Article Title: Metrics as tools for bridging climate science and applications

Authors:

Kevin A. Reed*

School of Marine and Atmospheric Sciences, State University of New York at Stony Brook, Stony Brook, NY, USA

ORCID: 0000-0003-3741-7080

Naomi Goldenson

Department of Atmospheric and Oceanic Sciences, University of California, Los Angeles, CA, USA

Richard Grotjahn

Department of Land, Air and Water Resources, University of California, Davis, CA, USA

William J. Gutowski

Department of Geological and Atmospheric Sciences, Iowa State University, Ames, IA, USA

Kripa Jagannathan

Climate and Ecosystem Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

Andrew D. Jones

Climate and Ecosystem Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA;

Energy and Resources Group, University of California, Berkeley, CA, USA

L. Ruby Leung

Atmospheric Sciences and Global Change Division, Pacific Northwest National Laboratory, WA, Richland, USA

Seth A. McGinnis

National Center for Atmospheric Research, Boulder, CO, USA

Sara C. Pryor

Department of Earth and Atmospheric Sciences, Cornell University, Ithaca, NY, USA

Abhishekh K. Srivastava

Department of Land, Air and Water Resources, University of California, Davis, CA, USA

Paul A. Ullrich

Department of Land, Air and Water Resources, University of California, Davis, CA, USA

Colin M. Zarzycki

Department of Meteorology and Atmospheric Science, Penn State University, State College, PA, USA

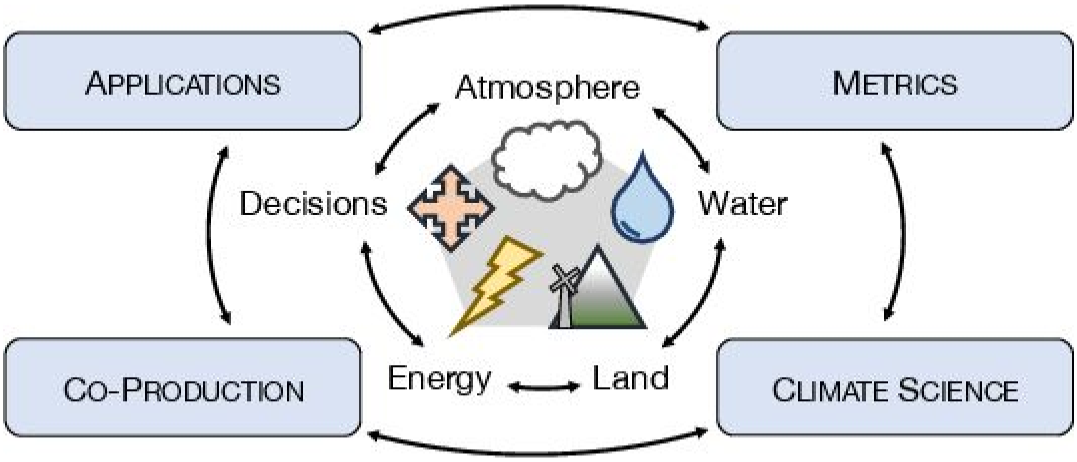
Conflict of Interest

The authors have declared that there is no conflict of interests.

Abstract

In climate science and applications, the term "metric" is used to describe the distillation of complex, multi-faceted evaluations to summarize the overall quality of a model simulation, or other data product, and/or as a means to quantify some response to climate change. Metrics provide insights into the fidelity of processes and outcomes from climate models and can assist with both differentiating models' representation of variables or processes and informing whether models are "fit for purpose." Metrics can also provide a valuable reference point for co-production of knowledge between climate scientists and climate impact practitioners. While continued metric developments enable model developers to better understand the impacts of decisions made in the model design process, metrics also have implications for the characterization of uncertainty and facilitating analyses of underlying physical processes. As a result, comprehensive evaluation with multiple metrics enhances usability of climate information by both scientific and stakeholder communities. This paper presents examples of insights gained from the development and appropriate use of metrics, and provides examples of how metrics can be used to engage with stakeholders and inform decision-making.

Graphical/Visual Abstract and Caption



Visual Abstract Caption: Metrics are essential tools within climate science. When climate scientists and climate impact practitioners co-design metrics then climate science and decision-making can both be advanced.

1. INTRODUCTION

A crucial challenge for bridging the gap between climate science and its use in applications is often termed as the "Practitioner's Dilemma" of how to synthesise the vast amount of climate information, assess its credibility, characterise associated uncertainties, and use the information appropriately for specific management needs (Barsugli et al., 2013, Briley et al., 2020, Jagannathan et al., 2020, Moss et al., 2019). Suitably derived metrics that synthesise complex numerical information, can assist with this bridging, though one should recognize that the term "metric" has a variety of uses and applications. For our purposes here, a metric is defined as a quantifiable measure that distills a complex, multi-faceted climate simulation or dataset into a small set of numbers or categories. These metrics can subsequently be used for (a) summarizing impact-relevant quantities, (b) evaluating features or aspects of the simulation, or (c) assessing changes and variability within a dataset or differences across multiple datasets. For example, the metric "total precipitation in a region" can be quantified, compared to observations, or its evolution assessed in light of climate change. Quantities such as statistical skill scores, which assess model performance against some benchmark dataset, are also often considered as metrics. Notably, terms like skill-score, metric, indicator, statistic, etc. are used interchangeably.

Climate metrics can serve varied purposes for both practitioners and scientists (Jagannathan et al., 2020). Some metrics can quantitatively describe key climatic features, phenomena, and processes that are relevant both from a scientific and a management perspective. Other metrics allow for efficient comparison of many climate simulations and evaluation of their fitness for different purposes, which can both push scientific improvements in climate modelling, and at the same time assist practitioners in selection of the most appropriate climate information for their decision-contexts (Briley et al. 2020, Jagannathan et al., 2020). Often, a single metric is not sufficient. Climate scientists and practitioners alike typically use collections of metrics that have multiple components or dimensions, such as spatial maps, vectors, or temporal series, to probe spatiotemporal characteristics and relationships. Summary metrics such as skill scores or statistics can then be derived through the application of dimension reduction to these metric collections (Taylor, 2001; Wehner, 2013; Collier et al., 2018).

Since metrics have utility for both scientists and practitioners, they can also serve as a crucial boundary object – that is, a common interface in knowledge co-production processes (Shackley and Wyne 1996, Sarkki et al., 2020). Co-production is a process where scientists work together with users of science, to iteratively and collaboratively develop actionable knowledge and practises (Wyborn et al., 2019, Mach et al., 2020). Boundary objects are described as graphs, maps, scenarios, indicators, or other concepts that allow for diverse groups to simultaneously project disparate interpretations (i.e., they have interpretive flexibility). Thus they can provide a nexus for communication and collaboration for diverse groups to effectively work together (Franco-Torres 2020, Sarkki et al., 2020, Turnhout 2009). By distilling complex climate model outputs into a succinct set of perspectives, metrics can act as effective boundary objects, and allow scientists and practitioner groups to collaboratively explore where, when, and how climate information can be improved to address needs and concerns of stakeholders.

This paper explores and discusses how different metrics can be developed and employed using examples drawn from three U.S. Department of Energy multidisciplinary projects: Hyperion, FACETS, and HyperFACETS. In these examples, we also describe how these metrics were used to engage with practitioners and inform decision-making, especially for managing water and energy resources. The remainder of the perspective is structured as follows. Section 2 provides an overview of different metric types within the context of science and applications. Section 3 details four example metrics and their relevance for specific decision applications. Finally, summary remarks on the desirable properties of metrics and a discussion of lessons learned are presented in Section 4.

2. FEATURES AND USES OF METRICS

Metrics can be used to perform analysis in a uniform, standardized, comparable, and reproducible way, for purposes such as informing model development via identification of areas where model skill is lacking, or evaluating the degree to which physical processes and phenomena are represented (Ekstrom et al., 2018; Zarzycki et al., 2021; Srivastava et al., 2020; Pryor & Schoof, 2019, 2020; Coburn & Pryor 2021; Xue and Ullrich, 2021). Metrics can also quantify the degree to which a data set is credible or “fit for purpose” for particular applications (Jagannathan et al., 2020, Briley et al., 2020).

Metrics can take many forms, and can be phenomena-based and/or derived from statistics that describe temporal, geospatial, or distributional properties of the data. For example, metrics can describe statistical properties of a variable at different timescales and locations (e.g., CLIMDEX;Alexander et al., 2011). They can transform model output into a form that has impact relevance e (e.g., Rx5day for flooding). Additional metrics for precipitation (e.g., Pendergrass et al., 2020) include fitting of probability distributions to daily precipitation amounts as generated from climate models and determining the degree to which the resulting distributions approximate observations using skill scores (Kjellstrom et al., 2010) to summarize the fidelity of models and forecasts.

Metrics can also be derived to describe fidelity with respect to spatial features and variability. For example, spatial metrics of extreme precipitation events over the contiguous United States (CONUS) have been developed and offer a potential extension to these simple metrics in that they can be used to evaluate a model's representation of the location, scale, and magnitude of extreme precipitation events, with the fidelity represented using standard statistical skill scores (Mondal et al., 2020). Two examples of such object-based diagnostic evaluation of precipitation are the Method for Object-Based Diagnostic Evaluation (Bullock et al., 2016) and TempestExtremes (Ullrich & Zarzycki, 2017, Ullrich et al., 2021).

Phenomena-based metrics can be employed that require fidelity for the parameter of interest (e.g., the spatial field of seasonally accumulated precipitation) and the process(es) responsible for that outcome. One such example is predicated on the process-level connection between moisture advection by the low-level jet and precipitation associated with the southwestern monsoon (Bukovsky et al., 2013, 2015).

Just as there are many types of metrics, there are many ways they are used in practice. Model developers might be more interested in whether the underlying processes are properly represented and the model adheres to principles of physical realism such as conservation laws. Users of climate model output may be more concerned with the ability to reproduce particular historical statistics with high fidelity. Indeed, the appropriate evaluation of model credibility is important before users turn their attention to model projections. However, a sole focus on historical reproduction of a small number of outcome metrics can be misleading since it is possible for models to "get the right answers for the wrong reasons" due to tuning or compensating errors.

Users of climate projections can vary based on sectors, regions, decision-context, etc. Hence, the metrics that may be relevant to potential users can be extremely context-specific, even to describe the same broad phenomenon such as extreme precipitation. In addition, specific-user context also dictates how the information contained in the metric would need to be aggregated. Both spatial and temporal aggregation can be inherent to the definitions of certain metrics, or a part of a subsequent distillation. For example, a user interested in extreme precipitation first needs to choose the metric they prefer (e.g., the maximum 24-hour precipitation in a year) from a large suite of possibilities. The metric can be calculated for many spatial locations, perhaps in gridded data, and visualized on a map. Whether further distillation is required depends on the specific application. A city planner might scan the map for the value over their town, while regional stakeholders looking to evaluate relative model skill, might compare the spatial pattern of this metric over a larger area using a pattern correlation, Root Mean Square Error, or other statistical measures. For the latter user, this distillation step is necessary for efficient comparison of different simulations, drawing into question the boundary between a metric and a tailored analysis. Similar issues arise for time-aggregation. Making these metrics easily accessible to users presents a challenge for data developers, especially when some summary metrics require "on-the-fly" calculations, like aggregations over stakeholder-defined regions and time periods, using multiple archived variables.

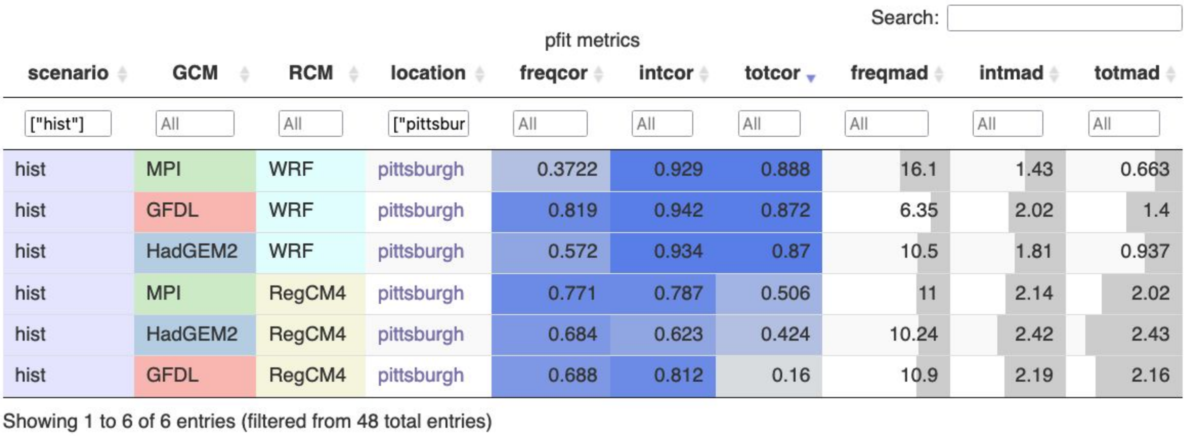
To navigate the seemingly ever-expanding sets of available metrics, many stakeholders might find it helpful to consult experts who are more familiar with the models and metrics in question. One example of potential misuse is the evaluation of a statistical model using metrics the model was trained to reproduce, which would not be appropriate. Similarly, not all users may understand how to evaluate data from climate simulations that are or are not bias-corrected without expert guidance. Ideally, non-expert users would have access to assistance in selecting metrics to best inform their judgments of fitness for purpose. In practice, many rely on metrics packages and public data, making clear definitions and tools all the more important.

3. EXAMPLE METRICS

We now provide examples of different metrics to gain insights from the appropriate use of metrics and context regarding how metrics can be developed and applied with stakeholder input to inform decision-making.

3.1 Automated Diagnostic Comparison Metrics

Process-based model evaluation is a critical tool for determining whether a climate simulation produces credible results that represent the underlying process(es) properly. However, it requires significant time, effort, and expertise, and the number of processes and process interactions that could be examined is very large. For FACETS, we developed a set of statistical metrics for targeting and accelerating process-based model evaluation. These metrics summarize a suite of diagnostic analyses of temperature and precipitation (McGinnis, 2019). We apply these analyses in an automated fashion at multiple locations across an ensemble of different simulations, then collect the metrics into a table that can be dynamically sorted and filtered for comparison. Examining the metrics this way allows easy identification of cases where the differences between simulations matter to the analysis. Examining the associated analyses then provides insight into the nature of the differences, which can guide a process-level analysis, as shown in the following example.



Search:

				pfit metrics					
scenario	GCM	RCM	location	freqcor	intcor	totcor	freqmad	intmad	totmad
["hist"]	All	All	["pittsbur	All	All	All	All	All	All
hist	MPI	WRF	pittsburgh	0.3722	0.929	0.888	16.1	1.43	0.663
hist	GFDL	WRF	pittsburgh	0.819	0.942	0.872	6.35	2.02	1.4
hist	HadGEM2	WRF	pittsburgh	0.572	0.934	0.87	10.5	1.81	0.937
hist	MPI	RegCM4	pittsburgh	0.771	0.787	0.506	11	2.14	2.02
hist	HadGEM2	RegCM4	pittsburgh	0.684	0.623	0.424	10.24	2.42	2.43
hist	GFDL	RegCM4	pittsburgh	0.688	0.812	0.16	10.9	2.19	2.16

Showing 1 to 6 of 6 entries (filtered from 48 total entries)

Figure 1. Summary metrics for the automated "pfit" diagnostic analysis of the annual precipitation cycle for Pittsburgh. Detailed descriptions of the correlation (corr) and median absolute deviation (MAD) metrics can be found in the text. The table has been sorted by the values in the total amount correlation (tot corr) column, which shows a clear distinction between results from the WRF simulations (correlation 0.87 and above) and the RegCM4 simulations (correlation 0.51 and below). The software that produces this analysis and calculates the metrics also generates an HTML table with color-coding and dynamic filtering and sorting to highlight these kinds of patterns as shown here. See text for more information.

Figure 1 shows metrics for an analysis that diagnoses the annual precipitation cycle of Pittsburgh for six simulations from NA-CORDEX (Mearns, 2017). These simulations form a mini-ensemble of two regional climate models (RCMs: WRF and RegCM4) downscaling three global climate models (GCMs: GFDL-ESM2M, MPI-ESM-LR, and HadGEM2-ES) from the Coupled Model Intercomparison Project Phase 5 (CMIP5; Taylor et al., 2012). The analysis shows the multi-year averages of daily precipitation frequency, intensity, and total amount using a 30-day moving window for a historical period and a future period and compares them to observations. This analysis is summarized by six values: the Pearson parametric correlation (corr) with observations and the median absolute deviation (MAD) from observations for the frequency (freq), intensity (int), and total (tot) precipitation curves.

The clear differences in two correlation metrics for the RCMs suggests that a comparison of this diagnostic analysis for these two models is warranted. Figure 2 shows results for the simulations driven by HadGEM2-ES, which are representative. The average annual cycle of total precipitation displays a dramatic increase in bias vs observations in the summer months for one model but not the other, which suggests that there is an important difference between these two RCMs with regard to how they simulate precipitation in the summer. The differences between RCMs are similar regardless of the driving GCM, and summer precipitation in this region is mostly convective; two RCMs also use different convective parameterization schemes. Altogether, this suggests that a process-based analysis of model bias in warm-season precipitation in this region should focus on the effects of convective parameterization and the skill enhancement that can be achieved by convection-permitting grid spacing (Lucas-Picher et al., 2021).

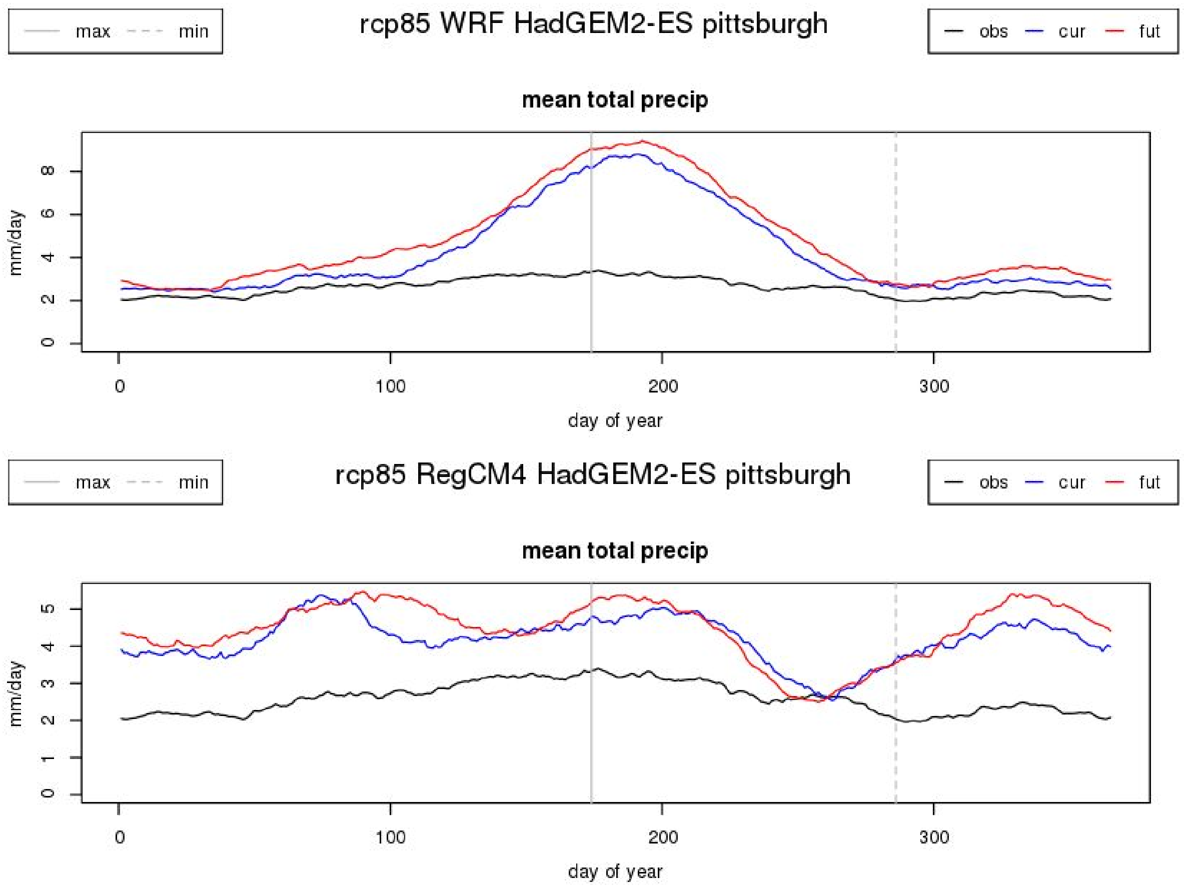


Figure 2. Excerpted results from the automated precipitation diagnostic analysis for Pittsburgh showing daily mean total precipitation for a 30-day running window across multiple years for outputs from WRF (top) and RegCM4 (bottom) RCMs, both driven by HadGEM2-ES. Observations (from Livneh et al., 2015) are shown in black, and historical (future) simulations in blue (red). The WRF simulation shows a dramatic positive bias in the warm season (n.b. the differing vertical scales). Note that this figure shows results from the automated diagnostic analysis "as-is," not polished for publication. Other outputs of this analysis (not shown) analyze precipitation frequency and intensity. Summary metrics are shown in Figure 1.

This example focuses on a difference relevant to model evaluation, but the same procedure can target process-based analysis for understanding climate change signals and evaluating their credibility. Using metrics to summarize automated diagnostic analyses allows the user to rapidly constrain and target an in-depth process analysis. Presenting metrics in an interactive color-coded table can also assist users in rapidly assessing model performance for a particular application (Pryor & Schoof, 2020; Zarzycki et al., 2021).

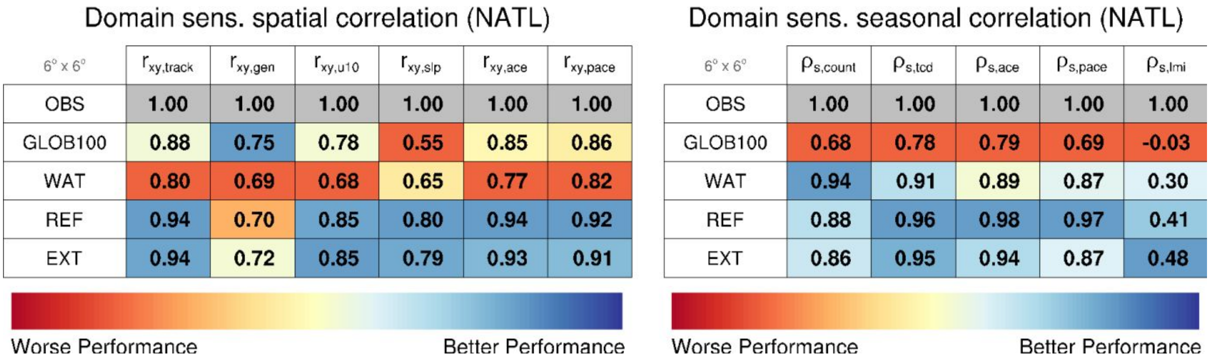
Although process-based model evaluation may not be thought of as relevant to stakeholders, during 5+ years of co-production engagement in Hyperion and HyperFACETS, practitioners often indicated a desire to better understand processes driving management-relevant climatic changes in their domain, and the extent to which these processes are represented in climate models. This was particularly true for practitioners in regions where precipitation projections were highly uncertain (e.g., Upper Colorado and South Florida regions). As one water manager stated:

"We need to push the models to look at processes. What are the processes behind the projected changes for my region? How well are models able to capture these? How might they change over time? How can I track their changes over time?"

The other benefit of metrics to stakeholders highlighted by this example is the fact that metrics are standardized, and typically designed to be automatable and reusable across different contexts. This allows the use of a metric "off the shelf" rather than developing a custom evaluation, knowing that the metric was created by experts to capture salient features of the data. We caution that "off the shelf" metrics does not imply that the metrics can be trusted blindly as measures of quality; co-production and carefully crafted guidance on proper metric use, can guard against misuse. Although every stakeholder has different needs, metrics provide a common starting point for understanding that can accelerate the assessment process.

3.2 Phenomena-Specific Storm Metrics

Regional policy is often shaped by historical events of significant impact – for instance, memorable droughts, floods, or extreme storms. And so, beyond an understanding of climatological fields such as temperatures and precipitation, stakeholders are often interested in how well models predict the frequency, intensity, spatial scale or other characteristics of particular weather features, and subsequently, what the most credible models say about their future change. Atmospheric phenomena such as tropical cyclones (TCs), windstorms resulting from intense extratropical cyclones, mesoscale convective systems (MCSs), and atmospheric rivers, are among the most important drivers for extreme and hazardous precipitation. Understanding the atmospheric context for such phenomena, the effect of climate change on these processes, and their relative contributions to annual total and extreme precipitation can allow scientists to better understand likely future extreme and hazardous precipitation regimes. Phenomena-specific storm metrics provide a window into whether or not models simulate these features realistically. Exhaustive investigation of the characteristics of atmospheric features can give rise to an understanding of the upstream drivers responsible for biases in feature climatology (e.g., errors in TC frequency due to biases in sea-surface temperature), and can allow scientists to understand whether models preserve known functional relationships between feature characteristics (e.g., wind-pressure relationship in TCs; Chavas et al. 2017).



Domain sens. spatial correlation (NATL)

6° x 6°	$r_{xy,track}$	$r_{xy,gen}$	$r_{xy,u10}$	$r_{xy,slp}$	$r_{xy,ace}$	$r_{xy,pace}$
OBS	1.00	1.00	1.00	1.00	1.00	1.00
GLOB100	0.88	0.75	0.78	0.55	0.85	0.86
WAT	0.80	0.69	0.68	0.65	0.77	0.82
REF	0.94	0.70	0.85	0.80	0.94	0.92
EXT	0.94	0.72	0.85	0.79	0.93	0.91

Domain sens. seasonal correlation (NATL)

6° x 6°	$\rho_{s,count}$	$\rho_{s,tcd}$	$\rho_{s,ace}$	$\rho_{s,pace}$	$\rho_{s,lmi}$
OBS	1.00	1.00	1.00	1.00	1.00
GLOB100	0.68	0.78	0.79	0.69	-0.03
WAT	0.94	0.91	0.89	0.87	0.30
REF	0.88	0.96	0.98	0.97	0.41
EXT	0.86	0.95	0.94	0.87	0.48

Figure 3. Stoplight diagrams assessing the climatology of tracked North Atlantic tropical storms in different variable-resolution configurations of the Community Earth System Model in Zarzycki et al. (2021). Shown are both spatial (left) and temporal (right) correlations for four different grids: uniform 1º, 0.25º western Atlantic only, 0.25º entire North Atlantic, and 0.25º entire North Atlantic plus northern Africa. Each column (i.e., metric) is color-coded by skill relative to a reference (observations) to reflect the different baseline correlations associated with different metrics.

Several CONUS examples of phenomena-specific storm metrics have been developed as part of the HyperFACETS project. TC metrics developed by Zarzycki et al. (2021) have provided insights into model design decisions that are important to correctly capturing TC characteristics such as the stoplight diagrams shown in Figure 3. In related work, Stansfield et al. (2020) investigated the contribution of TCs to precipitation totals. A phenomena-specific evaluation for warm-season MCSs and associated large-scale meteorological patterns has also recently been developed by Feng et al. (2021). Additional foci include storm-level snowfall and its proximity to population centers (Zarzycki, 2018), characterization of composite extreme events (e.g., windstorms with co-occurring frozen precipitation) and the structure and spatial scales of extreme wind speeds (Letson et al., 2021). From the scientist's perspective, such metrics allow for the identification of datasets that represent characteristics of specific phenomena sufficiently well to then be used to quantify changes in these characteristics on longer timescales. From practitioners' perspectives, such phenomena-specific metrics allow them to explore critical "what if scenario" important for planning and regulation applications. As one practitioner from South Florida suggested:

> *"Using climate models and the ability to estimate the contribution of TCs brings significant progress. It is valuable to know that for a future hurricane event similar to Irma, intensified rainfall per storm hour could be seen. Both accumulated rainfall and rainfall intensity matter."*

Another practitioner further elaborated that:

> *"It helps us think about questions like: can Harvey-like storms and associated rainfall happen here? Can the land-falling hurricanes stall in South Florida? What is the most vulnerable track for flooding in South Florida? How much extreme rainfall per storm will increase under future conditions?"*

3.3 Use-Inspired IDF Curves

Metrics can aid in answering decision-relevant questions about the effects of climate change on the hydrological cycle. For instance, a variety of stakeholders are interested in probable changes in the intensity-duration-frequency (IDF) curves of precipitation in a non-stationary climate (Cheng & AghaKouchak, 2014). IDF curves form the basis of engineering design standards for urban infrastructure and stormwater drainage systems. The utility of information about these changes is dependent on the specific decision, infrastructure, or region under consideration. Hence stakeholders are concerned with changes in precipitation associated with different combinations of duration and return period (e.g., 24-hr precipitation over a 25-year return period). But in all cases, stakeholders have a more general objective: how to reduce uncertainty in IDF estimates, and in projected IDF changes. The uncertainty in IDF estimates arises from multiple sources, including lack of long-term observational and model records, systematic model biases, and choice of extreme value distribution. Therefore, a focus of the Hyperion and HyperFACETS projects has been investigating methods to reduce uncertainty in IDF estimates. A “unified” framework for estimating IDF curves was applied over the Sacramento-San Joaquin and Kissimmee-Southern Florida watersheds (Srivastava et al., 2019) using L-moments based on the regional frequency analysis (RFA) method proposed by Hosking and Wallis (1997). This method uses the data from nearby *homogeneous* stations to enhance the sample size of the target station; the workflow involves station-specific estimation of IDF curves, use of kriging for spatial interpolation of the point estimates, and use of a z-score based metric to statistically compare two IDF estimates (and so taking into account uncertainty around the estimates).

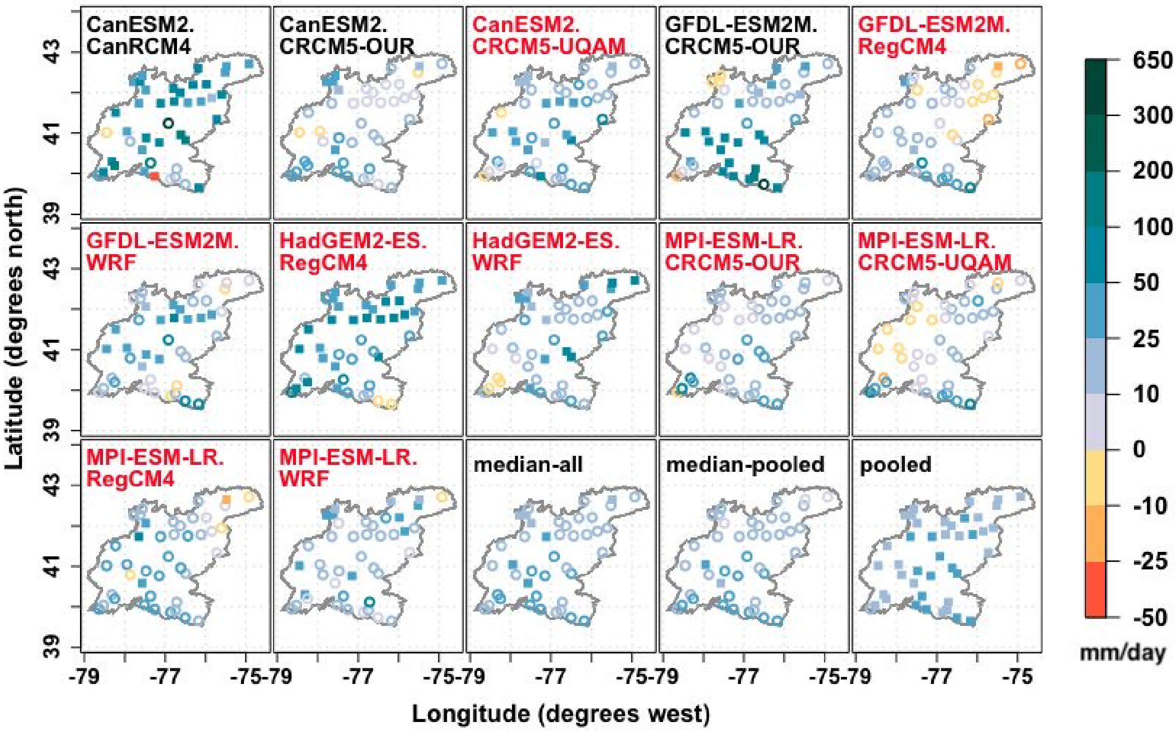


Figure 4. Changes in 24-hr precipitation for 25-year return period over the Susquehanna watershed estimated from NA-CORDEX simulations. Differences significant at the 5% significance level are shown as solid squares and those not significant at 5% are shown as open circles. The significance is computed from the z-statistic as described in Srivastava et al., 2021. Models shown in red are the models that perform reasonably in simulating the mean and variability of the observed annual maximum precipitation using the criteria defined in Srivastava et al., 2021, hence, these models are used for pooling. The "median-all" panel shows the median of 24-hr precipitation estimates from all models, while the "median-pooled" shows the median of 24-hr precipitation estimates from models that are used for pooling. "Pooled" shows changes in 24-hr precipitation estimates computed from pooled models. Units are in mm/day.

Stakeholders showed concern about the large estimation uncertainty in multimodel IDF estimates. Consequently, a novel methodology was developed for reducing uncertainty in the multi-model IDF estimates, which consists of three steps: (i) historical evaluation of climate models, (ii) bias correct the reasonably performing models, and (iii) pooling the bias-corrected model data (Srivastava et al., 2021). Since climate model performance is highly dependent on region (Srivastava et al., 2020; Srivastava et al., 2021), models should be assessed separately for each region so that underperforming models can be excluded for IDF estimation in that region. Using Monte Carlo simulations it was shown that multi-model IDF estimates based upon pooling of model data have smaller biases (difference between model and observation-based IDF estimates) and uncertainty (confidence interval) than traditional median-based multi-model IDF estimates. When applied to simulations from the NA-CORDEX project (Mearns, 2017), the proposed pooling-based method projects statistically significant changes in 24-hr precipitation estimates at more stations in the Susquehanna watershed than the individual model based estimates or the traditional median-based multi-model estimates (Fig. 4). We do not recommend automization of the above procedure, and emphasize that sufficient caution be exercised when selecting models for comparing IDF curves for two nearby geographic regions.

Considering the vast number of applications in stormwater management and flood protection that rely on IDF curves, this novel methodology for better characterising uncertainties in models was greatly appreciated by water managers - particularly for assessing storms with longer return periods where lack of long-term records of data can become limiting. One manager stated that:

"For IDF curves, the discussions about selecting best model performance, and pooled models/ ensembles was very important, and will inform how we will be assessing climate data that we are currently working in partnership with USGS."

Another practitioner stated that:

"These new results have reinforced the value of the approach of pooling best performer climate models together to reduce uncertainty."

3.4 LSMP-based Metrics for Visualization, Understanding, and Model Selection

Metrics using Large-scale meteorological pattern (LSMP) approaches enable co-production through effective visualization, improved physical understanding, and informing model selection. Early work on LSMPs by Grotjahn & Faure (2008) was coordinated with forecasters at a local National Weather Service forecast office (WSFO). Through interaction between scientists and WSFO forecasters, we improved composite maps consulted when forecasting extreme weather by introducing statistical identification of those parts of the pattern that are important. Some pattern parts of consequence were unknown to them before our work, such as: a ridge in the southeastern US occurs for a cold air outbreak on the US West Coast.

Extreme weather maps are a natural bridge between scientists and practitioners. The circulation patterns associated with extremes are useful to examine, and have included individual maps (e.g., Gutowski et al., 2008) or groupings of maps using composites (e.g., Gutowski et al., 2010; Schoof et al., 2019), Self-Organizing Maps (e.g., Glisan et al., 2016; Smalley et al., 2019; Song et al., 2019; Dong et al., 2021), or clustering (e.g., Lee & Grotjahn, 2016). LSMP metrics have since expanded beyond a forecast tool to explore drivers of heat waves (Grotjahn, 2011) and extreme event categories, including heavy precipitation events (Grotjahn & Faure, 2008).

Identifying LSMPs is a two-step procedure. First, an average pattern is formed for the type of extreme of interest for a region (Lackmann & Gyakum, 1999; Grotjahn et al., 2016). Second, the meaningful aspects of the pattern are identified; notably, LSMPs differ from composites or other averaged patterns because they are statistically significant (Agel et al., 2019; Hu et al., 2019; Marquardt Collow et al., 2016) and/or consistent (Gao et al., 2014; Gao & Schlosser, 2019) – preferably, both.

To illustrate how consistency matters, Grotjahn and Faure (2008) showed composite averages for extreme precipitation in Sacramento, California which includes a strong atmospheric ridge over western Alaska and a cut-off low off the northern California coast. All such extreme precipitation events there have that low (Grotjahn and Faure, 2008; Chen et al., 2021). While both the ridge and the trough are highly significant, only the trough occurs with high consistency. Combining the LSMP approach with clustering reveals four clusters of patterns during northern California extreme precipitation (Fig. 5) but only two clusters have a ridge near Alaska. Figure 5 shows that different clusters have significantly different dynamical and thermodynamical evolution.

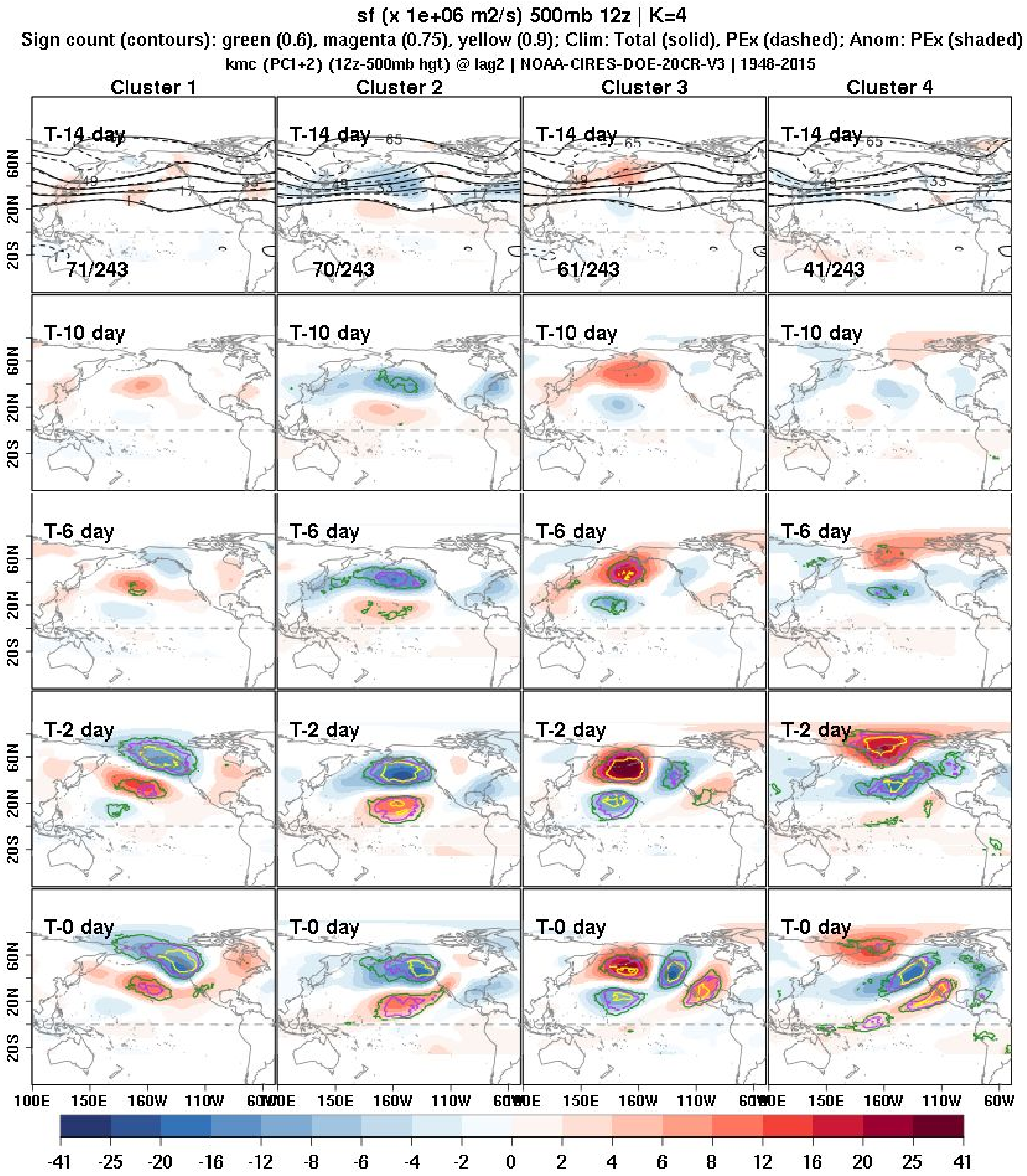

sf (x 1e+06 m2/s) 500mb 12z | K=4
Sign count (contours): green (0.6), magenta (0.75), yellow (0.9); Clim: Total (solid), PEx (dashed); Anom: PEx (shaded)
kmc (PC1+2) (12z-500mb hgt) @ lag2 | NOAA-CIRES-DOE-20CR-V3 | 1948-2015
Cluster 1
Cluster 2
Cluster 3
Cluster 4
T-14 day
T-10 day
T-6 day
T-2 day
T-0 day
71/243
70/243
61/243
41/243
60N
20N
20S
100E
150E
160W
110W
60W
-41
-25
-20
-16
-12
-8
-6
-4
-2
0
2
4
6
8
12
16
20
25
41

Figure 5. 500mb streamfunction (sf500) anomaly composites for the four clusters obtained by *k*-means clustering of NOAA-CIRES-DOE 20th Century Reanalysis V3 500 hPa daily geopotential height anomalies (Za500) at 12z on extreme precipitation days over Northern California in winter (NDJFM) season during 1948-2015. Daily anomalies are with respect to the long term daily mean. Extreme precipitation days are defined as days when daily precipitation spatially-averaged over Northern California exceeds the 95th percentile of the spatially-averaged precipitation (P) on days when P > 0 mm/day. Only sf500 anomaly values passing a bootstrap significance test at the 5% significance level are shaded (variable interval in the colorbar, units m2/s). Solid black contours in top panels show the composites of total sf500 fields for all NDJFM days in the 1948-2015 period. The dashed black contours in top panels show composites of total sf500 fields for each cluster. Colored contours indicate *sign counts* of sf500 anomaly at each grid point. The green contour is 0.6 sign count meaning 80% of the cluster members have the same sign there. Similarly: the magenta contour means 0.75, 87.5% ; the yellow contour means 0.9, 95%. T-0 indicates the day of the extreme event onset (bottom row). Corresponding plots at 2-14 days (T-2 to T-14) before onset show the evolution towards the pattern at onset.

The ratio at bottom left of each panel indicates the total number of events in the cluster out of the total extreme events considered for compositing (243).

Using only highly significant and consistent parts of the pattern, one can project LSMPs from one or more variables onto corresponding instantaneous weather conditions and formulate an index or metric, LSMPi. (Grotjahn (2011) calculated a LSMPi "Circulation Index" by multiplying composite and daily anomaly fields at each key region grid point then summing the resultant values, finally normalizing by a corresponding sum of composite values squared. Grotjahn (2011) used 700 hPa meridional wind and 850 hPa temperature.). The LSMPi can be applied in several practitioner-relevant ways: (i) showing basic statistics of extremes (Grotjahn, 2016), (ii) distinguishing extremes with similar impacts but potentially different large-scale drivers (Lee & Grotjahn, 2016), (iii) ensuring that climate models produce the correct statistics of extremes for the right reasons (Grotjahn, 2013; Gao et al., 2014), (iv) objectively weighing model skill (Grotjahn & Lee, 2016; Palipane & Grotjahn, 2018), or (v) addressing broad dynamical questions like whether future occurrence properties of the extreme arise from more variability or a time mean change (Palipane & Grotjahn, 2018). Practitioners using climate model data as input to their own models are keen to know what climate models are best for their application. When used as evaluation tools, LSMPi data are relevant for selecting global modelling systems to use for dynamical downscaling.

4. CONCLUSIONS AND DISCUSSION

Metrics play an important role in (i) identifying deficiencies in climate simulations or datasets, and (ii) identifying processes, regions, or phenomena for which datasets and models are particularly credible. Ensuring the credibility of these products, even if only for certain applications, is essential for both advancing the science and ensuring that stakeholders are using the best products for their needs and in a manner consistent with the data quality. In our experience, data users remain frustrated by insufficient or incomplete product evaluation and a lack of expert guidance (as discussed in, e.g., Briley et al., 2020). The development of robust, relevant, and intuitive metric packages provides a means forward for addressing this gap. At the same time, these metric packages are also highly relevant for advancing climate science, including building an understanding of the systems being represented by these models. This motivates the need for greater collaboration between scientists and stakeholders to identify quantities of relevance, and to explain why success in representing those quantities may vary across data products.

Several of the metric applications discussed in Section 3 can be used by stakeholders to guide the selection of climate model outputs for use in impacts modeling or decision-making. We highlight a few approaches taken within U.S. Department of Energy projects to enhance co-production of climate knowledge using metrics that have high value to practitioners. We anticipate new metrics will emerge over time as the context changes, the science and datasets improve, and the user and application needs evolve. This is particularly important because metrics can be invalidated by changes in climate; for example, an empirical-statistical metric based on the duration and timing of seasonal snow cover will cease to be relevant if the region shifts to ephemeral snow that melts completely between snowfall events. This highlights the need for iterative metric-based model evaluation frameworks, and for monitoring systems that measure change in the climate.

In our experience with knowledge co-production, some lessons learned for use and development of metrics have emerged:

1. A wide variety of different kinds of metrics are critical for adequately assessing climate simulations or datasets.

2. Metrics that are more easily understood or can be explained clearly and transparently are better at stimulating discourse between scientists and practitioners.

3. A diverse group of scientists and practitioners is ideal for the co-production of metric suites, and for extracting meaning from subsequent analysis with those suites.

This approach to metrics has been effective in guiding work with stakeholders on decision-relevant science in the Hyperion and HyperFACETS projects. While the design and selection of metrics enables the continued improvement of climate model performance, these metrics are also useful for the characterization of uncertainty and improved understanding of underlying physical processes to enhance usability of climate simulations or datasets by broader communities.

Not all metrics are equally useful - indeed, the utility of a particular suite of metrics can be highly dependent on the experiment being performed or the questions being asked. For example, if a climate simulation is being used by a stakeholder to assess the return times of a particular extreme precipitation event, metrics should likely be focused on how well a model simulates the range of intensities of precipitation and the associated processes for extreme precipitation. On the other hand, if a simulation will be employed for regional downscaling to examine extremes, then it's more important for the simulation to correctly capture large-scale patterns and associated teleconnections. A remaining challenge is the effective communication of fidelity as measured by standardized metrics across a model ensemble. Visualizations for metrics of relative model skill using a portrait diagram (e.g., Srivastava et al., 2020) or a stoplight color-coding of credibility have been developed and applied to allow synthesis of different aspects from a suite of models or model realizations (Pryor & Schoof, 2020; Zarzycki et al., 2021).

Finally, this work has provided some insights into what are desirable properties of successful metrics and collections of metrics. We have summarized some desirable properties of metrics and collections of metrics in Table 1. In general, metrics should be easily interpretable, informative for evaluating the dataset of need or interest, and capable of distinguishing statistically significant differences. Collections of metrics should consist of distinct metrics, should be comprehensive for the need or interest, and should leverage all available observations of sufficient quality. Techniques like principal feature analysis or principal component analysis may be used to identify relationships between metrics within collections (Xue & Ullrich, 2021).

Metrics should be...
...interpretable and intuitive: Metrics should be easy to understand and should relate to processes or variables that have physical meaning. It should be clear how they are computed (i.e., the procedure to obtain a metric should not be a "black box").
...informative: Metrics should convey useful/actionable information about models and data products. If employed to differentiate models or data products there should be a way to understand what choices may be responsible for those differences. Further, if essentially all input datasets can produce the same value for a given metric, that metric conveys little information of value.
...significant: The reference data used as a benchmark for a given metric should be sufficiently constrained so that differences between models and the reference are significant, in a statistical sense. If observations are poorly constrained then it's difficult to ascertain if a given data product is inconsistent with those observations.
Collections of metrics should be...
...distinct: Metric collections should describe quantities that are not, in effect, duplicates of each other. Along these lines, it is generally valuable to assess correlations among metrics within a given collection to understand how many unique metrics are present, and to identify quantities that may have a common upstream process or are otherwise related.
...comprehensive: Metric collections should address many, if not all, possible aspects of a particular process or feature (as long as individual metrics meet the criteria above).
...using all available observations: Metric collections should leverage all available observations, both to gauge observational uncertainty and to ensure the collection is comprehensive.

Table 1. Desirable properties of metrics and collections of metrics.

When considering these desirable properties of successful metrics, the scientific and stakeholder communities should also prioritize reusability (documentation, software, etc.) of metric frameworks among groups developing and evaluating metrics. Further, note that open-sourced technological toolsets like Coordinated Model Evaluation Capabilities and TempestExtremes (Ullrich et al., 2021) make it easy to evaluate and visualize multiple metrics and their relationships for broad community use. Finally, more co-production and collaboration among science and user communities is needed within the climate impacts, adaptation, and mitigation fields.

Funding Information

This work was funded with the support of the Department of Energy Office of Science award number DE-SC0016605, "A Framework for Improving Analysis and Modeling of Earth System and Intersectoral Dynamics at Regional Scales (HyperFACETS)". Some of this material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement No. 1852977.

Acknowledgments

KAR provided overall leadership of the manuscript, including the general drafting of the manuscript and supervising the review and editing process. WJG, LRL, KJ and CMZ led the drafting of Section 1. NG, SCP and ADJ led the drafting of Section 2. SAM, AKS, PAU, RG, KJ and CMZ led the drafting, analysis and visualizations of Section 3. KAR and PAU led the drafting of Section 4. All authors were involved in the editing and review of the full draft manuscript. PAU leads the project administration and supervision of HyperFACETS (http://climate.ucdavis.edu/hyperfacets/index.php).

For data from NA-CORDEX, we acknowledge the World Climate Research Programme's CORDEX Science Advisory Team, the Working Group on Regional Climate (former coordinating body of CORDEX), and the Working Group on Coupled Modelling (responsible panel for CMIP5). We also thank the climate modelling groups from the National Center for Atmospheric Research, Iowa State University, and the University of Arizona for producing and making available their model output. We also acknowledge the U.S. Department of Defense ESTCP for its support of the NA-CORDEX data archive.

References

Agel, L., Barlow, M., Colby, F., Binder, H., Catto, J., Hoel, A., & Cohen, J. (2019). Dynamical analysis of extreme precipitation in the US northeast based on large-scale meteorological patterns. Climate Dynamics, 52 , 1739-1760.

Alexander, L., Donat, M., Takayama, Y., & Yang, H. (2011). The climdex project: creation of long-term global gridded products for the analysis of temperature and precipitation extremes. In WCRP Open Science Conference, Denver.

Barsugli, J. J., Guentchev, G., Horton, R. M., Wood, A., Mearns, L. O., Liang, X. Z., ... & Ammann, C. (2013). The practitioner's dilemma: How to assess the credibility of downscaled climate projections. Eos, Transactions American Geophysical Union, 94(46), 424-425

Briley, L., Kelly, R., Blackmer, E. D., Troncoso, A. V., Rood, R. B., Andresen, J., & Lemos, M. C. (2020). Increasing the Usability of Climate Models through the Use of Consumer-Report-Style Resources for Decision-Making. Bulletin of the American Meteorological Society, 101(10), E1709-E1717.

Bukovsky, M. S., Carrillo, C. M., Gochis, D. J., Hammerling, D. M., McCrary, R. R., & Mearns, L. O. (2015). Toward assessing narccap regional climate model credibility for the north american monsoon: future climate simulations. Journal of Climate, 28 (17), 6707-6728.

Bukovsky, M. S., Gochis, D. J., & Mearns, L. O. (2013). Towards assessing narccap regional climate model credibility for the north american monsoon: Current climate simulations. Journal of Climate, 26 (22), 8802-8826.

Bullock, R., Brown, B., & Fowler, T. (2016). Method for object-based diagnostic evaluation. NCAR Technical Notes NCAR/TN-532+STR.

Chavas, D. R., Reed, K. A., & Knaff, J. A. (2017). Physical understanding of the tropical cyclone wind-pressure relationship. Nature Communications, 8 (1), 1-11.

Chen, D., Norris, J., Goldenson, N., Thackeray, C., & Hall, A. (2021). A distinct atmospheric mode for california precipitation. Journal of Geophysical Research: Atmospheres, e2020JD034403.

Cheng, L., & AghaKouchak, A. (2014). Nonstationary precipitation intensity-duration-frequency curves for infrastructure design in a changing climate. *Scientific reports*, *4*(1), 1-6.

Coburn, J.J., and Pryor S.C. (2021): Differential credibility of climate modes in CMIP6. Journal of Climate **34** 8145-8164 doi: 10.1175/jcli-d-21-0359.1

Collier, N., Hoffman, F. M., Lawrence, D. M., Keppel-Aleks, G., Koven, C. D., Riley, W. J., . . . Randerson, J. T. (2018). The International Land Model Benchmarking (ILAMB) system: design, theory, and implementation. Journal of Advances in Modeling Earth Systems, 10 (11), 2731-2754.

Dong, L., Leung, L. R., Qian, Y., Zou, Y., Song, F., & Chen, X. (2021). Meteorological environments associated with california wildfires and their potential roles in wildfire changes during 1984-2017. Journal of Geophysical Research: Atmospheres, 126 (5), e2020JD033180.

Ekstrom, M., Gutmann, E. D., Wilby, R. L., Tye, M. R., & Kirono, D. G. (2018). Robustness of hydroclimate metrics for climate change impact research. WIREs Water, 5 (4), e1288. doi: 10.1002/wat2.1288

Feng, Z., Song, F., Sakaguchi, K., & Leung, L. R. (2021). Evaluation of mesoscale convective systems in climate simulations: Methodological development and results from mpas-cam over the United States. Journal of Climate, 34 (7), 2611-2633.

Franco-Torres, M., Rogers, B. C., & Ugarelli, R. M. (2020). A framework to explain the role of boundary objects in sustainability transitions. Environmental Innovation and Societal Transitions, 36, 34-48.

Gao, X., & Schlosser, A. (2019). Mid-western US heavy summer-precipitation in regional and global climate models: the impact on model skill and consensus through an analogue lens. Climate Dynamics, 52 , 1569-1582.

Gao, X., Schlosser, A., Xie, P., Monier, E., & Entekhabi, D. (2014). An analogue approach to identify heavy precipitation events: Evaluation and application to CMIP5 climate models in the united states. Journal of Climate, 27 (15), 5941-5963.

Glisan, J. M., Gutowski Jr, W. J., Cassano, J. J., Cassano, E. N., & Seefeldt, M. W. (2016). Analysis of WRF extreme daily precipitation over Alaska using self organizing maps. Journal of Geophysical Research: Atmospheres, 121 (13), 7746-7761.

Grotjahn, R. (2011). Identifying extreme hottest days from large scale upper air data: a pilot scheme to   nd California Central Valley summertime maximum surface temperatures. Climate Dynamics, 37 (3-4), 587-604. doi:10.1007_s00382-011-0999-z

Grotjahn, R. (2013). Ability of CCSM4 to simulate California extreme heat conditions from evaluating simulations of the associated large scale upper air pattern. Climate Dynamics, 41 (5-6), 1187-1197. doi: 10.1007/s00382-013-1668-1

Grotjahn, R. (2016). Western North American extreme heat, associated large scale synoptic-dynamics, and performance by a climate model. In J. Li, R. Swinbank, R. Grotjahn, & H. Volkert (Eds.), Dynamics and predictability of large- scale, high-impact weather and climate events (p. 198-209). Cambridge, England: Cambridge University Press, (ISBN 978-1-107-07142-1)

Grotjahn, R., Black, R., Leung, R., Wehner, M. F., Barlow, M., Bosilovich, M., . . . Prabhat (2016). North American extreme temperature events and related large scale meteorological patterns: a review of statistical methods, dynamics, modeling, and trends. Climate Dynamics, 46 , 1151-1184. doi: 10.1007/s00382-015-2638-6

Grotjahn, R., & Faure, G. (2008). Composite predictor maps of extraordinary weather events in the Sacramento California region. Weather and Forecasting,, 23 , 313-335. doi: 10.1175/2007WAF2006055.1

Grotjahn, R., & Lee, Y.-Y. (2016). On climate model simulations of the large-scale meteorology associated with California heat waves. Journal of Geophysical Research: Atmospheres,, 121 , 18-32. doi: 10.1002/2015JD024191

Gutowski, W. J., Arritt, R. W., Kawazoe, S., Flory, D. M., Takle, E. S., Biner, S., . . . others (2010). Regional extreme monthly precipitation simulated by narccap rcms. Journal of Hydrometeorology, 11 (6), 1373-1379.

Gutowski Jr, W. J., Willis, S. S., Patton, J. C., Schwedler, B. R., Arritt, R. W., & Takle, E. S. (2008). Changes in extreme, cold-season synoptic precipitation events under global warming. Geophysical Research Letters, 35 (20).

Hosking, J. R. M., & Wallis, J. R. (1997). Regional frequency analysis: An approach based on l-moments. Cambridge University Press. doi: 10.1017/CBO9780511529443.

Hu, Y., Deng, Y., Zhou, Z., Li, H., Cui, C., & Dong, X. (2019). A synoptic assessment of the summer extreme rainfall over the middle reaches of Yangtze River in CMIP5 models. Climate Dynamics, 53 , 2133-2146.

Jagannathan, K., Jones, A. D., & Ray, I. (2020). The making of a metric: Coproducing decision-relevant climate science. Bulletin of the American Meteorological Society, 1 - 33. doi: 10.1175/BAMS-D-19-0296.1

Kjellstrom, E., Boberg, F., Castro, M., Christensen, J. H., Nikulin, G., & Sanchez, E. (2010). Daily and monthly temperature and precipitation statistics as performance indicators for regional climate models. Climate Research, 44 (2-3), 135-150.

Lackmann, G., & Gyakum, J. (1999). Heavy cold-season precipitation in the northwestern United States: Synoptic climatology and an analysis of the flood of 17-18 january 1986. Weather and Forecasting, 14 , 687-700.

Lee, Y.-Y., & Grotjahn, R. (2016). California Central Valley summer heat waves form two ways. Journal of Climate, 29 , 1201-1217. doi: 10.1175/JCLI-D-15-0270.1

Letson, F. W., Barthelmie, R. J., Hodges, K. I., & Pryor, S. C. (2021). Intense windstorms in the northeastern united states. Natural Hazards and Earth System Sciences, 21 (7), 2001-2020.

Livneh, B., Bohn, T. J., Pierce, D. W., Munoz-Arriola, F., Nijssen, B., Vose, R., . . . Brekke, L. (2015, Aug 18). A spatially comprehensive, hydrometeorological data set for mexico, the U.S., and southern canada 1950-2013. Scientific Data, 2 (1), 150042. doi: 10.1038/sdata.2015.42

Lucas-Picher, P., Argüeso, D., Brisson, E., Tramblay, Y., Berg, P., Lemonsu, A., ... & Caillaud, C. (2021). Convection-permitting modeling with regional climate models: Latest developments and next steps. Wiley Interdisciplinary Reviews: Climate Change, 12(6), e731.

Mach, K. J., Lemos, M. C., Meadow, A. M., Wyborn, C., Klenk, N., Arnott, J. C., ... & Wong-Parodi, G. (2020). Actionable knowledge and the art of engagement. Current Opinion in Environmental Sustainability, 42, 30-37.

Marquardt Collow, A., Bosilovich, M., & Koster, R. (2016). Large-scale influences on summertime extreme precipitation in the northeastern United States. Journal of Hydrometeorology, 17 , 3045-3061. doi: https://doi.org/10.1175/JHM-D-16-0091.1

McGinnis, S. (2019, September). sethmcg/climod. Zenodo. Retrieved from https://doi.org/10.5281/zenodo.3461259 doi: 10.5281/zenodo.3461259

Mearns, e. a., L.O. (2017). The NA-CORDEX dataset. NCAR Climate Data Gateway. doi: 10.5065/D6SJ1JCH

Mondal, S., Mishra, A. K., & Leung, L. R. (2020). Spatiotemporal characteristics and propagation of summer extreme precipitation events over united states: A complex network analysis. Geophysical Research Letters, 47 (15), e2020GL088185.

Moss, R. H., Avery, S., Baja, K., Burkett, M., Chischilly, A. M., Dell, J., ... & Zimmerman, R. (2019). Evaluating knowledge to support climate action: A framework for sustained assessment. Report of an independent advisory committee on applied climate assessment. *Weather, Climate, and Society*, *11*(3), 465-487.

Palipane, E., & Grotjahn, R. (2018). Future projections of the large-scale meteorology associated with California heat waves in CMIP5 models. Journal of Geophysical Research: Atmospheres, 123 , 8500-8517. doi: 10.1029/2018JD029000.

Pendergrass, A. G., Gleckler, P. J., Leung, L. R., & Jakob, C. (2020). Benchmarking simulated precipitation in earth system models. *Bulletin of the American Meteorological Society*, *101*(6), E814-E816.

Pryor, S. C., & Schoof, J. T. (2019). A hierarchical analysis of the impact of methodological decisions on statistical downscaling of daily precipitation and air temperatures. International Journal of Climatology, 39 (6), 2880-2900.

Pryor, S. C., & Schoof, J. T. (2020). Differential credibility assessment for statistical downscaling. Journal of Applied Meteorology and Climatology, 59 (8), 1333- 1349.

Sarkki, S., Heikkinen, H. I., Komu, T., Partanen, M., Vanhanen, K., & Lépy, É. (2020). How boundary objects help to perform roles of science arbiter, honest broker, and issue advocate. Science and Public Policy, 47(2), 161-171.

Schoof, J. T., Pryor, S. C., & Ford, T. W. (2019). Projected changes in united states regional extreme heat days derived from bivariate quantile mapping of cmip5 simulations. Journal of Geophysical Research: Atmospheres, 124 (10), 5214-5232.

Smalley, K. M., Glisan, J. M., & Gutowski Jr, W. J. (2019). Alaska daily extreme precipitation processes in a subset of cmip5 global climate models. Journal of Geophysical Research: Atmospheres, 124 (8), 4584-4600.

Shackley, S., & Wynne, B. (1996). Representing uncertainty in global climate change science and policy: Boundary-ordering devices and authority. Science, Technology, & Human Values, 21(3), 275-302.

Song, F., Feng, Z., Leung, L. R., Houze, Jr., R. A., Wang, J., Hardin, J., & Homeyer, C. (2019. Contrasting the spring and summer large-scale environments associated with mesoscale convective systems over the U.S. Great Plains. Journal of Climate, 32, 6749-6767, doi: 10.1175/JCLI-D-18-0839.1.

Srivastava, A., Grotjahn, R., & Ullrich, P. (2020). Evaluation of historical cmip6 model simulations of extreme precipitation over contiguous us regions. Weather and Climate Extremes, 29 , 100268. doi: https://doi.org/10.1016/j.wace.2020.100268

Srivastava, A., Grotjahn, R., Ullrich, P. A., & Risser, M. (2019). A unified approach to evaluating precipitation frequency estimates with uncertainty quantification: Application to florida and california watersheds. Journal of Hydrology, 578 , 124095. doi: 10.1016/j.jhydrol.2019.124095.

Srivastava, A. K., Grotjahn, R., Ullrich, P. A., & Sadegh, M. (2021). Pooling data improves multimodel idf estimates over median-based idf estimates: Analysis over the susquehanna and orida. Journal of Hydrometeorology, 22 (4), 971-995. doi: 10.1175/JHM-D-20-0180.1.

Stansfield, A. M., Reed, K. A., Zarzycki, C. M., Ullrich, P. A., & Chavas, D. R. (2020). Assessing tropical cyclones' contribution to precipitation over the eastern united states and sensitivity to the variable-resolution domain extent. Journal of Hydrometeorology, 21 (7), 1425-1445.

Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. Journal of Geophysical Research: Atmospheres, 106 (D7), 7183-7192.

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of cmip5 and the experiment design. Bulletin of the American Meteorological Society, 93 (4), 485-498.

Turnhout, E. (2009). The effectiveness of boundary objects: the case of ecological indicators. Science and public policy, 36(5), 403-412.

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

Ullrich, P. A., Zarzycki, C. M., McClenny, E. E., Pinheiro, M. C., Stansfield, A. M., & Reed, K. A. (2021). Tempestextremes v2.1: a community framework for feature detection, tracking, and analysis in large datasets. Geoscientific Model Development, 14 (8), 5023-5048. doi: 10.5194/gmd-14-5023-2021.

Wehner, M. F. (2013). Very extreme seasonal precipitation in the NARCCAP ensemble: Model performance and projections. Climate Dynamics, 40 (1-2), 59-80.

Wyborn, C., Datta, A., Montana, J., Ryan, M., Leith, P., Chaffin, B., ... & Van Kerkhoff, L. (2019). Co-producing sustainability: reordering the governance of science, policy, and practice. Annual Review of Environment and Resources, 44, 319-346.

Xue, Z., & Ullrich, P. A. (2021). A comprehensive intermediate-term drought evaluation system and evaluation of climate data products over the conterminous United States. Journal of Hydrometeorology, 22 (9), 2311-2337. doi: 10.1175/JHM-D-20-0314.1

Zarzycki, C. M., (2018). Projecting changes in societally impactful northeastern U.S. snowstorms. Geophysical Research Letters, 450 (21), 12,067-12,075. doi: 10.1029/2018GL079820

Zarzycki, C. M., Ullrich, P. A., & Reed, K. A. (2021). Metrics for evaluating tropical cyclones in climate data. Journal of Applied Meteorology and Climatology, 60 (5), 643-660.