BAMS Article

The Model Evaluation Tools (MET)

More than a Decade of Community-Supported Forecast Verification

Barbara Brown, Tara Jensen, John Halley Gotway, Randy Bullock, Eric Gilleland, Tressa Fowler, Kathryn Newman, Dan Adriaansen, Lindsay Blank, Tatiana Burek, Michelle Harrold, Tracy Hertneky, Christina Kalb, Paul Kucera, Louisa Nance, John Opatz, Jonathan Vigh, and Jamie Wolff

ABSTRACT: Forecast verification and evaluation is a critical aspect of forecast development and improvement, day-to-day forecasting, and the interpretation and application of forecasts. In recent decades, the verification field has rapidly matured, and many new approaches have been developed. However, until recently, a stable set of modern tools to undertake this important component of forecasting has not been available. The Model Evaluation Tools (MET) was conceived and implemented to fill this gap. MET (https://dtcenter.org/community-code/model-evaluation-tools-met) was developed by the National Center for Atmospheric Research (NCAR), the National Oceanic and Atmospheric Administration (NOAA), and the U.S. Air Force (USAF) and is supported via the Developmental Testbed Center (DTC) and collaborations with operational and research organizations. MET incorporates traditional verification methods, as well as modern verification capabilities developed over the last two decades. MET stands apart from other verification packages due to its inclusion of innovative spatial methods, statistical inference tools, and a wide range of approaches to address the needs of individual users, coupled with strong community engagement and support. In addition, MET is freely available, which ensures that consistent modern verification capabilities can be applied by researchers and operational forecasting practitioners, enabling the use of consistent and scientifically meaningful methods by all users. This article describes MET and the expansion of MET to an umbrella package (METplus) that includes a database and display system and Python wrappers to facilitate the wide use of MET. Examples of MET applications illustrate some of the many ways that the package can be used to evaluate forecasts in a meaningful way.

KEYWORDS: Model evaluation/performance; Statistics; Software

https://doi.org/10.1175/BAMS-D-19-0093.1

Corresponding author: Barbara Brown, bgb@ucar.edu

In final form 3 October 2020

AFFILIATIONS: Brown, Jensen, Halley Gotway, Bullock, Gilleland, Fowler,* Newman, Adriaansen, Blank, Burek, Harrold, Hertneky, Kalb, Kucera, Nance, Opatz, Vigh, and Wolff—Research Applications Laboratory, National Center for Atmospheric Research, Boulder, Colorado
* Retired

In the last several decades, forecast evaluation has become a central cog in the process of developing, improving, and applying complex numerical weather and climate modeling and prediction systems in both research and operational contexts, as well as in the process of monitoring and improving operational predictions. Until recently, modern tools were not available to meet the needs for evaluating high-resolution forecasts (e.g., Mass et al. 2002; Davis et al. 2006a), which often led to misleading (or noninformative) results from traditional verification analyses. Hence, the requirements for forecast verification tools and advanced evaluation methods increased dramatically as models and forecasts moved to higher resolution and the desire for more informative verification analyses grew (Casati et al. 2008). As a result, verification became an important area of research and development (e.g., Casati et al. 2008; Ebert et al. 2013).

However, a consolidated set of modern tools was not available for widespread use in both operational and research settings. Thus, in 2007, in response to this need, the U.S. Developmental Testbed Center (DTC; https://dtcenter.org/), with support from the U.S. Air Force (USAF), initiated an effort to create a state-of-the-art verification software package. The expectation was that the package would be available and supported for a wide set of users and would specifically meet the needs for evaluation of mesoscale weather prediction models, which was the primary focus of the DTC at the time.

The targeted users included research scientists [e.g., from the National Center for Atmospheric Research (NCAR), universities, other research laboratories] and staff at operational centers [e.g., the centers that are part of the National Centers for Environmental Prediction (NCEP)] focused on development or application of numerical weather prediction (NWP) models. The outcome of this initial work was the Model Evaluation Tools (MET) software package, which has since become a widely used community verification package in the short- and medium-range weather and climate prediction communities. MET also is used by a diverse set of Earth system modeling communities (e.g., space weather, polar, atmospheric composition prediction), in both research and operational environments.

This paper discusses MET's evolution and current capabilities, as well as the extensive community support for MET applications provided by the DTC. Connections with the verification research community, along with the operational verification, forecasting, and model development communities are also described. The new umbrella "METplus" framework is presented and examples of its capabilities are provided. A list of abbreviations is provided in the appendix.

History and user community engagement

Forecast verification/evaluation has been a subject of research and also applied to operational forecasts for more than a century (e.g., Finley 1884; Gilbert 1884; Brier 1950; Murphy and Winkler 1987; Murphy et al. 1989). Allan Murphy and others invested significant efforts in verification research and applications during the latter part of the twentieth century (e.g., Murphy and Daan 1985; Murphy 1986; Murphy and Winkler 1987; Brown and Murphy 1987; Ehrendorfer and Murphy 1988; Murphy et al. 1989; Stanski et al. 1989; Doswell et al. 1990; Murphy 1991; Murphy and Winkler 1992; Nurmi 1994; Brooks and Doswell 1996; Briggs and Levine 1997; Marzban 1998; Murphy and Wilks 1998; Hamill 1999; Wilson et al. 1999). However, verification emerged as a significant research topic during the last two decades (e.g.,

Ebert and McBride 2000; Stephenson 2000; Atger 2001; Smith and Hansen 2005; Baldwin and Kain 2006; Mason 2008; Roberts and Lean 2008; Ahijevych et al. 2009; Jolliffe and Stephenson 2012; Mittermaier et al. 2016; Wilks 2018).

In fact, this recent period can be viewed as a renaissance in the development and understanding of forecast evaluation methods, which arose in part due to the emergence of high-resolution NWP models and nowcasts (Mass et al. 2002), the increasing prevalence of ensemble predictions, and community interest in more meaningful and statistically valid approaches (Casati et al. 2008; Ebert et al. 2013; Dorninger et al. 2018a). Within this context, the dearth of applications of modern methodologies was recognized early in the twenty-first century when several organizations, including the National Oceanic and Atmospheric Administration (NOAA), NCAR, the USAF, and the World Meteorological Organization (WMO) encouraged efforts toward improving verification methods and expanding the testing of NWP models. For example, the WMO established the Joint Working Group on Forecast Verification Research (JWGFVR; www.wmo.int/pages/prog/arep/wwrp/new/Forecast_Verification.html) in 2002; and NOAA, NCAR, and the USAF established the DTC in 2003 (Bernardet et al. 2008) with a focus on improving NWP in the United States, including model testing and evaluation.

In 2006, the DTC was tasked by the USAF with building a community verification package. The vision for this package was "A world class, state of the art verification system for evaluating high-resolution forecast systems.... In addition, the DTC verification system will become a central feature of services the DTC provides to all WRF users. The package will be made available to all WRF users." As stated, the original intention was to meet the needs of Weather Research and Forecasting (WRF; Powers et al. 2017) Model users and developers, including university researchers, operational forecasters, NOAA, USAF, and the private sector, which was the purview of the DTC at the time. Since then, the vision of both community modeling and MET has expanded to include other types of modeling systems—including global, tropical cyclone (TC), space weather, climate, and hydrologic models.

MET's initial development was guided by many stakeholders, including the NWP community and verification method experts. Much of the input from these groups was obtained through yearly workshops during MET's early years (2007–10). Participants in these workshops included NCAR scientists and engineers, international verification experts, MET users, and NWP/verification experts from government agencies, universities, and research organizations. In addition, funding agencies (e.g., USAF, NOAA, NCAR) prioritized their specific needs, with many additional capabilities (e.g., tools for evaluation of TC forecasts) included in MET motivated by the needs of these agencies.

During MET's early years, and the period preceding its initiation, many new verification methods were developed and tested, and became accepted, including a variety of spatial verification techniques (e.g., Brown et al. 2012; Casati et al. 2004; Ebert and McBride 2000; Davis et al. 2006a,b; Gilleland et al. 2009; Roberts and Lean 2008), application and development of methods to measure the uncertainty associated with verification results (e.g., Jolliffe 2007; Gilleland 2010; Gilleland et al. 2018), and development and application of new graphical approaches to display verification results (Brown and Murphy 1987; Roebber 2009; Taylor 2001). In addition, the JWGFVR promoted the development and application of statistically valid, meaningful, and useful verification methodologies.

MET has expanded and changed as a result of the evolving state-of-the-art of forecast evaluation and in response to users' needs (e.g., for tools to evaluate TC forecasts). In the early years, this evolution focused on expansion of the number and types of verification tools in MET (e.g., inclusion of additional traditional and spatial tools). More recently, the enhancements have also focused on infrastructure and graphical tools.

MET itself includes very limited graphical capabilities (e.g., to display input datasets). However, early in MET's development, the MET team recognized that the availability of tools to display MET results would be extremely important to MET users, and would make MET results more meaningful and useful. This recognition led to development of an interactive database and display system (METviewer). METviewer provides access to databases created from MET output and was initially used by DTC staff as an essential tool for their many and varied testing and evaluation activities (see https://dtcenter.org/testing-evaluation). As part of this development, a database system (METdatadb) was created to store MET output and make it available to METviewer and other applications.

More recently, as new domestic and international partnerships evolved with NOAA, Department of Defense (DOD), and other organizations, the need for a broader and more flexible platform for MET development and application emerged and led to the creation of METplus (Fig. 1). METplus facilitates the application of the tools to a broader set of forecast and observation types (e.g., for air quality, space weather, climate), joint development of the packages, and implementation of more advanced structures to organize the MET tools.

As with other DTC efforts, *community engagement and support* has been—and continues to be—a fundamental aspect of MET development and application. This engagement included the verification workshops in 2007–10, training events and tutorials, and the 2018 DTC Community Unified Forecast System Test Plan and Metrics Workshop (www.dtcenter.org/events/workshop/2018/2018-dtc-community-unified-forecast-system-test-plan-metrics-workshop). In addition, several verification researchers and practitioners have participated in the DTC's Visitor Program (https://dtcenter.org/visitor-program), with at least four methods [*Wavelet-Stat*, Fractions Skill Score (FSS), High-Resolution Analysis (HiRA), and distance map metrics] added to MET through the years via this program. Hence, the operational and research communities have both been engaged with MET's evolution throughout its history.

Technical support for users' applications of MET has been ongoing since MET's inception. This support has included 15 tutorials on the application of MET (starting in 2008). These tutorials have also provided basic information on forecast verification methods, to ensure meaningful application of the tools. The DTC also supports a "help-desk" function for MET, to facilitate users' applications of the tools. The help-desk staff provide prompt responses via e-mail to users' problems/questions/issues; in recent years they responded by e-mail to more than 350 requests per year. The MET users' page (https://dtcenter.org/community-code/model-evaluation-tools-met/) provides more information about these and other resources, including MET documentation and links to other information sources. MET also can be downloaded from links on this page. Currently the MET community includes more than 3,700 researchers and operational users from 124 countries,

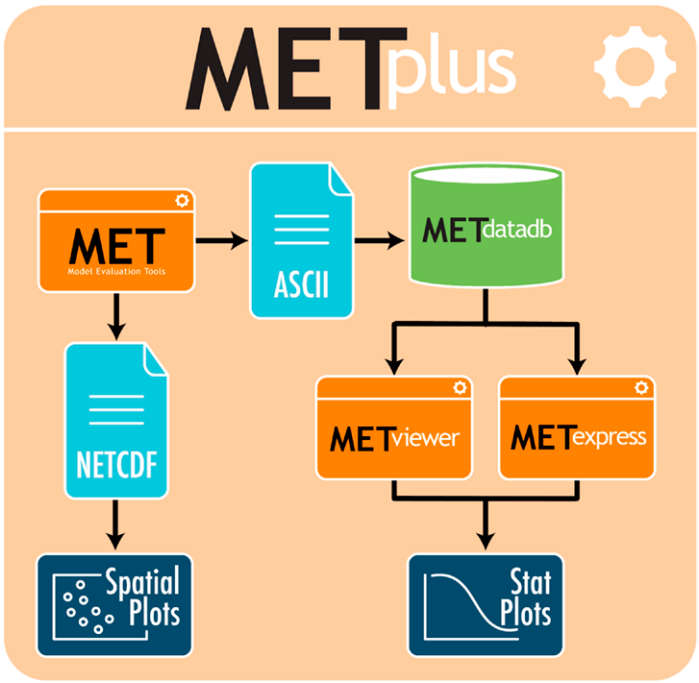

Fig. 1. Schematic diagram of the structure of METplus. METplus includes MET and the standard output of MET in ASCII and netCDF formats, and input of ASCII-formatted MET results into METdatadb (the MET database) and then to METviewer to produce custom statistical plots. METexpress is a streamlined dashboard version of METviewer. The netCDF output from *MODE* and other tools also is used to create spatial plots. Python wrappers (represented by the black arrows and orange background) tie all of the pieces together.

from a variety of disciplines and work sectors (universities, government, private companies, and nonprofit organizations).

MET support is a major effort for MET developers, software specialists, and verification experts. However, the benefits of directly interacting with MET users are extensive. For example, these interactions broaden the reach of the package, enable the collection of ideas for enhancements to the system, ensure that the tools are functioning as expected, and lead to improvements in the usability and technical aspects of the MET tools.

Although originally designed primarily for research applications, in recent years, MET and METplus have been adopted operationally by many governmental organizations, including NOAA research laboratories and several operational prediction centers [e.g., the Environmental Modeling Center (EMC) and Weather Prediction Center (WPC)] within the U.S. National Weather Service (NWS). MET has also been adopted for forecast verification activities by the USAF Operational Center and the Naval Research Laboratory (NRL). In addition, MET is being applied by numerous international prediction centers, including centers in the United Kingdom, Taiwan, South Africa, and China; and is widely applied by researchers at universities and research laboratories around the world. These agencies and organizations have adopted the MET package because it uniquely has a variety of desirable characteristics not available with other verification software, including the following:

- extensive catalog of traditional statistics;
- single source for many modern and alternative methods (e.g., spatial, ensemble);
- flexible options to meet a wide variety of verification problems;
- extensibility to new methods and fields using Python embedding;
- robust software development process including nightly builds, continuous integration, and cyber-security testing;
- open GitHub repository and community package that ensures all users have access to the same software and methods; and
- extensive community support.

Moreover, the MET system is not static and continues to grow to incorporate new capabilities as the verification community develops new tools. Furthermore, because MET has been extensively used and tested by the DTC throughout its history, it is a hardened and vetted system.

In summary, MET development was initiated during a period that was ripe for new tools to facilitate the development of new and improved modeling/forecasting systems. Community engagement and support have been at the forefront of this development. MET and METplus, and examples of their application, are described in greater detail in the remainder of this paper.

METplus: MET, METviewer, and more

METplus (Fig. 1; https://dtcenter.org/community-code/metplus) was established as a way to organize and connect various components, including MET (which could be thought of as the verification engine in METplus), METviewer (the MET visualization platform), METexpress (the simplified desktop version of METviewer for computing traditional verification statistics), and METdatadb (the database capability, which also connects the other three components). METplus also incorporates Python wrappers to aid in creation and management of specific workflows (e.g., running MET, aggregation and analysis, plotting and diagnostics) and Python embedding to allow users to easily integrate new datasets and methods into MET. The METplus tools represented in Fig. 1 are designed to run alone or interface with these wrappers.

As an open-source software package, all components of METplus are free for users to clone or download (at https://github.com/DTCenter/METplus). The METplus repository is equipped to

pull in the other framework components, including dependent packages. If a user desires to only download MET or METviewer, these may be obtained at https://github.com/DTCenter/MET, and https://github.com/DTCenter/METviewer. METexpress was released to the community (at https://github.com/DTCenter/METexpress) in the fall of 2020. Users are encouraged to reach out to DTC support (met-help and forums) with questions after first reviewing the online resources (user's guides, tutorials, and met-help archives).

A suite of configurations for the METplus wrappers and MET tools are bundled together as "use cases" or "examples" and included on the DTC's GitHub repository (at www.github.com/DTCenter/METplus). In addition, software containers (e.g., Docker and Singularity) are being used to facilitate implementation of the tools. These capabilities decrease the overhead required by users to apply MET, METviewer, and METplus, and—as a result of the enhanced partnerships associated with development of METplus—will lead to tools that are widely useful in the weather, climate, space-weather, hydrometeorology, and other communities. A few example use cases are described in a later section of this paper.

MET. The current configuration of the MET package (version 9.1) is illustrated in Fig. 2. MET consists of five functional layers: (i) input, (ii) reformatting, (iii) plotting, (iv) statistics, and (v) analysis. The first two components (input and reformatting) represent MET's data handling and

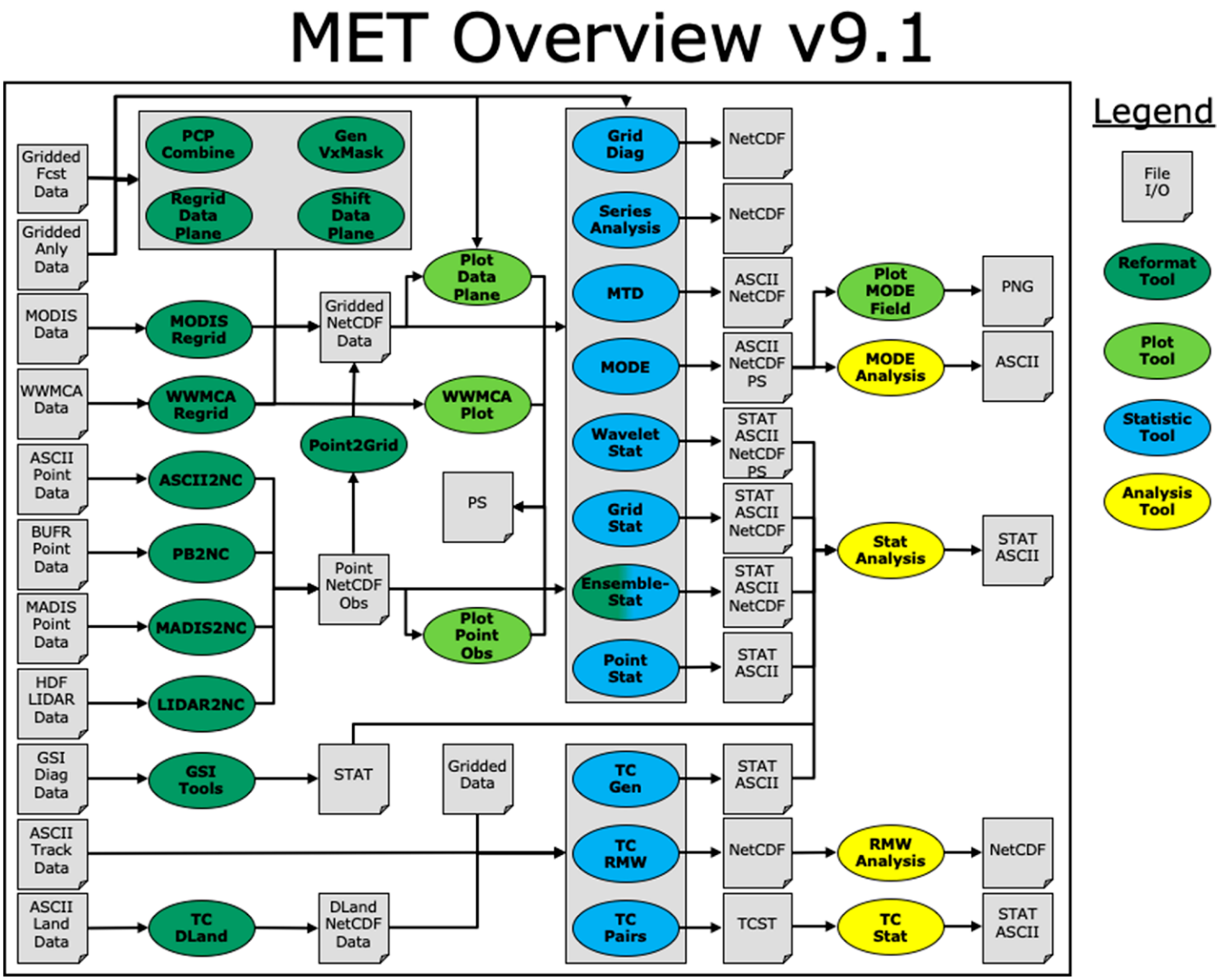

Fig. 2. Overview of the structure of the MET package (version 9.1).

preparation capabilities, while the remaining components are associated with specific verification activities (i.e., comparisons of forecasts and observations and computation of a wide variety of statistics). The MET Users Guide and portions of the METplus online tutorial describe all of these components in detail (https://dtcenter.org/community-code/model-evaluation-tools-met/documentation and https://dtcenter.org/community-code/https://dtcenter.org/community-code/metplus/online-tutorial); they are considered briefly in the following subsections.

Input, reformatting, and plotting components. MET's flexibility regarding forecast and observation formats and its tools for reformatting make it possible for MET to meet a wide range of user requirements. In particular, MET is designed to ingest a wide variety of forecast and observation types,[1] which makes it able to work with many different models (e.g., the many operational NWP models supported by prediction centers around the world). In general, forecasts are expected to be on a regular grid, with some exceptions (e.g., for TCs), and typically they are anticipated to be in a Gridded Binary (GRIB) version 1 or version 2 format or in Climate and Forecast (CF) Network Common Data Form (netCDF). However, MET's extensive reformatting functions make it possible to reformat most types of forecasts (e.g., deterministic, ensemble members, probabilistic, multicategory) into MET-usable formats. Recent incorporation of Python embedding (i.e., the ability to call a Python script from a MET tool) has expanded the number of supported forecast file formats and provides the ability to derive additional fields prior to evaluation.

At the outset of MET development, the observations used in MET were expected to be in Prepared Binary Universal Form for Representation of Meteorological Data (PrepBUFR) format,[2] because that format is commonly used by the NWS for many types of weather observations. Moreover, in initial implementations, MET primarily relied on point-based observations—that is, measurements from individual observing locations (e.g., surface stations, rawinsondes)—except for precipitation, which was anticipated to be gridded in at least some situations (e.g., based on radar mosaics). Since then, options for other point-observation formats [e.g., Meteorological Assimilation Data Ingest System (MADIS), Aerosol Robotic Network (AERONET), Surface Radiation (SURFRAD), and *Cloud–Aerosol Lidar and Infrared Pathfinder Satellite Observations* (*CALIPSO*)] have been incorporated. Additionally, Geostationary Satellite data from *GOES-16/17* (now East/West) are stored as a dense network of nongridded point observations [e.g., aerosol optical depth (AOD)] and support for these types of datasets also is included. Finally, as MET usage and capabilities expanded, support for gridded dataset types beyond radar/satellite quantitative precipitation estimate (QPE) mosaics were incorporated into MET's portfolio of observations.

[1] Note that MET does not in general provide forecasts or observation datasets; the datasets must be provided by the user.

[2] www.emc.ncep.noaa.gov/mmb/data_processing/prepbufr.doc/document.htm.

The observation datasets that are of interest for verification analyses often must be reformatted to be used by the MET tools; thus, MET includes an expanding variety of reformatting capabilities. For example, point observations are converted into a netCDF format using various tools, depending on the original observation format (e.g., *ASCII2NC* for ASCII-formatted observations). The *Point2Grid* tool makes use of nonparametric density estimation with a Gaussian kernel (e.g., Brooks et al. 1998; Hitchens et al. 2013) to create "practically perfect" gridded analyses from a set of point observations such as local storm reports. It also provides a data-thinning capability when placing dense datasets (e.g., GOES-East/West) on a user-defined grid.

Because it often is beneficial to examine datasets visually before applying verification analyses, MET includes several data inspection tools (e.g., *Plot-Data-Plane*, *WWMCA-plot*, *Plot-Point-Obs*) that provide visualizations of the datasets to be used by the statistical tools. Such plots can provide a sanity check, to ensure that the data are being correctly read by MET before an extensive analysis is undertaken. The output of these tools consists of images in postscript format.

MET also provides flexible options for a user to specify a subset, or area of interest, of a model grid to be used for the verification analysis (using the *Gen-Vx-Mask* tool), as well as other tools for data specification. The idea of a mask in this context is to identify, spatially, the area of interest for analysis and mask out the points that are not of interest.

Statistical tools. *Point-Stat, Grid-Stat and Ensemble-Stat.* *Grid-Stat* and *Point-Stat* are MET's main statistical "workhorses." These tools provide flexible capabilities to evaluate forecasts that are defined (i) on a continuous scale (e.g., temperature), (ii) in a categorical format (e.g., precipitation occurrence), or (iii) as a probabilistic forecast. MET also makes it possible for users to create categorical forecasts and observations from continuous forecasts/observations (e.g., by applying one or more thresholds to the forecasts and observations) and to perform conditional verification (i.e., evaluate the performance of subsets of the forecasts).

Grid-Stat assumes that the observations and forecasts being examined are on matching grids, whereas *Point-Stat* assumes that observations are located at a discrete set of point locations, not necessarily collocated with the forecast grid. Thus, *Grid-Stat* assumes that forecast and observation points can be directly matched and compared, whereas application of *Point-Stat* requires some form of spatial interpolation or matching protocol to create forecast-observation pairs associated with the observing locations. MET provides a wide variety of grid-to-point matching options, ranging from nearest neighbor to least squares and mass-conservation approaches. Three of MET's 15 options for interpolating from a model grid to other points in the region are illustrated in Fig. 3.

More than 85 traditional measures, including the measures recommended by the WMO (WMO 2010), are computed by *Point-Stat* and *Grid-Stat* (Fig. 4 shows a small subset). Hence, users have the opportunity to select the measures that are most relevant to answer the verification questions of interest (e.g., How accurate are the forecasts? How big is the bias?). These statistics are described in Appendix C of the MET User's Guide (https://dtcenter.org/community-code/model-evaluation-tools-met/documentation). *Point-Stat* and *Grid-Stat* compute almost identical sets of statistics, including categorical measures [e.g., equitable threat score (ETS), probability of detection (POD)] for 2 × 2 or multicategory contingency tables; measures designed for continuous forecast variables [e.g., root-mean-square error (RMSE), mean absolute error (MAE)]; and specific statistics for probabilistic forecasts (e.g., Brier score, reliability), as defined in texts on forecast evaluation (e.g., Jolliffe and Stephenson 2012; Wilks 2019). Small differences between the statistics computed by *Point-Stat* and *Grid-Stat* are associated with a few of the spatial verification methods (described in the "Spatial methods" section); for example, the

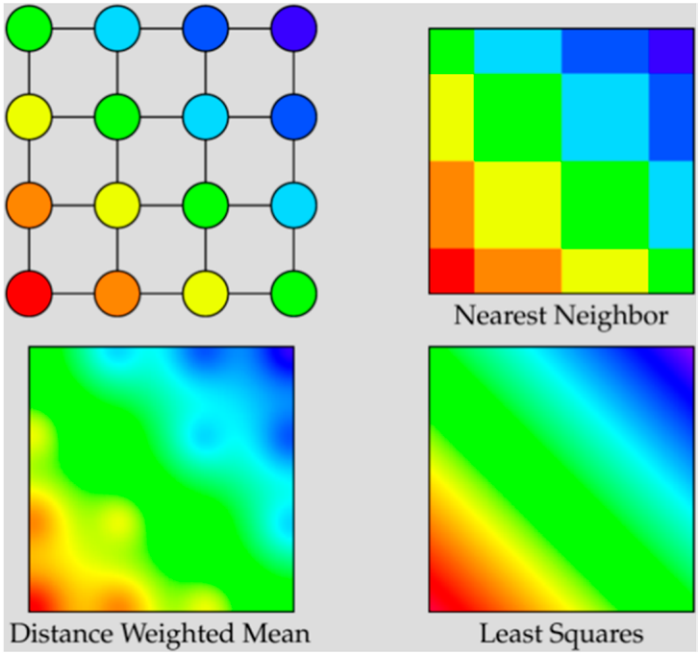

Fig. 3. Some of the methods included in MET to interpolate model grid values to a point observation location not coincident with a grid point are illustrated by the colors in these diagrams. The illustration in the upper-left corner represents forecast values on a grid. The results of applying a nearest neighbor, distance weighted mean, and least squares approach to interpolate the model grid values to sub-grid locations to pair with point observations, are illustrated by the colors shown for each interpolation method. (Figure generated by R. Bullock using his software.)

HiRA approach (a spatial approach described later) applies only to point observations (and thus is included in *Point-Stat*) and the Fractions Skill Score (FSS) (also a spatial approach) pertains to gridded observations (and is included in *Grid-Stat*).

Categorical and probabilistic forecasts can be evaluated directly by these tools. Alternatively, the tools can convert continuous forecasts into categorical forecasts (multicategory or binary) by applying thresholds defined in the user's implementation of *Grid-Stat* and *Point-Stat*. Similarly, continuous observation values can be converted to categorical values. In addition, special methods are provided for evaluation of wind vectors and gradients.

Point-Stat and *Grid-Stat* essentially produce tables of statistics for the forecasts being evaluated, without providing summary information (e.g., averaging across cases). This level of granularity allows users to undertake a variety of specific analyses that are meaningful for their particular application. For example, users may wish to aggregate information across forecasts, locations, times, or subset results based on other factors. The *Stat-Analysis* tool (described later) provides this capability.

Traditional	
Grid-Stat, Point-Stat, Series-Analysis *Contingency table statistics (CTS)* *Continuous statistics* *Probability forecast statistics* *Confidence intervals*	**Ensemble-Stat** *CRPS, CRPSS* *Rank prob., Prob. Integral Transform (PIT), and Relative Position histograms* *Spread/Skill* *Ignorance* *Confidence intervals*
Spatial	
MODE *Location differences* *Geometric attribute differences* *Intersection area* *Intensity distributions & differences* *CTS measures*	**MODE-TD** *Time and location differences* *Volume differences* *Velocity differences* *Intersection volume* *Intensity distributions & differences*
Wavelet-Stat *MSE by scale* *Energy by scale* *Intensity-scale skill score*	**Grid-Stat and Point-Stat** *FSS, HiRA* *Distance Measures: MED, Baddeley, Hausdorff, Zhu, etc.*
Tropical Cyclones and Diagnostics	
MET-TC *Track error (along, cross, total)* *Intensity errors (pressure, wind)* *Rapid intensification/weakening errors* **TC-GEN** *CTS measures of TC genesis*	**Grid-Diag** *Distributions of fields for use in contour plots* **TC-RMW** *Radius of maximum wind errors and metrics*

Fig. 4. MET statistical tools and statistics/metrics and diagnostics. This list is representative but not comprehensive.

Ensemble-Stat consists of a variety of statistical tools designed specifically to examine characteristics of ensemble forecasts (e.g., spread) and to evaluate their performance (e.g., rank histograms). It also provides the ability to preprocess ensemble predictions from a set of forecast files to produce derived fields (e.g., mean, probabilities). Statistics produced by *Ensemble-Stat* include standard ensemble verification measures, such as the rank histogram, continuous ranked probability score (CRPS), and spread/skill comparisons. A method to take into account observation error is also provided (e.g., Candille and Talagrand 2008).

Spatial methods. MET includes several modern tools that treat forecasts spatially, representing four categories of approaches (neighborhood, feature-based, scale separation, and distance metrics; Dorninger et al. 2018b); a fifth category (field deformation; Gilleland et al. 2010) is not yet included in MET. These spatial methods were developed over the last two decades in response to common difficulties inherent in the ability of more traditional approaches (i.e., most of the statistics included in *Grid-Stat* and *Point-Stat*) to provide meaningful information about forecast performance in many situations. While spatial methods are often more difficult to apply than traditional approaches, they can provide useful diagnostic information about forecast performance that may not be attainable from traditional approaches. Except for one approach (HiRA, described below), all of MET's spatial methods require gridded forecasts and observations; thus, these approaches are generally not appropriate for use with point observations.

MET's spatial methods include HiRA, FSS, *Wavelet-Stat*, *MODE*, *MTD*, and distance mapping methods. As noted earlier, HiRA is incorporated into *Point-Stat*, and FSS and distance maps methods are applied by *Grid-Stat*. *Wavelet-Stat*, *MODE*, and *MTD* methods are provided by MET tools named after them. HiRA and FSS are categorized as "neighborhood" or "filter" methods because they allow forecast points to be counted as correct if the forecast at a point or in an area matches an observation in its neighborhood (Ebert 2009). *MODE* and *MTD* are categorized as object-based methods, and *Wavelet-Stat* is categorized as a scale-separation approach (Brown et al. 2012; Gilleland et al. 2009). The distance-mapping methods form their own category and are incorporated into *Grid-Stat*.

More specifically,

- FSS (Roberts and Lean 2008) evaluates forecast–observation matches across a range of thresholds and horizontal scales. FSS applies thresholds to identify binary "events" and then compares the frequency of events in the forecast and observed fields as the size of the region (i.e., the scale) changes. Essentially, the forecast and observed fields are smoothed as more and more grid points are included in the "neighborhoods." Similarly, HiRA (Mittermaier 2014; Mittermaier and Csima 2017) interprets forecast values surrounding each point observation as an ensemble and provides a neighborhood-based assessment for point-based observations and has been applied to a variety of types of forecasts (e.g., Crocker et al. 2020). Figure 5 shows some of the intermediate steps involved in applying *Grid-Stat* to compute FSS. It should be noted that the data shown in Fig. 5 can be written out by *Grid-Stat* if it is configured to do so.
- Distance mapping methods (e.g., Gilleland 2011, 2017; Gilleland et al. 2020) focus on summarizing overall distances between forecast and observed event areas. The distance values depicted in Fig. 5d, when combined with the distance map for a corresponding observation

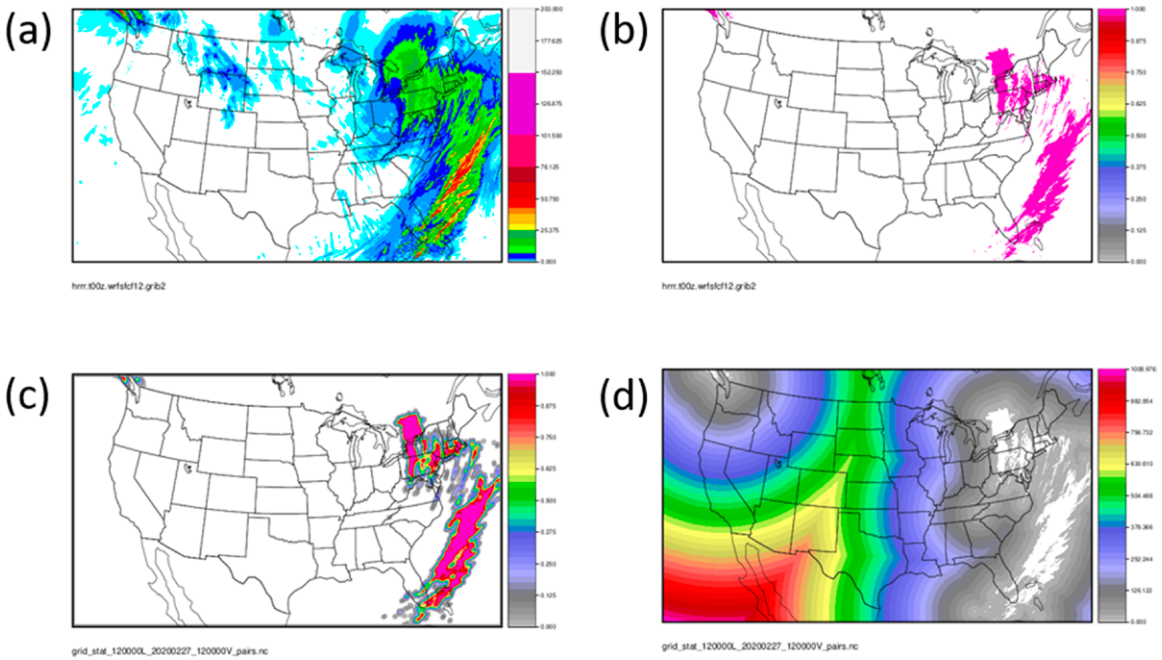

Fig. 5. (a)–(c) Example of some of the intermediate steps involved in applying FSS and distance maps to a sample of 3-km model-based precipitation predictions. In this case, precipitation values in (a) are thresholded using a threshold of 0.50 in. to produce the 0/1 values in (b). Step (c) shows the fractional coverage of values of 1 based on averaging the values in (b) across a neighborhood with a circular diameter of 21 grid points. The thresholded and smoothed forecast in (c) would be compared to a similarly treated observation field and the differences would contribute to computation of FSS as a function of scale and threshold. (d) A distance map showing the shortest distance from every grid point to the nearest 1-valued grid point in (b). (All graphs generated using MET's Plot-Data-Plane.)

field, would be used to compute many distance metrics that evaluate overall distances between forecast and observation fields using several different paradigms [e.g., mean error distance (MED), Hausdorff metric, Zhu's measure; Gilleland et al. 2020]. As with FSS, the intermediary data can be written out by *Grid-Stat* if configured to do so.

- MET's *Wavelet-Stat* tool (Casati et al. 2004) evaluates forecast performance as a function of the intensity values and the spatial scale of the error, and aims to determine the scales at which a forecast is "best" at reproducing the observed field, and to identify specific types of errors in a forecast (e.g., large displacements). A wide variety of statistics can be computed by *Wavelet-Stat*, consistent with the statistics produced by *Grid-Stat* and *Point-Stat*.
- *MODE* and *MTD* are designed to identify and compare relevant "objects" in gridded forecast and observation fields based on criteria selected by the user. In the case of *MODE* (Davis et al. 2006a,b; Davis et al. 2009; Bullock et al. 2016), the objects are static (i.e., objects in sequential fields are evaluated independently), whereas *MTD* evaluates forecasts in three dimensions, with time being the third dimension. *MODE* and *MTD* outputs include many geometric and intensity characteristics of forecast and observed objects such as object displacements and areas, intensity distributions within objects, and measures of the "quality" of the object matches. *MTD* also measures several attributes related to the evolution of objects across time, such as object volume, duration, and velocity.

It is important to note that FSS, HiRA, *Wavelet-Stat*, *MODE*, and *MTD* represent only a fraction of the many spatial verification methods that have been developed in the last two decades (Brown et al. 2012). The possibility of including additional spatial verification approaches in future versions of MET is considered in the summary.

MET-TC. *MET-TC* (consisting of *TC-Gen*, *TC-RMW*, *TC-Pairs*, and *TC-Stat* in Fig. 2) was first included in MET in 2013, initially to meet the needs of the U.S. Hurricane Forecast Improvement Project (HFIP; Gall et al. 2013) and more recently NOAA's National Hurricane Center (NHC). MET's TC tools are specifically designed to evaluate the unique characteristics of TC forecasts, which are very different from the typical gridded model output handled by the other MET tools. In particular, TC forecasts and observations are usually provided in a text format that lists the location and intensity, as well as other characteristics, of a particular storm at a specific time; in the United States, this format is called the Automated Tropical Cyclone Forecast (ATCF) file format. TC forecasts evaluated by *MET-TC* may be produced manually (e.g., by NHC forecasters) or by applying a vortex tracker algorithm to gridded model output.

MET-TC includes tools to (i) determine the location of coastlines and islands (e.g., to subset forecasts by the proximity of TCs to land), (ii) match and compare pairs of TC track forecasts and best track observations for corresponding storms, and (iii) provide summary statistics for forecast comparisons. The *MET-TC* summary tools produce a variety of statistics, including frequency of superior performance (e.g., to meet one of HFIP's goals to compare the performance of different TC modeling systems), time series independence calculations, and confidence intervals (CIs) on mean differences. In addition, *MET-TC* includes tools to evaluate rapid intensification/weakening (RI/RW) events, with flexible options for selecting thresholds to define their occurrence. The tool identifies RI/RW events in the forecast and observation datasets and derives contingency table statistics to evaluate them.

Two relatively new MET tools extend *MET-TC*'s capabilities: *TC-Gen* enables evaluation of TC genesis predictions, and *TC-RMW* provides diagnostic information to model developers and users regarding the radius of maximum winds (RMW) associated with a TC. Specifically, *TC-Gen* computes contingency table counts to evaluate genesis forecasts, and *TC-RMW* regrids TC model predictions onto a moving range–azimuth grid centered on points along the storm track, which facilitates estimation of RMW.

Analysis and diagnostic tools. Most of the output from the MET statistical tools is in the form of text files containing lists of numbers, with verification results for individual cases. MET's "analysis" tools are designed to filter, summarize, and analyze results produced by *Point-Stat*, *Grid-Stat*, *Ensemble-Stat*, *MODE*, *Wavelet-Stat*, and *MET-TC* based on a user's requests. These tools also make it possible to perform conditional verification and to aggregate results across a set of cases.

In particular, *Stat-Analysis* computes summary statistical information for results produced by *Point-Stat*, *Grid-Stat*, and some results from *MET-TC* (e.g., from *TC-Gen*). *Stat-Analysis* allows users to (i) aggregate results over a user-specified time period; (ii) stratify statistics based on time of day, model initialization time, lead-time, model run identifier, output filename, or wavelet decomposition scale; (iii) compute summary statistics and associated statistical CIs; and (iv) compute specific "NWP indices." These indices—including the generalized operations (GO) index used by the USAF and the NWP index used by the Met Office in the United Kingdom—are weighted averages of several verification measures (e.g., RMSE) across a variety of variables, forecast levels, and lead times. *Stat-Analysis* also reads the output of MET's *GSI-Tools* to compute statistics for the Gridpoint Statistical Interpolation (GSI) data assimilation system. Finally, *Stat-Analysis* computes several statistics for wind direction using two approaches to calculate the forecast error: (i) computing the error for each matched pair and averaging over the sample and (ii) computing the error of the aggregated forecast and observed vectors (tip to tail) across the sample.

Stat-Analysis also computes summary statistics, including mean, standard deviation, minimum, maximum, various percentiles, interquartile range, range, and weighted and unweighted means. Figure 6 presents an example map of *Stat-Analysis* results

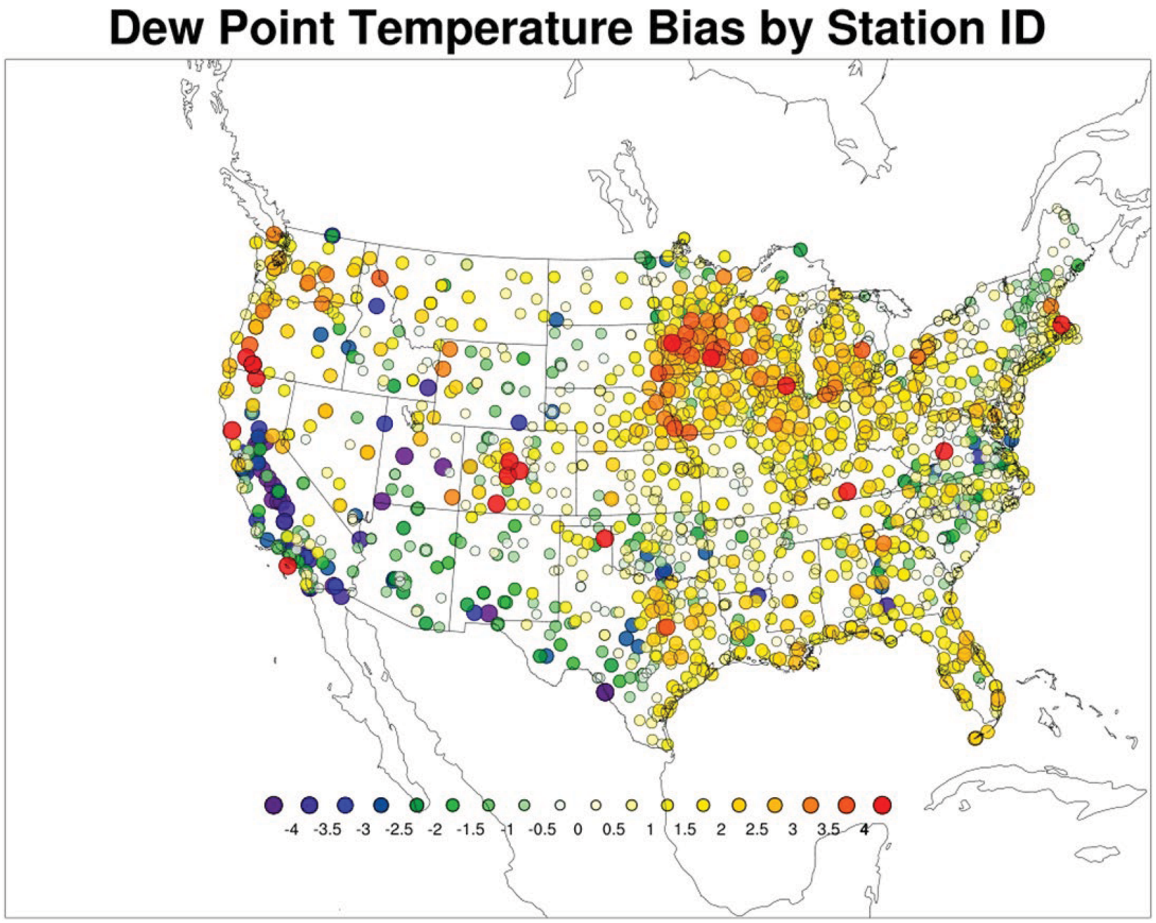

Fig. 6. Application of *Stat-Analysis* to represent geographical variations in forecast performance across a large set of observation locations, with results rendered graphically using a plotting script. [Figure generated using NCAR Command Language (NCL).]

showing geographical variations in a verification statistic (temperature bias in this case). CIs (parametric and nonparametric; Gilleland 2010) produced by *Stat-Analysis* on the various statistics facilitate model comparisons (see the first use case, on traditional verification applications). The ability for users to compute and output in the correct format statistics required for WMO reporting (WMO 2010) is also included. In addition, *Stat-Analysis* provides an approach for users to evaluate changes in forecast phenomena over time (e.g., ramps in wind speed). Future functionality will include application of user-provided climatological information to compute skill scores [e.g., Stable Equitable Error in Probability Space (SEEPS); Rodwell et al. 2010].

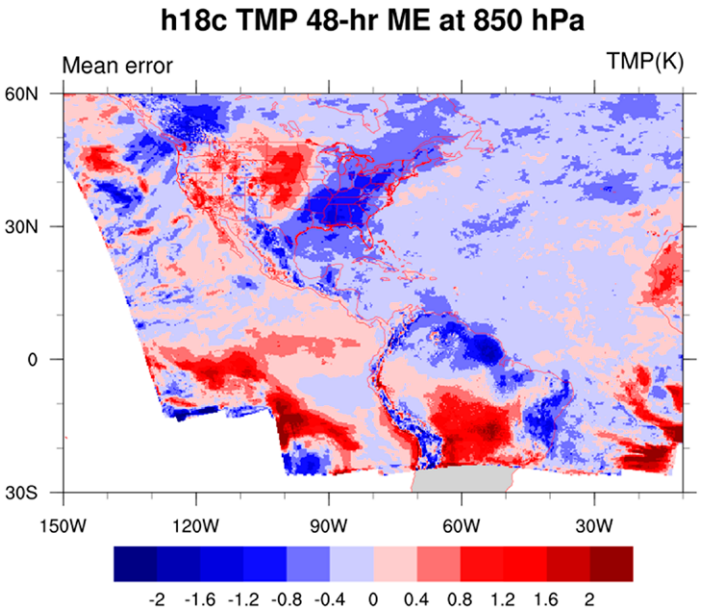

Fig. 7. Example application of *Series Analysis*, showing gridded mean error values for a set of 850-hPa model-based temperature predictions. (Figure generated using Matplotlib, a Python utility.)

Similarly, *MODE-Analysis* aggregates and summarizes results produced by *MODE*. Examples of *MODE-Analysis* capabilities include aggregation of results across cases and computation of summary statistics. In this case, the statistics are based on attributes—and comparisons of attributes—of objects identified and evaluated for individual cases or across sets of cases.

Series-Analysis provides a useful capability for analyzing and understanding the geographical representation of errors and forecast performance. In contrast to the bulk statistics computed by *Grid-Stat*, *Series-Analysis* computes user-selected summary statistics across time, or some other type of series, at individual points across a grid. The summary can include all of the categorical and continuous statistics produced by *Grid-Stat*. An example application of *Series-Analysis* to temperature forecasts is shown in Fig. 7.

Finally, *Grid-Diag* computes one- or two-dimensional probability density functions (PDFs) across forecast and observation grids. These PDFs can facilitate the development of climatologies for use in percentile thresholding or exploring the relationships between two model fields to contribute to process-oriented diagnostic studies.

METviewer. METviewer, the second major component of METplus, interacts with METdatadb, where MET results are stored, and provides extensive graphical and analysis capabilities (Fig. 1). Figure 8 shows an example of METviewer's interface, which enables application of many analysis and plotting options and provides extensive flexibility to examine MET verification results. The granularity of MET output allows METviewer users to explore verification results from many perspectives, with a wide variety of options for aggregation, stratification, display, and comparison. New METviewer capabilities include methods to create scorecards to summarize large quantities of verification information in a form that facilitates quick interpretation and comparisons of forecasting systems, the ability to create contour plots (aka heat maps or quilt plots), and numerous plotting templates. An example of a scorecard generated by METviewer (described in the use case discussion for traditional/operational verification) is shown below (see Fig. 12).

In addition to plotting MET output, METviewer includes expanded capabilities for analysis and evaluation, including visualization of distributions of measures across sets of cases (e.g., using boxplots). These distribution diagrams provide visual information about the

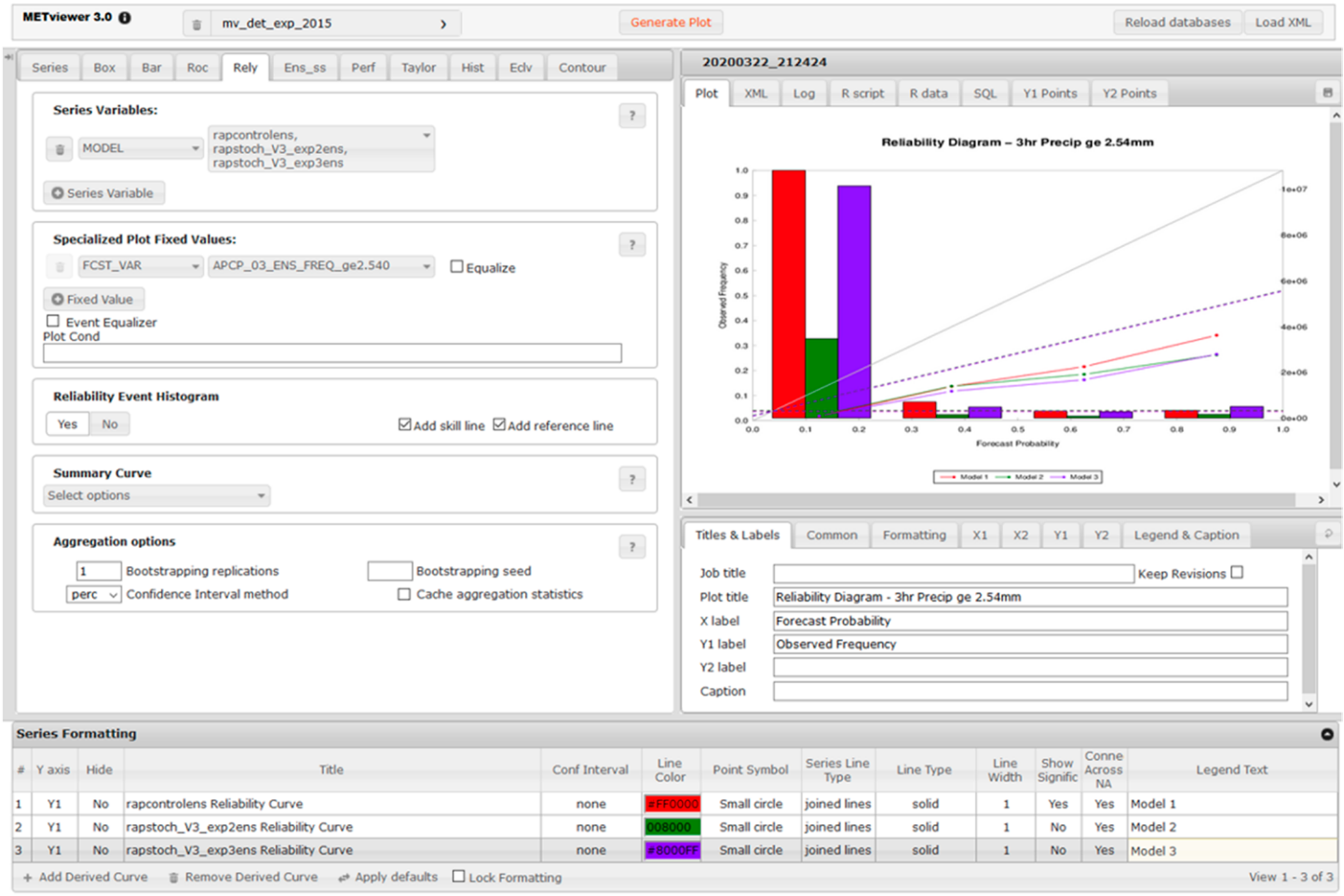

Fig. 8. Example of the METviewer user interface. METviewer allows users to interrogate MET results, with many choices regarding confidence intervals, aggregations, analyses, and plotting options (e.g., line plots, reliability diagrams, histograms, boxplots, performance diagrams). Scripts to create specific types of graphs can be saved for later reuse. (Plot generated from screen capture of METviewer interface.)

variability of the verification statistics for the samples being evaluated and the frequency of extreme (e.g., outlier) values. METviewer also includes specialized tools for summarizing verification measures, such as Taylor diagrams (Taylor 2001) and performance diagrams (Roebber 2009).

METviewer also computes and displays statistical CIs for MET's verification measures, which allow efficient, statistically valid, and fair comparisons of forecasting systems. The CIs make it possible to estimate the sampling uncertainty associated with the many different statistical measures produced by MET, and to compare the results from different samples. CIs are of particular importance when comparing the performance of one forecasting system to another (e.g., Hamill 1999; Jolliffe 2007).

METviewer and MET include options to apply both nonparametric methods based on a bootstrap procedure (Efron and Tibshirani 1993) and parametric methods based on the normal distribution to compute CIs. As described in Gilleland (2010), parametric approaches are only appropriate for a subset of metrics; MET and METviewer automatically select the appropriate approach. Specifically, METviewer enables fair comparisons of forecasting systems via (i) computation of CIs for each modeling system using the same sets of events ("event equalization"); (ii) computation of CIs on paired differences between the verification measures associated with two forecasting systems; and (iii) application of bootstrapping techniques to compute CIs when a Gaussian distribution cannot be assumed.

Example use cases: Sample applications of METplus

This section provides a few examples of applications of METplus, with a focus on the types of analyses that can be undertaken by MET and graphical results that can be obtained from METviewer. Note that for most of these cases, the forecast sources are not listed because the purpose of the discussion is to demonstrate the METplus capabilities rather than to evaluate particular forecasts.

Traditional/operational verification. The first use case focuses on the kind of verification analyses that might be applied in operational settings, relying primarily on traditional verification metrics (e.g., continuous, categorical statistics), and where forecasts from two models are being compared. Figure 9 shows the METplus workflow that might be applied in such an analysis, with an application of *Grid-Stat* to output from a model matched to gridded observations. MET output is stored by METdatadb and provided to METviewer for analysis and display of verification results. Examples of this output can include performance diagrams, boxplots, and scorecards, as well as a variety of other types of graphical information and summary statistics. For this example, the forecasts and observations are of categorical events (e.g., precipitation > 2.54 mm). Forecasts from two models (Model 1 and Model 2) are compared.

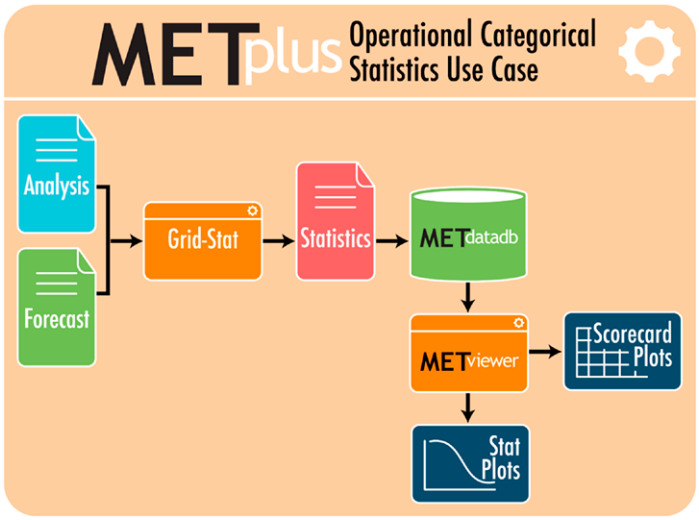

Fig. 9. Example METplus workflow for application to operational verification using traditional metrics.

Figure 10 provides an example of a performance diagram produced by METviewer for this case, with a precipitation threshold of 2.54 mm. In this diagram, the points associated with the best-performing model are located toward the upper-right corner. Points below the diagonal from (0,0) to (1,1) have a low bias, while those above the

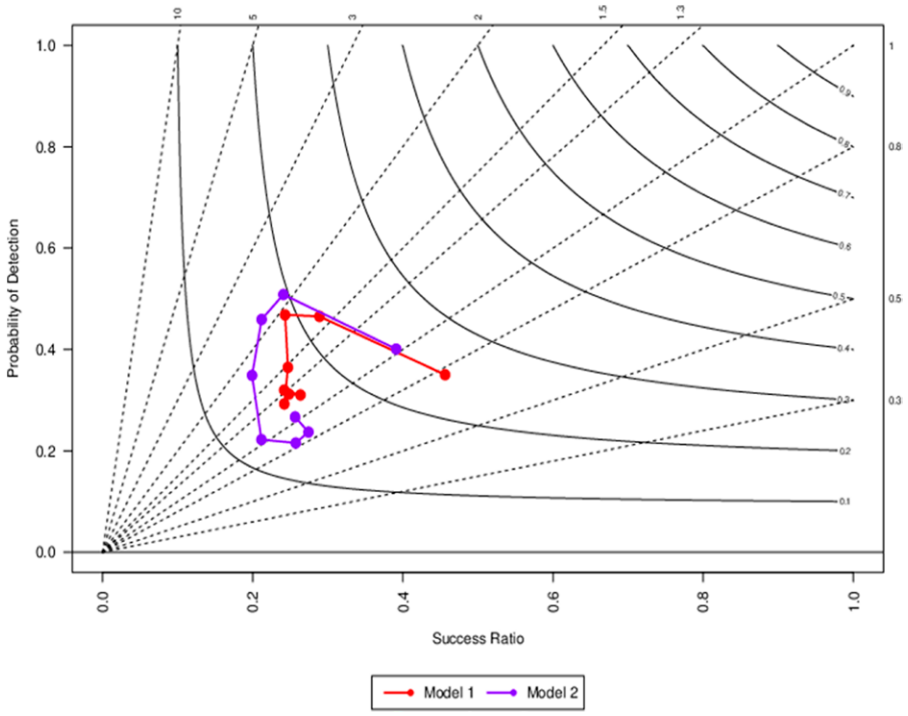

Fig. 10. Performance diagram (Roebber 2009) showing several verification statistics for model-based precipitation forecasts by two models for the event "3-h accumulated precipitation > 2.54 mm." Probability of detection (POD) is shown on the vertical axis, and success ratio [= 1 – false alarm ratio (FAR)] is on the horizontal axis; frequency bias (FBIAS) is represented by the slanted lines from the lower-left corner, to the right; and critical success index (CSI; also known as "threat score") is represented by the curved lines. The point for a "perfect forecast" would be located in the upper-right corner. Points for Model 1 (red) and Model 2 (purple) represent lead times for 3–24 h by 3-h increments. The 3-h lead time is the rightmost point, and 24-h lead times are near the FBIAS = 1 line. See text for details. (Figure generated using METviewer.)

diagonal have a high bias. This summary diagram indicates that both models have a critical success index (CSI) between 0.2 and 0.3 for the shortest lead time, but Model 1 has a low bias (less than 1) while the bias for Model 2 is approximately 1 (i.e., unbiased). For longer lead times, the performance of both models degrades, with the degradation somewhat slower for Model 1.

Figure 11 illustrates the benefits of applying CIs to paired differences in verification statistics. In this example, CIs are computed for Gilbert Skill Score (GSS) values associated with individual models, and for the differences in the statistics between the two models, for the same event (3-h precipitation > 2.54 mm) as considered in Fig. 10. The CIs for Models 1 and 2 overlap, which might suggest that the results for the two models are not significantly different. However, the CIs for the average pairwise differences in the scores (in green) do not intersect the zero line for some lead times (3, 9, 12, 15, and 18 h), which indicates that the differences are statistically significant. The apparent conflict between the conclusions of the two approaches illustrated in Fig. 11 is simply due to the greater statistical efficiency associated with estimating the uncertainty for the paired differences as opposed to examining and comparing the statistics for the two models individually (Wilks 2019).

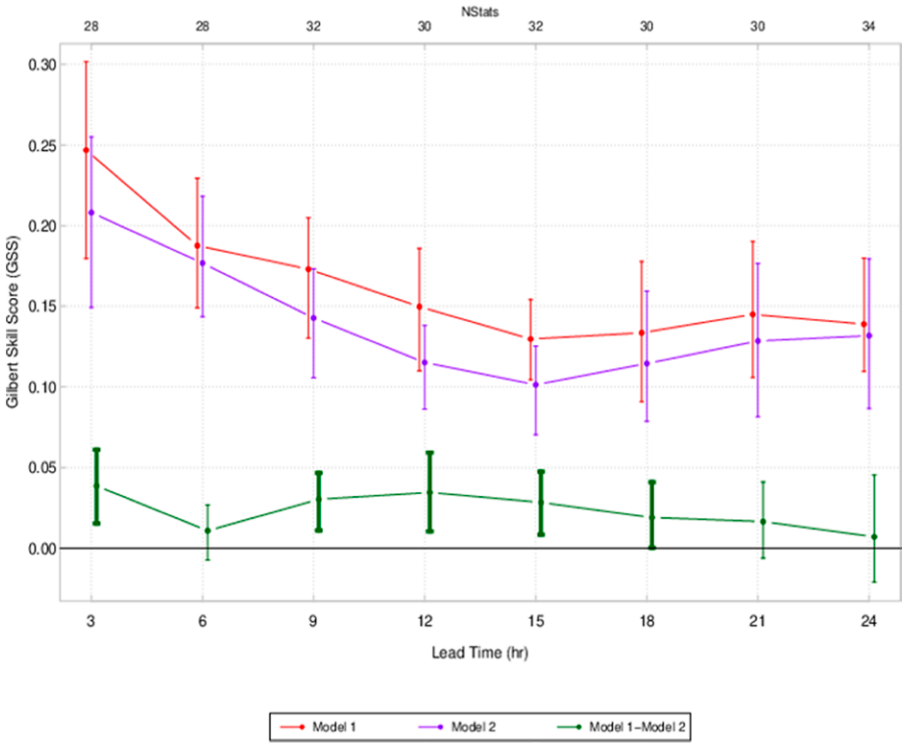

Fig. 11. Example of an evaluation of 3-h accumulated precipitation, showing Gilbert Skill Score (GSS) for Model 1 (red) and Model 2 (purple) for the event "3-h accumulated precipitation > 2.54 mm." The average pairwise differences between the scores (computed across the sample of forecasts) are shown in green. Bootstrapped confidence intervals (CIs) are shown as vertical lines for both the paired and unpaired cases. (Figure generated using METviewer.)

Finally, Fig. 12 shows an example of a scorecard produced by METviewer which summarizes a large amount of information in a small space, providing succinct comparisons of forecasts from different models. In this example, several thresholds are applied and the performance by Models 1 and 2 is compared for a variety of precipitation thresholds. The score card shows comparative performance as a function of variations in precipitation thresholds, lead times, accumulation periods, and region. A quick look at results in Fig. 12 suggest that Model 1 is frequently better than Model 2 when evaluated using CSI, but often has worse performance for FBIAS.

Ensemble forecasts. This use case represents the kinds of analyses that might be undertaken by an NWP model developer examining the performance of an ensemble prediction system. Several steps are required in applying METplus (Fig. 13), including conversion of PrepBUFR observations to netCDF and application of *Ensemble-Stat* to produce verification statistics. *Stat-Analysis* then summarizes this information to produce a variety of ensemble verification statistics, and plots can be created to present these results.

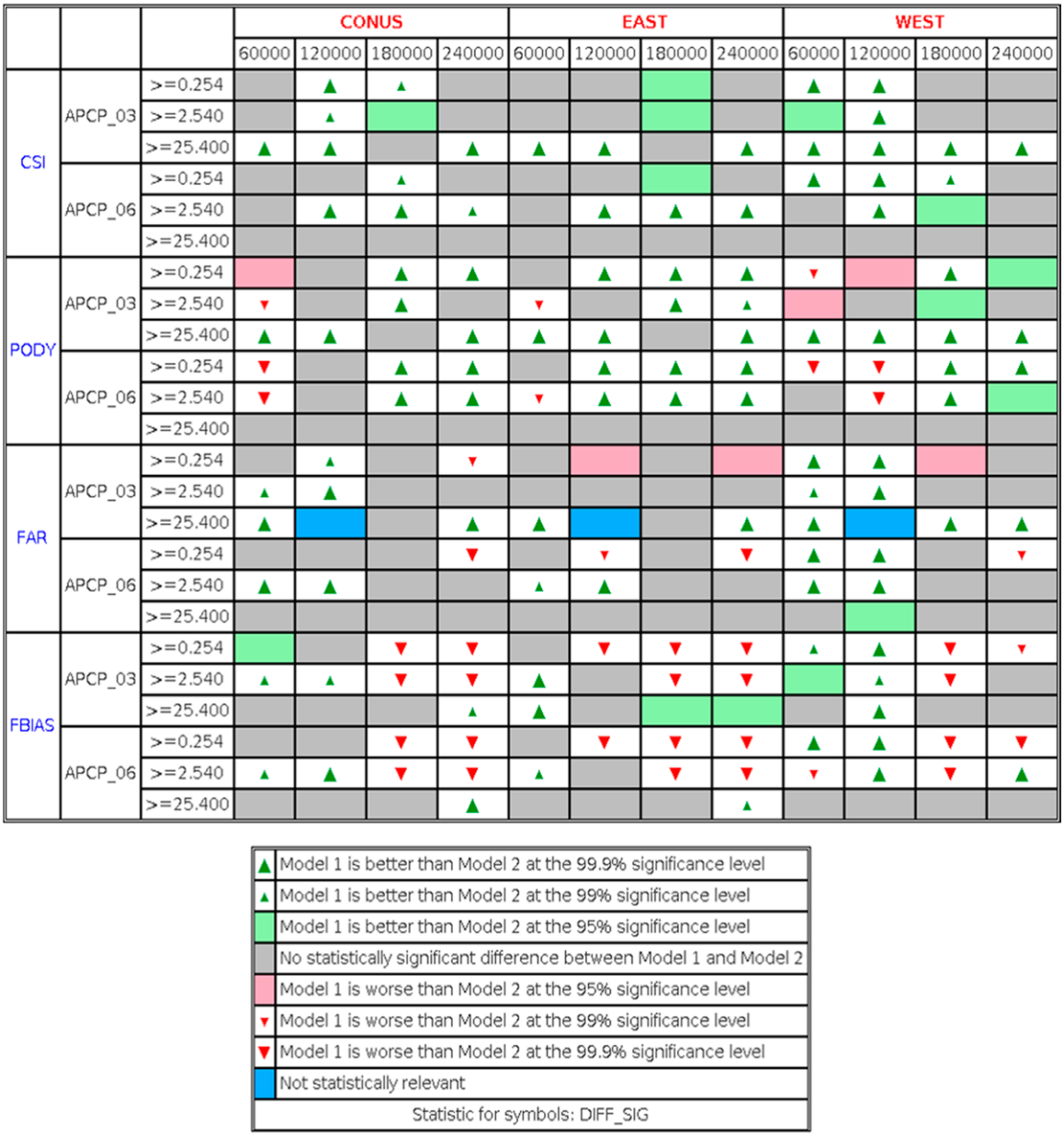

Fig. 12. Scorecard summarizing differences in performance between models across multiple forecast attributes. The events evaluated in this example are precipitation accumulated over 3 and 6 h (APCP_03, APCP_06) for three thresholds (precipitation > 0.254, 2.54, and 25.4 mm). Other attributes include the aggregation variables: region (CONUS, EAST, WEST) and lead time (6, 12, 18, 24 h). The statistics computed for this example are CSI, PODY (POD for "yes" events), FAR, and FBIAS (frequency bias). Symbols and colors are used to compare performance between Models 1 and 2. (Figure produced by METviewer.)

Figure 14 is an example application of this workflow to climate model predictions of monthly precipitation for a single month and year (with current climate). The top diagram shows ensemble mean values computed by *Ensemble-Stat* for 55,296 points around the globe, and the bottom diagram presents the corresponding rank histogram generated by METviewer. While the mean field does not represent verification per se, this field can be passed back into MET tools for further evaluation using both traditional and spatial methods. The field also provides some diagnostic information about the forecasts and could be compared to the observed mean field. The rank histogram indicates that the ensemble was underdispersive (i.e., the observed monthly precipitation often was smaller or larger than any of the ensemble members).

Spatial verification. The spatial verification example use case represents part of an analysis that might be undertaken by a researcher or model developer interested in the details of a model's performance in replicating features of a precipitation field, or they could be of interest to a hydrologist aiming to understand errors in streamflow forecasts. The use case shown here focuses on application of *MODE* for this purpose, with the METplus diagram for this application shown in Fig. 15.

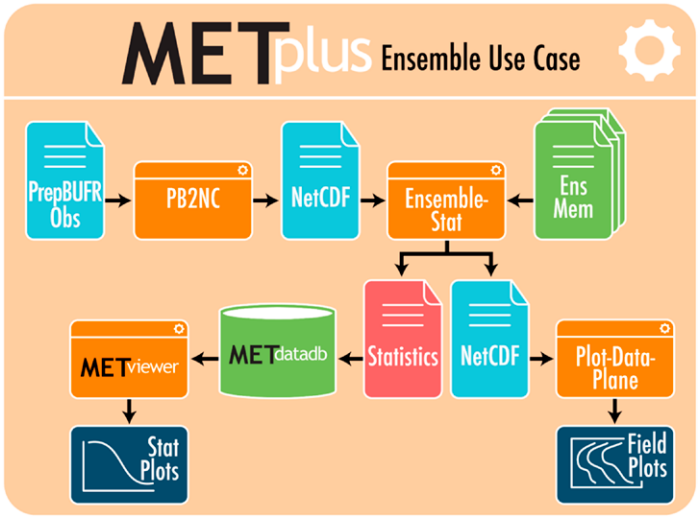

Fig. 13. Example METplus workflow for verification of ensemble forecasts.

Figure 16 shows an example of graphical output from *MODE*, comparing a high-resolution gridded precipitation forecast to an observed precipitation grid. *MODE* identified two objects that matched between the forecast and observation fields (the green and red objects), as well as a few smaller objects (in blue) that were not matched to objects in the other field. A small subset of the *MODE* attributes for this case are shown in Table 1. These attributes indicate that the forecasted red cluster was 3 times larger than the observed cluster but predicted approximately the correct average and extreme intensity values. In contrast, the green cluster forecast was half as large as the observed cluster, and the median (50th percentile) and near peak (90th percentile) intensity values were about one-half (0.47) and one-third (0.30) as big as the values for the observed object, respectively. Information like this can be summarized across larger samples of cases to obtain more general information about forecast performance using either *MODE-Analysis* in concert with user-generated plots, or METviewer (as shown in Fig. 15).

Tropical cyclones. This use case represents the kinds of analyses that might be undertaken in comparing predictions from multiple models (e.g., in the HFIP model intercomparisons), by a model developer aiming to improve TC forecast performance, or by a user interested in selecting the "best" prediction system for a particular application. The METplus workflow for evaluation of basic characteristics of a set of TC forecasts is presented in Fig. 17. This diagram shows the input of A-deck (forecast) and B-deck (observed best track) TC

Table 1. *MODE* example results for the case shown in Fig. 16 (*F* = forecast; *O* = observation). Bold lines represent "paired" attributes (e.g., area differences, intersection area).

Attribute	Cluster 1 (red)	Cluster 2 (green)
Centroid distance (grid squares)	**31.4**	**20.9**
Forecast area (grid squares)	3,802	3,187
Observed area (grid squares)	1,208	7,168
Area difference (*F* – *O*; grid squares)	**2,594**	**–3,981**
Intersection area (grid squares)	**1,080**	**2,436**
Forecast 50th percentile intensity	1.12	1.13
Observed 50th percentile intensity	1.00	2.40
50th percentile intensity ratio (*F*/*O*)	**1.12**	**0.47**
Forecast 90th percentile intensity (mm)	2.68	4.61
Observed 90th percentile intensity (mm)	2.10	15.20
90th percentile intensity ratio (*F*/*O*)	**1.27**	**0.30**

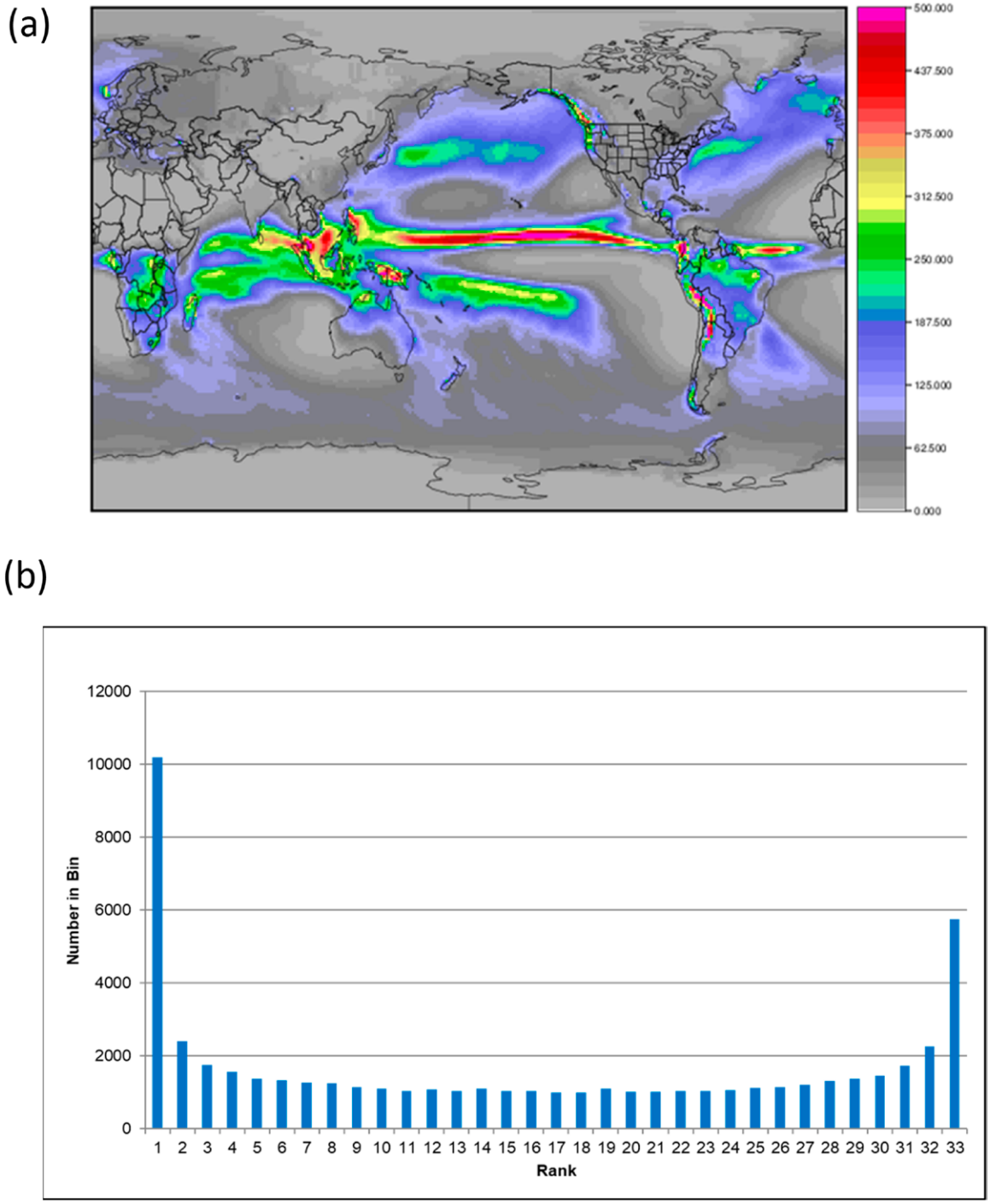

Fig. 14. Example application of *Ensemble-Stat* to predictions of monthly precipitation amount (mm) from a climate model with 32 ensemble members: (a) ensemble mean values and (b) rank histogram. (Top figure generated using Plot-Data-Plane. Bottom figure generated using METviewer.)

information, including the storms' center locations, intensity, and other parameters, followed by matching the forecast and observed values using *TC-Pairs*, computation of verification measures using *TC-Stat*, and user plotting of results, culminating in statistical plots of the verification results.

Figure 18 shows an example of *MET-TC* results from an evaluation of predictions of TC intensity from HFIP. In this example, the performance of intensity predictions from an experimental model (E2) is compared to corresponding forecasts from a "baseline" model (B1). The boxplots in Fig. 18a show results of a *MET-TC* comparison of observed "best track"

TC intensity to model-based TC intensity predictions for a large set of cases, collected over three years of retrospective predictions by models E2 and B1. The boxplots indicate that B1 tended to have somewhat larger errors than E2 for lead times between 36 and 96 h, but that E2 also had larger outlier errors than B1 for some of these lead times. Figure 18b suggests that the experimental model tended to have superior performance for a significant number of cases for lead times of 48, 72, 84, and 96 h. Information like this has informed decisions in HFIP regarding which models might be useful to demonstrate to NHC forecasters.

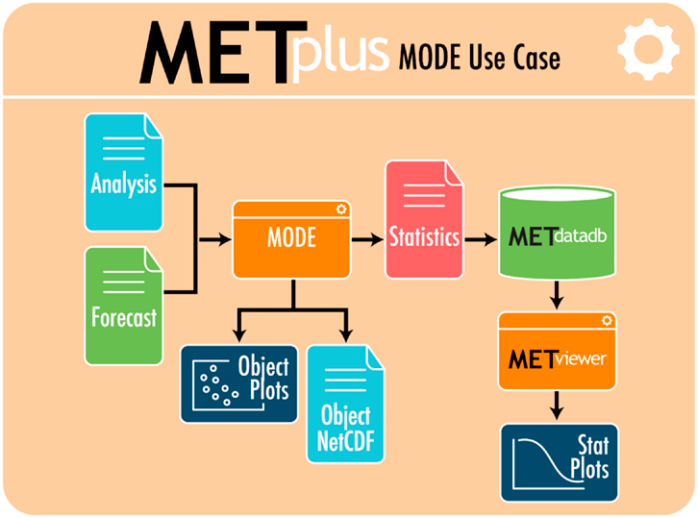

Fig. 15. Example METplus workflow for verification of spatial forecasts using *MODE*.

MET-TC's ability to evaluate predictions of RI/RW events for a different model and set of cases is illustrated in Table 2. This table demonstrates the flexible nature of the RI/RW tool to identify different types of RI/RW events through variations in the threshold for the amount

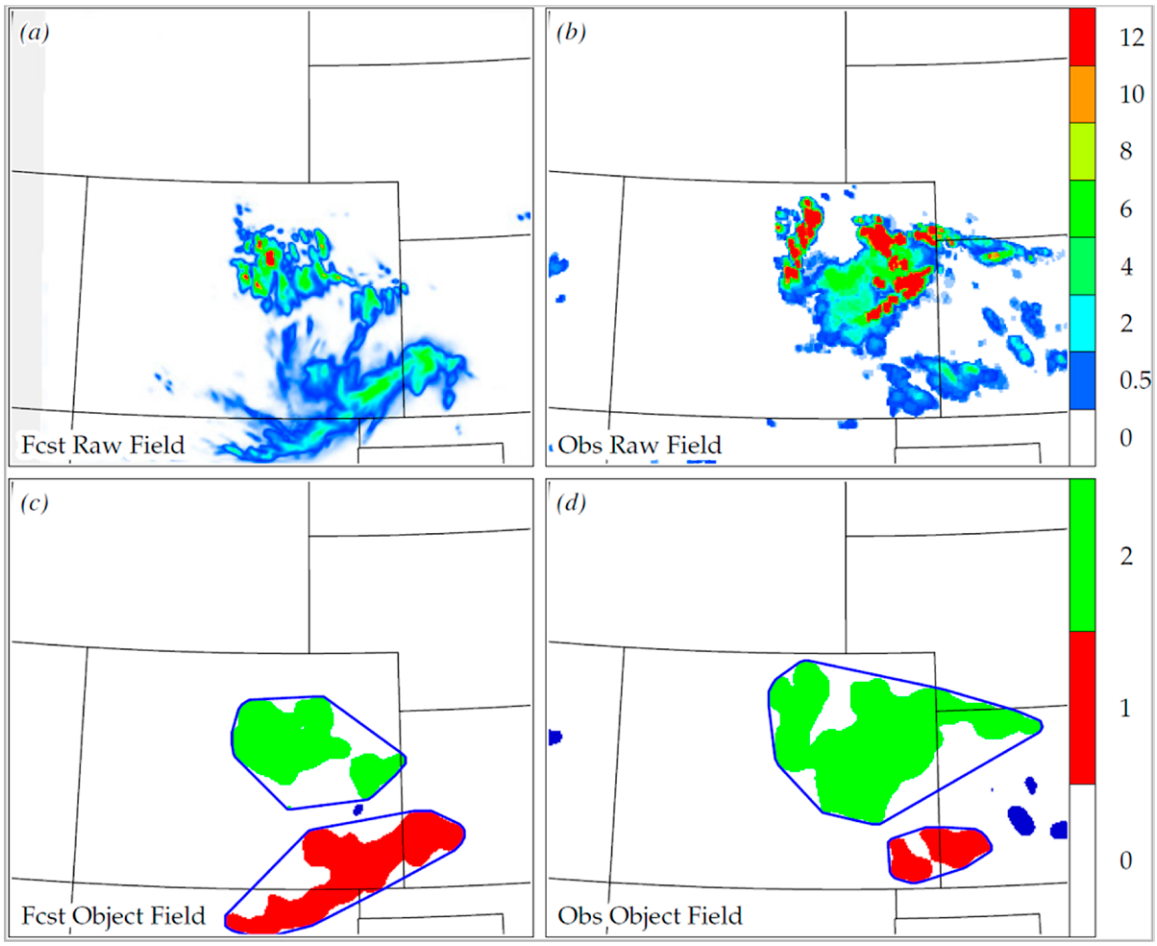

Fig. 16. Example application of *MODE* to a 4-km model-based 6-h precipitation forecast in the Colorado–Kansas region: (a) forecast precipitation (mm), (b) observed precipitation (mm), (c) forecast clusters identified by *MODE*, and (d) observed clusters identified by *MODE*. Green (red) forecast and green (red) observed clusters in (c) and (d) are identified as matched by *MODE*. Blue objects in (c) and (d) are unmatched. (Figures generated by *MODE*.)

of intensification/weakening and the time window for the event occurrence. The example shows results for three thresholds and three time windows. The results in Table 2 indicate that the example model strongly under-forecasted the frequency of RI events.

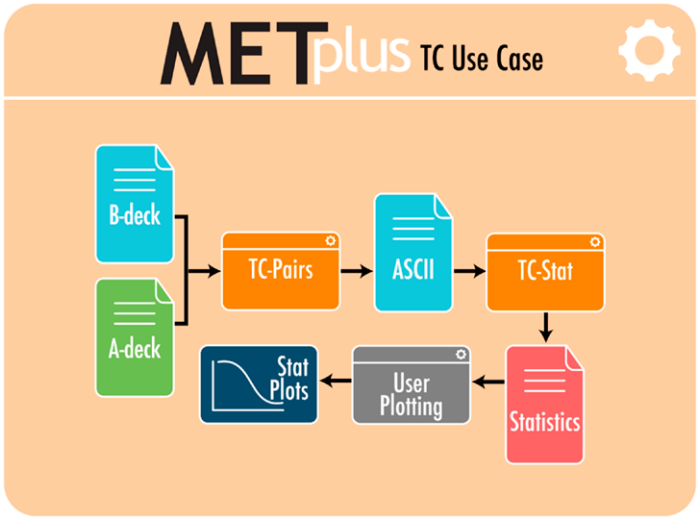

Fig. 17. Example METplus workflow for verification of TC forecasts.

Summary and outlook

MET and METplus represent unique capabilities in the world of forecasting and model evaluation, incorporating an extensive list of methods and tools that is too long to fully describe in this paper. The MET suite of tools is a comprehensive package of modern model evaluation capabilities that is freely available, with extensive user support. These unique qualities have led to broad adoption of MET by operational and research centers both in the United States (e.g., USAF, NOAA, NASA, and NRL) and internationally (including the Met Office in the United Kingdom and the South African Weather Service) and have contributed to the expansion of MET capabilities beyond weather forecasts to include space weather, energy applications, climate predictions, and more.

As MET and METplus have gained larger numbers of users, it has become incumbent on the developers to find ways to streamline its development, implementation, and application, and to increase the ease of use. Hence, the development has evolved toward the use of container technology and a more distributed framework for development, which will increase the number of contributors to the package and lead to expanded use of MET. Engagement of different user groups in the development process will ensure that MET continues to be a state-of-the-art package for forecast evaluation for a wide variety of users and across a broad range of forecast types.

Table 2. Example of verification results for an evaluation of rapid intensification (RI) predictions for a large set of TC cases. POD is probability of detection of events, PODN is probability of detection of nonevents, FAR is false alarm ratio, bias is the frequency bias, and CSI is the critical success index (see Wilks 2019) (1 kt ≈ 0.51 m s^{-1}).

Threshold (kt)	Total count	POD	PODN	FAR	Obs RI event rate	Fcst RI event rate	Bias	CSI
18 h								
25	43,066	0.06	0.99	0.77	0.04	0.01	0.26	0.05
30	43,066	0.02	1.00	0.79	0.03	0.00	0.11	0.02
35	43,066	0.01	1.00	0.84	0.02	0.00	0.05	0.01
24 h								
25	39,658	0.13	0.98	0.64	0.07	0.02	0.35	0.10
30	39,658	0.07	0.99	0.66	0.04	0.01	0.22	0.06
35	39,658	0.03	1.00	0.65	0.03	0.00	0.09	0.03
30 h								
25	36,392	0.19	0.98	0.54	0.10	0.04	0.40	0.15
30	36,392	0.11	0.99	0.62	0.06	0.02	0.29	0.09
35	36,392	0.07	1.00	0.61	0.04	0.01	0.17	0.06

Although MET was initiated with the idea of creating a state-of-the-art software package for forecast evaluation, the field of forecast verification has continued to evolve since 2007, with frequent new advances (Ebert et al. 2013; Dorninger et al. 2018a,b). Hence, to maintain relevance, it will be critical for MET development to endeavor to incorporate these new ideas and approaches. Specific areas of focus could include the implementation of additional spatial methods, such as image warping (Gilleland et al. 2010), a multivariate version of *MODE*, and other object-based approaches (Brown et al. 2012). MET development will strive to maintain consistency with the methods used operationally (e.g., by NCEP, Air Force, other operational prediction agencies), and to also include tools used in research (e.g., process-based methods; e.g., Maloney et al. 2019). In summary, METplus will continue to facilitate generalization and consistency in the application of the various tools included in MET for both researchers and operational users—enabling more meaningful comparisons of operational and research-based products.

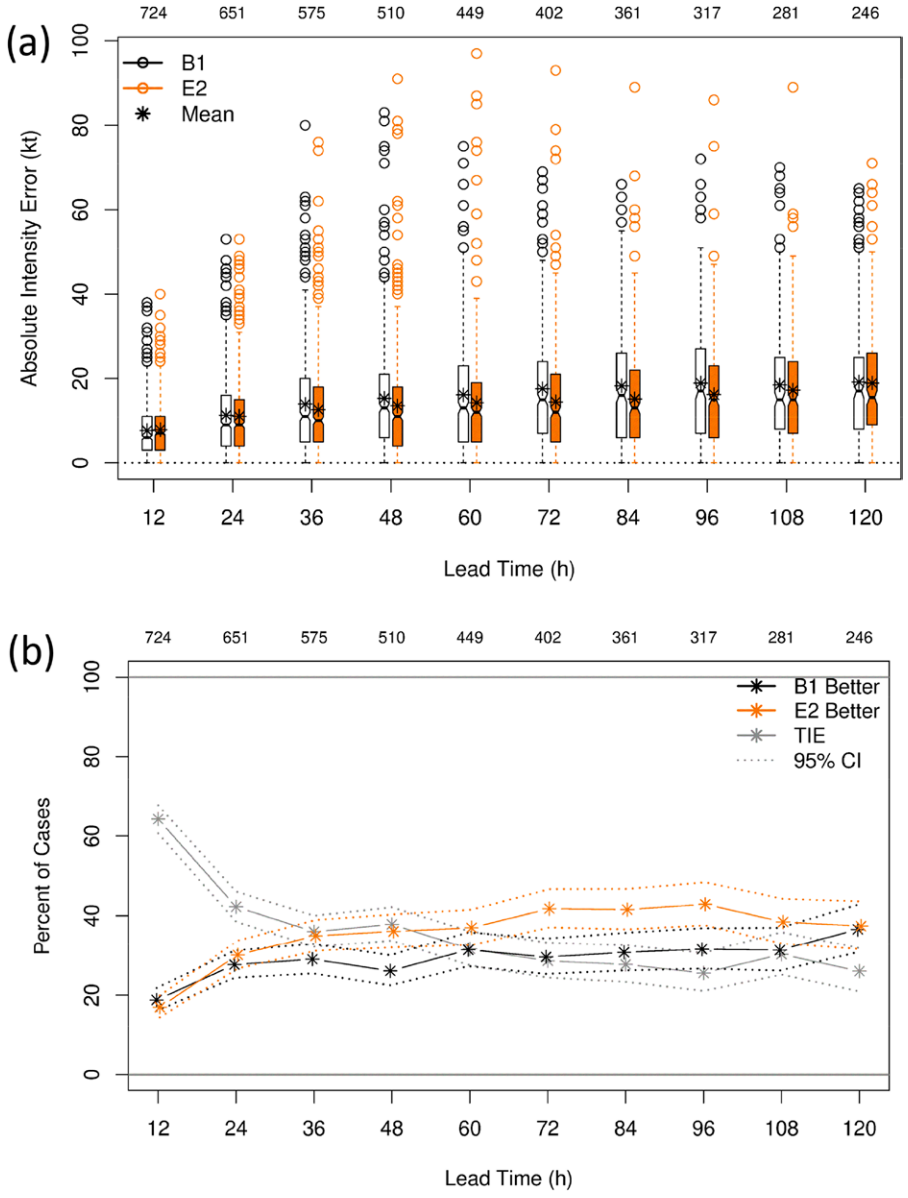

Fig. 18. TC verification results for comparison of predicted TC intensity (experimental model E2) compared to results for a baseline model (B1): (a) boxplots of TC intensity error values for the two models and (b) frequency of superior performance by E2 compared to B1. In (a), the boxes show the 0.25th, 0.50th, and 0.75th quantile values, asterisks show the means, ends of the dashed "whiskers" above and below the central box areas represent the expected extreme values, and outliers are represented by the circles above the box-and-whisker areas. In (b), orange (blue) points indicate the frequency of cases for which E2 (B1) performed better than the other model, with ties indicated by the gray lines, and 0.95 confidence intervals indicated by the dotted lines. Sample sizes are shown above the graphs. (Figures generated using R.)

New domestic and international partnerships have resulted in development of new tools such as feature-centric model evaluation capabilities (e.g., for cyclones, convective systems, droughts) and flexible designs for scorecards, which can summarize results across a wide sample of verification analyses. The MET development team looks forward to additional partnerships in the future, and to learning from users regarding their needs for forecast verification capabilities. Additionally, METplus community support will continue through the DTC but will evolve into a community forum

rather than a help desk. It is an exciting time in the advancement of verification tools worldwide, and MET is poised to take advantage of those new ideas and capabilities and make them available to the community.

Acknowledgments. MET and METplus development has always been a team effort, and would not have been possible without the significant contributions of many individuals, including T. Arbetter, D. Fillmore, H. Fisher, G. McCabe, J. Prestopnik, H. Soh, and M. Win-Gildenmeister, (all of NCAR); J. Frimel, B. Strong, J. Hamilton, J. Duda, R. Pierce, M. Smith, V. Hagerty, K. Searight, and I. McGinnis [all of Cooperative Institute for Research in the Atmosphere (CIRA) at Colorado State University and NOAA Global Systems Laboratory (GSL)]; and M. Marquis and D. Turner (both of NOAA GSL). We are indebted to the USAF, NOAA, and NCAR for their financial support and partnerships in MET development, especially through the DTC. The many contributions of ideas for new capabilities by the DTC Science Advisory Board, DTC visitors, participants in MET workshops and METplus tutorials, and individual suggestions via the MET helpdesk have led to many important enhancements to MET and METplus. We are indebted to E. Tollerud (retired), M. Mittermaier (Met Office), and two anonymous reviewers who provided very helpful and constructive reviews of this paper. We also are grateful to C. Halley Gotway for her extensive help with the graphics. Finally, we express our gratitude to staff at the U.S. Naval Research Laboratory, the U.S. National Aeronautical and Space Administration, the U.S. Department of Energy, State University of New York at Stony Brook, the University of Illinois at Urbana–Champaign, the Cooperative Institute for Meteorological Satellite Studies at the University of Wisconsin–Madison, the Cooperative Institute for Mesoscale Meteorological Studies at University of Oklahoma, Cooperative Institute for Research in the Atmosphere at Colorado State University, Cooperative Institute for Research in Environmental Sciences at the University of Colorado Boulder, and the MET Office in the United Kingdom for their interest in METplus and for contributing many ideas for METplus development.

Appendix: Abbreviations

AERONET	Aerosol Robotic Network
AOD	Aerosol optical depth
ASCII	American Standard Code for Information Interchange
ATCF	Automated tropical cyclone forecast
CALIPSO	Cloud–Aerosol Lidar and Infrared Pathfinder Satellite Observations
CF	Climate forecast
CI	Confidence interval
CRPS	Continuous ranked probability score
CRPSS	Continuous ranked probability skill score
CSI	Critical success index
CTS	Contingency table statistics
DOD	Department of Defense
DTC	Developmental Testbed Center
EMC	Environmental Modeling Center (of the NWS)
ETS	Equitable threat score
FBIAS	Frequency bias
FSS	Fractions skill score
GO	Generalized operations
GOES	Geostationary satellite
GRIB	Gridded binary
GSI	Gridpoint Statistical Interpolation
GSS	Gilbert skill score
HFIP	Hurricane Forecast Improvement Project

HiRA	High-resolution analysis
JWGFVR	Joint Working Group on Forecast Verification Research
MADIS	Meteorological Assimilation Data Ingest System
MAE	Mean absolute error
MED	Mean error distance
MET	Model Evaluation Tools
METdatadb	METplus database
METexpress	Simplified desktop version of METviewer
METplus	Infrastructure for MET tools
METviewer	MET visualization platform
MODE	Method for Object-Based Diagnostic Evaluation
MSE	Mean-square error
MTD	MODE time domain
NASA	National Aeronautics and Space Administration
NCAR	National Center for Atmospheric Research
NCEP	National Centers for Environmental Prediction
netCDF	Network Common Data Form
NHC	National Hurricane Center
NOAA	National Oceanic and Atmospheric Administration
NRL	Navy Research Laboratory
NWP	Numerical weather prediction
NWS	National Weather Service
PDF	Probability density function
PIT	Probability integral transform
POD	Probability of detection
PODN	Probability of detection of “no” event
PODY	Probability of detection of “yes” event
PrepBUFR	Prepared Binary Universal Form for the Representation of Meteorological Data
QPE	Quantitative precipitation estimate
RI	Rapid intensification
RMSE	Root-mean-square error
RMW	Radius of maximum winds
RW	Rapid weakening
STAT	Name applied to MET statistical output files (e.g., from *Grid-Stat*, *Point-Stat*)
SURFRAD	Surface Radiation
TC	Tropical cyclone
USAF	U.S. Air Force
WMO	World Meteorological Organization
WPC	Weather Prediction Center of the NWS
WRF	Weather Research and Forecasting Model

References

Ahijevych, D., E. Gilleland, B. G. Brown, and E. E. Ebert, 2009: Application of spatial verification methods to idealized and NWP-gridded precipitation forecasts. *Wea. Forecasting*, **24**, 1485–1497, https://doi.org/10.1175/2009WAF2222298.1.

Atger, F., 2001: Verification of intense precipitation forecasts from single models and ensemble prediction systems. *Nonlinear Processes Geophys.*, **8**, 401–417, https://doi.org/10.5194/npg-8-401-2001.

Baldwin, M. E., and J. S. Kain, 2006: Sensitivity of several performance measures to displacement error, bias, and event frequency. *Wea. Forecasting*, **21**, 636–648, https://doi.org/10.1175/WAF933.1.

Bernardet, L., and Coauthors, 2008: The Developmental Testbed Center and its Winter Forecasting Experiment. *Bull. Amer. Meteor. Soc.*, **89**, 611–628, https://doi.org/10.1175/BAMS-89-5-611.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Briggs, W. M., and R. A. Levine, 1997: Wavelets and field forecast verification. *Mon. Wea. Rev.*, **125**, 1329–1341, https://doi.org/10.1175/1520-0493(1997)125<1329:WAFFV>2.0.CO;2.

Brooks, H. E., and C. A. Doswell III, 1996: A comparison of measures-oriented and distributions-oriented approaches to forecast verification. *Wea. Forecasting*, **11**, 288–303, https://doi.org/10.1175/1520-0434(1996)011<0288:ACOMOA>2.0.CO;2.

——, M. Kay, and J. A. Hart, 1998: Objective limits on forecasting skill of rare events. Preprints, *19th Conf. on Severe Local Storms*, Minneapolis, MN, Amer. Meteor. Soc., 552–555.

Brown, B. G., and A. H. Murphy, 1987: Quantification of uncertainty in fire-weather forecasts: Some results of operational and experimental forecasting programs. *Wea. Forecasting*, **2**, 190–205, https://doi.org/10.1175/1520-0434(1987)002<0190:QOUIFW>2.0.CO;2.

——, E. Gilleland, and E. E. Ebert, 2012: Forecasts of spatial fields. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, 2nd ed., I. T. Jolliffe and D. B. Stephenson, Eds., Wiley, 95–117.

Bullock, R. G., B. G. Brown, and T. L. Fowler, 2016: Method for Object-Based Diagnostic Evaluation. NCAR Tech. Note NCAR/TN-532+STR, 84 pp., https://doi.org/10.5065/D61V5CBS.

Candille, G., and O. Talagrand, 2008: Impact of observational error on the validation of ensemble prediction systems. *Quart. J. Roy. Meteor. Soc.*, **134**, 959–971, https://doi.org/10.1002/qj.268.

Casati, B., G. Ross, and D. B. Stephenson, 2004: A new intensity-scale approach for the verification of spatial precipitation forecasts. *Meteor. Appl.*, **11**, 141–154, https://doi.org/10.1017/S1350482704001239.

——, and Coauthors, 2008: Forecast verification: Current status and future directions. *Meteor. Appl.*, **15**, 3–18, https://doi.org/10.1002/met.52.

Crocker, R., J. Maksymczuk, M. Mittermaier, M. Tonani, and C. Pequignet, 2020: An approach to the verification of high-resolution ocean models using spatial methods. *Ocean Sci.*, **16**, 831–845, https://doi.org/10.5194/os-16-831-2020.

Davis, C. A., B. G. Brown, and R. G. Bullock, 2006a: Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas. *Mon. Wea. Rev.*, **134**, 1772–1784, https://doi.org/10.1175/MWR3145.1.

——, ——, and ——, 2006b: Object-based verification of precipitation forecasts. Part II: Application to convective rain systems. *Mon. Wea. Rev.*, **134**, 1785–1795, https://doi.org/10.1175/MWR3146.1.

——, ——, ——, and J. Halley Gotway, 2009: The Method for Object-Based Diagnostic Evaluation (MODE) applied to numerical forecasts from the 2005 NSSL/SPC spring program. *Wea. Forecasting*, **24**, 1252–1267, https://doi.org/10.1175/2009WAF2222241.1.

Dorninger, M., P. Friederichs, S. Wahl, M. P. Mittermaier, C. Marsigli, and B. G. Brown, 2018a: Forecast verification methods across time and space scales—Part I. *Meteor. Z.*, **27**, 433–434, https://doi.org/10.1127/metz/2018/0955.

——, E. Gilleland, B. Casati, M. P. Mittermaier, E. E. Ebert, B. G. Brown, and L. Wilson, 2018b: The setup of the MesoVICT project. *Bull. Amer. Meteor. Soc.*, **99**, 1887–1906, https://doi.org/10.1175/BAMS-D-17-0164.1.

Doswell, C. A., III, R. Davies-Jones, and D. L. Keller, 1990: On summary measures of skill in rare event forecasting based on contingency tables. *Wea. Forecasting*, **5**, 576–585, https://doi.org/10.1175/1520-0434(1990)005<0576:OSMOSI>2.0.CO;2.

Ebert, E. E., 2009: Neighborhood verification: A strategy for rewarding close forecasts. *Wea. Forecasting*, **24**, 1498–1510, https://doi.org/10.1175/2009WAF2222251.1.

——, and J. L. McBride, 2000: Verification of precipitation in weather systems: Determination of systematic errors. *J. Hydrol.*, **239**, 179–202, https://doi.org/10.1016/S0022-1694(00)00343-7.

——, and Coauthors, 2013: Progress and challenges in forecast verification. *Meteor. Appl.*, **20**, 130–139, https://doi.org/10.1002/met.1392.

Efron, B., and R. J. Tibshirani, 1993: *An Introduction to the Bootstrap*. Chapman and Hall, 416 pp.

Ehrendorfer, M., and A. H. Murphy, 1988: Comparative evaluation of weather forecasting systems: Sufficiency, quality, and accuracy. *Mon. Wea. Rev.*, **116**, 1757–1770, https://doi.org/10.1175/1520-0493(1988)116<1757:CEOWFS>2.0.CO;2.

Finley, J. P., 1884: Tornado predictions. *Amer. Meteor. J.*, **1**, 85–88.

Gall, R., J. Franklin, F. D. Marks, E. N. Rappaport, and F. Toepfer, 2013: The Hurricane Forecast Improvement Project. *Bull. Amer. Meteor. Soc.*, **94**, 329–343, https://doi.org/10.1175/BAMS-D-12-00071.1.

Gilbert, G. K., 1884: Finley's tornado predictions. *Amer. Meteor. J.*, **1**, 166–172.

Gilleland, E., 2010: Confidence intervals for forecast verification. NCAR Tech. Note NCAR/TN-479+STR, 78 pp., http://doi.org/10.5065/D6WD3XJM.

——, 2011: Spatial forecast verification: Baddeley's delta metric applied to the ICP test cases. *Wea. Forecasting*, **26**, 409–415, https://doi.org/10.1175/WAF-D-10-05061.1.

——, 2017: A new characterization in the spatial verification framework for false alarms, misses, and overall patterns. *Wea. Forecasting*, **32**, 187–198, https://doi.org/10.1175/WAF-D-16-0134.1.

——, D. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert, 2009: Intercomparison of spatial forecast verification methods. *Wea. Forecasting*, **24**, 1416–1430, https://doi.org/10.1175/2009WAF2222269.1.

——, J. Lindström, and F. Lindgren, 2010: Analyzing the image warp forecast verification method on precipitation fields from the ICP. *Wea. Forecasting*, **25**, 1249–1262, https://doi.org/10.1175/2010WAF2222365.1.

——, A. S. Hering, T. L. Fowler, and B. G. Brown, 2018: Testing the tests: What are the impacts of incorrect assumptions when applying confidence intervals or hypothesis tests to compare competing forecasts? *Mon. Wea. Rev.*, **146**, 1685–1703, https://doi.org/10.1175/MWR-D-17-0295.1.

——, G. Skok, B. G. Brown, B. Casati, M. Dorninger, M. P. Mittermaier, N. Roberts, and L. J. Wilson, 2020: A novel set of geometric test fields with application to distance measures. *Mon. Wea. Rev.*, **148**, 1653–1673, https://doi.org/10.1175/MWR-D-19-0256.1.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155–167, https://doi.org/10.1175/1520-0434(1999)014<0155:HTFENP>2.0.CO;2.

Hitchens, N. M., H. E. Brooks, and M. P. Kay, 2013: Objective limits on forecasting skill of rare events. *Wea. Forecasting*, **28**, 525–534, https://doi.org/10.1175/WAF-D-12-00113.1.

Jolliffe, I. T., 2007: Uncertainty and inference for verification measures. *Wea. Forecasting*, **22**, 637–650, https://doi.org/10.1175/WAF989.1.

——, and D. B. Stephenson, 2012: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Wiley-Blackwell, 274 pp.

Maloney, E. D., and Coauthors, 2019: Process-oriented evaluation of climate and weather forecasting models. *Bull. Amer. Meteor. Soc.*, **100**, 1665–1686, https://doi.org/10.1175/BAMS-D-18-0042.1.

Marzban, C., 1998: Scalar measures of performance in rare-event situations. *Wea. Forecasting*, **13**, 753–763, https://doi.org/10.1175/1520-0434(1998)013<0753:SMOPIR>2.0.CO;2.

Mason, S. J., 2008: Understanding forecast verification statistics. *Meteor. Appl.*, **15**, 31–40, https://doi.org/10.1002/met.51.

Mass, C. F., D. Ovens, K. Westrick, and B. A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? The results of two years of real-time numerical weather prediction over the Pacific Northwest. *Bull. Amer. Meteor. Soc.*, **83**, 407–430, https://doi.org/10.1175/1520-0477(2002)083<0407:DIHRPM>2.3.CO;2.

Mittermaier, M., 2014: A strategy for verifying near-convection-resolving model forecasts at observing sites. *Wea. Forecasting*, **29**, 185–204, https://doi.org/10.1175/WAF-D-12-00075.1.

——, and G. Csima, 2017: Ensemble versus deterministic performance at the kilometer scale. *Wea. Forecasting*, **32**, 1697–1709, https://doi.org/10.1175/WAF-D-16-0164.1.

——, R. North, A. Semple, and R. Bullock, 2016: Feature-based diagnostic evaluation of global NWP forecasts. *Mon. Wea. Rev.*, **144**, 3871–3893, https://doi.org/10.1175/MWR-D-15-0167.1.

Murphy, A. H., 1986: A new decomposition of the Brier score: Formulation and interpretation. *Mon. Wea. Rev.*, **114**, 2671–2673, https://doi.org/10.1175/1520-0493(1986)114<2671:ANDOTB>2.0.CO;2.

——, 1991: Forecast verification: Its complexity and dimensionality. *Mon. Wea. Rev.*, **119**, 1590–1601, https://doi.org/10.1175/1520-0493(1991)119<1590:FVICAD>2.0.CO;2.

——, and H. Daan, 1985: Forecast evaluation. *Probability, Statistics, and Decision Making in the Atmospheric Sciences*, A. H. Murphy and R. W. Katz, Eds., Westview Press, 379–437.

——, and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.*, **115**, 1330–1338, https://doi.org/10.1175/1520-0493(1987)115<1330:AGFFFV>2.0.CO;2.

——, and ——, 1992: Diagnostic verification of probability forecasts. *Int. J. Forecasting*, **7**, 435–455, https://doi.org/10.1016/0169-2070(92)90028-8.

——, and D. S. Wilks, 1998: A case study of the use of statistical models in forecast verification: Precipitation probability forecasts. *Wea. Forecasting*, **13**, 795–810, https://doi.org/10.1175/1520-0434(1998)013<0795:ACSOTU>2.0.CO;2.

——, B. G. Brown, and Y.-S. Chen, 1989: Diagnostic verification of temperature forecasts. *Wea. Forecasting*, **4**, 485–501, https://doi.org/10.1175/1520-0434(1989)004<0485:DVOTF>2.0.CO;2.

Nurmi, P., 1994: Recommendations on the verification of local weather forecasts. ECMWF Rep., 14 pp.

Powers, J. G., and Coauthors, 2017: The Weather Research and Forecasting Model: Overview, system efforts, and future directions. *Bull. Amer. Meteor. Soc.*, **98**, 1717–1737, https://doi.org/10.1175/BAMS-D-15-00308.1.

Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97, https://doi.org/10.1175/2007MWR2123.1.

Rodwell, M. J., D. S. Richardson, T. D. Hewson, and T. Haiden, 2010: A new equitable score suitable for verifying precipitation in numerical weather prediction. *Quart. J. Roy. Meteor. Soc.*, **136**, 1344–1363, https://doi.org/10.1002/qj.656.

Roebber, P. J., 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting*, **24**, 601–608, https://doi.org/10.1175/2008WAF2222159.1.

Smith, L. A., and J. A. Hansen, 2005: Extending the limits of ensemble forecast verification with the minimum spanning tree. *Mon. Wea. Rev.*, **132**, 1522–1528, https://doi.org/10.1175/1520-0493(2004)132<1522:ETLOEF>2.0.CO;2.

Stanski, H. R., L. J. Wilson, and W. R. Burrows, 1989: Survey of common verification methods in meteorology. 2nd ed. WMO Rep. WMO/TD-358, 81 pp., www.cawcr.gov.au/projects/verification/Stanski_et_al/Stanski_et_al.html.

Stephenson, D. B., 2000: Use of the "odds ratio" for diagnosing forecast skill. *Wea. Forecasting*, **15**, 221–232, https://doi.org/10.1175/1520-0434(2000)015<0221:UOTORF>2.0.CO;2.

Taylor, K. E., 2001: Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.*, **106**, 7183–7192, https://doi.org/10.1029/2000JD900719.

Wilks, D. S., 2018: On assessing calibration of multivariate ensemble forecasts. *Quart. J. Roy. Meteor. Soc.*, **143**, 164–172, https://doi.org/10.1002/qj.2906.

——, 2019: Forecast verification. *Statistical Methods in the Atmospheric Sciences*, 4th ed., Elsevier, 369–483.

Wilson, L. J., W. R. Burrows, and A. Lanzinger, 1999: A strategy for verification of weather element forecasts from an ensemble prediction system. *Mon. Wea. Rev.*, **127**, 956–970, https://doi.org/10.1175/1520-0493(1999)127<0956:ASFVOW>2.0.CO;2.

WMO, 2010: Global aspects. Manual on the global data-processing and forecasting system, Vol. I, WMO Rep. WMO-485, 196 pp., www.wmo.int/pages/prog/www/DPFS/Manual/GDPFS-Manual.html.