Comparison of the NCEP and DTC Verification Software Packages

Point of Contact: Michelle Harrold
September 2011

1. Introduction

The National Centers for Environmental Prediction (NCEP) and the Developmental Testbed Center (DTC) have both developed software packages with the capability to verify model forecasts using several verification methods. NCEP employs both the NCEP Verification System (NVS) and the Quantitative Precipitation Forecast (QPF) Verification System (QVS), while the DTC uses the Model Evaluation Tools (MET) verification package. Both verification packages provide a means to generate partial sums, or raw numbers from which many standard statistical results can be computed and used for evaluating model performance. Since the NCEP verification packages and MET are concurrently being used in testing and evaluating efforts within the broad user community, it is important to ascertain the two systems are producing congruous output. Therefore, a study was performed to establish MET produces output functionally similar to the NCEP verification packages.

2. Verification Packages

2.1 NCEP Verification

NCEP performs model verification with NVS, which verifies surface and upper air forecasts, and QVS, which verifies precipitation forecasts. A brief summary of their functions and capabilities is given below.

NVS

NVS (v1.0; Chuang et al. 2004) computes grid-to-point verification and is comprised of three parts: ***editbufr***; ***prepfits***; and ***gridtobs***. The final output of the verification system generates a Verification Statistics Data Base (VSDB) file for each model forecast. The VSDB file is a record of the partial sums computed over a user-defined domain(s) and period of time. The computation and visualization of verification metrics is completed in the Forecast Verification System (FVS), which will not be discussed in this report; for this study, the computation of verification metrics from partial sum values is computed using a program in MET, ***Stat-Analysis***, which will be discussed later in the report.

The ***editbufr*** program reads and retains the observations required for verification from PREPBUFR files, which contain point observations, associated quality mark information, as well as a complete history of the observation (i.e., tracks any quality control processing). The ***prepfits*** program uses bilinear interpolation to interpolate the model forecast data to the observation location. The ***gridtobs*** program, the final step of NVS, generates the partial sums. NVS requires several user-defined inputs, including: grid number defining the mask region or area of retention, a time window of data to be retained, the type and level of observations to be retained, and verification specifications for aggregation (e.g., variable, initialization hour, forecast lead time, vertical level, domain, and precipitation threshold).

QVS

QVS performs grid-to-grid verification and is comprised of several programs. Prior to performing any verification, both the model forecasts and the precipitation analyses must be interpolated to a common, user-specified grid, which is done with ***copygb***, an additional, stand-alone software package also developed by NCEP. The gridded model forecasts are read in and converted to NCEP's regular GRIB format in the ***brkout_fcst*** and ***pcpconform*** programs, respectively. The ***eta_stage4_acc*** program accumulates the model precipitation to the desired increment (e.g., 3 or 24 hours), with the ***diffpcp*** program available to perform

subtraction of files, when necessary. Statistical output is created by the ***verfgen*** program and, like NVS, is in VSDB format. QVS requires several user-defined inputs, including: grid number defining the region over which verification will be performed and verification specifications (e.g., initialization hour, forecast lead time, and domain). Output consists of FHO lines, or lines which provide information regarding the forecast, hit, and observed rates, for specified thresholds. This information can be used to calculate a variety of contingency table statistics (e.g., false alarm ratio and frequency bias). Similar to NVS, the computation and visualization of QPF verification metrics is completed in the FVS.

2.2 DTC Verification

MET

MET (v3.0.1; Fowler et al. 2010) is a highly-configurable verification toolkit that includes the computation of traditional verification statistics as well as several advanced verification techniques such as object-based and neighborhood methods. A full overview of the capabilities and components of MET can be accessed on the DTC website (http://www.dtcenter.org/met/users/). Essentially, the main purpose of the components is to ingest and reformat model forecasts and observational data, from which verification output is computed. MET verification statistics are output in a format from which the data can be easily aggregated, plotted, and analyzed.

For this study, the MET tools primarily used to generate the model verification output included ***PB2NC***, ***Point-Stat***, and ***Grid-Stat***. The ***PB2NC*** tool was used to create a file containing point observations from the PREPBUFR files to be used in the verification process. ***Point-Stat*** was used in verifying surface and upper air model output (verifying gridded model output against a point observation), while ***Grid-Stat*** was used for verifying the precipitation forecasts (verifying gridded model output against a gridded observational field.) The ***Point-Stat*** configuration file specifies a host of fields, levels, and masking regions to be verified in addition to interpolation method and type of statistical information to output. The ***Grid-Stat*** configuration file contains similar information to the ***Point-Stat*** file; in addition, the user can define precipitation thresholds to verify. In essence, ***PB2NC*** and ***Point-Stat*** perform similar functions to NVS, while ***Grid-Stat*** is similar to QVS. One difference, however, is that ***Point-Stat*** and ***Grid-Stat*** offer the ability to output several types of statistical data (e.g., in addition to partial sums, contingency table counts and continuous statistics are also available).

3. Methodology of Comparing Verification Packages

In 2007, the DTC performed an evaluation of forecasts run with the two dynamic cores of the Weather Research and Forecasting (WRF) model. The model verification for the Core Test (hereafter, CT2007) was done using NVS (v1.0) and the QVS. A complete overview of the experiment, including the datasets that were used is provided in Bernardet et al. (2008).

For this study, the inter-comparison of the verification systems was accomplished by running MET v3.0.1 with a subset of the data used in CT2007. In order to replicate the user-defined specifications within the NCEP verification packages, the default MET configuration files were modified prior to running the CT2007 data through the system. MET output was then compared against the same subset of statistical output from CT2007, generated with NVS and QVS. The specific subset of data used for this comparison consisted of 30 00 UTC initializations and 29 12 UTC initializations, spanning all seasons. Model output files used for this analysis were generated using the Non-hydrostatic Mesoscale Model core of the WRF model.

As noted earlier, output from both the NVS and QVS only contain partial sums (i.e., scalar partial sums (SL1L2 lines), vector partial sums (VL1L2 lines), and FHO lines). The ***Stat-Analysis*** tool available in MET was used to convert the SL1L2 and FHO lines output from the NCEP verification packages to continuous and contingency table statistics, respectively. Due to the format of the verification output from CT2007, the VL1L2 lines could not be converted to wind vector statistics. The inter-comparison of the verification

systems performed here was accomplished by assessing the consistency between select continuous and contingency table statistics as well as the direct output from the VL1L2 lines.

Verification was performed over several domains; surface and upper air variables were verified over the contiguous United States (CONUS), Western CONUS, and Eastern CONUS (Fig. 1), while precipitation verification was computed over the CONUS and 14 sub-domains (Fig. 2). Due to the similarity of results between the CONUS domain and the individual sub-domains, for both surface and upper air variables, only results for the CONUS domain will be discussed in this report. Upper air verification statistics were stratified by vertical level and lead time for the combined set of 00 and 12 UTC initializations. Surface fields were stratified by both forecast lead time and precipitation thresholds for the 00 UTC initializations and the 12 UTC initializations separately.

Confidence intervals, at the 99% level, accompany each of the verification metrics that were computed by both verification packages. Since MET and NCEP verification packages used the same cases, a pair-wise difference methodology was applied, where appropriate. Using the statistical output generated by the verification systems, for every individual forecast lead time and for the metrics that will be discussed in detail in Section 4, differences between the two verification metrics were computed by subtracting the values generated from the NCEP verification packages by the values generated from MET. From these differences, both the median value and the magnitude of the maximum difference were calculated. CIs were computed for the median of the pair-wise differences, allowing for an objective determination of whether the differences between the two verification packages were statistically significant (SS). If the CIs of the pair-wise differences encompass zero, the differences are not considered SS. The pair-wise difference methodology is not applied to frequency bias due to the non-linear attributes of the metric. For frequency bias, differences were deemed SS if there was no overlap of the CIs between MET and QVS. Evaluation of the maximum differences between the two systems was done to expose any outliers not seen in the median pair-wise difference values and the associated CIs.

A number of SS pair-wise differences were anticipated, due both to the size of the dataset and the inherent introduction of rounding errors from running the verification systems. These differences, however, may not be practically meaningful when distinguishing discrepancies between MET and NVS. Therefore, practical significance (PS) was established by censoring the data to only consider the SS pair-wise differences that were greater than the operational measurement uncertainty requirements and instrument performance as defined by the World Meteorological Organization (WMO). As per the WMO, the following criteria was used to establish PS: a) temperature differences greater than 0.1 K, b) relative humidity (RH) differences greater than 1.0%, and c) wind speed differences greater than 0.5 m s^{-1}. PS was not established for QPF due to the nature of the verification metrics used in this comparison (i.e., skill scores).

Verification output from both the MET and the NCEP verification packages was loaded into a MySQL database where aggregated statistics were computed and plotted using the statistical programming language, R.

3.1 Surface and Upper Air

Surface and upper air forecasts of temperature, RH, and mean components of the vector winds were bilinearly interpolated to the location of the point observations within the PREPBUFR file. Verification of the surface variables commenced at the 3-h lead time and extended out to 60 h, in 3-h intervals. For the upper air, verification was computed for temperature and vector wind at all mandatory levels, with exception of the 1000 and 100 hPa levels; verification for upper air RH was done exclusively at 850, 700, and 500 hPa due to low reliability of RH observations above 500 hPa. The verification for upper air variables was computed in 12-h intervals (i.e., times valid at 00 and 12 UTC, due to availability of radiosonde data) out to 60 h.

During the testing and evaluation of the verification packages, it was discovered that MET and NVS use

different approaches for calculating RH, which is not a directly observed quantity. Both packages use the forecast values of RH directly from the post-processed model output; however, different methods are used when calculating the observed values, which leads to an inherent, systematic error. While both methods of calculating RH are correct, the equations and assumptions that are used by each package lead to differences in this comparison.

To evaluate the temperature and RH variables, this study used mean error (i.e., bias) and bias-corrected root-mean-square-error (BCRMSE). For evaluating the vector wind, due to the format of the original CT2007 statistical output, only a direct comparison with VL1L2 lines can be accomplished; therefore, the mean forecast and observed values of the U- and V-component of the vector wind are used in this evaluation. Pair-wise differences were computed and CIs were applied for each metric by computing the standard error estimates about the median. Maximum differences between the two systems were calculated for temperature and wind; differences were not computed for RH, due to the differing RH calculations that MET and NVS utilize.

3.2 Precipitation

Accumulated precipitation verification was done over 3- and 24-h periods. Both the model forecasts and the analysis fields were interpolated to NCEP Grid 218, a 12-km grid over the CONUS. After interpolating, precipitation verification was done by using a grid-to-grid comparison. The observational dataset used for the 3-h accumulations was NCEP Stage II analysis; Climate Prediction Center (CPC) analyses were used for the 24-h accumulations. CPC analyses are valid at 12 UTC; therefore, verification for 00 UTC initializations were done for the 36- and 60-h lead times, while 12 UTC initializations were verified at the 24- and 48 hour lead times.

For comparing output from the two verification packages, this study focuses on the traditional verification metrics of frequency bias and Gilbert Skill Score (GSS). Pair-wise differences were computed for GSS only, and CIs were computed for each metric using a bootstrapping method for both the 3- and 24-h QPF.

4. Results

The motivation behind this study was to evaluate the median pair-wise differences and associated CIs and the maximum pair-wise differences in order to assess the similarity of the verification systems.

4.1 Upper Air

Temperature

For temperature bias (Fig. 3) and BCRMSE (Fig. 4), MET and NVS produce similar verification results for the entire set of cases over the CONUS domain (only the 36-h lead time shown). This result is consistent for all lead times and at all vertical levels. For bias, there are no SS pair-wise differences seen at any forecast lead time or vertical level. Only one SS pair-wise difference is noted for BCRMSE (at the 36-h lead time at 200 hPa); however, it is not PS. Table 1 provides the maximum differences for bias and BCRMSE for all lead times and vertical levels. The largest difference is 0.5 K, and there is no distinct signal in relation to lead time or vertical level, showing the verification packages are essentially providing congruous output.

Relative humidity

Due to the different methods by which RH is derived in MET and NVS, as described earlier, there are obvious, systematic differences that are seen when evaluating bias (Fig. 5). MET consistently produces lower median bias values than NVS, signaling a systematic difference between the two packages; in general, this trend is seen at all lead times and vertical levels. While most lead times and vertical levels have SS pair-wise differences, none are PS (only the 36-h lead time is shown). When evaluating BCRMSE, the two verification systems produce similar results (Fig. 6); there are no PS pair-wise differences. These results are consistent for all lead times and at all vertical levels.

Vector wind

The median of the mean forecast and observed values of the U- and V-component of the vector wind are similar for both MET and NVS at all lead times and vertical levels (e.g., Figs. 7-8). For the wind vector variables evaluated in the report, there are no SS pair-wise differences (not all shown). A majority of the maximum differences between the MET and NCEP verification packages are below 1.0 m s^{-1} (Table 2). Those that are greater than 1.0 m s^{-1} typically occur between 300 – 200 hPa or at later lead times for the U-component of the vector wind.

4.2 Surface

Temperature

For temperature bias (Fig. 9a) and BCRMSE (Fig. 10), MET and NVS produce similar verification results for the entire set of cases over the CONUS domain; results are consistent for both the 00 and 12 UTC initializations and at all lead times. While temperature bias has SS pair-wise differences at most lead times valid between 12 and 18 UTC, none are PS (only the 00 UTC initialization is shown). Temperature BCRMSE also had no PS pair-wise differences. In fact, all maximum differences between MET and NVS, for both metrics and initialization times are less than 0.1 K (Table 3). A box plot of the range of maximum differences can provide graphical support to the information in the table; Fig. 9b shows that the largest outliers for temperature bias are below 0.05 K.

Relative humidity

Similar to the upper air verification, for surface RH (Figs. 11 – 12), there is a systematic difference when evaluating the verification metrics produced by the two systems. Even when considering this difference, both bias and BCRMSE values for both the 00 and 12 UTC initialization times show consistency between the two verification packages. For BCRMSE, a majority of lead times for both the 00 and 12 UTC initialization times have SS pair-wise differences (only the 00 UTC initialization is shown); however, none are PS.

Vector wind

The median of the mean forecast and observed values of the U- and V-component of the vector wind for MET and NVS show consistency between the two packages at both the 00 and 12 UTC initializations and at all lead times (e.g., Figs. 13-14). For all wind vector variables evaluated, there are no SS pair-wise differences (only the 00 UTC initialization is shown). Table 4 illustrates that any differences between the two systems are minimal; the maximum differences between MET and NVS considering all wind vector variables is about 0.1 m s^{-1}.

3-hr QPF

In general, for both GSS (Fig. 15) and frequency bias (Fig. 16), MET and QVS produce similar results. This is consistent for all initializations and lead times. QVS does produce GSS values that are SS greater than MET for both 00 and 12 UTC initializations and at all lead times at thresholds of 0.05" and below. A majority of lead times for both the 00 and 12 UTC initializations at the 1" threshold are also SS, with QVS producing greater GSS values than MET (only the 00 UTC initialization is shown). While there are a number of SS pair-wise differences, the largest maximum difference GSS between the two verification packages is 0.021 (Table 5). There is no distinct pattern in the maximum differences in relation to initialization, lead time, or threshold. For frequency bias, there are no SS pair-wise differences between MET and QVS; however, there are several large maximum differences (greater than 10), most notably at the higher thresholds. These differences were further investigated and were found to occur in situations where there were a small number of hits (less than ~5 grid points over a domain with approximately 73000 grid points) and misses (less than ~10 grid points). Due to the low coverage area of observed events, a small discrepancy in contingency table counts between the two verification system can cause large differences in the calculated frequency bias.

24-h QPF
Figures 17-18 illustrate the ability of MET to reproduce 24-h QPF verification results similar to QVS. These findings hold true for both 00 and 12 UTC initializations and for all lead times. Similar to the 3-h QPF, there are several SS pair-wise differences at the lower thresholds, with QVS producing SS higher GSS values than MET (only the 36-h lead time for the 00 UTC initialization is shown). However, the maximum differences for 24-h QPF GSS show most differences are minimal (Table 6). For frequency bias, there are no SS pair-wise differences between MET and QVS. As with the 3-h QPF, there are several larger maximum differences (i.e., greater than 15), and these are seen at thresholds of 1.5" and greater; these diffr

5. Summary

Both NCEP and the DTC provide software to verify model forecasts; it is advantageous to compare the verification output from both community verification systems to ensure the packages are producing functionally similar results. By using a subset of the CT2007 data, this evaluation successfully demonstrated the ability of MET to reproduce verification metrics similar to the metrics generated by the NCEP verification packages. Consistency between the two verification systems was observed for both upper air and surface verification statistics, stratified over different domains, initialization hour, forecast lead time, and precipitation thresholds. Consistency was demonstrated by assessing statistical and practical significance as well as the magnitude of the maximum differences between the two verification systems. For temperature, RH, and vector winds, there were no PS pair-wise differences between MET and NVS. When examining GSS for 3- and 24-h QPF, there were several SS pair-wise differences at lower thresholds, but the magnitudes of the maximum differences between the two systems are small in these instances. No SS pair-wise differences were noted for 3- and 24-h frequency bias. With the exception of frequency bias at higher thresholds for 3- and 24-h QPF, maximum differences for the calculated metrics display minimal deviations between the two systems.

6. Resulting Changes in MET

Several modifications were implemented in METv3.0.1 as a result of this testing and evaluating effort. A brief overview is given below.

PB2NC bug fix
A bug was discovered in the PB2NC tool, which is used to create a NetCDF file from a PRPBUFR file. The bug, which has been fixed, caused observations to be sorted incorrectly by time, keeping the oldest, least recent observation value.

Bilinear interpolation
METv3.0.1 now offers the opportunity to use bilinear interpolation.

Radius of the earth
Previous versions of MET used 6367.47 km as the radius of the earth, which is defined in "NCEP Office Note 388: GRIB specifications;" however, the NCEP w3 library, which contains utilities to encode and decode GRIB1 data, uses 6371.20 km as the radius of the earth. Due to the fact that WRF and port-processors (e.g., WRF Post-Processor (WPP) and Universal Post-Processor (UPP)) use the w3 libraries, METv3.0.1 was modified to use the same value for the radius of the earth as WPP and UPP. While the change appears to be small, there are apparent differences when converting latitudes and longitudes to x and y values.

Relative humidity calculation
It was noted that MET and NVS use differing calculations in the calculation of RH. Moisture variables (e.g.,

RH and dew point) are not often directly measured, and, therefore, need to be mathematically calculated. While both MET and NVS employ two different means of calculating RH, both are accepted ways. Due to the fact that a majority of MET users post-process data with WPP and UPP, the MET RH calculation was modified to be consistent with them.

7. References

Bernardet and coauthors, 2008: The Developmental Testbed Center 2007 Dynamic Core Test Report. Available online: http://verif.rap.ucar.edu/eval/ext_ct/CT2007_Report_Update.pdf

Chuang, H-Y, G. DiMego, M. Baldwin, and WRF DTC team, 2004: NCEP's WRF Post-processor and verification systems. *5th WRF / 14th MM5 Users' Workshop*, Boulder, CO. Available online: http://www.mmm.ucar.edu/mm5/workshop/ws04/Session7/Chuang.Hui-Ya.pdf

Fowler, T. L., T. Jensen, E. I. Tollerud, J. Halley Gotway, P. Oldenburg, R. Bullock, 2010: New Model Evaluation Tools (MET) Software Capabilities for QPF Verification. *Preprints*, 3rd Intl. Conf. on QPE, QPF 18-22 October 2010.

Table 1. Maximum difference between the MET and NCEP verification packages for upper air temperature (K) BCRMSE and bias by pressure level and forecast lead time for the 00 UTC and 12 UTC initializations combined over the full integration domain and over all the cases.

		f12	f24	f36	f48	f60
BCRMSE	850	0.069	0.059	0.147	0.163	0.101
	700	0.126	0.064	0.104	0.056	0.156
	500	0.099	0.079	0.112	0.094	0.090
	400	0.095	0.082	0.126	0.067	0.251
	300	0.098	0.089	0.106	0.111	0.125
	250	0.096	0.076	0.095	0.061	0.136
	200	0.148	0.108	0.106	0.059	0.540
	150	0.060	0.101	0.073	0.059	0.055
Bias	850	0.041	0.068	0.104	0.120	0.070
	700	0.047	0.043	0.063	0.046	0.202
	500	0.046	0.043	0.055	0.055	0.063
	400	0.050	0.025	0.074	0.056	0.139
	300	0.035	0.054	0.051	0.058	0.176
	250	0.045	0.059	0.060	0.055	0.101
	200	0.064	0.067	0.063	0.060	0.544
	150	0.038	0.055	0.053	0.046	0.145

Table 2. Maximum difference between the MET and NCEP verification packages for the mean forecast (F) and observed (O) U-component (U) and V-component (V) of the upper air wind (m s^{-1}) by pressure level and forecast lead time for the 00 UTC and 12 UTC initializations combined over the full integration domain and over the entire set of cases.

	UF					UO					VF					VO				
	f12	**f24**	**f36**	**f48**	**f60**	**f12**	**f24**	**f36**	**f48**	**f60**	**f12**	**f24**	**f36**	**f48**	**f60**	**f12**	**f24**	**f36**	**f48**	**f60**
850	0.600	0.313	0.433	0.309	0.793	0.614	0.304	0.378	0.400	1.218	0.312	0.690	0.330	0.346	2.4342	0.408	0.701	0.410	0.310	0.376
700	0.480	0.443	0.494	0.256	0.795	0.359	0.509	0.437	0.300	1.386	0.428	0.511	0.394	0.407	0.7418	0.371	0.640	0.345	0.370	0.576
500	0.593	0.458	0.613	0.395	1.761	0.510	0.513	0.704	0.455	1.915	0.514	0.338	0.513	0.394	0.6352	0.582	0.522	0.507	0.439	0.522
400	0.682	0.665	0.568	0.647	1.194	0.640	0.682	0.471	0.635	1.261	0.907	0.433	0.627	0.457	1.0550	0.896	0.783	0.811	0.650	0.804
300	1.207	0.909	0.863	0.718	2.659	1.314	0.882	1.110	0.811	2.150	0.866	1.505	0.884	1.118	0.934	1.073	1.346	1.049	1.073	1.057
250	1.386	0.902	0.839	1.007	2.695	1.350	1.032	0.965	1.016	2.731	1.096	1.019	1.005	0.889	1.272	1.156	1.086	1.446	0.884	1.441
200	1.206	1.219	0.717	1.041	3.697	1.015	1.149	0.962	1.035	4.093	1.200	0.587	0.897	0.726	1.985	1.199	0.600	0.937	1.108	0.853
150	1.249	0.852	0.843	0.690	2.697	0.986	0.882	0.686	0.869	3.121	0.488	0.574	0.568	0.533	0.840	0.689	0.712	0.662	0.689	0.833

Table 3. Maximum difference between the MET and NCEP verification packages for surface temperature (K) BCRMSE and bias by forecast lead time for the 00 UTC and 12 UTC initializations separately over the full integration domain and over the entire set of cases.

		f03	f06	f09	f12	f15	f18	f21	f24	f27	f30	f33	f36	f39	f42	f45	f48	f51	f54	f57	f60
BRCMSE	00 UTC	0.039	0.059	0.027	0.047	0.041	0.149	0.043	.0362	0.032	0.028	0.035	0.031	0.054	0.026	0.037	0.026	0.029	0.026	0.026	0.028
	12 UTC	0.043	0.028	0.037	0.029	0.032	0.035	0.024	0.029	0.041	0.023	0.033	0.042	0.035	0.041	0.045	0.031	0.025	0.033	0.029	0.073
Bias	00 UTC	0.024	0.033	0.022	0.014	0.022	0.016	0.016	0.018	0.016	0.021	0.024	0.016	0.018	0.040	0.015	0.011	0.015	0.022	0.020	0.014
	12 UTC	0.021	0.015	0.014	0.017	0.013	0.025	0.018	0.021	0.021	0.019	0.048	0.042	0.037	0.031	0.035	0.032	0.026	0.021	0.026	0.050

Table 4. Maximum difference between the MET and NCEP verification packages for the mean forecast (F) and observed (O) U-component (U) and V-component (V) of the wind (m s^{-1}) by forecast lead time for the 00 UTC and 12 UTC initializations separately over the full integration domain and over the entire set of cases.

		f03	f06	f09	f12	f15	f18	f21	f24	f27	f30	f33	f36	f39	f42	f45	f48	f51	f54	f57	f60
00 UTC	UO	0.044	0.051	0.045	0.049	0.056	0.052	0.071	0.072	0.062	0.034	0.037	0.043	0.051	0.060	0.058	0.042	0.036	0.045	0.043	0.034
	UF	0.058	0.065	0.063	0.058	0.061	0.067	0.089	0.099	0.083	0.056	0.054	0.060	0.056	0.067	0.044	0.041	0.046	0.049	0.062	0.035
	VO	0.041	0.036	0.032	0.032	0.044	0.040	0.038	0.047	0.051	0.048	0.029	0.033	0.033	0.041	0.049	0.036	0.033	0.049	0.033	0.032
	VF	0.067	0.042	0.041	0.033	0.039	0.044	0.047	0.092	0.079	0.072	0.051	0.043	0.041	0.049	0.059	0.044	0.038	0.054	0.040	0.035
12 UTC	UO	0.051	0.060	0.035	0.042	0.036	0.050	0.043	0.034	0.038	0.041	0.039	0.037	0.039	0.046	0.034	0.031	0.036	0.044	0.100	0.058
	UF	0.061	0.073	0.048	0.063	0.051	0.060	0.059	0.043	0.047	0.047	0.056	0.057	0.063	0.064	0.043	0.039	0.051	0.070	0.087	0.052
	VO	0.038	0.041	0.049	0.074	0.048	0.058	0.033	0.035	0.038	0.048	0.043	0.034	0.035	0.034	0.032	0.032	0.044	0.040	0.056	0.070
	VF	0.045	0.047	0.058	0.062	0.059	0.056	0.045	0.047	0.054	0.060	0.053	0.049	0.065	0.054	0.041	0.038	0.058	0.058	0.055	0.047

Table 5. Maximum difference between the MET and NCEP verification packages for 3-hour QPF GSS and frequency bias by forecast lead time and threshold for the 00 UTC and 12 UTC initializations separately over the full integration domain and over the entire set of cases.

		00 UTC Initializations									12 UTC Initializations								
		>0.01	**>0.02**	**>0.05**	**>0.1**	**>0.15**	**>0.25**	**>0.35**	**>0.5**	**>1**	**>0.01**	**>0.02**	**>0.05**	**>0.1**	**>0.15**	**>0.25**	**>0.35**	**>0.5**	**>1**
GSS	**f12**	0.010	0.007	0.006	0.005	0.009	0.011	0.007	0.004	0.000	0.012	0.009	0.007	0.005	0.006	0.012	0.015	0.009	0.023
	f24	0.015	0.011	0.006	0.019	0.019	0.011	0.009	0.021	0.009	0.008	0.008	0.008	0.005	0.007	0.007	0.010	0.005	0.002
	f36	0.017	0.013	0.012	0.012	0.005	0.004	0.011	0.002	0.000	0.012	0.008	0.006	0.006	0.005	0.009	0.010	0.018	0.000
	f48	0.007	0.007	0.007	0.007	0.050	0.012	0.007	0.009	0.016	0.013	0.008	0.005	0.007	0.005	0.008	0.011	0.014	0.000
	f60	0.013	0.011	0.010	0.007	0.007	0.005	0.013	0.002	0.000	0.016	0.012	0.004	0.008	0.008	0.009	0.014	0.005	0.014
Frequency Bias	**f12**	0.071	0.112	0.158	0.513	0.713	2.627	8.400	2.036	0.844	0.095	0.104	0.125	0.874	1.192	0.800	0.500	0.725	23.333
	f24	0.095	0.204	0.267	0.085	0.215	0.100	0.412	0.694	6.300	0.106	0.117	0.224	2.780	3.970	4.000	0.763	2.512	3.020
	f36	0.079	0.093	0.119	0.306	0.702	1.938	16.033	16.875	5.489	0.102	0.111	0.192	0.821	0.929	1.133	10.459	0.774	15.867
	f48	0.107	0.110	0.086	0.067	0.359	1.375	10.500	0.213	20.167	0.106	0.114	0.152	0.611	1.527	8.300	0.517	0.550	1.500
	f60	0.041	0.046	0.049	0.120	0.092	0.273	0.880	34.333	0.857	0.094	0.093	0.115	0.087	0.188	4.119	0.758	0.393	0.473

Table 6. Maximum difference between the MET and NCEP verification packages for 24-hour QPF GSS and frequency bias by forecast lead time and threshold for the 00 UTC and 12 UTC initializations separately over the full integration domain and over the entire set of cases.

		GSS									Frequency Bias								
		>0.01	**>0.1**	**>0.25**	**>0.50**	**>0.75**	**>1.00**	**>1.50**	**>2.00**	**>3.00**	**>0.01**	**>0.1**	**>0.25**	**>0.50**	**>0.75**	**>1.00**	**>1.50**	**>2.00**	**>3.00**
00 UTC	**f36**	0.005	0.008	0.008	0.015	0.010	0.025	0.022	0.017	0.014	0.017	0.009	0.030	0.133	3.916	4.167	5.589	34.000	1.833
	f60	0.005	0.006	0.006	0.018	0.008	0.006	0.005	0.002	0.017	0.036	0.386	3.000	0.088	0.170	0.760	15.056	21.500	29.500
12 UTC	**f24**	0.006	0.008	0.010	0.012	0.016	0.005	0.028	0.004	0.012	0.043	0.022	0.104	0.029	0.477	0.896	5.000	4.000	9.500
	f48	0.005	0.005	0.005	0.011	0.014	0.008	0.021	0.007	0.003	0.023	0.010	0.179	0.087	0.910	0.920	4.262	0.683	0.600

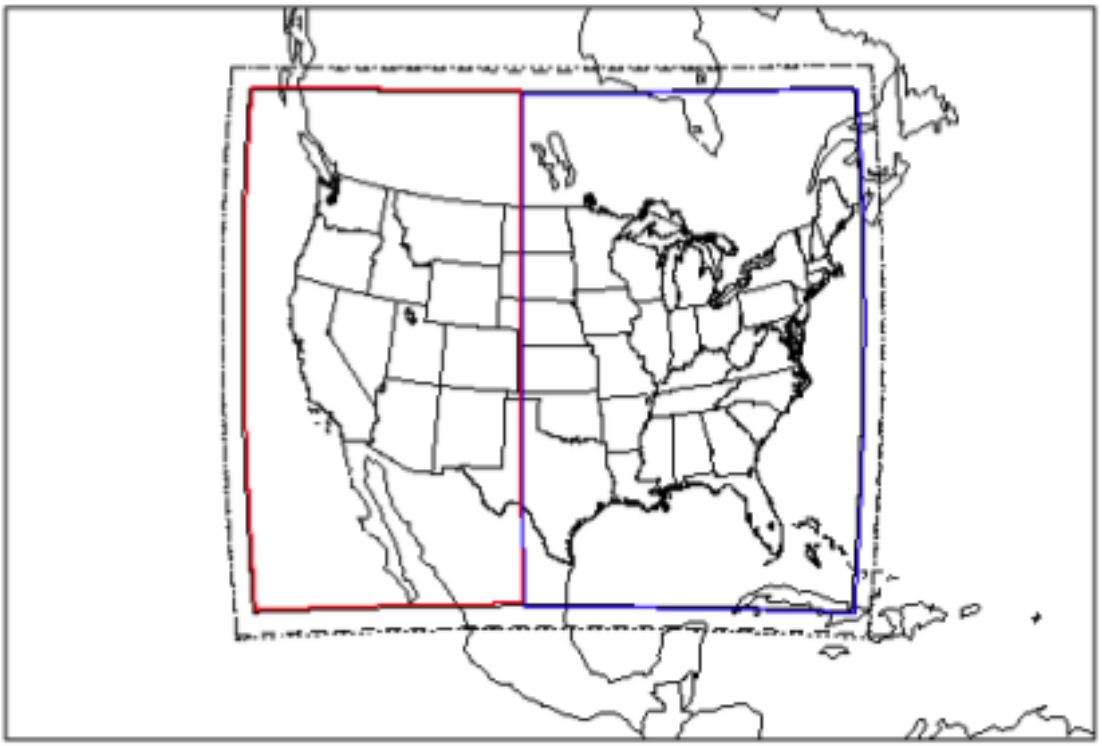

Figure 1. Map showing the CONUS, (solid black line), Western, (red line) and Eastern (blue line) domains.

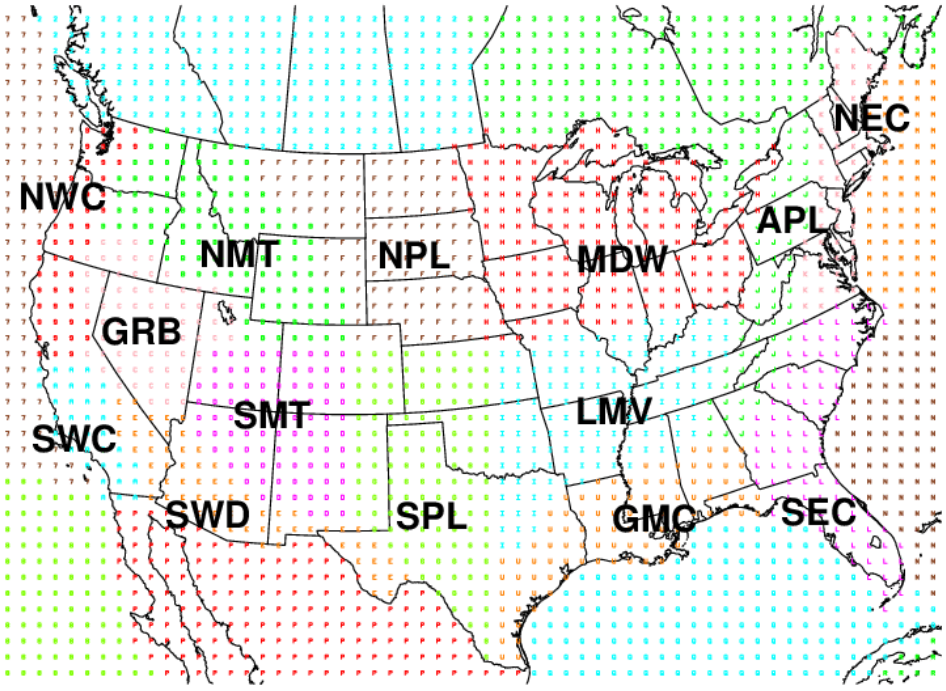

Figure 2. Map showing the domain for QPF verification; the full CONUS domain consists of the 14 sub-domains.

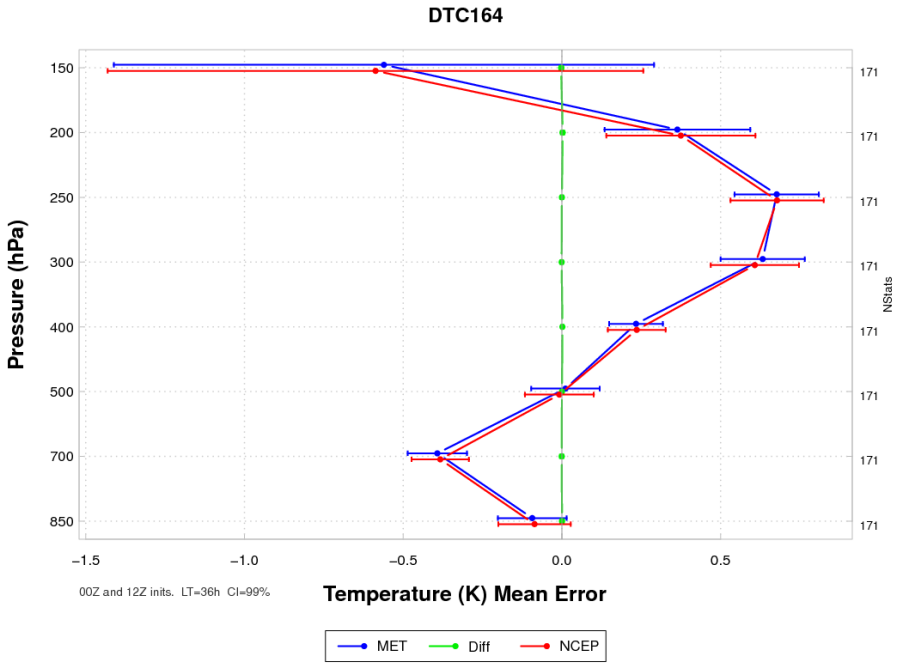

Figure 3. Vertical profile of the median bias for temperature (K) for the full domain (DTC164) for the 36-h lead time aggregated over the entire year of cases. MET is shown in blue, NCEP is shown in red, and the differences (MET-NCEP) are shown in green. The horizontal bars represent the CIs at the 99% confidence level.

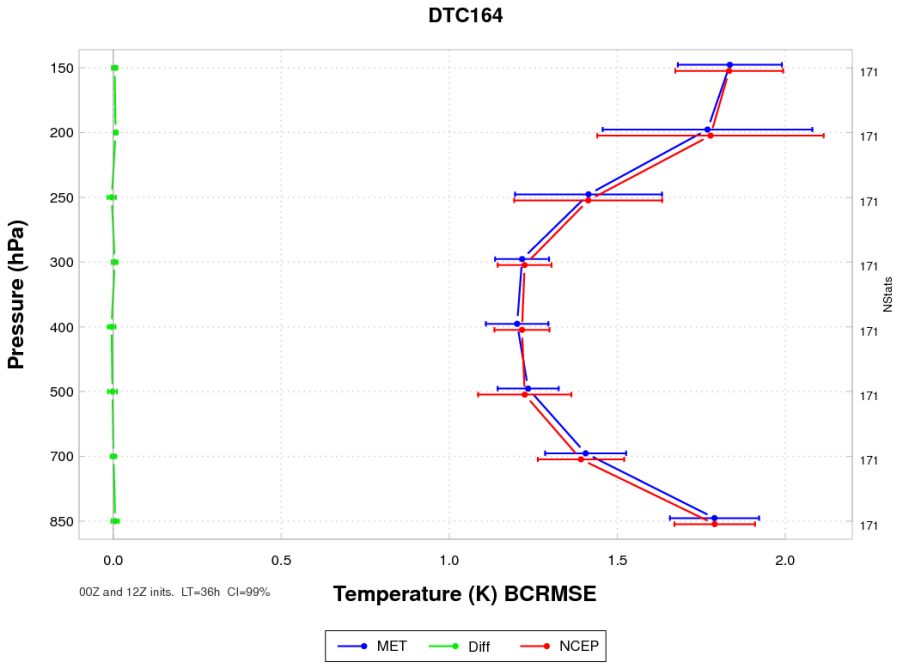

Figure 4. Vertical profile of the median BCRMSE for temperature (K) for the full domain (DTC164) for the 36-h lead time aggregated over the entire year of cases. MET is shown in blue, NCEP is shown in red, and

the differences (MET-NCEP) are shown in green. The horizontal bars represent the CIs at the 99% confidence level.

Figure 5. Vertical profile of the median bias for relative humidity (%) for the full domain (DTC164) for the 36-h lead time aggregated over the entire year of cases. MET is shown in blue, NCEP is shown in red, and the differences (MET-NCEP) are shown in green. The horizontal bars represent the CIs at the 99% confidence level.

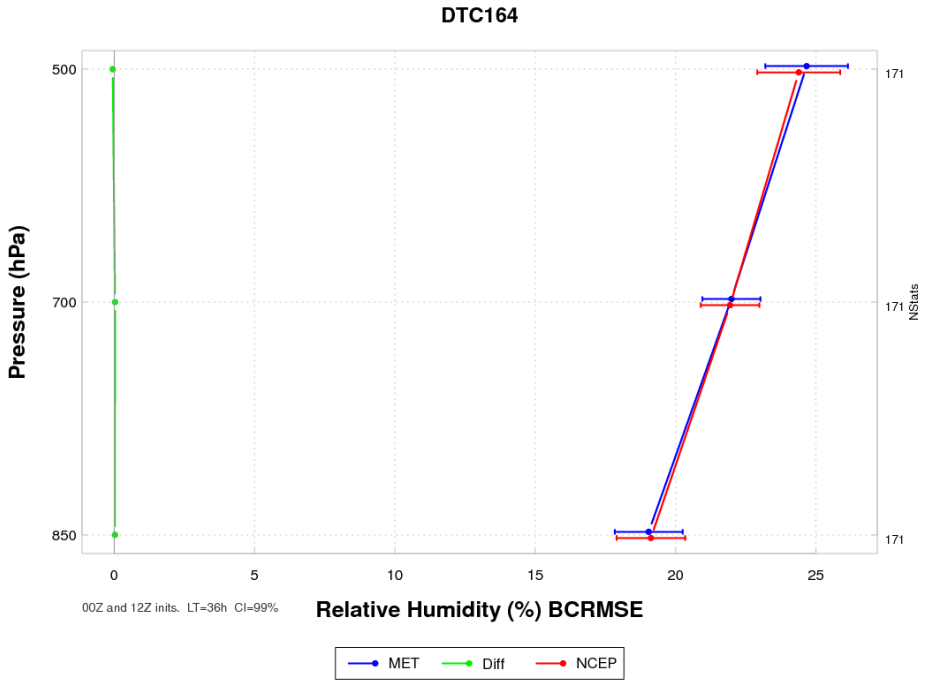

Figure 6. Vertical profile of the median BCRMSE for relative humidity (%) for the full domain (DTC164) for the 36-h lead time aggregated over the entire year of cases. MET is shown in blue, NCEP is shown in red, and the differences (MET-NCEP) are shown in green. The horizontal bars represent the CIs at the 99% confidence level.

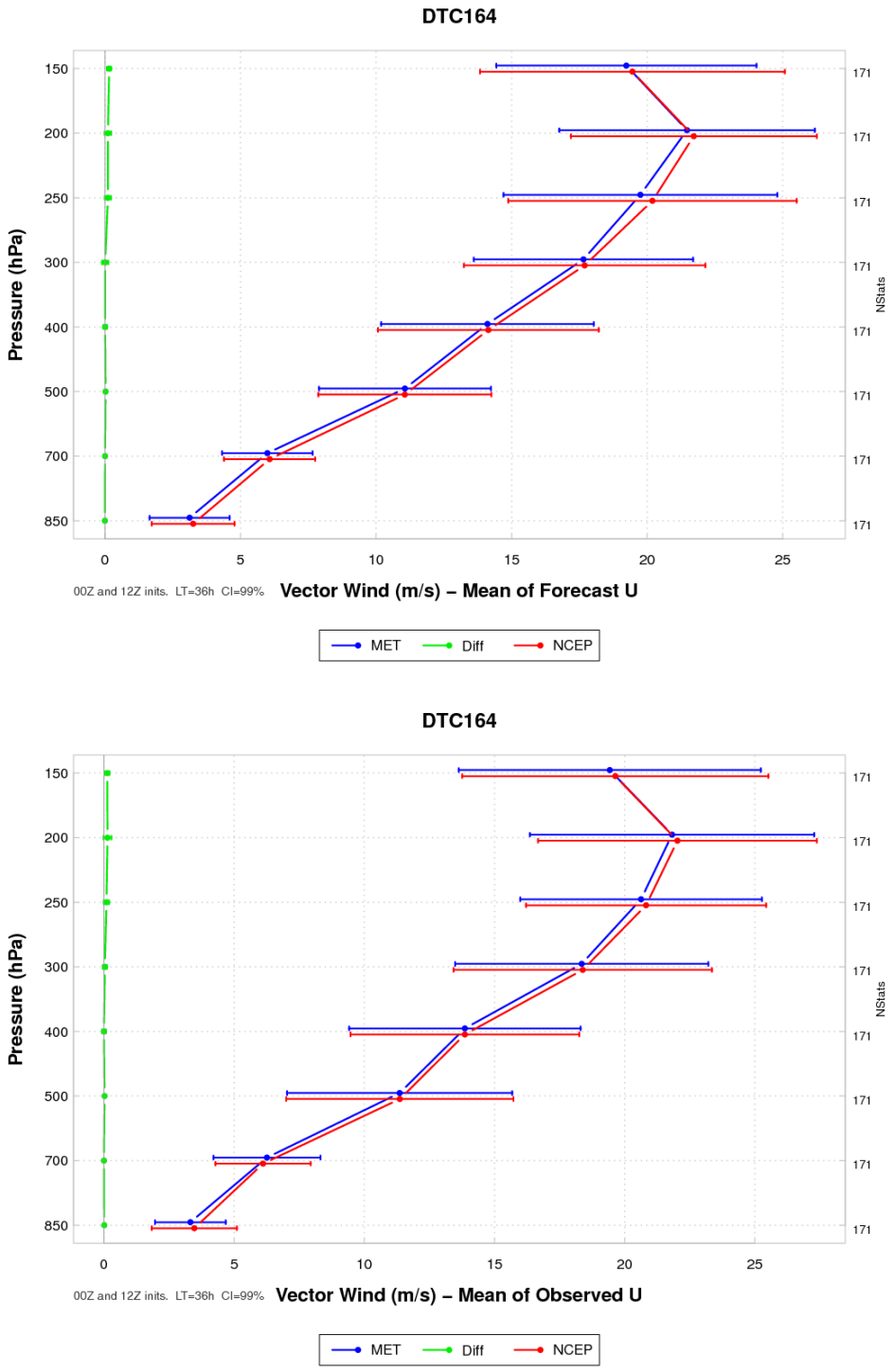

Figure 7. Vertical profile of the median mean of the (a) forecast and (b) observed U-component of the vector wind for the full domain (DTC164) for the 36-h lead time aggregated over the entire year of cases. MET is shown in blue, NCEP is shown in red, and the differences (MET-NCEP) are shown in green. The horizontal bars represent the CIs at the 99% confidence level.

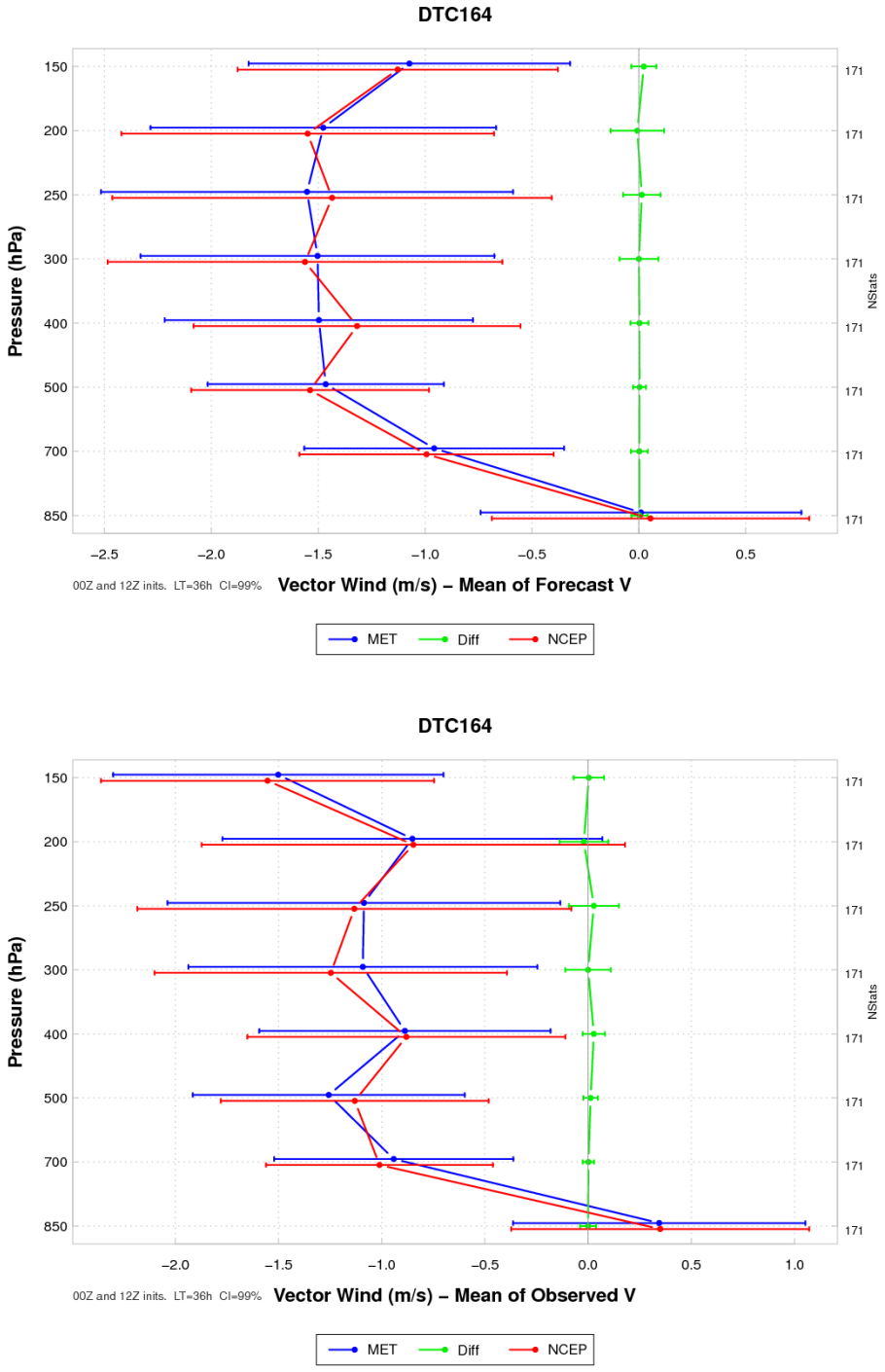

Figure 8. Vertical profile of the median mean of the (a) forecast and (b) observed V-component of the vector wind for the full domain (DTC164) for the 36-h lead time aggregated over the entire year of cases. MET is shown in blue, NCEP is shown in red, and the differences (MET-NCEP) are shown in green. The horizontal bars represent the CIs at the 99% confidence level.

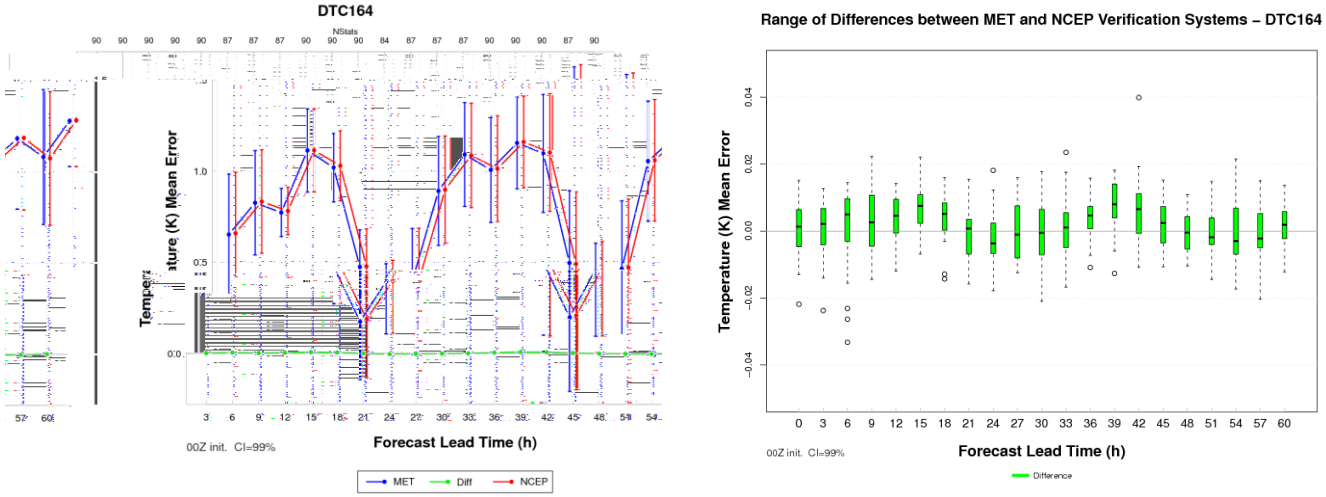

Figure 9. Time series plot of the 2 m AGL temperature (K) for (a) median bias and aggregated across all of the cases over the full domain (DTC164) and (b) the range of maximum differences between MET and NVS for the 00 UTC initializations. In (a), MET is shown in blue, NCEP is shown in red, and the differences (MET-NCEP) are shown in green. The vertical bars represent the CIs at the 99% confidence level. In (b), the horizontal black lines denote the median value, the top and bottom of the box correspond to the 25th and 75th percentile, respectively, and the black circles denote the outliers.

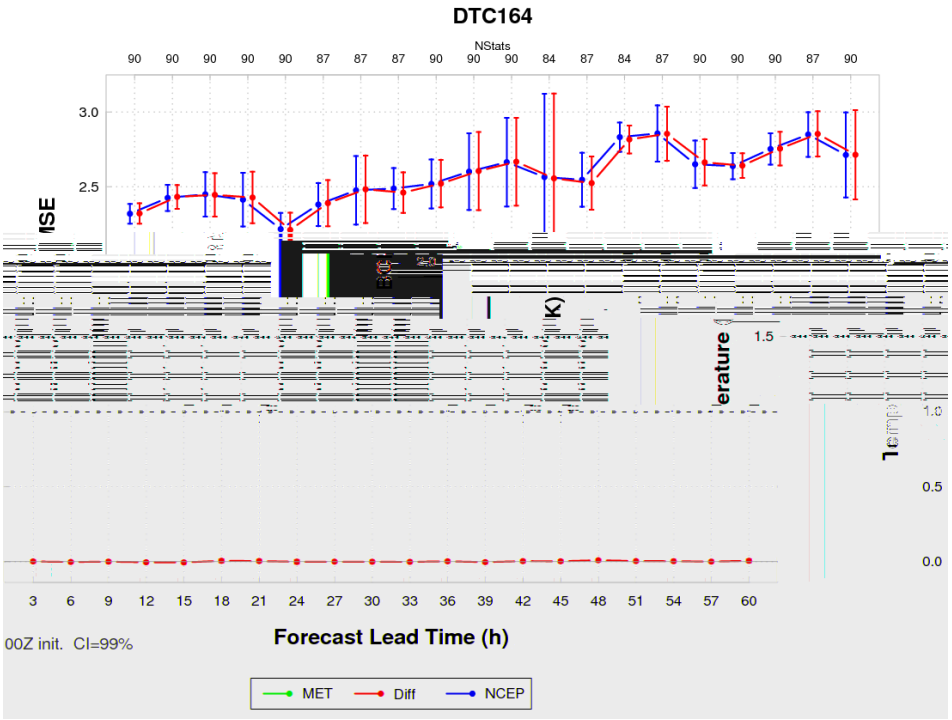

Figure 10. Time series plot of the 2 m AGL temperature (K) for median BCRMSE aggregated across all of the cases over the full domain (DTC164) for the 00 UTC initialization. MET is shown in blue, NCEP is shown in red, and the differences (MET-NCEP) are shown in green. The vertical bars represent the CIs at the 99% confidence level.

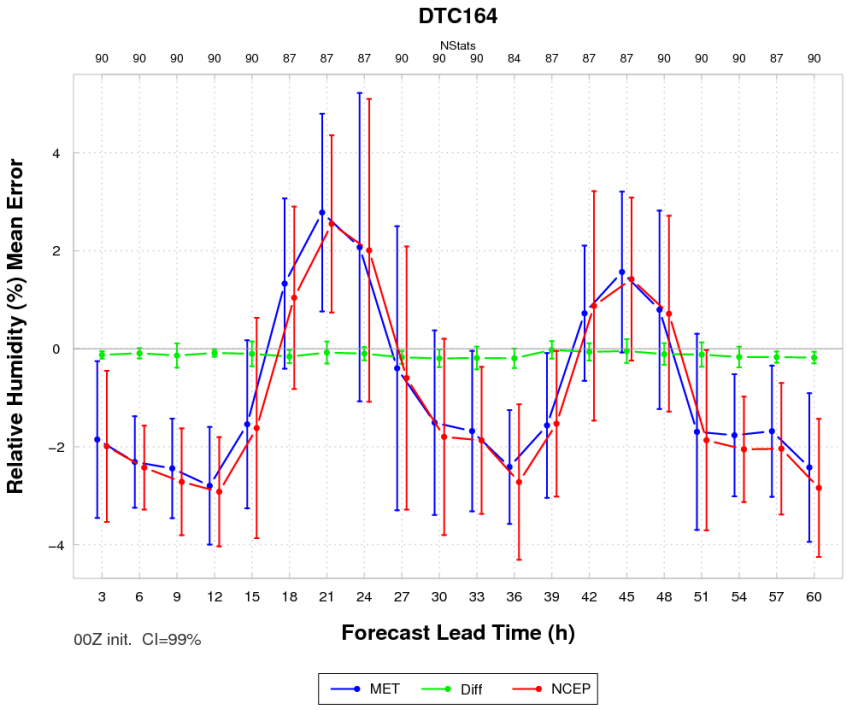

Figure 11. Time series plot of the 2 m relative humidity (%) for median bias aggregated across all of the cases over the full domain (DTC164) for the 00 UTC initialization. MET is shown in blue, NCEP is shown in red, and the differences (MET-NCEP) are shown in green. The vertical bars represent the CIs at the 99% confidence level.

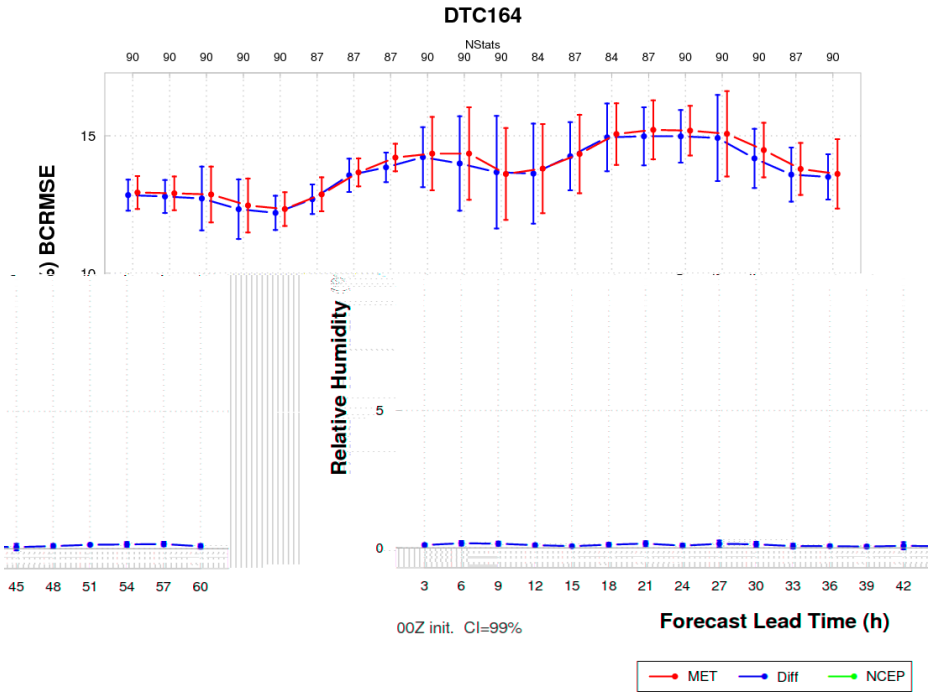

Figure 12. Time series plot of the 2 m relative humidity (%) for median BCRMSE aggregated across all of the cases over the full domain (DTC164) for the 00 UTC initialization. MET is shown in blue, NCEP is shown in red, and the differences (MET-NCEP) are shown in green. The vertical bars represent the CIs at the 99% confidence level.

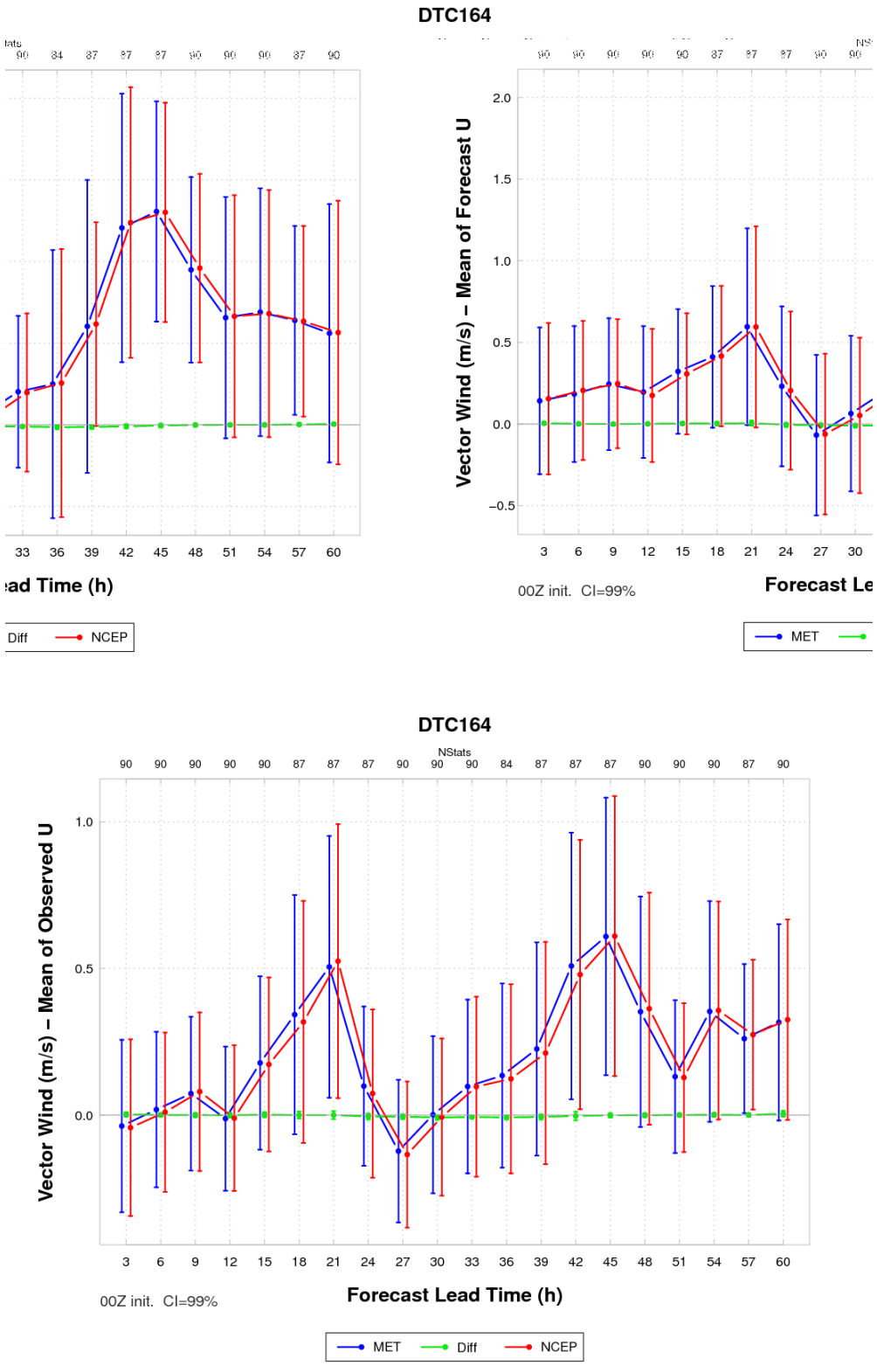

Figure 13. Time series plot of the median mean of the (a) forecast and (b) observed 10 m U-component of the vector wind aggregated across all of the cases for the 00 UTC initialization over the full domain (DTC164). MET is shown in blue, NCEP is shown in red, and the differences (MET-NCEP) are shown in green. The vertical bars represent the CIs at the 99% confidence level.

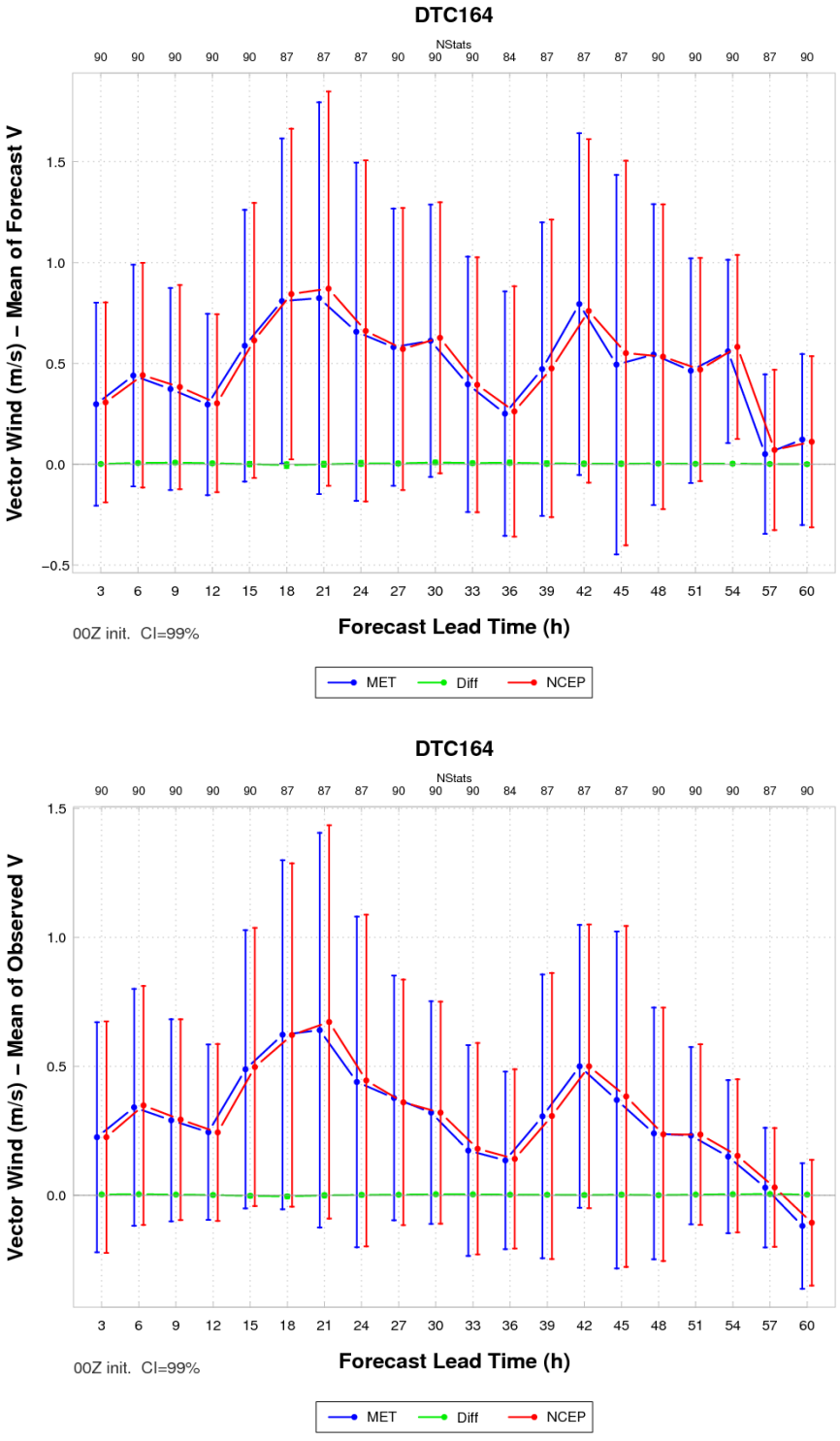

Figure 14. Time series plot of the median mean of the (a) forecast and (b) observed 10 m V-component of the vector wind aggregated across all of the cases for the 00 UTC initialization over the full domain (DTC164). MET is shown in blue, NCEP is shown in red, and the differences (MET-NCEP) are shown in green. The vertical bars represent the CIs at the 99% confidence level.

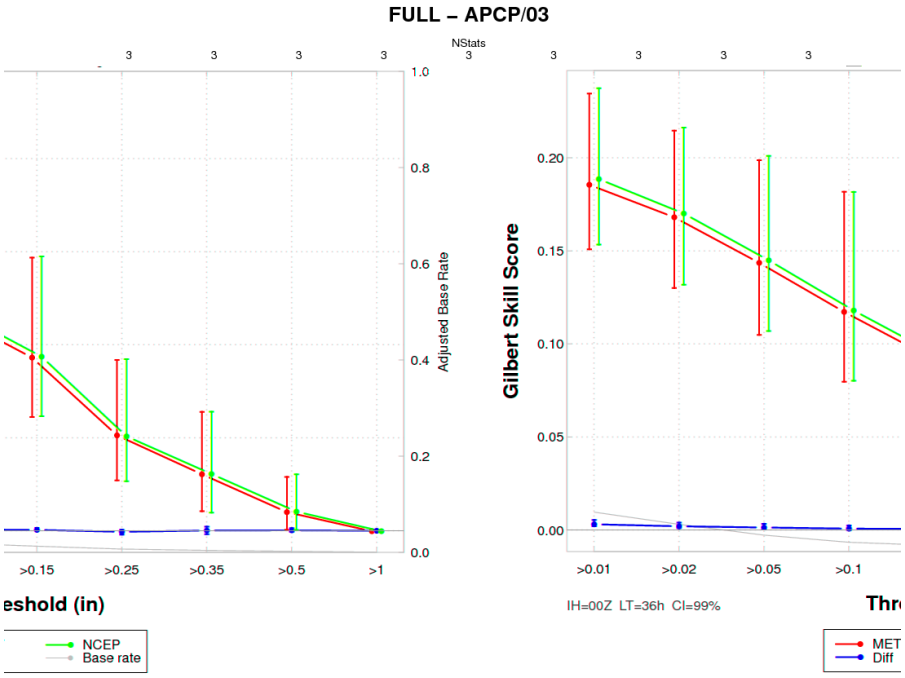

Figure 15. Threshold series plot of 3-h accumulated precipitation (in) for median GSS aggregated across all of the cases over the full domain (DTC164) for the 00 UTC initializations for 36-h lead time. MET is shown in blue, NCEP is shown in red, and the differences (MET-NCEP) are shown in green. The vertical bars represent the CIs at the 99% confidence level. The adjusted base rate is shown in grey on the second y-axis.

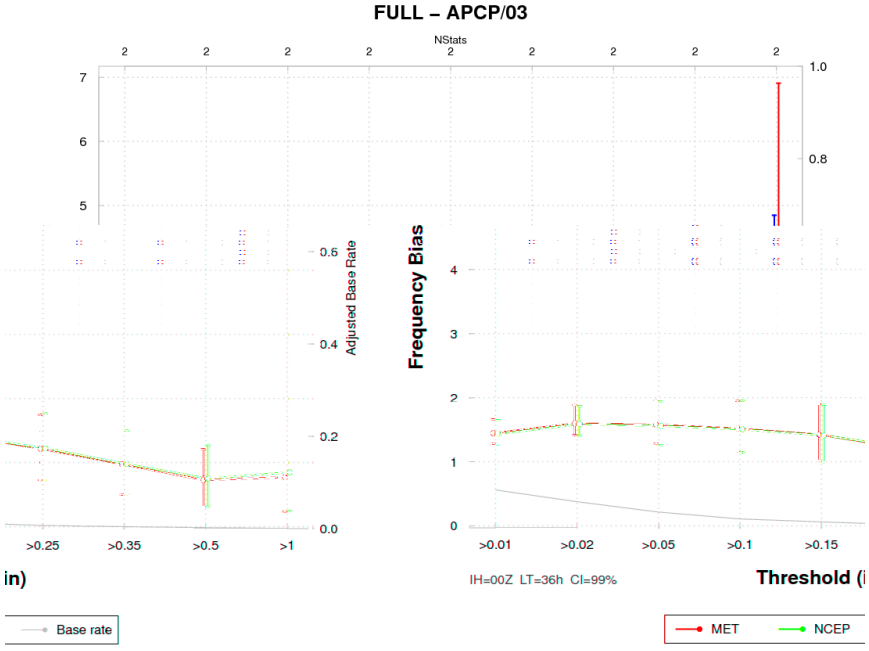

Figure 16. Threshold series plot of 3-h accumulated precipitation (in) for median frequency bias aggregated across all of the cases over the full domain (DTC164) for the 00 UTC initializations for the 36-h lead time. MET is shown in blue and NCEP is shown in red. The vertical bars represent the CIs at the 99% confidence level. The adjusted base rate is shown in grey on the second y-axis.

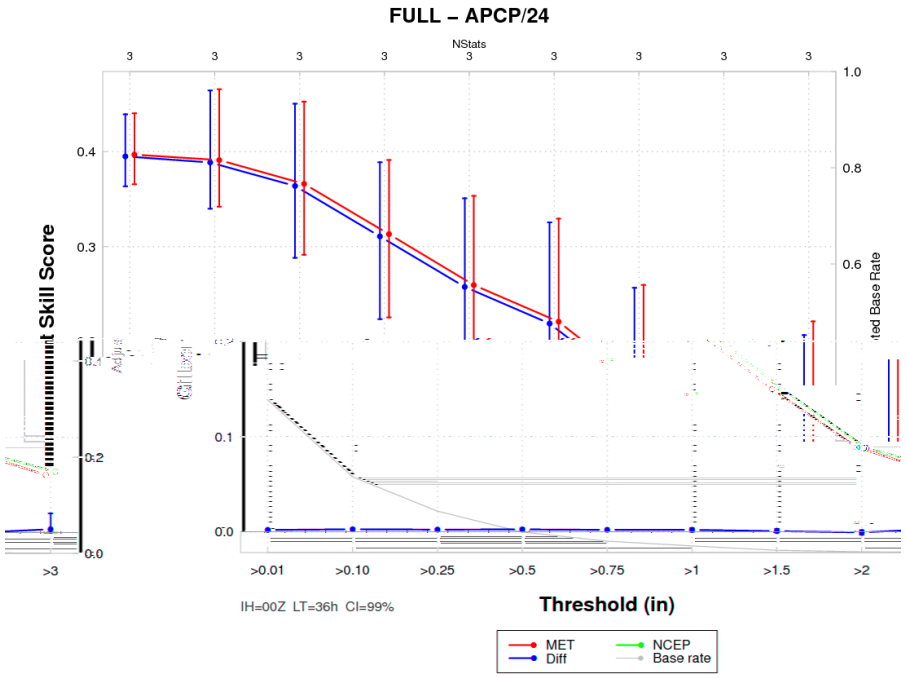

Figure 17. Threshold series plot of 24-h accumulated precipitation (in) for median GSS aggregated across all of the cases for the 00 UTC initialization for the 36-h lead time. MET is shown in blue, NCEP is shown in red, and the differences (MET-NCEP) are shown in green. The vertical bars represent the CIs at the 99% confidence level. The adjusted base rate is shown in grey on the second y-axis.

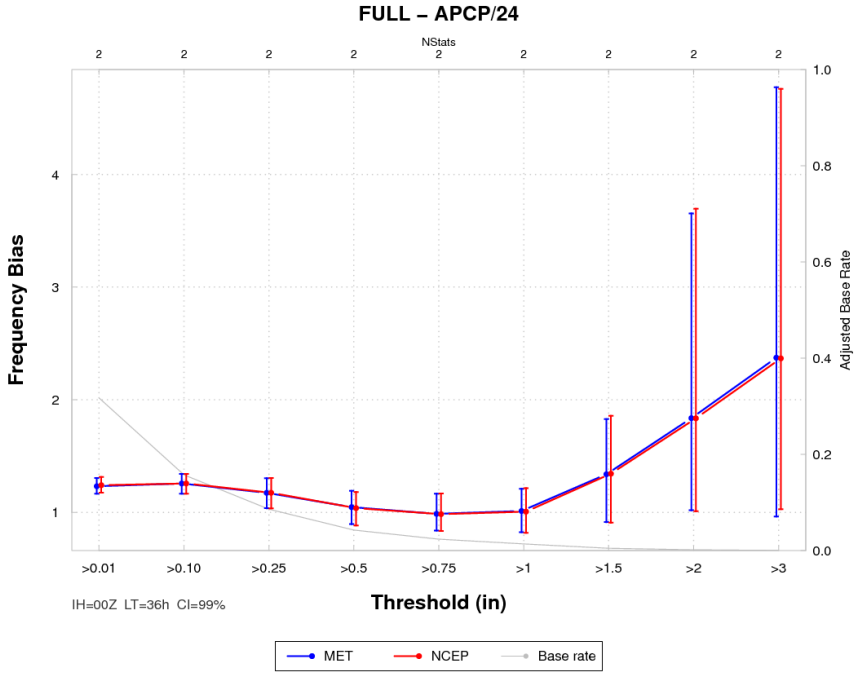

Figure 18. Threshold series plot of 24-h accumulated precipitation (in) for median frequency bias aggregated across all of the cases for the 00 UTC initialization for the 36-h lead time. MET is shown in blue, NCEP is shown in red, and the differences (MET-NCEP) are shown in green. The vertical bars represent the CIs at the 99% confidence level. The adjusted base rate is shown in grey on the second y-axis.