

Bayesian prediction of rainfall records using the generalized exponential distribution

Mohamed T. Madi^{1*,†} and Mohammad Z. Raqab²

¹Department of Statistics, UAE University, Al Ain, United Arab Emirates

²Department of Mathematics, University of Jordan, Amman 11942, Jordan

SUMMARY

The Los Angeles rainfall data are found to fit well to the two-parameter generalized exponential (GE) distribution. A Bayesian parametric approach is described and used to predict the behavior of further rainfall records. Importance sampling is used to estimate the model parameters, and the Gibbs and Metropolis samplers are used to implement the prediction procedure. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: generalized exponential distribution; record statistics; Bayesian estimation; Bayesian prediction; Gibbs and Metropolis sampling; importance sampling

AMS SUBJECT CLASSIFICATION: 62F10; 62F15; 62F25; 62E25

1. INTRODUCTION

Let X_1, X_2, \dots be a sequence of independent and identically distributed (iid) random variables (r.v.s) with cumulative distribution function (cdf) $F(x)$ and probability density function (pdf) $f(x)$. An observation X_j will be called an upper (lower) record value if its value exceeds (is lower than) that of all previous observations. That is, X_j is an upper (lower) record if $X_j > (<) X_i$ for every $i < j$. If $\{L(n), n \geq 1\}$ is defined by $L(1) = 1$, $L(n) = \min\{j | j > L(n-1), X_j < X_{L(n-1)}\}$ for $n \geq 2$, then $\{X_{L(n)}, n \geq 1\}$ provides a sequence of lower record statistics. The sequence $\{L(n), n \geq 1\}$ represents the record times.

Record values are defined as a model of successive extremes in a sequence of iid random variables (see, e.g., Glick, 1978; Gulati and Padgett, 1995). It is of interest to note that there are situations in which only records are observed, such as, in meteorology, hydrology, seismology, and mining. Predictions of future records based on observed records do arise naturally in this context.

Several authors have considered prediction problems involving record values. Ahsanullah (1980) obtained linear prediction of records from the exponential distribution. Awad and Raqab (2000) conducted a comparison study between several prediction intervals for future records from the exponential

*Correspondence to: M. T. Madi, Department of Statistics, UAE University, Al Ain, United Arab Emirates.

†E-mail: mmadi@uaeu.ac.ae

distribution. Raqab (2001) established the highest conditional density (HCD) prediction intervals for future records from the exponential distribution. Raqab (2002) considered the best linear unbiased estimation of the location and scale parameters of the generalized exponential (GE) model and discussed the prediction of future records. Dunsmore (1983) addressed the problem of Bayesian prediction of future records from the exponential distribution. Using the Laplace approximation method, Basak and Bagchi (1990) determined the predictive distribution of a future observation as developed by Tierney and Kadane (1986). Smith and Miller (1986) presented a class of non-Gaussian steady state models and discussed the prediction of records based on censored and uncensored observations. Recently, Al-Hussaini and Ahmad (2003) considered the problem of Bayesian interval prediction of future records using a general class of distributions. Ahsanullah and Bhatti (2003) discussed the problem of predicting olympic records using the extreme value distribution. Madi and Raqab (2004) used a Bayesian approach to establish future predictors for the Pareto model. Jaheen (2003) obtained Bayes estimators for the two parameters of the Gompertz model and determined future upper record predictors. Malinowska and Szyal (2004) derived a family of Bayesian estimators and predictors for the Gumbel model based on lower records. Raqab (2006) proposed a non-parametric procedure for predicting future rainfall records. Jaheen (2004) derived Bayes and Empirical Bayes estimators for the one-parameter GE model and obtained prediction bounds for future lower records.

Because of the nature of the GE model and to avoid the mathematical complication involved with the joint distribution of the upper records, we will consider only lower record statistics. Note that upper records can be transformed to lower records by replacing the original sequence of random variables by $\{-X_j, j \geq 1\}$, or by $\{1/X_j, j \geq 1\}$ if $P(X_j > 0) = 1$.

Let us denote the n th lower record value by R_n . The pdf of R_n is:

$$f_{R_n}(x) = \frac{[-\log F(x)]^{n-1}}{(n-1)!} f(x)$$

The joint density of the m lower records is:

$$f_{R_1, \dots, R_m}(x_1, \dots, x_m) = \prod_{i=1}^{m-1} h(x_i) f(x_m)$$

where $h(x) = f(x)/F(x)$ represents the reverse hazard rate. For more details about records, one can refer to Arnold *et al.* (1998) and Ahsanullah (2004).

Assume that X_1, X_2, \dots follow the GE distribution with cdf and pdf, respectively,

$$F(x; \alpha, \lambda) = (1 - e^{-\lambda x})^\alpha, \quad \alpha, \lambda > 0 \quad (1)$$

and

$$f(x; \alpha, \lambda) = \alpha \lambda (1 - e^{-\lambda x})^{\alpha-1} e^{-\lambda x}, \quad \alpha, \lambda > 0 \quad (2)$$

where α is the shape parameter and λ is the reciprocal of a scale parameter. The two-parameter GE distribution has been introduced and studied quite extensively by Gupta and Kundu (1999, 2001a, b). When $\alpha = 1$, the GE coincides with the exponential distribution. Due to the simple structure of its distribution function, the GE can be used quite effectively in analyzing any lifetime data, especially in the presence of censoring or if the data are grouped. For this distribution, the hazard function, which plays an important role in life testing and reliability problems, can be increasing, decreasing, or constant

depending on the shape parameter α . For any λ , the hazard function is non-decreasing if $\alpha > 1$, and it is non-increasing if $\alpha < 1$. For $\alpha = 1$, it is constant. The GE has a unimodal density function and for fixed scale parameter, as the shape parameter increases, it becomes more and more symmetric.

In this paper, we present a Bayesian approach to predict the behavior of further records from the same distribution. Namely, given m lower records $r_1 \geq r_2 \geq \dots \geq r_m$, of a sequential sample, we predict the future records $r_{m+1} \geq r_{m+2} \geq \dots \geq r_n$, ($n > m$). In Section 2 of this article, we derive the Bayes estimators of the parameters α and λ based on the observed records. In Section 3, we obtain the predictive density function and the corresponding Bayes predictors of the future n th record. We also use the Gibbs and Metropolis samplers to estimate the predictive distribution of future records. Section 4 includes an illustration using the rainfall data set.

2. POSTERIOR DISTRIBUTIONS AND BAYES ESTIMATION

In this section, we present the posterior densities of the parameters α and λ based on m observed records. A natural choice for the priors of α and λ would be to assume that the two quantities are independent and that

$$\alpha \sim G(a, b) \quad \text{and} \quad \lambda \sim G(c, d) \quad (3)$$

where a, b, c, d are chosen to reflect prior knowledge about α and λ and where $G(a, b)$ denotes the gamma distribution with density function

$$g(\alpha) = \frac{b^a}{\Gamma(a)} \alpha^{a-1} e^{-b\alpha}, \quad \alpha > 0$$

The likelihood function of α and λ for the given observed sequence sample $\mathbf{R} = (R_1, R_2, \dots, R_m)$ is given by:

$$L(\mathbf{r} | \alpha, \lambda) \propto \alpha^m \lambda^m \exp\{-m\lambda\bar{r} - \alpha\tau_\lambda(r_m) + D_\lambda(\mathbf{r})\} \quad (4)$$

where

$$D_\lambda(\mathbf{r}) = -\sum_{i=1}^m \ln(1 - e^{-\lambda r_i}), \quad \bar{r} = \frac{1}{m} \sum_{i=1}^m r_i \quad \text{and} \quad \tau_\lambda(r_i) = -\ln(1 - e^{-\lambda r_i})$$

By combining Equations (3) and (4), we obtain the joint posterior density of α and λ

$$\pi(\alpha, \lambda | \mathbf{r}) \propto g_\alpha(n + c, n\bar{r} + d) g_\lambda(n + a, \tau_\lambda(r_i) + b) \exp\{D_\lambda(\mathbf{r})\}$$

where g_α and g_λ are gamma densities for α and λ . The posterior model is essentially an updated version of our prior knowledge about α and λ in light of the sample data.

It is clear that the marginal posterior density functions of λ and α cannot be expressed in closed forms. To obtain the Bayes estimates of λ and α , we use importance sampling as follows:

The marginal posterior density of λ is given by

$$\pi(\lambda | \mathbf{r}) \propto g_\lambda(n + c, n\bar{r} + d) T(\lambda) \quad (5)$$

where $T(\lambda) = (\tau_\lambda(r_n) + b)^{-(n+a)} \exp\{D_\lambda(\mathbf{r})\}$. From Equation (5) we can express the Bayes estimate of λ as

$$E(\lambda | \mathbf{r}) = \frac{E_\lambda^{(1)}[\lambda T(\lambda)]}{E_\lambda^{(1)}[T(\lambda)]} \quad (6)$$

where $E_\lambda^{(1)}$ denotes the expectation with respect to $G(n + c, n\bar{r} + d)$. Using g_λ as the importance sampling function, the procedure is implemented by generating $\lambda_1, \lambda_2, \dots, \lambda_k$ from the gamma distribution $G(n + c, n\bar{r} + d)$ and then averaging the numerator $\lambda T(\lambda)$ and the denominator $T(\lambda)$ of Equation (6) with respect to these simulations. The Bayes estimate is then expressed as

$$E(\lambda | \mathbf{r}) = \frac{\sum_{i=1}^k \lambda_i T(\lambda_i)}{\sum_{i=1}^k T(\lambda_i)}$$

To find the Bayes estimate of α , we note that $\alpha | \lambda, \mathbf{r}$ is $G(n + a, \tau_\lambda(r_n) + b)$ and that $\pi(\alpha | \mathbf{r}) \propto E_{\lambda|\mathbf{r}}[\pi(\alpha | \lambda, \mathbf{r})]$. It follows that

$$\pi(\alpha | \mathbf{r}) = \frac{E_\lambda^{(1)}[T(\lambda) g_\alpha(n + a, \tau_\lambda(r_n) + b)]}{E_\lambda^{(1)}[T(\lambda)]}$$

Using $E(\alpha | \lambda, \mathbf{r}) = (n + a)/(\tau_\lambda(r_n) + b)$, we obtain the Bayes estimate of α as

$$E(\alpha | \mathbf{r}) = \frac{E_\lambda^{(1)}[T(\lambda) (n + a)/(\tau_\lambda(r_n) + b)]}{E_\lambda^{(1)}[T(\lambda)]} \quad (7)$$

The reliability function for the GE distribution is given by $\Lambda(x_0 | \alpha, \lambda) = 1 - (1 - e^{-\lambda x_0})^\alpha$. It can be checked that

$$\begin{aligned} E_{\alpha|\lambda, \mathbf{r}}[\Lambda] &= \int_0^\infty [1 - (1 - e^{-\lambda x_0})^\alpha] g_\alpha(n + a, \tau_\lambda(r_n) + b) d\alpha \\ &= 1 - \left(\frac{\tau_\lambda(r_n) + b}{\tau_\lambda(r_n) + \tau_\lambda(x_0) + b} \right)^{n+a} \end{aligned}$$

It follows that the Bayes estimate of Λ can be expressed as

$$E(\Lambda | \mathbf{r}) = \frac{E_{\lambda}^{(1)} \left\{ T(\lambda) \left[1 - \left(\frac{\tau_{\lambda}(r_n) + b}{\tau_{\lambda}(r_n) + b + \tau_{\lambda}(x_0)} \right)^{n+a} \right] \right\}}{E_{\lambda}^{(1)} [T(\lambda)]} \quad (8)$$

The Bayes estimates in Equations (7) and (8) are computed by averaging the numerators and the denominators in the two expressions with respect to the simulations $\lambda_1, \lambda_2, \dots, \lambda_k$.

3. SEQUENCE SAMPLE-BASED PREDICTION OF RECORDS

In this section we present the Bayesian predictive distributions for future records based on the observed m records $\mathbf{R} = (R_1, R_2, \dots, R_m)$. Our interest is in predicting the future n th record ($n > m$).

The extended likelihood function of \mathbf{R} and a future record R_n is given by

$$L(\mathbf{r}, r_n | \alpha, \lambda) \propto \alpha^{n+1} \lambda^{m+1} \exp\{-\alpha[\tau_{\lambda}(r_n)]\} \exp\{-\lambda[m\bar{r} + r_n + D_{\lambda}(\mathbf{r}) + \tau_{\lambda}(r_n)]\} \\ + [\tau_{\lambda}(r_n) - \tau_{\lambda}(r_m)]^{n-m-1}$$

By forming the product of the above extended likelihood function and the joint prior, the full Bayesian model is found to be

$$\pi(r_n, \alpha, \lambda | \mathbf{r}) = \frac{\xi \Gamma(m+c+1)}{(m\bar{r} + r_n + d)^{m+c+1}} \alpha^{n+a-1} \exp\{-\alpha[\tau_{\lambda}(r_n) + b]\} \\ \times \exp\{-\lambda[D_{\lambda}(\mathbf{r}) + \tau_{\lambda}(r_n)]\} [\tau_{\lambda}(r_n) - \tau_{\lambda}(r_m)]^{n-m-1} \\ \times g_{\lambda}(m+c+1, m\bar{r} + r_n + d)$$

where

$$\xi = \frac{b^a d^c}{\Gamma(b) \Gamma(d) (n-m-1)!}$$

The predictive density function of R_n , given $\mathbf{R} = \mathbf{r}$, is obtained to be

$$p(r_n | \mathbf{r}) = \frac{\xi \Gamma(m+c+1) \Gamma(n+a)}{(m\bar{r} + r_n + d)^{m+c+1}} E_{\lambda}^{(2)}(\varphi(\lambda)) \quad (9)$$

where

$$\varphi(\lambda) = [\tau_{\lambda}(r_n) - \tau_{\lambda}(r_m)]^{n-m-1} [\tau_{\lambda}(r_n) + b]^{-(n+a)} \exp\{-\lambda[\tau_{\lambda}(r_n) + D_{\lambda}(\mathbf{r})]\}$$

and $E_{\lambda}^{(2)}$ is the expectation with respect to $G(m+c+1, m\bar{r} + r_n + d)$.

Clearly the predictive estimates $E(R_n | \mathbf{R} = \mathbf{r})$ cannot be computed directly from Equation (9). Therefore we opt for stochastic simulation procedures, namely, the Gibbs and Metropolis samplers

(Gilks *et al.*, 1995) to generate samples from the predictive distributions. For this, let us consider the problem of predicting $\mathbf{V} = (R_{m+1}, \dots, R_n)$. The extended likelihood function of \mathbf{R} and future records \mathbf{V} is given by

$$L(\mathbf{r}, \mathbf{v} | \alpha, \lambda) \propto \alpha^n \lambda^n \exp\{-\alpha[\tau_\lambda(r_n)] - \lambda n \bar{w} + D_\lambda(\mathbf{r})\}$$

where $\bar{w} = \sum_{i=1}^n r_i / n$.

By forming the product of the extended likelihood and the joint prior of α , and λ , the full Bayesian model is expressed as

$$\begin{aligned} \pi(\alpha, \lambda, \mathbf{v} | \mathbf{r}) &\propto \alpha^{n+a-1} \lambda^{n+c-1} \exp\{-\alpha[\tau_\lambda(r_n) + b]\} \\ &\times \exp\{-\lambda[n\bar{w} + d]\} \exp\{D_\lambda(\mathbf{r})\} \end{aligned}$$

Setting $\mathbf{V}_k = (R_{m+1}, \dots, R_{k-1}, R_{k+1}, \dots, R_n)$, the full conditional distribution of R_k ($m+1 \leq k \leq n$), is found to be

$$\pi(r_k | \mathbf{r}, \mathbf{v}_k, \alpha, \lambda) = \begin{cases} \frac{\lambda e^{-\lambda r_k} I[r_{k+1} < r_k < r_{k-1}]}{(1 - e^{-\lambda r_k})(\tau(r_{k+1}) - \tau(r_{k-1}))}, & k = m+1, \dots, n-1 \\ \frac{\lambda \alpha e^{-\lambda r_n} (1 - e^{-\lambda r_n})^{\alpha-1} I(r_n < r_{n-1})}{(1 - e^{-\lambda r_{n-1}})^\alpha}, & k = n \end{cases}$$

The full conditional distribution of λ is

$$\pi(\lambda | \mathbf{y}, \alpha) \propto \lambda^{n+c-1} e^{-\lambda[n\bar{w}+d]} (1 - \exp\{-\lambda r_n\})^\alpha \exp\{D_\lambda(\mathbf{r})\}$$

and the conditional distribution of $\alpha | \mathbf{y}, \lambda$ is $G(n + a, \tau_\lambda(r_n) + b)$.

Using the Gibbs sampler to estimate the posterior distribution requires being able to sample from the full conditional distributions for each quantity involved. This is the case for α and R_k but not for λ . Consequently, Metropolis–Hastings (M–H) steps are introduced into the Gibbs sampler so that α and R_k are sampled directly from their full conditional distributions, whereas λ is updated via a M–H step as explained in Tierney (1994), using $G(n + c, n\bar{w} + d)$ as a proposal distribution. The M–H step proceeds as follows:

Given $\lambda^{(i-1)}$,

- (i) sample y from $G(n + c, n\bar{w} + d)$ and u from $U(0, 1)$;
- (ii) if $u < \min(1, \kappa)$ then let $\lambda^{(i)} = y$ else go to (i), where

$$\kappa = \frac{e^{D_y[1 - e^{-y r_n}]^\alpha}}{e^{D_{\lambda^{(i-1)}}[1 - e^{-\lambda^{(i-1)} r_n}]^\alpha}}$$

The records R_k are generated using the inverse cdf transformation method via the expressions

$$r_k = -\frac{1}{\lambda} \log [1 - \exp\{u(\tau(r_{k+1}) - \tau(r_{k-1})) - \tau(r_{k+1}))\}]$$

for $k = m + 1, \dots, n - 1$ and

$$r_n = -\frac{1}{\lambda} \log[1 - u^{\frac{1}{\alpha}} (1 - e^{-\lambda r_{n-1}})]$$

4. ILLUSTRATIVE EXAMPLE

In this section we present the analysis of a real life data. The computations are performed using Visual Fortran 5.0 and Minitab 14. The graphs are drawn using S-PLUS 6.0. The procedures can be applied easily for any data set.

Table 1. Simulated percentiles of the estimated distributions of α , λ , and $X_{L(i)}(i = 7, 8)$

p	0.005	0.025	0.05	0.5	0.95	0.975	0.995
α	0.9917	1.4822	1.8170	4.6386	10.0104	11.4392	14.8883
λ	0.0187	0.0275	0.0351	0.1156	0.2439	0.2721	0.3273
$X_{L(7)}$	0.6646	1.4824	2.0005	4.5326	5.5001	5.5383	5.5713
$X_{L(8)}$	0.2129	0.6788	1.0279	3.4811	5.0675	5.2447	5.4287

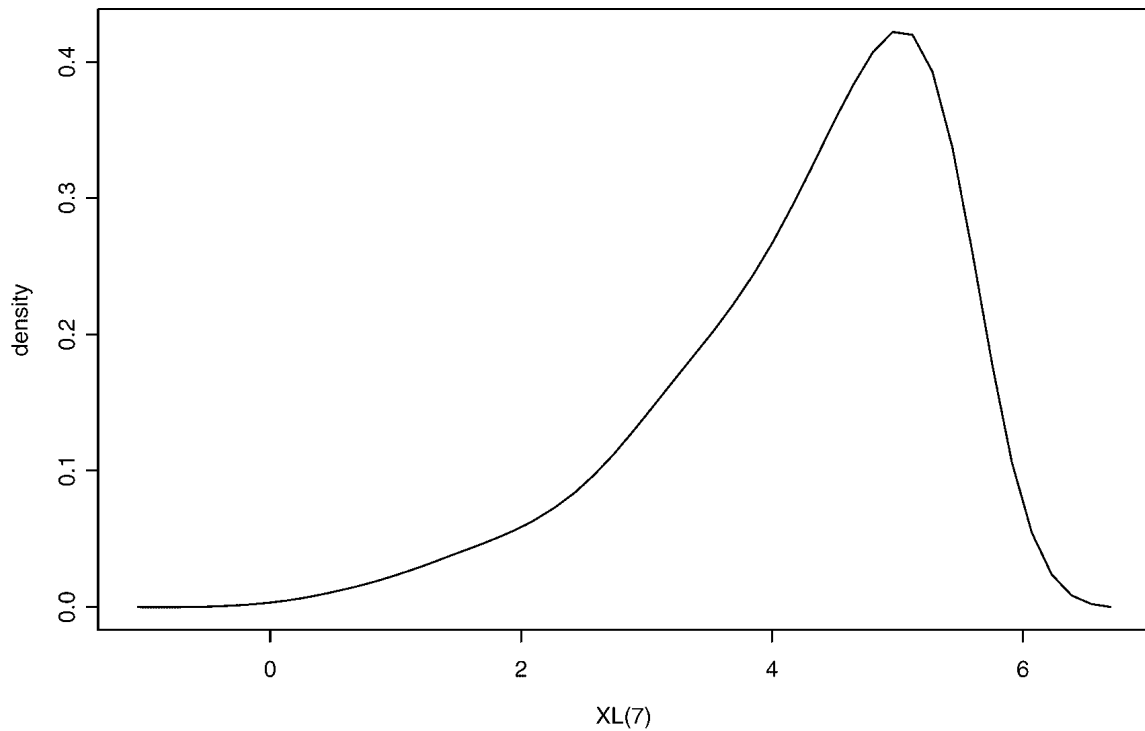


Figure 1. Estimate of the predictive density function of the next lower record

Example: In this example we present a data analysis of the amount of annual rainfall (in inches) recorded at the Los Angeles Civic Center for 127 years, from 1878 to 2005 (season July 1–June 30). We checked the validity of the GED model based on the parameters $\lambda = 0.15$ and $\alpha = 5.0$, using the Kolmogorov–Smirnov (K–S) test. It is observed that the K–S distance is $K - S = 0.0642$ with a corresponding p -value = 0.6718. This indicates that the GE model provides a good fit to the above data.

We used the lower records up to 1959 which were as follows:

11.35 10.40 9.21 6.73 5.59 5.58

to predict the next two lower records set in 1960 and 2001 and which were 4.85 and 4.35.

Very small values were given to the prior parameters to reflect little prior information. Namely, we assumed that $a = c = 0.25$ and $b = d = 0.5$. Ten thousand λ values are generated from $G(n + c, n\bar{r} + d)$ and the numerators and the denominators of Equations (6), (7), and (8) are averaged with respect to these simulated values. The resulting Bayes estimates for λ and α and the reliability at $x_0 = 0.5$ (for example) are found to be:

$$\lambda_B = 0.1162, \quad \alpha_B = 4.912, \quad \Lambda_B(0.5) = 0.9527$$

To predict the next records, 10 000 draws of equally spaced variates were collected for the parameters α and λ as well as for the next lower records $R_i (i = 7, 8)$. From Table 1, we can see that 95% confidence intervals for these quantities are respectively (1.4822, 11.4392), (0.0275, 0.2721), (1.4824, 5.5383), and (0.6788, 5.2447).

Moreover, the kernel method as a method of density estimation is used to estimate the predictive density functions of the future records. For example, Figure 1 shows the predictive density of R_7 . Note that the 95% prediction intervals for the next two lower records include the real values.

ACKNOWLEDGMENTS

The authors are thankful to two referees for their valuable comments that improved the original version of the manuscript.

REFERENCES

- Ahsanullah M. 1980. Linear prediction of record values for the two parameter exponential distribution. *Annals of the Institute of Statistical Mathematics* **32**: 363–368.
- Ahsanullah M, Bhatti MI. 2003. On prediction of Olympic records using extreme value distribution. *Journal of Statistical Theory and Applications* **2**: 85–95.
- Ahsanullah M. 2004. *Record Values-Theory and Applications*. University Press of America, Inc.: New York.
- Al-Hussaini, Ahmad AEA. 2003. On Bayesian interval prediction of future records. *Test* **12**(1): 79–99.
- Arnold BC, Balakrishnan N, Nagaraja HN. 1998. *Records*. Wiley: New York.
- Awad AM, Raqab MZ. 2000. Prediction intervals for the future record values from exponential distribution: comparative study. *Journal of Statistical Computation and Simulation* **65**(4): 325–340.
- Basak P, Bagchi P. 1990. Application of Laplace approximation to record values. *Communications in Statistics- Theory and Methods* **19**(5): 1875–1888.
- Dunsmore IR. 1983. The future occurrence of records. *Annals of the Institute of Statistical Mathematics* **35**: 267–277.
- Gilks WR, Richardson S, Spiegelhalter DJ. 1995. *Markov Chain Monte Carlo in Practise*. Chapman & Hall: London.
- Glick N. 1978. Breaking records and breaking boards. *American Mathematical Monthly* **8**: 2–26.

- Gulati S, Padgett WJ. 1995. Nonparametric function estimation from inversely sampled record-breaking data. *Canadian Journal of Statistics* **23**: 359–368.
- Gupta RD, Kundu D. 1999. Generalized exponential distribution. *Australian and New Zealand Journal of Statistics* **41**(2): 173–188.
- Gupta RD, Kundu D. 2001a. Exponentiated exponential distribution, an alternative to gamma and weibull distributions. *Biometrical Journal* **43**(1): 117–130.
- Gupta RD, Kundu D. 2001b. Exponentiated exponential distributions, different methods of estimations. *Journal of Statistical Computation and Simulation* **69**(4): 315–338.
- Jaheen ZF. 2003. A Bayesian analysis of record statistics from the Gompertz model. *Applied Mathematics and Computation* **145**: 307–320.
- Jaheen ZF. 2004. Empirical Bayes Inference for generalized exponential distribution based on records. *Communications in Statistics-Theory and Methods* **33**(8): 1851–1861.
- Madi MT, Raqab MZ. 2004. Bayesian prediction of temperature records using the Pareto Model. *Environmetrics* **15**: 701–710.
- Malinowska I, Szynal D. 2004. On a family of Bayesian estimators and predictors for a Gumbel model based on the k th lower records. *Applicationes Mathematicae* **31**: 107–115.
- Raqab MZ. 2001. Optimal prediction intervals for the exponential distribution based on generalized order statistics. *IEEE Transactions on Reliability* **50**: 112–115.
- Raqab MZ. 2002. Inferences for generalized exponential distribution based on record statistics. *Journal of Statistical Planning and Inference* **104**(2): 339–350.
- Raqab MZ. 2006. Nonparametric prediction intervals for the future rainfall records. *Environmetrics* **17**(5): 457–464.
- Smith RL, Miller JE. 1986. A non-Gaussian state space model and application to prediction of records. *Journal of the Royal Statistical Society Series B* **48**: 79–88.
- Tierney L. 1994. Markov chains for exploring posterior distributions. *Annals of Statistics* **22**: 1701–1762.
- Tierney L, Kadane JB. 1986. Accurate approximations for posterior moments and marginal densities. *Journal of American Statistical Association* **81**: 82–86.