Extremes for data with phase shifting seasons – some observations from marine climate

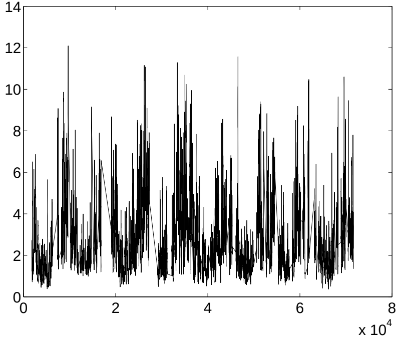
Georg Lindgren[1]
with input from Helena Olsson

[1] Mathematical Statistics, Lund University

Extreme events in climate and weather, Banff, August, 2010

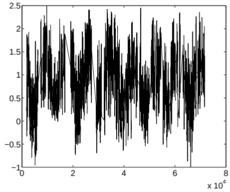
The problem

- Estimation of weather related return periods in the presence of strong seasonal effects
- Example: Significant wave height H_s (4 times standard deviation of sea surface height) in the North Sea March 1980 – March 1988, 3 hour sampling inteval

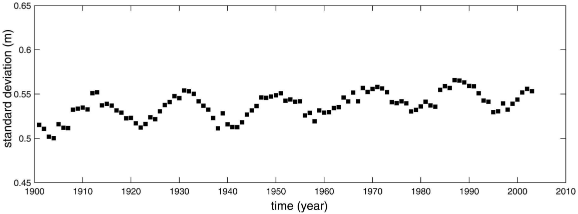


The problem, continued

- Strong seasonal effect for location in log H_s, i.e. in scale of H_s

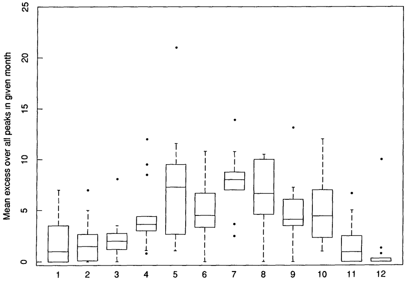


- Standard deviation of hourly sea level – astronomical cycle with period 18.61 year; Méndez et al, J. Atmospheric and Oceanic Technology, 2007



Common practise – I

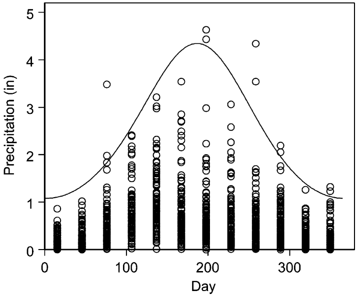
- Divide the year into "homogeneous" intervals, e.g. months, and make separate analyses for each period
- Smith: Extreme value analysis of environmental time series ..., Statistical Science, 1989



Central bar at median; upper and lower bars at quartiles;
whiskers extend to 1.5 times inter-quartile range

Common practise – II

- Estimate parameteric models for time dependent parameters in extreme value analysis
- Katz et al: Statistics of extremes in hydrology, Advances in Water Resources, 2002 – upper curve shows separate 100-year return level for GEV model for monthly maximum of Fort Collins precipitation.



The fixed periodic model

- For seasonal GEV analysis for monthly maximum, the maximum value over month $t = 1, \ldots, 12$ is assumed to have a GEV distribution with location $\mu(t)$, scale $\sigma(t)$ and shape parameter $\gamma(t)$ dependent on t:
- Katz, Méndez, and many others, assume ("sum of") harmonics:

$$\begin{aligned}
\mu(t) &= \mu_0 + \alpha_1 \cos 2\pi f_0 t + \beta_1 \sin 2\pi f_0 t \\
&= \mu_0 + A_1 \cos(2\pi f_0 t + \phi_1) \\
\log \sigma(t) &= s_0 + A_2 \cos(2\pi f_0 t + \phi_2)
\end{aligned}$$

and possibly a constant $\gamma(t)$.
Frequency $f_0 = 1/12$ for monthly maxima, and ϕ_k are fixed phases for the harmonic change in location and scale.

But – is the "fixed phase" model realistic?

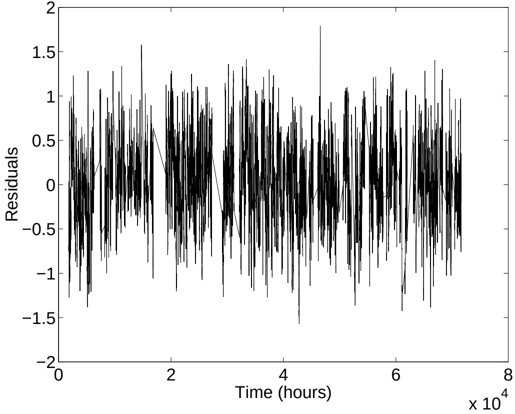
- Does the "stormy season" occur at a fixed date every year?
- Or does it come with some phase variation?
- And, does that really matter,
- so it can systematically influence estimation of return levels?
- Take a look at the log H_s data from the North Sea, centered around 0.
- Fit a fixed phase cosine to y_t, the centered $\log H_s(t)$ data from the North Sea, minimizing

$$\sum (y_j - \alpha \cos 2\pi f_0 t_j - \beta \sin 2\pi f_0 t_j)^2$$

- and take $\widehat{A} = \sqrt{\alpha^2 + \beta^2}$, $\widehat{\phi} = -\arctan \beta/\alpha$

A look at the North Sea significant wave height

- Residuals, $y_t - \widehat{A}\cos(2\pi f_0 t + \widehat{\phi})$, around the mean value function still contain a seasonal component!



The stormy season may come at different dates!

- Assume a slowly varying random amplitude A_t and phase $\phi(t)$

$$m(t) = A_t \cos(2\pi f_0 t + \phi_t)$$

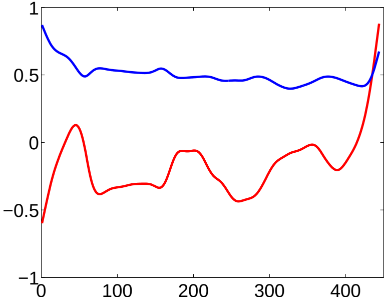
- Estimate A_t and ϕ_t by local least squares, minimizing, for fixed t,

$$\sum_{|t_j - t| < h} (y_j - \alpha_t \cos 2\pi f_0 t_j - \beta_t \sin 2\pi f_0 t_j)^2 \, \frac{K((t_j - t)/h)}{h}$$

- and take $\widehat{A}_t = \sqrt{\alpha_t^2 + \beta_t^2}$, $\widehat{\phi}_t = -\arctan \beta_t / \alpha_t$

Amplitude and phase are not constant

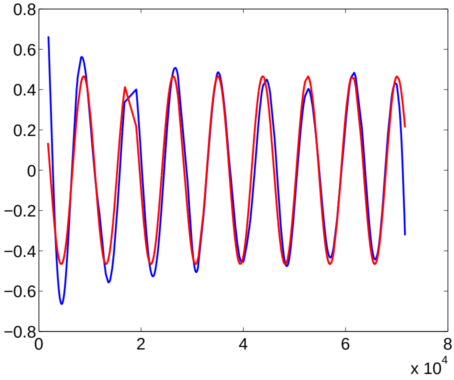
- Amplitude and phase vary slowly over the 8 years of North Sea data: plot of amplitude $\widehat{A}_t$ and phase $\widehat{\phi}_t$



- Phase varies by ± 0.2 around its average value, i.e. the time for peak H_s may shift back and forth with about 12 days between years!

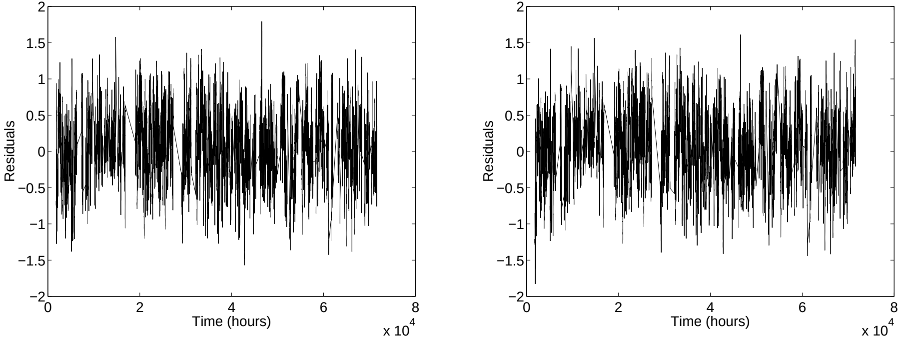
Does the phase shift have any effect on the extreme value analysis?

- Visible effect on seasonal pattern:



Fixed and shifting seasonal effects

- Small, but clear (?) difference between residual patterns for fixed (left) and shifting (right) phases



Simulation study: does random phase affect extreme value analysis?

Gaussian time series	sample interval	3 hours
Mean value	yearly cosine	variable A and ϕ
Residual standard deviation	constant	σ_y
Estimation assumption I	constant A and ϕ	WRONG!
Estimation assumption II	variable A and ϕ	CORRECT!

Estimate residual distribution

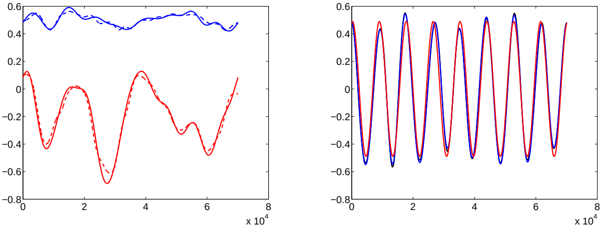
Question: are extreme quantiles correctly estimated?

The model

- Independent residuals Y_t around a random season:

$$\begin{aligned}
Y_t &= m_t + N(0, \sigma_y^2) \\
m_t &= m_0 + A_t \cos(2\pi f_0 t + \phi_t), \quad \mathrm{Var}(\mathrm{m_t}) = \sigma_\mathrm{m}^2 \\
A_t &= 0.5 + \widetilde{A}_t, \quad \text{stationary Gaussian, psd } S_a(\omega) \\
\phi_t &= -0.2 + \widetilde{\phi}_t, \quad \text{stationary Gaussian, psd } S_f(\omega)
\end{aligned}$$

- Left: examples of A_t and ϕ_t with estimates; Right: example of m_t (black), $\widehat{m}_t$ fixed (red), $\widehat{m}_t$ flexible (blue)



Estimate residual distribution under the two assumptions when in fact model II is correct?

- Generate data (8 year) from Model II (flexible) and estimate season m_t under Assumption I and Assumption II
- Residuals

$$x_t^I = y_t - \widehat{m}_t^I$$
$$x_t^{II} = y_t - \widehat{m}_t^{II}$$

- Compute empirical quantiles in the two sets of residuals
- Compare the two sets of quantiles

Results – 1a

- The results will depend on the relation between the residual standard deviation σ_y and the variability of the season measured by its standard deviation σ_m.
- Upper quantile ratio $\lambda_q^I/\lambda_q^{II}$ based on 100 replicates

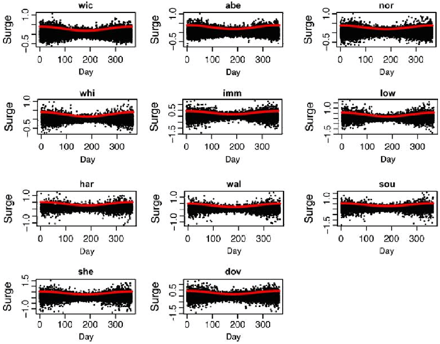
q =		0.9	–	0.99999
$\sigma_y/\sigma_m =$	1	1.05	–	1.04
	0.75	1.08	–	1.08
	0.6	1.12	–	1.11
	0.5	1.17	–	1.14
	0.4	1.27	–	1.25
	0.3	1.31	–	1.26
	0.2	1.59	–	1.45
	0.1	2.74	–	2.10

Results – 1b

- If residual and season have the same standard deviation, $\sigma_y = \sigma_m$, the effect of phase mismatch is small ($< 5\%$)
- If $\sigma_y = \sigma_m/2$ high residual quantiles may be overestimated by 15%
- If $\sigma_y = \sigma_m/5$ high residual quantiles may be overestimated by 50%

Peaks over threshold analysis

- For POT analysis with seasonal effect, select a variable high level and analyse exceedances
- Coles and Tawn: Seasonal effects of extreme surges. *Stoch Environ Res Risk Assess* (2005):



Threshold selection

- Coles and Tawn select a seasonal quantile curve,

$$Q(t) = a + b\cos(2\pi f_0 t + c)$$

  exceeded by 5% of all data, both summer and winter (actually C&T take $c = 0$) and make an extreme value analysis of the exceedances
- We compare with a similar analysis with a seasonal quantile curve with flexible phase
- Estimate by local quantile regression a quantile curve

$$Q(t) = A(t) + B(t)\cos(2\pi f_0 t) + C(t)\sin(2\pi f_0 t)$$

  with slowly varying $A(t), B(t), C(t)$ – this is a standard quantile regression problem

Local quantile regression

Recursion: for a small ϵ,

$$\theta^{m+1} = \arg\min \sum_{|t_j - t| < h} (y_j - A(t) - B(t)\cos 2\pi f_0 t_j - C(t)\sin 2\pi f_0 t_j)^2 \times \frac{v(y_j)\frac{K((t_j - t)/h)}{h}}{\sqrt{(y_j - Q(t)^m)^2 + \epsilon^2}}$$

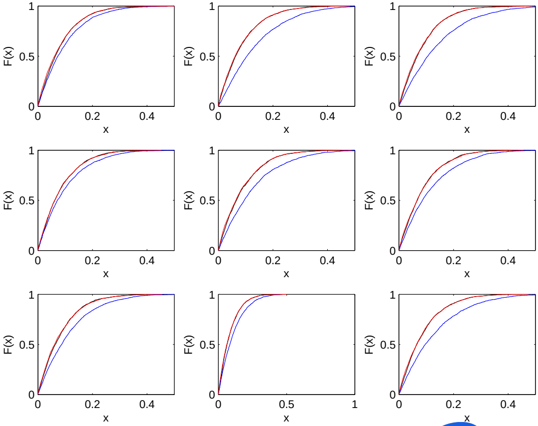
optimize for $A(t), B(t), C(t)$

Results – 2

- Simulate 8 year of independent normal data, around a varying mean, sampled 8 times a day
- Yearly season with flexible phase similar to H_s North Sea model
- Estimate upper quantile function with "fixed phase" and with "flexible phase"
- Compute exceedance distributions

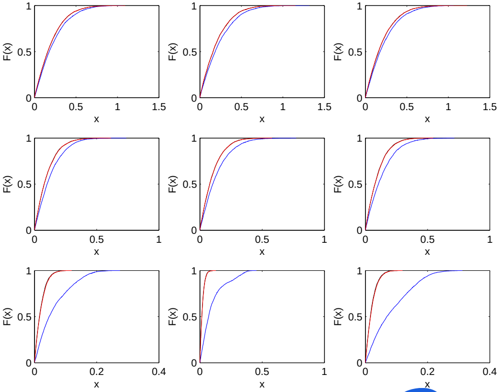
Results 2a – variability due to phase variation

Nine simulations of empirical CDF (blue=fixed, red=flexible) for exceedances over 90% quantile curve and true distribution (black) – phase STD twice that in "Results 1", $s_y = s_m/2$



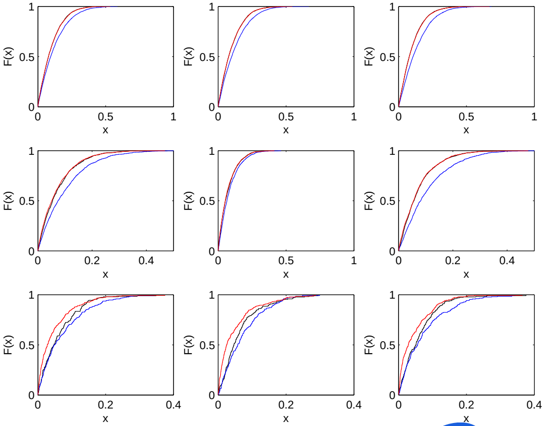
Results 2b – dependence on "signal-to-noise" ratio

Results depend on the "signal-to-noise" ratio, i.e. the ratio between the variance of the residuals and the season. Empirical CDF of exceedances over 80% limits. Upper row: $s_y = s_m$, Middle row: $s_y = s_m/2$, Bottom row: $s_y = s_m/10$,



Results 2c – dependence on quantile level

Results depend on the quantile level. Empirical CDF of exceedances over $80, 95, 99\%$ quantiles. $s_y = s_m/2$. Upper row: $p = 0.80$, Middle row: $p = 0.95$, Bottom row: $p = 0.99$,

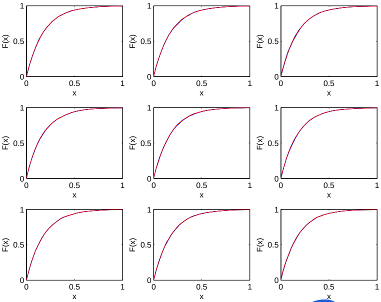


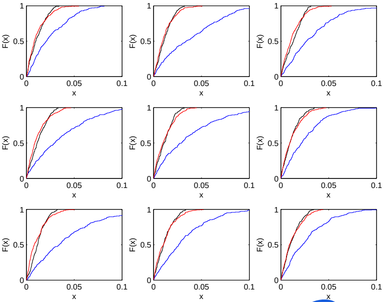
Result 2d – dependence on exceedance distribution

- Examples have illustrated normal data with almost exponential exceedance distribution
- Seasonal baseline function + exponential residuals gives (trivially) identical results for fixed and flexible quantile estimation
- Seasonal baseline function + GPD residuals follow the main pattern, with "fixed phase" assumption overestimating the excesses when phase shifts are present

Result 2d – examples

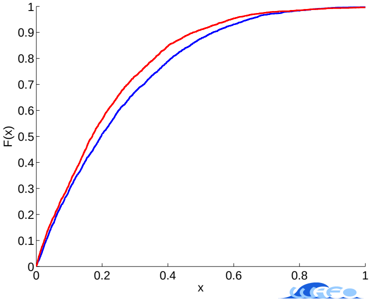
Left: exponential 80%; Right: GPD 99%





Back to the North Sea wave height

Based on available data with large gaps, estimate fixed (blue) and flexible (red) seasonal 80% quantile curves and obtain exceedance CDF – results are consistent with simulations (h=1600 hours). The overestimation in the center of the exceedance distribution is about 0.05, which implies a 5% overestimation of the true significant wave height return value



New problem: Should season amplitude be allowed to vary over years?

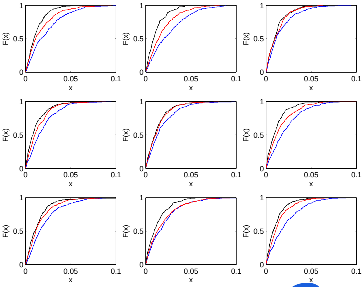
- The seasonal variation of the amplitude in Result 2 is confounded with the exceedance variation.
- Therefore, estimate quantile curve with fixed amplitude – allowing only the phase to vary between years
- Technically, first estimate variable amplitude and phase model
- Then, accept the estimated phase shifts $\phi(t)$ and estimate a new constant amplitude quantile function of the form

$$m_t = m_0 + A\cos(2\pi f_0 t + \phi(t))$$

Estimation of m_0 and A is now a linear quantile estimation problem.

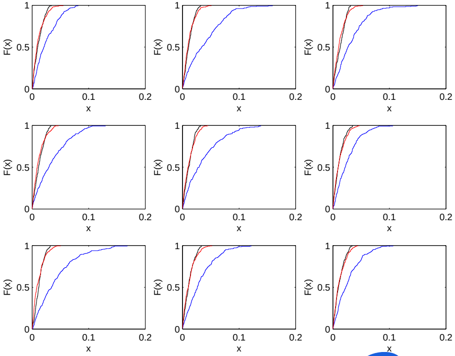
Results 3a – main results remain - normal residuals

Nine simulations of constant amplitude-variable phase model. Normal residuals with $s_y = 0.1 s_m$. Estimation of CDF from constant amplitude (red) and constant amplitude and phase (blue) compared to true exceedance CDF (black). 99% quantile. Note: the "flexible phase-constant amplitude" falls between the "fixed phase" CDF and the true CDF



Results 3b – main results remain - GPD

Nine simulations of constant amplitude-variable phase model. GPD residuals with shape parameter 0.5 and scale $s_y = 0.5 s_m$. Estimation of CDF from constant amplitude (red) and constant amplitude and phase (blue) compared to true exceedance CDF (black). 99% quantile.



Discussion 1

- Seasonal parameters in EVD or in "normal" models are not unreasonable in weather and climate models
- Formal non-parametric quantile regression can accurately estimate a amplitude/phase-shifting seasonal trend
- Trend estimation allowing for phase (and amplitude) shifts seem to "often" agree with the true shifts – when they are present
- Exceedances over an estimated fixed-phase season are (in simulations) systematically larger than exceedances over an estimated flexible phase model in the presence of flexible season
- Difference in "Significant wave height" example is of the order of 5% in moderately high extremes

Discussion 2

- Difference seems to be smaller for more extreme quantiles
- Difference depends on the "signal-to-noise" ratio
- Exceedance models from flexible phase model can be combined with a fixed phase model for prediction puposes
- Gaps with missing values may cause problems
- Effects of the variable amplitude have not been studied
- CONCLUSION: If one prefers a POT analysis in the presence of strong seasonal effects, the problem with modelling possible phase variations should be considered – return values may otherwise be biased in an unconservative way
- Block maxima (e.g. yearly) may be to prefer

References

- Helena Olsson: A study of extreme significant wave heights in the Norwegian Sea. Masters Thesis in Mathematical statistics, Lund University, 1994.
- G. Lindgren and H. Olsson: The effect of seasonal phase mismatch in extreme value analysis of environmental variables. In preparation.
- R. Katz, M.B. Parlange and P. Naveau: Statistics of extremes in hydrology. *Advances in Water Resources*, 25 (2002) 1287–1304.
- F.J. Méndez et al.: Analyzing monthly extreme sea levels with a time-dependent GEV model. *J. Atmospheric and Oceanic Research*, 24 (2007) 894–911.
- R.L. Smith: Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone. *Statistical Science*, 4 (1989) 367–377.