BAMS Article

Observational Data for Next-Generation Climate Model Evaluation: Requirements, Considerations, and Best Practices

Rebecca L. Beadling,[a] Ranjini Swaminathan,[b] Romain Beucher,[c] Ed Blockley,[d] Swen Brands,[e] Birgit Hassler,[f] Dora Hegedűs,[g,h] Forrest M. Hoffman,[i] Jiwoo Lee,[j] Jared Lewis,[k] Jianhua Lu,[l] Elizaveta Malinina,[m] Brian Medeiros,[n] Enrico Scoccimarro,[o] Jerry Tjiputra,[p] Briony Turner,[h] and Duncan Watson-Parris[q]

KEYWORDS: Climate records; Satellite observations; Surface observations; Climate models; Diagnostics; Model evaluation/ performance

ABSTRACT: Climate model simulations are an important source of information about our planet's climate system and also enable informed decision-making under different future scenarios. As a new archive of results from the next generation of climate models is anticipated to become available with the Coupled Model Intercomparison Project phase 7 (CMIP7), the need to develop efficient and robust methods to evaluate models is paramount. Observations are an integral part of model evaluation, providing a means to quantify and understand the degree to which climate models can faithfully reproduce Earth system processes. Such analysis is critical for constraining climate projections, identifying areas of focus for model development, and assisting analysts in deciphering the utility of models for specific applications. Observations of Earth system come from a diversity of sources, span different space–time domains, and are produced by different communities, and each dataset features different data structures and formats, metadata standards, and its own unique uncertainties. Uncertainties in an observational dataset may stem from gaps in temporal and spatial coverage, instrumentation errors, or assumptions in retrieval and processing methods. How then does one ensure that observational data are ready for use and utilized in the most appropriate way for robust, rapid, and routine climate model evaluation? The CMIP7 Model Benchmarking Task Team with input from the broader climate modeling, model evaluation, and observational data communities present a vision and considerations for best practices toward the optimal and appropriate use of observational data to support next-generation climate model evaluation.

DOI: 10.1175/BAMS-D-25-0079.1
Corresponding author: Ranjini Swaminathan, r.swaminathan@reading.ac.uk
Rebecca L. Beadling and Ranjini Swaminathan co-lead authors.
Manuscript received 21 March 2025, in final form 3 February 2026, accepted 12 March 2026

SIGNIFICANCE STATEMENT: Computer models that simulate Earth system, known as climate models, are important tools for understanding how climate processes work and provide estimates of the climate system in the past and the future. Policy decisions regarding how to adapt to and limit the impact of climate change rely on these models, and it is, therefore, important to know how well they capture the real world. Real-world observations play an important role in understanding how well climate models represent Earth's climate system. We describe various aspects of using observations for model evaluation and suggest best practices to make this process more efficient, accurate, and inclusive of a wider number of observed phenomena. Considerations and best practices are presented for the use of observations in climate model evaluation. The discussion centers on practices to support next-generation evaluation for Coupled Model Intercomparison Project simulations.

AFFILIATIONS: [a] Department of Earth and Environmental Science, Temple University, Philadelphia, Pennsylvania; [b] Department of Meteorology and National Centre for Earth Observation, University of Reading, Reading, United Kingdom; [c] Australian Climate Simulator (ACCESS-NRI), Canberra, Australian Capital Territory, Australia; [d] Met Office Hadley Centre, Exeter, United Kingdom; [e] Instituto de Física de Cantabria (CSIC-UC), Santander, Spain; [f] Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institut für Physik der Atmosphäre, Oberpfaffenhofen, Germany; [g] Rutherford Appleton Laboratory, RAL Space, Science and Technology Facilities Council, Harwell Campus, Didcot, United Kingdom; [h] CMIP International Project Office, European Space Agency, Harwell Campus, Didcot, United Kingdom; [i] Oak Ridge National Laboratory, Oak Ridge, Tennessee; [j] Lawrence Livermore National Laboratory, Livermore, California; [k] Climate Resource, Melbourne, Victoria, Australia; [l] School of Atmospheric Sciences, Sun Yat-sen University and Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Zhuhai, China; [m] Canadian Centre for Climate Modeling and Analysis, Environment and Climate Change Canada, Victoria, British Columbia, Canada; [n] NSF National Center for Atmospheric Research, Boulder, Colorado; [o] CMCC Foundation - Euro-Mediterranean Center on Climate Change, Lecce, Italy; [p] NORCE Norwegian Research Centre, Bjerknes Centre for Climate Research, Bergen, Norway; [q] Scripps Institution of Oceanography and Halıcıoğlu Data Science Institute, University of California San Diego, La Jolla, California

1. Introduction

Building climate models that are able to realistically simulate Earth system processes and feedbacks, and thereby be useful for societal decision-making, depends on the rigorous evaluation of simulated results against the observed world. As climate models increase in complexity and spatial resolution, and the sheer size of their archived output grows, it is becoming necessary to consider and develop best practices and standards agreed upon by the modeling, evaluation, and observational communities to facilitate efficient as well as robust model evaluation. We define efficient evaluation as methods that allow the community to go from raw model output to scientific insight at a faster rate than the present. Robust evaluation methods are those that rigorously assess climate model performance by ensuring accurate and reproducible comparisons with observations, while also accounting for observational uncertainty and model variability, to identify the most representative measures of model fidelity. This requires intentional coordinated, transparent, and sustainable efforts between model developers, analysts, software developers, and observational data scientists.

The World Climate Research Programme's (WCRP) Coupled Model Intercomparison Project (CMIP) provides a key framework necessary for large-scale climate and Earth system

model (ESM) simulation and evaluation efforts (Durack et al. 2025c). Looking ahead, the WCRP established Task Teams (TTs) to support the upcoming CMIP phase 7 (CMIP7). Recognizing the need for routine and rapid model evaluation capabilities to advance CMIP's scientific goals, the CMIP7 Model Benchmarking TT (MB-TT) was formed. The MB-TT was tasked with delivering guidelines and technical specifications for the infrastructure necessary for robust and rapid climate model assessment fully and seamlessly integrated into the CMIP7 framework. The MB-TT, in extensive collaboration with other CMIP7 TTs, WCRP activities, and the broader modeling and observational communities, seeks to achieve this nontrivial goal through scientific publications that describe the climate model evaluation landscape (Hassler et al. 2026) and assess and communicate considerations and cross-community needs, in addition to the development and implementation of the Rapid Evaluation Framework (REF; Hoffman et al. 2025). In tandem with the REF development, and addressing the WCRP's Earth System Modeling and Observations (ESMO) project objectives, the MB-TT also aims to synthesize best practices for integrating observational data for model evaluation.

In this paper, we focus on the central role that observational data play in model evaluation and present a vision for considerations and best practices for observation–model comparisons in the assessment of climate model performance. We frame these considerations in the context of advancing large-scale multimodel evaluation capabilities for CMIP7, but these can be applied more broadly to the assessment of single models, in-house evaluation routines applied during model development, and across CMIP generations. Many of the ideas developed herein are the result of cross-community discussions initiated during the first phase of REF development. Through this paper, we hope to actively promote engagement across modeling and observational data communities as cross-community collaborative efforts can both motivate the need for novel and sustained observations and aid in the full utilization of existing products for model evaluation and development. Every effort has been made to ensure these best practices not only just address observational data requirements from the modeling centers and model analysts but also highlight the challenges faced by observational data providers and the critical role they play in the process of model evaluation.

Before framing our considerations and best practices, which is the main focus of the paper, we begin in section 2 by presenting a brief discussion on past and ongoing efforts that have targeted improving the efficiency of large-scale multimodel evaluation frameworks. Throughout this manuscript, we refer collectively to both purely physical coupled model configurations and their more complex extensions, ESMs, simply as "climate models." We focus on climate models contributing to CMIP and on model evaluation, rather than "benchmarking," as described in Hassler et al. (2026).

As we define "climate model evaluation" as the process of assessing simulations against observations (Hassler et al. 2026), in framing our considerations and best practices, we begin section 3 with a discussion on the development of appropriate metrics for observation–model comparison to facilitate meaningful measurements of climate model fidelity (Gleckler et al. 2008; Flato et al. 2013). Section 3a serves to address the many factors that must be considered, treated, and communicated appropriately in striving for a fair assessment of model performance.

In subsequent sections (sections 3b–h), we suggest best practices to ensure appropriate comparisons and that these assessments can be done in such a way to support efficient large-scale climate model evaluation in the face of growing model complexity and data volume. Section 3i presents a brief discussion on the issue of underutilized observational datasets and the opportunity to expand the existing archive of ready-to-use datasets for evaluation. Section 3j briefly discusses the evolving role of artificial intelligence and machine learning

in methods in the climate model evaluation workflow. We close the paper by proposing a call for cross-community action to address the issues presented here and a vision for future-ready observations to unlock the next generation of robust, rapid, and routine climate model evaluation. Best practices from each section are summarized in Table 1.

Finally, we note that although the focus of this paper is on model evaluation, this discussion is also relevant to other phases of model development where observations play a key role (Fig. 1). This includes model initialization [e.g., using observed hydrography for ocean initial conditions (Griffies et al. 2016); prescribing boundary conditions and external forcing (Meinshausen et al. 2017; Durack et al. 2025a), as proxies for parameters

Table 1. Best practices and guidance for the use of observational datasets for climate model evaluation. The relevant section that discusses each recommendation is listed in the right column.

No.	Guidance	Section
1	Use multiple observations to minimize observational uncertainty, identifying dependencies between them	Section 3a: Ensuring the appropriate model to observational comparisons in metric development
2	Ensure variables being compared represent the same quantity	
3	Account for differences in forcings and required length of observational records	
4	Ensure evaluation is tailored to processes (adequately) represented in models	
5	Consider the experimental design (e.g., historical, AMIP, OMIP, DCPP) used to run the models when comparing them with observations	
6	Perform regridding or sampling only when absolutely essential. Where needed, this is best done by observational data providers or with guidance provided by providers so uncertainties are quantified appropriately	Section 3b: Spatial resolution and sampling frequency
7	Observational datasets with and without gap filling to be made available for evaluation and where possible gap filling and up- or downscaling to be done by the data providers with associated uncertainties provided for evaluation	Section 3c: Approaches for scaling, gap-filling, and extrapolating data
8	Vocabulary around what quantity is represented by a certain variable name and the uncertainty fields pertaining to the variable (e.g., standard error) to be clearly documented	Section 3d: Quantifying data uncertainties better
9	Where observations may be proxies or processed, care to be taken to evaluate and account for uncertainties appropriately	
10	Clear and concise documentation on uncertainties to be provided by observational data providers and to be incorporated in evaluation	
11	Continuous cross-community engagement, including updates on uncertainty guidance based on feedback received from the community or when novel data and processing methods are introduced for observations	
12	Point source to gridded data conversion only to be done when essential and where possible by data providers. Care is to be taken to account for differences between model and observational point sources such as with regards to topography and landscape representation	Section 3e: Point observations
13	Construction of time series and climatology for observational and model data to follow the same process and associated information used for this such as grid resolution and gaps be included.	Section 3f: Time series and climatologies
14	Observational datasets to follow data standards and conventions such as CMOR and CF for technical alignment with climate model data and easy access for evaluation	Section 3g: Data formats and conventions
15	Observational datasets should include important information such as digital object identifiers, version, and provenance and be easily accessible through open-source projects such as Obs4MIPs	Section 3h: Versioning, archiving, and distributing observational data
16	Community engagement activities should be planned around identifying under and unutilized datasets for model evaluation as well as to identify potential climate variables as part of new satellite missions or data collection projects that may benefit new processes developed in the next generation of climate models	Section 3i: Opportunities to expand the ready-to-use observational dataset archive
17	Include information regarding uncertainty and artifacts in the data due to novel methods used	Section 3j: AI, ML, and hybrid observations

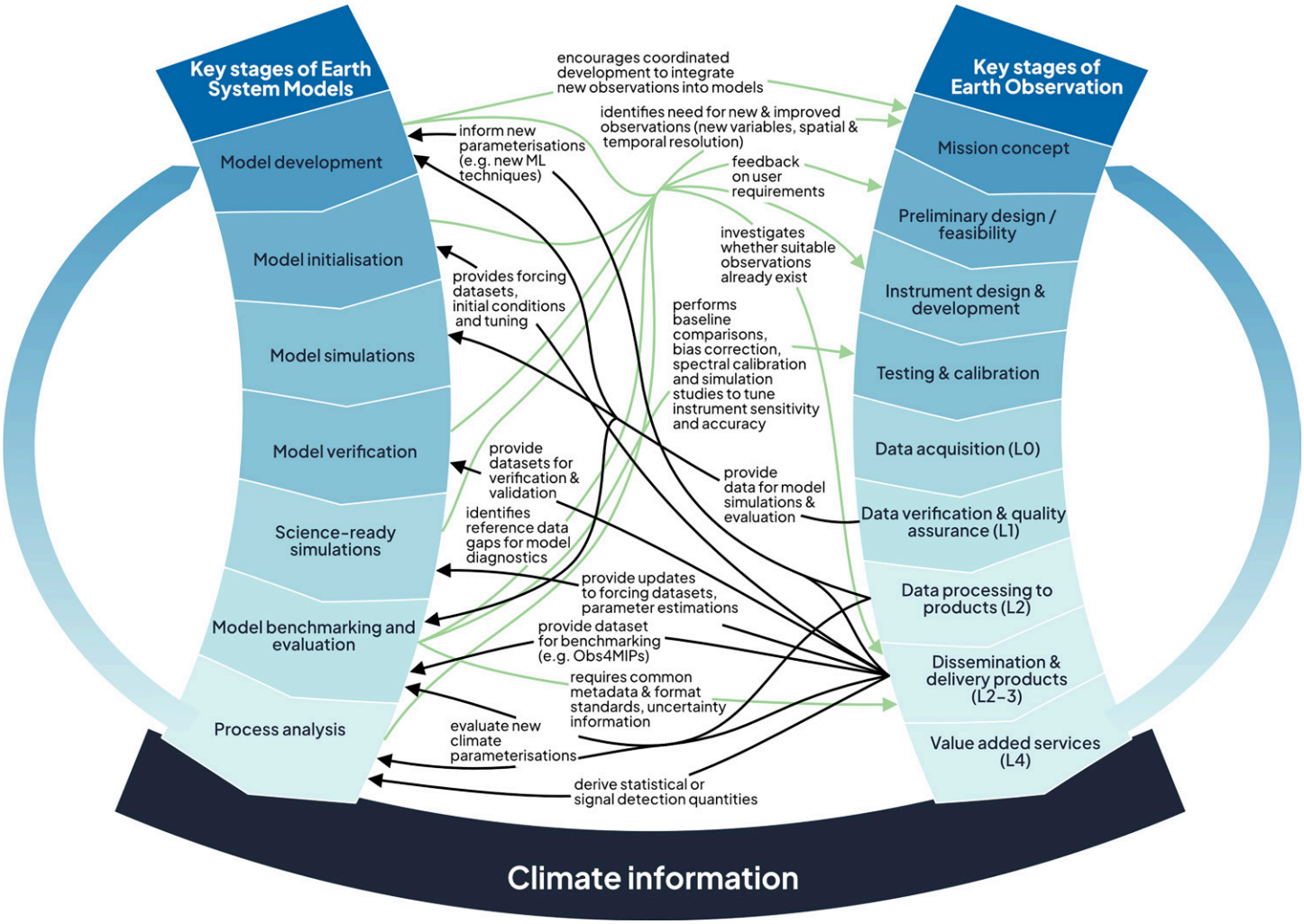

Fig. 1. An illustration of the different ways in which model development and model data analysis interface within various phases of the development of observational data products. Black lines show how observations feed into the model development process, and green lines show where model processes can influence observational data production and development (Hegedűs et al. 2026).

or human activities (Bontemps et al. 2012); and informing process-oriented model validation (Eyring et al. 2004, 2005)]. The model development process also integrates into and motivates many phases of observational product development (Mueller et al. 2013; Poschlod and Koh 2024) (Fig. 1).

2. Current landscape to support efficient and robust evaluation

The key to supporting efficient and robust model evaluation efforts is ensuring that well-established and documented observational products are readily accessible and technically aligned with climate model output. Specifically for multimodel CMIP evaluation, where many climate model simulations are assessed in tandem, it is ideal that the observational datasets are also readily available on the Earth System Grid Federation (ESGF) platform (Williams et al. 2016; Petrie et al. 2021), where CMIP output is archived and distributed. Meeting these two requirements improves efficiency by reducing the overhead of having analysts independently search for, download, and preprocess the datasets themselves. These coordinated practices can help limit the potential for errors and inconsistencies to propagate into the evaluation framework resulting from analysts making independent choices on the treatment of observational datasets in their assessments.

Recognizing existing barriers and the requirements noted above, strong foundations have been laid by existing projects that have targeted advancing climate model evaluation

capabilities, including the Observations for Model Intercomparison Project (Obs4MIPs; Teixeira et al. 2014; Ferraro et al. 2015; Waliser et al. 2020) and the Collaborative Reanalysis Technical Environment–Intercomparison Project (CREATE-IP; Potter et al. 2018—previously Ana4MIPs). These efforts have provided technically aligned, well-documented, and mostly global, gridded observational datasets (Obs4MIPs) and output from atmosphere and ocean reanalysis products (CREATE-IP) available on ESGF. This has supported routine model evaluation efforts, and Obs4MIPs and CREATE-IP datasets have been successfully incorporated into the workflow of model evaluation tools including the ESM Evaluation Tool (ESMValTool; Righi et al. 2020; Eyring et al. 2020), the Program for Climate Model Diagnosis and Intercomparison (PCMDI) Metrics Package (PMP; Lee et al. 2024), the International Land Model Benchmarking (ILAMB; Collier et al. 2018), and International Ocean Model Benchmarking (IOMB; Fu et al. 2022) packages and other packages compatible with the Coordinated set of Model Evaluation Capabilities (CMEC; https://cmec.llnl.gov/).

The Obs4MIPs and CREATE-IP efforts have introduced successful frameworks that can be further developed and improved, but considerable work is needed to expand these archives to include a larger suite of observational datasets covering more climate realms and Earth system variables. For example, most Obs4MIPs datasets are global atmospheric fields or limited to the ocean's surface layer, with a notable lack of in situ, regional, and station-based observations that capture local-scale variability. The observational data available are also mostly at monthly mean resolution and on grids with approximately 100-km grid spacing, with a lack of inclusion of higher-frequency or finer horizontal-resolution datasets. Additionally, an ongoing challenge is the ability to produce updates and versioning of the available products, which requires human resources and expanded archive space.

While Obs4MIPs is actively accepting and publishing new gridded data products and versions, with plans for an updated and improved submission workflow, the CREATE-IP archive on ESGF is no longer being updated with reanalyses since 2020. However, renewed efforts are currently under way through the World Meteorological Organization Integrated Processing and Prediction System (WIPPS) program to ensure that global climate reanalysis data are continuously generated and made available to users, accompanied by clear documentation and visualization tools. Currently, there are three WIPPS Designated Centres for global climate reanalysis data: ERA5 data from the European Centre for Medium-Range Weather Forecasts (ECMWF), CMA Global Atmospheric Reanalysis-40 (CRA-40) data from the China Meteorological Administration (CMA), and MERRA-2 data from the U.S. National Aeronautics and Space Administration (NASA). These three datasets along with the National Aeronautics and Space Administration (JRA-3Q) data from the Japan Meteorological Agency (JMA) will ultimately be made available on one platform, thus facilitating intercomparison of global reanalysis products including the provision of data on comparable grids.

With a growing diversity of Earth system observations available and lengthening records, there is an opportunity to leverage and contribute to the existing Obs4MIPs framework and other ongoing efforts such as those discussed above to improve evaluation capabilities in preparation for CMIP7.

3. Observational data considerations for evaluation

a. Ensuring appropriate model-to-observational comparisons in diagnostic development. In building a performance metric to quantify model fidelity and provide insight into a model's fitness for purpose (i.e., whether it is an appropriate tool for a specific scientific or decision-making objective), one must first ensure that suitable observational datasets exist for the climate variable or process of interest. Achieving appropriate "like-for-like" comparisons presents significant challenges both philosophically and technically. Philosophically, evaluating model fidelity through comparisons with observations rests on the assumption

that available observational datasets provide a reliable reference for a given climate phenomenon. Ideally, such comparisons enable an objective assessment of how well a model represents key aspects of the real climate system—i.e., whether or not a model represents the mean state, variability, trends (Simpson et al. 2025), or processes of the climate system as captured in instrumental records. However, observational records themselves may be imperfect for a variety of reasons, precluding the assumption that they objectively represent the actual climate state. Imperfections may stem from a diversity of sources including but not limited to, measurement errors and uncertainties, gaps in spatial and temporal sampling or coverage, errors introduced in retrieval or refinement of raw datasets, or uncertainties associated with the underlying numerical models used in data assimilation and state estimation approaches.

In some cases, despite identical naming conventions and units, in model output and observational products, two variables may not actually be representing the same property or process. There are two distinct issues here; one being that the observational and modeling community might have different working definitions of the same variable—i.e., the climate variable name provided in CMIP's Controlled Vocabulary (CV; Durack et al. 2025b) is identical to that used by the observational community but represents different physical processes (e.g., leaf area index; Fang et al. 2019). The second issue arises from the fact that what climate models may be simulating for a particular variable and what the observational instruments are measuring are not the same, despite both sharing the same name (e.g., precipitation/clouds; Stephens and Kummerow 2007). We discuss specific examples of these issues later when presenting considerations for quantifying uncertainties (see section 3d).

Evaluation diagnostics also must consider differences in climate forcings such as regional aerosols, emission-driven versus prescribed carbon dioxide (CO_2), land-use changes across models and observations (Eyring et al. 2016), and differences in intrinsic variability simulated by the model versus that experienced in reality. Importantly, in historical simulations initialized from preindustrial control runs, the simulated state of the main modes of internal (or unforced) climate variability is expected to be out of phase with the observed state over any period of comparison (Meehl et al. 2014; Fasullo et al. 2020). Similarly, the observed world in which the datasets for comparison were collected represents only one potential reality (i.e., one instance) within a dynamically chaotic climate system (Notz 2015). Additionally, different climate phenomena evolve on diverse time scales, and thus, different observational record lengths are required to estimate the mean state and variability against which to compare model results.

Missing or incomplete representation of climate processes or Earth system components in climate models can also complicate comparisons. For example, until recently, freshwater input from ice sheet dynamical changes has not been represented in climate model simulations due to the lack of incorporation of interactive ice sheet model components (Schmidt et al. 2023), even though there has been an observed acceleration of land–ice mass loss (Fox-Kemper et al. 2021). Similarly, uncoupled simulations forced with prescribed surface boundary conditions such as atmosphere (AMIP; Gates 1992; Gates et al. 1999) or ocean–sea ice-only configurations [Ocean Model Intercomparison Project (OMIP); Griffies et al. 2016] are, by design, lacking full Earth system process representation and thus limiting important interactions and feedbacks between climate realms. Evaluating coupled and uncoupled configurations in tandem to isolate potential sources of model biases is common (Walsh et al. 2002; Wang et al. 2009; Chen and Schneider 2014; Anandh et al. 2021; Mizuochi et al. 2021); however, developing diagnostics must be approached differently for the two configurations, specifically when interpreting reasonable expectations of agreement with observed phenomena, given different measures of realism. For example,

sometimes uncoupled configurations may be closer to the observed state of a given climate phenomenon because the forcing (e.g., prescribed sea surface temperatures) constrains the model state more than a freely evolving coupled configuration.

In acknowledging the imperfect nature of both observational datasets and climate model simulations as illustrated above, it is critical to include as many observational datasets and climate model ensembles as possible in diagnostic development to ensure robust evaluation. While helpful for a multiple lines of evidence approach, dependencies between different products for a single variable should be considered in the interpretation. Developing comparison frameworks that utilize multiple independent observational estimates of a single climate variable will help to capture the potential range of uncertainty (Gómez-Navarro et al. 2012; Gibson et al. 2019; Zumwald et al. 2020; Lauer et al. 2023; Evans and Imran 2024). Additionally, utilizing ensembles of climate model simulations provide a better estimate of the potential range of given climate variable due to simulated internal climate variability and model uncertainty (Kay et al. 2015; Deser et al. 2020; Maher et al. 2021).

The caveats noted here do not preclude the utility of using observational datasets to assess climate model simulations. Instead, they call attention to important considerations diagnostic developers should be aware of to ensure that robust assessments are made and that uncertainties are being carefully considered and communicated appropriately.

b. Spatial resolution and sampling frequency. Model resolution has increased across successive CMIP generations, with a greater number of participating models approaching 25–50 km in the atmosphere and 10–25 km in the ocean (Roberts et al. 2025). Comparing CMIP3 with CMIP6, the average ocean horizontal grid spacing employed in coupled configurations has approximately halved from ~133 km in CMIP3 to ~58 km in CMIP6, with several models entering the eddy-permitting regime and allowing for the explicit simulation of the ocean mesoscale (Hewitt et al. 2020). CMIP6 also featured models with nested and flexible ocean grid meshes that allow for better representation of specific climate relevant ocean processes such as storms, boundary currents, and coastal–open-ocean interactions (Sidorenko et al. 2015; Semmler et al. 2020). On the atmospheric side, there have been significant advancements in the vertical resolution across CMIP generations, resulting in a greater number of models employing atmospheric components with more vertical levels and higher model tops (Hardiman et al. 2012; Yu et al. 2024). Enhancement of vertical resolution aims to improve near-surface processes in the atmospheric boundary layer and/or the vertical propagation of waves through the troposphere and stratosphere (Smalley et al. 2023; Wicker et al. 2023). Higher-altitude atmospheric model tops provide for improved representation of the stratosphere which has implications for climate variability and predictability associated with, for example, the quasi-biennial oscillation and sudden stratospheric warming events (Hardiman et al. 2012).

The results from high-resolution simulations are key to constraining connections between large-scale climate dynamics and local impacts, which is necessary for climate risk assessments at regional scales (Roberts et al. 2018). One of CMIP7's guiding research questions aims to constrain how dangerous weather patterns will evolve in a warming climate (Dunne et al. 2025). Answering this with confidence relies on being able to evaluate extreme weather events simulated by climate models—spanning tropical and extratropical cyclones, extreme precipitation events, droughts, storm surges, and heat waves (Williams et al. 2024). Constraining the "climate of the extremes" requires the availability of fine spatial resolution and high-frequency (daily or subdaily) observational datasets to compare simulations against (Kotlarski et al. 2019; Trenberth et al. 2017; Gervais et al. 2014; Wehner et al. 2021). Such datasets can also aid in evaluating results from regional

downscaling experiments [e.g., Coordinated Regional Climate Downscaling Experiment (CORDEX); Diez-Sierra et al. 2022].

High spatial and temporal resolution observational datasets are rare at the long time scales needed to produce frequency distributions of large sample sizes to assess extreme events, as well as to assess fine-scale features such as ocean mesoscale eddy activity in high-resolution models. Although reliant on imperfect numerical simulations themselves, data assimilation products (e.g., reanalyses) and observationally constrained state estimates are available at higher spatial and temporal resolution and may be considered for comparing against climate model results. In some cases, atmospheric reanalysis products must be relied on to increase the time frequency of data to compare against (e.g., global hourly) for certain process-based diagnostics such as tropical cyclone–ocean interactions (Scoccimarro et al. 2017). However, when used for evaluation, individual product ensembles (e.g., the ERA5 10-member ensemble; Hersbach et al. 2020) and multiple independent reanalyses or state estimation products must be included with appropriate representation of uncertainty for the particular diagnostics of interest. A common approach in multimodel evaluation is to spatially regrid both the models and the observational data to a common resolution, which may introduce uncertainty and the loss of information. Additionally, certain quantities such as those related to ocean velocities and transport should not be regridded (Griffies et al. 2016), and thus, model evaluation workflows must be adapted to be able to handle finer-resolution model and observational datasets and different native grid structures. Similarly, information is lost through temporal averaging when the observational or model output must be upscaled or downscaled to align for comparison or data sharing purposes. The loss of variance going from model time steps to the monthly or annual averaged output typically requested for CMIP is an issue for many climate variables—particularly those that evolve on rapid time scales such as atmospheric wind or ocean velocities.

Model evaluation workflows will benefit enormously from having a diversity of observational product resolutions readily available and technically aligned with CMIP output to avoid excessive upscaling and downscaling of model output or observational datasets. Depending on the climate metric, regridding is often unavoidable when working across large multimodel ensembles, and analysts need to be aware of potential consequences, make appropriate choices for upscaling and downscaling routines, and quantify and report the associated uncertainties introduced. In the case of multiple observational products available at different resolutions, documentation and metadata that describes upscaling and downscaling methods and associated uncertainty should be provided with the dataset.

c. Approaches for scaling, gap-filling, and extrapolating data. Long-term, continuous availability of data with minimal gaps or missing data is necessary for evaluating climate simulations at decadal-to-centennial time scales (Henson et al. 2016; Karl et al. 1995). This especially poses a challenge for the evaluation of climate model representation of extreme events for which long time-scale observations are necessary to record conditions under which these relatively rarer events might occur and to discern signal from noise (Alexander et al. 2016; Alexander 2016; Zwiers et al. 2013). Global reanalysis and observationally constrained state estimates are commonly used in climate model evaluation workflows for quantities that suffer from gaps. However, these products suffer from temporally and spatially sparse observational datasets that are used to constrain the underlying numerical models, resulting in increased uncertainty in regions that are less constrained (e.g., have fewer observations) (Sterl 2004; Nakamura et al. 2025). As such, increasing the spatial coverage, securing long-term data collection, and improving the quality of anchor observations

underpinning reanalysis products (e.g., radiosondes; Haimberger et al. 2012, 2024) are important in the broader context of climate model evaluation.

Data paucity and difficulties in collecting data are not uniform across the globe, with polar regions, high-altitude zones, and the subsurface ocean presenting unique challenges for data collection. Polar regions in particular have an extreme paucity of long-term observations with insufficient space–time coverage for climate model assessment due to the presence of ice, snow, clouds, light availability, and harsh environmental conditions (Smith et al. 2019). Satellite products at high latitudes, for example, cannot easily distinguish snow and ice from clouds, and measurement accuracy can be seasonally dependent (Castro et al. 2023). Additionally, remote access and harsh conditions including the presence of sea ice in polar regions makes it difficult to collect in situ observations and deploy Argo floats for continuous data collection in the subsurface ocean (Roemmich and Gilson 2009). The Southern Ocean, for example, had extremely sparse observations of ocean physical and biogeochemical properties until the development and deployment of autonomous floats with ice-avoidance software (Sarmiento et al. 2023). Additionally, in situ ocean observations are often biased toward summer seasons when research cruise campaigns are regularly carried out.

Despite many of the end delivery products being global in nature, satellite data naturally suffer from space–time gaps due to reasons including coverage limitations, occlusions (physical obstructions blocking measurements such as clouds or snow), perturbations from emissions (e.g., smoke, emitted gases), or instrument irregularities. Interpolation or gap filling, therefore, becomes a necessary operation to deliver full global coverage for satellite-derived products. Some commonly used gap-filling methods applied to both in situ and satellite data include Kriging, pixel interpolation, regression, sampling-based approaches (Yin et al. 2017; Covey et al. 2016), and, more recently, machine learning approaches (discussed in section 3j). Beyond the issues of gap-filling missing data, both in situ and satellite observations also pose challenges related to representation error (Schutgens et al. 2017) when translating from finer-resolution or point-source measurements to gridded products for more standardized evaluation with climate model output (Kondrashov and Ghil 2006; Good et al. 2013). Issues related to point-source datasets are discussed further in section 3e.

Several methods have been proposed to quantify uncertainties resulting from interpolation and gap-filling observational data (Lepot et al. 2017; Richardson and Hollinger 2007; Herrera et al. 2019; Longman et al. 2020). We do not endorse nor recommend any specific approach, as different approaches are appropriate for different observational datasets. Instead, we strongly recommend that gap-filling, extrapolation, and upscaling and downscaling methods should be performed by the respective observational dataset provider with the requisite expertise on the data and methods. Providers should clearly document the approaches used and quantify the best possible estimate for the uncertainties introduced with minimizing, resolving, and documenting data gaps to ensure data integrity. We recommend that where possible, both filled and unfilled data should be provided to analysts to make decisions on how best to deal with gaps in their routines for model evaluation, minimizing uncertainty.

d. Quantifying data uncertainties better. Observational data collection has made significant technological and operational advances in recent years and serves the dual role of documenting climate change as well as of evaluating increasing complexity in climate models (Collins et al. 2013). For the purpose of climate model evaluation, we broadly characterize observational uncertainty as resulting from misrepresentation in the qualitative or quantitative aspect of the observed quantity. This could pertain to the semantic meaning of the quantity (i.e., is it the quantity itself or a proxy?) or the value of the measurement

(i.e., measured versus theoretical or actual value). Observational data uncertainties play a key role in our ability to evaluate models effectively and can arise for several reasons (Matthews et al. 2013; Merchant et al. 2017). Systemic and random instrumentation errors (Kennedy 2014), algorithmic biases (Slivinski et al. 2019; Yuan et al. 2010), processing raw data to a gridded product (Herrera et al. 2016; Mittaz et al. 2019), gap-filling methods (Yin et al. 2017), or instrument change can all contribute to observational uncertainty.

As a first step toward understanding uncertainty in observational datasets, we acknowledge that the modeling and observational science communities may use different vocabularies for uncertainty [see discussion in Elipot et al. (2022)]. It is, therefore, important that the vocabulary around uncertainty is unambiguously defined and differences between terms such as uncertainty, quality, and error are articulated in both the technical documentation accompanying datasets and in the scientific publications resulting from their use in model evaluation. As mentioned previously, another point where additional uncertainty may emerge in a model–observation comparison as a result of different vocabularies is when two variables compared against each other are not actually representing the same physical property or process (i.e., the CMIP CV variable name is the same as the name used by the observational community, but they refer to different underlying physical processes). This is less of an uncertainty associated with the observational dataset itself but an inconsistent use of observations in model evaluation due to different working definitions of the same variable. For instance, the amount of leaf area in an ecosystem is measured using the leaf area index (LAI), which in canopy reflectance models can be equivalent to the green LAI (GLAI) but not so in other studies (Fang et al. 2019). Clear descriptions of what the observation represents and guidance on how it should be used for comparisons with models can help minimize inconsistencies in evaluation.

In some cases, the variables that climate models are simulating do not align with what the observational instruments are actually measuring, despite sharing the same name. For example, there are a suite of Earth-observing satellites designed to provide measurements of the global atmosphere including the thermal structure, cloud distribution, precipitation, and surface and near-surface processes (Merchant and Embury 2014; Li et al. 2013). One might be tempted to use these satellite-derived datasets "out of the box" for model evaluation purposes for assessment of clouds or temperatures, for example. Yet, this is not possible given that it is not the vertical atmospheric temperature, precipitation, or cloud properties that are measured directly by the satellites, but these quantities are derived from radiation measurements during the retrieval process (Stephens and Kummerow 2007). These same issues also arise when evaluating chlorophyll concentration in ESMs against satellite observations of ocean color. Satellite sensors measure radiance, which is then used to derive estimates of chlorophyll concentration, whereas in ESMs, chlorophyll is often derived diagnostically through a prognostic phytoplankton variable [Séférian et al. 2020; see discussion in Clow et al. (2024)].

To deal with this inability to make direct comparisons between the satellite-derived data and that simulated by climate models, significant effort has gone into the development of "satellite simulators" (e.g., Klein and Jakob 1999; Klein et al. 2013; Eliasson et al. 2019)—software that allows for the simulation of observational "data" as observed by Earth system satellites, imitating the satellite observation and retrieval process. The use of satellite simulators has been adopted to evaluate cloud processes in climate models by the Cloud Feedback Model Intercomparison Project (CFMIP; Webb et al. 2017) where the CFMIP Observation Simulator Package (COSP) has been developed to allow for comparison of satellite data from a suite of instruments to direct model output (Bodas-salcedo et al. 2011). Similar efforts are under way to develop and utilize satellite simulators to assess model-simulated ocean chlorophyll against satellite observations (Clow et al. 2024). Such packages should be leveraged and

further developed to ensure appropriate comparison and uncertainty treatment for remotely sensed observational datasets.

Community efforts have emerged which center on the need to better quantify uncertainty, clarify vocabulary, and improve communication around uncertainty in climate records. Such efforts include the European Space Agency's Climate Change Initiative (Merchant et al. 2017) project, the Climate Model User Group report for Obs4MIPs (CMUG 2025), the Ocean Uncertainty Quantification (OceanUQ) Working Group under the U.S. Climate Variability and Predictability Program (Elipot et al. 2022), and the Information Quality Cluster (IQC) of the Earth Science Information Partners (ESIP) (Moroni et al. 2019), as well as other domain-specific studies (McMillan et al. 2012; Sayer et al. 2020). It is, therefore, recommended that uncertainty handling broadly follows guidelines drawn through community engagement, as described in the documents listed above. Furthermore, such efforts need to be ongoing and updated with new developments and novel datasets entering observational climate science. As complex as the issue of uncertainty can be, following guidelines from the community efforts outline above, it is recommended that the observational data providers include uncertainty information in as concise and clear a manner as possible with the dataset itself and not just as part of the references to relevant literature.

Finally, the use of multiple observations is also encouraged as studies also show the value of using an ensemble of observational datasets (Prein and Gobiet 2017; Zumwald et al. 2020; Lauer et al. 2023; Evans and Imran 2024) and reanalysis products (Buizza et al. 2005; Langland et al. 2008; Wei et al. 2010) to better understand and constrain uncertainty.

e. Point observations. Point-source observations are characteristic of in situ as well as satellite measurements and are almost always processed into gridded datasets through interpolation and sometimes extrapolation methods, with implications for uncertainty associated with the datasets (Haylock et al. 2008). Studies show that such gridded averages frequently underestimate observed variability, particularly in the extremes (Cavanaugh and Shen 2015), and more generally have statistical properties quite different from the original observations (Ensor and Robeson 2008). Reliability of such gridded data is further eroded when the methods fail to take into account geographical features such as terrain and elevation (Daly 2006) or land–sea boundaries. Approaches to overcome these drawbacks include comparing multiple gridding schemes for specific applications or climate variables (Hofstra et al. 2008) or regions (Avila et al. 2015; Abbasnezhadi and Wang 2024).

While there is uncertainty introduced in refinement of an observational dataset from point locations to discrete grid cells, gridded products are both analyst friendly and more appropriate for comparison with climate model output. Climate variables simulated by models exist on kilometer-scale grid boxes and are only able to represent behavior over an area much larger than the geographical extent that point observations represent, thus one would not expect agreement between gridboxed average quantities and point observations (Schutgens et al. 2016; Schutgens et al. 2017). Point observations can be used directly in model evaluation; however, a like-for-like direct comparison of data collected from a single geographical point to a kilometer-scale grid cell will be impossible, and caution is warranted. For example, large differences in the representation of topography in climate models relative to the real world require one to consider lapse-rate adjustments when comparing atmospheric fields. The land type represented in models may also be starkly different than the real world, where urbanization, for example, can exert a strong impact on local surface energy fluxes and thus surface observations.

When discrete observations (satellite swathes, in situ measurements, point observations) are processed to the final gridded product released to the community, it is important to be aware of and account for uncertainties introduced. We recommend that the original

creator of the observational product be consulted in the design of the evaluation metric to ensure the comparison is as accurate as possible. The code, process description, and any other algorithms detailing how the refinement is done should be openly available with clear guidance for users on how to account for associated uncertainties in their model comparisons.

f. Time series and climatologies. Univariate time series and climatologies allow one to understand key statistical properties of climate variables such as average conditions (means) or variability (standard deviations) over a given time period such as 30 years. Time series from observations with sufficiently long temporal coverage are a valuable resource for evaluating long-term transient changes simulated by climate models. The observed global surface temperature anomaly (NASA global land–ocean temperature index; Hansen et al. 2010; Lenssen et al. 2024) and measured atmospheric CO_2 concentrations (Keeling et al. 2001) are perhaps the most familiar and influential time series in the field of climate science. The time series of global-mean surface temperature over the historical period is often the first variable used in model evaluation.

While these types of datasets are valuable for climate model evaluation, in many cases, there may be a lack of observations of a particular climate variable for a time period long enough to subsample the range of intrinsic climate variability (Simpson et al. 2025). The differing time scales of variability of a given climate variable mean that differing observational record lengths are required to develop a more refined estimate of the mean state and variability against which to compare model simulations. Due to this issue, for variables that have shorter record length, care must be taken in the interpretation and communication of model performance. Similarly, if data assimilation products are used to extend records to construct time series or climatologies, multiple products should be used to account for uncertainty.

As time series and climatologies are often constructed from aggregated spatial data, it is critical that the construction of these datasets be well documented, including any spatiotemporal weighting, gap-filling, or regridding schemes applied. This will aid in ensuring that the model output and observational data are constructed through the same process for effective comparison between the two. For observational datasets used as time series and climatologies for evaluation, the following are recommended as a minimal set of metadata or supporting information that should be provided, some of which are already covered by Climate and Forecast (CF) conventions: grid resolutions, spatial and temporal coverage, as well as gaps, area-averaging methods, units of the quantity being evaluated, data availability, choice of reference period, and smoothing or scaling factors.

g. Data formats and conventions. Without having certain standards in data format, the diversity of observational datasets would make their use in model evaluation and software very challenging. The issue of data format and conventions is also important for open-source software to extract, process, and analyze the data for model evaluation in a collective and systematic way. While Network Common Data Form (netCDF) has been widely adapted for climate model output, especially CMIP, it is yet to be applied to most observational datasets. To address the issue of Variety (5 Vs of Big Data) and adhere to Findable, Accessible, Interoperable, and Reusable (FAIR) data principles (Wilkinson et al. 2016), it is recommended that observational data should be openly available and also follow CF metadata conventions (Eaton et al. 2003; Hassell et al. 2017). The field construct used within CF conventions associates the data with sufficient metadata such that the dataset itself is self-describing, i.e., it is clear what physical quantity the data represent, its units, standard name, and space–time coordinates. Observational datasets that conform to CF conventions

will help reduce uncertainty due to misinterpretation of variables (section 3d) and further facilitate comparisons with CF-compliant model data.

Obs4MIPs (Teixeira et al. 2014; Waliser et al. 2020) provides observational datasets technically aligned with CMIP model data by using the Climate Model Output Rewriter (CMOR), the same software used by modeling groups contributing to CMIP. Obs4MIPs is open to community contributions, which is a way to guide development better aligned with new datasets and constraints including for data from point sources such as station data. At a minimum, the use of CF conventions is recommended for optimal compatibility of observational data with climate model simulations. CMORization of observational datasets (e.g., standardization of variable/dimension/coordinate naming conventions, with respect to CMIP standards) is recommended to support routine and rapid evaluation such that the observations can be easily ingested into mature model evaluation packages. However, we recognize that the use of CMOR might be a high bar for many observational providers to meet and that other issues arise around backward compatibility of software used to CMORize and versioning of datasets that needs to be addressed.

The use of either netCDF or analysis-ready cloud-native file formats (e.g., Zarr) would promote ease of use and more technical alignment with climate model output, facilitating more efficient evaluation in the context of CMIP multimodel evaluation efforts. We are not endorsing one format over the other as we recognize that there is an advantage to future output being stored and delivered in a cloud-native way as the archive size grows (Abernathey et al. 2021; Stern et al. 2022).

h. Versioning, archiving, and distributing observational data. To maintain a baseline of model performance and ensure appropriate comparisons across ensembles and model generations, it is important to version observational datasets and maintain public access to prior versions. Flexibility to assess model performance to prior and current revisions of observational datasets is desirable, particularly if prior versions were used in published analyses. Version information and provenance should be made readily available in a single location with directions on where to find prior versions and future releases. Additionally, each dataset should contain clear instructions for a requirement of data citation [e.g., with a unique digital object identifier (DOI)] and the reporting of errata so that it is clear which versions are being used in any published analyses and if there are any known issues. Static versions of this information can be maintained in a dataset's metadata; however, accompanying documentation should be openly available and continuously updated.

Archiving datasets in free and open public databases is now common practice and increasingly required by science funding agencies, peer-reviewed journals, and government sponsors (Wilkinson et al. 2016). However, maintaining (e.g., providing and communicating timely updates, versioning, monitoring errata reports) large observational datasets is a challenge, given the often transient nature of research funding. Archiving multiple versions or resolutions of observational datasets also becomes cumbersome to host on a single platform given storage capabilities. Obs4MIPs and CREATE-IP projects provide a framework for publishing ready-to-use observational and reanalysis datasets in a single location (at ESGF) with technical documentation. Having the various versions hosted and disseminated via ESGF and technically aligned with CMIP model output is ideal; however, the ever-growing data storage footprint associated with CMIP model output alone is pushing the limits of the ESGF infrastructure. Thus, in the long term, it may be more feasible to have ready-to-use observational datasets and their documentation provided from the data providers themselves via free and openly accessible platforms. Existing operational service repositories such as the Climate Data Store through the Copernicus Climate Change Service (C3S; Buontempo et al. 2023), the Copernicus Marine Environment Monitoring Service (CMEMS; Le Traon et al. 2019),

or the National Centers for Environmental Information (NCEI) archive hosted through the National Oceanic and Atmospheric Administration (NOAA) should be supported and leveraged in efforts to archive and distribute data. Additionally, data redundancy should be practiced to avoid single point failures, and data integrity should be ensured through practices such as hashing.

i. Opportunities to expand the ready-to-use observational dataset archive. There is a wealth of in situ observational data and gridded observational products that remain underutilized in model evaluation. This underutilization results from multiple factors that inhibit ease of use by analysts, which may include formatting that is not technically aligned with CMIP (i.e., not CMORized), unclear treatment of observational uncertainty (which may include measurement uncertainty or that associated with processing of the data to its final form), licenses and registrations required to access and download data (i.e., restrictions), lack of informative metadata or instructions for use, or a lack of global awareness that the product is available (i.e., not well advertised or cited).

Many packages used for model evaluation require that the observational datasets are CMORized and comply with CF metadata conventions. Otherwise, preprocessing scripts are required to get the data into a usable format, which require the user to have a robust understanding of the observational dataset at the hand including associated uncertainties. Given that preprocessing adds an additional step to the model evaluation workflow, observational datasets already in CMIP compatible formats are often prioritized and "harder-to-use" products get left out. This results in a positive feedback where the "easy-to-use" products are cited more and thus more likely to be used in future analyses.

Addressing this issue of the underutilization of observational datasets requires a concerted effort from modelers, analysts, and the observational community to work together to identify useful products and package them into easy-to-use formats for model evaluation.

j. The role of artificial intelligence and machine learning in developing observations for climate model evaluation. Artificial intelligence (AI) and in particular data-driven ML approaches are starting to play an important role in the workflow of climate model evaluation in different ways. These include both AI-informed evaluation methodologies (Gibson et al. 2017; Nowack et al. 2020) and AI-informed hybrid observational products (Tselioudis et al. 2021; Kaps et al. 2023). While a detailed discussion of how AI is used in observations is beyond the scope of this work, we highlight a key area of application where ML algorithms have had a significant impact, which is in addressing data paucity of observations due to poor coverage of satellite and in situ observations or gaps in time series. ML methods have been applied to upscale point-source data to global gridded datasets (Tramontana et al. 2016), to produce spatiotemporal interpolations (Landschützer et al. 2015; Kim et al. 2024), and for gap filling to produce continuous data streams (Sloyan et al. 2023; Jung et al. 2025). However, it is important to bear in mind that these approaches may further add hard to quantify uncertainties (Singh et al. 2024) and also have other limitations such as accurately representing interannual variability (Jung et al. 2020), the ability to extrapolate to out of distribution data (Schneider et al. 2022), and spatial heterogeneity (Ghorbanpour et al. 2021). Dealing with these issues as they pertain to individual datasets will enable their appropriate use for model evaluation.

4. A Call for cross-community action: Future ready observations to unlock the next generation of climate model evaluation

The challenge of achieving efficient and robust climate model evaluation capabilities is nontrivial and will require human and infrastructure resources, transparency, and sustained

cross-community partnerships. The synthesis presented here and the suggested best practices (Table 1) are the result of collaborative discussions and an attempt to distill cross-community needs. Many of the needs discussed are the result of bringing climate modelers, model analysts, software developers, and observational dataset providers together during the initial development of the REF (Hoffman et al. 2025) led by the MB-TT. Bringing together communities to practically implement a working framework to advance multimodel climate evaluation provided us with an improved understanding of the landscape of existing capabilities, their strengths and weaknesses, and awareness of existing barriers for both analysts and observational dataset providers in facilitating robust and efficient model evaluation.

To achieve a future with a greater number and diversity of observational datasets readily available to meet growing model evaluation needs, we should begin to organize meetings to get communities on the same page with respect to barriers and best practices. There should be broad participation in this process across communities, career level, climate realms, and geographic locations. Specific goals must be identified and categorized, and practical hands-on workshops should then be organized to produce deliverables that will advance capabilities. As an example, we have identified that Obs4MIPs provides an optimal framework for both producing and hosting gridded observational datasets to facilitate multimodel climate model evaluation. Yet, there are many ways that this framework can be advanced with community input which would allow for the integration of a greater number and diversity of existing datasets. A series of workshops could be organized to 1) collate gridded observational datasets that are available but yet to be CMORized across climate realms, 2) partner CMOR-aware software developers with the observational dataset providers to submit dataset proposals and produce Obs4MIPs-compliant products, and then 3) integrate these new datasets into existing or novel model evaluation routines. Another set of workshops could then be organized on identifying nongridded observational datasets (e.g., in situ, point source) or producing blended products and advancing methods and routines that would allow for their use in routine model evaluation.

Many activities can be envisioned to target the specific nuances and considerations discussed throughout this manuscript. The key is that these efforts will need to be carefully organized with specific tangible deliverables to progress forward. Attention must be given to not duplicate efforts or create more barriers that observational providers feel they must jump over to make their datasets more user-friendly. Design of campaigns, instrument development, measurement collection, data quality management, and refinement to the final available products is already an enormous undertaking. We should be leveraging cross-community expertise to bring these products and infrastructure to fruition, rather than asking the observational providers to do more themselves. Strategies for the modeling community to provide output that is more compatible with observational datasets as a means to advance evaluation capabilities must also be explored. Modelers and observational communities could also work together to advance and produce more efficient satellite simulator packages—efforts which are current under way across several communities (e.g., Bodas-Salcedo et al. 2011; Clow et al. 2024). There are also other examples, outside of Obs4MIPs, of viable activities where the observational and modeling communities are coming together to improve the usability of observational datasets by communicating where to find them, providing expert guidance on their use in evaluation and discussing new tools and methods for evaluation, such as the Climate Data Guide efforts (https://climatedataguide.ucar.edu/climate-data; Schneider et al. 2013).

Cross-community activities should be designed with an awareness of the increase in model diversity, complexity, archive size, and resolution expected. Additionally, future-ready observations should align with CMIP research questions and CMIP data specifications/requests through cross-community dialogue and coordination. Examples of efforts that have centered

on guidance for diagnostics required for model evaluation include the OMIP (Griffies et al. 2016; Orr et al. 2017) and the Coupled Climate–Carbon Cycle Model Intercomparison Project (C4MIP; Jones et al. 2016). Stakeholders also must be invited into these discussions and efforts as it is important that priorities for model output and new and sustained observational campaigns align with stakeholder needs.

Looking forward, CMIP7 will feature a prioritization of emission-driven simulations (Sanderson et al. 2024) as opposed to simulations driven by prescribed greenhouse gas concentrations. This will require the community to reckon with how to compare model output against observations when the atmospheric CO_2 concentrations may be significantly different from observed due to physical model bias and carbon cycle feedbacks. The CMIP7 archive will also feature models with a greater number of represented processes including carbon cycle processes and with new Earth system components including interactive ice sheets. To address CMIP7 research questions, particularly those aimed to constrain the water-carbon nexus and improve understanding of weather-scale and extreme events, higher spatial-resolution models and higher frequency output will be required, and there must be a sufficient suite of observational datasets ready to interrogate these new simulations.

Modeling advancements motivate the need for continued and new observational strategies and campaigns to sustain and develop the observational datasets required to evaluate state-of-the-art climate models. Supporting efforts aimed to fill gaps in the observational record in data-sparse regions such as World Meteorological Organization's (WMO) Global Basic Observing Network (GBON; WMO 2021) and the Systematic Observations Financing Facility (SOFF; WMO 2021) are important to maintain continuous observational records. A gap analysis for the next decade is going to be crucial to highlight which observations the community may lose soon via funding lapses, instrument retirements, mission/operational shutdowns, etc., which would lead to terminations in critical time series of key climate variables needed for monitoring and continued model evaluation. Maintaining in situ measurements of the subsurface ocean, which cannot be measured by satellites, will be critical toward constraining the planetary heat and carbon budget and providing a continuous time series against which to evaluate models. If funding lapses occur, there is a real risk of space–time gaps emerging within such ocean in situ records with the limited lifetime of Argo floats and the high cost of maintaining a global fleet of research vessels. Future gaps in the instrument record would also impact the quality of reanalysis products and ocean state estimates which rely on observations for data assimilation—impacting the ability to use such global products in evaluation efforts. Thus, as a community, we must collaborate to keep our eyes forward to foresee and plug potential observational gaps.

Building sustainable cross-community connections, whether that means through the organization of workshops or virtual working groups and advancing and expanding existing infrastructure to support future-ready observations to advance climate model evaluation, will require funding. This will inevitably require the scientific and broader community to recognize the importance of these efforts in reducing uncertainty in projected climate change. There is an urgent need to kick start these discussions. Our intention is that this manuscript stimulates efforts to organize and provides guidance on areas of focus for further development and project initiatives.

Acknowledgments. R. L. Beadling was supported under NSF Division of Polar Programs Grant NSF2319828 and NOAA Award NA24OARX431C0057-T1-01. R. Swaminathan was funded by the U.K. Research Infrastructure Natural Environment Research Council (UKRI-NERC) funded TerraFIRMA: Future Impacts, Risks and Mitigation Actions in a changing Earth system Grant (NE/W004895/1) and by ESA from the project Exploiting Satellite Observations for climate model analysis (ESO4clima) under Contract 4000147360/25/I-LR. F. M. Hoffman was supported by the Reducing Uncertainties

in Biogeochemical Interactions through Synthesis and Computation (RUBISCO) Science Focus Area, which is sponsored by the Regional and Global Model Analysis (RGMA) activity of the Earth and Environmental Systems Modeling (EESM) Program in the Earth and Environmental Systems Sciences Division (EESSD) of the Office of Biological and Environmental Research (BER) in the U.S. Department of Energy Office of Science. Oak Ridge National Laboratory (ORNL) is managed by UT-Battelle, LLC, for the U.S. Department of Energy under Contract DE-AC05-00OR22725. Work of J. Lee is performed under auspices of the U.S. Department of Energy (DOE) by Lawrence Livermore National Laboratory (LLNL) under Contract DE-AC52-07NA27344, and effort was supported by the RGMA program of the U.S. DOE's Office of Science (OS), BER program. E. Blockley was supported by the Met Office Hadley Centre Climate Programme funded by DSIT. B. Hassler was supported by the European Union's Horizon 2020 research and innovation programme under Grant Agreement 101003536 (ESM2025—Earth system models for the Future) and by ESA from the project ESO4clima under Contract 4000147360/25/I-LR. B. Medeiros acknowledges support by the U.S. DOE RGMA under Award DE-SC0022070 and NSF IA 1947282 and support by the National Center for Atmospheric Research, which is a major facility sponsored by the NSF under Cooperative Agreement 1852977. S. Brands was supported by the Spanish "Generación de Conocimineto, Convocatoria 2024" project "Contribución Española al Atlas del IPCC-AR7: Desarollo y Problemas Científicos" (PID2024-162703OB-I00), funded by MCIN/AEI/10.13039/501100011033 and by ERDF/EU. E. Scoccimarro gratefully acknowledges the support of the ObsSea4Clim, Ocean observations and indicators for climate and assessments project. Grant Agreement 101136548. J. Tjiputra acknowledges Research Council of Norway funded infrastructure project Infrastructure for Norwegian Earth System Modelling phase 2 (INES2; 350390). D. Watson-Parris acknowledges funding from U.S. National Science Foundation Award 2441832. B. Turner is with the CMIP IPO which is hosted by the European Space Agency, with staff provided on contract by HE Space Operations Ltd. The work of D. Hegedus was supported by the Science and Technology Facilities Council through a graduate placement at the CMIP IPO. The Earth System Grid Federation (ESGF) is an international consortium of individually funded data provider institutions; the ESGF2-U.S. Project in the United States of America is sponsored by the Data Management Program in EESSD of BER in the U.S. Department of Energy Office of Science, and the ESGF activity in the United Kingdom is supported by the Centre for Environmental Data Analysis (CEDA), which is sponsored by the Science and Technology Facilities Council (STFC) and the Natural Environment Research Council (NERC). We thank the following community reviewers who provided feedback prior to manuscript submission: John Krasting, Peter Gleckler, and Yuhan Douglas Rao. The paper concept was initiated in February 2024 as part of the work of the CMIP Model Benchmarking Task Team. The members of the task team at this point were Rebecca Beadling, Ed Blockley, Birgit Hassler, Forrest Hoffman, Jiwoo Lee, Valerio Lembo, Jared Lewis, Jianhua Lu, Luke Madaus, Elizaveta Malinina, Brian Medeiros, Wilfried Pokam, Enrico Scoccimarro, and Ranjini Swaminathan.

Data availability statement. No datasets were generated or analyzed during the current study.

References

Abbasnezhadi, K., and X. L. Wang, 2024: Comparison of gridding methods for precipitation over Canada and assessment of station and data density effects on gridding results. *Atmos.–Ocean*, **62**, 320–346, https://doi.org/10.1080/07055900.2024.2394829.

Abernathey, R. P., and Coauthors, 2021: Cloud-native repositories for big scientific data. *Comput. Sci. Eng.*, **23**, 26–35, https://doi.org/10.1109/MCSE.2021.3059437.

Alexander, L. V., 2016: Global observed long-term changes in temperature and precipitation extremes: A review of progress and limitations in IPCC assessments and beyond. *Wea. Climate Extremes*, **11**, 4–16, https://doi.org/10.1016/j.wace.2015.10.007.

——, and Coauthors, 2016: Implementation plan for WCRP grand challenge on understanding and predicting weather and climate extremes–the "Extremes Grand Challenge". 14 pp., https://www.wcrp-climate.org/images/documents/grand_challenges/WCRP_Grand_Challenge_Extremes_Implementation_Plan_v20150203.pdf.

Anandh, T. S., B. K. Das, J. Kuttippurath, and A. Chakraborty, 2021: A comparative analysis of the Bay of Bengal Ocean state using standalone and coupled numerical models. *Asia-Pac. J. Atmos. Sci.*, **57**, 347–359, https://doi.org/10.1007/s13143-020-00197-z.

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

Bodas-Salcedo, A., and Coauthors, 2011: COSP: Satellite simulation software for model assessment. *Bull. Amer. Meteor. Soc.*, **92**, 1023–1043, https://doi.org/10.1175/2011BAMS2856.1.

Bontemps, S., M. Herold, L. Kooistra, A. van Groenestijn, A. Hartley, O. Arino, I. Moreau, and P. Defourny, 2012: Revisiting land cover observation to address the needs of the climate modeling community. *Biogeosciences*, **9**, 2145–2157, https://doi.org/10.5194/bg-9-2145-2012.

Buizza, R., P. L. Houtekamer, G. Pellerin, Z. Toth, Y. Zhu, and M. Wei, 2005: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Wea. Rev.*, **133**, 1076–1097, https://doi.org/10.1175/MWR2905.1.

Buontempo, C., and Coauthors, 2023: The Copernicus Climate Change Service: ECMWF's C3S puts climate science to work. *Bull. Amer. Meteor. Soc.*, **104**, 804–806, https://doi.org/10.1175/BAMS-D-21-0315.A.

Castro, S. L., G. A. Wick, S. Eastwood, M. A. Steele, and R. T. Tonboe, 2023: Examining the consistency of sea surface temperature and sea ice concentration in Arctic satellite products. *Remote Sens.*, **15**, 2908, https://doi.org/10.3390/rs15112908.

Cavanaugh, N., and S. Shen, 2015: The effects of gridding algorithms on the statistical moments and their trends of daily surface air temperature. *J. Climate*, **28**, 9188–9205, https://doi.org/10.1175/JCLI-D-14-00668.1.

Chen, H., and E. K. Schneider, 2014: Comparison of the SST-forced responses between coupled and uncoupled climate simulations. *J. Climate*, **27**, 740–756, https://doi.org/10.1175/JCLI-D-13-00092.1.

Clow, G. L., N. S. Lovenduski, M. N. Levy, K. Lindsay, and J. E. Kay, 2024: The utility of simulated ocean chlorophyll observations: A case study with the Chlorophyll Observation Simulator Package (version 1) in CESMv2.2. *Geosci. Model Dev.*, **17**, 975–995, https://doi.org/10.5194/gmd-17-975-2024.

CMUG, 2025: Obs4MIPs user requirements and gap analysis report. Climate Modelling User Group: Deliverable 5.7f. ESA Climate Change Initiative, Contract 4000125156/18/I-NB, 52 pp., https://climate.esa.int/media/documents/CMUG_FutureEvolutionofObs4MIPs_D5.7f_v1.2_ieS8qqu.pdf.

Collier, N., F. M. Hoffman, D. M. Lawrence, G. Keppel-Aleks, C. D. Koven, W. J. Riley, M. Mu, and J. T. Randerson, 2018: The International Land Model Benchmarking (ILAMB) system: Design, theory, and implementation. *J. Adv. Model. Earth Syst.*, **10**, 2731–2754, https://doi.org/10.1029/2018MS001354.

Collins, M., K. AchutaRao, K. Ashok, S. Bhandari, A. K. Mitra, S. Prakash, R. Srivastava, and A. Turner, 2013: Observational challenges in evaluating climate models. *Nat. Climate Change*, **3**, 940–941, https://doi.org/10.1038/nclimate2012.

Covey, C., P. J. Gleckler, C. Doutriaux, D. N. Williams, A. Dai, J. Fasullo, K. Trenberth, and A. Berg, 2016: Metrics for the diurnal cycle of precipitation: Toward routine benchmarks for climate models. *J. Climate*, **29**, 4461–4471, https://doi.org/10.1175/JCLI-D-15-0664.1.

Daly, C., 2006: Guidelines for assessing the suitability of spatial climate data sets. *Int. J. Climatol.*, **26**, 707–721, https://doi.org/10.1002/joc.1322.

Deser, C., and Coauthors, 2020: Insights from Earth system model initial-condition large ensembles and future prospects. *Nat. Climate Change*, **10**, 277–286, https://doi.org/10.1038/s41558-020-0731-2.

Diez-Sierra, J., and Coauthors, 2022: The worldwide C3S CORDEX grand ensemble: A major contribution to assess regional climate change in the IPCC AR6 atlas. *Bull. Amer. Meteor. Soc.*, **103**, E2804–E2826, https://doi.org/10.1175/BAMS-D-22-0111.1.

Dunne, J. P., and Coauthors, 2025: An evolving Coupled Model Intercomparison Project Phase 7 (CMIP7) and fast track in support of future climate assessment. *Geosci. Model Dev.*, **18**, 6671–6700, https://doi.org/10.5194/gmd-18-6671-2025.

Durack, P. J., and Coauthors, 2025a: Earth system forcing for CMIP7 and beyond. *Bull. Amer. Meteor. Soc.*, **106**, E1580–E1588, https://doi.org/10.1175/BAMS-D-25-0119.1.

——, K. E. Taylor, M. Mizielinski, C. Doutriaux, D. Nadeau, and M. Juckes, 2025b: CMIP6 Controlled Vocabularies (CVs) Version 6.2.58.79. Zenodo, https://doi.org/10.5281/zenodo.15243879.

——, and Coauthors, 2025c: The Coupled Model Intercomparison Project (CMIP): Reviewing project history, evolution, infrastructure and implementation. EGUsphere, https://doi.org/10.5194/egusphere-2024-3729.

Eaton, B., and Coauthors, 2003: NetCDF Climate and Forecast (CF) metadata conventions. 183 pp., https://cfconventions.org/Data/cf-conventions/cf-conventions-1.8/cf-conventions.pdf.

Eliasson, S., K. G. Karlsson, E. van Meijgaard, J. F. Meirink, M. Stengel, and U. Willén, 2019: The Cloud_cci simulator v1.0 for the Cloud_cci climate data record and its application to a global and a regional climate model. *Geosci. Model Dev.*, **12**, 829–847, https://doi.org/10.5194/gmd-12-829-2019.

Elipot, S., K. Drushka, A. Subramanian, and M. Patterson, 2022: Overcoming the challenges of ocean data uncertainty. *Eos*, **103**, https://doi.org/10.1029/2022EO220021.

Ensor, L., and S. Robeson, 2008: Statistical characteristics of daily precipitation: Comparisons of gridded and point datasets. *J. Appl. Meteor. Climatol.*, **47**, 2468–2476, https://doi.org/10.1175/2008JAMC1757.1.

Evans, J. P., and H. M. Imran, 2024: The observation range adjusted method: A novel approach to accounting for observation uncertainty in model evaluation. *Environ. Res. Commun.*, **6**, 071001, https://doi.org/10.1088/2515-7620/ad5ad8.

Eyring, V., and Coauthors, 2004: Comprehensive summary of the workshop on process-oriented validation of coupled chemistry-climate models. *SPARC Newsletter*, No. 23, SPARC Office, Toronto, ON, Canada, 5–11.

——, and Coauthors, 2005: A strategy for process-oriented validation of coupled chemistry–climate models. *Bull. Amer. Meteor. Soc.*, **86**, 1117–1134, https://doi.org/10.1175/BAMS-86-8-1117.

——, S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor, 2016: 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.

——, and Coauthors, 2020: Earth System Model Evaluation Tool (ESMValTool) v2.0—An extended set of large-scale diagnostics for quasi-operational and comprehensive evaluation of Earth system models in CMIP. *Geosci. Model Dev.*, **13**, 3383–3438, https://doi.org/10.5194/gmd-13-3383-2020.

Fang, H., F. Baret, S. Plummer, and G. Schaepman-Strub, 2019: An overview of global leaf area index (LAI): Methods, products, validation, and applications. *Rev. Geophys.*, **57**, 739–799, https://doi.org/10.1029/2018RG000608.

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

Ferraro, R., D. Waliser, P. Gleckler, K. Taylor, and V. Eyring, 2015: Evolving Obs4MIPs to support phase 6 of the Coupled Model Intercomparison Project (CMIP6). *Bull. Amer. Meteor. Soc.*, **96**, ES131–ES133, https://doi.org/10.1175/BAMS-D-14-00216.1.

Flato, G., and Coauthors, 2013: Evaluation of climate models. *Climate Change 2013: The Physical Science Basis*, T. F. Stocker et al., Eds., Cambridge University Press, 741–866.

Fox-Kemper, B., and Coauthors, 2021: Ocean, cryosphere and sea level change. *Climate Change 2021: The Physical Science Basis*, V. Masson-Delmotte et al., Eds., Cambridge University Press, 1211–1362, https://doi.org/10.1017/9781009157896.011.

Fu, W., J. K. Moore, F. Primeau, N. Collier, O. O. Ogunro, F. M. Hoffman, and J. T. Randerson, 2022: Evaluation of ocean biogeochemistry and carbon cycling in CMIP Earth system models with the International Ocean Model Benchmarking (IOMB) software system. *J. Geophys. Res. Oceans*, **127**, e2022JC018965, https://doi.org/10.1029/2022JC018965.

Gates, W. L., 1992: An AMS continuing series: Global change—AMIP: The Atmospheric Model Intercomparison Project. *Bull. Amer. Meteor. Soc.*, **73**, 1962–1970, https://doi.org/10.1175/1520-0477(1992)073<1962:ATAMIP>2.0.CO;2.

——, and Coauthors, 1999: An overview of the results of the Atmospheric Model Intercomparison Project (AMIP I). *Bull. Amer. Meteor. Soc.*, **80**, 29–55, https://doi.org/10.1175/1520-0477(1999)080<0029:AOOTRO>2.0.CO;2.

Gervais, M., L. B. Tremblay, J. R. Gyakum, and E. Atallah, 2014: Representing extremes in a daily gridded precipitation analysis over the United States: Impacts of station density, resolution, and gridding methods. *J. Climate*, **27**, 5201–5218, https://doi.org/10.1175/JCLI-D-13-00319.1.

Ghorbanpour, A. K., T. Hessels, S. Moghim, and A. Afshar, 2021: Comparison and assessment of spatial downscaling methods for enhancing the accuracy of satellite-based precipitation over Lake Urmia Basin. *J. Hydrol.*, **596**, 126055, https://doi.org/10.1016/j.jhydrol.2021.126055.

Gibson, P. B., S. E. Perkins-Kirkpatrick, P. Uotila, A. S. Pepler, and L. V. Alexander, 2017: On the use of self-organizing maps for studying climate extremes. *J. Geophys. Res. Atmos.*, **122**, 3891–3903, https://doi.org/10.1002/2016JD026256.

——, D. E. Waliser, H. Lee, B. Tian, and E. Massoud, 2019: Climate model evaluation in the presence of observational uncertainty: Precipitation indices over the contiguous United States. *J. Hydrometeor.*, **20**, 1339–1357, https://doi.org/10.1175/JHM-D-18-0230.1.

Gleckler, P. J., K. E. Taylor, and C. Doutriaux, 2008: Performance metrics for climate models. *J. Geophys. Res.*, **113**, D06104, https://doi.org/10.1029/2007JD008972.

Gómez-Navarro, J. J., J. P. Montávez, S. Jerez, P. Jiménez-Guerrero, and E. Zorita, 2012: What is the role of the observational dataset in the evaluation and scoring of climate models? *Geophys. Res. Lett.*, **39**, L24701, https://doi.org/10.1029/2012GL054206.

Good, S. A., M. J. Martin, and N. A. Rayner, 2013: EN4: Quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates. *J. Geophys. Res. Oceans*, **118**, 6704–6716, https://doi.org/10.1002/2013JC009067.

Griffies, S. M., and Coauthors, 2016: OMIP contribution to CMIP6: Experimental and diagnostic protocol for the physical component of the Ocean Model Intercomparison Project. *Geosci. Model Dev.*, **9**, 3231–3296, https://doi.org/10.5194/gmd-9-3231-2016.

Haimberger, L., C. Tavolato, and S. Sperka, 2012: Homogenization of the global radiosonde temperature dataset through combined comparison with reanalysis background series and neighboring stations. *J. Climate*, **25**, 8108–8131, https://doi.org/10.1175/JCLI-D-11-00668.1.

——, F. Ambrogi, and U. Voggenberger, 2024: Bias adjustments for the global historical radiosonde network in preparation for ERA6. *EGU General Assembly 2024*, Vienna, Austria, European Geophysical Union, EGU24-12929, https://doi.org/10.5194/egusphere-egu24-12929.

Hansen, J., R. Ruedy, M. Sato, and K. Lo, 2010: Global surface temperature change. *Rev. Geophys.*, **48**, RG4004, https://doi.org/10.1029/2010RG000345.

Hardiman, S. C., N. Butchart, T. J. Hinton, S. M. Osprey, and L. J. Gray, 2012: The effect of a well-resolved stratosphere on surface climate: Differences between CMIP5 simulations with high and low top versions of the Met Office climate model. *J. Climate*, **25**, 7083–7099, https://doi.org/10.1175/JCLI-D-11-00579.1.

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

Hassler, B., and Coauthors, 2026: Systematic benchmarking of climate models: Methodologies, applications, and new directions. *Rev. Geophys.*, **64**, e2025RG000891, https://doi.org/10.22541/essoar.174196646.65056548/v1.

Haylock, M. R., N. Hofstra, A. M. G. Klein Tank, E. J. Klok, P. D. Jones, and M. New, 2008: A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. *J. Geophys. Res.*, **113**, D20119, https://doi.org/10.1029/2008JD010201.

Hegedűs, D., B. Turner, S. Ferdini, and C. Macintosh, 2026: The climate modelling-observation interface. Zenodo, https://doi.org/10.5281/zenodo.14886515.

Henson, S. A., C. Beaulieu, and R. Lampitt, 2016: Observing climate change trends in ocean biogeochemistry: When and where. *Global Change Biol.*, **22**, 1561–1571, https://doi.org/10.1111/gcb.13152.

Herrera, S., J. Fernández, and J. M. Gutiérrez, 2016: Update of the Spain02 gridded observational dataset for EURO-CORDEX evaluation: Assessing the effect of the interpolation methodology. *Int. J. Climatol.*, **36**, 900–908, https://doi.org/10.1002/joc.4391.

——, R. M. Cardoso, P. M. Soares, F. Espírito-Santo, P. Viterbo, and J. M. Gutiérrez, 2019: Iberia01: A new gridded dataset of daily precipitation and temperatures over Iberia. *Earth Syst. Sci. Data*, **11**, 1947–1956, https://doi.org/10.5194/essd-11-1947-2019.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146**, 1999–2049, https://doi.org/10.1002/qj.3803.

Hewitt, H. T., and Coauthors, 2020: Resolving and parameterising the ocean mesoscale in Earth system models. *Curr. Climate Change Rep.*, **6**, 137–152, https://doi.org/10.1007/s40641-020-00164-w.

Hoffman, F. M., and Coauthors, 2025: Rapid evaluation framework for the CMIP7 assessment fast track. EGUsphere, https://doi.org/10.5194/egusphere-2025-2685.

Hofstra, N., M. Haylock, M. New, P. Jones, and C. Frei, 2008: Comparison of six methods for the interpolation of daily, European climate data. *J. Geophys. Res.*, **113**, D21110, https://doi.org/10.1029/2008JD010100.

Jones, C. D., and Coauthors, 2016: C4MIP—The Coupled Climate–Carbon Cycle Model Intercomparison Project: Experimental protocol for CMIP6. *Geosci. Model Dev.*, **9**, 2853–2880, https://doi.org/10.5194/gmd-9-2853-2016.

Jung, M., and Coauthors, 2020: Scaling carbon fluxes from eddy covariance sites to globe: Synthesis and evaluation of the FLUXCOM approach. *Biogeosciences*, **17**, 1343–1365, https://doi.org/10.5194/bg-17-1343-2020.

Jung, S., J. Gil, M. Lee, C. Betancourt, M. Schultz, Y. Choi, T. Joo, and D. Kim, 2025: Interpolation of missing ozone data using graph machine learning and parameter analysis through eXplainable artificial intelligence comparison. *Environ. Modell. Software*, **190**, 106466, https://doi.org/10.1016/j.envsoft.2025.106466.

Kaps, A., A. Lauer, and V. Eyring, 2023: CCClim - A machine-learning powered cloud class climatology. Zenodo, https://doi.org/10.5281/zenodo.8369202.

Karl, T. R., and Coauthors, 1995: Critical issues for long-term climate monitoring. *Climatic Change*, **31**, 185–221, https://doi.org/10.1007/BF01095146.

Kay, J. E., and Coauthors, 2015: The Community Earth System Model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability. *Bull. Amer. Meteor. Soc.*, **96**, 1333–1349, https://doi.org/10.1175/BAMS-D-13-00255.1.

Keeling, C. D., S. C. Piper, R. B. Bacastow, M. Wahlen, T. P. Whorf, M. Heimann, and H. A. Meijer, 2001: Exchanges of atmospheric CO_2 and $^{13}CO_2$ with the terrestrial biosphere and oceans from 1978 to 2000. I. Global aspects. SIO Reference Series, No. 01-06, Scripps Institution of Oceanography, 88 pp., https://escholarship.org/uc/item/09v319r9.

Kennedy, J. J., 2014: A review of uncertainty in in situ measurements and data sets of sea surface temperature. *Rev. Geophys.*, **52** (1), 1–32, https://doi.org/10.1002/2013RG000434.

Kim, S., J. Nathaniel, Z. Hou, T. Zheng, and P. Gentine, 2024: Spatiotemporal upscaling of sparse air-sea pCO2 data via physics-informed transfer learning. *Sci. Data*, **11**, 1098, https://doi.org/10.1038/s41597-024-03959-w.

Klein, S. A., and C. Jakob, 1999: Validation and sensitivities of frontal clouds simulated by the ECMWF model. *Mon. Wea. Rev.*, **127**, 2514–2531, https://doi.org/10.1175/1520-0493(1999)127<2514:VASOFC>2.0.CO;2.

——, Y. Zhang, M. D. Zelinka, R. Pincus, J. Boyle, and P. J. Gleckler, 2013: Are climate model simulations of clouds improving? An evaluation using the ISCCP simulator. *J. Geophys. Res. Atmos.*, **118**, 1329–1342, https://doi.org/10.1002/jgrd.50141.

Kondrashov, D., and M. Ghil, 2006: Spatio-temporal filling of missing points in geophysical data sets. *Nonlinear Processes Geophys.*, **13**, 151–159, https://doi.org/10.5194/npg-13-151-2006.

Kotlarski, S., and Coauthors, 2019: Observational uncertainty and regional climate model evaluation: A pan-European perspective. *Int. J. Climatol.*, **39**, 3730–3749, https://doi.org/10.1002/joc.5249.

Landschützer, P., and Coauthors, 2015: The reinvigoration of the Southern Ocean carbon sink. *Science*, **349**, 1221–1224, https://doi.org/10.1126/science.aab2620.

Langland, R. H., R. N. Maue, and C. H. Bishop, 2008: Uncertainty in atmospheric temperature analyses. *Tellus*, **60**, 598–603, https://doi.org/10.1111/j.1600-0870.2008.00336.x.

Lauer, A., L. Bock, B. Hassler, M. Schröder, and M. Stengel, 2023: Cloud climatologies from global climate models—A comparison of CMIP5 and CMIP6 models with satellite data. *J. Climate*, **36**, 281–311, https://doi.org/10.1175/JCLI-D-22-0181.1.

Lee, J., and Coauthors, 2024: Systematic and objective evaluation of Earth system models: PCMDI Metrics Package (PMP) version 3. *Geosci. Model Dev.*, **17**, 3919–3948, https://doi.org/10.5194/gmd-17-3919-2024.

Lenssen, N., G. A. Schmidt, M. Hendrickson, P. Jacobs, M. Menne, and R. Ruedy, 2024: A GISTEMPv4 observational uncertainty ensemble. *J. Geophys. Res. Atmos.*, **129**, e2023JD040179, https://doi.org/10.1029/2023JD040179.

Lepot, M., J.-B. Aubin, and F. H. L. R. Clemens, 2017: Interpolation in time series: An introductive overview of existing methods, their performance criteria and uncertainty assessment. *Water*, **9**, 796, https://doi.org/10.3390/w9100796.

Le Traon, P. Y., and Coauthors, 2019: From observation to information and users: The Copernicus Marine Service perspective. *Front. Mar. Sci.*, **6**, 234, https://doi.org/10.3389/fmars.2019.00234.

Li, Z. L., B. H. Tang, H. Wu, H. Ren, G. Yan, Z. Wan, I. F. Trigo, and J. A. Sobrino, 2013: Satellite-derived land surface temperature: Current status and perspectives. *Remote Sens. Environ.*, **131**, 14–37, https://doi.org/10.1016/j.rse.2012.12.008.

Longman, R. J., A. J. Newman, T. W. Giambelluca, and Lucas, M., 2020: Characterizing the uncertainty and assessing the value of gap-filled daily rainfall data in Hawaii. *J. Appl. Meteor. Climatol.*, **59**, 1261–1276, https://doi.org/10.1175/JAMC-D-20-0007.1.

Maher, N., S. Milinski, and R. Ludwig, 2021: Large ensemble climate model simulations: Introduction, overview, and future prospects for utilising multiple types of large ensemble. *Earth Syst. Dyn.*, **12**, 401–418, https://doi.org/10.5194/esd-12-401-2021.

Matthews, J. L., E. Mannshardt, and P. Gremaud, 2013: Uncertainty quantification for climate observations. *Bull. Amer. Meteor. Soc.*, **94**, ES21–ES25, https://doi.org/10.1175/BAMS-D-12-00042.1.

McMillan, H., T. Krueger, and J. Freer, 2012: Benchmarking observational uncertainties for hydrology: Rainfall, river discharge and water quality. *Hydrol. Processes*, **26**, 4078–4111, https://doi.org/10.1002/hyp.9384.

Meehl, G., T. H. Teng, and J. Arblaster, 2014: Climate model simulations of the observed early-2000s hiatus of global warming. *Nat. Climate Change*, **4**, 898–902, https://doi.org/10.1038/nclimate2357.

Meinshausen, M., and Coauthors, 2017: Historical greenhouse gas concentrations for climate modelling (CMIP6). *Geosci. Model Dev.*, **10**, 2057–2116, https://doi.org/10.5194/gmd-10-2057-2017.

Merchant, C. J., and O. Embury, 2014: Simulation and inversion of satellite thermal measurements. *Exp. Methods Phys. Sci.*, **47**, 489–526, https://doi.org/10.1016/B978-0-12-417011-7.00015-5.

——, and Coauthors, 2017: Uncertainty information in climate data records from Earth observation. *Earth Syst. Sci. Data*, **9**, 511–527, https://doi.org/10.5194/essd-9-511-2017.

Mittaz, J., C. J. Merchant, and E. R. Woolliams, 2019: Applying principles of metrology to historical Earth observations from satellites. *Metrologia*, **56**, 032002, https://doi.org/10.1088/1681-7575/ab1705.

Mizuochi, H., and Coauthors, 2021: Multivariable evaluation of land surface processes in forced and coupled modes reveals new error sources to the simulated water cycle in the IPSL (Institute Pierre Simon Laplace) climate model. *Hydrol. Earth Syst. Sci.*, **25**, 2199–2221, https://doi.org/10.5194/hess-25-2199-2021.

Moroni, D., and Coauthors, 2019: Understanding the various perspectives of Earth science observational data uncertainty. ESIP Rep., 34 pp., https://doi.org/10.6084/m9.figshare.10271450.v1.

Mueller, B., and Coauthors, 2013: Benchmark products for land evapotranspiration: LandFlux-EVAL multi-data set synthesis. *Hydrol. Earth Syst. Sci.*, **17**, 3707–3720, https://doi.org/10.5194/hess-17-3707-2013.

Nakamura, H., and Coauthors, 2025: Toward future reanalyses that meet evolving needs in science, public services, policymaking, and socioeconomic activity. *Bull. Amer. Meteor. Soc.*, **106**, E1445–E1453, https://doi.org/10.1175/BAMS-D-25-0126.1.

Notz, D., 2015: How well must climate models agree with observations? *Philos. Trans. Roy. Soc.*, **A373**, 20140164, https://doi.org/10.1098/rsta.2014.0164.

Nowack, P., J. Runge, V. Eyring, and J. D. Haigh, 2020: Causal networks for climate model evaluation and constrained projections. *Nat. Commun.*, **11**, 1415, https://doi.org/10.1038/s41467-020-15195-y.

Orr, J. C., and Coauthors, 2017: Biogeochemical protocols and diagnostics for the CMIP6 Ocean Model Intercomparison Project (OMIP). *Geosci. Model Dev.*, **10**, 2169–2199, https://doi.org/10.5194/gmd-10-2169-2017.

Petrie, R., and Coauthors, 2021: Coordinating an operational data distribution network for CMIP6 data. *Geosci. Model Dev.*, **14**, 629–644, https://doi.org/10.5194/gmd-14-629-2021.

Poschlod, B., and J. Koh, 2024: Convection-permitting climate models can support observations to generate rainfall return levels. *Water Resour. Res.*, **60**, e2023WR035159, https://doi.org/10.1029/2023WR035159.

Potter, G. L., L. Carriere, J. Hertz, M. Bosilovich, D. Duffy, T. Lee, and D. N. Williams, 2018: Enabling reanalysis research using the Collaborative Reanalysis Technical Environment (CREATE). *Bull. Amer. Meteor. Soc.*, **99**, 677–687, https://doi.org/10.1175/BAMS-D-17-0174.1.

Prein, A. F., and A. Gobiet, 2017: Impacts of uncertainties in European gridded precipitation observations on regional climate analysis. *Int. J. Climatol.*, **37**, 305–327, https://doi.org/10.1002/joc.4706.

Richardson, A. D., and D. Y. Hollinger, 2007: A method to estimate the additional uncertainty in gap-filled NEE resulting from long gaps in the CO_2 flux record. *Agric. For. Meteor.*, **147**, 199–208, https://doi.org/10.1016/j.agrformet.2007.06.004.

Righi, M., and Coauthors, 2020: 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.

Roberts, M. J., and Coauthors, 2018: The benefits of global high resolution for climate simulation: Process understanding and the enabling of stakeholder decisions at the regional scale. *Bull. Amer. Meteor. Soc.*, **99**, 2341–2359, https://doi.org/10.1175/BAMS-D-15-00320.1.

——, and Coauthors, 2025: High-Resolution Model Intercomparison Project Phase 2 (HighResMIP2) towards CMIP7. *Geosci. Model Dev.*, **18**, 1307–1332, https://doi.org/10.5194/gmd-18-1307-2025.

Roemmich, D., and J. Gilson, 2009: The 2004–2008 mean and annual cycle of temperature, salinity, and steric height in the global ocean from the Argo Program. *Prog. Oceanogr.*, **82**, 81–100, https://doi.org/10.1016/j.pocean.2009.03.004.

Sanderson, B. M., and Coauthors, 2024: The need for carbon-emissions-driven climate projections in CMIP7. *Geosci. Model Dev.*, **17**, 8141–8172, https://doi.org/10.5194/gmd-17-8141-2024.

Sarmiento, J. L., and Coauthors, 2023: The Southern Ocean carbon and climate observations and modeling (SOCCOM) project: A review. *Prog. Oceanogr.*, **219**, 103130, https://doi.org/10.1016/j.pocean.2023.103130.

Sayer, A. M., and Coauthors, 2020: A review and framework for the evaluation of pixel-level uncertainty estimates in satellite aerosol remote sensing. *Atmos. Meas. Tech.*, **13**, 373–404, https://doi.org/10.5194/amt-13-373-2020.

Schmidt, G. A., and Coauthors, 2023: Anomalous meltwater from ice sheets and ice shelves is a historical forcing. *Geophy. Res. Lett.*, **50**, e2023GL106530, https://doi.org/10.1029/2023GL106530.

Schneider, D. P., C. Deser, J. Fasullo, and K. E. Trenberth, 2013: Climate data guide spurs discovery and understanding. *Eos, Trans. Amer. Geophys. Union*, **94**, 121–122, https://doi.org/10.1002/2013EO130001.

Schneider, R., and Coauthors, 2022: ESA-ECMWF report on recent progress and research directions in machine learning for Earth system observation and prediction. *npj Climate Atmos. Sci.*, **5**, 51, https://doi.org/10.1038/s41612-022-00269-z.

Schutgens, N., S. Tsyro, E. Gryspeerdt, D. Goto, N. Weigum, M. Schulz, and P. Stier, 2017: On the spatio-temporal representativeness of observations. *Atmos. Chem. Phys.*, **17**, 9761–9780, https://doi.org/10.5194/acp-17-9761-2017.

Schutgens, N. A. J., E. Gryspeerdt, N. Weigum, S. Tsyro, D. Goto, M. Schulz, and P. Stier, 2016: Will a perfect model agree with perfect observations? The impact of spatial sampling. *Atmos. Chem. Phys.*, **16**, 6335–6353, https://doi.org/10.5194/acp-16-6335-2016.

Scoccimarro, E., P. G. Fogli, K. A. Reed, S. Gualdi, S. Masina, and A. Navarra, 2017: Tropical cyclone interaction with the ocean: The role of high-frequency (subdaily) coupled processes. *J. Climate*, **30**, 145–162, https://doi.org/10.1175/JCLI-D-16-0292.1.

Séférian, R., and Coauthors, 2020: Tracking improvement in simulated marine biogeochemistry between CMIP5 and CMIP6. *Curr. Climate Change Rep.*, **6**, 95–119, https://doi.org/10.1007/s40641-020-00160-0.

Semmler, T., and Coauthors, 2020: Simulations for CMIP6 with the AWI climate model AWI-CM-1-1. *J. Adv. Model. Earth Syst.*, **12**, e2019MS002009, https://doi.org/10.1029/2019MS002009.

Sidorenko, D., and Coauthors, 2015: Towards multi-resolution global climate modeling with ECHAM6–FESOM. Part I: Model formulation and mean climate. *Climate Dyn.*, **44**, 757–780, https://doi.org/10.1007/s00382-014-2290-6.

Simpson, I. R., and Coauthors, 2025: Confronting Earth System Model trends with observations. *Sci. Adv.*, **11**, eadt8035, https://doi.org/10.1126/sciadv.adt8035.

Singh, G., G. Moncrieff, Z. Venter, K. Cawse-Nicholson, J. Slingsby, and T. B. Robinson, 2024: Uncertainty quantification for probabilistic machine learning in Earth observation using conformal prediction. *Sci. Rep.*, **14**, 16166, https://doi.org/10.1038/s41598-024-65954-w.

Slivinski, L. C., and Coauthors, 2019: Towards a more reliable historical reanalysis: Improvements for version 3 of the Twentieth Century Reanalysis system. *Quart. J. Roy. Meteor. Soc.*, **145**, 2876–2908, https://doi.org/10.1002/qj.3598.

Sloyan, B. M., C. C. Chapman, R. Cowley, and A. A. Charantonis, 2023: Application of machine learning techniques to ocean mooring time series data. *J. Atmos. Oceanic Technol.*, **40**, 241–260, https://doi.org/10.1175/JTECH-D-21-0183.1.

Smalley, M. A., M. D. Lebsock, and J. Teixeira, 2023: Quantifying the impact of vertical resolution on the representation of marine boundary layer physics for global-scale models. *Mon. Wea. Rev.*, **151**, 2977–2992, https://doi.org/10.1175/MWR-D-23-0078.1.

Smith, G. C., and Coauthors, 2019: Polar ocean observations: A critical gap in the observing system and its effect on environmental predictions from hours to a season. *Front. Mar. Sci.*, **6**, 429, https://doi.org/10.3389/fmars.2019.00429.

Stephens, G. L., and C. D. Kummerow, 2007: The remote sensing of clouds and precipitation from space: A review. *J. Atmos. Sci.*, **64**, 3742–3765, https://doi.org/10.1175/2006JAS2375.1.

Sterl, A., 2004: On the (in)homogeneity of reanalysis products. *J. Climate*, **17**, 3866–3873, https://doi.org/10.1175/1520-0442(2004)017<3866:OTIORP>2.0.CO;2.

Stern, C., R. Abernathey, J. Hamman, R. Wegener, C. Lepore, S. Harkins, and A. Merose, 2022: Pangeo forge: Crowdsourcing analysis-ready, cloud optimized data production. *Front. Climate*, **3**, 782909, https://doi.org/10.3389/fclim.2021.782909.

Teixeira, J., D. E. Waliser, R. Ferraro, P. Gleckler, T. Lee, and G. Potter, 2014: Satellite observations for CMIP5: The genesis of Obs4MIPs. *Bull. Amer. Meteor. Soc.*, **95**, 1329–1334, https://doi.org/10.1175/BAMS-D-12-00204.1.

Tramontana, G., and Coauthors, 2016: Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms. *Biogeosciences*, **13**, 4291–4313, https://doi.org/10.5194/bg-13-4291-2016.

Trenberth, K. E., Y. Zhang, and M. Gehne, 2017: Intermittency in precipitation: Duration, frequency, intensity, and amounts using hourly data. *J. Hydrometeor.*, **18**, 1393–1412, https://doi.org/10.1175/JHM-D-16-0263.1.

Tselioudis, G., W. B. Rossow, C. Jakob, J. Remillard, D. Tropf, and Y. Zhang, 2021: Evaluation of clouds, radiation, and precipitation in CMIP6 models using global weather states derived from ISCCP-H cloud property data. *J. Climate*, **34**, 7311–7324, https://doi.org/10.1175/JCLI-D-21-0076.1.

Waliser, D., and Coauthors, 2020: Observations for Model Intercomparison Project (Obs4MIPs): Status for CMIP6. *Geosci. Model Dev.*, **13**, 2945–2958, https://doi.org/10.5194/gmd-13-2945-2020.

Walsh, J. E., V. M. Kattsov, W. L. Chapman, V. Govorkova, and T. Pavlova, 2002: Comparison of Arctic climate simulations by uncoupled and coupled global models. *J. Climate*, **15**, 1429–1446, https://doi.org/10.1175/1520-0442(2002)015<1429:COACSB>2.0.CO;2.

Wang, B., X. Xie, and L. Li, 2009: A review on aspects of climate simulation assessment. *Adv. Atmos. Sci.*, **26**, 736–747, https://doi.org/10.1007/s00376-009-9038-y.

Webb, M. J., and Coauthors, 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.

Wehner, M., J. Lee, M. Risser, P. Ullrich, P. Gleckler, and W. D. Collins, 2021: Evaluation of extreme sub-daily precipitation in high-resolution global climate model simulations. *Philos. Trans. Roy. Soc.*, **A379**, 20190545, https://doi.org/10.1098/rsta.2019.0545.

Wei, M., Z. Tóth, and Y. Zhu, 2010: Analysis differences and error variance estimates from multi-centre analysis data. *Aust. Meteor. Oceanogr. J.*, **59**, 25–34, https://doi.org/10.22499/2.5901.005.

Wicker, W., I. Polichtchouk, and D. I. V. Domeisen, 2023: Increased vertical resolution in the stratosphere reveals role of gravity waves after sudden stratospheric warmings. *Wea. Climate Dyn.*, **4**, 81–93, https://doi.org/10.5194/wcd-4-81-2023.

Wilkinson, M. D., and Coauthors, 2016: The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*, **3**, 160018, https://doi.org/10.1038/sdata.2016.18.

Williams, D. N., and Coauthors, 2016: A global repository for planet-sized experiments and observations. *Bull. Amer. Meteor. Soc.*, **97**, 803–816, https://doi.org/10.1175/BAMS-D-15-00132.1.

Williams, E., C. Funk, P. Peterson, and C. Tuholske, 2024: High resolution climate change observations and projections for the evaluation of heat-related extremes. *Sci. Data*, **11**, 261, https://doi.org/10.1038/s41597-024-03074-w.

WMO, 2021: World Meteorological Congress: Abridged final report of the extraordinary session. WMO-1281, 247 pp., https://library.wmo.int/.

Yin, G., G. Mariethoz, Y. Sun, and M. F. McCabe, 2017: A comparison of gap-filling approaches for Landsat-7 satellite data. *Int. J. Remote Sens.*, **38**, 6653–6679, https://doi.org/10.1080/01431161.2017.1363432.

Yu, A., and Coauthors, 2024: Northern Hemisphere stratosphere-troposphere circulation change in CMIP6 models: 2. Mechanisms and sources of the spread. *J. Geophys. Res. Atmos.*, **129**, e2024JD040823, https://doi.org/10.1029/2024JD040823.

Yuan, W., and Coauthors, 2010: Global estimates of evapotranspiration and gross primary production based on MODIS and global meteorology data. *Remote Sens. Environ.*, **114**, 1416–1431, https://doi.org/10.1016/j.rse.2010.01.022.

Zumwald, M., B. Knüsel, C. Baumberger, G. Hirsch Hadorn, D. N. Bresch, and R. Knutti, 2020: Understanding and assessing uncertainty of observational climate datasets for model evaluation using ensembles. *Wiley Interdiscip. Rev.: Climate Change*, **11**, e654, https://doi.org/10.1002/wcc.654.

Zwiers, F. W., and Coauthors, 2013: Climate extremes: Challenges in estimating and understanding recent changes in the frequency and intensity of extreme climate and weather events. *Climate Science for Serving Society: Research, Modeling and Prediction Priorities*, G. R. Asrar and J. W. Hurrell, Eds., Springer, 339–389.