Stochastic modeling of daily summertime rainfall over the southwestern U.S. Part I: interannual variability

Jingyun Wang, Bruce T. Anderson, and Guido D. Salvucci,
Department of Geography, Boston University
675 Commonwealth Ave.
Boston MA, 02215-1401
Abstract:

The interannual variability of summertime daily precipitation at 78 stations in the southwestern U.S. is studied using chain-dependent models and nonparametric empirical distributions of daily rainfall amounts. Modeling results suggest that a second-order chain-dependent model can optimally portray the temporal structure of the summertime daily precipitation process over the southwestern U.S. The unconditioned second-order chain-dependent model, in turn, can explain approximately 75% of the interannual variance in the seasonal total wet days over the region and 83% of the interannual variance in the seasonal total precipitation. In addition, only a small fraction (generally smaller than 20%) of the observed years at any given station show statistically significant changes in the occurrence and intensity characteristics, either related to the number of seasonal total wet days or the distributions of daily rainfall amounts. Investigations of the year-to-year variations in the occurrence and intensity characteristics indicate both variations are random (on interannual time-scales), and they display similar significance in explaining the remaining 17% of interannual variance of seasonal total precipitation over the region. However, numerical tests suggest that the interannual variations of the two are not independent for the summertime monsoon precipitation, and that complex co-variability that cannot be described with simple stochastic statistical models may exist between them.
1. Introduction

Summertime rainfall over the southwestern U.S. and northern Mexico, which is generally controlled by the North American Monsoon System (NAMS), contributes a large fraction to the total annual precipitation (Douglas et al. 1993; Farrara and Yu 2003; Higgins et al., 1998). For instance, nearly 35% of the annual precipitation in Arizona and approximately 60% in the northwestern Mexico is received from July to September (Higgins et al. 1999). Thus, the NAMS precipitation pattern significantly influences the local climate and hydrologic system.

An extensive literature has developed around NAMS since the beginning of last century (Adams and Comrie 1997), most of which focuses on the intraseasonal and interannual modifications of the summertime precipitation (Douglas et al., 1993; Adams and Comrie, 1997). In general interannual variations in rainfall are related to the intraseasonal structure of daily precipitation processes (Katz et al., 1993, 1998; Wilks, 1999). However, it is not fully understood how the interannual variability of NAMS precipitation in particular is statistically and physically related to its intraseasonal structure. Studies to date have been limited by using observations for small regions (such as single-station observations) or large-scale meteorological analyses. Few detailed regional studies of the daily rainfall system have been reported for this region. To better understand the NAMS precipitation process in the southwestern U.S., the statistical characteristics of summertime precipitation are investigated using chain-dependent models. This study consists of two parts: the first part focuses on the interannual characteristics and is presented here; the second part focuses on the intraseasonal characteristics and will be summarized in a following paper (Wang et al. 2005).

Treating the daily precipitation time series as a chain-dependent process has been previously used due to the simple form and easy computation of the stochastic model (e.g., Todorovic and
Woolhiser, 1975; Katz, 1977b). However, a well known deficiency of this simple stochastic model is that it tends to underestimate the interannual variance of monthly (or seasonal) total precipitation, which is termed “overdispersion” (Katz and Parlange, 1998; Wilks, 1999). Two types of variations are usually invoked to explain the discrepancy between the model and observed variances (Katz and Zheng, 1999): 1) the high-frequency variation of the daily rainfall series is beyond the simple model’s intrinsic range (Gregory et al., 1993); and 2) the low-frequency variation of statistical characteristics from year to year (also referred to as potential predictability by some researchers) is ignored by the model (Singh and Kripalani, 1986; Shea and Madden, 1990).

Efforts have been carried out to identify and eliminate both types of overdispersions. To ameliorate the first deficiency, previous studies suggest extending the statistical characteristics of the model to better portray the structure of daily precipitation, such as using higher order chain models (Gates and Tong, 1976; Wilks, 1998), adopting intraseasonal varying model parameters (Coe and Stern, 1982), or assuming auto-correlated structure of the rainfall amount series (Buishand, 1978; Katz and Parlange, 1998). These techniques will reduce, but not necessarily eliminate, the overdispersion in interannual variance. Another perspective is to introduce indices into the models to reduce overdispersion: Katz (1993) adopted conditional parameters indexed on atmospheric circulation patterns in his model; Grondona et al. (2000) conditioned their model on the ENSO phase; and some other researchers applied hidden indices estimated from the Expectation-Maximization (EM) algorithm (Dempster et al., 1977; McLachlan and Krishnan, 1997) or others methods. These can further reduce overdispersion due to low frequency (i.e. interannual) variations in the statistical characteristics of the daily rainfall patterns (Sansom and Thomson, 1992; Sansom, 1995; Sansom, 1998). Finally, through re-sampling of the daily rainfall
series using a nearest neighbor bootstrap and assigning individual characteristics to each re-
sampled rainfall series, stochastic generators can reflect in their simulation both the short-term
(daily) and long-term (annual) variability existing in the historical dataset (Harrold et. al., 2003a;
2003b), which may further reduce overdispersion.

In this paper, we will use chain-dependent stochastic models to study the interannual
characteristics of summertime precipitation over the southwestern U.S. Our goal here is not to
capture the intraseasonal evolution of rainfall and its metrics (for instance as given by dry-spell
and wet-spell length); these characteristics will be examined in a second paper (Wang et al.
2005). Instead we wish to construct a stochastic model system in order to better understand 1) how much of the interannual variance in seasonal rainfall is related to the random evolution of
daily rainfall events, thereby isolating the potentially-predictable variance that may be related to
year-to-year variations in event frequency and rainfall amount characteristics that are not
incorporated into the model; 2) identify the contributions year-to-year variations in these
characteristics make in explaining the remaining interannual variance overdispersion generated
in the simulations; and 3) whether these year-to-year variations have longer time-scale (i.e.
multi-annular) structures, again related to potentially-predictable variance in the system. In
section 2, the historic observation dataset is described. In section 3, chain dependent models with
varying complexities are defined. Section 4 describes simulated results by the models depicted in
Section 3 and analyzes their overdispersions. Section 5 investigates the influence of low
frequency variations in the occurrence and intensity processes on the interannual variance of
seasonal total variables using conditional models. Section 6 gives some conclusions.

2. Observed data set
The historical observed dataset used in this paper is the serially complete set of daily maximum and minimum temperatures and precipitation compiled by Eischeid et al. (2000). The latest version of this dataset comprises daily precipitation observations from at least January 1948 to August 2003 for 14,317 sites, although most stations include observations prior to 1948. A sub-sample of summertime daily precipitation at 78 stations covering a region from 115W to 102W and 30N to 42N (Figure 1) of the southwest U.S. is extracted to examine the statistical characteristics for this region. All data series of the sub-sample have a time span of more than 70 years with full observations from July 1st to September 30th. Years with omitted observations during the summer season are removed from the dataset.

3. Stochastic models

Chain-dependent models treat the occurrence and intensity of daily rainfall events separately (Katz, 1977a, b; Swift and Schreuder, 1981; Wilks, 1998, 1999). The term “chain-dependence” reflects the statistical structure of the occurrence sequence. For instance, for a first-order chain-dependent (also termed as Markov Chain) process, the “chain-dependence” means that the state at time \( t \) only depends on the state at time \( t - 1 \), and is independent of states at other times. If the present state depends on more than one previous state, the time sequence is said to follow a higher order chain-dependent process and the number of related previous states is termed the chain order.

3.1 Sub-models for rainfall occurrences

The daily rainfall occurrence series is generally assumed to follow a chain-dependent process in stochastic weather models. The occurrence of rainfall can be defined as:

\[ J_t = 1, \text{ if it is wet on day } t (\text{i.e. measurable precipitation, i.e. } > 0.25 \text{ mm, occurs}) \]
\[ J_t = 0, \text{ otherwise} \]

Once the model order is determined, the occurrence process can be completely characterized by the transition probabilities. For instance, the occurrence process of a two-state first-order chain series is completely described with the function:

\[
p_{ij} = \Pr(J_{t+1} = j \mid J_t = i, J_{t-1} = i_{t-1}, \ldots, J_1 = i_1) = \Pr(J_{t+1} = j \mid J_t = i) i, j, i_{t-1}, \ldots, i_1 = 0, 1 \quad (1)
\]

and can be extended to the second-order chain series as:

\[
p_{ijk} = \Pr(J_{t+1} = k \mid J_t = j, J_{t-1} = i, J_{t-2} = i_{t-2}, \ldots, J_1 = i_1) = \Pr(J_{t+1} = k \mid J_t = j, J_{t-1} = i) i, j, k, i_{t-1}, \ldots, i_1 = 0, 1 \quad (2)
\]

Here \( \Pr(\cdot) \) is the probability, \( p_{ij} \) and \( p_{ijk} \) the transition probabilities with constraints that \( p_{i0} + p_{i1} = 1, i = 0, 1 \) and \( p_{i0} + p_{ij} = 1, i, j = 0, 1 \).

Although the first-order two-state chain-dependent (Markov Chain) model is the most commonly used for studying daily precipitation, it often underestimates variance and extremes. In an effort to eliminate these differences, studies suggest that extending the model’s complexities can greatly improve the approximating and reproducing capability of the model (Gates and Tong, 1976; Buishand, 1978; Coe and Stern, 1982; Katz and Parlange; 1998; Madden et al., 1999). As such, here we test chain-dependent models ordered from zero to three for their overall fit to the summer daily rainfall events.

### 3.2 Estimated Transition Probabilities

All model parameters are estimated from the observed dataset by using Maximum Likelihood Estimators (MLE): \( p_{ij} = \frac{n_{ij}}{n_i} \) and \( p_{ijk} = \frac{n_{ijk}}{n_{ij}} \), in which \( n_{ij} \) is a count of events transferring from state \( i \) to state \( j \), subscript \( \cdot \) the sum operator, and \( n_i \) (and \( n_{ij} \)) the marginal total.
The time homogeneity of the transition probabilities is examined first using a \( x^2 \) test at daily and annual timescales (Anderson and Goodman, 1957) with Degree of Freedom of \( df = s^{r-1}(s-1)^2(T-1) \). Here \( s \) represents the occurrence states (dry or wet), \( r \) the order of the model and \( T \) the tested sample size. Test results indicate that all transition probabilities vary both from year-to-year and from day-to-day. However, no clear trend pattern is observed in the annual series (see Section 5). In the daily series, one of the transition probabilities displays a clear polynomial-dependent evolution across the summer season for each model (Figure 2) – namely \( p_1 \) for the null model, \( p_{0,1} \) for first-order model, \( p_{0,0,1} \) for the second-order model, and \( p_{0,0,0,1} \) for the third-order model. This parameter accounts for the occurrence of rainfall events following dry states. Numerical tests indicate that a third-order polynomial fit curve can significantly reduce the residuals of this parameter compared to the mean value line (i.e. by more than 15%). In addition, 25 stations show reductions in the residuals of this parameter larger than 50%. In contrast, when using a third-order polynomial fit curve the reductions of residuals for the intraseasonal evolution of the other transition probabilities are smaller than 15% at most stations.

Therefore, two types of models are constructed based on the above analyses: 1) models with constant intraseasonal transition probabilities, estimated as the averaged values over the whole summer season for a given station (termed the non-seasonal model); and 2) models with intraseasonal time-dependent transition probabilities (termed the seasonal model). Since no clear intraseasonal evolution pattern is observed for other transition probabilities, only the occurrence transition probabilities (i.e. \( p_1 \), \( p_{0,1} \), \( p_{0,0,1} \), and \( p_{0,0,0,1} \), respectively) are allowed to vary in the seasonal models. These are estimated using a third degree polynomial curve fitted across the summer season for a given station. In addition, all transition probabilities are assumed to be
constant on an interannual basis (i.e. from year to year). This produces eight chain-dependent occurrence sub-models (four with non-seasonal parameters and four with seasonal $P_{0...1}$ parameters). The overall fitness of all eight models will be tested at all stations to obtain the optimal model for daily rainfall occurrence processes.

### 3.3 Sub-models for rainfall amounts

Both empirical (nonparametric) and theoretical (parametric) distributions have been used in the literature to fit daily rainfall amounts. The empirical distribution can reproduce the statistics of observed rainfall amounts exactly with only a small discrepancy due to the binning of observations. But a weakness of the empirical distribution is that it produces a noisy probability density function (PDF) curve. With decreasing bin width, the empirical curve becomes much noisier and is much more heterogeneous than smooth, function-based, theoretical distributions (Keatinge, 1999). Another weakness of the empirical distribution is that it can only capture local (i.e. observed) extreme events, which may underestimate the extreme values present in the potential distribution.

To overcome these weaknesses of the empirical distributions, theoretical distributions are used to fit daily rainfall amounts, such as gamma (Katz, 1977 b; Richardson, 1981; Wilks, 1989, 1992; Salvucci and Song, 2000), lognormal (Swift and Schreuder, 1981), and mixed exponential (Woolhiser and Roldan, 1982; Foufoula-Georgiou and Lettenmaier, 1987; Wilks, 1998, 1999). Swift and Schreuder (1981) also examined the fitness of Weibull and Beta distributions to daily rainfall amounts. In previous studies, the gamma distribution is most frequently used and provides a good approximation to the empirical distribution in many regions. Under some situations, however, the Weibull distribution has a better fit to the daily observations (Swift and Schreuder, 1981).
Hence, in addition to the empirical distribution, two theoretical distributions – gamma and Weibull - are also tested for their fit to daily rainfall amounts. All the functions’ parameters are estimated from observed data by using MLE at each station. For the gamma function, the maximum likelihood estimators are approximated with the Thom methodology (1958).

The identification of rainfall amounts conditioned on rainfall persistence (one-day or multiday) or on previous occurrence state (rain or not) are also examined with a Kolmogorov-Smirnov (K-S) test at each station. This test is one of the most commonly used methods to compare distributions (Wilks, 1995). Since it uses the cumulative distribution function (CDF) curves, the test can compare datasets with different sample sizes. Here we use a 5% significance level to reject (or confirm) the null hypothesis that two datasets are drawn from the same distributions. No significant difference is found between the conditioned distributions (i.e. conditioned on storm persistence or on previous occurrence state), suggesting the daily rainfall amounts can be assumed to be an independent identically distributed (IID) variable. Hence we will use a single, non-varying distribution for each station for this study.

4. Results

All statistical generators depicted in Section 3 are run $10^4$ times to create the model dataset. The overall fits of simulations to surface observations are then evaluated with the variance overdispersion:

$$r_i = \frac{\sigma_i^2 - \sigma_f^2}{\sigma_o^2}$$  \hspace{1cm} (3)

Where, $r_i$ is the variance overdispersion, $\sigma_i^2$ is the interannual variance based on data series generated by the $i^{th}$ model, and $\sigma_o^2$ is the interannual variance based on the observation dataset.
We can also estimate the overdispersion for various sub-components of the chain-dependent weather models, which we present next.

**4.1 Daily rainfall amount distribution**

In this section, the three daily rainfall distributions – gamma, Weibull, and empirical – are tested for their fit to the observed daily rainfall amounts. First, a numerical test is performed to examine the sensitivity of the empirical distribution to the bin width. Test results indicate that the approximation of the empirical distribution is insensitive to bin widths from 0.25 to 1.00 mm and suggests that the empirical CDF curve is smooth enough for our study purpose even with bin widths of 0.25 mm. In addition, it was noted that the empirical distribution is limited in its ability to capture unrealized (i.e. unobserved) yet potential extreme events. This type of deficiency, while important for intraseasonal studies of extreme-event frequency, is less important when considering seasonal mean values and their interannual variance, which is the focus of this study. Hence, the simulations using the empirical distribution are compared with those using theoretical distributions.

Based upon the observed datasets, all statistical generators can reproduce the observed mean of the daily rainfall amounts well. However, their abilities to approximate the daily variance are quite different. Figure 3 gives the relationship between observed and modeled variances. The empirical distribution fits the observed dataset well (which is to be expected), with most simulations located between the observed 95% confidence intervals (CI) lines and with a slight overestimation with respect to the observed value resulting in a small negative overall overdispersion. In comparison, the Weibull and gamma distributions underestimate the variance. The Weibull distribution provides a better simulation than the gamma distribution. However, it still produces substantial overdispersions compared with the nonparametric distribution and all
the simulated variances lie outside of the 95% CI line. These discrepancies compared with observations are large enough to introduce extra uncertainty into the interannual variability study. Hence, the empirical distribution is chosen as the optimal distribution in this study and is used for the rest of the paper.

4.2 Number of seasonal total wet days

Generally, the random sum over some period of a random variable is also a random variable, and its statistical characteristics will depend upon the underlying random variable’s structure and temporal autocorrelation. Hence, it is suggested that a simple daily precipitation model can be used to estimate the interannual variation of variables summed over some time periods with fixed length of $T$ days. In turn, a critical test for stochastic precipitation models, which has been recommended by many researchers, is whether they can reproduce the interannual variance in the variable of interest (Gregory et al., 1993; and Wilks, 1999). In this section the reproduction of interannual variance of seasonal total wet days is analyzed in order to diagnose the occurrence sub-models.

As mentioned, eight sub-models for the occurrence process are described in Section 3; here these are assessed using the area-averaged variance overdispersion (Eq. 3) of the seasonal total wet days. Each model is run $10^4$ times and produces a $92 \times 10^4$ matrix in which the number 92 represents the number of days in the summer season, and the number $10^4$ is the sample size of yearly series.

Figure 4 gives the relationship between the observed and modeled interannual variances of the seasonal total wet days from the eight models. Also shown on the figure are the 95% CI lines for the observed variance estimates along with the area-averaged variance overdispersions over all stations; 95% CIs for the model estimates are smaller than the marker and hence are not shown.
The chain-dependent assumption of the summer daily precipitation sequence over the southwestern U.S. is confirmed through the rejection of the 0-order model. Unlike other chain-dependent models, the 0-order models generate similar variance of total wet days at all stations and produce large discrepancies compared with the observed variance. Compared to the 0-order models, the first-order models greatly reduce the overdispersion but do not evaluate as well compared with the second-order models. In addition, as the model complexity increases to third order, the discrepancies (overdispersion) increase, indicating that the state’s memory associated with a temporal lag longer than two days is insignificant and may even introduce extra noise into the interannual sequence. In addition to comparing the order of the models, Figure 4 also allows a comparison between the non-seasonal and seasonal versions of the models. Similar results are obtained from both models at each order, with a slightly smaller overdispersion by the non-seasonal models. This result implies that the assumption of constant transition probabilities across the season is a reasonable simplification for interannual variation studies in this region. However, as will be shown in the companion paper (Wang et al. 2005), the intraseasonal variation of transition probabilities is important for capturing intraseasonal variations in the daily rainfall sequence (i.e. daily rainfall probability -- not shown). However, based upon their simulation of seasonal total wet-days, for this study the second-order non-seasonal model appears to be the most appropriate.

It should be noted that even the second-order non-seasonal model underestimates the observed interannual variances (based upon the observed 95% CI line) and creates a positive area-averaged overdispersion. Although about 75% of the area-averaged observed variance is associated with stochastic variations of daily rainfall occurrence as captured by the model, 25% is related to other processes not represented in the simple model (Figure 4 and Table 1). This
result implies that simply extending the model’s complexity to second order cannot fully explain interannual variance in the number of seasonal wet days and that year-to-year variations of the intraseasonal transition probabilities may influence interannual variability in this region (see Section 5).

Figure 5 gives the spatial distribution of interannual variance overdispersion produced by the second-order non-seasonal model. This distribution figure provides an indication of the spatial structure of the sub-model’s approximating capability. It appears that there are geographic variations in the amount of unexplained variance. For instance, the model performs better over the southern regions than it does over the northern regions; if we subjectively divide the domain at 37N, 61% of southern stations and 19% of northern stations display insignificant overdispersion (i.e. within the observed 95% CI lines). In addition, five of the stations in the south display negative overdispersions (indicating interannual variance is completely described by the constant characteristics of the stochastic model) while none show negative overdispersion in the northern portion of the domain.

4.3 Seasonal total precipitation

The seasonal total precipitation is a sum variable produced by the combination of the occurrence and intensity sub-models. With the simplifying assumptions that the occurrence process is independent from the intensity process, and statistical characteristics in both processes are constant from year to year, the interannual variance of the seasonal total precipitation can be conditioned on the seasonal total wet days and can be approximated with (Gregory et al., 1993; Katz and Parlange, 1998; Wilks, 1999):

\[
Var[S_{yr}(yr)] = E[N_{yr}(yr)]\sigma^2 + Var[N_{yr}(yr)]\mu^2
\]  

(4)
where $T$ represents the number of days in the summation period (here 93 days), $yr$ is the given year, $S_T(yr)$ is the seasonal total precipitation, $N_T(yr)$ the total wet days in the summer season, $\text{Var}[]$ the variance operator applied across all years, and $E[]$ the expectation operator also applied across all years (Wilks, 1999). The $\mu$ and $\sigma^2$ in equation 4 are the mean and variance of daily rainfall amounts respectively. They are assumed to be constant on an interannual basis.

While recognizing that Eq.4 is an approximation of the overall variance, by subdividing the overall variance into the various terms, we will be able to identify the main causes of the model overdispersion in reproducing the interannual variance of seasonal total precipitation (seen below). In our study, all sub models can reproduce the two mean terms ($E[N_T(yr)]$ and $\mu^2$) on the right side of equation 4 well. The empirical distribution can also reproduce the variance of rainfall amount by definition ($\sigma^2$). As such, the interannual variation of seasonal total precipitation mainly depends on interannual variances in the seasonal total wet days ($\text{Var}[N_T(yr)]$). However, it should be noted that equation 4 represents a simplified version of the total variance. As will be seen later, other processes, such as yearly variations in the statistical characteristics of daily rainfall or covariance between occurrence and intensity processes, are left out of equation 4, and may also influence the interannual variance of seasonal total precipitation.

Figure 6 gives the relationship between simulated and observed interannual variance of the seasonal total precipitation for each of the eight chain-dependent sub-models combined with the empirical distribution. Also shown on the figure are the 95% CI lines for the observations along with the area-averaged overdispersion. Similar to the seasonal total wet days, the optimal model for seasonal total precipitation is the second-order non-seasonal chain-dependent model (Figure 6). From Figure 4 and 6, the overdispersions for the two sum variables (total wet days and total precipitation) display analogous variations between the occurrence sub-models. This confirms
the strong relationship between the interannual variances of total precipitation and total wet days, which is expressed in equation 4. However the stochastic precipitation models exhibit a better overall fit to total precipitation than the occurrence sub-models did to the total wet days (although the improved fit by the stochastic model does not hold at all individual stations). For instance, the second-order non-seasonal model can capture more than 83% of the area-averaged interannual variance for the total precipitation, while the occurrence sub-model can only capture about 75% of the area-averaged interannual variance for the total wet days. Part of the remaining 17% unexplained variance in total rainfall may arise from unexplained variance in the number of wet days as seen earlier. However, other additional processes, such as year-to-year variations of the daily rainfall distribution, may also be partial sources for the unexplained interannual variance of total precipitation (discussed in Section 5).

Figure 7 shows the geographic distribution of variance overdispersion for seasonal total precipitation by the second-order non-seasonal model. Overall, 63 out of 78 (81%) stations have overdispersions smaller than 0.3, indicating that more than 70% of the interannual variance in seasonal rainfall at these locations is captured by the stochastic evolution of daily rainfall alone. Similar to Figure 5, the model performs a little better over the south region: about 80% of the stations in the south have simulations within the 95% CI line and 24% of them are above the 1:1 line (i.e. underdispersion), while only about 50% of stations in the north have simulated variances within the 95% CI lines and 9% of them have underdispersion. However there are also different regional patterns of the unexplained variance found in Figure 7 as compared with Figure 5. For instance, over eastern Arizona where the unexplained seasonal wet day variance is fairly large, the seasonal precipitation variance is small. This discrepancy in the variance overdispersions for the two seasonal sum variables indicates that models accounting for both
occurrence and amount processes can explain additional variance compared with estimations from component processes only (this effect is also reflected in the difference in the area-average variance explained for each sum variable).

5. Interannual variations

In Section 4, the estimates of seasonal total precipitation by eight stochastic precipitation models are analyzed and the second-order non-seasonal model is selected as the optimal model. While the 2nd order chain-dependent model with constant frequency and intensity characteristics can capture 75% of the observed interannual variance in seasonal wet days and 83% of the observed interannual variance in total precipitation, it cannot fully eliminate the interannual variance overdispersion. This result implies that interannual (i.e. year to year) climate variations, which can influence the daily characteristics of summertime precipitation, may be ignored by stochastic models that assume constant precipitation characteristics for all years. To better understand the nature of these year to year variations and their contributions to the interannual overdispersion of total precipitation for this region, interannual variations of the occurrence characteristics and daily rainfall amount distributions are investigated respectively in this section.

5.1 Interannual variations of the occurrence characteristics

As mentioned in Section 4, by decomposing the interannual variance of seasonal total precipitation into the four terms in equation 4, we determined that the discrepancy in reproducing interannual variance of seasonal wet days is a leading cause for the overdispersion of seasonal total precipitation. Thus, the interannual overdispersion of the seasonal total wet days and their relation to interannual variations of the occurrence characteristics are analyzed first.
A straightforward method to discriminate these interannual variations is to classify (or condition) years based upon observed seasonal total wet days and/or total precipitation. Such a method for conditioning the stochastic precipitation models on total precipitation has been used by Wilks (1989). In that study, the author categorized transition probabilities into three subsets – dry, near-normal, and wet – which were grouped by thresholds for the lower 30- and upper 30-percentile of monthly total precipitation. In contrast to Wilks’ work, we used a different threshold technique to identify years with smallest or largest numbers of seasonal total wet days. This is motivated by detailed analyses of the observed interannual sequence and the model’s reproducing capability. In section 4, the optimal model can only portray part of the interannual variation in the seasonal total wet days. For most stations, several years are located outside of the models’ bounds on the dry and wet tails. By successively removing those driest and wettest years from the observed dataset at each station, and thus creating three sub datasets - dry, normal, and wet - the variance of seasonal total wet days for the normal sub-dataset will eventually be completely portrayed by the second-order occurrence sub-model. In addition, the removal of dry/wet years that are out of the model’s capability is used to identify years that have different occurrence characteristics from normal years (see below). To perform this selection at each station, the deviation distribution is first calculated for each year:

$$D = x - \mu$$ (5)

Where $x$ is the observed total wet days for each year, and $\mu$ the observed mean of the seasonal total variable. By then iteratively removing those years with the largest/smallest $D$ values, the variance of the seasonal total wet days within the normal subset at each station will decrease continually. Once the variance based on the new normal subset is no larger than the modeled
variance, the corresponding $D$ value is considered as a threshold to obtain the dry and wet sub-datasets for that station.

This procedure allows us to objectively identify those years that lie outside the occurrence model predictive capability at a given station without resorting to a subjective probability threshold. As such, the driest and wettest years at a given station, as removed by this iterative procedure, are assumed to be statistically different from a normal year, as represented by a stochastic model with constant transition probabilities. From these three subsets, new transition probabilities at each station can then be obtained.

Using this iterative method, statistically-different wet years are successfully discriminated at all 72 stations (six stations display no interannual variations in Section 4), and statistically-different dry years are discriminated at 58 stations in this region. To investigate the heterogeneity of the transition probabilities of the three sub-datasets, the $x^2$ test is used. Test results indicate all transition probabilities for the wet or dry years are statistically ($p<0.01$) different from those for the normal years.

Once the three transition probabilities are determined at each station, a new conditioned occurrence sub-model (for reference, the models developed in Section 4 are termed the unconditioned models) is developed for each station assuming that the transition probabilities can vary from year to year based upon three indexes:

Index I: corresponding to a dry year (i.e. relatively small number of seasonal total wet days);

Index II: corresponding to a normal year;

Index III: corresponding to a wet year (i.e. relatively large number of seasonal total wet days).

To determine how frequently the model should move from one state (i.e. index) to another, the fractions of the dry and wet years at a given station can be determined with:
\[ P = \frac{\text{number}}{\text{year}} \]  

Where \text{number} is the sample size of the sub-dataset and \text{year} the sample size of the full observation dataset. These represent how many of the dry and wet years lie outside the unconditioned-model reproducing capability. The sum of the two values gives the fraction of years that accounts for the unexplained interannual variance in seasonal total wet days by the unconditioned occurrence sub-model.

Performing this procedure we find that most stations’ fractions for the dry and wet years are smaller than 0.2; in addition, the area average fraction of dry (and wet) years is smaller than 0.1. These results suggest that only a small fraction of the years at a given station are out of the reproducing capability of the unconditioned stochastic occurrence model, and hence may be influenced by statistically-significant variations of the occurrence characteristics for particular years. In general, the wet years are more abundant than dry years (for all stations). This is reasonable for dry regions where wet years comprise the tail of the distribution.

To check whether the year-to-year variations in total wet days is captured by the classifying procedure, we constructed a conditioned occurrence sub-model dataset in which \(10^4\) years of summer rainfall series are generated first by classifying each year at a given station as one of the three states – dry, normal, or wet – using the probability given by equation 6; the time series (92 days) of daily rainfall occurrence is then simulated using the transition probabilities for the given state. Here, in contrast to the Harrold et al. (2003) studies that assumed previous seasonal (or yearly) precipitation has a persistent influence on the rainfall characteristics and thus accounts for long-term dependence, we assume that the state during a given year is the result of a random process that will only influence the rainfall characteristics in the given year and is independent from precipitation in other years. To see whether interannual variability of the observed seasonal
total wet days is in fact related to longer (i.e. multi-annual to multi-decadal) time-scale processes or instead has a stochastic year-to-year evolution, the area-averaged spectra of the observed and simulated number of seasonal wet days (Figure 8) are computed. Also shown in the figure are the 90% and 95% significance level as determined by the distributions of the simulated spectra. Overall, the observed spectra displays comparable amplitude to the model values; in addition, no statistically significant spectral peak is observed for the observed number of wet days over this region with the exception of the lowest-frequency amplitude which is slightly above the 90% level (Figure 8). This result suggests the observed inter-annual variations in the seasonal total wet days are temporally independent in this region, with a possible exception at multi-decadal time-scales (>60-70 years), possibly representative of a long-term trend in this region. In addition, while spectral analysis at some individual stations do have peaks that lie outside the 95% CI line (22 stations out of 78 in total), they contain a mixed spatial structure (not shown), suggesting the spectral characteristics at a given station are a localized feature rather than a regional phenomenon. Hence, for this region, it does not appear that introducing a model with year-to-year dependence of the occurrence characteristics is appropriate.

Using the conditioned occurrence sub-model (which accounts for year-to-year variations in the occurrence process) completely eliminates the interannual variance overdispersion for seasonal total wet days at all stations (not shown), indicating the importance of incorporating year to year variations in the annual occurrence series. Next, the overall overdispersion level for the seasonal total precipitation produced by the conditioned occurrence sub-model combined with the unconditioned intensity sub-model is analyzed (Figure 9). The area-averaged overdispersion is much smaller than that produced by the unconditioned second-order non-seasonal model in Section 4.3 (Figure 6), with most of the simulated interannual variance located between the
observed 95% CI lines, indicating that the year-to-year variations in the occurrence characteristics have a significant impact upon the interannual variance of seasonal total precipitation in this region.

5.2 Interannual variations in the daily rainfall amount distribution

While above it was shown that accounting for year-to-year variations in the occurrence characteristics could remove the area-averaged overdispersion in seasonal rainfall amounts, it is also reasonable to assume the daily rainfall amounts in a rainy (or dry) year may follow a distribution statistically different from that in general years, which in turn could also impact the seasonal rainfall amount for that year. Based upon this assumption, the statistical characteristics of the daily rainfall distribution in each year are tested for each station. Those years that have a distribution statistically different from the station-based mean (i.e. the mean for the individual station) and with a smaller daily-averaged amount than general years are considered light-rain years, while those that have statistically different distributions and a larger daily-averaged amount than general are considered heavy-rain years. Similar to the interannual variations in the occurrence process, these light/heavy-rain years are considered to be related to interannual variations in the rainfall intensity process. As such, a conditioned intensity sub-model is constructed by adopting three indexes for each station:

Index I: corresponding to a light-rain year (i.e. a daily rainfall distribution statistically different from normal years with a smaller mean value);

Index II: corresponding to a normal year;

Index III: corresponding to a heavy-rain year (i.e. a daily rainfall distribution statistically different from normal years with a larger mean value).
The Kolmogorov-Smirnov test is used to determine whether a given year has a statistically different daily rainfall distribution compared with the station’s mean at the 5% significance level. With this test, the light-rain and heavy-rain years are separated from general years. Again, the selection is performed separately at each station.

Hence, three distributions – representing light-rain, normal, and heavy-rain years – are obtained at each station. Similar to the studies of interannual variations in the occurrence process, the fractions of light-rain and heavy-rain years are determined with equation 6. Only a small fraction (smaller than 0.2 at 90% of the stations) of the years displays daily rainfall distributions that are statistically different from the general distribution. Also, the heavy-rain years are more abundant than light-rain in this region.

Having identified years with significant variations in the intensity characteristics and using these to construct three separate intensity distributions, the influence of interannual variations in daily rainfall amounts are examined by combining the conditioned distributions with the unconditioned occurrence sub-model. $10^4$ years series are generated with the new conditioned-intensity precipitation model. Each simulation year is assigned one of the three conditioned distributions – light-rain, normal, and heavy-rain – based upon the percentages calculated from equation 6. Simulations of the rainfall amounts in that year will then use the same distribution derived from the observations for the given index value. As before, it assumed here that the daily rainfall distribution in a given year is independent from precipitation during other years (see below).

Simulated results of the conditioned-intensity precipitation model are given in Figure 10. Adopting conditioned distributions allows the statistical model to capture all the area-averaged interannual variance not explained by the unconditioned model in Section 4.3 (Figure 6). This
result indicates that year-to-year variations in the intensity processes are of similar significance to those in the occurrence processes for producing interannual variations in seasonal total precipitation.

Next the area-average observed spectra for the seasonal total precipitation is compared with the simulated spectra, along with the 90% and 95% CI of the model spectra distribution (Figure 11). Overall, the observed spectra display similar density levels to the model value at most frequencies. Only one point, at a period of approximately 20 years, is above the 90% CI, although it is still smaller than the 95% CI, suggesting again that interannual variability in the observed rainfall amount for this region is principally related to stochastic year-to-year variations and not to a systematic, low-frequency evolution of these variations. Significant spectral peaks exist at single stations (12 stations out of 78) however these again appear to be mainly localized features.

5.3 Full model including interannual variations in both occurrence and intensity processes

The two types of conditioned models depicted in Section 5.1 and 5.2 can essentially explain, within the confidence intervals of the observations, all of the interannual variance in seasonal total precipitation on a regional level in the southwestern U.S. However, at any given station, the improvements of the two types of conditioned models can differ markedly. In addition, only partial overlap is observed between years with anomalous occurrence characteristics and years with anomalous daily rainfall distributions, suggesting the two types of interannual variations over the southwestern U.S. may be attributable to different climatic conditions. It is of interest to investigate whether a modeling system that incorporates both conditioned sub-models produces similar estimates of total rainfall variability. Hence, a full conditioned model including both the
conditioned occurrence and conditioned intensity sub-models is created to investigate the contribution of year-to-year variations in both the occurrence and intensity characteristics to the interannual variations of total precipitation. Like any general chain-dependent precipitation model, the conditioned-occurrence/conditioned-intensity model treats the occurrence and intensity processes as independent from one another. Thus, the total influence in the full model will be a linear addition of individual influences from the interannual variations in each component process. Overall, the full model further reduces the overdispersions at all stations and in fact creates a large negative area-averaged overdispersion (\( r = -0.19 \)). This negative overdispersion indicates that the simple combination of the conditioned-occurrence and conditioned-intensity sub-models will overestimate the interannual variance of seasonal total precipitation in this region, suggesting that the interannual variations in the characteristics of each process may not be independent of one another.

6. Conclusions and discussions

This study investigated the interannual variability of summertime daily rainfall over the southwestern U.S. using chain dependent statistical precipitation models based on historical observed data. Study results indicate that an unconditioned second-order chain dependent occurrence model combined with a nonparametric empirical daily-rainfall distribution can optimally portray the intraseasonal structure in the daily rainfall time-series. In turn, the model can explain about 75% of the area-averaged interannual variance for the seasonal total wet days and 83% for the seasonal total precipitation. The remaining overdispersions in both the seasonal total wet days (25%) and total precipitation (17%) indicate that interannual variations may exist in this region that are not accounted for by the unconditioned stochastic model. Two types of
interannual variations – in the occurrence and the intensity processes – are suggested in previous studies and investigated further here.

The interannual variations in the occurrence process are studied first. By using indices conditioned on the observed seasonal total wet days, the year-to-year variations in the occurrence process are iteratively identified at each station. Sub-datasets consisting of years that are beyond the unconditioned occurrence model predictive capability are then assigned separate transition probabilities based upon the occurrence characteristics during the wet, normal, and dry years. Similarly, the interannual variations in the intensity process are identified with a K-S test. By determining years with daily rainfall distributions that are statistically different from the station’s mean, sub-datasets for light-rain and heavy-rain years are separated out at each station. In general, the fractions of dry and wet years, as well as the fractions of light-rain/heavy-rain years, for most stations are smaller than 0.2, indicating that only a small fraction of the years at a given station are influenced by interannual variations in rainfall characteristics (occurrence or intensity). In addition, these interannual variations – both in the occurrence and intensity processes – appear to be stochastic from year to year based upon the regional spectral characteristics.

Finally, by combining the conditioned sub-model of occurrence with the unconditioned sub-model of intensity (and vice versa), the contributions of the two types of year-to-year variations upon total precipitation are examined separately. Simulation results indicate that for most stations both conditioned models can capture the interannual variance of total precipitation at the 95% significant level and create a near zero area-averaged overdispersion, suggesting both interannual variations in the occurrence and intensity processes are significant contributors to the interannual variability of seasonal precipitation. However, the two conditioned models exhibit
different approximation capabilities at individual stations. Further tests with a full model consisting of both conditioned transition probabilities in the occurrence sub-model and conditioned distributions in the intensity sub-model indicate that the linear addition of the two types of interannual variations within the full model creates a large negative area-averaged overdisperion. This suggests that simple linear addition of the two sub-models will overestimate the interannual variability, and that the interannual variations in the occurrence and intensity processes are not two independent variables but have covariance between them.

From these results we conclude that a significant fraction of the area-average rainfall variance can be explained by the random evolution of daily stochastic models, which are inherently unpredictable, and that most of the unexplained variance is related to a small fraction of years in which there are systematic variations in the seasonal mean occurrence and/or intensity characteristics. Here we suggest that variations during these years may in turn be related to climatic anomalies in the NAMS that then influence the daily rainfall series across the summer season. As such, further studies of the year-to-year variations that account for the anomalous characteristics of daily precipitation are needed. The identified wet/dry years, as well as the light-rain/heavy-rain years, can form the basis for studying the effects of climatic conditions upon the frequency and intensity of rainfall events, possibly associated with a northerly displacement of the subtropical ridge (Comrie and Glenn 1998; Ellis and Hawkins 2001; Hawkins et al. 2002), previous winter (spring) rainfall conditions (Higgins et al. 1998; Zhu et al. 2004), boundary condition forcing in the equatorial and sub-tropical Eastern Pacific (Carleton et al., 1990), or gulf surge activity (Adams and Comrie 1997). In addition, it is likely that these slowly-evolving climate variations could also influence rainfall characteristics during other times of years, which also warrants further investigation.
Acknowledgement:

The authors wish to thank Jon Eischeid at NOAA’s Climate Diagnostic Center for producing and providing the station-based precipitation data products.

This research was funded by a cooperative agreement from NOAA-NA040AR431002. The views expressed here are those of the authors and do not necessarily reflect the views of NOAA.
Reference:


Figures:

Fig. 1 Average observed seasonal total precipitation (mm) over the southwest US. The area and shading of the dots are proportional to the amount of averaged seasonal total precipitation.

Fig. 2 Time-dependence of the transition probabilities of chain-dependent models, \( p_{0,1} \) for the 1st order model (a & b) and \( p_{0,0,1} \) for the 2nd order model (c & d), plotted as a function of Julian day. The triangles are observed transition probabilities. The solid lines are regression curves returned by least-square polynomial fit. The dashed lines are the averaged transition probabilities over the summer season: (a) \( p_{0,1} \) at GILA BEND; (b) \( p_{0,1} \) at TOMBSTONE; (c) \( p_{0,0,1} \) at GILA BEND; and (d) \( p_{0,0,1} \) at TOMBSTONE.

Fig. 3 Relationship between observed and modeled variances of daily rainfall amount based on the unconditioned distribution models at all stations. The solid lines are the 1:1 line, and the dotted lines are the observed 95% confidence intervals (CI).

Fig. 4 Relationship between modeled (vertical) and observed (horizontal) interannual variances for seasonal total wet days with area-averaged overdispersion shown in the left-hand corner. The solid line is the 1:1 line, and the dotted lines are the observed 95% confidence intervals. a) 0-order, non-seasonal model; b) 0-order, seasonally varying \( p_{1} \) model; c) 1-order, non-seasonal model; d) 1-order, seasonally varying \( p_{01} \) model; e) 2-order, non-seasonal model; f) 2-order, seasonally varying \( p_{001} \) model; g) 3-order non-seasonal model; h) 3-order, seasonally varying \( p_{0001} \) model.

Fig. 5 Spatial distribution of variance overdispersions for seasonal total wet days using the second-order non-seasonal occurrence model. The solid dots represent positive overdispersions and empty circles represent negative overdispersions. The area and shading of the dots are proportional to the absolute value of overdispersion.

Fig. 6 Same as Fig. 4 but for seasonal total precipitation estimated using the occurrence sub-models from Fig. 4 and the empirical daily rainfall distribution model for each station.

Fig. 7 Same as Fig. 5 but for seasonal total precipitation variance overdispersion estimated using the second-order non-seasonal occurrence sub-model combined with unconditioned empirical daily rainfall distributions for each station.
**Fig. 8** Spectral densities in the seasonal total wet days. The solid line represents the averaged spectral densities in the observed interannual time-series. The thick dotted line is the mean spectral density of the modeled interannual time-series. The thin dotted lines are the 90%, and 95% confidence interval of the spectral densities as derived from the modeled interannual time-series.

**Fig. 9** Relationship between the observed and modeled interannual variance for the seasonal total precipitation by the combination of the conditioned-occurrence sub-model which incorporates interannual variations in transition probability characteristics associated with dry, normal, and wet years (see text) and the unconditioned empirical daily rainfall distributions.

**Fig. 10** Relationship between the observed and simulated interannual variance for the seasonal precipitation by the combination of the unconditioned second-order occurrence sub-model and the conditioned intensity sub-model incorporating interannual variations in the intensity distribution characteristics.

**Fig. 11** Spectral densities in the seasonal total precipitation. The solid line represents the averaged spectral densities in the observed interannual time-series. The thick dotted line is the mean spectral density of the modeled interannual time-series. The thin dotted lines are the 90%, and 95% confidence interval of the spectral densities as derived from the modeled interannual time-series.
**Fig. 1** Average observed seasonal total precipitation (mm) over the southwest US. The area and shading of the dots are proportional to the amount of averaged seasonal total precipitation.
Fig. 2 Time-dependence of the transition probabilities of chain-dependent models, $p_{0,1}$ for the 1st order model (a & b) and $p_{0,0,1}$ for the 2nd order model (c & d), plotted as a function of Julian day. The triangles are observed transition probabilities. The solid lines are regression curves returned by least-square polynomial fit. The dashed lines are the averaged transition probabilities over the summer season: (a) $p_{0,1}$ at GILA BEND; (b) $p_{0,1}$ at TOMBSTONE; (c) $p_{0,0,1}$ at GILA BEND; and (d) $p_{0,0,1}$ at TOMBSTONE.
Fig. 3 Relationship between observed and modeled variances of daily rainfall amount based on the unconditioned distribution models at all stations. The solid lines are the 1:1 line, and the dotted lines are the observed 95% confidence intervals (CI).
Fig. 4 Relationship between modeled (vertical) and observed (horizontal) interannual variances for seasonal total wet days with area-averaged overdispersion shown in the left-hand corner. The solid line is the 1:1 line, and the dotted lines are the observed 95% confidence intervals. a) 0-order, non-seasonal model; b) 0-order, seasonally varying $p_1$ model; c) 1-order, non-seasonal model; d) 1-order, seasonally varying $p_{01}$ model; e) 2-order, non-seasonal model; f) 2-order, seasonally varying $p_{001}$ model; g) 3-order non-seasonal model; h) 3-order, seasonally varying $p_{0001}$ model.
Fig. 5 Spatial distribution of variance overdispersions for seasonal total wet days using the second-order non-seasonal occurrence model. The solid dots represent positive overdispersions and empty circles represent negative overdispersions. The area and shading of the dots are proportional to the absolute value of overdispersion.
Fig. 6 Same as Fig. 4 but for seasonal total precipitation estimated using the occurrence sub-models from Fig. 4 and the empirical daily rainfall distribution model for each station.
Fig. 7 Same as Fig. 5 but for seasonal total precipitation variance overdispersion estimated using the second-order non-seasonal occurrence sub-model combined with unconditioned empirical daily rainfall distributions for each station.
Fig. 8 Spectral densities in the seasonal total wet days. The solid line represents the averaged spectral densities in the observed interannual time-series. The thick dotted line is the mean spectral density of the modeled interannual time-series. The thin dotted lines are the 90%, and 95% confidence interval of the spectral densities as derived from the modeled interannual time-series.
Fig. 9 Relationship between the observed and modeled interannual variance for the seasonal total precipitation by the combination of the conditioned-occurrence sub-model which incorporates interannual variations in transition probability characteristics associated with dry, normal, and wet years (see text) and the unconditioned empirical daily rainfall distributions.
Relationship between the observed and simulated interannual variance for the seasonal precipitation by the combination of the unconditioned second-order occurrence sub-model and the conditioned intensity sub-model incorporating interannual variations in the intensity distribution characteristics.

Fig. 10
Fig. 11 Spectral densities in the seasonal total precipitation. The solid line represents the averaged spectral densities in the observed interannual time-series. The thick dotted line is the mean spectral density of the modeled interannual time-series. The thin dotted lines are the 90%, and 95% confidence interval of the spectral densities as derived from the modeled interannual time-series.
Tables:

Table 1. Area-averaged interannual variance overdispersions over all stations for eight models. Minimum absolute value in each row is shaded

Table 2. Area-averaged interannual variance overdispersion over all stations for the seasonal total precipitation.
Table 1. Area-averaged interannual variance overdispersions over all stations for eight models. Minimum absolute value in each row is shaded

<table>
<thead>
<tr>
<th></th>
<th>Model</th>
<th>Null</th>
<th>Order 1</th>
<th>Order 2</th>
<th>Order 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total wet days</td>
<td>Non-seasonal</td>
<td>0.5684</td>
<td>0.3123</td>
<td>0.2431</td>
<td>0.3503</td>
</tr>
<tr>
<td></td>
<td>Seasonal</td>
<td>0.5772</td>
<td>0.3309</td>
<td>0.2626</td>
<td>0.3675</td>
</tr>
<tr>
<td>Total precipitation</td>
<td>Non-seasonal</td>
<td>0.3326</td>
<td>0.2029</td>
<td>0.1652</td>
<td>0.2474</td>
</tr>
<tr>
<td></td>
<td>Seasonal</td>
<td>0.3368</td>
<td>0.2108</td>
<td>0.1757</td>
<td>0.2571</td>
</tr>
</tbody>
</table>
Table 2. Area-averaged interannual variance overdispersion over all stations for the seasonal total precipitation.

<table>
<thead>
<tr>
<th></th>
<th>Unconditioned optimal model</th>
<th>Conditioned (occurrence) model</th>
<th>Conditioned (intensity) model</th>
<th>Conditioned (occurrence and intensity) model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1652</td>
<td>0.0032</td>
<td>-0.0256</td>
<td>-0.1932</td>
</tr>
</tbody>
</table>