Modeling extreme values of processes observed at irregular time steps.
Application to significant wave height

Nicolas Raillard

Laboratoire Comportement des Structures en Mer
IFREMER, Brest

Joint work with Pierre Ailliot (UBO)

Introduction

- Hs : significant wave height, parameter related to wave energy in a sea state
  - Historical definition: mean height of the one-third highest waves
  - Current definition: four times the standard deviation of the sea surface elevation
- Aim: develop a method to estimate the extremal properties of Hs
  - Quantities of interest: return levels, storm durations,...
  - Example: 20-year return levels in the North Atlantic (method explained hereafter)
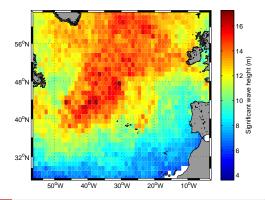
Introduction

- Hs : significant wave height, parameter related to wave energy in a sea state
  - Historical definition: mean height of the one-third highest waves
  - Current definition: four times the standard deviation of the sea surface elevation
- Aim: develop a method to estimate the extremal properties of Hs
  - Quantities of interest: return levels, storm durations,...
  - Example: 20-year return levels in the North Atlantic (method explained hereafter)

Introduction

- Hs : significant wave height, parameter related to wave energy in a sea state
  - Historical definition: mean height of the one-third highest waves
  - Current definition: four times the standard deviation of the sea surface elevation
- Aim: develop a method to estimate the extremal properties of Hs
  - Quantities of interest: return levels, storm durations,...
  - Example: 20-year return levels in the North Atlantic (method explained hereafter)

Introduction   Data available   Method
00000   Analysis of Hs data
00   Conclusion

Plan of the talk

Plan of the talk

1. **Data available**

2. Method
   - Usual methods in the iid case
   - Generalization to dependent sequences

3. Analysis of Hs data
   - Data available
   - A few results

4. **Conclusion**

Introduction
What are the data available to estimate the extremal properties of Hs

- Reanalysis data
    - Rich space-time sampling
    - Standard statistical method apply!
    - Underestimation for high Hs?
- Buoy data
    - Sparse spatial sampling
    - Rich time sampling
    - Most reliable for high Hs?
- Satellite altimeter data
    - Sparse space-time sampling
    - Good spatial coverage
    - Quality of the data for high Hs?

Figure : Data available on 15/12/2002.

Introduction
What are the data available to estimate the extremal properties of Hs

- Reanalysis data
  - Rich space-time sampling
  - Standard statistical method apply!
  - Underestimation for high Hs?
- Buoy data
  - Sparse spatial sampling
  - Rich time sampling
  - Most reliable for high Hs?
- Satellite altimeter data
  - Sparse space-time sampling
  - Good spatial coverage
  - Quality of the data for high Hs?

Figure : Data available on 15/12/2002.

Introduction
What are the data available to estimate the extremal properties of Hs

- Reanalysis data
  - Rich space-time sampling
  - Standard statistical method apply!
  - Underestimation for high Hs?
- Buoy data
  - Sparse spatial sampling
  - Rich time sampling
  - Most reliable for high Hs?
- Satellite altimeter data
  - Sparse space-time sampling
  - Good spatial coverage
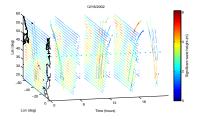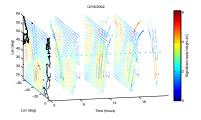 - Quality of the data for high Hs?

Figure : Data available on 15/12/2002.

Hs data

- Single site analysis
  - Buoy Brittany (47.5 N, 8.5 W)
    - 10 years of data available, $\Delta t = 1h$
    - 7.7% of missing data
    - Breakdowns linked to extreme events?
  - Reanalysis data: ERA interim (ECMWF)
    - Same years than buoy, $\Delta t = 6h$
  - Satellite data
    - 15 years of data available, 7 different satellites
    - Closest observations in the satellite tracks which intersect a $3^o \times 3^o$ box

Figure : Hs data for dec. 1999

Hs data

- Single site analysis
  - Buoy Brittany (47.5 N, 8.5 W)
    - 10 years of data available, $\Delta t = 1h$
    - 7.7% of missing data
    - Breakdowns linked to extreme events?
  - Reanalysis data: ERA interim (ECMWF)
    - Same years than buoy, $\Delta t = 6h$
  - Satellite data
    - 15 years of data available, 7 different satellites
    - Closest observations in the satellite tracks which intersect a $3^o \times 3^o$ box

- How can we analyze the extremal behavior of such time series?
- Need for methods which can deal with missing data and irregular time sampling
- Can we get useful information on extreme Hs from satellite data?

Plan of the talk

1. Data available

2. **Method**
   - Usual methods in the iid case
   - Generalization to dependent sequences

3. Analysis of Hs data
   - Data available
   - A few results

4. Conclusion

Probabilistic background (iid case)

- Let $M_n = \max\limits_{i=1,\ldots,n} X_i$ with (X_i) i.i.d. sample with c.d.f. G ;

- If $\mathbb{P}\left\{\frac{M_n - b_n}{a_n} \leq x\right\} = G^n(a_n x + b_n) \to F(x)$, then F is a max-stable distribution ie

$$F^n(\alpha_n x + \beta_n) = F(x)$$

- Fisher-Tippett (1928), Gnedenko (1943): Max-stable distributions have cdf

$$F(x; \mu, \sigma, \xi) = \exp\left[-\left(1 + \xi\frac{x - \mu}{\sigma}\right)_+^{-1/\xi}\right] \in \text{GEV}(\mu, \sigma, \xi)$$

  This GEV distribution includes the Weibull (finite upper bound, $\xi < 0$), Fréchet (heavy tail, $\xi > 0$) and Gumbel ($\xi = 0$) distributions

- The GEV is a natural distribution for modeling the maximum of a large number of i.i.d. random variables

Probabilistic background (iid case)

- Let $M_n = \max\limits_{i=1,\ldots,n} X_i$ with (X_i) i.i.d. sample with c.d.f. G ;
- If $\mathbb{P}\left\{ \frac{M_n - b_n}{a_n} \leq x \right\} = G^n(a_n x + b_n) \to F(x)$, then F is a max-stable distribution ie

$$F^n(\alpha_n x + \beta_n) = F(x)$$

- Fisher-Tippett (1928), Gnedenko (1943): Max-stable distributions have cdf

$$F(x; \mu, \sigma, \xi) = \exp\left[-\left(1 + \xi \frac{x - \mu}{\sigma} \right)_+^{-1/\xi} \right] \in \text{GEV}(\mu, \sigma, \xi)$$

This GEV distribution includes the Weibull (finite upper bound, $\xi < 0$), Fréchet (heavy tail, $\xi > 0$) and Gumbel ($\xi = 0$) distributions

- The GEV is a natural distribution for modeling the maximum of a large number of i.i.d. random variables

Probabilistic background (iid case)

- Let $M_n = \max\limits_{i=1,\ldots,n} X_i$ with (X_i) i.i.d. sample with c.d.f. G ;

- If $\mathbb{P}\left\{\frac{M_n - b_n}{a_n} \le x\right\} = G^n(a_n x + b_n) \to F(x)$, then F is a max-stable distribution ie

$$F^n(\alpha_n x + \beta_n) = F(x)$$

- Fisher-Tippett (1928), Gnedenko (1943): Max-stable distributions have cdf

$$F(x; \mu, \sigma, \xi) = \exp\left[-\left(1 + \xi\frac{x - \mu}{\sigma}\right)_+^{-1/\xi}\right] \in \text{GEV}(\mu, \sigma, \xi)$$

This GEV distribution includes the Weibull (finite upper bound, $\xi < 0$), Fréchet (heavy tail, $\xi > 0$) and Gumbel ($\xi = 0$) distributions

- The GEV is a natural distribution for modeling the maximum of a large number of i.i.d. random variables

Probabilistic background (iid case)

- Let $M_n = \max\limits_{i=1,\ldots,n} X_i$ with (X_i) i.i.d. sample with c.d.f. G ;

- If $\mathbb{P}\left\{\frac{M_n - b_n}{a_n} \leq x\right\} = G^n(a_n x + b_n) \to F(x)$, then F is a max-stable distribution ie
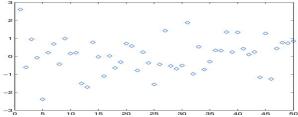
$$F^n(\alpha_n x + \beta_n) = F(x)$$

- Fisher-Tippett (1928), Gnedenko (1943): Max-stable distributions have cdf

$$F(x; \mu, \sigma, \xi) = \exp\left[-\left(1 + \xi \frac{x - \mu}{\sigma}\right)_+^{-1/\xi}\right] \in \mathrm{GEV}(\mu, \sigma, \xi)$$

This GEV distribution includes the Weibull (finite upper bound, $\xi < 0$), Fréchet (heavy tail, $\xi > 0$) and Gumbel ($\xi = 0$) distributions

- The GEV is a natural distribution for modeling the maximum of a large number of i.i.d. random variables

Usual methods for analyzing extremes of i.i.d. sequences

- **Block maxima method**

  - Group the data into blocks of equal lengths (usually one year to remove seasonal effects)
  - Fit a GEV distribution to the sample of block maxima
  - Waste of data (only the yearly maxima are kept to fit the GEV)
    - Only 10 observations to fit the GEV for buoy data although more storms are observed!

- Peaks Over Threshold (POT)
  - Choose a high threshold u
  - Fit a GPD distribution to the exceedances above u

  $$X_i - u | X_i > u \sim_{i.i.d.} GPD(\theta, \xi)$$

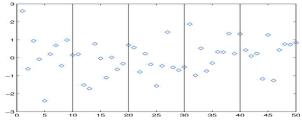 with $\lambda = \mathbb{P}(X_i > u)$ an extra unknown parameter
  - Similar results when fitting a seasonal GEV distribution

  $$u\mathbb{1}[X_i \leq u] + X_i \mathbb{1}[X_i > u] = u\mathbb{1}[Y_i \leq u] + Y_i\mathbb{1}[Y_i > u] \text{ with } Y_i \sim_{i.i.d.} GEV(\mu, \sigma, \xi)$$

  - GEV and GPD distributions lead to similar tail approximations. If $x > u$ then

  $$\exp\left[-\left(1 + \xi \frac{x-\mu}{\sigma}\right)_+^{-1/\xi}\right] \approx 1 - \left(1 + \xi \frac{x-u}{\sigma}\right)_+^{-1/\xi}$$

Usual methods for analyzing extremes of i.i.d. sequences

- **Block maxima method**
  - Group the data into blocks of equal lengths (usually one year to remove seasonal effects)

  - Fit a GEV distribution to the sample of block maxima
  - Waste of data (only the yearly maxima are kept to fit the GEV)
    - Only 10 observations to fit the GEV for buoy data although more storms are observed!
- Peaks Over Threshold (POT)
  - Choose a high threshold u
  - Fit a GPD distribution to the exceedances above u

$$X_i - u | X_i > u \sim_{i.i.d.} GPD(\delta, \xi)$$

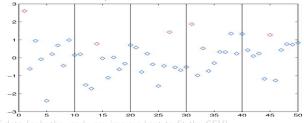 with $\lambda = \mathbb{P}(X_i > u)$ an extra unknown parameter

  - Similar results when fitting a censored GEV distribution

$$u\mathbb{1}[X_i \leq u] + X_i\mathbb{1}[X_i > u] = u\mathbb{1}[Y_i \leq u] + Y_i\mathbb{1}[Y_i > u] \text{ with } Y_i \sim_{i.i.d.} GEV(\mu, \sigma, \xi)$$

    - GEV and GPD distributions lead to similar tail approximations. If $z \mapsto \infty$ then

$$\exp\left[-\left(1 + \xi \frac{x - \mu}{\sigma}\right)^{-1/\xi}\right] \approx 1 - \left(1 + \xi \frac{x - \mu}{\sigma}\right)^{-1/\xi}$$

  - Successive exceedances are generally dependent

Usual methods for analyzing extremes of i.i.d. sequences

- Block maxima method
  - Group the data into blocks of equal lengths (usually one year to remove seasonal effects)
  - Fit a GEV distribution to the sample of block maxima

- Waste of data (only the yearly maxima are kept to fit the GEV)
  - Only 10 observations to fit the GEV for buoy data although more storms are observed!
- Peaks Over Threshold (POT)
  - Choose a high threshold u
  - Fit a GPD distribution to the exceedances above u

$$X_i - u | X_i > u \sim_{i.i.d.} GPD(\beta, \xi)$$

  with $\lambda = \mathbb{P}(X_i > u)$ an extra unknown parameter
  - Similar results when fitting a censored GEV distribution

$$u \mathbb{1}[X_i \leq u] + X_i \mathbb{1}[X_i > u] = u \mathbb{1}[Y_i \leq u] + Y_i \mathbb{1}[Y_i > u] \text{ with } Y_i \sim_{i.i.d.} GEV(\mu_i, \sigma, \xi)$$

  - GEV and GPD distributions lead to similar tail approximations. If $z > u$, then

$$\exp\left[-\left(1 + \xi \frac{z - \mu}{\sigma}\right)_+^{-1/\xi}\right] \approx 1 - \left(1 + \xi \frac{z - u}{\tilde{\sigma}}\right)_+^{-1/\xi}$$

  - Successive exceedances are generally dependent

Usual methods for analyzing extremes of i.i.d. sequences

- **Block maxima method**
  - Group the data into blocks of equal lengths (usually one year to remove seasonal effects)
  - Fit a GEV distribution to the sample of block maxima
  - Waste of data (only the yearly maxima are kept to fit the GEV)
    - Only 10 observations to fit the GEV for buoy data although more storms are observed!

- Peaks Over Threshold (POT)
  - Choose a high threshold u
  - Fit a GPD distribution to the exceedances above u

  $$X_t - u | X_t > u \sim_{i.i.d.} GPD(\delta, \xi)$$

  with $\lambda = P(X_t > u)$ an extra unknown parameter
  - Similar results when fitting a censored GEV distribution

  $$u \mathbb{1}[X_t \le u] + X_t \mathbb{1}[X_t > u] = u \mathbb{1}[Y_t \le u] + Y_t \mathbb{1}[Y_t > u] \text{ with } Y_t \sim_{i.i.d.} GEV(\mu, \sigma, \xi)$$

    - GEV and GPD distributions lead to similar tail approximations. If $x \to \infty$ then

    $$\exp\left[-\left(1 + \xi \frac{x - \mu}{\sigma}\right)_+^{-1/\xi}\right] \simeq 1 - \left(1 + \xi \frac{x - \mu}{\sigma}\right)_+^{-1/\xi}$$

  - Successive exceedances are generally dependent
    - A "declustering" step is generally applied to keep only one value per "storm"
    - Difficulties: arbitrary rules to define the "storms", waste of data, loss of "storm" profile
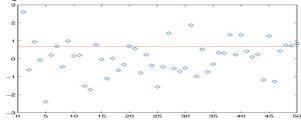   - Alternative: model $\{Y_t\}$ as a max-stable process and keep all exceedances

Usual methods for analyzing extremes of i.i.d. sequences

- Block maxima method
  - Group the data into blocks of equal lengths (usually one year to remove seasonal effects)
  - Fit a GEV distribution to the sample of block maxima
  - Waste of data (only the yearly maxima are kept to fit the GEV)
    - Only 10 observations to fit the GEV for buoy data although more storms are observed!
- Peaks Over Threshold (POT)
  - Choose a high threshold u

- Fit a GPD distribution to the exceedances above u

$$X_i - u | X_i > u \sim_{i.i.d} GPD(\tilde{\sigma}, \xi)$$

with $\lambda = \mathbb{P}(X_i > u)$ an extra unknown parameter
- Similar results when fitting a censored GEV distribution

$$u\mathbf{1}[X_i \leq u] + X_i\mathbf{1}[X_i > u] = u\mathbf{1}[Y_i \leq u] + Y_i\mathbf{1}[Y_i > u] \text{ with } Y_i \sim_{i.i.d} GEV(\mu, \sigma, \xi)$$

- GEV and GPD distributions lead to similar tail approximations. If $x \to x+$ then

$$\exp\left[-\left(1 + \xi\frac{x-\mu}{\sigma}\right)_+^{-1/\xi}\right] \approx 1 - \left(1 + \xi\frac{x-\mu}{\sigma}\right)_+^{-1/\xi}$$

- Successive exceedances are generally dependent

Usual methods for analyzing extremes of i.i.d. sequences
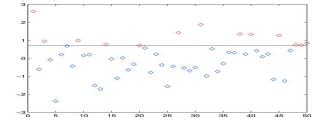
- Block maxima method
  - Group the data into blocks of equal lengths (usually one year to remove seasonal effects)
  - Fit a GEV distribution to the sample of block maxima
  - Waste of data (only the yearly maxima are kept to fit the GEV)
    - Only 10 observations to fit the GEV for buoy data although more storms are observed!
- Peaks Over Threshold (POT)
  - Choose a high threshold u
  - Fit a GPD distribution to the exceedances above u

$$X_i - u | X_i > u \sim_{i.i.d} GPD(\tilde{\sigma}, \xi)$$

  with $\lambda = \mathbb{P}(X_i > u)$ an extra unknown parameter

- Similar results when fitting a censored GEV distribution

  $u\mathbf{1}[X_i \leq u] + X_i\mathbf{1}[X_i > u] = u\mathbf{1}[Y_i \leq u] + Y_i\mathbf{1}[Y_i > u]$ with $Y_i \sim_{i.i.d} GEV(\mu, \sigma, \xi)$

  - GEV and GPD distributions lead to similar tail approximations. If $x \to x+$ then

$$\exp\left[-\left(1 + \xi\frac{x-\mu}{\sigma}\right)_+^{-1/\xi}\right] \approx 1 - \left(1 + \xi\frac{x-\mu}{\sigma}\right)_+^{-1/\xi}$$

- Successive exceedances are generally dependent

Usual methods for analyzing extremes of i.i.d. sequences

- **Block maxima method**
  - Group the data into blocks of equal lengths (usually one year to remove seasonal effects)
  - Fit a GEV distribution to the sample of block maxima
  - Waste of data (only the yearly maxima are kept to fit the GEV)
    - Only 10 observations to fit the GEV for buoy data although more storms are observed!

- **Peaks Over Threshold (POT)**
  - Choose a high threshold u
  - Fit a GPD distribution to the exceedances above u

$$X_i - u | X_i > u \sim_{i.i.d} GPD(\tilde{\sigma}, \xi)$$

  with $\lambda = \mathbb{P}(X_i > u)$ an extra unknown parameter
  - Similar results when fitting a censored GEV distribution

$$u\mathbf{1}[X_i \leq u] + X_i\mathbf{1}[X_i > u] = u\mathbf{1}[Y_i \leq u] + Y_i\mathbf{1}[Y_i > u] \text{ with } Y_i \sim_{i.i.d} GEV(\mu, \sigma, \xi)$$

  - GEV and GPD distributions lead to similar tail approximations. If $x \to x+$ then

$$\exp\left[-\left(1 + \xi\frac{x-\mu}{\sigma}\right)_+^{-1/\xi}\right] \approx 1 - \left(1 + \xi\frac{x-\mu}{\sigma}\right)_+^{-1/\xi}$$

  - Successive exceedances are generally dependent
    - A "declustering" step is generally applied to keep only one value per "storm";
    - Difficulties: arbitrary rules to define the "storms", waste of data, loss of "storm" profile...
    - Alternative: model $\{Y_i\}$ as a max-stable process and keep all exceedances.

Usual methods for analyzing extremes of i.i.d. sequences

- Block maxima method
  - Group the data into blocks of equal lengths (usually one year to remove seasonal effects)
  - Fit a GEV distribution to the sample of block maxima
  - Waste of data (only the yearly maxima are kept to fit the GEV)
    - Only 10 observations to fit the GEV for buoy data although more storms are observed!
- Peaks Over Threshold (POT)
  - Choose a high threshold u
  - Fit a GPD distribution to the exceedances above u

  $$X_i - u | X_i > u \sim_{i.i.d} GPD(\tilde{\sigma}, \xi)$$

  with $\lambda = \mathbb{P}(X_i > u)$ an extra unknown parameter
  - Similar results when fitting a censored GEV distribution

  $$u\mathbf{1}[X_i \leq u] + X_i\mathbf{1}[X_i > u] = u\mathbf{1}[Y_i \leq u] + Y_i\mathbf{1}[Y_i > u] \text{ with } Y_i \sim_{i.i.d} \text{GEV}(\mu, \sigma, \xi)$$

    - GEV and GPD distributions lead to similar tail approximations. If $x \to x+$ then

    $$\exp\left[-\left(1 + \xi\frac{x-\mu}{\sigma}\right)_+^{-1/\xi}\right] \approx 1 - \left(1 + \xi\frac{x-\mu}{\sigma}\right)_+^{-1/\xi}$$

  - Successive exceedances are generally dependent
    - A "declustering" step is generally applied to keep only one value per "storm";
    - Difficulties: arbitrary rules to define the "storms", waste of data, loss of "storm" profile...
    - Alternative: model $\{Y_i\}$ as a max-stable process and keep all exceedances.

Max-stable processes

- Max-stable processes are natural approximations of processes over high thresholds;
  - Similar to the GEV in the i.i.d. case;
  - No unified parametric representation for max-stable processes.
- A particular max-stable process: the Gaussian extreme value process;
  - Assume that $\{Y_t\}$ has GEV marginal distribution distribution with parameter (μ, σ, ξ).
  - Let

$$Z_t = -\frac{1}{\log(F(Y_t; \mu, \sigma, \xi))},$$

  $\{Z_t\}$ has unit Fréchet marginal distribution. Assume that

$$Z_t = \max\left\{ \frac{\zeta_i}{\nu\sqrt{2\pi}} \exp\left(-\frac{(s_i - t)^2}{2\nu^2} \right) \right\},$$

  with $\{(\zeta_i, s_i), i \geq 1\}$ the points of a Poisson process with intensity $\zeta^{-2}d\zeta \times ds$.
  - The full joint distribution is not tractable but bivariate distributions are

$$P(Y_{t_1} \leq y_{t_1}, Y_{t_2} \leq y_{t_2}) = \exp\left[-\frac{1}{z_{t_1}} \Phi\left(\frac{a}{2} + \frac{1}{a}\log\frac{z_{t_2}}{z_{t_1}} \right) - \frac{1}{z_{t_2}} \Phi\left(\frac{a}{2} + \frac{1}{a}\log\frac{z_{t_1}}{z_{t_2}} \right) \right]$$

  with $a = \frac{|t_2 - t_1|}{\nu}$, Φ the cdf of the standard normal distribution and $z_{t_i} = \frac{-1}{\log F(y_{t_i}; \mu, \sigma, \xi)}$
  - Continuous time process ... permits to deal with irregular time sampling

Max-stable processes

- Max-stable processes are natural approximations of processes over high thresholds;
  - Similar to the GEV in the i.i.d. case;
  - No unified parametric representation for max-stable processes.
- A particular max-stable process: the Gaussian extreme value process;
  - Assume that $\{Y_t\}$ has GEV marginal distribution distribution with parameter (μ, σ, ξ).
  - Let
    $$Z_t = -\frac{1}{\log(F(Y_t; \mu, \sigma, \xi))},$$
    $\{Z_t\}$ has unit Fréchet marginal distribution. Assume that
    $$Z_t = \max\left\{\frac{\zeta_i}{\nu\sqrt{2\pi}}\exp\left(-\frac{(s_i - t)^2}{2\nu^2}\right)\right\},$$
    with $\{(\zeta_i, s_i), i \geq 1\}$ the points of a Poisson process with intensity $\zeta^{-2}d\zeta \times ds$.
  - The full joint distribution is not tractable but bivariate distributions are
    $$P(Y_{t_1} \leq y_{t_1}, Y_{t_2} \leq y_{t_2}) = \exp\left[-\frac{1}{z_{t_1}}\Phi\left(\frac{a}{2} + \frac{1}{a}\log\frac{z_{t_2}}{z_{t_1}}\right) - \frac{1}{z_{t_2}}\Phi\left(\frac{a}{2} + \frac{1}{a}\log\frac{z_{t_1}}{z_{t_2}}\right)\right]$$
    with $a = \frac{|t_1 - t_2|}{\nu}$, Φ the cdf of the standard normal distribution and $z_{t_i} = \frac{-1}{\log F(y_{t_i}; \mu, \sigma, \xi)}$.
  - Continuous time process... permits to deal with irregular time sampling
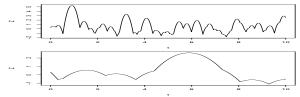
Max-stable processes

- Max-stable processes are natural approximations of processes over high thresholds;
  - Similar to the GEV in the i.i.d. case;
  - No unified parametric representation for max-stable processes.
- A particular max-stable process: the Gaussian extreme value process;
  - Assume that $\{Y_t\}$ has GEV marginal distribution distribution with parameter (μ, σ, ξ).
  - Let

$$Z_t = -\frac{1}{\log(F(Y_t; \mu, \sigma, \xi))},$$

  $\{Z_t\}$ has unit Fréchet marginal distribution. Assume that

$$Z_t = \max\left\{ \frac{\zeta_i}{\nu\sqrt{2\pi}} \exp\left(-\frac{(s_i - t)^2}{2\nu^2}\right) \right\},$$

  with $\{(\zeta_i, s_i), i \geq 1\}$ the points of a Poisson process with intensity $\zeta^{-2}d\zeta \times ds$.

- $\nu \to 0$ [resp. $\nu \to +\infty$] corresponds to independence [resp. perfect dependence]
- The full joint distribution is not tractable but bivariate distributions are

$$P(Y_{t_1} \leq y_{t_1}, Y_{t_2} \leq y_{t_2}) = \exp\left[-\frac{1}{z_{t_1}}\Phi\left(\frac{a}{2} + \frac{1}{a}\log\frac{z_{t_2}}{z_{t_1}}\right) - \frac{1}{z_{t_2}}\Phi\left(\frac{a}{2} + \frac{1}{a}\log\frac{z_{t_1}}{z_{t_2}}\right)\right]$$

Max-stable processes

- Max-stable processes are natural approximations of processes over high thresholds;
  - Similar to the GEV in the i.i.d. case;
  - No unified parametric representation for max-stable processes.
- A particular max-stable process: the Gaussian extreme value process;
  - Assume that $\{Y_t\}$ has GEV marginal distribution distribution with parameter (μ, σ, ξ).
  - Let
  $$Z_t = -\frac{1}{\log(F(Y_t; \mu, \sigma, \xi))},$$
  $\{Z_t\}$ has unit Fréchet marginal distribution. Assume that
  $$Z_t = \max \left\{ \frac{\zeta_i}{\nu \sqrt{2\pi}} \exp \left(-\frac{(s_i - t)^2}{2\nu^2} \right) \right\},$$
  with $\{(\zeta_i, s_i), i \geq 1\}$ the points of a Poisson process with intensity $\zeta^{-2} d\zeta \times ds$.
  - The full joint distribution is not tractable but bivariate distributions are
  $$P(Y_{t_1} \leq y_{t_1}, Y_{t_2} \leq y_{t_2}) = \exp \left[-\frac{1}{z_{t_1}} \Phi \left(\frac{a}{2} + \frac{1}{a} \log \frac{z_{t_2}}{z_{t_1}} \right) - \frac{1}{z_{t_2}} \Phi \left(\frac{a}{2} + \frac{1}{a} \log \frac{z_{t_1}}{z_{t_2}} \right) \right]$$
  with $a = \frac{|t_1 - t_2|}{\nu}$, Φ the cdf of the standard normal distribution and $z_{t_i} = \frac{-1}{\log F(y_{t_i}; \mu, \sigma, \xi)}$.
  - Continuous time process... permits to deal with irregular time sampling

Max-stable processes

- Max-stable processes are natural approximations of processes over high thresholds;
  - Similar to the GEV in the i.i.d. case;
  - No unified parametric representation for max-stable processes.
- A particular max-stable process: the Gaussian extreme value process;
  - Assume that $\{Y_t\}$ has GEV marginal distribution distribution with parameter (μ, σ, ξ).
  - Let
  $$Z_t = -\frac{1}{\log(F(Y_t; \mu, \sigma, \xi))},$$
  $\{Z_t\}$ has unit Fréchet marginal distribution. Assume that
  $$Z_t = \max\left\{\frac{\zeta_i}{\nu\sqrt{2\pi}} \exp\left(-\frac{(s_i - t)^2}{2\nu^2}\right)\right\},$$
  with $\{(\zeta_i, s_i), i \geq 1\}$ the points of a Poisson process with intensity $\zeta^{-2}d\zeta \times ds$.
  - The full joint distribution is not tractable but bivariate distributions are
  $$P(Y_{t_1} \leq y_{t_1}, Y_{t_2} \leq y_{t_2}) = \exp\left[-\frac{1}{z_{t_1}}\Phi\left(\frac{a}{2} + \frac{1}{a}\log\frac{z_{t_2}}{z_{t_1}}\right) - \frac{1}{z_{t_2}}\Phi\left(\frac{a}{2} + \frac{1}{a}\log\frac{z_{t_1}}{z_{t_2}}\right)\right]$$
  with $a = \frac{|t_1 - t_2|}{\nu}$, Φ the cdf of the standard normal distribution and $z_{t_i} = \frac{-1}{\log F(y_{t_i}; \mu, \sigma, \xi)}$.
  - Continuous time process... permits to deal with irregular time sampling

Proposed methodology

- Let $\{X_t\}_{t \in \{t_1, ..., t_n\}}$ be a stationary process observed at time $(t_1, ..., t_n)$
- Chose a high threshold u
- Model the process censored at u as a censored max-stable process

$$u\mathbf{1}[X_{t_i} \leq u] + X_{t_i}\mathbf{1}[X_{t_i} > u] = u\mathbf{1}[Y_{t_i} \leq u] + Y_{t_i}\mathbf{1}[Y_{t_i} > u]$$

  with $\{Y_t\}$ a Gaussian extreme value process with parameter $\theta = (\mu, \sigma, \xi, \nu)$
  - Process with marginal distribution $GEV(\mu, \sigma, \xi)$, ν describes the time structure
- Estimate θ by maximizing the pairwise likelihood function

$$PL(\theta; \tilde{y}_{t_1}, ..., \tilde{y}_{t_n}) = \prod_{i=1}^{n-1} p_{\tilde{Y}}(\tilde{y}_{t_i}, \tilde{y}_{t_{i+1}}; \theta)$$

  of the censored sequence $(\tilde{y}_{t_1}, ... \tilde{y}_{t_n})$ with $\tilde{y}_t = u\mathbf{1}[y_t \leq u] + y_t\mathbf{1}[y_t > u]$
  - Consistent estimator
  - Better results when using only closest neighbors in the pairwise likelihood function
- Use simulations of the fitted model to estimate quantities of interest
  - Distribution of the estimator (parametric bootstrap)
  - Return values, length of the storms, number of storm per year,...

Proposed methodology

- Let $\{X_t\}_{t \in \{t_1, ..., t_n\}}$ be a stationary process observed at time $(t_1, ..., t_n)$
- Chose a high threshold u
- Model the process censored at u as a censored max-stable process

$$u\mathbf{1}[X_{t_i} \leq u] + X_{t_i}\mathbf{1}[X_{t_i} > u] = u\mathbf{1}[Y_{t_i} \leq u] + Y_{t_i}\mathbf{1}[Y_{t_i} > u]$$

  with $\{Y_t\}$ a Gaussian extreme value process with parameter $\theta = (\mu, \sigma, \xi, \nu)$
    - Process with marginal distribution $GEV(\mu, \sigma, \xi)$, ν describes the time structure
- Estimate θ by maximizing the pairwise likelihood function

$$PL(\theta; \tilde{y}_{t_1}, ..., \tilde{y}_{t_n}) = \prod_{i=1}^{n-1} p_{\tilde{Y}}(\tilde{y}_{t_i}, \tilde{y}_{t_{i+1}}; \theta)$$

  of the censored sequence $(\tilde{y}_{t_1}, ... \tilde{y}_{t_n})$ with $\tilde{y}_t = u\mathbf{1}[y_t \leq u] + y_t\mathbf{1}[y_t > u]$
    - Consistent estimator
    - Better results when using only closest neighbors in the pairwise likelihood function
- Use simulations of the fitted model to estimate quantities of interest
    - Distribution of the estimator (parametric bootstrap)
    - Return values, length of the storms, number of storm per year,...

Proposed methodology

- Let $\{X_t\}_{t \in \{t_1, ..., t_n\}}$ be a stationary process observed at time $(t_1, ..., t_n)$
- Chose a high threshold u
- Model the process censored at u as a censored max-stable process

$$u\mathbf{1}[X_{t_i} \leq u] + X_{t_i}\mathbf{1}[X_{t_i} > u] = u\mathbf{1}[Y_{t_i} \leq u] + Y_{t_i}\mathbf{1}[Y_{t_i} > u]$$

with $\{Y_t\}$ a Gaussian extreme value process with parameter $\theta = (\mu, \sigma, \xi, \nu)$

  - Process with marginal distribution $GEV(\mu, \sigma, \xi)$, ν describes the time structure

- Estimate θ by maximizing the pairwise likelihood function

$$PL(\theta; \tilde{y}_{t_1}, ..., \tilde{y}_{t_n}) = \prod_{i=1}^{n-1} p_{\tilde{Y}}(\tilde{y}_{t_i}, \tilde{y}_{t_{i+1}}; \theta)$$

of the censored sequence $(\tilde{y}_{t_1}, ... \tilde{y}_{t_n})$ with $\tilde{y}_t = u\mathbf{1}[y_t \leq u] + y_t\mathbf{1}[y_t > u]$

  - Consistent estimator
  - Better results when using only closest neighbors in the pairwise likelihood function

- Use simulations of the fitted model to estimate quantities of interest

  - Distribution of the estimator (parametric bootstrap)
  - Return values, length of the storms, number of storm per year,...

Proposed methodology

- Let $\{X_t\}_{t \in \{t_1, ..., t_n\}}$ be a stationary process observed at time $(t_1, ..., t_n)$
- Chose a high threshold u
- Model the process censored at u as a censored max-stable process

$$u\mathbf{1}[X_{t_i} \leq u] + X_{t_i}\mathbf{1}[X_{t_i} > u] = u\mathbf{1}[Y_{t_i} \leq u] + Y_{t_i}\mathbf{1}[Y_{t_i} > u]$$

  with $\{Y_t\}$ a Gaussian extreme value process with parameter $\theta = (\mu, \sigma, \xi, \nu)$

  - Process with marginal distribution $GEV(\mu, \sigma, \xi)$, ν describes the time structure
- Estimate θ by maximizing the pairwise likelihood function

$$PL(\theta; \tilde{y}_{t_1}, ..., \tilde{y}_{t_n}) = \prod_{i=1}^{n-1} p_{\tilde{Y}}(\tilde{y}_{t_i}, \tilde{y}_{t_{i+1}}; \theta)$$

  of the censored sequence $(\tilde{y}_{t_1}, ... \tilde{y}_{t_n})$ with $\tilde{y}_t = u\mathbf{1}[y_t \leq u] + y_t\mathbf{1}[y_t > u]$
  - Consistent estimator
  - Better results when using only closest neighbors in the pairwise likelihood function
- Use simulations of the fitted model to estimate quantities of interest
  - Distribution of the estimator (parametric bootstrap)
  - Return values, length of the storms, number of storm per year,...

Proposed methodology

- Let $\{X_t\}_{t \in \{t_1, ..., t_n\}}$ be a stationary process observed at time $(t_1, ..., t_n)$
- Chose a high threshold u
- Model the process censored at u as a censored max-stable process

$$u\mathbf{1}[X_{t_i} \leq u] + X_{t_i}\mathbf{1}[X_{t_i} > u] = u\mathbf{1}[Y_{t_i} \leq u] + Y_{t_i}\mathbf{1}[Y_{t_i} > u]$$

  with $\{Y_t\}$ a Gaussian extreme value process with parameter $\theta = (\mu, \sigma, \xi, \nu)$

  - Process with marginal distribution $GEV(\mu, \sigma, \xi)$, ν describes the time structure
- Estimate θ by maximizing the pairwise likelihood function

$$PL(\theta; \tilde{y}_{t_1}, ..., \tilde{y}_{t_n}) = \prod_{i=1}^{n-1} p_{\tilde{Y}}(\tilde{y}_{t_i}, \tilde{y}_{t_{i+1}}; \theta)$$

  of the censored sequence $(\tilde{y}_{t_1}, ...\tilde{y}_{t_n})$ with $\tilde{y}_t = u\mathbf{1}[y_t \leq u] + y_t\mathbf{1}[y_t > u]$
  - Consistent estimator
  - Better results when using only closest neighbors in the pairwise likelihood function
- Use simulations of the fitted model to estimate quantities of interest
  - Distribution of the estimator (parametric bootstrap)
  - Return values, length of the storms, number of storm per year,...

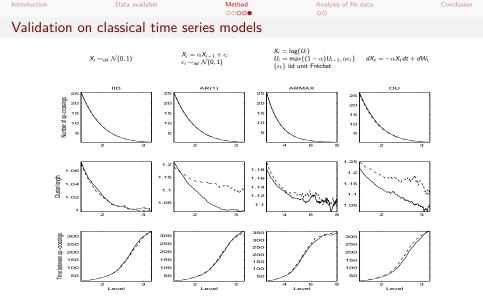## Validation on classical time series models

Figure : Comparison of the extremal behavior of classical processes (solid line) against the fitted Gaussian extreme value process (dashed line). Results obtained by simulating 1000 years of each model (one observation per day).
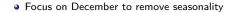
Plan of the talk

1. Data available

2. Method
   - Usual methods in the iid case
   - Generalization to dependent sequences

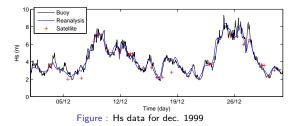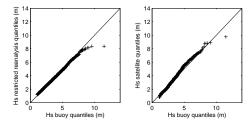3. Analysis of Hs data
   - Data available
   - A few results

4. Conclusion

Data analysis

- Focus on December to remove seasonality

Figure : Hs data for dec. 1999

- The quantiles of the satellite data are the closest to the ones of the buoy

A few results

- The model has been fitted to the different data sets
- It seems to provide a realistic extrapolation of the extremal properties of the data
  - Comparison of statistics for the buoy data and the fitted model

- Comparison of parameters and return values for the different data sets
- 20 years return levels in the North Atlantic computed using satellite data
  - Independent analysis at each location, no spatial information, $\beta = 0$

A few results

- The model has been fitted to the different data sets
- It seems to provide a realistic extrapolation of the extremal properties of the data
- Comparison of parameters and return values for the different data sets

	Reanalysis	Buoy	Satellite
	Threshold		
u (m)	6	8	6
No. obs $> u$	74	59	48
	Parameter values		
μ	2.19 [-6.85,3.80]	5.12 [-7.3,6.67]	3.80 [2.23,4.47]
σ	1.70 [0.81,8.83]	0.50 [0.11,7.23]	1.33 [0.86,2.69]
ξ	-0.17 [-0.6,-0.02]	0.07 [-0.40,0.32]	0.01 [-0.25,0.17]
ν	0.12 [0.07,0.18]	1e-3 [8e-4,2e-3]	0.05 [0.03,0.08]
	Return levels		
q_{10}	8.2[7.3,9.1]	10.4[9.3,12.2]	12.1[10.7,14.6]
q_{20}	8.5[7.5,9.6]	10.8[9.4,13.8]	12.9[11.1,16.4]
q_{50}	8.8[7.6,10.1]	11.4[9.6,16.0]	14.0[11.5,19.1]
q_{100}	9.0[7.7,10.6]	11.8[9.7,18.8]	14.7[11.7,21.6]

- Confidence intervals (computed using parametric bootstrap) are wide!
- Gumbel type distributions ($\xi = 0$) except for the reanalysis data ($\xi < 0$)
    - Lower tail for the reanalysis data?
- Higher value of ν for the reanalysis data (smoother, longer storms)
- Buoy and satellite lead to similar fitted models
- 20 years return levels in the North Atlantic computed using satellite data
    - Independent analysis at each location, no spatial information, $\xi = 0$.

A few results

- The model has been fitted to the different data sets
- It seems to provide a realistic extrapolation of the extremal properties of the data
- Comparison of parameters and return values for the different data sets
- 20 years return levels in the North Atlantic computed using satellite data
  - Independent analysis at each location, no spatial information, $\xi = 0$.

Plan of the talk

1. Data available

2. Method
   - Usual methods in the iid case
   - Generalization to dependent sequences

3. Analysis of Hs data
   - Data available
   - A few results

4. **Conclusion**

Conclusion

- The proposed methodology permits to analyze the extremal behavior of dependent sequences
- Still valid in presence of missing data or irregular time sampling
- Theoretical results show that the estimator is consistent
- The method has been successfully validated on classical time series models
- Similar models are identified on satellite and buoy data
- Longer storm and lighter tails are identified on reanalysis data
- Preprint available:
  Raillard N., Ailliot P., Yao J.F., Modeling extreme values of processes observed at irregular time steps. Application to significant wave height. *To appear in the Annals of Applied Statistics*.

N. Raillard, P. Ailliot, and J.F. Yao.
Modelling extreme values of processes observed at irregular time step. application to significant wave height.
Submitted, 2011.

R. L. Smith.
Max-stable processes and spatial extremes.
Unpublished, 1990.

Estimation des paramètres
Propriétés

Théorème

Soit $\hat{\nu} = \underset{CL_1(x_{t_1}, \ldots, x_{t_n}; \theta)}{\operatorname{argmax}}$. Si $\nu^* \in \Theta := [\nu_-, \ \nu_+]$ où $0 < \nu_- < \nu_+$, alors $\hat{\nu}$ est un

estimateur convergeant de ν^*.

Idées de la preuve: Estimation en deux étapes.

1. Estimation des paramètres marginaux ;
2. Estimation du paramètre de dépendance.

Estimation des paramètres
Propriétés

Théorème

Soit $\hat{\nu} = \underset{CL_1(x_{t_1},\ldots,x_{t_n};\theta)}{\text{argmax}}$. Si $\nu^* \in \Theta := [\nu_-,\ \nu_+]$ où $0 < \nu_- < \nu_+$, alors $\hat{\nu}$ est un estimateur convergeant de ν^*.

Idées de la preuve: Estimation en deux étapes.

1. Estimation des paramètres marginaux ;
2. Estimation du paramètre de dépendance.

Estimation des paramètres
Propriétés

Théorème

Soit $\hat{\nu} = \underset{CL_1(x_{t_1},\ldots,x_{t_n};\theta)}{\arg\max}$. Si $\nu^* \in \Theta := [\nu_-, \ \nu_+]$ où $0 < \nu_- < \nu_+$, alors $\hat{\nu}$ est un estimateur convergeant de ν^*.

Idées de la preuve: Estimation en deux étapes.

1. Estimation des paramètres marginaux ;
2. Estimation du paramètre de dépendance.

Estimation des paramètres
Propriétés

Théorème

Soit $\hat{\nu} = \underset{CL_1(x_{t_1}, \ldots, x_{t_n}; \theta)}{\text{argmax}}$. Si $\nu^* \in \Theta := [\nu_-, \ \nu_+]$ où $0 < \nu_- < \nu_+$, alors $\hat{\nu}$ est un estimateur convergeant de ν^*.

Idées de la preuve: Estimation en deux étapes.

1. Estimation des paramètres marginaux ;
2. Estimation du paramètre de dépendance.

Estimation des paramètres
Propriétés

Théorème de Pfzangal

Soit $\{Z_i\}_{i=1,\ldots,n}$ un processus stationnaire ergodique dont la loi dépend d'un paramètre $\nu^* \in \Theta$ avec Θ un compact de $\mathbb{R}$ et soit Q_n un contraste tel que
$Q_n(\nu) = \frac{1}{n} \sum_{i=1}^{n-1} f(Z_i, Z_{i+1}; \nu)$, où f est une fonction mesurable continue en ν. Si

1. $E \inf_{\nu \in \Theta} f(Z_1, Z_2; \nu) > -\infty$;

2. $\nu \mapsto Ef(Z_1, Z_2; \nu)$ possède un unique minimum fini en ν^*.

alors l'estimateur au minimum de contraste $\hat{\nu}_n = \arg\min_{\nu \in \Theta} Q_n(\nu)$ est fortement consistant:
$$\lim_{n \to \infty} \hat{\nu}_n = \nu^* \ p.s.$$

Estimation des paramètres
Propriétés

Un processus Z extrémal gaussien, est ergodique (Stoev, 2008).

Définitions

- $F_Z(z_1, z_2; \nu) := \mathbb{P}(Z_1 < z_1, Z_2 < z_2) = \exp\{-V(z_1, z_2; \nu)\}$
- $V(z_1, z_2; \nu) := \dfrac{\Phi(a/2 + 1/a \log \frac{z_2}{z_1})}{z_1} + \dfrac{\Phi(a/2 + 1/a \log \frac{z_1}{z_2})}{z_2}$
- $p(Z_1, Z_2; \nu) := \dfrac{\partial^2}{\partial z_1 \partial z_2} F_Z(z_1, z_2; \nu)$;
- $f(Z_1, Z_2; \nu) := -\log p(Z_1, Z_2; \nu)$;

Expression du contraste

$f(z_1, z_2; \nu) = V(z_1, z_2; \nu) - \log\left[\dfrac{\Phi(w)\Phi(v)}{a^2 z_1 z_2} + \dfrac{\varphi(w)}{a z_1^2 z_2}\right]$

avec

- $w := w(z_1, z_2; \nu) := a/2 + 1/a \log \dfrac{z_2}{z_1}$
- $v := v(z_1, z_2; \nu) := a/2 - w$

Estimation des paramètres
Propriétés

Preuve de (1) (Minoration)

On a d'après le slide précédent: $f(z_1, z_2; \nu) = V(z_1, z_2; \nu) - \log\left[\frac{\Phi(w)\Phi(v)}{a^2 z_1 z_2} + \frac{\varphi(w)}{a z_1^2 z_2}\right]$.

- On a grâce à la borne de Fréchet (1951): $\inf_\nu \{V(z_1, z_2; \nu)\} \geq \min\left(-\frac{1}{z_1}, -\frac{1}{z_2}\right)$;

- Si $\nu \in [\nu_-, \nu_+]$, $\log\left[\frac{\Phi(w)\Phi(v)}{a^2 z_1 z_2} + \frac{\varphi(w)}{a z_1^2 z_2}\right] \geq 1 - \frac{\nu_+^2}{z_1 z_2}$

Si $Z \sim$ Fréchet(1), $1/Z \sim \mathcal{E}(1)$, donc $E\frac{1}{Z} = 1 < \infty$;

Donc $E \inf_{\nu \in \Theta} f(Z_1, Z_2; \nu) > -\infty$

Estimation des paramètres
Propriétés

Preuve de (2) (Identifiabilité)

Distance de Kullback-Leibler: $K(P_{\nu^*}, P_\nu) := E_{\nu^*}\left[-\log \frac{p(Z_1, Z_2; \nu)}{p(Z_1, Z_2; \nu^*)}\right]$

- $K = 0 \Leftrightarrow \forall z_1, z_2 : p(z_1, z_2; \nu^*) = p(z_1, z_2; \nu)$
- En particulier: pour $z_1 = z_2 = z$, on a:

$$\exp\left[-\frac{2}{z}\ \Phi\left(\frac{1}{2\nu}\right)\right]\left[\frac{\Phi\left(\frac{1}{2\nu}\right)^2}{z^2}\nu^2 + \frac{\varphi\left(\frac{1}{2\nu}\right)}{z^3}\nu\right]$$

$$= \exp\left[-\frac{2}{z}\Phi\left(\frac{1}{2\nu^*}\right)\right]\left[\frac{\Phi\left(\frac{1}{2\nu^*}\right)^2}{z^2}\nu^{*2} + \frac{\varphi\left(\frac{1}{2\nu^*}\right)}{z^3}\nu^*\right]$$

- Quand $z \to 0$: $\Phi\left(\frac{1}{\nu}\right) = \Phi\left(\frac{1}{\nu^*}\right)$