Generalized Linear Models for Site-Specific Density Forecasting of U.K. Daily Rainfall

MAX A. LITTLE AND PATRICK E. MCSHARRY

Department of Engineering Science, University of Oxford, Oxford, United Kingdom

JAMES W. TAYLOR

Saïd Business School, University of Oxford, Oxford, United Kingdom

(Manuscript received 1 April 2008, in final form 30 June 2008)

ABSTRACT

Site-specific probability density rainfall forecasts are needed to price insurance premiums, contracts, and other financial products based on precipitation. The spatiotemporal correlations in U.K. daily rainfall amounts over the Thames Valley are investigated and statistical Markov chain generalized linear models (Markov GLM) of rainfall are constructed. The authors compare point and density forecasts of total rainfall amounts, and forecasts of probability of occurrence of rain from these models and from other proposed density models, including persistence, statistical climatology, Markov chain, unconditional gamma and exponential mixture models, and density forecasts from GLM regression postprocessed NCEP numerical ensembles, at up to 45-day forecast horizons. The Markov GLMs and GLM processed ensembles produced skillful 1-day-ahead and short-term point forecasts. Diagnostic checks show all models are well calibrated, but GLMs perform best under the continuous-ranked probability score. For lead times of greater than 1 day, no models were better than the GLM processed ensembles at forecasting occurrence probability. Of all models, the ensembles are best able to account for the serial correlations in rainfall amounts. In conclusion, GLMs for future site-specific density forecasting are recommended. Investigations explain this conclusion in terms of the interaction between the autocorrelation properties of the data and the structure of the models tested.

1. Introduction

The source of most atmospheric rainwater is the sea, with rain forming when large droplets eventually become heavy enough to fall to the ground. Rainfall over land eventually flows back to the sea, completing the cycle (Brutsaert 2005). Water is vital to life, but also immensely destructive: understanding the movement of atmospheric water is critical. Forecasting rainfall is therefore important in many disciplines, for example, economics and finance, hydrology, meteorology, ecology, agriculture, and renewable energy.

The U.K. climate is temperate and strongly influenced by the oceans, with cool summers and mild winters (Barry and Chorley 2003). Rainfall forecasting has taken on new urgency in the United Kingdom due to recent flooding caused by extreme rainfall: evidence exists that such extremes may increase in frequency with global temperature increases (Easterling et al. 2000). The Thames River runs directly through London and several major midland and southern towns. Flooding on this river has significant costs to the U.K. economy, and insurance premiums have increased substantially because of these recent severe events.

Rainfall forecasts can therefore help quantify the risk of floods and droughts with which to price products such as flood and crop insurance, weather derivatives, and other commodities (Cao et al. 2004; Diebold et al. 1998; Taylor and Buizza 2006). All forecasts have errors because of the combined uncertainty in observational data and model structure. Comprehensive quantification of this forecast uncertainty is critical to risk assessment, motivating interest in density forecasting producing a distribution over all possible future rainfall events, rather than a single-point forecast misrepresenting these

Corresponding author address: Max Little, Department of Engineering Science, University of Oxford, Parks Road, Oxford OX1 3PJ, United Kingdom.
E-mail: max.little@eng.ox.ac.uk

DOI: 10.1175/2008MWR2614.1

© 2009 American Meteorological Society

uncertainties. Density forecasts are particularly flexible, allowing the calculation of the probability of any event of interest, such as the probability of occurrence of rain or of extreme rainfall above any threshold. Complementing density forecast comparisons between models, we can also issue the density median as a point forecast.

Numerical weather predictions (NWPs) are highly complex, nonlinear systems producing a single or a set (ensemble) of point forecasts, allowing the anticipation of distinct meteorological events. Statistical time series models are mathematically simple and produce full-density forecasts capturing statistical properties of the data. NWP now routinely outperforms purely statistical methods for medium-range (1 day to 1 week ahead) operational forecasts, but for very short term (a few hours) and very long term (greater than 10 days ahead), statistical approaches remain competitive (Wilks 2006).

Density prediction for product pricing requires accurate forecasts at specific locations, for a wide range of forecast lead times, for which unified approaches to site-specific density forecasts seamlessly covering short- to long-range time scales are needed. For product pricing applications, medium-range global forecasts are currently more useful than short-range regional forecasts, because short-range forecasts do not have sufficiently long forecast horizons.

Because the primary application is the precision quantification of the probability of rain rather than the anticipation of particular meteorological events, dynamical and statistical forecasting, although fundamentally different in approach, can be combined to produce accurate, site-specific density forecasts. Statistical postprocessing (ensemble calibration) (Applequist et al. 2002; Wilks 2006) is one such combination approach.

Precipitation measurement is mostly by ground-based rain gauge measurement of total rainfall depth (Upton et al. 2005). Strong evidence exists that total rainfall amount distributions are discontinuous at zero depth (no rainfall) motivating separate modeling of occurrence (rain/no rain) and intensity (nonzero amount; Cao et al. 2004; Grunwald and Jones 2000; Wilks 1998). These separate models are combined in a mixture density of daily rainfall totals. Similarly, rainfall is nonnegative and non-Gaussian. A wide range of proposed statistical rainfall probability models could be used and it is instructive to test as many as possible. This includes powerful generalized linear models (GLMs; Grunwald and Jones 2000) that allow flexible, nonlinear, non-Gaussian regression, but any new method must demonstrate performance superiority over existing, simpler approaches before being considered successful.

To build time series density forecast models, we explore spatiotemporal correlation and seasonality properties that could be captured and hence exploited. Here we analyze the Thames Valley time series and construct Markov chain GLMs incorporating information from neighboring gauges and past time steps (Grunwald and Jones 2000). We test these models against simple benchmarks, and ensemble NWP forecasts postprocessed with GLMs.

Our main contribution here is to conduct direct tests of a range of methods proposed in the existing literature for producing site-specific, full-density forecasts against novel GLM methods, and hence to suggest an improved NWP postprocessing method. Our investigations explain the performance of these different methods in terms of the autocorrelation properties of the data and the structure of each model, for full-density and point forecasts.

The paper is organized as follows. Section 2 reviews the current state of rainfall measurement, modeling, and forecasting. Section 3 describes the data used and details the correlation structure analysis. Section 4 describes model construction and forecast performance comparison methods used in this study. Section 5 discusses the results of the forecast comparison, and finally section 6 summarizes the paper and concludes with the relevance of these results for future site-specific rainfall forecasting.

2. Review of rainfall measurement and forecasting methods

Ground-based rain gauges capture precipitation, recording the total amount as rainfall depth, usually in millimeters (Upton et al. 2005). The temporal measurement resolution can be high, but accuracy can be site dependent and unreliable in extreme weather and can include melted snow or hail in addition to rain. Radar measurements, by contrast, detect low-altitude atmospheric water content and have excellent spatiotemporal resolution, but current geographic coverage is restrictive and the historical record is short (Upton et al. 2005).

Rainfall forecasting models depend on the application. Thunderstorms implicated in flash floods typically take place on scales of minutes to hours (Battan 1984), requiring forecasts on the shortest time scales. Localized flooding often occurs when medium to heavy rain falls in the same location over several days, inundating rivers and urban drains, requiring forecasts from hours to days. Predicting droughts requires forecasts on longer time scales of weeks to months.

NWP solves equations of atmospheric dynamics and produces rainfall predictions. Calibrated against atmospheric measurements, they vary in spatial scale from synoptic (on the order of 1000 km) to mesoscale

(approximately 50 km); current operational models have minimum resolutions of approximately 1.3 km (limited-area mesoscale), forecasting days to a few weeks ahead (Buizza 2003). Sophisticated NWP systems generate an ensemble of predictions by varying the model initial conditions and/or by varying physical parameterization schemes. The frequency distribution of the ensembles estimates the probability density (Buizza 2003; Buizza et al. 1998; Molteni et al. 1996; Palmer et al. 1993). The high computational complexity of running many different parallel forecasts limits most operational NWP systems to single-point forecasts or low spatial resolution ensembles, and important precipitation sources such as isolated thunderstorms and small-scale details are not well resolved.

Classical statistical forecasting identifies relationships between past observations and their temporal successors, using observations at the current forecast origin as predictors for the future state of the atmosphere based solely on these relationships and not on explicit meteorological information (Wilks 2006). Methods include conditional climatology (issuing successors of past observational data closest to the current state as a forecast for the future state), and applications of more sophisticated multiple nonlinear regression such as neural networks (Moura and Hastenrath 2004). Included in this category are statistical time series models that can forecast at the spatiotemporal resolution of rainfall measurements, and are univariate or multivariate (comprising a vector of rainfall measurements from a number of sites simultaneously), producing density forecasts. They are either temporally unconditional or conditional on past time steps.

Statistical postprocessing methods such as the analog method or model output statistics (Applequist et al. 2002; John 2003; Wilks 2006) determine the statistical relationships between forecast NWP variables and actual observations, acting as postprocessors to improve NWP forecasts.

Unconditional daily rainfall models are commonly split into occurrence and intensity (Cao et al. 2004; Grunwald and Jones 2000; Wilks 1998), with the occurrence often modeled as a Bernoulli random variable (Grunwald and Jones 2000). Intensity models usually use exponential family distributions, for example, gamma densities (Grunwald and Jones 2000; Hyndman and Grunwald 2000), exponential mixtures (Cao et al. 2004; Wilks 1998), or truncated normals (Sanso and Guenni 1999). In nonparametric methods, kernel density techniques can model intensity (Cao et al. 2004).

For conditional occurrence models, first-order Markov chains are common (Cao et al. 2004; Wilks 1998). For conditional intensity, generalized linear and generalized additive Markov chain regression models (GLM/GAM) have been used (Grunwald and Jones 2000; Hyndman and Grunwald 2000).

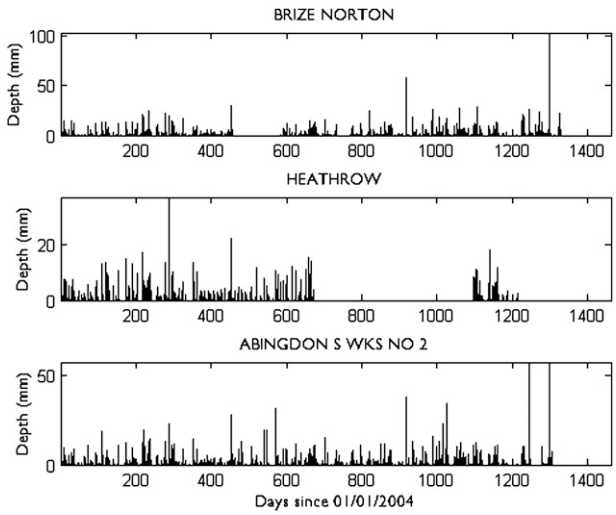

FIG. 1. Three selected rainfall time series from the Thames Valley catchment. Vertical axis is rainfall depth (mm) and horizontal axis is the number of days since 1 Jan 2004.

3. Data

The data comprises daily rainfall depth measurements from all 295 Met Office Integrated Data Archive System (MIDAS) WADRAIN rain gauges in the Thames Valley, United Kingdom, within the square grid 51°–52.5°N, −2°–0.5° E, covering an area approximately 400 km by 400 km. Observations cover the time period from 2004 to late September 2007. Figure 1 shows selected examples of the rainfall time series.

Spatiotemporal correlation structure for all 295 gauges is tested. However, only a few of the sites have sufficiently complete records overlapping the available NWP ensemble forecasts, so that, in forecast comparisons, a much smaller subset (10 sites) of this data was selected (Fig. 2). These sites are chosen as a compromise between minimizing the number of missing observations, economic relevance (Heathrow in London), and hydrological interest (Brize Norton received some of the highest rainfall totals during the recent flooding). Table 1 lists the location and number of rainfall observations available for these selected sites (the inset in Fig. 5 shows their physical layout). All gauges have missing measurements and consistent procedures, described below for each forecast model, ensure fair comparisons.

In this section we explore the spatiotemporal correlations in the data. We denote the total rainfall amount on day $n = 1, 2, \ldots, N$ for each site $m = 1, 2, \ldots, M$ as x_n^m, where N is the maximum length in days of the time series, and $M = 10$ is the number of selected sites

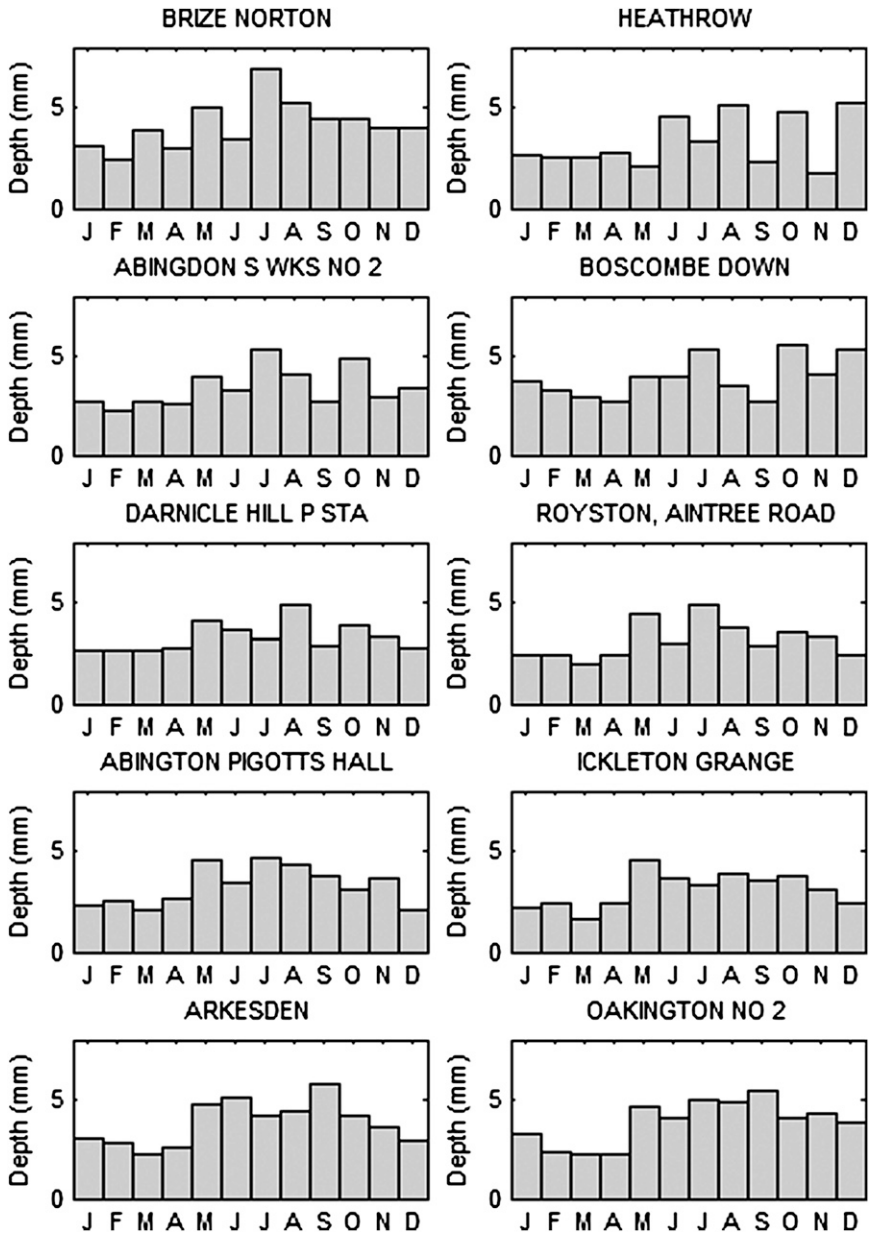

FIG. 2. Monthly average rainfall depth on rainy days for the 10 gauges selected for the modeling part of the study (see Table 1). The horizontal axis is the month and the vertical axis is the average rainfall depth.

for modeling. We denote occurrence with the indicator variable q_n^m, which is 0 for dry days and 1 for days where rainfall x_n^m is nonzero.

We first calculate the autocorrelation function, and the standard Bartlett 95% confidence intervals. Although the autocorrelation function tests the dependence up to second-order statistical moments, such highly non-Gaussian rainfall data could have nonzero higher-order moments. A more general test of independence at different time lags is the time-delayed mutual information (TDMI; Kantz and Schreiber 2004; Little et al. 2006):

$$I(\tau) = \int_0^\infty \int_0^\infty P(x_n^m, x_{n+\tau}^m) \log \frac{P(x_n^m, x_{n+\tau}^m)}{P(x_n^m) P(x_{n+\tau}^m)} dx_n dx_{n+\tau}. \quad (1)$$

Marginal densities $P(x_n^m), P(x_{n+\tau}^m)$ and joint densities $P(x_n^m, x_{n+\tau}^m)$ are estimated using histograms of x_n^m. We assume that x_n^m are weakly stationary stochastic processes; then only the relative time lag τ is important and

TABLE 1. Selected Thames Valley catchment rain gauge stations used in the modeling part of the study.

Rain gauge No.	Rain gauge station name	Lat (fractional °N)	Lon (fractional °E)	Height above sea level (m)	Available observations in days (missing)	Percentage of dry days (%)	Avg rainfall depth on rainy days (mm)
1	Brize Norton	51.76°	−1.58°	81	1168 (293)	48.9	4.1
2	Heathrow	51.48°	−0.45°	25	791 (670)	51.7	3.1
3	Abingdon Sewerage Works No. 2	51.65°	−1.29°	50	1280 (181)	50.4	3.4
4	Boscombe Down	51.16°	−1.75°	126	1222 (239)	47.9	3.9
5	Darnicle Hill Pumping Station	51.73°	−0.10°	73	1186 (275)	46.5	3.2
6	Royston Aintree Road	52.05°	−0.01°	78	1309 (152)	46.2	3.1
7	Abington Pigotts Hall	52.08°	−0.10°	30	1309 (152)	50.0	3.2
8	Ickleton Grange	52.06°	0.13°	76	1309 (152)	48.2	3.0
9	Arkesden	51.98°	0.15°	114	1309 (152)	50.0	3.7
10	Oakington No. 2	52.26°	0.07°	12	1309 (152)	56.5	3.8

we can assume that $P(x_n^m)$, $P(x_{n+\tau}^m)$ are the same. The joint density is estimated using histograms formed by counting the number of times that x_n^m falls into the same histogram bin as $x_{n+\tau}^m$. The integrals are approximated using summations. TDMI significance tests use the null hypothesis of no mutual information using bootstrap independent, identically distributed (i.i.d.) time series, generated by randomly permuting individual days' measurements and destroying any original temporal ordering. The TDMI for each bootstrap is compared with that of the original series to try to reject the null hypothesis at each time lag. If, for a significance probability of $\alpha = 0.05$, on generating $2/\alpha - 1 = 39$ bootstraps, the TDMI at any time lag of the original data is either the smallest or the largest value among all the bootstraps, then we can reject the null hypothesis at that time lag.

Although using histograms to estimate densities is simplistic, we are only interested in the TDMI relative to the i.i.d. bootstraps, and, as such, these inaccuracies are relatively unimportant to the question of whether significant nonlinear/non-Gaussian temporal dependence exists in the time series.

For the spatial correlation analysis of total rainfall amount, the pairwise correlation coefficients for all 295 locations are plotted against physical distance between sites. The no-correlation null hypothesis uses the standard asymptotic normal distribution of Fisher's z transformation of the correlation coefficient to estimate the p value, tested at 5% significance. Finally, for spatial correlations in occurrence, we calculate the pairwise conditional probability of nonoccurrence of rain for all 295 locations. We take each pair of locations, calculate the *conditional* probability of nonoccurrence at one location, given that rainfall did/did not occur at the other, and plot these probabilities against physical distance between locations. We can then see the contribution of occurrence of rainfall to the spatial correlation of the total amount.

Turning to the results of this correlation analysis, Table 1 shows that, any one day taken at random, is as likely to be dry as wet. This is to be expected for this generally temperate climate, and the average rainfall intensity is small. Regarding spatial correlation, from Fig. 3, all locations show similar rainfall patterns, and the maximum correlation falls slowly with increasing distance. At any given distance, there is an apparent maximum and minimum correlation between gauges. Figure 4 shows that this effect is even stronger for the nonoccurrence. This is confirmation that most rainfall events are on the meso-alpha scale: resulting from widespread cloud cover due to warm fronts or convective complexes spread over hundreds of square kilometers. Similarly, if dry at one location, then it is highly likely to be dry at all the other locations in the catchment. This physical effect provides some justification for modeling techniques that try to capture catchment-wide spatial cross correlations (Fig. 5).

Regarding temporal correlation, Fig. 6 shows autocorrelation decaying extremely rapidly, and although significant up to 4 days ahead, the TDMI in Fig. 7 shows a slightly different story with significant mutual information to only 3 days. Similar results were obtained for the other selected gauges. The results show that the rapid decay of autocorrelation or TDMI is not just due to linearity limitations of the autocorrelation function; weather systems move rapidly across the United Kingdom and normally dissipate within a few hours. However, the lack of obvious annual periodicity is in need of explanation. Because of lowering temperatures, all other things being equal, lower saturation vapor causes higher average and maximum U.K. rainfall totals during winter (Brutsaert 2005) and increasing frequency of extremes in early autumn. Figure 2 shows some seasonal variation in the rainfall intensity for each month, but this variation is very slight. This slight seasonality is not the strict repetitiveness detectable by autocorrelation

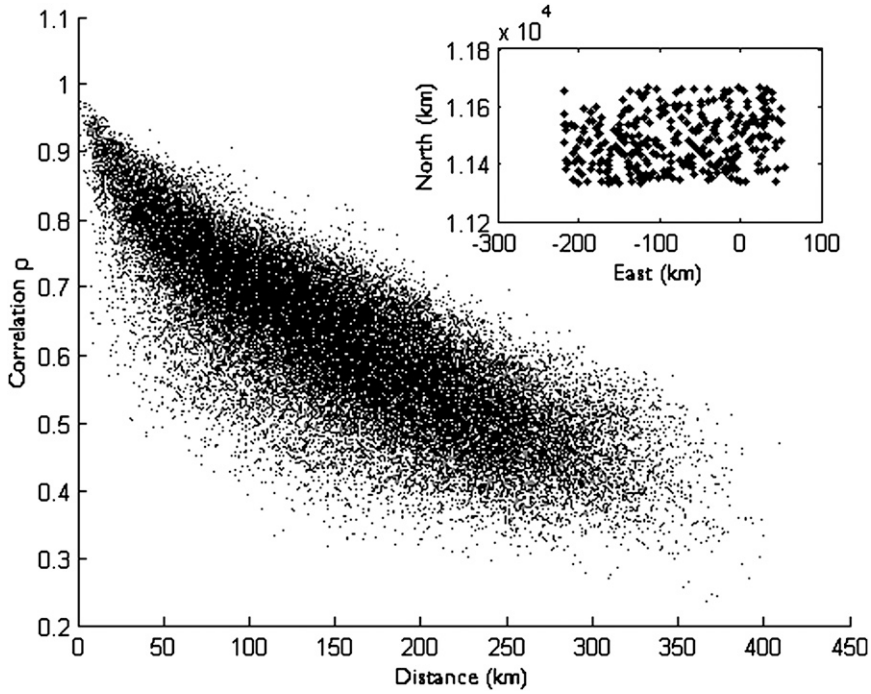

FIG. 3. Pairwise correlation coefficients and (inset) position of all 295 gauging stations in the Thames Valley. For the inset, the horizontal axis is the horizontal location (kilometers east of Greenwich), and the vertical axis is the vertical location (kilometers north of the equator).

and TDMI—exact annual pattern start/end days are ill-defined, varying from year to year. For statistical modeling, rapid exponential decay of mutual information justifies models with very short memory, and regressing on past rainfall at annual time lags is unlikely to provide significant model improvements, particularly for the short lead times tested here. Instead, regressing on a seasonal variable should allow exploitation of these slight seasonal variations.

Taking these cross-correlation and temporal correlation results together, there may be some time-delayed cross correlation that could be captured by multisite models (that use past information from many nearby or distant sites). Nonetheless spatial correlation at time lag zero dominates over temporal correlation at any time lag.

4. Methods

This section details spatiotemporal statistical models of rainfall totals at the 10 selected Thames Valley locations, and the forecast performance comparisons of these models against unsophisticated benchmarks and postprocessed ensemble NWP.

a. Forecasting models

The time series models for daily rainfall compared in this paper can be grouped into simple nonparametric benchmarks and sophisticated parametric/numerical methods. For all the models, any missing rainfall observations for each site are ignored in the model parameter estimation and forecast comparisons, such that comparisons are only made on days for which forecasts from all models are available. Parameters are estimated using observations in the years 2004 and 2005. Model performance is tested on a hold-out sample of the years 2006 and 2007. There are two simple benchmarks used in this study:

1) *Persistence forecast*: This is just the rainfall total of the day prior to the forecast origin. The persistence is not a density forecast, only a point forecast. This forecast is a baseline to assess the point forecast performance of the more sophisticated models.
2) *Climatology forecast*: This is the unconditional empirical density of rainfall amount on each day. Any missing observations in the estimation period are excluded from the empirical cumulative density [which is used to calculate the z-series probability integral transform (PIT) histogram, see below].

In addition, there are seven models of increasing sophistication used:

1) *Unconditional gamma/Bernoulli density*: The shape and scale parameters for the gamma intensity model

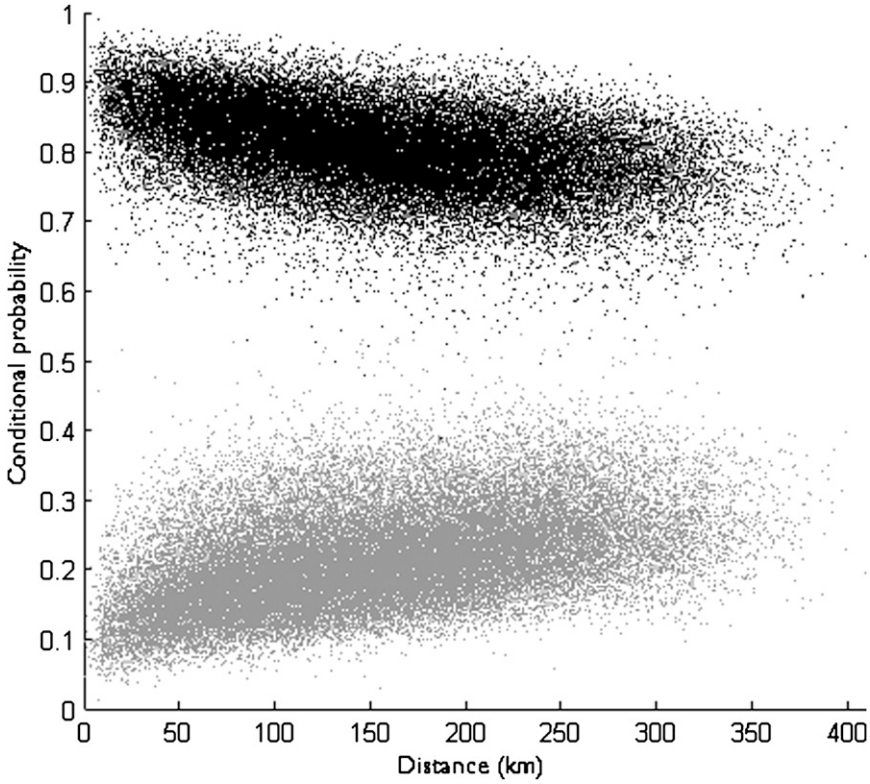

FIG. 4. Variation of the conditional probability of nonoccurrence of rainfall against the pairwise distance between all locations, given nonoccurrence/occurrence at the other location. The black dots show the probability of nonoccurrence of rainfall at a location, given that rainfall did not occur at the other location. Gray dots show the probability of nonoccurrence of rainfall at a location, given that rainfall *did* occur at the other location. The horizontal axis is the distance between locations (km) and the vertical axis is the conditional probability.

were estimated using the maximum likelihood method, with a single parameter Bernoulli model for occurrence. Any missing observations in the estimation samples are excluded from the gamma and Bernoulli parameter estimation. The combined intensity and occurrence model (a Bernoulli–gamma mixture density) is constructed and used to make forecasts.

2) *Cao–Li–Wei model* (Cao et al. 2004): An unconditional mixture of two exponential densities was fitted to the intensity of the estimation samples:

$$f(x_{n+1}^m) = \frac{d}{g_1}\exp\left(-\frac{x_{n+1}^m}{g_1}\right) + \frac{1-d}{g_2}\exp\left(-\frac{x_{n+1}^m}{g_2}\right). \quad (2)$$

The parameters d, g_1, and g_2 were found using an iterative maximum likelihood method (Agha and Ibrahim 1984). For occurrence, this model fits a first-order Markov chain to the occurrence in the estimation samples by estimating the transition density matrix from counts (Cao et al. 2004). Missing observations are handled as per the above gamma model. The combined intensity and occurrence mixture model is constructed and used to make forecasts.

3) *Generalized linear Markov model (Markov GLM)*: This model is described in (Grunwald and Jones 2000). The Markov transition density has the following form:

$$f(x_{n+1}^m | x_n^m) = [1 - p(x_n^m)]\delta_0(x_{n+1}^m) + p(x_n^m)\mathrm{Gamma}(x_{n+1}^m | x_n^m), \quad (3)$$

where δ_0 is the Dirac delta function. The transition density for the intensity of the estimation samples is a conditional gamma generalized linear model with log link function, with conditional mean μ:

$$\log[\mu(x_n^m)] = b_0 + b_1 \log(x_n^m + c). \quad (4)$$

The constant shape parameter for this gamma density is estimated using the maximum likelihood method (Venables and Ripley 2002). To improve the model fit, the logarithm of the past rainfall amount with a small, additive constant c is used instead of the

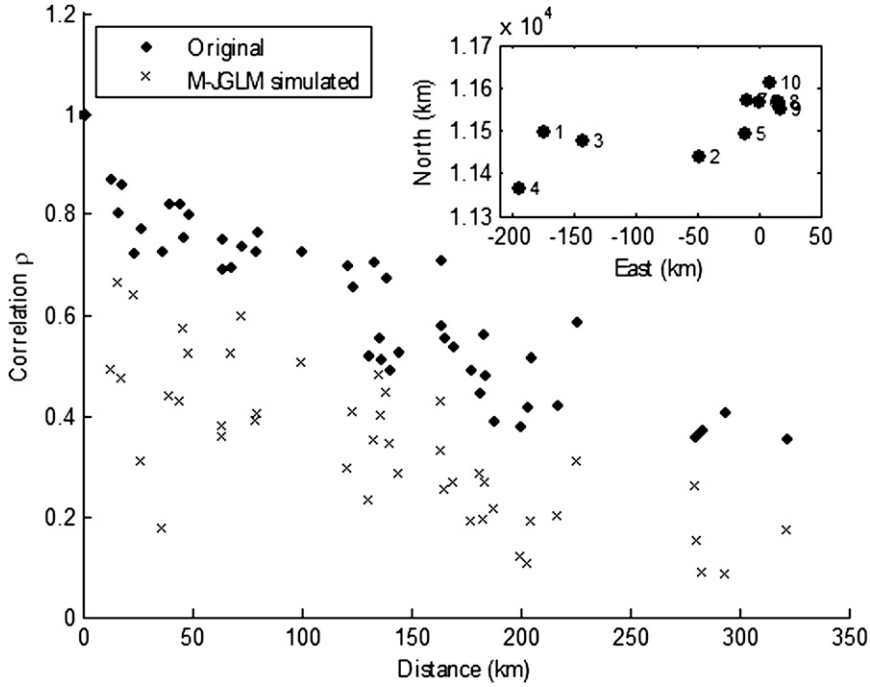

FIG. 5. Pairwise correlation coefficient of selected gauges used in the modeling against distance. Also shown are the simulated Markov JGLM rainfall correlation coefficients for the same gauges. (Inset) The position of modeled gauges (refer to Table 1 for the gauge numbering), the horizontal axis is the horizontal location (km east of Greenwich), and the vertical axis is the vertical location (km north of the equator).

actual rainfall depth. Similarly, the conditional Bernoulli density for the occurrence $p(x_n^m)$ is the following generalized linear model using the inverse log link function l:

$$p(x_n^m) = l[a_0 + a_1 \log(x_n^m + c)],\ l(u) = \frac{\exp(u)}{1+\exp(u)}. \quad (5)$$

4) *Joint generalized linear Markov model (Markov JGLM)*: This model has a similar structure to the previous model, except that the joint distribution of each site is captured by sequentially conditioning on adjacent gauges. This exploits the chain rule for probabilities that relates the joint probability of all gauges to the conditional probabilities:

$$P(x^1, x^2, \ldots, x^M) = P(x^1)P(x^2|x^1)P(x^3|x^2, x^1)\cdots \times P(x^M|x^{M-1}, x^{M-2}, \ldots, x^1). \quad (6)$$

Thus, it is possible to reproduce the entire joint density by sequentially modeling the conditionals (i.e., first modeling the marginal density of the first time series, followed by the second conditional on the first, followed by the third conditional on the second and first and so on). The Markov transition density of this model is

$$f(x_{n+1}^m|\mathbf{x}_n^m) = [1 - p(\mathbf{x}_n^m)]\delta_0(x_{n+1}^m) + p(\mathbf{x}_n^m)\mathrm{Gamma}(x_{n+1}^m|\mathbf{x}_n^m), \quad (7)$$

and for this model, the vector $\mathbf{x}_n^m = (x_n^m, \hat{x}_{n+1}^{m-1}, \ldots, \hat{x}_{n+1}^1)^{\mathrm{T}}$ contains the past rainfall of the time series m, and the forecast rainfall $\hat{x}_{n+1}^{m-1}, \ldots, \hat{x}_{n+1}^1$, produced sequentially by this model, of the $m-1$ adjacent time series, in the following way: To produce the forecast $\hat{x}_{n+1}^1$, only x_n^1 is used. Next, to produce the forecast $\hat{x}_{n+1}^2$, x_n^2 *as well* as the newly produced forecast $\hat{x}_{n+1}^1$ is used. Similarly, to produce a forecast $\hat{x}_{n+1}^3$, the three values $x_n^3, \hat{x}_{n+1}^2$, and $\hat{x}_{n+1}^1$ are used, and so on. Thus, only historical information is used to produce forecasts for each time series. Similar to the above, the intensity transition density is

$$\log[\mu(\mathbf{x}_n^m)] = b_0 + b_1 \log(x_n^m + c) + \sum_{i=1}^{m-1} b_{i+1} \log(\hat{x}_{n+1}^{m-i} + c). \quad (8)$$

The constant gamma shape parameter is estimated using the maximum likelihood method as above.

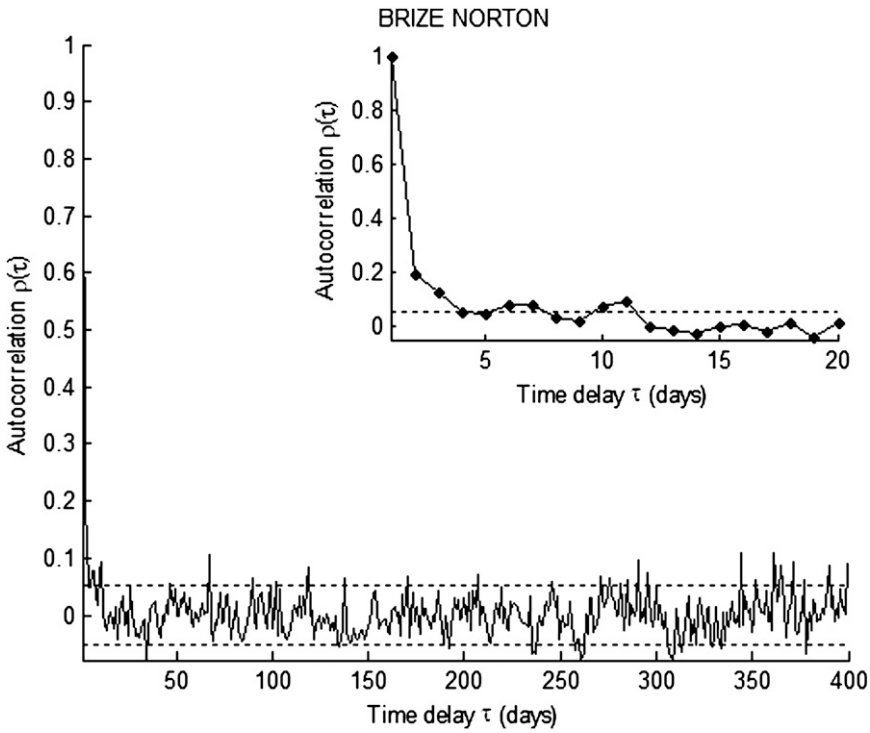

FIG. 6. Autocorrelation function for the Brize Norton gauge, from τ = 1- to τ = 400-day time lags. The dotted lines are the 95% Bartlett confidence intervals; the blue line is the autocorrelation coefficient. (inset) The short-range zoom for τ = 1- to τ = 20-day time lags.

Similarly, the conditional Bernoulli density for the occurrence $p(\mathbf{x}_n^m)$ is

$$p(\mathbf{x}_n^m) = l\left[a_0 + a_1 \log(x_n^m + c) + \sum_{i=1}^{m-1} a_{i+1} \log(\hat{x}_{n+1}^{m-i} + c)\right]. \quad (9)$$

5) *Generalized linear Markov multisite (Markov MGLM) model*: This model is similar to the model in (3) as it uses the same distribution, but for each site the regressors are the past rainfall value from all sites rather than just that particular site. Thus, if there is any time-delayed cross correlation, as we might expect for catchment-wide events that last for more than 1 day, this model should be able to capture them. This forms an alternative approach to the model in (4), which regresses on the predicted rainfall for the other sites, and models (noontime delayed) cross correlations. The Markov transition density is

$$f(x_{n+1}^m|\mathbf{x}_n) = [1 - p(\mathbf{x}_n)]\delta_0(x_{n+1}^m) + p(\mathbf{x}_n)\mathrm{Gamma}(x_{n+1}^m|\mathbf{x}_n), \quad (10)$$

where the vector $\mathbf{x}_n = (x_n^1, x_n^2, \ldots, x_n^M)^{\mathrm{T}}$ contains the past rainfall amount of all the sites. As above, the intensity transition density is

$$\log[\mu(\mathbf{x}_n)] = b_0 + \sum_{i=1}^{M} b_i \log(x_n^m + c). \quad (11)$$

Again, the constant gamma shape parameter is estimated using the maximum likelihood method. The conditional Bernoulli density for the occurrence $p(\mathbf{x}_n)$ is

$$p(\mathbf{x}_n) = l\left[a_0 + \sum_{i=1}^{M} a_i \log(x_n^m + c)\right]. \quad (12)$$

Note that this model differs from the model in (4) in that it uses information from all sites on the previous day to make a forecast at each individual site, whereas the model in (4) uses past information from only the *first* site to make a forecast for that site, whereupon this forecast is used to make the forecast for the next site, and so on. Therefore, we expect this model to reproduce spatial cross correlations where the time series have been shifted by 1 day relative to each other, whereas the model in (4) will reproduce

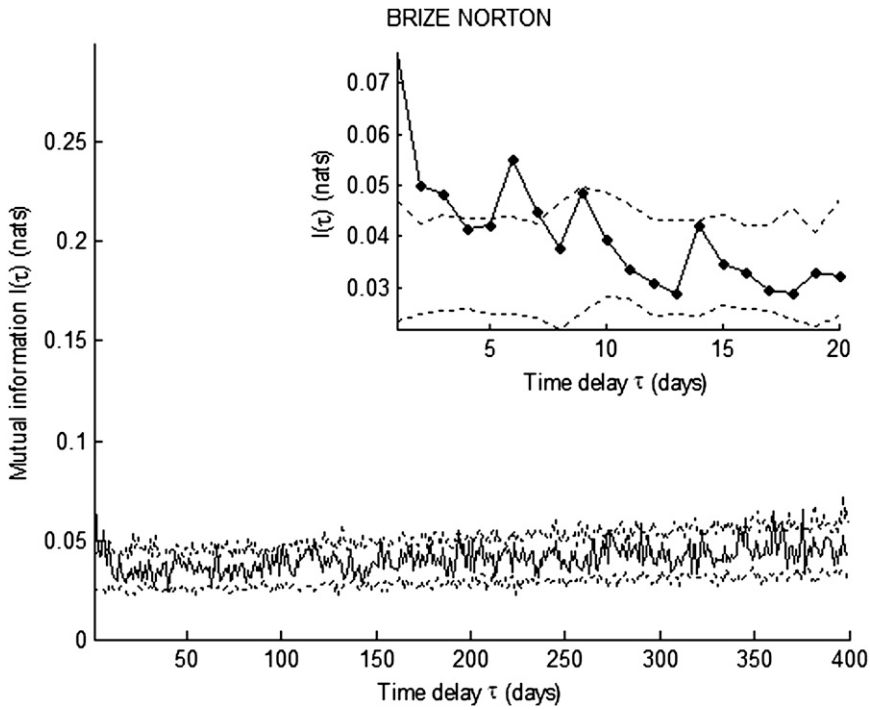

FIG. 7. Time-delayed mutual information for the Brize Norton gauge, from $\tau = 1$- to $\tau = 400$-day time lags. The dotted lines are the maximum and minimum mutual information over all the bootstraps and over all time lags; the blue line is the mutual information for the original time series. (inset) The short-range zoom for $\tau = 1$- to $\tau = 20$-day time lags.

spatial cross correlations where there is no time shift between sites.

6) *Generalized linear Markov seasonal multisite (Markov SMGLM) model*: This model is similar to the model in (5), except that it also regresses on seasonal variables, to attempt to exploit any slight seasonal variations. These variables have the form $s_n^w = \cos(2\pi w d_n/365)$, where d_n is the day number in the year, running from 0 to 364, and $w = 1, 2, 3, \ldots$ is the harmonic number. We use three harmonics; more did not lead to any significant improvements. These variables are appended to the end of the vector $\mathbf{x}_n$ in Eqs. (10), (11), and (12), but without a logarithmic transformation. Note that this model ceases to be strictly Markovian, as the transition function depends on the day in the year. The transition density has the same form as the model in (5).

(For notational brevity, we are using the same functions f, p, and μ to represent the unconditional or conditional model density, Bernoulli probability, and gamma means functions, respectively, and the parameters a, b, and c for all the models. In practice they are different functions and parameters, but they serve analogous roles in each model.)

To clarify further the GLMs above, the transition density in Eq. (3) and the conditional means specified in Eqs. (4) and (5) can be explained by considering the analogous situation for linear Gaussian autoregressive (AR) models. Informally, the transition density in Eq. (3) describes how the probability of any given forecast rainfall depth x_{n+1}^m depends on the rainfall depths x_n^m. In the perhaps more familiar context of the linear AR model, the transition density is a conditional Gaussian with mean that is a linear combination of past observations. The GLM framework extends this idea in two ways: first by generalizing the Gaussian density to the more general exponential family (of which the Gaussian, gamma and Bernoulli densities are special cases), and second by allowing a nonlinear transformation of the density mean, this transformation being called the link function. In the current GLM, the transition density in Eq. (3) is a mixture of two densities: the Bernoulli occurrence density with conditional mean $p(x_n^m)$, and the gamma intensity model. The Dirac delta function encodes the fact that, with probability $1 - p(x_n^m)$, zero rainfall depths will be forecast by the model, and alternately, with probability $p(x_n^m)$, nonzero rainfall depths will be

produced. In this linear mixture combination, the transition density is appropriately normalized.

7) *Postprocessed NCEP ensembles*: Finally, we use postprocessed National Centers for Environmental Prediction (NCEP) Global Ensemble (GENS) ensemble NWP model outputs. This uses supercomputing resources producing 10 different forecasts of the rainfall amounts, each forecast generated by a different perturbation of the initial atmospheric state assimilated from observations (Buizza et al. 2005). An additional unperturbed control forecast makes a total of 11 forecasts. The spatial resolution is 1° latitude–longitude, corresponding to approximately 110 km. Forecasts are available at 6-hourly intervals out to 16-days forecast horizon. Here, because of data size constraints, we have been able to access forecasts up to 8 days ahead.

The ensembles are first downscaled to the location of each site, using bilinear interpolation (linear in both north–south and east–west directions). After interpolation, the forecasts are calibrated using a mixture of generalized linear models (Sloughter et al. 2007):

$$f(x^m_{n+t}|y^m_{n,k,t}) = \frac{1}{11}\sum_{k=1}^{11}\{[1-p(y^m_{n,k,t})]\delta_0(x^m_n) + p(y^m_{n,k,t})\mathrm{Gamma}(x^m_n|y^m_{n,k,t})\}, \quad (13)$$

where $y^m_{n,k,t}$ is the interpolated ensemble member $k = 1, 2, \ldots, 11$, for site m on day n, for forecast horizons $t = 1, 2, \ldots, 8$, forming the density forecast for the rainfall sample x^m_{n+t}. The density for the intensity of rain is a conditional gamma generalized linear model with log link function, with conditional mean μ, such that $\log[\mu(y^m_{k,n,t})] = b_{0,k,t} + b_{1,k,t}y^m_{k,n,t}$. Also, the probability of occurrence of rain is given by $p(y^m_{n,k,t}) = l(a_{0,k,t} + a_{1,k,t}\sqrt[3]{y^m_{n,k,t}})$. Thus, any miscalibration due to bias in any ensemble member at any forecast horizon is removed by regression with the ensembles as predictors and rainfall intensity and occurrence as predictands, using the same GLM parameter estimation as described for the Markov models above. The cube root of rainfall amount in the probability of occurrence was found to improve the model fit (Sloughter et al. 2007).

b. Comparing daily rainfall point forecasts

Here we compare the ability of the models to produce point forecasts of daily total rainfall amount, which for the appropriate models is the combined model of occurrence/intensity. The point forecast from each model is the median of the model's forecast density. The mean absolute error (MAE) score is used:

$$E^m = \frac{1}{L}\sum_{n=1}^{L}|\hat{x}^m_n - x^m_n|, \quad (14)$$

where $\hat{x}^m_n$ is a forecast of total rainfall amount and L is the test data length. This score is proper (Gneiting and Raftery 2007) meaning that lower MAE scores imply more accurate forecasts, and the score is minimized by the perfect forecast.

c. Comparing daily rainfall density forecasts

The assessment of density forecasts is somewhat more complex than point forecasts. Specifically, it is important that the forecast produces the correct density of the observations: it must be well calibrated, and at the same time maximize sharpness; each forecast density must have a high probability around the actual observations (Diebold et al. 1998; Gneiting et al. 2007). Here we use the continuous ranked probability score (CRPS; Gneiting and Raftery 2007) which is also proper, and it can be shown decomposable into separate components of both calibration and sharpness. Thus, small values indicate forecasts that are both well calibrated and sharp. We use the empirical form (Gneiting and Raftery 2007):

$$\mathrm{CRPS}^m = \frac{1}{L}\sum_{n=1}^{L}(E|X - x^m_n| + \frac{1}{2}E|X - X'|), \quad (15)$$

where X and X' are independent random variables drawn from model's forecast density function p, and E denotes expectation.

We also perform diagnostic checks of the forecast calibration using the probability integral transform (Diebold et al. 1998; Gneiting et al. 2007):

$$z_n = \int_0^x p(u)du \quad (16)$$

(for notational clarity we have $x = x_n$). Here the function p is the (unconditional or conditional) forecast density function (or transition function) of each of the models, at forecast lead time of 1 day. For the perfectly calibrated model, z_n will be i.i.d. with uniform density in the interval between 0 and 1. Therefore, measuring calibration requires assessing the extent of deviation from uniformity of this time series. Typically, if the histogram is "U shaped" it will be because the spread of the forecasts is too narrow. Conversely, a humped-shaped histogram will indicate overdispersed forecasts (i.e., their range is too large; Gneiting et al. 2007).

Similarly, if the model captures the serial dependence in the time series, then z_n will be serially independent.

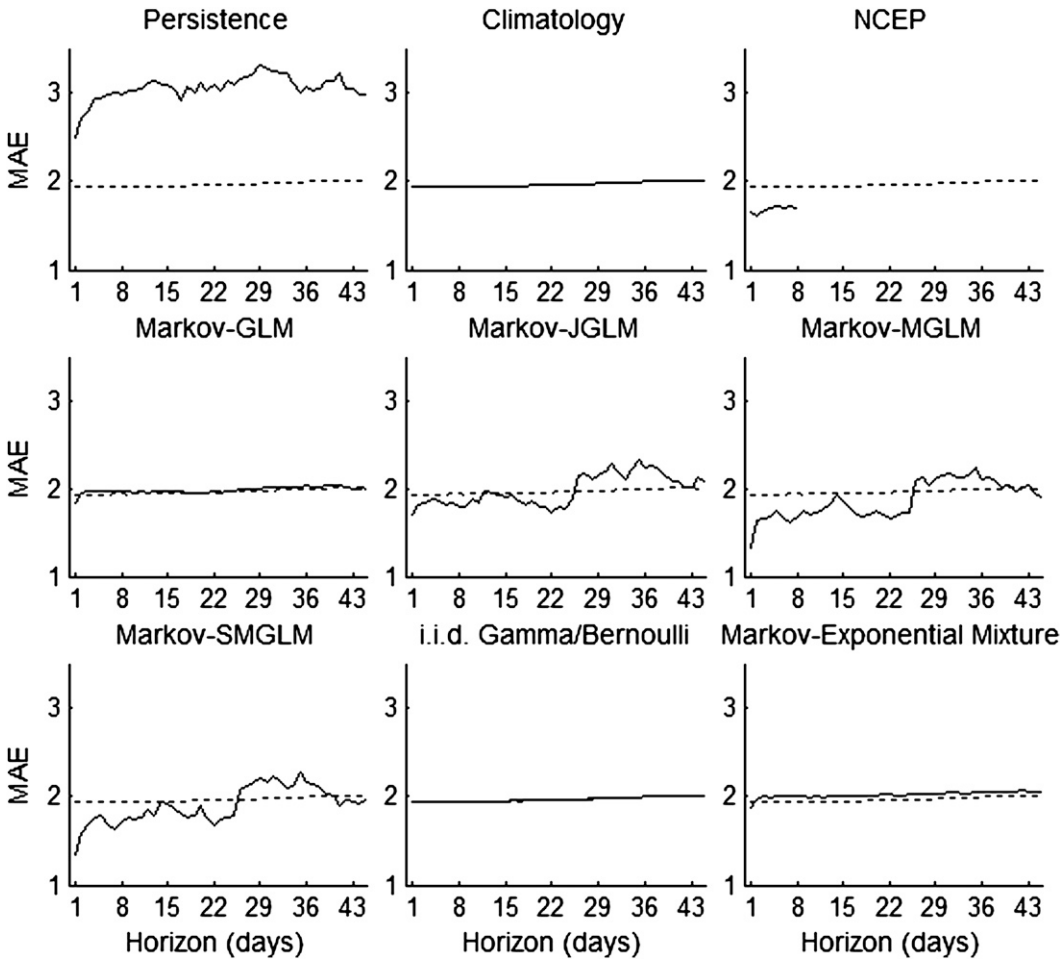

FIG. 8. Forecast MAE for all models, out to a forecast horizon of 45 days, averaged over the 10 gauges selected for modeling. The horizontal axes are forecast horizon in days, and the vertical axes are MAE. The dashed line on each plot is the climatological forecast MAE for comparison.

Tests for serial independence using the autocorrelation are most often applied in this context, and we follow this practice here (Gneiting et al. 2007), displaying the standard Bartlett 95% autocorrelation confidence intervals. We employ the stochastic interpolation method to calculate the PIT, by drawing 1000 samples from each predicted density and constructing the empirical cumulative density function. This defines a discrete distribution that approximates the underlying mixed discrete continuous density function (see Smith 1985 for further details).

d. Comparing daily rainfall occurrence probability forecasts

We compare probability forecasts of occurrence using the Brier score, which is also a proper score (Brier and Allen 1951):

$$B^m = \frac{1}{L}\sum_{n=1}^{L}(\hat{q}_n^m - q_n^m)^2, \tag{17}$$

where $\hat{q}_n^m$ is the forecast probability of occurrence and L is the test data length.

5. Results

For point forecasting performance MAE of rainfall totals, Fig. 8 shows that at lead times of 1–8 days ahead, the multisite nonseasonal/seasonal Markov SMGLM ranks slightly better than the processed ensembles. At lead times of between 9 and 25 days ahead, the Markov MGLM is best. The postprocessed ensembles have skill over climatology at the available lead times. The i.i.d. gamma–Bernoulli model does not have skill at any forecast horizon. The Markov GLM is an improvement

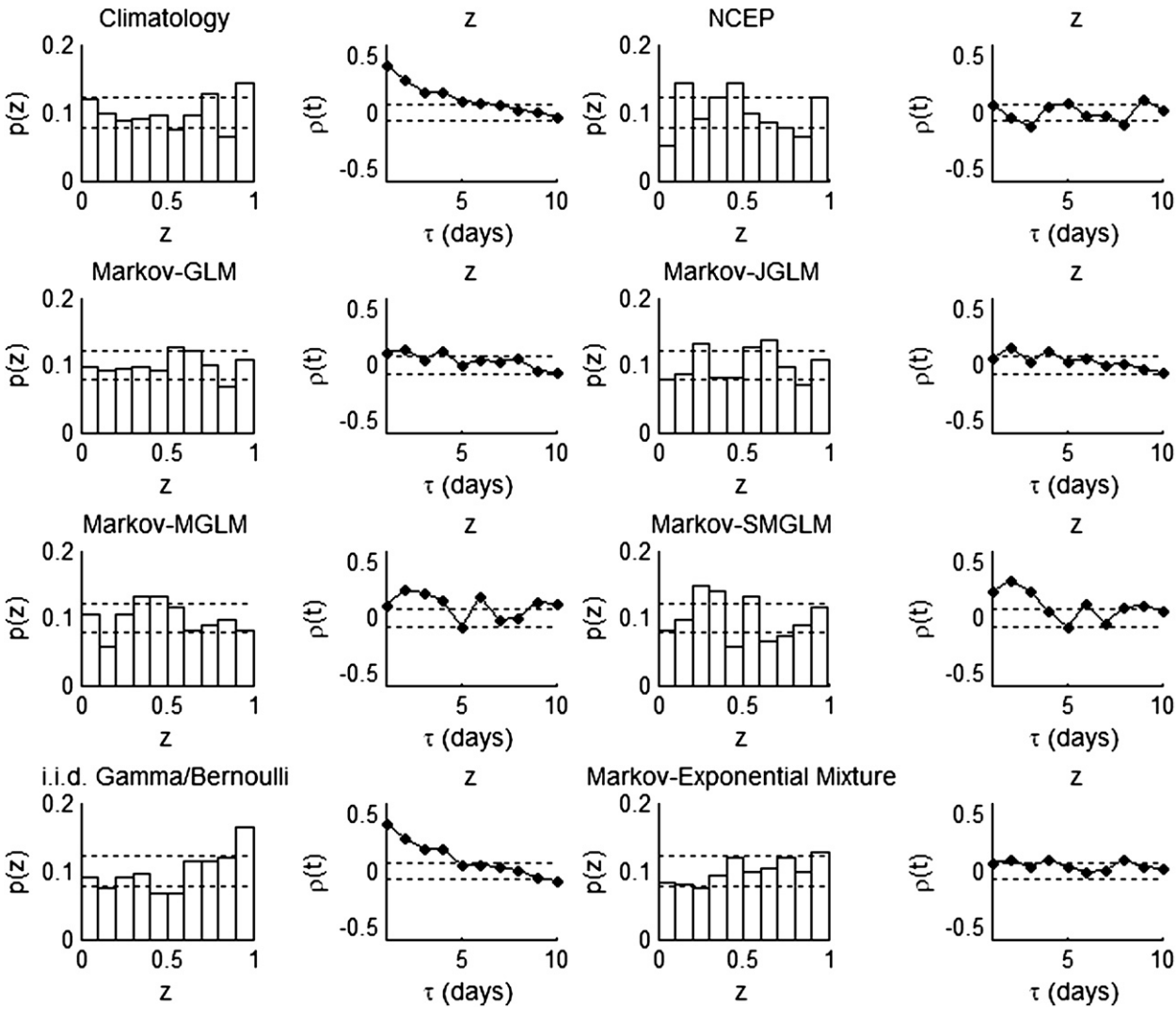

FIG. 9. PIT z series for the Brize Norton gauge, for the 1-day-ahead forecast horizon. (from left to right) The estimated distribution of the z series and the autocorrelation function for z, associated with the bar plot on the left. The dotted horizontal lines are the estimated 95% confidence intervals, the bars are the estimated distributions, and the black lines are the autocorrelation at time lag τ.

over the climatology for the first day, but afterward, ceases to have skill. The exponential mixture model only has skill on the first day, and thereafter loses skill. The persistence forecast is consistently the worst forecast over all horizons.

Turning to the density forecasts, the diagnostic checks in Fig. 9 show that all of the models are reasonably well calibrated, although there is some residual over- and underdispersion in most models. From the autocorrelation functions of the z series, as expected the unconditional climatology and i.i.d. gamma–Bernoulli models fail to capture the small amount of serial correlation in the data for the first 4 or 5 days time lag. The conditional models naturally fare better in this regard, and the postprocessed ensembles perform best. The results are consistent with these findings for the other gauges, with some negligible differences. The diagnostic check of Fig. 5 shows that the Markov JGLM is capable of reproducing the spatial correlations to a reasonable extent, although the correlations are smaller than those in the original time series because the mixed GLM density model of the actual probability densities of rainfall at each site is not perfect.

Regarding combined calibration and sharpness of the density forecasts, Fig. 10 shows that the nonseasonal/seasonal multisite Markov GLMs are the best performers on the first day. The joint-site GLM has some skill at 1 day ahead, but thereafter lacks appreciative skill. At 2–8 days ahead, the postprocessed ensembles

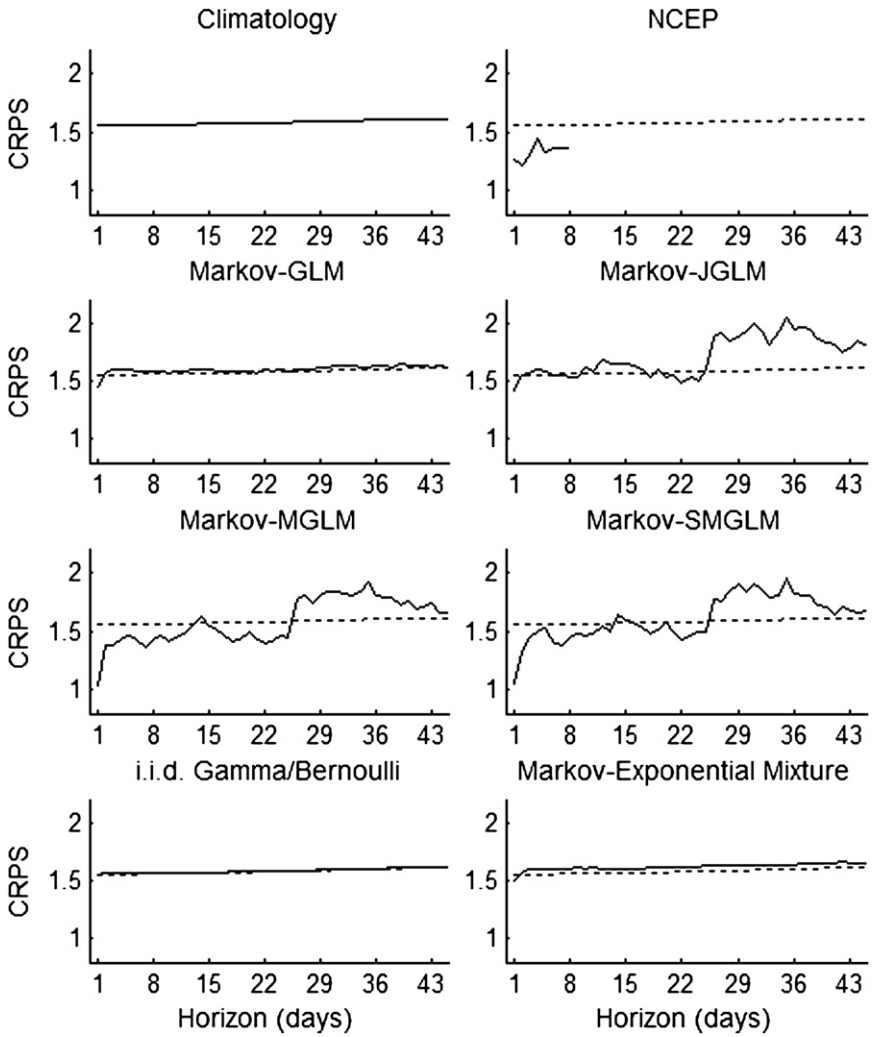

FIG. 10. CRPS results over all density forecast models averaged over all 10 gauges selected for the modeling part of the study. The horizontal axes are forecast horizon in days, and the vertical axes are CRPS. The dashed line on each plot is the climatology CRPS for comparison.

rank first, just ahead of the Markov MGLM/SMGLMs, which have some skill for some forecast horizons between 9 and 25 days ahead.

The i.i.d. gamma–Bernoulli model does not have skill over climatology at any horizon. The rest of the conditional models show slight improvements at 1 day ahead, but then show similar performance to the unconditional models.

For the occurrence Brier score, the postprocessed ensembles have the best score for forecast horizons of 2–8 days, outperforming all other models. However, the conditional models, in particular the Markov MGLM/SMGLMs all show skill relative to climatology for the first day, thereafter they lose skill. The i.i.d. gamma–Bernoulli model does not have skill at any forecast horizon, and the persistence forecast is the worst at every horizon.

It is worth noting that Markov GLMs involve highly nonlinear feedback mechanisms, particularly noticeable when propagating information from many neighboring sites. Although often drifting to zero, unlike the simple, unconditionally stable Markov chains such as the Cao–Li–Wei model, it is possible for Markov GLMs to produce growing responses as well. This is noticeable in the MAE, CRPS, and Brier scores, where the performance of some of the more complex Markov models varies somewhat with forecast horizon. The other, simpler models produce smoother results. Another note is that experiments with shuffling the order of sites used in the Markov JGLM method did not lead to substantial differences in performance, either in the MAE, CRPS, or Brier scores.

These forecasting results raise the question of why some of the statistical methods have comparable or

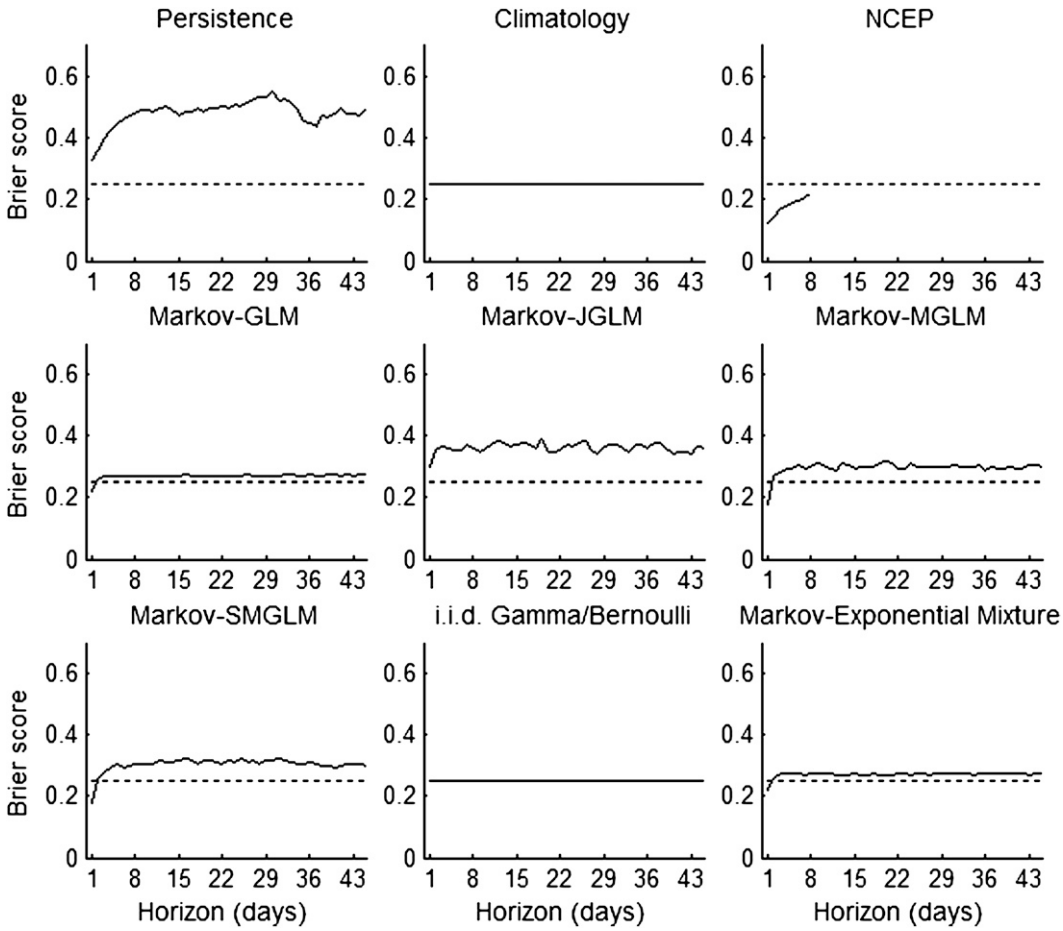

FIG. 11. The Brier score of forecast of the probability of occurrence of rainfall with the results averaged over all 10 gauges selected for the modeling part of the study. The horizontal axes are the forecast horizon (days) and the vertical axes are the Brier score. The dashed line is the climatology Brier score for comparison.

better performance than the postprocessed ensembles in some aspects as previously described. Turning first to point forecast performance, temporal correlations in the data beyond 1 day ahead are very small. Nonetheless, the Markov GLMs are well equipped to exploit this small, 1-day-ahead autocorrelation.

Second, with regard to density forecasts, the CRPS results show that incorporating all the information from every site on the previous day when forecasting each site individually, improves the calibration of the Markov multisite models relative to all the other models (including the Markov JGLM joint site method). Therefore, neighboring sites do contain useful information that can be exploited to improve forecasts at short lead times.

Regarding the occurrence forecast performance, the ensemble calibration regression method is successfully able to remove bias in the interpolated ensembles to produce the best forecasts. However, the training data is very similar to the test data, both having long, consecutive runs of dry days, followed by shorter, consecutive runs of wet days. Thus, the occurrence time series is highly autocorrelated 1 day ahead, diminishing rapidly with increasing forecast horizon. The conditional Markov MGLM and Markov SMGLMs use the most information from the past in order to produce forecasts. As can be seen in Fig. 11, these conditional models, which are designed to capture temporal autocorrelation, do very well in exploiting this 1-day-ahead autocorrelation.

6. Conclusions

In this paper, we investigated the autocorrelation and cross-correlation structure of a large number of rain gauges in the Thames Valley, United Kingdom, and demonstrated that while autocorrelation in the rainfall depth amount is of minor importance, spatial cross

correlation is highly dominant. We also showed some slight seasonal variations in the mean intensity of rainfall on wet days. We used this information to produce a set of new, site-specific statistical density forecast models in this spatial area, based on variations of a Markov GLM, non-Gaussian regression method in a couple of different configurations. We tested these models against a set of simple benchmarks and some more sophisticated models proposed in the literature, and against ensemble NWP forecasts combined with GLM regression in a postprocessing approach. The tests involved the comparison of rainfall forecast performance of all the models for each rain gauge, up to 45 days ahead. The tests demonstrated that Markov GLMs can be configured to produce good 1-day-ahead forecasts and reasonably skillful short-term forecasts up to a couple of weeks ahead. They also show that combining GLM regression with ensembles can effectively calibrate the ensembles to produce skilful density forecasts up to a week ahead. The results do not support the use of any of the other proposed models.

In terms of overall density forecasting, all the models were well calibrated, but in summary, the GLMs, either alone or in combination with ensembles, performed best when both calibration and sharpness were considered simultaneously. In terms of ability to forecast occurrence of rain, except at lead times of 1 day, no models were capable of bettering the postprocessed ensembles. We also demonstrated the superior ability of the postprocessed ensembles to reproduce the (small) serial correlation in the rainfall data.

A similar study (Taylor and Buizza 2004) compared temperature forecasts from simple autoregressive time series models, postprocessed European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble mean, and a high-resolution point NWP; it was found that the ensemble mean was the best under the MAE at up to 10 days ahead. Similarly, Campbell and Diebold (2005) found that point forecasts from time series models could not outperform NWP forecasts. Our findings disagree as we have found it possible to produce time series point forecasts slightly better than calibrated NWP forecasts up to 8 days ahead. We believe this is because precipitation is notoriously difficult to predict, particularly at local sites, and time series models exploiting correlations can contribute to making useful forecasts.

The results lead us to suggest ways in which ensemble forecasts might best be calibrated for full-density forecast applications. Contrary to other reports (Cao et al. 2004; Robertson et al. 2004), we believe that this study can act as a caution against the use of simple unconditional density models and the more elaborate two-state Markov chains combined with exponential mixtures for this purpose. In particular, we believe our results suggest that Markov GLMs could be effective new techniques in this regard, which concurs with other studies (Sloughter et al. 2007).

Acknowledgments. The rainfall data were supplied by the British Atmospheric Data Centre. This work was carried out with funding from the Natural Environment Research Council (NERC), United Kingdom, under the Flood Risk from Extreme Events Programme, NERC Grant NE/E002013/1. The comments of the reviewers also helped to improve the quality of the manuscript.

REFERENCES

Agha, M., and M. T. Ibrahim, 1984: Algorithm AS 203: Maximum likelihood estimation of mixtures of distributions. *Appl. Stat.,* **33,** 327–332.

Applequist, S., G. E. Gahrs, R. L. Pfeffer, and X. F. Niu, 2002: Comparison of methodologies for probabilistic quantitative precipitation forecasting. *Wea. Forecasting,* **17,** 783–799.

Barry, R. G., and R. J. Chorley, 2003: *Atmosphere, Weather, and Climate*. 8th ed. Routledge, 421 pp.

Battan, L. J., 1984: *Fundamentals of Meteorology*. 2nd ed. Prentice-Hall, 304 pp.

Brier, G. W., and R. A. Allen, 1951: Verification of weather forecasts. *Compendium of Meteorology,* T. F. Malone, Ed., Amer. Meteor. Soc., 841–848.

Brutsaert, W., 2005: *Hydrology: An Introduction*. Cambridge University Press, 605 pp.

Buizza, R., 2003: Weather prediction: Ensemble prediction. *Encyclopedia of Atmospheric Sciences*, J. R. Holton, J. Pyle, and J. A. Curry, Eds., Academic Press, 2546–2557.

——, T. Petroliagis, T. Palmer, J. Barkmeijer, M. Hamrud, A. Hollingsworth, A. Simmons, and N. Wedi, 1998: Impact of model resolution and ensemble size on the performance of an ensemble prediction system. *Quart. J. Roy. Meteor. Soc.,* **124,** 1935–1960.

——, P. L. Houtekamer, Z. Toth, G. Pellerin, M. Z. Wei, and Y. J. Zhu, 2005: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Wea. Rev.,* **133,** 1076–1097.

Campbell, S. D., and F. X. Diebold, 2005: Weather forecasting for weather derivatives. *J. Amer. Stat. Assoc.,* **100,** 6–16.

Cao, M., A. Li, and J. Z. Wei, 2004: Precipitation modeling and contract valuation: A frontier in weather derivatives. *J. Altern. Invest.,* **Fall,** 93–99.

Diebold, F. X., T. A. Gunther, and A. S. Tay, 1998: Evaluating density forecasts with applications to financial risk management. *Int. Econ. Rev.,* **39,** 863–883.

Easterling, D. R., J. L. Evans, P. Ya. Groisman, T. R. Karl, K. E. Kunkel, and P. Ambenje, 2000: Observed variability and trends in extreme climate events: A brief review. *Bull. Amer. Meteor. Soc.,* **81,** 417–425.

Gneiting, T., and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *J. Amer. Stat. Assoc.,* **102,** 359–378.

——, F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc. B,* **69,** 243–268.

Grunwald, G. K., and R. H. Jones, 2000: Markov models for time series with mixed distribution. *Environmetrics,* **11,** 327–339.

Hyndman, R. J., and G. K. Grunwald, 2000: Generalized additive modelling of mixed distribution Markov models with application to Melbourne's rainfall. *Aust. N. Z. J. Stat.,* **42,** 145–158.

John, S. E., 2003: Site-specific probability of precipitation forecasting. M.Sc. thesis, Department of Meteorology, University of Reading, Reading, United Kingdom, 93 pp.

Kantz, H., and T. Schreiber, 2004: *Nonlinear Time Series Analysis*. 2nd ed. Cambridge University Press, 388 pp.

Little, M. A., P. E. McSharry, I. M. Moroz, and S. J. Roberts, 2006: Testing the assumptions of linear prediction analysis in normal vowels. *J. Acoust. Soc. Amer.,* **119,** 549–558.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.,* **122,** 73–119.

Moura, A. D., and S. Hastenrath, 2004: Climate prediction for Brazil's Nordeste: Performance of empirical and numerical modeling methods. *J. Climate,* **17,** 2667–2672.

Palmer, T. N., F. Molteni, R. Mureau, R. Buizza, P. Chapelet, and J. Tribbia, 1993: Ensemble prediction. *Proc. ECMWF Seminar on Validation of Models over Europe: Vol. I,* ECMWF, 21–66.

Robertson, A. W., S. Kirshner, and P. Smyth, 2004: Downscaling of daily rainfall occurrence over northeast Brazil using a hidden Markov model. *J. Climate,* **17,** 4407–4424.

Sanso, B., and L. Guenni, 1999: A stochastic model for tropical rainfall at a single location. *J. Hydrol.,* **214,** 64–73.

Sloughter, J. M., A. E. Raftery, T. Gneiting, and C. Fraley, 2007: Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Mon. Wea. Rev.,* **135,** 3209–3220.

Smith, J. Q., 1985: Diagnostic checks of nonstandard time-series models. *J. Forecasting,* **4,** 283–291.

Taylor, J. W., and R. Buizza, 2004: Comparing temperature density forecasts from GARCH and atmospheric models. *J. Forecasting,* **23,** 337–355.

——, and ——, 2006: Density forecasting for weather derivative pricing. *Int. J. Forecasting,* **22,** 29–42.

Upton, G. J. G., A. R. Holt, R. J. Cummings, A. R. Rahimi, and J. W. F. Goddard, 2005: Microwave links: The future for urban rainfall measurement? *Atmos. Res.,* **77,** 300–312.

Venables, W. N., and B. D. Ripley, 2002: *Modern Applied Statistics with S*. 4th ed. Springer, 512 pp.

Wilks, D. S., 1998: Multisite generalization of a daily stochastic precipitation generation model. *J. Hydrol.,* **210,** 178–191.

——, 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. Academic Press, 648 pp.