The Pennsylvania State University

The Graduate School

College of Earth and Mineral Sciences

STATISTICAL GUIDANCE METHODS FOR PREDICTING SNOWFALL ACCUMULATION IN THE NORTHEAST UNITED STATES

A Thesis in

Meteorology

by

Tyler McCandless

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

May 2010

The thesis of Tyler McCandless was reviewed and approved* by the following:

Sue Ellen Haupt
Professor of Meteorology
Thesis Advisor

George Young
Professor of Meteorology and GeoEnvironmental Engineering

Hampton N. Shirer
Associate Professor of Meteorology

Paul G. Knight
Senior Lecturer in Meteorology

Richard Grumm
Science Operations Officer, National Weather Service, State College, PA

William H. Brune
Professor of Meteorology
Head of the Department of Meteorology

*Signatures are on file in the Graduate School

ABSTRACT

Accurate forecasting of snowfall accumulation has widespread economic and safety consequences. Owing to the complex characteristics and dynamics inherent in winter weather systems, snowfall accumulation forecasts tend to have a large degree of uncertainty associated with them. Numerical Weather Prediction (NWP) ensemble prediction systems were developed to address the uncertainty in weather forecasts. A deterministic forecast is of utmost importance to the public; therefore, several post-processing methods for combining ensemble members have been developed. This study examines the use of several statistical guidance methods for post-processing the Global Ensemble Forecast System (GEFS) output in order to predict 24-hour snowfall accumulation. Out of the seven methods—an artificial neural network, linear regression, least median squares regression, support vector regression, radial basis function network, conjunctive rule, and k-nearest neighbor—the k-nearest neighbor method produces significantly more accurate forecasts, ones with lower mean absolute error, and the best calibrated ensemble spread as measured by rank histograms, spread-skill plots, and quantile-quantile plots.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This research is supported by the Educational and Foundational Program of the Applied Research Laboratory of the Pennsylvania State University. Thanks are due to Dr. Bob Glahn and the rest of the National Weather Service Meteorological Development Laboratory for access to the data as well as helpful guidance and suggestions. Thanks to my thesis advisor, Sue Ellen Haupt, and all of the committee members for helpful ideas, knowledgeable discussions, and support for this interesting research project. Thanks are also due to my colleagues, Andrew Annunzio, Luna Rodriguez, Jared Lee, Kerie Long, and Walter Kolczynski, who have helped with discussions, coding issues, and suggestions. I also wish to thank my parents, Ralph and Brenda McCandless, for their unwavering support in all of my academic endeavors. Finally, I wish to thank Penn State Track and Field Coach Beth Alford-Sullivan for supporting my academics while a member of the track and cross country teams.

Chapter 1

Introduction

Meteorologists face the difficult task of forecasting complex winter storm systems that can affect millions of people. More than 85,000 automobile crashes occurred on average each year for the period 1995-2001 nationwide when the road conditions were reported as either snowy and slushy or icy (Kocin and Uccellini 2005: L. Good-win, Mitretek Systems Inc, unpublished manuscript). An average of 1270 fatalities per year possibly result from snow/ice road conditions (Kocin and Uccellini 2005). These major weather storms also can have an extensive impact on the economy. The National Climate Database Center (NCDC) estimated the March 1993 and January 1996 storms each resulted in billions of dollars in damage (Kocin and Uccellini 2005). Accurate winter weather forecasts provide state and local departments of transportation the information required to prepare for these winter weather events efficiently and safely. At the 2005 USWRP Workshop, Ralph et al. (2005) stated "one effort should focus on winter storms along the East Coast of the United States, with freezing rain, coastal cyclones (e.g., nor'easters), heavy snow, and lake effect snow as priorities." Thus, accurate winter weather forecasts are valuable for public safety and more research is warranted to improve these predictions.

Forecasting snowstorms is a multifaceted problem presenting many challenges. One challenge is the difficulty of obtaining accurate and precise snowfall measurement, owing to the location and the surface of the observation, blowing and drifting, melting, compaction, mixed precipitation events, and how often the measurement is taken (Doesken and Leffler 2000). Without consistent and accurate snow accumulation observations, it is difficult to evaluate the performance of any forecast. Another challenge in snowfall forecasting is that both small and

large snowfall events can have similar indicators of snowfall. Evans and Jurewicz (2009) show that frontogenetical forcing, weak moist symmetric stability, saturation, and microphysical characteristics favorable for the production of dendritic snow crystals each of which having been identified as being critical indicators for heavy and/or banded snowfall in major storms, are often found in smaller magnitude snowfall events. Other challenges include the different spatial distribution of snowfall among various intensity snowfall events and the difference in spatial distribution among snowfalls of similar intensity. For example, the Blizzard of 1996 produced heavy snowfall in a wide area along the eastern seaboard, yet a lake-effect snow band may produce heavy snow in Erie, PA while the sky is clear in nearby Bradford, PA.

Another issue with winter weather forecasting is the large local differences in snowfall accumulation and precipitation type that are created by differing elevations and topographies among sites. Miller (1946) examines 200 cases of East Coast snowstorm development and classifies them into two categories: the classical polar front wave cyclone depicted by the Norwegian Cyclone model and the complex cyclonic development that Miller describes as a phenomenon unique to the East Coast. The distinctions between these two types of snowfall create different spatial and temporal snowfall accumulations. Snowstorms also have various economic and transportation impacts based on the duration of snowfall, propagation rates, whether there are winds that can cause significant blowing and drifting, and the population density where the snowfall occurs.

There have been significant improvements during the past century in winter weather prediction. Kocin and Uccellini (2005) state that, "one can conclude that the accuracy of weather forecasting, even for a rare event such as a major Northeast snowstorm, can be attributed to the introduction of numerical models into the forecast process beginning in the 1950s, the continued improvements made to the numerical models and global data, and the overall professional development of forecasters whose training and education are based heavily on understanding the

strengths and weaknesses of the models." In addition to improvements made to the numerical models, improvements in weather forecasting have come from the development of statistical post-processing methods of weather forecasting. Glahn and Lowry (1972) implement a post-processing technique named Model Output Statistics (MOS), which is an objective weather forecasting technique that consists of determining a statistical relationship between a predictand and variables forecast by a numerical model at various projection times.

Another development in weather forecasting is the advent of meteorological ensembles. Ensemble forecasts represent possible realizations of future states of the atmosphere. Grimmit and Mass (2002) show that a correlation exists between ensemble spread and forecast uncertainty, thus providing the forecaster with valuable uncertainty information. Advanced statistical post-processing techniques recently have been developed and implemented to order to improve the calibration, or uncertainty information, and the accuracy of NWP ensembles. An ensemble approach should provide more stable forecast solutions, thus improving the confidence in forecasters from the public and emergency management community when the potential for a snowstorm arises. Tracton and Kalnay (1993) applied the ensemble approach operationally for medium- and extended range predictions. Significant model variations reduce forecasters' confidence and accuracy.

Several studies examine different methods of post-processing ensemble forecasts in order to improve weather prediction. Raftery et al. (2005) use Bayesian Model Averaging (BMA) to improve 48-hr surface temperature forecasts in the Pacific Northwest. Their BMA forecasts were are calibrated much better than the raw ensemble, have prediction intervals 60% narrower than climatology, and yield a deterministic point forecast with 7% lower error than the best of the ensemble members and 8% lower than the ensemble mean. Greybush et al. (2008) use performance-weighted windows and a k-means regime clustering technique of optimally weighting ensemble temperature forecasts from the University of Washington Mesoscale

Ensemble to improve 48-hr surface temperature forecasts as compared with an equal-weighted deterministic forecast. Glahn et al. (2009) apply the error estimation capabilities of a linear regression framework together with the kernel density fitting applied to individual and aggregate ensemble members of the Global Ensemble Forecast System of the National Centers for Environmental Prediction to develop forecast probability density functions and cumulative density functions. This method produces reliable temperature, dewpoint, daytime maximum temperature, and nighttime minimum temperature forecasts with an accuracy exceeding that of the raw ensembles.

Although many of these advanced statistical post-processing methods have been shown to improve general forecasting, only recently have there been attempts to use post-processing to improve snowfall accumulation predictions. Cosgrove and Sfanos (2004) apply the MOS post-processing technique to forecast the conditional probability of snow and the snowfall amount exceeding a certain threshold, given that snowfall occurs, using the Global Forecast System (GFS) model.

Several studies attempt to improve snowfall forecasting by more accurately predicting the snow density. Roebber et al. (2003) conduct a principal component analysis of radiosonde and surface data and identify seven factors that influence snow ratio diagnosis: solar radiation per month, low- to mid-level temperature, mid- to upper-level temperature, low- to mid-level relative humidity, mid-level relative humidity, upper-level relative humidity, and external compaction as measured by surface wind speed and liquid equivalent precipitation amount. Using these seven predictors, they develop a ten-member ensemble of artificial neural networks that improve the snow-ratio class when compared with using a 10:1 ratio, sample climatology, and National Weather Service new-snowfall-to-estimated-meltwater conversion table. In 2007, Roebber et al. tested this ensemble of artificial neural networks on the 2004/05 and 2005/06 cold seasons and they find that this neural network approach substantially improves upon previous methods.

Baxter et al. (2005) develop a climatology of snow-to-liquid ratio for the contiguous United States and find that a mean snow-to-liquid equivalent of 13 is more typical than the often-assumed value of 10; there is considerable spatial variation in the mean however.

The goal of this study is to use advanced statistical guidance methods to post-process forecasts from the Global Ensemble Forecast System (GEFS) in order to improve both the accuracy and ensemble calibration, i.e. the forecast uncertainty information, for 24-hour snowfall accumulation predictions. Seven different statistical guidance methods are tested for producing 24-hour snowfall accumulation forecasts from the Global Ensemble Forecast System direct model output. These methods are trained to reduce the error of the control ensemble member and then applied to each ensemble member individually. These individual ensemble members are then used to produce a single consensus forecast of snowfall accumulation. Several techniques are used to examine the ensemble spread or calibration.

In chapter 2, the Global Ensemble Forecast System and the cooperative observing network are described. The statistical guidance methods used in this study are explained in chapter 3. In the following chapter results are summarized and discussed. Conclusions and prospects for future research are discussed in chapter 5.

Chapter 2

Data

This chapter introduces the data used for this thesis: first the observing network is described, and then the ensemble forecast system is explained.

2.1 Cooperative Observing Network

In order to test the validity of any forecast or create a consistent, reliable statistical post-processing, an accurate observing system is necessary (Allen 2001). The National Climatic Data Center (NCDC) Cooperative Summary of the Day (co-op) reports were used as the snowfall observing network. The co-op reports are produced daily by volunteers at their home or workplace and then sent to NCDC, which collects and processes the data. The variables reported once per 24 hour period were precipitation amount, snowfall accumulation, maximum temperature, and minimum temperature. Co-op stations are established, closed, supervised, and inspected by National Weather Service (NWS) personnel in annual visits to ensure observer proficiency, adherence to instrument and exposure standards, and network integrity (NWS 2000). The co-op stations do not report at a fixed time each day; thus in order to maintain consistency and to provide a valid test of the statistical guidance methods as described by Allen (2001), only those observations between 11 UTC and 17 UTC are retained. This is done in order to compare forecasts valid for 12 UTC. Not only are most of the observations recorded between 11 UTC and 17 UTC, but this time period corresponds with a lead time of 12 hours from the Global Ensemble Forecast System. This methodology is the same as that used by Cosgrove and Sfanos (2004).

Only the co-op observations with a snowfall measurement of a trace or more are retained in the dataset that spans the period from October 1, 2006 to March 31, 2007.

2.2 Global Ensemble Forecast System

The National Center for Environmental Prediction (NCEP) Global Ensemble Forecast System is an ensemble forecast system that uses the Global Spectral Model. Owing to several changes in the model configuration and number of members, the longest cold season consistent dataset available was October 1, 2006 to March 31, 2007. During this time period, the GEFS consisted of 15 total ensemble members (individual NWP forecasts): one high resolution control run and fourteen perturbations using the NCEP Ensemble Transform Bred Vector. The high resolution control run is the Global Forecast System (GFS) model at 35 km resolution and 0 degrees, while the fourteen other ensemble members are 105 km resolution and 0 degrees. The initialization time of the forecasts was 00 UTC each day, and each forecast was archived at 95.25-km resolution. The GEFS direct model output consists of forecasts for every six hours from 0-364 hours. There are 31 forecast elements from the GEFS, which are listed in Table 2-1. There are four forecasted wind speeds for both the U- and V-wind components, temperature forecast at five levels, four categorical precipitation variables: rain, freezing rain, ice pellets, and snow, four geopotential height fields, six-hour maximum temperature, six-hour minimum temperature, six-hour accumulated precipitation, relative humidity forecast at four levels, and two pressure fields.

Table 2-1. GEFS archived elements.

Element	Element Description [Units]	Levels
Press	Pressure [hPa]	Surface
PRMSL	Pressure reduced to MSL [hPa]	Mean Sea Level
RH	Relative humidity [%]	2-M, 925hPa, 850hPa, 700hPa, 500hPa
TMP	Temperature [K]	2-M, 1000hPa, 850hPa, 700hPa, 500hPa
TMAX	Maximum temperature in 6-hr period [K]	2-M
TMIN	Minimum temperature in 6-hr period [K]	2-M
U GRD	U-comp of wind [m/s]	10-M, 850hPa, 700hPa, 500hPa
V GRD	V-comp of wind [m/s]	10-M, 850hPa, 700hPa, 500hPa
HGT	Geopotential height [gpm]	1000hPa, 850hPa, 700hPa, 500hPa
FRZR	Categorical Freezing Rain [1=yes;0=no]	2-M
ICEP	Categorical Ice Pellets [1=yes;0=no]	2-M
SNOW	Categorical Snow [1=yes;0=no]	2-M
RAIN	Categorical Rain [1=yes;0=no]	2-M
PRCP	6-hr Total Precipitation Accumulation [kg/m^2]	2-M

Some of the GEFS output variables exhibit collinearity, thus three variables are deleted: surface pressure because it is collinear with mean sea level pressure, 1000 hPa temperature because it is collinear with 2-M temperature, and 1000 hPa height because it is collinear with mean sea level pressure. Although the GEFS predictions and co-op data cover the entire United States, the Northeast United States was the focus of this study. Thus, the dataset only included Northeast observation sites. Lake Ontario and Lake Erie help generate lake-effect snow, which has a completely different weather pattern than synoptic winter weather storms. Therefore, the

Northeast United States dataset was structured so that it did not include locations that generally receive direct lake effect snow.

The dataset was split into levels based on elevation, owing to the fact that certain variables would be potentially below ground level if they were kept in the dataset. At an elevation of 760 m, or approximately 2500 ft, the 925 hPa relative humidity is potentially below ground level; thus, the Northeast dataset was split into locations below 760 m, labeled the level-one dataset, and locations above 760 m, labeled the level-two dataset. For the locations above 760 m, the dataset does not include the 925 hPa relative humidity for each six-hour forecast interval. There are 10418 snowfall observations in the level-one dataset and 762 observations in the level-two dataset. Figure 2-1 plots the observations for both levels in the northeast with black asterisks marking the locations of the level-one observations and the red plus signs marking the locations of the level-two observations. There are 417 observation sites in the level-one dataset and 16 observation sites in the level-two dataset.

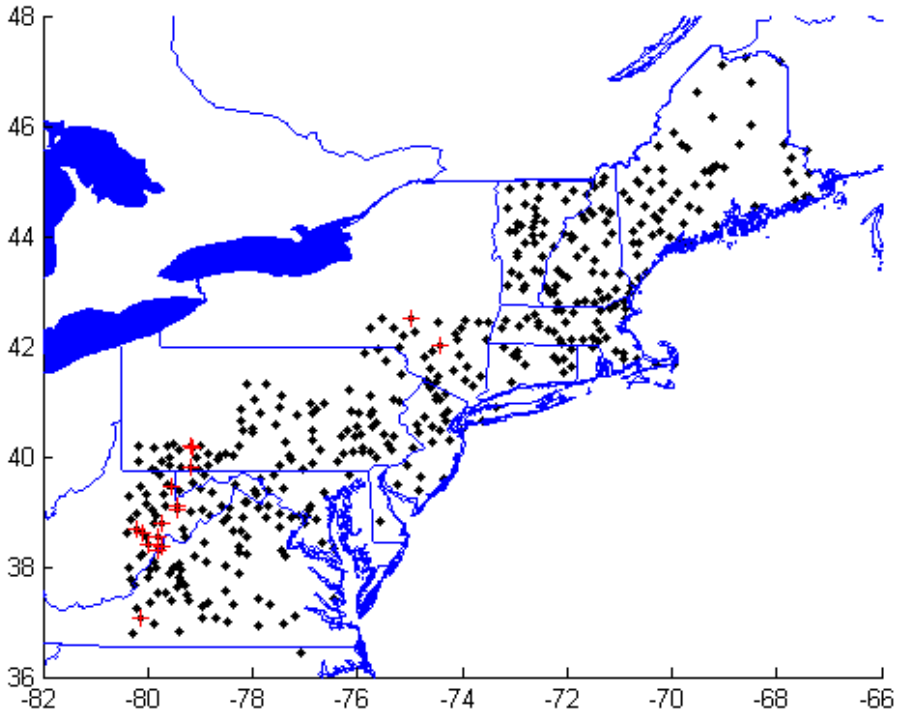

Figure 2-1. Snowfall observation sites with black asterisks marking the locations of the level-one observations and the red plus signs marking the locations of the level-two observations.

In order to compare the GEFS forecasts on a 95.25 km grid with individual co-op reporting sites, a nearest-neighbor weighting method was used to interpolate the GEFS gridded forecasts to the co-op locations. First, the interpolation process converted the grid point locations and co-op reporting sites from spherical to Cartesian coordinates. Then, the respective distances between the three nearest grid points and the co-op reporting sites were computed. Finally, a distance-weighted average of the three nearest neighbor grid points was used to calculate a forecast for the co-op location.

The datasets consisted of GEFS predicted weather variables for each level at forecast valid times of 12-18 hours, 18-24 hours, 24-30 hours, and 30-36 hours. These variables were

combined with latitude, longitude, and elevation of each station as predictors for the statistical guidance model as shown in Figure 2-2.

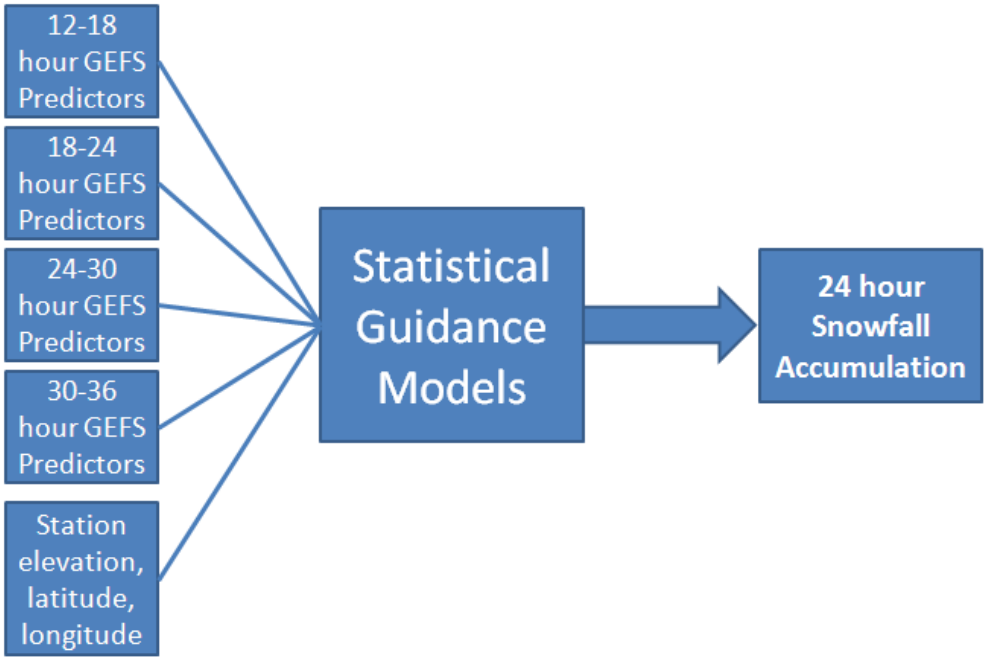

Figure 2-2. Schematic of the predictors that the statistical guidance models use to predict the 24-hr snowfall accumulation.

Thus, the statistical guidance methods used these variables to predict the total 24-hour snowfall accumulation, conditional on snow occurring. This methodology is displayed in Figure 2-3. The 11 UTC-17 UTC co-op observation, which is approximately 7 AM for observations in the Eastern Time Zone, is compared with a 12-36 hour prediction from the GEFS.

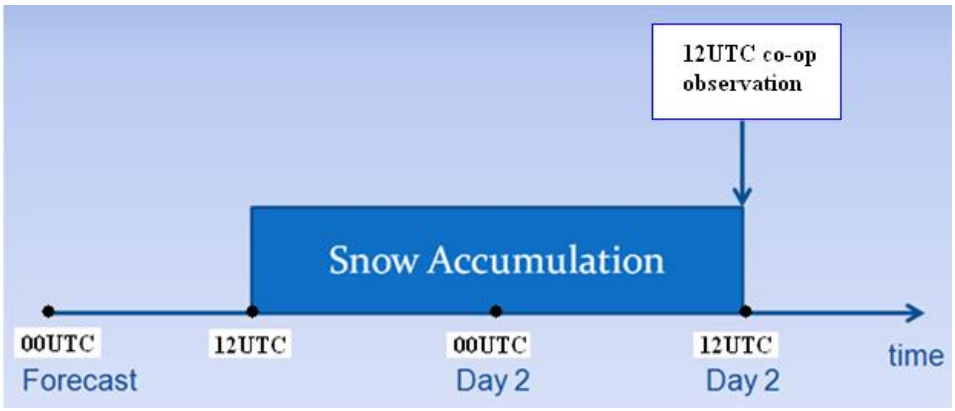

Figure 2-3. Process layout for the 12-36 hour snowfall accumulation forecast and corresponding observation.

Chapter 3

Statistical Guidance Methods

In order to forecast 24-hour snowfall accumulation, a method must be devised that translates the predicted weather variables from the GEFS, as well as the latitude, longitude, and elevation from the co-op sites, into a 24-hour snowfall accumulation prediction. These are referred to as statistical guidance methods because they use statistics to post-process the direct model output into weather variables not directly output by the numerical weather prediction model.

3.1 Linear Regression (LR)

MOS is the statistical technique used in operational forecasting at the NWS to predict numeric variables not explicitly forecast in the model (Glahn and Lowry 1972). MOS uses linear regression equations to relate the predicted variables to the predictands; thus, linear regression (LR) is the baseline method for comparison.

3.2 Least Median Squares (LMS)

In a slight variation of linear regression, the Least Median Squares (LMS) method is used (Rousseeuw and Annick 1987). This method is a robust linear regression that minimizes the median of the squares of residuals instead of minimizing the mean of the residuals from the regression line. It repeatedly applies ordinary linear regression to subsamples of the data and outputs the solution that has the smallest median-squared error.

3.3 Artificial Neural Network (ANN)

The first nonlinear statistical guidance method used here is an Artificial Neural Network (ANN), which is depicted in Figure 3-1 (Rosenblatt 1958). The goal is to improve upon LR by

capturing nonlinear relationships among the predictors. This simplified diagram shows four predictors fed into one hidden layer consisting of five nodes. These five nodes are connected to the output layer, or prediction, which in this case is the 24-hour snowfall accumulation forecast.

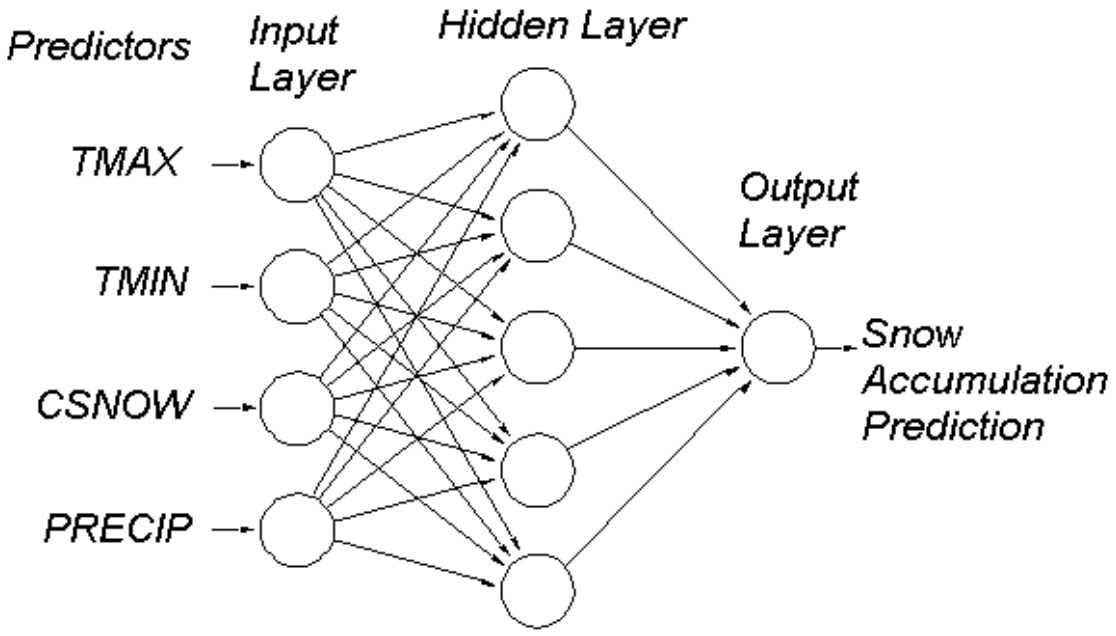

Figure 3-1. Schematic of an Artificial Neural Network.

The ANN used in this study is a feed-forward neural network trained by a backpropagation algorithm, also known as a multi-layer perceptron (Rosenblatt 1958). This ANN configuration is one hidden layer, a learning rate of 0.1 and a momentum of 0.1. The back-propagating algorithm goes through 500 training cycles to find the optimal set of model weights. For level-one, the hidden layer contains 30 nodes, while for level-two the ANN contains 58 nodes. These configurations are chosen because they produce the lowest Root Mean Square Error (RMSE) on the control ensemble member using a three-fold cross-validation with 50 training cycles. Increasing the momentum or learning rate increased the RMSE, and decreasing these parameter

values produced statistically insignificant improvements in RMSE. The addition of a second hidden layer or decay function increased the RMSE; therefore, neither were used in the final configuration.

The activation function (1) is the standard sigmoid function,

$$Y(x) = \frac{1}{1+e^{-x}} \tag{1}$$

where *Y* is the snowfall prediction for instance *x*. The predictor values are scaled to range from -1 to +1. The activation function defines the output of the node given a set of predictors. This output node is linear for the numerical regression task of snowfall accumulation prediction.

3.4 Radial Basis Function Network (RBF)

A method similar to the Artificial Neural Network used is a Radial Basis Function (RBF) network (Frank and Witten 2005). The key difference between an ANN and an RBF network is the way in which the hidden layers perform computation. The RBF network uses Gaussian radial basis functions (2) as the activation functions, and the ANN uses the sigmoid function (1) as the activation function. A sigmoid function divides the pattern space using planes (or hyperplanes), but the Gaussian radial basis function uses circles (or hyperspheres). Equation (2) is the Gaussian radial basis function *Y*

$$Y(x - c_i) = \frac{1}{e^{\beta (x-c_i)^2}} \quad for\ \beta > 0 \tag{2}$$

in which c_i is the center vector for neuron i, and β is a weight. The radius, $x - c_i$ is the distance from the center of the hypersphere to the instance *x*. The Gaussian radial basis function

$$P(x) = \sum_{i=1}^{N} a_i Y(x - c_i) \tag{3}$$

then predicts the output *P(x)*, or the snowfall accumulation, via equation (3). *N* is the number of neurons in the hidden layer and the three weights a_i c_i, and β, are tuned to optimize by fit between the predictions and the training data. In this RBF algorithm, the k-means clustering

algorithm is used to provide the basis functions and the algorithm learns a linear regression on top of that (Witten and Frank 2005). The activations of the basis functions are normalized to sum to one before they are fed into the linear models.

The RBF network with three-fold cross validation on the control ensemble member for level-one tested various cluster sizes from 2 to 300 in order to determine the optimal configuration. Three-fold cross validation is used to assess how the results of a method will generalize on a test dataset. Three-fold cross validation splits the dataset into three equal subsets. Then, the method is trained on two of the subsets and predicts the third subset. This is repeated two more times to predict the other two subsets. Finally, the results are averaged from the three rounds. Although ten-fold cross validation is generally used (Witten and Frank 2005), three-fold cross validation is used for finding the optimal configuration because it still generalizes well and is more computationally efficient. Figure 3-2 plots the RMSE of the RBF network for the different cluster numbers.

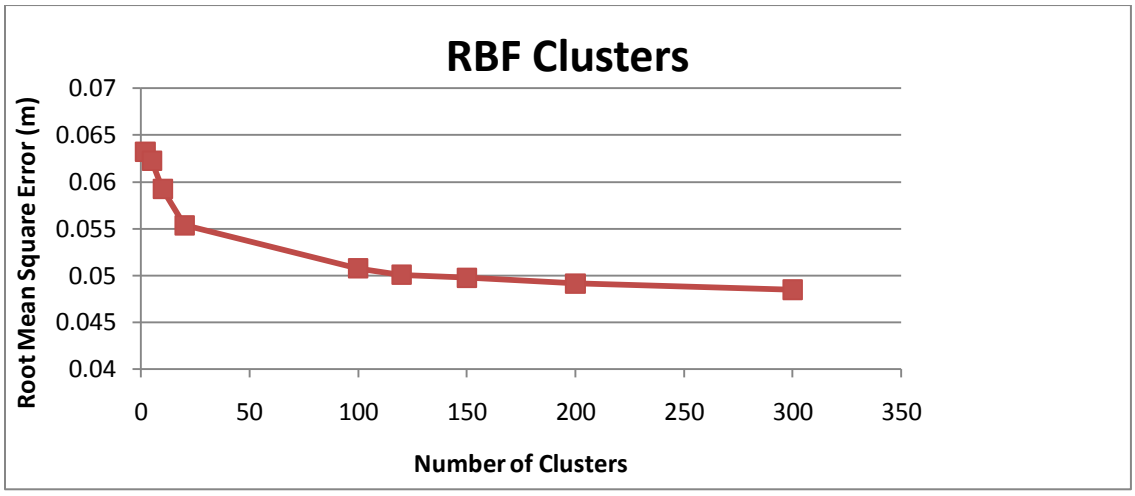

Figure 3-2. Sensitivity study to determine the optimal cluster size for the RBF network.

The RMSE decreases noticeably until approximately 120 clusters and then is approximately level. Therefore, 120 clusters were used in the RBF network.

The minimum standard deviation for the clusters was set to 0.1 and the iterations stop when the value for the ridge regression is less than 1.0×10^{-8}, because these values produced the lowest RMSE for the RBF configuration with 120 clusters. Ridge regression was used because it avoids the issue of predictor collinearity by using a penalized least squares procedure (Witten and Frank 2005). For level two, 75 clusters were used in the RBF network because it produced the lowest RMSE with a three-fold cross validation on the level two dataset for the control ensemble member. More clusters than 75 results in an exponentially increase of the RMSE, which was likely the result of overfitting.

3.5 Conjunctive Rule (CR)

A Conjunctive Rule (CR) is a machine learning algorithm that bases its predictions on setting rules for the predictors (Witten and Frank 2005). Conjunctive rules are learned by determining conditions shared by the examples. A limitation of conjunctive rules occurs when problems do not have a single set of necessary and sufficient conditions. A rule consists of antecedents "AND"ed together and the consequent for the regression (Witten and Frank 2005). The consequent is the mean for the numeric predictors in the dataset. Uncovered test instances are assigned the default mean value of the uncovered training instances. The algorithm selects an antecedent by computing the information gain from each antecedent and then prunes the generated rule. Pruning is done with a reduced-error pruning technique that uses the weighted average of the mean-squared errors on the pruned data to determine the amount of pruning required. In regression problems like snowfall accumulation predictions, the information gain is the weighted average of the mean-squared errors of both the data covered and the data not covered by the rule. A simple example of a conjunctive rule technique is predicting whether

freezing rain is possible. If the temperature is less than 0°C and the minimum relative humidity is greater than 99%, then freezing rain would be predicted. The parameter that was tuned in the conjunctive rule was the number of rules to be used. The results for three-fold cross validation on the control ensemble member with the level-one dataset produces no significant difference among 2, 3, 5, 10, and 25 rules, with all RMSE values between 0.0566 and 0.0571 meters. Thus, the default of three rules was used in the final configuration.

3.6 Support Vector Regression (SVR)

Another method tested was a type of support vector machine for regression (SVR) (Smola and Scholkopf, 2004). An issue with many linear models is that they only can represent linear boundaries between classes; a SVR method uses linear models to produce nonlinear class boundaries, however (Witten and Frank 2005). The key in SVR is to transform the input, or predictors, using nonlinear mapping that transforms the predictor space into a new space. A linear model, specifically a maximum margin hyperplane, developed in this new space can represent a nonlinear decision boundary in the original space (Witten and Frank 2005).

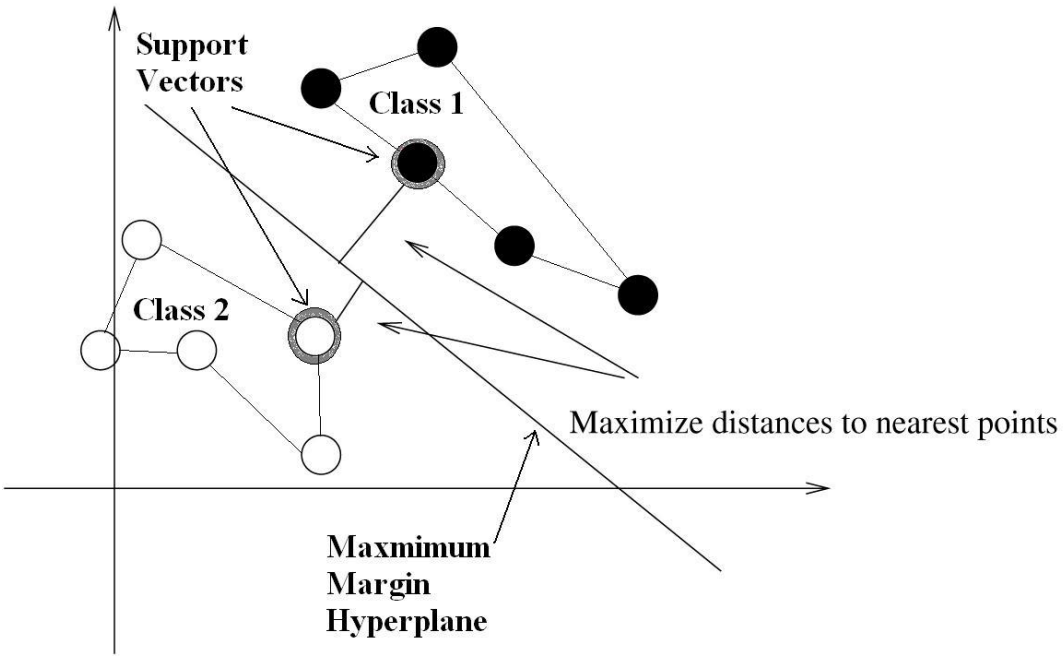

Figure 3-3. Schematic of a maximum margin hyperplane and support vectors separating two classes in support vector classification.

Figure 3-3 describes the design of a support vector classifier. In this two-class example, the instances are separated into two classes, labeled Class 1 and Class 2. The maximum margin hyperplane is the linear model that separates the classes while maximizing the separation between classes. The instances, i.e. the individual observations and corresponding predictors mapped on a high-dimensional space, that are closest to the maximum margin hyperplane are called the support vectors. These support vectors uniquely define the maximum margin hyperplane for the learning algorithm. This example also shows how overfitting is rarely an issue with support vectors because the two support vectors describe the maximum margin hyperplane and the other instances do not affect the hyperplane. Thus, the maximum margin hyperplane remains relatively stable even when moving or deleting the non-support vector training instances. In order to form a prediction, a dot product between the test instance and every support vector is calculated. This

process can be very computationally expensive when there are many predictors, as is the situation with the snowfall prediction dataset of over 100 predictors.

Although the concept of maximum margin hyperplane applies to classification, it serves as a basis for numerical regression. Both classification and numerical regression produce a model represented by support vectors that can be applied to nonlinear problems through kernel functions. There are three main differences between classification and numerical regression in support vector machines. First, the deviations up to a certain user-specified threshold are discarded in numerical regression. This threshold determines how well the kernel functions are fit to the dataset. Second, when minimizing the error, the risk of overfitting is reduced by simultaneously trying to maximize the flatness of the function kernel. The third difference is that the error that is minimized is the absolute error of the prediction, instead of the squared error.

This support vector regression method uses a dot kernel type, a convergence epsilon of 0.0010, and a maximum of 100000 iterations. This configuration is chosen because it provides the lowest RMSE on the three-fold cross validation compared to other convergence epsilons and maximum number of iterations. A dot kernel type also produces lower RMSE than a radial kernel type.

3.7 K-Nearest Neighbor (kNN)

The final method tested was a k-nearest neighbor (kNN) algorithm that finds the closest k training instances in Euclidean distance to the given test instance (Witten and Frank 2005). A distance-weighted average of these k-nearest neighbors was used to predict the snowfall accumulation. Thus, the algorithm finds the closest k neighbors and averages their corresponding snowfall accumulation by 1/distance. The number of k training instances was selected using leave-one-out cross-validation, given an upper limit for k of 100. Leave-one-out cross validation

determined the optimal number of k to be 6 for the level one dataset and 5 for the level two dataset.

Figure 3-4 illustrates how the k-nearest neighbor algorithm works. All dots correspond to instances mapped on a high-dimensional space. Instances are the GEFS interpolated forecasts, latitude, longitude, and elevation valid for each snowfall observation. The center orange dot corresponds to the test instance that the k-nearest neighbor method predicts. The algorithm searches for the closest ten instances in this space, which is represented by all the dots inside the circle. The method then computes the distance-weighted average snowfall accumulation for these ten instances. In this diagram, out of the ten closest instances, eight have 0.2 m accumulations, one has a 0.1 m accumulation, and one has a 0.3 m accumulation. Thus, the forecast given by the k-nearest neighbor method would be 0.2 m, assuming the 0.1 m and 0.3 m observations are the same distance from the test instance.

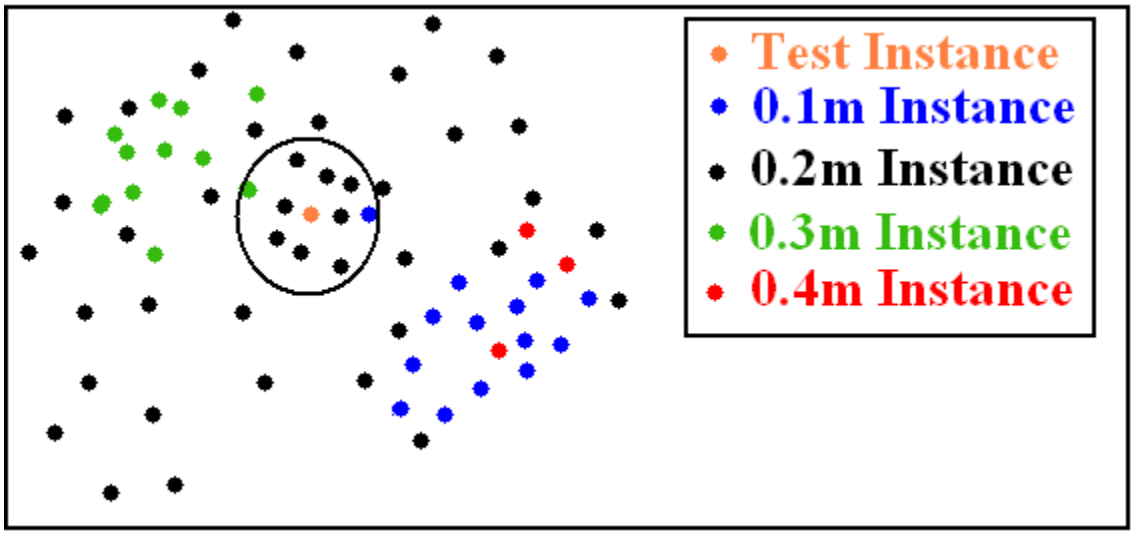

Figure 3-4. Illustration of the k-nearest neighbor algorithm. All dots correspond to instances mapped on a high dimensional space. For k = 10, the distance weighted average of the ten instances inside the circle are used to predict the test instance in orange.

All of the statistical guidance methods were trained on the high resolution ensemble member using a ten-fold cross validation (Wilks 2005). Ten-fold cross validation uses the same methodology as three-fold cross validation, but instead of splitting the dataset into three equal subsets, it splits the dataset into ten equal subsets. After the Root Mean Square Error (RMSE) of the 24-hour snowfall accumulation prediction is minimized in the cross validation, the statistical guidance models are saved. The models were next applied to each ensemble member individually to form a 15-member ensemble of 24-hour snowfall accumulation predictions. After the predictions were made by all statistical guidance methods, all forecasts below a trace, which are reported as 0.04 inches, or 0.001016 m, are set to 0.001016 m. In addition, all forecasts greater than 36 inches, or 0.9144 m, are reset to that value in order to provide an upper boundary of snowfall accumulation, and therefore to eliminate outliers from skewing the results of the methods.

Chapter 4

Results

The accuracy of the methods is evaluated using the Mean Absolute Error (MAE) of the consensus forecast. Ensemble spread is assessed with rank histograms, spread-skill relationships, and quantile-quantile plots. Lastly, the value of splitting the datasets is tested by training the kNN method on a combined dataset and analyzing the results of the level-two instances.

4.1 Accuracy Testing

The ensemble mean consensus forecast was calculated by averaging the forecasts from the 15 individual ensemble members in order to test the deterministic forecast accuracy of the statistical guidance methods,. The Mean Absolute Error (MAE) of the consensus forecast is the mean absolute difference between the consensus forecast and the observation. The MAE is averaged over all instances for each level in the Northeast separately. Table 4-1 shows the MAE results for the statistical guidance methods for both levels. The k-nearest neighbor method produced the lowest MAE of all the methods tested. The ANN was the second-best method and the SVR was the third best method. Linear regression performed in the middle of the techniques for both levels. The least accurate methods were the CR, LMS, and RBF.

Table 4-1. Mean absolute error for all statistical guidance methods on both the level-one and level-two datasets.

	Level 1			Level 2	
Method	MAE (m)	Ranking	Method	MAE (m)	Ranking
ANN	0.0236	2	ANN	0.0232	2
CR	0.0327	6	CR	0.0327	5
kNN	0.0168	1	kNN	0.0186	1
LMS	0.0318	5	LMS	0.0353	6
LR	0.0298	4	LR	0.0326	4
RBF	0.0370	7	RBF	0.0426	7
SVR	0.0267	3	SVR	0.0323	3

A paired two-sample student t-test with unequal variances was performed to examine if the results were significantly different among the statistical guidance methods. The null hypothesis is that the two samples of forecast errors, one from each of two methods, have the same mean value. Stated another way, the null hypothesis is that the two methods produce the same mean absolute error. A p-value less than 0.05 rejects the null hypothesis. Thus, values less than 0.05 indicate that the mean absolute error of the forecasting method in the column is significantly different from the mean absolute error of the forecasting method in the row. The p-value results for the t-test for level-one are shown in Table 4-2 and the results for level-two are shown in Table 4-3. For the level one analysis, all methods produce significantly different mean absolute errors than all other methods at the 95% confidence level. For the level two dataset, there were three combinations of methods that produced insignificantly different mean absolute errors because the null hypothesis could not be rejected: LMS and LR; SVR and LR; and CR and RBF.

Thus, the most accurate method, kNN, produced significantly better MAEs from all other methods on both datasets. For both levels, three methods produced more accurate forecasts than linear regression; kNN, ANN, and SVR; the MAE for SVR was not significantly different from that for LR on the level two dataset, however.

Table 4-2. Paired two-sample student t-test results for the level-one dataset. Values less than 0.05 are statistically significant at the 95% level.

Level One	ANN	LR	K	LMS	RBF	SVR	CR
ANN	-	0	0	0	0	0	0
LR	0	-	0	0	0	0	0
kNN	0	0	-	0	0	0	0
LMS	0	0	0	-	0	0	0.0156
RBF	0	0	0	0	-	0	0
SVR	0	0	0	0	0	-	0
CR	0	0	0	0.0156	0	0	-

Table 4-3. Paired two-sample student t-test results for the level-two dataset. Values less than 0.05 are statistically significant at the 95% level.

Level Two	ANN	LR	K	LMS	RBF	SVR	CR
ANN	-	0	0.0062	0	0	0	0
LR	0	-	0	0.0796	0	0.8409	0
kNN	0.0062	0	-	0	0	0	0
LMS	0	0.0796	0	-	0	0	0
RBF	0	0	0	0	-	0	0.1025
SVR	0	0.8409	0	0	0	-	0
CR	0	0	0	0	0.1025	0	-

4.2 Ensemble Spread Tests

It is also important to evaluate the ensemble spread, or the forecast uncertainty, given by the statistical guidance methods. One method to display ensemble spread is rank histograms. The rank histogram was developed independently by Anderson (1996), Hamill and Colucci (1996, 1997), and Talagrand et al. (1997) 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 to highest (Hamill 2001). For example, if the verifying snow accumulation observation is lower than the lowest ensemble member forecast, then bin one, the leftmost bin, would get a tally. Likewise, if the verifying observation is higher

than the lowest ensemble member forecast, but lower than the second highest ensemble member, then the second bin from the left would get a tally. There are thus 16 bins for the 15-member ensemble studied here because the verifying observation could be lower than all ensemble members, higher than all ensemble members, or in between any of the 15 sorted members.

The rank histograms for all statistical guidance methods are displayed in Figures 4-1 and 4-2 for level-one and level-two respectively. Almost all of these rank histograms have a U-shape for all statistical guidance methods, which was traditionally identified as the ensemble being underdispersive. An underdispersive ensemble has a spread that is not great enough to adequately capture the uncertainty in the forecast. This result is evidenced by bins one (leftmost) and 16 (rightmost) having higher tallies than the other bins, meaning that the snowfall accumulation observation tended to be less than all of the ensemble member forecasts or greater than all of the ensemble member forecasts. The lone exception is for the ANN method for level two that has more of an under-predictive rank histogram, that is, the snowfall accumulation observation tends to be greater than most or all ensemble member predictions. These rank histograms show that none of the methods produced a calibrated or flat ensemble. However, Marzban et al. (2010) showed that a U-shaped rank histogram does not necessarily mean that the ensemble is underdispersive because correlations among ensemble members can also produce a U-shaped rank histogram.

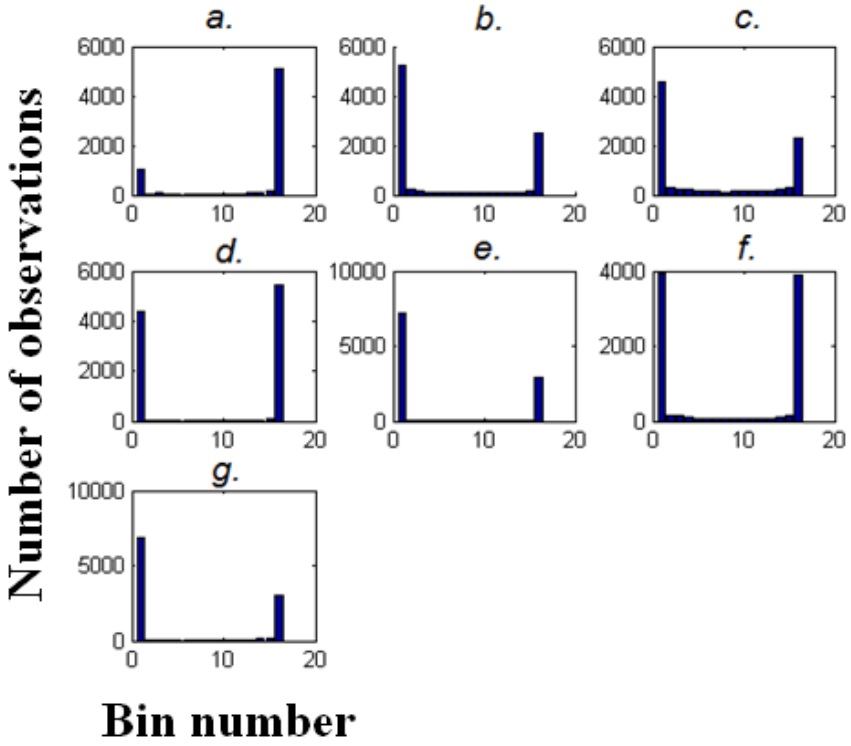

Figure 4-1. Rank histograms for all statistical guidance methods on the level-one dataset. Here, a. is ANN, b. is LR, c. is kNN, d. is LMS, e. is RBF, f. is SVR, and g. is CR. All methods show a U-shaped rank histogram.

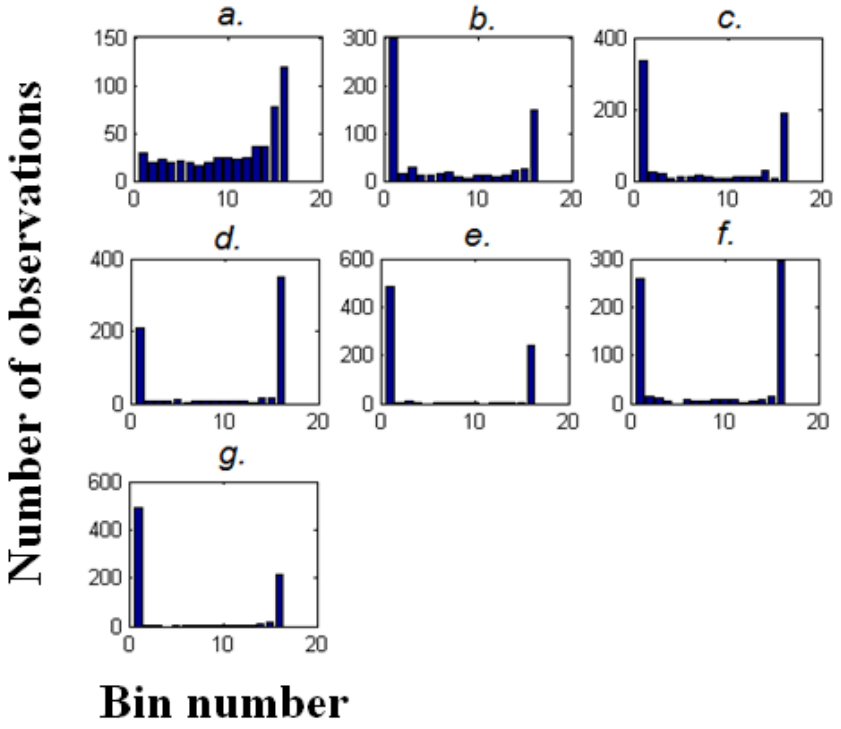

Figure 4-2. Rank histograms for all statistical guidance methods on the level-two dataset. Here, a. is ANN, b. is LR, c. is kNN, d. is LMS, e. is RBF, f. is SVR, and g. is CR. All methods except the ANN show a U-shaped rank histogram.

A simple prediction method using the GEFS direct model output is used in order to test whether the statistical guidance methods are the cause of the U-shaped rank histograms, or if the cause of the U-shaped rank histograms is from the GEFS direct model output. This method sums the accumulated precipitation from each 6-hr interval into a 24-hr total accumulated precipitation. Then, the liquid water equivalent in kg/m2 is converted to inches of rain, and finally a snowfall accumulation prediction is made using a snow-to-rain ratio of 10:1. All forecasts that are greater than 36 inches of snowfall accumulation are given the value of 36 inches and all forecasts less than a trace, or 0.04 inches, are given the value of a trace. The rank histogram for this method, Figure 4-3, shows this simple prediction method still produces a U-shaped rank histogram. The first bin is higher than the last bin, which is expected because it is possible that some of the

precipitation fell as rain, sleet, or freezing rain during the 24-hr period. This analysis shows that the under-variability in the forecasts, as shown by the U-shape rank histogram in Figure 4-3, is likely a result of the direct model output rather than the statistical guidance methods.

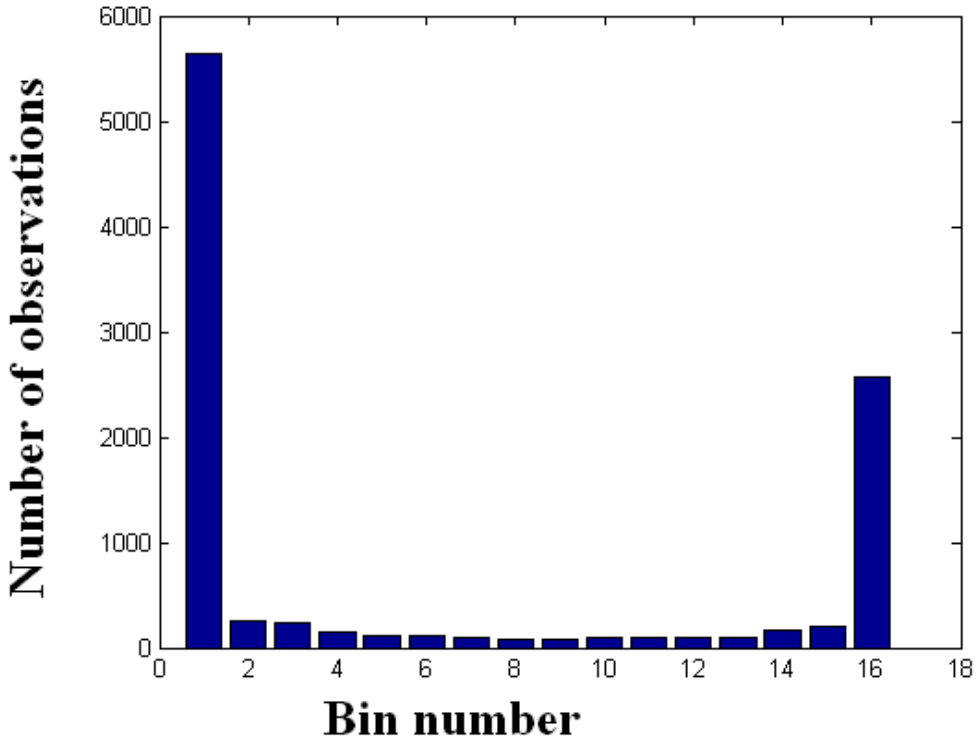

Figure 4-3. Rank histograms for the simple prediction method based off the GEFS direct model output accumulated precipitation. This method also produces a U-shaped rank histogram.

Another method for examining ensemble spread or calibration is analyzing the spread-skill relationship, which is a measure of the correlation between the ensemble spread and the ensemble mean error (Whitaker and Loughe 1998). The ensemble spread is calculated by finding the standard deviation of the ensemble member forecasts. The ensemble error is the absolute difference between the ensemble mean consensus forecast and the snowfall accumulation observation. A calibrated ensemble should show a y = x correlation, or a slope of unity, between

the ensemble error and the ensemble spread. Tables 4-4 and 4-5 show the correlation coefficient, slope, and intercept for a linear fit on the level-one and level-two datasets respectively.

Table 4-4. Correlation coefficient (R), slope and intercept for all statistical guidance methods on the level-one dataset. The highest correlation coefficient value is for the kNN method.

Method	R	Slope	Intercept
ANN	0.33	1.12	0.02
CR	0.17	0.52	0.03
kNN	0.80	1.66	0.00
LMS	0.06	0.97	0.03
LR	0.25	1.20	0.02
RBF	0.10	1.28	0.03
SVR	0.32	3.28	0.01

Table 4-5. Correlation coefficient (R), slope and intercept for all statistical guidance methods on the level-two dataset. The highest correlation coefficient value is for the kNN method.

Method	R	Slope	Intercept
ANN	0.55	0.61	0.01
CR	0.26	0.51	0.04
kNN	0.90	2.30	0.00
LMS	0.10	0.24	0.03
LR	0.44	0.78	0.02
RBF	-0.07	1.28	0.03
SVR	0.12	0.29	0.03

The method with the highest correlation coefficient values for both levels is the k-nearest neighbor method. This indicates that the kNN method produced the best correlation between ensemble spread and ensemble error. Figures 4-4 and 4-5 show the spread-skill plots for the k-nearest method for levels one and two respectively. The plots indicate that ensemble for both levels, the slope is greater than one, which means the ensemble mean error was on average greater than the ensemble spread; that is, the ensemble is underdispersive.

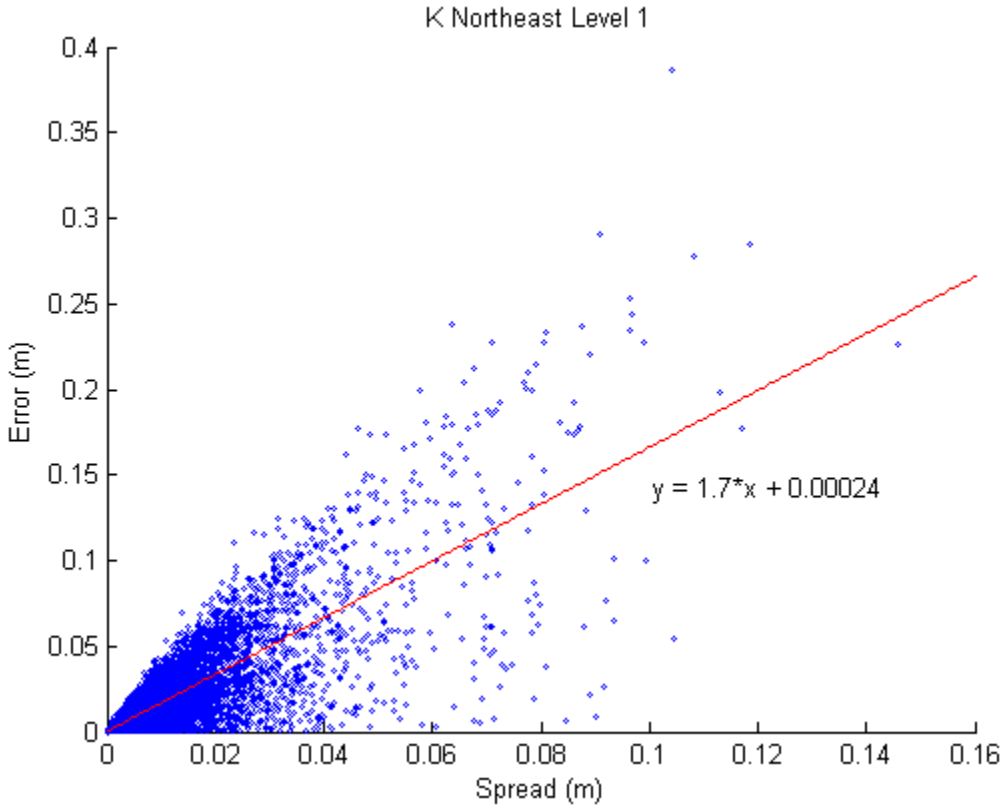

Figure 4-4. Spread-skill plot for the k-nearest neighbor method for the level-one dataset.

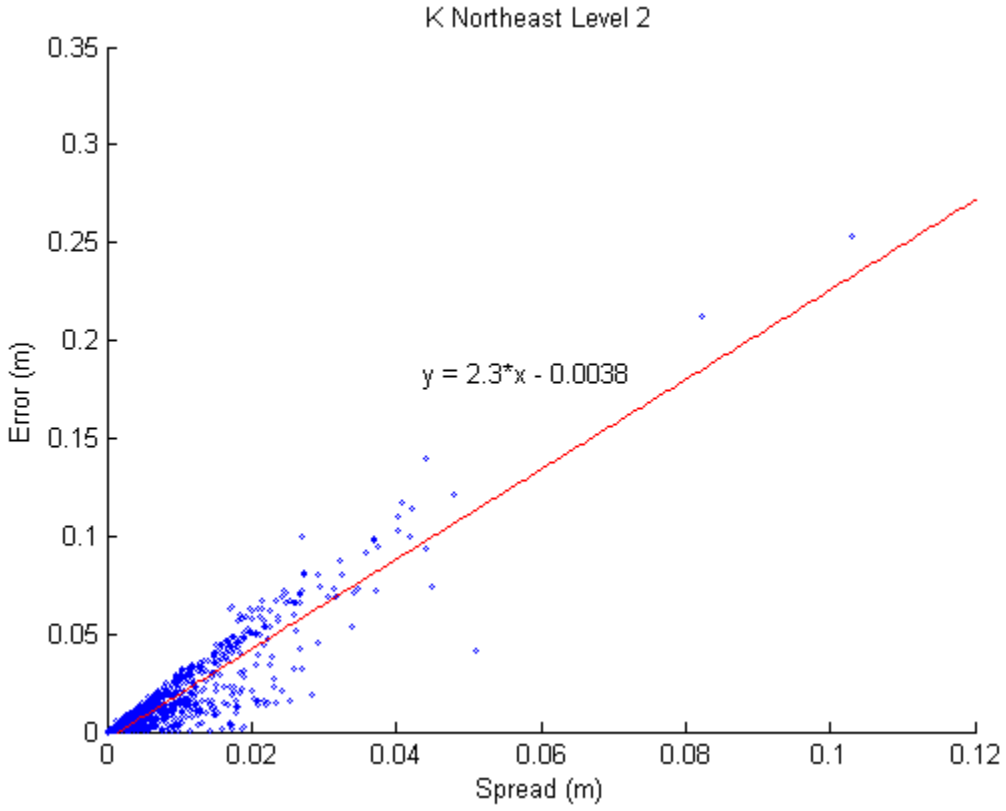

Figure 4-5. Spread-skill plot for the k-nearest neighbor method for the level-two dataset

The high correlation coefficient value indicates that the correlation between ensemble error and spread is high and that the relationship can be accurately calibrated. Other methods have slopes closer to unity but correlation coefficients much less than one, such as the LMS method for level-one that has a slope of 0.9692 but a correlation coefficient of only 0.0593. This means that the slope of the best-fit line is close to unity, but is a poor representation of the relationship between spread and error because the correlation coefficient value is low. These spread-skill plots and correlation analyses show that the kNN method produces an ensemble spread that best predicts the ensemble error.

Another technique for evaluating the forecast uncertainty is Quantile-Quantile, or QQ, plots (Wilks 2005). A QQ plot is a method for comparing two probability distributions by

plotting the quantiles of the two variables against each other. A QQ plot is computed by sorting the observations from lowest to highest and the forecasts from lowest to highest independently. These independently sorted observations are then paired with these independently sorted ensemble forecasts. These pairs are then plotted with the observations on the abscissa versus the ensemble member forecasts on the ordinate, as shown by the + signs on Figures 4.6 and 4.7. The different colors represent the different ensemble members. The dashed lines join the first and third quartiles of each distribution and are extended out to the ends of the sample to help evaluate the linearity of the data. A calibrated ensemble will show a 1-1 relationship between the ensemble member forecasts and the observations. Thus, the closer the dashed lines to the solid line, the more calibrated the ensemble. Figures 4-6 and 4-7 display all of the statistical guidance methods QQ plots for level-one and level-two, respectively.

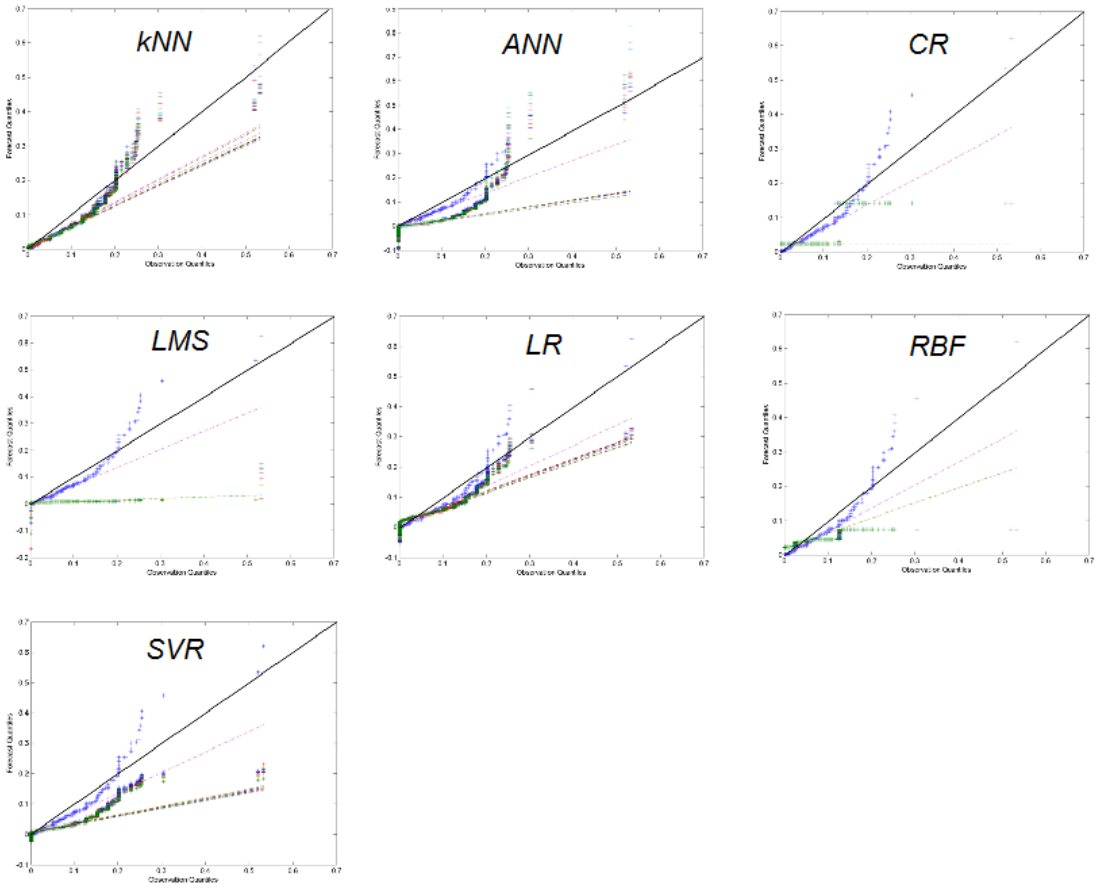

Figure 4-6. QQ plots for all statistical guidance methods on the level-one dataset. The dark line is unity, the colored plus signs represent the different ensemble member quantiles, and the colored dashed lines connect the first and third quartiles of each ensemble member.

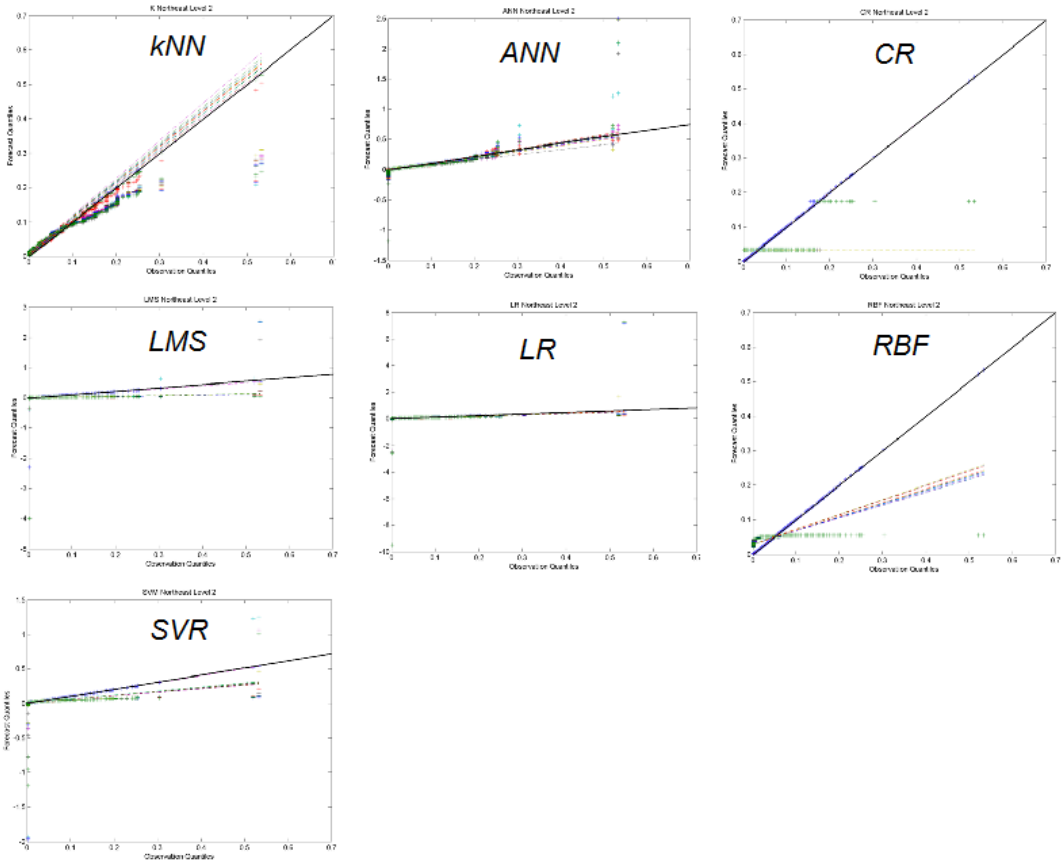

Figure 4-7. QQ plots for all statistical guidance methods on the level-two dataset. The dark line is unity, the colored plus signs represent the different ensemble member quantiles, and the colored dashed lines connect the first and third quartiles of each ensemble member.

There are two distinct features of the QQ plots that show the kNN method produces the best ensemble spread estimates of the methods tested. First, the first to third quartile lines for the kNN method are closest to the y = x line. Second, within the same plot, each of the first to third quartile lines are only slightly different, unlike other methods that have some outlier ensemble members The QQ plots confirm that the kNN method produces the most calibrated ensemble. Shown in Figures 4-8 and 4-9 are larger QQ plots for the kNN method on level-one and level-two datasets.

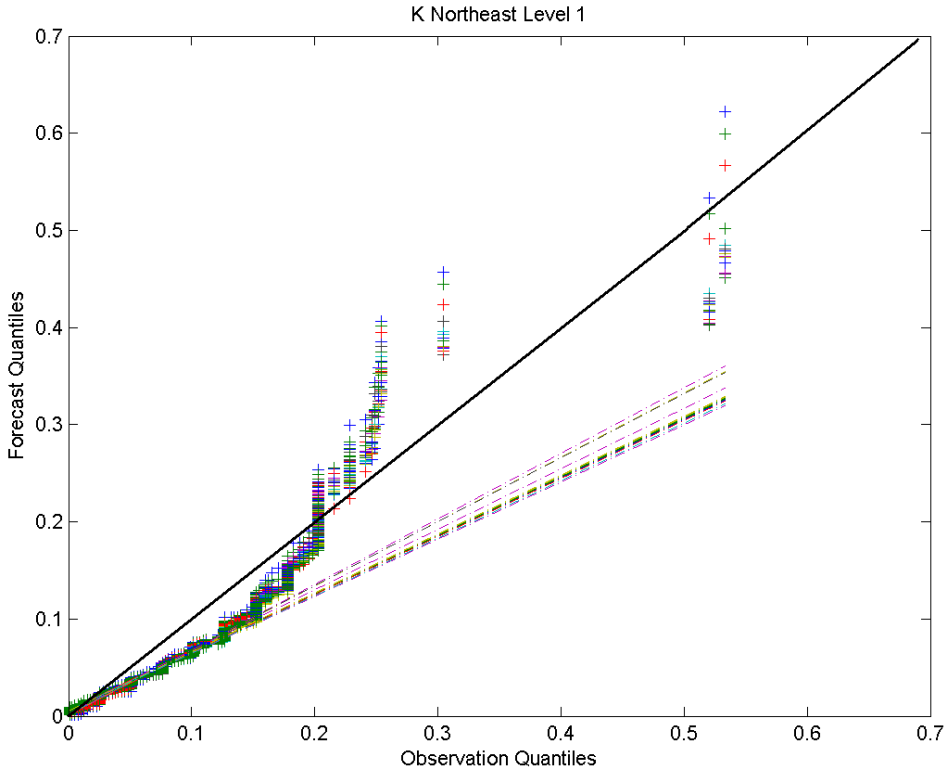

Figure 4-8. QQ plot for kNN method on the level-one dataset. The dark line is y = x, the colored plus signs represent the different ensemble member quantiles, and the colored dashed lines connect the first and third quartiles of each ensemble member.

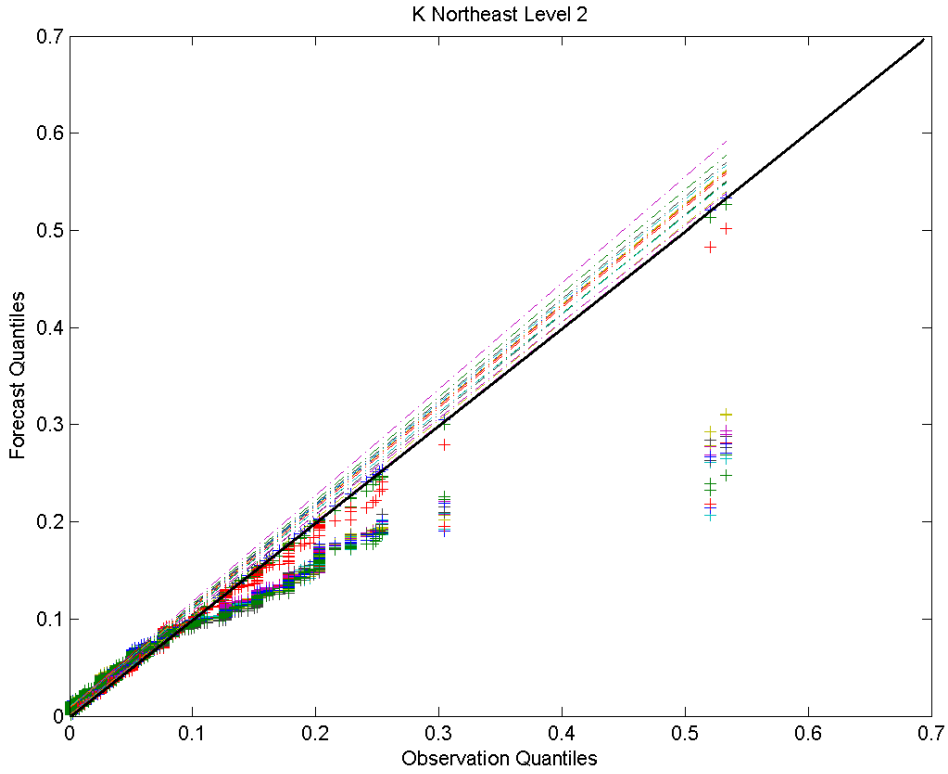

Figure 4-9. QQ plot for kNN method on the level-two dataset. The dark line is y = x, the colored plus signs represent the different ensemble member quantiles, and the colored dashed lines connect the first and third quartiles of each ensemble member.

The QQ plot for kNN method on the level-one dataset, Figure 4-8, had a convex shape compared to the y = x line. This indicates that the kNN method produced ensemble members with probability density function that is negatively skewed. This means that the probability density function of the ensemble forecasts is skewed low compared to the probability density function of the observations. For the level-two QQ plot, the plotted quantiles below the y = x line indicate that the kNN method had a low forecasting bias for snowfall observations greater than 0.1 m. Although the kNN QQ plots are not ideal, they are better than the other methods. In addition, these QQ plots can be used to improve operational forecasting. For example, at a location site above 760 m, if the kNN method was predicting 0.2 m for an impending snowstorm,

the QQ plot shows that the forecasts tend to be biased low, and therefore, the forecaster should predict snowfall to be greater than 0.2 m. Thus, the QQ plots reaffirm that the kNN method produced the best ensemble spread, or calibration.

4.4 Robustness

In addition to the ensemble mean accuracy and ensemble spread metrics, the k-nearest neighbor method was tested on a combined dataset. This approach combined the data from the two levels and kept the 925 hPa relative humidity for the original level-two instances. This dataset had 11180 instances of snowfall accumulation. The goal of this analysis was twofold: to determine if the kNN method was robust enough to handle the combined dataset and to determine if splitting into levels was significant for prediction. As was done for finding the optimal number of k-nearest neighbors on the elevation split datasets, leave-one-out cross validation technique is used and the optimal number of k-nearest neighbors is determined to be six. This kNN model is trained on the high resolution ensemble member using ten-fold cross validation and then applied to each ensemble member individually.

Since there are 10418 observations in the level-one dataset and only 762 in the level-two dataset, the question is whether the skill of the kNN method is degraded on the level-two dataset when training the kNN on the combined dataset. The MAE of these 762 instances from the level-two trained kNN is 0.019 m, while the MAE of these 762 instances from the kNN trained on the combined dataset is 0.022 m. A two-sample student t-test with unequal variances proved that the MAE for the kNN method on the combined dataset is significantly different at the 95% level.

For the spread-skill analysis, the level-two trained kNN produces a correlation coefficient of 0.90 with a slope of 2.30 when the kNN method trained on the combined dataset produces a correlation coefficient of 0.87 and a slope of 2.36. Both the correlation coefficient and slope

show decreased forecasting performance of the kNN method when trained on the combined analysis.

The QQ plot for the kNN method trained on the combined dataset confirms the degraded skill of the kNN method by training on the combined dataset. The QQ plot, Figure 4-10, shows that there are some outlier ensemble members, as evidences by their first to third quartile lines further from the y = x line. In addition, none of the ensemble member lines that connect the first and third quartiles are as close to the y = x line as they are in the level-two trained kNN method, Figure 4-9.

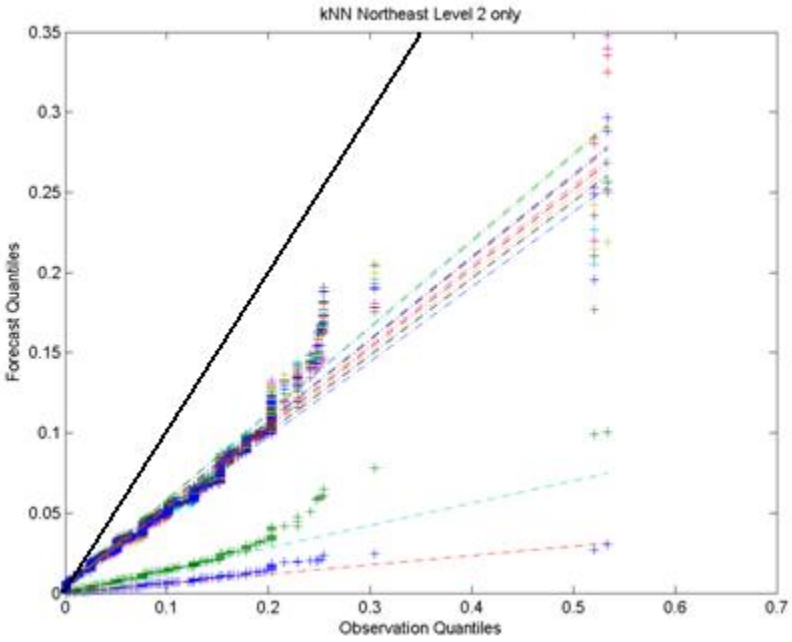

Figure 4-10. QQ plot for kNN trained method on the combined dataset with only the level-two instances plotted. The dark link is y = x, the colored plus signs represent the different ensemble member quantiles, and the colored dashed lines connect the first and third quartiles of each ensemble member.

Thus, the kNN method on the combined analysis provides evidence that the kNN method is not robust enough to handle not splitting the dataset into elevation subsets. In addition, a dataset that included the Rocky Mountains with higher elevations that have different

climatologies or datasets that have a fairly equal number of observations may degrade the forecasting ability of the kNN method even more.

Chapter 5

Conclusions

Seven different statistical guidance methods were tested for producing 24-hour snowfall accumulation forecasts from the Global Ensemble Forecast System direct model output. These methods were trained to reduce the error of the control ensemble member and then applied to each ensemble member individually. An average of these individual ensemble members into a single consensus forecast produced a deterministic snowfall accumulation forecast. Several techniques were used to examine the ensemble spread or calibration. The linear regression method was used as the baseline for comparison because it is the method closest to the National Weather Service standard operational statistical guidance method, Model Output Statistics. Three methods, SVR, ANN, and kNN, showed potential improvements over linear regression in terms of accuracy with the kNN method producing the best ensemble spread calibration, or uncertainty estimate, of all the methods tested.

The results demonstrate that the kNN statistical guidance method of predicting 24 hr snowfall accumulation provides more accurate deterministic forecasts than the linear regression method or any other method. The MAE of the kNN method was significantly smaller than that for the second most accurate method, the ANN. The RBF network and SVR did not perform as well as did the ANN or the kNN. The disadvantage of the RBF network and SVR is that both methods give every predictor the same weight because they are equally valued in the distance computation; thus, these methods cannot ignore irrelevant predictors.

To evaluate the ensemble spread or calibration, three methods were used: rank histograms, spread-skill relationships, and Quantile-Quantile (QQ) plots. The rank histograms provided evidence that nearly all statistical guidance methods produced U-shaped ranked histograms,

which can indicate underdispersive or correlated ensemble members. Examining the correlation coefficients for the spread-skill relationships showed that the kNN method produced the highest correlation between ensemble spread and ensemble error for both level-one and level-two. Thus, its spread can be calibrated via a linear transformation to produce useful error estimates. In addition, the QQ plots confirmed that the kNN method produced the most appropriately calibrated ensemble.

In conclusion, the kNN statistical guidance method outperforms not only linear regression, but all other methods in terms of the both deterministic forecast accuracy and the ensemble spread. Thus, the kNN method appears to be the best statistical guidance method for forecasting 24-hour snowfall accumulation for this dataset.

This study examined a 24 hour snowfall accumulation with a forecast lead time of 12 hours. Examining these methods at longer lead times would help confirm the rankings of the methods. The support vector regression technique required extensive computer resources. There are many versions of support vector regression techniques, but the extensive computer time required to test each limited the number of different variations tested. It may be possible to combine support vector regression with a technique that preprocesses the data to allow for faster computations and potentially more accurate predictions. In addition, a performance, regime, or elevation-weighted average of these statistical guidance methods may improve the forecast accuracy and spread of the ensemble.

Future work will involve calibrating the kNN method to produce better ensemble spread estimates. The QQ plot for the kNN method on the level-one dataset showed that the probability density function of the forecasts is negatively skewed compared to the probability density function of the observations. A post-processing method could be devised to transform the probability density function of the forecasts to more accurately match the probability density function of the observations, which would lead to better ensemble spread estimates. For the

level-two trained kNN method, the QQ plot showed a low forecasting bias for instances greater than 0.1 m. An equation could be developed to post-process the kNN method to increase the forecasts that were already previously greater than 0.1 m. The spread-skill relationships for the kNN method on both levels datasets produced slopes greater than one; thus, a post-processing method that increases the ensemble spread compared to the ensemble error would provide better forecast uncertainty estimates.

Bibliography

Allen, R. L. 2001: Observational data and MOS: The challenges in creating high-quality guidance. *Preprints 18th Conference on Weather Analysis and Forecasting*, Ft. Lauderdale, Amer. Meteor. Soc., 322-326.

Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecasts from Ensemble Model Integrations. *J. Climate*, **9**, 1518–1530.Baxter, M.A., C.E. Graves, and J.T. Moore, 2005: A Climatology of Snow-to-Liquid Ratio for the Contiguous United States. *Wea. Forecasting*, **20**, 729–744.

Cosgrove, R. L., and B. Sfanos, 2004: Producing MOS Snowfall Amount Forecasts from Cooperative Observer Report. Preprints AMS 20th Conference on Weather Analysis and Forecasting, January 11-15, 2004, Seattle, WA.

Evans, M., and M.L. Jurewicz, 2009: Correlations between Analyses and Forecasts of Banded Heavy Snow Ingredients and Observed Snowfall. *Wea. Forecasting*, **24**, 337–350.

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.

Glahn, B., M. Peroutka, J. Wiedenfeld, J. Wagner, G. Zylstra, B. Schuknecht, and B. Jackson, 2009: MOS Uncertainty Estimates in an Ensemble Framework. *Mon. Wea. Rev.*, **137**, 246–268.

Gneiting, T. and Raftery, A. E. (2005). Weather forecasting with ensemble methods. *Science*, **310**, 248-249.

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.

Grimit, E.P., and C.F. Mass, 2002: Initial Results of a Mesoscale Short-Range Ensemble Forecasting System over the Pacific Northwest. *Wea. Forecasting*, **17**, 192–205.

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

_____, 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.

_____, and _____, 1997: Verification of Eta-RSM short-range ensemble forecasts. *Mon. Wea. Rev.*, **125,** 1312–1327.

_____, 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.

Kocin P. J., and L. W. Uccellini, 2005: *Northeast Snowstorms*. Vols. 1 and 2, *Meteor. Monogr.*, No. 54, Amer. Meteor. Soc., 818 pp.

Marzban C, Wang R, Kong F, Leyton S (2010) On the Effect of Correlations on Rank Histograms: Reliability of Temperature and Wind-speed Forecasts from Fine-scale Ensemble Reforecasts. Monthly Weather Review: In Press

Miller, J.E., 1946: Cyclogenesis in the Atlantic Coastal Region of the United States. *J. Atmos. Sci.*, **3**, 31–44.

National Climatic Data Center, 2000: *Surface Land Daily Cooperative Summary of the Day TD-3200*, NOAA, U.S. Department of Commerce, 23pp.

National Operational Hydrologic Remote Sensing Center. 2004. *Snow Data Assimilation System (SNODAS) data products at NSIDC*. Boulder, CO: National Snow and Ice Data Center. Digital media.

National Weather Service, 2000: *Cooperative Observer Program*, [available at http://www.weather.gov/om/coop/Publications/coop.PDF]

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.

Ralph, F.M., R.M. Rauber, B.F. Jewett, D.E. Kingsmill, P. Pisano, P. Pugner, R.M. Rasmussen, D.W. Reynolds, T.W. Schlatter, R.E. Stewart, S. Tracton, and J.S. Waldstreicher, 2005: Improving Short-Term (0–48 h) Cool-Season Quantitative Precipitation Forecasting: Recommendations from a USWRP Workshop. *Bull. Amer. Meteor. Soc.*, **86**, 1619–1632.

Roebber, P.J., S.L. Bruening, D.M. Schultz, and J.V. Cortinas, 2003: Improving Snowfall Forecasting by Diagnosing Snow Density. *Wea. Forecasting*, **18**, 264–287.

Roebber, P.J., M.R. Butt, S.J. Reinke, and T.J. Grafenauer, 2007: Real-Time Forecasting of Snowfall Using a Neural Network. *Wea. Forecasting*, **22**, 676–684.

Rousseeuw, P. J., and L. M. Annick, 1987: *Robust Regression and Outlier Detection,* John Wiley & Sons Inc.

Rosenblatt, F., 1958: The Perceptron:A Probabilistic Model for Information Storage and Organization in the Brain. *In. Psychological Review,* **65(6)**, 386-408.

Smola, A. J., and B. Scholkopf, 2004: A Tutorial on Support Vector Regression. *Statistics and Computing* **14(3)**, 199-222.

Tracton, M.S., and E. Kalnay, 1993: Operational Ensemble Prediction at the National Meteorological Center: Practical Aspects. *Wea. Forecasting*, **8**, 379–398.

Ware, E.C., D.M. Schultz, H.E. Brooks, P.J. Roebber, and S.L. Bruening, 2006: Improving Snowfall Forecasting by Accounting for the Climatological Variability of Snow Density. *Wea. Forecasting*, **21**, 94–103.

Whitaker, J.S., and A.F. Loughe, 1998: The Relationship between Ensemble Spread and Ensemble Mean Skill. *Mon. Wea. Rev.*, **126**, 3292–3302.

Wilks, D.S., 2005: *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.