ELSEVIER

Journal of Hydrology 239 (2000) 69–84

Journal of Hydrology

www.elsevier.com/locate/jhydrol

A large-sample investigation of statistical procedures for radar-based short-term quantitative precipitation forecasting

M. Grecu, W.F. Krajewski*

Iowa Institute of Hydraulic Research, The University of Iowa, 300 South Riverside Drive, Iowa City, IA 52242, USA

Received 13 October 1999; revised 25 September 2000; accepted 25 September 2000

Abstract

We present the results of an extensive evaluation of radar-based quantitative precipitation forecasting techniques. Using a large data set of radar observations from the Tulsa, Oklahoma, WSR-88D radar we evaluate several techniques, including persistence, advection, and neural-network-based schemes. The scope of our study is limited to very-short-term forecast lead-times of up to three hours. We consider several spatial resolutions ranging from 4×4 km^2 to 32×32 km^2. Performance of the schemes is evaluated using several popular criteria that include correlation coefficient, multiplicative bias, and probability of detection. We discuss the effects of average storm intensity and rainfall intensity integration on the predictability limits. The most significant conclusions from the study are: (1) advection is the most important physical process that impacts useful predictions; (2) larger and more intense storms are easier to forecast; and (3) both spatial and temporal integration significantly extends the predictability limits.

Keywords: Rainfall; Weather radar; Forecasting

1. Introduction

Quantitative forecasts of precipitation are essential for practical applications in numerous areas including flood protection, water resources, agriculture, transportation, energy production, and industry. In some regions of US, the demand for water for domestic, industrial and agricultural purposes exceeds the available supply (US Global Research Program, 1997). Excess rainfall, on the other hand, can cause floods and flash-floods, some of the most damaging and deadly weather-related hazards. Obviously then, reliable quantitative forecasts of precipitation are needed if we are to operate and manage successfully the relevant infrastructure.

In this paper we focus attention on short-term (minutes to hours), high spatial resolution (about 4×4 km^2) forecasts of rainfall fields. The advent of instruments such as radars and satellites, combined with fast communication systems, increasingly complex numerical models, and high performance computers marks a significant advance in quantitative precipitation forecasting (QPF). However, the physical processes associated with rain are extremely complex and, at present, not completely understood. Consequently, these physical processes cannot be fully replicated by numerical models for predictive purposes. On the other hand, radar systems offer high spatial and temporal resolution observations, becoming attractive tools for flood forecasting applications.

* Corresponding author. Fax: +1-319-335-5238.
E-mail address: witold-krajewski@uiowa.edu (W.F. Krajewski).

0022-1694/00/$ - see front matter
PII: S0022-1694(00)00360-7

In this paper we limit the scope of our investigation to the radar-based, statistical QPF procedures. This paper does not address an alternative approach, that of incorporation (also referred to as "assimilation") of radar data into complex physically based numerical weather prediction models. Despite the fact that statistical, radar-based QPF procedures have been developed and tried in practice for over three decades (Austin and Bellon, 1974; Bellon and Austin, 1978; Tsonis and Austin, 1980; Bellon and Austin, 1984; Dixon and Wiener, 1993; Bellon and Zawadzki, 1994; Zawadzki et al., 1994), their comparative studies can provide important input to hydrologists. The objective of our paper is to evaluate these procedures for forecast lead times up to three hours.

It has been indicated in a series of studies (see Wilson et al., 1998 for references) that there is little skill in predicting precipitation for lead times greater than 30 min. However, the uncertainty of radar-based QPF, which depends on the temporal and spatial scale (resolution) of the predicted rainfall fields, has not yet been rigorously quantified. Such quantification is a necessary condition for effective flood and flash flood mitigation. Several hydrologic studies (e.g. Ogden and Julien, 1993; Singh, 1997) indicated a big influence of rainfall spatial and temporal variability on stream flow. This implies that the errors in rainfall forecasting have a big impact on flood forecasting.

Given the significantly nonlinear character of hydrologic processes, the existence of uncertainties, and the possibility of multiple performance criteria, the best precipitation forecasts do not provide necessarily the best flood forecasting. In this context, ensemble flood forecasts yield more reliable information than does a deterministic forecast. However, useful ensemble flood forecasting cannot be performed without a good characterization of the precipitation forecasting uncertainty. This justifies the need for a comprehensive investigation of the performance of current radar-based QPF procedures.

To achieve this goal, we formulated a radar-based general statistical procedure (GSP) for QPF, and investigated its performance using a large radar database. We assumed, in the GSP formulation, that the reflectivity fields are advected while undergoing dynamic changes. The advection velocity was estimated from two successive radar scans, using only reflectivity data. We modeled the dynamic component using a neural network. Because of its characteristics, using horizontal advection and flexible modeling of intensity changes, the procedure may be considered representative of a large category of statistical approaches.

We were motivated to use a large database by the Law of Large Numbers (Feller, 1968), which states that in any chance event, when the event happens repeatedly, the actual results will tend to be the calculated results. Thus, according to the law, the confidence in the results increases with the size of the database used in analysis. Our radar database consists of over a year's worth of WSR-88D Level II (Klazura and Imy, 1993) radar reflectivity data collected from Tulsa, OK, in the summers of 1994 and 1995. We evaluated the GSP performance based on a set of seven measures widely used in hydrology. These measures characterize the quantitative and qualitative (in terms of patterns) agreement between the forecasts and observations.

The GSP formulation allows us to investigate the effect of various components of the current QPF procedures. That is because two popular procedures, namely Eulerian and Lagrangian persistence (in stationary and moving systems of reference), are particular cases of the GSP. Also, various current formulations for modeling the reflectivity changes in time may be easily mimicked using the GSP. We consider three different types of forecasts: neural network-based (NN), Eulerian persistence (P), and Lagrangian persistence (LP) at several resolutions ranging from 4×4 km^2 to 32×32 km^2.

The paper is organized as follows. First, we present the GSP formulation and follow it with a description of the procedure used for velocity estimation from two consecutive radar scans. We devote the next section to the use of a neural network for modeling the dynamic changes of the reflectivity fields. We then define the performance measures and present the results, and close with the conclusions and a summary.

2. General statistical procedure

2.1. Formulation

Most of the statistical, radar-based QPF procedures

may be described by the following equation

$$\frac{\Delta Z^n}{\Delta t} + U_x \frac{\Delta Z^n}{\Delta x} + V_y \frac{\Delta Z^n}{\Delta y} = f(Z^n, \ldots, Z^{n-k}, \mathbf{a}) + w \tag{1}$$

where Z^n is the reflectivity at time n, U_x and V_y are the velocity components of the reflectivity field, $f(Z^n, \ldots, Z^{n-k}, \mathbf{a})$ is a function of parameters $\mathbf{a} = (a_1, a_2, \ldots, a_m)$ that are to be determined, and w is the noise component. The number of preceding scans taken into account for fitting the function f is denoted as k. Operator "Δ" is used instead of "∂" to make the notations consistent with the discrete nature of variable Z. Variables U_x, V_y, and Z are functions of x and y (the Cartesian coordinates of a grid point in the radar domain), but to simplify the notation we omit these arguments in equations. For example, notation U_x is used instead of $U_x(x, y)$.

Formulation (1) only partially models the modification of rainfall areas in time, as term $f(\cdot)$ may account for reductions, but not for expansions, of an area. This is clear if Eq. (1) is interpreted in a Lagrangian system of reference (moving with the storm) where a reflectivity pixel undergoes only intensity changes. There is no zero reflectivity pixel that can become strictly positive, because this would imply $f(0, \ldots, 0, \mathbf{a}) > 0$, and consequently, all zero pixels in a scan will become strictly positive by the integration of Eq. (1) for one time-step. This could lead to large errors, and therefore, only the non-zero pixels are used in forecasting, as we explain in detail later. More elaborate modeling of the modification of rainfall areas would be beneficial only for lead times shorter than 30 min. This is because the usual lifetime of rain cells is about 30 min (Wilson et al., 1998). Since we were concerned with lead times that are significantly longer (up to three hours), we made no attempt to further consider this aspect of the problem.

Model (1) may be formulated in terms of rainfall as well. In fact, our variable of interest is rainfall and not reflectivity, and, therefore, the performance should be evaluated relative to rainfall. However, validation of radar-rainfall estimation is a difficult problem in itself, and methods of uncertainty determination are not well established (Ciach and Krajewski, 1999; Anagnostou and Krajewski, 1999). Therefore, it is difficult to partition the total forecast error into those components that are due to radar-rainfall estimation and those due to a particular forecasting method. For this reason, we formulated most of our results in terms of reflectivity. In other words, we focused on the ability of the formulation given by Eq. (1) to predict the radar-reflectivity patterns. Still, one should keep in mind that the reflectivities are observed with some measurement error. Assuming the radar is well maintained and calibrated (i.e. there is no systematic bias in its measurements), the magnitude of the random component of this error is about 1–2 dBZ.

To apply model (1) for short term precipitation forecasting, the following steps are necessary: (a) quality control of the radar data; (b) estimation of radar echo velocity components U_x and V_y; (c) estimation of parameters $\mathbf{a}$ in $f(\cdot)$; (d) reflectivity forecasting using Eq. (1); (e) transformation of reflectivity into rainfall rates. The purpose of the quality control is to detect and remove non-rain echoes, mostly due to ground clutter and anomalous propagation (AP). To mitigate the negative effect of AP on QPF, we applied the AP procedure developed by Grecu and Krajewski (2000).

2.2. Estimation of radar echo velocity

The benefit of including velocity estimates in radar-based, short-term rainfall forecasting was recognized quite early (Austin and Bellon, 1974). Since then, several attempts have been made to refine the echo movement estimation (see, for example, Tuttle and Foote, 1990; Laroche and Zawadzki, 1994; Wu, 1995). Two distinct types of procedures may be used to estimate velocity from two consecutive images. The first is based on the analysis of the correlation between regions in the two images. This type of approach is robust, relatively accurate, but quite intensive computationally. The second approach is based on the assumption that the first image is subjected to a certain physical law (involving advection) that results in the second image. The velocity field is determined by minimizing the difference between the second image and the output obtained by applying the physical law to the first image. This second alternative provides a less computer-intensive solution. We used a procedure of the second type to estimate the radar echo movement in this study.

Eq. (1) is used to describe the reflectivity transformation between two consecutive scans. However, it would be too complicated to consider the advection and the source/sink term at the same time. Therefore, function f is considered null for the velocity estimation. Thus, Eq. (1) becomes

$$\frac{\Delta Z}{\Delta t} + U_x \frac{\Delta Z}{\Delta x} + V_y \frac{\Delta Z}{\Delta y} = w \tag{2}$$

The derivatives in Eq. (2) can be estimated from two consecutive radar scans using finite difference schemes. However, this is not sufficient to estimate the velocity field, because there is only one equation (Eq. (2)) for two unknowns (U_x and V_y). The velocity estimation may be reformulated as an optimization problem.

We first assumed that, due to the velocity field smoothness, the residual in Eq. (2) obtained by substituting the real values of the velocity with values corresponding to neighboring locations is small. Thus, we defined the following objective function at each location (x,y):

$$E(x,y) = \sum_{(i,j)\in\Omega(x,y)} \lambda(i,j) \times \left(\frac{\Delta Z_{i,j}^n}{\Delta t} + U_x \frac{\Delta Z_{i,j}^n}{\Delta x} + V_y \frac{\Delta Z_{i,j}^n}{\Delta y} \right)^2 \tag{3}$$

where λ is a positive parameter that controls the residual influence as a function of location. Velocities U_x and V_y in Eq. (2) were assumed to be constant in a neighborhood, Ω (x, y), centered on location (x, y). This assumption allows an analytical solution for the minimization of Eq. (3), but it does not imply that the velocity estimates are constant in Ω (x, y). Therefore, it is not as restrictive as it may seem initially. This formulation, although theoretically less accurate than that of Horn and Shunk (1981), is considerably faster and less computer-intensive.

The solution (U_x, V_y) that minimizes the function in Eq. (3) may be obtained by a weighted least-squared (WLS) regression. The time derivative is evaluated using two consecutive radar scans, while the spatial derivatives are obtained from a finite difference scheme.

Parameters $\lambda(i, j)$ were assumed to follow an exponential decay with distance

$$\lambda(i,j) = \exp\left(-\frac{\sqrt{(i-x)^2 + (j-y)^2}}{d} \right) \tag{4}$$

where d is a positive constant. Eqs. (2) and (3) require a Cartesian grid representation of the data. Therefore, the radar data originally represented in polar coordinates ($1\ \text{km} \times 1^\circ$) were projected on an orthogonal, uniform, $4 \times 4\ \text{km}^2$ grid.

We determined d and Ω (x, y) heuristically rather than based on mathematical considerations. For example, the smaller $\Omega(x, y)$, the better the optimal solution of Eq. (3) satisfies Eq. (2). However, the problem of minimizing Eq. (3) is ill-conditioned for a small Ω (x, y) because the x and y derivatives of Z do not vary much in a small neighborhood. We chose $\Omega(x, y)$ arbitrarily as a square whose size was determined by sensitivity studies. Distance d was initially set to 60 km, an arbitrary choice. We have observed by sensitivity studies that squares smaller than 40 km led to unnaturally large variability in the velocity estimates. Based on this observation and earlier considerations, $\Omega(x, y)$ was set to be a $40 \times 40\ \text{km}^2$ square.

2.3. Dynamic modeling

A back-propagation neural network (BPNN) was used to model function $f(\cdot)$ in Eq. (1). A BPNN is a multiparametric nonlinear function that can reproduce mappings between multidimensional real spaces. For a detailed description of neural networks, see, for example, Pao (1989). An algorithm can be easily derived to calibrate a BPNN to reproduce as accurately as possible an arbitrary function, if the problem is viewed in terms of a nonlinear regression solved using optimization techniques. We used an efficient quasi-Newton optimization method (Bertsekas, 1995) for the BPNN training, i.e. for the determination of parameters **a** in Eq. (1).

To develop function $f(\cdot)$ using a neural network modeling framework we considered time evolution changes in the reflectivity fields. In essence, we assumed that these changes could be predicted based on the local characteristics, both spatial and temporal, of the reflectivity fields. Thus, the first step in setting up the neural network was the selection of the

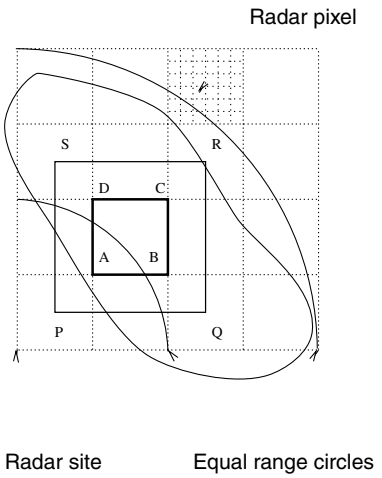

Fig. 1. Determination of BPNN training set.

predictors. Here, these were time series of reflectivity evolution.

Given a pixel in the reflectivity field, its corresponding past positions were determined using the velocity information. A time series was constructed for each pixel location (k, l). Let $\{Z^i_{k,l}, \hat{Z}^{i-1}_{k,l}, \ldots, \hat{Z}^{i-n}_{k,l}\}$ be the time series, where i is the time index. The symbol "^" is used to distinguish between the reflectivity at location (k, l) and time i–j, and the value at the same time and the position that translates to (k, l) at time i.

The time series was derived as follows. A pixel at time i was displaced back in time, and the reflectivity values at its corresponding locations were determined. Next, sequences of m, $2 \leq m \leq n$ consecutive elements in the time series were used to define the BPNN training set. The first element in a sequence designated the BPNN output, and the others its inputs. The sequences with the last $m-1$ elements null were not included in the training set. Due to limited duration of storms, the time series are not long enough to provide sufficient information for training a neural network. Therefore, the training sets of several pixels were concatenated and a unique BPNN was determined for these pixels. The criterion for choosing the pixels was purely geometric. The radar domain was divided into rectangular windows, and a BPNN was trained and used for forecasting in each window. The windows that defined the BPNNs did not overlap (ABCD in Fig. 1). The training set was defined based on a larger window that includes the one for which the BPNN was defined (PQRS in the Fig. 1). That is, the BPNN was trained using the data provided by square PQRS, but was applied only to the pixels in square ABCD. This led to a set of overlapping windows for BPNN training. We used overlapping windows in BPNN training to prevent significantly different forecasts among neighboring pixels that belong to different BPNN definition windows. Another option would be the use of only one BPNN for the entire scan. This alternative, although less computer-intensive, might be less accurate because dynamics of the same type are implicitly imposed on two or more distant regions. These distant regions might be characterized by different phases of the storm, e.g. decay for one and intensification for another. As information, the trend in reflectivity is not reliable enough to identify the phase of the storm. By training different BPNNs for different regions, the information in the reflectivity is better exploited. Consequently, BPNNs were trained for regions small enough to be of the same type but with a number of pixels sufficient for training. The exact figures we used in the BPNN structure determination and the training are given in the next section.

We made the forecasts in a Lagrangian system of reference. That is, forecasts were made for the time series defined in each pixel of the current scan for the lead times of up to three hours. Then, we advected the forecasted fields accordingly.

3. Results and discussion

The GSP was applied to a large radar data set consisting of 145 days of full volume scans (6439 scans, one taken every six minutes) collected by a WSR-88D system in Tulsa, Oklahoma in the summer months of 1994 and 1995. The data were quality controlled, pixel-by-pixel, using the procedure developed by Grecu and Krajewski (2000).

3.1. Performance measures

We used the following seven performance measures to evaluate the procedure: coefficient of correlation, coefficient of efficiency, index of agreement, bias, probability of detection, false alarm ratio,

Table 1
Classification of events in the forecast contingency table

Event	Condition
Success	$F = 0$ and $O = 0$
False alarm	$F > 0$ and $O = 0$
Failure	$F = 0$ and $O > 0$
Success	$F > 0$ and $O > 0$

and critical success index. The measures were defined based on the agreement between the forecasts F and observations O. The number of forecasts and observations is N (the notation follows that of Legates and McCabe, 1999). The correlation coefficient describes the proportion of the total variance in the observed data that can be explained by the forecasts. Its formula is

$$C = \frac{\sum_{i=1}^{N}(O_i - \bar{O})(F_i - \bar{F})}{\left(\sum_{i=1}^{N}(O_i - \bar{O})^2\right)^{0.5}\left(\sum_{i=1}^{N}(F_i - \bar{F})^2\right)^{0.5}} \quad (5)$$

where the bar indicates the mean value. It is well known that the correlation coefficient is not sensitive to proportional differences between the observations and forecasts. For this, and other reasons (see Legates and McCabe, 1999), the correlation coefficient is not a sufficient measure of the agreement between observations and forecasts.

The coefficient of efficiency E is defined as

$$E = 1 - \frac{\sum_{i=1}^{N}(O_i - F_i)^2}{\sum_{i=1}^{N}(|O_i - \bar{O})^2} \quad (6)$$

E ranges from $-\infty$ to 1, and a large value indicates a good agreement between the observations and forecasts. If the mean squared error is as large as the variance of the observed data, then $E = 0$. If $E < 0$, then the observations are better described by their mean than by the forecasts.

The index of agreement, d, was introduced to overcome the insensitivity of correlation-based measures to differences between the forecasted and observed means (Legates and McCabe, 1999)

$$d = 1 - \frac{\sum_{i=1}^{N}(O_i - F_i)^2}{\sum_{i=1}^{N}(|O_i - \bar{O}| + |F_i - \bar{O}|)^2} \quad (7)$$

The index of agreement takes values from 0.0 to 1.0, greater values corresponding to better predictions.

The bias (multiplicative) is defined as the ratio of the mean forecast to the mean observation

$$B = \frac{\bar{F}}{\bar{O}} \quad (8)$$

It ranges from $-\infty$ to ∞. In a general context, the bias might not be an important measure, but for hydrologic applications it is, as the rainfall-runoff models often amplify it.

The four measures presented above characterize globally the agreement between forecasted and observed rain fields, but do not provide information about their match in terms of patterns. To quantify the pattern match, we used measures based on contingency tables. Constructing the contingency tables requires defining concepts such as *success*, *failure*, and *false alarm*. Table 1 contains these definitions. We counted the numbers of successes n_s, failures n_f, and false alarms n_{fa} in a forecasted field, and, based on them, defined the probability of detection, false alarm ratio, and critical success index as follows:

$$\mathrm{POD} = \frac{n_s}{n_s + n_f} \quad (9)$$

$$\mathrm{FAR} = \frac{n_{fa}}{n_s + n_{fa}} \quad (10)$$

and

$$\mathrm{CSI} = \frac{n_s}{n_s + n_f + n_a} \quad (11)$$

3.2. Experimental setup

To make forecasts in a set-up that mimics a real-time operational environment we applied the statistical procedure presented above. The radar scans were read sequentially, and the reflectivity fields within a range of 200.0 km from the radar projected

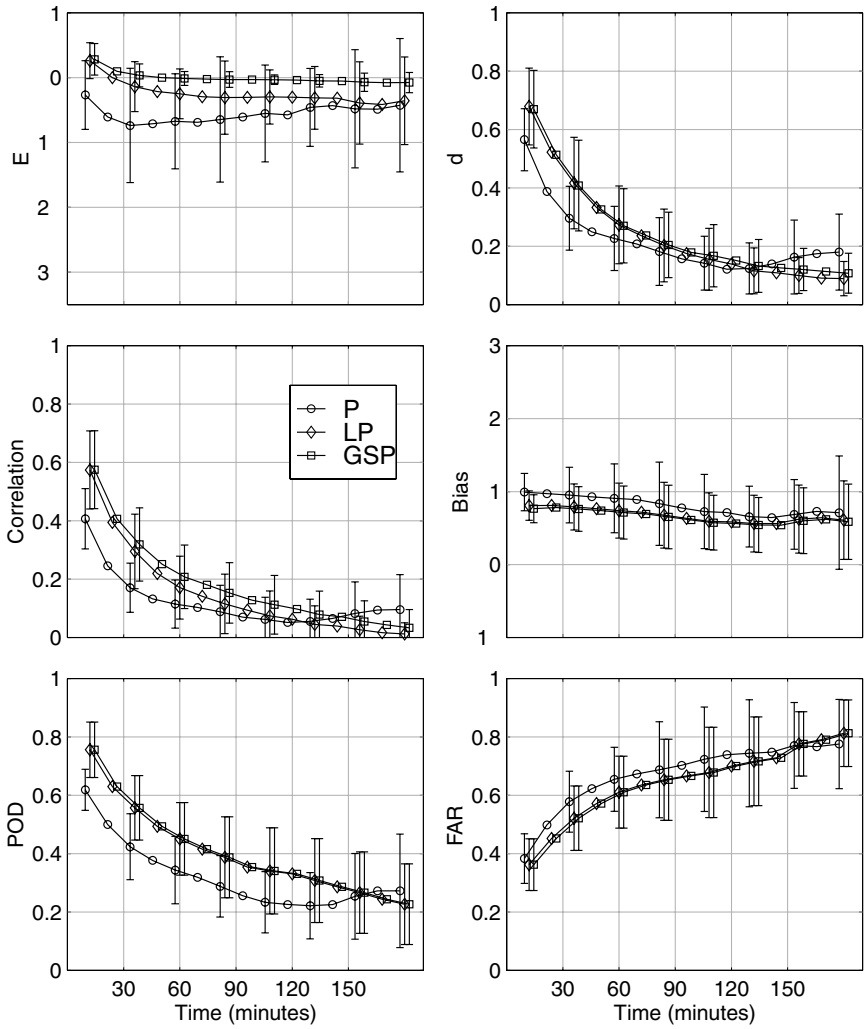

Fig. 2. Skill measures as a function of forecasting lead-time. Instantaneous fields, $Z_{av} < 15$ dBZ, 4 km resolution. Measures shown: efficiency coefficient (E), coefficient of correlation, bias, probability of detection (POD), and false alarm ration (FAR). The vertical bars indicate plus/minus one standard deviation.

on an orthogonal grid with size of 4.0 km. We used the current and previous scans to estimate the velocity of the reflectivity field, saving the velocity fields in a circular queue of size 10. That is, only the last ten velocity fields were saved, and the eleventh overwrote the first. The reflectivity scans (in Cartesian representation) were saved in a queue of size 11. This size corresponds roughly to an hour of reflectivity data (for the WSR-88Ds the lag between scans varies, but on the average it is about six minutes). Then, based on the saved velocity and reflectivity fields, we derived time series for each grid point. These time series were used for the BPNN training. Window size for a BPNN (square ABCD in Fig. 1) was 40 km (10 grid points). The larger window that defined the training set (square PQRS in Fig. 1) was 80 km (20 grid points). The length of the sequences in the time series was four, implying that the change in a reflectivity value was forecasted based on its three predecessors. We chose six as the number of neurons in the hidden layer (Pao, 1989), though other choices could be used as well.

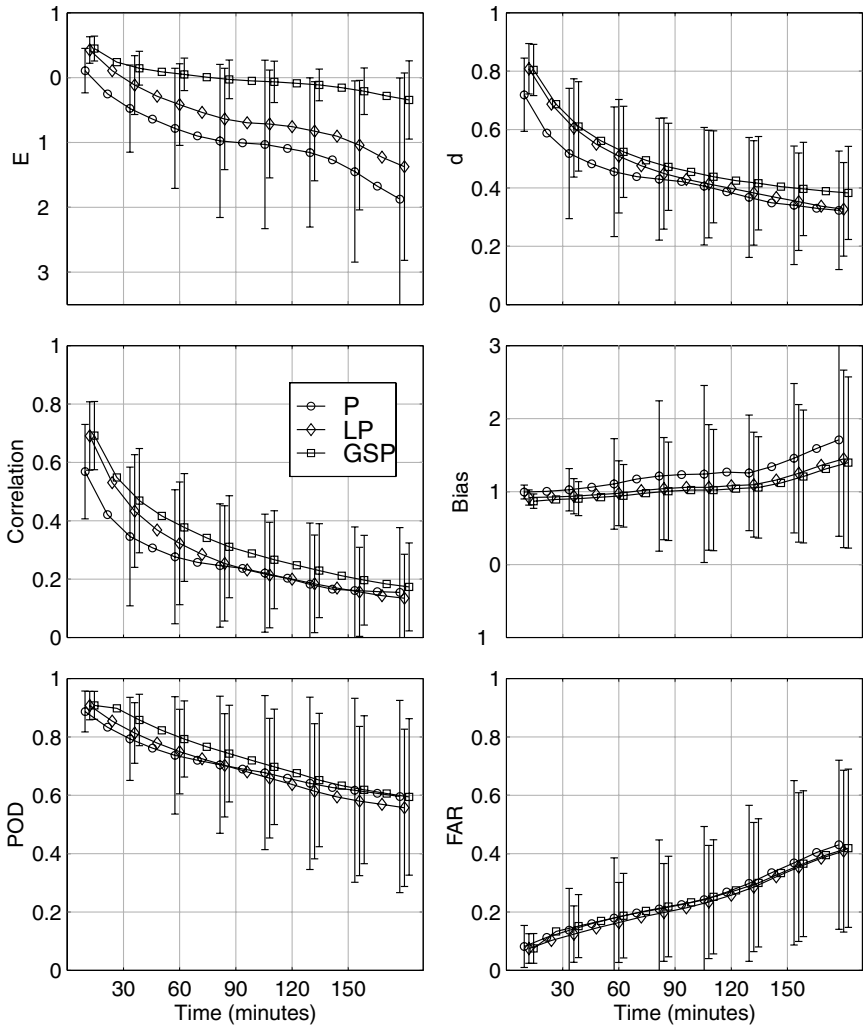

Fig. 3. Same as for Fig. 2, but for $Z_{av} > 25$ dBZ and 4 km resolution.

In principle, the increase in numbers of neurons in the hidden layer determines performance improvement. However, our tests, made on a small portion of radar data, showed no significant improvement. If the training set determined by a window contained less than 200 elements, then no training was done. The above number corresponds roughly to six times the number of BPNN parameters. According to an empirical rule, a training set approximately five times bigger than the number of BPNN parameters is required to assure the significance of results (Pao, 1989). In our study, if this condition was not met, no BPNN was trained and the Lagrangian persistence was applied instead. The forecasts, for lead times up to three hours, were saved every 12 min. At the same time, the other two types of forecasts, namely persistence and Lagrangian persistence were made and saved. Later, we compared the three types of forecasts off-line with the observed reflectivity fields.

The forecasts were made in terms of reflectivity for the reasons mentioned earlier. However, we evaluated them in terms of rain, which is the variable of interest for hydrologic applications. The Z–R (where Z is the reflectivity and R the corresponding rain rate) relationship used was arbitrarily chosen as $Z = 300R^{1.4}$, the same as the one used operationally in

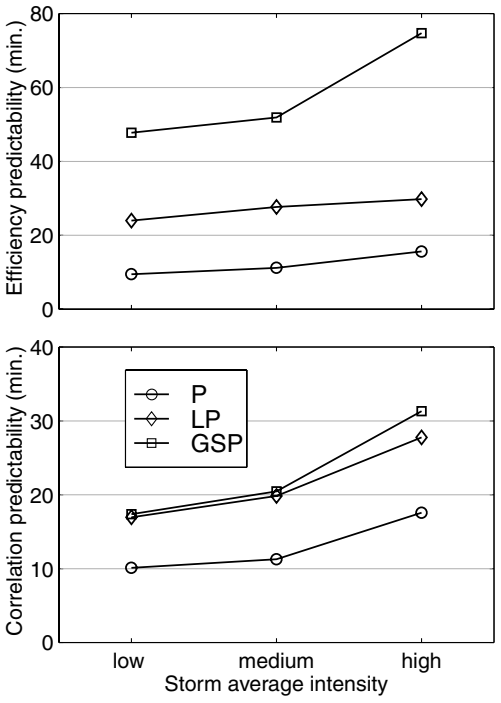

Fig. 4. Predictability as a function of storm average intensity. Category "low" corresponds to $0 < Z_{av} < 15$ dBZ, category "medium" to 15–25 dBZ and category "high" to $Z_{av} > 25$ dBZ.

the WSR-88D system. In operational use, the particular Z–R relationship used could make a big difference (Anagnostou and Krajewski, 1998), but for the current study it was not essential to find the most appropriate parameters, because the errors in rainfall estimation affect both the forecasts and observations equally. For example, an error in the multiplicative coefficient of the Z–R relationship did not affect at all the performance measures defined earlier. Moreover, the errors in the forecasts were likely to be substantially greater than those in the estimates.

3.3. Large sample results

We determined the statistics of the measures defined in Eqs. (5)–(11) as a function of the storm intensity. That is, we calculated the average reflectivity of non-zero cells for each scan. The scans were then classified as a function of their average reflectivity in three categories determined by the intervals: (1) below 15 dBZ; (2) between 15 and 25 dBZ; and (3) above 25 dBZ. The skill measures were also classified in these three categories and characterized by their means and standard deviations.

The results for the average reflectivity Z_{av} less than 15.0 dBZ are given in Fig. 2 (for the sake of plot symmetry we omit the CSI, as it only synthesizes the information contained in FAR and POD.) A degradation of performance with time is obvious. We note no significant difference between the GSP performance and that of the LP. The simple persistence, P, shows worse results, as expected. The GSP indicates significantly better results only for the efficiency coefficient E. As is obvious from its definition, a negative E indicates that the variance of errors is larger than that of observations. For the situation in Fig. 2, the means of the forecast squared errors for P and LP become bigger than the variance of observations for relatively short lead times (12 min for P and 24 for LP). The GSP shows positive E for 36 min. The time at which E becomes negative may be considered a predictability limit. Another predictability measure is the decorrelation time defined arbitrarily as the moment the correlation becomes less than 0.5 (Zawadzki et al., 1994). The two values agree quite well for P and LP, while for GSP, they are quite different. This indicates that E alone is not a sufficient measure for characterizing the rainfall predictability. It is likely that the GSP forecasts are, in general, spatially smoother than those of P, and LP (property expected given the increased volume of numerical calculations involved), and this leads to smaller sums of the squared errors. However, this does not provide overall better performance of the GSP relative to LP; a point also clearly indicated by the other measures we used. Nevertheless, the improvement in performance produced by considering the advection is obvious. It is also worth pointing out that the standard deviation of all but one of the performance measures is similar for the three types of forecasts. The only notable difference occurs in terms of coefficient of efficiency.

The skill measures for the third category of scans, $Z_{av} \geq 25$ dBZ, are given in Fig. 3. An improvement may be observed in terms of all measures. This supports the conclusion that rainfall is more predictable for intense events than for weak ones. This is in agreement with Zawadzki et al. (1994), who related the decorrelation time to large-scale properties that determine the precipitation intensity. They also

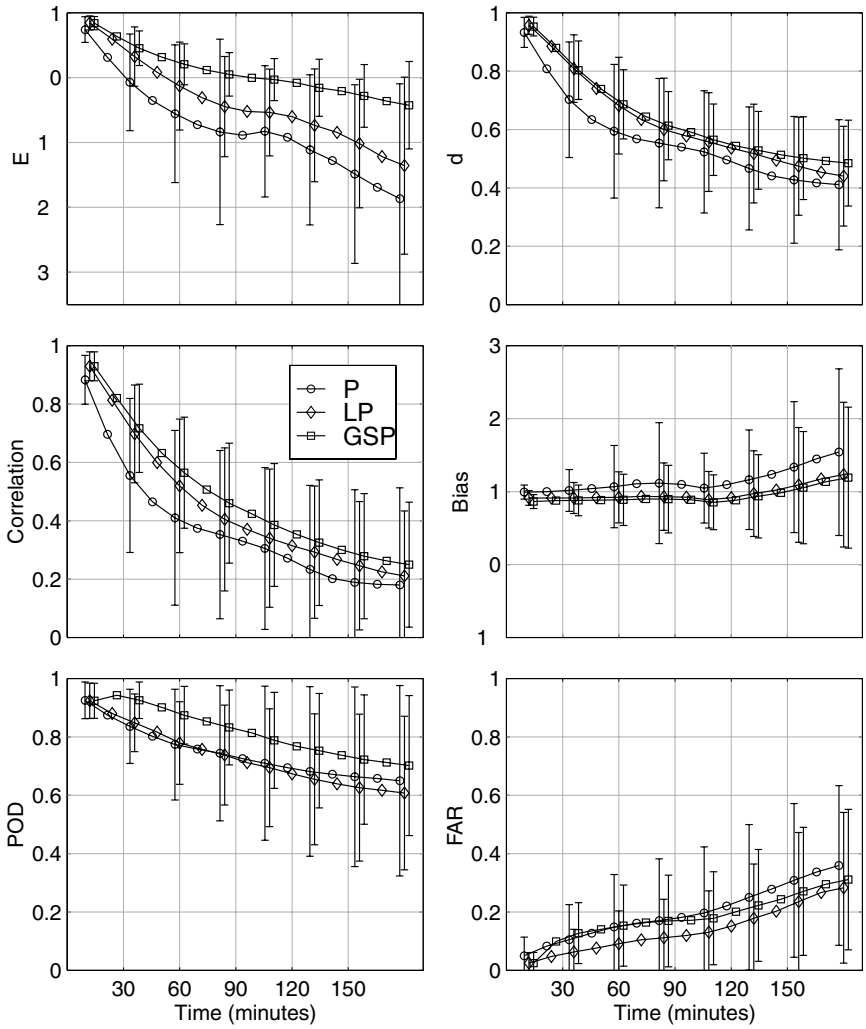

Fig. 5. Same as for Fig. 2, but for $Z_{av} > 25$ dBZ and 32 km resolution.

found longer decorrelation times for stronger events. Similar to $Z_{av} < 15$ dBZ, both the GSP and the LP perform better than the P, and the difference between the GSP and the LP is mainly in the efficiency coefficient. Most likely, the effect suggested in case $Z_{av} < 15$ dBZ is responsible for the difference.

We synthesize the results in terms of the predictability as a function of average reflectivity in Fig. 4 using two measures: the forecast lead-time at which E becomes negative, and the lead-time at which the correlation coefficient becomes less than 0.5. The predictability increase with the average reflectivity, seen in Figs. 2 and 3, is evident in Fig. 4 as well. The predictability defined by correlation is, probably, a better indication concerning the statistical QPF limitations. This is because the coefficient of efficiency, E, shows significantly better skills for QPF than for P and LP, behavior that is not seen in other measures. The two predictability measures are determined from Figs. 2 and 3, and the corresponding results for 15 dBZ $\leq Z_{av} < 25$ dBZ (not shown in the paper).

To investigate the effect of the forecast resolution on its performance (Bellon and Zawadzki, 1994), the results obtained for the 4×4 km^2 grid size were

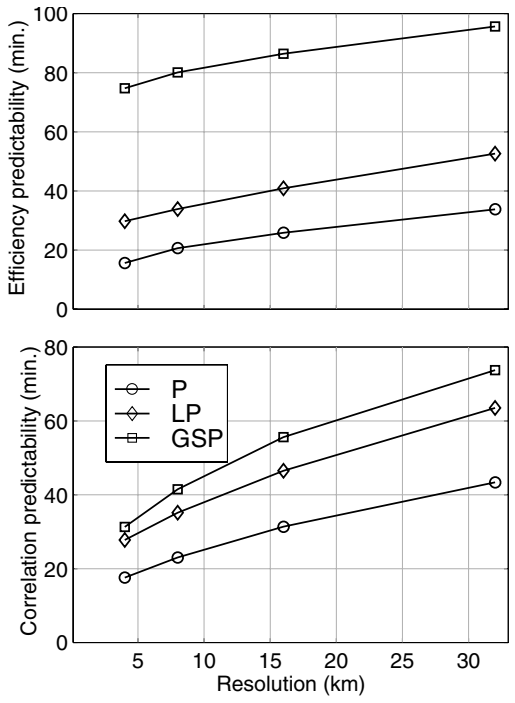

Fig. 6. Predictability as a function of resolution.

aggregated on grids with sizes 8×8, 16×16, and 32×32 km^2. That is, the forecasts were made on the 4×4 km^2 grid, and averaged at 8, 16 and 32 km resolution. One motivation for expressing the GSP performance at different scales is because numerical weather prediction models, including the mesoscale models, are run at similar resolutions. For example, the mesoscale model MM5 runs at a 36 km resolution (White et al., 1999), and the ETA model, at 32 km (Black, 1994). Unfortunately, to the best of our knowledge, the literature lacks a comprehensive evaluation of the mesoscale model performance for short-time QPF, therefore, we cannot make a direct comparison. We hope this situation will change in the near future.

The analysis of GSP performance at different resolutions has other benefits as well. It is possible from the multiscale analysis to identify the limitations of the statistically based forecasting procedures and characterize better the rainfall predictability. Fig. 5 represents the performance measures for $Z_{av} >$ 25 dBZ, at 32 km resolution. There is a significant improvement in all measures relative to the case of 4×4 km^2 resolution. This indicates that the large-scale precipitation features are more persistent (both in Eulerian and Lagrangian systems of reference) than are the small ones. However, in time, the mass distribution within such features changes rather quickly. The figures show good skills ($E > 0.0$, correlation $>$ 0.5) for about one hour, relative to just 30 min as in the case of the 4 km resolution. But, one hour duration is longer than the lifetime of an individual storm cell, which is generally less than 30 min. This indicates a complex interaction among storm cells responsible for the mass organization in large-scale precipitation volumes. That is, a cell dying in less than 30 min at a certain location must be accompanied by storm intensification at some relatively close location. Even in the case of complex (multicellular) storms, small scale features are not expected to show predictability beyond 30 min (Fig. 3), but their changes are strongly coupled at larger scales where the predictability extends over longer time scales. These considerations allow us to understand the GSP characteristics better. The GSP models the storm evolution in a weakly coupled manner, i.e. the precipitation dynamics at a certain location is not significantly affected by what happens at relatively close locations. That is why the GSP skills become poor beyond 30–45 min (for $Z_{av} > 25$ dBZ). However, there is still large-scale useful information beyond that limit. An improvement in performance with the increase of average reflectivity and the decrease of resolution is evident.

In Fig. 6 we summarize predictability as a function of the forecast resolution. This figure provides a quantitative description of the performance improvement with the resolution decrease, an implication observed in the analysis of Fig. 3 relative to Fig. 5. To summarize our results further, we show the error histograms (Fig. 7). Error is defined as the difference between the forecast and observation. Smooth symmetric distributions characterize P and LP at all lead-times. The GSP shows the asymmetric character of the error with more frequent overestimation. Another important error characteristic is the so-called conditional bias (e.g. Ciach et al., 2000). Fig. 8 shows the three methods and the three forecast lead-times. The conditional bias is defined as the expectation of the error conditional on the indicated observed value. The results clearly demonstrate that the forecasts

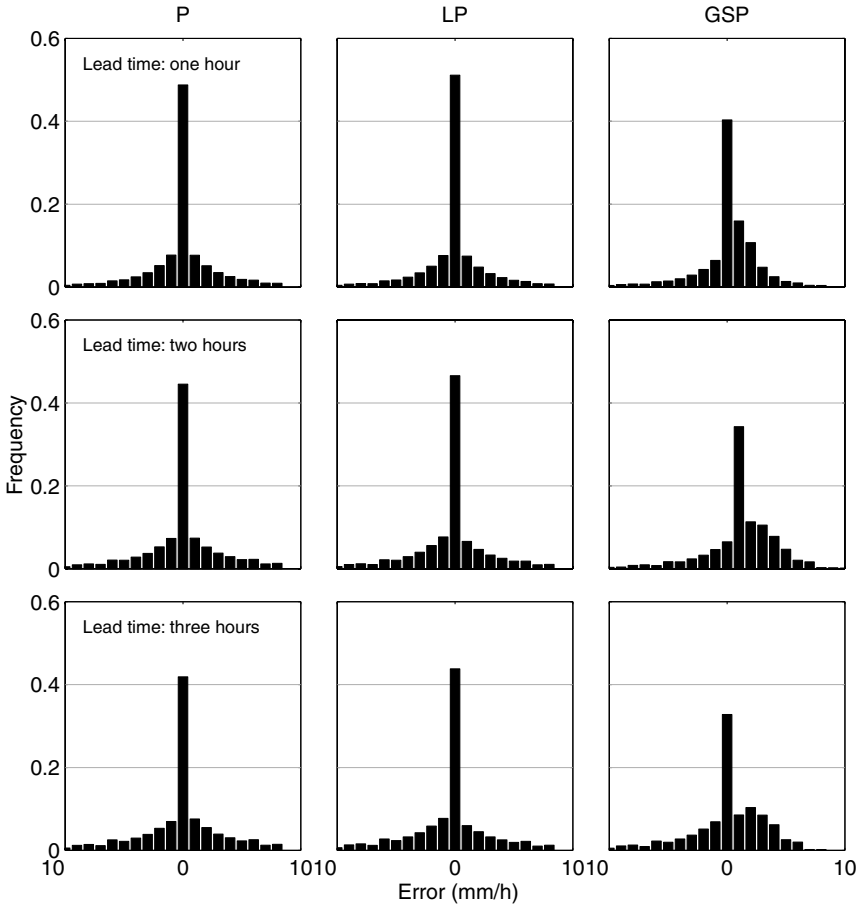

Fig. 7. Error histogram. First row corresponds to one hour lead-time, second row to two hours, and third row to three hours.

overestimate low intensity rain while underestimating rain of high intensity. This is not a desirable characteristic, but we see few options to improve it.

The results become better if we consider rainfall accumulation patterns. Performance of GSP, P and LP is given in Fig. 9 as a function of lead-time for strong storms ($Z_{av} > 25$ dBZ) and 4 km resolution. The accumulation is not forecasted, but derived from the rain rate forecasts. The rain accumulation represents useful information for hydrologic applications, where the water budgets require integrated rain-rates, i.e. accumulations. The figure shows considerably better skills than in the case of rain-rates. This is because for short lead-times the error in rain-rate forecasts is small, which positively affects the accumulation forecasts. The P yields significantly worse results than the other two methods. This is a consequence of its not considering the advection, which makes the rain accumulate at the same locations as it was last observed. Fig. 9 suggests the size of the negative impact had by the use P, instead of advection-based schemes, on hydrologic applications. Even though in terms of rain rates, the improvement provided by advection may sometimes be considered marginal, it is important to account for advection if accumulated quantities are to be used. The advection is responsible for distributing the rain mass over larger areas, compared to the P that concentrates in a smaller area with obvious negative consequences.

Similar to the rain-rate case, the performance improves with decrease in resolution. Fig. 9 shows the pattern agreement measures as a function of lead-time. Considerably lower rate of performance

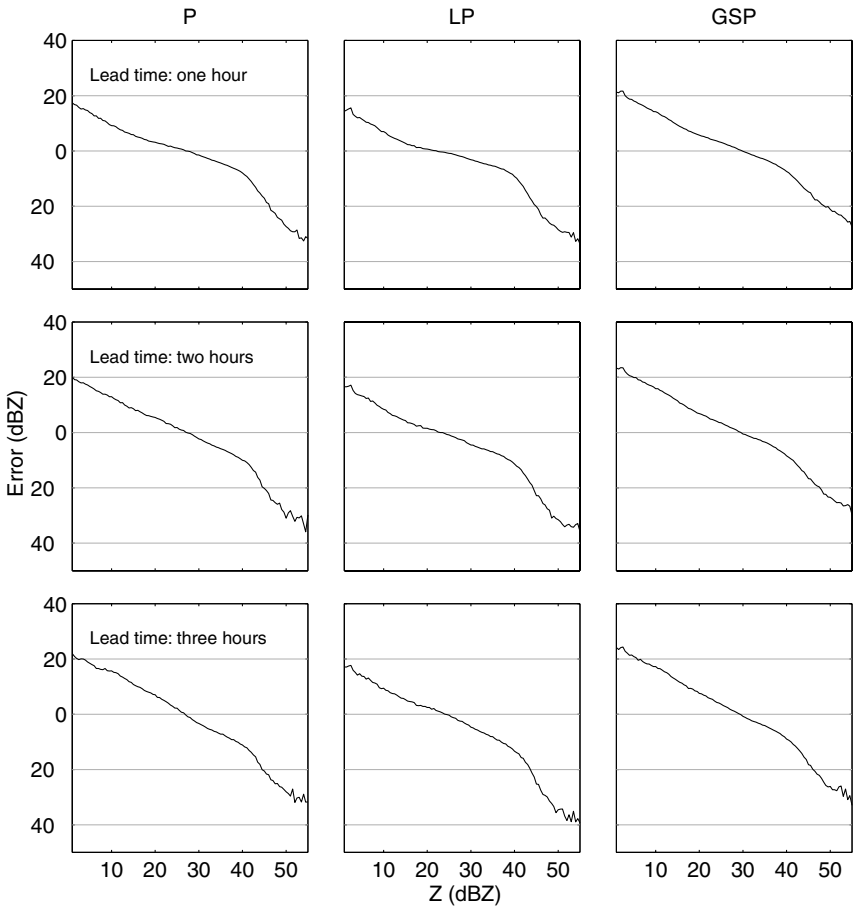

Fig. 8. Conditional bias. First row corresponds to one-hour lead-time, second row to two hours, and third row to three hours.

degradation with the lead-time, compared with that of the instantaneous fields, is apparent.

4. Summary and conclusions

We presented a comprehensive investigation concerning the performance of a general statistical procedure for radar-based QPF. The procedure is representative of a large category of approaches currently used for QPF. The necessity to quantify the QPF errors is dictated by the major role of precipitation in flood and flash-flood forecasting.

One important finding of our study is that there is no significant difference between the neural-network-based procedure and the relatively simple advection schemes, and that both perform significantly better than the persistence scheme. Thus, it is beneficial for QPF to consider advection, but statistical modeling helps little. Still, considering the high non-linearity of the system, even this small improvement may warrant further investigation of the GSP approach. Arguably, the GSP presented in this study cannot account for all possible formulations of neural networks. However, the results suggest that statistical modeling does not guarantee better QPF than do simpler schemes. Considering that statistical modeling is computer-intensive for real-time applications, then simple, advection-based schemes may be preferable in some application.

We also succeeded in quantifying the performance improvement for all three procedures due to the increase of average storm precipitation intensity and the spatial aggregation scale. The results show

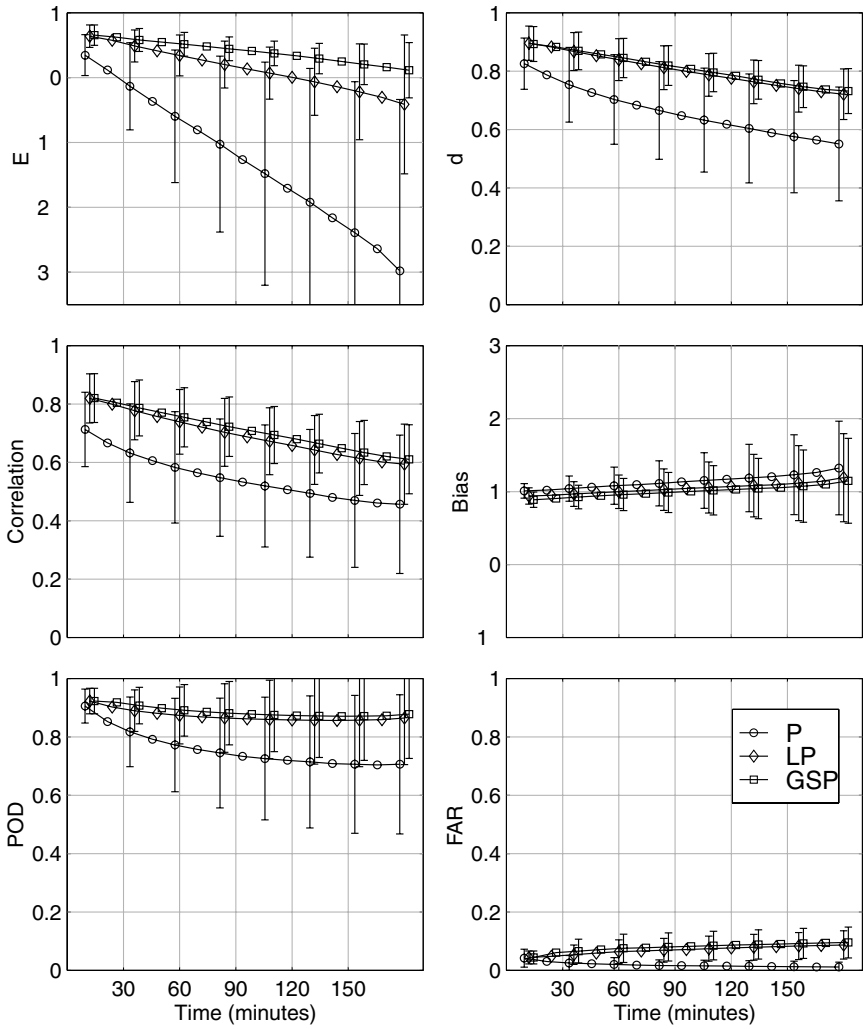

Fig. 9. Skill measures as a function of forecast lead-time. Accumulation fields, $Z_{av} > 25$ dBZ, 4 km resolution.

performance improvement with the decrease of resolution, which supports the conclusion that large-scale rain features are characterized by better Lagrangian persistence. The ability of the GSP to model small-scale features seems limited by the fact that the interaction between vertical and horizontal developments is not explicitly considered. Even procedures that may be considered "two and a half"-dimensional because of their parameterizations of the vertical precipitation structure (Lee and Georgakakos, 1996; French and Krajewski, 1994) are probably not flexible enough to capture the small-scale interactions

We evaluated the forecasting schemes in terms of rainfall accumulation as well. In this case, persistence is significantly worse than the other procedures, mainly because it accumulates the precipitation in a constant area. This might have strong negative impact on hydrologic applications. The improvement of performance as event intensity increases and spatial resolution decreases, evident in rain rate forecasting, is present in accumulations as well.

Perhaps the most important conclusion of this study is that the QPF procedures need to account explicitly for the three-dimensional evolution of phenomena to capture properly the interaction among small spatial and temporal scales (few kilometers, and below

30 min). The interaction is dictated mainly by vertical developments, about which radar provides information but we chose to ignore. One possibility of accounting for these developments is the use of cloud models. However, cloud-model-based QPF involves the development of mathematically complex, computer-intensive procedures for assimilation of radar data (Sun and Crook, 1997). It follows that a compromise between the accuracy of physical representation, observational capabilities, and mathematical complexity needs to be sought for a significant improvement in real-time QPF.

Acknowledgements

This study was supported by NASA Grant NAG8-1425 under the auspices of the US Weather Research Program. The radar data were provided under the Cooperative Agreement between the Office of Hydrology of the National Weather Service and the Iowa Institute of Hydraulic Research (NA47WH0495). We thank Drs Konstantine Georgakakos and Mark French for useful discussions, and Dr Isztar Zawadzki for his comments on the original manuscript.

References

Anagnostou, E.N., Krajewski, W.F., 1998. Calibration of the WSR-88D precipitation processing subsystem. Weather and Forecasting 13 (2), 396–406.

Anagnostou, E.N., Krajewski, W.F., 1999. Real-time radar rainfall estimation. Part II: case study. Journal of Atmospheric and Oceanic Technology 16 (2), 189–197.

Austin, G.L., Bellon, A., 1974. The use of digital weather records for short-term precipitation forecasting. Quarterly Journal of the Royal Meteorological Society 100, 658–664.

Bellon, A., Austin, G.L., 1978. The evaluation of two years of real-time operation of a short-term precipitation forecasting procedure (SHARP) (Short-term automated radar prediction). Journal of Applied Meteorlogy 17, 1778–1787.

Bellon, A., Austin, G.L., 1984. The accuracy of short-term radar-rainfall forecasts. Journal of Hydrology 70, 35–49.

Bellon, A., Zawadzki, I., 1994. Forecasting of hourly accumulations by optimal extrapolation of radar maps. Journal of Hydrology 157, 211–233.

Bertsekas, D.P., 1995. Nonlinear Programming. Athena Scientific, Belmont, MA (646pp.).

Black, T.L., 1994. The new NMC mesoscale ETA-model — description and forecast examples. Weather and Forecasting 9 (2), 265–278.

Ciach, J.G., Krajewski, W.F., 1999. On the estimation of radar rainfall error variance. Advances in Water Resources 22 (6), 585–595.

Ciach, J.G., Morrissey, M.L., Krajewski, W.F., 2000. Conditional bias in radar rainfall estimation, Journal of Applied Meteorology 39 (11), 1941–1946.

Dixon, M., Wiener, G., 1993. TITAN — thunderstorm identification, tracking, analysis, and nowcasting — a radar-based methodology. Journal of Atmospheric and Oceanic Technology 10 (6), 785–797.

Grecu, M., Krajewski, W.F., 2000. An efficient methodology for detection of anomalous propagation echoes in radar reflectivity data using neural networks. Journal of Oceanic and Atmospheric Technology 17 (2), 121–129.

Feller, W., 1968. An Introduction to Probability Theory and its Applications. Wiley, New York (516pp.).

French, M.N., Krajewski, W.F., 1994. A model for real-time quantitative rainfall forecasting using remote-sensing. 1, Formulation. Water Resources Research 30 (4), 1075–1083.

Horn, B.K., Schunk, P.B., 1981. Determining optical flow. Artificial Intelligence 17, 185–204.

Klazura, G.E., Imy, D.A., 1993. A description of the initial set of analysis products available from the NEXRAD WSR-88D system. Bulletin of the American Meteorological Society 74 (7), 1293–1311.

Laroche, S., Zawadzki, I., 1994. A variational analysis method for retrieval of three-dimensional wind-field from single Doppler radar data. Journal of the Atmospheric Sciences 51 (18), 2664–2682.

Lee, T.-H., Georgakakos, K.P., 1996. Operational rainfall prediction on meso-gamma scales for hydrologic applications. Water Resources Research 32 (4), 987–1003.

Legates, D.R., McCabe, G.J., 1999. Evaluating the use of goodness-of-fit measures in hydrologic and hydroclimatic model validation. Water Resources Research 35 (1), 233–241.

Ogden, F.L., Julien, P.Y., 1993. Runoff sensitivity to temporal and spatial rainfall variability at runoff plane and small basin scales. Water Resources Research 31 (6), 1533–1541.

Pao, Y.H., 1989. Adaptive Pattern Recognition and Neural Networks. Addison-Wesley, Readings (309pp.).

Singh, V.P., 1997. Effect of spatial and temporal variability in rainfall and watershed characteristics on stream flow hydrograph. Hydrological Processes 11 (12), 1649–1669.

Sun, J.-Z., Crook, N.A., 1997. Dynamical and microphysical retrieval from Doppler radar observations using a cloud model and its adjoint. 1. Model development and simulated data experiments. Journal of the Atmospheric Sciences 54 (12), 1642–1661.

Tsonis, A.A., Austin, G.L., 1980. An evaluation of extrapolation techniques for the short-term prediction of rain amounts. Atmosphere-Ocean 19 (1), 54–65.

Tuttle, J.D., Foote, G.B., 1990. Determination of the boundary-layer air-flow from a single Doppler radar. Journal of Atmospheric and Oceanic Technology 7 (2), 218–232.

US Global Change Research Program, 1997. Workshop on Climate

Variability and Water Resources Management in the Southeastern US, US Global Change Research Program, 74pp.

Wilson, J.W., Crook, N.A., Mueller, C.K., Sun, J., Dixon, M., 1998. Nowcasting thunderstorms: a status report. Bulletin of the American Meteorological Society 79 (10), 2079–2099.

White, B.G., Paegle, J., Steenburgh, W.J., Horel, J.D., Swanson, R.T., Cook, L.K., Onton, D.J., Miles, J.G., 1999. Short-term forecast validation of six models. Weather and Forecasting 14 (1), 84–108.

Wu, Q.X., 1995. A correlation–relaxation–labeling framework for computing optical flow — template matching from a new perspective. IEEE Transactions on Pattern Analysis and Machine Intelligence 17 (9), 843–853.

Zawadzki, I., Morneau, J., Laprise, R., 1994. Predictability of precipitation patterns — an operational approach. Journal of Applied Meteorology 33 (12), 1562–1571.