January 1993

Simulation of Daily Precipitation from a Nonparametric Renewal Model

Balaji Rajagopalan

Upmanu Lall

David G. Tarboton

Utah Water Research Laboratory

Simulation of Daily Precipitation from A Nonparametric Renewal Model

Balaji Rajagopalan, Upmanu Lall
and
David G. Tarboton

Working Paper WP-93-HWR-UL/003

November 1993

Utah Water Research Laboratory, Utah State University, Logan UT 84322-8200, USA
(801)-750-3155, FAX : (801)-750-3663

SIMULATION OF DAILY PRECIPITATION FROM A NONPARAMETRIC RENEWAL MODEL

Balaji Rajagopalan, Upmanu Lall and David G. Tarboton
Utah Water Research Laboratory
Utah State University, Logan, UT - 84322-8200

Abstract

Wet/dry spell characteristics of daily precipitation are of interest for a number of hydrologic applications (e.g., flood forecasting or assessment of erosion potential). Here, we examine issues related to designing an appropriate nonparametric scheme that focuses on spell characteristics for resampling historical daily precipitation data. A subset of the nonparametric wet/dry spell model presented in Lall et al (1993) is tested with synthetic data to justify the strategy proposed for applications. An application of the nonparametric wet/dry spell model to a Utah data set follows. Performance is judged on a set of statistical measures. Numerical comparisons of model performance with a parametric alternating renewal model are also offered for this data set. Our presentation stresses data exploratory aspects rather than formal hypothesis testing.

1. INTRODUCTION

Many may be the unknown gods of daily precipitation at a site of interest. With this refrain, we (Lall et al (1993)) motivated our pantheistic (a.k.a. nonparametric) approach to a stochastic model for daily precipitation. The salient features of this model were the consideration of alternating wet and dry spells and of a daily rainfall structure within the wet spell. Kernel density estimates (k.d.e.'s) were espoused as effective methods for recovering all univariate, multivariate or conditional, discrete and/or continuous probability densities that were needed directly from the historical record. A kernel based procedure for disaggregating spell precipitation into daily precipitation was also provided. Here, we complement the largely theoretical discussion in Lall et al (1993) with numerical, illustrative examples. The purpose of these examples is twofold. First, we shall explore some of the issues relevant to the implementation of the proposed kernel density estimates. These are (a) the specification of the bandwidth of the kernel estimator for the continuous case, (b) the role of boundary effects in kernel estimation, and (c) the selection of the estimator in the discrete case. The intent is to justify our recommended procedures by example, and to provide a comparison of the major estimation schemes available in the literature. Second, we pursue an application with a daily precipitation record from a station in Utah to evaluate the performance of the proposed model through Monte Carlo simulation relative to historical statistics. The relative performance of traditional models such as a first order, two state Markov chain (MC) and an alternating renewal model with the same data set was also of interest. We recognize from Trivedi (1982, p.325) that if we define an 'event' to be a change of state (in our case wet to dry or dry to wet), then the successive interevent times (i.e the wet spell length or the dry spell length) of a Markov chain are independent, geometrically distributed random variables. An alternating renewal model with geometrically distributed wet and dry spells is thus equivalent to a Markov Chain. Consequently we offer comparisons of the Nonparametric Renewal model (NPR) with Parametric alternating renewal model (PAR), that uses geometric wet and dry spell length distributions and an exponential distribution for the precipitation amount each wet day.

To keep this presentation simple and informative we consider a subset of the Lall et al (1993) model. Wet and dry spells are assumed to be independent, allowing the use of a univariate description for the spell length distributions. Daily precipitation amounts within a wet spell are also assumed to be independent. This leads to a univariate description of wet day precipitation amount. Correlation statistics computed for the data sets analyzed supported these assumptions.

We begin with a brief description of the resulting model (section 2). Examples motivating our recommendations for model estimation follow in section 3. Performance measures and an

experiment design to test application(s) is presented next in section 4. The utility of the model developed is then illustrated through an application to data collected at Woodruff, Utah in section 5. Musings on the results and pointers to related work in progress conclude the presentation.

2. NONPARAMETRIC RENEWAL MODEL (NPR)

The year is divided into four seasons:- (1) Winter (January - March), (2) Spring (April - June), (3) Summer (July - September), and (4) Fall (October - December). The precipitation process is assumed to be stationary within these seasons. The random variables of interest are the wet spell length, w days, dry spell length, d days, and wet day precipitation amount, p inches. Variables w and d are defined through the set of integers greater than 1 (and less than season length), and p is defined as a continuous, positive random variable. Precipitation data are usually rounded to measurement precision (e.g., 0.01 inch increments). We do not expect the effect of such quantization of the data to be significant relative to the scale of the precipitation process, and treat precipitation as a continuous random variable. A mixed set of discrete and continuous random variables is thus considered.

The model is applied to daily precipitation for each season. The probability density functions (p.d.f.'s) of wet day precipitation amount f(p) and the probability mass functions (p.m.f.'s) of wet spell length f(w) and dry spell length f(d) are estimated for each season using kernel density estimates based on the observed data. These estimates comprise the model.

Synthetic precipitation sequences are generated following the strategy indicated in Figure 1. A dry spell is first generated using f(d). Then a wet spell is generated using f(w). Precipitation for each of the w wet days is then generated from f(p). The process is repeated with the generation of another dry spell. If a season boundary is crossed, the p.d.f.'s used for generation are switched to those for the new season. This procedure continues until a synthetic sequence of the desired length has been generated. As is usual practice in such cases we discard the early part of each sequence generated to avoid startup bias.

For the discrete variables (w and d), the kernel estimator developed by Hall and Titterington (1987) (HT) is recommended. The HT kernel estimator for the p.m.f., $f_n(x)$, where x is either w or d, and n is the corresponding sample size is given as:

$$f_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} W((x - x_i)/h) \qquad (1)$$

where, $W((x - x_i)/h) = \frac{K((x - x_i)/h)}{\sum_{j=-\infty}^{j=\infty} \frac{1}{h} K(j/h)}$, with $h \geq 1$

and K(.) is any suitable continuous variate kernel function satisfying properties of positivity, integrating to unity, symmetry and finite variance (see Lall et al. (1993), Equation 3.2), with compact support over the interval [-1,1].

The bandwidth h is selected as a minimizer of a Least Squares Cross Validation (LSCV) function suggested in Hall and Titterington (1987), over a suitable range for h. We have used the bisquare kernel ($K(t) = 15/16(1-t^2)^2$) for our analysis. This estimator is defined over the set of integers. However wet and dry spell lengths are counting numbers (integers greater than 1). To avoid the problem of the estimator assigning probability to integers less than 0, the so called boundary problem, we have used boundary kernels developed by Dong and Simonoff (1994, in press) as given in Lall et al. (1993), Table 1. Section 3.1 compares three approaches to estimating discrete p.m.f's. The HT estimator proved best among these.

For the p.d.f of wet day precipitation, a kernel density estimator applied to log transformed data was used. The log transformation provides bandwidth adaptation(in real space), alleviating the need to choose variable bandwidths with heavily skewed data (such as wet day precipitation). The boundary bias issues are also alleviated through use of the log transform. The resulting estimator is given as:

$$f_n(p) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{hp} K\left(\frac{\ln(p) - \ln(p_i)}{h}\right) \qquad (2)$$

where, h is the bandwidth in log space and K(.) is the kernel function.

The bandwidth h is chosen for the log transformed data using a recursive approach of Sheather and Jones (1991) to minimize the Mean Integrated Square Error (MISE). This procedure is described in Lall et al. (1993), section 3.5. The Epanechnikov kernel ($K(t) = 3/4(1-t^2)$, with $t=(\ln(p)-\ln(p_i))/h$) was used. Comparisons of this strategy with kernel density estimators defined in real space using boundary kernels, and for different bandwidth selection methods are offered in the next section.

3. COMPARING CHOICES FOR KERNEL DENSITY ESTIMATION

We compared three nonparametric estimators for estimating the p.m.f. for the discrete

variables (w or d), including the one recommended in section 2. The other two estimators are formally defined in Appendix 1. We also fit a parametric (geometric) distribution. Two examples of the application of these estimators to data from known populations are presented in section 3.1. For the continuous case, we considered four methods of bandwidth selection and estimation in real and in log space. Two examples comparing these estimators to data from known populations are presented in section 3.2.

3.1 Discrete variable(s)

The three methods considered are (1) HT - a kernel estimator suggested by Hall and Titterington (1987), using a Bisquare kernel with boundary modifications by Dong and Simonoff (1994), and bandwidth selected by LSCV (2) WV - a kernel estimator suggested by Wang and Van Ryzin (1981) using a geometric kernel with bandwidth selected by minimizing global or local mean square error (MSE) developed by Wang and Van Ryzin (1981), (3) MPLE - a maximum penalized likelihood estimator developed by Simonoff (1983). We refer the reader interested in "average" performance and properties to Monte Carlo analyses demonstrating the effectiveness of each estimator and some comparisons with other estimators to the authors cited above. It is instructive to see how these estimators do where a single sample is available. Consequently, the thrust of the presentation here is on a comparison of these estimators with data sets generated specifically from situations that may be of interest in our particular context.

Two examples are provided. First we use a sample (D1) from a geometric p.m.f. with π=0.2. The second sample (D2), was generated from a mixture of two geometric p.m.f.'s defined as (0.3G(π=0.9)+0.7G(π=0.2)). In each case a sample of size 250 was used. For the rainfall data we subsequently analyse, sample sizes were typically around 400 to 500, based on 40 years of data for the season of interest. We also fitted a geometric distribution (GP) to D1 and D2 using the method of moments. Sample statistics and values of the key parameters in each case are summarized in table 1. The corresponding p.m.f.'s estimated by each method for D1 and D2 are shown in Figure 2. Note that results from WV with local bandwidths are not shown, since they are very close to those obtained using a global bandwidth.

The following observations are apparent from the figures:

1. The WV procedure does not smooth the sample proportions (r_i) properly. In most cases, there is very little smoothing. In cases where there is some smoothing (e.g., Figure 2a, in the range x=4 to 6), the resulting estimate is rather unsatisfactory, and is inconsistent with the underlying population. We feel that part of this behavior is due to the rapid "drop off" of weight associated

with the Geometric kernel, and part due to the method used for selecting the bandwidth h.

2. On the other hand, since the roughness penalty tries to make the p.m.f. uniform, MPLE emphasizes smoothness. Consequently, when the true p.d.f has a high second derivative (e.g., near the origin), MPLE has difficulty distinguishing between "true" curvature and observed variation. The resulting estimate often has a strong downward bias near the origin (Figure 2a). The MPLE is also sensitive to the value specified for x_{max}, the longest spell length considered. As x_{max} is increased, the downward bias at the origin is increased and the entire p.m.f. is "flattened".

3. The GP fit is very good (estimated π=0.1956) for D1 where the true distribution was geometric. As expected, a large bias is incurred near the origin for D2 (see figures 2b and 2d), where the estimated π was 0.2554.

4. Figures 2b and 2d indicate that HT is the best of the estimators considered here. The estimated p.m.f is smooth, and it also exhibits the least pointwise bias. The HT estimator automatically adapts to a large range of density variation, providing optimal smoothness in finite samples. The HT estimates are robust to outliers, as opposed to the parametric fits. To illustrate this see Figure 2e. Outliers were added at 45, 50, 75 and 100. These could be generated if the data were contaminated by a few large values (e.g. from a Geometric distribution with π=0.01). The fitted Geometric distribution, i.e (GP) is very much affected by the outliers and deviates from the true distribution, especially near the mode (i.e 1.). While the HT estimator still follows the data closely.

The rationale for recommending HT is obvious from the above observations. Further reassurance of its asymptotic performance in terms of MSE and its relative MSE efficiency (compared to the MLE) may be found in Dong and Simonoff (1994).

3.2 Continuous variable p

The most critical aspect of developing the k.d.e. in this situation is the specification of the bandwidth. A second factor is the need for specialized treatment near p=0 (i.e., the boundary problem). We consider six estimators for $f_n(p)$. These are: (1) (PR-N) parametric reference assuming the underlying probability density function to be $N(0,\hat{\sigma}^2)$, (2) (PR-M), parametric reference assuming the underlying probability density function to be a Gaussian mixture 0.5N(-2,1)+0.5N(2,1), (3) (PR-E) parametric reference assuming the underlying probability density function to be $Exp(\hat{\alpha})$, (4) (LSCV) Least squares cross validation, (5) (MLCV) Maximum likelihood cross validation, (5) (SJ) Sheather and Jones (1991) procedure, and (6) (SJL) Sheather Jones (1991) procedure applied to log transformed data. Table 2 summarizes the bandwidth

estimation procedures. Further details of these procedures are given in Lall et al. (1993). In the first three methods the term parametric reference means the bandwidth is chosen to be optimal with reference to an assumed underlying parametric distribution. The first five methods, which consider untransformed real space data also use Abramson's method (Lall et al, 1993, section 3.5) to specify a local rather than a fixed global bandwidth. Boundary kernels as defined by Müller (1991) were used to adjust the density estimates near the lower boundary ($p \geq 0$), but were not used during bandwidth estimation. The SJL procedure, eliminated the boundary problem and provides some local bandwidth adaption, so no local bandwidth adjustment and no boundary kernels were used.

Two examples are provided. First we sample (C1) from a Gaussian mixture(0.5N(-2,1)+0.5N(2,1)), to demonstrate estimatibility with location mixtures. The second sample (C2), was generated from an Exponential distribution with mean 0.15. In each case a sample of size 250 was used. For the wet day precipitation data we subsequently analyse, sample sizes were typically around 600 to 800, based on 40 years of data for the season of interest. Sample statistics and values of the key parameters in each case are summarized in table 3. The corresponding p.d.f.'s estimated by each method for C1 and C2 are shown in Figure 3. For data set C1 we used methods PR-N, PR-M,LSCV, MLCV, SJ, while, for data set C2 we used PR-E, LSCV, MLCV, SJ and SJL.

The following observations are apparent from the figures:

1. The parametric reference (PR) procedures work very well as expected when the assumed p.d.f. matches the underlying p.d.f. However, under mis-specification, performance suffers. In case of C1, the bandwidth from the true reference (PR-M) is 1.0, while from using the normal distribution (i.e mis-specification) as the reference (PR-N) the bandwidth is 1.76. This results in gross oversmoothing of the two modes present in C1 (see Figure 3a). The bandwidth from PR is the best possible estimate of h provided f(x) is known. Of course, one reason we pursue nonparametric estimates of the p.d.f. is lack of knowledge of the underlying model. In this case PR estimates with the correct f(x) are useful as a benchmark to compare the performance of fully data driven methods.

2. LSCV and MLCV are prone to undersmoothing especially when the data exhibits fine structure (e.g bimodes) and is long tailed (see, Hall and Marron (1987)). Also the cross-validation functions (which are minimized for the bandwidth estimation) have spurious local optima (corresponding to clustering of data at different scales) at small values of bandwidth, (see Hall and Marron (1991)). Thus, we expect small bandwidths from LSCV and MLCV which leads to an undersmoothed density estimate. This can be seen from Figures 3b and 3d, where the estimates from LSCV and

MLCV are very rough suggesting that the variance is high.

3. SJ has been shown to have a better mean integrated square error (MISE) convergence rate than CV methods (see Sheather and Jones (1991)) and hence should lead to a better estimate. This is borne out in figures 3a and 3e, and table 2. Note that the SJ optimal bandwidth for C1 is close to the PR optimal bandwidth where the use of PR is justified, while for C2 it is not. This is due to the fact that the boundary effect is not considered while estimating the bandwidth, which is a problem in case C2 but not in C1. In both cases the SJ bandwidth is superior to those chosen by MLCV and LSCV.

Note that in all these cases, the optimal h is determined without using the boundary kernels, and is perhaps smaller than it would be (to reduce the effect of leakage across the boundary) if boundary kernels were used during bandwidth estimation. This emphasizes the need for proper treatment of the boundary of the domain during all phases of k.d.e.. We expect to pursue modifications of the SJ estimator to account for boundaries during bandwidth selection.

4. For C2, in Figures 3c and 3d, we use the Müller boundary kernels (except when using SJL) to reduce the bias at the boundary. Despite this a considerable bias can be observed near the origin in these figures, for each of these estimators. This is a consequence of the high curvature of the target density near the origin, and the "leakage" from the kernels across the boundary at x=0. Figure 3e for the case C2 includes a p.d.f estimated without using boundary kernels (SJ-NBK) along with those from SJ and SJL. The inclusion of boundary kernels in SJ offers only a marginal improvement over SJ in this case, since it still suffers from a bias due to the high curvature of f(x) in this area. SJL, on the other hand does not suffer as much from this problem and hence, performs better.

5. For data sets with a heavy concentration of data near the origin, a log transformation is an attractive choice. We see from Figure 3e that the SJL procedure provides a very competitive k.d.e in this situation. Note that SJL provides local bandwidth adaptation in real space. For the wet day precipitation data, that is usually modeled using an Exponential, or a Gamma distribution, this may be a natural transformation to consider.

Our recommendation of SJL is motivated largely by a desire to deal with the boundary effect issue and local bandwidth adaptation in a natural way given the nature of the precipitation data. Where boundary effects are not of concern (e.g., C1) a direct application of SJ would be preferred. Once a modification of SJ to account for boundary effects during bandwidths estimation is successful, SJL need not be the method of choice even in this situation.

4. PERFORMANCE MEASURES AND EXPERIMENT DESIGN

To demonstrate the utility of the NPR model for simulation of daily precipitation, the model was applied to daily rainfall data from the station Woodruff in Utah. Forty two years of daily rainfall data was available from the period 1948-1989. Woodruff is at 41°32'N latitude, 111°9'W longitude and at an elevation of 6320 ft. Most of the precipitation comes in the form of snow. Rainfall occurs mainly in Spring, with some in Fall. We can see from Table 4 that season 3 (Summer) has the highest mean wet day precipitation and maximum wet day precipitation, while season 2 (Spring) has the highest percentage of yearly precipitation. Season 2 (Spring) has the highest average wet spell length and the longest wet spell length. For the dry spells, season 3 (Summer) has the highest average dry spell length and the longest dry spell length. It can be observed from Table 4 that Seasons 2 and 3 (i.e Spring and Summer) are the active seasons and show marked persistence. Usually Winter and Fall are the active seasons in this region (mainly due to snow fall). The unusual behaviour seen at Woodruff is due to the fact that it is in the rain shadow region. Convective precipitation in summer seems to be the dominant mechanism.

We shall first list some measures of performance that can help us judge the utility of the model, and then outline the experimental design.

4.1 Performance measures

The following statistics were considered to be of interest in comparing the historical record and the model simulated record.

1. Probability distribution function of wet spell length, dry spell length and wet day precipitation in each season.
2. Mean of wet spell length, dry spell length and wet day precipitation in each season.
3. Standard deviation of wet spell length, dry spell length and wet day precipitation in each season.
4. Length of longest wet spell and dry spell in each season
5. Maximum wet day precipitation in each season.
6. Percentage of yearly precipitation in each season.
7. Fraction of wet and dry days in each season.

4.2 Experiment design

Our purpose here is to test the utility of the NPR model in modeling the daily precipitation process. The main steps involved in accomplishing this are:

1. Each year is divided into four seasons and the wet and dry spells for each season are determined from the daily precipitation data. Spells that cross seasonal boundaries are truncated at the season boundary and included in the appropriate seasons. We recognize that this could have the effect of introducing a small bias in the spell characteristics for a given season. Missing data are skipped, and the spell count is restarted with the next event.
2. Probability density/mass functions are fitted for the wet day precipitation, wet spell lengths and dry spell lengths for each season using the kernel estimators recommended in section 2.
3. Twenty five synthetic records of forty two years each (i.e the historical record length) are simulated using the NPR model .
4. The statistics of interest are computed for each simulated record, for each season and compared to statistics of the historical record using boxplots.

Comparisons with a parametric alternating renewal (PAR) model are also made. Here, Geometric distributions are fitted for the wet spell length and dry spell length for each season, and an Exponential distribution is fitted for the precipitation amount. Simulations are made from the fitted parametric distribution functions, following the scheme in Figure 1. Once again, twenty five records of forty two years are simulated. The statistics listed in section 4.1 are computed for the simulated record and compared with those of the historical record and those from NPR simulations.

5. RESULTS

In this section we present comparative results (using the performance measures listed in section 4.1) of the NPR and PAR model for the Woodruff data. The statistics of interest calculated from the simulations are compared with those for the historical record using boxplots. The box in the boxplots (e.g Figure 5) indicates the interquartile range of the statistic computed from twenty five simulations, the line in the middle of the box indicates the median simulated value. The solid lines correspond to the statistic of the historical record. The boxplots show the range of variation in the statistics from the simulations and also show the capability of the simulations to reproduce historical statistics. The plots of the pdf's and pmf's are truncated to show a common range across seasons and to highlight differences near the origin (mode).

Figure 4 shows that fitted kernel densities for wet day precipitation amount are close to the

histogram in all the four seasons. They differ from the fitted exponential distribution, particularly in season 3 (Summer). These differences are manifested in the simulations. Simulations based on the NPR model reproduce the observed standard deviations and maximum daily precipitation well (Figure 5) while simulations from the parametric (Exponential distribution) model are too low in these statistics (Figure 6). The mean wet day precipitation and the fraction of yearly precipitation are comparable in both cases.

Figure 7 shows that the wet spell length p.m.f.'s estimated by HT and the fitted geometric distribution are very close. In this case one could argue for using the Geometric distribution rather than HT. But the "loss" in using HT is small and for uniform application across sites, HT may still be a better choice. Figures 8 and 9 show that for the mean wet spell length, standard deviation of wet spell length and for the fraction of wet days per season, results between simulations from HT and the Geometric distribution are similar with somewhat tighter boxes for the Geometric distribution, as is to be expected. However the longest wet spell length simulated by the Geometric distribution has greater variation. The Geometric distribution and the HT behaviour is really quite close overall. Differences are largely accounted for by the fact that a fixed bandwidth, rather than a local bandwidth is used in HT, as was discussed in section 4.1. This leads to increased variation in the simulations for small wet spell values (near the mode) and reduced variance in the tails (longest wet spell lengths). We expect local bandwidths to be larger in the tails and smaller near the mode compared to the fixed bandwidth used now.

Figure 10 shows that the dry spell length p.m.f.'s estimated by HT and the fitted Geometric distribution are generally similar with the most difference in season 3 (Summer), which we noted as being the most "active" with regard to dry spell length extremes. The difference between Geometric and HT p.m.f estimates is similar to the difference we saw in Figure 2d where these procedures were applied to the synthetic Geometric mixture data D2. The dry spell length data for season 3 (Summer) behaves as though it was from a mixed distribution.

Observationally we know that there are dry summers with little rainfall activity and other summers with intermittent, stagnating precipitation systems in this area. Thus we would expect a mixture mechanisms generating dry spells to show up in this season. Of course if this structure is really present for a while season our model with independencies of sequential dry spells would not properly reproduce it. A more general model that considers dependence between sequential spell length (Lall et. al. 1993) may also be useful in such cases.

The reader might be tempted to suggest formal tests to check for a mixture of the geometric distributions in this case as an alternative to the kernel density estimate. While this may be a fruitful activity (we did consider it), it gets harder to perform and/or justify as we consider arbitrary, finite component mixtures. An advantage of the kernel density estimator (HT) employed here is that it

readily admits such mixtures without requiring that they be hypothesized or formally identified. We feel that this provides a more direct and parsimonious representation of this sort of structure if present in the data.

Figures 11 and 12 show that in terms of summary statistics both models perform equally well. These summary statistics do not appear to be stringent discriminators between models. The p.m.f.'s of wet spell length, dry spell length and p.d.f.'s of the wet day precipitation of the historical record are well reproduced by the simulations (figures are not shown for brevity).

6. SUMMARY AND CONCLUSIONS

A subset of the Nonparametric wet/dry spell model described in Lall et.al (1993) was compared with an alternating renewal model for real precipitation data. The primary difference here was in the parametric versus nonparametric estimation (using kernel density estimators) of the p.d.f/p.m.f needed. Issues in estimating parameters for continuous and discrete kernel density estimators were discussed and recommended procedures were developed through examples.

Three estimators were tested for the discrete p.m.f estimation (namely, the Hall and Titterington (1989) estimator (HT), Wang and Van Ryzin (1981) estimator (WV) and Maximum Penalized Likelihood Estimator, Simonoff (1983) (MPLE)). The HT estimator was recommended. Several methods for the continuous kernel estimation of p.d.f.'s in our context were also tested and we recommend the Sheather and Jones (1991) (SJ) procedure when there was no boundary and the Sheather and Jones (1991) procedure on log transformed data (SJL) when data was clustered near the boundary (the case for wet day precipitation). The NPR model was implemented with HT and SJL procedures and compared with an alternating renewal model with Geometric and Exponential distributions.

We found that where the parametric procedure was appropriate, the nonparametric procedure worked nearly as well. Where the parametric model was inappropriate, the nonparametric kernel density estimators were suitable. Given that the nonparametric procedures are robust and reproduce different parametric alternatives without prior assumptions, they offer a very general procedure for uniform application across a variety of sites and processes. Simulations effectively demonstrated the ability of kernel methods to reproduce moments as well as historical extremes.

Problems with kernel density estimates are high relative bias and variance in the tail of the density if local adaption of the bandwidth is not used. Ability to extrapolate is limited to one bandwidth of the maximum observed value. Where a local bandwidth is used, the local bandwidth

at the extreme point of observation is usually quite large and this problem is ameliorated.

We are working on improved kernel methods for estimation of probability mass function of discrete variables, with p.m.f's of the type considered here. Initial results from our new univariate kernel density estimator are quite promising, and work on a bivariate estimator is in progress.

Thus, the nonparametric modeling framework provides a promising alternative to parametric approaches. The assumption free, data adaptiveness and robust nature of the nonparametric estimators makes the model attractive in a broad class of situations. Our analyses also demonstrated the utility of the approach for judging the performance of a candidate parametric model with simulations from a purely data driven resampling procedure.

A number of issues of interest to stochastic precipitation modelers were not discussed here. The foremost is the behaviour of the proposed model at different time scales. We view our developments as "operational" and relevant to the time scale of the data which was daily. Spell definitions are tenuous at best at finer time scales and sample sizes drop rapidly as longer time scales (e.g. monthly or annual) are considered. Thus while the scaling issue is of theoretical and practical interest, it is difficult to formally assess how such a model may fit in. It is an issue we expect to explore in due course. A second issue is the need to incorporate climatic or precipitation "types" (e.g. Bogardi et al 1993, Wilson and Lattenmaier, 1993) into the daily precipitation model. We feel that implicit consideration of some of these factors is provided by our model by admitting an arbitrary mixture of generating mechanisms. Transitions between generating mechanisms are not explicitly modeled. However, their relative frequencies ought to be reproduced. Given limited data sets and the relatively large number of generating mechanisms (e.g. related to Pacific North American (PNA) pattern, North American Oscillation (NAO), El-Nino etc.) this may be all that is reliably feasible in a number of cases. Finally, there is the question of regionalization and /or portability of the method. The nonparametric approach clearly enjoys broader applicability than its parametric competitors. On the other hand, it may be less amenable to direct regionalization as is sometimes done in terms of the parameters of a parametric distribution. It is meaningless to talk of a regional bandwidth. It may be more fruitful to develop a space-time nonparametric precipitation model with a nonhomogenous point process structure that is inferred from the data. Preliminary work on such a model is in progress.

ACKNOWLEDGEMENTS

Partial support of this work by the U.S. Forest Service under contract notes, INT-915550-RJVA and INT-92660-RJVA, Amend #1 is acknowledged. We are grateful for discussions with D.S. Bowles, the principal investigator of the project. We thank S.J. Sheather and J.Simonoff for providing codes for implementing the SJ procedure and HT estimator with boundary modification respectively. Finally, we thank H.G. Muller, J. Dong, M.C. Jones and M. Wand for stimulating discussions and provision of relevant manuscripts. The work reported here was also supported in part by the USGS through their funding of the second author's 1992-93 sabbatical leave, when he worked with BSA,WRD,USGS, National center, Reston, VA.

Table 1

Statistics (Sample size =250 for each) and methods for Figure 2

Figure	Data	Estimator	Kernel used	Method of Bandwidth Selection
2a	D1 ($\bar{x}$ = 4.8, s = 4.23)	WV	Geometric kernel	MSE
		MPLE	- - -	- - -
2b	D1	HT	Bisquare kernel	LSCV
2c	D2 ($\bar{x}$ = 3.92, s = 4.02)	WV	Geometric kernel	MSE
		MPLE	- - -	- - -
2d	D2	HT	Bisquare kernel	LSCV

Note: $\bar{x}$ is sample mean and s is sample standard deviation

Table 2

Choices of bandwidth selection for kernel estimators of continuous variables

Method	Equation	Criteria/Remarks
PR-M	$h_{opt} = 3.03n^{-0.5}$	Based on minimization of MISE, given Epanechnikov kernel and assuming underlying probability density function to be 0.5N(-2,1)+0.5N(2,1).
PR-N	$hopt = 2.13\hat{\sigma}n^{-0.5}$	Based on minimization of MISE, given Epanechnikov kernel and assuming the underlying probability density function to be $N(0,\hat{\sigma}^2)$. $\hat{\sigma}$ is the sample standard deviation.
PR-E	$hopt = 1.97\hat{\sigma}n^{-0.5}$	Based on minimization of MISE, given Epanechnikov kernel and assuming the underlying probability density function to be $Exp(\hat{\alpha})$. $\hat{\sigma}$ is the sample standard deviation.
LSCV	$LSCV(h) = \int \hat{f}^2 - 2n^{-1}\sum_{i=1}^{n} \hat{f}_{-i}(x_i)$	Based on minimizing the LSCV(h) function. $\hat{f}_{-i}$ represents the k.d.e constructed by dropping the i^{th} observation.
MLCV	$MLCV(h) = n^{-1}\sum_{i=1}^{n} \log(\hat{f}_{-i})$	Based on maximizing the MLCV(h) function.
SJ	refer to equations, 3.18-3.20, of Lall et.al (1993)	Based on recursive estimation of MISE
SJL	Same as SJ, but applied to log transformed data	

Note: (For details on these methods, see Lall et.al (1993) section 3.5)

PR Parametric reference

LSCV Least squares cross validation

MLCV Maximum likelihood cross validation

SJ Sheather and Jones (1991) procedure

SJL Sheather and Jones (1991) procedure applied to log transformed data

MISE Mean integrated squared error

Table 3

Statistics (Sample size =250 for each) and methods for Figure 3

Data	Method (corresponding to Appendix 2)	Global Bandwidth
C1 ($\bar{x}$ = 0.00, s = 2.26)	PR-M	1.00
	PR-N	1.76
	LSCV	0.48
	MLCV	0.53
	SJ	1.03
C2 ($\bar{x}$ = 0.16, s = 0.18)	PR-E	0.11
	LSCV	0.015
	MLCV	0.02
	SJ	0.04
	SJL	0.77 (in log space)

Note:

$\bar{x}$ is sample mean and s is sample standard deviation

The SJL estimator is, (Equation 2)

$f_n(p) = \frac{1}{n}\sum_{1}^{n} \frac{1}{hp} K(\frac{\ln(p) - \ln(p_i)}{h})$ with Epanechnikov kernel

The Parametric reference, LSCV, MLCV and SJ all use,

$f_n(p) = \sum_{i=1}^{n} \frac{1}{nh_i} K(\frac{x-x_i}{h_i})$ with Epanechnikov kernel and Müller boundary kernels.

Local bandwidths h_i are given by, $h_i = h(f(p_i)/g)^{-1/2}$, where h is global bandwidth, $f(p_i)$ is the kernel density estimate at p_i using the global bandwidth h and g is the geometric mean of $f(p_i)$. These estimators only differ in the procedure used to obtain global bandwidth.

Table 4

Statistics from the historical precipitation record at Woodruff, UT, 1948-1989

Statistic	Season 1 (Jan - Mar)	Season 2 (Apr - Jun)	Season 3 (Jul - Sep)	Season 4 (Oct - Dec)
Avg. wet spell length	1.39 days	1.90 days	1.59 days	1.55 days
Std. dev. of wet spell length	0.81 days	1.36 days	1.06 days	0.90 days
Fraction of wet days	0.25	0.28	0.23	0.24
Longest wet spell length	6 days	10 days	8 days	7 days
Avg. dry spell length	4.12 days	4.82 days	5.45 days	5.08 days
Std. dev. of dry spell length	3.91 days	4.46 days	5.58 days	4.88 days
Fraction of dry days	0.75	0.72	0.77	0.76
Longest dry spell length	26 days	25 days	39 days	38 days
Avg. wet day precip.	0.09 in.	0.13 in.	0.15 in.	0.12 in.
Std. dev. of wet day precip.	0.10 in.	0.16 in.	0.21 in.	0.15 in.
Fraction of yearly precip.	0.19	0.31	0.27	0.23
Max. wet day precip.	0.87 in.	1.43 in.	1.65 in.	1.11 in.

Appendix 1
Discrete Density Estimators

1. Wang and Van Ryzin (1981) estimator (WV)

A kernel estimator for p.m.f. of discrete variable x, (here the length of wet or dry spell with n sample values x_i) as given by Wang and Van Ryzin (1981) is:

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} K(x, x_i, h) \qquad \text{(A1)}$$

h is the bandwidth and K(.) is the geometric kernel given as:

$$K(x,x_i,h) = 0.5(1-h)h^{|x-x_i|} \quad \text{if } |x-x_i| \geq 1 \qquad h \in [0,1]$$

$$= (1-h) \qquad \text{if } x = x_i \qquad \text{(A2)}$$

The bandwidth h can be a constant parameter, in which case it is called a global bandwidth, or it can vary locally from point to point, denoted $h(x_i)$.

Wang and Van Ryzin (1981) derived optimal global and local bandwidths to minimize the MSE (Mean Square Error = $E[(f(x)-f_n(x))^2]$). They estimate the local bandwidths $h(x_i)$ by minimizing the approximate MSE of $f_n(x_i)$, by truncating the geometric kernel at $x_i \pm 2$. The resulting expressions are in terms of the unknown true probabilities $f(x_i)$. They show that substitution of the relative frequencies of x_i, estimated from the sample as r_i ($r_i = n_i/n$) in the expressions leads to a strongly consistent procedure. An optimal global bandwidth is obtained by minimizing the average MSE (i.e., $1/n\sum_i MSE(x_i)$) over the data. Expressions for the optimal global and local bandwidths are given below.

Global bandwidth $$h = \beta_1\{3/2 + B_1 - B_2 + (n-1)\beta_{10}\}^{-1} \qquad \text{(A3)}$$

Local bandwidth $$h(x_i) = d_i\{p_i + \frac{1}{4}E_i + F_i - G_i + (n-1)e_i\}^{-1} \qquad \text{(A4)}$$

where,

$$\beta_1 = 1 - \sum_i^n r_i^2 + \frac{1}{2}B_1,\ B_1 = \sum_i^n r_i(r_{i-1} + r_{i+1}),\ B_2 = \sum_i^n r_i(r_{i-2} + r_{i-2}),$$

$$b_{10} = \sum_i^n r_i^2 - B_1 + \frac{1}{4}B_0$$

$$G_i = r_i(r_{i-2} + r_{i+2}),\ F_i = r_i(r_{i-1} + r_{i+1}),\ E_i = (r_{i-1} + r_{i+1}),\ d_i = r_i(1 - r_i) + \frac{1}{2}F_i,$$

$$e_i = (r_i - \frac{1}{2}E_i)^2,\ B_0 = \sum_i^n (r_{i-1} + r_{i+1})^2$$

Note that for small values of h, the estimator is close to the naive maximum likelihood estimator (MLE) (i.e p_i), and for p_i small, h is larger, leading to a higher smoothing, or larger "smearing" of the relative frequencies. An improved extrapolation in the tail of the density results through the use of the local bandwidths.

2. Maximum Penalized Likelihood Estimator (MPLE)

The MPLE was first introduced by Good and Gaskins (1971) for continuous variables, and was later extended to the density estimation for discrete variables by Simonoff (1983). Simonoff (1983) proposes a solution for the "category" probabilities $\hat{f}_i$ that maximizes a penalty function given by,

L = Log likelihood - roughness penalty (A5)

The idea is to balance the goodness-of-fit of the estimate (i.e., likelihood) with its smoothness (i.e., roughness penalty). The smoothest estimate is obtained if all cell probabilities are equal over the range of cells considered. With this in mind, the penalized likelihood function is defined as:

$$L = \sum_{i=1}^{x_{max}} n_i \log(\hat{f}_i) - \beta \sum_{i=1}^{x_{max}} \{\log(\hat{f}_i/\hat{f}_{i+1})\}^2 \quad \text{(A6)}$$

where $\sum_{i=1}^{x_{max}} \hat{f}_i = 1,$ (A7)

$\beta \geq 0$, is a smoothing parameter, and x_{max} is the longest spell length considered.

The smoothing parameter β controls the relative weight assigned to smoothness and consequently has the same role as the bandwidth used in kernel estimation. Here a data dependent β is used through the following procedure which minimizes asymptotic mean square error.

1. An initial β is chosen as $0.009N(x_{max})^{0.6}(\log(x_{max}))^{0.4})$, where N is the sample size.
2. Given this β, the penalized likelihood (equation A6) is maximized with respect to $\hat{f}_i$, $i = 1,...,x_{max}$ using the method of Lagrange multipliers.
3. An optimal β is now estimated by minimizing an asymptotic MSE, defined as an asymptotic approximation to $\sum_{i=1}^{x_{max}} (\hat{f}_i - \pi_i)^2$. Simonoff (1983) develops this asymptotic MSE expression in terms of the sample relative frequencies r_i ($r_i = n_i/n$), β and the unknown probability π_i. For π_i he uses the estimates $\hat{f}_i$ from step 2.
4. Steps 2 and 3 are repeated till convergence is achieved.

Simonoff (1983) argues that although a formal proof of the convergence of this procedure is not available, extensive computations have indicated that the scheme does converge. The need to specify x_{max} (in excess of the longest observed spell) detracts from the use of this method. We would prefer a natural extension of the tail of the p.m.f. by the method used, rather than a prior specification of its extent.

Figure Captions

Figure 1: Structure of the renewal model used for daily precipitation.

Figure 2a: Plot of p.m.f's estimated from WV (h = 0.43), MPLE (β=30.25), the true underlying p.m.f and observed proportions, for the data set D1.

Figure 2b: Plot of p.m.f's estimated from HT (h=7), GP (p=0.1956), the true underlying p.m.f and observed proportions, for the data set D1.

Figure 2c: Plot of p.m.f's estimated from WV (h=0.08), MPLE (β=28.25), the true underlying p.m.f and observed proportions, for the data set D2.

Figure 2d: Plot of p.m.f's estimated from HT (h=3), GP (p=0.2554), the true underlying p.m.f and observed proportions, for the data set D2.

Figure 2e: Plot showing the effect of outliers on fitted Geometric distribution (GP), and HT estimate. Outliers at 45,50,75,100 in the data set D1.

Figure 3a: Plot of p.d.f's estimated from PR-M (h=1), PR-N (h=1.76), SJ (h=1.03), the true underlying p.d.f, observed data and histogram of observed data, for the data set C1.

Figure 3b: Plot of p.d.f's estimated from LSCV (h=0.48), MLCV (h=0.53), the true underlying p.d.f, observed data and histogram of observed data, for the data set C2.

Figure 3c: Plot of p.d.f's estimated from PR-E (h=0.11), SJ (h=0.04), SJL (h=0.77), the true underlying p.d.f, observed data and histogram of observed data, for the data set C2.

Figure 3d: Plot of p.d.f's estimated from LSCV (h=0.015), MLCV (h=0.02), the true underlying p.d.f, observed data and histogram of observed data, for the data set C2.

Figure 3e: Plot of p.d.f's estimated from SJ, SJL, and SJ-NBK (Band width chosen from SJ procedure but boundary kernels are not used). Along with true underlying p.d.f, observed data and histogram of observed data, for the data set C2.

Figure 4: Plots of p.d.f's of wet day precipitation for the four seasons at Woodruff, UT, estimated using SJL procedure, the fitted Exponential distribution and histogram of the observed data.

Figure 5: Boxplots of mean, standard deviation, fraction of yearly precipitation and maximum precipitation of wet day precipitation in each season, for simulations made from the NPR model along with the historical values.

Figure 6: Boxplots of mean, standard deviation, fraction of yearly precipitation and maximum precipitation of wet day precipitation in each season, for simulations made from the Alternating Renewal Model (i.e using fitted Exponential distribution).

Figure 7: Plots of p.m.f's of wet spell length for the four seasons at Woodruff, UT, estimated using HT estimator. Along with the fitted Geometric distribution and observed proportions.

Figure 8: Boxplots of mean, standard deviation, fraction of wet days and longest wet spell length

in each season, for simulations made from the NPR model along with the historical values.

Figure 9: Boxplots of mean, standard deviation, fraction of wet days and longest wet spell length in each season, for simulations made from the Alternating Renewal Model (i.e using fitted Geometric distribution).

Figure 10: Plots of p.m.f's of dry spell length for the four seasons at Woodruff, UT, estimated using HT estimator. Along with the fitted Geometric distribution and observed proportions.

Figure 11: Boxplots of mean, standard deviation, fraction of dry days and longest dry spell length in each season, for simulations made from the NPR model along with the historical values.

Figure 12: Boxplots of mean, standard deviation, fraction of dry days and longest dry spell length in each season, for simulations made from the Alternating Renewal Model (i.e using fitted Geometric distribution).

Figure 4

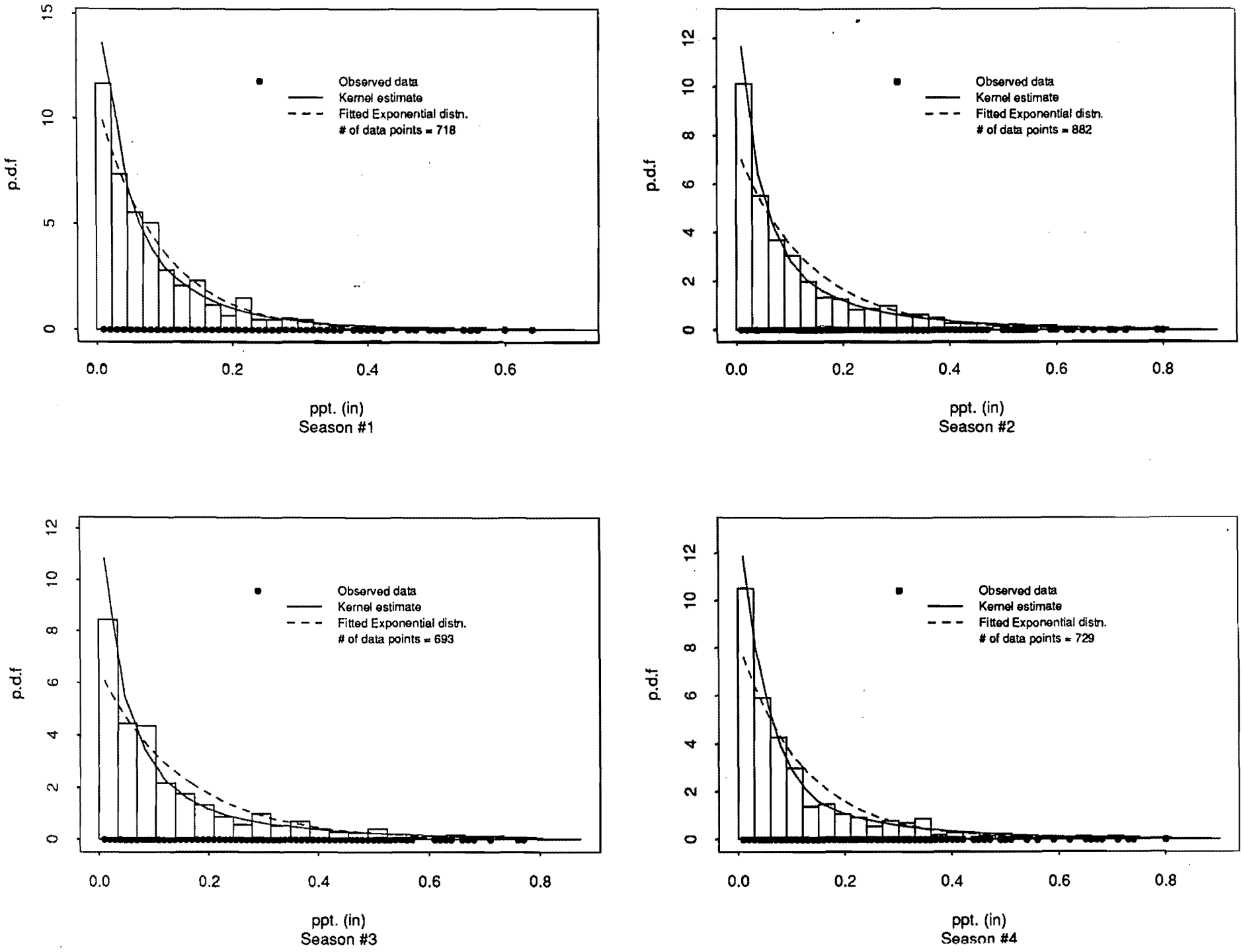

p.d.f
ppt. (in)
Season #1
Observed data
Kernel estimate
Fitted Exponential distn.
# of data points = 718
p.d.f
ppt. (in)
Season #2
Observed data
Kernel estimate
Fitted Exponential distn.
# of data points = 882
p.d.f
ppt. (in)
Season #3
Observed data
Kernel estimate
Fitted Exponential distn.
# of data points = 693
p.d.f
ppt. (in)
Season #4
Observed data
Kernel estimate
Fitted Exponential distn.
# of data points = 729

Figure 5

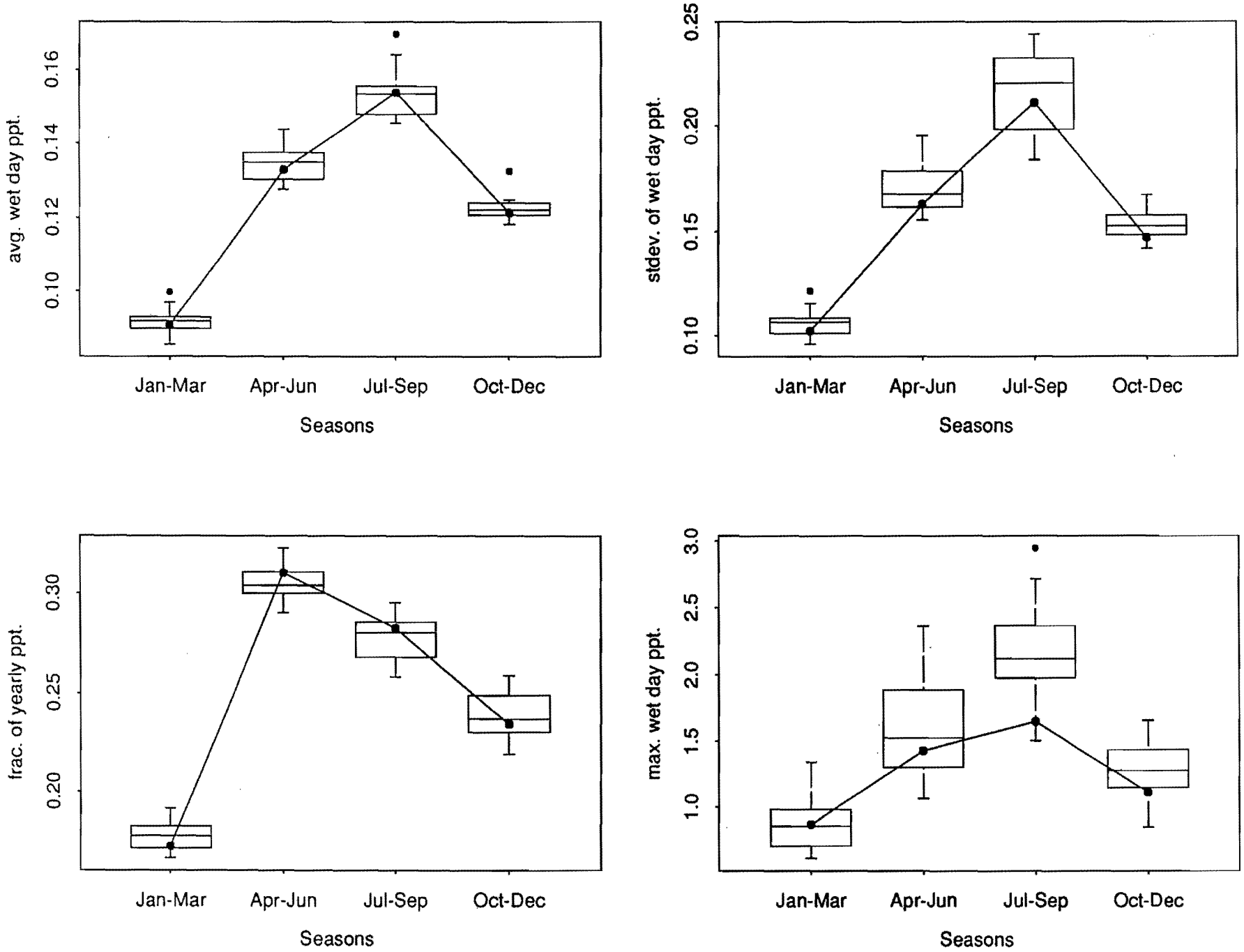

avg. wet day ppt.
0.10
0.12
0.14
0.16
Jan-Mar
Apr-Jun
Jul-Sep
Oct-Dec
Seasons
stdev. of wet day ppt.
0.10
0.15
0.20
0.25
Jan-Mar
Apr-Jun
Jul-Sep
Oct-Dec
Seasons
frac. of yearly ppt.
0.20
0.25
0.30
Jan-Mar
Apr-Jun
Jul-Sep
Oct-Dec
Seasons
max. wet day ppt.
1.0
1.5
2.0
2.5
3.0
Jan-Mar
Apr-Jun
Jul-Sep
Oct-Dec
Seasons

Figure 6

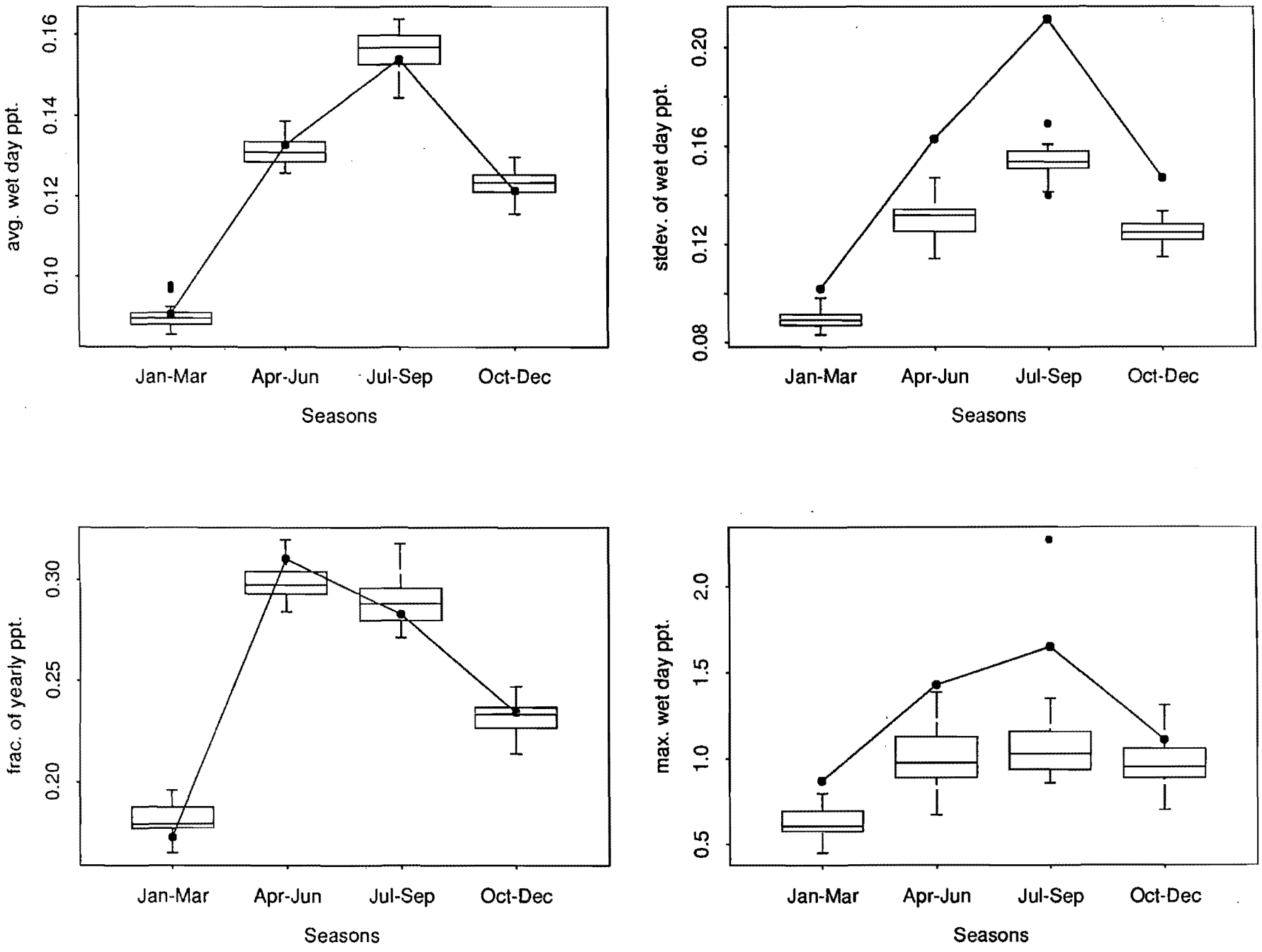

avg. wet day ppt.
0.10
0.12
0.14
0.16
Jan-Mar
Apr-Jun
Jul-Sep
Oct-Dec
Seasons
stdev. of wet day ppt.
0.08
0.12
0.16
0.20
Jan-Mar
Apr-Jun
Jul-Sep
Oct-Dec
Seasons
frac. of yearly ppt.
0.20
0.25
0.30
Jan-Mar
Apr-Jun
Jul-Sep
Oct-Dec
Seasons
max. wet day ppt.
0.5
1.0
1.5
2.0
Jan-Mar
Apr-Jun
Jul-Sep
Oct-Dec
Seasons

Figure 7

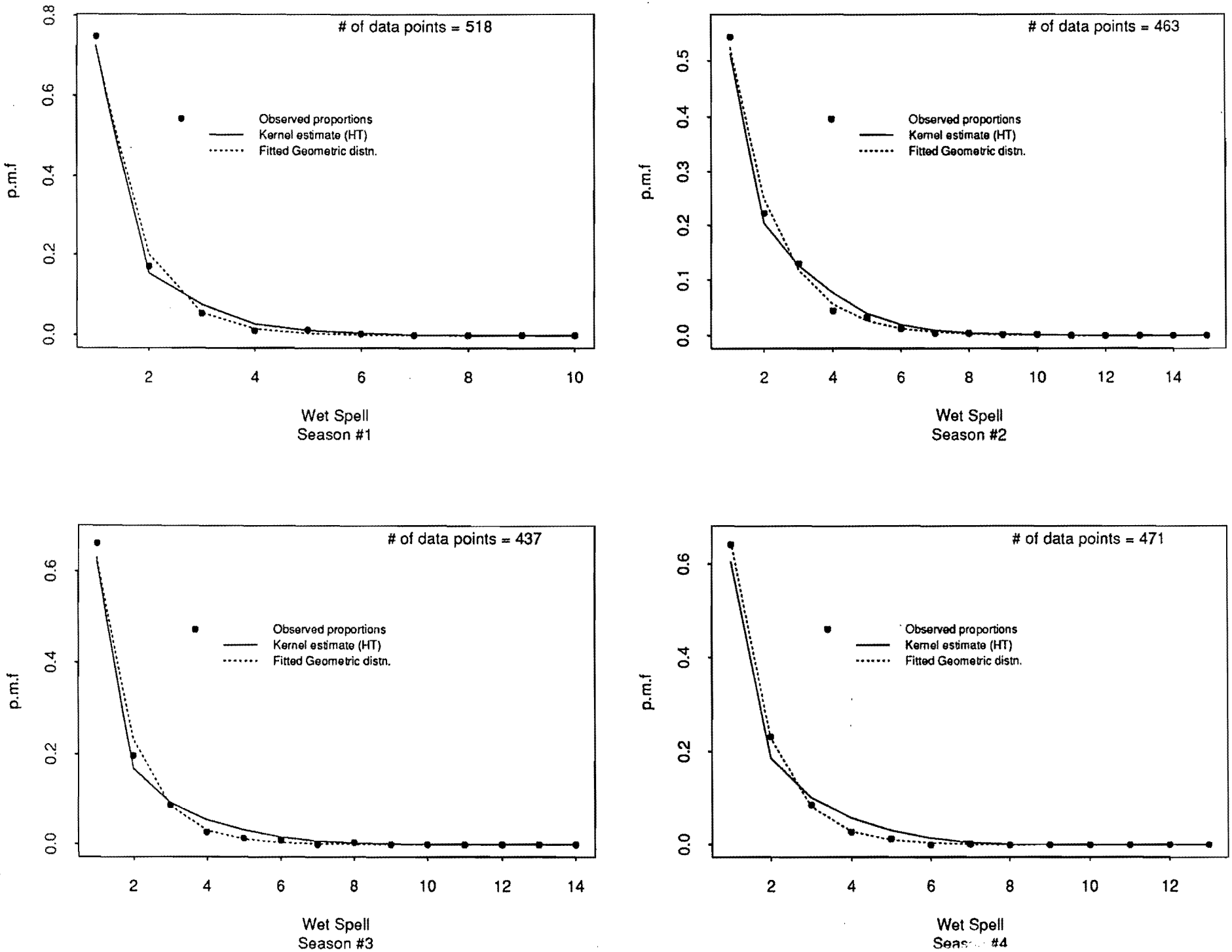
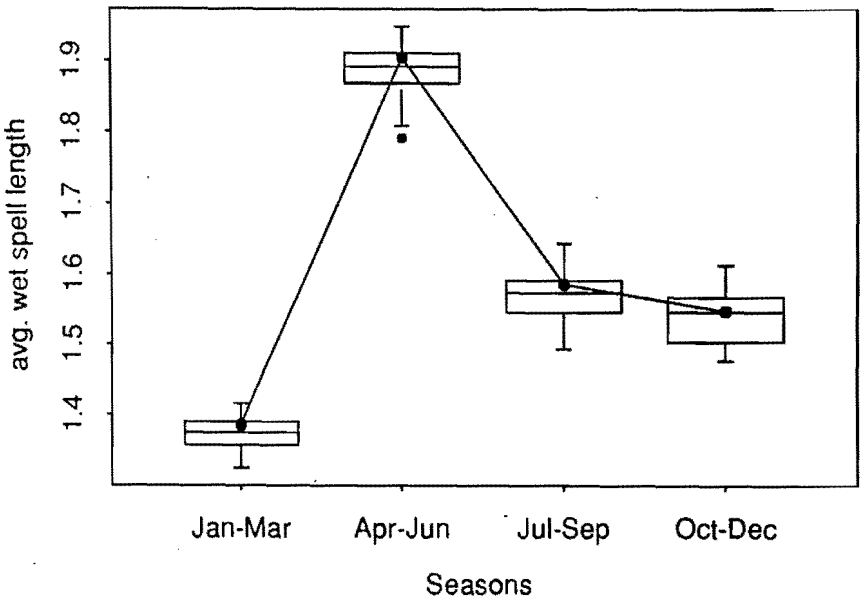
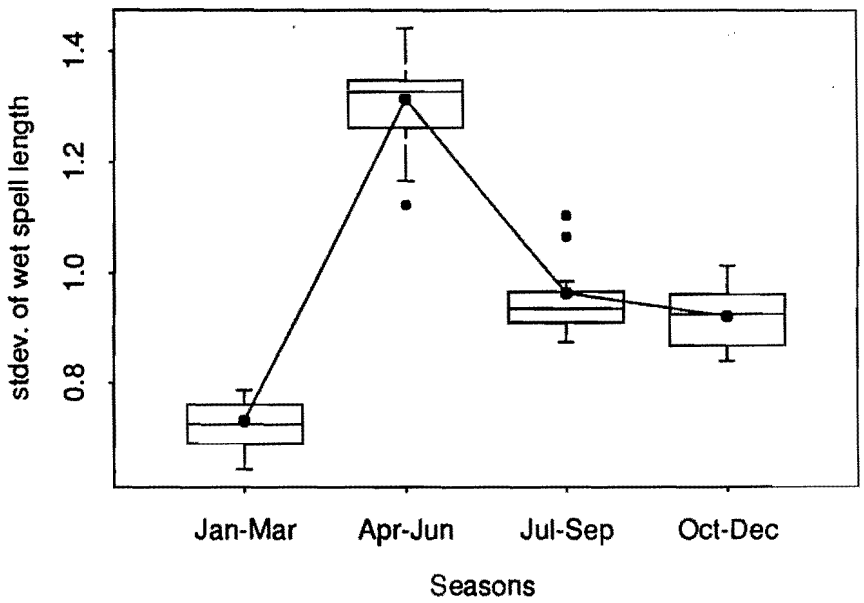
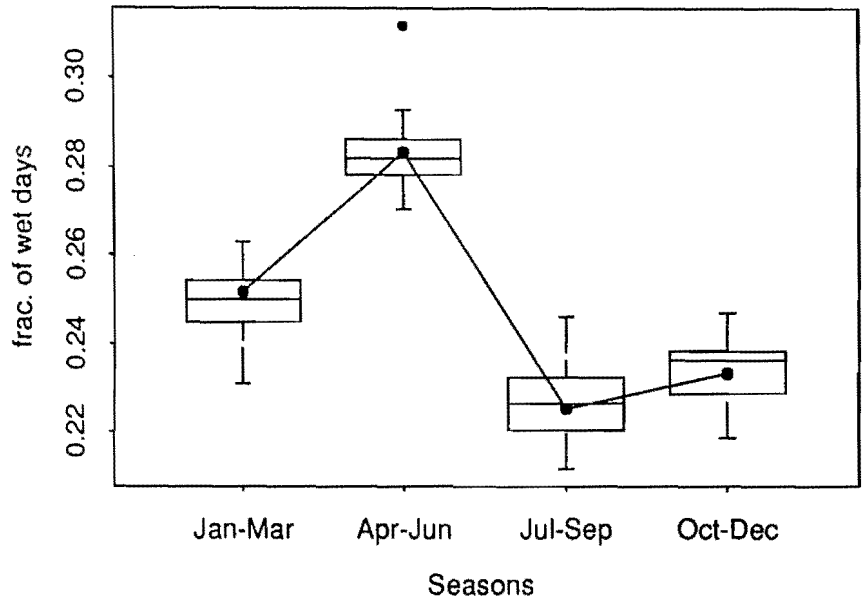

# of data points = 518
Observed proportions
Kernel estimate (HT)
Fitted Geometric distn.
p.m.f
Wet Spell
Season #1
# of data points = 463
Observed proportions
Kernel estimate (HT)
Fitted Geometric distn.
p.m.f
Wet Spell
Season #2
# of data points = 437
Observed proportions
Kernel estimate (HT)
Fitted Geometric distn.
p.m.f
Wet Spell
Season #3
# of data points = 471
Observed proportions
Kernel estimate (HT)
Fitted Geometric distn.
p.m.f
Wet Spell
Seas #4

Figure 8


avg. wet spell length
1.4
1.5
1.6
1.7
1.8
1.9
Jan-Mar
Apr-Jun
Jul-Sep
Oct-Dec
Seasons


stdev. of wet spell length
0.8
1.0
1.2
1.4
Jan-Mar
Apr-Jun
Jul-Sep
Oct-Dec
Seasons


frac. of wet days
0.22
0.24
0.26
0.28
0.30
Jan-Mar
Apr-Jun
Jul-Sep
Oct-Dec
Seasons

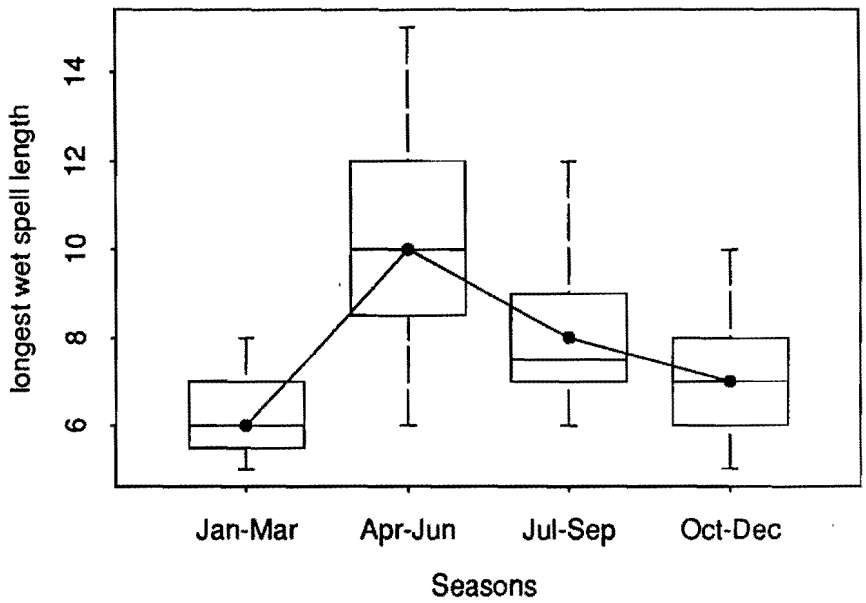

longest wet spell length
6
8
10
12
14
Jan-Mar
Apr-Jun
Jul-Sep
Oct-Dec
Seasons

Figure 9

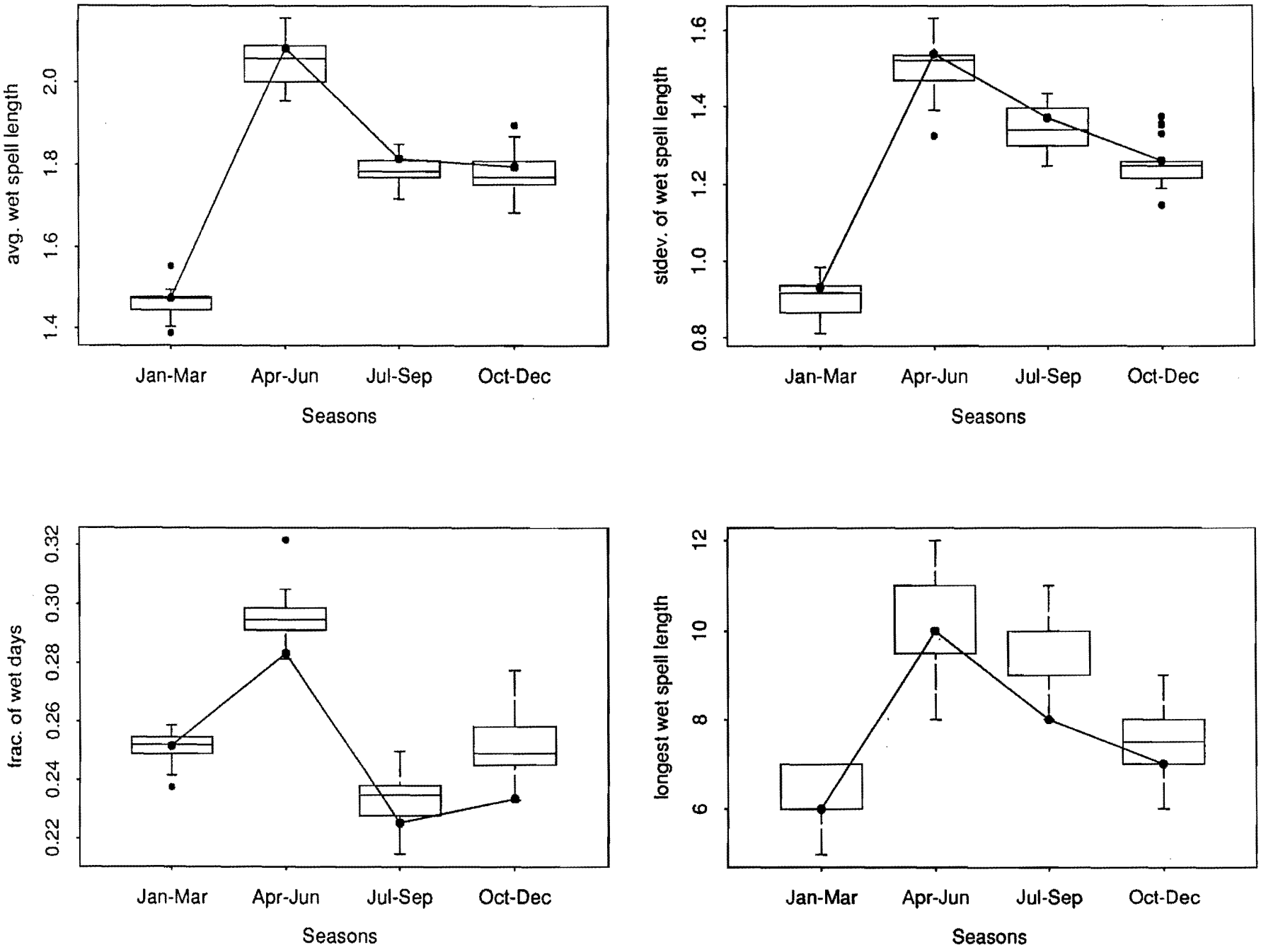
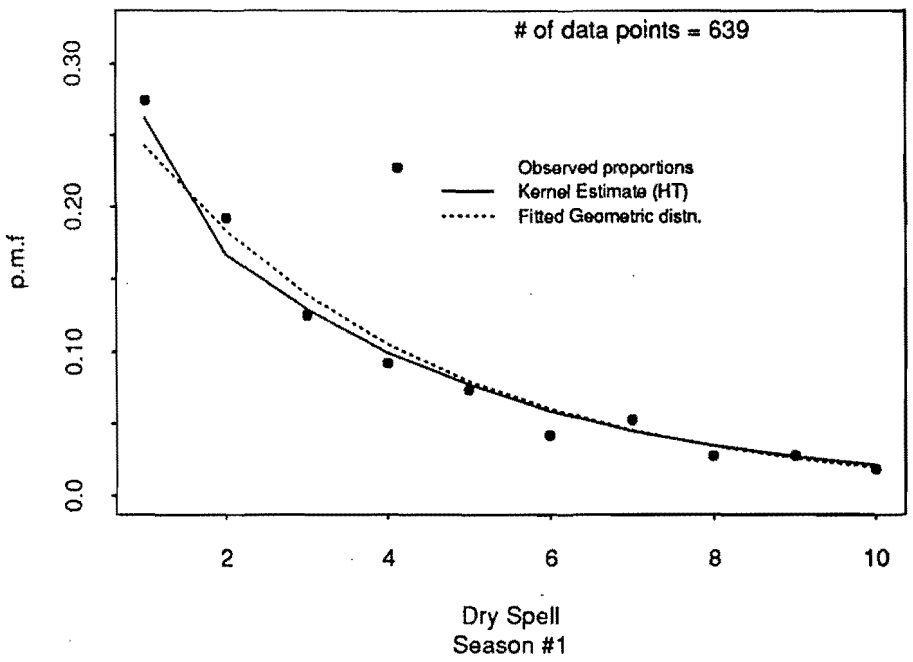
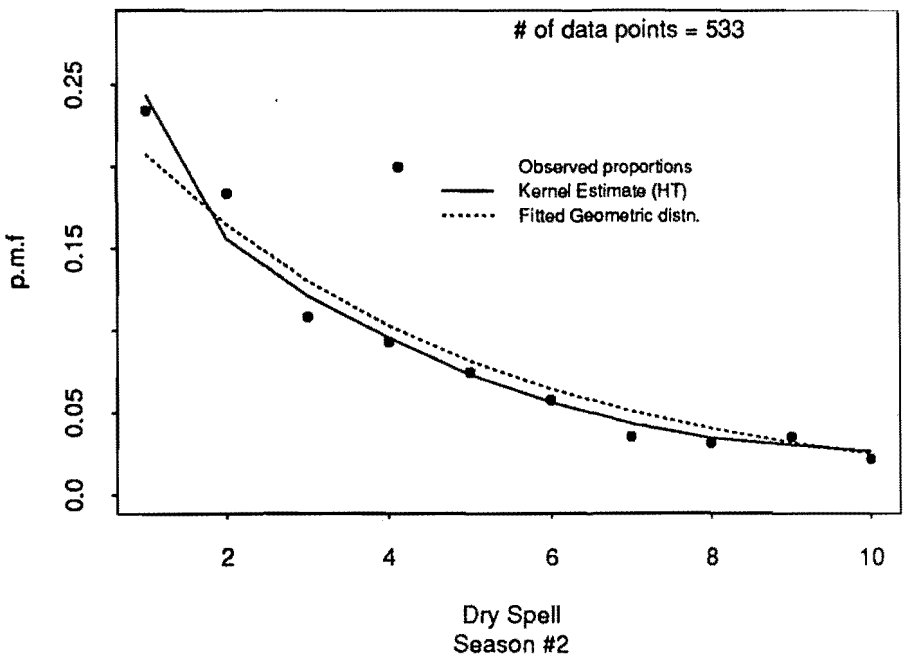
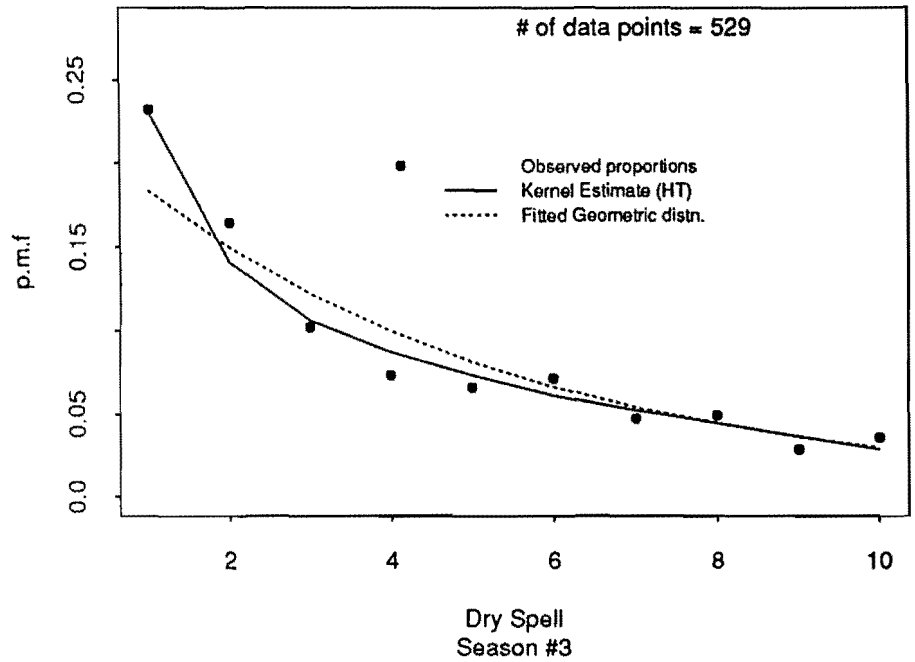

avg. wet spell length
1.4
1.6
1.8
2.0
Jan-Mar
Apr-Jun
Jul-Sep
Oct-Dec
Seasons
stdev. of wet spell length
0.8
1.0
1.2
1.4
1.6
Jan-Mar
Apr-Jun
Jul-Sep
Oct-Dec
Seasons
frac. of wet days
0.22
0.24
0.26
0.28
0.30
0.32
Jan-Mar
Apr-Jun
Jul-Sep
Oct-Dec
Seasons
longest wet spell length
6
8
10
12
Jan-Mar
Apr-Jun
Jul-Sep
Oct-Dec
Seasons

Figure 10


# of data points = 639
Observed proportions
Kernel Estimate (HT)
Fitted Geometric distn.
p.m.f
0.0
0.10
0.20
0.30
2
4
6
8
10
Dry Spell
Season #1


# of data points = 533
Observed proportions
Kernel Estimate (HT)
Fitted Geometric distn.
p.m.f
0.0
0.05
0.15
0.25
2
4
6
8
10
Dry Spell
Season #2


# of data points = 529
Observed proportions
Kernel Estimate (HT)
Fitted Geometric distn.
p.m.f
0.0
0.05
0.15
0.25
2
4
6
8
10
Dry Spell
Season #3

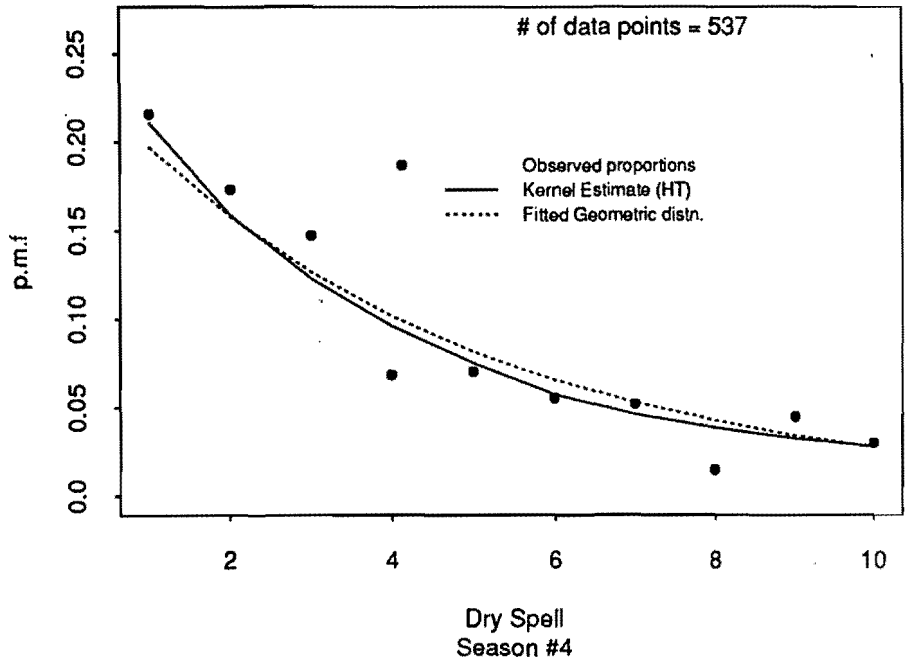

# of data points = 537
Observed proportions
Kernel Estimate (HT)
Fitted Geometric distn.
p.m.f
0.0
0.05
0.10
0.15
0.20
0.25
2
4
6
8
10
Dry Spell
Season #4

Figure 11

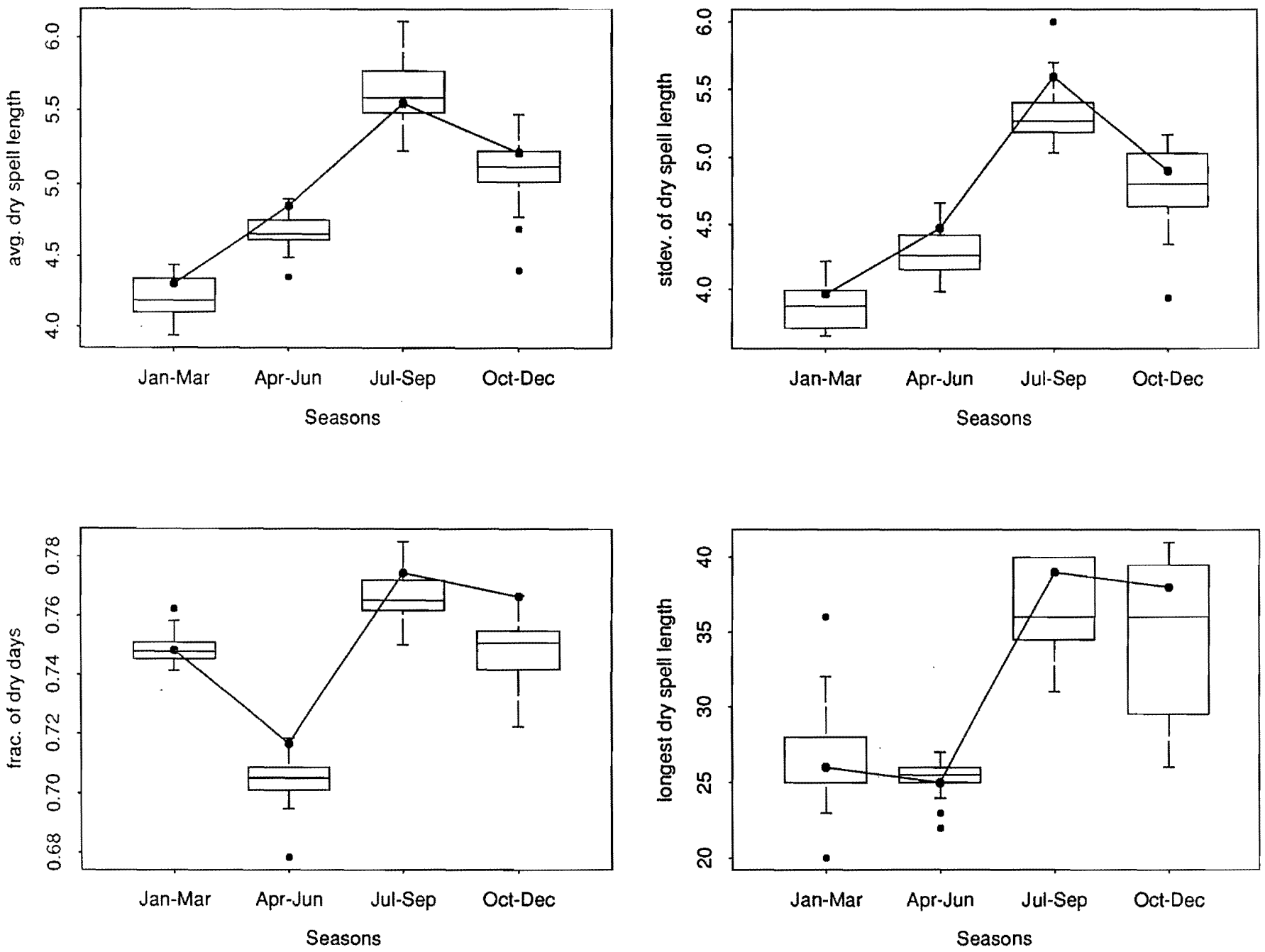
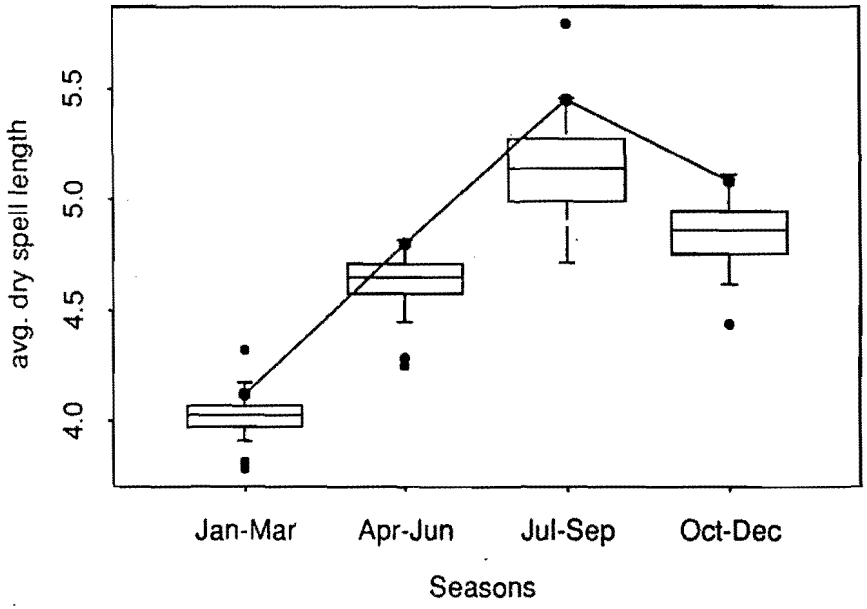
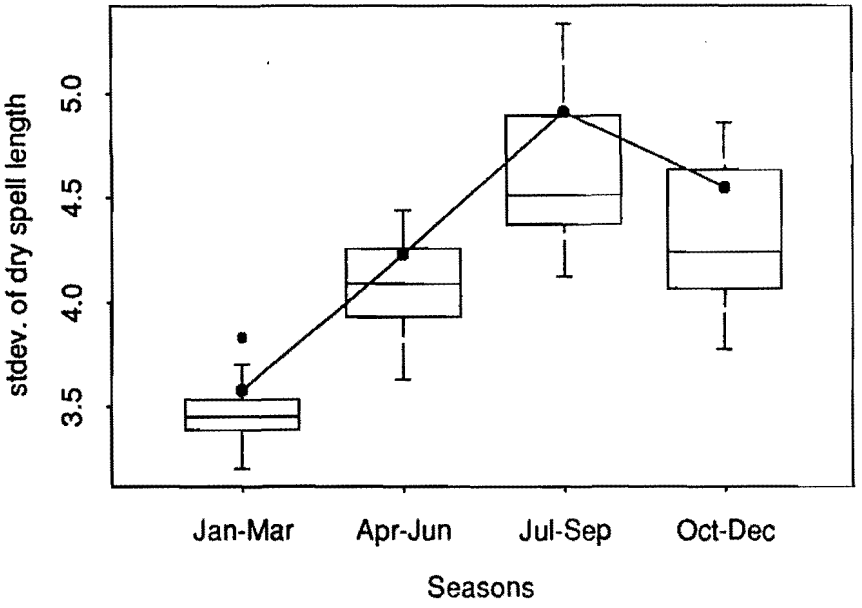
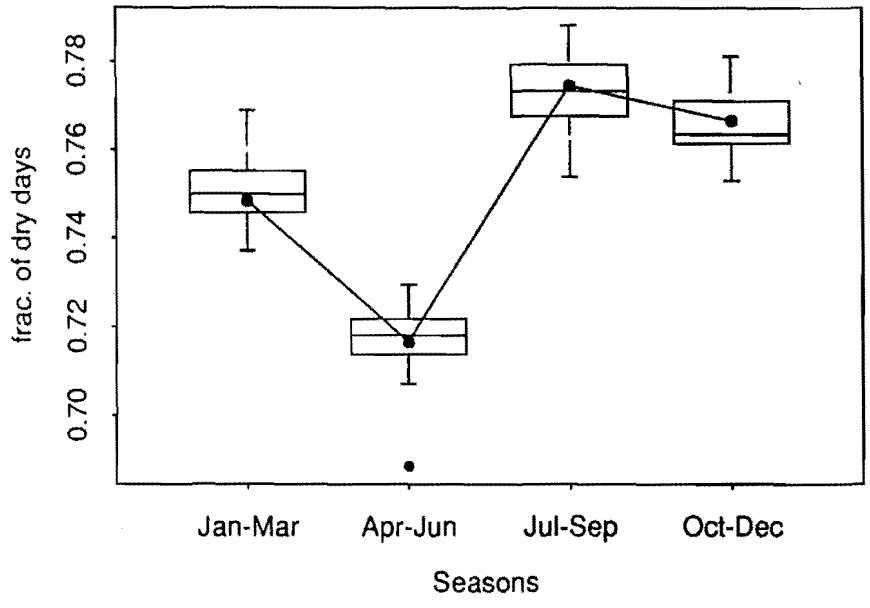
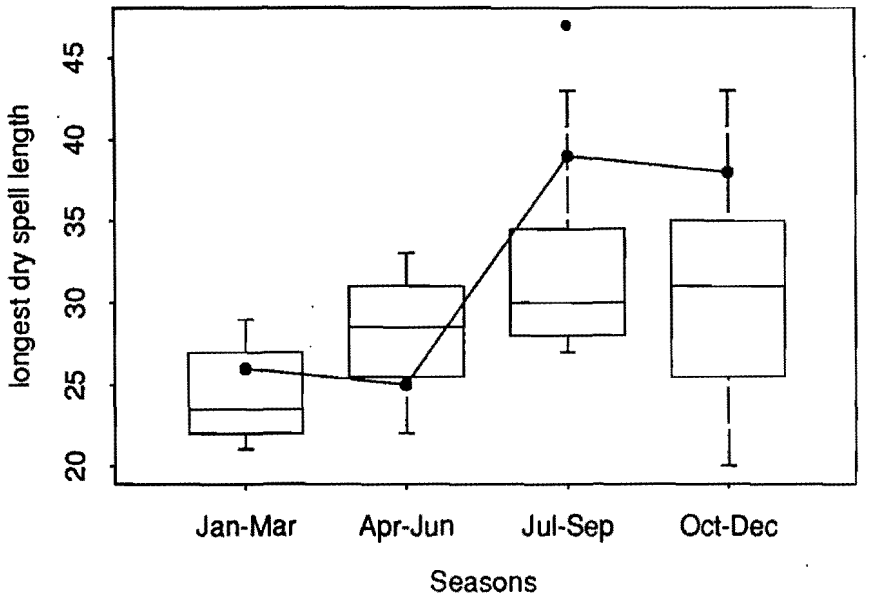

avg. dry spell length
4.0
4.5
5.0
5.5
6.0
Jan-Mar
Apr-Jun
Jul-Sep
Oct-Dec
Seasons
stdev. of dry spell length
4.0
4.5
5.0
5.5
6.0
Jan-Mar
Apr-Jun
Jul-Sep
Oct-Dec
Seasons
frac. of dry days
0.68
0.70
0.72
0.74
0.76
0.78
Jan-Mar
Apr-Jun
Jul-Sep
Oct-Dec
Seasons
longest dry spell length
20
25
30
35
40
Jan-Mar
Apr-Jun
Jul-Sep
Oct-Dec
Seasons

Figure 12


avg. dry spell length
4.0
4.5
5.0
5.5
Jan-Mar
Apr-Jun
Jul-Sep
Oct-Dec
Seasons


stdev. of dry spell length
3.5
4.0
4.5
5.0
Jan-Mar
Apr-Jun
Jul-Sep
Oct-Dec
Seasons


frac. of dry days
0.70
0.72
0.74
0.76
0.78
Jan-Mar
Apr-Jun
Jul-Sep
Oct-Dec
Seasons


longest dry spell length
20
25
30
35
40
45
Jan-Mar
Apr-Jun
Jul-Sep
Oct-Dec
Seasons