Determining an Optimal Decay Factor for Bias-Correcting MOS Temperature and Dewpoint Forecasts

BOB GLAHN
Meteorological Development Laboratory, NOAA/National Weather Service/Office of Science and Technology, Silver Spring, Maryland

(Manuscript received 24 October 2013, in final form 17 April 2014)

ABSTRACT
Model output statistics (MOS) forecast relationships for temperature and dewpoint developed with least squares regression and put into operation by the National Weather Service (NWS) are unbiased over the sample period of development. However, short-term biases within that period can exist, and application of the regression equations to new data may produce forecasts with short- or long-term biases. Because NWP models undergo changes over time, MOS forecasts can be biased because of these changes, and also possibly because of local environmental changes. These biases can be largely eliminated. In the decaying average method, a “decay factor” is used. This value affects not only the short- and long-term bias characteristics, but also other accuracy measures of the forecasts. This paper shows how different values of the decay factor affect MOS temperature and dewpoint forecasts, and the range of factors that would be appropriate for bias correcting those forecasts. Biases and other quality measures are shown for both cool and warm season samples before and after various values of the decay factor have been applied.

1. Introduction
The National Weather Service (NWS) has continuously provided model output statistics (MOS; Glahn and Lowry 1972) dry-bulb temperature and dewpoint temperature forecasts to customers and partners since their introduction in 1978 (TPB 1978). Initially, the NWP products on which MOS forecasts were based did not contain forecasts of “surface” weather elements, such as 2-m temperature and dewpoint and 10-m wind, and postprocessing was crucial in the use of those NWP products. Today, NWP does produce these weather elements directly, and their accuracy makes them prime predictors in the MOS regression equations. Even so, NWP forecasts can be improved by post-processing, and in particular the NWP forecasts have biases. NWP upper-atmospheric variables and some near the surface have been bias corrected at the National Centers for Environmental Prediction (NCEP) since 2006 by a method called “decaying average” (Cui et al. 2012).

MOS forecast relationships developed by the Meteorological Development Laboratory (MDL) have usually been based on so-called cool (October–March) and warm (April–September) season samples. This separation accounts to some degree for the annual biases and inaccuracies that would exist because of different relationships between predictors and predictands in the warm and cool months. Regression equations that produce MOS forecasts give unbiased estimates over the developmental sample period, but they may have biases over intervals within that sample and over other samples, including future forecasts. Biases in operational forecasts can exist because of NWP model changes, which render the developed relationships less than optimal; changing weather regimes that the equations do not handle adequately (e.g., blocking highs); or local environmental changes. A source of MOS error is the inability to “keep up” with the operational model changes. Changes in a model, without redeveloping MOS on
an adequate sample, can create larger errors in MOS forecasts.\(^2\) Short-sample bias correction is seen as a way of correcting such biases, and now that NWP models are much better than a few years ago, possibly bias-corrected raw NWP 2-m temperature is of comparable accuracy to MOS temperature. However, Cheng and Steenburgh (2007) found that in 2003 the MOS forecasts were better than bias-corrected model forecasts of temperature except in periods of “quiescent large-scale patterns.”

According to the usual definition of bias used in meteorology (Wilks 2011, p. 258; Jolliffe and Stephenson 2003, 99–100; Murphy and Daan 1985, p. 385), if the average of a consecutive set of errors is sufficiently different from zero, the forecasts would be considered to be biased.\(^3\) However, most of the authors discussing bias do not specifically state over what period of time the errors would have to be consistently above or below zero for them to consider the forecasts biased.

Routine verification has shown that the MOS temperature and dewpoint forecasts currently being made have some small but consistent long-term bias (Glahn et al. 2009, their Fig. 11). This paper shows the results of investigating the adaptability and effectiveness of applying the decaying average method to those forecasts.

### 2. Decaying average algorithm

NCEP implemented in 2006 a decaying average algorithm to reduce bias (Cui et al. 2012). Other authors have investigated other methods of correcting bias (e.g., Roeger et al. 2003; Homleid 1995; Galanis and Anadranistakis 2002; Crochet 2004; Cheng and Steenburgh 2007), success usually being measured as an average over a relatively long period (e.g., Yussouf and Stensrud 2007). The NCEP algorithm is attractive because it is very easy and cost effective to implement.

To implement the decaying average algorithm, one has only to compute a delta based on the current and past errors and carry it forward to add to the forecast made at the next forecast cycle. The delta \(d\) is computed by

\[
  d_{t+1} = (1 - a)d_t + a(F - O)_t,
\]

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 \(a\) 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 and forecast projection. If \((F - O)\) is occasionally missing, zero can be assumed (see the next section for more discussion of this situation).

NCEP has implemented the algorithm at grid points. The error \(F - O\) is the difference between the forecast at...
a grid point and an objective analysis of observations at that grid point. This is an attempt to make the forecast be the same as the analysis, which can itself have considerable error and bias. In the application to MOS, the $F - O$ is the difference between the MOS forecast and the verifying observation at a station.

As indicated above, a “data sample” is not used except as it is represented by the delta. The “decaying” of the influence of the error $F - O$ at a particular time is indicated in Fig. 1 of Cui et al. (2012). A larger decay factor will cause the delta used at one forecast time to decay faster than will a smaller one. There is no developmental sample or separate test (or independent) sample as there is in most postprocessing systems.

### 3. Performance of the algorithm applied to MOS forecasts

Before implementing bias correction into the MOS system, several questions have to be addressed, including not only whether the accuracy and bias of the forecasts are improved, but also what effect this might have on the customers and partners who use the “forecasts” directly or manipulate them by either automated or manual means to achieve a “final” forecast. In the NWS, this further processed guidance would be the “official” forecast.

#### a. Data and processing

I used the operational Global Forecast System [GFS (Caplan et al. 1997)]-based MOS temperature and dewpoint forecasts for projections every 12 h out to 264 h (11 days) made at 0000 UTC over the period from 1 January 2011 through 31 May 2012. This provided a sufficient sample on which to investigate the bias correction method.

Testing on temperature was done separately on the traditional MDL warm (March–September) and cool (October–March) seasons; however, for dewpoint, all 17 months of data were processed together. The comparison of MOS forecasts and bias-corrected (BC) forecasts for a particular projection involved matching samples. The verification was done with forecasts rounded to whole degrees Fahrenheit, the same precision as the observations. The deltas, however, were carried to three decimal places.

Errors in automated, data-driven processes occasionally occur. An extremely erroneous forecast or observation could cause a very large change in a delta. To address this possibility, a cap on the error can be imposed, and the testing reported here was done with

---

4 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 corrections were implemented, the temperature/dewpoint check would logically 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 for consistency are rather small and do not materially affect the temporal characteristics of the forecasts.
such a cap. The magnitude of the error allowed for a 24-h forecast was 20°F; for an 11-day forecast it was 40°F and it varied linearly in between by projection. For instance, if an error of 90°F were calculated for an 11-day forecast, likely due to an incorrect observation, it would be capped at 40°F, which is still a sizeable error, but the effect would not be disastrous. In an operational setting, it is likely that large errors would have already been culled out with quality control procedures.

b. Performance of different decay factors

NCEP and MDL have investigated different decay factors up to at least 0.1; NCEP uses 0.02 for all projections for raw model data (B. Cui 2012, personal communication.). Testing at MDL has indicated that a higher value might be better for MOS. I tested four different values: 0.025, 0.05, 0.075, and 0.1. I used 1319 stations in the conterminous United States (CONUS), the same stations used in routine MDL verifications. Of...
course, a few observations, and even forecasts, may be missing in the sample. The adjustment algorithm can deal with occasional missing data; if there is no forecast or matching observation for a particular projection, no change in the delta is necessary, and on the next cycle for which there is a forecast and observation, the delta last calculated could be used. However, a station may miss reporting for an extended period or stop altogether. MOS forecasts will likely still be made because the model data are available. To perpetuate a particular delta indefinitely would not be prudent, so a better alternative would be to use the difference as zero and let the delta decay gradually to zero. That is, given no recent history of the station’s bias, a correction is not made. In either case, no elaborate backup software is necessary to accommodate one or more missing pairs of data. For the testing reported here, the decay toward zero was not used.

1) BIAS

Figure 1 shows the bias for the original MOS temperature and the BC forecasts with all four decay factors tested for all stations and projections for the 2011–12 cool season. The forecast process was “cold started” with a delta of zero on 1 January 2011, and ran continuously. It can be seen that there was a significant cold bias in the MOS forecasts that varied each 12 h and generally was more pronounced with increasing projection. Forecasts verifying at some times of the day have larger errors and different error characteristics than forecasts verifying at other times—that is what causes the sawtooth effect. It can also be seen that the BC forecasts are better in terms of bias. They are slightly positive at 24 h, near zero at 72 h, and have a cold bias approaching 0.4°F at later projections. It is interesting that for the early projections, the simulations with the largest negative bias for MOS have the largest positive bias for BC MOS. This mirror performance lasts until about 228 h, when it reverses and the two curves come into phase. The bias was improved substantially for all decay factors. The largest improvement was for 0.1 and the smallest for 0.025 for all projections, but the differences are small, especially compared to the MOS bias.

The bias varies considerably over the CONUS. Figure 2 is the same as Fig. 1, but for the NWS Central Region, the region for which the MOS bias was greatest, and similarly, Fig. 3 shows biases for the Western Region, the region for which the MOS bias was least (note the different ordinate scales in these figures). Figure 2 shows

---

5 Because the cool season started on 1 October 2011, the process was well established. However, MOS development is done in 6-month periods, so the MOS equations used on 1 October are not those used on 30 September. Therefore, a short adjustment period would have occurred.

6 The NWS has six operational regions, four in the CONUS plus Alaska and the Pacific. In the CONUS, the Western Region consists generally of the Rocky Mountains and westward; the Central Region is the north-central CONUS, including as far south as Missouri; the Southern Region lies just south of the Central Region; and the Eastern Region consists of the eastern CONUS including Ohio on the west and South Carolina to the south.
a remarkably negative bias of 3°F at 264 h; the BC forecasts are better for every projection, but still drift systemically down to −0.6°F at 264 h for \( a = 0.5 \). On the other hand, the biases for the Western Region range generally from −0.5°F to +0.8°F, and the BC forecasts are better for most projections, especially for the larger values of \( a \).

Figure 4 is the same as Fig. 1, but for the 2011 warm season, April–September. The pattern of MOS error by projection is dissimilar to the cool season, and the improvement is questionable except for projections ≤72 h. For projections of 84 h and beyond, the MOS bias was quite small, and the correction was not, in general, helpful. As with the cool season, \( a = 0.1 \) was the best and 0.025 the least helpful.

Rather than separate the dewpoint data into seasons, the whole sample from January 2011 through May 2012 was used. For each test, the process was cold started with a \( \delta = 0 \) on 1 January. Because of this, the verification period for dewpoint started on 15 January, allowing a stabilizing period. Figure 5 shows the results. Again, MOS bias is improved, and all decay factors produce a very small positive bias; the small values were best, but the differences are minuscule. Probably one reason the biases of the corrected forecasts do not vary much and are so close to zero is because of the long averaging period over all 16.5 months (15 January 2011–31 May 2012).

2) MEAN ABSOLUTE ERROR

Figures 6 and 7 show mean absolute errors (MAEs) for all stations for MOS and the four values of decay factor tested for temperature for the cool and warm seasons, respectively. Only forecasts verifying at 0000 UTC are shown. Forecasts verifying at 1200 UTC show the same pattern, but the errors are considerably larger, and if both are shown on the same graph, the sawtooth pattern would present a less clear picture. One might hope with the improvement in bias, that the MAEs would also improve (decrease). Indeed they do, but not with all values of decay factor. For the lower values of 0.025 and 0.05, there is consistent improvement, but with the two higher values of 0.075 and 0.1 for projections >144 h, there are larger MAEs in the cool season. Because the improvement with the smaller values of decay factor is consistent for all projections, especially for the warm season, and because the paired Student’s \( t \) tests for each projection show very high significance, some slight improvement can be expected in the future.

Figure 8 is similar to Figs. 6 and 7, but for dewpoint. The conclusion for dewpoint is the same as for temperature; the improvement with bias correction is consistent for all projections, on the order of 0.3°F at lower projections; improvement is small at higher

---

7 The paired \( t \) tests were corrected for first-order autoregression (Katz 1985, p. 275; Wilks 2011, p. 147). The errors of the MOS and the bias-corrected forecasts are highly autocorrelated; the MAE pairs are also autocorrelated, but less so. Even allowing a reduction in degrees of freedom by a factor of 10 for spatial correlation, the paired \( t \) tests show high significance.
projections. The improvement does not vary much with decay factor, but the lower values are better at higher projections.

3) SMALL AND LARGE ERRORS

Figures 9 and 10 show the percentages of small errors, those <5°F, for temperature for the cool and warm seasons, respectively. These small errors are more frequent with all decay factors, but the smaller values give slightly better results. The improvement is roughly equal to a 24-h improvement for the warm season. The same general conclusion is reached from Fig. 11, which is for dewpoint; the relative frequencies of small errors are higher for BC forecasts than for MOS, and lower decay factors are slightly better.

Figures 12 and 13 are similar to Figs. 9 and 10, but they are for large errors, those ≥15°F. There are very few such errors for the short projection times, and errors
reach 1% at 11 days for the cool season. For that season, \( a = 0.025 \) is the only scenario that actually has fewer large errors than MOS, but \( a = 0.05 \) does not have more.

For the warm season, \( a = 0.1 \) is the only case that does not improve on MOS for the higher projections, with 0.025 and 0.05 being of about equal value according to this score.

Figure 14 shows the percentage of large errors for dewpoint. Improvement over MOS holds for all projections, but only for \( a = 0.025 \) and 0.05 at longer projections.

4) CONSISTENCY OF FORECASTS OVER PROJECTIONS

Long-projection forecasts have more error than short-projection 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. The convergence score (Ruth et al. 2009) measures the tendency of the forecasts to march “consistently” from the longer-range forecast toward the final short-range forecast, a higher score being better.

![Figure 9](https://example.com/Figure9.png)

**Fig. 9.** The percentage of temperature errors <5°F for the cool season.

![Figure 10](https://example.com/Figure10.png)

**Fig. 10.** As in Fig. 9, but for the 2011 warm season.
with a possible maximum of 1.0. Figures 15 and 16 show this score for the four NWS regions and overall for temperature for the cool and warm seasons, respectively. For the cool season, the higher decay factors give worse results than MOS for all except the Western Region; \( a = 0.025 \) and 0.05 have essentially the same or better scores than MOS. For the warm season, \( a = 0.1 \) gives worse results than MOS except in the Western Region. The lower decay factors are generally the best, the results for \( a = 0.025 \) and 0.05 being essentially indistinguishable.

Figure 17 shows the convergence scores for dewpoint for the warm and cool seasons combined. Only \( a = 0.025 \) is able to be about as good as MOS for all regions, but \( a = 0.05 \) is about the same overall. The scores for \( a = 0.075 \) and 0.1 are consistently worse than for MOS.

5) BIAS BY AVERAGING TIME

The biases over the sample have been shown in previous sections. However, biases can be present over shorter periods, may be negative for one period and
positive for another period, and may cancel out over the sample. To see what effect the bias correction has on biases of a shorter time period, running means of forecasts and observations were computed over a 20-day period—long enough that someone might consider consistently high (or low) forecasts to be “biased.” Then, the MAEs of these running averages were computed. This measures the biases, both positive and negative, over 20-day periods. Figure 18 shows the results for the warm season temperature. All values of the decay factor show improvement over MOS, with the larger values giving the smaller biases, consistent with the biases over the whole samples. The results, not shown, are similar for the cool season, as they are for dewpoint for the cool and warm seasons combined as shown in Fig. 19.

Figure 20 shows the 72-h forecast MAEs with running means of various lengths up to 30 days. The conclusion is the same, regardless of averaging time; short-term local biases are less for BC forecasts than for MOS for all decay factors. This means that if a user is interested in mean forecasts, as opposed to daily nonaveraged forecasts, bias-corrected forecasts are better than uncorrected, and larger decay factors, at least up to 0.1, the
maximum value tested, are better. The value at 1 day would be exactly the same as shown in Fig. 6 for 72 h if the sample periods were the same; for averaged forecasts, the ends of the period are slightly different. It is emphasized that the values in Fig. 20 account for the bias being either positive or negative over the $x$-day period, and are not the same as the overall biases shown in other figures.

6) PERFORMANCE ON INDIVIDUAL DAYS

Figure 21 shows the MOS 48-h forecasts for Appleton, Minnesota, and the deltas for the four values of $a$ for the cool season forecasts from 1 February through 28 November 2011, with the warm season forecasts from April through September omitted. The deltas operative on 31 March were used on 1 October. During

![Figure 15](image1.png)

**FIG. 15.** Convergence scores for the four NWS CONUS regions and overall for the cool season. The score for $a = 0.025$ is very close to that for $a = 0.05$, and can hardly be seen.

![Figure 16](image2.png)

**FIG. 16.** As in Fig. 15, but for the warm season. The score for $a = 0.025$ is essentially the same as for $a = 0.05$, and can hardly be seen.
March, the forecasts were generally too high. As this period started in late February, the deltas became negative, and for \( a = 0.1 \) reached a low of \(-5.0^\circ F\). For the first few days in October, the MOS forecasts were decidedly too low. The deltas rose rather quickly, and for \( a = 0.1 \) became as high as \(+2.5^\circ F\). But by the time this peak was reached, the period of consistently low forecasts was over, and then a march back toward zero delta began.

The delta for \( a = 0.025 \) was much more conservative, never reaching a negative value of more than about \(2^\circ F\). It, too, rose starting 1 October, but never peaked above zero. The other decay factors performed at an intermediate manner as expected.

4. Discussion

It is clear from the data presented that the more volatile nature of larger decay factors, while improving bias, hurt the overall quality of the MOS forecasts. Also, the deltas (corrections) to apply to the MOS forecasts can vary quite dramatically over a few days with large
decay factors; it is doubtful that users would welcome this volatility. The decay factor $a = 0.025$ is quite conservative and is competitive with $a = 0.05$, although 0.05 is slightly better in providing accuracy as judged by some scores for some projections. A decay factor in that range (0.025–0.05) seems to be best overall.

5. Conclusions

The decaying average algorithm has been tested with decay factors ranging from 0.025 to 0.1. It has been found that all values improve the bias of MOS temperature and dewpoint forecasts, but $a = 0.025$ and 0.05 consistently provide more improvement in the MAEs. These results hold for both warm and cool season temperature and for dewpoint over a period consisting of both warm and cool seasons. It is noted that the biases are not only reduced over a several-month period (where the pluses and minuses can cancel out), but also over periods from a few days to up to a month (see Figs. 18–20). The lower decay factors also provide more forecasts of less than 5°F error and fewer forecasts with

---

**Fig. 19.** As in Fig. 18, but for the cool and warm seasons combined for dewpoint.

**Fig. 20.** MAEs of 72-h forecasts as a function of averaging period ($a = 0.075$ is not shown).
>15°F error. In addition, only the lower factors improve the consistency of the MOS forecasts from the longer range to the shorter range, as shown by the convergence score. Plots of the actual performance in terms of the corrections and their volatility indicate that use of the larger decay factors might not be acceptable to the users of MOS forecasts. All of these results lead to the strong indication that bias correction would improve the MOS temperature and dewpoint forecasts if implemented with a decay factor in the range 0.025–0.05, the exact value within that range not being very important. Testing was done here on data from 2011 and 2012. It is likely that any medium- to long-sample MOS could be improved by this decaying average method. The method is very easy to implement, and has been shown by Glahn (2013) to be more robust and to give more accurate forecasts than a regression method used in many NWS Weather Forecast Offices in the Advanced Weather Interactive Processing System.

While the decaying average method could logically be applied to any quasi-continuous variable, such as wind speed, it should be used with caution. This and other successful bias-correcting methods generally produce smoother forecasts than the forecasts to which they are applied. Smoother forecasts may not be desired even if the bias is decreased, because the forecasts will have less variance and the extremes will be reduced. For instance, the usefulness of wind speed forecasts depends critically on strong winds being reliably forecasted, and in some operational situations (e.g., for orchard growers at near-freezing temperatures), calm or very weak winds. MOS forecasts, for instance, have themselves been postprocessed to increase their variance above the mean with partial inflation (Schwartz and Carter 1982; Jacks et al. 1990; Glahn and Allen 1966).8

Acknowledgments. I am indebted to Tamarah Curtis for furnishing me data and scripts, and for tutoring me in generating the graphs; to Valery Dagostaro for assistance in retrieving verification data from MDL’s verification program; to Stephen Olson and Brian Zeigler for arranging for and setting up a script to access MDL’s IBM computers from my PC; to Yun Fan, Eric Eagle, and David Rudack for their previous work and for intensifying my interest in this subject; and to Kathryn Gilbert for careful proofreading of an early draft. This paper is the responsibility of the author and does not necessarily represent the views of the NWS or any other governmental agency.

8 Inflation was proposed by I. Enger and first applied by Klein et al. (1959). Inflated forecasts are obtained by subtracting from the regression estimate the developmental sample mean, dividing the difference by the (multiple) correlation coefficient, and adding the result to the sample mean. MDL found that this worked well for wind forecasts above the mean, but those below the mean were too weak, so the established practice is to “partially inflate” by using the procedure on only those regression estimates above the mean. Inflation, either full or partial, will increase the variance of the forecasts and increase the mean square error [see Glahn and Allen (1966) for details].
REFERENCES


