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

MDL OFFICE NOTE 13-1

A COMPARISON OF TWO METHODS OF BIAS CORRECTING MOS TEMPERATURE AND DEWPOINT FORECASTS

Bob Glahn

January 2013

A COMPARISON OF TWO METHODS OF BIAS CORRECTING MOS TEMPERATURE AND DEWPOINT FORECASTS

Bob Glahn

1. INTRODUCTION

MOS temperature and dewpoint forecasts tend to have some bias on both a monthly and seasonal basis. Regression equations that produce MOS forecasts give unbiased estimates over the period of the developmental sample, but they may have bias over intervals within that developmental sample and over other samples including future forecasts. Given there are no changes in the system producing the forecasts (e.g., numerical model, method of data collection, etc.), the bias on a seasonal basis is expected to be small, but the overall synoptic patterns are different from year to year, so biases may occur.

Several bias correction methods have been reported in the literature (e.g., Yussouf and Stensrud 2007; Woodcock and Engel 2005) that largely eliminate forecast bias, but a modification to the forecasts to correct bias can lead to larger mean absolute errors (MAE) and degrade other metrics used to evaluate the quality of the forecasts. Most studies have not dealt with MOS operational forecasts. Glahn (2012) tested a method that has been used at the National Centers for Environmental Prediction (NCEP) for several years (Cui et al. 2012) called decaying average and found that it could not only improve bias but either improved or did not degrade the other performance metrics. A method of correction is employed within the Boise Verify software that is widely used at National Weather Service (NWS) field offices. It essentially fits a linear line to the last 30 days of forecasts and observations from which a corrected forecast can be deduced. While the decaying average could be applied successfully to MOS forecasts, it seemed prudent to compare it with the method predominately used at NWS Weather Forecast Offices (WFO). This office note describes the two methods and documents the results of the comparison.

2. DECAYING AVERAGE ALGORITHM

Cui et al. (2012, abstract) call the decaying average implemented at NCEP in 2006 a "Kalman filter type algorithm," but do not state the assumptions made in deriving it. The use of the Kalman filter as originally proposed (Kalman 1960) requires estimates of several constants and parameters. Kalman's original work (1960) has been described in several publications in the meteorological literature (e.g., Roeger et al. 2003; Homleid 1995; Galanis and Anadranistakis 2002; Cheng and Steenbaugh 2007) and won't be repeated here.

To implement the algorithm, one has only to carry forward a delta and apply it to the current forecast. Then to prepare for the next forecast cycle, the delta d would be updated by:

$$d(t+1) = (1-\alpha)\, d(t) + \alpha\, (F - O)(t)$$

where d(t+1) is the delta to apply at time t+1, d(t) is the delta applied at time t, F is the forecast "verified" by the observation O at time t, and α is the weight to apply to the most recently calculated forecast error F - O at time t. There would optimally be a specific delta for each station as well as forecast projection. When F - O is missing, zero can be assumed (see below).

In an operational setting, the modification to the MOS forecast does not have to be made until the observation is available. Therefore, the delta for the next forecast can incorporate the error of the forecast verifying at that time.

A choice of α has to be made. NCEP uses $\alpha = 0.02$.[1] Glahn (2012) tested α values of 0.025, 0.050, 0.075, and 0.1 and found that values of 0.025 and 0.050 were overall best, and the difference in results for those values was not great. While higher values were more effective in removing the bias both long term and short term, the other performance metrics were generally degraded for higher values. For the tests reported here, $\alpha = 0.04$ was used.

As a practical matter, there are situations where a station will not report for a considerable length of time, or may stop altogether. MOS forecasts continue because they are based on model data that are available. If the delta computed from the time of the last observation were continued, it would likely become inappropriate and be different from surrounding stations so that a spatial discontinuity would be created. To address this potential problem, when the current error could not be computed, it was considered to be zero, so that the decayed average would drift toward zero. In other words, if the past short-term bias is not known, there can be no correction for it. (Biases from surrounding stations could be consulted, but that is not addressed here and is likely not worth the effort.)

As another practical matter, in an operational, automated system, the unexpected can happen, and in the situation studied here the MOS forecast or the observation could be highly erroneous, making the computed error disastrous. (There may be other error checks in the system, but they are likely not stringent enough to alleviate a problem here.) In order to avert potential disaster, a cap is put on the forecast error that is exactly 20 degrees for a 24-h forecast and 40 degrees for an 11-day forecast, with the cap defined by a linear line between those two projections and extending on either end as necessary. (For the projections used here, there are no projections outside those limits.) The large differences are still used, but they are capped. That is, if a 55 degree error occurred at 11 days, it would be used as 40 degrees. This would still be a shock to the correction algorithm, but manageable.

3. BOISE VERIFY ALGORITHM

This algorithm is simple in concept. A single predictor equation is fit on data from the previous 30 days with the dependent variable being the observation and the predictor being the forecast. Then for the new forecast, the corrected forecast is determined from the regression line. A "delta" is not actually calculated, but the difference between the forecast and the corrected forecast corresponds to the delta in the decaying average algorithm. There can be a difficulty when the new forecast falls outside the data domain of the regression line. That is, the new forecast is quite different from the past month. In that case, the regression line can be extended outside the past data, but that is dangerous and can lead to large unwarranted corrections. To address this problem, when the forecast is more than 3.0 standard deviations away from the mean of the 30-day sample, the regression is not used, and the forecast is corrected by the bias in the 30-day sample. In addition, there is a smooth transition between 1.5 and 3.0 standard deviations from full regression to full bias correction, by weighting the regression result and the bias

[1] Chi 2012, personal communication.

correction according to the closeness of the forecast to 1.5 and 3.0 standard deviations, respectively. Boise Verify allows options including the length of the regression sample and even the inclusion of data from previous years.[2] I have tested only the default values.

To make the test as comparable as possible, although not part of the Boise Verify algorithm, the error cap used in the decaying average was used here. If the cap was exceeded in the regression sample, that case was omitted to protect against outliers. In addition, I required ≥ 5 days on which to base a regression. That is, if there were < 5 cases during the past month that could be used in the regression (because of missing forecasts or observations), the regression and bias were not computed, and no correction was made to the forecast. As mentioned previously, this could happen if the station stopped reporting for a long period of time.

As one last feature in the testing, it was noted that there can be an abrupt change from a full regression correction to a full bias correction. This can happen when the range of the forecasts was small during the past 30 days, and the new forecast can easily exceed 3 standard deviations from the mean. The sample plotted as observations vs. forecasts may look like a blob, and the regression will have a very low reduction of variance. In fact, the slope of the regression line can be negative (correlation < 0). To try to mitigate the effect of this possibility, I used a variation that required the correlation to be > 0.44 for the regression to be used. That is, for all regressions with a correlation less than that value, the simple bias was used. The 0.44 is, of course, somewhat arbitrary, but corresponds to a t-test level of significance of 0.1 assuming 15 degrees of freedom in the 30-day sample. I tested the Boise Verify method both with and without this feature.

4. DATA USED

I had available operational GFS-based MOS 2-m temperature and 2-m dewpoint forecasts made at 0000 UTC over the period January 1, 2011, through May 31, 2012,[3] for projections every 12 hours out to 264 hours (11 days). This provided a sufficient sample on which to investigate the bias correction methods. A change was made to the Global Forecast System (GFS) at NCEP during the summer of 2011 that had some major negative effects on MOS in some portions of the United States, but primarily to wind rather than temperature and dewpoint. Previous testing with the decaying average method (Glahn 2012) showed that testing projections every 24 h was sufficient and adding the intermediate projections did not change the conclusions, so only the projections evenly divisible by 24 are tested here.

The testing was done over the whole sample period. Separation was not made between cool (October-March) and warm (April-September) seasons, even though MOS forecasts are developed that way. The results will be more prone to reflect the cool season because the variability and errors are higher there. The changing on April 1 and October 1 between warm and cool season

---

[2] Tim Barker, email.

[3] Being operational forecasts, the dewpoint forecasts had been checked with temperature forecasts, and if a forecast was greater than the temperature, it was set to the temperature. If bias correction were implemented centrally, the temperature/dewpoint check would come *after* the bias correction. Using the checked values of dewpoint for this study instead of unchecked ones is not seen as a problem, because the number and magnitude of the changes are rather small. Also, Boise Verify uses the checked forecasts at WFOs, because those are what are transmitted.

MOS equations can create a somewhat abrupt change in bias characteristics and has been recognized as a problem by field forecasters. If the decaying average bias correction method were essentially restarted on April 1 (the beginning of the warm season), it could (1) assume a cold start (no history of errors, d = 0), (2) use the d from September 30 of the previous warm season, or (3) use the d from the March 31 (the end of the cool season) `of the current year. It is not obvious which would be best. The regression method runs year round, so that the past 30 days are always used, corresponding to (3) above. Running a bias correction method continuously does not eliminate the shock at the boundary times created by MOS development being done by season, but adaption comes fairly rapidly.

5. PERFORMANCE

The performance metrics bias, MAE, fraction of large and small errors, ratio of variance of forecasts to variance of observations, and consistency over projections are presented here for the decaying averaging method with $\alpha = 0.04$ and for the regression method both with and without the check on the regression correlation being > 0.44. Also, a time plot of the performance for Crescent City, California, is shown.

The verifications are over the period February 15, 2011 through May 15, 2012, even though the data available covered the period January 1, 2011, through May 31, 2012. This provided enough data at the ends of the sample for calculations on previous days and verification out to 11 days. All figures show verifications for projections 24 hours out to 11 days at 24-h increments, where applicable. Each graph shows the uncorrected MOS forecasts and the MOS forecasts corrected with the decaying average method, the regression method, and the regression method with the check on correlation. The bias correction method was cold started on February 15, even though the regression method used the past 30 days; this gave a slight advantage to the regression method for the first few days. All forecasts were rounded to tenths of degrees F for verification, as were the observations.

Verification was done on 1,319 stations, the same ones used in routine MDL verification that have been judged to have reliable and accurate observations. Over the period, there were about 574, 500 cases verified, the exact number depending on projection. Of that number for the modified Boise Verify method, approximately 71% had corrections determined from the regression equation, 7% were corrected by the bias, 21% were corrected by the weighted regression and bias method, and less than 1% had no bias correction made.

A. Bias

Figures 1 and 2 show the bias for temperature and dewpoint, respectively. MOS forecasts had considerable more bias than any of the bias corrected ones. The bias is quite small for short projections and is higher for longer projections. Overall, the decaying average method was the best of the correction methods. The check on regression did not improve the regression method. Overall, all methods gave biases within $\pm$ 0.2 degrees F. Biases over shorter periods of time can be larger for MOS forecasts, and may be up to 5 degrees F or more.

B. Mean Absolute Error

Figures 3 and 4 show the MAE for temperature and dewpoint, respectively. MOS is improved on by the decaying average by about 0.25 to 0.30 degrees at the shorter projections to about 0.20 degrees at the longer projections. There is little difference in performance among the three bias corrections methods at the shorter projections, but the decaying average method is better at the longer projections where the check on regression correlation improves the results without the check. At projections longer than 168 hours for temperature, the basic regression method has higher MAEs than the uncorrected MOS.

C. Large Errors

Figure 5 shows the relative frequency of errors > 15 degrees F for temperature. The MOS and the decayed average bias corrected MOS are nearly identical for all projections. The regression method has more large errors, especially past 72 hours; the check on the correlation of the regression helped somewhat at the longer projections.

Figure 6 is the same as Fig. 5 but for dewpoint. The decaying average improves on uncorrected MOS at all projections. The regression method improves equally well as the decaying average at short projections but is slightly worse than uncorrected MOS at the longer projections, where the check on correlation helped only a bit.

D. Small Errors

Figures 7 and 8 show the relative frequency of errors < 5 degrees F for temperature and dewpoint, respectively. All methods of bias correction gave a higher frequency of small errors than MOS equally well for the short projections. However, the decaying average is best at longer projections for temperature; the basic regression almost ties with MOS and the check on correlation gives slightly better results. For dewpoint at longer projections, regression improved on MOS with the check on correlation being slightly better and essentially tied with the decaying average method.

E. Variance of Forecasts and Observations

Figure 9 shows the average ratio of the variance of the temperature forecasts to the variance of the temperature observations. Note this is not the ratio of all stations together, but the individual station ratios averaged over all stations. Regression produces forecasts that have less variability than the observations over the developmental sample, the ratio being the reduction of variance afforded by the equation. Figure 9 shows about the expected ratios over the range of projections for MOS. The decayed average did not change the ratio by a large amount, but tended to lower it as expected by the "smoothing" aspect of the algorithm. The regression method, on the other hand, gave somewhat larger ratios, averaged over all stations. Surprisingly, the corrected forecast ratios with the check on correlation were higher than those without the check for the longer projections. This could be caused by the bias correction being larger than the regression equation would have given in those cases when the correlation was < 0.44.

Figure 10 is the same as Fig. 9 except for dewpoint. The most striking thing is that the ratio for MOS approaches unity for the shortest projections. Because MOS "explains" only a portion of the predictand variance, the ratio is expected to be substantially below unity. This high ratio is probably attributable to the change(s) in the GFS model between the development of the equations and this sample. All bias correction methods performed about the same and the ratios were lower than MOS, as expected. Again, the correction with the check on correlation gave higher average variance than without the check.

F. Consistency over Projections

Long-projection forecasts have more error than short-protection forecasts. As the forecasts for a particular verifying time are improved with time, they should be as consistent as possible, and not "bounce around" from forecast to forecast and converge toward the verifying observation. The Convergence Score (Ruth et al. 2009) measures the tendency of the longer range forecast to march "consistently" toward the final short range forecast, a higher score being better with a possible maximum of 1.0. The parameters of the score used here were that (1) there was no penalty if the change was < 3 degrees F or the change was in the direction of the next forecast, and (2) the observation was not used as an anchor point.[4] With the latter option, the score in no way measured the accuracy, but only the consistency of forecasts from day to day.

Figure 11 shows the Convergence Score for temperature for each NWS coterminous region and overall. The decaying average method shows improvement over MOS for the Western and Central Regions and overall. The regression method, whether there is a check on the correlation or not, does not give as consistent forecasts; that is, they bounce more, given the definition stated above.

G. Performance on Individual Days

Figure 12 shows for the MOS 72-h temperature forecast the difference between the observation and the forecast for individual days over the period August 1 through December 1 of 2011 for Crescent City, California. (This is the negative of what we think of as the forecast error. It is plotted this way so it is apparent how the changes follow the differences.) Also plotted are the changes to the MOS forecasts for all three bias correction methods discussed above. This station was chosen for display only because there were some differences between the basic regression method and the variation that checked the correlation before the regression was used.

The MOS forecasts through much of August had little average error, and the corrections hovered around zero. Starting around September 1, the MOS forecasts were not high enough (had a negative bias), and the correction methods increased them by as much as 3 or 4 degrees during October. Because the bias was over several months, the correction methods improved the overall accuracy, as well as bias. The characteristics of the three methods are clear. The decaying average is conservative and changes little from day to day; see expanded view in Fig. 13.

---

[4] Options for the score are that the difference defining a change can be other than 3 degrees, and the last (most recent) "forecast" value be the observation itself. This latter extends the projections to 0 hours and introduces some measure of accuracy for the short projection forecasts. That is, if the difference between the 24-h forecast and the observation were ≥ 3 degrees, the score would be lower than if the observation were not used.

It seems throughout early- and mid-October, the decaying average method did not increase the forecasts enough, whereas the regression method was better. However, the opposite was true for several days shortly thereafter when the MOS errors became less, but were still of the same sign.

It is evident the regression method has more fluctuation day to day in the corrections, although for the data shown the differences day to day were not large. For about half of the period, the correlations were above 0.44 (the green and red lines overlap). However, for much of mid-September through October, the check on regression modified the corrections and smoothed them out, although the overall corrections were not necessarily for the better.

The corrections were positive during a period starting around September 1, which was a period of rather warm temperatures compared to the average (not shown). The forecasts were too low, and the corrected forecasts were higher. Because this was a high temperature period, the contribution to the variance of the corrected forecasts was then higher than the contribution to the variance of the uncorrected forecasts. This may help to explain why, as shown in Figs. 9 and 10, the correction methods sometimes produced variance ratios greater than MOS.[5]

6. DISCUSSION AND CONCLUSIONS

It was shown in MDL Office Note 12-1 (Glahn 2012) that the decaying average method of correcting MOS temperature and dewpoint forecasts could improve the bias and accuracy when an α the range 0.025 to 0.5 is used. This is the method used by NCEP for model forecasts with an α of 0.02. Many WFOs use a method of bias correction of various sources of forecasts provided in the software called Boise Verify. This is a regression method relating observations to forecasts over some period of days, 30 days being the default. If MDL were to implement bias correction of MOS forecasts, there might be an advantage to use either the method used at NCEP or at the WFOs. These two methods were compared for temperature and dewpoint on a sample of MOS data covering about 15 months. Because some of the regression relationships had low correlations, a modified procedure was also used in which the correlation had to be > 0.44 or the correction would fall back to the monthly bias. I did not investigate other alternatives within Boise Verify, assuming the default values were thought to be adequate.

Any bias correction method that will do well for overall accuracy cannot "track" the errors closely; the errors are not consistent enough for that. Rather, longer term biases on the order of a couple of weeks can be improved (lessened). The decaying average method for the $\alpha = 0.04$ used is fairly slow in responding to persistent errors, but larger values of ≥ 0.075, while improving biases, gave less accurate forecasts (see MDL Office Note 12-1). The regression method is appealing in that the correction for a specific day depends in a major way on the forecast for that day as well as the forecasts during the past month. In essence, the correction depends on the regression line and can be more (or less) for a high forecast than a low forecast. On the other hand, for the decaying average method, the effect of today's forecast error is limited to the amount specified by the α used. For this reason, the changes to the forecast are not as consistent day to day for the regression method as for the decaying average method. With the options built into Boise Verify and used here, the changes are kept within reasonable bounds.

---

[5] As further elaboration, positive corrections during times of high temperature and negative corrections during times of low temperature may make the variance of the forecasts larger than the variance of the observations.

The differences in metrics between the decaying bias and regression approaches are quite small except for the longer range projections, where the overall differences may be meaningful. For instance, at 11 days, the difference in the MAE of the two methods is on the order of two or three tenths of a degree F. Even the very small differences, because they are computed on a large number of cases (15 months and 1,319 stations) are highly significant. For instance, at the 72-h projection, the mean difference in MAE was only .041 degrees F, and this is statistically significant at the 5% level even if there are only ~ 2600 degrees of freedom in the sample of ~ 574,500 cases.[6]

The overall conclusions are that the decaying average method is better for correcting MOS temperature and dewpoint forecasts than the regression method. In addition, the decaying average is easier to implement because it only requires carrying along one value (per station per projection) and not the past 30 days (also per station and per projection) needed for the regression method. The only fail safe procedures needed for the decaying average method are to guard against very large errors (that might be caused by an erroneous observation or other errors that might crop up in an automated system) and a station that stops reporting for a long period of time. In the latter situation, the daily errors cannot be computed. To keep the same delta from being used indefinitely, zero error can be assumed and the delta will drift toward zero. For the regression method, checks are needed (1) to determine whether the past month's data are appropriate for using regression or whether the overall bias over the previous month is to be used, and (2) whether or not no correction is to be made at all if there are too few cases over the recent period to even furnish a bias.

These conclusions are based on forecasts made from 0000 UTC data valid at 0000 UTC 1 to 11 days in the future. It was shown in Office Note 12-1 that forecasts verifying at 1200 UTC led to the same conclusion, although the magnitude of the errors, etc. were different. Also, these results apply to the aggregation of 1,319 stations, and the characteristics for individual stations vary considerably.

ACKNOWLEDGMENTS

I am indebted to Tim Barker for not only furnishing me details of Boise Verify, without which I could not have made this comparison, but also for offering important suggestions on a draft. Also, I thank Jeff Craven for important information, Tamarah Curtis who was very helpful in providing details on graph production, and to David Rudack and Bruce Veenhuis for helpful discussions.

REFERENCES

Cheng, W. Y., and W. J. Steenbaugh, 2007: Strengths and weaknesses of MOS, running-mean bias removal, and Kalman filter techniques for improving model forecasts over the western United States. *Wea. Forecasting*, **22**, 1304-1318.

Cui, B., Z. Toth, Y. Zhu, and Hou, D., 2012: Bias correction for global ensemble forecast. *Wea. Forecasting*, **27**, 396-410.

[6] This calculation extends to spatial correlation the logic in Wilks (2011, p. 147) where the "effective sample size" for use in calculating the sampling variation of the sample mean is based on the autocorrelation of the variable.

Galanis, G., and M. Anadranistakis, 2002: A one-dimensional Kalman filter for the correction of near surface temperature forecasts. *Meteorol. Appl.*, **9**, 437-441.

Glahn, B., 2012: Bias correction of MOS temperature and dewpoint forecasts. *MDL Office Note* **12-1**, Meteorological Development Laboratory, National Weather Service, NOAA, U.S. Department of Commerce, 33 pp.

Homleid, M., 1995: Diurnal corrections of short-term surface temperature forecasts using the Kalman Filter. *Wea. Forecasting,* **10**, 689-707.

Kalman, R. E., 1960: A new approach to linear filtering and prediction problems. *Trans. ASME. J. Basic Eng.* **82**, 35-45.

Roeger, C., R. Stull, D. McClung, J. Hacker, X. Deng, and H. Modzelewski, 2003: Verification of mesoscale numerical weather forecasts in mountainous terrain for application to avalanche prediction. *Wea. Forecasting,* **18**, 1140-1160.

Woodcock, F., And C. Engel, 2005: Operational consensus forecasts. *Wea. Forecasting* **20**, 101-111.

Yusssouf, N., and D. J. Stensrud, 2007: Bias-corrected short-range ensemble forecasts of near-surface variables during the 2005/06 cool season. *Wea. Forecasting*, **22**, 1274-1286.

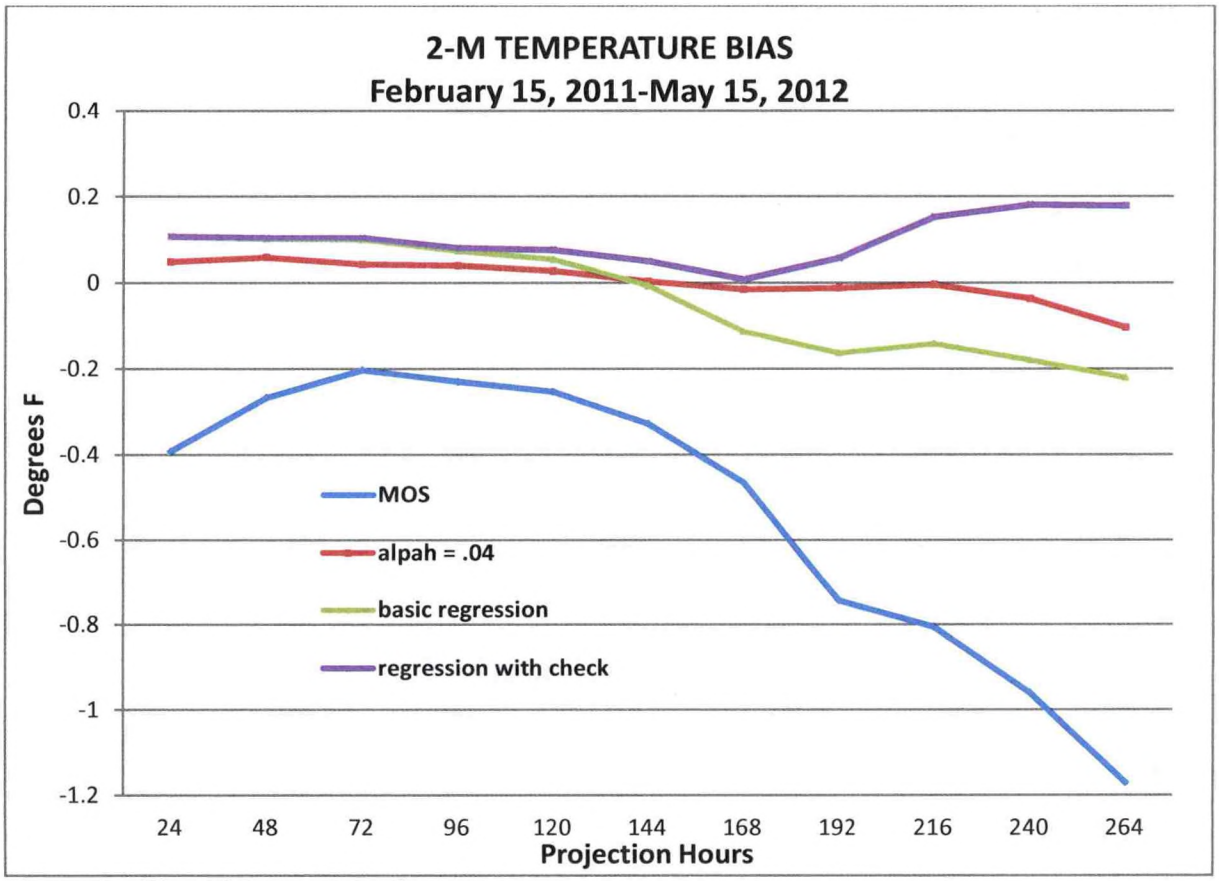

Figure 1. Bias of temperature forecasts.

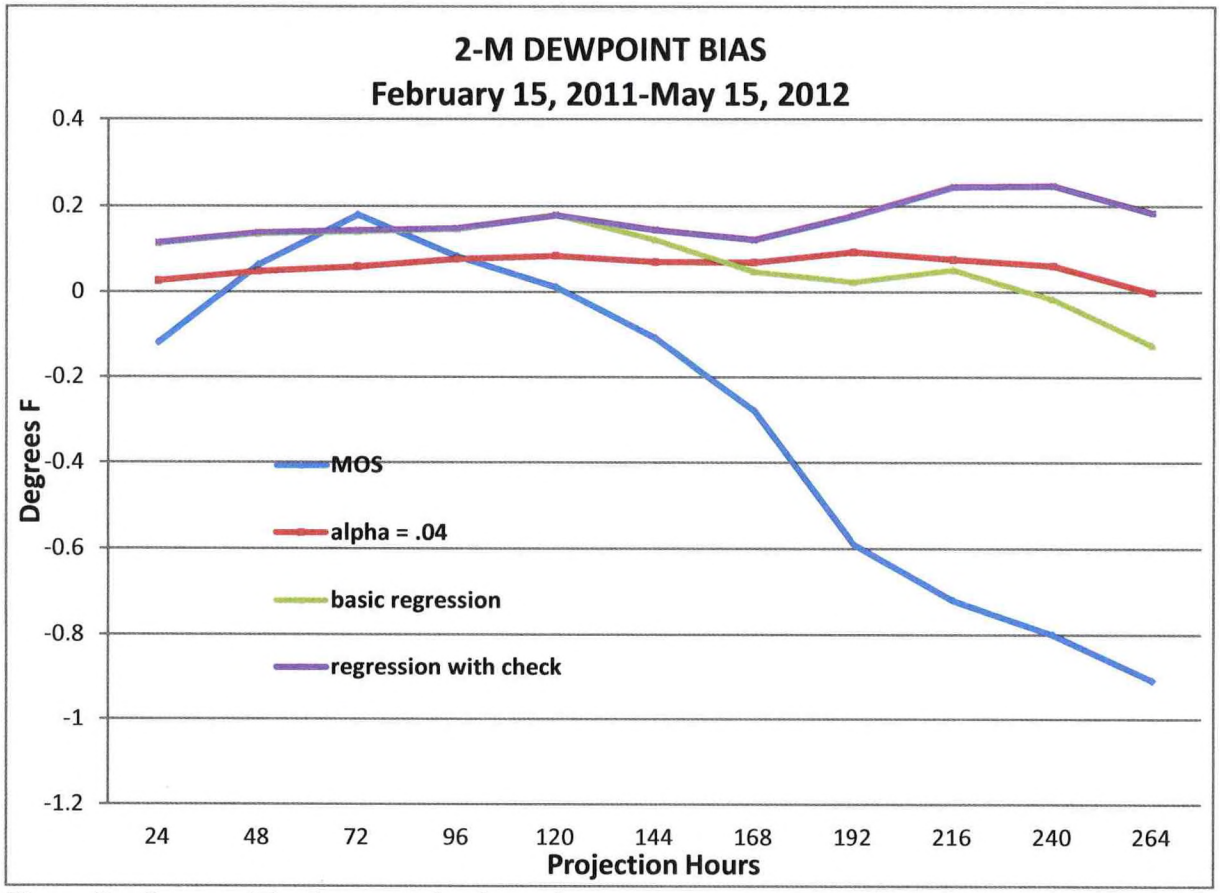

Figure 2. Same as Fig. 1, except for dewpoint.

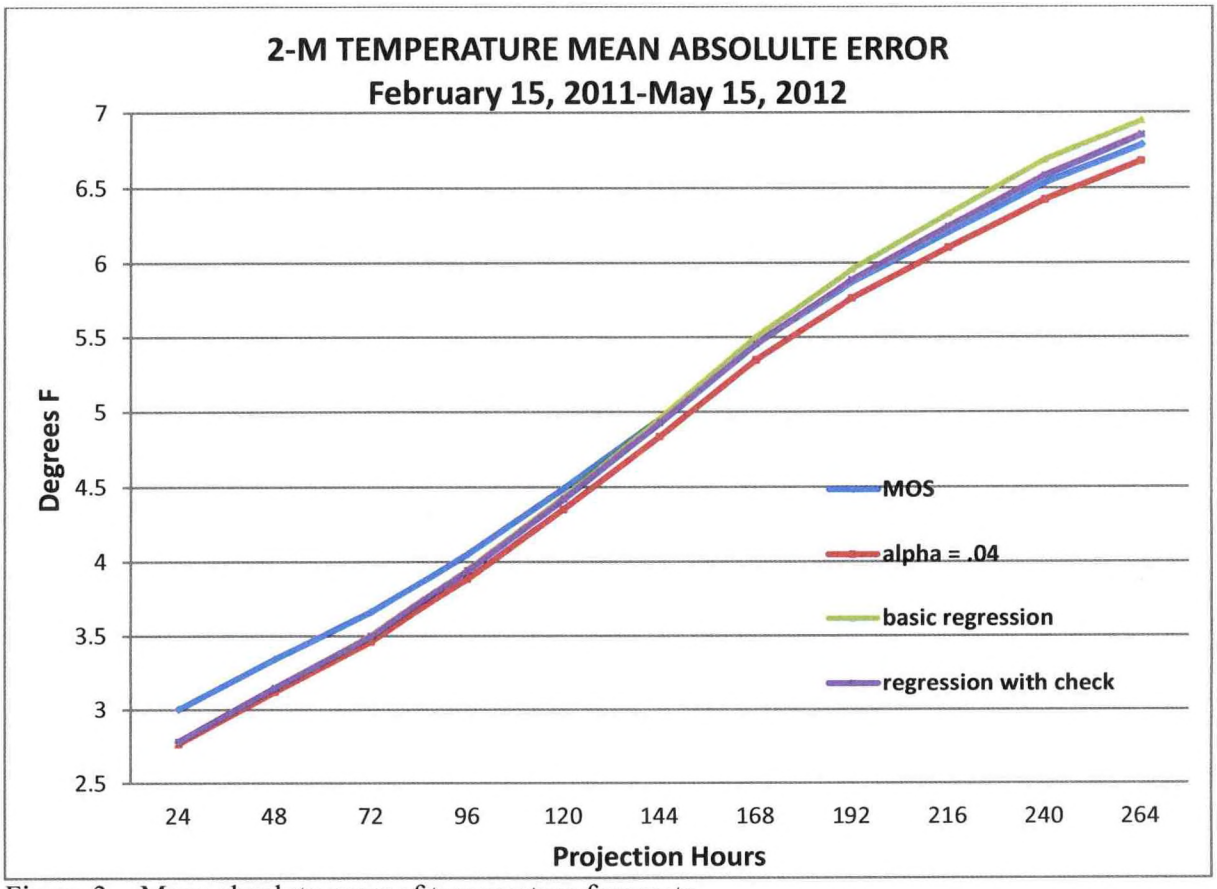

Figure 3. Mean absolute error of temperature forecasts.

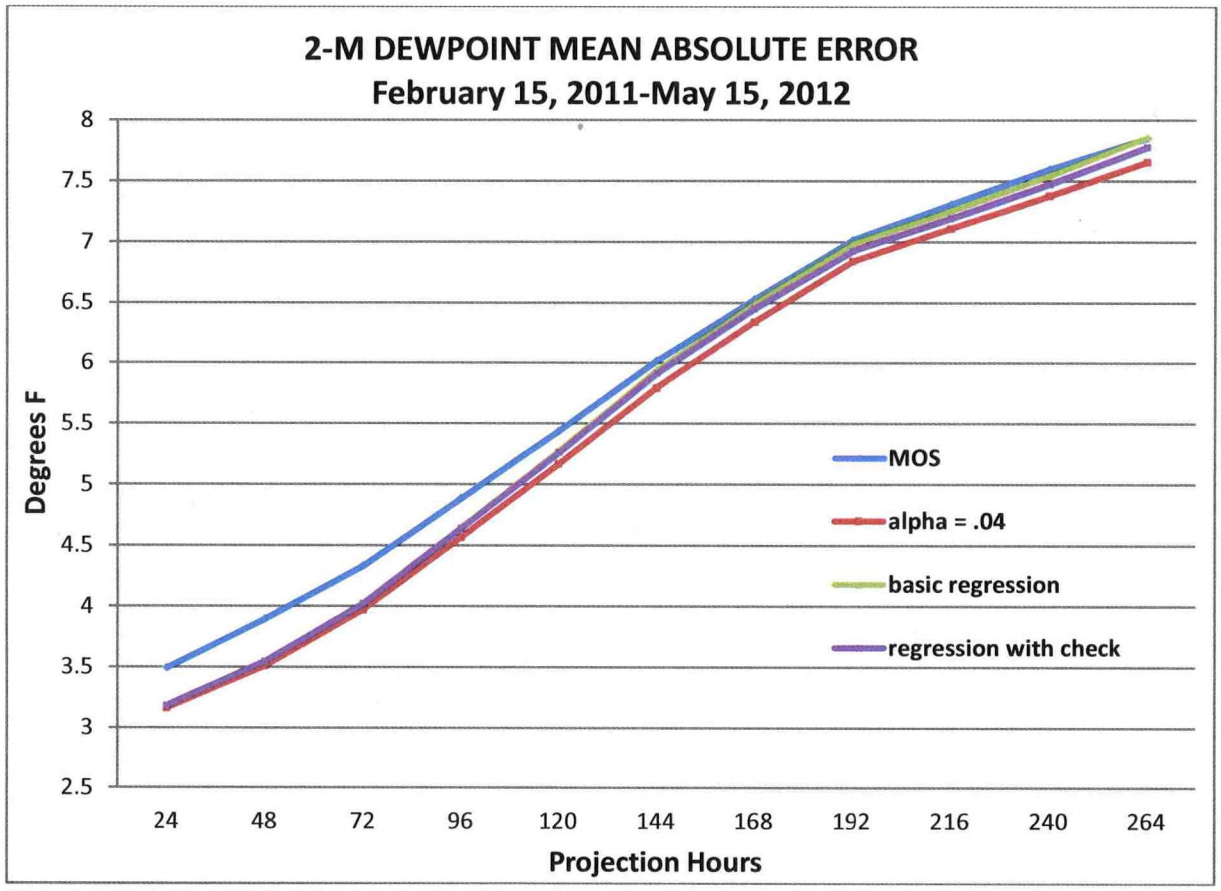

Figure 4. Same as Fig. 3, except for dewpoint.

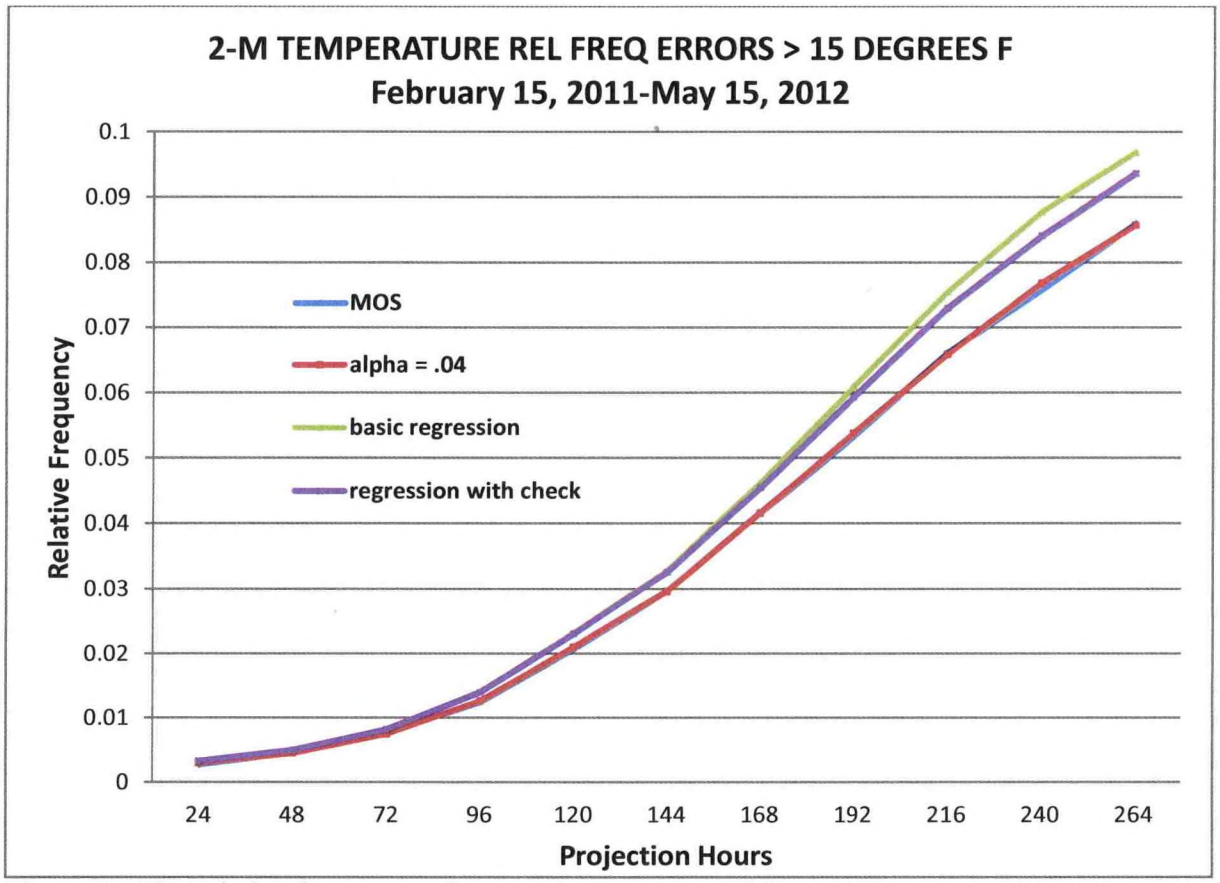

Figure 5. The relative frequency of temperature errors > 15 degrees F.

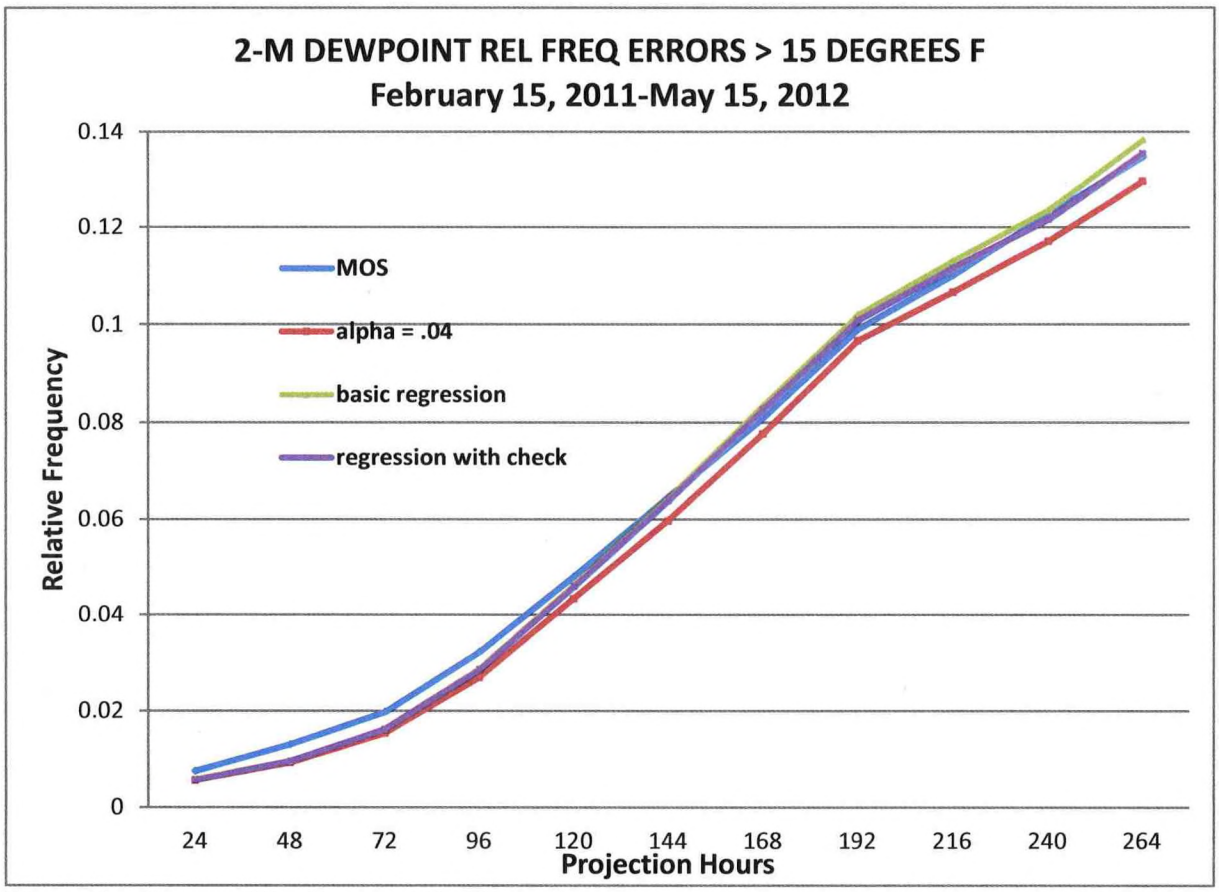

Figure 6. Same as Fig. 5, except for dewpoint.

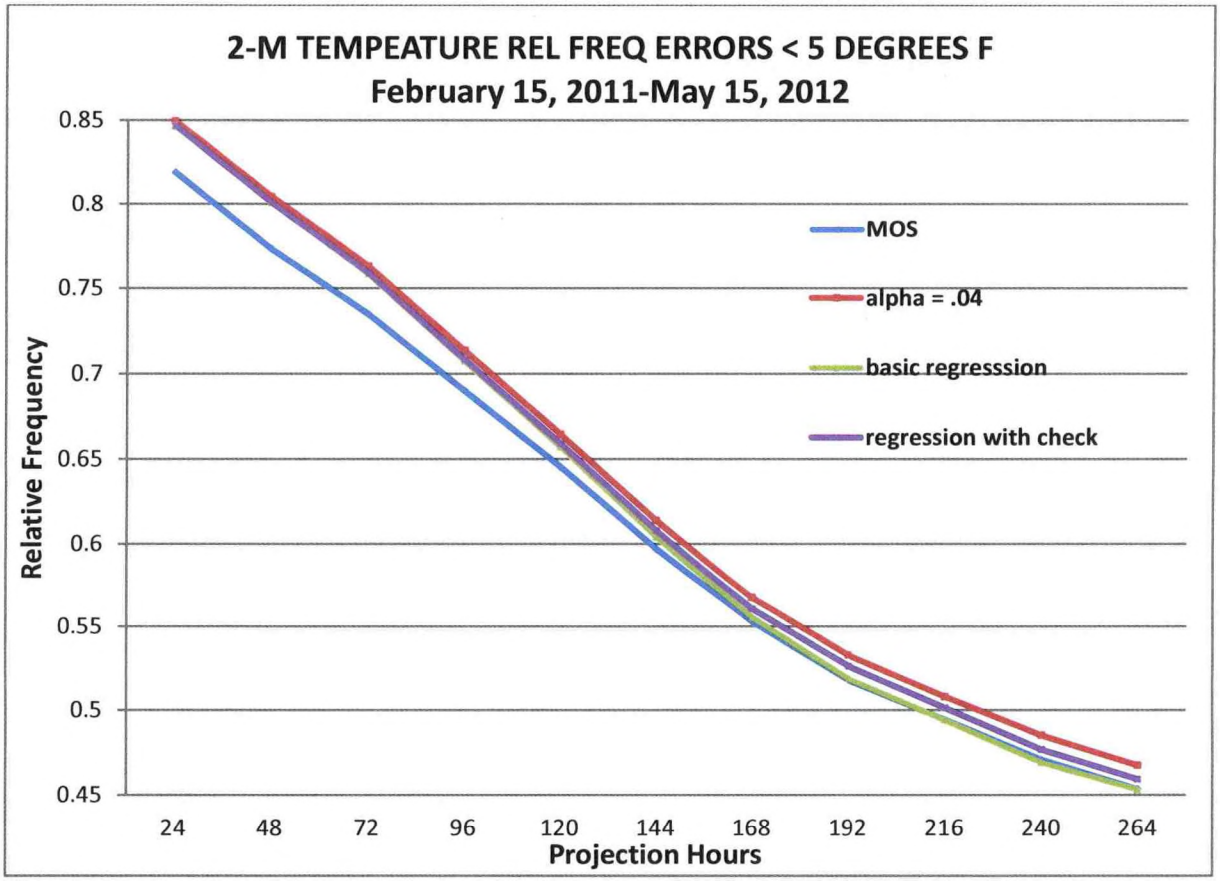

Figure 7. Relative frequency of temperature errors < 5 degrees F.

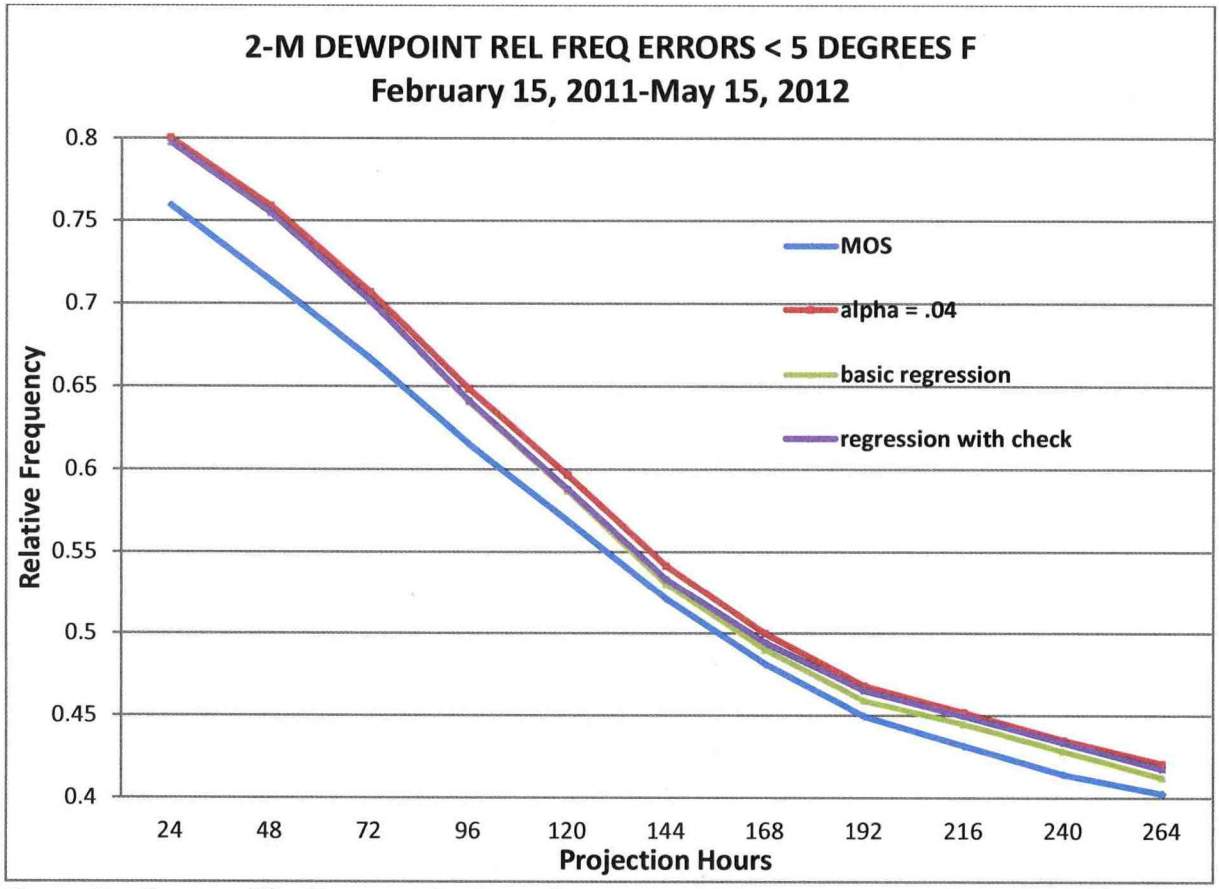

Figure 8. Same as Fig. 7, except for dewpoint.

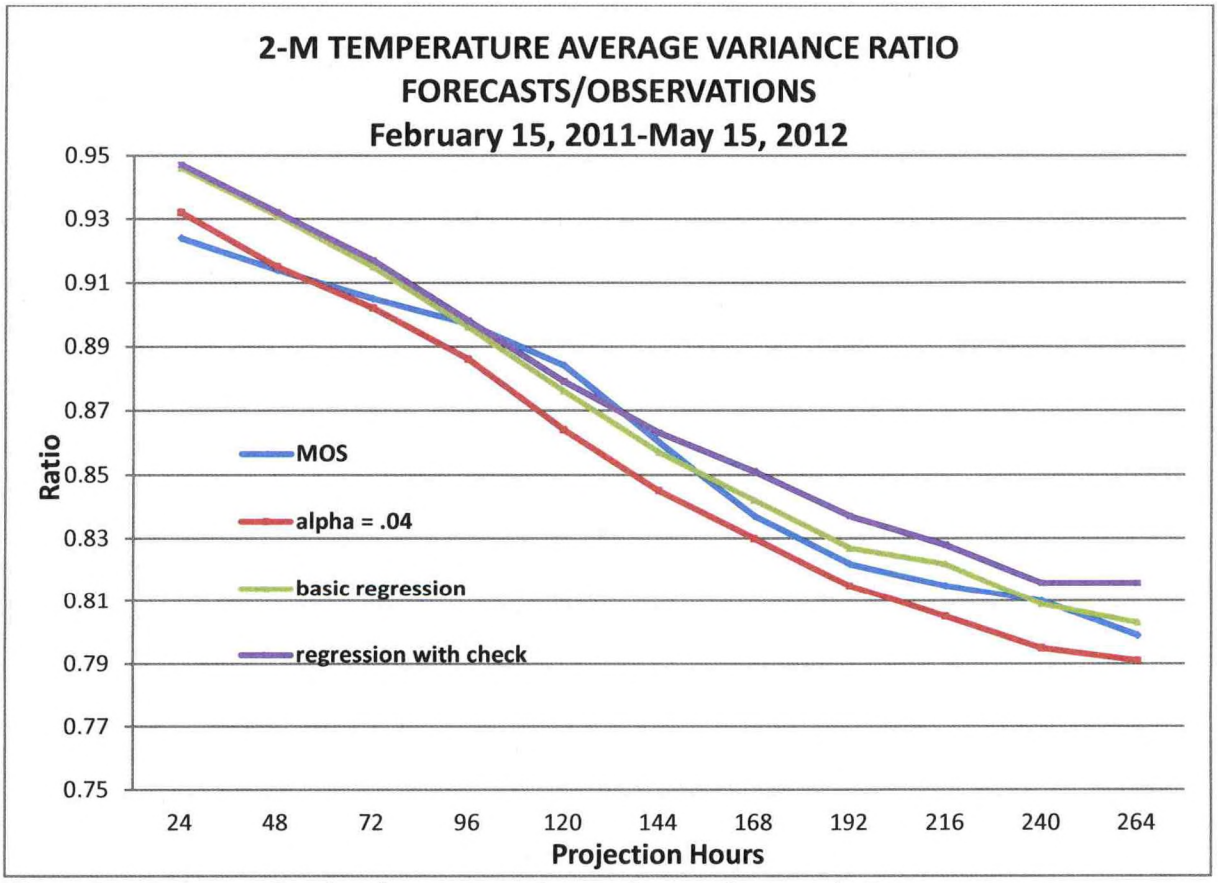

Figure 9. Average of station forecasts to observations variance ratio for temperature.

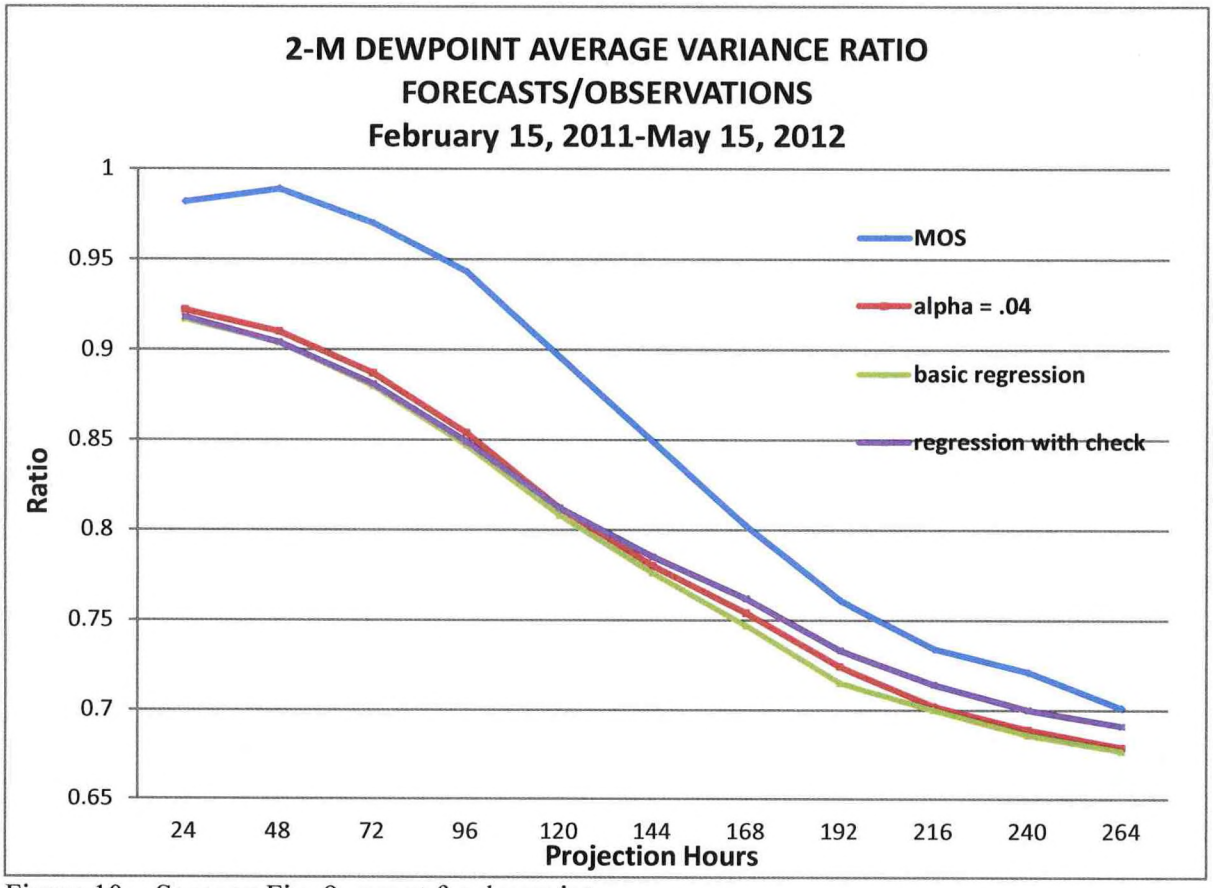

Figure 10. Same as Fig. 9 except for dewpoint.

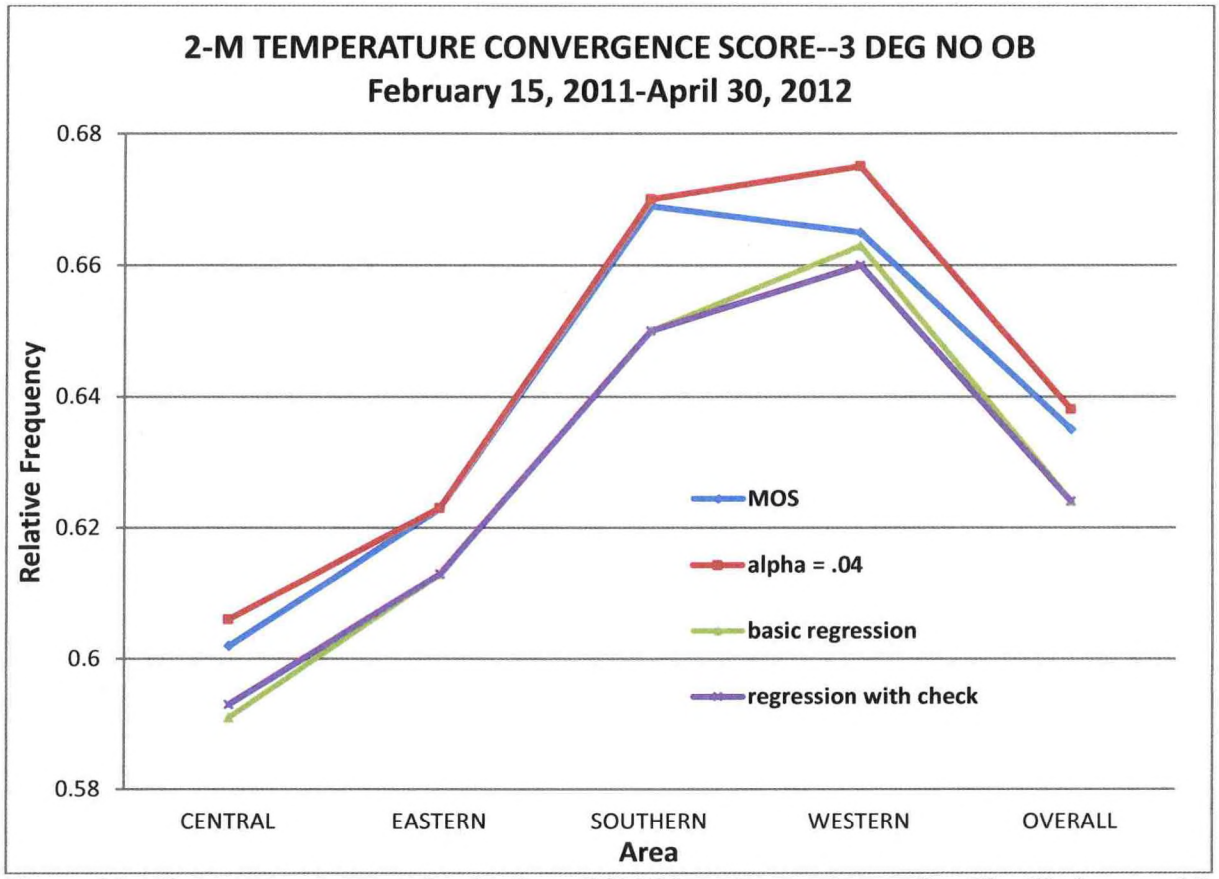

Figure 11. Convergence scores for the four CONUS regions and overall. The "3 deg no ob" refers to parameters for the score. Here, there is no penalty to the score if the change is < 3 degrees F, and the observation to which the forecasts aspire is not used as the 0-h anchor point.

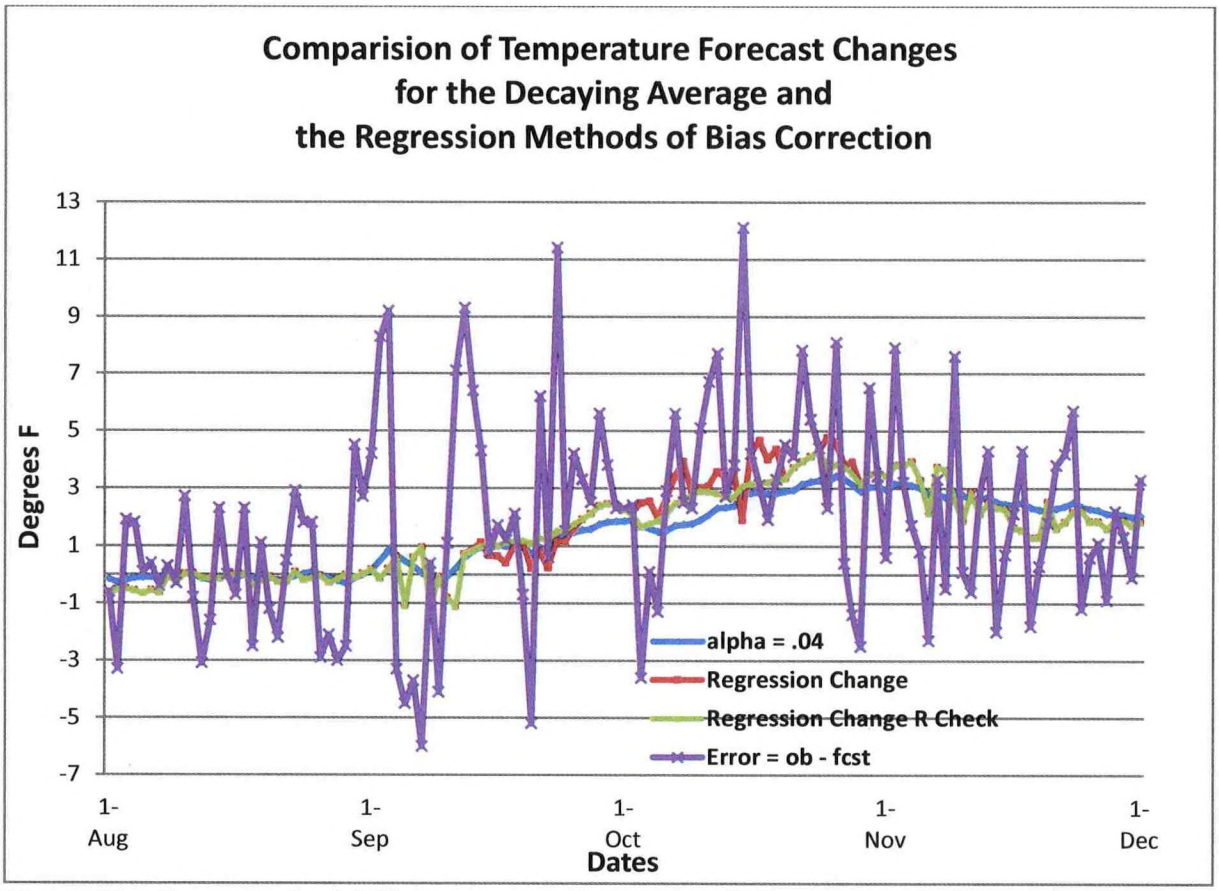

Figure 12. Errors in the MOS temperature forecasts and the associated changes for the three correction methods for Crescent City, California. The errors are plotted on the dates at which they occur, and the changes to the forecasts are those made to the 72-h forecast made on that date.

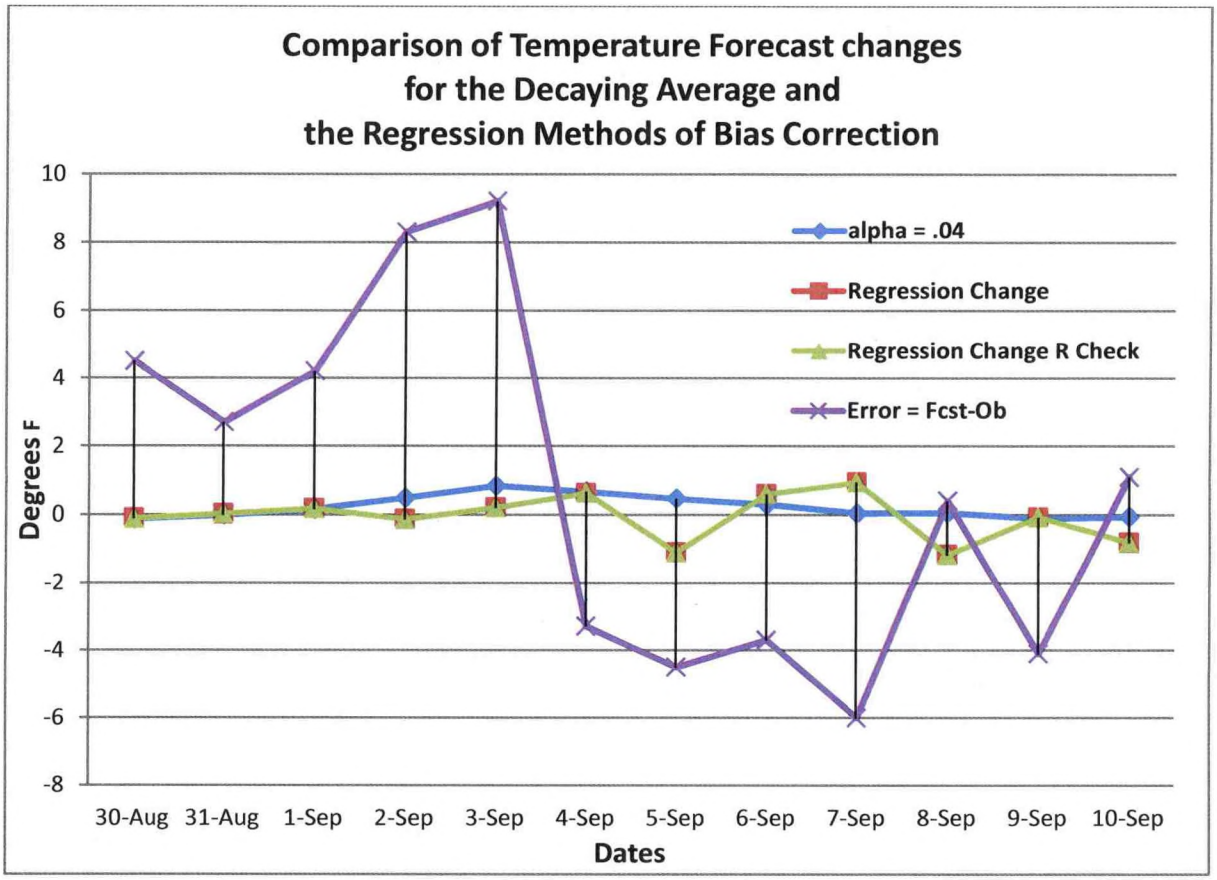

Figure 13. Expanded view of errors in the MOS temperature forecasts and the associated changes for the three correction methods for Crescent City, California. The errors are plotted on the dates at which they occur, and the changes to the forecasts are those made to the 72-h forecast made on that date.