Influence of Large Scale Circulation Measures on Precipitation at Local Stations in the South East of the US

Claudia Tebaldi

NRCSE

Technical Report Series

NRCSE-TRS No. 055

September 14, 2000

The **NRCSE** was established in 1996 through a cooperative agreement with the United States
Environmental Protection Agency which provides the Center's primary funding.

INFLUENCE OF LARGE SCALE CIRCULATION MEASURES ON PRECIPITATION AT LOCAL STATIONS IN THE SOUTH EAST OF THE US

Claudia Tebaldi

December 13, 2000

The rainfall process at local stations in the area of the South-East US is analyzed both in its occurrence and intensity components. The goal of the analysis is to assess how much information is contained in measures of large scale circulation over the area, that may be of value in modelling and describing the statistical characteristics of the rainfall process. Exploratory Data Analysis (EDA), chain dependent processes and homogeneous and nonhomogeneous Hidden Markov Models (HMM) are used to measure the strength of the signal in the rainfall occurrence and amounts. The results hint at the necessity of considering contemporaneous, high frequency information (i.e. daily pressure fields as predictors of daily rainfall) in order to achieve substantial explanatory power, for both components of the process, consistently at all the 8 locations analyzed. The analysis by means of monthly measures of pressure and other teleconnection indices (SOI, NAO, BH) reveals how the daily variation overcomes any significant contribution from the lower frequency signals. We conclude this analysis by discussing potential ways of improving the modeling effort.

Keywords: *Rainfall process; Occurrence; Intensity; Markov Chain; Chain Dependent Process; Hidden Markov Models (HMM); Nonhomogeneous Hidden Markov Models (NHMM).*

Claudia Tebaldi, PhD, is visiting scientist with the Geophysical Statistics Project at the National Center for Atmospheric Research (NCAR), Boulder, CO.

1 Introduction

The relation between large scale circulation patterns and daily rainfall precipitation is of interest when hydrologists tackle the question of predicting droughts or floods or simply the day-to-day management of water supply, having as input coarser (in time and space) forecast of atmospheric quantities.

Another area sensitive to the issue of correctly and effectively modeling such relation is the assessment of climate change consequences. Studies of the impact of climate change may want to look at the translation into daily process statistics of lower frequency changes in the circulation. Such changes in fact are believed to be well forecasted by General Circulation Models, when varying initial conditions and inputs like greenhouse gases production (Grotch and MacCracken,1991). Conversely, General Circulation Models have failed to reproduce the high frequency, smaller scale processes in a reliable fashion, and that is where statistics can help by modeling the dependence of the latter, poorly reproduced quantities on the former, more exact and realistic ones.

In particular this study was initiated in connection to a larger study concerned with the impact of changes in the frequency and intensity of teleconnection patterns (i.e. el niño) on crops' yield. For such study an input in terms of a realistic rainfall process has to bridge the gap between what the General Circulation Model predicts in terms of large scale changes and what the agricultural sector would see as a consequence. The rainfall process parameters may or may not be influenced by large scale/low frequency variations and our analysis wants to address this issue.

A preliminary analysis, not reported in this paper, ruled out the existence of a significant signal in either processes (intensity and occurrence) of such patterns like SOI, NPA, NAO, when considering a number of isolated stations in the South East of the US. This paper presents results from a subsequent choice of explanatory variables, found to be more significantly related to the precipitation statistics in the area.

Section 2 describes the data on which the results of our analysis are predicated. Section 3 briefly fills in some notational details and presents results from a nonparametric exploratory data analysis stage and different fits of chain dependent processes to the local stations' data, comparing the performance of several choices for the dependence of the parameters on a dicotomous index of pressure over the area. Section 4 lists the results from fitting the same process parameters in a way that allows dependence on monthly/daily values of the index, this time letting the index assume values in a continuous interval. Section 5 reports the results of fitting hidden Markov models using no circulation information, monthly or daily pressure indices. Discussion and further directions follow.

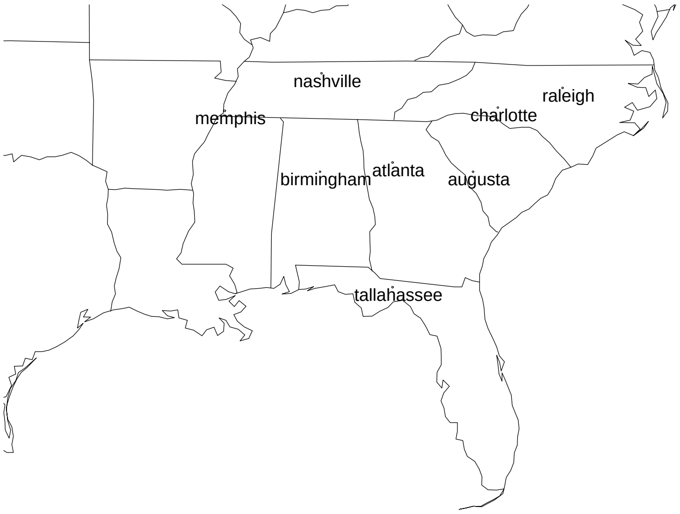

Figure 1: *The 8 stations under analysis.*

2 Data Specifications

The rainfall data span the period 1949-1996. Daily time series of precipitation occurrence and amounts at 8 stations in the South-East region of the US have been recorded. Our analysis isolates the three winter months of each year (December, January and February), the period when the precipitation mechanisms are less influenced by local, convective phenomena, and rather respond to larger scale circulation features over the area. The total number of daily observations for each station - not accounting for a few missing values - is 4242. The rainfall process constitutes our predictand. As predictor(s) we choose summary values of the mean sea level pressure (*mslp*) field over the area. The pressure data come from a coarse grid of nodes over the globe, and monthly or daily mean values at each grid-point are available for the period under study. The gridbox size is roughly 1 by 1 degrees. We choose a subregion of the grid that covers a large area over and around the stations' locations (we refer to Figure 1 for the stations' location and to the contours of correlations in Figure 2 for the extent of the pressure field under analysis), due to the fact that previous studies (Wilby, 1998) have found the centers of higher correlation between circulation and rainfall to be sometimes far removed from the areas where the precipitation is recorded.

A naive approach to the computation of an index of *mslp* could proceed as follows:

- compute the correlation between the time series of *mslp* values at each grid point and the precipitation time series at a single station;
- repeat for all the grid points and all the stations;
- locate the two areas on each contour map (one map for each station) showing the highest and lowest correlation values;
- use the value of pressure in those areas at each time step to build a summary of the relevant features in the pressure field, with regard to the precipitation process.

In our case the results of computing the 8 correlation maps are shown in Figure 2. Clearly, two areas of low and high correlation (both in absolute value roughly 0.3) are located in the Carribean and New Orleans regions, consistently for all the 8 stations. We can thus choose two points in these two geographic regions and use the values of pressure at those points to build an index by subtracting the value in the lower correlation region from the value in the higher correlation region.

A more rigorous approach is in fact taken, by computing a singular value decomposition of the 260×8 matrix of correlations (260 grid points, 8 stations). and using the first few left eigenvectors of the matrix as loadings (coefficients of a linear combination) to be applied to the values of the pressure field at each time. This is done for the first four eigenvectors (accounting for 99% of the total variance), thus computing 4 scalar indices from the 20×13 grid of pressure values at each time.

We do this for both monthly and daily mean pressure values.

Figure 3 shows the linear relation between the index computed by the naive approach and the first index computed by the singular value decomposition technique.

We will be using the first index computed by the svd approach both as a discrete factor (below or above its climatological mean), and as a continuous variable (once standardized to mean zero and variance one). Both the monthly mean values and the daily values will be separately considered as predictors in order to quantifying the gain in the explanatory power of including high frequency information rather than just the lower frequency signal, which was the original focus of the analysis.

The single index models will be compared to those using 4 indices, again in order to quantify the gain in adding dimensions to the pressure field summary.

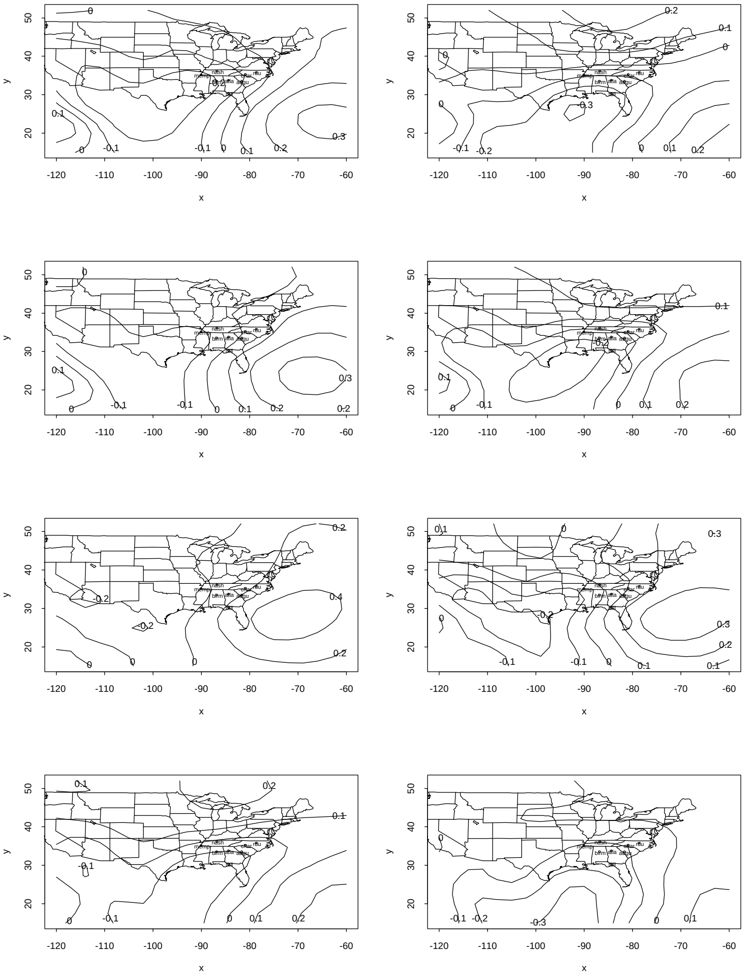

Figure 2: *The 8 plots show contours of correlation values between each station – in the order: Atlanta, Augusta, Birmingham, Charlotte, Memphis, Nashville, Raleigh, Tallahassee – and each point of the grid at which the pressure values are observed. These correlations are computed between daily rainfall values* (0 *or positive), and the monthly values of pressure replicated for each day of the month.*

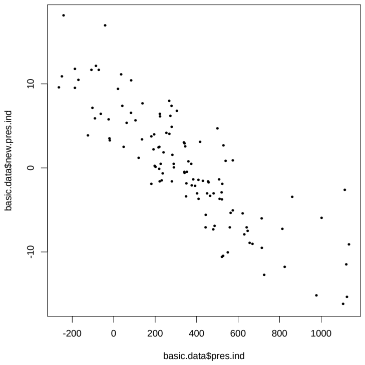

Figure 3: *The naive pressure index's values plotted against those of the first svd-derived monthly index.*

3 First Results from Exploratory Data Analysis and AIC/BIC model selection

3.1 Chain Dependent Processes

We give here the definition of a chain dependent process (Katz and Parlange, 1996), which is implicitly chosen as the paradigm of reference in the EDA part of the analysis and will be explicitly fitted in the next section, using generalized linear models estimation.

The precipitation process is divided into two components

- the occurrence process - i.e. the time series of 0/1 corresponding to dry/wet day occurrence;
- the intensity process - i.e. the sequence of amounts in wet days.

If we call S_t, $t = 1, 2, \ldots$ the occurrence process, $S_t = 0$ if the t^{th} day is dry, $S_t = 1$ if it is wet. This process follows a first-order Markov chain, and can be characterized by the 2×2 transition matrix $\{p_{ij},\ i, j = 0, 1\}$, where $p_{ij} = P(S_t = j | S_{t-1} = i)$ or equivalently by the following two parameters: $\pi = P(S_t = 1)$ and $\rho = \text{Corr}(S_{t-1}, S_t)$. These parameters have the following relation to the transition probabilities:

$$\pi = \frac{p_{01}}{p_{01} + p_{10}}$$

and

$$\rho = p_{11} - p_{01}.$$

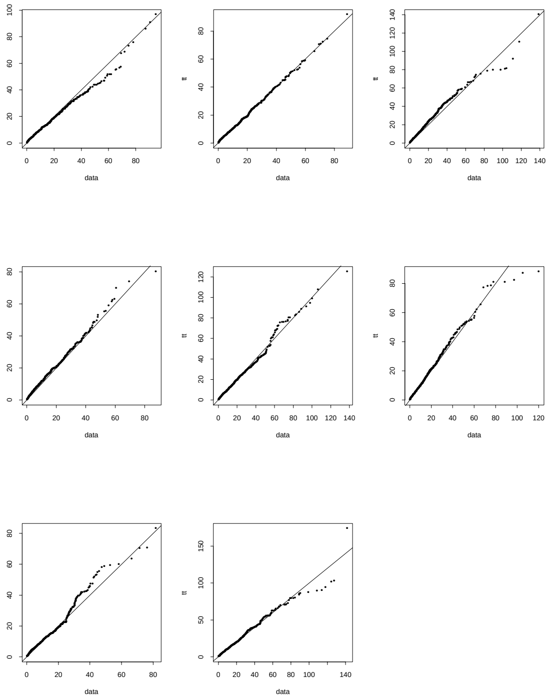

Figure 4: *Quantile-quantile plots of the distributions of amounts in wet days. Compared are the empirical distribution and the Gamma distribution whose parameters are estimated from the data by the method of moments. One plot for each station. Roughly, each plot shows* 1400 *points,* 95% *of which assume values lower than* 40

The intensity process X_k, $k = 1, 2, \dots$ is assumed to have a distribution independent of k, conditionally on $S_k = 1$

The Gamma distribution is a popular choice for modeling wet days amount (...), and we adopt it, since a simple comparison of the empirical quantiles to the quantiles of the gamma distribution with mean and variance estimated from the data show good agreement. Figure 4 shows such plots.

3.2 Exploratory Data Analysis

The findings from an exploratory data analysis are limited to considering monthly values of the single index, in their discretized version and in their original continuous version. We looked at several statistics of the daily and monthly rainfall process and their dependence on the index, by conditional distributions and nonparametric fits. Figure 5 through 8 show that a weak signal may be detected in the occurrence process, when looking at the

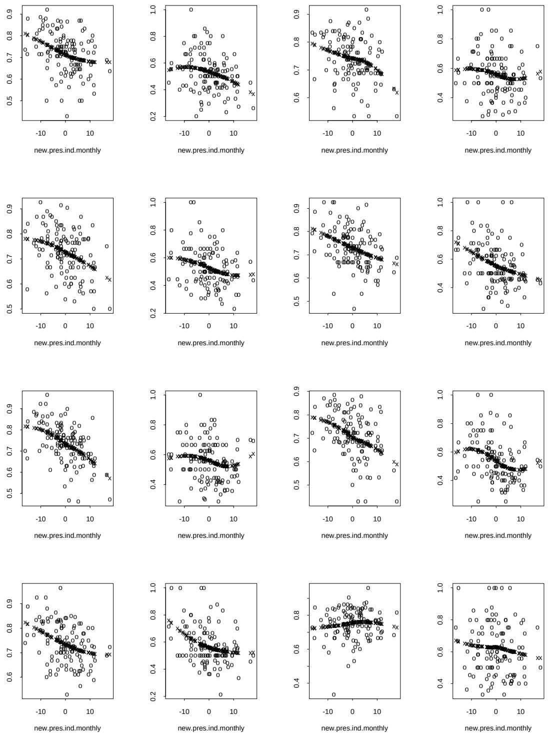

Figure 5: *Each pair of plots (proceeding row by row) correspond to an individual station. The first shows estimated values (over a month) of the probability of a dry day following a dry day, plotted against the value of the pressure index that month. The second shows values of the probability of a dry day following a wet day. Superimposed are nonparametric estimate of the underlying relation obtained by the* `loess()` *routine in S-PLUS (Cleveland et al. 1994).*

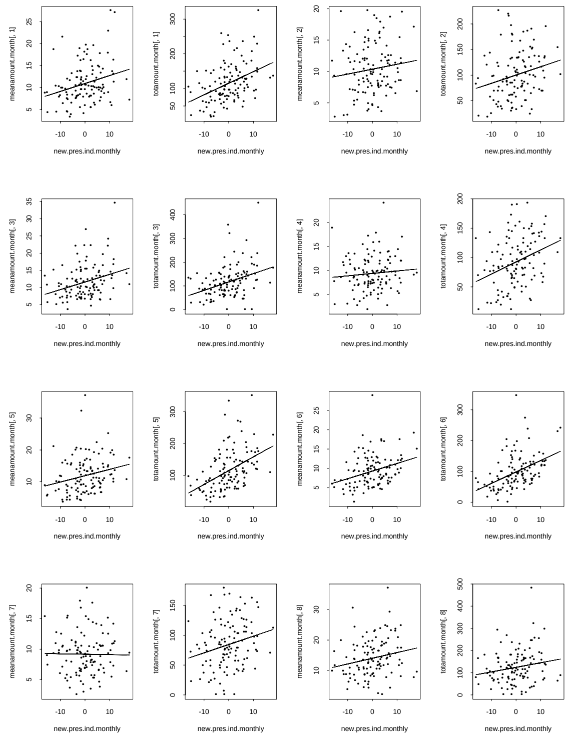

Figure 6: *Same as in Figure 5, but now the values correspond to monthly mean amounts in wet days (first plot of each pair) and monthly total amounts (second plot). The values have been log-transformed, and the superimposed lines are estimated by least squares.*

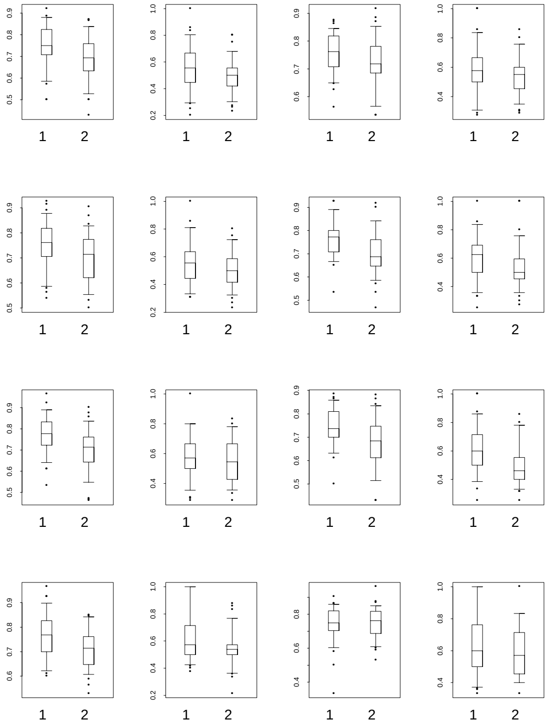

Figure 7: *Similar to Figure 5, but discretizing the index into below (1) or above (2) its long-term mean and looking at the conditional distributions of the values of the probability of a dry day, following a dry (first plot of each pair) or a wet day (second plot).*

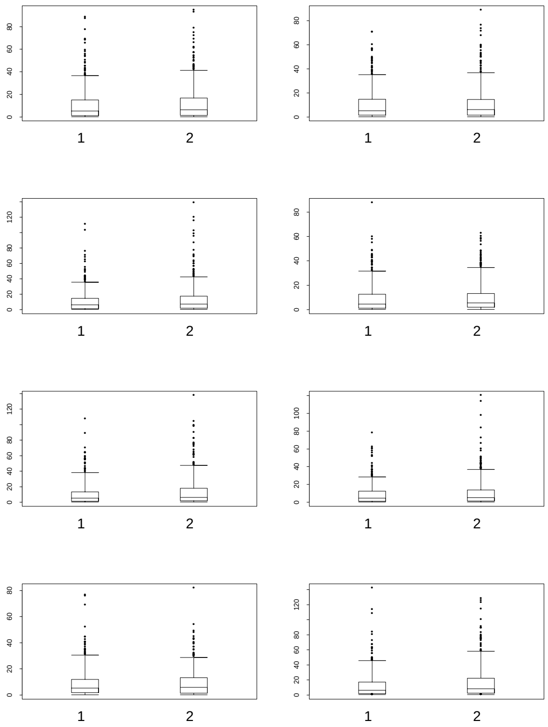

Figure 8: *Similar to Figure 7, but here the conditional distributions are of the amounts in wet days.*

Table 1: *AIC and BIC choices among different models for the occurrence and intensity process at each station. The parameters indicated at the right of each stations are the ones that according to the model selection criteria should be allowed to vary with the two-level pressure factor.*

station	AIC	BIC
Atlanta	π	π
Augusta	π, ρ, μ	ρ
Birmingham	π, ρ, σ	π
Charlotte	π, μ	π
Memphis	π	π
Nashville	π	π
Raleigh	μ	none
Tallahassee	π, μ, σ	π

distribution of the monthly estimates of probability of a wet day following a dry or a wet day. It is a weaker signal when it comes to the intensity process, i.e. when looking at mean amounts in wet days or monthly total amounts.

3.3 AIC/BIC model selection

To substantiate these qualitative findings we estimate the parameters of the chain dependent process at their climatological values (by the method of moments applied to the entire series, unconditionally) and conditionally on the monthly index value being above or below its climatological mean. We then apply AIC (see for example Encycl. of Statistical Sciences, 1982) and BIC criteria (Kass and Raftery, 1995) to choose among different models obtained by letting subsets of the parameter vector vary with the pressure factor. Table 1 reports the choice for each station of the model by indicating which parameter of the process should be allowed to vary with the pressure factor. Notice that

- μ is the mean of the distribution of amounts in a wet day, hypothesized to be a Gamma distribution
- σ is its standard deviation

and of course μ and σ are just an alternative parametrisation with respect to the more traditional Gamma parameters (shape α and scale β), i.e. $\mu = \alpha/\beta$, $\sigma = \sqrt{\alpha/\beta^2}$. The table indicates that overall the part of the process that seems to be more consistently dependent on the pressure index is the occurrence (π) and only AIC - inherently the least conservative between the two criteria - in four cases prescribes varying the intensity parameters with the pressure factor.

An alternative way of comparing the two models (conditional and unconditional estimation) is to simulate synthetic time series of rainfall according to the estimated process parameters and compare the resulting statistics to the observed ones. Figure 9 shows the distribution of several statistics derived from simulated series of rainfall occurrence and amounts at one arbitrary station (results for the other stations are comparable), for both a chain-dependent process whose parameters are kept fixed at their climatological values and a process where the parameters are estimated separately for the days when the index was above its mean, and the days when the index was below its mean. The statistics shown are

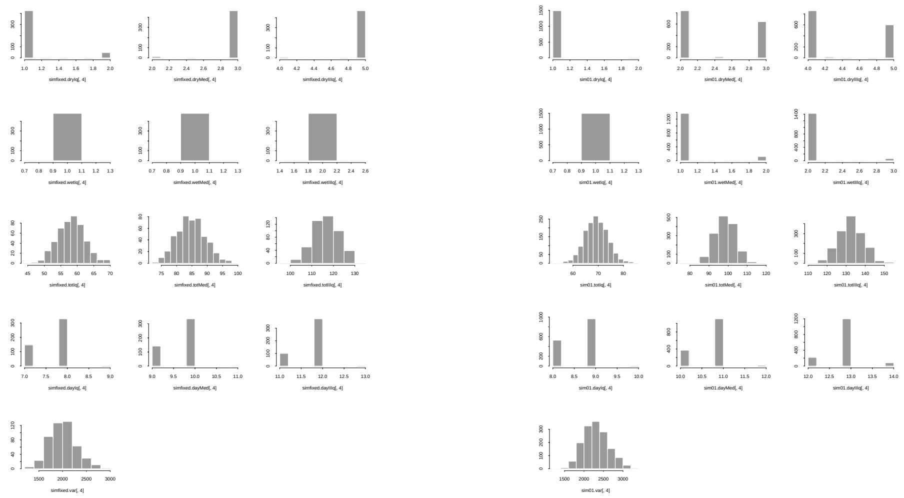

Figure 9: *Results from 5000 independent simulations of a rainfall process of the same length as the data set. The chain dependent process parameters are kept fixed at their estimated "climatological" values (estimated on the basis of the entire time series of observations) in the left panel, and assume two values conditionally on the pressure index being above and below its mean, in the right panel. Each plot compares the simulation-based distribution of a particular statistic of the process to its observed value (the mark on the x-axis). See text for a list of the* 13 *statistics shown.*

- I,II (median) and III quartile of the distribution of dry spells;
- I,II (median) and III quartile of the distribution of wet spells;
- I,II (median) and III quartile of the distribution of monthly total amounts;
- I,II (median) and III quartile of the distribution of monthly number of wet days;

- variance of monthly total amounts.

Consistently across the 8 stations, the 'weather generator' with fixed parameter reproduces well most of the observed statistics (indicated by a mark on the x axis). The only weak feature may originate from the age-old problem of overdispersion (i.e. a tendency of the models to underestimate the variance of the monthly totals; Katz and Parlange, 1998). Adding the dependence of the process on the pressure factor doesn't seem to improve the mimicking skills significantly. It rather introduces a tendency to overestimate the total number of wet days in a month and - perhaps as a consequence - the total rainfall amount in a month.

4 Generalized Linear Models' Estimation

We estimate several alternative models by estimating the parameters' dependence on the continuous pressure index using generalized linear models (McCullagh and Nelder,1983).

In order to evaluate the amount of information in different predictors, we apply a chi-square test to the values of the difference between null deviance (i.e. residual deviance of the constant model) and residual deviances of the alternative models. Table 2 compares:

- the model that treats the pressure index as a factor ('$0-1$ monthly' in the table),
- the model that uses only monthly means of the index ('continuous monthly' in the table),
- the model that uses daily values of the index ('daily' in the table),
- the model that uses all four svd-derived, daily indices as predictors (see Section2; 'daily (4svd)' in the table).

It collects the results for both the occurrence process fits, and the intensity process fits.

The series of values in the 'occurrence' part of the table shows that for all the stations bu Tallahassee the inclusion of the pressure index (even in its coarser form of a two-level factor with monthly frequency) contributes significantly to the regression model. Tallahassee needs at least the monthly continuous information as a significant regressor. In the 'intensity' part of the table things are less obvious: Atlanta, Augusta, Charlotte and Raleigh need daily information to make the regression significantly better than a simple constant fit. The other four stations respond instead to the monthly index as a two-level factor (and obviously to the further "higher resolution" predictors). It is true that the p-values are less "extreme" than the ones computed for the occurrence process, for the indices on a monthly frequency, confirming the results from the preliminary analysis that indicated how the occurrence process 'responds' better to the signal of the large scale circulation than the

Table 2: *Chi-square test for several models, for occurrence and intensity, each station separately. Those values not significant at the* .05 *level are highlighted by an asterisk*

	Occurrence			
	0-1 monthly	continuous monthly	daily	daily (4 svd)
Atlanta	$2.0e-07$	$7.7e-12$	0	0
Augusta	$8.2e-04$	$8.6e-07$	0	0
Birmingham	$1.0e-05$	$4.5e-11$	0	0
Charlotte	$1.2e-07$	$1.7e-12$	0	0
Memphis	$3.9e-08$	$2.5e-12$	0	0
Nashville	$4.8e-08$	$2.8e-12$	0	0
Raleigh	$4.6e-06$	$1.3e-09$	0	0
Tallahassee	$2.6e-01^*$	$3.0e-02$	0	0

	Intensity			
	0-1 monthly	continuous monthly	daily	daily (4 svd)
Atlanta	$3.6e-01^*$	$8.0e-02^*$	0	0
Augusta	$7.7e-01^*$	$3.8e-01^*$	0	0
Birmingham	$4.8e-03$	$4.5e-04$	0	0
Charlotte	$8.5e-01^*$	$3.5e-01^*$	0	0
Memphis	$3.4e-05$	$2.3e-05$	0	0
Nashville	$5.6e-06$	$8.5e-09$	0	0
Raleigh	$2.1e-01^*$	$7.0e-01^*$	0	0
Tallahasse	$2.0e-03$	$1.7e-03$	0	0

intensity process. The 0's in the table are to be interpreted as values not distinguishable from zeroo up to the 12th decimal point. The same results can be summarized graphically by looking at the frontier of performance for models that include monthly pressure and models that include daily pressure, when it comes to predicting wet or dry days: at different threshold values for the probability of a wet day, the Receiver Operating Characteristic curves look very similar to a straight line for the monthly models, while the curve is pushed towards the upper right corner when including the daily pressure index. See Figure 10. As for the fit of the amounts in rainy days, the images in Figure 11 compare the predicted values (just about the mean value) for the monthly model and the predicted values (more spread out) for the daily model, plotted against the observed ones. The different shading hints to the distribution of the points predicted on the plane. In this case as well, only the inclusion of daily circulation information makes the predicted values closer to the observed ones.

A further representation is given in Figure 12 where monthly statistics derived from three of the four models described (all but the 'daily (4 svd)') are compared. The plots shown are for Atlanta, but are representative of what happens at the other 7 stations as well. We first fit the model to the first half of the series, and use the parameters to predict the second half. The first row of plots compares the monthly total number of wet days as predicted by the models to the observed. The second row compares the monthly total amounts as predicted by the model (conditionally on the true number of wet days observed) to the observed. THe third row uses the of wet days together with the intensity parameters as predicted by the three models, to produce the same quantity and compares it to the

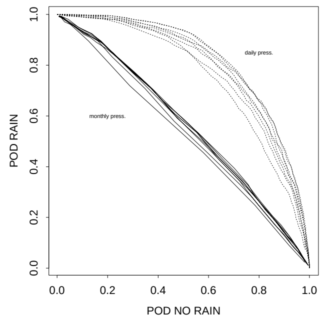

Figure 10: *ROC curves showing the predictive performance of the models with monthly pressure index (almost straight lines) and daily pressure index (curves) as covariate. One line for each station. The different points on a line are obtained by varying the threshold for the probability of a wet day used to predict a wet day – one day ahead.*

observed. As it is clear from the first and third row the models that use low frequency information have a poor performance in terms of predicting an independent set of test data, while the inclusion of daily information achieves a better reproduction of reality.

5 Hidden Markov Models' estimation

Results consistent with the chain dependent process estimation were obtained by fitting homogeneous hidden Markov models (HMM) and nonhomogeneous HMM with monthly or daily pressure as covariates. The specifics of these models are as follows (Hughes and Guttorp, 1994 and Hughes et al. 1999): it is hypothesized that the atmospheric conditions determine an unobserved, latent weather state; then the rainfall outcome is a stochastic result of the underlying weather state. Consider a first-order Markov chain $\{S_t, t = 1, 2, \ldots\}$ Associate to each state S_t a random variable R_t with distribution P_{R_t}. Each P_{R_t} is different and independent of P_{R_k}, conditionally on the states S_t, S_k. We observe a sequence of realizations $\mathbf{R}$ (the rainfall outcome at each station through time), whose components are independent of each other, conditionally on the corresponding, unobservable $\mathbf{S}$ (the sequence of weather states). The time-dependence acts only through $P(S_t = S_i|S_{t-1} = S_j)$, the transition matrix of the hidden Markov process. In the case of nonhomogeneous HMM the transition probabilities depend on a set of covariates associated to each t in the process (atmospheric variables). More specifically,

$$P(S_t = i|S_{t-1} = j, \mathbf{X}_t) \propto \gamma_{ij} \cdot (\mathbf{X}_t - \mu_{ij})V^{-1}(\mathbf{X}_t - \mu_{ij})'$$

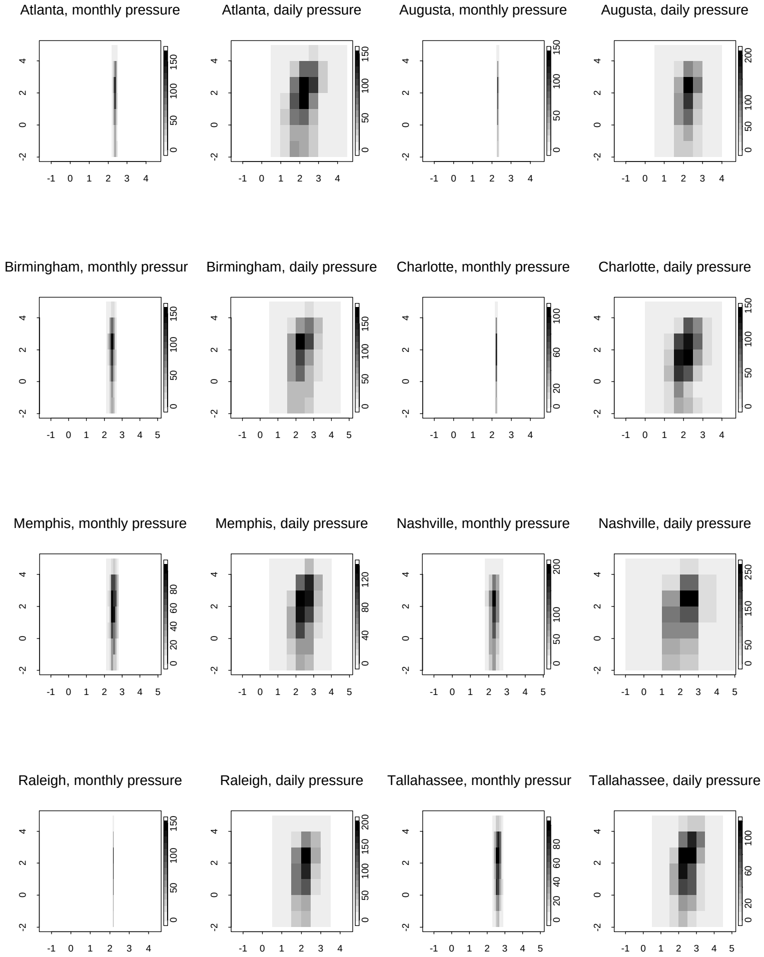

Figure 11: *Fitted versus observed values of amounts in wet days. Each pair of plots (row-wise) correspond to one of the stations. The first plot of each pair compare the fits of the model with monthly values of the pressure index as covariate to the observed values. The second compares the fits of the model with daily values of the pressure index as covariate.*

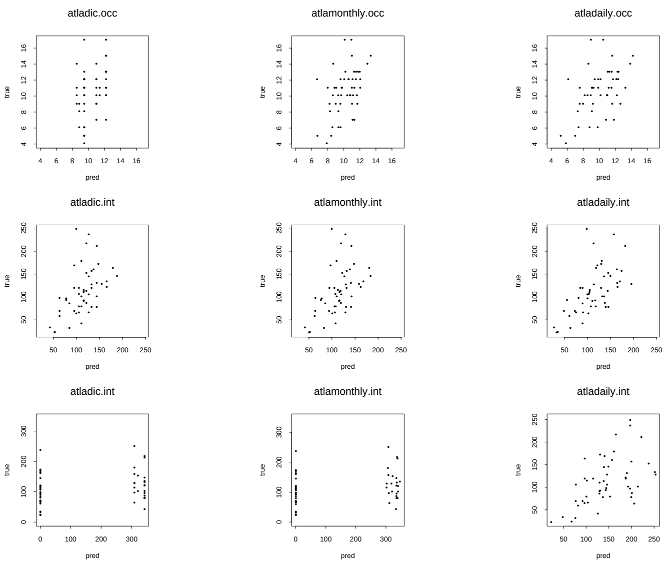

Figure 12: *Atlanta. Predicted versus observed number of wet days in a month (first row) and total monthly amounts (second and third row). Column one corresponds to the '0-1 monthly' model; column two to the 'continuous monthly' and column three to the 'continuous daily' models.*

Notice that the form of this transition consists of a baseline transition probability, γ_{ij}, and a normal kernel which involves the atmospheric variables.

Conditionally on S_t the observed process is just a sequence of independent random variables distributed as either Gamma or Negative Binomial (after transforming the amounts into integers), when the intensity is modelled, or as Bernoulli when just the occurrence is modelled. The parameters' estimation is performed via an EM/MCML (Monte Carlo Maximum Likelihood) algorithm (Dempster et al.,1977; Geyer and Thompson, 1992).

We present results from a suite of different models.

- We first fit three models for only the occurrence process, with three hidden states (the number of states is a free parameter, chosen by trial and error by the modeler), without covariates (homogeneous HMM), with monthly pressure index as covariate (monthly NHMM) and with daily pressure index as covariate (daily NHMM). The value of BIC shows a significant drop only when the daily model is compared to the two others.

- Then two models with four states are fitted, only to the occurrence process, comparing HMM and daily NHMM, and the drop in BIC value is still significant.

- The same is true when models for both occurrence and intensity are fitted using three

Table 3: *BIC values for different HMM/NHMM models fitted.*

Occurrence		
3 States (records 58-96)		
	BIC	number of parameters
Homogeneous	20370	27
Monthly Pressure	20350	30
Daily Pressure	19152	30
3 States (records 49-96)		
	BIC	number of parameters
Homogeneous	25072	27
Daily Pressure	23601	30
4 States (records 49-96)		
	BIC	number of parameters
Homogeneous	23798	40
Daily Pressure	22225	44

Occurrence & Intensity		
3 States (records 49-96)		
	BIC	number of parameters
Homogeneous	175178	69
Daily Pressure	173660	72
4 States (records 49-96)		
	BIC	number of parameters
Homogeneous	173030	96
Daily Pressure	171394	100

and four states, and HMM versus daily NHMM.

A detailed list of results is presented in Table 3

Even if the gain in the penalized loglikelihood is significant, pictures of the resulting fits from the HMM and NHMM with four states and daily pressure don't show a striking improvement with respect to any of the diagnostic quantities: the meaning of the four states is the same, and so are the model-based simulated values of rainfall amount at the stations, consistent with the observed for both models and all the stations. The pairwise dependence between the stations is fairly well reproduced, even if on average underestimated, by both models almost identically. The model itself doesn't contain spatial dependence features and all what is observed in the synthetic series of multivariate rainfall processes is induced by the underlying common atmospheric states.

The only evident improvement is the better fit of the NHMM to the observed wet spells distributions, as shown by their observed and estimated survival curves.

Figure 13 through 16 detail the above described results.

6 Discussion

The exercise performed in this paper clearly demonstrate the need for high frequency circulation information when trying to downscale large scale pattern to local rainfall processes, in the region and at the stations under study.

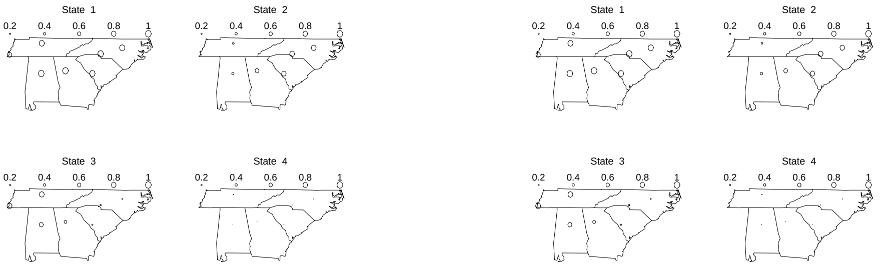

Figure 13: *The four states identified by an HMM and a NHMM, in terms of probability of a wet day at the stations.*

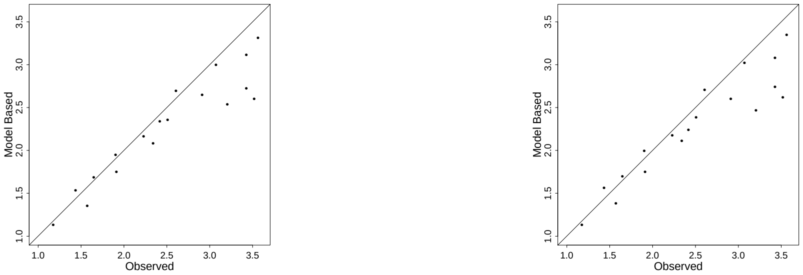

Figure 14: *Consider the quantity* $\log \frac{n_{00}n_{11}}{n_{01}n_{10}}$. *Here* n_{ij} *refers to the entry of a contingency table that for a given pair of stations collects the number of days that at one the occurrence process assumed value* i *and at the other station value* j. *Each point on the graph compares observed values of this quantity to one obtained by simulation. In the left panel the simulation is based on the model estimated by the four-state HMM; in the right panel on the model estimated by the four state NHMM that uses daily pressure as a covariate.*

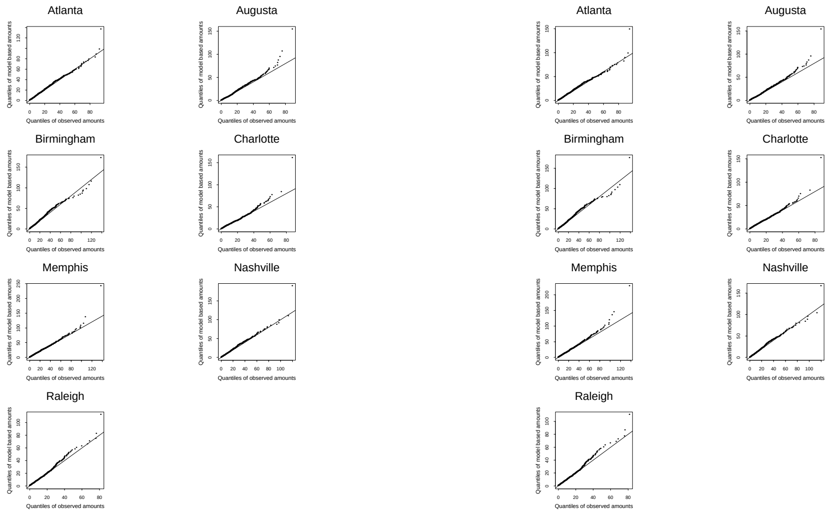

Figure 15: *Quantile-quantile plots of observed vs. model-based simulated amounts in a wet day. One plot for each station. The model is a four-states HMM in the left panel and a four-states NHMM with daily pressure as covariate in the right panel.*

We propose a few directions for further work on the issue, and for this specific dataset, in particular. Given that models with more svd-derived covariates show consistent improvement in their explanatory power, it is worth pursuing a search for more/better predictors. Other information about the circulation over the area may be of value, like temperature, or relative humidity or some measure of vertical motion (though the latter is probably more interesting in seasons of convective activity). One can also use the output of an HMM (a model with no covariates) to derive a classification of days into different unobserved states. The days belonging to the same state may be analysed/averaged in order to detect interesting features in the circulation that may explain the different characteristics of rainfall. We performed this exercise using a 3 state HMM and deriving three indices of pressure "dictated" by the proximity of each daily pressure pattern to the average ones in the three states. Models fitted by using the three HMM-dictated indices as covariates delivered results similar in explanatory power to the model with a single svd-derived covariate.

The analysis performed in the EDA stage and by GLM estimation has not tackled the issue of spatial dependence among stations. The larger values of correlations are observed in the number of wet days in a month, while the actual amounts don't show correlations above .5 between any couple of stations. It would be interesting to build a model able to induce the right amount of correlation by the action of the index of pressure. Right now, predicting occurrence and intensity at each station separately by using the models fitted independently produce "perfectly correlated sequences" in the sense that when it rains at one station it rains everywhere else and viceversa. When generating synthetic sequences

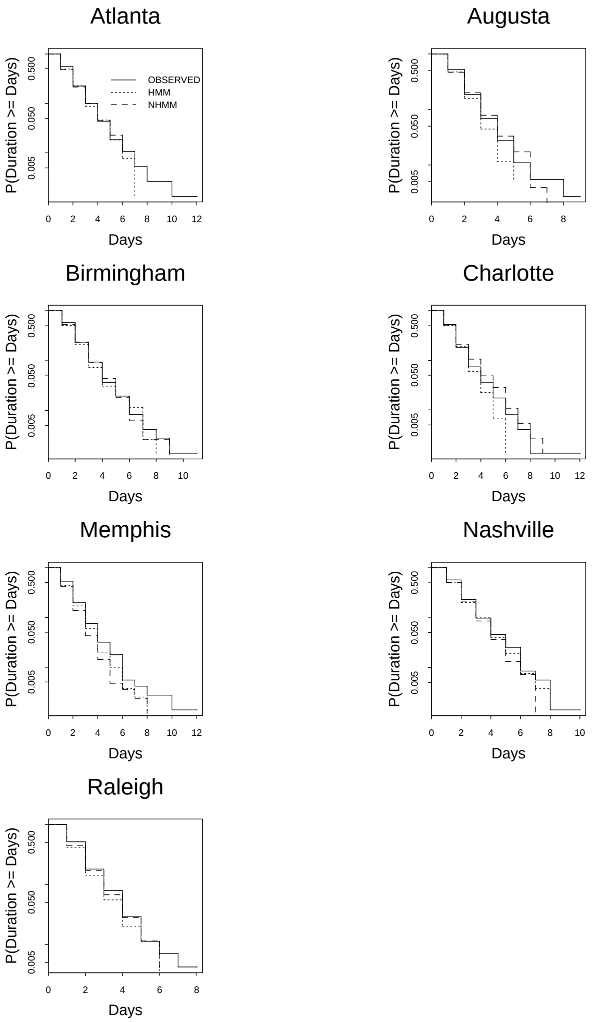

Figure 16: *Survival curves for the length of a wet spell; solid line is the observed, the dotted lines correspond to model-simulated series, for both HMM and daily NHMM.*

from the independent models at each station, with daily pressure driving the parameters, the amount of correlation induced between couple of stations is on the order of $.5, .6$, too uniformly high compared to the observed one. On the other hand, generating the same series from models using monthly pressure (continuous and discretized) produces series totally uncorrelated.

Conditionally on finding interesting and relevant circulation information for our rainfall process in the area, it can be worth investigating the relation between large scale circulation and teleconnection patterns. Quantities like pressure and temperature may in fact be more influenced by these low frequency signals than local intermittent processes like precipitation. A three-tier model can ensue, in which the teleconnections indices "drive" the large scale circulation over the area, whose signal, in turn, will impact the precipitation statistics at local stations.

7 Acknowledgements

I thank the National Research Center for Statistics and the Environment for its support during the two months I spent there, while working on this study; in particular, Peter Guttorp and Jim Hughes helped me with interesting suggestions and discussions. I'm expecially grateful to Enrica Bellone, who contributed the results of her Hidden Markov Model approach, along with considerable amounts of her time in a particularly busy moment of her PhD student's career.

References

Akaike's criterion, edited by S. Kotz, N. L. Johnson, and C. B. Read. in *Encyclopedia of Statistical Sciences* (Volume 1) John Wiley & Sons, New York (1982).

Cleveland, W.S., E. Grosse and W.M. Shyu: Local regression models, in *Statistical Models in S*. Chambers, J.M. and Hastie T.J. editors. Wadsworth and Brooks/Cole, Pacific Grove, CA (1994).

Dempster, A.P., N.M Laird and D.B. Rubin: Maximul Likelihood estimation from incomplete data via the EM algorithm (with discussion). *J.R. Statist. Soc. B*, **39**, 1-38 (1977).

Geyer, C.J. and E.A. Thompson: Constrained Monte Carlo maximum likelihood for dependent data. *J.R. Statist. Soc. B*, **54**, 657-699 (1992).

Grotch, S.L. and M.C. MacCracken: The use of general circulation models to predict climate change. *J. Climate*, **4**, 286-303 (1991).

Hughes, J.P. and P. Guttorp: A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena. *Water Resour. Res.*, **30**, 1535-1546 (1994).

Hughes, J.P., P. Guttorp and S.P. Charles: A nonhomogeneous hidden Markov model for precipitation. *Applied Statistics*, **48**, 15-20 (1999).

Kass, R.E. and A.E. Raftery: Bayes factors. *J. Am. Statist. Assoc.*, **90**, 773-795 (1995).

Katz, R.W., and M.B. Parlange: Mixtures of stochastic processes: Application to statistical downscaling. *Climate Research*, **7**, 185-193 (1996).

Katz, R.W and M.B. Parlange : Overdispersion phenomenon in stochastic modeling of precipitation. *Journal of Climate*, **11**, 591-601 (1998).

Mc Cullagh, P. and J.A. Nelder: *Generalized Linear Models*. Chapman and Hall, New York, NY (1983).

Wilby, R.L.: Statistical downscaling of daily precipitation using daily airflow and seasonal teleconnection indices. *Climate Research*, **10**, 163-178 (1998).