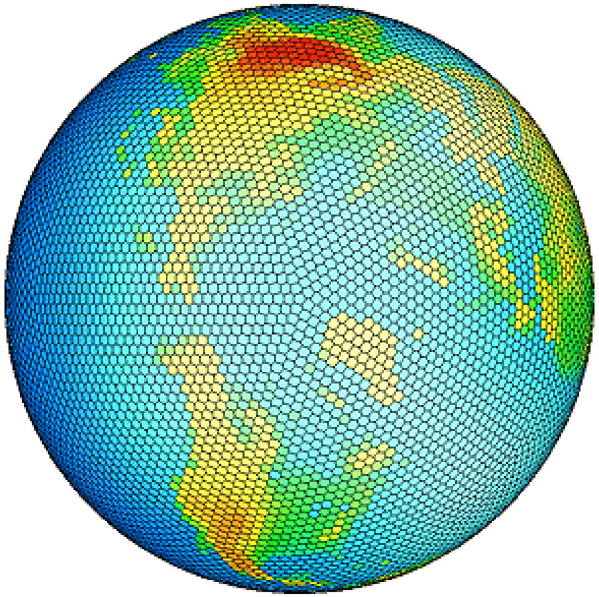
Granularity Issues on Climatic Time Series

A. Charpentier (UQAM & Université de Rennes 1)

joint work with J.C. Bouëttte, J.F. Chassagneux, D. Sibaï, R. Terron,



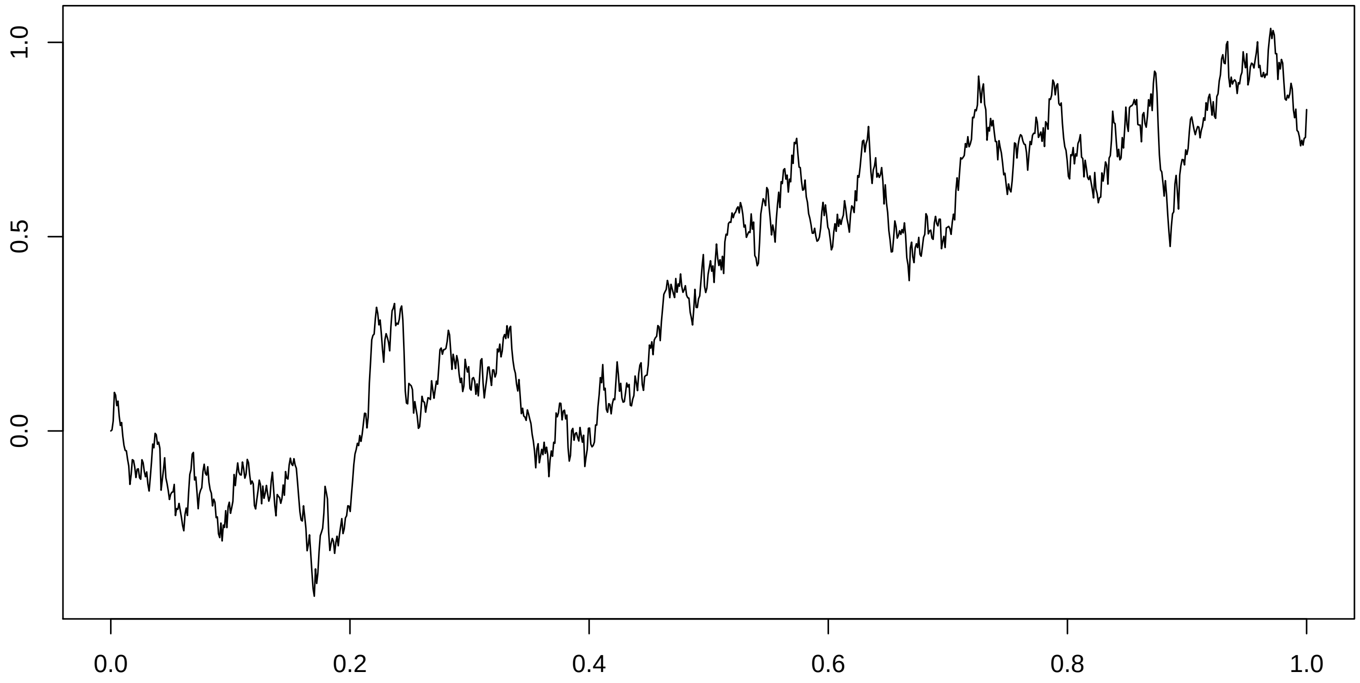
Buenos Aires, November 2015

Big Data & Environment Workshop

http://freakonometrics.hypotheses.org

Self-similar Time Series, and Granularity Issues

$Y_{at} \overset{\mathcal{L}}{=} a \cdot Y_t$, see Mandelbrot (1982) or Embrechts & Maejima (2002).



The more data we get, the better... But what about climate time series?

'Period of Return' in the context of Climate Data

1.2.2. The Distribution of Repeated Occurrences. To derive the notion of return period we construct a dichotomy for a continuous variate. First, we consider the observations equal to or larger than a certain large value x. (This exceedance is the event in which we are interested.) Second, we consider the observations smaller than this value. Let

$$q = 1 - p = F(x) \tag{1}$$

be the probability of a value smaller than x. Observations are made at regular intervals of time, and the experiment stops when the value x has been exceeded once. We ask for the probability $w(v)$ that the exceedance happens for the first time at trial v (geometric distribution).

Gumbel (1958). Statistics of Extremes. Columbia University Press

Let T be the time of first success for some events occuring with yearly probabiliy p, then

$$\mathbb{P}[T = k] = (1-p)^{k-1}p \text{ so that } \mathbb{E}[T] = \frac{1}{p}$$

(geometric distribution, discrete version of the exponential distribution).

Models for River Levels and Flood Events

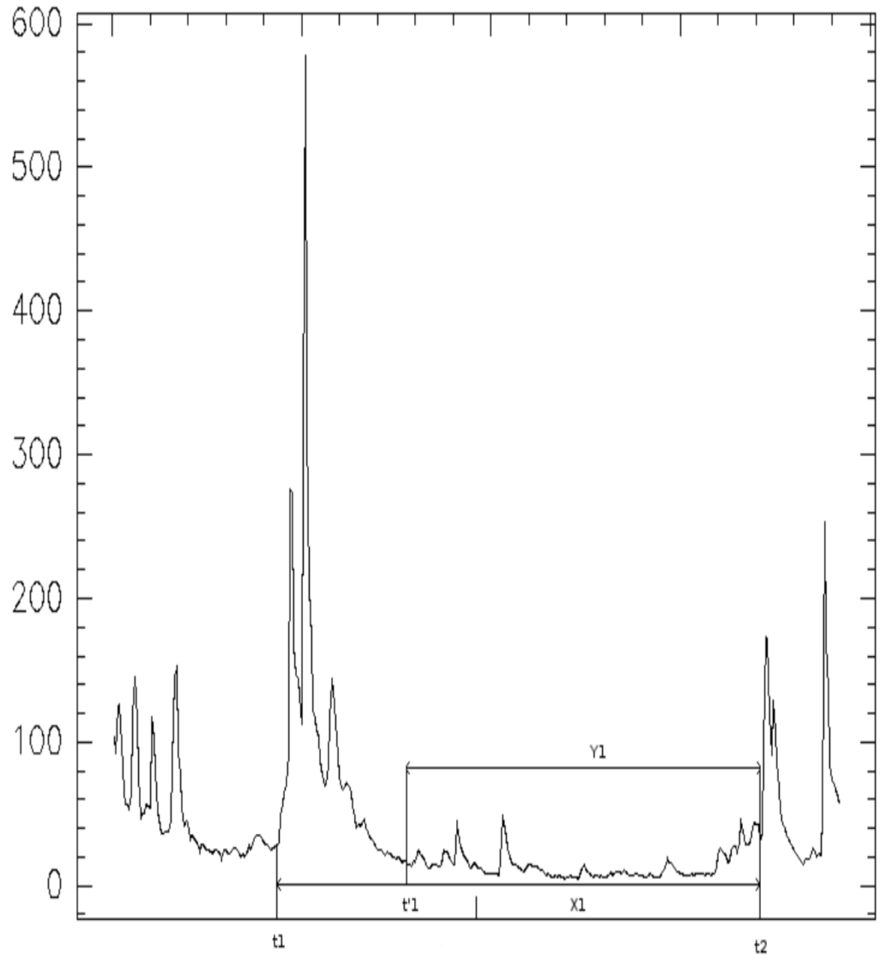
In hydrological papers, huge interest on Annual Maximum time series

- Hurst (1951) observed that annual maximum exhibit long-range dependence (so called Hurst effect),
- Gumbel (1958) observed that annual maximum were i.id with a similar distribution (so called Gumbel distribution)

How could it be identical series be at the same time independent and with long-range dependence? Hurst (1951) used 700 years of data on the Nile, Gumbel (1958) used European data, over less than a century.

Can't we use more data to model flood events?

Flood Events



High Frequency Models (for Financial Data)

On financial data,

- "*traditional*" approach (time series): consider the closing data price, X_t at the end of day t, i.e. regularly spaced observations,
- "*high frequency daya*": the price X is observed at each transaction: let T_i denote the data of the ith transaction, and X_i the price paid.

See e.g. ACD - Autoregressive Conditional Duration models, introduced par in Engle & Russell (1998).

In practice, three information are stored: (1) date of transaction, or time between two consecutive transactions, on the same stock; (2) the volume, i.e. number of stocks sold and bought (3) the price, i.e. individual stock price (or total price exchanged)

Flood Events

The analogous of a transaction is a flood event, where 4 variables are kept,

- time length of the flood event
- time between two consecutive flood events
- volume V_i
- peak P_i

Remark: see Todorovic & Zelenhasic (1970) and et Todorovic & Rousselle (1970) where marked Poisson processes were considered.

Some 'Optimal' Threshold

The choice of the threshold is crucial. Standard tradeoff

- should be low to have more events
- should be high to have significant flood events

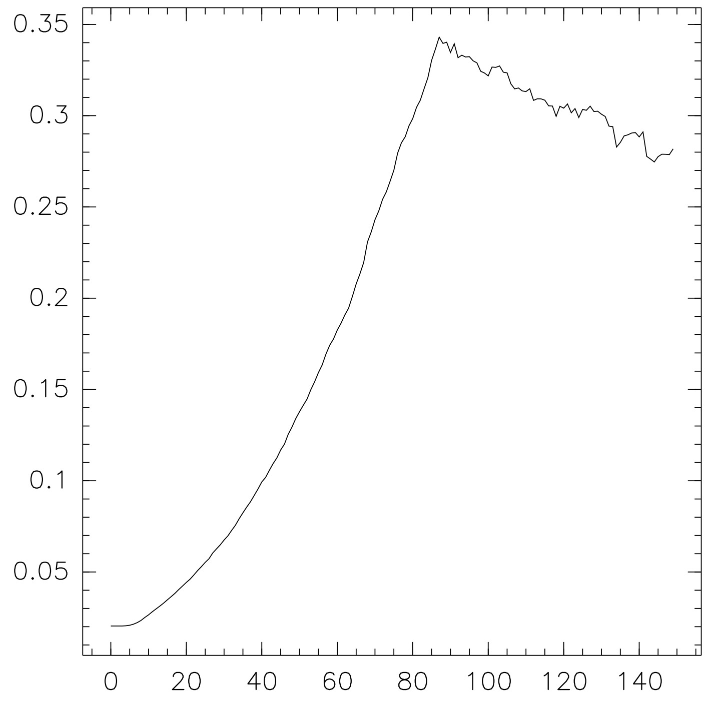
Standard technique in hydrology: given some function f (e.g. f affine), solve

$$u\star = \text{argmax}\{\mathbb{P}(X > f(u)|X > u)\}$$

or its empirical couterpart

$$u\star = \text{argmax}\left\{\frac{\#\{X_i > f(u)\}}{\#\{X_i > u\}}\right\}$$

with e.g. $f(x) = 1,5x + 5$.



A Two-Duration Model

Engle & Lunde (2003) in *trades and quotes: a bivariate point process*, consider a two duration model, that can be used here.

The two dates are T_i beginning of ith flood, and T_i' end of the flood. Set

- $X_i = T_{i+1} - T_i$ the time length between the begining of two consecutive floods
- $Y_i = T_{i+1} - T_i'$ the time length between the end of a flood and the begining of the next one

Engle & Russell (1998) ACD(p,q) Model

In the one-duration model, let X_i denote the time lengths ($X_i = T_i - T_{i-1}$), and $\mathcal{H}_i = \{X_1, ..., X_{i-1}\}$. Then

$$\begin{cases} X_i = \Psi_i \cdot \varepsilon_i, \text{ with } (\varepsilon_i) \text{ i.i.d. noise} \\ \mathbb{E}(X_i|\mathcal{H}_{i-1}) = \Psi_i = \omega + \displaystyle\sum_{k=1}^{p} \alpha_k X_{i-k} + \sum_{k=1}^{q} \beta_k \Psi_{i-k}, \end{cases}$$

i.e.

$$X_i = \omega + \sum_{k=1}^{\max\{p,q\}} (\alpha_k + \beta_k) X_k - \sum_{k=1}^{q} \beta_k \eta_{i-k} + \eta_i,$$

where $\eta_i = X_i - \Psi_i = X_i - \mathbb{E}(X_i|\mathcal{H}_{i-1})$ (ARMA($\max\{p,q\}, q$) representation of the ACD(p,q)).

In the Exponential ACD(1,1), (ε_i) is an exponential noise

$$\mathbb{E}(X_i|\mathcal{H}_{i-1}) = \Psi_i = \theta + \alpha X_i + \beta \Psi_{i-1}, \text{ with } \alpha, \beta \geq 0 \text{ and } \theta > 0,$$

Engle & Russell (1998) ACD(p, q) Model

More generally, the conditional density of X_i is

$$f(x|\mathcal{H}_i) = \frac{1}{\Psi_i(\mathcal{H}_i, \theta)} \cdot g_\varepsilon\left(\frac{-}{x}\Psi_i(\mathcal{H}_i, \theta)\right)$$

e.g. $g_\varepsilon(\cdot) = \exp[-\cdot]$, if $\varepsilon \sim \mathcal{E}(1)$.

Inference is very similar to GARCH(1,1), the proof being the same as the one in Lee & Hansen (1994) and Lumsdaine (1996).

The Two-Duration Model

As in Engle & Lunde (2003), consider some two-EACD model,

$$f(x_i|\mathcal{H}_i) = \frac{1}{\Psi_i(\mathcal{H}_i,\theta_1)}\cdot\exp\left(-\frac{x_i}{\Psi_i(\mathcal{H}_i,\theta_1)}\right)$$

where

$$\Psi_i(\mathcal{H}_i,\theta_1) = \exp\left(\alpha+\delta\log(\Psi_{i-1})+\gamma\frac{X_{i-1}}{\Psi_{i-1}}+\beta_1 P_{i-1}+\beta_2 V_{i-1}\right),$$

while

$$g(y_i|x_i,\mathcal{H}_i) = \frac{1}{\Phi_i(x_i,\mathcal{H}_i,\theta_2)}\cdot\exp\left(-\frac{y_i}{\Phi_i(x_i,\mathcal{H}_i,\theta_2)}\right)$$

where

$$\Phi_i(x_i,\mathcal{H}_i,\theta_2) = \exp\left(\mu+\rho\log(\Phi_{i-1})+\gamma\frac{Y_{i-1}}{\Phi_{i-1}}+\tau\frac{x_i}{\Psi_i}+\eta_1 P_{i-1}+\eta_2 V_{i-1}\right).$$

The Two-Duration Model

Define residuals

$$\varepsilon_i = \frac{X_i}{\Psi_i(\mathcal{H}_{i-1}, \theta_1)}.$$

Since there are two kinds of floods, ordinary ones and those related to snow melt, we should consider a mixture distribution for ε, a mixture of exponentials

$$f(x) = \alpha \cdot \lambda_1 \cdot e^{-\lambda_1 \cdot x} + (1-\alpha) \cdot \lambda_2 \cdot e^{-\lambda_2 \cdot x}, x > 0.$$

or a mixture of Weibull's

$$f(x) = \alpha \cdot \lambda_1 \cdot \theta_1^{-\lambda_1} \cdot x^{\lambda_1 - 1} \cdot e^{-(x/\theta_1)^{\lambda_1}} + (1-\alpha) \cdot \lambda_2 \cdot \theta_2^{-\lambda_2} \cdot x^{\lambda_2 - 1} \cdot e^{-(x/\theta_2)^{\lambda_2}}$$

Modeling Marks

Finally,

$$f(p_i, v_i, x_i, y_i | \mathcal{H}_{i-1}) = g(p_i, v_i | \mathcal{H}_{i-1}, x_i, y_i) \cdot h(x_i, y_i | \mathcal{H}_{i-1}).$$

which can be simplified using a triangle approximation,

$$\text{Volume} = V_i = P_i \cdot \frac{X_i - Y_i}{2} = \frac{\text{peak} \times \text{flood duration}}{2},$$

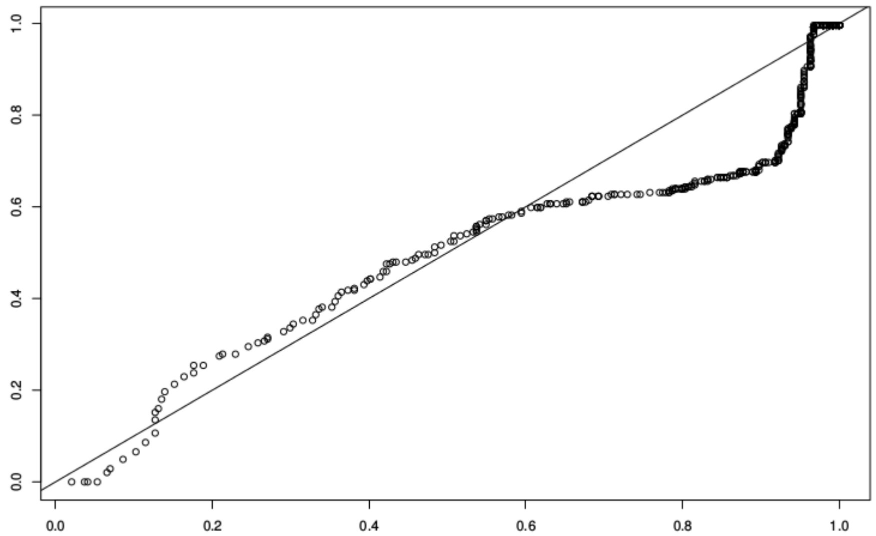
Modeling Peaks

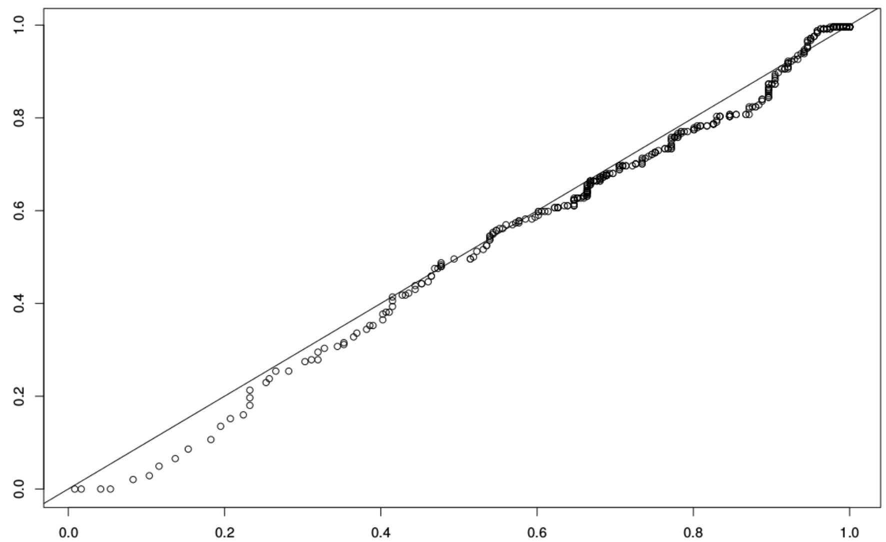
From the threshod based approach, use Pickands-Balkema-de Haan theorem and fit a Generalized Pareto distribution

$$h(p_i | \mathcal{H}_{i-1}, x_i, y_i) = \alpha \left(\frac{p_i + b(x_i - y_i) + d}{\sigma} \right)^{-(1+\alpha)}.$$

Application

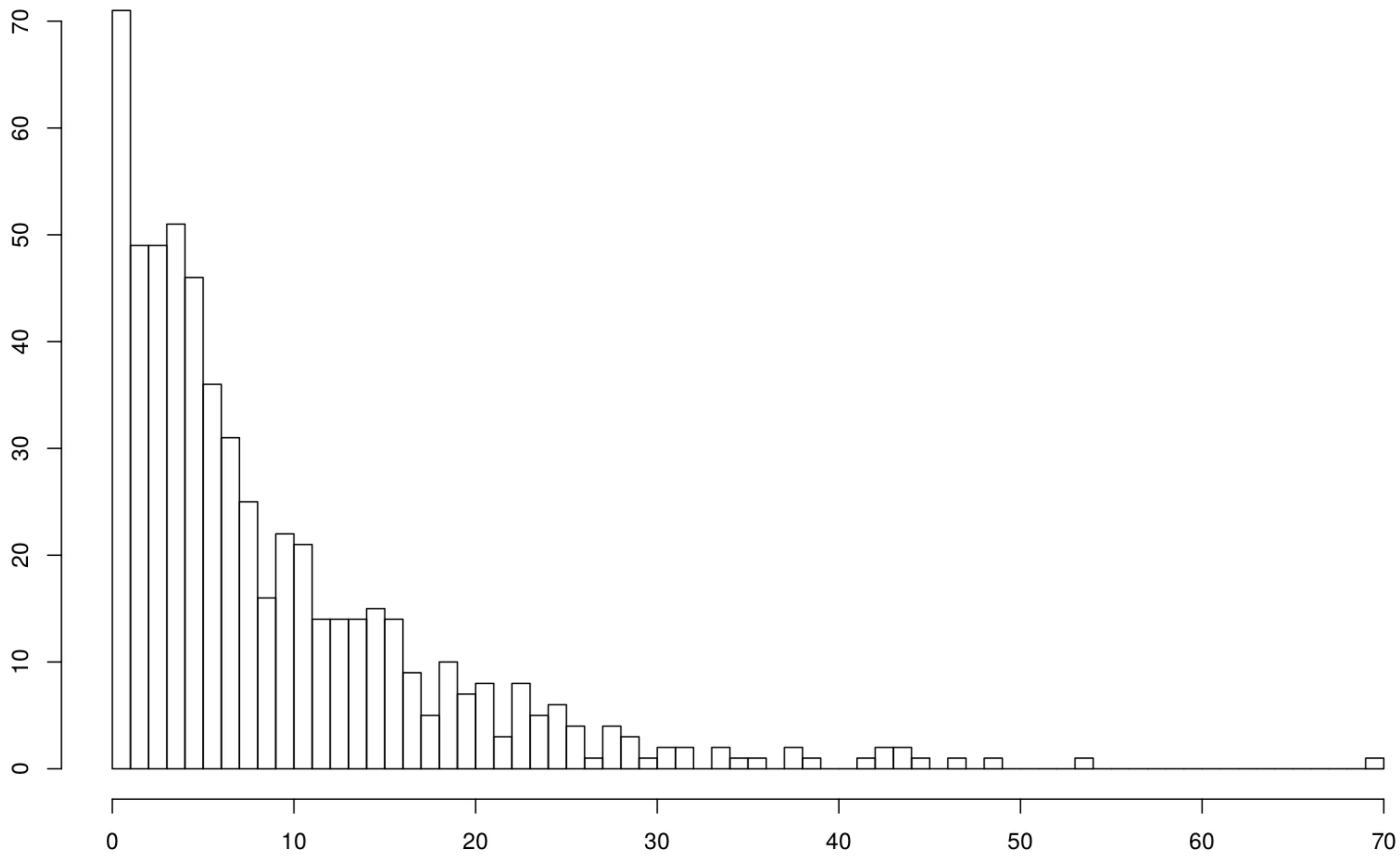
In Charpentier & Sibaï (2010), *Environmetrics*, we considered a mixture of Weibull distribution, fitted using EM algorithm, see (conditional) QQ plot, exponential vs. mixture of Weibull,





There is some dynamics, but not long memory here (from the EACD(1,1) processes).

Distribution of Time Before Next Flood Event



Long Memory and Wind Speed (very popular application)

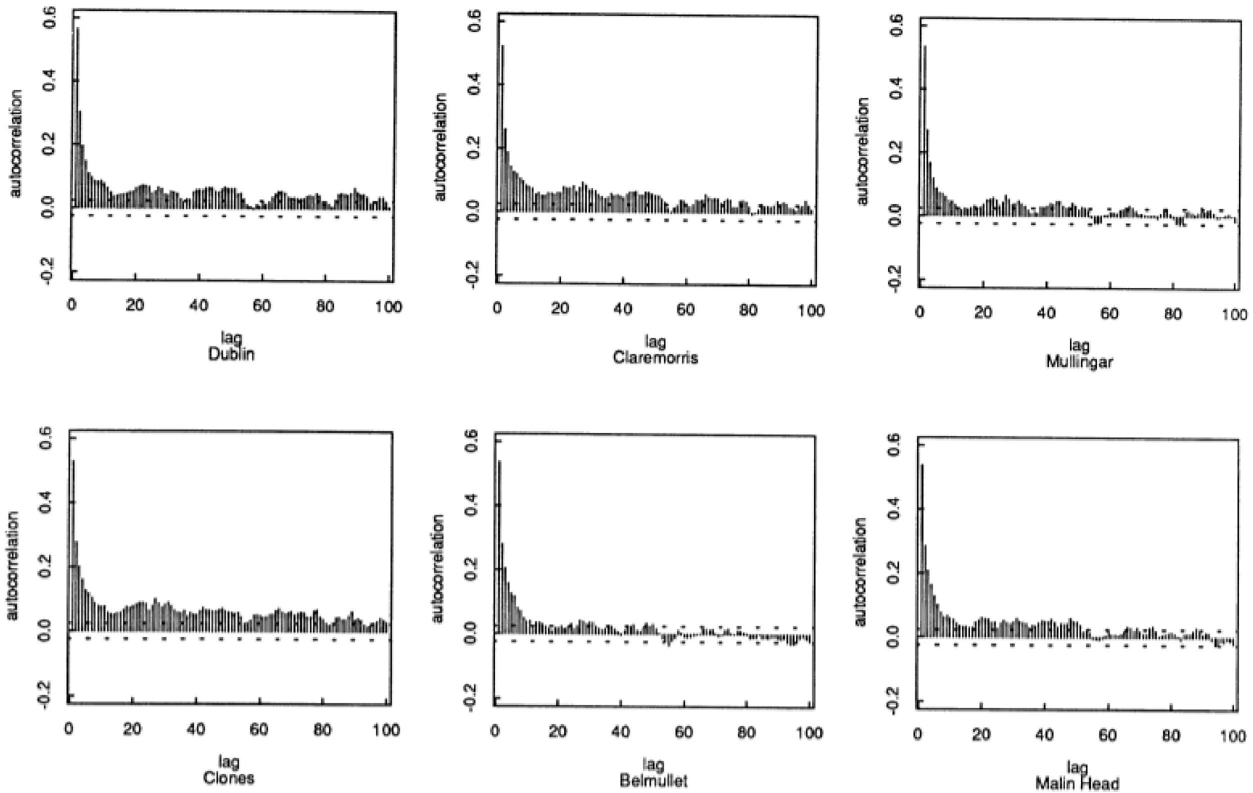
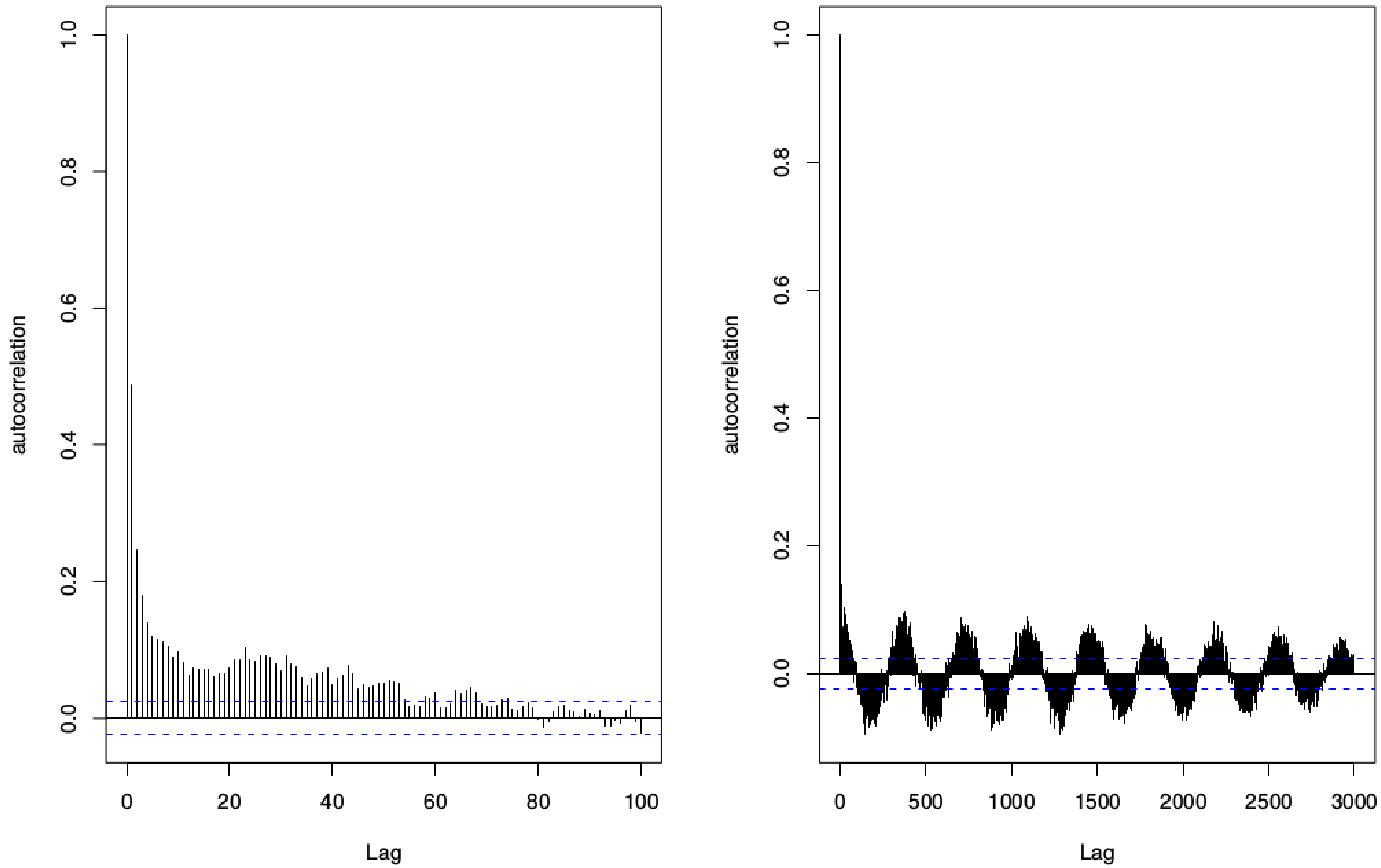


Fig. 4. Autocorrelation functions of the velocity measures for the 12 synoptic stations.

Haslett & Raftery (1989). Space-time modelling with long-memory dependence: assessing Ireland's wind power resource (with discussion). *Applied Statistics*. **38**. 1-50.

Daily Wind Speed in Ireland, long memory, really?



Modeling Stationary Time Series

Given a stationary time series (X_t), the autocovariance function, is

$$h \mapsto \gamma_X(h) = \text{Cov}(X_t, X_{t-h}) = \mathbb{E}(X_t X_{t-h}) - \mathbb{E}(X_t) \cdot \mathbb{E}(X_{t-h})$$

for all $h \in \mathbb{N}$, and its Fourier transform is the spectral density of (X_t)

$$f_X(\omega) = \frac{1}{2\pi} \sum_{h \in \mathbb{Z}} \gamma_X(h) \exp(i\omega h)$$

for all $\omega \in [0, 2\pi]$. Note that

$$f_X(\omega) = \frac{1}{2\pi} \sum_{h=-\infty}^{+\infty} \gamma_X(h) \cos(\omega h)$$

Let $\rho_X(h)$ denote the autocorrelation function i.e. $\rho_X(h) = \gamma_X(h)/\gamma_X(0)$.

Long-Range Dependence

Stationary time series (Y_t) has long range dependence if

$$\sum_{h=1}^{\infty} |\rho_X(h)| = \infty,$$

and short range dependence if the sum is bounded. E.g. ARMA processes have short range dependence since

$$|\rho(h)| \leq C \cdot r^h, \text{ for } h = 1, 2, \ldots$$

where $r \in (0,1)$.

A popular class of long memory processes is obtained when

$$\rho(h) \sim C \cdot h^{2d-1} \text{ as } h \to \infty,$$

where $d \in (0, 1/2)$. This can be obtained with fractionary processes

$$(1 - L)^d X_t = \varepsilon_t,$$

where (ε_t) is some white noise. Here, $(1-L)^d$ is defined as

$$(1-L)^d = 1 - dL - \frac{d(1-d)}{2!}L^2 - \frac{d(1-d)(2-d)}{3!}L^3 + \cdots = \sum_{j=0}^{\infty} \phi_j L^j,$$

where

$$\phi_j = \frac{\Gamma(j-d)}{\Gamma(j+1)\Gamma(d)} = \prod_{0<k\leq j}\left(\frac{k-1-d}{k}\right) \text{ for } j = 0, 1, 2, \ldots$$

If $Var(\varepsilon_t) = 1$, note that par

$$\gamma_X(h) = \frac{\Gamma(1-2d)\Gamma(h+d)}{\Gamma(d)\Gamma(1-d)\Gamma(h+1-d)} \sim \frac{\Gamma(1-2d)}{\Gamma(d)\Gamma(1-d)} \cdot h^{2d-1}$$

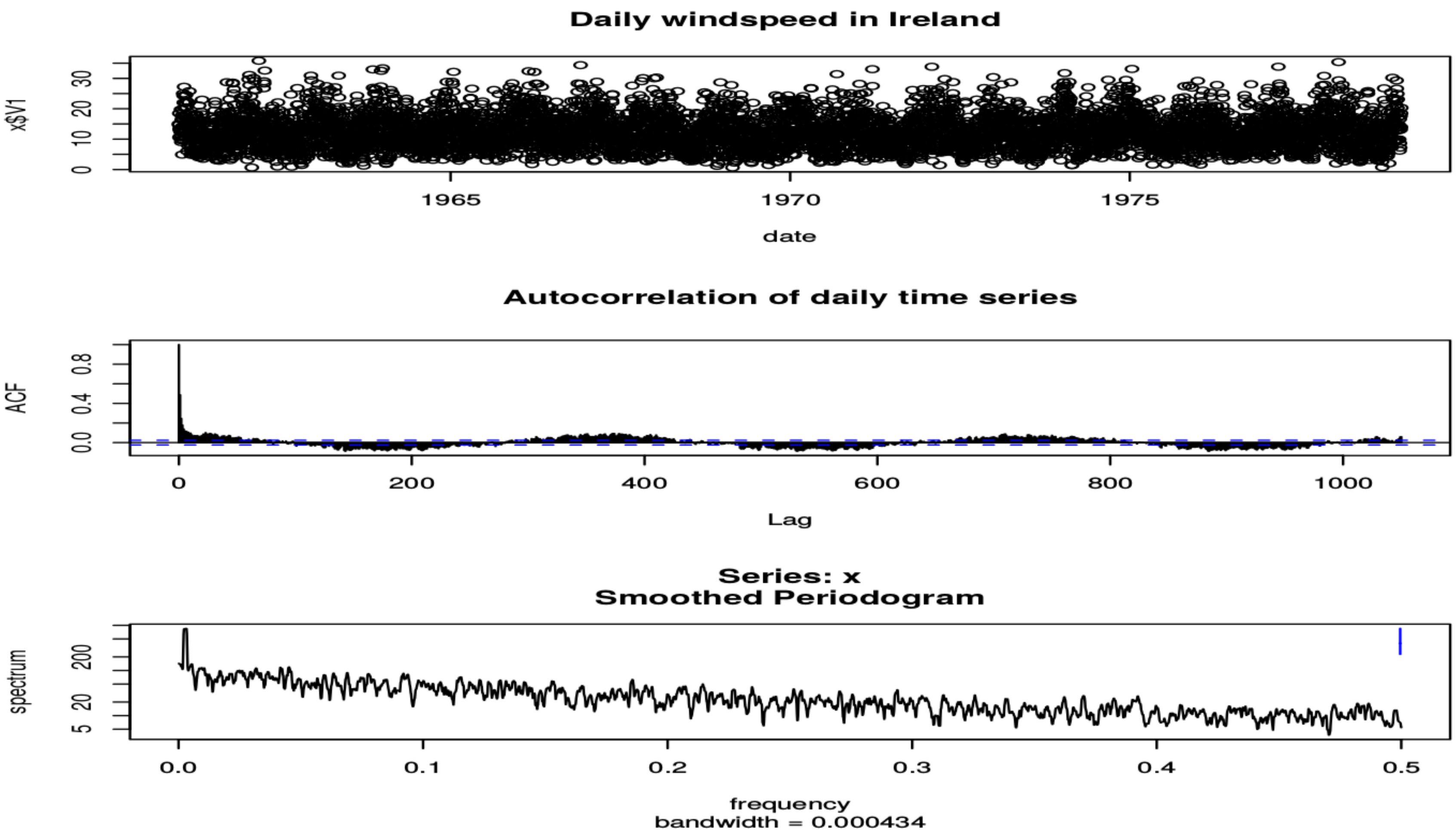
as $h \to \infty$, and

$$f_X(\omega) = \left(2\sin\frac{\omega}{2}\right)^{-2d} \sim \omega^{-2d}$$

as $\omega \to 0$.

See also Mandelbrot et Van Ness (1968) for the continuous time version, with the fractionary Brownian motion.

Daily Windspeed Time Series



Defining Long Range Dependence

Hosking (1981, 1984) suggested another definition of long range dependence: (X_t) is stationnary, and there is ω_0 such that $f_X(\omega) \to \infty$ as $\omega \to \omega_0$.

Such a ω_0 can be related to seasonality

Gray, Zhang & Woodward (1989) defined $GARMA(p, d, q)$ processes, inspired by Hosking (1981)

$$\Phi(L)(1 - 2uL + L^2)^d X_t = \Theta(L)\varepsilon_t$$

Hosking (1981) did not studied those processes since it is difficult to invert $(1 - 2uL + L^2)^d$.

Defining Long Range Dependence with Seasonality

This can be done using Gegenbauer polynomial: for $d \neq 0$, $|Z| < 1$ and $|u| \leq 1$,

$$(1 - 2uL + L^2)^{-d} = \sum_{i=0}^{\infty} P_{i,d}(u) L^n,$$

where

$$P_{i,d}(u) = \sum_{k=0}^{[i/2]} (-1)^k \frac{\Gamma(d+n-k)}{\Gamma(d)} \frac{(2u)^{n-2k}}{[k!(n-2k)!]}$$

If $|u| < 1$, the limit of $(\omega - \omega_0)^{2d} f(\omega)$ exists when $\omega \rightarrow \omega_0$, where $\omega_0 = \cos^{-1}(u)$.

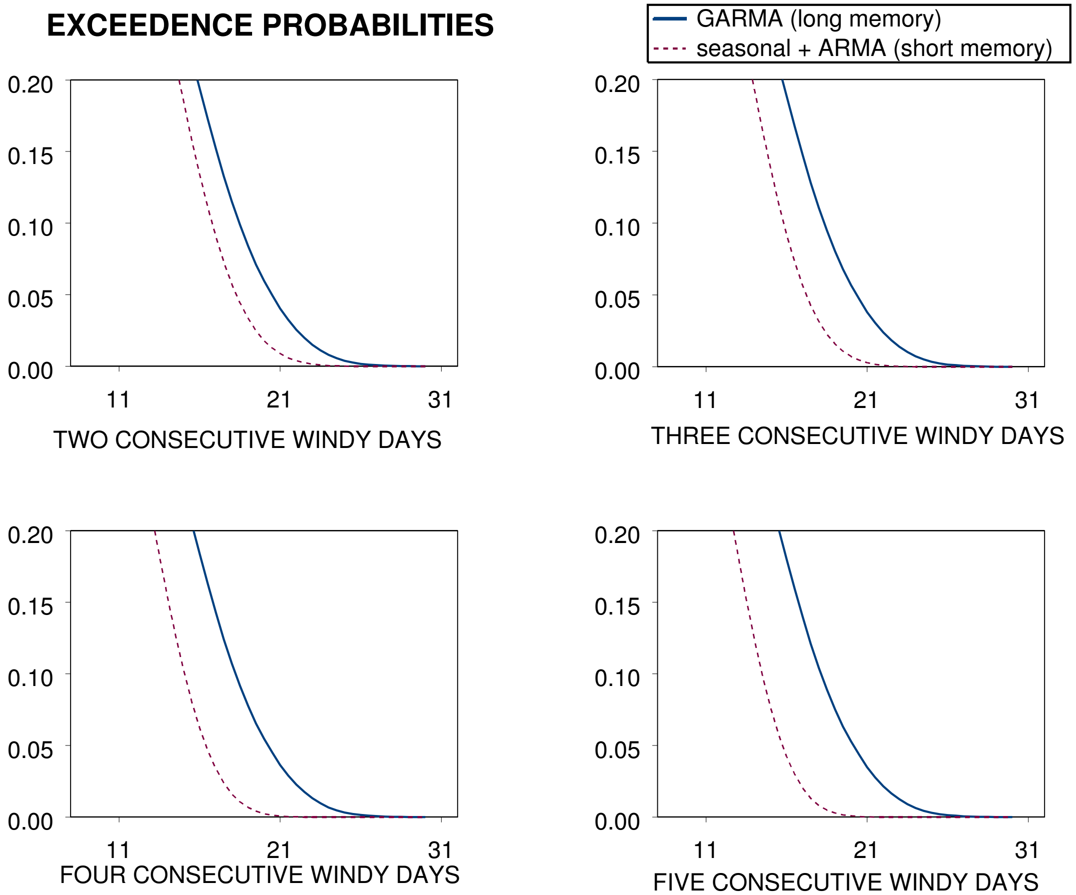
Further, if $|u| < 1$ and $0 < d < 1/2$, then

$$\rho(h) \sim C \cdot h^{2d-1} \cdot \cos(\omega_0 \cdot h) \text{ as } h \rightarrow \infty.$$

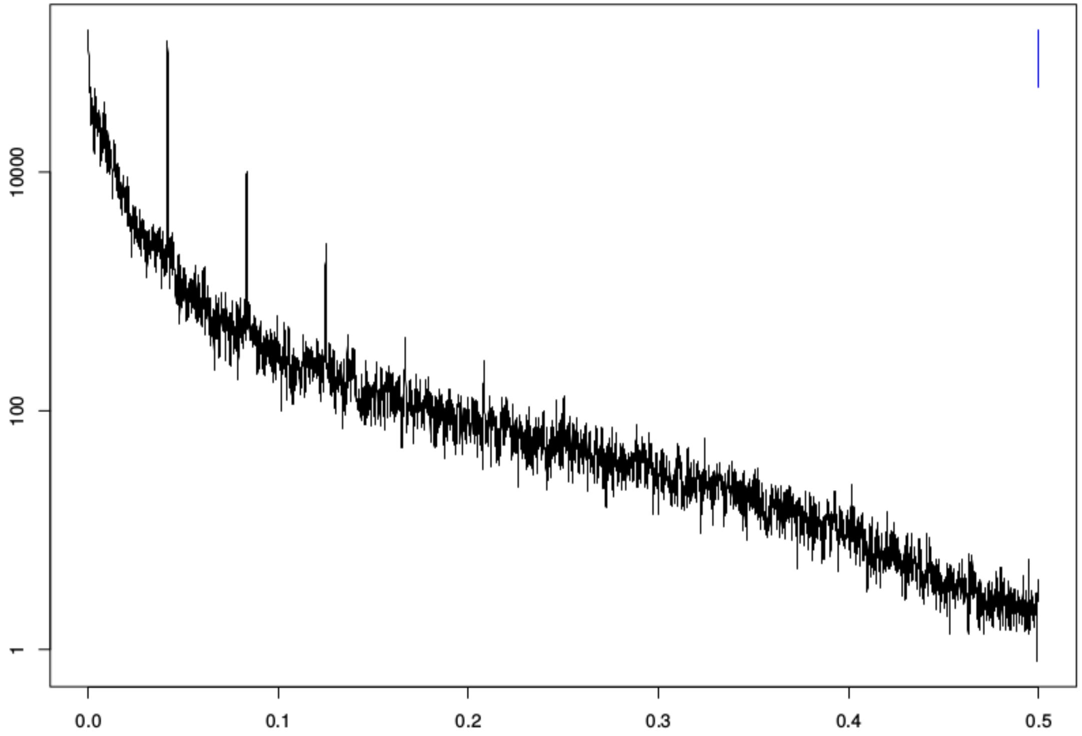
In Bouëtte *et al.* (2003) *Stochastic Environmental Research & Risk Assesment* we obtained on daily windspeed $\widehat{d} \sim 0,18$.

Estimation 'Return Periods'

Using Gray, Zhang & Woodward (1989), it is possible to simulate $GARMA$ processes, to estimate probabilities



Spectral Density of Hourly Wind Speed in the Netherlands



Some k factor GARMA should be considered, see (Bouëtte et al. (2003))

The European heatwave of 2003

Third IPCC Assessment, 2001: treatment of extremes (e.g. trends in extreme high temperature) is "*clearly inadequate*". Karl & Trenberth (2003) noticed that "*the likely outcome is more frequent heat waves*", "*more intense and longer lasting*" added Meehl & Tebaldi (2004).

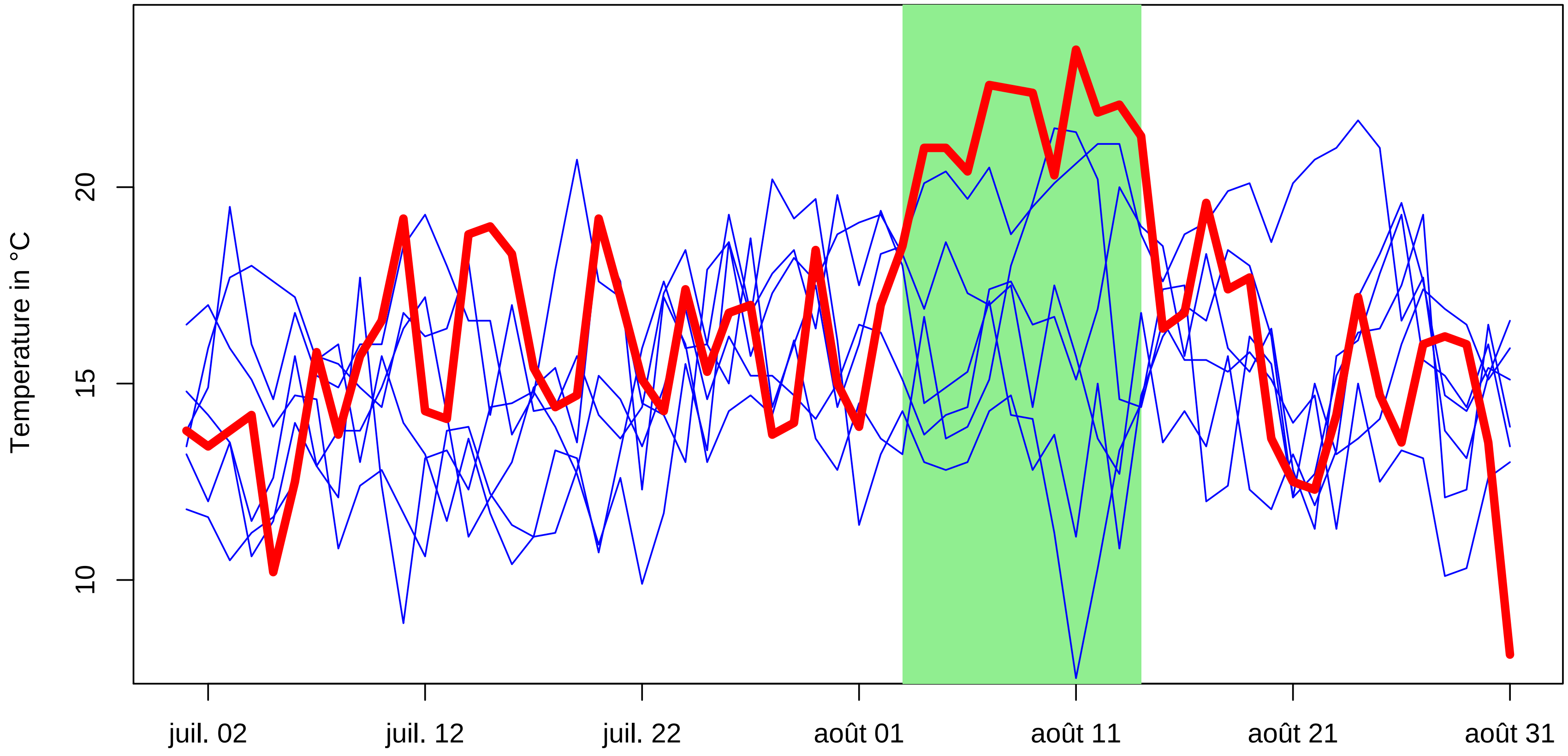
In Nîmes, there were more than 30 days with temperatures higher than 35° C (versus 4 in hot summers, and 12 in the previous heat wave, in 1947).

Similarly, the average maximum (minimum) temperature in Paris peaked over 35° C for 10 consecutive days, on 4-13 August. Previous records were 4 days in 1998 (8 to 11 of August), and 5 days in 1911 (8 to 12 of August).

Similar conditions were found in London, where maximum temperatures peaked above 30°C during the period 4-13 August

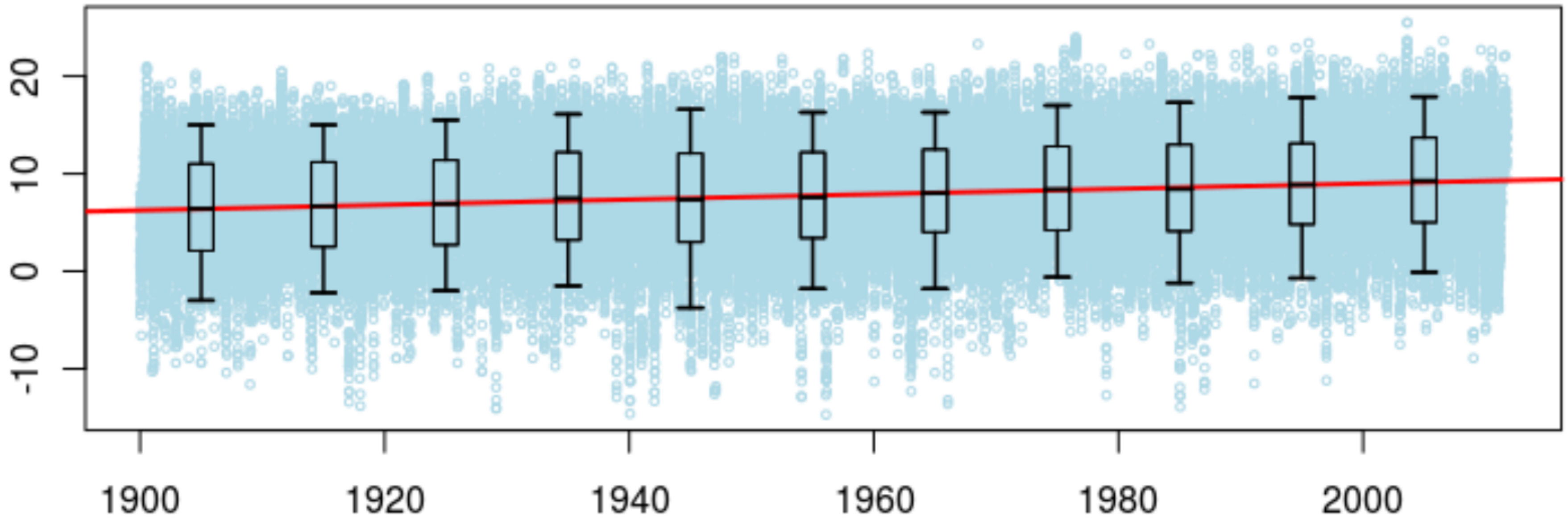
(see e.g. Burt (2004), Burt & Eden (2004) and Fink *et al.* (2004).)

Minimum Daily Temperature in Paris, France



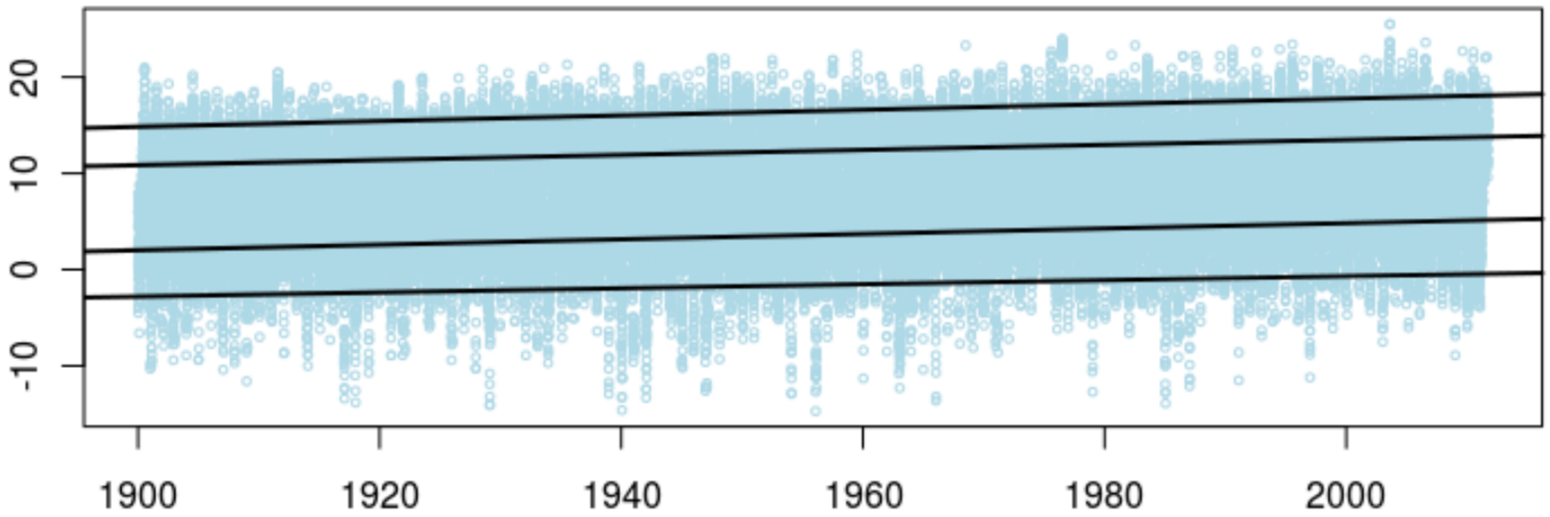
Modelling the Minimum Daily Temperature

Karl & Knight (1997) , modeling of the 1995 heatwave in Chicago: minimum temperature should be most important for health impact (see also Kovats & Koppe (2005)), several nights with no relief from very warm nighttime



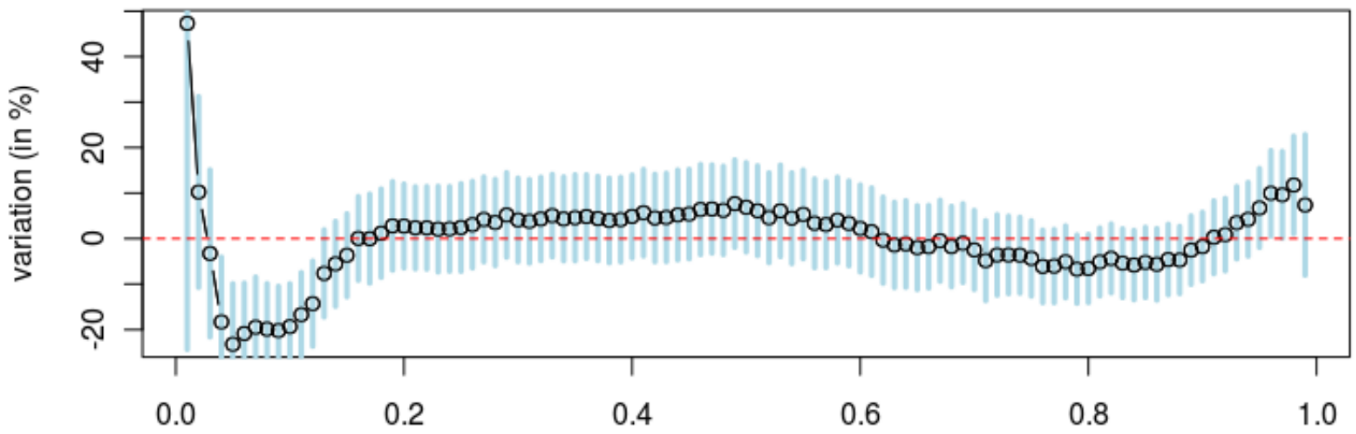
Modelling the Minimum Daily Temperature

Instead of boxplots, consider some quantile regression

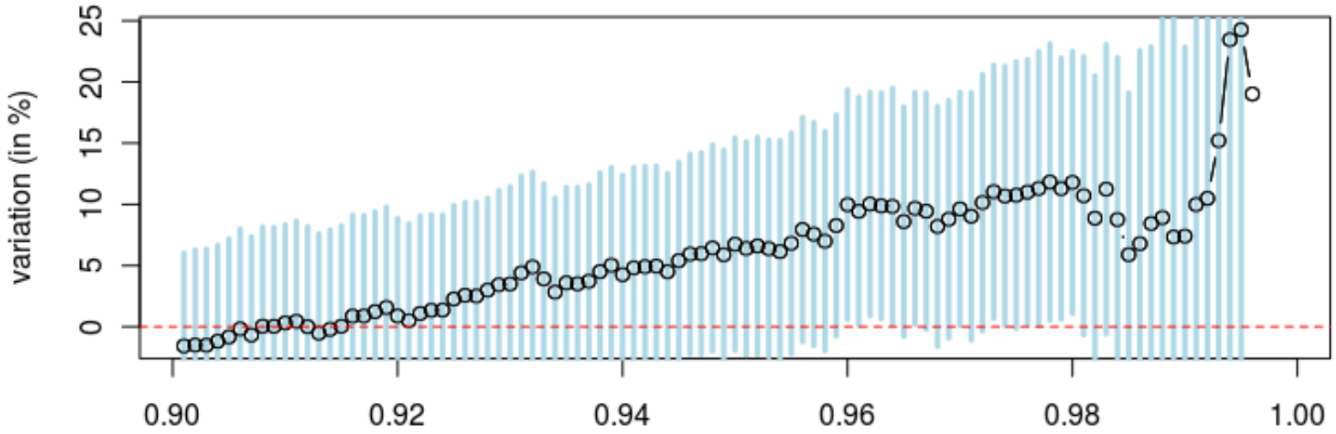


Modelling the Minimum Daily Temperature

Note that the slope for various probability levels is rather stable



unless we focus on heat-waves,



Which *temperature* might be interesting ?

Consider the following decomposition

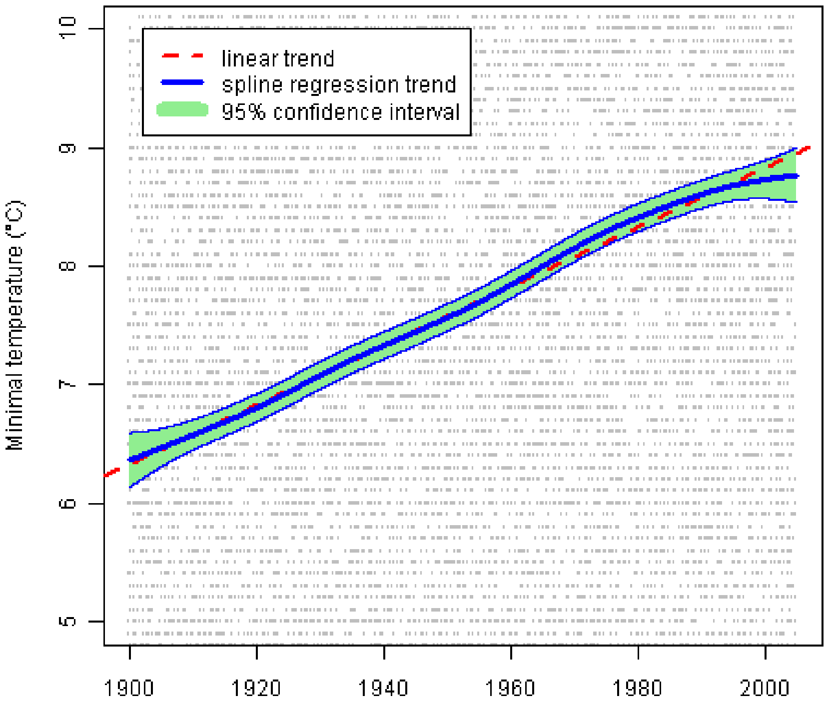
$$Y_t = \mu_t + X_t$$

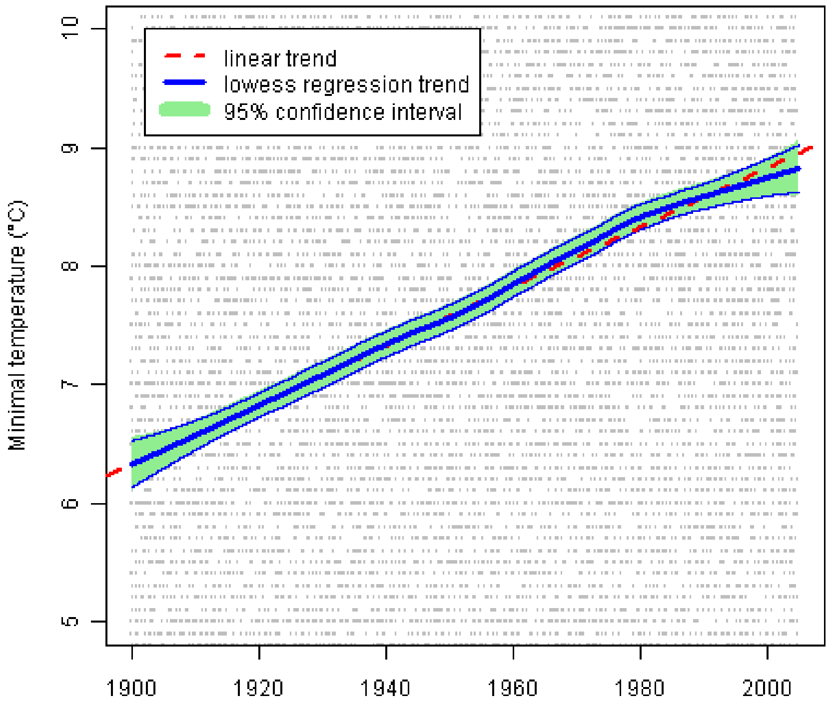
where

- μ_t is a (linear) general tendency
- X_t is the remaining (stationary) noise

Nonstationarity and *linear* trend

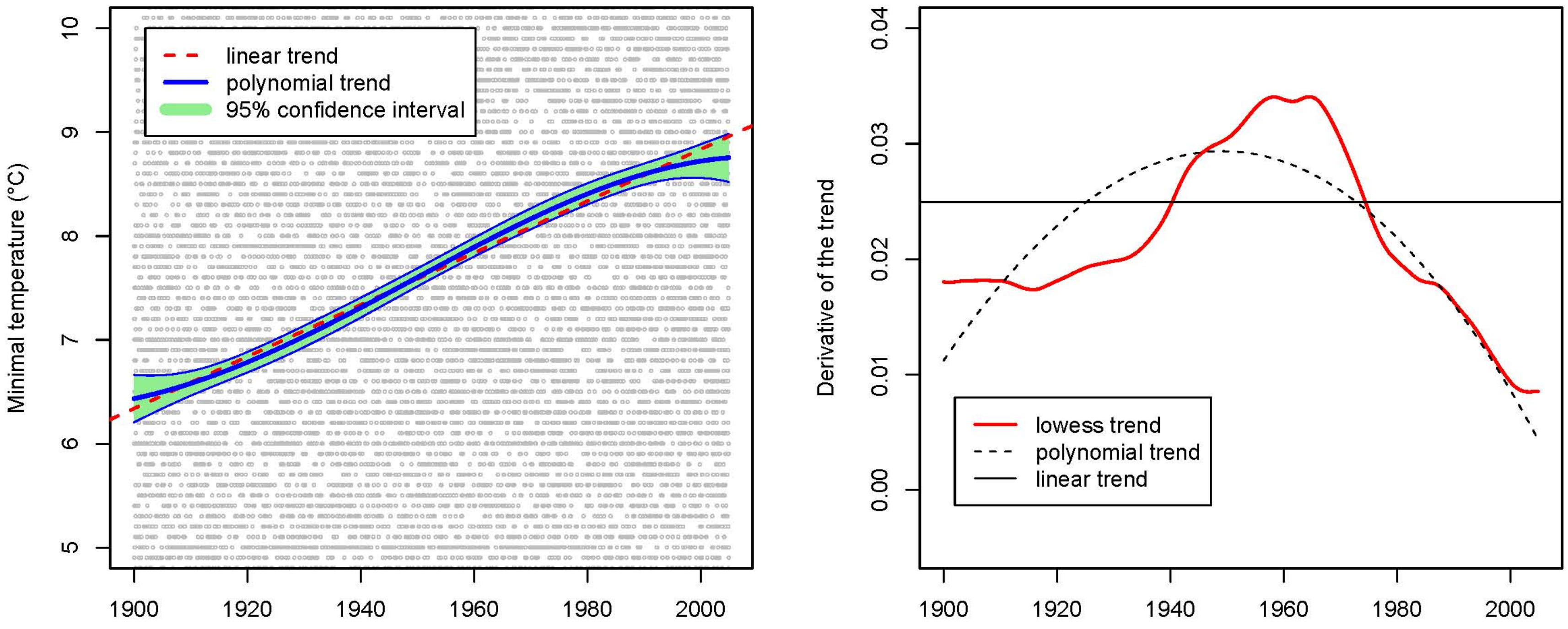
Consider a spline and lowess regression





Nonstationarity and *linear* trend

or a polynomial regression,and compare local slopes,



Linear trend, and Gaussian noise

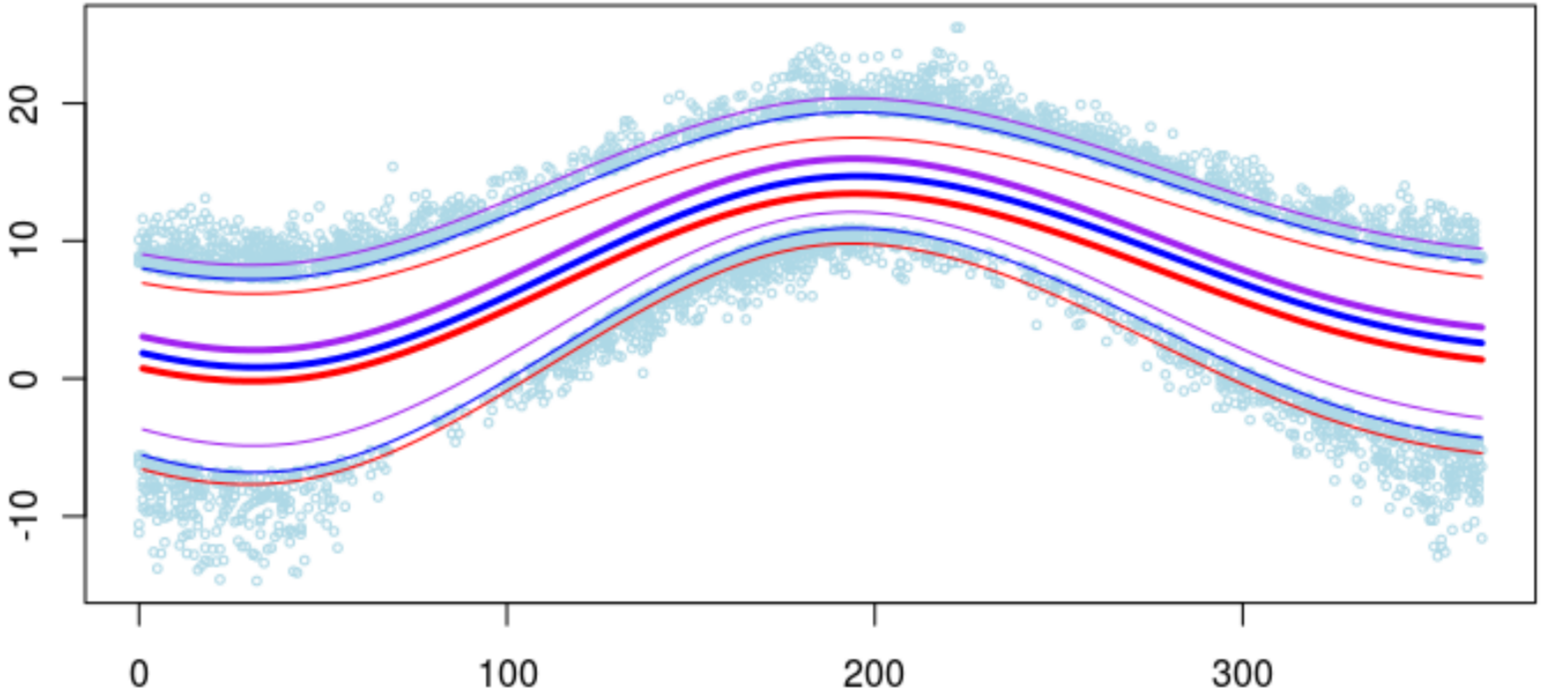
Benestad (2003) or Redner & Petersen (2006) suggested that temperature for a given (calendar) day is an "*independent Gaussian random variable with constant standard deviation* σ *and a mean that increases at constant speed* ν"

In the U.S., $\nu = 0.03°$ C per year, and $\sigma = 3.5°$ C

In Paris, $\nu = 0.027°$ C per year, and $\sigma = 3.23°$ C

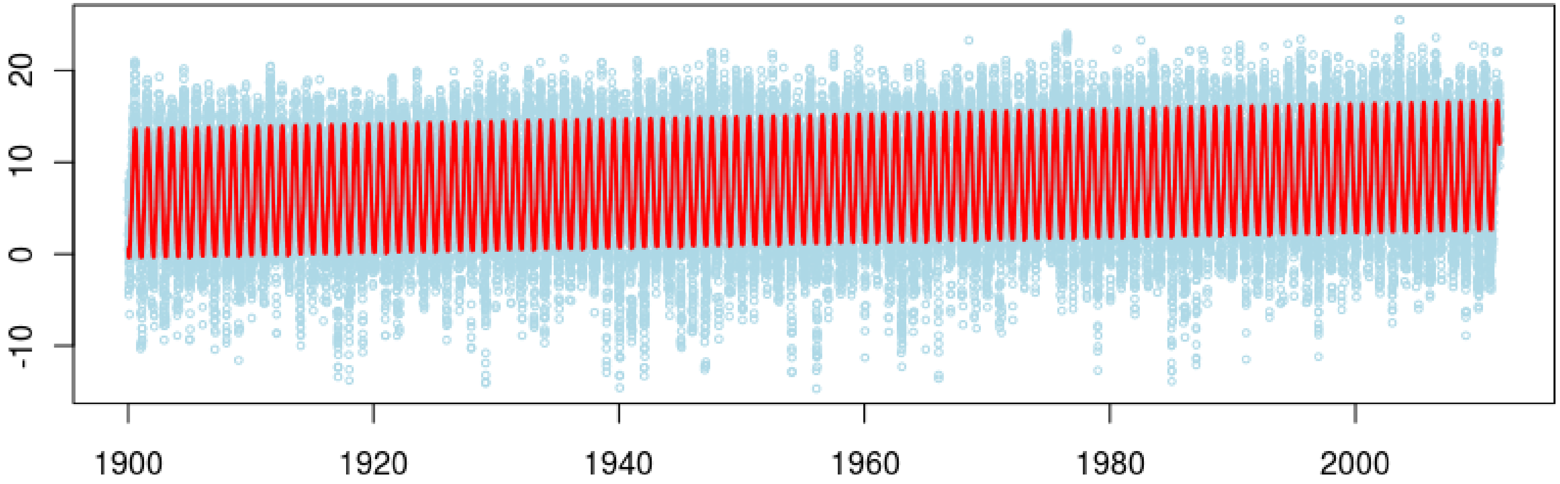
The Seasonal Component

There is a seasonal pattern in the daily temperature



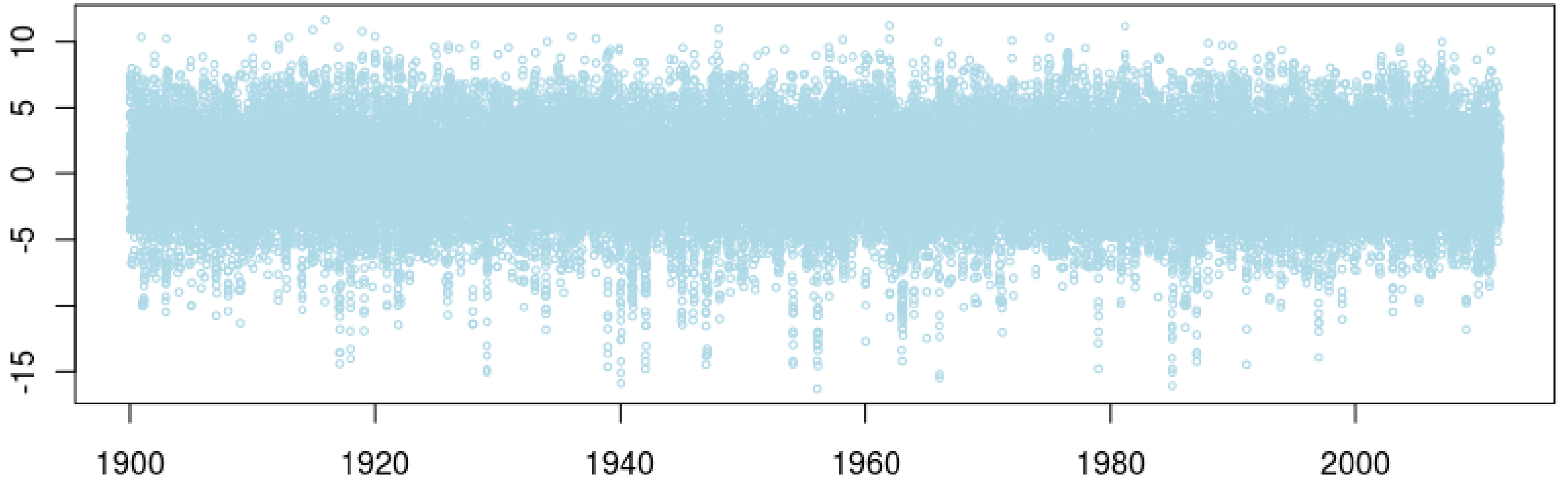
The Residual Part (or stationary component)

Let $\widehat{X}_t = Y_t - \left(\widehat{\beta}_0 + \widehat{\beta}_1 t + \widehat{S}_t\right)$



The Residual Part (or stationary component)

$\widehat{X}_t$ might look stationary,



One can consider some short-range dependence (ARMA) model, with either light or heavy tailed innovation process.

Long range dependence ?

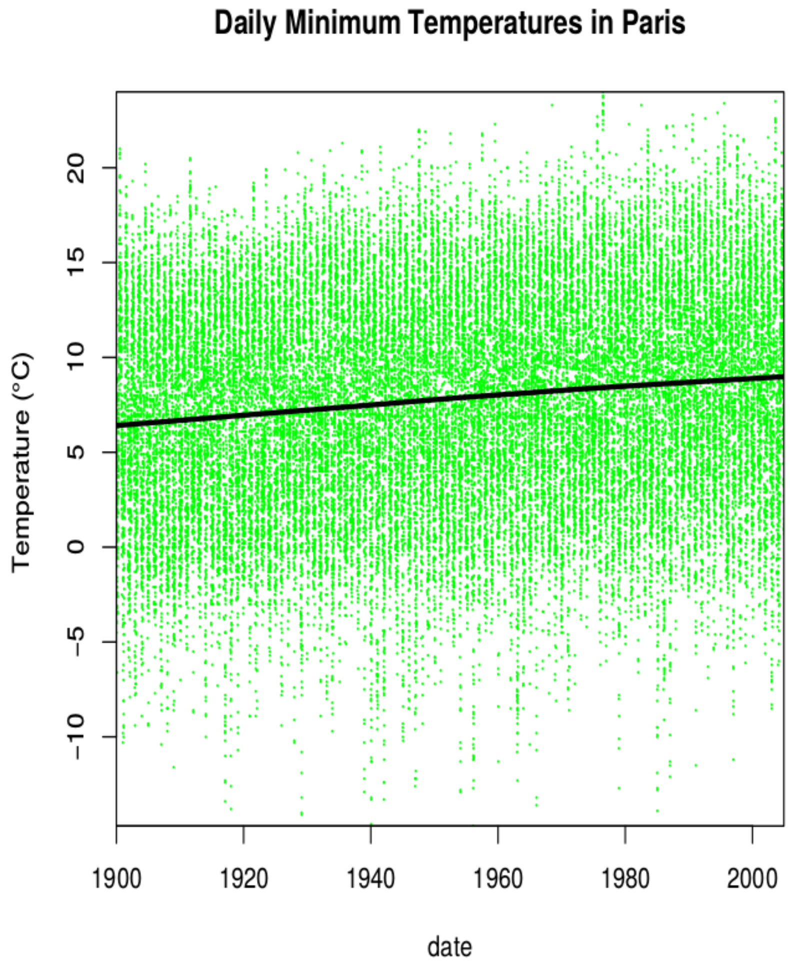
Smith (1993) "*we do not believe that the autoregressive model provides an acceptable method for assessing theses uncertainties*" (on temperature series)

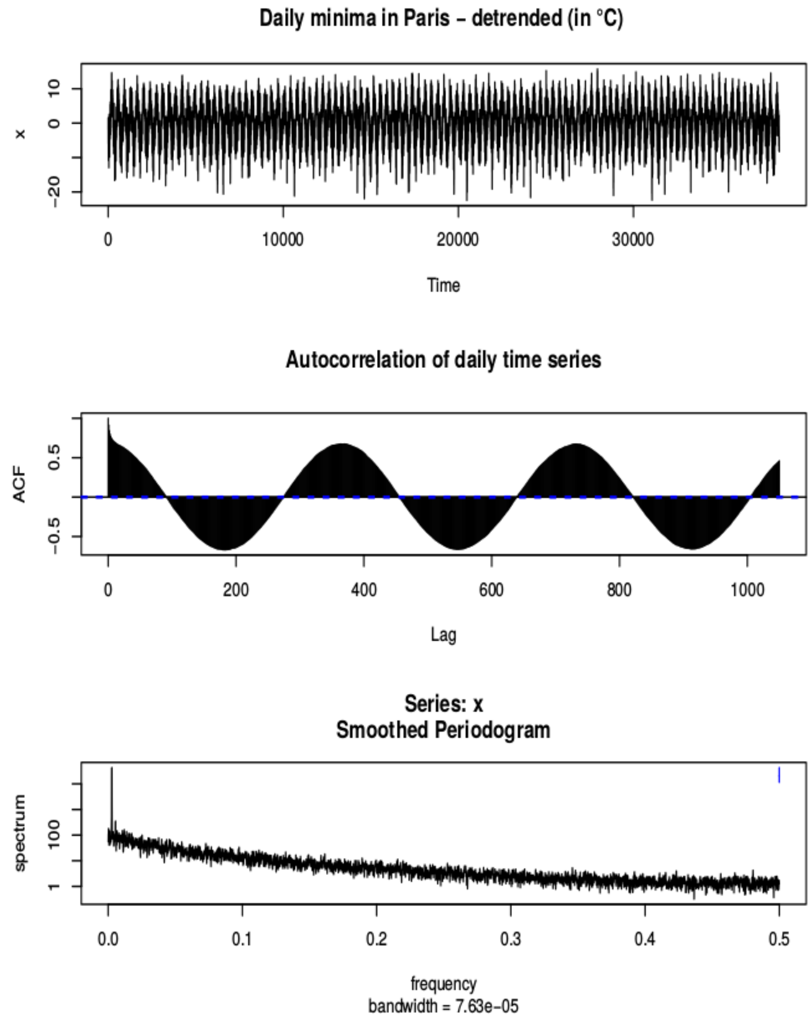
Dempster & Liu (1995) suggested that, on a long period, the average annual temperature should be decomposed as follows

- an increasing linear trend,
- a random component, with long range dependence.

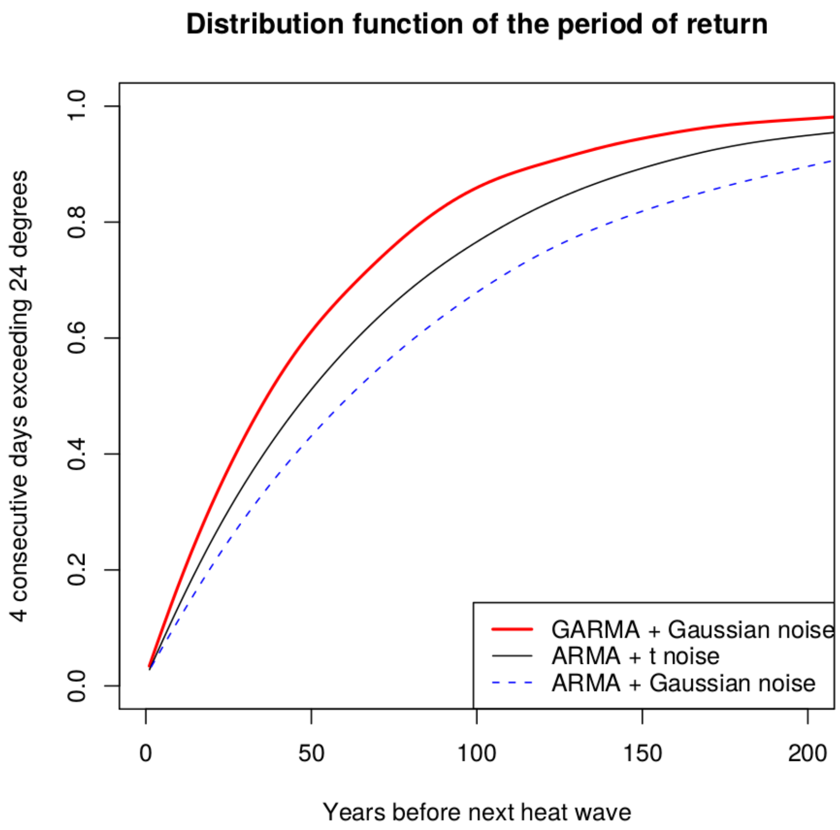
Consider GARMA time serie models, as in Charpentier (2011), *Climatic Change*.

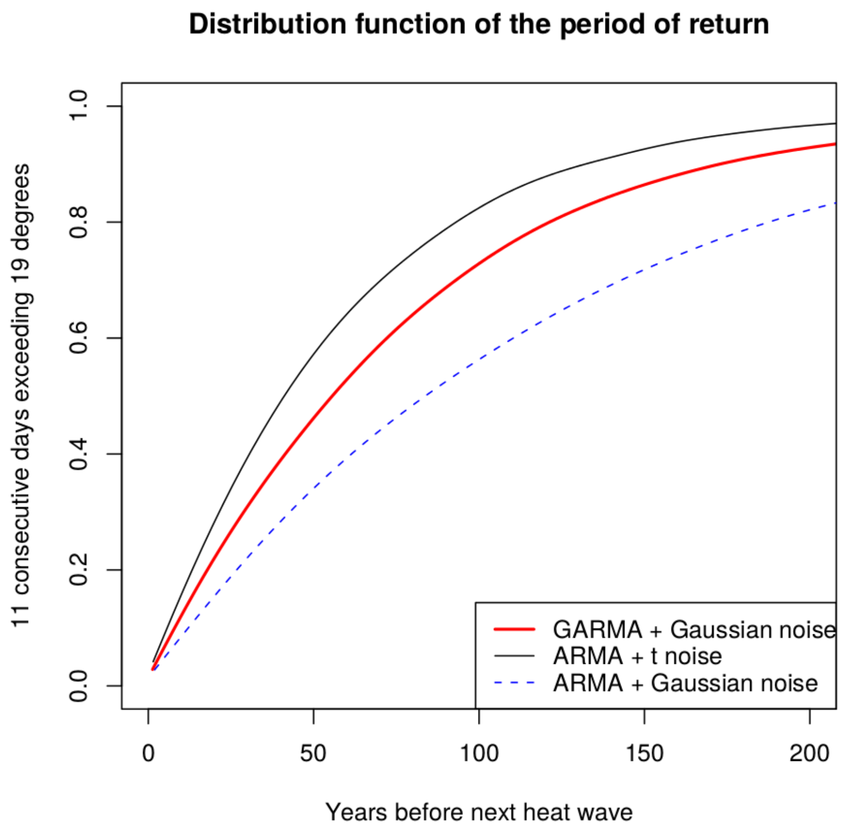
Long range dependence ?



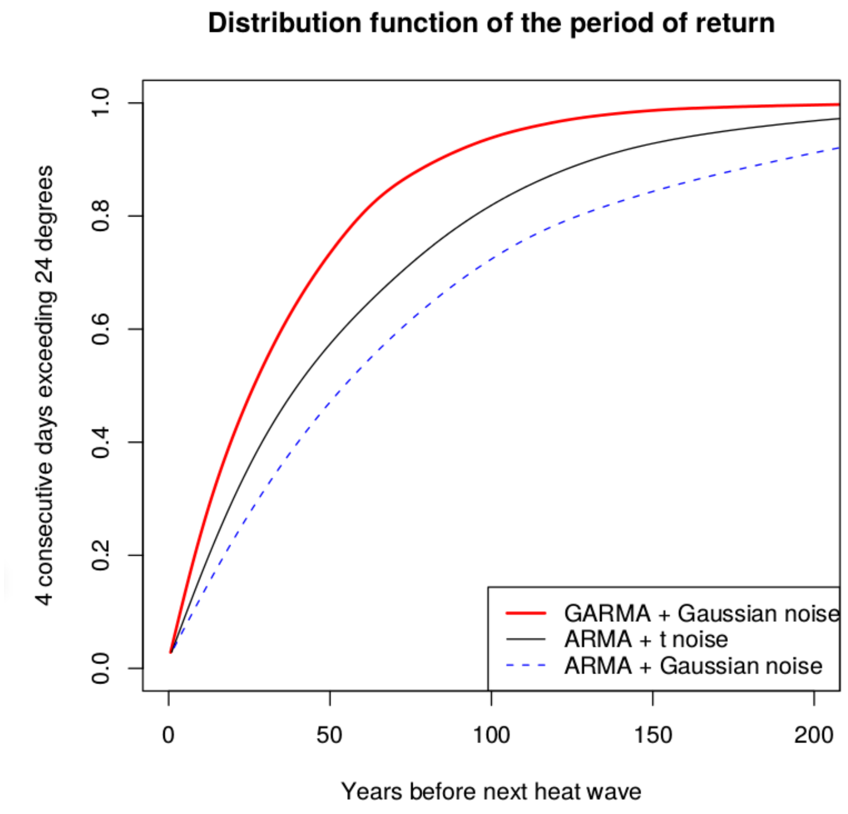


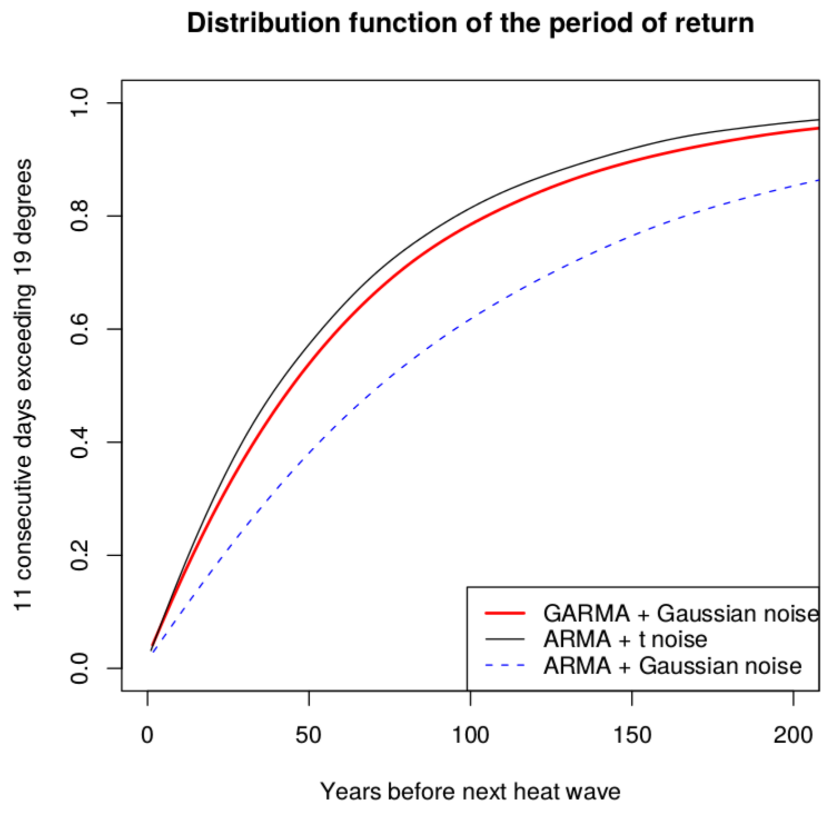
On return periods, optimistic scenario





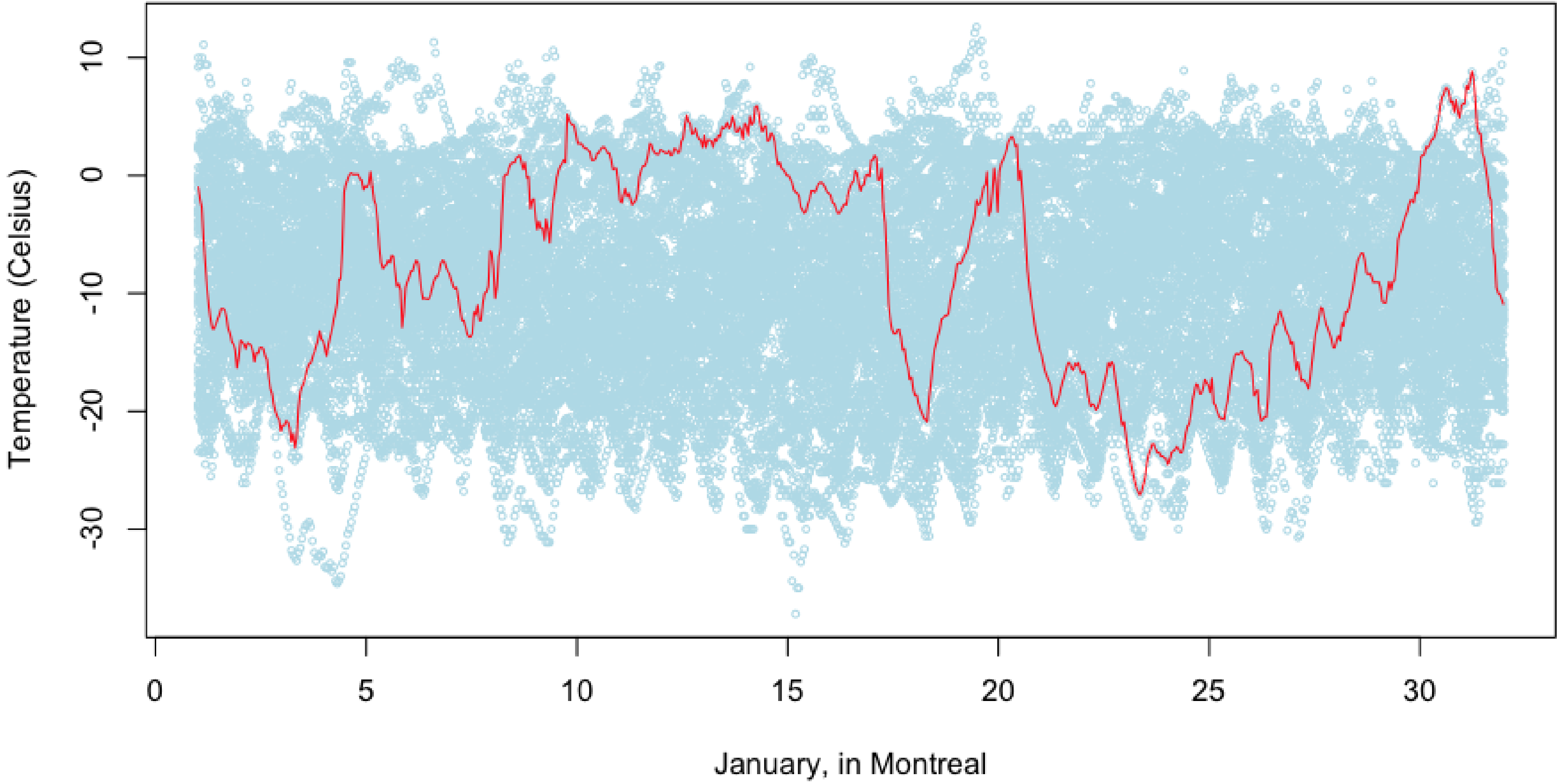
On return periods, pessimistic scenario





Long Memory, non Stationarity and Temporal Granularity

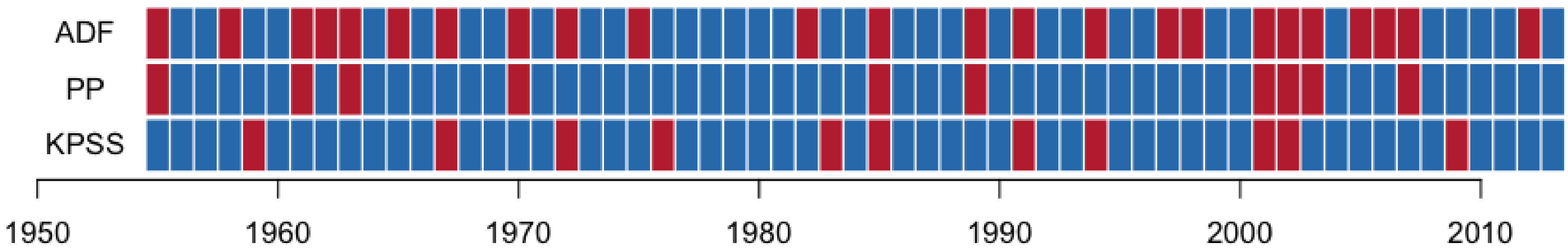
Hourly Temperature in Montreal, QC, in January,



Hourly Temperature as a Random Walk?

Use of various test to test for integrated time series (random walk)

- ADF, Augmented Dickey-Fuller, see Fuller (1976) and Said & Dickey (1984)
- KPSS, Kwiatkowski–Phillips–Schmidt–Shin, see Kwiatkowski *et al.* (1992)
- PP, Phillips–Perron, see Phillips & Perron (1988)



where ▌ random-walk vs. ▌ stationnary

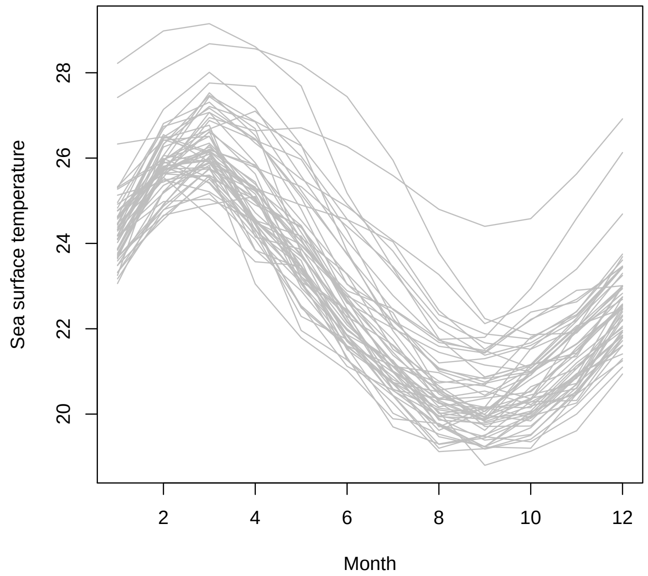
March in Montréal: Which Winter Was 'Abnormal'

Detecting Abnormalities and Outliers

Consider the case where $X_{i,t}$ denote the temperature at date/time t, for year i.

Let $\varphi_{1,t}, \varphi_{2,t}, \varphi_{3,t}, \cdots$ denote the principal components, and $Y_{i,1}, Y_{i,2}, Y_{i,3}, \cdots$ the principal component scores.

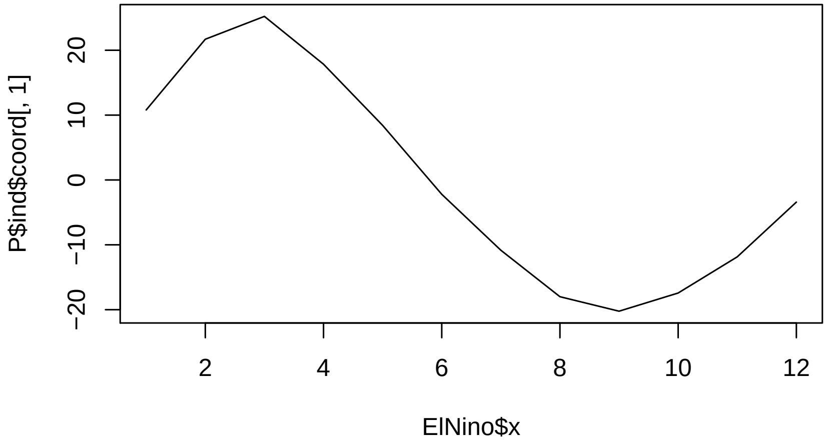
To detect outliers, see Jones & Rice (1992), Sood *et al.* (2009) or Hyndman & Shang (2010) use a bivariate depth plot on $\{(Y_{1,i}, Y_{2,i}), i = 1, \cdots, n\}$.

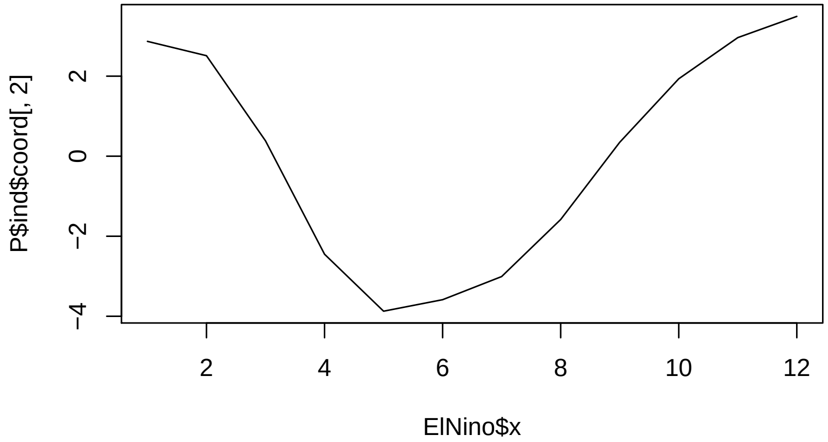


E.g. monthly sea surface temperatures, from January 1950 to December 2006

Detecting Abnormalities and Outliers

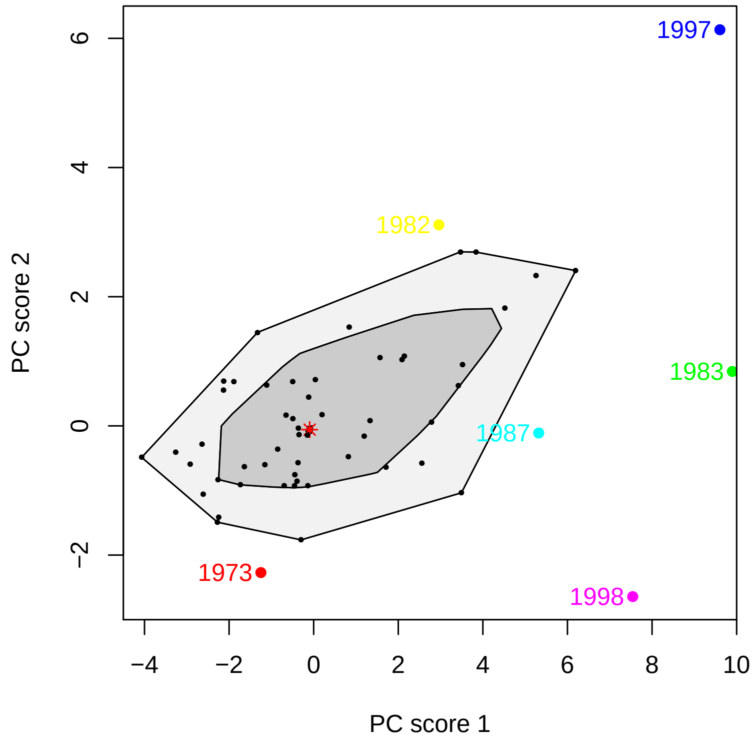
The first two components are

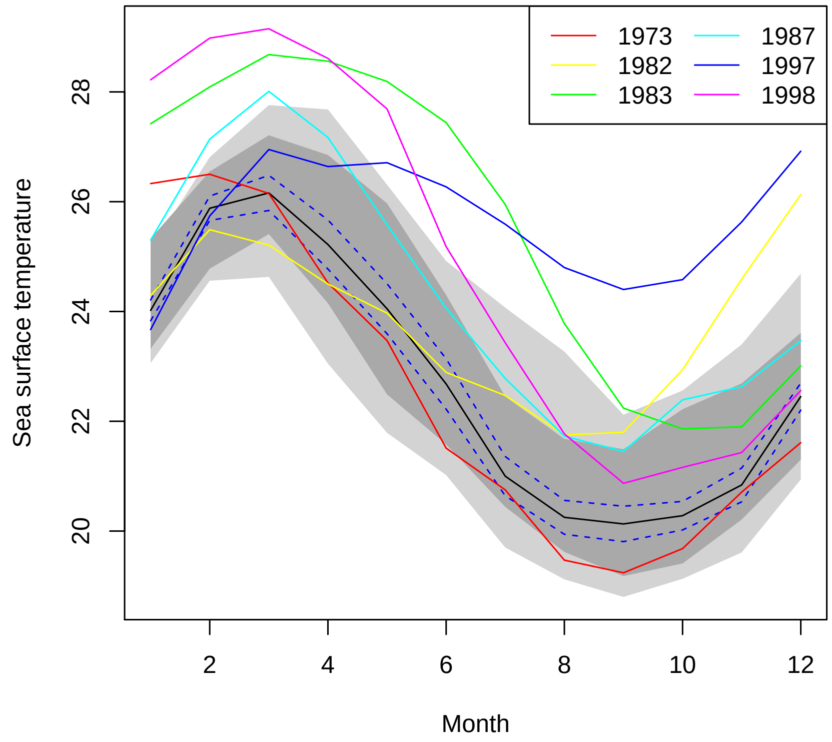




And we can use a depth plot on the first two principal component scores.

Detecting Abnormalities and Outliers





Depth Set and Bag Plot

Here we use Tukey's depth set concept. In dimension 1, define

$$\text{depth}(y) = \min\{F(y), 1 - F(y)\}$$

and the associated depth set of level $\alpha \in (0,1)$ as

$$D_\alpha = \{y \in \mathbb{R} : \text{depth}(y) \geq 1 - \alpha\}$$

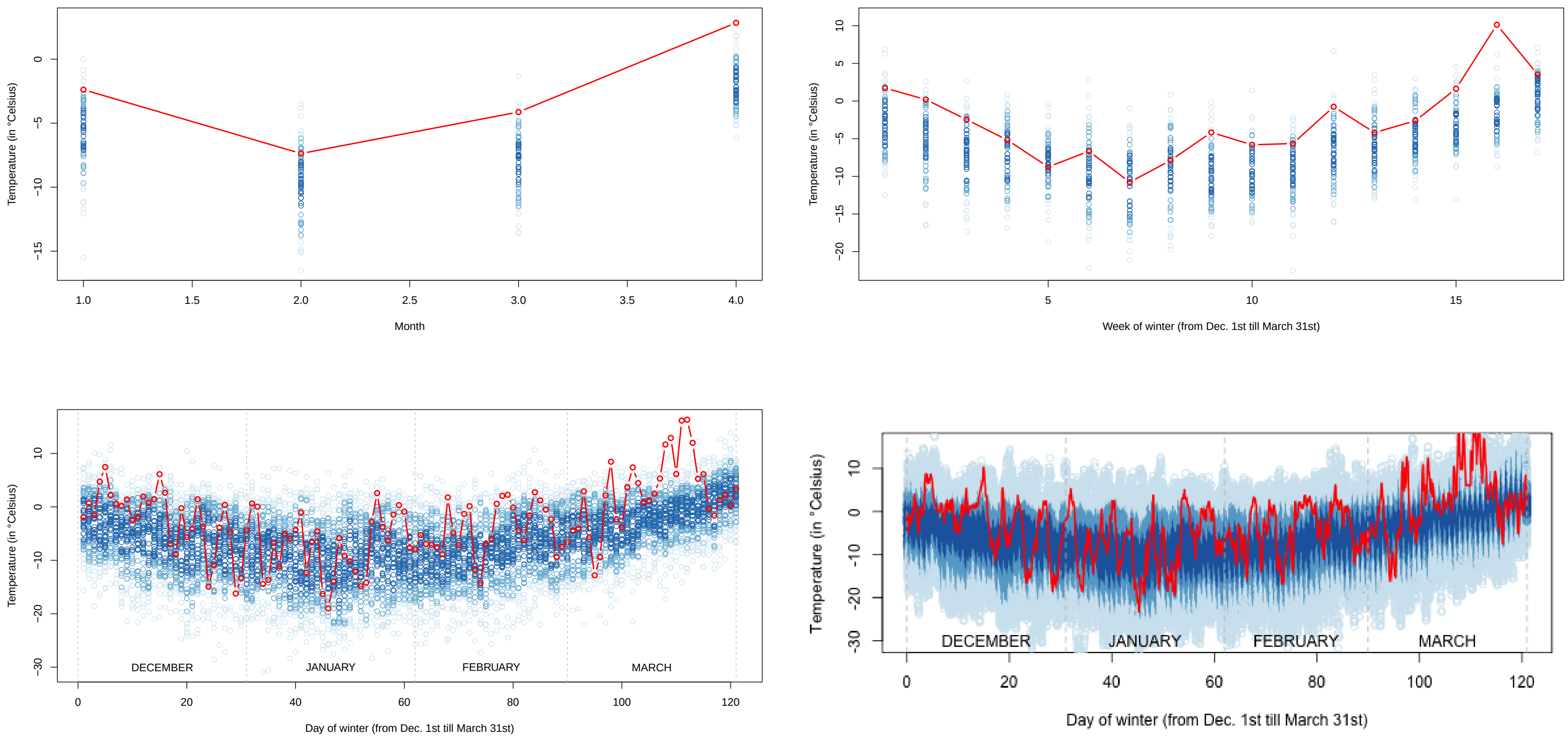
In higher dimension,

$$\text{depth}(\boldsymbol{y}) = \inf_{\boldsymbol{u}:\boldsymbol{u}\neq\boldsymbol{0}} \{\mathbb{P}[\mathcal{H}_{\boldsymbol{u}}(\boldsymbol{y})]\}$$

where $\mathcal{H}_{\boldsymbol{u}}(\boldsymbol{y}) = \{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{u}^\mathsf{T}\boldsymbol{x} \leq \boldsymbol{u}^\mathsf{T}\boldsymbol{y}\}$ and the associated depth set of level $\alpha \in (0,1)$ as

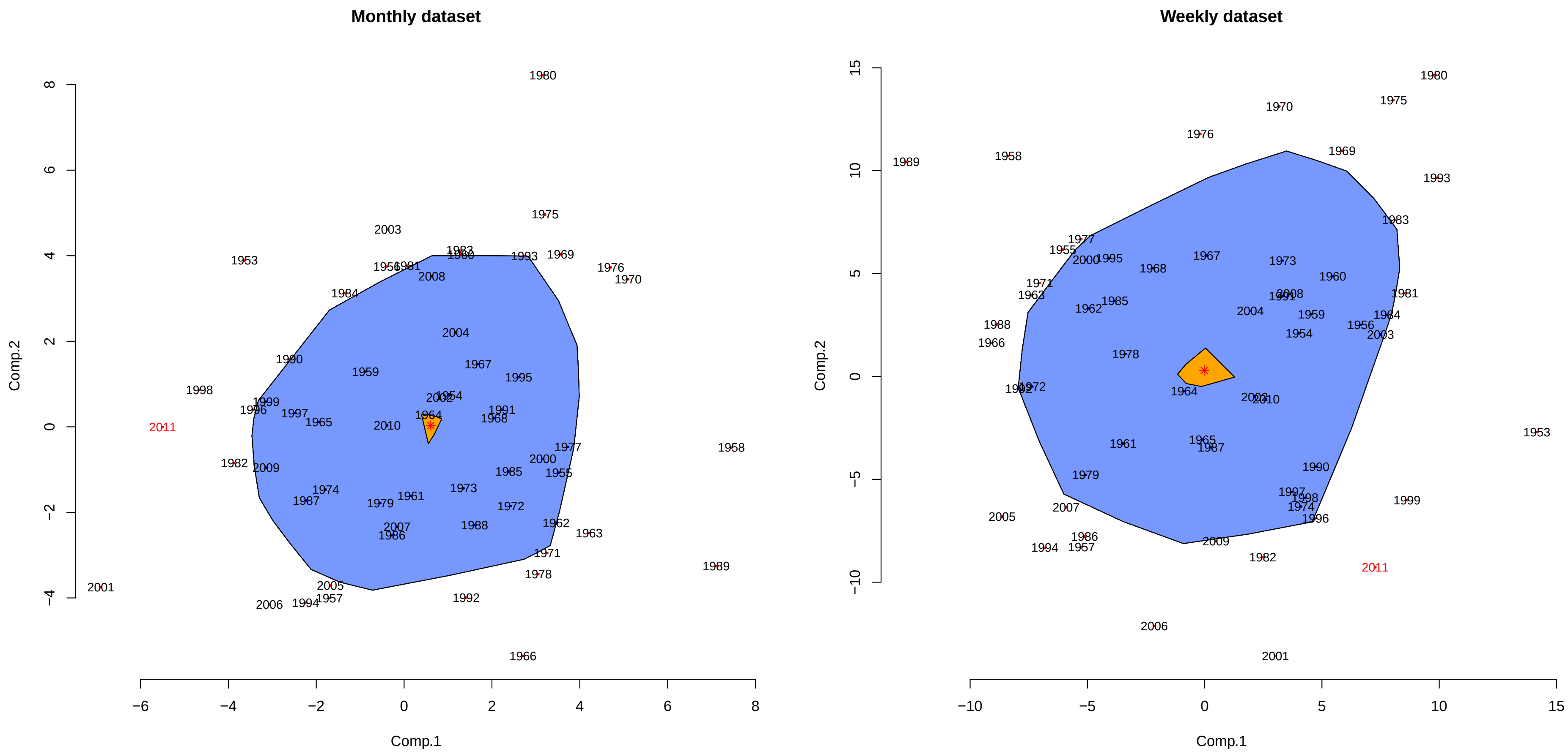
$$D_\alpha = \{\boldsymbol{y} \in \mathbb{R}^d : \text{depth}(\boldsymbol{y}) \geq 1 - \alpha\}$$

Winter Temperature in Montreal

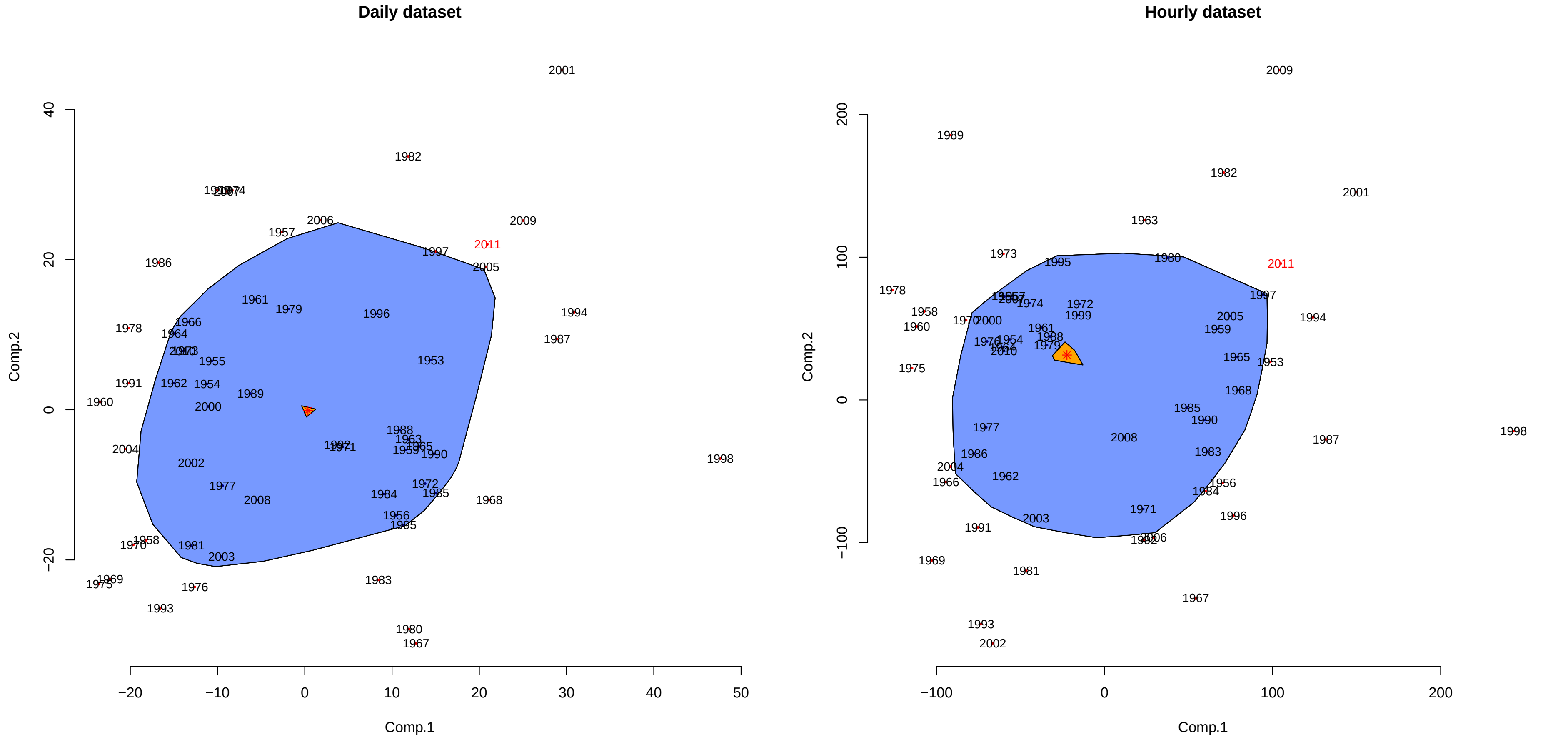


Winter temperature in Montréal, from December 1st till March 31st, with Monthly, Weekly, Daily and Hourly temperatures. Winter 2011 is in red.

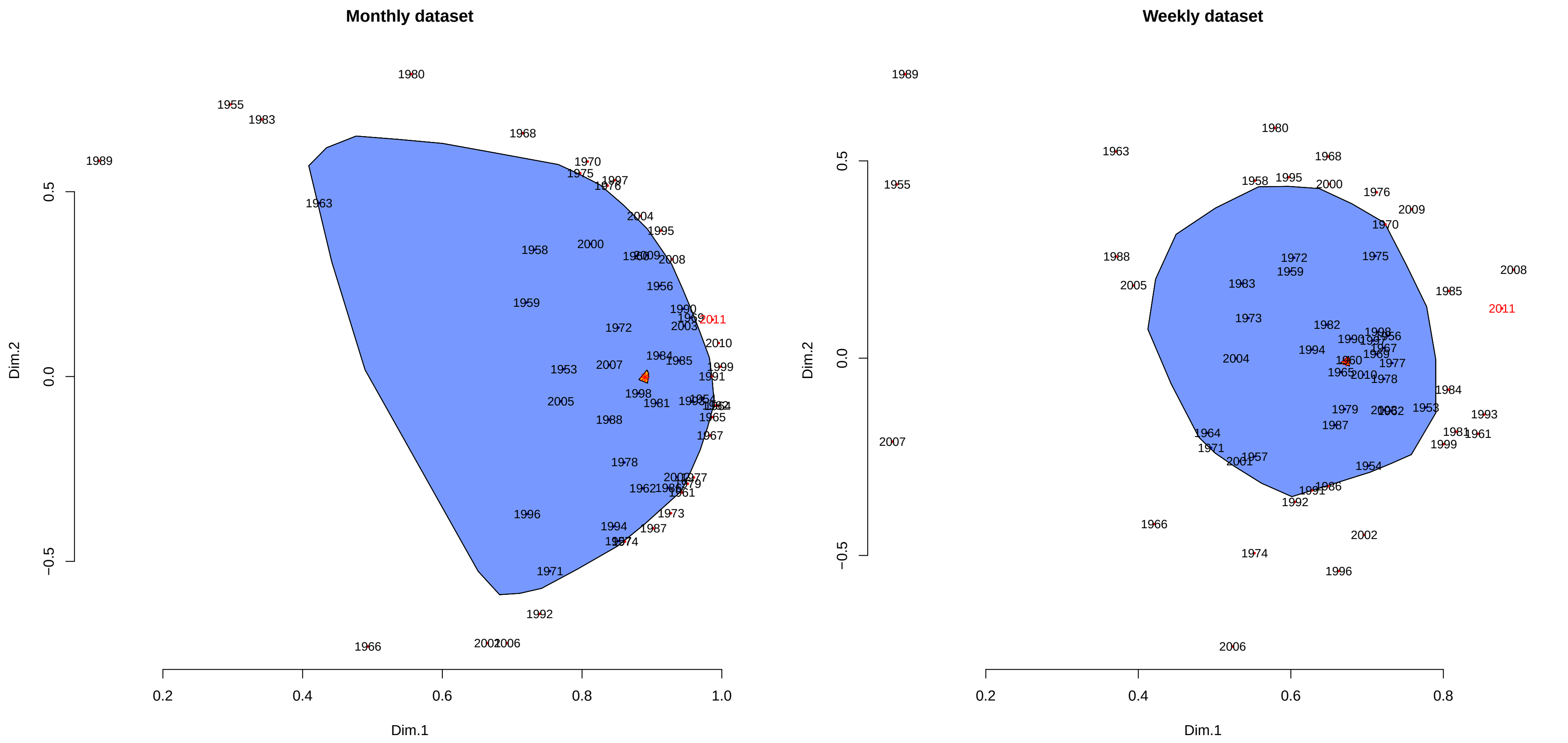
Robust ℓ_1 PCA Scores



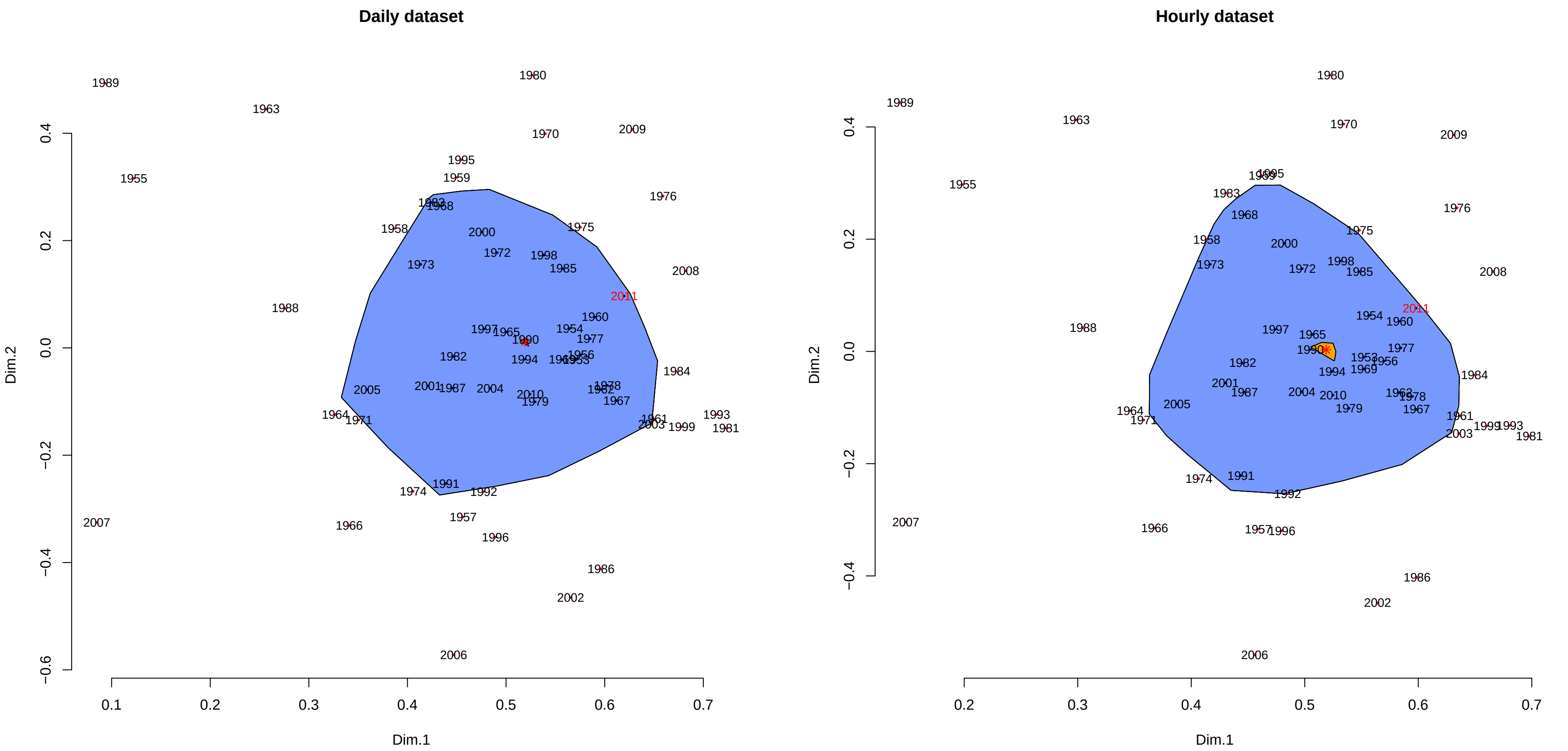
Robust ℓ_1 PCA Scores



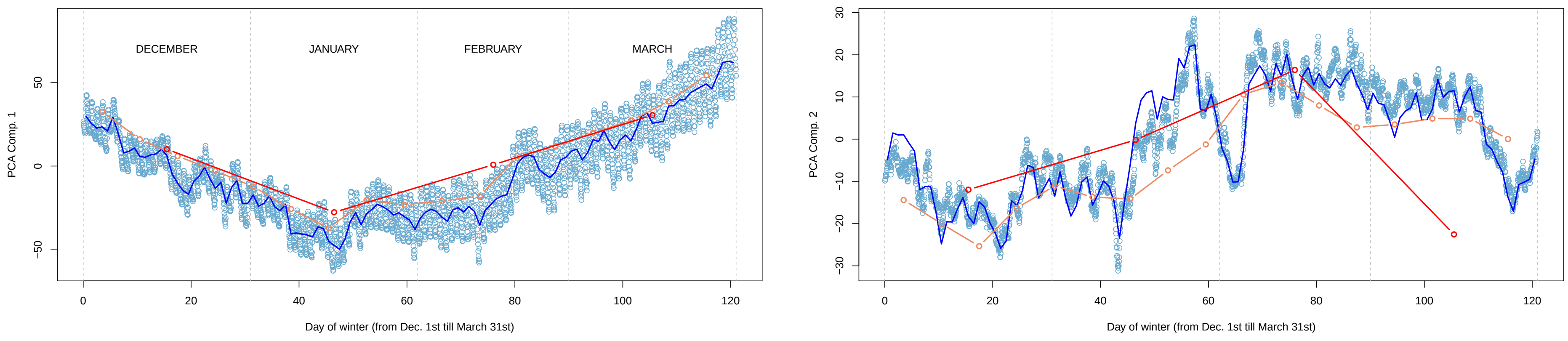
Standard ℓ_2 PCA Scores



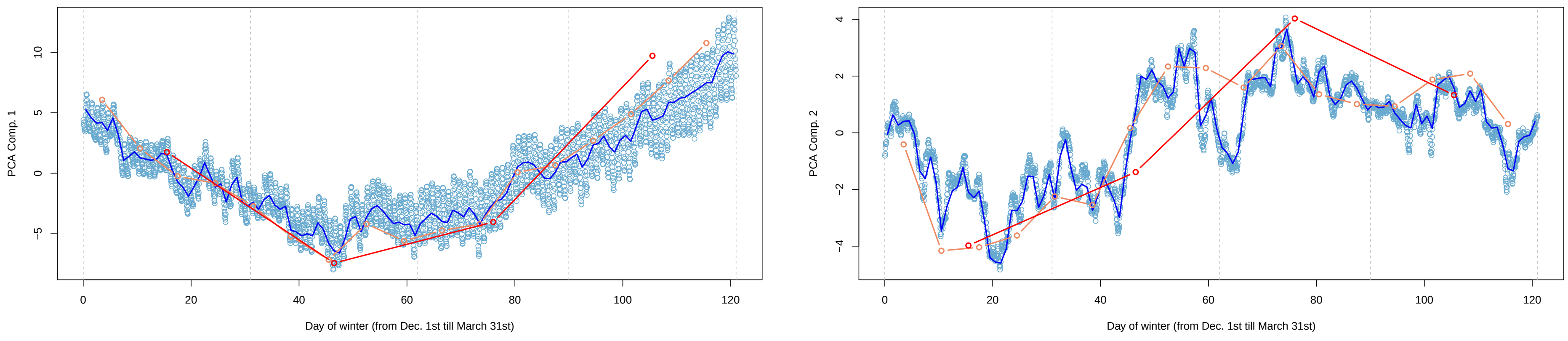
Standard ℓ_2 PCA Scores



Robust ℓ_1 PCA Principal Components



Standard ℓ_2 PCA Principal Components



Take-Home Message

When dealing with time series, having ‘big data’ with a more detailed granularity (higher frequency) looks nice (T is larger, higher accuracy) but usually leads to more complex models...

Still seems difficult to reconcile...

charpentier.arthur@uqam.ca

or @freakonometrics