6B.3

LARGE SAMPLE NONPARAMETRIC MODELING OF THE UNCERTAINTIES IN RADAR RAINFALL PRODUCTS

Grzegorz J. Ciach*, Witold F. Krajewski, and Gabriele Villarini
The University of Iowa, Iowa City, Iowa, USA

1. INTRODUCTION

The high level of uncertainties in radar rainfall (RR) estimates is a broadly acknowledged problem. However, comprehensive information about their mathematical structure is not available. In fact, the operational RR products delivered by the USA National Weather Service based on the WSR-88D stations still lack any information on their error bounds. In our opinion, the most complete description of RR uncertainties can be achieved by providing the products in a probabilistic rather than deterministic form. To advance this direction, we are developing an empirically based approach to the quantification of the functional-statistical error structure of RR products. Our prospective goal is to create a realistic mathematical model describing the dependence of the error frequency distribution on RR in different situations. We search for a parsimonious model that can have the same mathematical form under a broad range of conditions, and the possibly small set of parameters that can be estimated in each situation using the available data. At present, the considered conditions include different distances from the radar, seasons of the year, time-scales and Z-R relationships. In the future, when adequate data samples become available, we will investigate the sensitivity of the uncertainty model to different spatial resolutions, geographic locations and climatic regimes, precipitation types, and different RR estimation algorithms.

In this paper, we describe the selected results of our first large-sample modeling of RR uncertainty. We present a functional-statistical model of RR error in its structural form that was specifically designed for the purposes of the probabilistic quantitative precipitation estimation (PQPE) based on WSR-88D data. We also discuss briefly the applications of our modeling results to the PQPE.

Our model of RR error structure describes the combined effect of all the error sources in RR. Since the error structure identification applies to the final RR products, we call this a product-error driven (PED) approach. The analyses are based on six years (1998-2003) of Level II data from the Oklahoma City radar (KTLX). These data are uniformly processed with the Built 4 version of the Precipitation Processing System (PPS) of the NEXRAD (Fulton et al. 1998). The PPS-generated products are then compared with the ground reference (GR) based on raingauge data from two good quality surface networks: the Oklahoma Mesonet, and the Micronet of the Agricultural Research Service (ARS). The schematic of this data collection setup is shown in Figure 1.

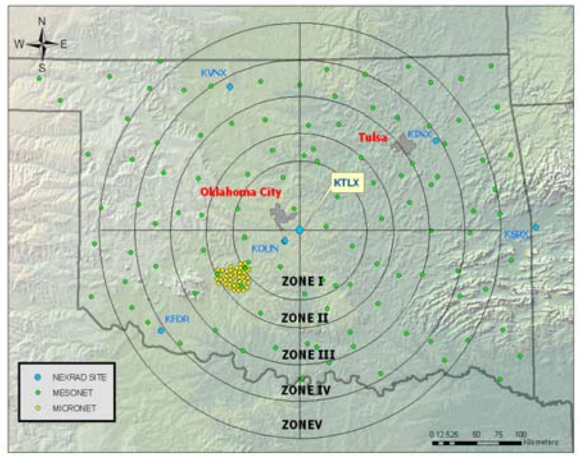

FIG.1. Locations of KTLX radar, and ARS Micronet and Oklahoma Mesonet stations. The circles show the five distance zones considered in this analysis.

We assume that, for this particular data sample, single raingauges provide sufficiently accurate approximations of the rainfall averaged over the PPS product grids (about 4 km by 4 km). It is justified because, for the time-scales considered here (hourly and longer), the spatial rainfall variability in Oklahoma is relatively small (Ciach and Krajewski 2006) and the area-point errors do not affect our results in a critical way. In Florida, for example, spatial rainfall variability is much stronger than in Oklahoma (Krajewski et al. 2003). In such regions, using single raingauges directly as the GR for the RR uncertainty modeling can yield spurious results.

The RR error model outlined below is based on a functional-statistical representation of the relationship between RR values and the corresponding true rainfall values. The concept of such a mathematical representation was used in Ciach and Krajewski (1999) in an idealized parametric model, later applied to study the possible effects of conditional biases in RR (Ciach et al. 2000). A preliminary nonparametric data analysis using this approach was performed by Ciach and Gebremichael (2004) based on a small data sample. Here, we briefly outline the selected parts of our first large-sample data analysis aimed at developing a general RR uncertainty model in a form suitable for the PQPE applications. More detailed reports can be found in Ciach et al. (2007), Villarini et al. (2007), and other forthcoming publications.

* *Corresponding author:* Grzegorz J. Ciach, University of Iowa, IIHR–Hydroscience&Engineering, 100 Stanley Hydraulics Laboratory, Iowa City, IA 52242-1585, USA; e-mail: g-ciach@uiowa.edu

2. MODELING AND ESTIMATION METHODS

We define true rainfall as the volume of rain-water falling on a specified area in a specified interval of time. Various rainfall estimates are just approximations of this physical quantity over given spatiotemporal domains. We define the uncertainties of RR as all discrepancies between the values of a RR product and the corresponding (concurrent and collocated) values of the true rainfall.

We assume that RR uncertainty can be fully described by the family of bivariate distributions of RR values, R_r, and corresponding true rainfall, R_{true}, conditioned on the major factors affecting the distributions. These factors include the spatiotemporal scale, the distance from the radar, the synoptic regime, and the specifics of a RR estimation system. The bivariate frequency distributions can be determined from large samples of RR products and corresponding raingauge data, if the raingauges can provide sufficiently accurate approximations of R_{true}. Next, we can identify a functional-statistical model of the R_r-R_{true} relationship that has well defined mathematical structure.

Lets consider RR products with specified resolution and other conditions. To describe the R_r-R_{true} relationship, we use the following functional-statistical representation:

$$R_{true} = h(R_r) \cdot e(R_r) \,, \tag{1}$$

where $h(\cdot)$ is a deterministic distortion function describing conditional biases depending on R_r, and $e(\cdot)$ is a random variable representing the combined effect of all random error sources. The distribution of the random component can also depend on R_r. In this specific representation, R_{true} is a random variable describing possible values of the true rainfall that can occur at a given value of the R_r product. Therefore, this model fits well the PQPE objectives. To identify the model based on a sample of the available R_r-R_{true} pairs, we need to estimate $h(R_r)$ and the distributions of $e(R_r)$ conditioned on R_r values. Then, the model can be applied to predict the distributions of the unknown R_{true}, given the R_r values.

Because all systematic biases can be described by the deterministic distortion component, we can assume without any loss of generality that $\mathbf{E}\{e(R_r)|R_r=r_r\}=1$ for any r_r value. For this to be true, it is sufficient that we determine $h(\cdot)$ formally as the conditional expectation function:

$$h(r_r) = \mathbf{E}\left\{R_{true} \middle| R_r = r_r\right\} , \tag{2}$$

In the conditional expectations above R_r is a random variable and r_r is its specified value, according to the commonly used statistical convention. The deterministic distortion function can be estimated using a nonparametric regression framework. The estimator used here is the kernel regression (e.g., Hardle 1990; Simonoff 1996) in the form of the following moving-window weighted averaging:

$$h(r_r) = \left. \frac{\sum_i \left(w_i R_{g,i}\right)}{\sum_i w_i} \right| \frac{r_r}{k} \le R_{r,i} \le k \cdot r_r \ , \tag{3}$$

where w_i are the weighting factors and k is a parameter that governs the size of the averaging window centered geometrically on r_r. The averaging weights, w_i, depend on the positions of the $R_{r,i}$ points within the moving window according to a parabolic function.

Once the $h(\cdot)$ function is known, we can determine the multiplicative random component as:

$$e(R_r) = \frac{R_{true}}{h(R_r)} \ . \tag{4}$$

Although its conditional mean is equal to one for each R_r=r_r, its distribution depends on r_r. The first step to identify this dependence is estimating the conditional variance, $\sigma_e^2(r_r)$=$\mathbf{Var}\{e(R_r)|R_r=r_r\}$, as a function of r_r. This is done similarly to estimating the $h(r_r)$ function. The specific nonparametric regression procedures used above are described in more detail in Ciach (2003) and Ciach et al. (2006).

To get more insight into the conditional distributions of the $e(R_r)$ component, we also estimate its conditional quantiles, q_p, at selected levels of the probability of non-exceedance, p. They can be defined formally through:

$$\mathbf{Pr}\left\{ e(R_r) \le q_p \ \middle| \ R_r = r_r \right\} = p \ , \tag{5}$$

where $\mathbf{Pr}\{\cdot|\cdot\}$ is the conditional probability and q_p depends on r_r for each p. Consistently with the $h(r_r)$ and $\sigma_e(r_r)$ functions, we estimate $q_p(r_r)$ functions using a nonparametric "weighted-point-counting" procedure (Ciach et al. 2006).

Apart from its probability distribution, another important characteristic of RR error is its spatiotemporal dependency structure. At this stage of our research, we address this aspect in a limited scope by estimating the spatial and temporal correlation functions of the $e(R_r)$ component.

3. SELECTED RESULTS

The key element of our RR uncertainty model development is the extensive exploratory data analysis based on a large sample of RR products and corresponding raingauge data. The 6-year-long sample allows us to estimate the model components for different seasons and distances from the radar. We partitioned the whole dataset into three seasons: cold (January, February, March, November, and December), warm (April, May, and October) and hot (June, July, August, and September). To capture the range effects, we divided the radar umbrella into the following five distance zones: 0-75, 70-105, 100-145, 140-185, 180-225 km.

3.1 *Deterministic Distortion Function*

The conditional expectation of the true rainfall depends on the RR magnitude. This behavior is called a conditional bias (Katz and Murphy 1997) and it is described by the deterministic distortion function in (1).

The nonparametric estimates of $h(r_r)$ were obtained at four time scales (1, 3, 6, 24 hours), for the five distance zones and the three seasons defined above. For brevity, we present in Figure 2 the estimates for the hourly scale only. However, the general shape of these functions holds also for the other time-scales.

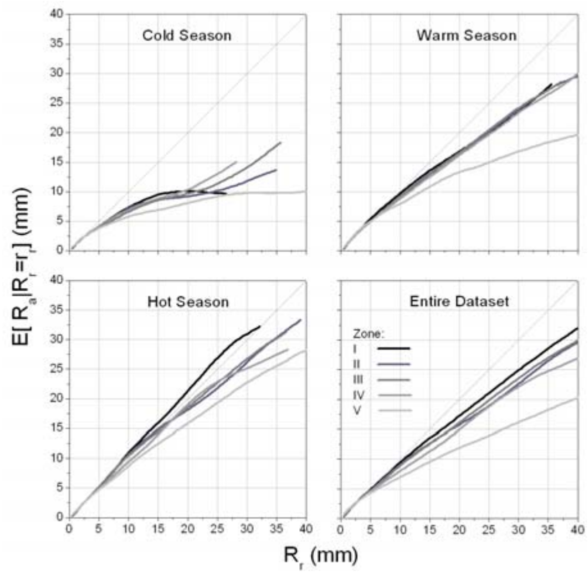

FIG. 2. The $h(r_r)$ function for three seasons and entire dataset, and in the five distance zones.

Figure 2 shows that the $h(r_r)$ curves tend to bend towards the r_r-axis for higher RR values. For the distances up to 180 km and for the warm and hot seasons, the conditional biases in Figure 1 do not show any significant range dependence, and the curves for Zones II-IV have no systematic arrangement.

3.2 *Standard Deviation of Random Component*

While the deterministic distortion function describe the systematic effects in the R_r-R_{true} relationship modeled by (1), the random component, $e(R_r)$, is a stochastic process accounting for the remaining random uncertainties. Its expectation is always equal to one, thanks to the definition of the deterministic component given by the expression (2). However, its standard deviation, $\sigma_e(r_r)$, is a function of the RR values. We estimated this function for different seasons, distance zones, and time-scales using the nonparametric regression estimator analogical to (3). The results for the 1-hour accumulation interval are presented in Figure 3.

In general, all the nonparametric estimates of the $\sigma_e(r_r)$ function exhibit a hyperbolic behavior growing to infinity for r_r closing to zero, and decreasing to a constant asymptotic level for growing RR values. For the warm and hot seasons, the distance dependences are as expected: $\sigma_e(r_r)$ becomes larger as the distance from the radar increases. An exception from this regular behavior is the cold season, where any clear distance pattern is not distinguishable.

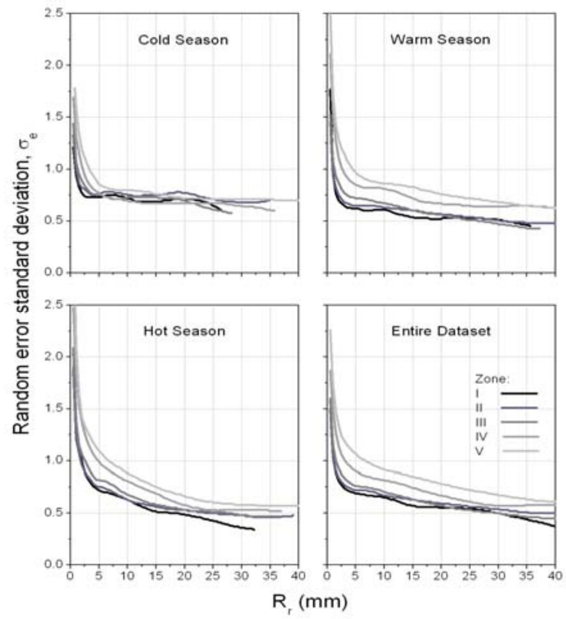

FIG. 3. The $\sigma_e(r_r)$ function for three seasons and entire dataset, and in the five distance zones.

3.3 *Effect of different Z-R Relationships*

Both the deterministic and random component in the RR uncertainty model (1) depend strongly on the Z-R relationship selected in the PPS processing of the raw reflectivity data. These effects are presented in Figures 4 and 5. They were computed for the Zone II only because it contains the largest number of raingauges.

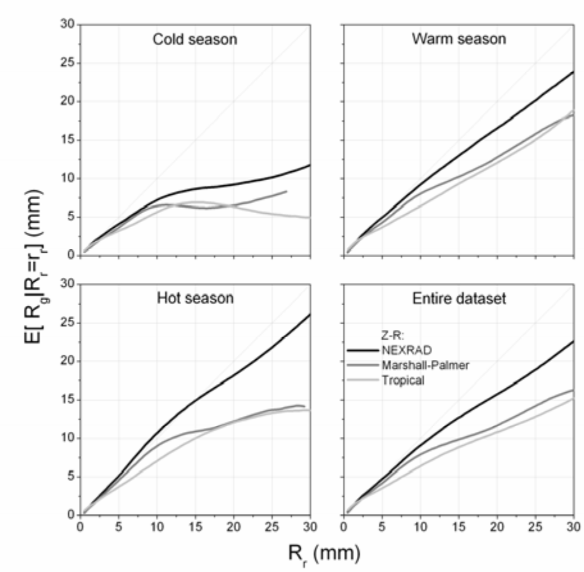

FIG. 4. The dependence of the $h(r_r)$ function on the Z-R relationship, for three seasons and entire dataset.

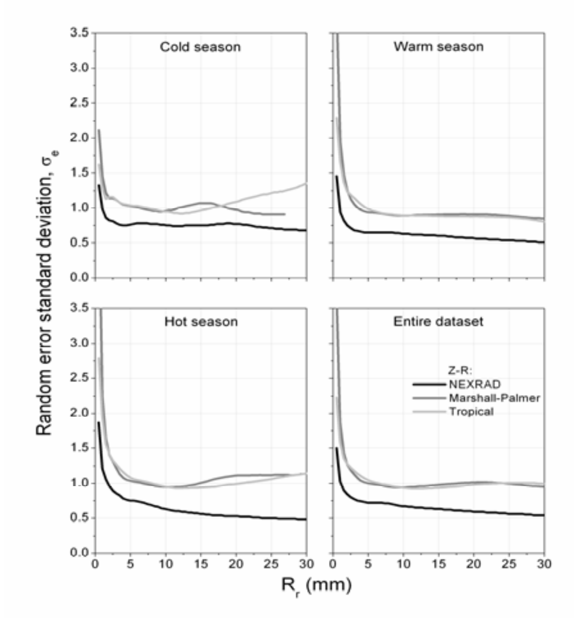

FIG. 5. The dependence of the $\sigma_e(r_r)$ function on the Z-R relationship, for three seasons and entire dataset.

The RR error model components for three commonly used power-law Z-R parameter sets are shown in these figures: the standard NEXRAD (*A=300, b=1.4*) discussed in the previous sections, as well as the Marshall-Palmer (*A=200, b=1.6*) and the "tropical" (*A=250, b=1.2*) relationships.

The surprising feature of these results is that both the Marshal-Palmer and the "tropical" Z-R conversion functions results in quite similar departures from the standard NEXRAD relationship. For the deterministic component in Figure 5, the change of the conditional bias in the same direction is counterintuitive. Based on our previous analytical studies (Ciach and Krajewski 1999, Ciach at al. 2000), we expected the $h(r_r)$ function for *b=1.2* to be above the standard NEXRAD curve in the region of moderate and strong RR values. The fact that the change is in the opposite direction is difficult to explain. One possible cause could be a dramatic increase of the uncertainty level in the PPS products for any departures from the standard Z-R relationship. In our opinion, this might indicate some inconsistencies in the PPS algorithm in respect to its parameter selection. This suspicion is supported by the much higher levels of the standard deviations of the random component for the Z-R's different from the standard (Figure 5). Based on our early study by Ciach et. al. (1997), we expected the level of the $\sigma_e(r_r)$ function to be lower for the Marshal-Palmer than for the standard Z-R relationship. More detailed investigation is needed to resolve this new "cognitive dissonance."

3.4 *Conditional Distributions of Random Component*

The conditional distributions of the random component, $e(R_r)$, can be described by the conditional quantile functions, $q_p(r_r)$, defined by (5) for any non-exceedance probability threshold. For graphical illustration we use five levels of this threshold (10%, 25%, 50%, 75% and 90%). An example of the results is shown in Figure 6.

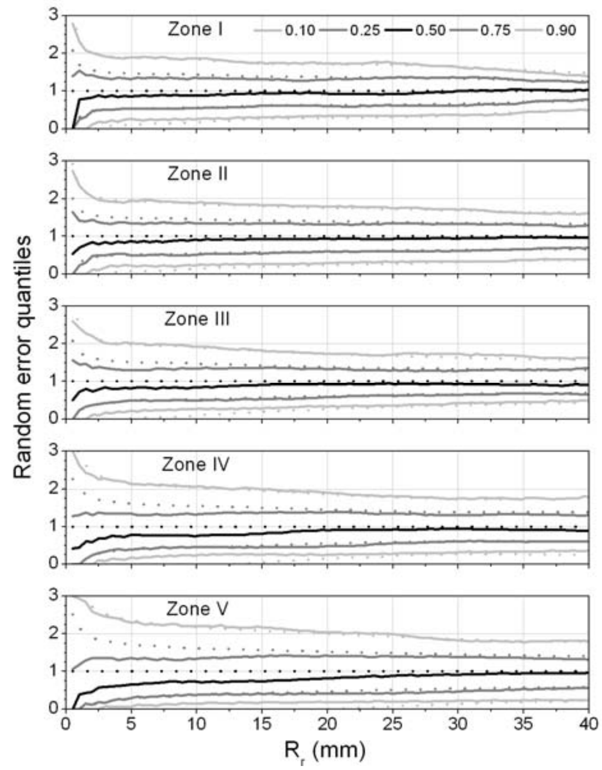

FIG. 6. The $q_p(r_r)$ functions for the entire dataset for five levels of non-exceedance probability, in five distance zones. The dotted curves are based on the empirical estimates, and the continuous curves are their approximations based on the Gaussian model.

These results are for the standard NEXRAD Z-R relationship and all three seasons ("entire dataset"). The data-based estimates are compared with their analytical approximations based on the normal distribution model. The mean of this distribution is set to one, whereas its conditional standard deviations are based on the empirical estimates in Figure 3. In Zones I-III, the accuracy of this simple approximation is quite good especially for moderate and strong RR values, which are most significant in hydrology. In Zones IV-V, the departures can be attributed to a change in the RR errors at large distances. But these samples are too small to make a conclusive inference on it.

4. PQPE APPLICATIONS

As mentioned before, the functional-statistical model of RR uncertainties described here can be used for the probabilistic quantitative precipitation estimation (PQPE) based on WSR-88D data. In the specific representation given by (1), the probable true rainfall corresponding to the observed RR is a random variable. Its distribution depends in a specific way on the given values of the RR product. This model not only accounts for the presence of conditional and unconditional biases, but also characterizes the random error component in terms of its statistical distribution and its spatiotemporal correlations (nor shown here). Thus, the PQPE objectives can be met when the two components of the RR error model (1) are known.

A description of two specific PQPE applications of the RR error model can be found in Villarini et al. (2007). The first is the conditional simulation of spatiotemporal rainfall ensembles, and the second is the estimation of true rainfall probability maps. In the former, a user can generate ensembles of probable true rainfall fields that are consistent with the uncertainty structure of given RR products. The simulation of the random component in this application is based on the Cholesky decomposition method. The method provides the flexibility to account for the spatiotemporal correlations in the random component, as well as for its dependence on the given RR. In the latter application, the static maps showing the exceedance probability of selected rainfall thresholds are produced, given a RR map and the information on its uncertainty. They provide the answer to the question typical in operational hydrology: "What is the probability that, for the RR values observed over an area, the corresponding actual rainfall values exceed a specified threshold?" In both of the above PQPE applications, the Gaussian distributions of the random component have to be truncated from below to avoid negative values.

5. CONCLUSIONS

We presented a RR uncertainty model that can be applied to the PQPE based on weather radar data. Its components are estimated based on a six-year-long sample of RR products and the GR based on the corresponding raingauge data. Up to date, the main findings in this prospective research (only small part of our results could be shown here) can be enumerated as follows:

1. The RR uncertainty model (1) is a two-component functional-statistical representation of the relationship between RR and the corresponding true. Its form is suitable for the PQPE applications.
2. The RR products contain considerable conditional biases (systematic distortions dependent on RR values) that have to be quantified in any complete RR error model.
3. The standard deviation of the random component in (1) is a decreasing function of RR that is converging to a constant non-zero asymptote for large RR values.
4. The two components of (1) can be parameterized using analytical approximations based on power-law functions.
5. The conditional distributions of the random component can be approximated with the Gaussian model.
6. The random component is correlated in time and space.

The fact that the random component in (1) is nearly Gaussian-distributed simplifies its tractability. Concerning the spatiotemporal correlations in the random component, the estimates obtained so far are fairly inaccurate. In our future research, we need to find better methods to describe the dependences in the RR errors. Although we believe that the general structure of our model is transferable to different RR products and radar sites, the specific effects of such changes need to be further investigated based on adequate data samples.

Acknowledgments. This work was supported by the NSF grant EAR-0309644 and the contract under the PQPE initiative by the National Weather Service. The opinions expressed here are of the authors and do not necessarily reflect those of NSF or NOAA or their sub-agencies.

REFERENCES

Ciach, G.J., W.F. Krajewski, E.N. Anagnostou, J.A. Smith, M.L. Baeck, J.R. McCollum, and A. Kruger, 1997: Radar rainfall estimation for Ground Validation studies of the Tropical Rainfall Measuring Mission. J. Appl. Meteor., 36, 735-47.

Ciach, G.J., and W.F. Krajewski, 1999: Radar-rain gauge comparisons under observational uncertainties. *Journal of Applied Meteorology*, **38**, 1519-1525.

Ciach, G.J., M.L. Morrissey, and W.F. Krajewski, 2000: Conditional bias in radar rainfall estimation. *Journal of Applied Meteorology*, **39**(11), 1941-46.

Ciach, G.J., 2003: Local random errors in tipping-bucket raingauge measurements, *Journal of Atmospheric and Oceanic Technology*, **20**, 752-759.

Ciach, G.J., and M.V. Gebremichael, 2004: Empirical modeling of the uncertainties in radar rainfall estimates. Preprints: 6th Int. Symp. on Hydrol. Appl. of Weather Radar, Melbourne, Australia, February 2-4, 2004.

Ciach, G.J., and W.F. Krajewski, 2006: Analysis and modeling of spatial correlation structure of small-scale rainfall in Central Oklahoma. *Adv. Water Resour.*, **29**, 1450-1463.

Ciach, G.J., W.F. Krajewski, and G. Villarini, 2007: Product-Error-Driven Uncertainty Model for Probabilistic Quantitative Precipitation Estimation with NEXRAD Data. *J. Hydrometeorology*, (in print).

Fulton, R.A., J.P. Breidenbach, D.J. Seo, and D.A. Miller, 1998: WSR-88D rainfall algorithm. *Weather and Forecasting*, **13**, 377-395.

Habib, E., G.J. Ciach, and W.F. Krajewski, 2004: A method for filtering out raingauge representativeness errors from the verification distributions of radar and raingauge rainfall *Adv. Water Resour.*, **27**(10), 967–98.

Hardle, W., 1990: *Applied Nonparametric regression*. Cambridge University Press, 333 pp.

Katz, R.W., and A.H. Murphy (Editors), 1997: *Economic Value of Weather and Climate Forecasts.* Cambridge University Press, 222 pp.

Krajewski, W.F., G.J. Ciach, and E. Habib, 2003: An analysis of small-scale rainfall variability in different climatic regimes. *Hydrol. Sci. J.*, **48**, 151-162.

Simonoff, J.S., 1996: *Smoothing Methods in Statistics*, Springer-Verlag, 338 pp.

Villarini, G., W.F. Krajewski, and G.J. Ciach, 2007: Product-Error-Driven Ensemble Generator of Spatiotemporal Rainfall Conditioned on WSR-88D Precipitation Estimates and its Application in Hydrologic Forecasting. (in preparation).