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

Tyler McCandless
The Pennsylvania State University
State College, Pennsylvania

Sue Ellen Haupt
National Center for Atmospheric Research
Boulder, Colorado

George S. Young
The Pennsylvania State University
State College, Pennsylvania

Abstract

The accuracy of snowfall accumulation forecasts has widespread economic and safety consequences. Due to the complex structure and dynamics of 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 specifically to address the uncertainty in weather forecasts. An accurate deterministic forecast as well as an estimate of the uncertainty is of utmost importance to the public; therefore, the ensemble mean is often used as the deterministic forecast and the ensemble spread as the basis for a forecast uncertainty estimate. This approach works for weather parameters directly forecast by the model; however, snowfall accumulation is not directly forecast by the Global Ensemble Forecast System (GEFS). Therefore this study examines nine artificial intelligence methods for producing a 24-h snowfall accumulation prediction from the parameters directly output from the GEFS. These methods are then examined by their deterministic forecast skill using the mean absolute error of the ensemble mean forecast as well as the degree to which the ensemble spread corresponds to forecast uncertainty, which is examined by spread-skill relationships and quantile-quantile plots. Out of the nine methods—an artificial neural network, linear regression, least median squares regression, support vector regression, radial basis function network, conjunctive rule, k-nearest neighbor, regression tree, and an average of these methods—the k-nearest neighbor method produces significantly more accurate forecasts as well as the best calibrated ensemble spread. This postprocessing method would be appropriate for operational forecasting.

Corresponding Author: Dr. Sue Ellen Haupt
Research Applications Laboratory; National Center for Atmospheric Research
3450 Mitchell Lane
Boulder, CO 80301. Email: haupt@ucar.edu

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 road conditions were reported as either snowy and slushy or icy (Kocin and Uccellini 2005). On average 1270 fatalities per year occur during snow/ice road conditions (Kocin and Uccellini 2005). The most severe winter storms also can have an extensive impact on the economy. The National Climatic Data Center (NCDC) estimated the March 1993 and January 1996 storms each resulted in billions of dollars in damage (Kocin and Uccellini 2005). Winter weather forecasts provide state and local departments of transportation with the information required to prepare for these winter weather events efficiently and safely. At the 2005 United States Weather Research Program 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, particularly for this region.

Forecasting snowstorms is a multifaceted problem presenting many challenges. One challenge is the difficulty of obtaining accurate and precise snowfall measurements, due to blowing and drifting that depends on the location and the surface of the observation, melting, compaction, mixed precipitation events, and how often the measurement is taken (Doesken and Leffler 2000). Without consistent and accurate snowfall accumulation observations, it is difficult to evaluate the performance of any forecast system. Another challenge in snowfall forecasting is that both small and large snowfall events can have similar weather conditions prior to accumulating snowfall.

A major development in weather forecasting is the advent of meteorological ensembles. Ensemble forecasts represent a set of possible realizations of future states of the atmosphere. The mean of the ensemble is typically a more accurate forecast than that of a single member (Woodcock and Engel 2005). Grimit 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 in order to improve both the calibration of uncertainty information and the overall accuracy of NWP ensembles (Kolczynski et al. 2009; Delle Monache et al. 2011).

Statistical post-processing methods such as model output statistics (MOS) typically improve the prediction of direct variables (Glahn and Lowry 1972) and has been applied to ensembles as well (Woodcock and Engle 2005). Here we wish to explore whether other statistical post-processing techniques would yield forecast improvement when applied to an ensemble for an indirect variable, specifically snowfall accumulation prediction. Different methods of post-processing ensemble forecasts in order to improve weather prediction has been studied in various contexts (Raftery et al. 2005; Greybush et al. 2008; Glahn et al. 2009). 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 MOS techniques to forecast the conditional probability of snow and the snowfall amount exceeding a specific threshold, given that snowfall occurs, using the Global Forecast System (GFS) model.

The goal of this study is to test the ability of advanced statistical post-processing methods to improve both the overall accuracy and the ensemble calibration of the forecast uncertainty information for 24-h snowfall accumulation predictions from the Global Ensemble Forecast System (GEFS). Improving the forecast accuracy and ensemble uncertainty calibration is vital for operational meteorologists who wish not only to predict snowfall accumulation more accurately, but also to quantitatively address the uncertainty in the prediction. Nine different statistical post-processing methods are tested for producing 24-h snowfall accumulation forecasts from the GEFS direct model output. These methods are trained to reduce the error of a single "control" ensemble member and then applied to each ensemble member individually[1]. These individual ensemble members are then used to produce a single consensus forecast of snowfall accumulation at specific points. The objective is to determine if it would be prudent to use any of these methods operationally. Several techniques are used to examine the calibration of ensemble spread. The accuracy of both results is then evaluated on two separate datasets that are split based on the observation's altitude.

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

[1] This approach is appropriate because all members of the GFS ensemble have the same physics and differ only in initial conditions.

2. Data

a. Cooperative Observing Network

In order to test the validity of any forecast system, or indeed, to create a consistent, reliable statistical post-processing system, an accurate observing system is necessary (Allen 2001). This issue poses particular challenges for snowfall accumulation prediction. In order to achieve adequate spatial coverage, the NCDC Cooperative Summary of the Day (co-op) reports are used as the snowfall observations. 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 sites when observations are taken between 1100 Coordinated Universal Time (UTC) and 1700 UTC are retained. This restriction allows us to compare the observations to 24-h snowfall accumulation forecasts valid for the period ending at 1200 UTC. Moreover, most of the observations are recorded between 1100 UTC and 1700 UTC so little data is lost by the imposition of this restriction 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. The resulting dataset spans the period from 1 October 2006 to 31 March 2007.

b. Global Ensemble Forecast System

The National Center for Environmental Prediction (NCEP) Global Ensemble Forecast System (GEFS) uses the Global Spectral Model (GFS). Due to several changes in the model configuration and number of members, the longest cold season consistent dataset available was 1 October, 2006 to 31 March 1, 2007. During this time period, the GEFS consisted of 15 total ensemble members (individual NWP forecasts): one control run and fourteen perturbations using the NCEP Ensemble Transform Bred Vector (NCEP 2010, Toth and Kalnay 1993). The control run is the GFS model at 3-km resolution at 0 degrees latitude, while the fourteen other ensemble members are 10- km resolution at 0 degrees latitude[2]. The initialization time of the forecasts was 0000 UTC each day, and each forecast was archived at 95.25-km resolution. The GEFS direct model output consists of forecasts for every 6 hr from 0-364 hrs. The GEFS yields 31 forecast variables (Table 1). Some of the GEFS output variables exhibit interdependence, thus three variables are deleted: surface pressure because it is interrelated with mean sea level pressure, 100- hPa temperature because it is interrelated with 2-m temperature, and 100- hPa height because it is interrelated with mean sea level pressure. None of these variables is listed in Table 1.

Variable	Variable 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/m2]	2-m

Table 1. Global Ensemble Forecast System archived variables.

[2] In an ideal setting we would have all ensemble members at the same resolution with consistent perturbations for each ensemble member. This would allow us to train the methods to account for each ensemble member's individual forecast tendencies. For point forecasts, however, interpolating to the same point is adequate.

Although the GEFS predictions and co-op data cover the entire United States, the Northeast United States was the focus of this study. Within this region, Lake Ontario and Lake Erie help generate lake-effect snow, which requires a completely different synoptic weather pattern than does snowfall caused by baroclinic winter storms such as nor'easters or Alberta clippers (Nizol et al. 1995). Therefore, the Northeast United States dataset was structured so that it did not include locations that generally receive appreciable lake-effect snowfall. The approach described herein could, of course, be applied separately to these locations or any other meteorologically homogeneous region.

The Northeast dataset was split into two subsets (herein called levels) based on elevation, with locations below 760 m composing level-one and those above 760 m composing level-two. For the locations above 760 m the dataset does not include the 92- hPa relative humidity for each 6-h forecast interval as that level is frequently below the surface. There are 10,418 snowfall observations in the level-one dataset and 762 observations in the level-two dataset. Figure 1 plots the observation sites 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.

In order to compare the GEFS forecasts on their 95.2 km grid with individual co-op reporting sites, the three-nearest neighbor weighting method was used to interpolate the GEFS gridded forecasts to the co-op locations, in compliance with the method used by NOAA's Meteorology Development Laboratory (2009 personal communication). 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, an inverse distance-weighted average of the three nearest neighbor grid points was used to calculate a forecast for the co-op location.

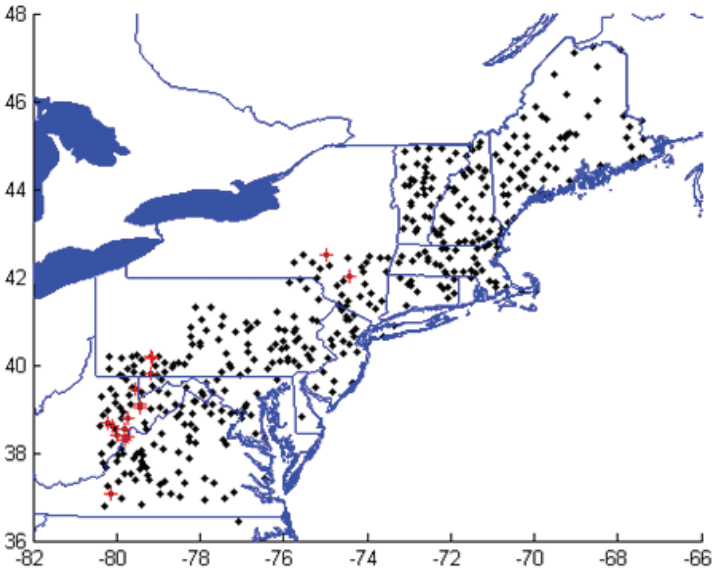

Fig. 1. Snowfall observation sites, with black asterisks marking the locations of the level-one observations and the red plus signs marking the locations of the level-two observations.

The datasets consist of GEFS predicted weather variables at forecast valid times of 12-18 hrs, 18-24 hrs, 24-30 hrs, and 30-36 hrs. These variables were combined with latitude, longitude, and elevation of each station as predictors for the statistical guidance model as shown in Fig. 2. The statistical guidance methods used these variables to predict the total 24 hr snowfall accumulation, conditional on snow occurring. The 1100 UTC o-1700 UTC co-op observation is compared with a 12-36ur prediction from the GEFS.

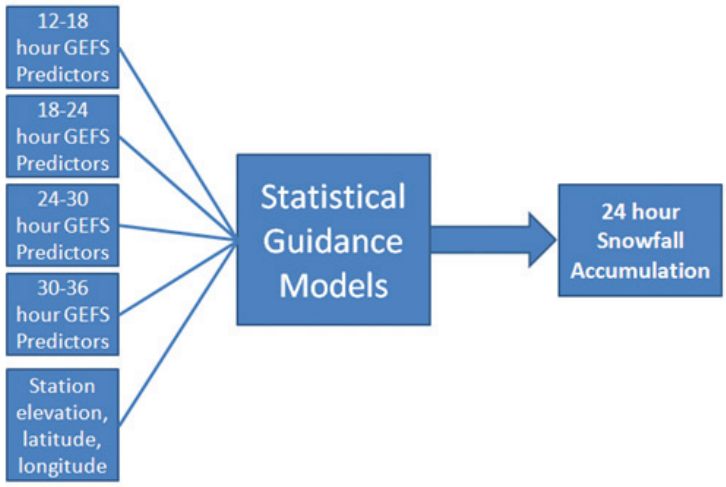

Fig. 2. Schematic of the predictors that the statistical guidance models use to predict the 24-h snowfall accumulation.

3. Statistical Guidance Methods

Eight different statistical post-processing methods as well as the average of the eight results are used to predict snowfall accumulation from the GEFS direct output parameters. These methods were selected because they represent a broad array of commonly used, but fundamentally different, approaches to capturing the relationship between a forecast variable and a set of predictors. We used the software package, RapidMiner (Rapid-i 2010), to configure and test each method. This process is similar to what could be done by operational forecasters. The specific configuration for each method was optimized by using standard cross-validation approaches. Descriptions of each method, the rationale for selecting it, and details of choosing specific configurations for application appear in the appendix.

A brief listing of the rational and configuration for each of the statistical post-processing methods appears in Table 2. Each was trained on the control ensemble member using a ten-fold cross validation. We set the configuration for each method as described above by minimizing the RMSE of the 24 hr snowfall accumulation prediction. The methods were next applied to each ensemble member individually to form a 15-member ensemble of 24 hr snowfall accumulation predictions. After the predictions were made by all statistical guidance methods, all forecasts below a trace were set to 0.001 m. In addition, all forecasts greater than 36 inches, or 0.91 m, were reset to that value in order to provide an upper limit for snowfall accumulation, thereby eliminating outliers that could skew the results.

4. Results

The accuracy of the methods is evaluated using the Mean Absolute Error (MAE) of the consensus forecast. Ensemble spread calibration is assessed with spread-skill relationships and quantile-quantile plots.

a. Accuracy Testing

The ensemble mean consensus forecast is calculated by averaging the forecasts from the 15 individual ensemble members in order to test the deterministic forecast accuracy of the statistical post-processing methods. The MAE is averaged over all instances for each altitude-based site subsets (i.e. levels). Table 3 shows the

Method	Rationale	Application Details
Linear Regression (LR)	Basic comparison method	Ordinary least squares (OLS)
Least Median Squares (LMS)	Basic, but more robust to outliers	Median-based error metric
Artificial Neural Network (ANN)	Captures nonlinear relationships	Multi-layer perceptron with 1 hidden layer, learning rate-0.1, momentum=0.1, sigmoid activation function
Radial Basis Function Network (RBF)	Captures nonlinear relationships with Gaussian radial basis functions	Cluster size of 120 instances with min stand dev=0.1, ridge regression with iteration to
Conjunctive Rule (CR)	Produces predictive rules that are understandable	3 rules
Support Vector Regression (SVR)	Captures nonlinear relationship by mapping to high dimensional space	Kernel operation with convergence level=0.001, max its=100,000
K-Nearest Neighbor (KNN)	Find analogues via clustering	6 clusters for level 1 5 clusters for level 2
Regression Tree (RT)	Forms decision nodes with OLS regression at each node	Min instance leaf=3, min number class variance=0.001, folds for error pruning=3
Consensus (AI)	Combine advantages of all techniques	Average all of the above

Table 2. Statistical post-processing methods used and their application information.

Table 3 (at right). Mean absolute error for all statistical guidance methods on both the level-one and level-two datasets.

	Level One		Level Two	
Method	MAE (m)	Ranking	MAE (m)	Ranking
LR	0.0298	6	0.0326	6
LMS	0.0318	7	0.0353	8
ANN	0.0236	2	0.0232	2
RBF	0.0370	9	0.0426	9
CR	0.0327	8	0.0327	7
SVR	0.0267	5	0.0323	5
kNN	0.0168	1	0.0186	1
RT	0.0262	4	0.0313	4
AI	0.0250	3	0.0286	3

MAE results for the statistical post-processing methods for both levels. The k-nearest neighbor method produces the lowest MAE of all the methods tested. The ANN is the second best method and the AI consensus is the third best method. SVR and LR rank in the middle of the pack for both levels. The least accurate methods are the CR, LMS, and RBF. A paired two-sample student t-test with unequal variances was performed to determine if the results are significantly different among the statistical post-processing methods. The null hypothesis is that the two sets of forecast errors, one from each of two methods, have the same MAE. A p-value less than 0.05 rejects the null hypothesis. Thus, values less than 0.05 indicate that the MAE of the forecasting method in the column is significantly different from the MAE of the forecasting method in the row. The p-value results for the t-test for level-one appear in Table 4 and the results for level-two are shown in Table 5. For the level one analysis, all methods produce significantly different MAEs than all other methods at the 95% confidence level. For the level-two dataset, there are only five combinations of methods that produced insignificantly different MAEs: LMS and LR; SVR and LR; CR and RBF; RT and LR; and RT and SVR.

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

b. Ensemble Spread Tests

It is also important to evaluate calibration of the ensemble spread given by the statistical post-processing methods. A simple method quantifying this calibration is the spread-skill relationship, which measures 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 consensus forecast and the snowfall accumulation observation. An ideally calibrated ensemble should show a $y = x$ relation,

Level Two	LR	LMS	ANN	RBF	CR	SVR	KNN	RT	AI
LR	-	0	0	0	0	0	0	0	0
LMS	-	-	0	0	0.0156	0	0	0	0
ANN	-	-	-	0	0	0	0	0	0
RBF	-	-	-	-	0	0	0	0	0
CR	-	-	-	-	-	0	0	0	0
SVR	-	-	-	-	-	-	0	0.0394	0
kNN	-	-	-	-	-	-	-	0	0
RT	-	-	-	-	-	-	-	-	0
AI	-	-	-	-	-	-	-	-	-

Table 4. Paired two-sample student t-test results for the level-one dataset. Values less than 0.05 are statistically significant at the 95% level. The 0 values indicate that the values are less than 0.0001.

Level Two	LR	LMS	ANN	RBF	CR	SVR	KNN	RT	AI
LR	-	0.0796	0	0	0	0.8409	0	0.2678	0
LMS	-	-	0	0	0.0156	0	0	0.0040	0
ANN	-	-	-	0	0	0	0.0062	0	0
RBF	-	-	-	-	0.1025	0	0	0	0
CR	-	-	-	-	-	0	0	0.3121	0
SVR	-	-	-	-	-	-	0	0	0
kNN	-	-	-	-	-	-	-	0	0
RT	-	-	-	-	-	-	-	-	0.0045
AI	-	-	-	-	-	-	-	-	-

Table 5. As for Table 4, except for the level-two dataset.

or a slope of unity, between the ensemble error and the ensemble spread. In practice, any ensemble whose spread x is related to the error y by a linear equation of the form $y = mx+b$, could be recalibrated by that equation to yield more accurate estimates of the ensemble error. Therefore, Tables 6 and 7 show the correlation coefficient, slope, and intercept for this linear fit for the level-one and level-two datasets respectively. Values of the correlation coefficient closer to 1 indicate a better fit.

The method with the highest correlation coefficient values for both levels is the k-nearest neighbor method. Several other methods have slopes closer to unity but with correlation coefficients much less than one, indicating a weaker relationship. These spread-skill plots and correlation analyses show that the kNN method produces the ensemble spread best able to predict the ensemble prediction uncertainty via a linear recalibration.

Another technique for evaluating both the forecast accuracy and uncertainty is the use of 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 distributions against each other. A QQ plot is computed by independently sorting the observations from lowest to highest and the forecasts from lowest to highest. The sorted observations are then paired with the 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 Fig. 3 and 4. The different colors represent the different ensemble members.

Method	R	Slope	Intercept
LR	0.25	1.20	0.02
LMS	0.06	0.97	0.03
ANN	0.33	1.12	0.02
RBF	0.10	1.28	0.03
CR	0.17	0.52	0.03
SVR	0.32	3.28	0.01
kNN	0.80	1.66	0.00
RT	0.37	0.89	0.02
AI	0.49	4.58	0.01

Table 6. 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
LR	0.44	0.78	0.02
LMS	0.10	0.24	0.03
ANN	0.55	0.61	0.01
RBF	-0.07	1.28	0.03
CR	0.26	0.51	0.04
SVR	0.12	0.29	0.03
kNN	0.90	2.30	0.00
RT	0.39	0.84	0.02
AI	0.47	0.98	0.02

Table 7. 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.

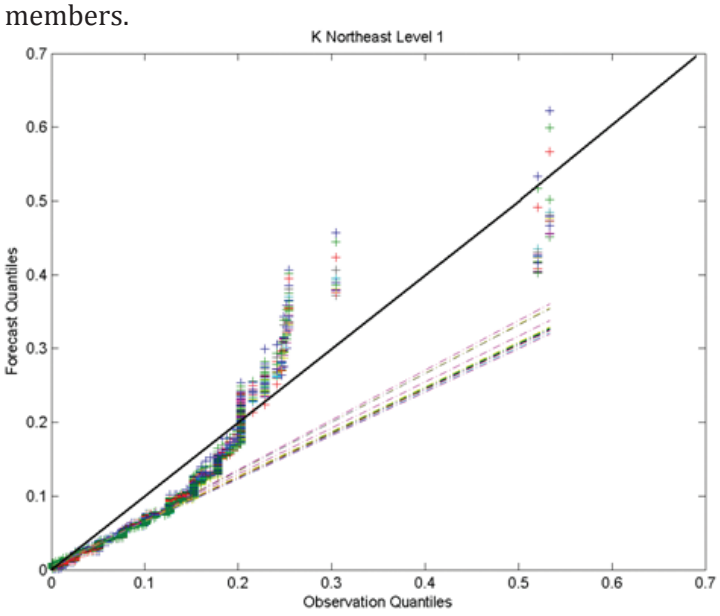

Fig. 3. QQ plot for kNN method on the level-one dataset. The dark line is $y = x$, the colored plus signs represent the different ensemble member quantiles, and the colored dashed lines connect the first and third quartiles of each ensemble member. The forecast and observation quantiles are plotted in meters.

The QQ plot for kNN method on the level-one dataset, Fig. 3, has a convex shape compared to the $y = x$ line. This indicates that the kNN method produces ensemble members with a probability density function that is positively skewed compared to the probability density function of the observations (Marzban et al. 2010), indicating an over-forecast of snowfall when the amount is substantial. For the level-two QQ plot, the plotted quantiles below the $y = x$ line indicate that the kNN method has a negative forecasting bias for snowfall observations greater than 0.1 m. Although the QQ plots for kNN are not ideal, they indicate better performance than those of the other methods, which are not shown for brevity. In addition, these QQ plots can be used to improve operational forecasting. For example, at a location site above 760 m (level two), if the kNN method predicts 0.2 m for an impending snowstorm, the QQ plot shows that the forecasts tend to be biased low, and therefore, the

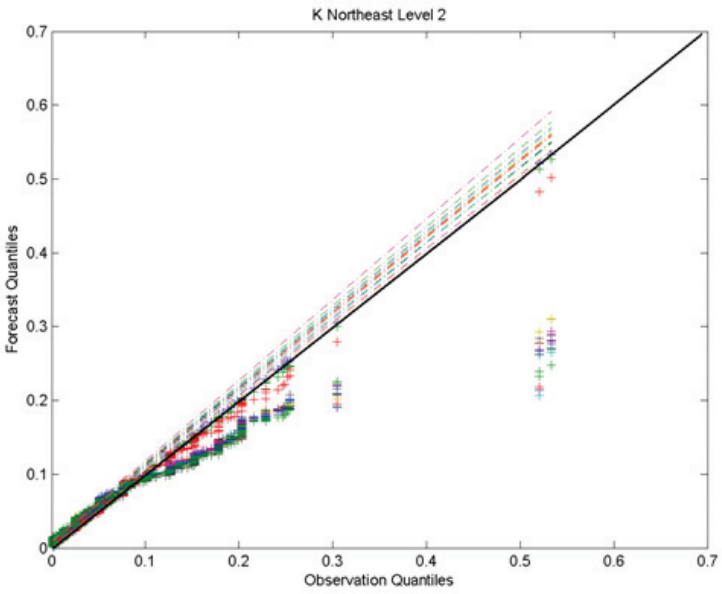

Fig. 4. QQ plot for kNN method on the level-two dataset. The dark line is $y = x$, the colored plus signs represent the different ensemble member quantiles, and the colored dashed lines connect the first and third quartiles of each ensemble member. The forecast and observation quantiles are plotted in meters.

forecaster should predict snowfall to be greater than 0.2m.

The QQ plots can also be used to assess uncertainty. The dashed lines connect the first and third quartiles of each ensemble member. An ideal ensemble would have these lines for each ensemble member lie parallel to and centered on the $y = x$ line. The QQ plots for the kNN on level-one show that the slope of these lines are less than $y = x$, but are clustered close together. For level-two, the kNN method produces lines clustered just above the $y = x$ line. Both indicate that there are differences among the ensemble members, yet they show similar relationships between the ensemble member forecasts and the observations. None of the other methods, which are not shown for brevity, produce these quartile lines as close to the $y = x$ line and clustered together.

5. Conclusions

Eight different post-processing statistical guidance methods and a consensus of these methods were tested for producing 24 hr snowfall accumulation forecasts from direct model output of the GEFS. These artificial intelligence 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 then forms a single consensus forecast that becomes a deterministic snowfall accumulation forecast. The MAE of the consensus forecast was used to assess forecast accuracy. Spread-skill relationships and QQ plots were used to examine the calibration of ensemble spread. Two methods, the kNN and ANN, showed potential improvements over the AI consensus in terms of accuracy, with the kNN method producing the best ensemble spread calibration (i.e. uncertainty estimate) of all the methods tested.

The results demonstrate that the kNN statistical post-processing method of predicting 2-hr snowfall accumulation provides more accurate deterministic forecasts than any of the other methods or the consensus of all methods. The MAE of the kNN method was significantly smaller than that for the second most accurate method, the ANN. None of the other methods, including the RBF network and SVR, performed as well as did the kNN or the ANN. 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. It is somewhat surprising that the consensus method (AI) did not perform better. That is likely due to the poor performance of RBF and CR that is averaged into it.

To evaluate calibration of the ensemble spread as a measure of forecast uncertainty, two methods were used: spread-skill relationships and quantile-quantile (QQ) plots. The kNN method produces 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 (Kolczynski et al. 2009). In addition, the QQ plots confirm that the kNN method produces the most appropriately calibrated ensemble and illustrates ways to calibrate the forecast more accurately.

In summary, the kNN statistical post-processing method outperforms all of the other methods in terms of both deterministic forecast accuracy and producing ensemble spread that is linearly related to forecast uncertainty. Thus, the kNN method appears to be the best statistical post-processing method for forecasting 24-h snowfall accumulation for this dataset. Note that this method essentially identifies the closest analogue events and use that information to correct the prediction for similar events by similar amounts. These findings are consistent with those of Delle Monache et al. (2011) for direct variables.

This study examined a 24-h snowfall accumulation with a forecast period of 12-36 h from model initialization. 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 the support vector regression technique, but the extensive computer time required to test each limited the number of different variations tested. It may be possible to combine SVR with

a technique that preprocesses the data to allow for faster computations and potentially more accurate predictions. In addition, a performance- or regime-weighted average of these statistical post-processing methods may improve the forecast accuracy and spread of the ensemble.

For an operational meteorologist, these results are significant because not only does the kNN method produce the most accurate results with the best ensemble calibration, but the method is also fast and efficient to implement for real-time applications. Free software tools such as Weka (Witten and Frank 2005) and RapidMiner (Rapid-i 2010) allow for local implementation with little or no programming. Alternatively this method could be implemented at the national level as part of the MOS suite.

Future work will involve nonlinearly calibrating the kNN method to produce better model error estimates. The QQ plot for the kNN method on the level-one dataset indicates that the probability density function of the forecasts is negatively skewed compared to the probability density function of the observations. A nonlinear calibration 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 negative forecasting bias for instances greater than 0.1 m. A non-linear equation could be developed to correct this deficiency as well.

Authors

Tyler C. McCandless is a Ph.D. 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 M.S. thesis was titled, *Statistical Guidance Methods for Predicting Snowfall Accumulation for the Northeast United States*. He earned a B.S. 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 and a professional long distance runner.

Sue Ellen Haupt is currently at the National Center for Atmospheric Research and Professor of Meteorology at The Pennsylvania State University. She earned a B.S. in meteorology and marine science certificate from The Pennsylvania State University in 1978, M.S. in engineering management from Western New England College in 1982, M.S. in mechanical engineering from Worcester Polytechnic Institute in 1984, and Ph.D. in atmospheric science from the University of Michigan in 1988. Her prior affiliations include University of Colorado/Boulder, U.S. Air Force Academy, Utah State University, University of Nevada/Reno, New England Electric System, and GCA Corporation. She has authored two books and over 250 book chapters, journal articles, conference papers, technical reports, and workshop proceedings. Her specialty is in applying novel numerical techniques to problems in fluid dynamics.

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 of the National Weather Association and the American Meteorological Society.

Acknowledgments

This research was supported by the Educational and Foundational Program of the Applied Research Laboratory of The Pennsylvania State University. Thanks are due to Dr. Harry R. Glahn and the rest of the National Weather Service Meteorological Development Laboratory for access to data as well as helpful guidance and suggestions. We also thank Rich Grumm, Dr. Hampton Shirer, and Paul Knight for knowledgeable discussions and advice. Finally, we thank two anonymous reviewers for their perceptive comments, which prompted us to make changes that have improved the manuscript.

References

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

Cosgrove, R. L., and B. Sfanos, 2004: Producing MOS snowfall amount forecasts from cooperative observer report. Preprints, *20th Conf. on Weather Analysis and Forecasting*, Seattle, WA, Amer. Meteor. Soc.

Delle Monache, L., T. Nipen, Y. Liu, G. Roux, R. Stull, 2011: Kalman filter and analog schemes to post-process numerical weather predictions. *Mon. Wea. Rev.*, in press.

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.

_____, 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.

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.

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

Kolczynski, W. C., D. R. Stauffer, S. E. Haupt, and A. Deng, 2009: Ensemble variance calibration for representing meteorological uncertainty for atmospheric transport and dispersion modeling, *J. Appl. Meteor. Climatol*, 48, 2001-2021.

National Centers for Environmental Prediction, 2010: *Global Ensemble Forecast System*. [Available online at http://www.emc.ncep.noaa.gov/GEFS/isched.php]

National Climatic Data Center, 2000: *Surface Land Daily Cooperative Summary of the Day TD-3200*, U.S. Department of Commerce, NOAA, 23 pp. {Available online at http://www.ncdc.noaa.gov/oa/documentlibrary/surface-doc.html#3200]

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. [Available online at http://nsidc.org/data/g02158.html]

National Weather Service, 2000: *Cooperative Observer Program*. U.S. Department of Commerce, NOAA, National Weather Service. [Available online at http://www.weather.gov/om/coop/Publications/coop.PDF]

Nizol, T. A., W. R. Snyder, and J. S. Waldstreicher, 1995: Winter weather forecasting throughout the eastern United States. Part IV: Lake effect snow. *Wea. Forecasting*, 10, 61–77.

Marzban, C., 2009: Basic statistics and basic AI: neural networks. In *Artificial Intelligence Methods in the Environmental Sciences*, S. E. Haupt, A. Pasini, and C. Marzban, Eds., Springer Publishing Company, 15-47.

_____, 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.

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.

Rapid-i, 2010: RapidMiner. [Available online at http://rapid-i.com/component/option,com_frontpage/Itemid,1/lang,en/]

Reed, R. D. and R. J. Marks II, 1999: Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks, The MIT Press, Cambridge, MA, 346 pp.

Richman, M. B., T. B. Trafalis, and I. Adrianto, 2009: Missing data imputation through machine learning algorithms. In *Artificial Intelligence Methods in the Environmental Sciences*, S. E. Haupt, A. Pasini, and C. Marzban, Eds., Springer Publishing Company, 133-169.

Rosenblatt, F., 1958: The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6), 386-408.

Rousseeuw, P. J., and L. M. Annick, 1987: *Robust Regression and Outlier Detection,* John Wiley & Sons, Inc., 329 pp.

Smola, A. J., and B. Scholkopf, 2004: A tutorial on support vector regression. *Statistics and Computing,* 14(3), 199-222.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: the generation of perturbations. *Bull. Amer. Meteor. Soc.*, 74, 2317-2330.

Trafalis, T. B., B. Santosa, and M. B. Richman, 2003: Prediction of rainfall from WSR-88D radar using kernel-based methods. *International Journal of Smart Engineering System Design*, 5, 429-438.

Vapnik, V., S. E. Golowich, and A. Smola, 1996: Support vector method for function approximation, regression estimation, and signal processing, Conference on Advances in Neural Information Processing Systems 9. [Available online at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.41.3139]

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 ed. Morgan Kaufmann, San Francisco, CA, 664 pp..

Woodcock, J. S., and A. F. Engel, 2005: Operational consensus forecasting. Wea. Forecasting, 20, 101–110.

Appendix

The details of why we chose each method, how it was optimized within RapidMiner and its application details are described in this appendix.

A. Linear Regression (LR)

Linear Regression (LR) is perhaps the simplest statistical post-processing method. It determines the relationship between predictand and predictors by fitting a hyperplane [i.e., multi-input linear equation (Glahn and Lowry 1972)] that minimizes the root mean squared error (RMSE). We test LR to provide a baseline for comparison of the remaining techniques, many of which support non-linear relationships.

B. Least Median Squares (LMS)

In a slight variation of linear regression, the Least Median Squares (LMS) method (Rousseeuw and Annick 1987) is assessed. It is expected to be more robust in the face of outliers than LR because it iteratively minimizes the median of the squares of residuals from the regression line instead of minimizing the mean of the squared residuals.

C. Artificial Neural Network (ANN)

The first nonlinear statistical guidance method used here is an Artificial Neural Network (ANN), which is depicted in Fig. A1 (Rosenblatt 1958). The goal is to improve upon LR and LMS by capturing nonlinear relationships between the predictand and the predictors. This simplified diagram shows four predictors fed into one hidden layer consisting of five nodes. Each node includes a LR whose output layer feeds an activation function that converts its output into the range from 0 to 1, much as in logistic regression (Reed and Marks 1990, Rosenblatt 1958). These five nodes are connected to the output layer, which combines these intermediate results using LR to produce the final prediction. This approach allows different nodes to focus on different aspects of the forecast problem with

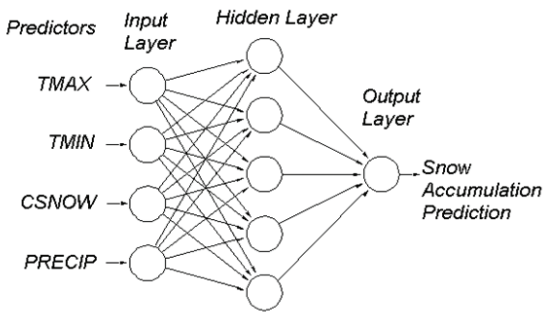

Fig. A1. Schematic of an Artificial Neural Network.

subsequent layers working to combine those elements into a final forecast. The use of an activation function after all but the final regression is critical to ANN's ability to fit nonlinear relationships. While Fig. A1 depicts an ANN with only four predictors and five nodes in the hidden layer, our ANN configurations include all of the predictors listed in Table 1 and incorporate many more hidden layer nodes.

The ANN used in this study is a feed-forward Neural Network trained by a back-propagation algorithm, also known as a multi-layer perceptron (Rosenblatt 1958). This ANN configuration includes 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 RMSE on the control ensemble member using a three-fold cross-validation with 50 training cycles. Although ten-fold cross validation with 500 training cycles is generally used (Witten and Frank 2005), three-fold cross validation with 50 training cycles is used here for finding the optimal configuration because it still generalizes well and is more computationally efficient. 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 (to drive the LR coefficients for unnecessary predictors towards zero) also increased the RMSE; therefore, neither were used in the final configuration. The activation function is the standard sigmoid function,

$$Y(x) = \frac{1}{1 + e^{-x}} \tag{A1}$$

where Y is the snowfall prediction for instance x.

D. Radial Basis Function Network (RBF)

Another method similar to the ANN is a Radial Basis Function (RBF) network (Witten and Frank 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 as the activation functions.

$$Y(x - c_i) = \frac{1}{e^{\beta (x - c_i)^2}} \quad \text{for } \beta > 0 \tag{A2}$$

in which Y is the function's output for input vector (i.e. predictor list) x and node i, c_1 is the center vector for node i (i.e. the cluster centroid for those training cases contributing to the node), and β is a weight. The radius, $x - c_1$, is the distance from the center of the hypersphere (i.e. cluster of cases contributing to that node) to the instance x. The Gaussian radial basis function then predicts the output $P(x)$, or the snowfall accumulation as

$$P(x) = \sum_{i=1}^{N} a_i Y(x - c_i) \tag{A3}$$

Here N is the number of nodes in the hidden layer and the three weights a_1, c_1, and β, are tuned to optimize the fit between the predictions and the training data. In this RBF algorithm, the k-means clustering algorithm computes the basis functions, which are normalized to sum to one before being fed to an LR model. (Witten and Frank 2005). The RBF network with three-fold cross validation on the control ensemble member for level-one was tested with various cluster sizes from 2 to 300 instances in order to determine the optimal configuration. Figure A2 plots the RMSE of the RBF network for the different cluster sizes. The RMSE decreases noticeably until approximately 120 instances per clusters and then becomes approximately

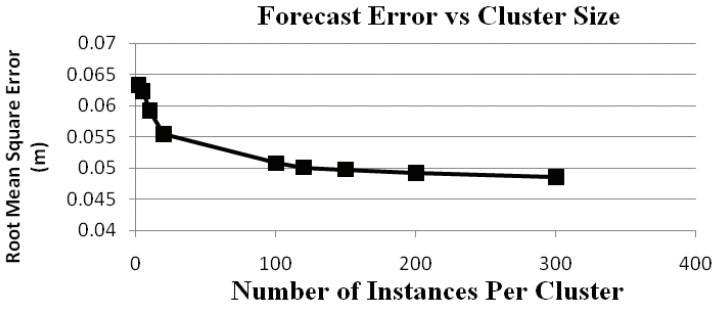

Fig. A2. Sensitivity study to determine the optimal cluster size for the RBF network.

The minimum standard deviation for the clusters was set to 0.1 and the training 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 a cluster size of 120 instances. 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 instances per cluster were used in the RBF network because that number produced the lowest RMSE with a three-fold cross validation on the level-two dataset for the control ensemble member. Cluster sizes greater than 75 instances result in an exponential increase of the RMSE, which was likely the result of overfitting.

E. Conjunctive Rule (CR)

A Conjunctive Rule (CR) is a machine learning algorithm that deduces a set of predictive rules involving

the predictors (Witten and Frank 2005). Conjunctive rules are learned by determining conditions shared by the examples. 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 such as snowfall accumulation prediction, 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. As a simple example, a conjunctive rule set for predicting whether freezing rain is possible is as follows. If the surface temperature is less than 0°C, the 850-hPA temperature is greater than 0°C, and the minimum relative humidity is greater than 99%, then freezing rain would be predicted.

The advantage of this approach is that the rules are readily understood by the human forecaster. A limitation of conjunctive rules occurs when particular outcomes do not have a single set of necessary and sufficient conditions. Because this situation arises with snowfall accumulation in the Northeast, conjunctive rules would not be expected to perform as well in this study as do some of the other non-linear methods.

The results for three-fold cross validation on the control ensemble member with the level-one dataset produces no significant difference between using 2, 3, 5, 10, and 25 rules, with all RMSE values between 0.0566 and 0.0571 meters. We opted to use three rules in our final configuration.

F. Support Vector Regression (SVR)

Another non-linear method tested here is support vector regression (SVR) (Smola and Scholkopf 2004; Trefalis et al. 2003; Richman et al. 2009). SVR is essentially a linear regression applied in a higher dimensional space that incorporates nonlinear relationships. The key to SVR is to transform the input (i.e. predictors) into a new space using a nonlinear mapping. One then must define the support vectors by fitting a maximum margin hyperplane that defines the support vectors (Witten and Frank 2005; Marzban 2009). These support vectors are data instances that are closest to this hyperplane. Figure A3 illustrates the concept of a support vector. In this example, the instances are separated into two classes, labeled Class 1 and Class 2. The maximum margin hyperplane separates the classes while maximizing the separation between the closest instances in the classes. These closest instances in this mapped space are the support vectors, which are then used to reduce the dimensionality of the problem. These support vectors constitute a reduced set of basis functions, that when convolved with an appropriate kernel function, allow compression of the problem into a smaller number of parameters. Thus, overfitting is rarely an issue with support vectors because the support vectors describe the maximum margin hyperplane and the other instances do not affect the hyperplane. In order to form a prediction, a dot product between the test instance and every support vector (projected onto the kernel function) is calculated to fit a linear regression in the higher dimensional nonlinearly mapped space. This mapping process can be computationally expensive when there are many predictors, as is the situation with the snowfall prediction dataset of over 100 predictors.

We apply the support vector regression method using a kernel operator that directly computes the dot product between data point vectors, a convergence epsilon of 0.0010, and a maximum of 100,000 iterations. This configuration is chosen because it provides the lowest RMSE on the three-fold cross validation compared to other convergence epsilons, maximum number of iterations, and types of kernels.

G. K-Nearest Neighbor (kNN)

A k-nearest neighbor (kNN) algorithm works by searching for *k* historical analogs to the current predictor values and then using the consensus predictand value from those analogs as the prediction. The kNN method first normalizes the predictors and then finds the *k* training instances that are closest in Euclidean distance

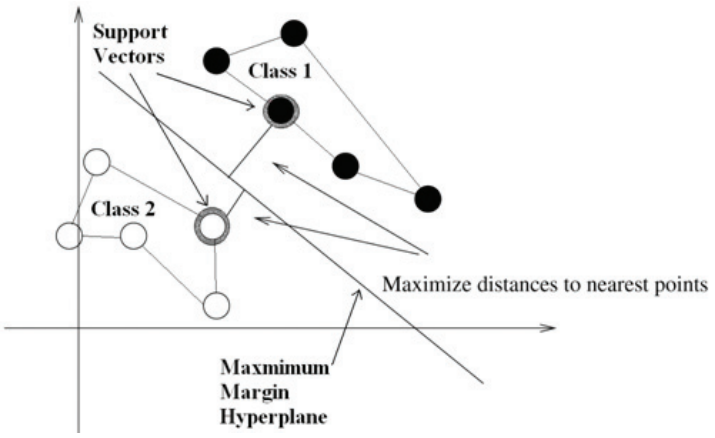

Fig. A3. Schematic of a maximum margin hyperplane and support vectors separating two classes in support vector classification. The axes depicted here are in the transformed space.

to the given test instance (Witten and Frank 2005). A consensus forecast is formed from the distance-weighted average of these k-nearest neighbor's predictand values. The number *k* of analogs is selected using leave-one-out cross-validation, given an upper limit for *k* of 100 with the optimal value being 6 for the level-one dataset and 5 for the level-two dataset.

Figure A4 illustrates how the k-nearest neighbor algorithm works. All dots correspond to instances mapped onto a high-dimensional space. The center orange dot corresponds to the test instance that the k-nearest neighbor method predicts. For *k = 10*, 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.

H. Regression Tree (RT)

The final non-linear method tested is the Regression Tree (RT, Witten and Frank 2005). An RT is formed by building a decision tree in which the leaf nodes contain the numeric value that is the average outcome, i.e. snowfall accumulation, of those instances falling in that leaf. The term regression signifies that the tree produces a numerical prediction rather than a categorical forecast as in a traditional decision tree. This algorithm uses information gain/variance reduction to select branches and prunes the tree using reduced-error pruning. The reduced-error pruning is performed with back-fitting.

A simple example of a regression tree is shown in Fig. A5. The regression tree in this figure has only three predictors; temperature, relative humidity, and accumulated precipitation. Each node, which are represented as boxes in the figure, determines the path to traverse down the tree. When the instance reaches the bottom of the tree a regression equation is used to predict the snowfall accumulation.

We determined a configuration that produced the lowest error on three-fold cross validation: the minimum number of instances per leaf was set at three, the value used to minimize the numeric class variance to determine

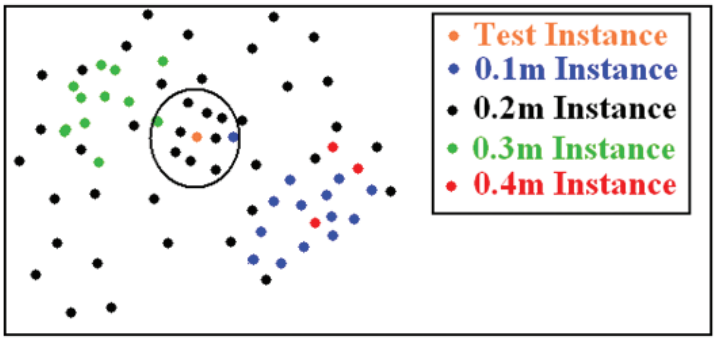

Fig. A4. Illustration of the k-nearest neighbor algorithm. All dots correspond to instances mapped on a high dimensional space. For k = 10, the distance weighted average of the ten instances inside the circle are used to predict the test instance in orange.

the appropriate split was determined to be 0.001, and the number of folds for reduced error pruning was found to be three.

I. Consensus (AI)

Finally, an average of the eight Artificial Intelligence (AI) methods is used to produce a consensus forecast. These average forecasts were produced by averaging each AI method forecast for each instance and each ensemble member.

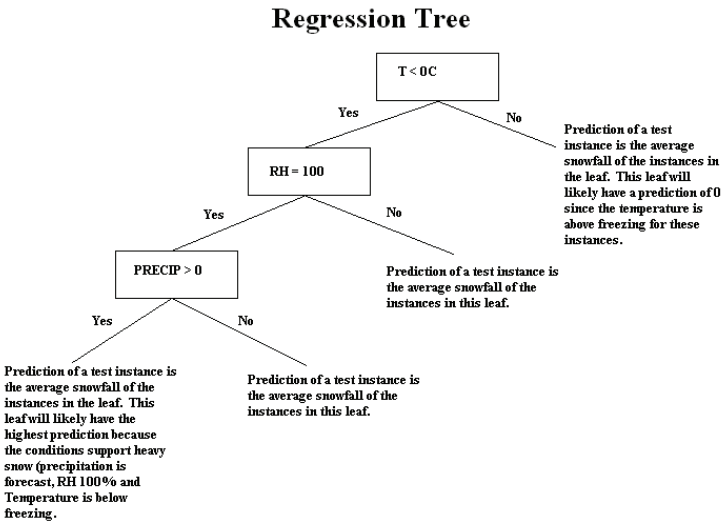

Fig. A5. Simple diagram of the process of a regression tree using three predictors; temperature, relative humidity, and predicted accumulated liquid equivalent rainfall. Regression equations are used at each leaf to predict snowfall accumulation.