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

Georg Lindgren¹
with input from Helena Olsson

¹Mathematical 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 interval
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

![Box plot showing mean excesses over all peaks in given month with central bar at median, upper and lower bars at quartiles, and whiskers extending to 1.5 times inter-quartile range.](image)
Common practise – II

- Estimate parameteric models for time dependent parameters in extreme value analysis
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:

$$
\mu(t) = \mu_0 + \alpha_1 \cos 2\pi f_0 t + \beta_1 \sin 2\pi f_0 t \\
= \mu_0 + A_1 \cos(2\pi f_0 t + \phi_1) \\
\log \sigma(t) = s_0 + A_2 \cos(2\pi f_0 t + \phi_2)
$$

and possibly a constant $\gamma(t)$.
Frequency $f_0 = 1/12$ for monthly maxima, and $\phi_k$ 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 $\hat{A} = \sqrt{\alpha^2 + \beta^2}$, $\phi = -\arctan \beta / \alpha$
A look at the North Sea significant wave height

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

![Graph showing residuals](image-url)
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 $\phi_t$ 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 $\hat{A}_t = \sqrt{\alpha_t^2 + \beta_t^2}$, $\hat{\phi}_t = -\arctan \beta_t/\alpha_t$
Amplitude and phase are not constant

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

![Amplitude and phase plot]

- Phase varies by $\pm 0.2$ 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 $\phi$
Residual standard deviation  constant  $\sigma_y$
Estimation assumption I  constant A and $\phi$  WRONG!
Estimation assumption II  variable A and $\phi$  CORRECT!

Estimate residual distribution
Question  are extreme quantiles correctly estimated?
The model

- Independent residuals $Y_t$ around a random season:
  \[ Y_t = m_t + N(0, \sigma^2_Y) \]
  \[ m_t = m_0 + A_t \cos(2\pi f_0 t + \phi_t), \quad \text{Var}(m_t) = \sigma^2_m \]
  \[ A_t = 0.5 + \tilde{A}_t, \quad \text{stationary Gaussian, psd } S_a(\omega) \]
  \[ \phi_t = -0.2 + \tilde{\phi}_t, \quad \text{stationary Gaussian, psd } S_f(\omega) \]

- Left: examples of $A_t$ and $\phi_t$ with estimates; Right: example of $m_t$ (black), $\hat{m}_t$ fixed (red), $\hat{m}_t$ 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
  \[ x_t^I = y_t - \hat{m}_t^I \]
  \[ x_t^{II} = y_t - \hat{m}_t^{II} \]
- Compute empirical quantiles in the two sets of residuals
- Compare the two sets of quantiles
The results will depend on the relation between the residual standard deviation $\sigma_y$ and the variability of the season measured by its standard deviation $\sigma_m$.

Upper quantile ratio $\lambda_q^l/\lambda_q^\|$ based on 100 replicates

<table>
<thead>
<tr>
<th>$\sigma_y/\sigma_m$</th>
<th>$q = 0.9$</th>
<th>$q = 0.99999$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.05</td>
<td>1.04</td>
</tr>
<tr>
<td>0.75</td>
<td>1.08</td>
<td>1.08</td>
</tr>
<tr>
<td>0.6</td>
<td>1.12</td>
<td>1.11</td>
</tr>
<tr>
<td>0.5</td>
<td>1.17</td>
<td>1.14</td>
</tr>
<tr>
<td>0.4</td>
<td>1.27</td>
<td>1.25</td>
</tr>
<tr>
<td>0.3</td>
<td>1.31</td>
<td>1.26</td>
</tr>
<tr>
<td>0.2</td>
<td>1.59</td>
<td>1.45</td>
</tr>
<tr>
<td>0.1</td>
<td>2.74</td>
<td>2.10</td>
</tr>
</tbody>
</table>
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.
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 $\epsilon$,

$$\theta^{m+1} = \arg \min_{|t_j - t| < h} \sum (y_j - A(t) - B(t) \cos 2\pi f_0 t_j - C(t) \sin 2\pi f_0 t_j)^2$$

$$\times \frac{\nu(y_j) K((t_j - t)/h)}{\sqrt{(y_j - Q(t)^m)^2 + \epsilon^2}}$$

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 $s_y = 0.1s_m$. 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 $s_y = 0.5s_m$. 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 purposes
- 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