The Effects of Imputing Missing Data on Ensemble Temperature Forecasts

Tyler C. McCandless
The Pennsylvania State University/Applied Research Laboratory & Meteorology Department
State College, PA, USA
Email: tcm5026@gmail.com

Sue Ellen Haupt
The Pennsylvania State University/Applied Research Laboratory, State College, PA, USA
Current Address: National Center for Atmospheric Research, Boulder, CO, USA
Email: haupt@ucar.edu

George S. Young
The Pennsylvania State University Meteorology Department, State College, PA, USA
Email: young@meteo.psu.edu

***Abstract*—A major issue for developing post-processing methods for NWP forecasting systems is the need to obtain complete training datasets. Without a complete dataset, it can become difficult, if not impossible, to train and verify statistical post-processing techniques, including ensemble consensus forecasting schemes. In addition, when ensemble forecast data are missing, the real-time use of the consensus forecast weighting scheme becomes difficult and the quality of uncertainty information derived from the ensemble is reduced. To ameliorate these problems, an analysis of the treatment of missing data in ensemble model temperature forecasts is performed to determine which method of replacing the missing data produces the lowest Mean Absolute Error (MAE) of consensus forecasts while preserving the ensemble calibration. This study explores several methods of replacing missing data, including ones based on persistence, a Fourier fit to capture seasonal variability, ensemble member mean substitution, three day mean deviation, and an Artificial Neural Network (ANN). The analysis is performed on 48-hour temperature forecasts for ten locations in the Pacific Northwest. The methods are evaluated according to their effect on the forecast performance of two ensemble post-processing forecasting methods, specifically an equal-weight consensus forecast and a ten day performance-weighted window. The methods are also assessed using rank histograms to determine if they preserve the calibration of the ensembles. For both post-processing techniques all imputation methods, with the exception of the ensemble mean substitution, produce mean absolute errors not significantly different from the cases when all ensemble members are available. However, the three day mean deviation and ANN have rank histograms similar to that for the baseline of the non-imputed cases (i.e. the ensembles are appropriately calibrated) for all locations, while persistence, ensemble mean, and Fourier substitution do not consistently produce appropriately calibrated ensembles. The three day mean deviation has the advantage of being computationally efficient in a real-time forecasting environment.**

***Index Terms*— ensemble forecasting; data imputation; Artificial Intelligence (AI), Artificial Neural Network (ANN); missing data; numerical weather prediction**

I. 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 [1], who assessed the impact of missing data on general circulation statistics by systematically decreasing the amount of available 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 as compared to being 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 found that highly correlated neighbor stations can be used to interpolate missing data in Canadian temperature datasets [2]. Schneider 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 [3]. Richman et al. showed that iterative imputation

doi:10.4304/jcp.6.2.162-171

techniques used to fill in systematically removed data from both linear and nonlinear synthetic datasets produced the lowest Mean Square Error (MSE) [4]. They also showed that three iterative imputation techniques[1] have similar results, and they all had lower errors than using the mean value, the non-imputed cases, or simple linear regression. However, Kemp et al. compared seven different methods[2] of replacing missing values and found that between-station regression yielded the smallest absolute errors (estimated – actual) in an analysis of temperature observing networks [5]. They found that averaging within-station and a four day moving average methods were not as accurate as between-station regressions, although linear averaging had smaller errors than the four day moving average.

No prior studies have focused on the unique issues associated with imputing missing data for ensemble forecast systems. Specifically, not only is the best forecast sought, but the ensemble spread, or calibration, should also be maintained in order to preserve the uncertainty information that is a major purpose of running ensemble forecast systems.

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 of NWP forecasts has been shown to decrease errors in forecasting since the introduction of Model Output Statistics (MOS) [6]. For these statistical post-processing techniques to make such improvements in forecast skill, however, requires training them on archived forecasts and validation data. Success depends on obtaining enough data to ensure that the resulting forecast technique is stable. Imputing any missing data will provide larger training and validation datasets, thus aiding in 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 forecast uncertainty [7]. Some statistical calibration techniques give performance-based weights to ensemble members, thereby giving more weight to an ensemble member that produces relatively higher skill than other ensemble members [8-9]. 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, or in the method used by Raftery et al. [9] and Fraley et al. [10], a probabilistic forecast is produced, which allows for uncertainty estimates as well as a 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 [11] or used an EM algorithm with Bayesian Model Averaging (BMA) to replace missing ensemble forecasts in the training data [10]. 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. As we move toward more complex post-processing methods, such an approach may no longer be feasible. Unlike replacing missing observation data, replacing missing forecast data must preserve each ensemble member's forecast characteristics in order to preserve the uncertainty information contained in the ensemble. This uncertainty information is due to the varying initial conditions, boundary conditions, physics, and parameterizations because these perturbations grow in time. Therefore, uncertainty is reflected in the spread in the ensemble. Thus, imputation should preserve the characteristics of each ensemble member so as to maintain the statistical consistency of the ensemble spread.

We wish to study several methods of imputing missing data in a real meteorological ensemble prediction system with post-processing to provide a consensus forecast. The objective of the ensemble prediction system is twofold: first, we wish to provide accuracy in the consensus forecast and second, we wish to use the ensemble spread to assess uncertainty in the forecast. Thus, any method of imputing data must address both objectives. To do that, we wish to use a real ensemble prediction system with a sufficient amount of data to make preliminary conclusions regarding the most likely imputation methods. Since most operational ensemble prediction systems change configurations relatively frequently, it is difficult to find a consistent system for a long enough time to train and judge the imputation methods.

The ensemble modeling system used in this study is the University of Washington Mesoscale Ensemble, a varied-model (differing physics and parameterization schemes) multi-analysis ensemble [12]. The University of Washington Mesoscale Ensemble used in this study has eight ensemble members. This study uses forty-eight hour two-meter temperature forecasts from the University of Washington Mesoscale Ensemble for five locations in Oregon and five locations in Washington. Three years are used in this study, 2004 through 2006, which is 1096 days. Twenty-eight of the days do not have any ensemble members available. These cases were not used in the results because missing an entire ensemble forecast is a completely different problem than missing a member of the ensemble due to the lack of relationships between ensemble members and the ensemble mean. Such relationships are vital for preserving the characteristics of the ensemble members.

Two post-processing methods for creating a deterministic forecast from the ensemble are used for post-processing the dataset in this study. The first post-

[1] The three iterative imputation techniques used in Richman et al. [4] were support vector regression, artificial neural networks, and stepwise-regression.

[2] The seven different methods in Kemp et al. [5] were within station linear interpolation, within station moving averages, between station additive procedure, regression with best r weighting coefficient, regression with weighting proportional to square of the between-station correlation values, and two methods that are the same as the previous two methods except that the techniques are pooled-across-year monthly data sets.

processing method is an equal-weight consensus forecast. This is the standard post-processing method that assumes all models have equal skill and are equally likely to produce the correct forecast; therefore, each ensemble member receives 1/8, or 12.5%, of the weight of the final deterministic forecast. The second post-processing method used is a performance-weighted average with a ten day sliding window. For this method, the relative performance of each of the eight ensemble members in the previous ten 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 [11].

The goal of this study is to determine the effects of imputing the missing data on these post-processing methods and on the ensemble calibration. Although the non-imputed cases could be useful for training a post-processing method if the dataset is sufficiently large, they are not usable for generating a real-time forecast that includes uncertainty information when required input data is missing. This can pose severe problems in operational forecasting unless ensemble member failures are rare. In an operational setting we must make the forecast and assess its uncertainty even in cases where an ensemble member is missing, so we must use one of the imputation methods to provide an estimate for the missing member forecasts. The goal of the imputation methods is thus to produce forecasts which match, as closely as possible, the error characteristics of those forecasts made when all ensemble members are available (i.e. the non-imputed cases). The Mean Absolute Error (MAE) is used here as the forecast accuracy metric while rank histograms are used here to assess ensemble spread calibration. The MAE is the average difference between the forecast and the observations. The ensemble spread, or difference among ensemble member forecasts, is important because it provides a useful proxy for forecast uncertainty [12].

Section 2 describes the methods used to replace the missing data and the baseline of the non-imputed cases. Section 3 shows the results of the various methods of replacing the missing data on an equal-weight consensus forecast and a ten day performance window forecast. Section 4 summarizes and analyzes the results.

II. METHODS

This study uses data from an eight member ensemble of daily 48-hour surface temperature forecasts for three years. Ten locations from the Pacific Northwest are used: Portland, Oregon; Astoria, Oregon; Eugene, Oregon; Ephrata, Oregon; Burns Municipal Air Force Base, Oregon; McChord Air Force Base (Tacoma), Washington; Seattle, Washington; Spokane, Washington; Olympia, Washington; and Redmond, Washington. Fig. 1 plots the locations and Table 1 provides details on the observation stations. Not including the 28 days where the entire ensemble was missing, 372 out of the 1068

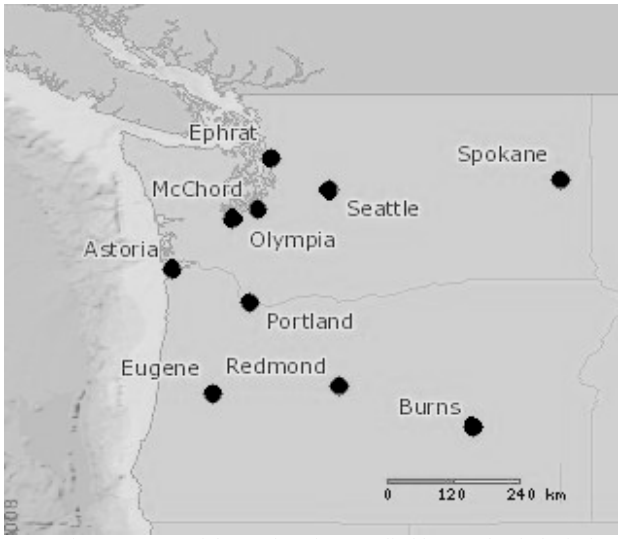

Figure 1. Map of the ten locations studied here. The dark circles mark the location of the observation sites.

days (or 34.8%) have at least one ensemble member missing. The number of such cases decrease with increasing number of missing ensemble members: 266 days have one ensemble member missing, 72 days have two ensemble members missing, 16 days have three ensemble members missing, six days have four ensemble members missing, six days have five ensemble members missing, two days have six ensemble members missing, and four days have seven ensemble members missing. There is no discernable pattern in the missing ensemble temperature forecasts for each location; however, the missing days for each ensemble member are consistent across the locations.

The verification data were obtained from the National Climate Data Center and then verified for consistency with historical data on wunderground.com. All of the verification data are available.

The full process of missing data imputation, bias-correction, and post-processing is shown as a data flowchart in Fig. 2. The missing data are replaced before performing simple linear bias-correction. After bias-correcting the individual ensemble members with simple linear regression, weights are assigned to the ensemble members. The equal-weight post-processing combines the ensemble members in order to produce a consensus forecast. This consensus forecast is compared to the verification surface data to produce a Mean Absolute Error (MAE). The ten day performance-weighted window post-processing method calculates separate performance weights for each ensemble member based on its ten day average MAE. These weights are proportional to the inverse MAE. The consensus forecast is the weighted average of the ensemble members and is compared to the verifying observations to produce a consensus forecast MAE for each day.

Finally, the performance of the post-processing is evaluated to quantify the benefit of missing data imputation. The data used to replace the missing ensemble member's temperature forecasts must mimic the statistical characteristics of that member in order to

Table 1. List of each observation location and the ICAO code.

Location	ICAO Code
Portland International Airport, OR	KPDX
Astoria Regional Airport, OR	KAST
Eugene Airport/Mahlon Sweet Field, OR	KEUG
Ephrata Municipal Airport, WA	KEPH
Burns Municipal Airport, OR	KBNO
McChord Air Force Base, Tacoma, WA	KTCM
Seattle-Tacoma International Airport, WA	KSEA
Spokane International Airport, WA	KWGEG
Olympia Airport, WA	KOLM
Redmond Robert's Field Municipal Airport, OR	KRDM

maintain the validity of the weights used to produce the deterministic forecast.

The methods used in this study were carefully selected to address a range of climate, weather, and statistical characteristics. Using the ensemble member mean provides a simple "no skill" imputation method for comparison. Persistence takes into account the most recent weather pattern. A Fourier fit approximates the annual temperature cycle. Three day deviation estimates an ensemble member's most recent behavior relative to the rest of the ensemble. An Artificial Neural Network (ANN) takes this approach several steps further, being an advanced statistical method capable of capturing the non-linear relationships among predictors. All of these methods can be used in a real-time ensemble forecasting environment; i.e., if an ensemble member is missing in the current forecast, these methods can be used to replace that missing ensemble member without re-running the ensemble member. Several other advanced statistical methods, such as multiple imputation, have been shown to be efficient at replacing missing data [13-14]. They, however, require iterative processing of large amounts of data.

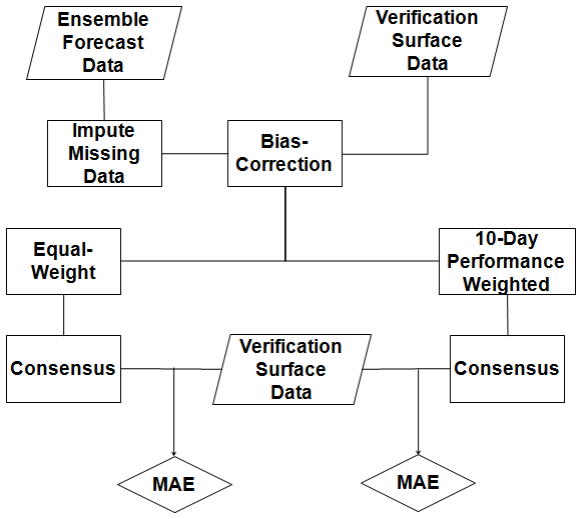

Figure 2. Flowchart of the process of imputing the missing data and the effects on the MAE of statistical post-processing techniques.

data.

Therefore, these methods would both be difficult to implement and computationally intensive in a real-time ensemble forecasting environment. Thus, the methods tested in this study are limited to those that could be used efficiently in real-time forecasting.

A. Non-imputed Cases, The Standard

The non-imputed cases serve as our standard for comparing the imputation methods. For both the ensemble mean and ten day performance weighted post-processors the non-imputed cases are simply the days when all ensemble members are available. In contrast, the existence of one or more missing ensemble members at some point during the previous ten days does not exclude a case from the non-imputed set. Rather, if an ensemble member forecast is missing at any time in the previous ten days, that day is excluded and the performance weight for that member is computed on the remaining days.

Thus, the non-imputed cases provide an estimate of what the post-processors' forecast MAE and spread should be if it is doing its job perfectly. We are able to use the non-imputed cases as a standard because neither of the post-processors uses more than 10 days of data rather than requiring a large training dataset. In contrast, various post-processing methods use prior distributions or some other statistical representation of climatology, requires a large training set in order to achieve accuracy. Hence, the need for imputation as discussed in the Introduction.

B. Mean Substitution

One of the simplest techniques for missing data imputation is replacing missing values with the three year mean forecast from the ensemble member. Using a daily mean temperature from the ensemble member is not feasible because we only have three years of data and using climatology of the observations would not retain the characteristics of the missing ensemble member, thereby not preserving the appropriate ensemble calibration. Although this method does not take into account the seasonality of temperatures, we test it here because it provides a computationally fast "no skill" baseline.

C. Fourier Fit

The next data imputation method corrects the problem noted above by using a Fourier series to directly fit the periodicity of the annual temperature cycle. Each ensemble member is fit with the first two Fourier harmonics. As an example, Figure 3 shows the Fourier fit (thick line) to the temperature forecasts (circles) from ensemble member #3 for Astoria. The Fourier fit is then used to impute missing data for each ensemble member.

D. Persistence

While the Fourier fit captures the annual cycle, it does not capture the shorter, aperiodic oscillations of synoptic weather patterns. When a weather pattern persists for several days, as is often the case, the temperature at a location is similar to that of the previous day. This persistence can be exploited by using the previous day's temperature forecast as the imputed value.

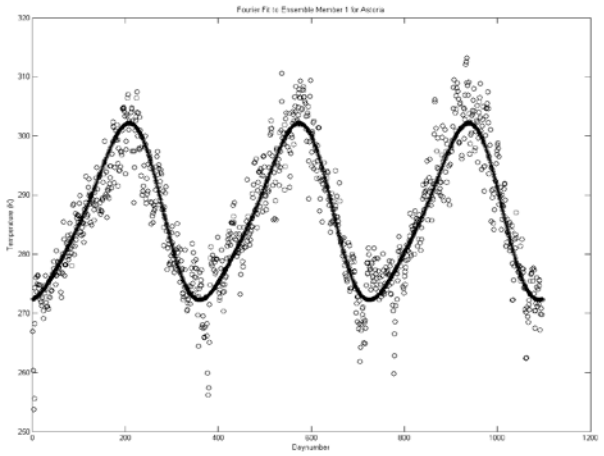

Figure 3. Fourier fit to ensemble member #3's temperature forecasts for all three years. The forecasts from ensemble member #3 are shown as circles. Missing data is replaced by the equation represented by the thick line.

Consecutive days of missing data are treated by using the value of the last previous day with an available temperature forecast. Persistence of the ensemble member's forecast is used to replace the missing data in order to retain the characteristics of the ensemble members. Replacing the missing data with persistence of the observation may lead to a more accurate forecast, but a mis-calibrated ensemble. Persistence is a computationally efficient method of replacing the missing data, but does not take into account the day-to-day changes in temperature predicted by the other ensemble members.

E. Three Day Deviation

The next method attempts to make use of the ensemble information about day-to-day weather changes while maintaining the statistical characteristics of the missing ensemble member. It does so by basing the imputation on the difference between the ensemble member forecast and the corresponding ensemble mean. 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 was calculated for several windows: the entire year, the previous five days, previous three days, and previous day. A sensitivity study on one year of the data for four locations determined that the mean absolute error did not change significantly as the averaging period for the deviation changed. Figure 4 illustrates the mean absolute error on the ten-day performance weighted window post-processing method for four locations with various averaging periods. The insignificant difference between results obtained with different averaging periods is likely a result of each ensemble member having a consistent forecast bias compared to the ensemble mean. The results for the equal-weighting post-processing method are not shown here, but they also showed no significant difference among the different averaging periods. The three day deviation is the shortest window that could still be used if there are two consecutive days of missing data. Although the five day and entire dataset mean deviation added a few more cases and had insignificantly different results, they are not as computationally efficient as the three day mean deviation and were therefore not used in the remainder of this study. Note that this method could be easily generalized to the averaging period most applicable for a specific dataset.

F. Artificial Neural Network

The final method tested is an Artificial Neural Network (ANN), an advanced statistical technique that is able to capture non-linear relationships in data [15]. ANNs have been used in a wide range of weather forecasting and modeling [16-17]. Here, we attempt to improve upon the three day ensemble member mean deviation by using an ANN to predict the missing member's forecast from those of the other ensemble members. A separate ANN is developed for each ensemble member and for each location with a goal of preserving the characteristics of the ensemble as well as the recent, or three day, weather pattern at each location. First, two separate datasets are created for each ensemble member at each location: one containing cases that can be used to train and test the ANN and the other for those cases actually requiring imputation. The first dataset consists of every four consecutive days of complete data for an ensemble member and the ensemble mean for each of those days. Therefore, the ANN is given seven predictors: the ensemble member's forecast for the first three days and ensemble means of all four days. The ensemble member's forecast for the fourth day is the predictand. The other dataset consists of the ensemble member forecasts and the ensemble means for the three previous days, and the ensemble mean for the day that ensemble member is missing. The ANN that has been developed and tested on subsets of the first dataset is then applied to the second dataset in order to predict that ensemble member's missing forecasts.

The ANN used here is a feed-forward neural network trained via back-propagation. It has twenty neurons with one hidden layer and one output layer. This configuration was chosen because it provides a balance of computational complexity and performance. Figure 5 illustrates the architecture of the ANN. The left column

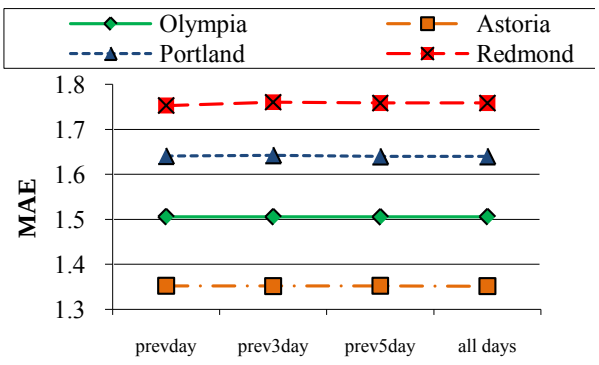

Figure 4. Sensitivity study for the optimal length of deviation from the ensemble mean with ten-day performance-weighted window post-processing.

of rectangles represents the predictors. The twenty vertical circles represent the neurons (adjustable weights), which connect to the final output layer and the forecast value. The development and testing dataset is randomly divided into three subsets with 60%, 20%, and 20% of the vectors. The 60% subset is used to train the network while one set of 20% is used to evaluate how well the network generalized and so control termination of the training cycle. The second 20% provides an independent test of the network generalization. Training continues until the network's error on the evaluation data does not improve. This technique avoids the problem of overfitting, which plagues many optimization and learning algorithms. The resulting ANNs are then used on the actual imputation datasets to predict the missing day.

III. RESULTS

A primary objective of ensemble forecasting is to provide uncertainty estimates along with an improved deterministic forecast. Since the goal of this study is to determine a best method for imputing missing ensemble forecast data so that they that can be used in real-time operational consensus post-processing, we must assess both the accuracy of the resulting forecasts and whether the ensemble maintains the same uncertainty characteristics.

The MAE resulting from each of the methods of imputing the missing data is shown in Tables 2 and 3, along with that for the standard of the non-imputed cases.

The non-imputed cases only include the days with a complete ensemble; thus, there are 696 days in the results for the standard. The MAE recorded for the other methods are computed for only the cases where ensemble member(s) were missing and imputed; therefore, there are 372 days in the results for the imputed methods. Although the imputation days and the days from the non-imputed cases are different, this dataset is sufficiently long to provide consistent results. A more accurate evaluation would be to test the methods on data that were synthetically removed. If the dataset was synthetic, we would have been able to randomly remove data and test on the same days; however, our real dataset did not permit this. Comparing the results of the imputation methods on the days with missing ensemble members to the days with a complete ensemble, or the non-imputed cases, is the appropriate technique for evaluating methods to impute data in real-time.

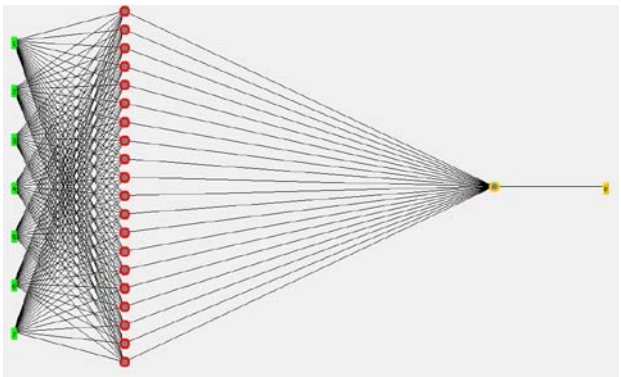

Figure 5. Neural network architecture.

Table 2. Mean absolute error for equal-weight consensus forecast. Bold indicates MAE closest to that of the non-imputed cases.

Equal-Weight		Temperature (K)				
City	Non-Imputed	Mean	Fourier	Persistence	Three Day	Neural Network
Astoria	1.45	1.63	1.33	**1.40**	1.35	1.34
Burns	2.08	3.07	2.13	2.12	2.03	**2.06**
Ephrata	5.37	6.39	5.64	5.68	5.64	**5.62**
Eugene	1.86	2.39	1.79	1.84	**1.87**	1.84
McChord	1.63	1.84	1.61	1.64	**1.63**	1.62
Olympia	1.63	2.04	1.61	**1.62**	1.51	1.52
Portland	1.82	2.34	**1.91**	1.97	1.97	2.00
Redmond	1.89	2.59	2.02	1.98	**1.86**	1.85
Seattle	1.60	1.82	1.62	1.63	**1.60**	1.58
Spokane	1.78	2.40	**1.70**	1.69	1.67	1.67

The temperature forecast MAEs for the standard and imputation methods with equal-weight consensus post-processing are shown in Table 2. The temperature forecast MAEs for the standard and imputation methods with ten-day performance-weighted window post-processing are shown in Table 3. The first column of each table corresponds to the non-imputed cases, the cases when all ensemble members were available. The MAE results for both the ten day performance-weighted window and equal-weighted forecasts indicate that ensemble member mean substitution yields poor performance in these two ensemble consensus methods. For each station in Table 2 and Table 3, the bold values denote which method produces an MAE closest to that for the standard of the non-imputed cases. For the equal-weight post-processing (Table 2); MAEs for the three day deviation were closest those of the non-imputed for four of the locations, while persistence, ANN, and Fourier did so for two of the locations each. For the ten day performance-weighted window post-processing (Table 3), MAEs for the three day deviation method were closest to this standard for four of the locations, while MAEs for the ANN and Fourier methods were closest to the standard for three of the locations each, and persistence produced MAEs closest to the standard in two of the locations. These counts include two ties. Based on the number of times that the each of the methods' MAE was closest to that for the non-imputed cases, the three day mean deviation method appears to marginally produce the best results, although, the ANN, persistence, and Fourier substitutions do better for some of the locations. This same analysis could be undertaken with the standard

Table 3. Mean absolute error for ten-day performance weighted window post-processing. Bold indicates MAE closest to that of the non-imputed cases.

10-Day PW		Temperature (K)				
City	Non-Imputed	Mean	Fourier	Persistence	Three Day	Neural Network
Astoria	1.45	1.59	1.34	**1.40**	1.34	1.34
Burns	2.01	2.79	2.06	**2.03**	1.96	**1.99**
Ephrata	5.33	6.28	5.59	5.62	5.59	**5.56**
Eugene	1.81	2.25	1.76	1.79	1.83	**1.80**
McChord	1.60	1.84	1.61	1.62	**1.60**	1.59
Olympia	1.62	1.95	**1.62**	1.61	1.51	1.52
Portland	1.81	2.30	**1.93**	1.97	1.96	1.99
Redmond	1.88	2.39	2.00	1.94	**1.85**	1.84
Seattle	1.57	1.80	1.61	1.62	**1.58**	**1.56**
Spokane	1.74	2.19	**1.70**	1.65	1.67	1.65

deviation of forecast error, although this approach would be more sensitive to outliers [18].

In order to quantify these results, a two sample Student t-test was performed. The null hypothesis is that the MAE for the imputation method is equal to that for the non-imputed cases, i.e., there is no significant difference between the ensemble's consensus forecast accuracy when all members are available and when missing data is imputed. The number of degrees of freedom is the minimum of the sample sizes, which is the non-imputed cases for this problem, less one (371). At the 95% confidence level, the null hypothesis could not be rejected for the following methods: three day mean deviation, persistence, Fourier, and neural network. Only the mean substitution method had MAE values significantly different from the standard for seven of the ten locations with both the ten-day performance weight window post-processing and the equal-weight post-processing. These results indicate that the three day mean deviation, persistence, Fourier, and neural network can all be used to replace missing ensemble member forecasts without significantly altering the forecast accuracy of the post-processing techniques.

The second major objective of our imputation methods also is to preserve the calibration of the ensemble so that its uncertainty information is maintained. Thus, the next step in testing an imputation method is to determine its effects on the calibration of the ensemble. When replacing the missing ensemble members, we wish to produce an ensemble whose ensemble member spread (i.e. dispersion) is the same as that which we would obtain for the case if all ensemble members are available. We assess this dispersion using a rank histogram. The rank histogram was developed independently by [19-21] to quantify ensemble dispersion. These rank histograms are created by tallying the rank of the verifying observation relative to the ensemble member forecasts, which are first sorted from lowest (coldest) to highest (warmest) [22]. For example, if the verifying observation is colder than the coldest ensemble member forecast, then bin one (the leftmost bin) would get a tally. Likewise, if the verifying observation is warmer than the coldest ensemble member forecast, but colder than the second coldest ensemble member, then the second bin from the left would get a tally. Thus, there are nine bins for the eight member ensemble studied here 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. Note that all of our rank histograms, when interpreted in the classical manner, denote an underdispersive ensemble. Recent studies, however, have cautioned users in interpreting such rank histograms that are produced by ensembles with either temporal correlations or correlations between members, which are certainly present in meteorological ensemble prediction systems [22-24].

The results show that although persistence, Fourier, three day mean deviation, and the ANN imputation methods all produced MAEs not significantly different from that for the non-imputed cases, the rank histograms for Fourier and persistence methods do not preserve the ensemble dispersion as well. Recall that the goal is not to create a uniform rank histogram (i.e. a calibrated ensemble) but rather a rank histogram that is similar to that of the non-imputed cases (i.e. consistent ensemble dispersion). Figure 6 shows the rank histograms for the standard and each data imputation method for Eugene, Oregon. This plot indicates that the rank histograms for three day mean deviation and ANN both closely resemble that for the non-imputed cases. In contrast, persistence, ensemble mean, and Fourier substitution methods do not produce rank histograms that closely resemble that for the non-imputed cases.

The same imputation methods did well at reproducing the rank histogram of non-imputed cases for Spokane, Washington (Figure 7). The most striking feature of this rank histogram is that bin nine has more than double the counts as bin eight. Thus, Spokane has an under-forecasting bias in the ensemble, i.e. the observation tends to fall on the warm side of the sorted forecasts. The rank histograms for the three day mean deviation and ANN both closely resemble the rank histogram for the non-imputed cases, as they each have a rank histogram with bin nine at least double the size of the bin eight. In contrast, the other methods do not display this feature.

Large tails to the rank histograms are also evident for the non-imputed cases for Portland, Oregon (Figure 8). Once again, the three day mean deviation and the ANN have the most similar rank histograms to the rank histogram for the non-imputed cases. For brevity, the rank histograms for the other seven are not shown here but produce similar results; the three day mean deviation and ANN have the most similar rank histograms to that for the non-imputed cases and thus best preserve the ensemble spread information.

In order to quantify these rank histogram results, a chi-square goodness of fit test was performed. The chi-square goodness of fit test distinguishes between true deviations of a sample histogram from a target

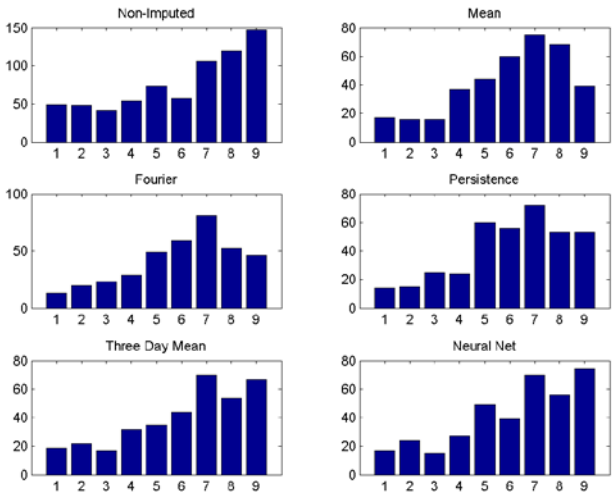

Figure 6. Rank histogram for Eugene, Oregon. The x-axis is the bin number and the y-axis is the number of forecasts in each of the respective bins.

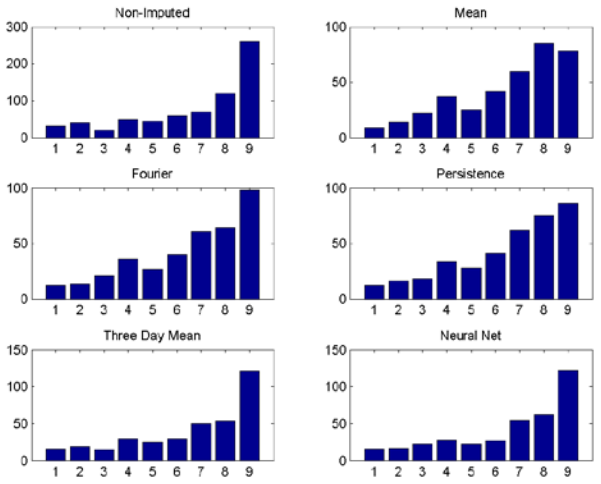

Figure 7. Rank histogram for Spokane, Washington. The x-axis is the bin number and the y-axis is the number of forecasts in each of the respective bins.

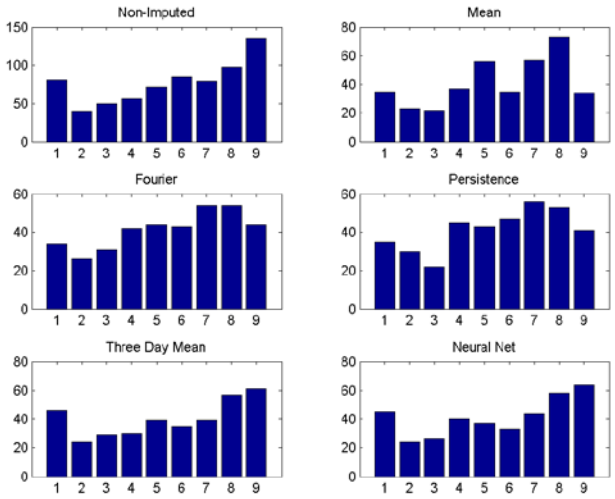

Figure 8. Rank histogram for Portland, Oregon. The x-axis is the bin number and the y-axis is the number of forecasts in each of the respective bins.

distribution and those caused by sampling variations [18].

Previous work compared rank histograms for non-imputed cases with the theoretical (flat) distribution for specific forecasting biases or inadequate sample size [25-26]. In our study, however, we are evaluating whether the rank histograms for the various imputation methods correspond to that for the non-imputed cases; thus the target distribution for the chi-square goodness of fit test is the rank histogram for the non-imputed cases. The null hypothesis is that the rank histogram of the imputation method is indistinguishable from the rank histogram tallied from the non-imputed cases, i.e. the cases repaired via imputation methods are drawn from the same ensemble member distribution as the non-imputed cases. If n_i is the observed count in the i^{th} bin and e_i is the corresponding expected count based on the rank histogram for the non-imputed cases, then the test statistic, T, for the χ^2 goodness of fit test is

$$T = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i}. \qquad (1)$$

The T statistic comes from (approximately) a χ^2 distribution with (k - 1) degrees of freedom where k is the number of bins in the rank histogram. The T values are listed in Table 4 for all ten locations.

For a goodness-of-fit test at the 95% confidence level, all T-values greater than 15.51 lead to rejection of the null hypothesis that the rank histogram from the imputation method is not significantly different from the rank histogram tallied from the non-imputed cases. Thus, there are only three instances where the imputation method creates a rank histogram that is not significantly different from the non-imputed cases at the 95% confidence level, two are from the neural network and one is from the three day mean deviation. Thus, none of the methods perfectly reproduces the rank histogram for non-imputed cases. The methods can, however, be ranked in how close they come to this goal. Thus, the final row shows the average T-statistic for each imputation method. The three day mean deviation imputation method produces rank histograms with the lowest average T-statistic, 31.18, indicating that, averaged across all 10 locations, this method produces rank histograms most similar to that for the non-imputed cases. The T-statistic for the ANN imputation method is only slightly larger at 33.02. Both the three day mean deviation and ANN imputation methods produce rank histograms with T-statistic values that are markedly smaller than those for the mean, Fourier, and persistence imputation methods, 65.89, 43.26, and 43.23 respectively. These results lead us to conclude that the three day mean deviation and ANN method best preserve the ensemble spread of the non-imputed dataset, with the ANN being only slightly worse. Note, however, that the three day mean deviation and ANN imputation methods both do a much better job of preserving the shape of the rank histogram, and thus, the ensemble characteristics, than do the other imputation methods tested. One must be careful in interpreting the results of the T-tests here, however, due to the likely correlations in our time series, which act to decrease the effective degrees of freedom. Thus, the ranking of the techniques is actually more uncertain. A longer period of data, and thus more missing values, would better enable the ranking of the methods.

IV. CONCLUSIONS

When training a statistical post-processing or artificial intelligence technique, developers require a dataset that is large enough to prevent overfitting. Thus, if the training dataset available is too small due to missing values, the values must be replaced in order to maximize the effectiveness of the post-processing or artificial intelligence technique. In addition, for use of the resulting post-processing methods in real-time forecasting a data imputation method is mandatory, since one must make a forecast even when ensemble members

Table 4. Chi-square goodness-of-fit T-statistic values for all ten locations. For eight degrees of freedom, a value greater than 15.51 means the imputation method produces a significantly different rank histogram from the non-imputed cases at the 95% confidence level. The bolded values correspond to rank histograms that are not significantly different from the non-imputed cases at the 95% confidence level. The final row is the average T-statistic for each of the locations.

	Mean	Fourier	Persistence	Neural Network	Three Day Mean
Astoria	44.47	19.46	35.52	24.36	19.39
Burns	77.24	41.47	27.78	21.49	20.23
Ephrata	42.23	55.47	58.40	73.20	67.28
Eugene	66.37	63.44	58.30	**15.35**	17.20
McChord	87.69	60.93	45.43	18.90	19.58
Olympia	56.82	53.93	55.66	34.89	17.63
Portand	48.05	23.79	32.02	88.75	52.47
Redmond	63.69	21.31	29.78	**12.35**	67.90
Seattle	100.11	45.75	37.79	17.62	19.47
Spokane	72.22	47.03	51.58	23.31	**10.67**
Average	65.89	43.26	43.23	33.02	31.18

may be missing. A requirement for replacing missing ensemble member temperature forecasts is to preserve the characteristics of that ensemble member in order to preserve ensemble spread and the calibration of the post processing method. Therefore, determining the best data imputation method based on these two criteria is important both for developing the statistical forecast methods and for operational use of these forecasting methods.

This study seeks to determine an optimal method to replace missing ensemble temperature data when producing a consensus forecast through statistical post-processing techniques. The results from a three year test on ensemble data from the Pacific Northwest show that imputing the missing data with the three day mean deviation from the ensemble mean or Artificial Neural Network produce mean absolute errors and rank histograms closest to those of the non-imputed cases. Both the ANN and three day mean deviation imputation methods produce mean absolute errors of the post-processed consensus forecasts that are insignificantly different from those of the non-imputed cases. Likewise both methods yield ensemble calibrations similar to that of the non-imputed cases. The ANN is, however, much more computationally expensive than the three day mean deviation method. This study has demonstrated that the three day mean deviation and ANN methods are the most effective methods for imputing missing ensemble temperature forecasts, with the three day mean deviation having the advantage of being time efficient in a real-time forecasting environment.

This study has not been exhaustive. There are other imputation methods that could also be explored. We attempted, however, to test representatives from the most widespread classes of techniques that meet our criterion of being applicable in real-time. For instance, although multiple imputation methods look promising, they would be difficult to apply real-time and are not as computationally efficient as some of the methods tested here. Furthermore, the three day mean deviation imputation method not only best preserves the post-processing accuracy and ensemble spread, but it is also very fast since it only requires an archive of forecasts from the previous three days. Likewise, many other non-linear statistical learning techniques are available besides the feed-forward ANN used here, but it is representative of the class. Several polynomial fit methods were also tested, but were not included here because the Fourier fit produced better results and represents the class of techniques using basis function decomposition.

Our study focused on one ensemble dataset for short range forecasts in the Pacific Northwest. It would be interesting to compare the effects of imputing missing ensemble member temperature forecasts for a different ensemble. It would also be interesting to test the methods for replacing missing forecast data in hourly and long range temperature forecast ensembles to determine if different lead times would require different imputation techniques.

ACKNOWLEDGEMENTS

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. Finally, we thank the anonymous reviewers for their helpful comments. Revisions resulting from these comments have strengthened the paper. The research was funded in part by the PSU Applied Research Laboratory Honors Program.

REFERENCES

[1] 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.

[2] 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.

[3] 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

[4] Richman, M. B., R, B. Trafalis, and I. Adrianto., 2009: Missing data imputation through machine learning algorithms. *Artificial Intelligence Methods in the Environmental Sciences,* S. E. Haupt, A. Pasini, and C. Marzban, Eds., Springer-Verlag, 153-169.

[5] 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.

[6] Glahn, H. R., and D. A. Lowry, 1972: The use of Model Output Statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203-1211.

[7] 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.

[8] Woodcock, F. and C. Engel, 2005: Operational consensus forecasts. *Wea. Forecasting* **20,** 101-111.
[9] 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.
[10] Fraley, C. A.E. Raftery, and T. Gneiting, 2010: Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian Model Averaging, Mon. Wea. Rev., 138, 190-202.
[11] 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,* **23**, 1146-1161.
[12] Grimit, E.P. and C.F. Mass, 2002: Aspects of effective mesoscale short-range ensemble forecasting system over the Pacific Northwest. *Wea. Forecasting*, **17,** 192-205.
[13] Rubin, D.B. (1976) Inference and missing data. *Biometrika*, **63**, 581-592.
[14] Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, 2nd edition, New York: John Wiley.
[15] Witten, I. H., and E. Frank., 2005: *Data mining:* p*ractical machine learning tools and techniques*, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.
[16] Krasnopolsky, V. M., 2009: Neural network applications to solve forward and inverse problems in atmospheric and oceanic satellite remote sensing. *Artificial Intelligence Methods in the Environmental Sciences,* S. E. Haupt, A. Pasini, and C. Marzban, Eds., Springer-Verlag, 191-205.
[17] Young, G. S., 2009: Implementing a neural network emulation of a satellite retrieval algorithm. *Artificial Intelligence Methods in the Environmental Sciences,* S. E. Haupt, A. Pasini, and C. Marzban, Eds., Springer-Verlag, 207-216.
[18] Wilks, D.S., 2005: *Statistical methods in the atmospheric sciences*, 2nd ed., Academic Press, 626 pp.
[19] Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecasts from Ensemble Model Integrations. *J. Climate*, **9**, 1518–1530.
[20] Hamill, T. M., and S. J. Colucci, 1996: Random and systematic error in NMC's short-range Eta ensembles. Preprints, *13th Conf. on Probability and Statistics in the Atmospheric Sciences,* San Francisco, CA, Amer. Meteor. Soc., 51–56.
[21] Talagrand, O., R. Vautard, and B. Strauss, 1997: Evaluation of probabilistic prediction systems. *Proceedings, ECMWF Workshop on Predictability,* ECMWF, 1–25.
[22] 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.,* **129,** 550-560.
[23] Marzban, C., R. Wang, F. Kong, S. Leyton, 2010: On the effect of correlations on rank histograms: Reliability of temperature and wind-speed forecasts from fine-scale ensemble reforecasts, Mon. Wea. Rev., in press.
[24] Saetra, O, H. Hersbach, J-R Bidlot and D.S. Richardson, 2004: Effects of observation errors on the statistics for ensemble spread and reliability. *Mon. Wea. Rev.*, **132**, 1487-1501.
[25] Elmore, K.L., 2005: Alternatives to the chi-square test for evaluating rank histograms from ensemble forecasts. *Wea. Forecasting,* **20,** 789-795.
[26] Jolliffe, I. T., and C. Primo, 2008: Evaluating rank histograms using decomposition of the chi-square test statistic, *Mon. Wea. Rev.,* **136,** 2133-2139.

Tyler C. McCandless is a PhD candidate in Meteorology at The Pennsylvania State University. He has worked on ensemble weather forecasting problems, particularly using artificial intelligence techniques for replacing missing data. His MS thesis was titled, *Statistical Guidance Methods for Predicting Snowfall Accumulation for the Northeast United States*. He earned a BS and M.S. in Meteorology from Penn State in the spring of 2010. Tyler has received the NASA Sylvia Stein Space Grant Scholarship, the American Meteorological Society Bob Glahn Scholarship for Statistical Meteorology, the Astronaut Scholarship, and recently was awarded the NCAA Postgraduate Scholarship for Men's Cross Country. He is a member of the American Meteorological Society.

Sue Ellen Haupt earned a B.S. in meteorology and marine science certificate from The Pennsylvania State University (University Park, PA) in 1978, M.S. in engineering management from Western New England College (Bedford, MA) in 1982, M.S. in mechanical engineering from Worcester Polytechnic Institute (Worcester, MA) in 1984, and Ph.D. in atmospheric science from the University of Michigan (Ann Arbor, MI) in 1988.

She is currently a Program Manager at the National Center for Atmospheric Research (Boulder, CO), a Senior Scientist in the Computational Mechanics Division of the Applied Research Laboratory, and Professor of Meteorology at The Pennsylvania State University (State College/University Park, PA). Her prior affiliations include National Center for Atmospheric Research, University of Colorado/Boulder, US Air Force Academy, Utah State University, University of Nevada/Reno, New England Electric System, and GCA Corporation. She is coauthor of *Practical Genetic Algorithms* (Wiley and Sons, NY, NY, 1998, second edition 2004) and is edited *Artificial Intelligence Methods in the Environmental Sciences* (Springer, 2009). She has authored over 200 book chapters, journal articles, conference papers, technical reports, and workshop proceedings. Her specialty is in applying novel numerical techniques to problems in fluid dynamics.

Dr. Haupt recently chaired the Committee on Artificial Intelligence Applications to Environmental Science of the American Meteorological Society (AMS), In addition to AMS she is a member of the American Society of Mechanical Engineers, Society for Industrial and Applied Mathematics, American Society of Engineering Educators, the American Geophysical Union, Society of Women Engineers, and three honor societies: Tau Beta Pi (Engineering), Phi Mu Epsilon (Mathematics), and Chi Epsilon Pi (Meteorology).

George S. Young earned a B.S. in meteorology from Florida State University (Tallahassee, FL) in 1979, M.S. in meteorology from Florida State University (Tallahassee, FL) in 1982, and Ph.D. in atmospheric science from Colorado State University (Fort Collins, CO) in 1986.

He is a Professor in the Meteorology Department at The Pennsylvania State University (University Park, PA) where he has been on the faculty since 1986. He has authored over 170 book chapters, journal articles, conference papers, technical reports, and workshop proceedings. His specialty is application of statistical and artificial intelligence methods to weather forecasting, decision making and satellite image analysis.

Dr. Young is a member the National Weather Association , the American Meteorological Society and four honor societies: Phi Beta Kappa (National Honor Society), Sigma Xi (Scientific Research Society), Phi Mu Epsilon (Mathematics), and Chi Epsilon Pi (Meteorology).