Process-oriented diagnostics: principles, practice, community development and common standards

J. David Neelin,[a] John P. Krasting,[b] Aparna Radhakrishnan[c], Jessica Liptak,[b] Thomas Jackson,[d] Yi Ming,[b,e] Wenhao Dong,[b,f] Andrew Gettelman,[g,h] Danielle R. Coleman,[g] Eric D. Maloney,[i] Allison A. Wing,[j] Yi-Hung Kuo,[a,c] Fiaz Ahmed,[a] Paul Ullrich,[k] Cecilia M. Bitz,[l] Richard B. Neale,[g] Ana Ordonez,[m] Elizabeth A. Maroon[n]

[a] *University of California, Los Angeles, Los Angeles, CA*

[b] NOAA/OAR *Geophysical Fluid Dynamics Laboratory, Princeton, NJ*

[c] *Program in Atmospheric and Oceanic Sciences, Princeton University, Princeton, NJ*

[d] *Science Applications International Corporation, Reston, VA*

[e] *Schiller Institute for Integrated Science and Society, Boston College, Boston MA*

[f] *Cooperative Programs for the Advancement of Earth System Science, UCAR, Boulder, CO*

[g] *National Center for Atmospheric Research, Boulder, CO*

[h] *Pacific Northwest National Laboratory, Richland, WA*

[i] *Colorado State University, Fort Collins, CO*

[j] *Florida State University, Tallahassee, FL*

[k] *University of California, Davis, Davis, CA*

[l] *University of Washington, Seattle, WA*

[m] *Lawrence Livermore National Laboratory, Livermore, CA*

[n] *University of Wisconsin-Madison, Madison, WI*

Corresponding author: J. David Neelin, neelin@atmos.ucla.edu

Early Online Release: This preliminary version has been accepted for publication in *Bulletin of the American Meteorological Society*, may be fully cited, and has been assigned DOI 10.1175/BAMS-D-21-0268.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

ABSTRACT

Process-oriented diagnostics (PODs) aim to provide feedback for model developers through model analysis based on physical hypotheses. However, the step from a diagnostic based on relationships among variables, even when hypothesis-driven, to specific guidance for revising model formulation or parameterizations can be substantial. The POD may provide more information than a purely performance-based metric, but a gap between POD principles and providing actionable information for specific model revisions can remain. Furthermore, in coordinating diagnostics development, there is a trade-off between freedom for the developer, aiming to capture innovation, and near-term utility to the modeling center. Best practices that allow for the former, while conforming to specifications that aid the latter, are important for community diagnostics development that leads to tangible model improvements. Promising directions to close the gap between principles and practice include the interaction of PODs with perturbed physics experiments and with more quantitative process models as well as the inclusion of personnel from modeling centers in diagnostics development groups for immediate feedback during climate model revisions. Examples are provided, along with best-practice recommendations, based on practical experience from the NOAA Model Diagnostics Task Force (MDTF). Common standards for metrics and diagnostics that have arisen from a collaboration between the MDTF and the Department of Energy's Coordinated Model Evaluation Capability are advocated as a means of uniting community diagnostics efforts.

CAPSULE

Experience-based recommendations are provided for ongoing community development of process-oriented diagnostics. These include promising directions for more actionable process information for model improvement and common standards for metrics and diagnostics framework

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-21-0268.1.

Introduction

In the quest to improve climate and weather model simulations, process-oriented diagnostics (PODs) have been advocated as a means of providing more information to model developers beyond performance-based metrics. A "process-oriented diagnostic" (Eyring et al. 2005; Sperber and Waliser 2008; Maloney et al. 2014; Kim et al. 2014; Eyring et al. 2019; Maloney et al. 2019), or POD, characterizes a physical process that is hypothesized to be related to the ability of a model to simulate an observed phenomenon. Evaluating a candidate model version against observations analyzed with such a POD can, in principle, give insight into whether a particular process is being well represented, focus model improvement on specific processes, and identify gaps in the understanding of phenomena. However, moving from principles to practice is a nontrivial step that requires thoughtful implementation. Here we draw on experience with the NOAA Model Diagnostics Task Force (MDTF) to provide best-practice recommendations for entraining diagnostics from a broad scientific community to make them available to model developers.

Developing a suitably comprehensive and useful package of diagnostics to enable effective climate model validation is a challenge. The set of phenomena that a numerical weather, climate, or Earth System model (ESM) is expected to capture is continually expanding, and the observational datasets to which the model can be compared continue to be expanded and improved (Teixeira et al. 2014; Eyring et al. 2016a). A model development team normally has a set of diagnostics from prior phases of model development but has limited resources for maintenance and further development of those diagnostics. Legacy diagnostics packages can quickly become outdated as new observational datasets are developed and are often associated with a particular developer who may have moved on from the organization. It can thus be highly advantageous to have a mechanism by which diagnostics development in a wider set of research groups can be brought into a coherent framework for use by the model development group.

Translating knowledge from a comprehensive diagnostics package into actionable information to engender model improvements is also a challenge. The bias associated with a particular misrepresented process may have expressions in the basic state climatology or statistics of variability (e.g., Hwang and Frierson 2013; Rosa and Collins 2013; Grose et al. 2014; Held et al. 2019; Voldoire et al. 2019; Tatebe et al. 2019; Danabasoglu et al. 2020; Kelley et al. 2020; Boucher et al. 2020). When a bias in the simulation of a particular

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-21-0268.1.

phenomenon has been identified in performance metrics (typically scalar metrics or bias plots assessing model performance compared to observations or reanalyses; Gleckler et al. 2008), it can be arduous to identify the underlying physical process. Working backward from a given error in a simulation to the process producing it is often a sustained effort involving a dialogue between theory for mechanisms and new diagnostics for observation-model comparison. Diagnostics that initially aim to provide new process information can become performance metrics as this dialogue moves to the next level of process evaluation. Some simulation issues can be stubbornly resistant to improvement (e.g., Eyring et al. 2018; Tian and Dong 2020), while other issues can be affected by multiple potential process revisions that are difficult to disambiguate using existing diagnostics. An added complication in translating process diagnosis to model improvement is that compensating errors in models often exist, particularly when different Earth System Model components are coupled together. For example, improving one aspect of model performance such as tropical precipitation variability or particular measures of cloud feedback through improved process representation can degrade more fundamental measures of model performance such as the tropical mean state (Kim et al. 2011) or equilibrium climates sensitivity (Zelinka et al. 2020).

The NOAA MDTF, renewed for its current phase in 2021-2024, was constituted to implement a process-based model evaluation diagnostics framework for use in national modeling center diagnostics packages, coordinating inputs from diagnostics developers at other institutions. This paper describes lessons learned and formulates a set of recommendations based on the experience of moving a heterogeneous set of diagnostics groups and modeling centers toward a community process-oriented diagnostics framework. The outline of the paper is as follows. We first provide more background on the MDTF, including initial successes and challenges in building this model diagnostics community. Examples of diagnostics used in model development led to a discussion of both expanding the scope of climate processes covered by the diagnostics and for directions in future diagnostic developments that interface more strongly with process models or perturbed physics experiments. MDTF efforts related to use of common community software such as feature trackers and diagnostics that require specialized output not available under standard output CMIP protocols are also highlighted. Current efforts to support common community standards and protocols that aid the dissemination of diagnostics across modeling centers are outlined, along with a set of recommendations to aid community development more broadly.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-21-0268.1.

Task force and diagnostics community development

MDTF Background.

The MDTF effort was launched in response to the need for modeling centers to expand the amount of process-oriented information in their diagnostics packages to foster model improvement. A description of activities related to the first phase of the task force can be found in Maloney et al. (2019). Time constraints on model development staff make it difficult to individually engage with multiple community members and diagnostic efforts, so the effort aims to streamline the process for entraining diagnostics from the academic community into modeling-center workflows, beginning with GFDL and NCAR, in a way that is repeatable across multiple model versions during the development process. Diagnostics development teams within MDTF, mandated to provide process-oriented information directly relevant to physical parameterizations in the model components, are selected by peer review in open funding calls.

The essential design feature of the MDTF framework (Fig. 1) is to provide a protocol for the POD developer teams, such that each POD can be ingested relatively easily into the development stream of the modeling centers.

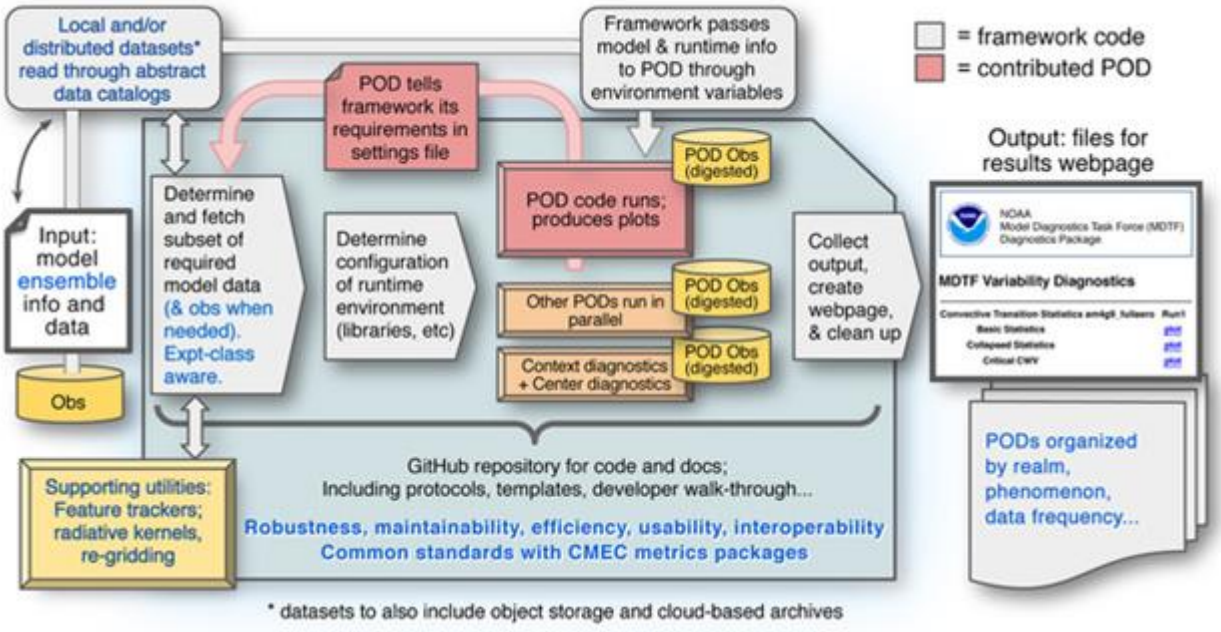

Fig. 1: Diagram of the MDTF framework evolving under the current phase of development. The framework manages a set of Process-Oriented Diagnostics modules (PODs) contributed by a variety of diagnostic development teams. Coordinated standards facilitate the inclusion/exchange of metrics and diagnostics with other parts of the US diagnostics community.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-21-0268.1.

Key features of this protocol include: A Python driver script sets up paths, variable names, etc. for the model data to be analyzed. It calls process-oriented diagnostics modules (PODs) contributed by various groups; these yield plots (and associated data for generating them); each group provides the observational comparison for its POD as a compact dataset that has been "digested", or processed, to make the comparison plots; these plots are then composed into web pages that permit easy comparison of the candidate model and observations. The details of the format vary according to the POD, but each provides the developer with model-to-observation comparisons for a process of interest. The PODs must be open source and follow the specified conventions for interacting with the framework. PODs can be repeatedly applied in modeling center workflows and are focused on model improvement. Community development makes use of GitHub in which POD developers download the framework, test a diagnostic, and then submit to the repository, contributing to the library of diagnostics. A streamlined process for contributing diagnostics to the GitHub repository with automated test suites and documentation as part of code reviews enables building a robust framework, while also fostering collaborations between the POD team and the MDTF framework team. Technical details of the updated MDTF framework are provided in the documentation linked from the sidebar "*Accessing and developing diagnostics for the MDTF framework*", which provides information on accessing and contributing to the diagnostics.

[Sidebar 1 near here]

Building the community.

The MDTF has had significant success bringing together participants from GFDL and NCAR model development groups and science-oriented POD developers in regular contact (monthly for the task force and weekly for the leadership team) in a bottom-up effort to improve model evaluation. The MDTF framework evolved from the very lightweight and idealized concept discussed in Maloney et al. (2019) where considerations of coding language, style, and other aspects were left to the POD developers, to one that is more standardized, efficient, and useful. Interaction with diagnostic development groups in the framework has been a two-way street. Adaptations and extensions to the framework have been needed to support capabilities requested by the POD developers, while the guidelines for formatting, documentation, and training from the leadership team in the key software

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-21-0268.1.

tools has helped improve the PODs. The science of the PODs has also improved through similar interactions.

Recognizing that other agencies have a need for broad evaluation of climate models and datasets, a key connection has been made to the US Department of Energy (DOE) Coordinated Model Evaluation Capabilities (CMEC) project. CMEC intends to harmonize DOE investments in model evaluation, providing access to DOE packages such as the PCMDI Metrics Package (PMP), Toolkit for Extreme Climate Analysis (TECA 2017), and the International Land Model Benchmarking Tool (ILAMB), along with other community-contributed metrics packages. Connectivity between MDTF and CMEC has supported the development of common standards aimed at broadening the community of users and diagnostic developers, as elaborated below.

However, the experience of the MDTF has also identified a set of issues to address in moving forward with a robust community-based POD development effort. These and the associated recommendations are documented here because they are relevant to the diagnostics and modeling community at large. Science-focused development teams may not be familiar with emerging coding practices and tools, nor do they necessarily have a critical need to optimize code for efficiency in a package running many modules, or to plan for future updates to the observational datasets and for scaling with increasing resolution. Protocols and tools that are standard at modeling centers may be relatively foreign to many science-focused POD developers. In addition to extensive documentation, tutorial materials and interactive training sessions prove necessary. These must be iterated due to continued onboarding of new personnel such as students in academic settings and the need for just-in-time training given the tendency for science personnel to pay attention to software considerations primarily at the time needed in their development trajectory.

A more fundamental aspect of building community comes in the science interactions, where there are several challenges. The fundamental motivations and incentives of academic investigators may differ from those of the modeling centers. While academic investigators recognize in principle that delivery of software is a useful contribution and are genuinely interested in improving climate models, academic incentives favor the prioritization of peer-reviewed papers. Standards for referencing the code contribution of PODs can be of benefit—*if* this is recognized by the larger community. While groups have an interest in seeing PODs used, maintaining code beyond the timeline of the funded project (and the departure of the

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-21-0268.1.

group member who developed the code) can be challenging for Principal Investigators (PIs). There can be tension between what is funded by peer review and the immediate needs for the current stage of model development. Even PODs that provide keen insight into dynamics of the climate system and model deficiencies in simulating these processes may not immediately translate into a path forward for model improvement. Uncertainty of observational products can be crucial, but diagnostics developers may have limited ability to characterize this for a specific diagnostic. Finally, journal articles are essentially a one-way communication from a POD development group to peers, but diagnostics can be most beneficial to model development when they evolve from a two-way interaction between POD developers and modeling groups that tightens the connection to the target processes and related parameterizations.

The following points have been found useful or are advocated to address some of these challenges, as elaborated in the recommendation section. Early career scientists benefit from task-force engagement in visibility to their science community; this can be enhanced by authorship in science task-force presentations and associated conference presentations, acknowledgments in POD documentation, and recognition for such activities as advocated in Recommendations. Clear documentation standards for PODs, assisted by templates, include a section in which developers summarize the embodied science. Liaisons from the MDTF leadership team or model development groups can provide an independent review of usability, ideally extending to the usefulness of the science insights from the PODs for model development and characterization of observational uncertainty. These connections help foster relationships between POD development groups and model developers. Funding calls that select POD developers can prioritize specific ties to model improvement in the science content of the contributed PODs.

Expanding directions

PODs in the development process

Examples of PODs used in the model development process are useful to illustrate the role we seek to broaden. These are chosen from the development of NCAR Community Atmosphere Model version (CAM) CAM6 (Simpson et al., 2020) and GFDL AM4 (Zhao et al. 2018a,b), CM4 (Held et al. 2019), and ESM4 (Dunne et al. 2020) models, and so are from early contributions to the framework.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-21-0268.1.

Figure 2 shows an example from the diurnal cycle of precipitation POD used in the CAM6 development cycle. This diagnostic is relatively simple in terms of detailed process information but helps highlight the role of time dependence in the development of precipitating systems. Figure 2 illustrates that tropical convection peaks in the early morning (blue colors) over the ocean and in the evening (green colors) over tropical land, with a unique cycle of sea-breeze circulations over the maritime continent. CAM5 (Fig. 2c) has a decent diurnal cycle over the ocean, but peaks over land are too early in the day. In CAM6, largely due to increased stability sensitivity in the deep convection scheme and the implementation of a new unified moist turbulence scheme, which has prognostic and time-evolving shallow convective motions, the diurnal cycle peak over land shifts several hours later, with an improvement in the maritime-continent land timing. There is also evidence of the sea breeze circulation. This diagnostic also detects a slight degradation/weakening (lighter colors) in the amplitude of the diurnal cycle over land.

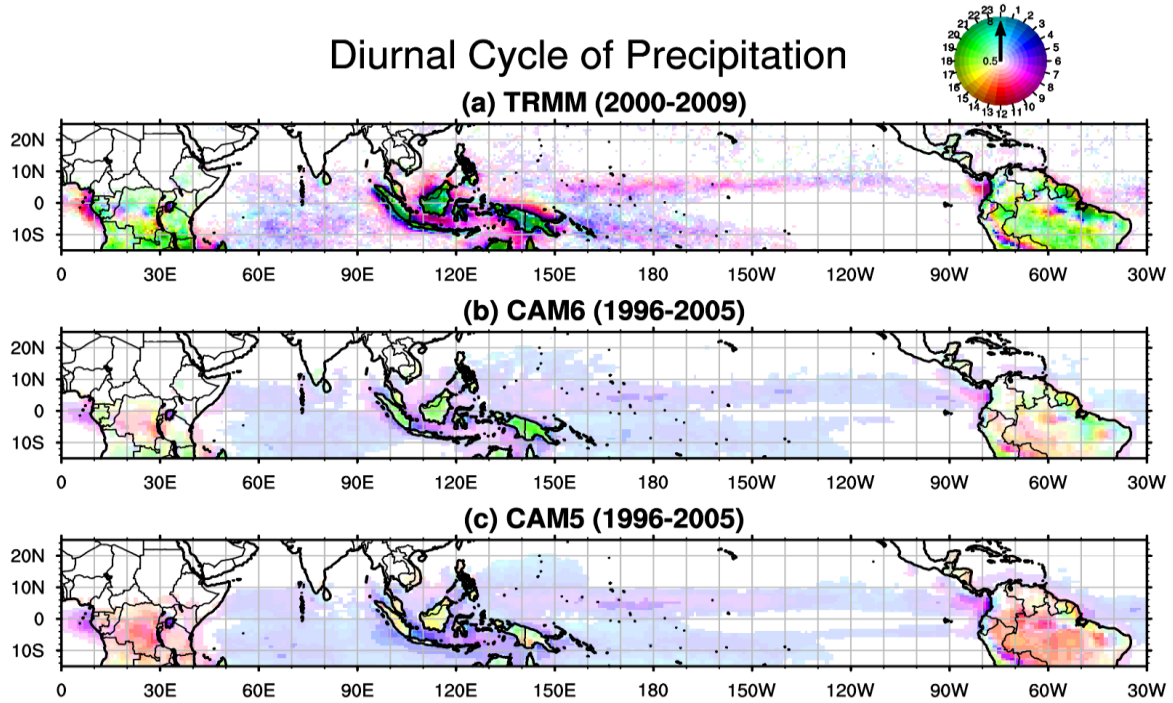

Fig. 2. Examples from the precipitation diurnal cycle POD showing phase-amplitude diagrams from (a) observational reference (using Tropical Rainfall Measurement Mission 3B42 data set); (b) Community Atmospheric Model (CAM6); (c) the earlier version, CAM5. Colors indicate the local hour of maximum precipitation with color intensity indicating diurnal cycle amplitude.

Fig. 3 shows an example of a POD for the Madden-Julian oscillation (MJO), applied to the GFDL model in atmosphere-only, coupled ocean-atmosphere and earth-system versions,

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-21-0268.1.

AM4.0, CM4.0, and ESM4.0, respectively. This POD was also used in the development of NCAR CAM6/CESM2 (Danabasoglu et al 2020). The POD relates propagation characteristics of the MJO to the spatial pattern of moisture in the climatology of the equatorial Indian Ocean-Western Pacific region shown in Fig. 3c. Moisture pattern skill is measured by pattern correlations of simulated 650–900 hPa averaged winter mean moisture against ERA-I reanalysis over the Maritime Continent following Gonzalez and Jiang (2017). MJO propagation skill is measured by pattern correlations of simulated anomalous rainfall Hovmöller diagrams, against their TRMM counterparts following Jiang (2017). MJO propagation skill based on TRMM and moisture pattern skill based on ERA-I is denoted by the black dot. Another two datasets (GPCP and MSWEPV2 precipitation datasets for MJO propagation skill and ERA5 and MERRA2 specific humidity for moisture pattern skill), indicated by square and triangle, respectively, are also included as an indicator of uncertainty in the target observational estimates. A pre-existing multimodel ensemble from the MJO Task Force (Jiang et al. 2015) is used to provide context. This example illustrates surprises in the use of PODs: improvement in MJO propagation skill is seen in coupled versions despite little change in moisture pattern skill.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-21-0268.1.

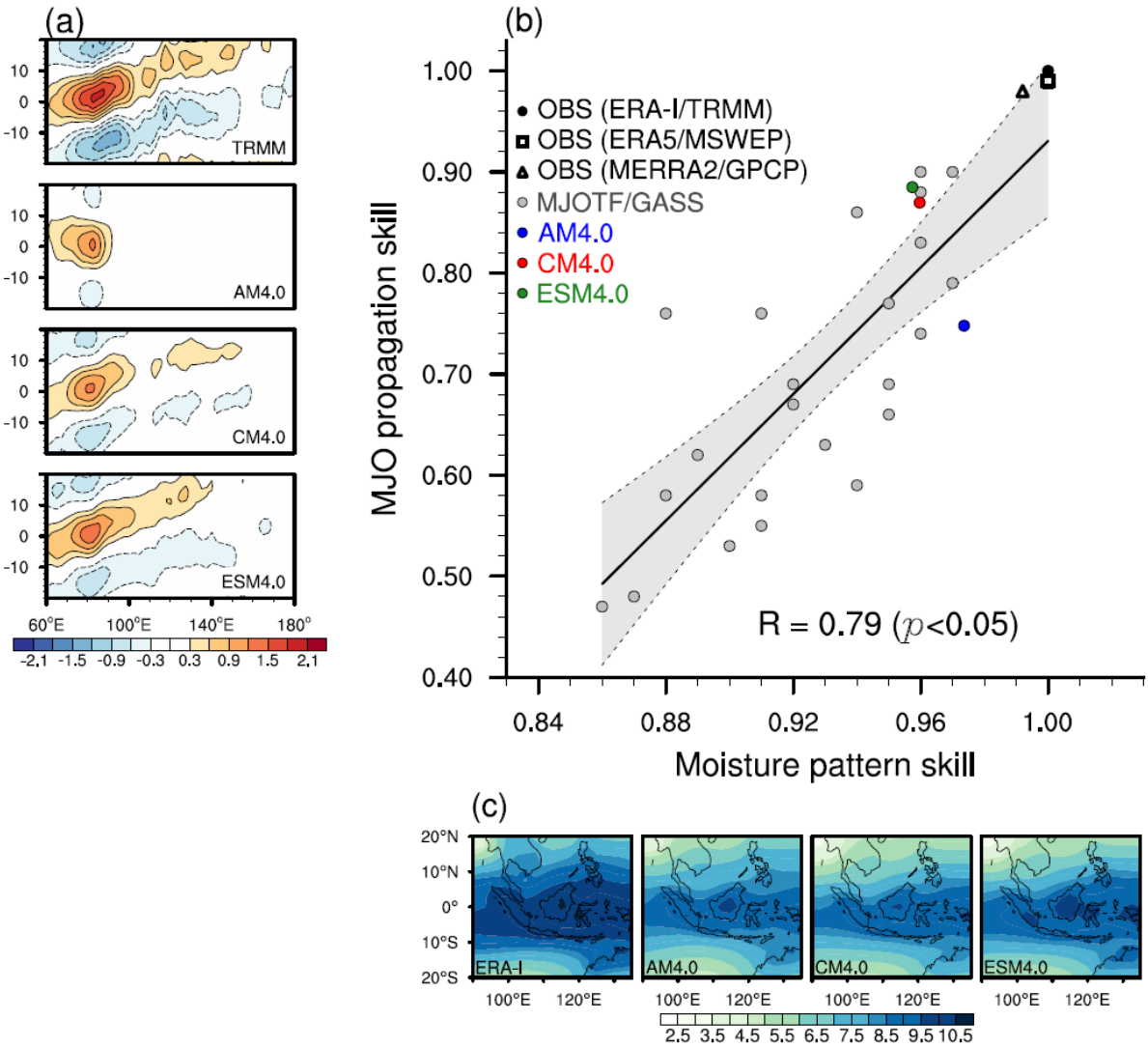

Fig. 3. (a) Hovmöller diagrams of rainfall anomalies (units: mm d^{-1}) averaged over 10°S–10°N based on (top-bottom) TRMM, GFDL AM4.0, GFDL CM4.0, and GFDL ESM4.0. Hovmöller diagrams are derived by lag-regressions of 20–100 day band-pass filtered anomalous rainfall against itself averaged over the equatorial Eastern Indian Ocean (75–85°E, 5°S–5°N) and normalized by its standard deviation over the regression base point. (b) Scatterplot between moisture pattern skill (*x*-axis corresponding to the bottom panel) and MJO propagation skill (*y*-axis corresponding to the left panel) based on multi- model simulations from the MJOTF/GASS project (gray dots) and three GFDL models (colored dots). A linear best-fit regression line by least squares means is shown in the gray line with the shading indicating the 95% confidence interval. The correlation coefficient is listed in the bottom right. (c) The winter (November–April) mean specific humidity (units: g kg^{-1}) averaged over 650–900 hPa based on (left-right) ERA-I, GFDL AM4.0, GFDL CM4.0, and GFDL ESM4.0 over Maritime Continent (20°–20°N, 90°E–135°E).

Broadening coverage of the climate system.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-21-0268.1.

Recent activities of the MDTF have expanded the suite of diagnostics to extend the range of processes beyond the atmosphere-centric suite described in Maloney et al. (2019). For example, a suite of Arctic-centric diagnostics have been entrained to diagnose reasons for the large spread in top-of-atmosphere shortwave radiative sensitivity to sea ice changes (Donohoe et al. 2020), as well as diagnose sea ice concentration and its variability and attribution of biases in the persistence of winter sea ice to oceanic mixed layer biases. PODs currently cover phenomena including ENSO and MJO and associated teleconnections, blocking, stratosphere-troposphere coupling, and tropical sea level variations. Additional PODs currently under development as part of the MDTF's third phase will extend the suite's coverage of diagnostics for air-sea interactions, oceanic processes, and terrestrial hydroclimate. These admittedly represent a fraction of the vast set of processes that must be represented in the system models, and further contributions from the community would be welcomed.

Interaction of PODs with perturbed-physics experiments.

A significant direction that can be more fully exploited is the coordination of perturbed-physics experiments (PPEs) with PODs targeting the same processes. PODs and PPEs provide a deeper understanding of a single model in terms of processes PODs and PPEs provide a complementary understanding of a single model in terms of processes and so can work well together. Changing parameters or formulation of a parameterized process occurs naturally in the model development cycle, and a POD can help direct and interpret these, or PPEs can help suggest the need for a POD (e.g., Tsushimaet al. 2020). Such experiments can also be conducted during the development of a POD to provide additional process information. Fig. 4 provides an example of a POD that displays the relationship of precipitation to two leading factors in convective buoyancy: a measure related to convective available potential energy ($CAPE_L$) of a convective plume traveling from the boundary layer through the lower troposphere in absence of entrainment, and a measure of the impact of subsaturation in the lower free troposphere ($SUBSAT_L$) due to effects of entrainment (Ahmed and Neelin, 2021). Conditional-average precipitation as a function of these two environmental variables is seen in the inset for CESM2. A gradient-like estimate of the precipitation sensitivity to these provides a pair of summary metrics shown in the polar plot for CMIP6 models and a series of CAM experiments perturbing the convective

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-21-0268.1.

parameterization. In the PPE experiments, as entrainment increases the magnitude of the precipitation pickup increases, but for large values the precipitation depends too heavily on near-saturation. The PPE thus helps to confirm the process linkage suggested by the POD and helps connect it to the specific model parameterization. Conversely, a POD-PPE combination could help to show if none of the model parameterizations represented in the PPE is sufficient to improve the process diagnosed by the POD, or to clarify if a particular process contributes to several relationships among climate variables (e.g., Schiro et al. 2020).

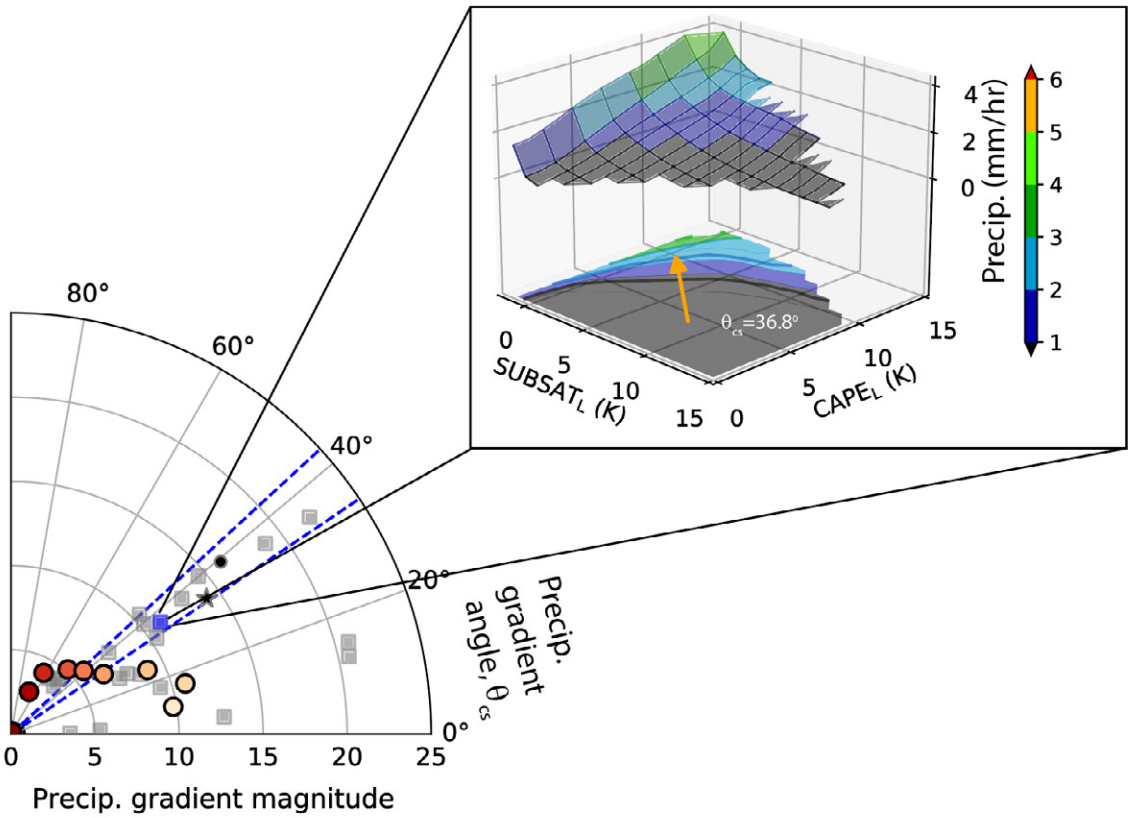

Fig. 4: The moisture-temperature sensitivity of precipitation in various climate models is shown in the polar plot. Models with angles closer to zero have greater sensitivity to moisture in the lower free troposphere.The blue dashed lines denote the uncertainty range generated using TRMM 3B42 and ERA5 data. The colored circles denote CAM5 fixed SST simulations with varying levels of entrainment, with darker colors indicating smaller entrainment. The gray squares denote different CMIP6 models. Inset: The precipitation thermodynamic relationship summarized by each marker is illustrated for one example CMIP6 model (CESM2) showing precipitation conditionally averaged on the $CAPE_L$ and $SUBSAT_L$ contributions to buoyancy (see text) with a measure of the magnitude and direction of the precipitation gradient in these variables (orange arrow).

Specialized output and timeslice experiments.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-21-0268.1.

While the standard output provided by CMIP6 is extensive (e.g,. Eyring et al. 2016a; Balaji et al. 2018; Juckes et al. 2020) and can be exploited for many PODs, some phenomena require higher temporal resolution, additional variables, or specific subsetting for more direct comparison to observational products or elucidation of the relevant physical processes. For example, the Cloud Feedback Model Intercomparison Project (CFMIP) recommends output from the CFMIP Observational Simulator Package (COSP) that allows a more equivalent comparison of modeled cloud processes to satellite observations (Webb et al. 2017). Given that output such as COSP is not uniformly provided across the CMIP ensemble of models, the MDTF is developing a protocol for timeslice experiments that provide higher resolution and specialized output. In these timeslice experiments, specialized model runs are conducted for common 20-year periods with specified SST and radiative forcing for atmosphere-only runs, or specified radiative forcing and initial conditions for coupled runs. A more extensive variable set than commonly provided by CMIP models can be output from these runs. Shorter 2-year and 5-year runs that allow higher frequency temporal output for 3-dimensional variables and for specialized output such as COSP are also proposed. Timeslice experiments within MDTF have been conducted with recent versions of the GFDL and NCAR climate models, although the timeslice experiment protocol may be applied more broadly across the community. A future publication will provide further details on this timeslice protocol.

Connectivity to scientific software such as feature trackers.

As the range of PODs advances, cases emerge that require "supporting utilities" (orange box in Fig. 1), i.e., additional tools such as feature trackers (Ullrich et al., 2021; Zhao, 2022), or interpolation packages that may be required for analysis but not part of PODs per se, and which may be externally maintained. These can be computationally heavy, so for a new model run being evaluated, the capability to initially output, e.g., a data set of tropical cyclone tracks that can be used for subsequent analysis, must be provided. Given the diversity of tools and output formats being maintained across the community (e.g. Shields et al. 2018), this area is ripe for standardization.

Future directions on the interaction of PODs with process models.

The use of simpler process models has been advocated to more clearly identify fundamental mechanisms in simulated phenomena (e.g., Held, 2005). The evolution of the

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-21-0268.1.

MDTF POD suite to more strongly interact with process models has the potential to provide deeper insight into reasons for model bias and ways to address those biases in model formulations and parameterizations. An example among PODs being developed exploits weak temperature gradient (WTG) theory for the tropics to present hypothesis-based diagnostics for convective moistening processes (Zhu et al (2017). Opportunities for additional integration of process models with diagnostics must be balanced with potential additional complexity. For instance, single-column modeling of diurnal cycle effects on precipitation (Tang et al. 2022) provides a more detailed process analysis than the POD in Fig. 2, but would be too complex for a package that is repeatedly run for new model versions. Process-model-based diagnostics that explicitly align with model parameterizations are encouraged—noting that this requires diagnostic developers to be knowledgeable about formulations and key variables of the parameterizations that are being developed and validated.

Common standards — CMEC-MDTF interoperability.

The diagnostics to metrics continuum.

Scalar evaluation metrics, commonly used to assess CMIP and development models (e.g., Gleckler et al. 2008; Reed et al. 2022), and the more complex process-oriented diagnostics can be viewed as opposite ends of a continuum of diagnostics. Scalar metrics represent the most condensed form but arise from underlying diagnostics that have representations in non-scalar form, such as maps, plots of conditional averages of one variable on another, and related statistics. For instance, the CLIVAR 2020 ENSO Metrics package (Planton et al. 2021) has "dive-down diagnostics"—such as the Pacific climatology of wind stress or regression patterns of ENSO sea surface temperature anomalies — which pop open from portrait plots from PMP (Gleckler et al. 2016; Lee et al. 2019). ILAMB provides maps of quantities underlying certain metrics. Conversely, PODs typically have a more complex representation of information as they seek insight into processes, but can often be condensed into scalar metrics, e.g., Leung et al. (2022) and examples below.

Fostering coordination.

Model evaluation is a common requirement among multiple federal agencies and a topic of scientific exploration for researchers. As such, agencies and researchers have, for decades, invested in the development of tools and capabilities for the evaluation of modeling systems.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-21-0268.1.

Collaborative efforts exist on a spectrum, ranging from more centralized efforts, where an individual agency brings in evaluation capabilities and maintains them in a single framework (e.g., ESMValTool; Eyring et al. 2016b; Righi et al. 2020), to more distributed efforts, where individual developers are tasked with software development and maintenance and collaboration occurs via standardization (e.g., CMEC). Of course, placement on this spectrum does not preclude further collaboration: distributed packages can be forked and brought into a centralized code base, and centralized packages can be wrapped for use in distributed frameworks.

The MDTF-CMEC coordination has been a bottom-up scientific collaboration that arose naturally from a hallway discussion at a conference. Commonalities in design made it straightforward to create a common standard, so this provides an example that can be built on by a broader community. Development of common standards and protocols between the DOE and NOAA efforts benefits the community by facilitating the flow of diagnostics among model development groups across agencies. This exchange can greatly amplify the value of targeted investments. One early success of this coordination is that MDTF PODs can be used as standalone climate evaluation codes within the CMEC framework. It is anticipated that most new PODs will be able to operate in the CMEC framework by default; when inconsistencies have occurred so far, they have been in minor details about the implementation of metadata files that are solvable with minor software modifications. Another example is the development of an oceanic water mass transformation utility (Tesdal et al. 2023) that makes use of the common CMEC standard. Similarly, evaluation packages that are compatible with CMEC (e.g., PMP, ILAMB, and community-contributions), should be straightforward to adapt for MDTF. The sidebar "Earth System Metrics and Diagnostics Standards (EMDS)" provides a summary of the current standards.

[Sidebar 2 near here]

Summary and Recommendations

The MDTF has demonstrated success in fostering a community of diagnostic developers probing different physical components and time scales of the coupled weather and climate system that has translated into diagnostics with pragmatic utility for model developers. Here we extract a set of recommendations based on the experience of moving a heterogeneous set

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-21-0268.1.

of diagnostics groups toward a community framework. We identify specific challenges that we anticipate will generalize to the larger community and propose some best practices that may address these. We view this as a collection of emerging opportunities in the realm of process-oriented diagnostics that could inspire future work by the community. These recommendations may be useful for developers proposing diagnostics integration, for current and future diagnostics framework teams, and for funding agencies setting proposal guidelines.

Community recommendations.

1. Continuity is essential to such efforts. For the diagnostics and framework developed here to outlive the participation of individual POD developers and the MDTF itself, an ongoing commitment to open-source diagnostics, community development on publicly accessible platforms (e.g.GitHub), and embrace of common community standards is necessary.
2. Fostering the development of PODs useful for model development should be a funding priority. More targeted guidance in the funding calls for POD developers with selected expertise, and having more model developers involved in panel reviews, would help optimize the utility of contributed PODs in the context of model improvement.
3. A sustained effort must be made to foster POD developer education of best software practices for robustness, maintainability, and efficiency and to enhance features of the framework that enable this. This may be accomplished through training materials, regular tutorials, and possibly more-focussed workshops where POD developers both within and outside the MDTF can interact with framework developers over a few days for a deeper dive into the operation of the framework.
4. With early examples and sustained communication, standards can be fostered for high-level uniformity among PODs, simplifying use, and facilitating coordination with other efforts, while assisting diagnostic development groups with common tools. Common standards will not only aid the continuity of the effort but also help expand the usefulness of PODs to modeling centers outside GFDL and NCAR.
5. The combined science and software intellectual contribution embodied in a diagnostic software module should be recognized in hiring/advancement practices as a

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-21-0268.1.

contribution beyond the associated journal articles. Early career scientists should be encouraged to include such items in their curriculum vitae; senior scientists familiar with such endeavors take these contributions into account, and we encourage the spread of such recognition. While we realize that this may involve a change in the culture of hiring, promotion, and tenure decisions across multiple independent institutions, providing community support and documentation of this support may help change the culture.

Development recommendations.

1. Documentation protocols:

    - Uniformity in the presentation of POD software usage and science content is important, aided by clear instructions and templates.
    - Besides references in the peer-reviewed literature, a detailed science summary providing context for suggested usage of the POD is advocated. This information provides useful context to the model developer. For example, providing interpretation of model values outside of observational uncertainty may provide insight into necessary parameterization modifications to improve model performance.
    - We recommend complementing peer-reviewed papers with other tools for the creation of citable resources for software code and datasets, specifically DOI assignment for a POD and digested analysis of observational data (with appropriate citation of original data sources). These can aid provenance and traceability, and promote recognition of software contributions.
    - Accessible and extensive framework documentation supporting ease of use by POD developers should be provided. Compared to their counterparts within the modeling centers, many POD developers have different software skills and performance requirements and often different assumptions of what constitutes efficient practices. Effective documentation from "quick-start" guides through materials that provide deeper dives into the framework will aid the successful integration of PODs while minimizing time commitment from framework and model developers.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-21-0268.1.

2. Code review: Similar to the peer-review process in scientific literature, diagnostic code benefits from the review practices that are commonplace in the open-source software community. POD developers should expect feedback from the framework team and their peers on code efficiency, style, documentation, and other factors that affect integrating their PODs in the framework. In addition, the POD code must be up to date with the latest development version of the MDTF-diagnostics framework before it is merged into the repository. Agencies supporting community development could encourage contributions to peer review of PODs' science within the diagnostics development community, while the framework team is primarily responsible for assessing the code quality.

3. Closer interactions between model developers and diagnostics developers should be facilitated to the largest extent possible. This will familiarize POD developers with efficient coding practices for use in their PODs, optimize the science insight and relevance of PODs for model development, as well as allow early incorporation of new PODs into routine diagnostics packages used by modeling centers.

4. Science feedback to POD developers

    - A process should be developed for peer feedback on POD code and documentation, akin to peer review for scientific manuscripts, that optimizes the usefulness of the PODs for model development. In addition to providing feedback from the model development community on the usefulness of PODs, POD developers must find this feedback useful and make good-faith efforts to modify their PODs accordingly. The following recommendations support these basic points.

    - POD developers should be encouraged to make extra effort to provide estimates of observational uncertainty (e.g. multiple reanalysis products or satellite retrievals; bootstrapping error bars for sensitivity to sampling). Links to the observational product development community are further encouraged to aid the characterization of systematic error including retrieval algorithm bias and prioritization of future observational products.

    - It should be consistently reinforced to POD developers to emphasize information that is pragmatic for the model developer. In addition to POD

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-21-0268.1.

developers being encouraged to invest in understanding existing parameterizations to help determine where process-oriented information can be most useful to model developers, peer review including that from model developers will reinforce this. The need to revisit PODs as deeper levels of information are required in subsequent model development should be expected.

Common standards recommendations.

In the US model and diagnostics development landscape, with different agencies focusing on their designated mission priorities, it seems unlikely that one framework "to rule them all" will emerge—or even be desirable. Further, it is unlikely that the variety of agencies interested in model evaluation capabilities would consolidate funding under a singular development effort. Instead, we advocate for a distributed approach and offer the CMEC-MDTF collaboration as an example that emerged naturally from science discussions. We encourage broader efforts to:

1. Pursue dialogue and maintain connections among agencies and scientists, while being aware of their priorities and needs. The benefit of these conversations is the ability to leverage the diversity of metrics and diagnostics packages already present in the US ESM evaluation ecosystem, as well as the potential to facilitate coordination with international efforts. While inevitably time-consuming in the early stages of development, such discussions enable the identification of common denominators among various organizations. This then allows us to avoid redundant and incompatible efforts, which would in turn require substantial time commitment later in the process to harmonize.
2. Develop and maintain common, flexible, and lightweight standards to facilitate interoperability. Building PODs that are interoperable under common standards will permit users to create workflows that bring more specialized diagnostics into the context of more general and widely understood metrics. For instance, users familiar with either PODs or CMEC packages will not need a large time investment to mix and match diagnostics and metrics from the two frameworks.
3. Organize a standards body to develop and maintain those standards. Such a body should draw both leadership and technical expertise from different agencies to ensure

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-21-0268.1.

that the standards conform to the goals of flexibility, interoperability, and utility. Governance of a standards committee with a modest initial purview is relatively straightforward (see link in Sidebar 2). Long-term, as the level of community coordination evolves, one could envision the standards body being folded into or bootstrapping toward a larger governance structure for diagnostics and related observational efforts.

4. Seek agreement among agencies to respect those standards internally and contribute meaningfully to the ecosystem of evaluation tools. Namely, it is important to seek a commitment to community development practices among participating organizations, rather than only focusing on internal needs. This recommendation is connected with the need for lightweight and flexible standards, which would allow for agency-specific interpretations of the standards to be made that do not corrupt the use of PODs by users outside of that agency.
5. Foster scientific coordination in the continuum between diagnostics and metrics. Common standards facilitate software connection but scientific attention to this connectivity is also required. This can be promoted by task force and funding agency emphasis, and by awareness in the diagnostic/metrics development community.

Acknowledgments.

This work was supported by National Oceanographic and Atmospheric Administration Modeling, Analysis, Predictions, and Projections (MAPP) program under Climate Program Office grants NA21OAR4310354 (JDN, JPK, AR, JL, TJ, YM, WD, YHK, FA), NA21OAR4310353 (EM), NA21OAR4310344 and NA18OAR4310270 (AAW), NA20OAR4310410 (EAM) and NA21OAR4310356 (PAU). The effort of AO was supported by Lawrence Livermore National Laboratory Contract DE-AC52-07NA27344 from the Regional and Global Modeling Analysis program area, under the auspices of the United States Department of Energy's Office of Science. We thank D. Barrie for his long-standing role in envisioning and supporting the MDTF, along with the MAPP program staff. We acknowledge past and current members of the MDTF for many productive interactions. We acknowledge the World Climate Research Programme Working Group on Coupled

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-21-0268.1.

Modelling for coordinating CMIP6. We thank the climate modeling groups for 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.

Data Availability Statement.

Supporting observational data and sample model data for the MDTF package are available via anonymous FTP at ftp://ftp.cgd.ucar.edu/archive/mdtf. Analysis code is available at https://github.com/NOAA-GFDL/MDTF-diagnostics (Liptak et al. 2022). AM4.0/LM4.0 code and selected model results are at http://data1.gfdl.noaa.gov/nomads/forms/am4.0/ (Guo et. al, 2018; Krasting et. al, 2018; Zhao et. al, 2018). Access to the MJOTF/GASS MJO model comparison project data can be found at https://earthsystemcog.org/projects/gassyotc-mip/ as described at http://www.ucar.edu/yotc/mjodiab.html. CMIP6 data is available at https://esgf-node.llnl.gov/projects/cmip6/. Pre-processed CAM5 parameter perturbation experiments used in Figure 3 are available at https://github.com/ahmedfiaz/MDTF-diagnostics/blob/master/cam5_perturbation_exps.zip.

Sidebar 1: Accessing and developing diagnostics for the MDTF framework

An overview of the MDTF package is found at https://www.gfdl.noaa.gov/mdtf-diagnostics, including resources such as video tutorials, with code available at https://github.com/NOAA-GFDL/MDTF-diagnostics for both download and contribution. The associated documentation is at https://mdtf-diagnostics.readthedocs.io with documentation for installing and running the framework, documentation of existing PODs and information for POD developers. The latter includes a developer QuickStart guide, POD development guidelines and summaries of coding best practices and Git-based development workflow, and a template for POD documentation that enables automatic updates of the manual as a new POD is accepted.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-21-0268.1.

Sidebar 2: Earth System Metrics and Diagnostics Standards (EMDS)

The collaboration on common standards between MDTF and Coordinated Model Evaluation Capabilities (CMEC) https://cmec.llnl.gov/ has been encoded as the Earth System Metrics and Diagnostics Standards, available at https://github.com/Earth-System-Diagnostics-Standards/EMDS/. This provides guidance for the design of Earth system metrics and diagnostics-based software utilities. The standards document provides simple requirements for software interfaces, metadata, and metrics output format to promote the interoperability of software packages and reproducibility of results, without being unduly burdensome to implement.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-21-0268.1.

REFERENCES

Ahmed, Fiaz, and J. David Neelin, 2021: A Process- Oriented Diagnostic to Assess Precipitation- Thermodynamic Relations and Application to CMIP6 Models. *Geophysical Research Letters* 48, e2021GL094108.

Balaji, V., and Coauthors, 2018: Requirements for a global data infrastructure in support of CMIP6, Geosci. Model Dev., 11, 3659–3680, https://doi.org/10.5194/gmd-11-3659-2018.

Boucher, O., Servonnat, J., Albright, A. L., Aumont, O., Balkanski, Y., Bastrikov, V., ... & Co-authors, 2020. Presentation and evaluation of the IPSL- CM6A- LR climate model. *Journal of Advances in Modeling Earth Systems*, 12(7), e2019MS002010.

Danabasoglu, G., Lamarque, J. F., Bacmeister, J., Bailey, D. A., DuVivier, A. K., Edwards, J., & Co-authors, 2020. The community earth system model version 2 (CESM2). *Journal of Advances in Modeling Earth Systems*, 12(2), e2019MS001916.

Donohoe, A., E. Blanchard-Wrigglesworth, A. Schweiger, and P. J. Rasch, 2020: The effect of atmospheric transmissivity on model and observational estimates of the sea ice albedo feedback. *J. Climate*, **33**, 5743–5765, https://doi.org/10.1175/JCLI-D-19-0674.1.

Dunne, J. P., and Coauthors, 2020: The GFDL Earth System Model version 4.1 (GFDL-ESM4.1): Model description and simulation characteristics. *J. Adv. Model. Earth Syst.*, **12**, e2019MS002015, https://doi.org/10.1029/2019MS002015.

Eyring, V., Harris, N. R. P., Rex, M., Shepherd, T. G., Fahey, D. W., Amanatidis, G. T., ... & Waugh, D. W., 2005. A strategy for process-oriented validation of coupled chemistry–climate models. *Bulletin of the American Meteorological Society*, 86(8), 1117-1134.

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

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., Gottschaldt, K.-D., Hagemann, S., Juckes, M., Kindermann, S., Krasting, J., Kunert, D., Levine, R., 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,

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-21-0268.1.

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

Eyring, V., Cox, P. M., Flato, G. M., Gleckler, P. J., Abramowitz, G., Caldwell, P., ... & Williamson, M. S., 2019. Taking climate model evaluation to the next level. *Nature Climate Change*, 9(2), 102-110.

Gleckler, P.J., Taylor, K.E. and Doutriaux, C., 2008. Performance metrics for climate models. *Journal of Geophysical Research: Atmospheres*, *113*, D6, https://doi.org/10.1029/2007JD008972.

Gleckler, P. et al., 2016: A more powerful reality test for climate models. *Eos*, https://doi.org/10.1029/2016eo051663.

Gonzalez, A.O. and Jiang, X., 2017. Winter mean lower tropospheric moisture over the Maritime Continent as a climate model diagnostic metric for the propagation of the Madden- Julian oscillation. *Geophysical Research Letters*, 44, 2588-2596, https://doi.org/10.1002/2016GL072430.

Guo, H. and and Coauthors; 2018. NOAA-GFDL GFDL-CM4 model output, Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.1402

Grose, M. R., Bhend, J., Narsey, S., Gupta, A. S., & Brown, J. R., 2014. Can we constrain CMIP5 rainfall projections in the tropical Pacific based on surface warming patterns?. Journal of Climate, 27(24), 9123-9138.

Held, I.M., 2005. The gap between simulation and understanding in climate modeling. *Bulletin of the American Meteorological Society*, *86*(11), pp.1609-1614, https://doi.org/10.1175/BAMS-86-11-1609.

Held, I. M., Guo, H., Adcroft, A., Dunne, J. P., Horowitz, L. W., Krasting, J., et al. (2019). 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

Hwang, Y. T., & Frierson, D. M., 2013. Link between the double-Intertropical Convergence Zone problem and cloud biases over the Southern Ocean. *Proceedings of the National Academy of Sciences*, 110(13), 4935-4940.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-21-0268.1.

Jiang, X., and Coauthors, 2015: Vertical structure and physical processes of the Madden–Julian oscillation: Exploring key model physics in climate simulations. J. Geophys. Res. Atmos., 120,4718–4748, doi:10.1002/2014JD022375.

Jiang, X., 2017. Key processes for the eastward propagation of the Madden- Julian Oscillation based on multimodel simulations. *Journal of Geophysical Research: Atmospheres*, 122,755-770, https://doi.org/10.1002/2016JD025955.

Juckes, M., Taylor, K. E., Durack, P. J., Lawrence, B., Mizielinski, M. S., Pamment, A., Peterschmitt, J.-Y., Rixen, M., and Sénési, S., 2020: The CMIP6 Data Request (DREQ, version 01.00.31), *Geosci. Model Dev.*, 13, 201–224, https://doi.org/10.5194/gmd-13-201-2020.

Kelley, Maxwell, Gavin A. Schmidt, Larissa S. Nazarenko, Susanne E. Bauer, Reto Ruedy, Gary L. Russell, Andrew S. Ackerman et al. "GISS- E2. 1: Configurations and climatology." Journal of Advances in Modeling Earth Systems 12, no. 8 (2020): e2019MS002025.

Kim, D., A. H. Sobel, E. D. Maloney, D. M. W. Frierson, and I.-S. Kang, 2011: A systematic relationship between intraseasonal variability and mean state bias. *J. Climate*, **24**, 5506-5520.

Kim, D., Xavier, P., Maloney, E., Wheeler, M., Waliser, D., Sperber, K., ... & Liu, H., 2014. Process-oriented MJO simulation diagnostic: Moisture sensitivity of simulated convection. *Journal of Climate*, 27(14), 5379-5395.

Krasting, J. P. and Coauthors, 2018 , NOAA-GFDL GFDL-ESM4 model output prepared for CMIP6 CMIP, Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.1407

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

Leung, L.R., Boos, W.R., Catto, J.L., A. DeMott, C., Martin, G.M., Neelin, J.D., O'Brien, T.A., Xie, S., Feng, Z., Klingaman, N.P. and Kuo, Y.H., 2022. Exploratory precipitation metrics: spatiotemporal characteristics, process-oriented, and phenomena-based. *Journal of Climate*, 35, 3659-3686, https://doi.org/10.1175/JCLI-D-21-0590.1.

Liptak, J., and Coauthors, 2022: MDTF-diagnostics framework code (v3.0). Zenodo.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-21-0268.1.

https://doi.org/10.5281/zenodo.7343282

Maloney, E. D., Gettelman, A., Ming, Y., Neelin, J. D., Barrie, D., Mariotti, A., ... & Zhao, M., 2019. Process-oriented evaluation of climate and weather forecasting models. Bulletin of the American Meteorological Society, 100(9), 1665-1686.

Planton, Y.Y., Guilyardi, E., Wittenberg, A.T., Lee, J., Gleckler, P.J., Bayr, T., McGregor, S., McPhaden, M.J., Power, S., Roehrig, R. and Vialard, J., 2021. 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.

Reed, K.A., Goldenson, N., Grotjahn, R., Gutowski, W.J., Jagannathan, K., Jones, A.D., Leung, L.R., McGinnis, S.A., Pryor, S.C., Srivastava, A.K. and Ullrich, P.A., 2022. Metrics as tools for bridging climate science and applications. *Wiley Interdisciplinary Reviews: Climate Change*, e799, https://doi.org/10.1002/wcc.799.

Rosa, D., & Collins, W. D., 2013. A case study of subdaily simulated and observed continental convective precipitation: CMIP5 and multiscale global climate models comparison. *Geophysical Research Letters*, 40(22), 5999-6003.

Schiro, K.A., Su, H., Ahmed, F. et al., 2022: Model spread in tropical low cloud feedback tied to overturning circulation response to warming. *Nature Communications* 13, 7119. https://doi.org/10.1038/s41467-022-34787-4.

Shields, C.A., Rutz, J.J., Leung, L.Y., Ralph, F.M., Wehner, M., Kawzenuk, B., Lora, J.M., McClenny, E., Osborne, T., Payne, A.E. and Ullrich, P., 2018. Atmospheric river tracking method intercomparison project (ARTMIP): project goals and experimental design. *Geoscientific Model Development*, 11, 2455-2474, https://doi.org/10.5194/gmd-11-2455-2018.

Sperber, K. R., & Waliser, D. E., 2008. New approaches to understanding, simulating, and forecasting the Madden–Julian oscillation. *Bulletin of the American Meteorological Society*, 89(12), 1917-1920.

Simpson, I. R., and co-authors, 2020: An evaluation of the large-scale atmospheric circulation and its variability in the Community Earth System Model version 2 (CESM2) and other CMIP models, *J. Geophys., Res.*, **125**, https://doi.org/10.1029/2020JD032835.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-21-0268.1.

Righi, M., and Coauthors.: Earth System Model Evaluation Tool (ESMValTool) v2.0 – technical overview, *Geosci. Model Dev.*, 13, 1179–1199, https://doi.org/10.5194/gmd-13-1179-2020, 2020.

Tang, S., Xie, S., Guo, Z., Hong, S.Y., Khouider, B., Klocke, D., Köhler, M., Koo, M.S., Krishna, P.M., Larson, V.E. and Park, S., 2022. Long- term single- column model intercomparison of diurnal cycle of precipitation over midlatitude and tropical land. *Quarterly Journal of the Royal Meteorological Society*, 148, 641-669, https://doi.org/10.1002/qj.4222.

Tatebe, H., Ogura, T., Nitta, T., Komuro, Y., Ogochi, K., Takemura, T., ... & Kimoto, M., 2019. Description and basic evaluation of simulated mean state, internal variability, and climate sensitivity in MIROC6. *Geoscientific Model Development*, 12(7), 2727-2765.

Teixeira, J., Waliser, D., Ferraro, R., Gleckler, P., Lee, T., & Potter, G., 2014. Satellite observations for CMIP5: The genesis of Obs4MIPs. *Bulletin of the American Meteorological Society*, 95(9), 1329-1334.

Tesdal, E., MacGilchrist, G. A., Beadling, R. L., Griffies, S. M., Krasting, J. P., & Durack, P. J. (2023). Revisiting Interior Water Mass Responses to Surface Forcing Changes and the Subsequent Effects on Overturning in the Southern Ocean. *Journal of Geophysical Research: Oceans*, *128*(3), e2022JC019105. https://doi.org/10.1029/2022JC019105

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

Tsushima, Y., Ringer, M.A., Martin, G.M. et al., 2020: Investigating physical constraints on climate feedbacks using a perturbed parameter ensemble. Climate Dynamics, 55, 1159–1185. https://doi.org/10.1007/s00382-020-05318-y

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

Voldoire, A., Saint- Martin, D., Sénési, S., Decharme, B., Alias, A., Chevallier, M., ... & Waldman, R., 2019. Evaluation of CMIP6 deck experiments with CNRM- CM6- 1. *Journal of Advances in Modeling Earth Systems*, 11(7), 2177-2213.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-21-0268.1.

Webb, M. J., Andrews, T., Bodas-Salcedo, A., Bony, S., Bretherton, C. S., Chadwick, R., Chepfer, H., Douville, H., Good, P., Kay, J. E., Klein, S. A., Marchand, R., Medeiros, B., Siebesma, A. P., Skinner, C. B., Stevens, B., Tselioudis, G., Tsushima, Y., and Watanabe, M., 2017: The Cloud Feedback Model Intercomparison Project (CFMIP) contribution to CMIP6, *Geosci. Model Dev*., 10, 359–384, https://doi.org/10.5194/gmd-10-359-2017.

Zelinka MD, Myers TA, McCoy DT, et al., 2020: Causes of Higher Climate Sensitivity in CMIP6 Models. *Geophys Res Lett* 47, e2019GL085782, https://doi.org/10.1029/2019GL085782.

Zhao, M. (2022). A Study of AR-, TS-, and MCS-Associated Precipitation and Extreme Precipitation in Present and Warmer Climates, *Journal of Climate*, 35, 479-497, https://doi.org/10.1175/JCLI-D-21-0145.1.

Zhao, M., and Coauthors, 2018a: The GFDL global atmosphere and land model AM4.0/LM4.0: 1. Simulation characteristics with prescribed SSTs. *J. Adv. Model. Earth Syst.*, **10**, 691–734, https://doi.org/10.1002/2017MS001208

Zhao, M., and Coauthors, 2018, NOAA-GFDL GFDL-AM4 model output, Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.1401

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-21-0268.1.