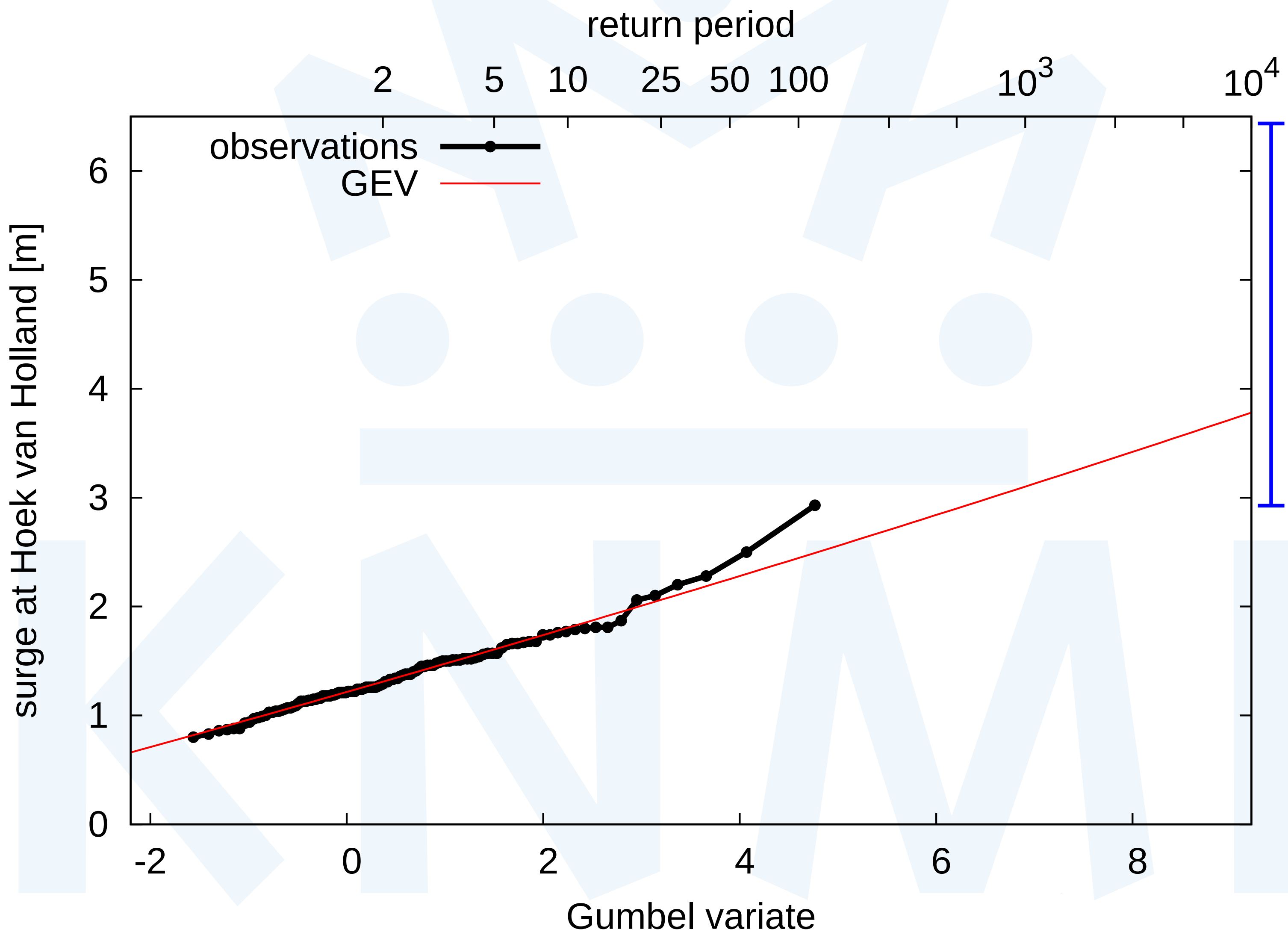
The statistical distribution of meteorological outliers

SEAMOCS - 23 October 2008

Henk van den Brink

KNMI

.... **Extrapolation:**



.... Use Regional Frequency Analysis

If we have n (independent) stations with a total of 10^4 years:

- 1 stations is expected to have a 10000-year event
- 2 stations are expected to have a 5000-year event
- 10 stations are expected to have a 1000-year event
- ...

Mathematical derivation (1)

Start with probability integral transform theorem:

$$\Pr(F(y) \le x) = x \qquad 0 \le x \le 1$$

Then:

$$\Pr(F(y) \le G(x)) = G(x)$$
$$\Pr(G^{-1}(F(y)) \le x) = G(x)$$

Holds for every F, every G!

.... Mathematical derivation (2)

n indepedent values $y_1 \leq y_2 \leq \ldots \leq y_n$ from $F(y)$:

$$F(y_n) = \Pr(y_n \leq y) = \Pr(y_1 \leq y) \cdots \Pr(y_n \leq y) = [F(y)]^n$$

$$\begin{aligned} \Pr(G^{-1}(F(y_n)) \leq x) &= \\ \Pr(G^{-1}(F^n(y)) \leq x) &= \\ (\Pr(G^{-1}(F(y)) \leq x))^n &= [G(x)]^n \end{aligned}$$

Holds for every F, every G, every n!

.... **Mathematical derivation (3)**

choose $G(x)$ to be the normalized Gumbel distribution:

$$\begin{aligned} G(x) &= e^{-e^{-x}} \\ \text{Then: } [G(x)]^n &= G(x - \ln(n)) \end{aligned}$$

and thus:

$$\begin{aligned} \Pr(-\ln(-\ln(F(y_n)) \leq x) &= G(x - \ln(n)) \\ \Pr(-\ln(-\ln(F(y_n)) - \ln(n) \leq x) &= G(x) \end{aligned}$$

Holds for every F, every n!

....

Mathematical derivation (4)

$$\Pr(-\ln(-\ln(F(y_n)) - \ln(n) \le x) \;=\; G(x)$$

define ΔX_n as:

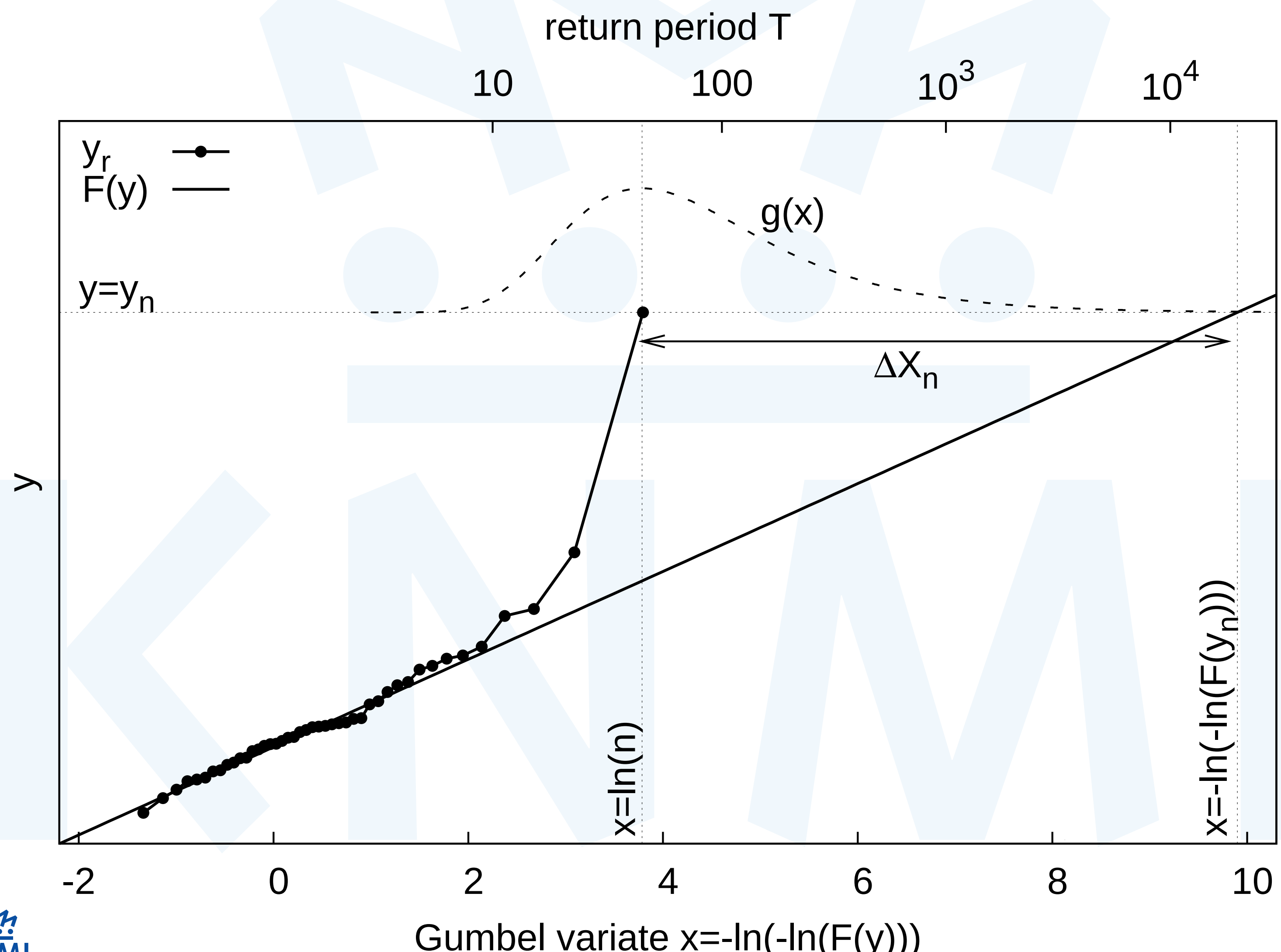
$$\Delta X_n = -\ln(-\ln(F(y_n))) - \ln(n)$$

then we have for the distribution of ΔX_n:

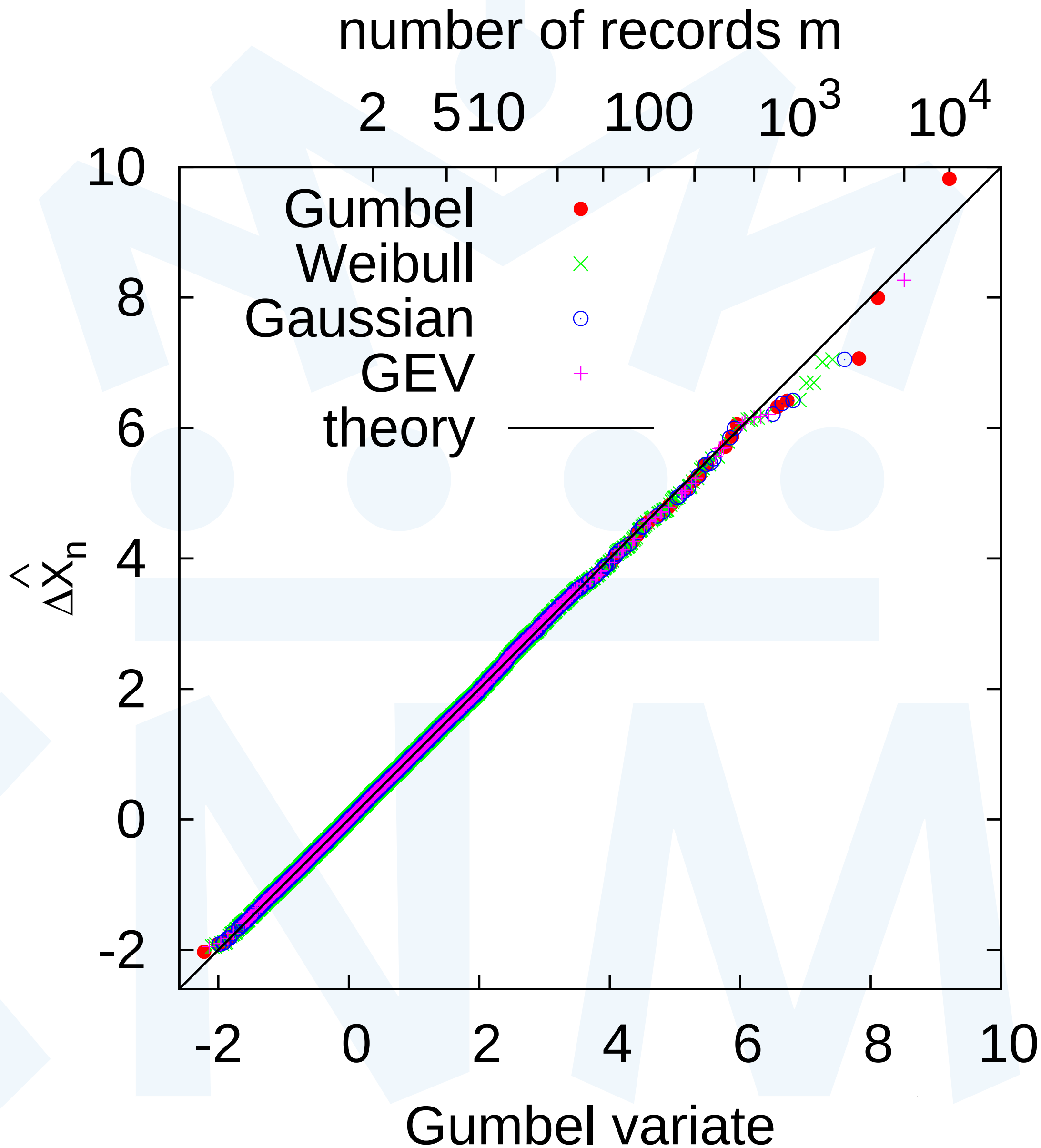
$$\Pr(\Delta X_n \le x) \;=\; G(x)$$

.... **Visualization on Gumbelplot**

$$\Delta X_n = -\ln(-\ln(F(y_n))) - \ln(n)$$



.... # Check with random generator



.... **Summary:**

Result is:

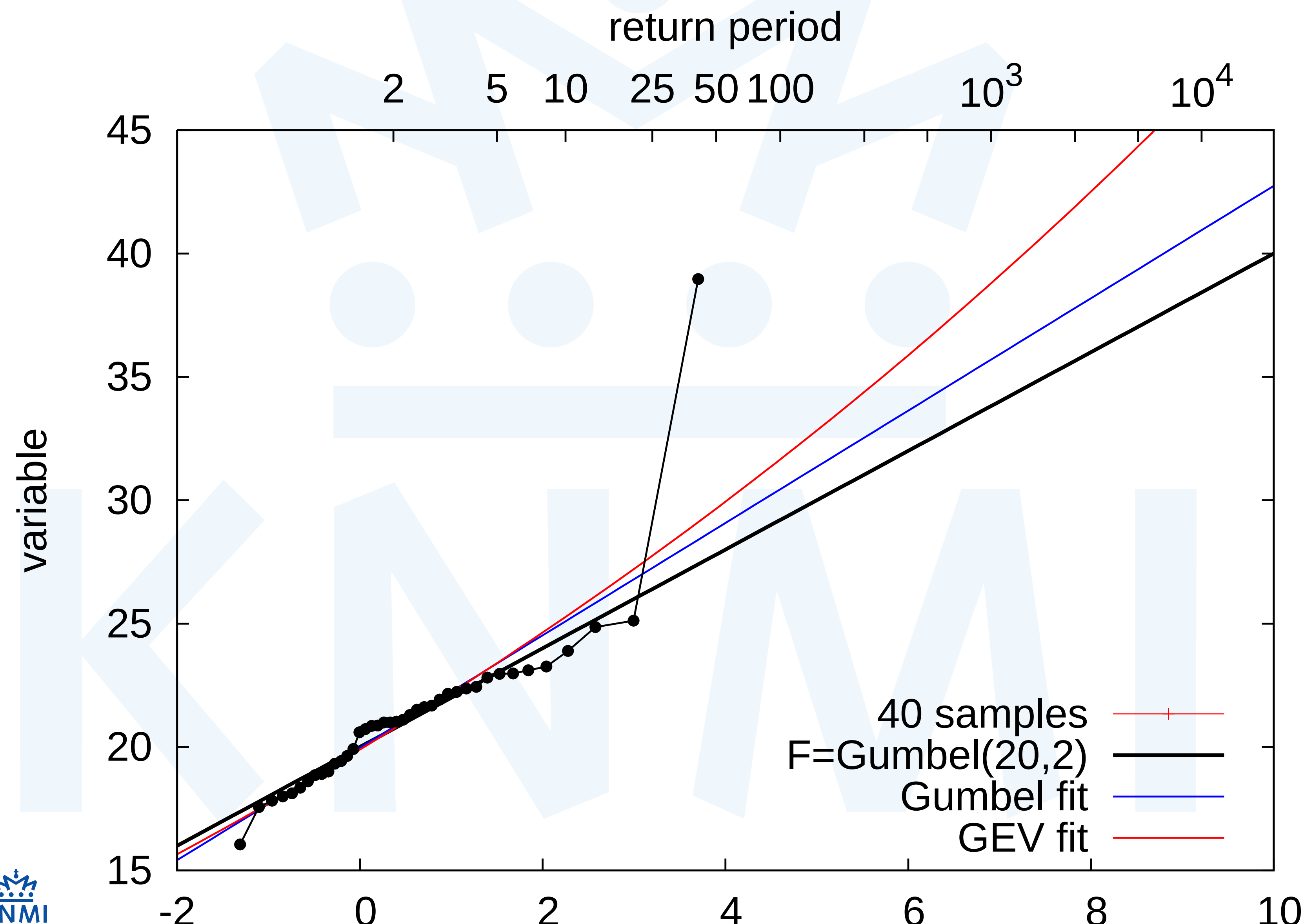
- independent of *known* F
- independent of *known* parameters of F
- independent of *known* n

.... **In practice:**

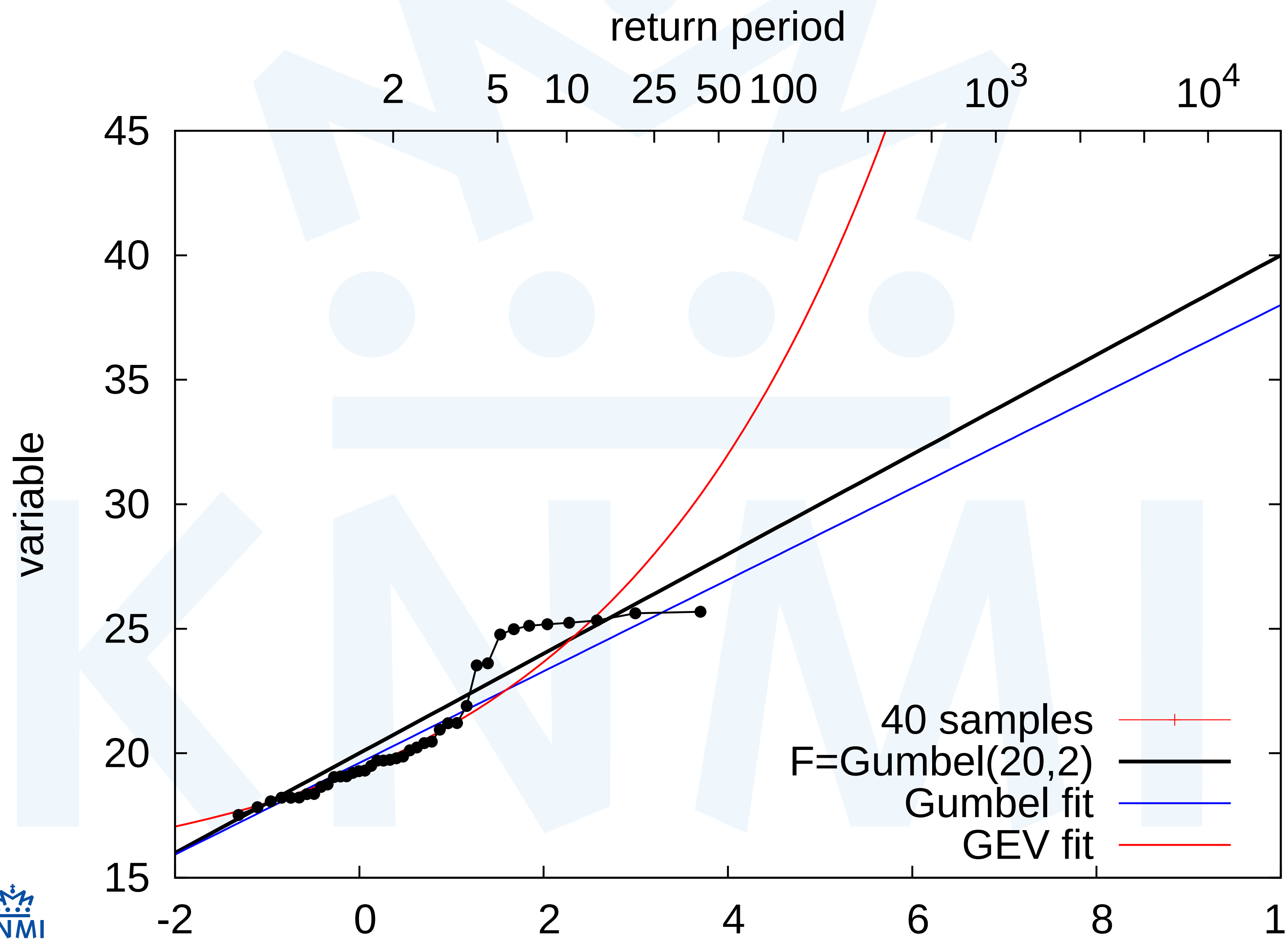
- F *unknown* $\longrightarrow$ assume $\tilde{F} = F$
- parameters $\boldsymbol{\theta}$ of F *unknown* $\longrightarrow$ fit $\tilde{F}_{\hat{\boldsymbol{\theta}}}$
  - sampling errors
  - highest value y_n influences $\hat{\boldsymbol{\theta}}$

.... **Example 1**

- F=Gumbel,$\tilde{F}$=GEV or Gumbel



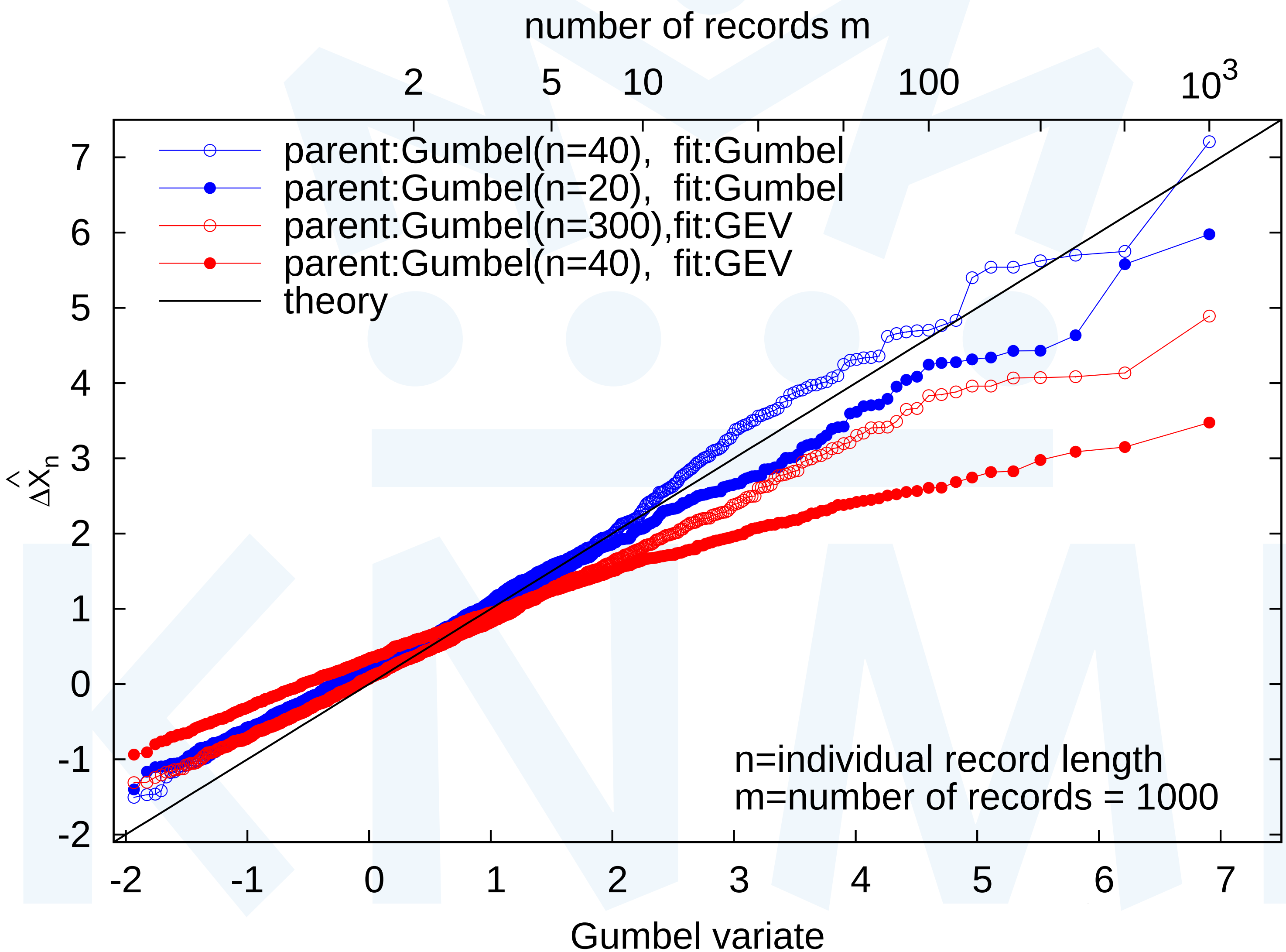
.... **Example 2**



return period
2
5
10
25
50
100
10^3
10^4
45
40
35
30
25
20
15
variable
-2
0
2
4
6
8
10
40 samples
F=Gumbel(20,2)
Gumbel fit
GEV fit

…. Result:

- F=Gumbel,$\tilde{F}$=GEV or Gumbel



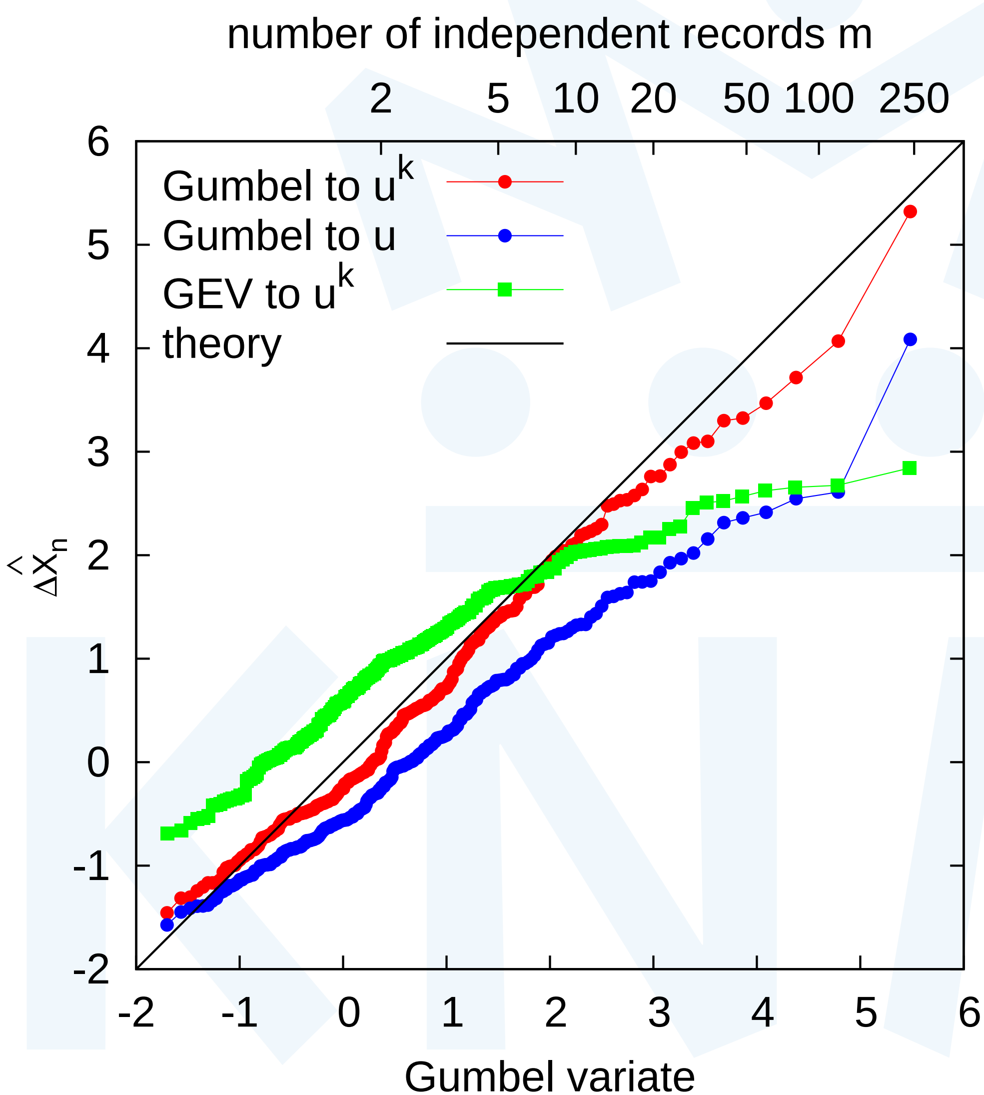
....

Application to ERA40-wind (1)

- 1958-2001 $\longrightarrow n = 44$ annual maxima
- North-Atlantic area (60W-10E,40N-70N)
- 1 degree resolution: 70×30=2100 grid points
- Determine distribution of $\Delta\hat{X}_n$
- require independence of outliers: 3-day interval (240 values of $\Delta\hat{X}_n$ left)

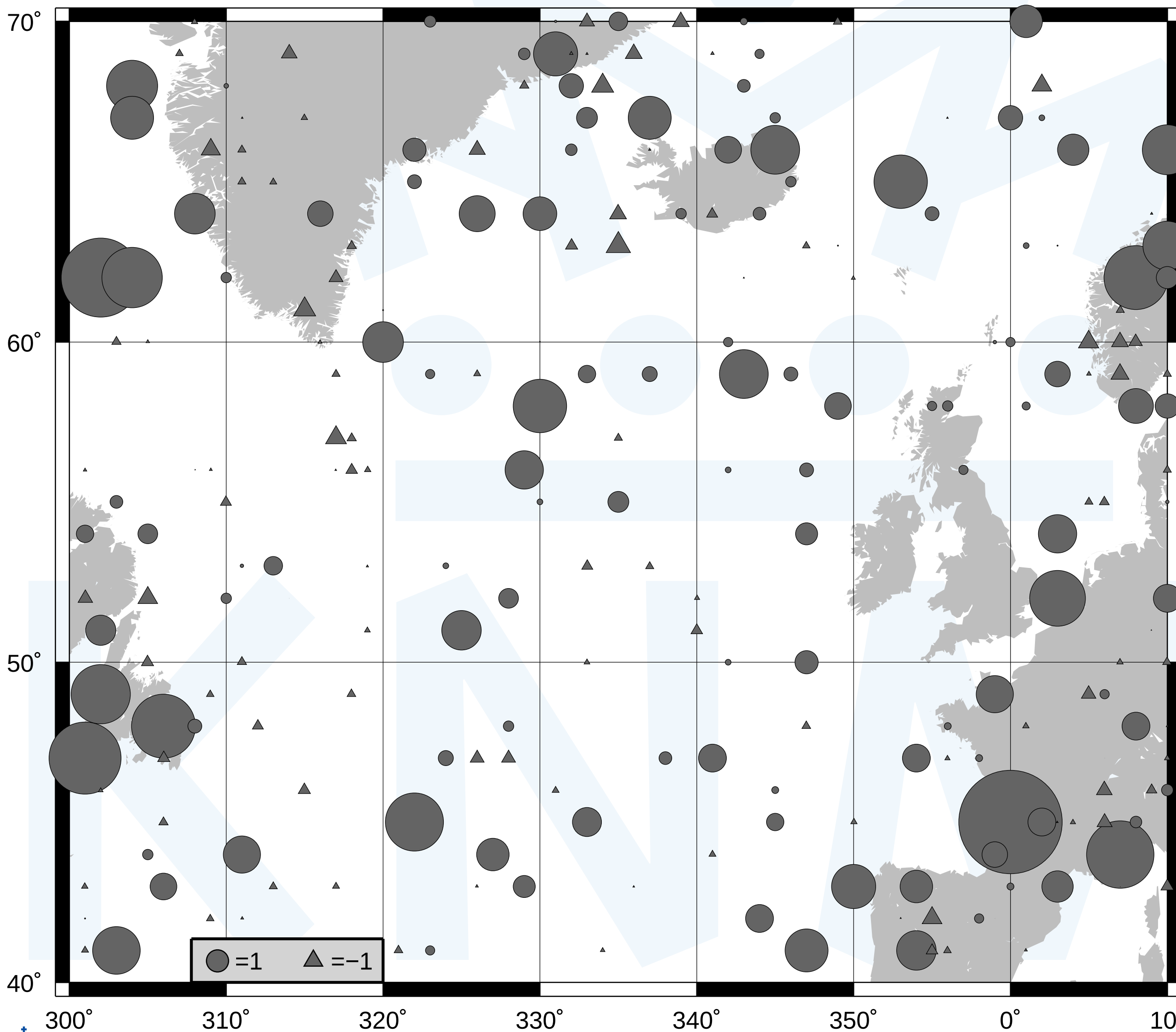
....

Application to ERA40-wind (2)



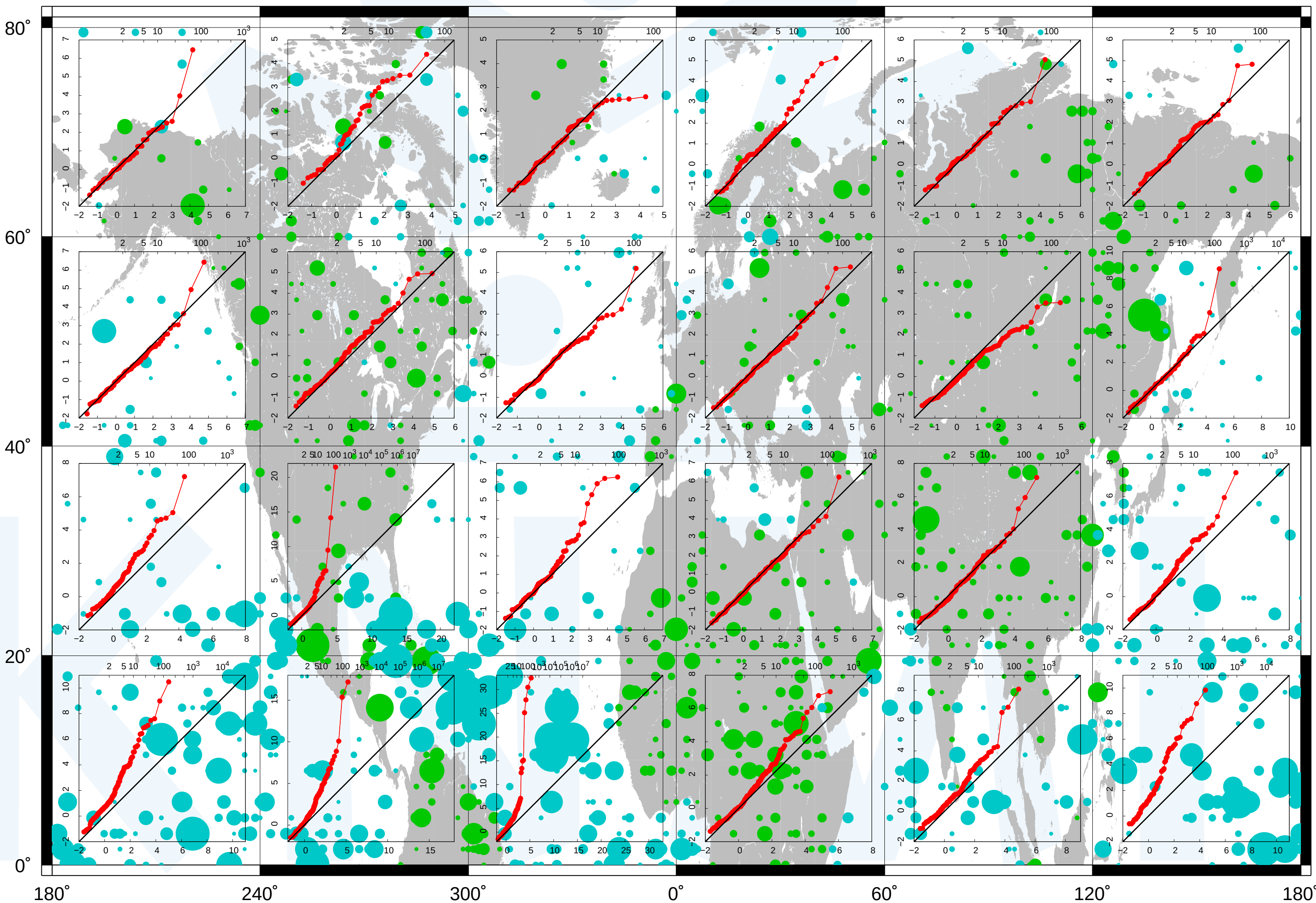
....

Application to ERA40-wind (3)



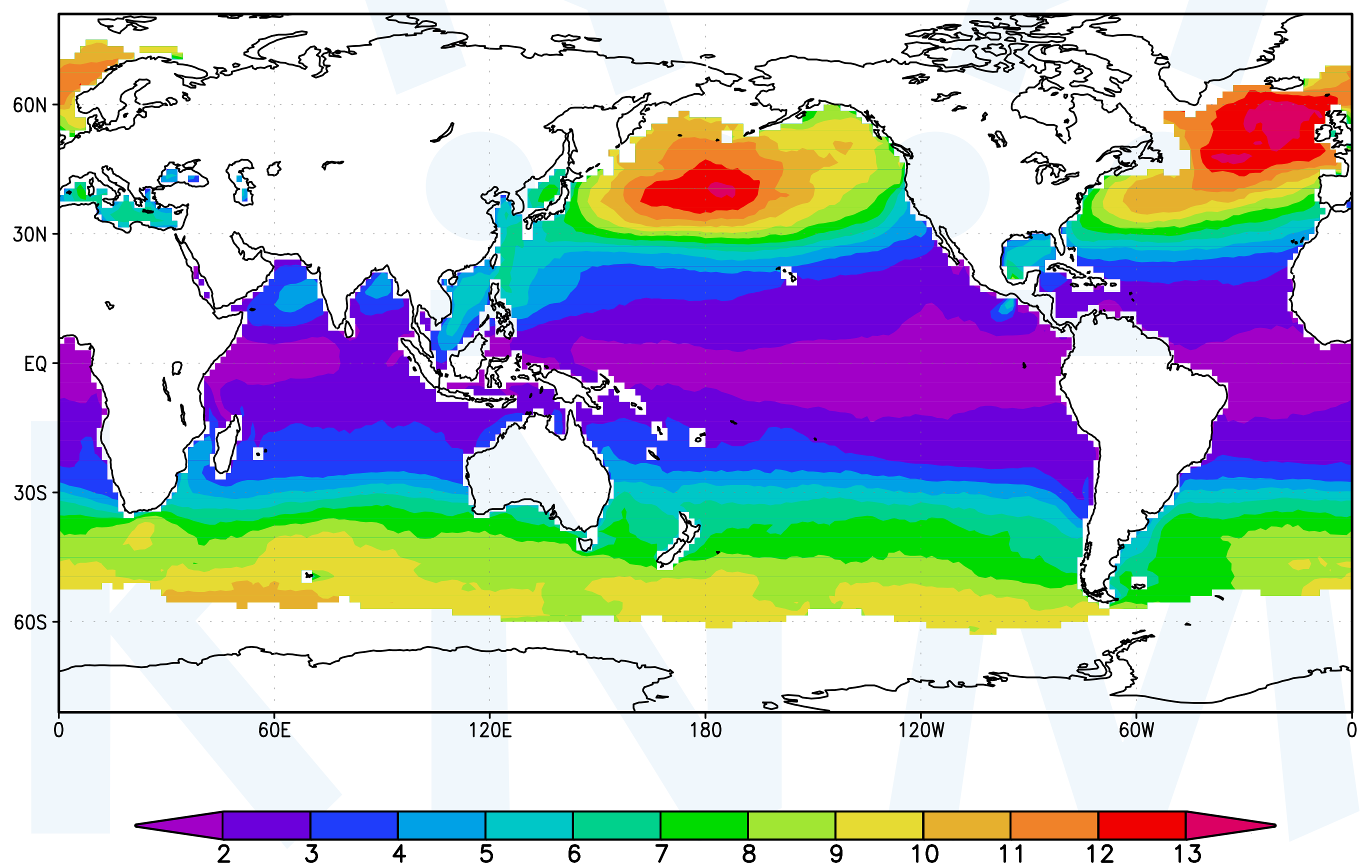
.... # Application to ERA40-wind (4)

Gumbel ERA40 45yr k=k



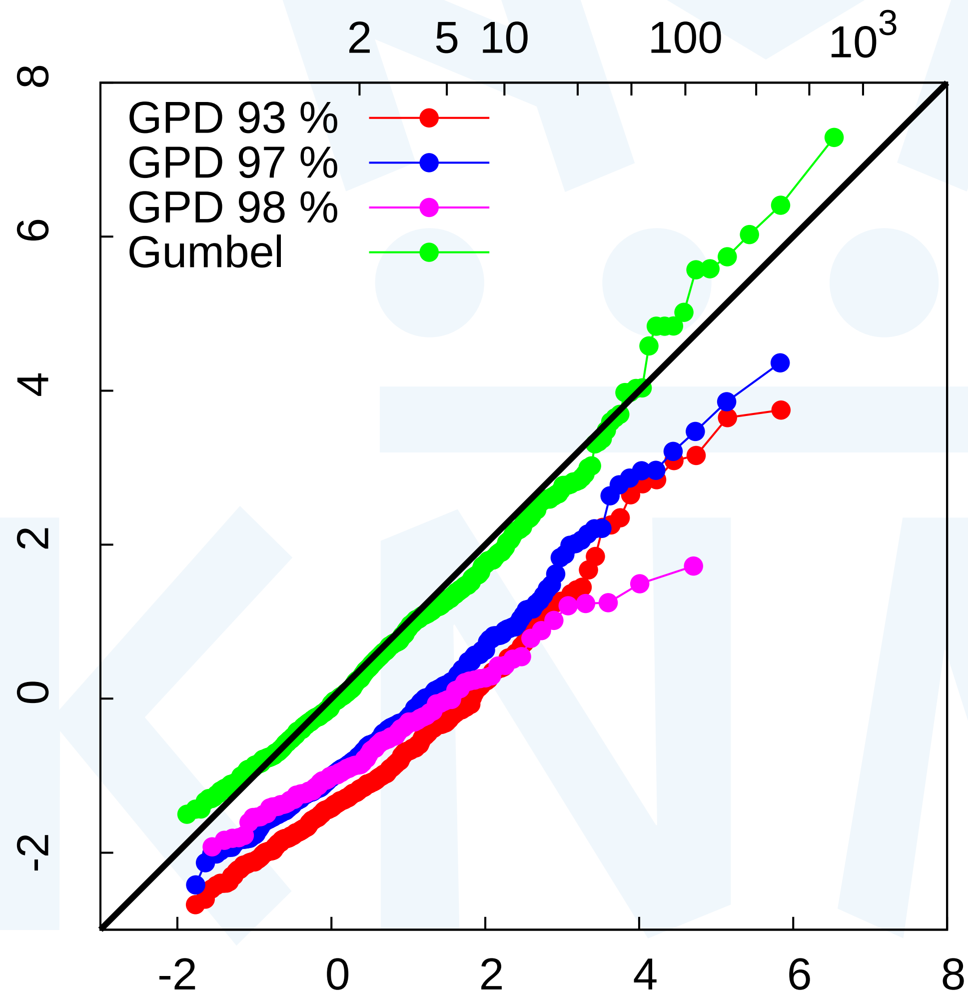
Application to waves (1)

100-year value from ERA40 (Caires and Sterl 2005):



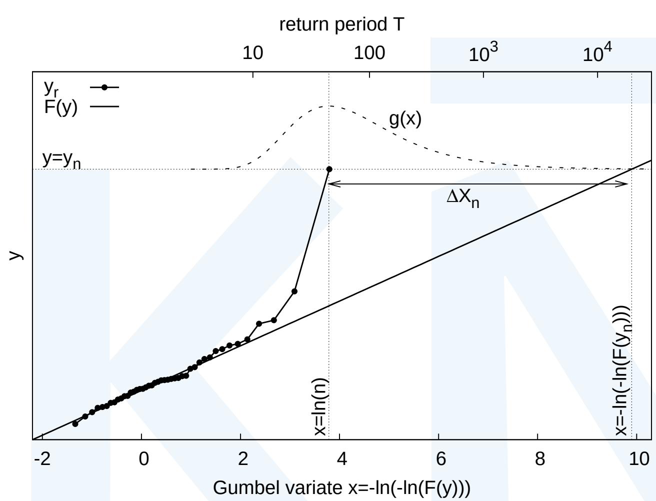
.... # Application to waves (2)

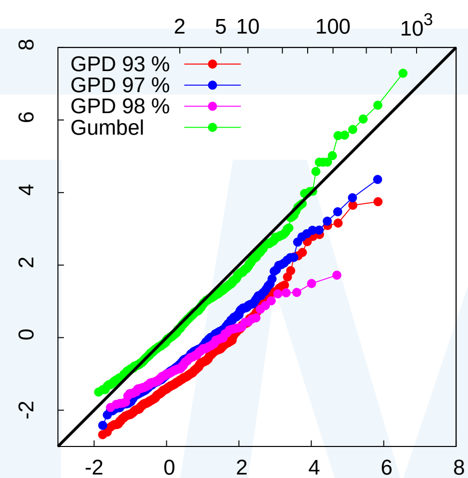
Distribution of $\Delta\hat{X}_n$:



.... **Application to waves (3)**

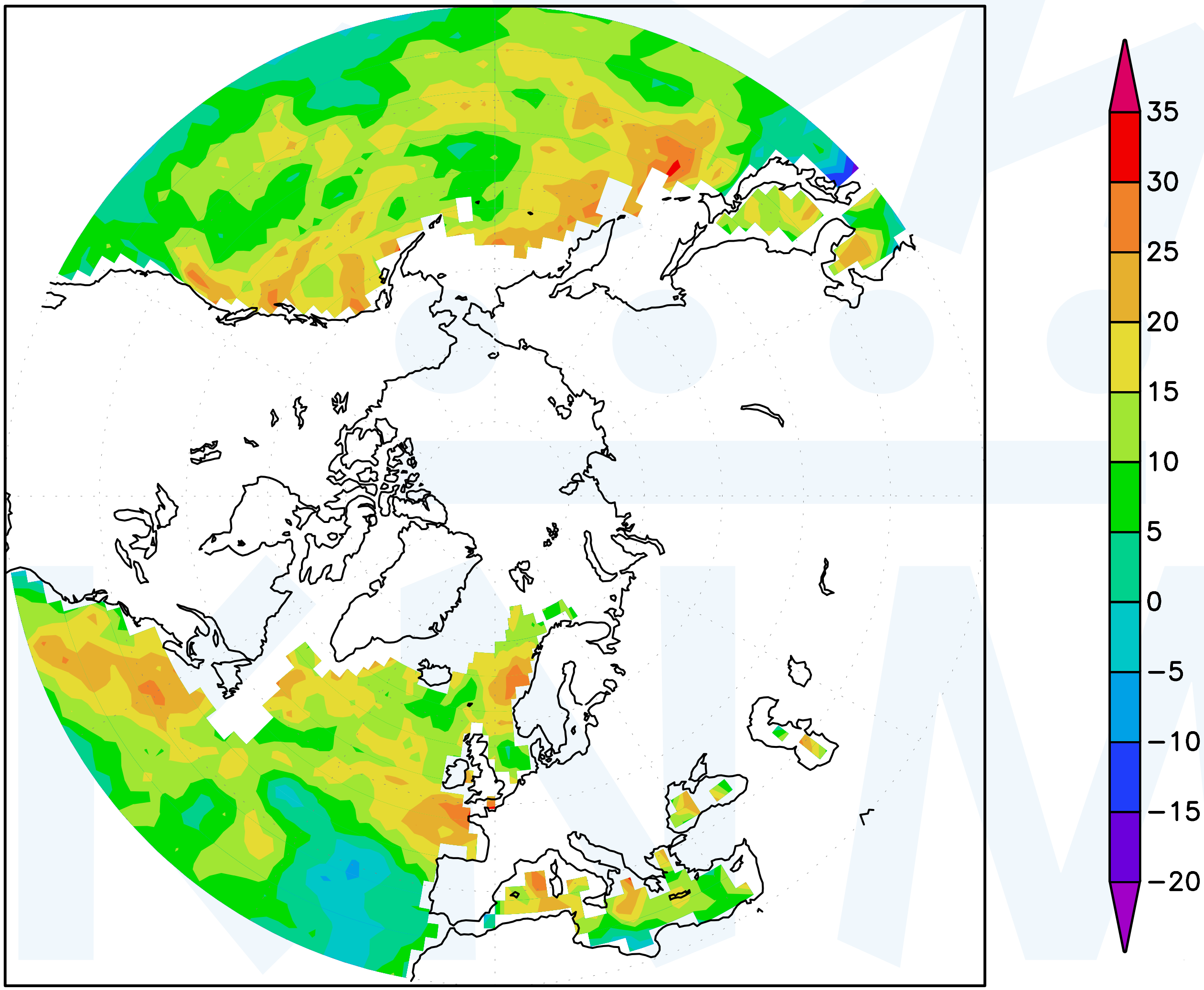
- 100-year value should have $\Delta \hat{X}_n = \ln(100) - \ln(44) = 0.82$
- previous graph gives 2.57
- belongs to return period of $\exp(\ln(44) + 2.57) = 575$ years!





.... Application to waves (4)

effect on Hs_{100} = 15% overestimation!

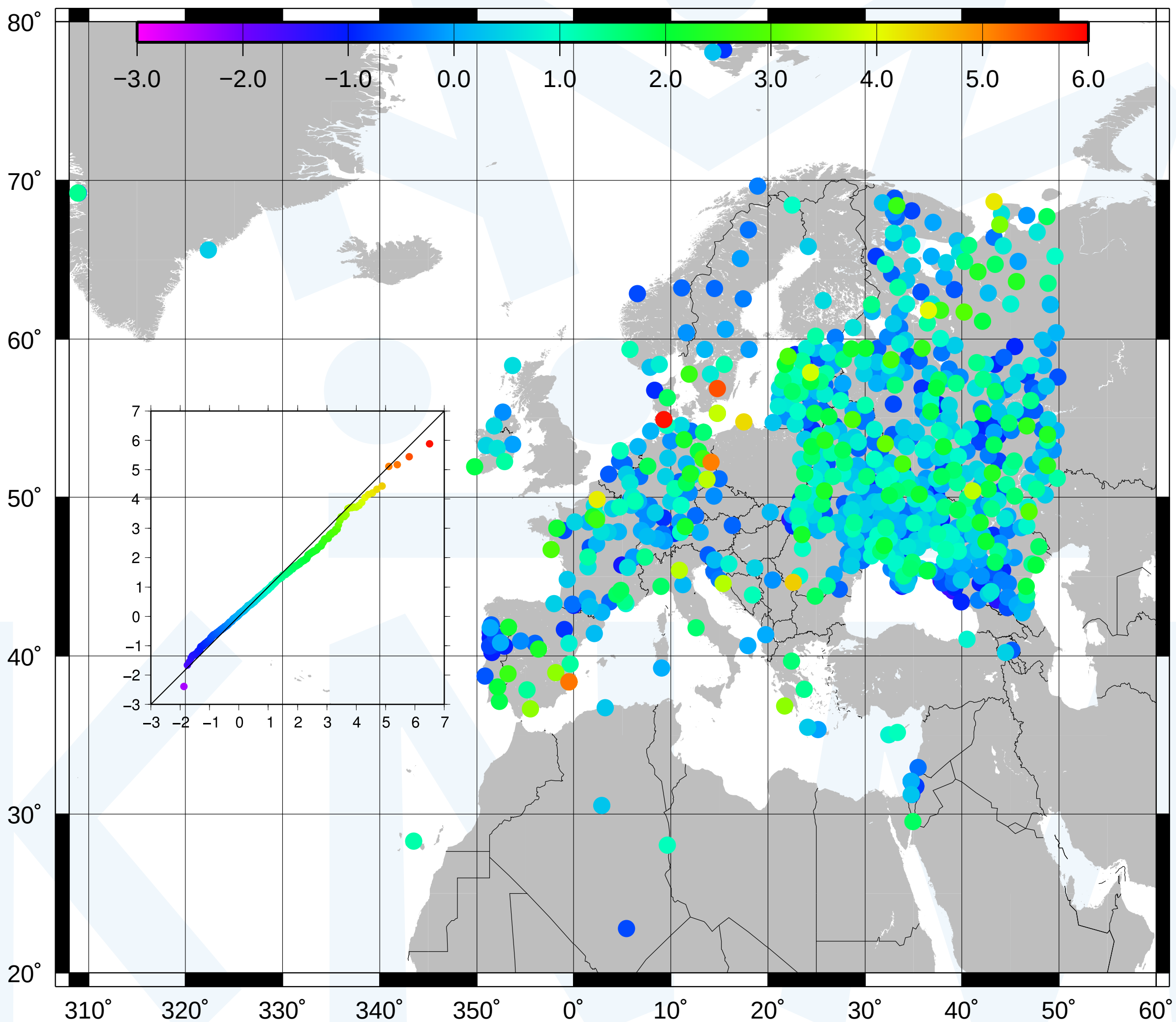


....

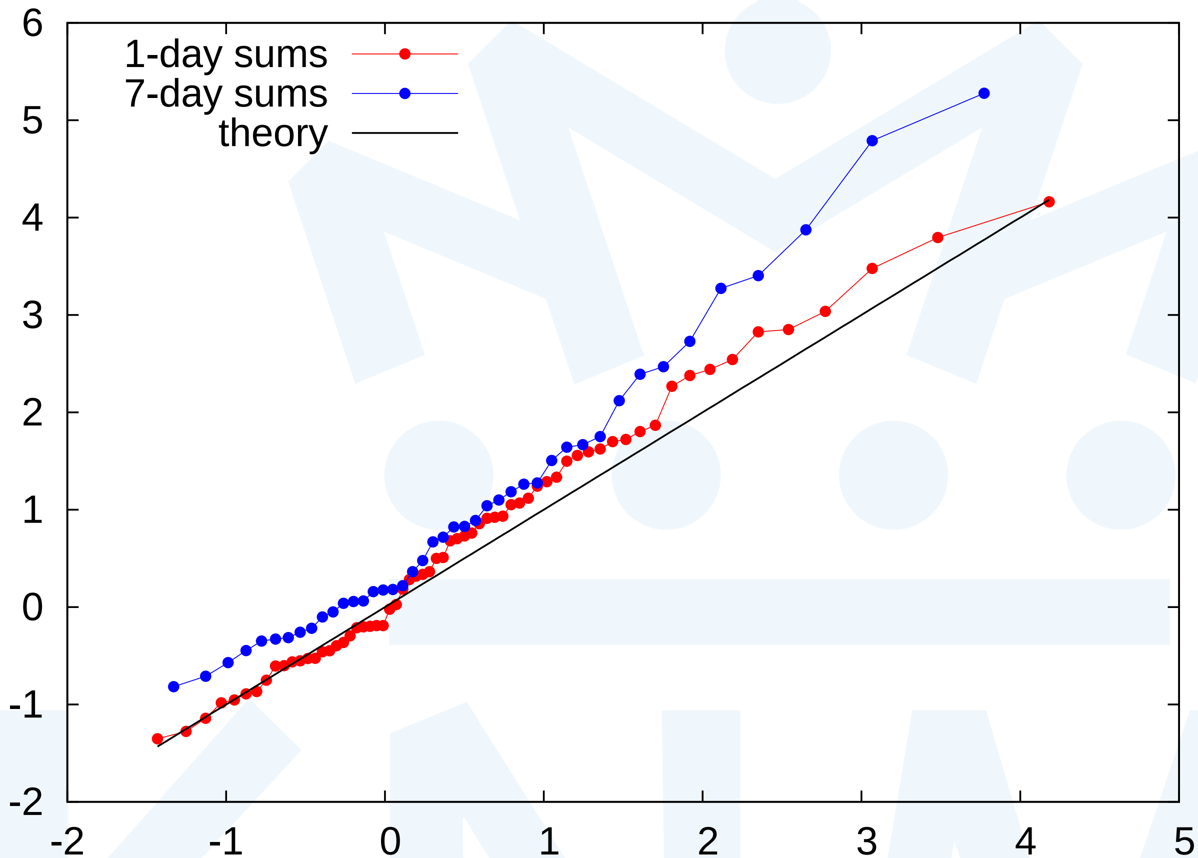
Application to precipitation (1)

- daily precipitation
- 1039 stations in Europe
- different lengths
- only homogeneous sets used (859)
- require 3 days interval (768)
- total length 41188 years
- assume GEV distribution with $\xi = 0.1$

.... Application to precipitation (2)



Application to precipitation (3)

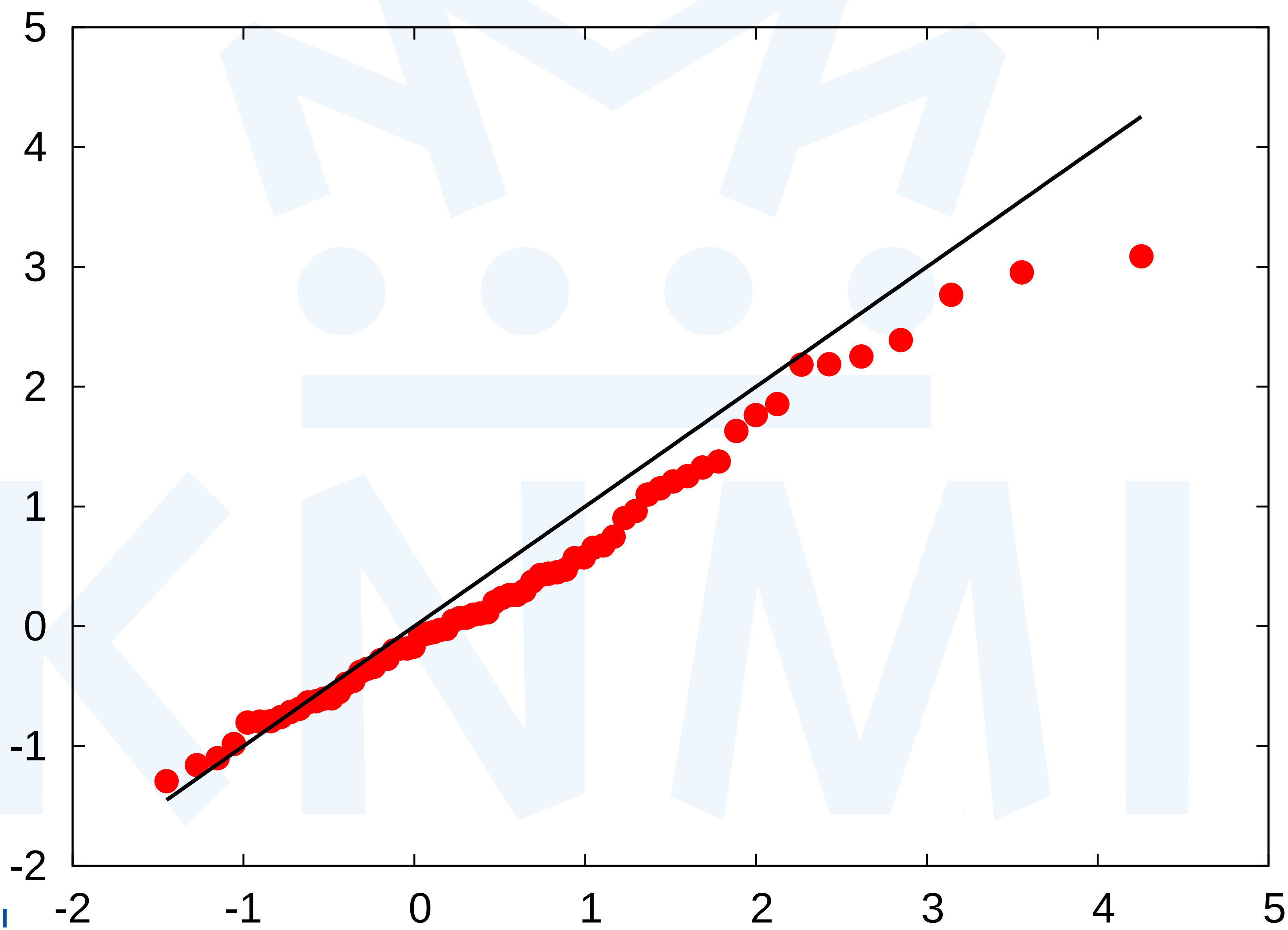


- 1-day sums are convective precipitations
- 7-day sums are mix of convective and large-scale precipitation

Application to surges (1)

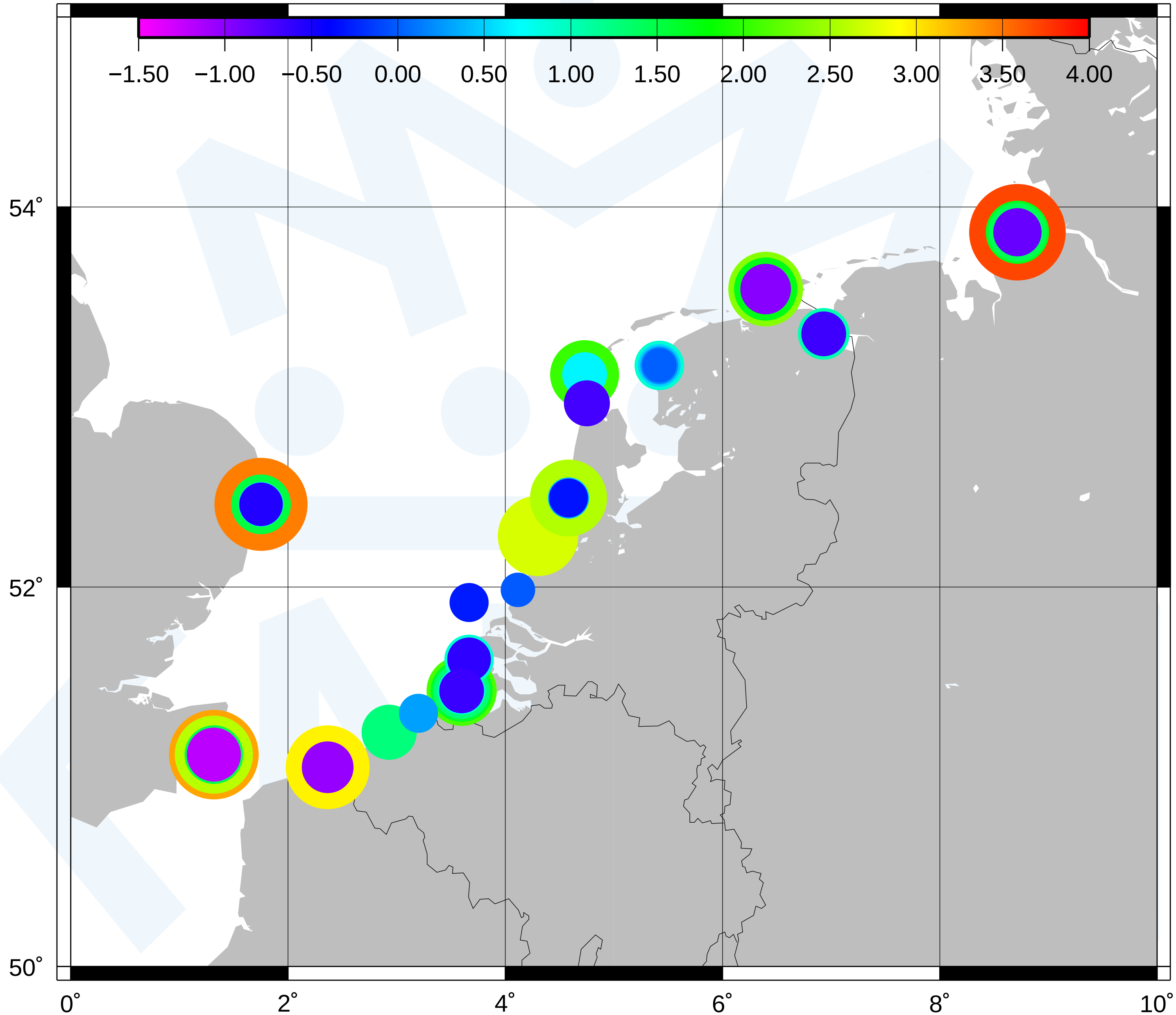
- 17×50 years with ECHAM5 model
- WAQUA used to derive surges for 19 locations
- assume GEV with constant $\xi = -0.03$
- calculate $\Delta\hat{X}_n$ for all 19×17 records
- require 3 days interval

.... **Application to surges (2)**



....

Application to surges (2)



Conclusing remarks

- tool especially interesting for extreme value distributions
- 'goodness-of-fit' tool for *extrapolation*
- only requirement are independent outliers
- GEV-distribution has too many free parameters to be applicable for extrapolation purposes
- can be used to detect:
  - overfitting
  - mixed distributions
  - convergence problems
  - check if constant ξ is justified

Questions?