9A.6

THE MODEL EVALUATION TOOLS (MET): COMMUNITY TOOLS FOR FORECAST EVALUATION

Barbara Brown[1], John Halley Gotway, Randy Bullock, Eric Gilleland,
Tressa Fowler, David Ahijevych, and Tara Jensen

National Center for Atmospheric Research, Boulder, Colorado

1. INTRODUCTION

Assessments of forecast quality are a critical component of the forecast development, improvement, and application processes. While some verification capabilities have been used in practice for many years, modern, state-of-the-art tools are especially needed to provide meaningful evaluations of high-resolution numerical weather prediction (NWP) model forecasts. The Model Evaluation Tools (MET) verification package has been developed to provide this capability and to aid the Developmental Testbed Center (DTC) in testing and evaluation of the Weather Research and Forecasting (WRF) model.

MET is a community developed package that is freely available and distributed through the DTC. The package includes numerous statistical tools for forecast evaluation, including traditional measures for categorical and continuous variables [e.g., Root-mean squared error (RMSE), Critical Success Index (CSI)]. In addition, MET provides advanced spatial forecast evaluation techniques. Two general categories of spatial methods are currently included in MET: (i) “object-based” and (ii) “neighborhood” techniques. An upcoming release will also include wavelet techniques that are able to estimate forecast performance capabilities at different scales.To account for the uncertainty associated with estimation of traditional verification measures, methods for estimating confidence intervals for the verification statistics are an integral part of MET.

This paper describes MET development, capabilities, and future plans.

2. MET HISTORY AND GOALS

MET development was initiated by the DTC in 2006 in response to the needs of the forecasting and research communities and the DTC for forecast evaluation tools that are appropriate for assessing the performance of forecasts produced by high-resolution NWP models. The major goals considered in the design of MET focused on (i) incorporating state-of-the-art capabilities; (ii) making the tools freely available to the operational, research, model development, verification, and user communities (hereafter, referred to as “the community”); (iii) enabling the community to help create the tools through contributed methods and display capabilities.

MET is designed to be modular so that new tools can be incorporated with relative ease. The tools generally are written in the C++ programming language. However, they are highly configurable through the use of ASCII configuration files and command-line arguments.

Version 1.0 of MET was released in January 2008. MET version 1.1 followed in July 2008, and we expect to announce the release of version 2.0 in late winter 2009.

3. MET COMMUNITY

As noted in Section 2, the MET community is quite diverse. It includes model developers working on the WRF model, staff at the DTC, university researchers, and operational centers. As of January 2009, the number of registered MET users is 265; of these users, 56% are from universities, 26% are from government organizations, 7% are associated with private corporations, and 10% are from non-profit organizations. About 43% of the users are located in the U.S.

The MET community has been involved in the development of MET from its inception. In particular, the original components of MET were discussed at town hall meetings in 2007, during which potential MET users were surveyed regarding their priorities for MET development. In addition, users, modeling experts, and verification experts participated in two MET workshops in 2007 and 2008. The presentations and discussions at these workshops provided guidance on the tools and methods that should be included in MET.

An important goal of MET development is to incorporate new tools and graphics capabilities that are contributed by members of the community, as time and resources allow. A process is being developed to provide guidance on when and whether new tools should be implemented. In addition, the annual MET workshops provide a venue for discussing and determining whether new techniques are mature enough to be included in the package.

4. MET AVAILABILITY AND TRAINING

MET is freely available for download at http://www.dtcenter.org/met/users/. This website also includes a link to the MET Users’ Guide which is available for download in Acrobat pdf format. In addition, the website includes a variety of information about running MET,

[1] *Corresponding author address*: Barbara G. Brown, National Center for Atmospheric Research (NCAR), PO Box 3000, Boulder CO 80307-3000; e-mail: bgb@ucar.edu

known issues, and frequently asked questions. The website also contains links to additional software (e.g., graphics scripts) that may be used to display and interpret MET output. A met_help e-mail address (identified on the web page) links users to support for questions regarding setting up and running MET, as well as the interpretation of MET output.

A link to an on-line training module is also available on the MET website. The on-line tutorial provides guidance on downloading, compiling, and running MET and its various components. MET tutorials have also been provided as part of the WRF tutorials which take place twice each year at the National Center for Atmospheric Research (NCAR). We anticipate that the MET tutorials will continue to be included with the WRF tutorials in future years.

5. MET COMPONENTS

The components of MET are illustrated in a system diagram in Fig.1[2] . Broadly speaking, the MET components include data re-formatting tools; statistics tools; and analysis tools. Each of these components is briefly described in the following subsections.

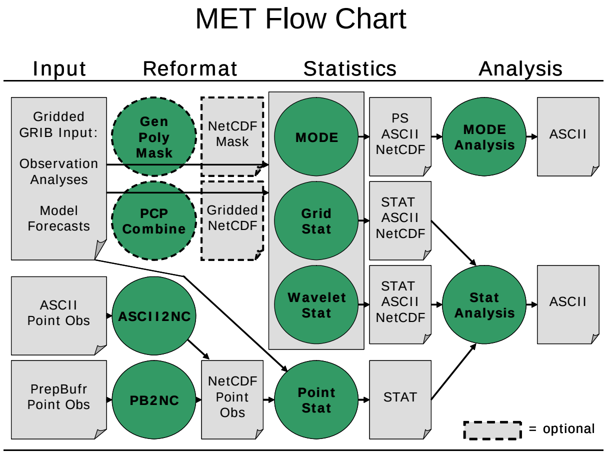

FIGURE 1. System diagram showing major components of MET and primary steps in MET processing for MET version 2.0.

5.1 Data requirements and reformatting

MET typically requires model output and observation analyses in GRIB version 1 format, but can also handle data in the appropriate NetCDF format. For comparisons of gridded forecasts and observations, both datasets must be on the same grid. Point observations in PrepBufr format are available through NOAA's National Centers for Environmental Prediction (NCEP); point observations also may be in ASCII format.

The PCP-Combine, PB2NC, and ASCII2NC tools write intermediate files in NetCDF format; NetCDF is used by MET as an intermediate file format. The PCP-Combine tool can be used to sum accumulated precipitation from several GRIB files into a single NetCDF file containing the desired accumulation period. It also can be used to add or subtract the accumulated precipitation in two GRIB files. The PB2NC tool is used to reformat the input PrepBufr files containing point observations. The NetCDF output of the PB2NC tool is used as input to the verification step performed in the Point-Stat tool. The ASCII2NC tool simply reformats ASCII point observations into the NetCDF format needed by the Point-Stat tool. Finally, the Gen Poly Mask tool applies a lat/lon polyline to generate a 0/1 mask field to be applied to gridded data; this mask is applied to the data files prior to running the Point-Stat or Grid-Stat tools.

5.2 Statistics modules

MET version 1.1 includes three modules that perform statistical verification computations: MODE, Grid-Stat, and Point-Stat. Version 2.0 will include a fourth module, the Wavelet Stat tool. These tools are described in the following subsections.

5.2.1 Gridded and point verification

MET computes a wide variety of traditional statistics for evaluations of forecasts at points and across grids. For example, the Point-Stat tool is useful for performing verification using observations at rawinsonde or surface observation locations. The Grid-Stat tool is appropriate for comparing a model grid to a gridded analysis (e.g., for precipitation).

Grid-Stat and Point-Stat compute verification measures for both categorical (e.g., 0/1) and continuous variables. A wide array of statistics is computed, including most of the measures typically presented in texts and other documents on forecast verification methods (e.g., Jolliffe and Stephenson 2003; Wilks 2006; WWRP/WGNE Joint Working Group on Verification 2009). For example, the basic statistics include measures such as the Critical Success Index (CSI), Gilbert Skill Score [also known as the Equitable Threat Score (ETS)], root-mean-squared error, correlation coefficient, and so on. In addition, the tools provide summary measures for the forecast and observation samples (e.g., mean, variance, percentiles).

The ability to compute confidence intervals for all of the verification measures is an important component of the Grid-Stat and Point-Stat tools. This capability adds meaning to the measures by quantifying the uncertainty associated with the estimates derived from a particular sample of forecasts. Confidence intervals are of particular importance when comparing the performance of two or more models. Comparisons of verification results that don't take into account the sampling variability can lead to incorrect conclusions regarding the improvement of one version of a model over another.

MET includes standard confidence interval methods based on application of the Gaussian (i.e., normal) distribution. However, because the distributions of many

[2] *Note that in addition to the tools included in MET, several freely-available libraries are required to run MET. These libraries are identified in the MET Users' Guide.*

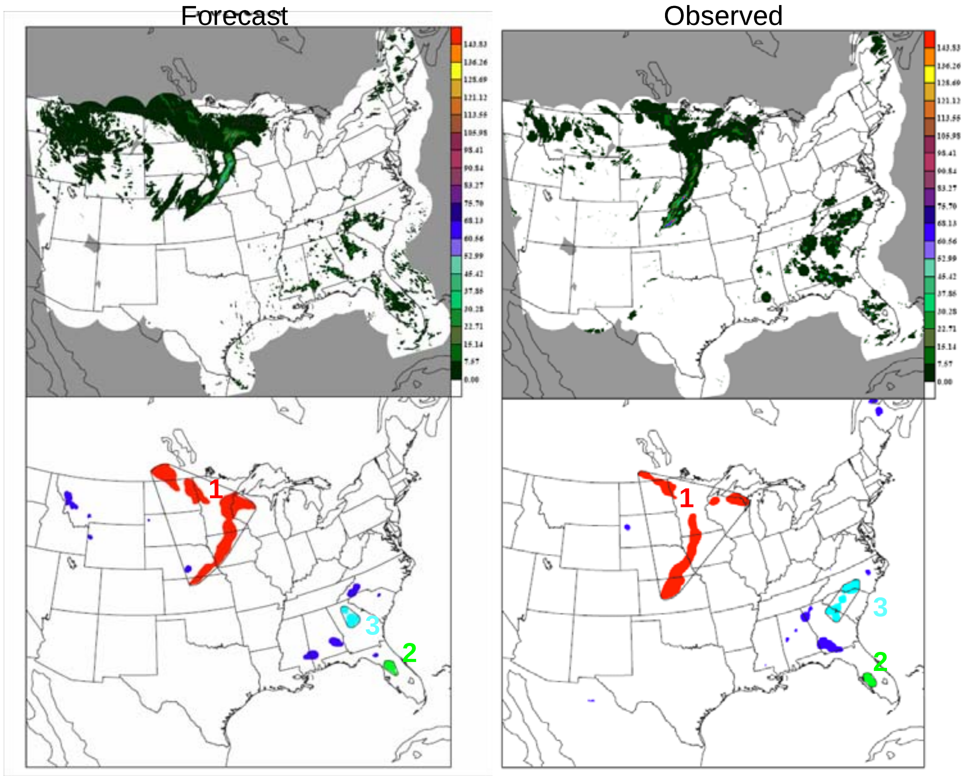

FIGURE 2: Example of MODE applied to WRF forecasts and Stage II precipitation observations. Colored objects represent matched pairs of object clusters. Dark blue objects are unmatched. The matched objects in this example indicate that the forecast was of fairly good quality. In contrast, traditional verification measures suggest that the forecast had little or no skill (e.g., CSI = 0.09).

verification measures cannot be represented by a Gaussian model, MET also provides a non-parametric approach for computing confidence intervals, using the so-called "bootstrap" procedure. Detailed explanations and references for the confidence interval methods are provided in Gilleland (2009).

5.2.2 Spatial verification methods

In recent years, it has become clear that traditional verification approaches are not as informative as might be desirable, especially for evaluation of forecasts from high-resolution NWP models. For example, these measures can indicate that a forecast was incorrect, but they may not indicate the relevance of the error (e.g., whether it was large or small) or the source of the error (e.g., displacement). In response to these issues, the modeling and verification communities have endeavored to develop new spatial methods that provide more meaningful information about forecast performance for gridded forecasts. Several categories of methods have been developed, including object- (or features-) based methods, neighborhood methods, and scale separation methods (Gilleland et al. 2009). MET version 1.1 includes tools for the first two of these categories; version 2.0 will also include wavelet methods that represent the third category.

The Method for Object-based Diagnostic Evaluation (MODE) is the features-based tool currently included in MET. This approach attempts to identify precipitation features in much the same way as they would be identified by a human analyst or forecaster (Davis et al. 2006, 2009). Forecast and observed objects are then matched and compared. An example of an application of MODE is shown in Fig. 2. The MET implementation of MODE provides some graphical output (partially presented in Fig. 2) as well as extensive statistics describing the objects and the strength of the forecast/observation matches.

Neighborhood (or "fuzzy") verification methods are included in MET as part of the Grid-Stat tool. These methods measure variations in forecast performance as the constraints required for matching forecast and observed grids are progressively relaxed (Ebert 2008). The Fractional Skill Score developed by Roberts and Lean (2008) is an example of a neighborhood method included in MET.

MET version 2.0 will also include the Wavelet-Stat tool. The methods included in this tool allow direct assessment of forecast performance at different spatial scales. The intensity-scale approach of Casati et al. (2004) is an example of a scale-separation method that will be included in MET.

Together, the spatial methods included in MET represent state-of-the-art methods for evaluation of these types of forecasts. These methods are most appropriate for application to fields that have coherent structures, such as precipitation or clouds.

5.2.3 Probabilistic

Version 2.0 of MET will include some measures of performance for probabilistic forecasts. These measures will include the Brier score and Brier skill score and their decompositions (i.e., resolution, reliability, and uncertainty); joint and conditional probabilities; discrimination measures; and relative operating curve (ROC) statistics and areas. Evaluation of probabilistic forecasts will be supported by both Point-Stat and Grid-Stat.

5.3 Analysis tools

MET is applied to individual forecasts and observations (e.g., at a single forecast time). The analysis tools in MET are designed to summarize the results across multiple times or to categorize them according to specified stratification criteria. For example, it frequently is of interest to examine verification results as a function of lead time or time of day. The analysis tools for MODE and for Grid-Stat and Point-Stat allow extensive flexibility in how the summaries and stratifications are applied.

6. FUTURE CAPABILITIES

MET is a work in progress. Many new capabilities are planned for the future; the selection of those to be implemented will be influenced by the desires and contributions of the MET community.

Specific plans for future implementation include (i) additional spatial methods [e.g., the Contiguous Rain Area approach of Ebert and McBride (2000)]; (ii) methods for ensemble forecasts; and (iii) methods for cloud forecast evaluation. A graphical interface will also be developed to make it easier to formulate configuration files. In addition, a database and display system is being developed to provide easier application of MET for some users.

7. SUMMARY AND CONCLUSIONS

The Model Evaluation Tools (MET) has been developed by the DTC for use by the community in evaluations of forecasts from high-resolution NWP models. The tools are equally applicable to many other types of forecasts, and have already been used in alternative applications (e.g., for space weather forecasts).

MET has been extensively applied over the last year, by many users around the world. An example application is the NOAA Storm Prediction Center (SPC) Spring Experiment, where MET was demonstrated in 2008. This venue provided an excellent opportunity to expose forecasters and researchers to the new tools included in MET and help them begin to understand their capabilities.

As a community tool, MET development depends on the contributions of the community, both to enhance the tools and to keep them relevant. In particular, the MET development team would appreciate contributions of new methods to consider for implementation. In addition, graphical methods that have been developed for the display of MET results are always welcome for posting on the MET website.

ACKNOWLEDGMENTS

NCAR is sponsored by the National Science Foundation. MET development has been sponsored by the Air Force Weather Agency and NOAA.

REFERENCES

Casati B., G. Ross, and D.B. Stephenson, 2004: A new intensity-scale approach for the verification of spatial precipitation forecasts. *Meteorol. Appl.,* **11**, 141-154.

Davis, C.A., B.G. Brown, and R.G. Bullock, 2006: Object-based verification of precipitation forecasts, Part I: Methodology and application to mesoscale rain areas. *Mon. Wea. Rev.*, **134**, 1772-1784.

Davis, C.A., B.G. Brown, R.G. Bullock and J. Halley Gotway, 2009: The Method for Object-based Diagnostic Evaluation (MODE) Applied to WRF Forecasts from the 2005 SPC Spring Program. Submitted to *Weather and Forecasting*; available at http://www.dtcenter.org/met/users/docs/overview.php.

Ebert E.E., 2008: Fuzzy verification of high resolution gridded forecasts: A review and proposed framework. *Meteorological Applications*, **15**, 51-64.

Ebert, E.E., and JL McBride, 2000. Verification of precipitation in weather systems: determination of systematic errors. *Journal of Hydrology*, **239**,179-202.

Gilleland, E., 2009: Confidence intervals for forecast verification. Available at http://www.dtcenter.org/met/users/docs/overview.php.

Gilleland, E., D. Ahijevych, B.G. Brown, B. Casati, and E.E. Ebert, 2009: Intercomparison of spatial forecast verification methods. Submitted to *Weather and Forecasting*; available at http://www.ral.ucar.edu/projects/icp/specialcollection.html.

Jolliffe, I.T., and D.B. Stephenson, 2003: *Forecast verification. A practitioner's guide in atmospheric science*. Wiley and Sons Ltd, 240 pp.

Roberts, N.M., and H.W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, **136**, 78-96.

Wilks, D., 2006: *Statistical methods in the atmospheric sciences*. Elsevier, San Diego.

WWRP/WGNE Joint Working Group on Verification, 2008: http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html