1.2 **REPLACING MISSING DATA FOR ENSEMBLE SYSTEMS**

Tyler McCandless*, Sue Ellen Haupt, George Young
[1] The Pennsylvania State University

1. INTRODUCTION

Missing data presents a problem in many fields, including meteorology. The data can be missing at random, in recurring patterns, or in large sections. Incomplete datasets can lead to misleading conclusions, as demonstrated by Kidson and Trenberth (1988), who assessed the impact of missing data on general circulation statistics by systematically decreasing the amount of available training data. They determined that the ratio of the Root Mean Square Error (RMSE) in the monthly mean to the daily standard deviation was two to three times higher when the missing data was spaced randomly compared to spaced equally, and RMSE increased by up to a factor of two when the missing data occurred in one block. Therefore, the spacing of the missing data can have an impact on statistical analyses.

It is useful to consider how to best replace, or impute, the missing data. Various methods have been considered, Vincent and Gullet (1999) found that highly correlated neighbor stations can be used to interpolate missing data in Canadian temperature datasets. Schneider (2001) used an expectation maximization (EM) algorithm for Gaussian data and determined that it is applicable to typical sets of climate data and that it leads to more accurate estimates of the missing values than a conventional non-iterative imputation technique. Richman et al. (2008) showed that iterative imputation techniques used to fill in systematically removed data from both linear and nonlinear synthetic datasets produced lower Mean Square Error (MSE) than other methods studied. Three iterative imputation techniques produced similar results, and they all had lower errors than using the mean value, case deletion, or simple linear regression. However, Kemp et al. (1983) compared seven different methods of replacing missing values and found that between-station regression yielded the smallest errors (estimated – actual) in an analysis of temperature observing networks. They found that linear averaging within station and a four-day moving average were not as accurate as between station regressions, although linear averaging had smaller errors than the four-day moving average.

* Corresponding author address: Applied Research Laboratory, P.O. Box 30, The Pennsylvania State University, State College, PA 16804-0030, email: tcm5026@only.arl.psu.edu

A major problem for statistical predictive system developers is obtaining enough input data from Numerical Weather Prediction (NWP) models for developing, or training, new post-processing methods. Such post-processing, or calibrating, of NWP forecasts has been shown to decrease errors in forecasting since the introduction of Model Output Statistics (MOS) (Glahn and Lowry 1972). For these statistical post-processing techniques to make these improvements in forecast skill, however, requires training them on archived forecasts and validation data. Enough data must be available for the resulting technique to be stable. Thus, imputing any missing data will provide larger training and validation datasets for this process.

The advent of ensembles of forecasts has led to new methods of post-processing. Post-processing ensemble temperature data can produce an even more accurate deterministic forecast than the raw model forecast, as well as provide information on uncertainty (Hamill et al. 2000). Some statistical calibration techniques give performance-based weights to ensemble members, thus giving more weight to an ensemble member that produces relatively higher skill than other ensemble members (Woodcock and Engel 2005, Raftery et al 2005). After computing performance weights from a set of previous ensemble forecasts, these weights are used to combine the current ensemble forecasts to form a single deterministic forecast. If some ensemble members are missing, however, a method must be devised to deal with the situation. Previous studies have simply excluded cases with missing ensemble forecast data (Greybush et al. 2008). In a real-time forecasting situation, however, excluding missing forecast data limits the forecast skill of the ensemble and limits the uncertainty information available from the ensemble. Currently, if an ensemble member is missing from a forecast, the National Weather Service (NWS) simply excludes it from the ensemble (Richard Grumm, 2008 personal communication), which limits the spread and utility of the ensemble. As we move toward more complex weighting methods, such an approach may no longer be feasible. Unlike replacing missing observation data, replacing missing forecast data must preserve each ensemble member's characteristics in order to portray its impact on the consensus forecast accurately, so that the computed performance based weight still applies.

This current study uses the same data and post-processing algorithms as Greybush et al. (2008). That study developed and evaluated regime dependent post-processing ensemble forecasting techniques. Greybush et al. (2008) used the University of Washington

Mesoscale Ensemble, a varied-model (differing physics and parameterization schemes) multi-analysis ensemble with eight members, to analyze several post-processing methods of forecasting two-meter temperature. A large amount of this data was missing, and Greybush et al. (2008) used case deletion in the training and analysis of the post-processing methods. The performance of the post-processing methods was tested on a year of data for four locations in the Pacific Northwest.

The first post-processing method used here is a performance-weighted average using a 10-day sliding window. The relative performance of each of the eight ensemble members in the previous 10-days is used to compute forecast weights. Although this is not directly a regime-dependent forecast method, a seven to ten day period reflects the typical persistence of atmospheric flow regimes in the Pacific Northwest (Greybush et al. 2008).

A second post-processing method used here is the K-means regime clustering method. Clustering techniques are applicable when the instances are to be divided into natural, or mutually exclusive, groups (Witten and Frank 2005). The parameter, K, is the number of clusters that are sought. Greybush et al. (2008) showed that the K-means regime clustering post-processing method produced the lowest MAE in two-meter temperature forecasts when compared to regime regression, a genetic algorithm regime method, and different windowed performance-weighted averages.

The goal of the current research is to determine the effects of replacing the missing data on these post-processing methods. The same dataset and model development code are used to isolate the effects of replacing the missing data. We test the data imputation methods for four point locations for which we have verification observations. Section 2 describes the methods used to replace the missing data. Section 3 shows the results of the various methods of replacing the missing data on a 10-day performance window post-processing technique and a K-means regime clustering method. Section 4 summarizes and analyzes the results.

2. METHODS

This study uses data from an eight member ensemble of daily 48-hour surface temperature forecasts for a year (365 consecutive days). Four point locations from the Pacific Northwest are used: Portland, Oregon; Astoria, Oregon, Olympia, Washington; and Redmond, Washington. All of the verification data are available. However, 260 of the 2920 (8.9%) temperature forecasts are missing from the eight member ensemble. Out of the 365 days, 151 (or 41.4%) have at least one ensemble member missing. Although there are often consecutive days of missing data for an ensemble member, the pattern of missing data appears to be random. The cases with missing data are deleted in the original research (Greybush et al. 2008). More sophisticated methods are tested here.

The missing data are replaced before performing simple linear bias-correction. The full process of missing data imputation, bias-correction, and regime-based performance weighted averaging is shown as a data flowchart in Figure 1.

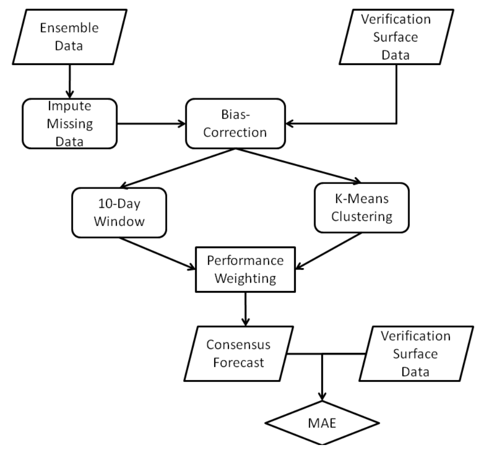

Figure 1. Flowchart of the process of imputing the missing data and the effects on the MAE of statistical post-processing techniques.

After bias-correcting the individual ensemble members with simple linear regression, weights are assigned to the ensemble members. The 10-day performance-weighted window and K-means regime clustering both calculate separate performance weights for the ensemble members. The deterministic forecast is a combination of the weighted ensemble members and is compared to the verifying observations to produce a consensus forecast MAE for each day. For a fair assessment of the methods, independent verification data is necessary; therefore, the data is randomly divided into a dependent (training) dataset (2/3 of the days) and an independent (verification) dataset (1/3 of the days). The performance weights are calculated on the training dataset and applied to the independent dataset to produce the mean MAE of the forecasts from the independent dataset. This process is repeated 50 times and averaged to avoid overfitting and biasing the results via sampling error.

Finally, the performance of two weighted average post-processing is evaluated to quantify the benefit of missing data imputation. The post-processing techniques calculate averaging weights for each ensemble member based upon that ensemble member's forecasting performance in a 10-day window or under a specific weather regime. Therefore, the data used to replace the missing ensemble member's temperature forecasts must mimic the statistical characteristics of that member in order to maintain the validity of the performance weights.

2.1 Case-Deletion

The first way to address the issue of missing data is to exclude it from the dataset. This method deletes the ensemble member that is missing. Therefore, if two ensemble members are missing for a given day, the performance weights are computed on the remaining ensemble members. In training the 10-day performance-weighted window, if an ensemble member forecast is missing at any time in the previous 10-days, that day is excluded and the performance weight for that member is computed on the remaining nine days. Note that this procedure reduces the amount of data available for training and the number of members used in the consensus forecast. This is the method used operationally and when training new statistical post-processing methods by the National Weather Service (Richard Grumm 2008, personal communication; Dr. Harry Glahn 2008, personal communication). Thus case deletion will be our baseline method for comparison. Note that if the training dataset is sufficiently large enough, post-processing methods can be trained with case deletion. However, this method is difficult to adapt to any missing ensemble data in a real-time forecast situation. All other methods of imputing missing data are compared to the baseline results provided by case deletion.

2.2 Mean substitution

One of the simplest techniques for missing data imputation is replacing missing values by the annual mean. Although this method does not take into account the seasonality of temperatures, we test it here because it is computationally fast.

2.3 Persistence

When a weather pattern persists, the temperature at a location is often similar to that of the previous day. Therefore, the first comparison method of predicting the missing data is persistence, which uses the previous day's temperature forecast. Consecutive days of missing data are treated by using the value of the last previous day with an available temperature forecast. Our data includes several periods of consecutive missing data, which limits the effectiveness of persistence, because the most recent available data is assigned to all consecutive missing days. Persistence is a computationally efficient method of replacing the missing data.

2.4 Polynomial Fit

Another approach to imputing data is to fit the time series of each ensemble member with a smooth curve that captures the seasonal variability of temperature. The resulting curve can then be used to predict missing members of the time series. A least squares polynomial fit has the advantage of being aperiodic and thus accounting for interseasonal variability.
A sensitivity study was conducted to determine the most appropriate degree polynomial for fitting the data. Neither a linear fit (first degree polynomial fit) nor a second degree polynomial fit (dotted line) adequately capture the seasonal cycle in the temperature data. This inadequacy is reflected in Figure 2.

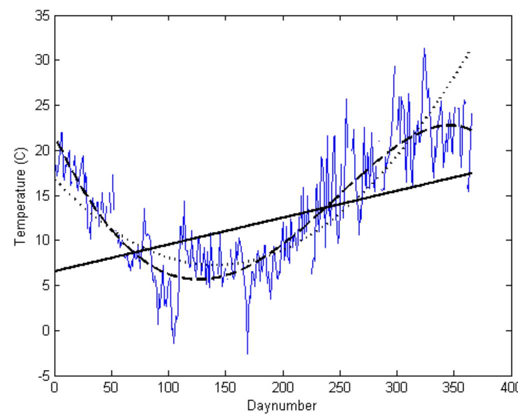

Figure 2. Comparison of a single polynomial fit (straight line), a second degree polynomial fit (dotted line), and a fifth degree polynomial fit (dashed line).

The polynomial fit of degree five (dashed line) is used in this study because it was the lowest degree polynomial that adequately fit the yearly temperature data. Polynomials with degrees greater than five did not produce lower MAEs.

2.5 Iterative Polynomial Imputation

Richmond et al (2008) developed an iterative imputation method that performed well on their synthetic data set. This method is based on the fifth polynomial fit to the time series described above. The iterative imputation method splits the ensemble forecast data into two separate datasets, with the split being based upon whether or not each case contains any missing forecasts. If a day is missing an ensemble member forecast, that day is put into dataset one. If the day has all ensemble members available, it is placed into dataset two. A polynomial fit similar to that described in the previous section is constructed for each ensemble member in dataset one. The resulting regression equations are then used to predict the missing values in the second dataset. These predicted values are imputed into the missing dataset and the two datasets are merged. New polynomial regression equations are computed for each ensemble member in the merged dataset. These regression equations are used to update the missing values in dataset two. These new estimated missing values are merged with the original data. This process is repeated a third time.

2.6 Fourier Fit

In contrast to the two polynomial methods described above, a Fourier series directly fits the periodicity of the annual temperature cycle with a Fourier series. Each of the ensemble members is fit with the first two Fourier harmonics (sines and cosines). As an example, Figure 3 shows the Fourier curve (dashed line) that fits the temperature forecasts (light solid line) from ensemble member #3. Then, the Fourier curve is used to impute missing data for each ensemble member.

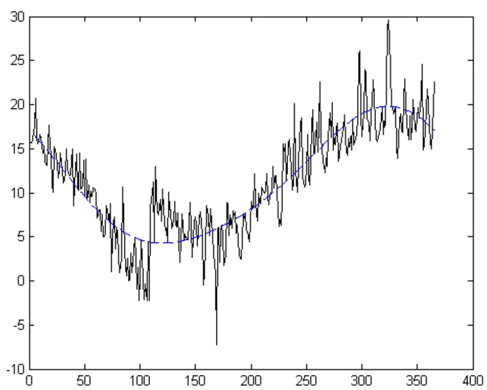

Figure 3. Fourier fit to ensemble member #3's temperature forecasts. The y-axis is the temperature in Celcius and the x-axis is the daynumber.

2.7 Three-day Deviation

The final method attempts to maintain the statistical characteristics of each ensemble member by computing the difference between the ensemble member forecast and the ensemble mean for each day. The deviation between an ensemble member and the ensemble mean reflects the recent behavior of that member compared to the other members. The average deviation for each ensemble member is calculated for several windows: the entire year, the previous five days, previous three days, and previous day.

Figures 4 and 5 show the results for the previous day, three-day, five day, and entire dataset (all days) mean deviation with a 10-day performance-weighted window and K-means regime clustering respectively. Based on these results, a three-day deviation was chosen for subsequent comparisons. Although the results for all four windows are similar, the three-day deviation is the shortest window that could still be used if there are two consecutive days of missing data. Therefore, the three-day deviation was useable in more of the cases than the previous day deviation. The five day and entire dataset mean deviation showed similar results, but were not as computationally efficient as the three-day mean deviation.

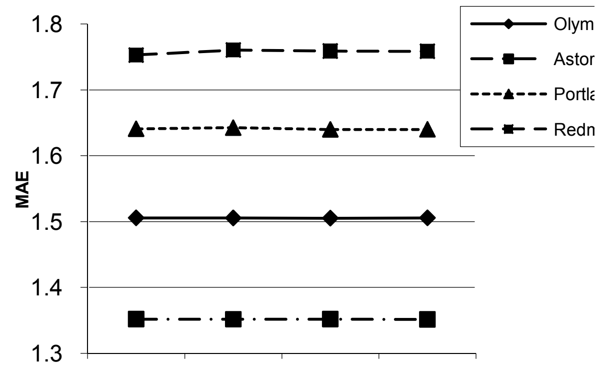

Figure 4. Sensitivity study for the optimal length of deviation from the ensemble mean with 10-day performance-weighted window post-processing.

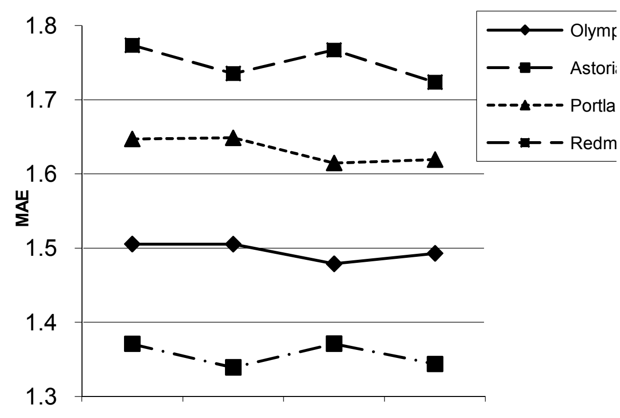

Figure 5. Sensitivity study for the optimal length of deviation from the ensemble mean with K-means regime clustering post-processing.

3. RESULTS

The goal of this study is to determine a best method for replacing missing ensemble forecast data to be used in the development or real-time operational use in a consensus post-processing technique. The Mean Absolute Error (MAE) is used as the accuracy metric while the verification rank histograms and ensemble spread are used to assess ensemble calibration. To demonstrate ensemble calibration, verification rank histograms are used (Hamill 2000).

The MAE for each of the methods of imputing the missing data is shown in Table 1, along with that for the baseline of case deletion. The case deletion method is expected to always have the lowest MAE since it only includes the days with a complete ensemble. Although that method could be useful for training a post-processing method, it is not feasible for generating a real-time forecast. Nonetheless, even if an ensemble member is missing, we must make the forecast, so we must use one of the implementation methods to provide an estimate for any missing values. Therefore, the goal of replacing missing ensemble data is to minimize the MAE of the forecasts in order to best match the baseline provided by the case deletion method.

The three-day mean deviation method produces the lowest MAE closest to that of the baseline for three of the four locations. The exception is for Portland, where the three iteration polynomial fit produced the lowest MAE. In that case, the MAE for the 10-day performance-weighted window was actually slightly lower than for the baseline method. Table 1 also lists the standard deviations of the MAE across the 50 tests. The standard deviations are less than or equal to 0.01°C for the 10-day performance-weighted window and less than 0.19°C for the K-means regime clustering method, which shows that there is more variability in the K-means regime clustering technique.

The next step in testing the methods of replacing the missing data is to determine the effects on the calibration of the ensemble. When replacing the missing ensemble members, we wish to produce ensembles whose dispersion is the same as those for the cases where all ensemble members are available. Verification rank histograms are used to test this correspondence. These rank histograms are created by iteratively tallying the rank of the verifying observation relative to the ensemble member forecasts, which are sorted from lowest to highest (Hamill 2000). If the verifying observation is colder than the coldest ensemble member forecast, then bin one would get a tally. If the verifying observation is warmer than the coldest ensemble member forecast, but colder than the second coldest ensemble member, then the second bin would get a tally. There are nine bins since the verifying observation could be colder than all ensemble members, warmer than all ensemble members, or in between any of the eight sorted members. Therefore the plots denote

Table 1. Mean Absolute Error (MAE) and mean Standard Deviation (STD) for all methods and locations.

Table 1	MAE		STD	
	Olympia			
Method	K-means	10-Day	K-means	10-Day
Case Deletion	1.49	1.51	0.09	0.00
Fourier	1.60	1.60	0.08	0.00
Persistence	1.58	1.60	0.09	0.00
Ensemble Member Mean	1.77	1.80	0.11	0.01
Polynomial	1.57	1.60	0.09	0.00
Polynomial 3-Iterations	1.56	1.59	0.10	0.00
3Day Mean Deviation	1.51	1.51	0.09	0.00
	Astoria			
Method	K-means	10-Day	K-means	10-Day
Case Deletion	1.34	1.35	0.07	0.00
Fourier	1.40	1.39	0.10	0.00
Persistence	1.38	1.41	0.09	0.00
Ensemble Member Mean	1.51	1.53	0.09	0.01
Polynomial	1.35	1.39	0.09	0.00
Polynomial 3-Iterations	1.38	1.39	0.10	0.00
3Day Mean Deviation	1.34	1.35	0.09	0.00
	Portland			
Method	K-means	10-Day	K-means	10-Day
Case Deletion	1.60	1.64	0.09	0.01
Fourier	1.69	1.66	0.13	0.01
Persistence	1.65	1.66	0.10	0.01
Ensemble Member Mean	1.86	1.91	0.15	0.01
Polynomial	1.68	1.65	0.12	0.00
Polynomial 3-Iterations	1.63	1.63	0.11	0.01
3Day Mean Deviation	1.65	1.64	0.11	0.01
	Redmond			
Method	K-means	10-Day	K-means	10-Day
Case Deletion	1.76	1.76	0.14	0.01
Fourier	1.83	1.83	0.14	0.01
Persistence	1.80	1.81	0.12	0.01
Ensemble Member Mean	2.13	2.09	0.18	0.00
Polynomial	1.84	1.83	0.14	0.01
Polynomial 3-Iterations	1.88	1.83	0.17	0.01
3Day Mean Deviation	1.74	1.76	0.12	0.01

the number of occurrences of the verifying observation that fall below (colder) all the member's forecasts, between any of the eight sorted member's forecasts, or higher (warmer) than all the member forecasts. The verification rank histograms in Figure 6 compare the calibration of all methods for replacing the missing forecasts for Portland. The histogram, (a), from the cases where no ensemble members are missing because of case deletion (baseline), illustrates that the verifying observation frequently fell in the middle of the ensemble or fell above (warmer) than the ensemble. This is indicated by the fact that bins five and nine had much higher counts than the rest. Of the imputation methods, the three-day mean deviation best showed relatively high bins for numbers five and nine while retaining similar bins for the rest. The three-day mean deviation did show bins one through three relatively higher than the cases without missing data; however, no other method kept bins five and nine distinctly higher.

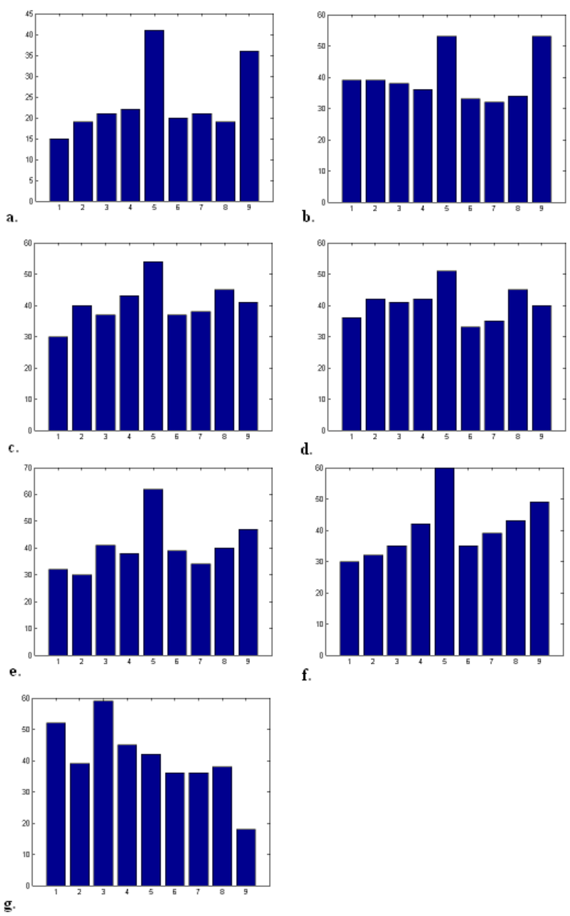

Figure 6. Verification rank histograms for all methods at Portland. (a) is case deletion, (b) is three-day mean deviation, (c) is polynomial fit with three iterations, (d) is a polynomial fit, (e) is persistence, (f) is Fourier, and (g) is ensemble member mean.

The rank histograms for Astoria (Figure 7) show clear underdispersion. Underdispersion occurs when the ensemble spread is too small, causing the verifying observation to tend to fall either warmer or colder than the warmest or coldest ensemble member respectively (Hamill 2000). We see such behavior in Figure 7a, for the baseline. When imputing the missing data with the three-day mean deviation (Figure 7b), the verification rank histogram is quite similar to that of the baseline. The polynomial fit and the three iteration polynomial fit also produce verification rank histograms that are similar to that of the baseline. The verification rank histograms for Olympia and Redmond (not shown) also indicate that the three-day mean deviation preserves the ensemble characteristics of their baselines.

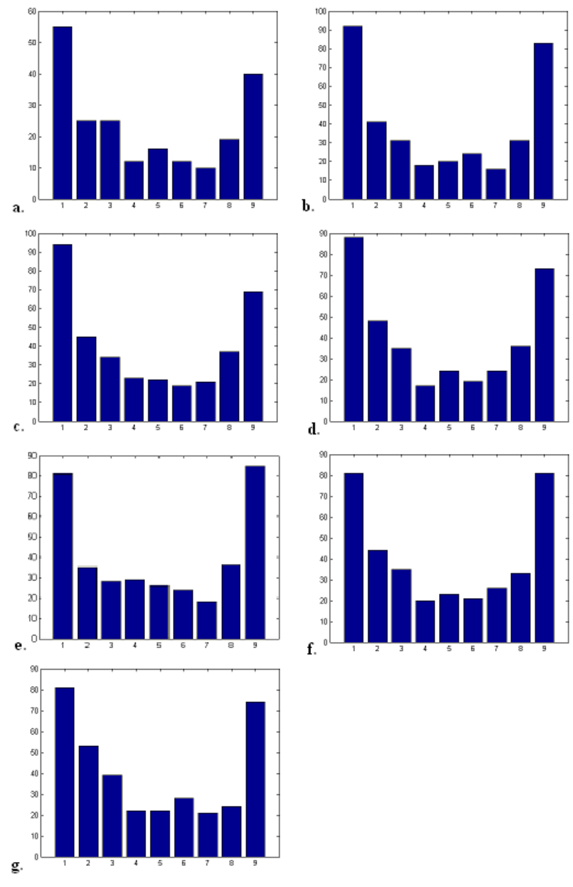

Figure 7. Verification rank histograms for all methods at Astoria. (a) is case deletion, (b) is three-day mean deviation, (c) is polynomial fit with three iterations, (d) is a polynomial fit, (e) is persistence, (f) is fourier, and (g) is ensemble member mean.

Figure 8 compares the rank histograms for the four different locations. The case deletion baseline method is the left column and the three-day mean deviation is the right column. From the histograms, it is evident that the calibration of the ensembles when missing forecasts are replaced with the three-day mean deviation is similar to that for the baseline. For Redmond, the histograms appear similar with the exception of the ninth bin, which is relatively higher when imputing the missing data. The verification rank histograms for Portland are also very similar, with relatively higher bins five and nine for both excluding and replacing the missing data. The verification rank histograms for Olympia and Astoria also show that the three-day mean deviation produces ensemble dispersion similar to that of the case deletion method.

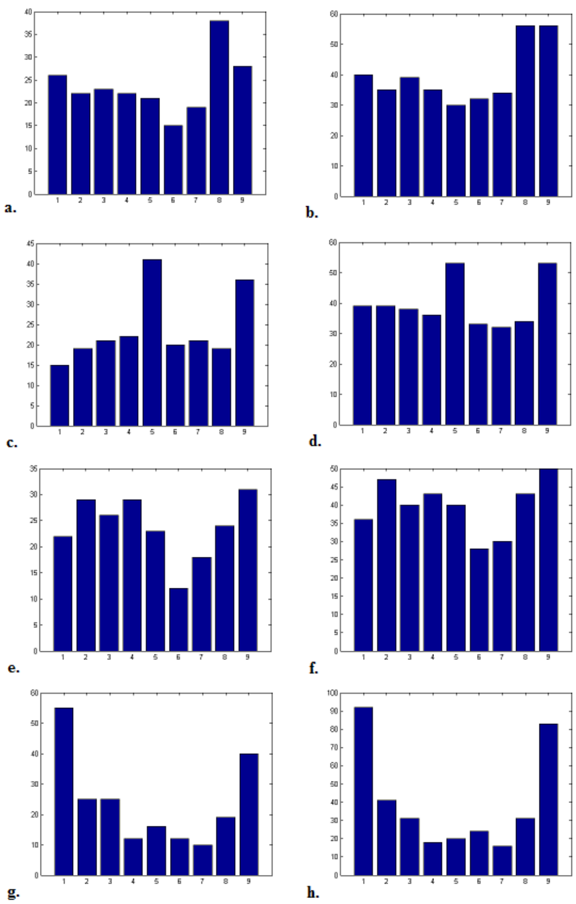

Figure 8. Verification rank histograms for case deletion (left column) and three-day mean deviation (right column) for four locations: Redmond; first row (a, b), Portland; second row (c, d), Olympia; third row (e, f), Astoria; last row (g, h)

Two final verification metrics are the mean and standard deviation of the ensemble spread. When replacing missing ensemble forecasts, we wish to preserve the ensemble spread and the variations in the ensemble spread. The ensemble spread is calculated by subtracting the lowest forecast temperature from the highest forecast temperature. Table 2 displays the results of the ensemble spread. For three of the four sites, the three-day mean deviation method yields means and standard deviations of the ensemble spread closest to that of the baseline.

Table 2. Mean ensemble spread and standard deviations (STD) for all methods. Underlined are the methods to replace the missing data with the lowest mean and standard deviation.

Olympia		
Method	**Mean**	**STD**
Case Deletion	2.08	1.65
Fourier	2.26	1.83
Ensemble Member Mean	2.31	2.12
Persistence	2.23	1.83
Polynomial	2.25	1.80
Polynomial 3-Iterations	2.23	1.78
3Day Mean Deviation	2.15	1.72
Astoria		
Case Deletion	1.04	1.40
Fourier	0.98	1.22
Ensemble Member Mean	1.04	1.28
Persistence	1.00	1.23
Polynomial	0.98	1.22
Polynomial 3-Iterations	0.99	1.22
3Day Mean Deviation	0.99	1.22
Portland		
Case Deletion	2.78	2.05
Fourier	2.98	2.21
Ensemble Member Mean	3.04	2.51
Persistence	2.91	2.15
Polynomial	2.97	2.18
Polynomial 3-Iterations	2.96	2.15
3Day Mean Deviation	2.84	2.10
Redmond		
Case Deletion	3.12	2.38
Fourier	3.08	2.39
Ensemble Member Mean	3.19	2.54
Persistence	3.22	2.46
Polynomial	3.16	2.39
Polynomial 3-Iterations	3.17	2.39
3Day Mean Deviation	3.08	2.26

4. CONCLUSIONS

For training a statistical post-processing and artificial intelligence techniques, researchers require a dataset that is sufficiently large to avoid overfitting. Thus, if the training dataset available is too small due to missing values, the values must be replaced. In addition, for use of the resulting post-processing methods in real-time forecasting, one must make a forecast even when ensemble members may be missing, so a data imputation method is mandatory. A requirement for replacing missing ensemble member temperature forecasts is to keep the characteristics of that ensemble member in the values that replace the missing data. Therefore, determining the best data imputation method is important for both developing the statistical forecast methods and operational use of these forecasting methods.

This study seeks to determine the optimal method to replace missing ensemble temperature data when producing a consensus forecast through statistical post-processing techniques. The results from a one year test in the Pacific Northwest show that imputing the missing data with the three-day mean deviation from the ensemble mean produces the lowest mean absolute error for three of the four locations and also produces similar ensemble dispersion characteristics to the case deletion method. When developing statistical post-processing methods, archived ensemble forecasts are necessary. If those forecasts are missing, some method is necessary to impute the missing data. This study has shown that the three-day mean deviation method reduces MAE and preserves ensemble spread and dispersion.

Our plans for extending this work includes statistical testing, experimenting with advanced statistical multiple imputation techniques, and using a longer dataset with more locations. It would be interesting to compare the effects of replacing missing ensemble member temperature forecasts for another region or to test the methods on additional forecast variables with different statistical properties. It would also be interesting to test the three-day mean deviation for missing forecast data in long range forecast ensembles.

***ACKNOWLEDGMENTS** – The authors would like to thank Steven Greybush for the use of his code, detailed documentation, and knowledgeable discussions. The authors wish to thank Pennsylvania State Climatologist, Paul Knight, for his insightful idea for the mean deviation method of replacing missing data. The authors also wish to thank Richard Grumm and Dr. Harry Glahn of the National Weather Service for providing valuable information for this project. Thanks are also due to the University of Washington for enabling public access to its ensemble data. The research was funded in part by the PSU Applied Research Laboratory Honors Program.*

REFERENCES:

Greybush, S. J., S.E. Haupt., and G. S. Young., 2008: The regime dependence of optimally weighted ensemble model consensus forecasts of surface temperature. *Wea. Forecasting,* accepted and posted by AMS.

Hamill, T. M., 2000: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.,* **129,** 550-560.

Hamill, T.M., S.L. Mullen, C. Snyder, Z. Toth, and D.P. Baumhefner, 2000: Ensemble forecasting in the short to medium range: report from a workshop, *Bull. Amer. Meteor. Soc.,* **81**, 2653-2664.

Kemp, W. P., D. G. Brunell, D.O. Everson, and A. J. Thomson, 1983. Estimating missing daily maximum and minimum temperatures. *J. Climate*, **22**: 1587-1593.

Kidson, J.W., and K.E. Trenberth, 1988: Effects of missing data on estimates of monthly mean general circulation statistics, *J. Climate*, **1**, 1261–1275.

Lorenz, E., 1963: Deterministic nonperiodic flow. *Jour. of Atmos. Science,* **20**: 130-141.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian Model averaging to calibrate forecast ensembles, *Mon. Wea. Rev.,* **133,** 1155-1174.

Richman, M. B., R, B. Trafalis, and I. Adrianto., 2008: Missing data imputation through machine learning algorithms. *Artificial Intelligence Methods in the Environmental Sciences,* Springer, S. E. Haupt, A. Pasini, and C. Marzban, Eds. In press.

Schneider, T., 2001: Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. *J. Climate*, **14**, 853–871.

Toth, Z. 2001: Ensemble forecasting in WRF; meeting summary. *Bull. Amer. Meteor. Soc.,* **82**, 695-697.

Toth, Z. and E. Kalnay, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.,* **125,** 3297-3319.

Vincent, L. A., and D. W. Gullet., 1999: Canadian historical and homogeneous temperature datasets for climate change analyses. *International Journal of Climatology.* **19**, 1375-1388.

Wilks, D.S., 2006: *Statistical methods in the atmospheric sciences*, 2nd ed., Academic Press, 626 pp.

Witten, I. H., and E. Frank., 2005: Data Mining: *Practical machine learning tools and techniques*, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

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