12-10-2014

Long Term Ground Based Precipitation Data Analysis: Spatial and Temporal Variability

Luciano Rodriguez
Chapman University, rodri178@mail.chapman.edu

Cyril S. Rakovski
Chapman University, rakovski@chapman.edu

Hesham el-Askary
Chapman University, elaskary@chapman.edu

Mohamed Allali
Chapman University, allali@chapman.edu

CHAPMAN UNIVERSITY
SCHMID COLLEGE OF SCIENCE AND TECHNOLOGY

Long Term Ground Based Precipitation Data Analysis: Spatial and Temporal Variability

Rodriguez, L., Rakovski, C., El-Askary, H., Allali, M.
Schmid College Of Science and Technology Chapman University, Orange, CA

Abstract

California is an area of diverse topography and has what many scientists call a Mediterranean climate. Various precipitation patterns exist due to El Niño Southern Oscillation (ENSO) which can cause abnormal precipitation or droughts. As temperature increases mainly due to the increase of CO2 in the atmosphere, it is rapidly changing the climate of not only California but the world. An increase in temperature is leading to droughts in certain areas as other areas are experiencing heavy rainfall/flooding. Droughts in return are providing a foundation for fires harming the ecosystem and nearby population. Various natural hazards can be induced due to the coupling effects from inconsistent precipitation patterns and vice versa. Using wavelets, we were able to identify anomalies of high precipitation and droughts within California's 7 climate divisions using NOAA's hourly precipitation data from rain gauges and compared the results with modeled data, SOI, and PDO. The identification of anomalies can be used to compare and correct remote sensing measurements of precipitation and droughts. Promising results show a possible connection with increasing tropical moisture activity.

Data and Study Area

Hourly precipitation data was used for the analysis from National Climatic Data Center (NCDC) and NOAA's Forecast Systems Laboratory (FSL) CD-ROM where there are more than 2500 active stations and 7000 total stations. The majority of the data ranges from 1948 through 1995, however, there are some stations that begin as early as 1900. The data extracted concentrates on all the stations in California. The data obtained needed a lot of corrections since there were many hours and days missing, accumulated and deleted data. Secondly, to apply the ARMA model, the data needs to be continuous in time. Therefore, we average the precipitation measurements of all the stations pertaining to a climate region (Figure 1) to get an adequate estimate of the climate per regions of California. Then fill in the missing data by imputing timely averages.

The second set of data is monthly modeled data provided by NOAA. The gauge data set will be used to compare the modeled data set to see how well the model is performing.

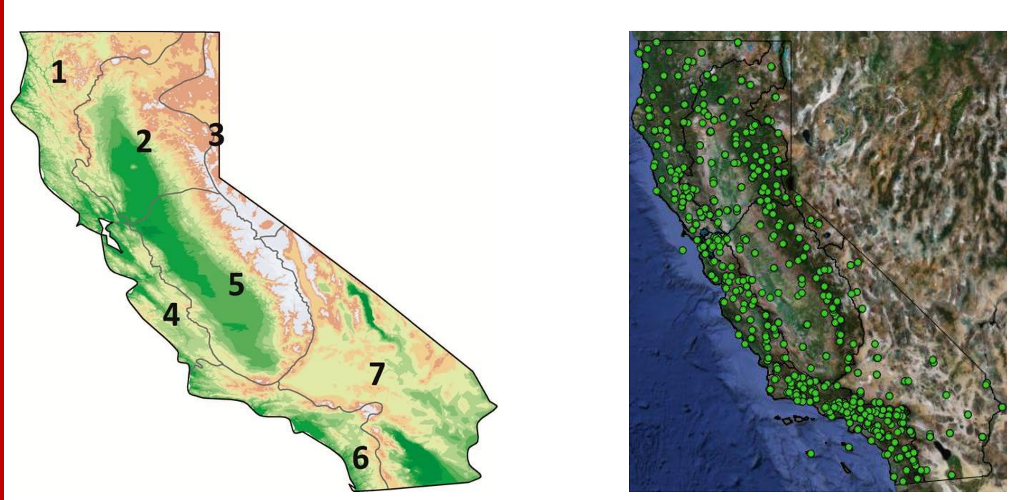


Figure 1-2. 7 climate divisions of California: 1 - North Coast Drainage, 2 - Sacramento Drainage, 3 - Northeast Interior Basins, 4 - Central Coast Drainage, 5 - San Joaquin Drainage, 6 - South Coast Drainage, 7 - Southeast Desert Basin

Methodology

The monthly precipitation was calculated over a period of 47 years (1948 - 1995) and was implemented in a time series decomposition to remove seasonal and trend components. This work used local linear regression (loess) to remove such trends into a composition of seasonal, trend, and residual components (STL decomposition). It is given by:

$$Y_t = S_t + T_t + R_t$$

which decomposes the time series (Y_t) into three distinct components. The seasonal and trend components are of no interest to this study, but the residual (R_t) is due to its fluctuation of anomalies. Moreover, an implementation for analysis of the stationary time series autoregressive-moving average (ARMA) model was used to analyze the residual component. Autoregressive-moving average model is defined as:

Methodology (continued)

$$R_t = \sum_{i=1}^{p} \alpha_i R_{t-i} + \sum_{j=0}^{q} \beta_j \varepsilon_{t-j}$$

where α_i, β_j, and ε_{t-j} are the parameters of the ARMA model. The error to this model is given by the following properties: $cov(\varepsilon_t, y_{t-j}) = 0$, $\varepsilon_{t-j} \sim N(0, \sigma^2)$, and $cov(\varepsilon_t, \varepsilon_{t-j}) = 0$. This model can be rewritten to establish a maximum likelihood error: $\phi_\alpha(L) R_t = \phi_\beta(L) \varepsilon_t$

where $\phi_\alpha(z) = 1 - \sum_{i=1}^{p} \alpha_i z^i$, $\phi_\beta(z) = \sum_{j=0}^{p} \beta_j z^j$, and $LR_t = R_{t-1}$, which is the lag operator. Using the maximum likelihood estimator one can calculate the residuals (R). Approaches to fitting the ARMA models to the residual is based on Akaike Information Criterion (AIC) minimization. The AIC model is given by: $AIC(\phi_\alpha, \phi_\beta, \sigma^2) = -2\ln[L(\phi_\alpha, \phi_\beta, \sigma^2)] + p + q + 1$
The AIC model is optimal when the forecast mean square error (FMSE) measure is no more than a certain threshold.

Once the autocorrelation is calculated, a derivation of 95% confidence band is used to detect extreme values: $\hat{R}_t \pm 1.96\sqrt{MSE(\hat{R}_t)}$,

where $\hat{R}_t$ corresponds to the residual output from the ARMA model and MSE is the mean square error. This equation reflects the confidence intervals of anomalies beyond the common observations in the residual decomposition of the data. Thus the extreme observations (E) are given by:

$$E = \{y_t | r_t > \hat{R}_t + 1.96\sqrt{MSE(\hat{R}_t)}\} \bigcup \{y_t | r_t < \hat{R}_t - \sqrt{MSE(\hat{R}_t)}\}$$

Where y_t are time dependent observations and r_t are time dependent residual values from the maximum likelihood. This methodology was perform in El-Askary et al. article and was used in this project to model precipitation patterns and extract ENSO anomalies in California.

To compare the two time series data set (modeled and gauge), we subtract the two time series: $Y_t = X_t - Z_t$, where X_t, is the modeled time series and Z_t, is the gauge time series. We then performed a test for trend using Mann-Kendall Test:

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sign(x_j - x_i), \text{ where } sign(x_j - x_i) = \begin{cases} 1, & > (x_j - x_i) \\ 0, & = (x_j - x_i) \\ -1, & < (x_j - x_i) \end{cases}$$

If a trend is detected, one can test the subtracted time series for stationarity using the Augmented Dickey-Fuller Unit-Root Test:

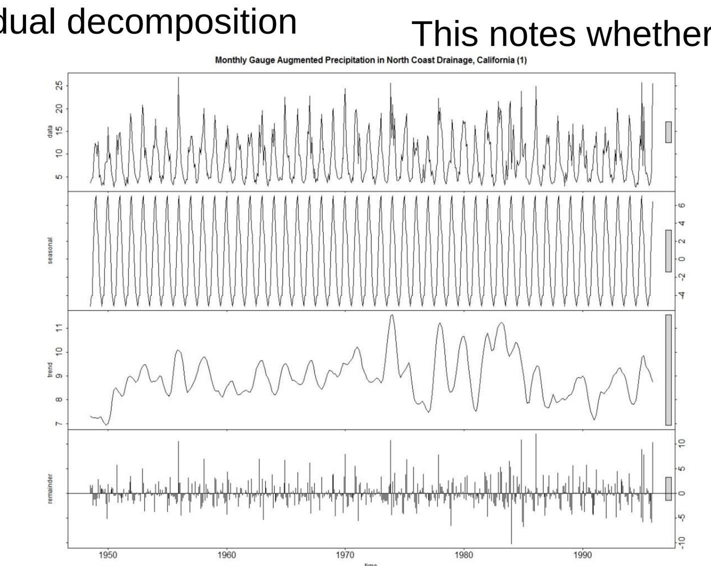
$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \cdots + \delta_{p-1} \Delta y_{t-p+1} + \epsilon_t.$$

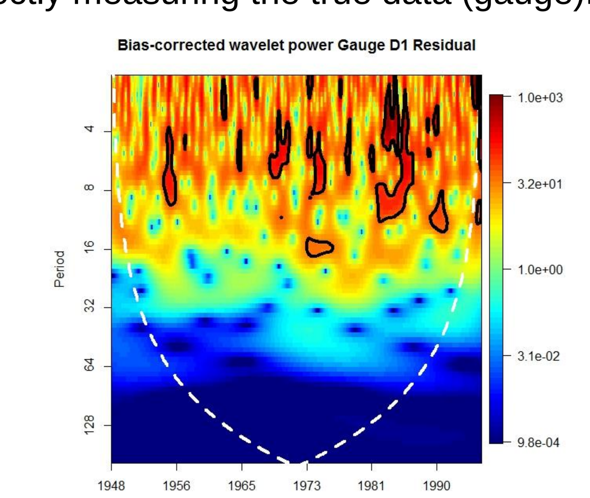
The null hypothesis $\gamma = 0$ against the alternative hypothesis $\gamma < 0$ contains the unit root test: $DF_t = \frac{\hat{\gamma}}{SE(\hat{\gamma})}$. If the time series is stationary, then one can perform a hypothesis testing based on the z-scores:

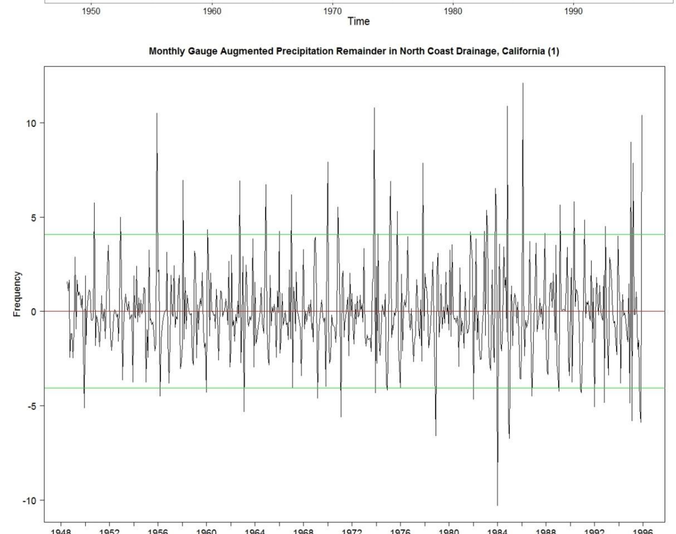
$$H_0 \; : \; \mu(t) = 0$$
$$H_A \; : \; \mu(t) \neq 0$$

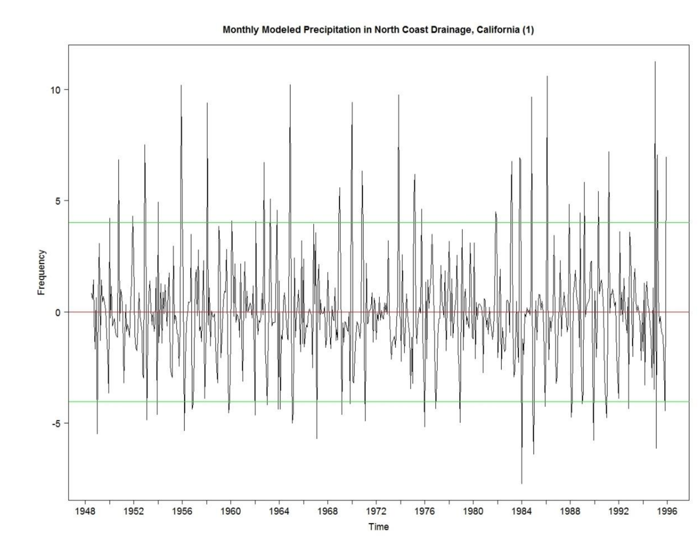
This notes whether the model is correctly measuring the true data (gauge).

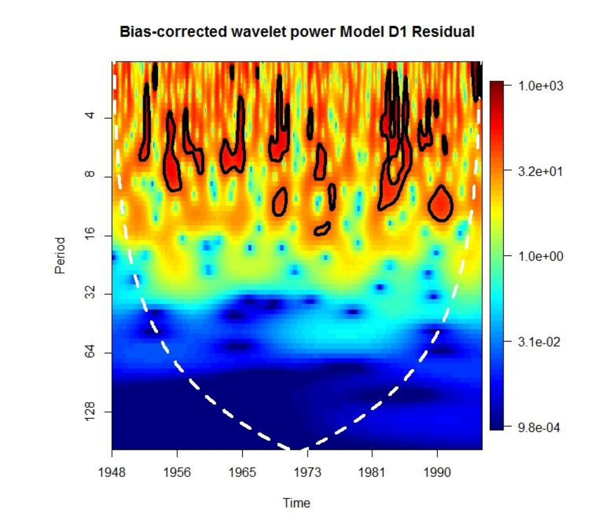
Results

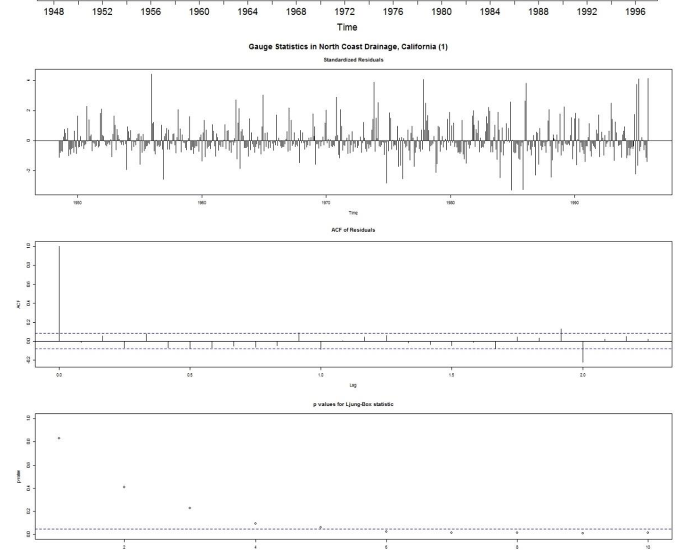


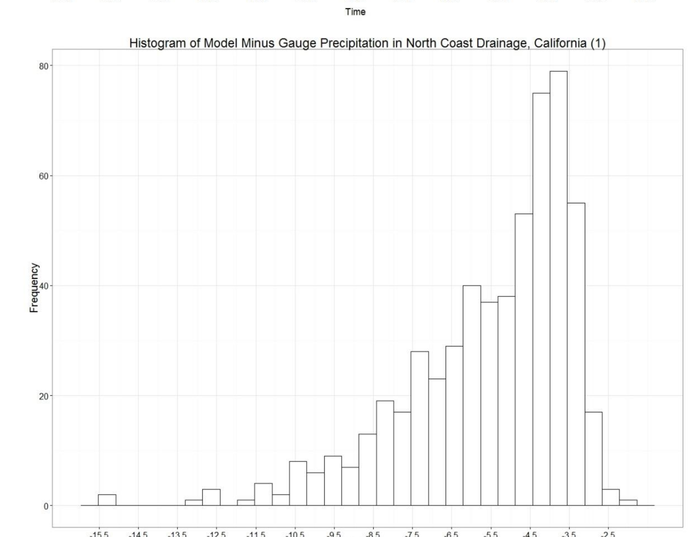


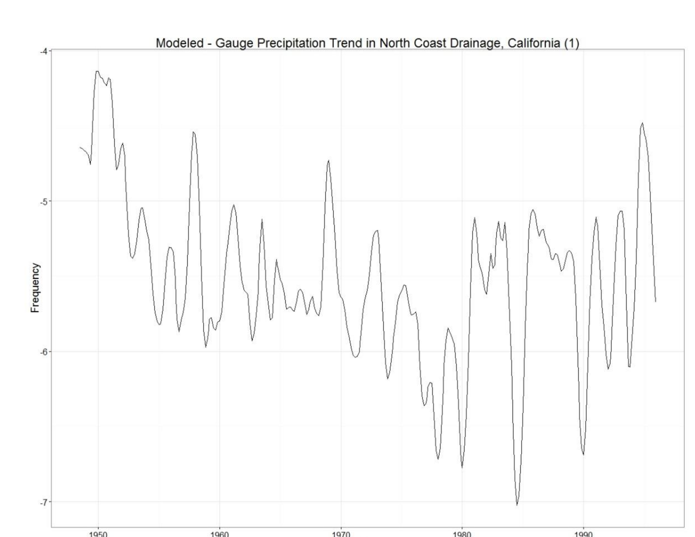


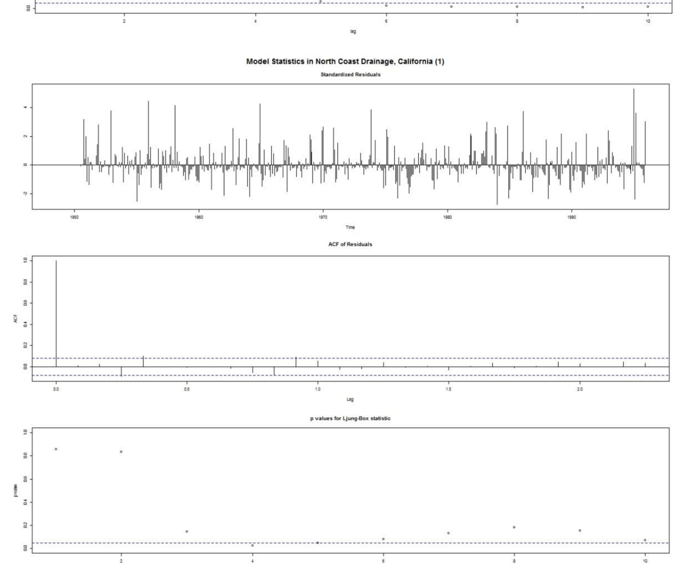












Monthly Gauge Augmented Excessive Rain in North Coast Drainage, California (1)

Date	CI	Remainder_Value	Date	CI	Remainder_Value
Oct-50	4.067634	5.757119	Oct-81	4.067634	4.23382
Dec-52	4.067634	4.980898	Dec-82	4.067634	4.239821
Dec-55	4.067634	10.5226	Feb-83	4.067634	5.356321
Feb-58	4.067634	6.942903	Mar-83	4.067634	4.307159
Feb-60	4.067634	4.348038	Nov-83	4.067634	6.528172
Oct-62	4.067634	6.927224	Dec-83	4.067634	5.318291
Dec-64	4.067634	6.721753	Nov-84	4.067634	10.87232
Jan-66	4.067634	4.258372	Feb-86	4.067634	12.10319
Jan-67	4.067634	6.189922	Dec-87	4.067634	4.125445
Jan-70	4.067634	7.924156	Mar-89	4.067634	5.649046
Nov-70	4.067634	5.542048	May-90	4.067634	5.816245
Nov-73	4.067634	10.79694	Mar-91	4.067634	4.845202
Mar-74	4.067634	4.11226	Dec-92	4.067634	4.491649
Mar-75	4.067634	6.896264	Jan-95	4.067634	8.969987
Oct-75	4.067634	5.288991	Mar-95	4.067634	7.866046
Nov-77	4.067634	7.849707	Dec-95	4.067634	10.39994

Monthly Modeled Excessive Rain in North Coast Drainage, California (1)

Date	CI	Remainder_Value	Date	CI	Remainder_Value
Jan-50	4.024495	4.196245	Nov-73	4.024495	9.755491
Oct-50	4.024495	6.839264	Feb-75	4.024495	5.138798
Dec-51	4.024495	4.298942	Mar-75	4.024495	6.180109
Dec-52	4.024495	7.498336	Oct-75	4.024495	4.623324
Jan-53	4.024495	4.323402	Nov-81	4.024495	4.491962
Jan-54	4.024495	4.9192	Dec-81	4.024495	4.150797
Dec-55	4.024495	10.17264	Feb-83	4.024495	4.414713
Jan-56	4.024495	4.561349	Mar-83	4.024495	6.765538
Feb-58	4.024495	9.390391	Nov-83	4.024495	6.929051
Feb-60	4.024495	4.080439	Dec-83	4.024495	6.838359
Feb-62	4.024495	4.074846	Nov-84	4.024495	9.648742
Oct-62	4.024495	6.713863	Feb-86	4.024495	10.59648
Apr-63	4.024495	5.073475	Dec-87	4.024495	4.839672
Nov-63	4.024495	4.577761	Nov-88	4.024495	4.458243
Dec-64	4.024495	10.21712	Mar-89	4.024495	5.825754
Dec-68	4.024495	4.045503	May-90	4.024495	5.401517
Jan-69	4.024495	5.576568	Mar-91	4.024495	7.189508
Dec-69	4.024495	5.388231	Jan-95	4.024495	11.24316
Jan-70	4.024495	9.401017	Mar-95	4.024495	7.048231
Nov-70	4.024495	6.329715	Dec-95	4.024495	6.952523

Results (continued)

The model and gauge data have a similar pattern in the residuals for all divisions. The months found to have excessive precipitation are overlapping between the two sets of data. One of the difference noted between the plots is that the model frequency is, at times, lower than the gauge signifying that the model is not capturing all precipitation fluctuations. The technique of taking the difference between the two datasets (model and gauge) is useful to find the differences between the two. Theoretically if both the model and gauge time series are the same it should equal zero or it would have a zero mean at worst case scenario. The histogram shows that the majority of the time the model underestimates the amount of precipitation. On another note, the difference of the two time series should diminish any traces for seasonality and trend. One can check if there is a trend present for all climate divisions using the Mann-Kendall Test. The results show that climate division 5 still has a trend present. This is the second time we noted that the model is not a good estimator of precipitation. Furthermore, if the differenced time series is stationary, which it is by the Augmented Dickey-Fuller Unit-Root Test, one can check if the time series has a mean of zero. We can check this using z-scores. The results show that for all climate divisions the time series is not zero mean. Overall, the model must be recalibrated.

Conclusions

Forecasting atmospheric hazards using historical precipitation data is what scientist/researchers are attempting to do well enough to be able to avoid catastrophic events. However, the usage is different from researcher and location. Therefore, this study focuses on historical hourly precipitation data to model extreme precipitation (mainly related to ENSO) to have a proper input for other models (climate, hydrological, crop, etcetera) to ensure the public with food, water, and shelter. The data consisted of 47 years of hourly observation at various regions in California. The data provided insight to the tropical storm patterns traveling west to east during the summer. Our results show that high precipitation, which may be caused by various components (i.e. El Niño, La Niña, etc.), may cause increase in precipitation as well as an increase in droughts and heat waves. The usage of 95% confidence interval bands were used to isolate the times that extreme precipitation and droughts were present. This method was consistent in portraying high precipitation from thunderstorms and heat waves related to the ENSO effects. This method allows a good detection to forecast high precipitation that may be a possible catalyst to tropical storms, which will allow authorities and civilians to properly prepare for safety and evacuation if needed in terms of flooding.

Future Research

The next step is to continue to map the ENSO patterns with the usage of weekly, daily and hourly dataset. Also see what and how one can improve the modeled data from NOAA .

Acknowledgements

The author would like to acknowledge the help received from Dr. El-Askary for providing the idea and data for the project, Dr. Cyril Rakovski for the time series lectures, and David Stack for the Python code suggestions.

References

1. El-Askary, H., Allali, M., Rakovski, C., Prasad, A., Kafatos, M., and Struppa, D. Computational methods for climate data. WIREs: Computational Statistics 4, 4 (Aug. 2012), 359.

2. EL-Askary, H., Sarkar, S., Chiu, L., Kafatos, M., and El-Ghazawi, T. Rain gauge derived precipitation variability over virginia and its relation with the el nino southern oscillation. Advances in Space Research 33, 3 (2004), 338-342.

3. Foundation, P. S. Python Programming Language. http://www.python.org/, 2012.

4. Prasad, A., and Singh, R. Extreme rainfall event of july 2527, 2005 over mumbai, west coast, india. Journal of the Indian Society of Remote Sensing 33 (2005), 365{370. 10.1007/BF02990007.

5. RDC, T. R: A language and environment for statistical computing. R Foundation for Statistical Computing, 2008.