Society, 97, 803–816, https://doi.org/10.1175/bams-d-15-00132.1, 2016.

Williams, D. N., Doutriaux, C., Drach, R., and McCoy, R.: The Flexible Climate Data Analysis Tools (CDAT) for Multi-model Climate Simulation Data, IEEE International Conference on Data Mining Workshops, 254–261, https://doi.org/10.1109/icdmw.2009.64, 2009.

Wong, P. C., Shen, H. W., Leung, R., Hagos, S., Lee, T. Y., Tong, X. and Lu, K.: Visual analytics of large-scale climate model data, in: 2014 IEEE 4th Symposium on Large Data Analysis and Visualization (LDAV), 85-92, https://doi.org/10.1109/LDAV.2014.7013208, 2014.

Xie, P., Joyce, R., Wu, S., Yoo, S.H., Yarosh, Y., Sun, F. and Lin, R.: Reprocessed, bias-corrected CMORPH global high-resolution precipitation estimates from 1998, Journal of Hydrometeorology, 18, 1617-1641, 2017.

Xue, Z. and Ullrich, P. A.: A Comprehensive Intermediate-Term Drought Evaluation System and Evaluation of Climate Data Products over the Conterminous United States, Journal of Hydrometeorology, https://doi.org/10.1175/jhm-d-20-0314.1, 2021.

Young, A. H., Knapp, K. R., Inamdar, A. K., Hankins, W., and Rossow, W. B.: The International Satellite Cloud Climatology Project H-Series climate data record product, Earth System Science Data, 10, 583–593, https://doi.org/10.5194/essd-10-583-2018, 2018.

Zelinka, M. D., Klein, S. A., Qin, Y., and Myers, T. A.: Evaluating climate models' cloud feedbacks against expert judgment, Journal of Geophysical Research: Atmospheres, 127, https://doi.org/10.1029/2021jd035198, 2022.

Zelinka, M. D., Myers, T. A., McCoy, D. T., Po-Chedley, S., Caldwell, P. M., Ceppi, P., Klein, S. A., and Taylor, K. E.: Causes of higher climate sensitivity in CMIP6 models, Geophysical Research Letters, 47, e2019GL085782, https://doi.org/10.1029/2019GL085782, 2020.

Zhang, C., Golaz, J.-C., Forsyth, R., Vo, T., Xie, S., Shaheen, Z., Potter, G. L., Asay-Davis, X. S., Zender, C. S., Lin, W., Chen, C.-C., Terai, C. R., Mahajan, S., Zhou, T., Balaguru, K., Tang, Q., Tao, C., Zhang, Y., Emmenegger, T., Burrows, S., and Ullrich, P. A.: The E3SM Diagnostics Package (E3SM Diags v2.7): a Python-based diagnostics package for Earth system model evaluation, Geosci. Model Dev., 15, 9031–9056, https://doi.org/10.5194/gmd-15-9031-2022, 2022.

Zhang, C. and Hendon, H. H.: Propagating and standing components of the intraseasonal oscillation in tropical convection, Journal of the Atmospheric Sciences, 54, 741–752, https://doi.org/10.1175/1520-0469(1997)054, 1997.

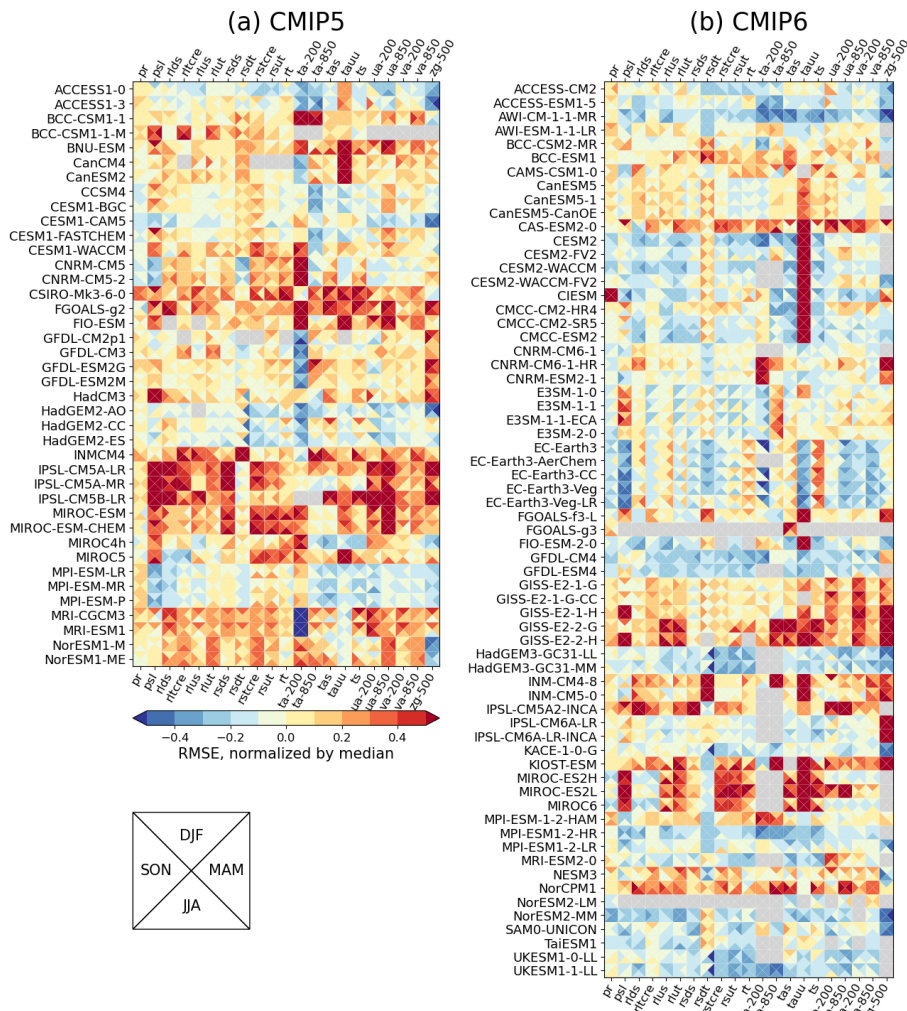
Zhang, C., Xie, S., Klein, S. A., Ma, H. Y., Tang, S., Van Weverberg, K., Morcrette, C. J. and Petch, J.: CAUSES: Diagnosis of the summertime warm bias in CMIP5 climate models at the ARM Southern Great Plains site. Journal of Geophysical Research: Atmospheres, 123, 2968-2992, https://doi.org/10.1002/2017JD027200, 2018.

Zhang, C., Xie, S., Tao, C., Tang, S., Emmenegger, T., Neelin, J. D., Schiro, K. A., Lin, W. and Shaheen, Z.: The ARM data-oriented metrics and diagnostics package for climate models: A new tool for evaluating climate models with field data, https://doi.org/10.1175/BAMS-D-19-0282.1, Bulletin of the American Meteorological Society, 101, E1619-E1627, 2020.

1616 Zhao, B., Lin, P., Hu, A., Liu, H., Ding, M., Yu, Z., and Yu, Y.: Uncertainty in Atlantic Multidecadal Oscillation
1617   derived from different observed datasets and their possible causes, Frontiers in Marine Science, 9,
1618   https://doi.org/10.3389/fmars.2022.1007646, 2022.
1619 Zhao, M., Golaz, J.-C., Held, I. M., Guo, H., Balaji, V., Benson, R., Chen, J. H., Chen, X., Donner, L. J., Dunne, J.,
1620   Dunne, K. A., Durachta, J., Fan, S.-M., Freidenreich, S. M., Garner, S. T., Ginoux, P., Harris, L., Horowitz,
1621   L. W., Krasting, J. P., Langenhorst, A. R., Zhi, L., Lin, P., Lin, S. J., Malyshev, S., Mason, E., Milly, P. C.
1622   D., Ming, Y., Naik, V., Paulot, F., Paynter, D., Phillipps, P. J., Radhakrishnan, A., Ramaswamy, V.,
1623   Robinson, T., Schwarzkopf, D., Seman, C. J., Shevliakova, E., Shen, Z., Shin, H. H., Silvers, L. G., Wilson,
1624   J. R., Winton, M., Wittenberg, A. T., Wyman, B., and Xiang, B.: The GFDL Global Atmosphere and Land
1625   Model AM4.0/LM4.0: 1. Simulation characteristics with prescribed SSTS, Journal of Advances in Modeling
1626   Earth Systems, 10, 691–734, https://doi.org/10.1002/2017ms001208, 2018.
1627

Table 1. List of variables and observation datasets used as reference datasets for the PMP's mean climate evaluation in this paper (Section 3.1 and Figs. 1-2). A ditto mark (") indicates the same as above.

Variable	Variable full name	Product	Reference
ps	Precipitation	GPCP-2-3	Adler et al. (2018)
psl	Sea level pressure	ERA-5	Hersbach et al. (2020)
rlds	Surface Downwelling Longwave Radiation	CERES-EBAF-4-1	Loeb et al. (2018)
rltcre	Longwave cloud radiative effect	"	
rlus	Surface Upwelling Longwave Radiation	"	
rlut	Upwelling longwave at the top of atmosphere	"	
rsds	Surface Downwelling Shortwave Radiation	"	
rsdt	TOA Incident Shortwave Radiation	"	
rstcre	Shortwave cloud radiative effect	"	
rsut	Upwelling shortwave at the top of atmosphere	"	
rt	Net radiative flux	"	
ta-200, ta-850	Air temperature at 850 and 200 hPa	ERA-5	Hersbach et al. (2020)
tas	2-m air temperature	"	
tauu	Surface zonal wind stress	ERA-INT	Dee et al. (2011)
ts	Surface temperature	ERA-5	Hersbach et al. (2020)
ua-200, ua-850	Zonal wind component at 850 and 200 hPa	"	
va-200, va-850	Meridional wind component at 850 and 200 hPa	"	
zg-500	Geopotential height at 500 hPa	"	

Formatted Table

1637
Figure 1. Portrait plot for spatial RMSE (uncentered) of global seasonal climatologies for (a)
CMIP5 (models ACCESS1-0 to NorESM1-ME on the ordinate) and (b) CMIP6 (models
ACCESS-CM2 to UKESM1-1-LL on the ordinate) for 1981-2005 epoch. The RMSE is calculated
for each season (shown as triangles in each box) over the globe including both land and ocean,
and model and reference data were interpolated to a common 2.5x2.5 degree grid. The RMSE
of each variable is normalized by the median RMSE of all CMIP5 and 6 models. A result of 0.2
(-0.2) is indicative of an error that is 20% greater (lesser) than the median RMSE across all
models. Models in each group are sorted in alphabetical order. Full names of variable names on
the abscissa and their reference datasets can be found in Table 1. Detailed information for

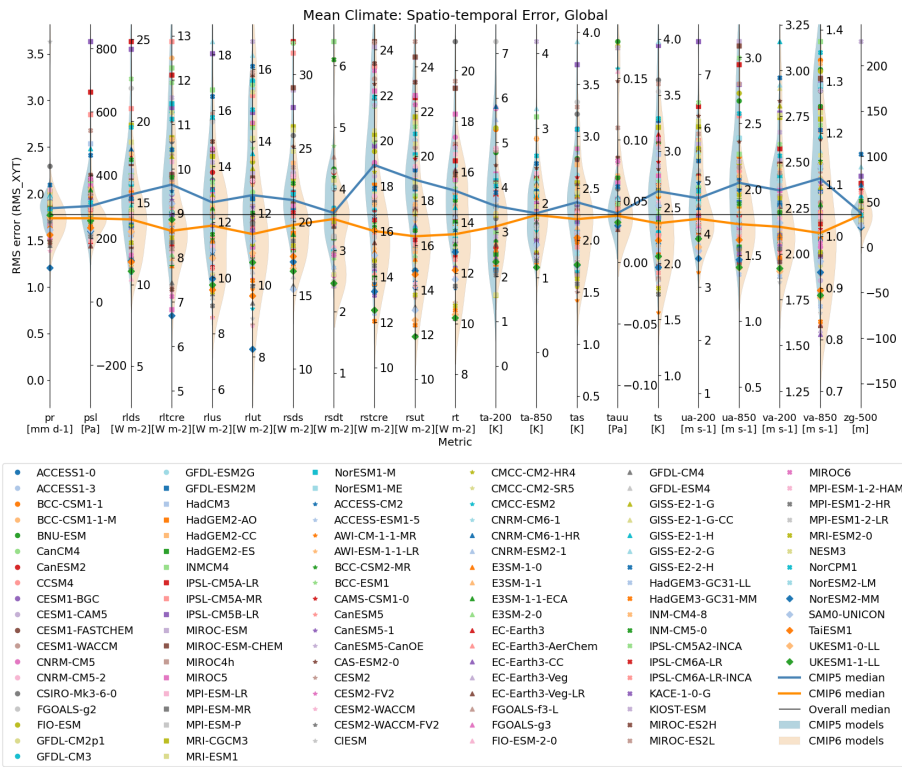
1647 models can be found at the *Earth System Documentation* (ES-DOC, https://search.es-doc.org/;
1648 Pascoe et al., 2020). The interactive version of the Portrait plot in this figure is available on the
1649 PMP result pages on the PCMDI website (https://pcmdi.llnl.gov/metrics/mean_clim/).

1651
1652
Figure 2. Parallel Coordinate Plot for spatio-temporal RMSE (Gleckler et al., 2008) from mean climate evaluation. Each vertical axis represents a different variable. Results from each model are displayed as symbols. Middle of each vertical axis is aligned with the median statistic of all CMIP5 and CMIP6 models. The cross-generation model distributions of model performance are shaded on the left (CMIP5, blue) and right (CMIP6, orange) sides of each axis. Also, medians from CMIP5 (blue) and CMIP6 (orange) model groups are highlighted as lines. Full names for model variables on the abscissa and their reference datasets can be found in Table 1. Time epoch used for this analysis is 1981-2005. Detailed information for models can be found at the *Earth System Documentation* (ES-DOC, https://search.es-doc.org/; Pascoe et al., 2020). The interactive version of the Portrait plot in this figure is available on the PMP result pages on the PCMDI website (https://pcmdi.llnl.gov/metrics/mean_clim/).

Figure 3. Application of ENSO metrics to CMIP6 models. Model names with an asterisk (*) indicate that 10 or more ensemble members were used in this analysis. Dots indicate metric values from individual ensemble members while bars indicate the average of metric values across the ensemble members. Bars colored for easier identification of model names at the bottom of the figure. Metrics were grouped into three *Metrics Collections:* (a-n) ENSO Performance, (o-r) ENSO Teleconnections, and (s-w) ENSO processes. Names of individual metrics and default reference datasets being used are noted on top of each panel, and observational uncertainty by applying the metrics for alternative reference datasets noted on the upper right of each panel is shown as gray-shaded. Detailed descriptions for each metric can be found at https://github.com/CLIVAR-PRP/ENSO_metrics/wiki.

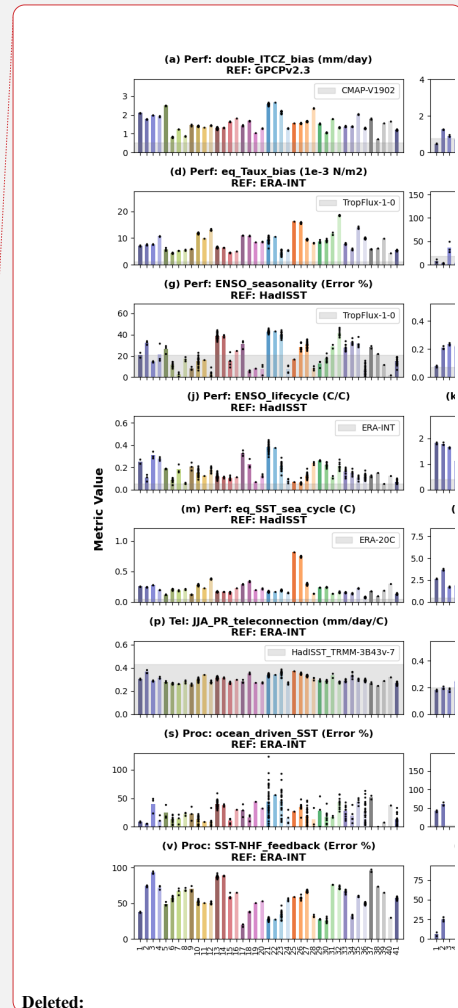
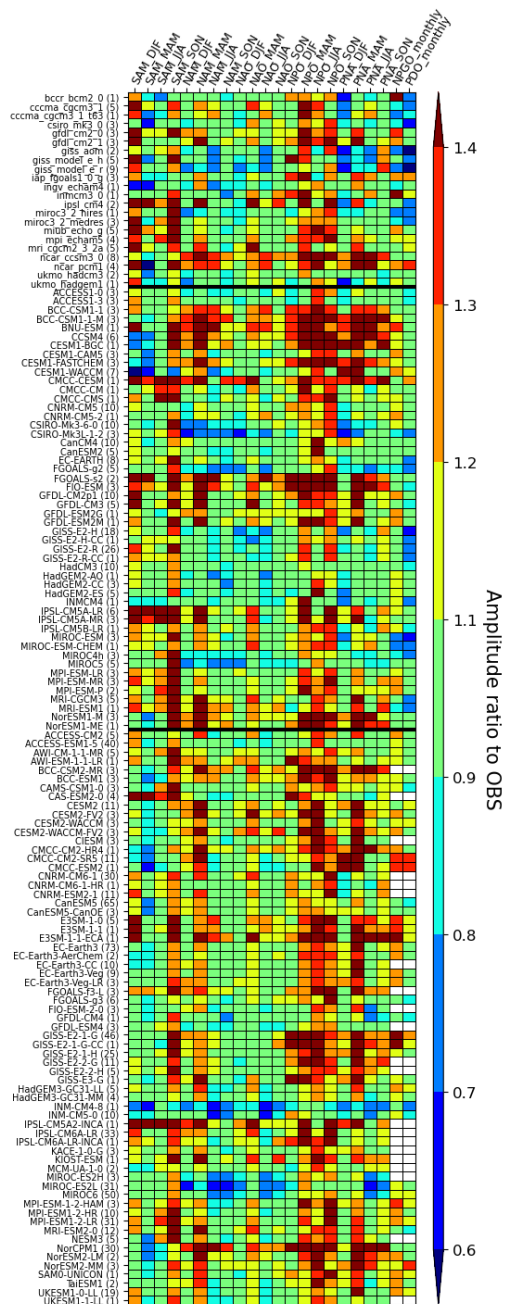



Figure 4. Portrait plots of the amplitude of extratropical modes of variability simulated by CMIP3, 5, and 6 models in their historical or equivalent simulations, as gauged by the ratio of spatiotemporal standard deviations of the model and observed PCs, obtained using the CBF method in the PMP. Columns (horizontal axis) are for mode and season, and rows (vertical axis) are for models from CMIP3 (top), CMIP5 (middle), and CMIP6 (bottom), separated by thick black horizontal lines. For sea level pressure–based modes (SAM, NAM, NAO, NPO, and PNA) in the upper-left hand triangle the model results are shown relative to NOAA-20CR. For SST-based modes (NPGO and PDO), results are shown relative to HadISSTv1.1. Numbers in parentheses following model names indicate the number of ensemble members for the model. Metrics for individual ensemble members were averaged for each model. White boxes indicate missing value.

Deleted: *<object>*¶

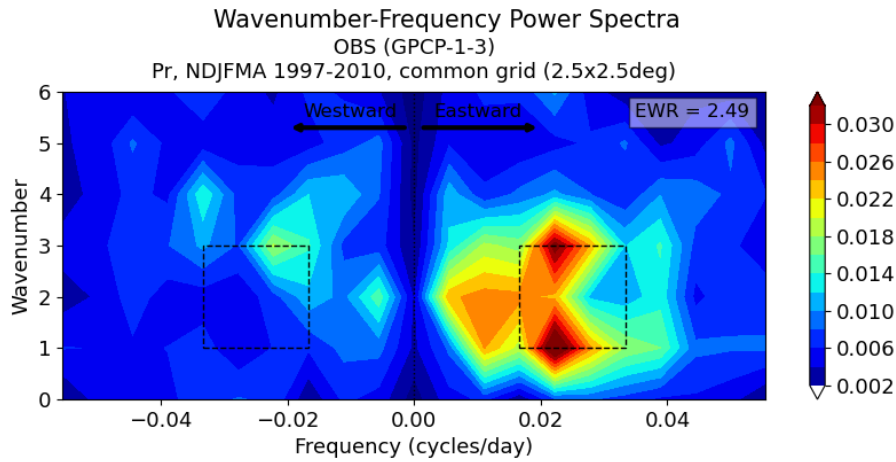
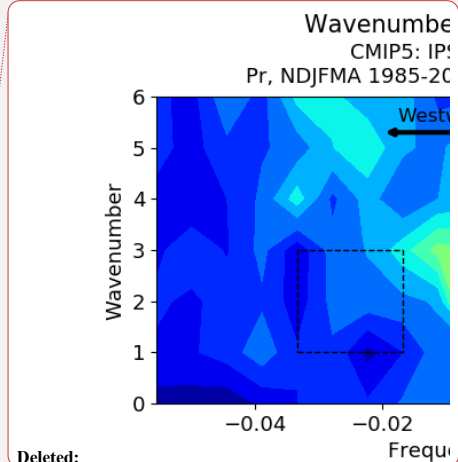
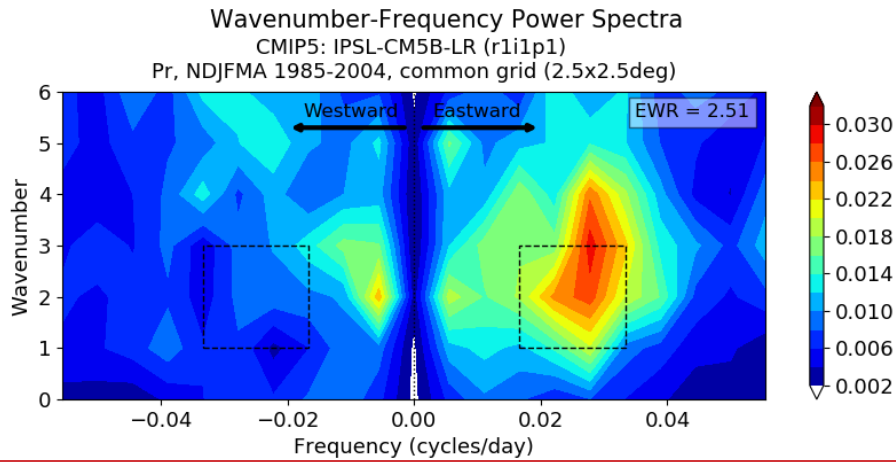
Deleted: rows of gray boxes.

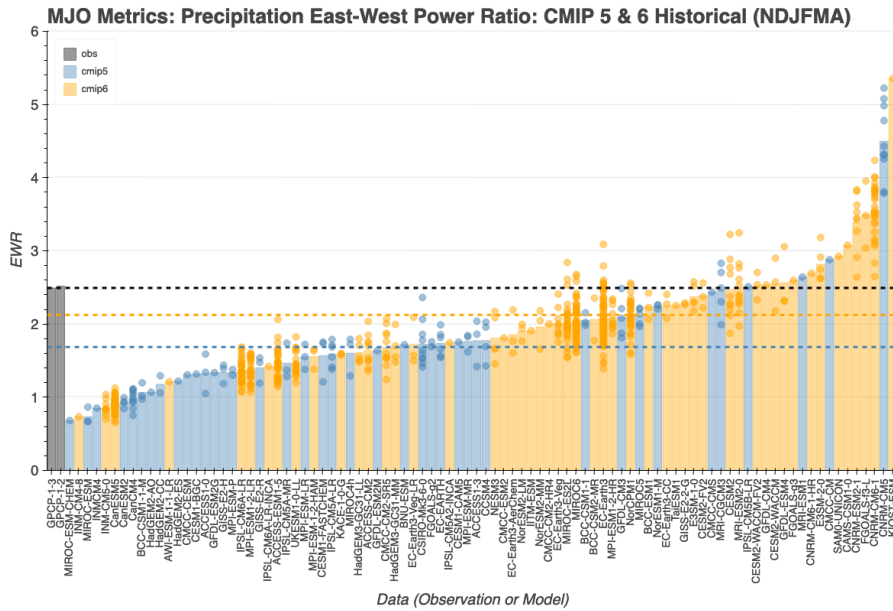
Deleted: whereas in the lower-right triangle, the model results are shown relative to the ERA-20C.

Deleted: (upper-left triangle) and HadISSTv2.1 (lower-right triangle).

Deleted: The figure is adapted from Lee et al. 2021b.

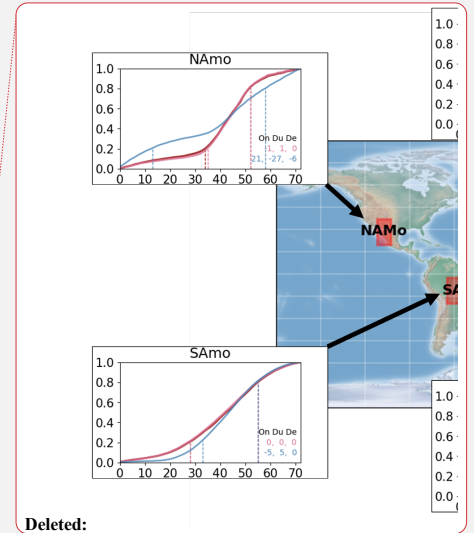
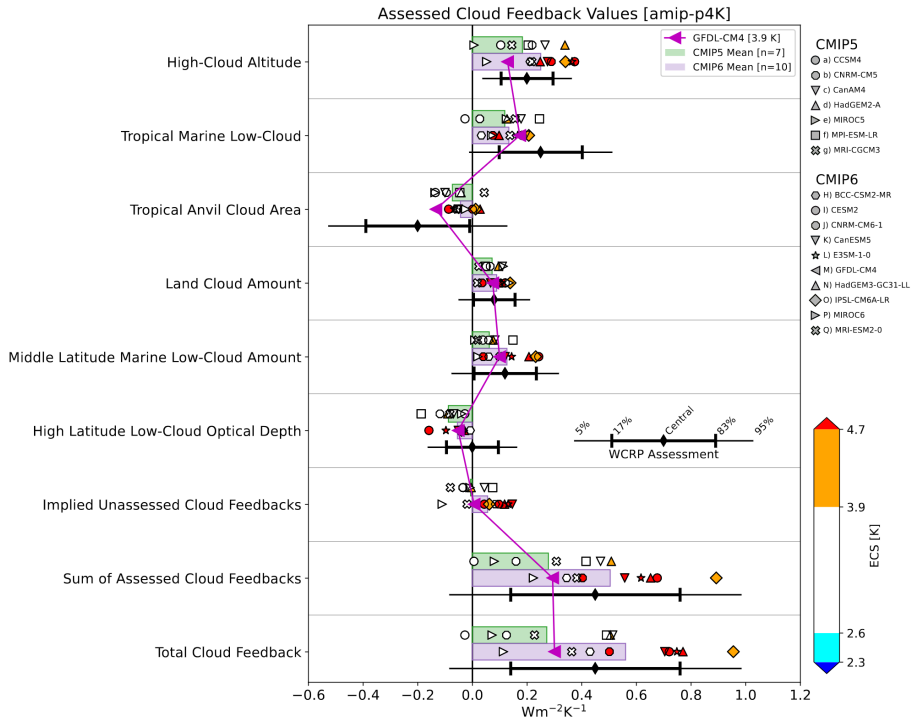
1711

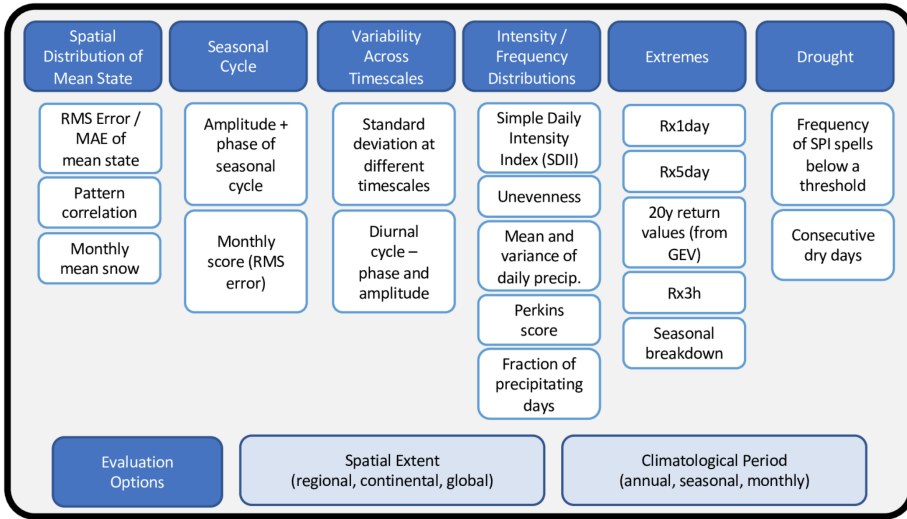
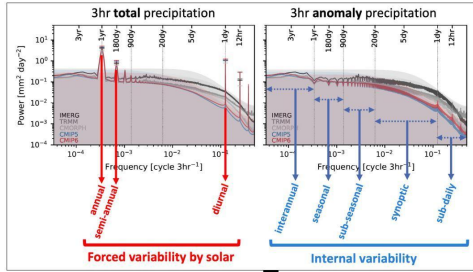
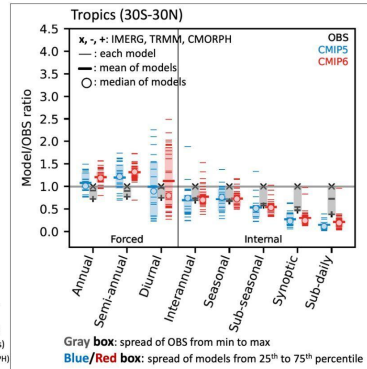
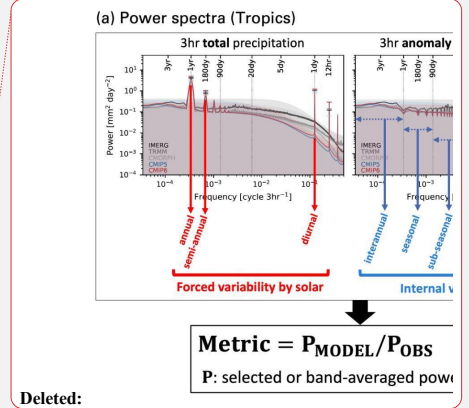
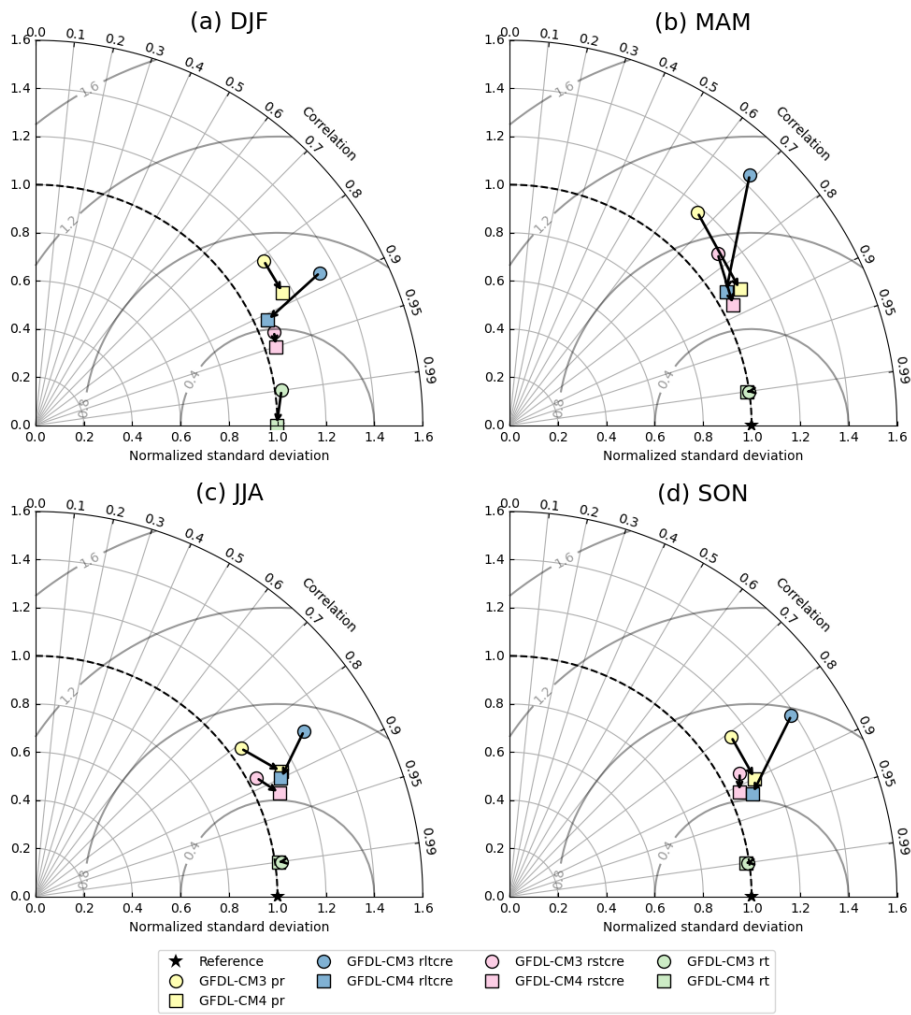
1712 (a) Observation

1713
1714 (b) Model

1715
1716
1717 **Figure 5.** MJO EWR diagnostics – wavenumber-frequency power spectra – from (a) GPCP v1.3
1718 (Huffman et al., 2001) and (b) IPSL-CM5B-LR model of CMIP5. The EWR is defined as the ratio
1719 of eastward power (averaged in the box on the right) to westward power (averaged in the box
1720 on the left) from the 2-dimensional wavenumber-frequency power spectra of daily 10°S–10°N
1721 averaged precipitation in November to April (shaded, mm^2 day^{-2}). Power spectra are calculated
1722 for each year and then averaged over all years of data. The units of power spectra for the
1723 precipitation is mm^2 day^{-2} per frequency interval per wavenumber interval.

MJO Metrics: Precipitation East-West Power Ratio: CMIP 5 & 6 Historical (NDJFMA)

1728
1729
1730   **Figure 6.** MJO East-West Power Ratio (EWR, *unitless*) from CMIP5 and CMIP6 models,
1731   models in two different groups (CMIP5: blue, CMIP6: orange) are sorted by the value of the
1732   metric and compared to two observation datasets (purple, GPCP v1.2 and v1.3; Huffman et al.,
1733   2001). Horizontal dashed lines indicate EWR from the default primary reference observation
1734   (i.e., GPCP v1.3, black), averages of CMIP5 and CMIP6 models. The interactive plot is
1735   available at https://pcmdi.llnl.gov/research/metrics/mjo/ where the horizontal axis can be
1736   resorted by CMIP group or model names as well. Hover mouse over boxes will show tooltips for
1737   metric values and a preview of dive-down plots that are shown in Figure 5.
1738

1739

Figure 7. Demonstration of the monsoon metrics obtained from observation datasets (GPCP
v1.3 and CMORPH v1.0 (Joyce et al., 2004; Xie et al., 2017)) and a CMIP6 model's Historical
simulation conducted using CNRM-CM6-1. The results are obtained for monsoon regions: All-
India Rainfall (AIR), Sahel, Gulf of Guinea (GoG), North American Monsoon (NAM), South
American Monsoon (SAM), and Northern Australia (AUS). The regions are defined in Sperber
and Annamalai (2014). Metrics for onset (On), Duration (Du), and Decay (De) derived as
differences to the default observation (GPCP v1.3) in pentad indices (observation minus model)
are shown at lower right of each panel. Pentad indices for onset and decay of each region are
also shown as vertical lines.

1740
1741
1742
1743
1744
1745
1746
1747
1748

1750

1751
Figure 8. Cloud feedback components estimated in amip-p4K simulations from CMIP5 and
CMIP6 models. Symbols indicate individual model values, while horizontal bars indicate multi-
model means. Each model is color-coded by its ECS, with color boundaries corresponding to the
likely and very likely ranges of ECS as determined in Sherwood et al (2020). Each component's
expert-assessed likely and very likely confidence intervals are indicated with black error bars. An
illustrative model (GFDL-CM4) is highlighted.

1758

1759
1760
1761
1762 **Figure 9.** Proposed suite of baseline metrics for simulated precipitation benchmarking (figure
1763 reprinted from workshop report; US DOE, 2020).
1764
1765

1766
1767
1768 **Figure 10.** Example (a) an underlying diagnostic and (b) its resulting metrics for precipitation
1769 variability across timescales. (a) Power spectra of 3-hourly total (left) and anomaly (right)
1770 precipitation from IMERG (black), TRMM (gray), CMORPH (silver), CMIP5 (blue), and CMIP6
1771 (red) averaged over the tropics (30°S-30°N). The colored shading indicates the 95% confidence
1772 interval for each observational product and for the CMIP5 and CMIP6 means. (b) Metrics for
1773 forced and internal precipitation variability based on power spectra. The reference observational
1774 product displayed is GPM IMERG (Huffman et al., 2015). The gray boxes represent the spread
1775 of the three observational products ("X" for IMERG, "-" for TRMM, and "+" for CMORPH) from
1776 the minimum to maximum values. Blue and red boxes indicate the 25th to the 75th percentile of
1777 CMIP models as a measure of spread. Individual models are shown as thin dashes, the
1778 multimodel mean as a thick dash, and the multimodel median as an open circle. Details for the
1779 diagnostics and metrics are described in Ahn et al. (2022).
1780

1782
1783 **Figure 11.** Taylor Diagram contrasting performance of an ESM in their two different versions
1784 (i.e., GFDL-CM3 from CMIP5 and GFDL-CM4 from CMIP6) in its Historical simulation for
1785 multiple variables (precipitation [pr], longwave cloud radiative effect [rltcre], shortwave cloud
1786 radiative effect [rstcre], and total radiation flux [rt]) in their climatology over the globe for (a) DJF,
1787 (b) MAM, (c) JJA and (d) SON seasons. The arrow is directed toward the newer version of the
1788 model from the older version (i.e., GFDL-CM3 → GFDL-CM4).

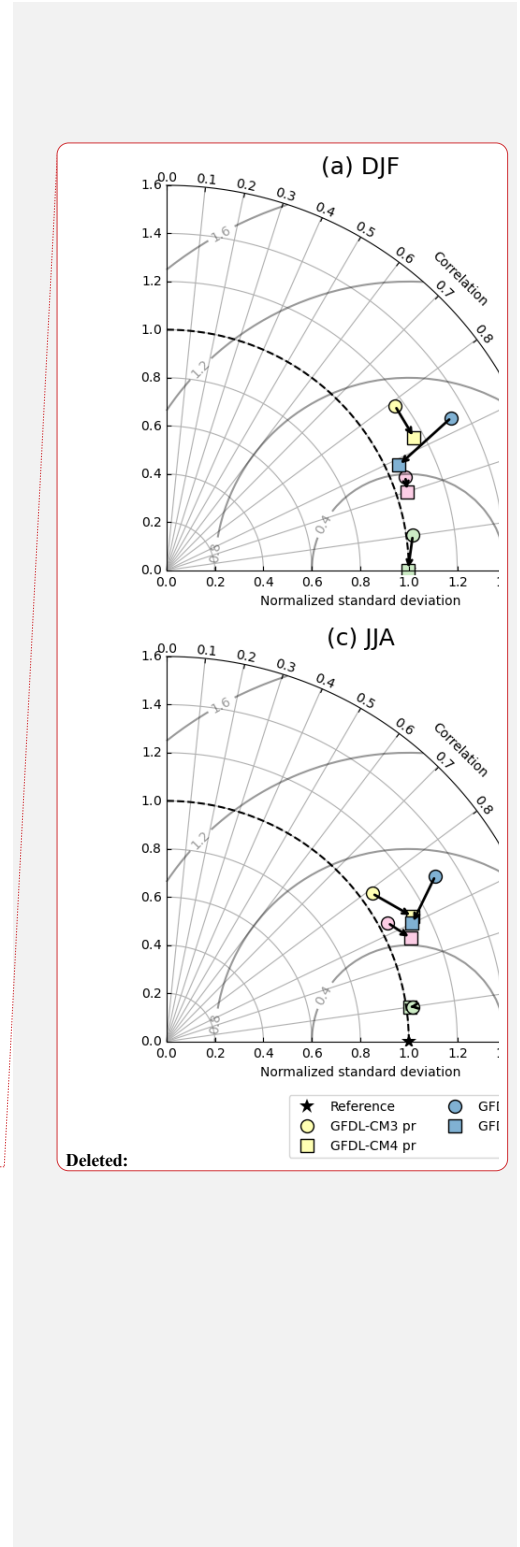
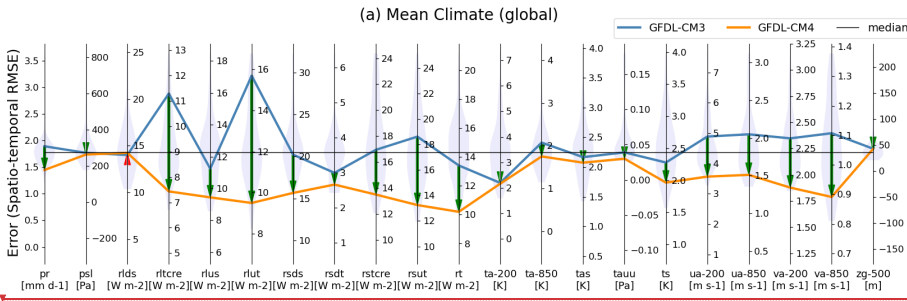
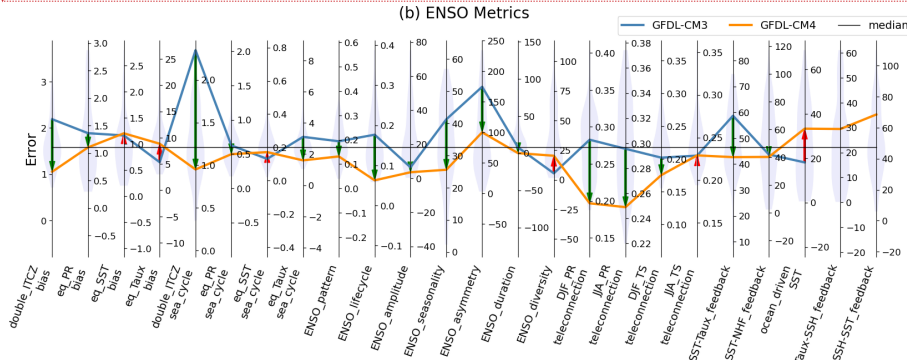
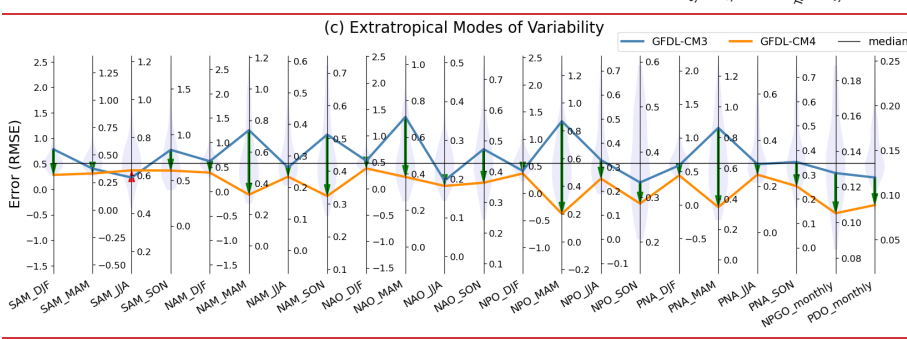
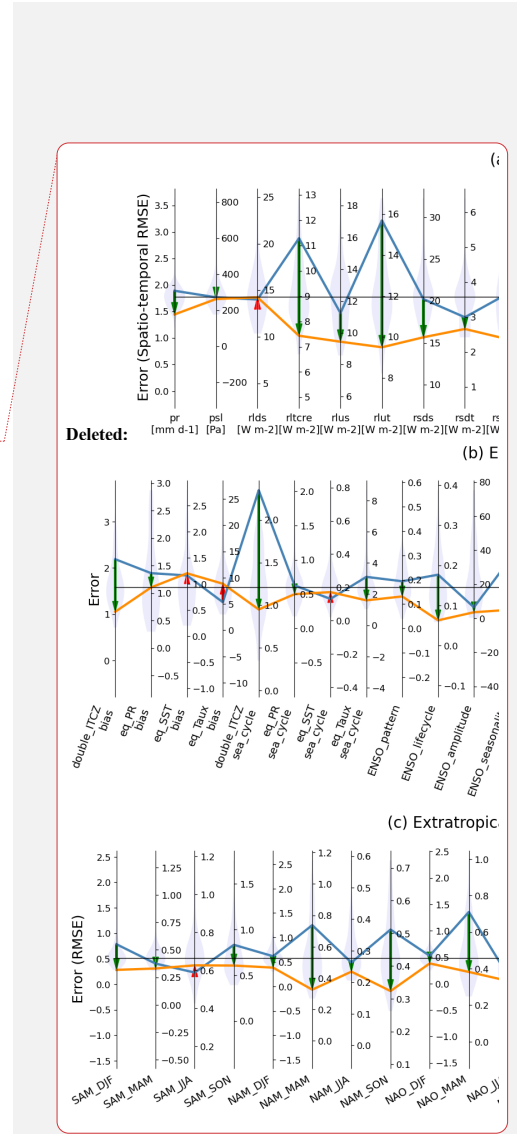



Figure 12. Parallel Coordinate Plot contrasting performance of two different versions of the GFDL model (i.e., GFDL-CM3 from CMIP5 and GFDL-CM4 from CMIP6) in their Historical experiment for errors from (a) mean climate, (b) ENSO, and (c) extratropical modes of variability. Improvement (degradation) in the later version of the model is highlighted as a downward green (upward red) arrow between lines. Middle of each vertical axis is set to the median of the group of benchmarking models (i.e., CMIP5 and CMIP6), with the axis range stretched to maximum distance to either minimum or maximum from the median for visual consistency. The inter-model distributions of model performance are shown as shaded violin plots along each vertical axis.

Page 17: [1] Deleted	Lee, Jiwoo	3/9/24 11:42:00 AM

Page 17: [2] Deleted	Lee, Jiwoo	3/9/24 11:42:00 AM

Page 17: [3] Deleted	Lee, Jiwoo	3/9/24 11:42:00 AM

Page 18: [4] Deleted	Lee, Jiwoo	3/9/24 11:42:00 AM

Page 18: [5] Deleted	Lee, Jiwoo	3/9/24 11:42:00 AM

Page 18: [6] Deleted	Lee, Jiwoo	3/9/24 11:42:00 AM

Page 18: [7] Deleted	Lee, Jiwoo	3/9/24 11:42:00 AM

Page 18: [8] Deleted	Lee, Jiwoo	3/9/24 11:42:00 AM

Page 18: [9] Deleted	Lee, Jiwoo	3/9/24 11:42:00 AM

Page 18: [10] Deleted	Lee, Jiwoo	3/9/24 11:42:00 AM

Page 18: [11] Deleted	Lee, Jiwoo	3/9/24 11:42:00 AM

Page 18: [12] Deleted	Lee, Jiwoo	3/9/24 11:42:00 AM

Page 18: [13] Deleted	Lee, Jiwoo	3/9/24 11:42:00 AM