Establishment of a Verification Methods Testbed at the WRF Developmental Testbed Center

Michael E. Baldwin
Department of Earth and Atmospheric Sciences, Purdue University
West Lafayette, IN 47907

October 14, 2008

Introduction

Traditional methods of measuring the performance of numerical, gridded forecasts (for example, threat scores applied to precipitation) fail to provide meaningful information when applied to forecasts containing realistic, small-scale features. Several researchers have described this problem in detail, while providing suggestions and alternative methods of measuring forecast accuracy (e.g., Anthes 1983; Davis and Carr 2000; Ebert and McBride 2000; Harris et al. 2001; Tustison et al. 2001; Baldwin et al. 2002; Davis et al. 2006). There is a great need within both the research and operational numerical weather prediction communities for new verification methods to verify the high-resolution forecasts that are currently being produced routinely by operational centers and many research groups around the world. Several researchers have proposed improved verification techniques, including object-oriented, fuzzy neighborhood, scale decomposition, and field comparison approaches. These methods have demonstrated the potential to provide users with a wealth of information regarding various aspects of forecast performance, such as direct measures of displacement errors (e.g., Ebert and McBride 2000), event-based errors for specific classes of meteorological phenomena (e.g., Nachamkin et al. 2005), and errors as a function of spatial scale (e.g., Casati et al. 2004). A formal evaluation of a variety of new verification techniques was recently initiated at NCAR (Spatial Forecast Verification Intercomparison Project). It is expected that several of these advanced methods for spatial verification will be incorporated into the WRF verification toolkit being that is currently under development, known as MET (Model Evaluation Toolkit).

While many of the negative aspects of traditional verification methods are widely known, such methods continue to be used extensively to assess the performance of forecast systems. For example, the Environmental Modeling Center (EMC) uses the equitable threat score and bias score (Mesinger 1996) as primary performance metrics when evaluating the accuracy of quantitative precipitation forecasts (QPF) in their operational and experimental forecast systems (e.g., the latest EMC briefing on the most recent NAM upgrade http://www.emc.ncep.noaa.gov/mmb/namchanges_winter2008/nam_upgrades.2008.html). One explanation for why the community continues to use these traditional methods is that these methods have been in use for a long period of time and the users of the verification information are familiar and comfortable with the behavior of those performance measures. A certain level of *credibility* has been established for these methods after measuring forecast performance over a period of several decades (e.g., Fig. 1, from Olson et al. 1995). Such credibility has not yet been established for any of the newer methods of spatial verification that have been recently developed. In addition, many of the results produced by newer methods are somewhat difficult to interpret. It is not clear how one should summarize the results from such methods over long periods of forecasts, such as over a season or multiple years. The *usability* of such methods has not been demonstrated in many cases.

One can envision multiple paths that verification researchers can follow in order to establish an acceptable level of credibility leading to the overall acceptance and adoption of new verification techniques by the operational and research numerical weather prediction (NWP) communities. One such path, which I would describe as the path the verification development community is currently following, consists of the following steps:

1. Develop a new technique
2. Test it on a relatively small number of cases
3. Publish those results and methodology
4. Apply the technique to operational or experimental forecasts on a routine basis
5. Collect verification information from the new technique over several years
6. Compare the information provided by the new and traditional methods
7. Once users have become comfortable with the behavior of the new technique, the new technique is accepted by the community.

Clearly, this path requires several years to complete, steps 4-6 in particular. However, one could also envision an alternate, *accelerated* path to establishing credibility for new verification

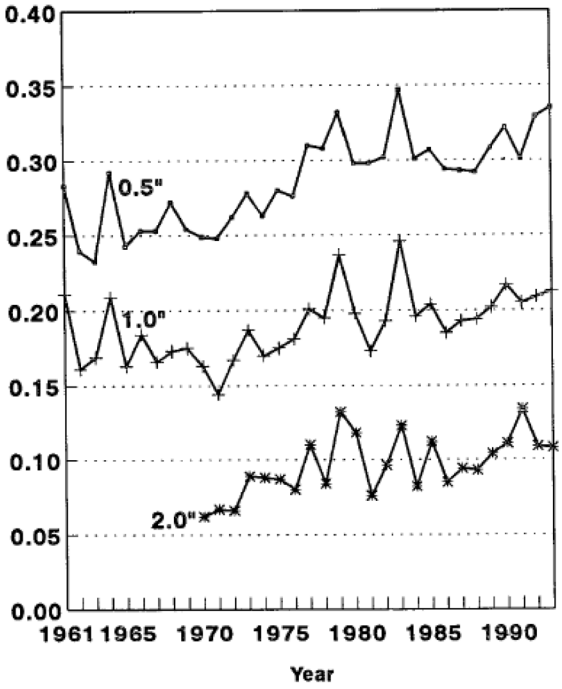

Figure 1: Annual threat scores from NMC (now HPC) forecasters' 0.5, 1.0, and 2.0 inch forecasts for the Day 1 period from 1961 through 1993. (Fig. 6 from Olson et al. 1995)

techniques. Specifically, this involves applying new techniques to a historical database of operational and experimental forecasts that covers a period of several years. This will speed up considerably the process of establishing credibility and eventual acceptance and use by the operational and research NWP communities. Rather than waiting 5-10 years to collect a set of information from new verification techniques in parallel with traditional methods, an archive of forecasts and observations have been collected that will allow researchers to generate results covering a ~7 year period. This archive currently resides at the WRF Developmental Testbed Center (DTC) and establishes a *testbed* for new verification methods.

The Verification Methods Testbed allows developers of verification methods to compare the results of traditional methods with those from advanced techniques, resulting in fast "benchmarking" of the newer techniques. Users can quickly determine how the new techniques detect and measure the degree of difference in performance, for example, between older and newer versions of operational models (GFS, NAM) before and after upgrades were implemented. These differences can easily be compared to those shown by the traditional measures to see if the new verification techniques are providing similar information as the traditional methods, or perhaps

that the new techniques are providing an alternate perspective on the relative performance of those models. Users can quickly determine how new verification methods measure the change in NWP model performance over time.

Verification Methods Testbed

The Verification Methods Testbed (VMT) consists of archives of operational and experimental NWP model output of quantitative precipitation forecasts (QPF), and analyses of observed precipitation, and in its current state is roughly ~15 GB in total size. These forecasts and analyses are collected with a standard naming convention, and have been remapped (interpolated) from their original grids to a common map projection to allow for quick and easy comparisons. These datasets can quickly be used by the MET verification toolkit (http://www.dtcenter.org/met/users/). The VMT datasets are on the NCAR mass store in the following directory: /DTCRT/TESTBED

There are several compressed (bzip2) tar files that are organized by model/analysis. The file names are as follows:

cpcday_g212.tar.bz2
eta_g212.tar.bz2
gfs_g212.tar.bz2
multi_g212.tar.bz2
nam_g212.tar.bz2
stage2.tar.bz2
wrfcaps_g240.tar.bz2
wrfncar_g240.tar.bz2
wrfncep_g240.tar.bz2
wrfnssl_g240.tar.bz2
eta_tar.bz2 (native grid original Eta forecasts)

These tar files can be unpacked on many machines using the following command:

```
tar -jxf FILENAME.tar.bz2
```

On machines that do not allow "tar -jxf", you must first decompress the tar files using:

```
bzip2 -d *.tar.bz2
```

and then untar them using the standard "tar -xf *.tar" commands.

These file names indicate which model or analysis is contained within, along with the grid number for those grids that have been remapped from their original native grid. The operational NCEP models (eta, gfs, nam) have been remapped to AWIPS grid #212, which is a 40km Lambert Conformal grid (see Appendix). The analyses that correspond to those NCEP models are multi_g212, which is a 3h accumulation of the NCEP Stage II hourly radar/gage multi-sensor analysis, and cpcday_g212, which is the NCEP/CPC analysis of 24h gage data remapped from their original 1/8 deg lat/lon grid. The experimental WRF models (wrfcaps, wrfncar, wrfncep, wrfnssl) have been remapped to grid #240, which is the same 4km polar stereographic grid used by the NCEP Stage II/IV national multi-sensor analysis. Detailed information regarding the configuration and execution of these experimental forecasts can be obtained from the SPC/NSSL Hazardous Weather Testbed (HWT) operations plans from the various years that are included here (available at http://hwt.nssl.noaa.gov/). The analyses corresponding with these WRF runs is called stage2, and has no _g*** suffix in the file name since that analysis has not been remapped from the original grid. The files within each tar file are organized by year and month, a subdirectory will be created for each model, and further directories exist for each year and month (YYYYMM).

The remapping was performed using *copygb*, which is distributed as part of the WRF Post Processor (can be obtained from: http://www.dtcenter.org/wrf-nmm/users/downloads/index.php). Interpolation option #6 (budget method) was used, which attempts to conserve the total water in the remapping.

Individual file names generally follow this naming convention:
MDL.YYYYMMDDHH.GRID.fXX
ANL.YYYYMMDDHH.GRID.grb

MDL or ANL are the model/analysis name prefixes, YYYYMMDDHH is the start time/date of each forecast, or the valid time/date of each analysis, GRID is either g212 or g240, fXX is the

forecast hour. All of these files are in GRIB1 format. Generally, forecasts are available from 00 and 1200 UTC cycles at 3h output intervals from the NAM and GFS, out to 84h. Forecasts from the high-resolution experimental WRF runs are generally available at 1h output intervals out to 36h from 0000 UTC cycles.

Future Plans

Future plans for the VMT include the creation of web-based access to the archive and the addition of other forecast variables. An OPeNDAP/THREDDS web server will be acquired to allow for easier access to these data. Additional variables will be added, such as a reflectivity diagnostic from the experimental WRF runs (and observed radar reflectivity for verification), 1h QPF from the operational NAM, surface weather parameters (2m temperature, 10m winds) from the various forecasts and RTMA analyses. When possible, the original forecasts and analyses on the native grids of each will also be made available, which would allow researchers to test methods that utilize native-grid output and also examine the impact of interpolation techniques on verification results. Also, forecasts and analyses from various operational centers from around the world will be included to allow for testing of datasets from outside of the U.S. Ensemble forecasts from the NCEP operational ensemble systems, plus the WRF ensembles that were generated to support the SPC/NSSL HWT during 2007 and 2008 will also be included in the future.

Acknowledgements

I would like to acknowledge Ying Lin at NCEP/EMC for her ongoing efforts in maintaining EMC's archive of QPF data, and Jack Kain at NSSL for his ongoing efforts in maintaining NSSL's archive of experimental WRF model data. Without their help, progress on this project would have been impossible. I would also like to thank everyone at the DTC for their support during this project, especially Pam Johnson for all of her hard work with travel arrangements and other logistics.

Appendix

Grid specifications:

Grid #212

Projection: Lambert Conformal

Nx: 185

Ny: 129

Lat 1: 12.190N

Lon 1: 226.541E = 133.459W

Lonv 265.000E = 95.000W

Dx 40.63525 km

Dy 40.63525 km

Latin1 25.000N

Latin2 25.000N (tangent cone)

Lat/Lon values of the corners of the grid

(1,1) 12.190N, 133.459W

(1,129) 54.536N, 152.856W

(185,129) 57.290N, 49.385W

(185,1) 14.335N, 65.091W

Pole point

(I,J) (105.000, 356.490)

Grid #240

Projection: Polar Stereographic

Nx 1121

Ny 881

La1 23.117N

Lo1 240.977E = 119.023W

Lov 255.000E = 105.000W

Dx 4.7625 km

Dy 4.7625 km

Lat/Lon values of the corners of the grid

(1,1) 23.117N, 119.023W

(1,881) 53.509N, 134.039W

(1121,881) 45.619N, 59.959W

(1121,1) 19.805N, 80.750W

Pole point

(I,J) (400.500, 1600.500)

References

Anthes, R. A., 1983: Regional models of the atmosphere in middle latitudes. *Mon. Wea. Rev.*, **111**, 1306–1335.

Baldwin, M. E**.**, S. Lakshmivarahan, and J. S. Kain, 2002: Development of an "events-oriented" approach to forecast verification. *Preprints, 15th Conference on Numerical Weather Prediction*, San Antonio, TX, Amer. Meteor. Soc., 210-213.

Baldwin, M. E., and M. S. Wandishin, 2002: Determining the resolved spatial scales of Eta model precipitation forecasts. *Preprints, 15th Conference on Numerical Weather Prediction*, San Antonio, TX, Amer. Meteor. Soc., 85-88.

Baldwin, M. E. and K. L. Elmore, 2005: Objective verification of high-resolution WRF forecasts during 2005 NSSL/SPC Spring Program. *Preprints, 17th Conference on Numerical Weather Prediction*, 1-5 August, Washington, DC, Amer. Meteor. Soc., paper 11B.4

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

Davis, C., B. Brown, and R. 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. and F. Carr, 2000: Summary of the 1998 Workshop on Mesoscale Model Verification. *Bull. Amer. Met. Soc*., **81**, 809-819.

Ebert, E.E. and J. L. McBride, 2000: Verification of precipitation in weather systems: Determination of systematic errors. *J. Hydrol.*, **239**, 179-202.

Harris, D., E. Foufoula-Georgiou, K.K. Droegemeier and J.J. Levit, 2001: Multiscale statistical properties of a high-resolution precipitation forecast. *J. Hydromet*., **2**, 406-418.

Mesinger, F., 1996: Improvements in quantitative precipitation forecasts with the Eta regional model at the National Centers for Environmental Prediction: The 48-km upgrade. *Bull. Amer. Meteor. Soc.*, **77**, 2637–2649.

Nachamkin, J. E., S. Chen, and J. Schmidt. 2005: Evaluation of heavy precipitation forecasts using composite-based methods: A distributions-oriented approach. *Mon. Wea. Rev:*, **133**, 2163–2177.

Olson, D. A., N. W. Junker, and B. Korty, 1995: Evaluation of 33 years of quantitative precipitation forecasting at NMC. *Wea. Forecasting*, **10**, 498–511.

Tustison, B., D. Harris, and E. Foufoula-Georgiou, 2001: Scale issues in verification of precipitation forecasts. *J. Geophys. Res*., **106** (D11), 11,775-11,784.