

**U.S. DEPARTMENT OF COMMERCE
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
NATIONAL WEATHER SERVICE
OFFICE OF SCIENCE AND TECHNOLOGY INTEGRATION
METEOROLOGICAL DEVELOPMENT LABORATORY**

MDL OFFICE NOTE 16-1

**VERIFICATION OF LAMP AND MOS FORECASTS
OF 10-M WIND SPEED AND COMPARISON
TO VALUES RETRIEVED FROM ANALYSES**

Bob Glahn

July 2016

VERIFICATION OF LAMP AND MOS FORECASTS OF 10-M WIND SPEED AND COMPARISON TO VALUES RETRIEVED FROM ANALYSES

Bob Glahn

1. INTRODUCTION

The Meteorological Development Laboratory (MDL) has been providing MOS forecasts of 10-m wind speed (so called surface wind) for many years. Since 1975 (Schwartz and Carter 1982), the forecasts have been “inflated” to produce more strong winds than regression produces.¹ They are produced either twice or four times per day, depending on the driving numerical model. These forecasts are currently provided for projections every 3 or 6 hours out to several days. More recently, but still for several years, MDL has provided LAMP (Ghirardelli and Glahn 2010) updates each hour and at hourly projection intervals. LAMP inputs include the most recent MOS forecasts, the latest available observations, and outputs from three advective models internal to LAMP.

To keep current, the regression equations, the means by which observations at stations are related to the numerical models and other predictors, should be rederived at intervals appropriate to changes to the inputs (e.g., numerical model changes). Refresh of the prediction system also allows new locations where there are observations to be included. The MOS system built on NCEP’s Global Forecast System (GFS) (Kalnay et al. 1990) furnishes the forecasts which go into LAMP. The latest refresh of the MOS wind equations was in 2015; the latest LAMP refresh was much earlier.

With the advent of IFPS (Interactive Forecast Preparation System; Ruth 2002) and the NDFD (National Digital Forecast Database) (Glahn and Ruth 2003), MOS and LAMP forecasts for station locations have been analyzed (gridded) with the BCDG method (Glahn et al. 2009; Im et al. 2010) and put into the National Digital Guidance Database, the partner of NDFD. Such grids can be used by forecasters in preparing the grids that go into the NDFD. In making such analyses, there is a tradeoff between fitting the data points closely, which tends to produce a map that may appear spotty, and smoothing to provide a map that does not emphasize detail.

There has been a move in recent years to verify the gridded MOS (GMOS), gridded LAMP (GLMP), and NDFD forecasts at gridpoints where the gridpoint values are determined by analyses of observations (and perhaps other information) at the resolution of the NDFD, nominally 2.5 km. The efficacy of this process depends, of course, on the quality of the verifying analysis. The URMA² has been proposed as the verifying analysis for this purpose.

¹ Inflated forecasts are obtained from the regression estimates by (1) subtracting the dependent sample mean, (2) dividing the result by the multiple correlation coefficient, and (3) adding back the sample mean. This increases the values above the mean, but also decreases the values below the mean. The practice has become in MDL to only modify the forecasts above the mean, which is called partial inflation.

² URMA is essentially the Real-Time Mesoscale Analysis (RTMA; De Ponca et al. 2007a, 2007b) system which is run several hours after observation time in order to include all relevant observations.

The tuning of BCDG to the characteristics of the MOS forecasts over the conterminous United States (CONUS) dates back several years (circa 2007), and it is planned to implement a retuned system in the near future. The LAMP BCDG system was implemented more recently, but, as stated above, is based on older LAMP prediction equations.

This office note presents the results of computing and comparing measures of accuracy of the LAMP and MOS forecasts, both their original values and values retrieved from their analyses, at points where the best measurements of 10-m winds exist. These measurement points were 1552 METAR sites over the CONUS. The sample consisted of one LAMP forecast cycle per day (06Z) and one MOS cycle per day (00Z) for each day in January and February in 2016. The MOS forecasts made 6 h earlier than LAMP were those available at the same time as LAMP. We also looked at how well point observations could be retrieved from the URMA analyses.

LAMP and MOS forecasts are made explicitly for the 1552 sites used in the verification. Point values from the gridded LAMP and MOS forecasts and the URMA analyses at the METAR sites were “interpolated” from the grid by taking the closest gridpoint value to the station location. At a grid resolution of only 2.5 km the closest gridpoint is always within about 2 km. In rough terrain, especially, the estimation at a measurement point (e.g., a METAR station) from a grid is not perfect, and should be considered in assessing the statistics, but the retrieval from a regular grid is a much simpler and accurate process than estimating gridpoint values from a much more sparse set of somewhat random points.

2. MEASURES OF QUALITY

All forecasts were rounded to whole kt before calculating metrics. Observations are reported and forecasts are provided to users at that resolution. Bias (forecast - observed) and mean absolute error (MAE) were calculated. However, it is recognized that most forecasts and observations are of very low speed and are relatively unimportant for aviation purposes. For presentation of bias and MAE, we stratified the sample into one where at least one of the forecast systems being comparatively verified or the verifying observation was $\geq X$ kt,³ where $X = 8, 15,$ and 20 . A 20-kt sustained wind will usually be gusty and be very important for aircraft operations. It is recognized that this sampling will give scores that vary depending on what and how many systems are being verified, but does provide a sample that excludes all relatively unimportant cases where all forecast systems and the verifying observation were < 8 kt.

We also looked at bias in terms of how many very low and high wind forecasts were made, in relation to the observations. We calculated the threat score (TS) for $\geq X$ kt, where $X = 10, 15,$ and 20 , and the Gerrity skill score. The so called Gerrity skill score (Gerrity 1992)⁴ is a measure of accuracy where the scoring matrix is calculated from the sample relative frequencies of observations. The score gives high weight for hitting or near-missing rare categories, and very little weight for the predominant category (see Table 1).

³ Meteorologists drafting the National Verification Plan (1982) for the NWS realized light winds were of lesser importance, and recommended “no comparison for wind speeds < 10 mph.”

⁴ The credit for the score should rightly be shared with Lev Gandin and Allan Murphy who discussed equitable skill scores for categorical variables (Gandin and Murphy 1992). Gerrity provided explicit formula for calculating the scoring matrix for an arbitrary number of categories.

3. BIAS ON STRATIFIED SAMPLES

Figures 1 and 2 show, respectively, the bias (forecast-observed) on the stratified samples when one or more winds were ≥ 8 kt, and when one or more winds were ≥ 20 kt. Note the vertical scales are different to emphasize differences among systems. The LAMP and MOS forecasts (solid lines) were roughly comparable, MOS having slightly less bias for the stronger winds. The biases were high (i.e., > 0), especially for the stronger winds, probably because they are partially inflated. Partial inflation (see footnote above) means the regression forecasts above the mean are increased to produce more strong winds. The LAMP forecast analyses were recently tuned to emphasize strong winds, and the GLMP analyses show more high bias than the forecasts on which they are based. The GMOS analysis values, however, are quite different from the MOS forecasts. This difference is probably due to the BCDG analysis not being adequately tuned for the data set being analyzed, which includes numerous mesonet sites. Surprisingly, the URMA is quite low biased. This is probably because there are many mesonet winds being considered, and they are, overall, low biased (Manikin and Pondeca 2009). Evidently, because of this difference in quality of measurement, the reliable METAR observations were not fit well (fit about 4.0 kt too low for the important ≥ 20 -kt winds).

4. MAE ON STRATIFIED SAMPLES

Figures 3 and 4 show, respectively, the MAE on the stratified samples when one or more winds were ≥ 8 kt, and when one or more winds were ≥ 20 kt. Again, note the vertical scales are different to emphasize differences among systems. The MOS forecasts had, in general, lower MAE than the LAMP forecasts, except for the 3-h projection. This is reasonable, given that MOS has been developed more recently than LAMP, and LAMP used the older MOS equations in development, but the newer ones in operations. This supports the intention to redevelop LAMP equations within the next few months.

The MOS and LAMP analyses of the forecasts had more error than the original station forecasts. This is not surprising, but does show that the modest smoothing inherent in the BCDG analysis does not improve the forecasts at data points. The LAMP values were fit better than the MOS values. It is surprising the URMA did not fit the winds observed at METAR sites any better than the LAMP or MOS forecasts, except on the stratified 8 kt-sample for the longest projections. However, the values recovered from URMA (at the valid time) generally had lower MAE than the values recovered from GLMP and GMOS, except for the LAMP 3-h projection.

5. NUMBER OF CALM WINDS FORECAST AND OBSERVED

Figure 5 shows the number of calm (speed = 0) winds forecast and observed. The percentage of calm winds in the sample varied diurnally from about 8% to 21%. None of the forecast systems approached that percentage, the closest being GLMP and the largest difference being for URMA and GMOS. LAMP and MOS were not far different. GLMP had more calm winds than LAMP. GMOS had less calm winds than MOS.

6. NUMBER OF VALUES FORECAST AND OBSERVED < 3 KT

Figure 6 shows the number of forecasts and observations of < 3 kt. For reference 20,000 is about 25% percent of the sample. There is not a great difference among the systems verified,

except URMA is considerably higher. In concert with above, URMA is probably heavily influenced by the large number of mesonet obs but is hesitant to give gridpoints the value of zero (see Fig. 5). The relationship of number of forecasts to obs varies diurnally, the forecasts underestimating the number of low values when the winds are higher in the middle of the day (low number of low values) and overestimating the number of low values when the winds are lower. By this measure, the fit between the analyses and original LAMP and MOS station forecasts was quite good, especially for GLMP.

7. NUMBER OF VALUES FORECAST AND OBSERVED ≥ 20 AND ≥ 25 KT

Figures 7 and 8 show, respectively, the number of forecasts and observations of ≥ 20 and ≥ 25 kt. The number of high LAMP winds exceeded MOS somewhat, but the two were relatively close. The LAMP analysis showed the intentional emphasizing of the stronger winds. The number of MOS forecasts and observations were very close, and given the needed emphasis of strong winds, both MOS and LAMP did well. The number of winds recovered from URMA analysis, was as low as 19% of the observed number for the ≥ 25 -kt winds.

8. THREAT SCORES OF STRONG WINDS

Figures 9, 10, and 11 show, respectively, the threat scores for winds ≥ 15 , ≥ 20 , and ≥ 25 kt. These figures show that LAMP and MOS are very close, MOS edging LAMP a bit, in agreement with the better MAE for MOS. GLMP is very close to LAMP and slightly lower as might be expected. Again, this shows the slight smoothing inherent in the BCDG analysis does not improve the forecasts. On the other hand GMOS is considerably lower than MOS indicating the necessity of retuning the analysis. URMA is the lowest of all. This means that for the sample in which either the forecast or the verifying observation indicated a threat, the forecasts had a higher percentage of being correct than URMA.⁵

9. GERRITY SCORE

The Gerrity skill score (Gerrity 1992) is computed on a contingency table of forecast/observed values. It is an equitable score that gives high weight to correctly forecasted rare categories and considers near misses. While this is not the only such equitable score, and the assumptions underlying the calculation of the scoring matrix it uses are somewhat arbitrary, it seems a reasonable attempt to measure the overall goodness of a set of forecasts, taking into account the importance of the event. The scoring matrix calculated on the observations at the valid time of the 3-h LAMP forecasts is shown in Table 1. As indicated, the strong winds get weighted much more heavily than the light winds. For this score, a 5-category contingency table was used as shown in Table 1.

Figure 12 shows LAMP, MOS, and GLMP to be of about equal skill, while GMOS and URMA are considerably lower. This means that, by this measurement, LAMP, MOS, and GLMP give better forecasts at METAR sites than would an URMA analysis made with data at the valid time.

⁵ In this comparison, the samples are not matched, because each system is treated separately and will not always agree on what is a threat.

10. COMPARISON WITH GRIDDED VERIFICATION

Figure 13 shows the verification of GLMP and GMOS at gridpoints where the URMA is used as the verifying analysis. Fig. 13 shows GMOS to have lower MAE than GLMP, but Fig. 3, where retrieved points from the analyses were verified with METAR observations, shows GLMP to have the lower MAE. We believe Fig. 3 gives a much truer assessment of GLMP and GMOS than Fig. 13. Fig. 3 is based on the gold standard wind observations as truth, while Fig. 13 is based on values at hundreds of thousands of points for which there are no close reliable observations, and other information that may be used in URMA cannot adequately compensate for lack of accurate measurements.

11. SUMMARY AND CONCLUSIONS

Conclusions drawn from the data and discussion above concerning the 10-m winds over the CONUS are:

- 1) The LAMP forecasts did not provide an improving update to MOS, except for the 3-h projection. MOS has been updated more recently than LAMP, and this indicates a LAMP update is needed. It is planned for the near future.
- 2) The LAMP analysis fit the forecasts better than did the MOS analysis. The BCDG analysis control parameters have been retuned more recently for LAMP than for MOS, and this indicates a GMOS update is needed. It is planned for the near future.
- 3) The GLMP update over GMOS provided considerable improvement, even though LAMP did not generally improve over MOS. This emphasizes the importance of a good analysis.
- 3) The URMA was disappointing for determining values at specific points where there are good measurements. The URMA, as do most numerical models and data assimilation (analysis) systems, undergoes improvements. These conclusions are based on what was running operationally in early 2016. URMA, and other analysis systems, provide a way of estimating values at points where there are no measurements. However, it is hard to see how a system that is poor for estimating values where there are reliable measurements would be useable for assessing the goodness of a gridded forecast, except, possibly, in very broad terms at long projections where detail is judged to be unimportant.

ACKNOWLEDGMENTS

Thanks go to Adam Schnapp and Chenjie Huang for providing me the data used in this study.

REFERENCES

- De Pondeca, M. S. F. V., and Coauthors, 2007a: The development of the Real Time Mesoscale Analysis System at NCEP. Preprints, *23rd Conf. On Interactive Information Processing Systems*, San Antonio, TX, Amer. Meteor. Soc., **P1.10**.
- _____, and Coauthors, 2007b: The status of the Real Time Mesoscale Analysis System at NCEP. Preprints, *22nd Conf. On Weather Analysis and Forecasting/18th Conf. On Numerical Weather prediction*, Park City, UT, Amer. Meteor. Soc., **4A5**.

- Gandin, L., and A. H. Murphy, 1992: Equitable skill scores for categorical forecasts. *Mon. Wea. Rev.*, **120**, 361-370.
- Gerrity, J. P., 1992: A note on Gandin and Murphy's equitable skill score. *Mon. Wea. Rev.*, **120**, 2709-2712)
- Ghirardelli, J. E., and B. Glahn, 2010: The Meteorological Development Laboratory's aviation weather prediction system. *Wea. Forecasting*, **24**, 1027-1051.
- Glahn, H. R., and D. P. Ruth, 2003: The new digital forecast database of the National Weather Service. *Bull. Amer. Meteor. Soc.*, **84**, 195-201.
- _____, K. Gilbert, R. Cosgrove, D. P. Ruth, and K. Sheets, 2009: The Gridding of MOS. *Wea. Forecasting*, **24**, 520-529.
- Im, J.-S., B. Glahn, and J. E. Ghirardelli, 2010: Real-time objective analysis of surface data at the Meteorological Development Laboratory. Preprints, *20th Conf. on Probability and Statistics in the Atmospheric Sciences*, Atlanta, GA, Amer. Meteor. Soc., **219**.
- Kalnay, E., M. Kanamitsu, and W. E. Baker, 1990: Global numerical weather prediction at the National Meteorological Center. *Bull. Amer. Meteor. Soc.*, **71**, 1410-1428.
- Manikin, G. S., and M. Pondeva, 2009: Challenges with real-time mesoscale analysis (RTMA). *23rd Conference on Weather Analysis and Forecasting/19th Conference on Numerical Weather Prediction*, Omaha, NE, Amer. Meteor. Soc., **1A1**.
- NWS, 1982: *National Verification Plan*. National Oceanic and Atmospheric Administration, U.S. Department of Commerce, 81pp.
- Ruth, D. P., 2002: Interactive Forecast Preparation - The future has come. Preprints, *Interactive Symposium on the Advanced Weather Interactive Processing System (AWIPS)*, Orlando, FL, Amer. Meteor. Soc., 20-22.
- Schwartz, B. E., and G. M. Carter, 1982: An evaluation of a modified speed enhancement technique for objective surface wind forecasting. *TDL Office Note 82-1*, National Weather Service, NOAA, U.S. Department of Commerce, 11 pp.

Table 1. The forecast/observed contingency table for the 9-h GMOS forecast verifying at 0900 UTC (top) and the Gerrity scoring matrix for this verification time (bottom). The scoring matrix is based on (only) the observed frequencies and is the same for all systems for this verifying time. The TS, probability of detection (POD) and false alarm ratio (FAR) for the higher three categories and the Gerrity score are also shown. A 9-h GMOS forecast corresponds to a 3-h GLMP forecast.

| 9-HR PROJECTION | | 79042 CASES | | | | | | | | |
|--|--------------------|-------------------|---------|------------------|--------|-------|--------|-------|-------|---|
| 9h GMOS | SPD (KT) | | | FORECAST | | | | | | |
| | CATEGORY | LOWER LIMITS (GE) | -99999. | 3. | 8. | 15. | 20. | | | |
| | | UPPER LIMITS (LT) | 3. | 8. | 15. | 20. | 99999. | TOTAL | | |
| | | | . | . | . | . | . | . | . | . |
| GE | 20. | | . | 10060 | 6506 | 309 | 4 | 3 | 16882 | |
| | THREAT SCORE | .266 | . | | | | | | | |
| | POD | .346 | . | 7955 | 23333 | 4432 | 135 | 6 | 35861 | |
| | FAR | .466 | . | | | | | | | |
| GE | 15. | | OBS. | 305 | 7824 | 11869 | 1306 | 119 | 21423 | |
| | THREAT SCORE | .361 | . | | | | | | | |
| | POD | .478 | . | 8 | 232 | 1943 | 1245 | 237 | 3665 | |
| | FAR | .403 | . | | | | | | | |
| GE | 8. | | . | 2 | 43 | 318 | 429 | 419 | 1211 | |
| | THREAT SCORE | .573 | . | | | | | | | |
| | POD | .680 | TOT. | 18330 | 37938 | 18871 | 3119 | 784 | 79042 | |
| | FAR | .215 | . | | | | | | | |
| | GERRITY SKILL SCOR | .462 | . | | | | | | | |
| | | | . | . | . | . | . | . | . | . |
| | | | | BIAS BY CATEGORY | 1.09 | 1.06 | .88 | .85 | .65 | |
| GERRITY NFCST X NOBS SCORING MATRIX. IT IS THE SAME FOR ALL SYSTEMS. | | | | | | | | | | |
| | 1.065 | -.105 | -.480 | -.746 | -1.000 | | | | | |
| | -.105 | .213 | -.162 | -.428 | -.682 | | | | | |
| | -.480 | -.162 | .590 | .323 | .069 | | | | | |
| | -.746 | -.428 | .323 | 4.376 | 4.122 | | | | | |
| | -1.000 | -.682 | .069 | 4.122 | 20.439 | | | | | |

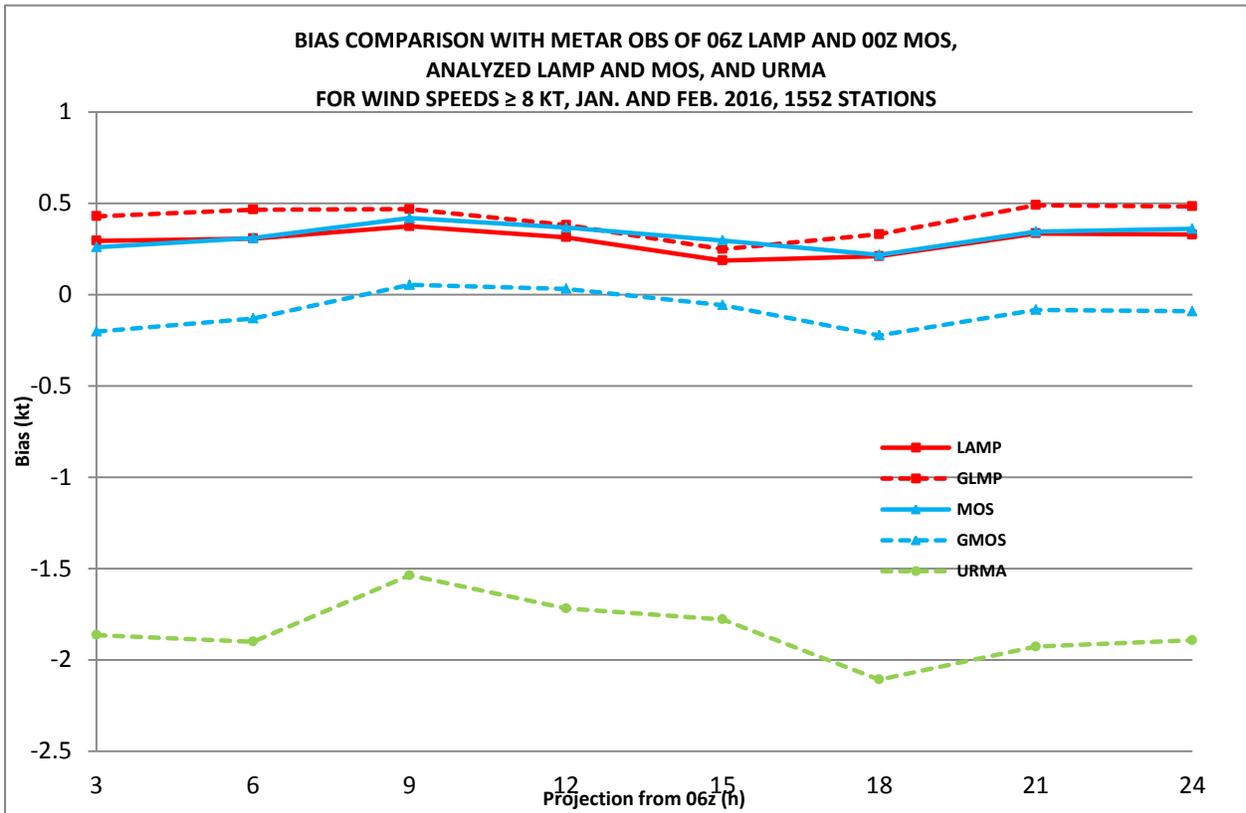


Figure 1. Bias of MOS, LAMP, GMOS, GLMP, and URMA for any wind speed \geq 8 kt.

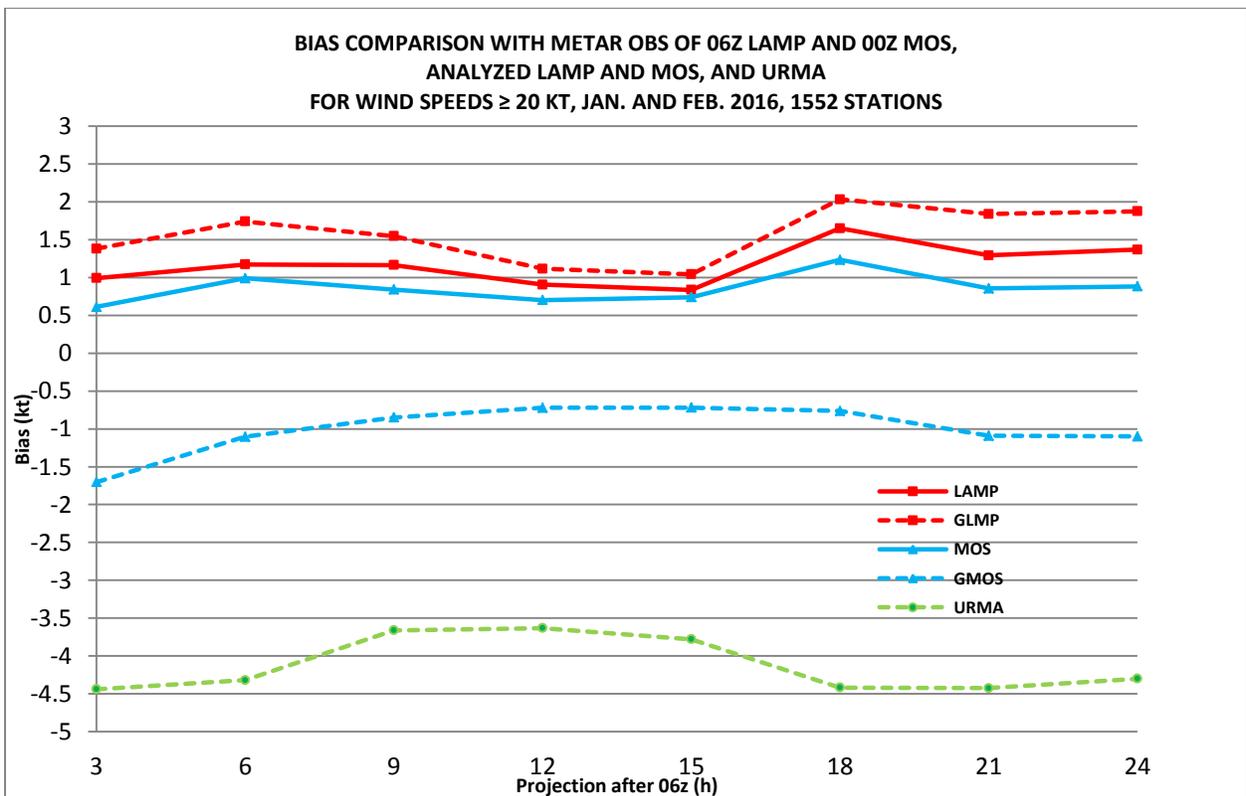


Figure 2. Same as Fig. 1 except for \geq 20 kt.

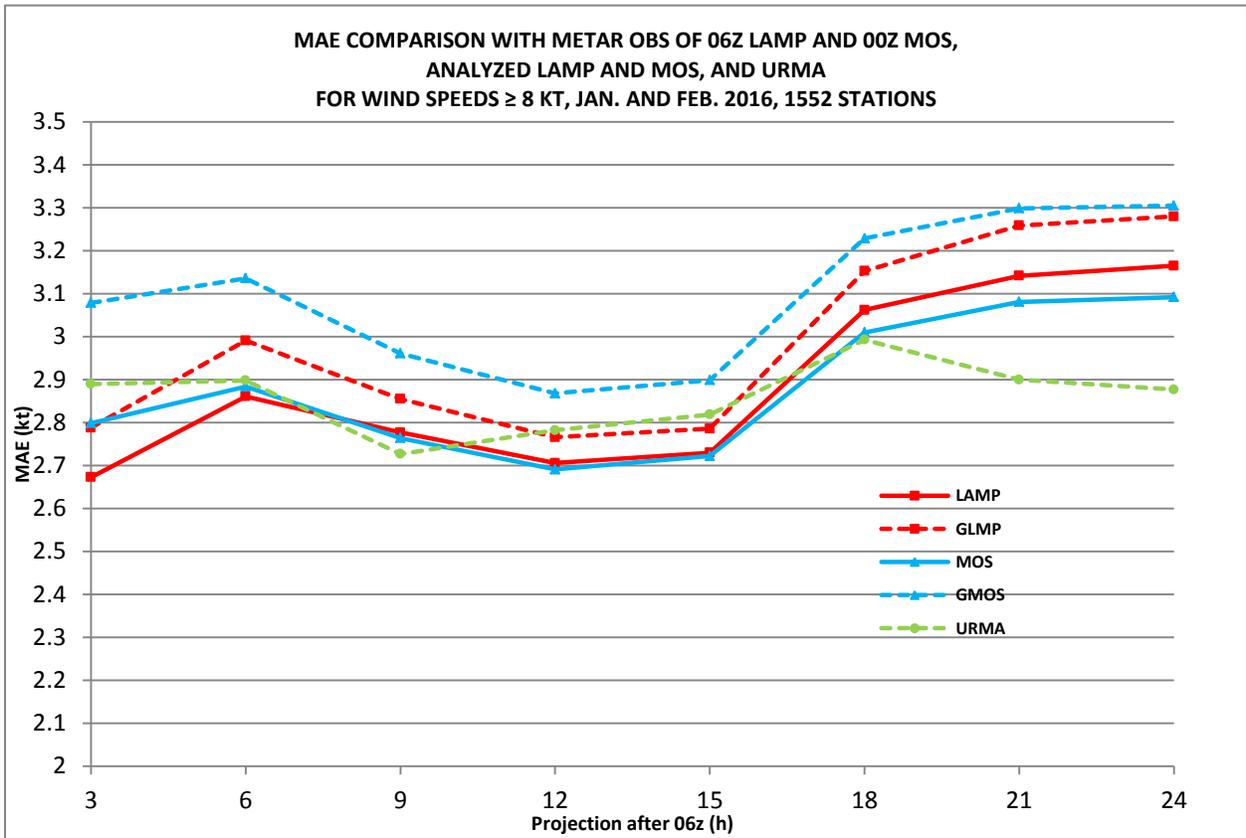


Figure 3. Same as Fig. 1 except for MAE.

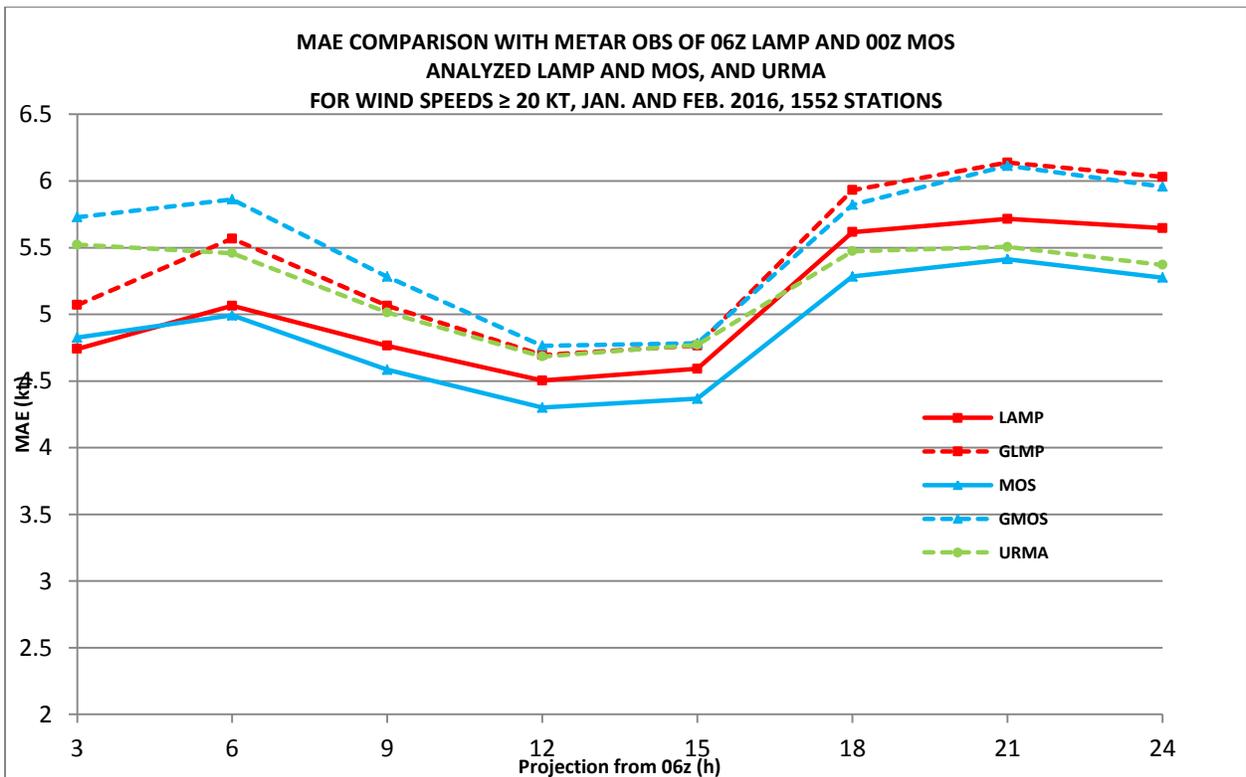


Figure 4. Same as Fig. 2 except for MAE.

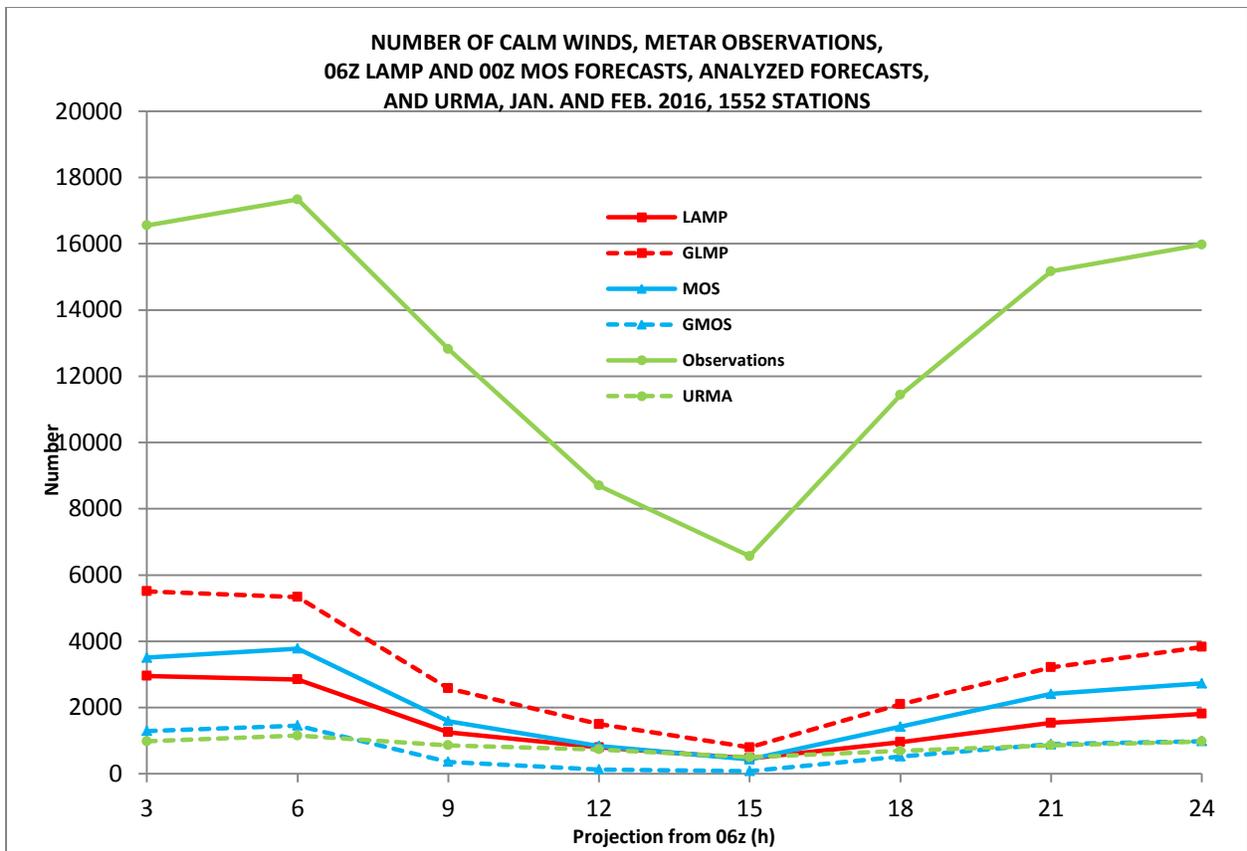


Figure 5. Number of calm winds in sample of approximately 80,000 cases.

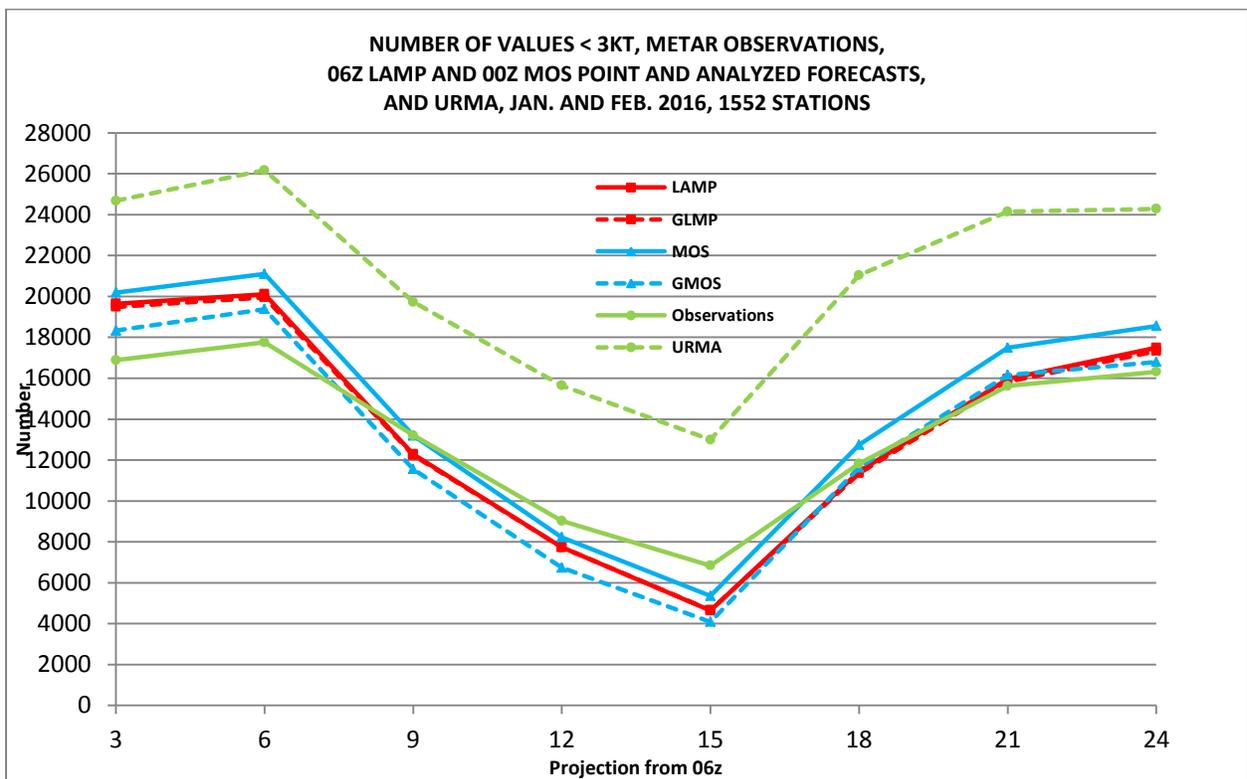


Figure 6. Number of winds < 3 kt in sample of approximately 80,000 cases.

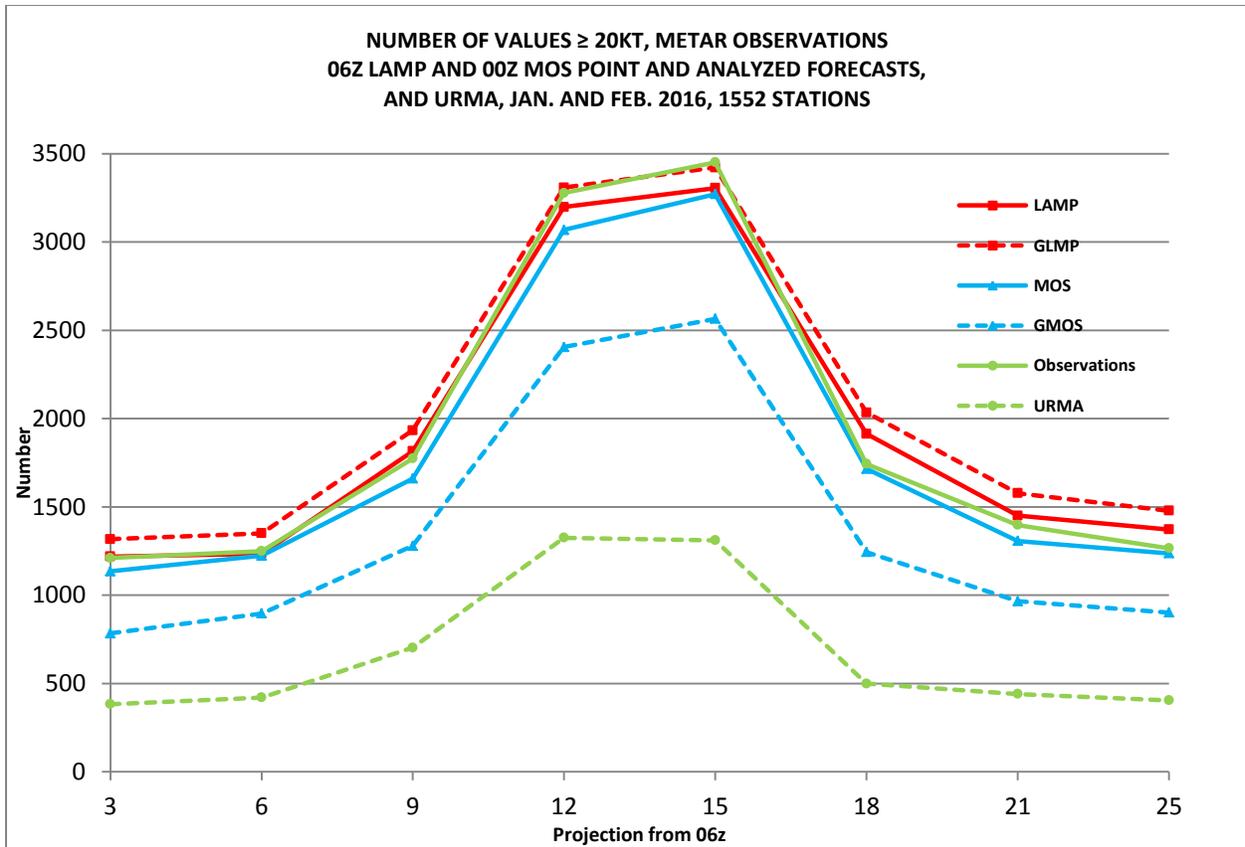


Figure 7. Same as Fig. 6 except ≥ 20 kt. Note different scales.

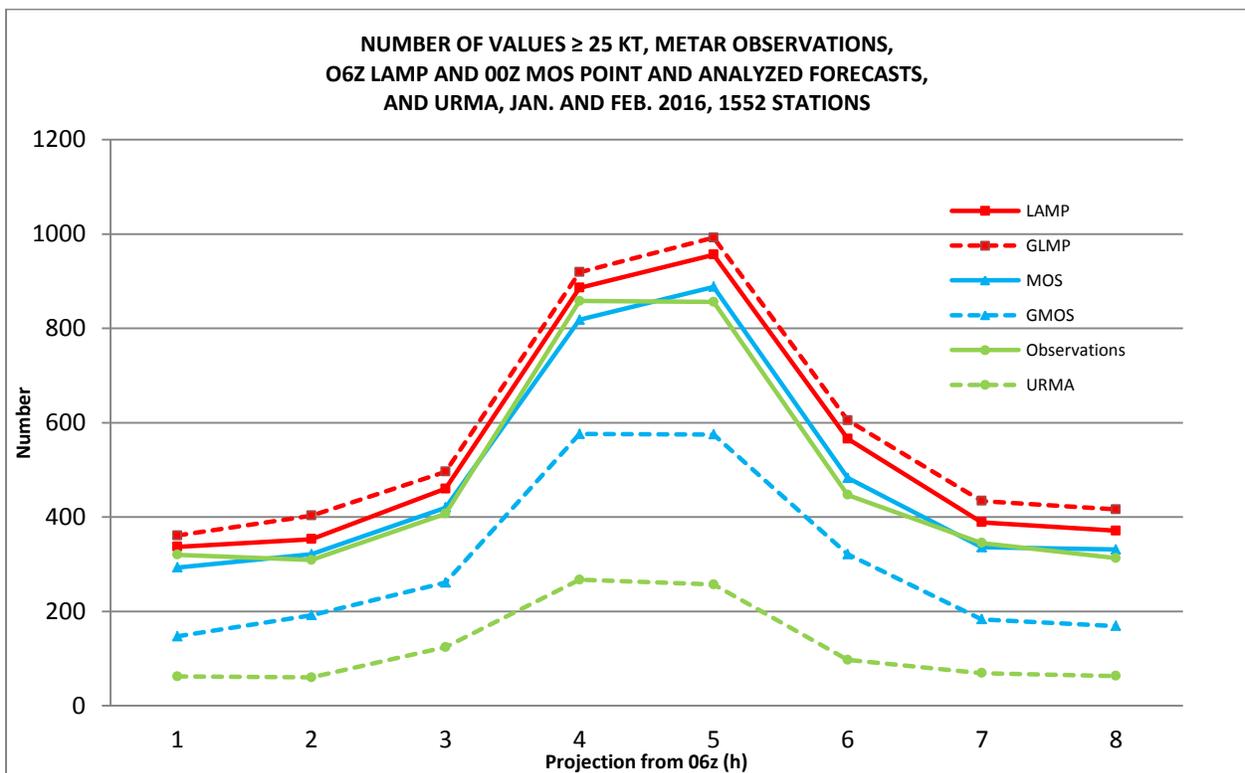


Figure 8. Same as Fig. 6 except ≥ 25 kt. Note different scales.

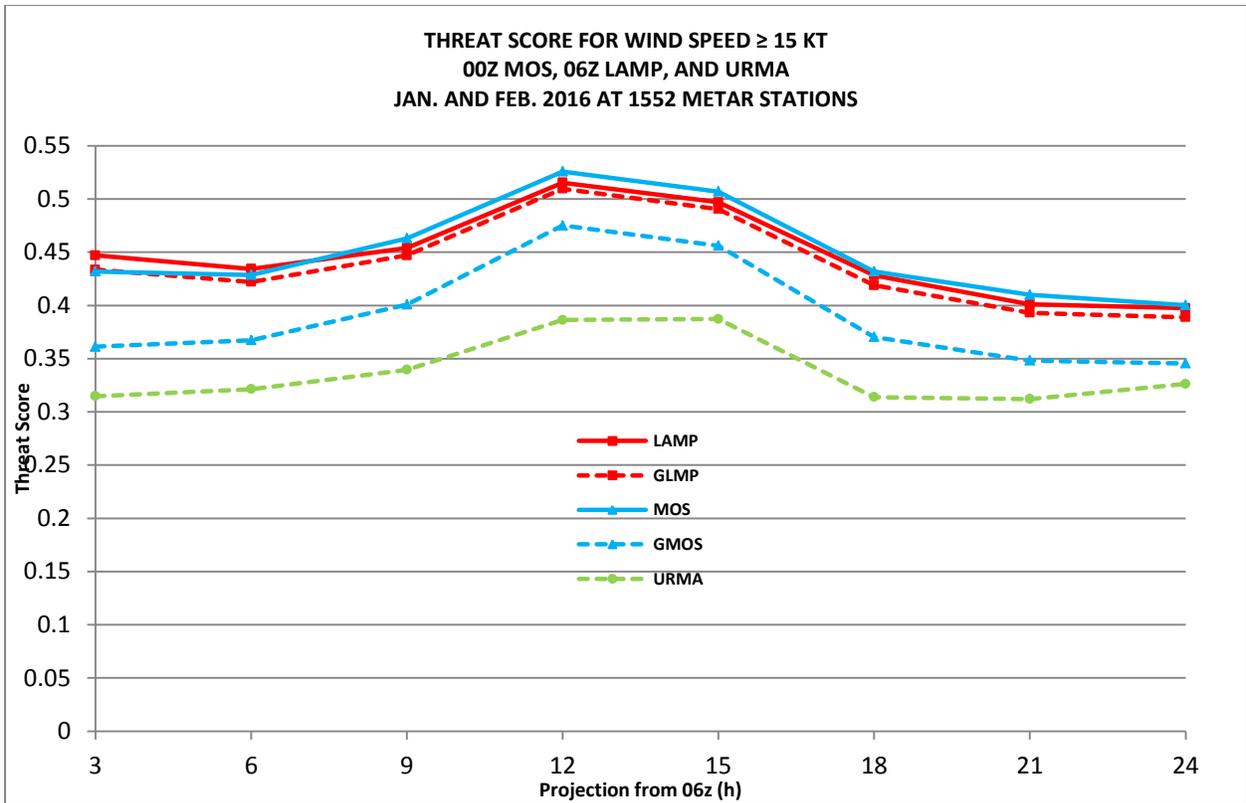


Figure 9. Threat score for winds \geq 15 kt.

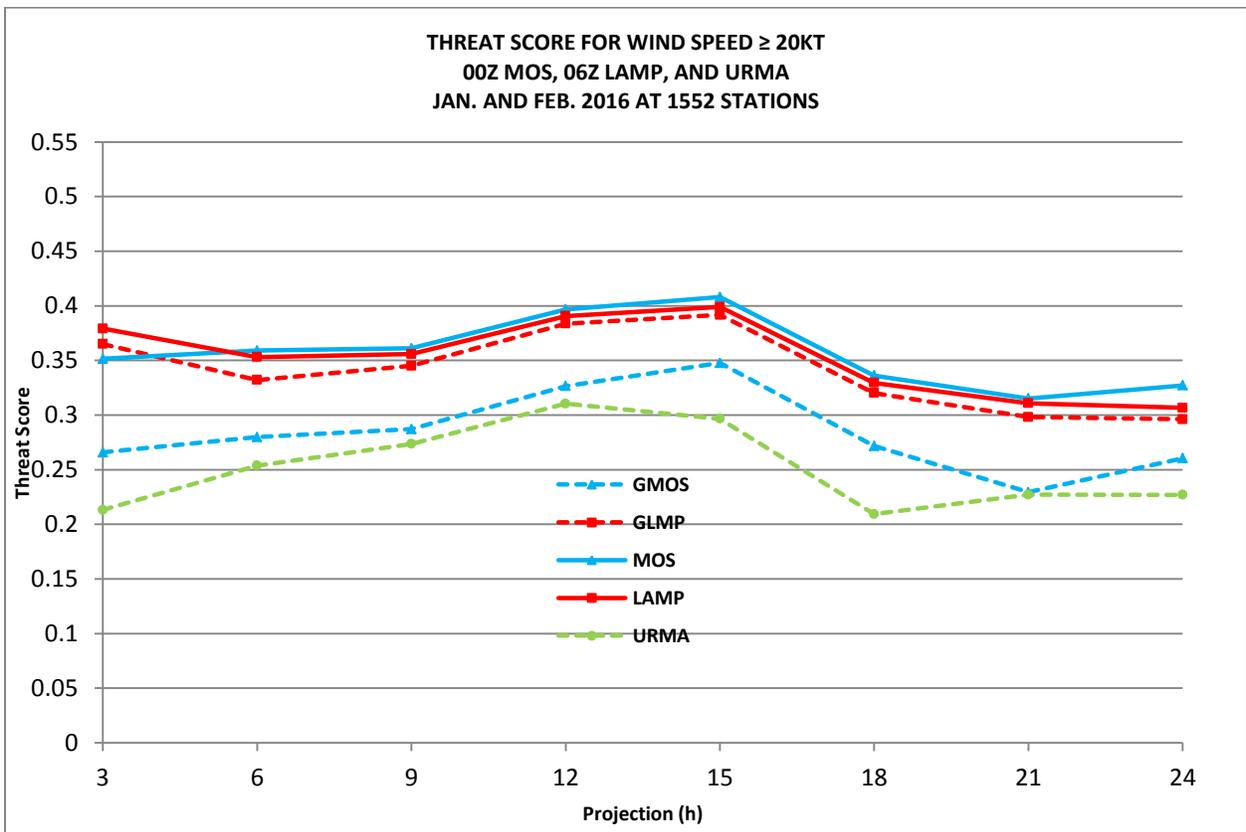


Figure 10. Threat score for winds \geq 20 kt.

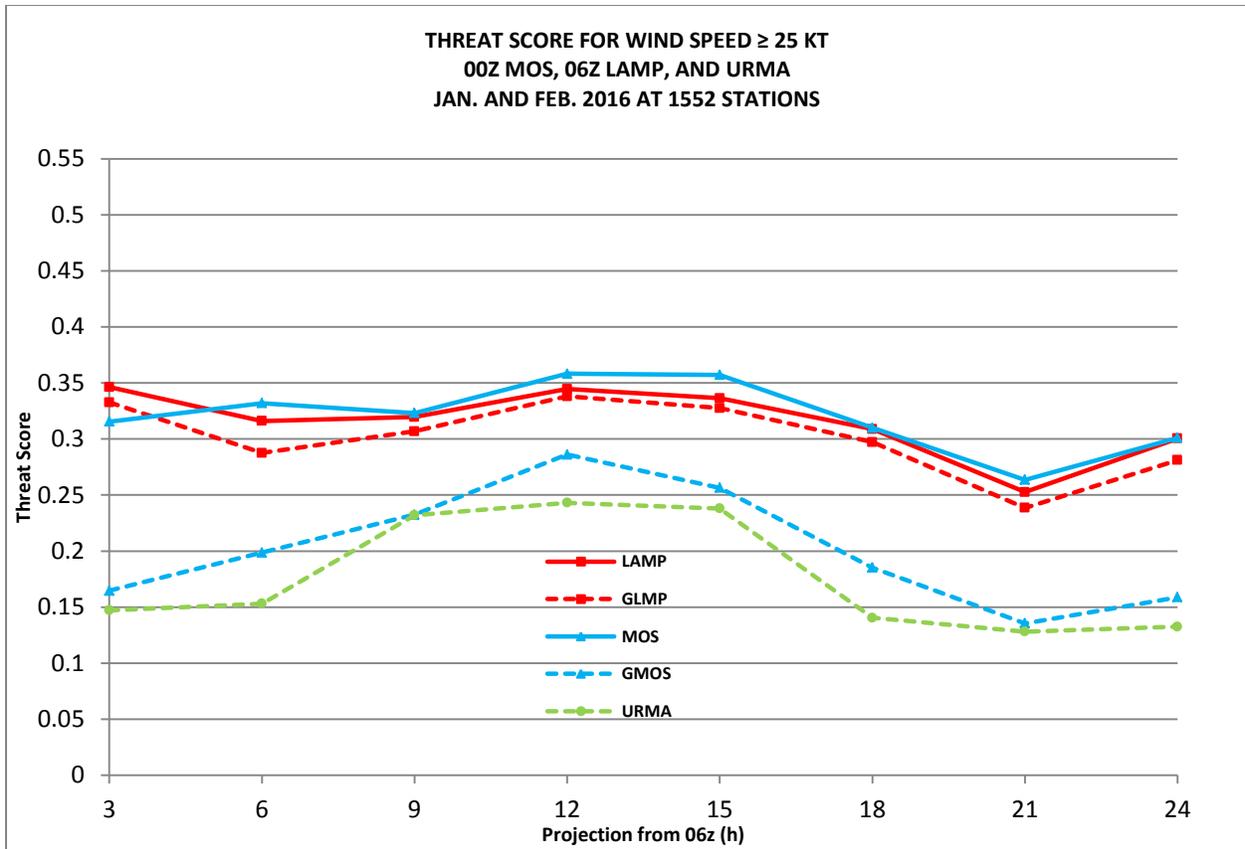


Figure 11. Threat score for winds \geq 25 kt.

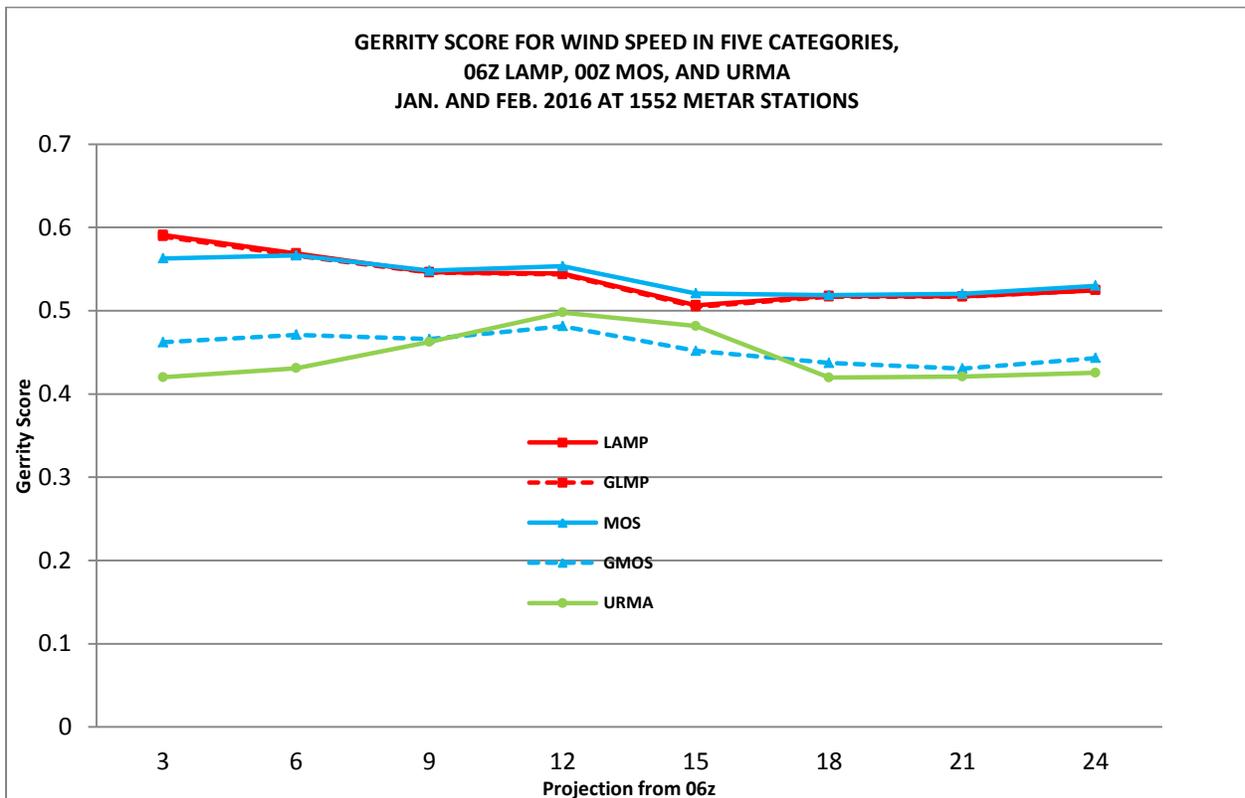


Figure 12. Gerrity skill score.

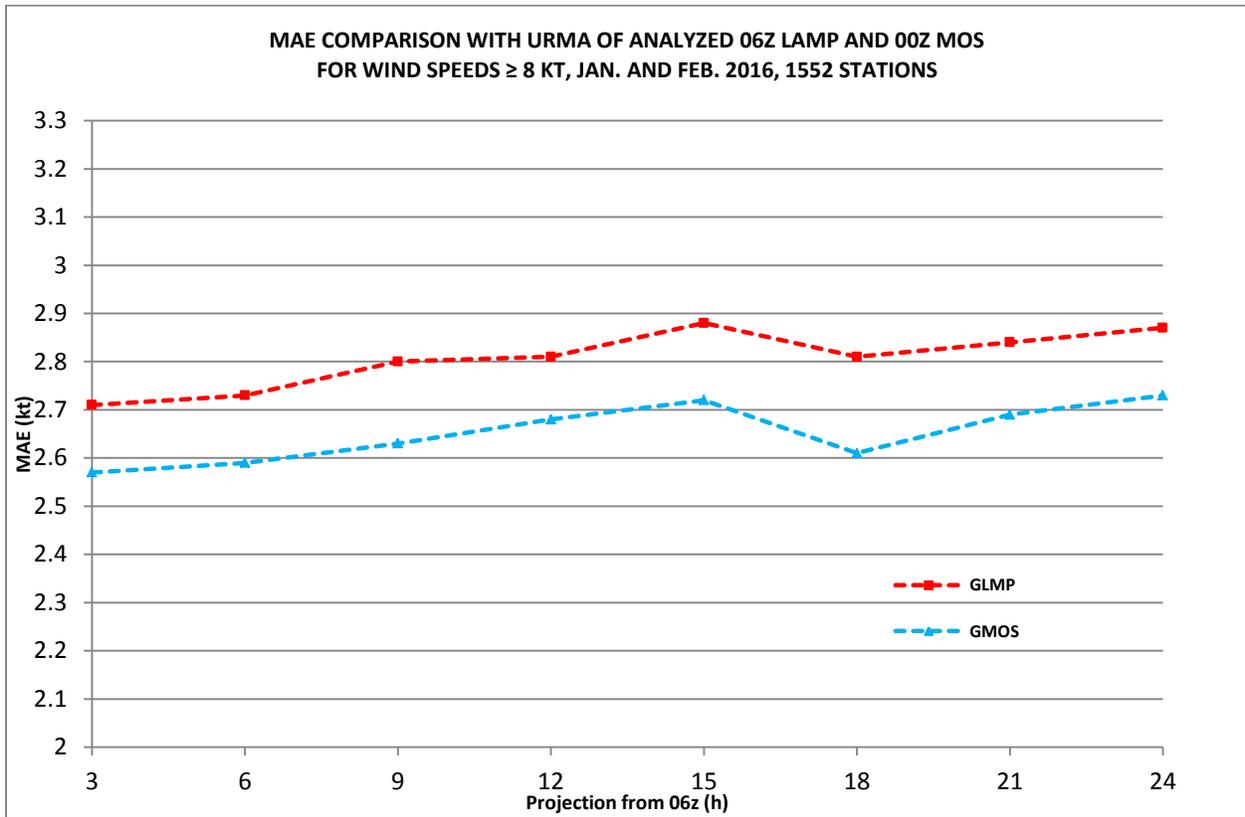


Figure 13. Gridded comparison of GMOS and GLMP.