Final Report: Development of a Large Member Ensemble Forecast System for Heavy Rainfall using Evolutionary Programming (EP)

1. Introduction

The advantage of the Evolutionary Programming (EP) method (see Roebber 2015a), in principle, is in its ability to generate large member ensembles (many thousands of members) with relative computational efficiency. Further, the method can use any combination of NWP model output and observations – it is up to the analyst to determine the most appropriate inputs and to develop the training and testing dataset, and then the method will produce the optimized ensemble equations, based on the specific cost function that is applied (typically root mean square error (RMSE) and Brier Skill Score (BSS)]. This allows the full probability distribution to be explored, compared to small member ensembles and easily can be applied to problems where NWP guidance is less useful.

Roebber (2010; 2013; 2015abc) did considerable work in developing the approach, specifically in the context of maximum and minimum daily temperature forecasting, including forecasts out to 156 hours, where presumably the forecast "signal" is weak. With that in mind, in this study, we intentionally applied the method to a "tough" forecast problem (next day convective rainfall exceeding one inch), with the goals of further developing some of the methodology and testing the limits of the approach. After these explorations, given sufficiently robust results, the hope was to develop a demonstration, operational EP-based heavy rainfall ensemble forecast system, in which estimates of the probability of convective rainfall exceeding specified thresholds would be produced for the next day forecast period for a specific location.

In order to accomplish these goals, a number of developmental issues still needed to be addressed, including a number of questions related to EP construction and training, as well as the evaluation of deterministic and probabilistic performance relative to alternative approaches. Further, some effort was devoted to the question as to how to downscale regional precipitation forecasts, which provide situational awareness, to the local scale, where the impacts of heavy rainfall are felt. In the next section, we will describe the results from these investigations, and in section 3, we will discuss what these findings suggest about potentially profitable future research and application directions.

2. Findings

a. How to improve the Bayesian Model Combination (BMC) calibration?

Unfortunately, while ensembles of numerical weather prediction (NWP) models are well suited to generating probabilistic information, and the resulting forecast distributions exhibit a characteristic sharpness, there has been a long-standing problem with NWP ensemble under dispersion (see Novak et al. 2008 for a discussion of the effect on operations). In essence, the NWP ensembles display "overconfidence."

Post-processing of ensembles can help in this regard. Roebber (2015b) showed that an approach known as Bayesian Model Combination (BMC) can improve both EP and traditional NWP (e.g., GFS 21-member MOS) ensemble deterministic and probabilistic performance, with the largest improvements accruing to the smaller member GFS ensemble. BMC is similar in many respects to the more familiar Bayesian Model Averaging (BMA; e.g., Raftery et al. 2005), except that whereas in BMA one assumes that a single ensemble member is the true data generating model (DGM) and thus the selection is organized to find this best model, in BMC one assumes that no single ensemble member is the DGM and thus the selection is based on the most optimal combination of ensemble members (see Monteith et al. 2011 for a full comparison of BMC versus BMA).

A limitation to both these approaches for large ensembles is that owing to computational constraints, one must sub select members from the overall population (e.g., Hoeting et al. 1999). For example, using BMC with 4 possible raw weights and 10 members (yielding normalized weight values ranging from $\frac{1}{37}$ to $\frac{4}{13}$) requires evaluation of 4^{10} or 1,048,576 possible combinations. While evaluating O(1,000,000) combinations is tractable, evaluating O(1,000,000,000) is not, and this latter would be required for only 15 members and the same number of weights.

For member sub selection, we have used an approach in which we rank all members according to their root mean square error (RMSE) on the training data and then compare the forecast variance between every ensemble member pair. The lower ranking member based on RMSE for any pair for which this variance is below a specified threshold (set to 0.098 F^2 for temperature) is then eliminated. The top-ten ranked remaining members

are then sub-selected for BMC. In this study, we tested an alternative approach, in which we use past analogs to select forecast members (after Delle Monache et al. 2013). Unfortunately, this approach offered no improvement relative to the above selection method for the same dataset. Similar lack of improvement resulted from experiments with the Du-Zhou ranking method (Du and Zhou 2011) and consequently these alternatives were not pursued further in this study.

Following BMC, an additional bias correction is applied. Previously, for temperature, we have used the weighted correction method of Cui et al. (2012). After discussion with NOAA scientists during Summer 2015, however, we have modified this approach to use the N-day running mean bias correction method now in wide use at NCEP (Fan and van den Dool 2011). However, for the precipitation forecast problem, we have opted for a modified approach in generating probabilities, motivated by logistic regression (see below), and for which this additional bias correction is not needed.

b. What relative performance advantages are provided by ensemble size?

Most of the EP ensemble work conducted by Roebber (2015abc) is based on a population "carrying capacity" of 10,000 members (i.e., maximum ensemble size, with resultant final ensembles typical being 10-20% of this maximum size). Evidence from nature suggests that population size, given sufficient resource availability, strongly increases genetic diversity (Stevens et al. 2007), and thus we wished to explore this effect in our method.

Roebber (2015a) used a chaotic dataset (broadly similar to 500 hPa height) obtained from a modified form of the Lorenz (2005) model to examine the effects of increased carrying capacity on rate of training. Increasing carrying capacity 25-fold shortened the number of generations to convergent solutions by 30% compared to the smaller populations, with the best solutions being of approximately equivalent performance in each approach. Roebber (2015a) also performed an experiment using actual temperature forecast data with a carrying capacity of 200,000 members and found that it shortened the number of training steps by 12%. However, neither of these experiments examined impacts on ensemble diversity and on overall probabilistic performance.

For that purpose, in this study we performed an additional experiment with the minimum temperature forecast dataset of Roebber (2015a), in which carrying capacity was increased 50-fold to 500,000 members. This large dataset required special handling with programming run on the Yellowstone computer.

A measure of population diversity is the Shannon index, where larger values indicate larger diversity. We found that for the ensemble carrying capacity of 10,000 members, which produced an ensemble of 2,713 members, the Shannon index is 2.05. In contrast, a lower Shannon index of 1.52 is obtained from the carrying capacity run of 500,000 members (producing an ensemble of size 41,831).

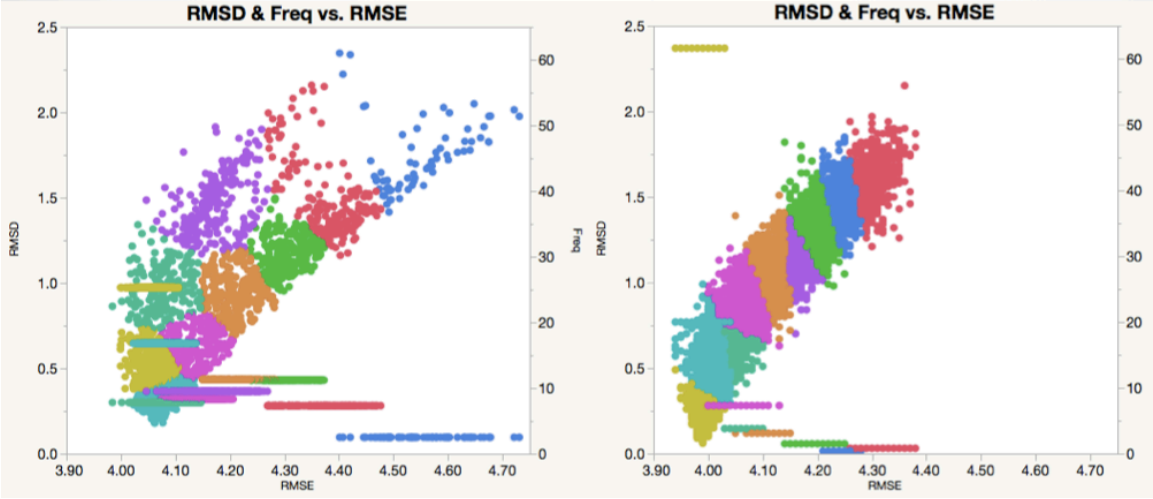

Figure 1: Self Organizing Map (SOM) cluster analysis for the 10,000 (left) and 500,000 (right) member carrying capacity EP ensembles. Shown are each ensemble member's location in RMSE-RMSD space colored by cluster membership, and also the relative proportion of that cluster in the overall population (right axis).

Nonetheless, the rationale as to why very large ensemble sizes might contribute to increased *absolute* diversity rests on a probability argument. Consider that the probability of a 5 sigma event is 0.00003. If the

sample size is 10,000 than the expected number of such events is less than one, whereas if the sample is 500,000 than one expects O(15). Consequently, the question arises as to how one might take advantage of that property and form a selective, diverse ensemble from the large population? One possible approach would be to choose members with low RMSE on the training data, but with moderate root mean square difference (RMSD) between the member and the ensemble mean. A cluster analysis of the data for these two ensembles (Figure 1) showed that there might be some opportunity to do that.

This procedure was then applied to the very large ensemble. The resultant performance showed marginal improvement over that of the standard-sized ensemble. Specifically, the RMSE decreases from 4.11 to 3.88 with a Brier Skill Score (BSS) of 9.4%. Notable, however, is the fact that application of BMC to the standard-sized ensemble reduces the RMSE to 3.94 and to an equivalent BSS of 9.4%. Thus, the gains after bias correction are relatively marginal relative to the substantial data handling and ensemble production procedures required to produce and manage such a large population.

c. *How can EP training be optimized (examination of EP architectures and cost functions)?*

The basic architecture of the EP algorithm takes the following "block" form (see Roebber 2015a):

	V_{1ij}	O_{rij}	V_{2ij}		C_{1ij}		V_{3ij}	O_{1ij}	C_{2ij}		V_{4ij}	O_{2ij}	C_{3ij}		V_{5ij}
IF	PRS_{SNW}	>	1	THEN	0.151	*	1	*	0.559	*	1	*	-0.567	*	1
IF	MSP	≤	1	THEN	0.138	*	WS_{DVN}	*	-0.393	*	STL	*	0.242	*	Sin (JD)
IF	$[PP]_{min}$	>	1	THEN	0.125	*	MSP	*	-0.267	*	PRS_{SNW}	+	-0.160	*	1
IF	Cos (JD)	≤	1	THEN	0.996	*	1	*	0.119	*	1	*	0.512	*	1
IF	MLR	≤	WS_{DVN}	THEN	0.315	*	PRS_{SNW}	*	0.606	*	1	*	-0.473	*	1
IF	$[T_{+24h}]$	≤	1	THEN	-0.114	*	[CC]	*	-0.361	*	1	+	0.899	*	F.T
IF	PRS_{SNW}	≤	1	THEN	0.055	*	1	*	0.467	*	1	*	0.531	*	1
IF	DSM	>	1	THEN	-0.867	*	DSM	*	0.408	*	$[PP]_{min}$	+	-0.160	*	1
IF	PRS_{PP}	≤	PP_{GRB}	THEN	-0.110	*	Cos (JD)	*	0.822	*	MLR	*	-0.805	*	1
IF	[CC]	≤	1	THEN	-0.972	*	F.WS	*	0.172	*	1	*	0.453	*	1

Figure 2: EP 10-line IF-THEN block architecture. Colors indicate adjustable variables, operators, and coefficients, while the black operators are fixed. See Roebber (2015a) for further details.

where there are up to 10 lines of IF-THEN conditionals. V_1 through V_5 represent any five of the variables contained in the list of potential predictors, O_r is always > or ≤, O_1 and O_2 are operators which can be either multiplication or addition, and C_1 through C_3 are coefficients.

All of the variables are normalized to the range of [0-1] (based on the minimum and maximum of each variable within the training data, and which would be set to 0 for any future value less than the minimum in the training data and 1 for any future value greater than the maximum in the training data). One additional variable added to the potential predictor list is unity. Thus, in the above example, the variable unity appears in the IF-THEN conditional lines 1, 2, 3, 4, 6, 7, 8, and 10. This allows the possibility of lines never being executed (in the above example, lines 1, 3, and 8) or always executed (lines 2, 4, 6, 7 and 10), such that a baseline multiple linear or non-linear (with up to cubic functions) equation is formed and for which IF-THEN modifiers can occur (in the above example, lines 5 and 9). Once the equation is computed, the output is then scaled back to the dimensional value based on the minimum and maximum of the output in the training dataset. This architecture turns out to be quite flexible. Experimentation with additional, more complicated forms has been conducted (e.g., Roebber 2015a explored the use of lines that could reference the results of prior lines), but these were not shown to provide additional performance and were not used in the present study.

However, two modifications to the above procedure were investigated. In the above form, one could imagine producing forecasts for specific precipitation values and then using the ensemble to produce probabilistic

predictions of rainfall exceeding a given threshold. For the precipitation domain, this was found to produce training results that were not as effective as desired. Another possibility is to code the precipitation in the training data as 0 when it is less than a desired threshold and 1 when it exceeds the desired threshold (thus, the same form as above but where the range of the output is 0-1 representing probability). This approach is similar to multiple linear regression using binary output. Again, this form was found to produce training results that were not as effective as desired.

Another possibility is to construct the EP architecture consistent with a decision tree format. This was accomplished using a structure with 28 rather than 10 lines, and for which the set of lines defines all the possible IF-THEN combinations for a set of five variables. Thus, only one line is true and each of the 28 lines has a result which is computed as: $a*V_1$ (+ or *) $b*V_2$ (+ or *) $c*V_3$ (+ or *) $d*V_4$ (+ or *) $e*V_5$, where V_1 to V_5 are variables, and *a* to *e* are coefficients. Since each line is composed of an IF-THEN conditional with five variables and five operators, and the result uses five variables, five coefficients, and four operators, each decision tree algorithm uses 393 elements. An experiment was run on the temperature data to see whether this approach could provide any advantage relative to the standard (10-line) approach, but it was found to be less effective for the training data. We did not pursue this line of research for the precipitation data, but based upon some of the precipitation downscaling results obtained at the end of the project (see section f), this may be worth pursuing further.

The structure that proved most effective borrows from the log-odds basis of multiple logistic regression. Thus, the training precipitation data were coded as 0 or 1 depending on whether a desired threshold was exceeded, and the standard IF-THEN structure was used to produce log-odds equations. Thus, the output from the EP algorithm is expressed as:

$$\mathrm{P\,(prob\ >\ N\ inches)} = \frac{1}{1+\exp(-R')}$$

where R' is the scaled output from the EP IF-THEN equation. Since the output from this approach is a probability, the EP training was used to find the "best" algorithm (as in Roebber 2010 for temperature) rather than to produce an ensemble of algorithm predictions (as in Roebber 2015abc), where best is defined based upon BSS performance on the training data.

An experiment with the cost function used to evaluate algorithm success in each generation was also conducted, for the EP architecture in which precipitation values were being produced. The experiment tested training success using RMSE for precipitation amount versus mean absolute error. The results were not substantially different but neither result was sufficiently successful to consider that approach further (as detailed above).

d. Which are the optimal input variables for heavy rainfall?

Some experimentation with input variables for prediction of heavy rainfall was required. Regression and EP approaches require a set of inputs to map to the output (probability of precipitation in excess of a threshold). However, the "curse of dimensionality" (Bellman 1961) requires that the list of inputs be as restrictive as possible, since the search space for multidimensional solutions increases exponentially with added predictors, with resultant increases in the amount of data needed for training. Thus, data sparsity quickly becomes a problem even for very large training datasets.

With this principle in mind, we identified a potential list of predictors based upon domain expertise. From this initial list, the following variables were found to be most useful: CAPE, time of maximum CAPE, deep layer shear (surface to 300 hPa), precipitable water, precipitable water anomaly (relative to climatology), upward 850 hPa vertical velocity, 700 hPa wind speed, observed prior day 24-hour precipitation at 7 upstream sites (see below), observed prior day maximum temperature at 7 upstream sites, forecast precipitation category, and cluster identity (obtained from a cluster analysis of the input data, see below). Synoptic-mesoscale arguments as to why these predictors might be useful for this purpose are easily constructed, thus assuring that the analysis is passing a basic sanity check.

Observed (gridded) precipitation data were obtained from the daily 0.25° x 0.25° unified precipitation dataset for 1985-2006 for the upper midwest region from 40.125 – 44.875°N and 91.875 – 87.125°W. Forecast (12 - 36 hour) values for the above input variables for the period 1200 UTC – 1200 UTC were obtained for the warm seasons (May – October) of 1985 – 2006 from the 0000 UTC Reforecast V2 (equivalent to the 2012 GEFS at ~ 50 km resolution), for all grid points in the region 40 – 45°N and 92 - 88°W. The upstream site data were

obtained from Chicago (ORD), Des Moines (DSM), Green Bay (GRB), Madison (MSN), Milwaukee (MKE), Minneapolis (MSP), and Saint Louis (STL) surface observations.

Two additional predictors suggested by the literature were evaluated during the course of the study but were not found to be useful. These were a soil moisture estimate (based on the antecedent precipitation) and near surface layer average relative humidity. We speculate that while these predictors make some sense for the general precipitation problem, since all else being equal higher soil moisture and higher near surface relative humidity can increase moisture accession and reduce rainfall evaporation, respectively, in cases of heavy rainfall in the upper midwestern U.S., the moisture content of the atmospheric column is already high (as measured by the precipitable water and the precipitable water anomaly). Consequently, these predictors did not add useful information that was not already contained in the other variables.

The Self Organizing Map (SOM) technique was used with the predictor training data to place each case into three objective clusters:

$$\begin{aligned} \text{CL} &= 1 \text{ if Min(a1, a2, a3)} = \text{a1} \\ &= 2 \text{ if Min(a1, a2, a3)} = \text{a2} \\ &= 3 \text{ if Min(a1, a2, a3)} = \text{a3} \end{aligned}$$

where

$$\text{a1}= \left(\frac{CAPEx-2474}{1430}\right)^2 + \left(\frac{Shear-18.2}{9.4}\right)^2 + \left(\frac{fPx-1.3}{1.1}\right)^2 + \left(\frac{PW-33.6}{9.7}\right)^2 + (PWa-3.0)^2 + \left(\frac{VV+0.024}{0.071}\right)^2 + \left(\frac{V7-11.5}{5.3}\right)^2 + \left(\frac{Px-14.5}{19.6}\right)^2$$

$$\text{a2}= \left(\frac{CAPEx-1886}{1430}\right)^2 + \left(\frac{Shear-26.0}{9.4}\right)^2 + \left(\frac{fPx-3.4}{1.1}\right)^2 + \left(\frac{PW-34.2}{9.7}\right)^2 + (PWa-3.4)^2 + \left(\frac{VV+0.095}{0.071}\right)^2 + \left(\frac{V7-16.0}{5.3}\right)^2 + \left(\frac{Px-21.6}{19.6}\right)^2$$

$$\text{a3} = \left(\frac{CAPEx-404}{1430}\right)^2 + \left(\frac{Shear-27.5}{9.4}\right)^2 + \left(\frac{fPx-1.1}{1.1}\right)^2 + \left(\frac{PW-18.9}{9.7}\right)^2 + (PWa-1.6)^2 + \left(\frac{VV-0.031}{0.071}\right)^2 + \left(\frac{V7-14.1}{5.3}\right)^2 + \left(\frac{Px-14.6}{19.6}\right)^2$$

Here, the subscript "x" refers to the maximum regional value of a quantity, and the other variables are regional average values (fP is the forecast precipitation category; PW is the precipitable water [mm]; VV is the upward 850 hPa vertical velocity [Pa/s, negative is upward]; V7 is the 700 hPa wind speed [m/s], and Px is the prior day observed rainfall [mm]). Note that PWa represents the precipitable water anomaly such that PWa=1 is the 25th percentile value, =2 is the 50th percentile, =3 is the 75th percentile, and =4 is the maximum. As can be seen from the above equations, clusters 1 and 2 are moist, convectively charged environments relative to cluster 3.

e. What are the performance characteristics relative to alternative methods?

There are multiple methods for producing probabilistic information. Here, we investigated the performance characteristics of several approaches using the collected data for training and testing. These methods, in addition to the "best algorithm" EP, included multiple logistic regression (MLR), single hidden layer artificial neural networks (ANN), and the raw output from the (coarse-grain) Reforecast model V2. The results for these methods (save the Reforecast, whose CSI was ~ 0.0) are plotted for the 1.5 inch threshold forecasts on the performance diagram (Roebber 2009) in Figure 3.

The MLR and EP approaches provide comparable skill at the 1.5 inch threshold, but interestingly, their performance characteristics are slightly different, with the MLR showing somewhat higher probability of detection (POD) and also more false alarms compared to the EP. One can envision using these equally skillful approaches in concert to give a better sense of what might happen in a given instance. In contrast, the ANN approach provided lower skill, primarily owing to a much lower POD with only a very slight reduction in false alarms relative to the EP.

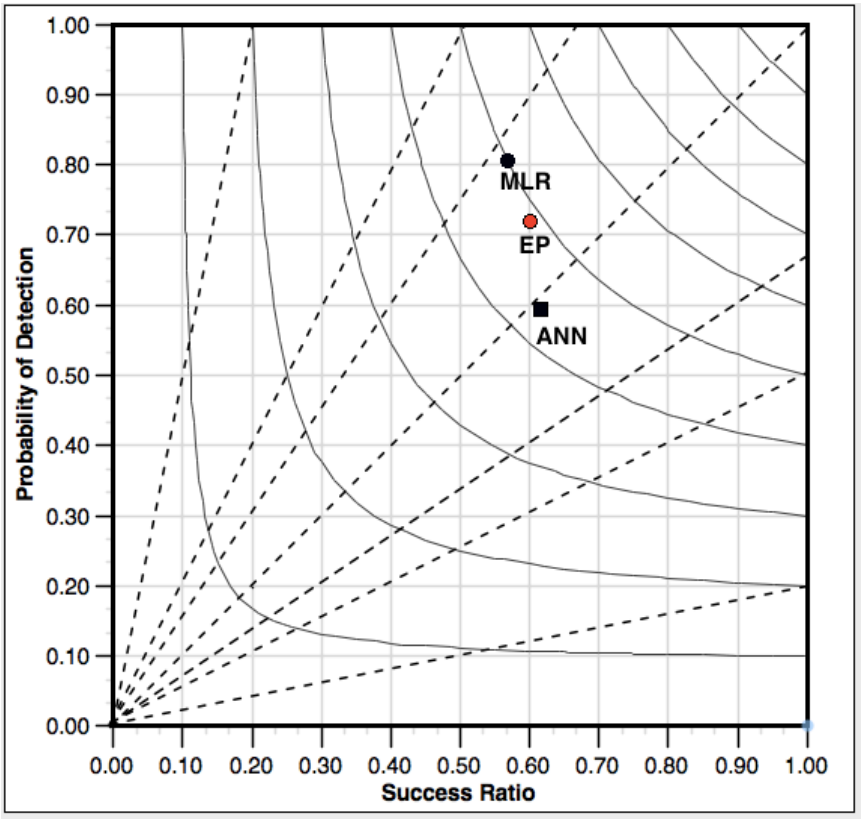

Figure 3: Performance diagram showing forecast metrics for the 1.5 inch threshold obtained from the artificial neural network (ANN), multiple logistic regression (MLR), and the best-member Evolutionary Program (EP) equations.

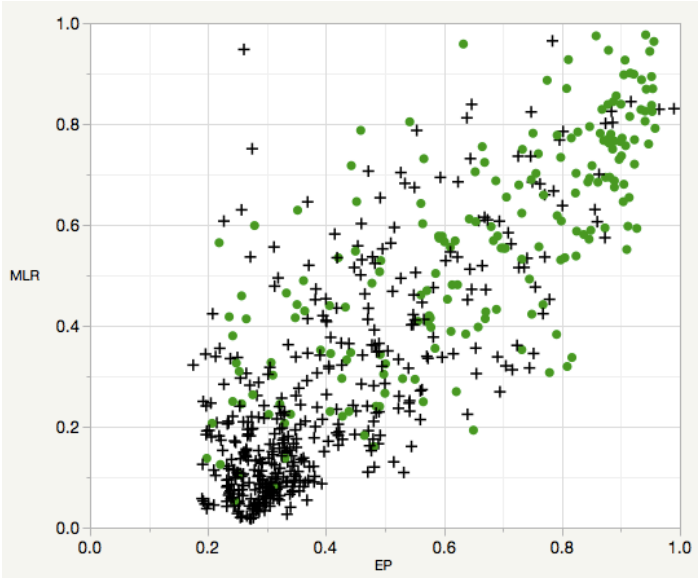

Figure 4: MLR versus EP probability of regional exceedance of 1.5 inches. Green denotes that precipitation verified in excess of 1.5 inches within the region of interest.

This differential response between the two methods can be seen in Figure 4. In general, when both

forecasts are high (low), a larger proportion of the events verify (do not verify), as one would hope. When there is some inconsistency between the two, however, this would be an indication to forecasters that the event has more uncertainty associated with it and that even more attention to the particulars of the case may be warranted.

f. How can the regional forecasts best be downscaled to the local scale?

While the regional forecasts appear relatively skillful for next day heavy rainfall forecasting, given the size of the region for which these predictions hold, these forecasts could only be used to raise situational awareness rather than provide information at the local scale where the impacts from such events are felt. Thus, there is a need to downscale this information to specific sites.

In order to test the ability to use this information for that purpose, we tested several approaches for these same data, but now applied to the maximum precipitation observed in the area covered by the Milwaukee Metropolitan Sewerage District (MMSD) network of stations. This area is depicted in Figure 5 as the grey-shaded region within Milwaukee County.

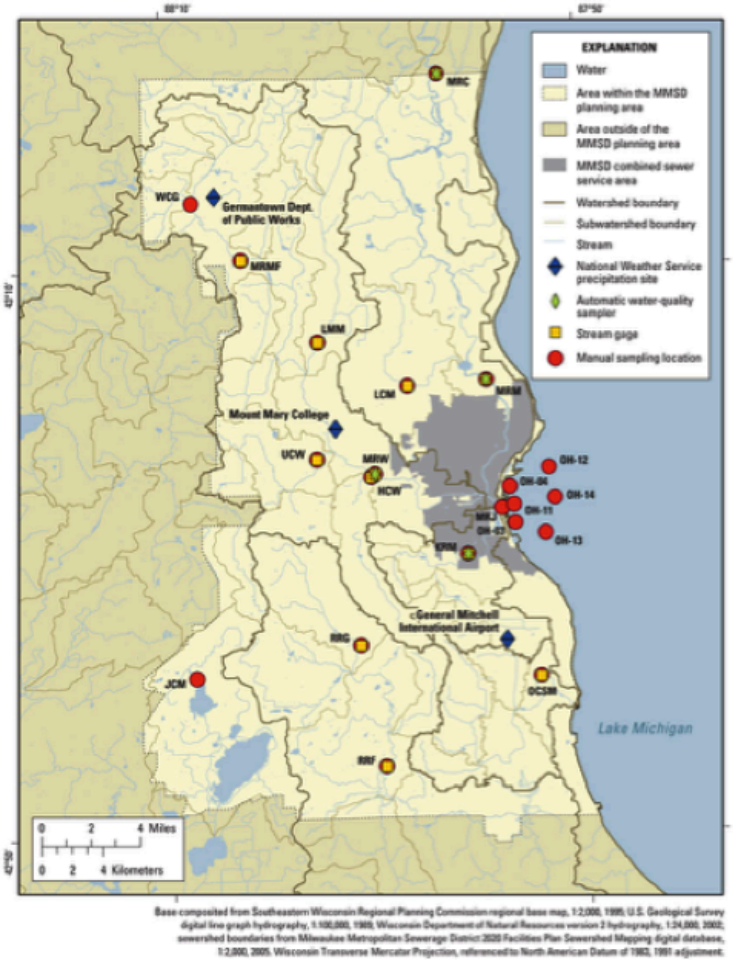

Figure 5: MMSD region with precipitation measurements (combined sewer area), denoted by the grey shading. The light yellow region denotes Milwaukee County.

Several approaches were attempted. First, we used the EP methodology directly as applied to the region but now focused specifically on the MMSD area of interest. This was attempted with the same data as previously used for the regional precipitation probability forecasts, and also with and without the EP and MLR precipitation probabilities. Similarly, we used MLR in this way, but now applied to the MMSD area of interest. Neither of these approaches provided very useful forecast information.

Additional methods tested and rejected based upon training data results included boosted trees, bootstrap forests, random forests, K nearest neighbors, and single hidden layer artificial neural networks (see Altman 1992; Principe et al. 2000; Hastie et. al., 2001; Breiman 2001). The method that provided the best result was a fairly simplistic decision tree structure that used prior day precipitation observations at MKE, cluster designations, and EP and MLR probabilities to define broad risk categories: very likely, likely, unlikely, and very unlikely, which roughly corresponded to probabilities of greater than 75%, greater than 50%, less than 20%, and less than 1%. Performance diagnostics were computed which suggest some skill for these predictions (CSI=0.47 for >1 inch and 0.37 for > 1.5 inches), but with relatively lower POD (0.54 and 0.44 for > 1 inch and > 1.5 inches, respectively).

3. Discussion and Future Directions

The scientific objectives of the project specified in the proposal were threefold. First, we sought to understand how the specific predictability constraints for heavy rainfall apparent in NWP ensembles can be improved using EP ensemble techniques. In the course of this study, we found that post-processing of coarse-grain data as provided by the Reforecast V2 was able to provide considerable additional *regional* insight into heavy rainfall occurrence. We found that local insight was also possible, but more limited than that obtained regionally, as one might expect from the basic characteristics of convection. We also found that the EP approach did not necessarily provide *better* information than that obtained from a traditional technique (multiple logistic regression), but it did provide *unique* information in some contexts, which might be leveraged advantageously by forecasters.

Second, we sought to explore how effective a large increase in population size is in improving EP ensemble under dispersion. We found that a 50-fold increase in population carrying capacity which led to a more than 15-fold increase in ensemble size *decreased relative ensemble diversity* even though absolute diversity was increased. Nonetheless, because absolute diversity was increased, it suggested the possibility that if training data could provide useful information on ensemble member selection, it might be possible to improve ensemble deterministic and probabilistic performance. We found that some of that advantage was lost owing to the inability to map precisely from the best and most diverse member performance in the training to the test data, but that the larger ensemble was able to provide some advantage. Unfortunately, all of the probabilistic advantage and a large fraction of the deterministic advantage was lost when bias correction was applied. Since the latter process can be computationally more efficient (within limits) than producing very large ensembles, this approach does not appear to be optimal.

This point connects to the third objective, which was to consider extensions to the existing procedures. Many experiments were performed in this regard. Approaches that might still benefit from further investigation include:

(1) Use of a decision tree EP structure. Here, we found that this structure was not more useful for temperature, but this approach has not been applied to other problems and the mapping of regional heavy precipitation to local scales benefited the most from decision tree processes;

(2) EP member selection procedures for use in Bayesian Model Combination bias correction. This represents an area of potentially substantial gain but experiments thus far have not suggested a better approach than current practice;

(3) New methodologies, most particularly approaches that leverage the spatial dimension of the meteorological fields rather than relying on site-specific forecasts. One method with great potential may be Deep Learning (unsupervised encoding of spatial and temporal patterns by multiple layer neural networks) and initial experiments with this approach are underway. Other related research includes adaptive optimization that might involve flow-dependent adjustments to training, especially during the “fast mode” which is currently fixed at 7 days, but also for the “slow-mode” (i.e., train ensemble members to optimize for particular flow conditions, which might be identified in a variety of ways, for example, cluster identification).

(4) Observational focus. To date, the EP methodology has been focused on using a combination of NWP and observations to leverage the NWP data. This sort of NWP “post-processing” has proven most effective at longer forecast ranges where the NWP forecast signal is weak (see Roebber 2015b). This suggests an opportunity to

leverage observations more directly by focusing on forecast problems at the nowcast range, where NWP guidance is less useful. For example, examination of the convective initiation problem (Burghardt et al. 2014) and subsequent convective modes might prove productive.

Project deliverables were to be:

(1) Two seminars at NCAR, one at project initiation and a second at project completion. Five seminars were delivered over the course of the project – two at project initiation at NCAR and at NOAA – Boulder, two mid-project during a DTC-sponsored visit by the PI to NCEP (one at EMC and one at MDL), and one at project close at NCAR. The discussions prior to and during these site visits were extremely valuable to the PI and have helped to shape the research during the conduct of this project and will help to shape the next steps.

(2) A project final report. This document. The PI believes that the research is not yet at the stage where another refereed publication can be produced but rather the project has laid the foundation for such an article, eventually to be produced for *Monthly Weather Review*. The expectation is that such a manuscript will be submitted within the next year.

(3) Computer code for the adaptive EP ensemble, such that, given the identified set of inputs, the ensemble forecast will be produced. Since the research focus was not continued in this direction, based on findings during the course of the work, this code was not developed in the context of heavy rainfall forecasting.

References

Altman, N.S., 1992: An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician,* **46**, 175–185.

Bellman, R.E., 1961: *Adaptive control processes: a guided tour.* Princeton University Press, 255 pp.

Breiman, L., 2001: Random Forests. *Machine Learning*, **45**, 5–32.

Burghardt, B., C. Evans, and P.J. Roebber, 2014: Assessing the predictability of convection initiation using an object-based approach. *Wea. Forecasting*, **29**, 403-418.

Cui, B., Z. Toth, Y. Zhu, and D. Hou, 2012: Bias correction for global ensemble forecast. *Wea. Forecasting*, **27**, 396–410.

Delle Monache, J., F. A. Eckel, D. L. Rife, B. Nagarajan, and K. Searight, 2013: Probabilistic weather prediction with an analog ensemble. *Mon. Wea. Rev*., **10**, 3498-3516.

Du, J. and B. Zhou, 2011: A dynamical performance-ranking method for predicting individual ensemble member performance and its application to ensemble averaging. *Mon. Wea. Rev*., **139**, 3284-3303.

Fan, Y., and H. van den Dool, 2011: Bias correction and forecast skill of NCEP GFS ensemble week-1 and week-2 precipitation, 2-m surface air temperature, and soil moisture forecasts. *Wea. Forecasting*, **26**, 355-370.

Hastie, T., R. Tibshirani, and J.H. Friedman, 2001: *The elements of statistical learning: Data mining, inference, and prediction.* New York: Springer Verlag.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky, 1999: Bayesian model averaging: A tutorial. *Stat. Sci*., **14**, 382–417.

Lorenz, E. N., 2005: Designing chaotic models. *J. Atmos. Sci.*, **62**, 1574–1587.

Monteith, K., J. Carroll, K. Seppi, and T. Martinez, 2011: Turning Bayesian model averaging into Bayesian model combination. *Proc. Int. Joint Conf. on Neural Networks (IJCNN'11)*, San Jose, CA, IEEE, 2657–2663.

Novak, D. R., D. R. Bright, and M. J. Brennan, 2008: Operational forecaster uncertainty needs and future roles. *Wea. Forecasting*, **23**, 1069–1084.

Principe, J.C., N.R. Euliano, and W.C. Lefebvre, 2000: *Neural and Adaptive Systems: Fundamentals through Simulations*. John Wiley and Sons, 656 pp.

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.

Roebber, P.J, 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting*, **24**, 601-608.

Roebber, P.J., 2010: Seeking consensus: A new approach. *Mon. Wea. Rev*., **138**, 4402-4415.

Roebber, P.J., 2013: Using evolutionary programming to generate skillful extreme value probabilistic forecasts. *Mon. Wea Rev*., **141**, 3170-3185.

Roebber, P.J., 2015a: Evolving ensembles. *Mon. Wea Rev*., **143**, 471-490.

Roebber, P.J., 2015b: Ensemble MOS and evolutionary program minimum temperature forecast skill. *Mon. Wea. Rev*., **143**, 1506-1516.

Roebber, P.J., 2015c: Adaptive evolutionary programming. *Mon. Wea. Rev*., **143**, 1497-1505.

Stevens, M. H. H., M. Sanchez, J. Lee, and S. E. Finkel, 2007: Diversification rates increase with population size and resource concentration in an unstructured habitat. *Genetics*, **177**, 2243–2250.