Climate data: Long range dependent or nonstationary?

Dependence, Stability, and Extremes Workshop
The Fields Institute
Toronto, Ontario, Canada

6 May 2016

Mark M. Meerschaert
Department of Statistics and Probability
Michigan State University

mcubed@stt.msu.edu
http://www.stt.msu.edu/users/mcubed

Partially supported by NSF FRG grant DMS-1462156, NSF INSPIRE grant EAR-1344280, and ARO MURI grant W911NF-15-1-0562.

Abstract

A North American Regional Climate Change Assessment Program (NARCCAP) climate model was used to generate 29 years of daily maximum temperature data at around 16,000 spatial locations in North America. Our ultimate goal is to develop accurate models for forecasting extreme weather conditions, e.g., heat waves. The goal of this talk is to develop a reasonable time series model for the temperature data at any given spatial location. Future work will also consider the spatial pattern. Since the data shows a strong seasonal variation, as well as long range correlations, it is not clear *a priori* whether to apply a model with long range dependence, or periodic stationarity, or both. It has been observed, e.g., in financial time series, that nonstationarity can mimic long range dependence. We will see that this is also the case for climate data. Hence a periodically stationary time series model may be indicated.

Acknowledgments

Paul L. Anderson, Maths and Computer Science, Albion College
Metin Eroglu, Statistics and Probability, Michigan State U
Joshua French, Math and Statistical Sci., U Colorado Denver
Piotr Kokoszka, Statistics, Colorado State University
Stilian Stoev, Statistics, University of Michigan

NARCCAP data set

Data is freely available at `http://www.narccap.ucar.edu/` from the North American Regional Climate Change Assessment Program (NARCCAP) .

Data Source: Canadian Regional Climate Model (CRCM) + Community Climate System Model (CCSM)

CRCM is a detailed local climate model from the Canadian Centre for Climate Modelling and Analysis.

CCSM is a global circulation climate model from the USA National Center for Atmospheric Research (NCAR) in Boulder CO.

A complete data set (no missing values) of maximum daily surface air temperature (°K) on a 140×115 grid over much of North America covers $N = 29$ years, 2041–2069.

NARCCAP data set

$X_n(\boldsymbol{s}_k, t_j) =$ max temperature at location $\boldsymbol{s}_k$ on day t_j of year n

Standardize: $Y_n(\boldsymbol{s}_k, t_j) = \dfrac{X_n(\boldsymbol{s}_k, t_j) - \widehat{\mu}(\boldsymbol{s}_k, t_j)}{\widehat{\sigma}(\boldsymbol{s}_k, t_j)}$ where

$$\widehat{\mu}(\boldsymbol{s}_k, t_j) = \frac{1}{N} \sum_{n=1}^{N} X_n(\boldsymbol{s}_k, t_j)$$

$$\widehat{\sigma}(\boldsymbol{s}_k, t_j)^2 = \frac{1}{N-1} \sum_{n=1}^{N} \Big(X_n(\boldsymbol{s}_k, t_j) - \widehat{\mu}(\boldsymbol{s}_k, t_j) \Big)^2 .$$

Heat wave statistic:

$$Z^*(\boldsymbol{s}_k, t_j) = \min_{k \in I_K(k)} \left[\frac{1}{\ell} \sum_{t_j - \ell < t_i \leq t_j} Y_n(\boldsymbol{s}_k, t_j) \right]$$

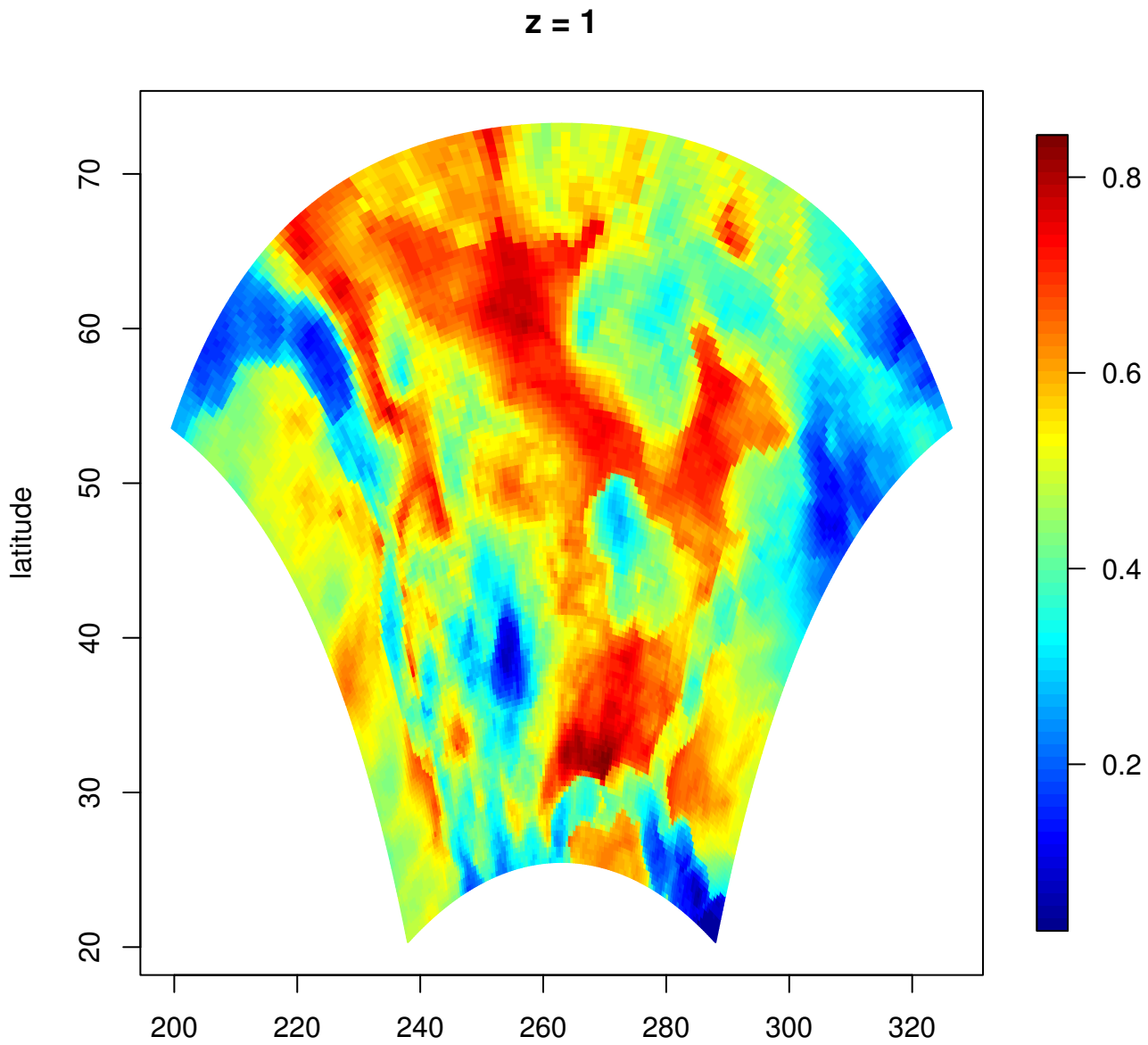
$K = 50$ nearest neighbors, $\ell = 9$ day moving average

Nearest neighbor map

Examples of 50 nearest neighbors

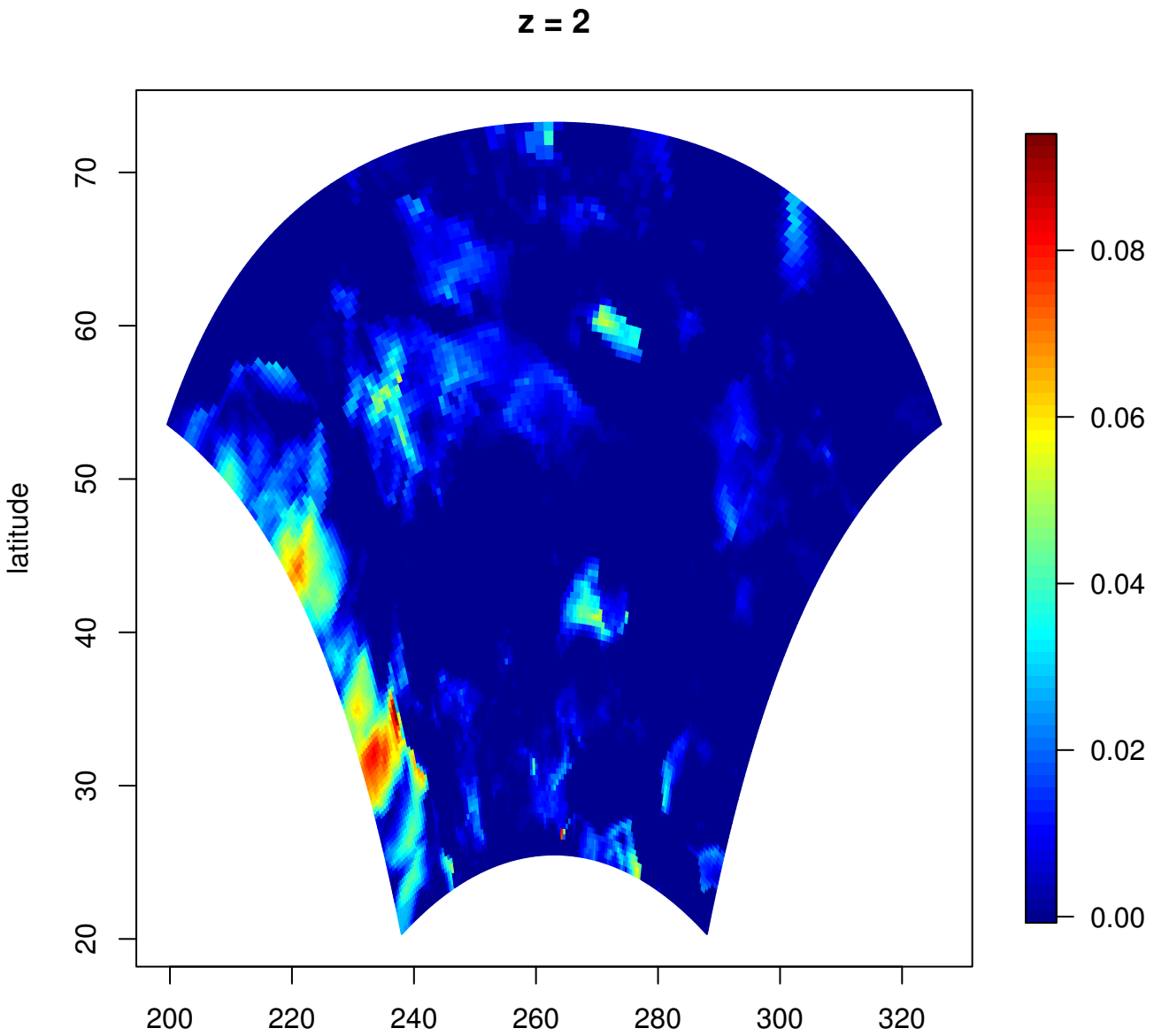
Heat wave map

Probability that $Z^*(\boldsymbol{s}_k, t_j) > 1$ (mild heat wave)



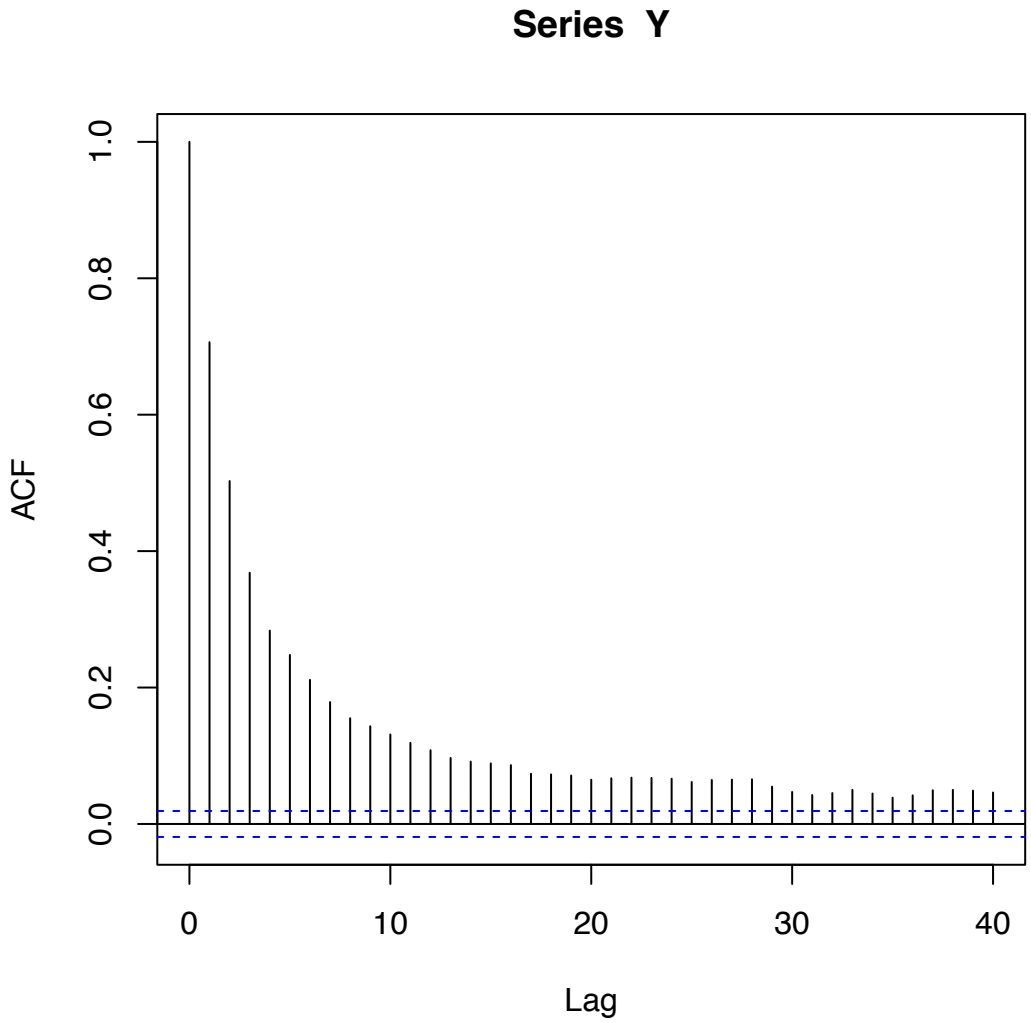
Heat wave map

Probability that $Z^*(\boldsymbol{s}_k, t_j) > 2$ (strong heat wave)



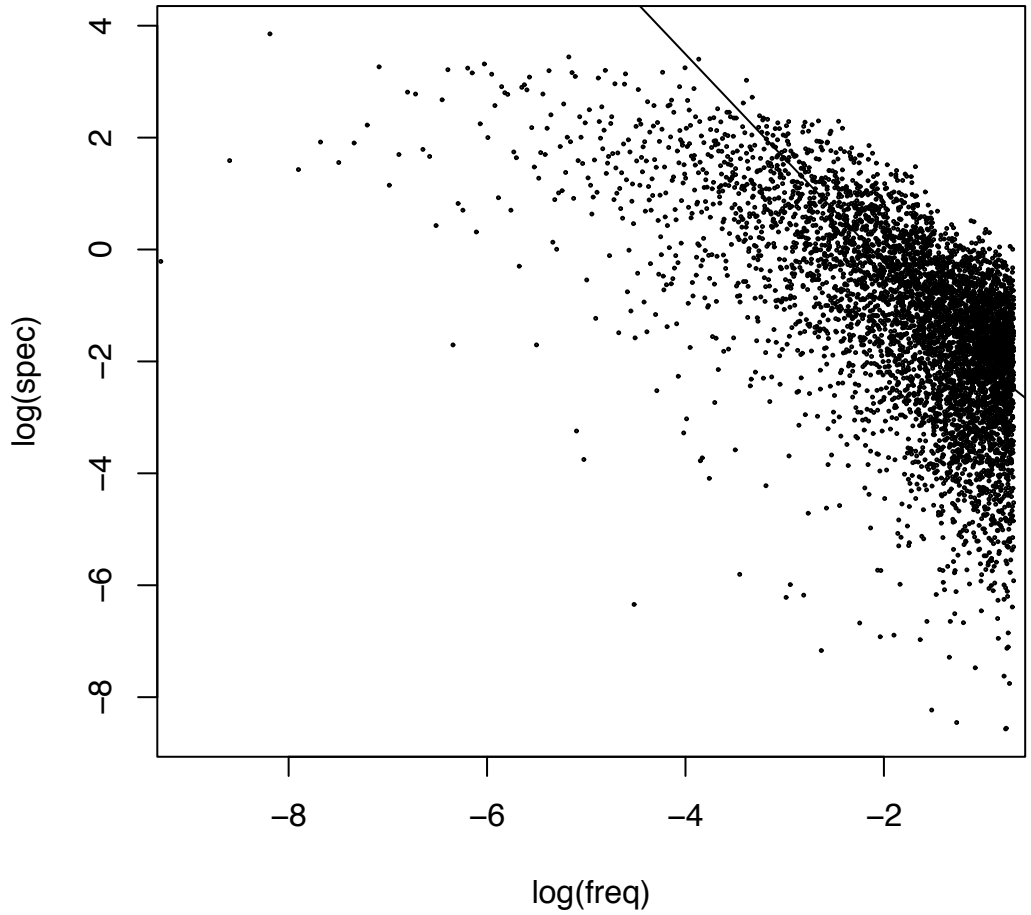
Long range dependence?

Autocorrelation at one site suggests LRD.



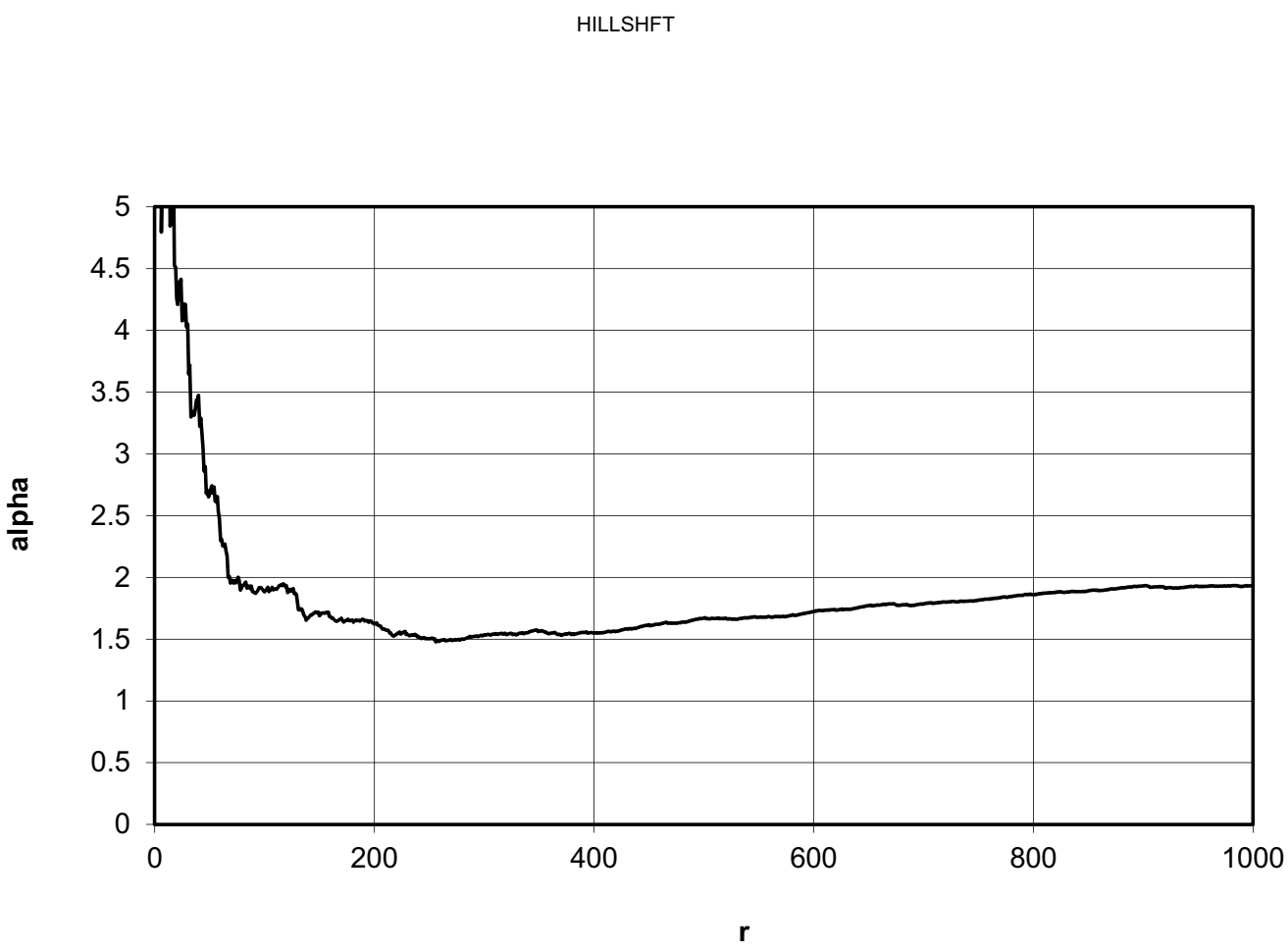
Check spectral density

Power law spectral density also suggests LRD $d = 0.9$.



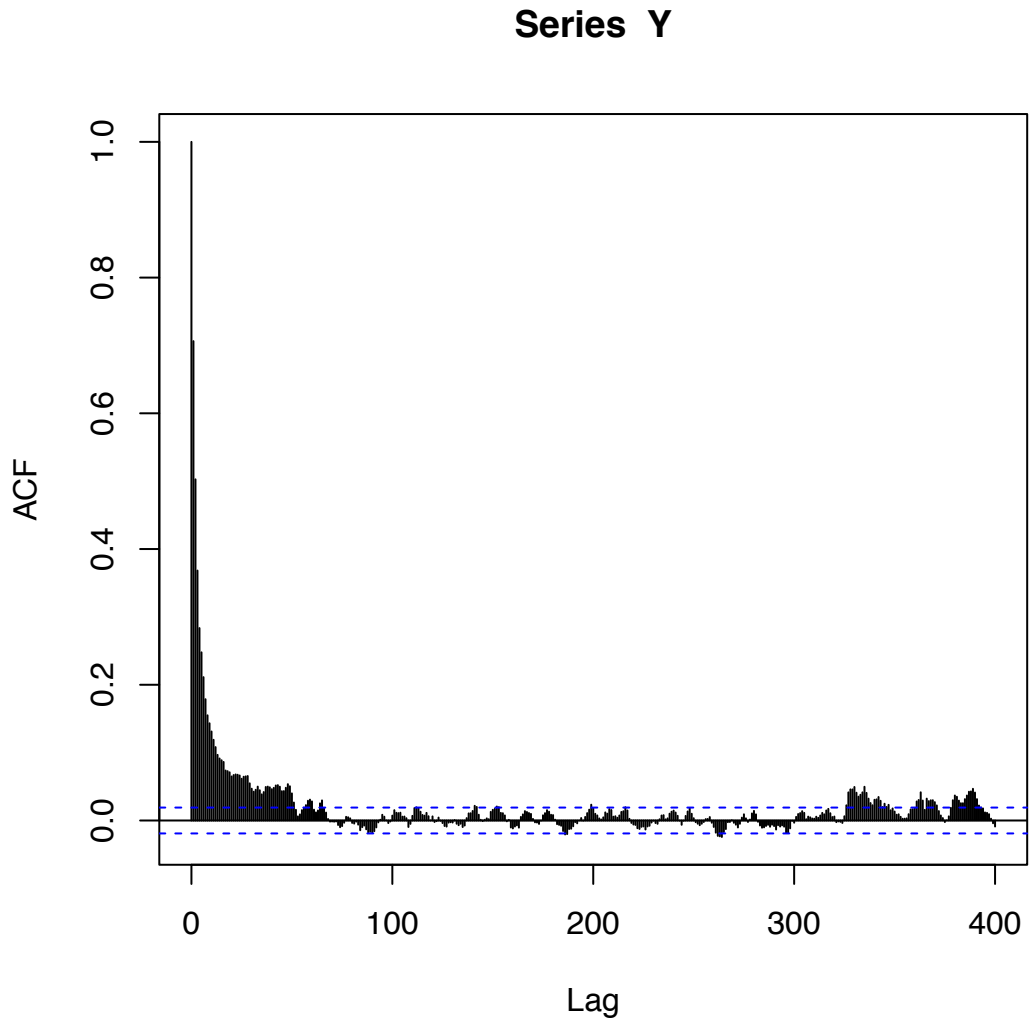
LRD and heavy tails

Hill estimator also indicates heavy tail with $\alpha \approx 1.5$.



Question: Is $Y_n(s_k, t_j)$ stationary?

Oscillation in the ACF suggests periodic behavior.



Periodic ARMA model

One spatial location: $\tilde{X}_{365\,(n-1)+t} = X_n(\boldsymbol{s}_k, t)$

PARMA$_S(p, q)$ model:

$$X_t - \sum_{j=1}^{p} \phi_t(j) X_{t-j} = \varepsilon_t + \sum_{j=1}^{q} \theta_t(j) \varepsilon_{t-j}$$

Mean-centered data $X_t = \tilde{X}_t - \mu_t$

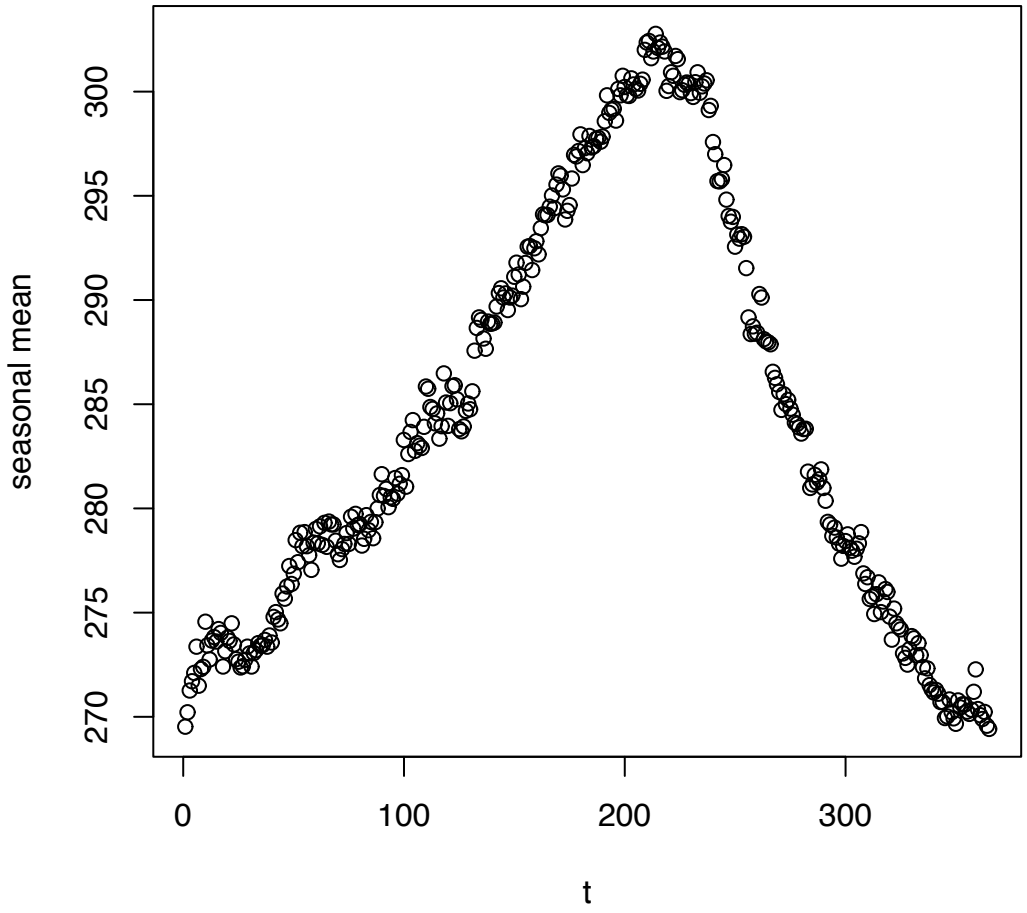
Standardized errors $\delta_t = \sigma_t^{-1} \varepsilon_t$ assumed IID

Parameters $\phi_t(j)$, $\theta_t(j)$, σ_t all periodic with period $S = 365$.

R package `perARMA` is quite convenient

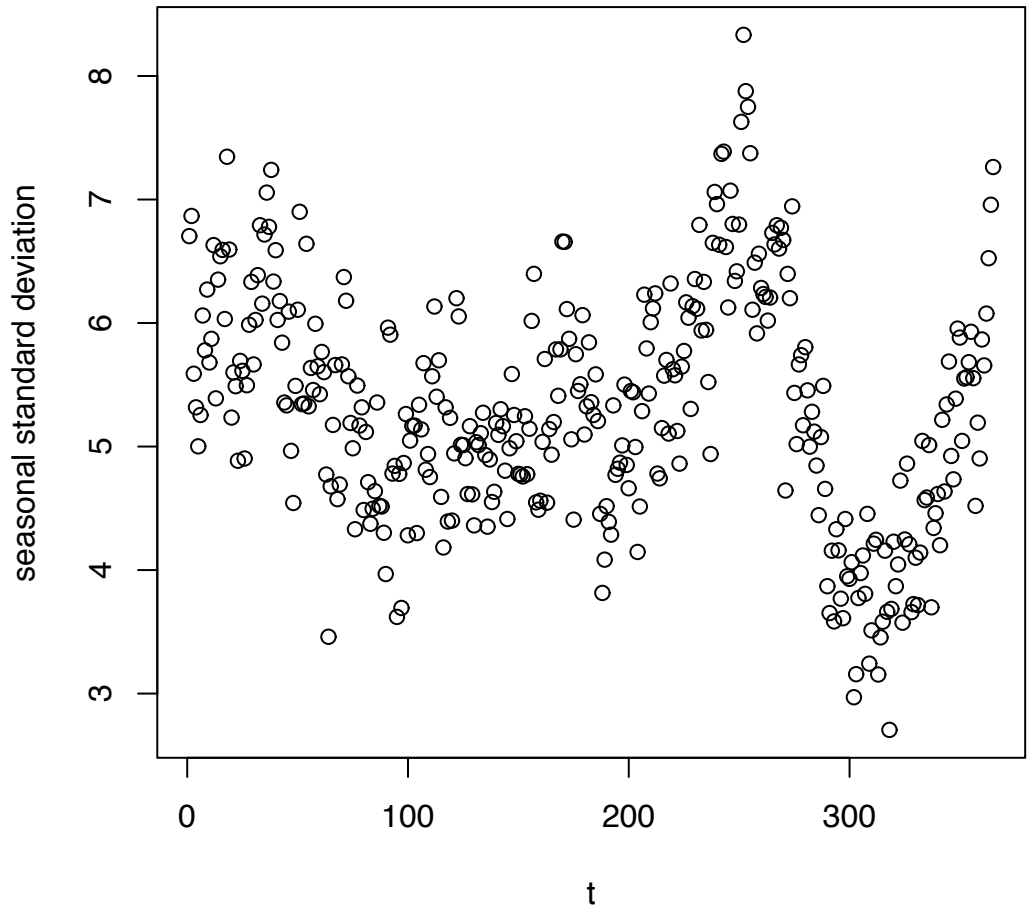
Periodic mean

Plot suggests some seasonal variation. Is SARIMA enough?



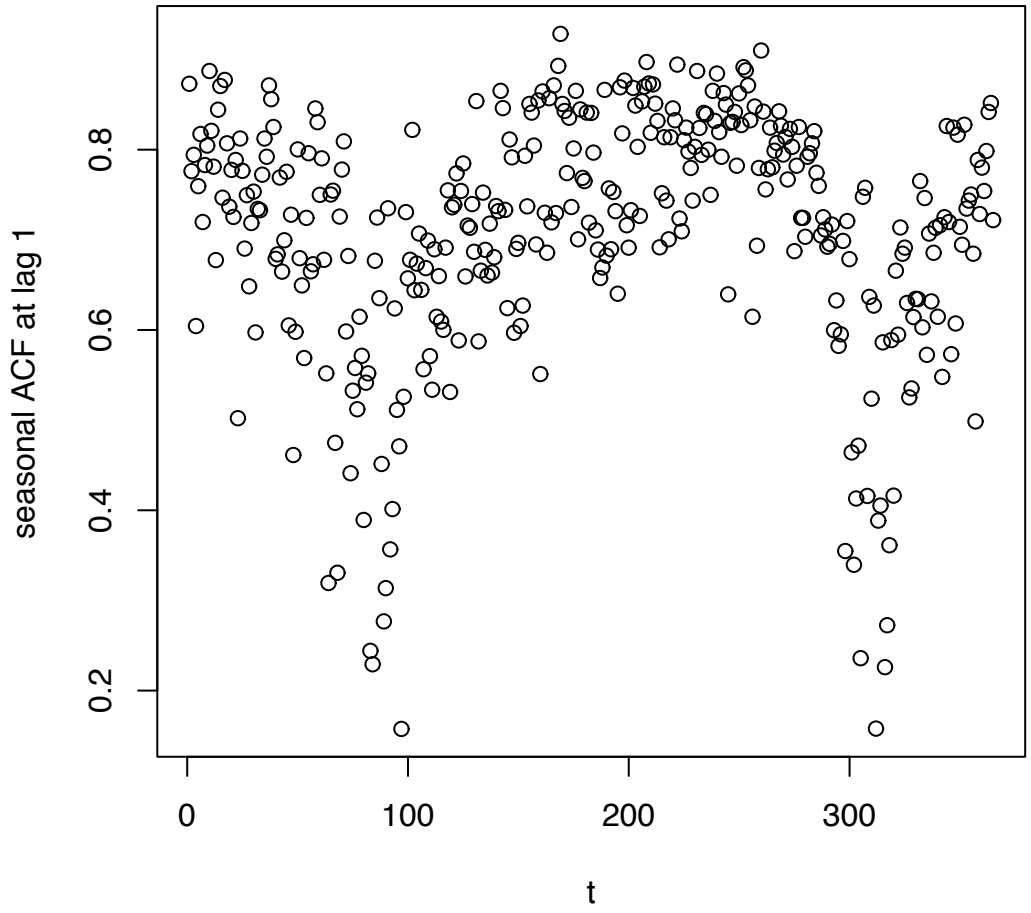
Periodic standard deviation

SARIMA needs σ_t constant. Is standardizing enough?



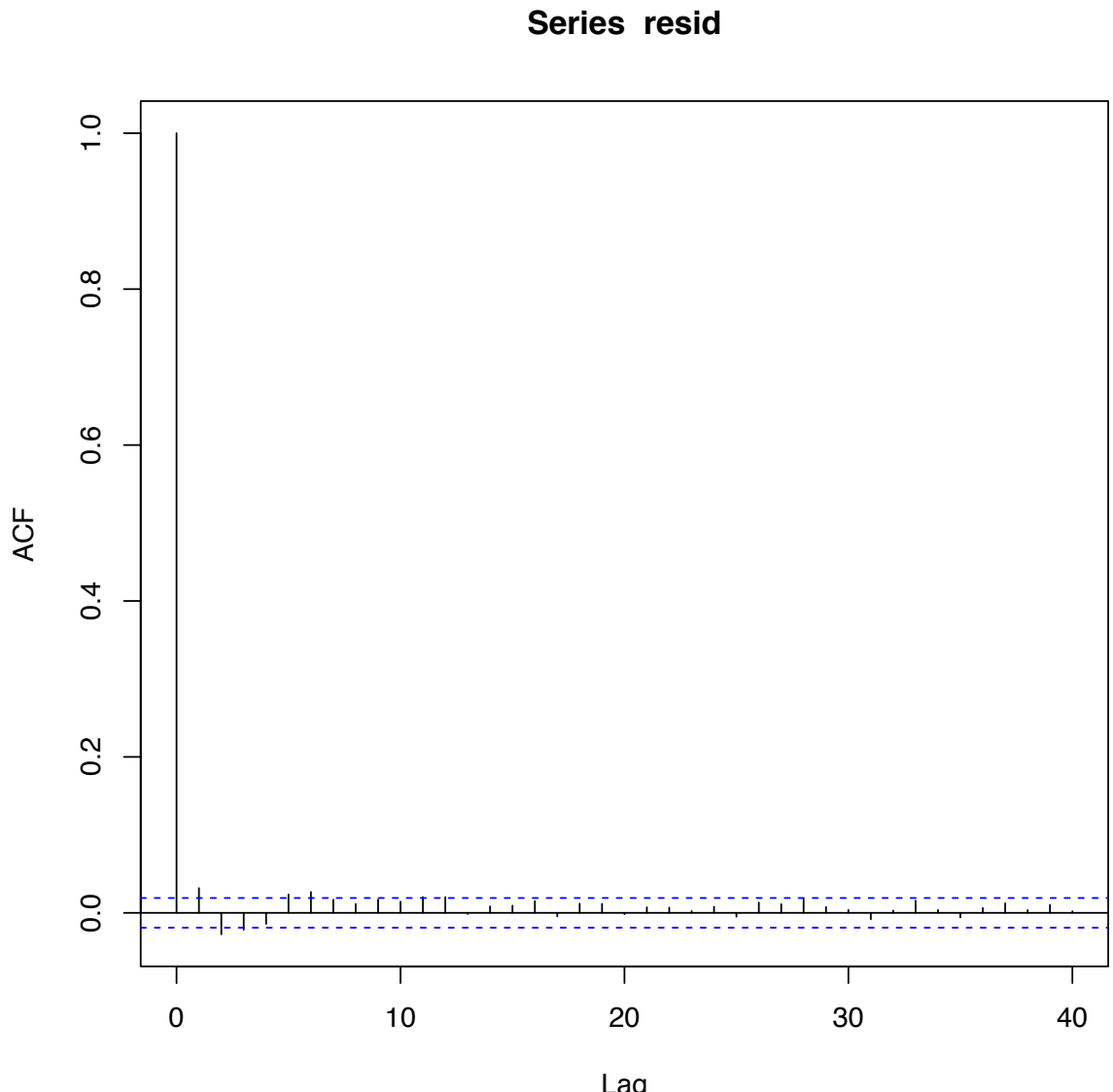
Periodic autocorrelation function at lag 1

Standardizing does not produce a stationary process.



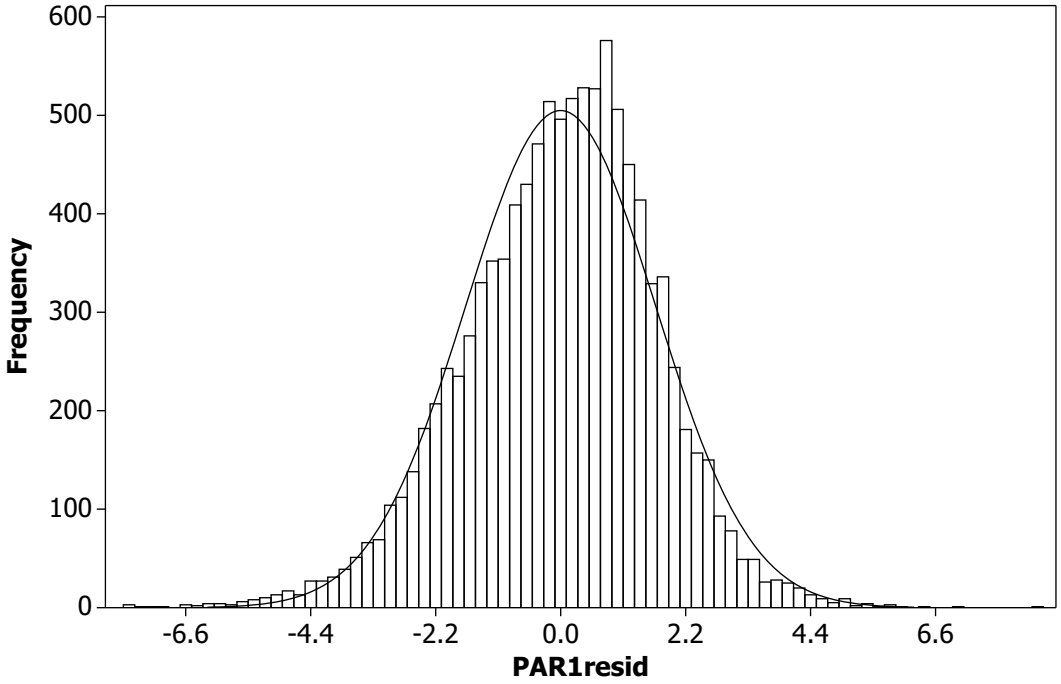
Fit a periodic AR(1) via perARMA

Residuals indicate a good fit. Nonstationary data mimics LRD!



Model residuals are approximately normal

PAR(1) residuals via `perARMA` close to a normal PDF.



A cautionary tale

Temperature data exhibits a heavy tail with infinite mean $\alpha < 2$

Standardized data exhibits LRD $d = 0.9$

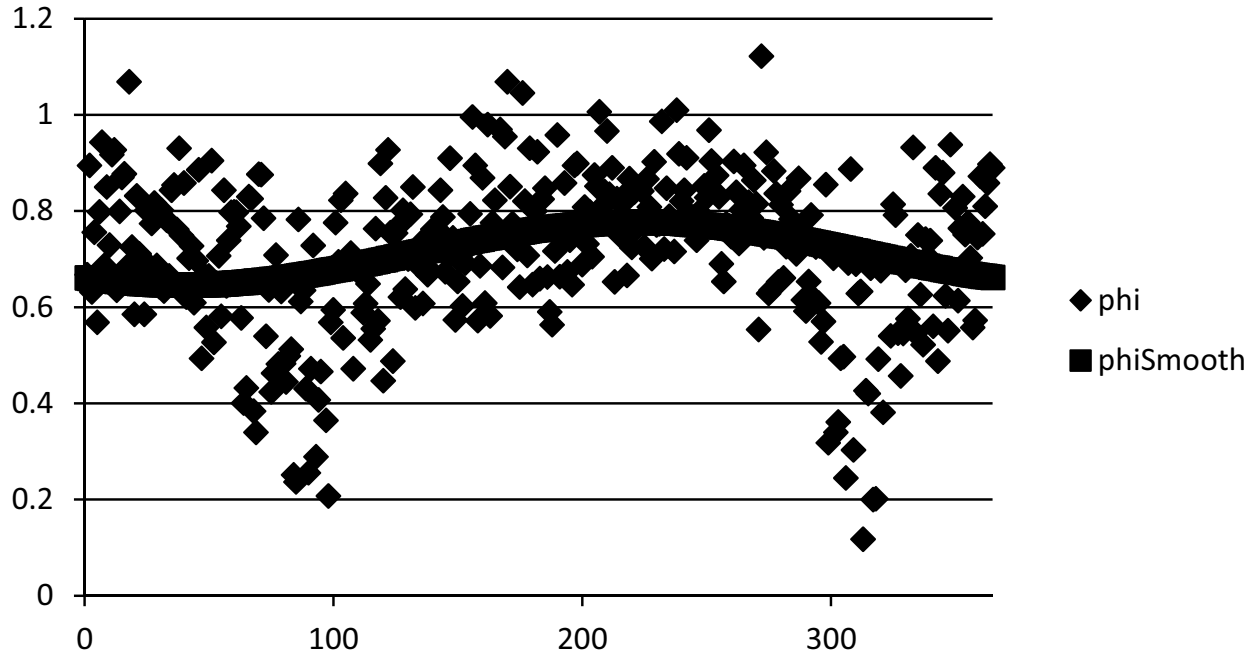
Periodic ARMA model removes the heavy tail

Periodic ARMA model removes the LRD

Nonstationary data mimics LRD and heavy tail!

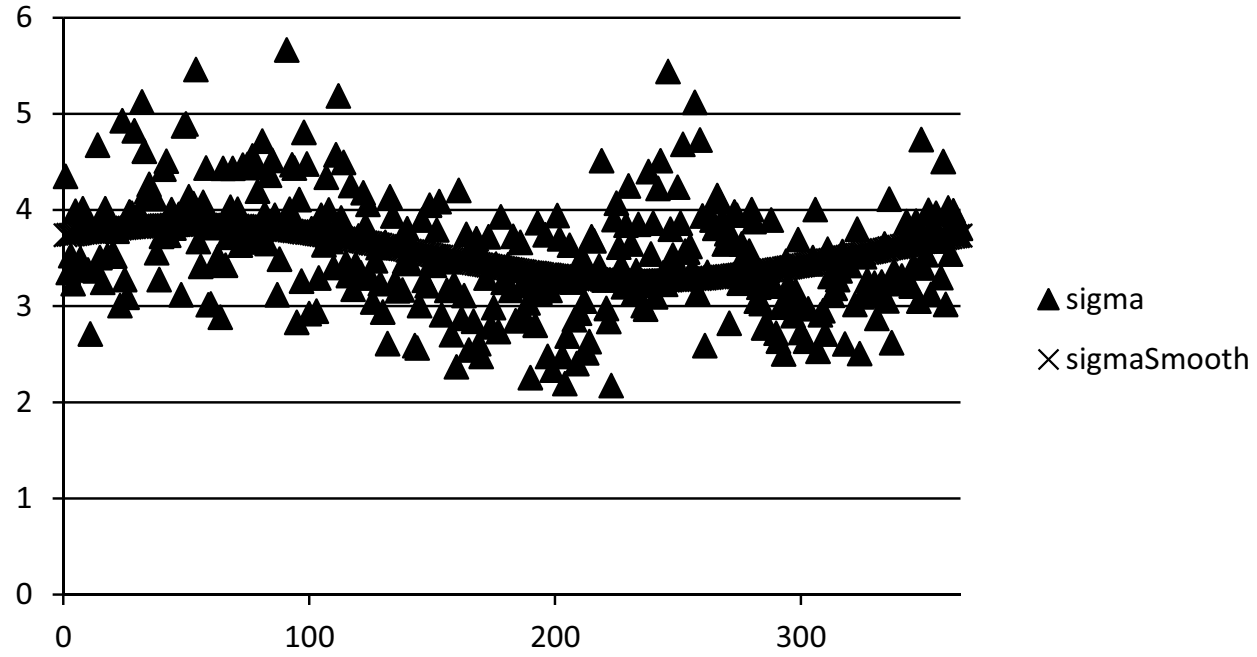
Smoothing the PARMA parameters

Discrete Fourier transform smooths PAR(1) coefficients $\phi_t(1)$



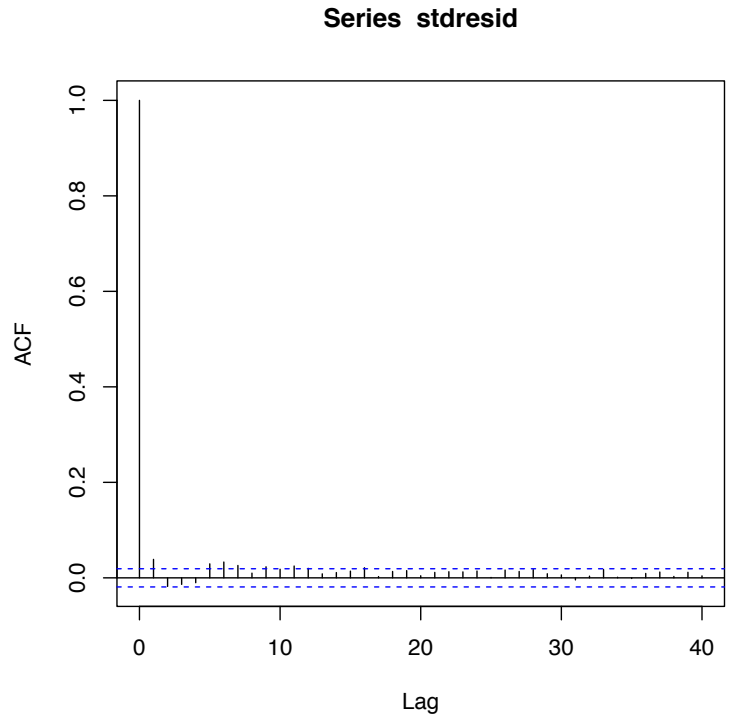
Smoothing the seasonal standard deviation

Discrete Fourier transform smooths PAR(1) coefficients σ_t



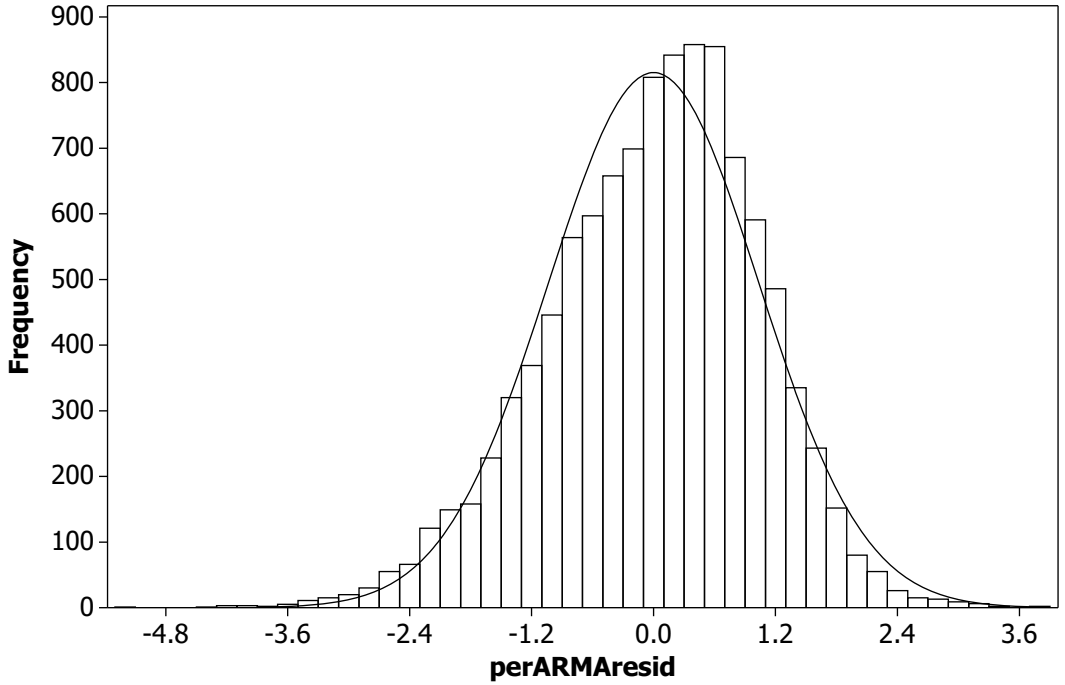
Residuals of the smoothed PAR(1) model

Smoothed model has only 4 parameters. Residuals still uncorrelated.



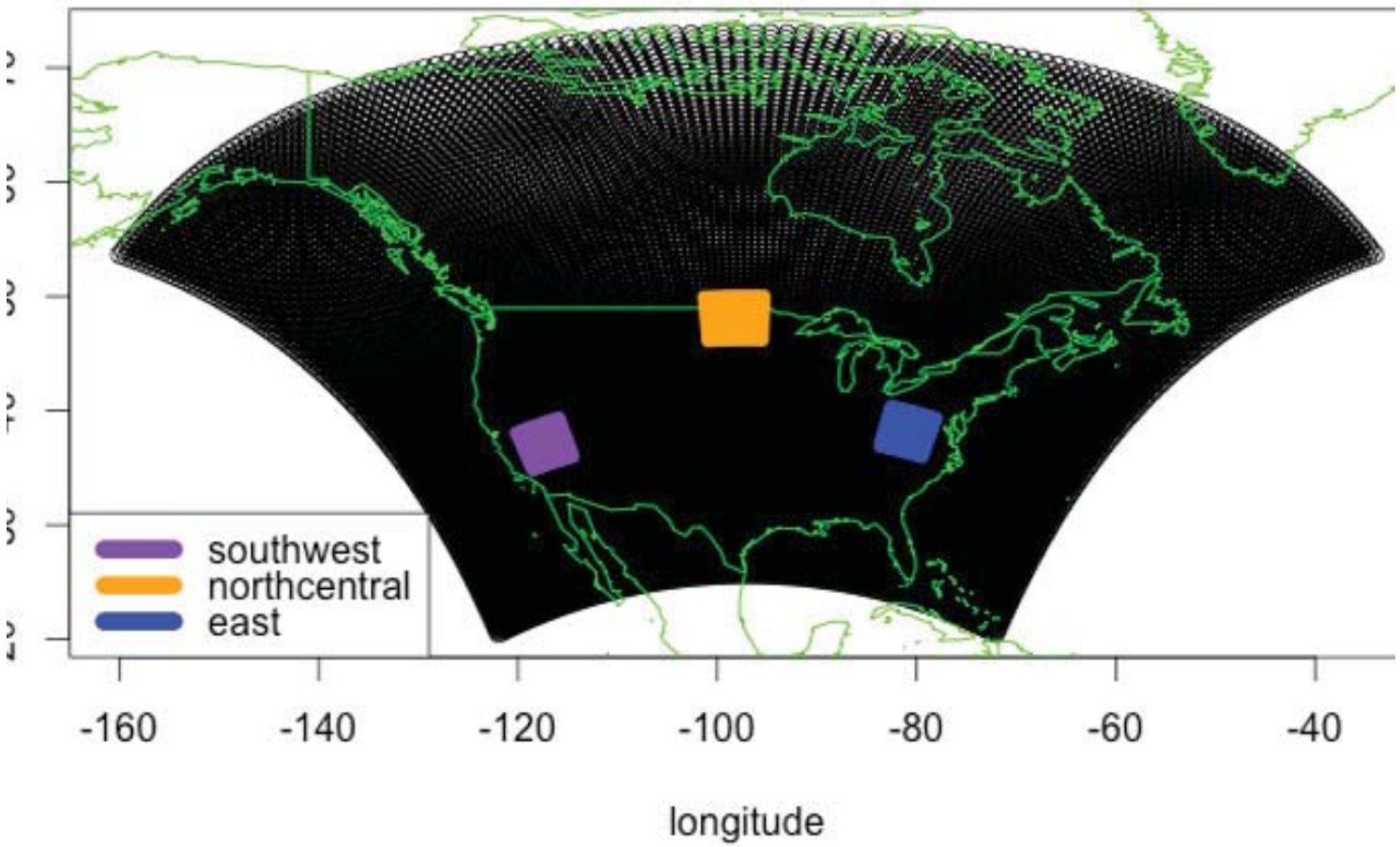
Residuals of the smoothed PAR(1) model

Histogram of smoothed PAR(1) residuals fits a normal PDF reasonably well, but with a bit of negative skewness.



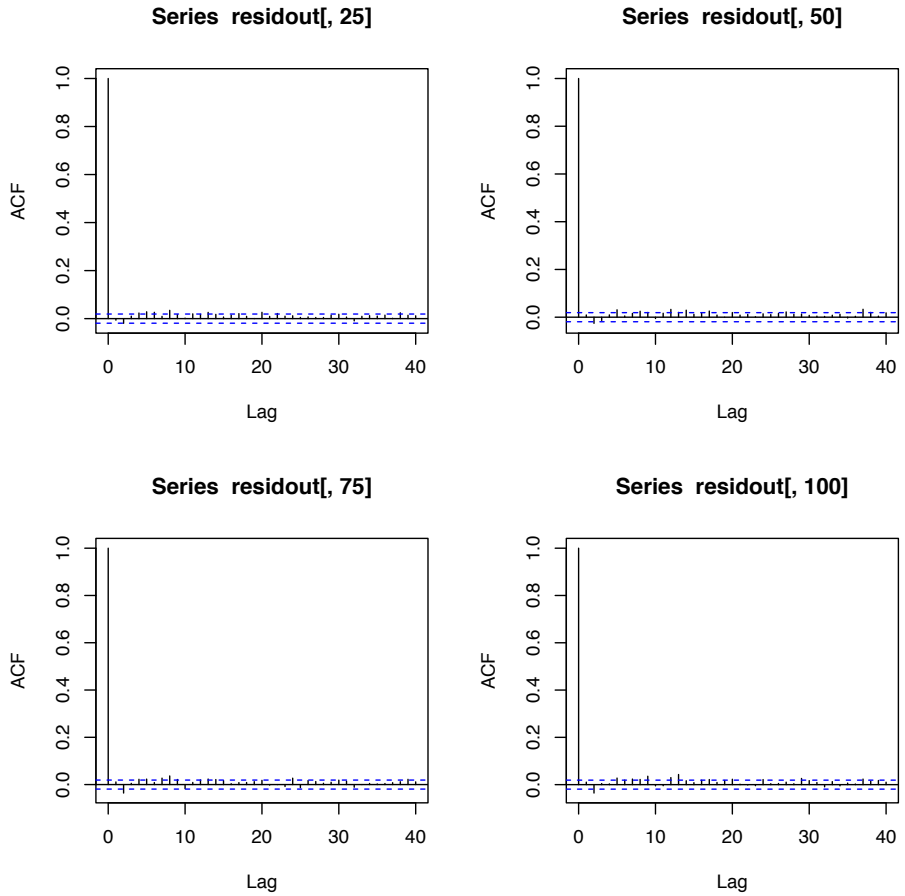
Test on three 10×10 grids

Try a PAR(p) model on three test grids. Find $p = 1$ is sufficient in SW grid, $p = 3$ enough for North-Central and Eastern grids.



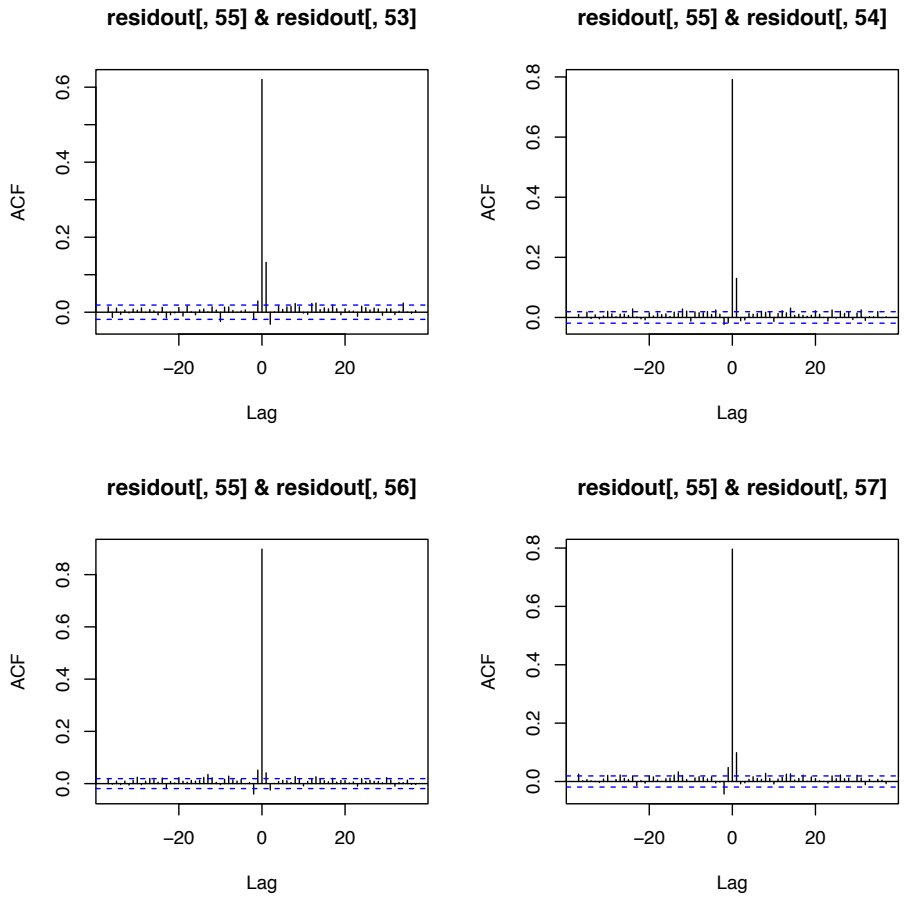
Check ACF for some randomly selected sites

Residual ACF plots, at four sites in the SW grid, PAR(1) model fit using Yule-Walker method via `perARMA`.



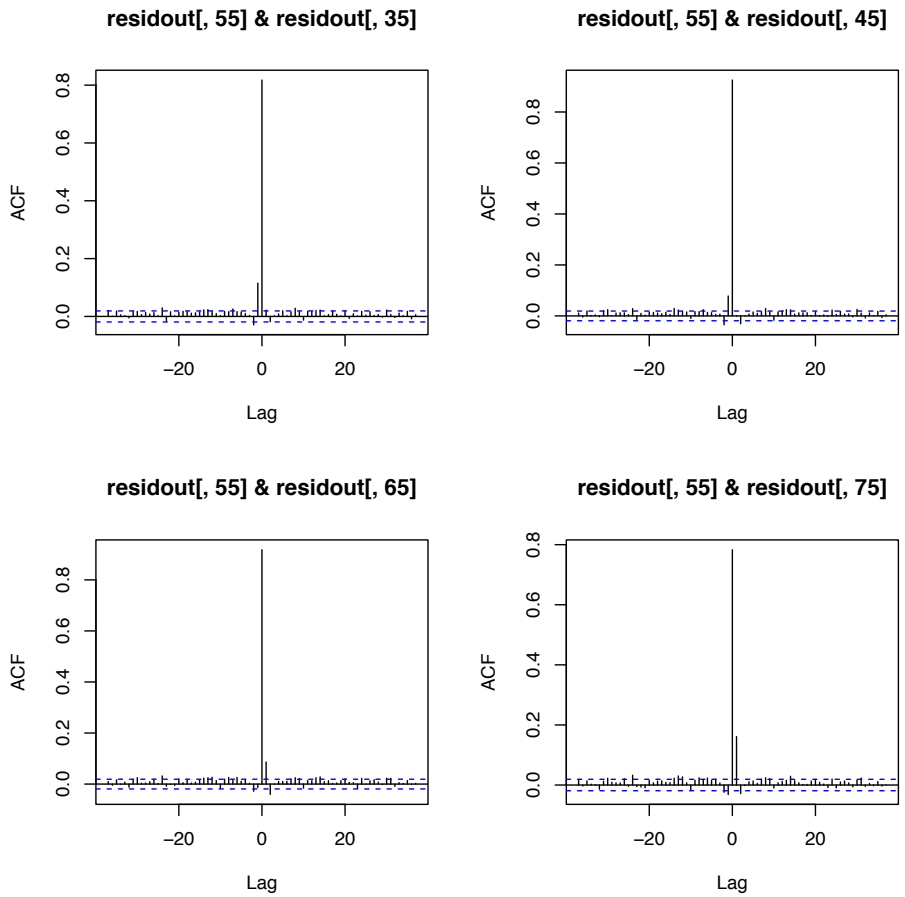
Spatial cross-correlation at nearby sites

Center site in SW grid and 4 nearest neighbors (in longitude)



Spatial cross-correlation at nearby sites

Center site in SW grid and 4 nearest neighbors (in latitude). Perhaps a spatial AR(1) will be sufficient?



References

1. I.B. Aban and M.M. Meerschaert, Shifted Hill's estimator for heavy tails. Communications in Statistics: Simulation and Computation, Vol. 30 (2001), No. 4, pp. 949–962.

2. I.B. Aban, M.M. Meerschaert, and A.K. Panorska, Parameter Estimation for the Truncated Pareto Distribution, Journal of the American Statistical Association: Theory and Methods, Volume 101 (2006), Number 473, pp. 270–277.

3. P.L. Anderson and M.M. Meerschaert, Periodic moving averages of random variables with regularly varying tails, The Annals of Statistics, Vol. 25 (1997), No. 2, pp. 771–185.

4. P.L. Anderson, M.M. Meerschaert, Y.G. Tesfaye, Fourier-PARMA Models and Their Application to Modeling of River Flows, Journal of Hydrologic Engineering, Vol. 12 (2007), No. 5, pp. 462–472.

5. Y.G. Tesfaye, P.L. Anderson, and M.M. Meerschaert, Asymptotic results for Fourier-PARMA time series, Journal of Time Series Analysis, Vol. 32 (2011), No. 2, pp. 157–174.