Institute of Geophysics
Polish Academy of Sciences

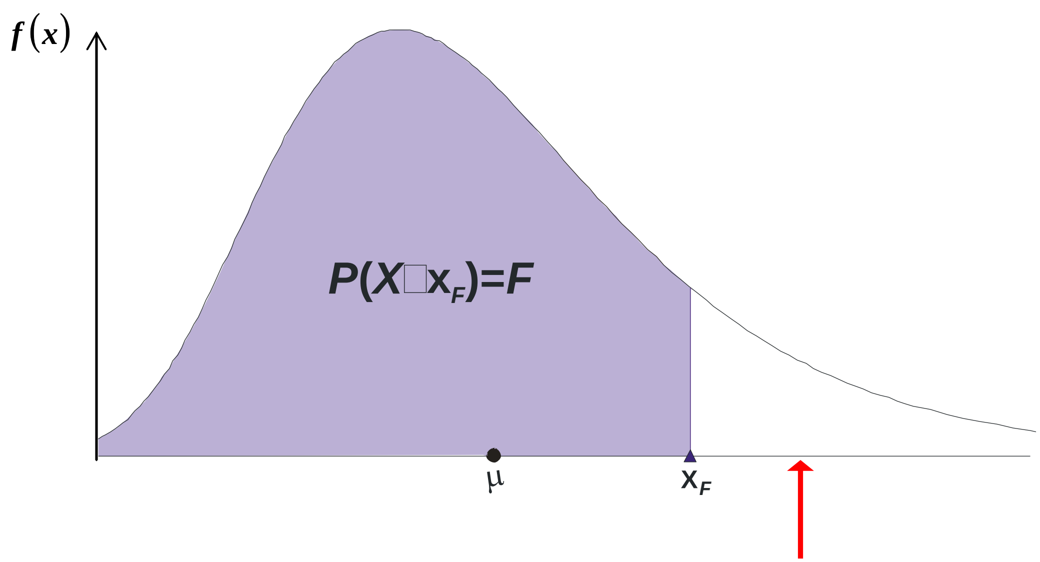
The estimation of annual peak flows

Iwona Markiewicz
Witold G. Strupczewski
Ewa Bogdanowicz

ESF-COST High-level Research Conference on Extreme Environmental Events, December 13-17, 2010 Cambridge, UK

Just to know what we are talking about

Flood Frequency Analysis (FFA) = estimation of upper quantiles of peak flows probability distribution, obtained from annual or partial duration series.



x_F - F quantile

$$P(X \leq x_F) = \int_{-\infty}^{x_F} f(x)\,dx = F$$

Return period *T*

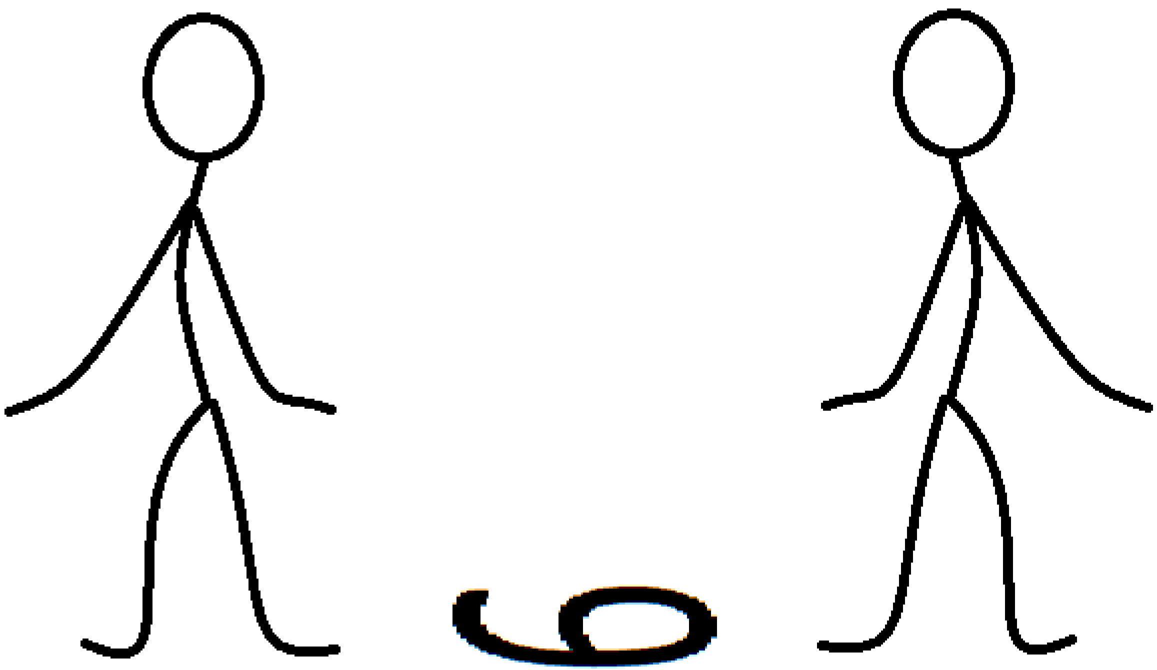
$$T = \frac{1}{1-F} = \frac{1}{P(X > x_F)}$$

$$T = 10 \leftrightarrow F = 0.9$$

$$T = 100 \leftrightarrow F = 0.99$$

$$T = 1000 \leftrightarrow F = 0.999$$

Sometimes the truth depends on the point of view



The true distribution

When we know it...

If we knew it...

If we know it...

„When we know the true distribution”

- **We can identify the theoretical properties of estimation methods**
- **But the hypothetical model differs from the true one!**
  - **upper part of PDF is outside the scope of actual observation range**
  - **peak flows are error-corrupted data and their quality of information is rather low**
  - **no simple statistical model can reproduce the data set in its entire range of variability**
  - **probability of correct identification of PDF on the basis of short hydrological samples is very low**

↓

Traditional approach based on the knowledge of theoretical distribution is not acceptable

So, „if we know the true distribution”

…and we try to estimate the parameters of the false one

- We can investigate the errors, which are due to applied estimation method and to the model misspecification

Probability distributions

Distribution	Probability density function (PDF)
Log-normal 3 (LN3) $\varepsilon = 0$: log-normal 2 (LN2)	$f(x) = \frac{1}{(x-\varepsilon)b\sqrt{2\pi}} \exp\left[-\frac{(\ln(x-\varepsilon)-m)^2}{2b^2}\right]$ m - scale, $b > 0$ - shape; $\varepsilon < x < \infty$
Generalized extreme values (GEV) $\varepsilon = 0$: log-Gumbel (LG)	$f(x) = \frac{1}{\alpha}\left[-\frac{\kappa}{\alpha}(x-\varepsilon)\right]^{1/\kappa - 1} \exp\left\{-\left[-\frac{\kappa}{\alpha}(x-\varepsilon)\right]^{1/\kappa}\right\}$ $\alpha > 0$ - scale, $\kappa < 0$ - shape; $\varepsilon < x < \infty$

Estimation methods

- **Method built on mean deviation - MDM**

MDM	Location	Dispersion	Skewness
Measure	μ	$\delta_\mu = \int_{-\infty}^{+\infty} \lvert x-\mu \rvert \, dF(x)$	$\delta_S = \mu - x_{0.5}$
Dimensionless measure	-	$\delta C_V = \dfrac{\delta_\mu}{\mu}$	$\delta C_S = \dfrac{\delta_S}{\delta_\mu}$

Markiewicz, I. and Strupczewski ,W.G. (2009). Dispersion measures for flood frequency analysis. *Physics and Chemistry of the Earth*, 34: 670-678. DOI 10.1016/j.pce.2009.04.003.
Markiewicz, I., Strupczewski, W.G., Kochanek, K. and Singh, V.P. (2006). Relations between three dispersion measures used in flood frequency analysis. *Stochastic Environamental Research and Risk Assessment*, 20: 391-405. DOI 10.1007/s00477-006-0033-x.

True hypothetical distribution

- **Two-parameter distributions**
  ***T* = LN2, *H* = LN2 and *T* = LG, *H* = LG**

log-normal2, log-Gumbel
$\mu > 0$
$C_V = 0.2, 0.6, 1.0$
$N = 20\,(10)\,100$

MC = 20,000

$\longrightarrow$ $\delta RMSE\left(\hat{x}_{0.99}\right), \delta B\left(\hat{x}_{0.99}\right)$

- **Three-parameter distributions**
  ***T* = LN3, *H* = LN3 and *T* = GEV, *H* = GEV**

log-normal3, GEV
$\mu = 0,\ \sigma = 1$
$C_S = 2.0, 4.0$
$N = 20\,(10)\,100$

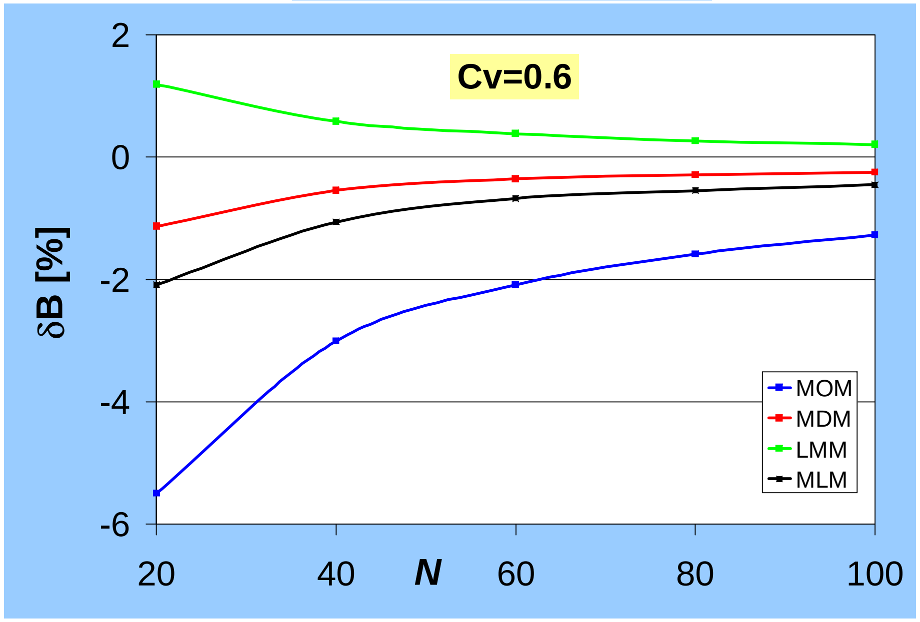
MC = 20,000

$\longrightarrow$ $\delta RMSE\left(\hat{x}_{0.99}\right), \delta B\left(\hat{x}_{0.99}\right)$

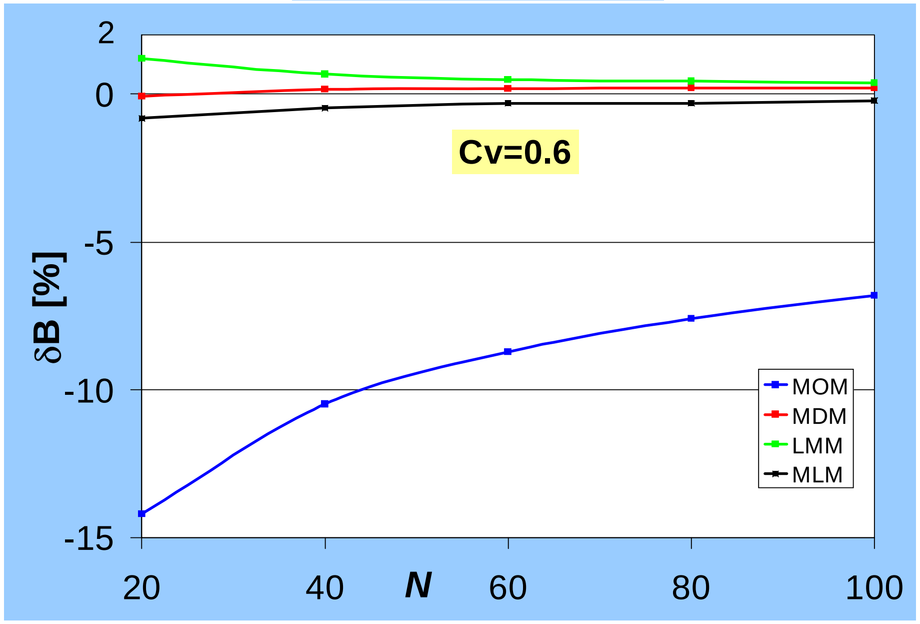
Accuracy of upper quantile estimates two-parameter PDFs

$$\delta B(\hat{x}_{0.99}) = \frac{E(\hat{x}_{0.99} - x_{0.99})}{x_{0.99}}$$

***T* = LN2, *H* = LN2**



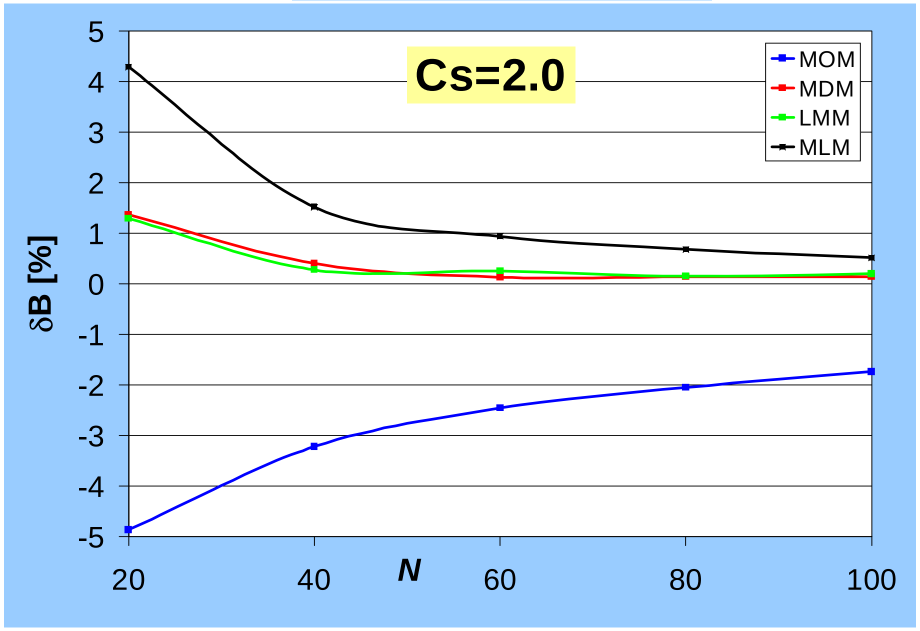
***T* = LG, *H* = LG**



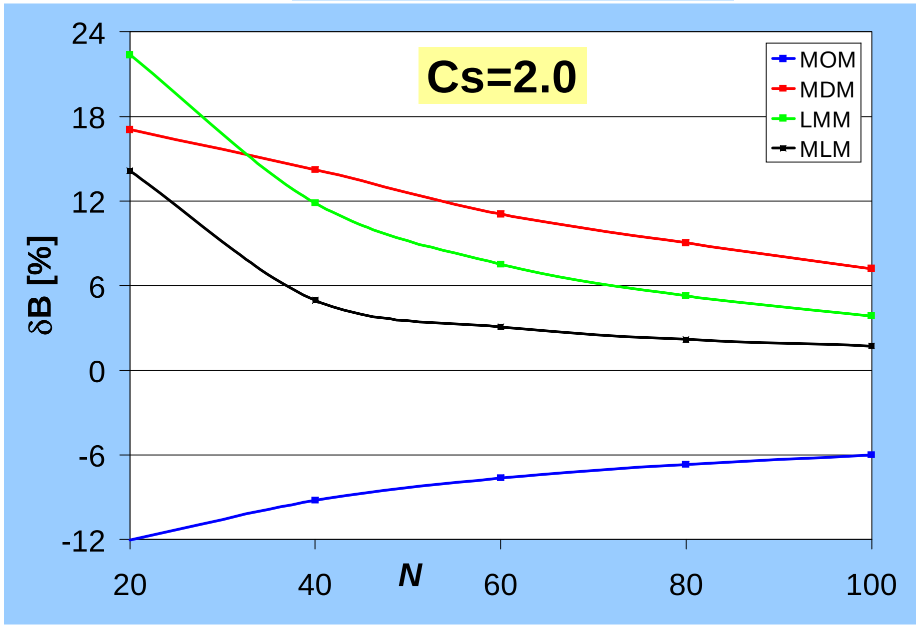
Accuracy of upper quantile estimates
three-parameter PDFs

$$\delta B\left(\hat{x}_{0.99}\right)=\frac{E\left(\hat{x}_{0.99}-x_{0.99}\right)}{x_{0.99}}$$

T = LN3, H = LN3



T = GEV, H = GEV



False hypothetical distribution but we know the true one

- **Two-parameter distributions**
  ***T* = LN2, *H* = LG and *T* = LG, *H* = LN2**

log-normal2, log-Gumbel
$\mu > 0$
$C_V = 0.2, 0.6, 1.0$
$N = 20\,(10)\,100$

MC = 20,000

$\longrightarrow$ $\delta RMSE(\hat{x}_{0.99}), \delta B(\hat{x}_{0.99})$

- **Three-parameter distributions**
  ***T* = LN3, *H* = GEV and *T* = GEV, *H* = LN3**

log-normal3, GEV
$\mu = 0, \ \sigma = 1$
$C_S = 2.0, 4.0$
$N = 20\,(10)\,100$

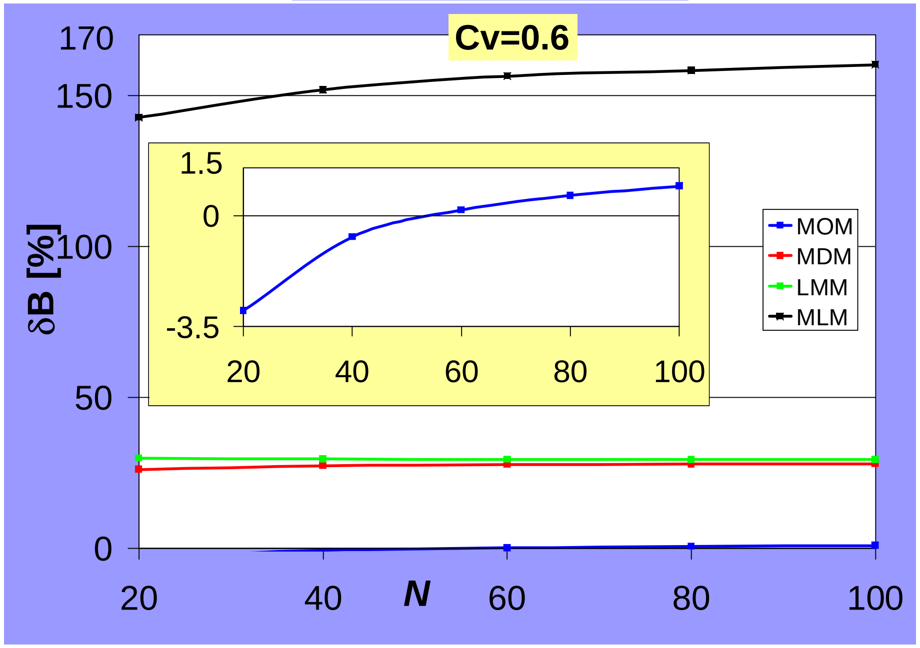
MC = 20,000

$\longrightarrow$ $\delta RMSE(\hat{x}_{0.99}), \delta B(\hat{x}_{0.99})$

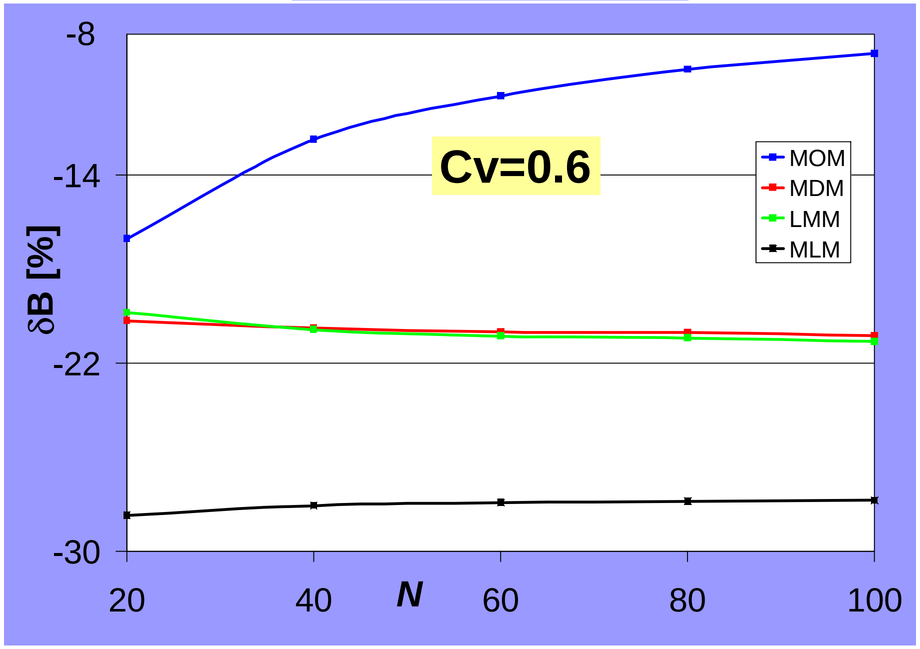
Accuracy of upper quantile estimates two-parameter PDFs

$$\delta B\left(\hat{x}_{0.99}\right) = \frac{E\left(\hat{x}_{0.99} - x_{0.99}\right)}{x_{0.99}}$$

T = LN2, H = LG



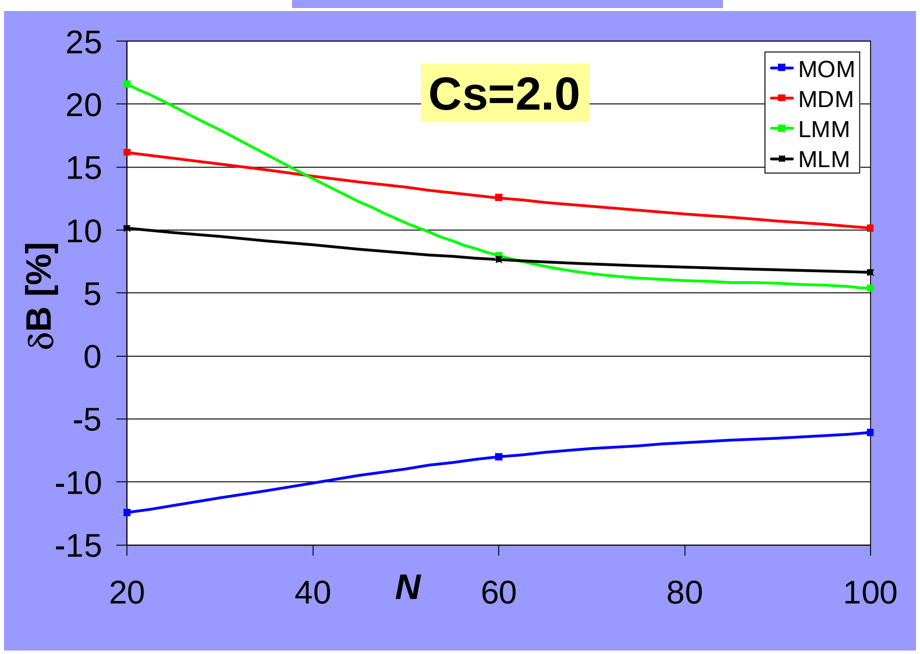
T = LG, H = LN2



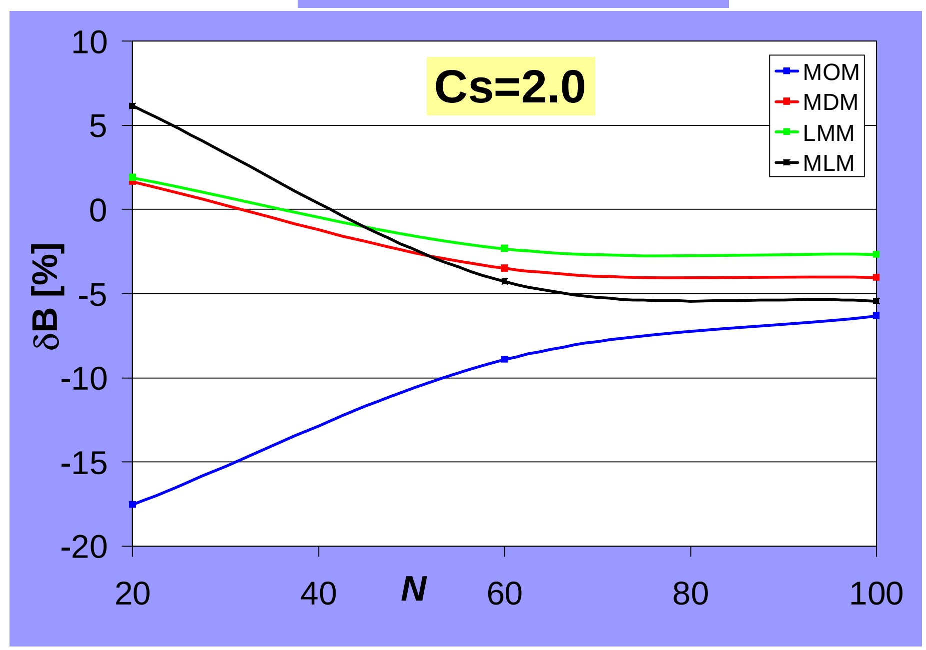
Accuracy of upper quantile estimates three-parameter PDFs

$$\delta B\left(\hat{x}_{0.99}\right)=\frac{E\left(\hat{x}_{0.99}-x_{0.99}\right)}{x_{0.99}}$$

***T* = LN3, *H* = GEV**



***T* = GEV, *H* = LN3**



We don't know the true distribution and we want to choose among the candidate-distributions

AKAIKE information criterion

$$AIC = -2\ln\left(L\left(g\left(x|\hat{\theta}\right)\right)\right) + 2K$$

$$AICc = -2\ln\left(L\left(g\left(x|\hat{\theta}\right)\right)\right) + 2K + \frac{2K(K+1)}{n-K-1}$$

The best (true?) model = this one of the lowest AIC value

But… there are some doubts

- ✓Differences between AIC values for different models are small in context of data accuracy
- ✓Consequences of the best distribution type changes, when the length of observation series increases
- ✓Number and type of candidate distributions

The solution is to use the information provided by the cadidate distributions

and aggregate the results obtained from different models

$\hat{D}_1$

$\hat{D}_2$

$\vdots$

$\hat{D}_m$

$$\hat{D} = \Psi\left(\hat{D}_1, \hat{D}_2, \dots, \hat{D}_m\right)$$

Aggregation of quantiles

$$\bar{x}_F = \Psi\left(\hat{x}_{F_1}, \dots, \hat{x}_{F_m}\right) = \sum_{i=1}^{m} w_i \cdot \hat{x}_{F_i}$$

Conditional expected value

$$w_i = \frac{L_i}{\sum_{k=1}^{m} L_k}$$

Conditional probability of the adequacy of i-th model

$$w_i = \frac{exp\left(-\frac{1}{2}\delta_i\right)}{\sum_{k=1}^{m} exp\left(-\frac{1}{2}\delta_k\right)}$$

$$\delta_i = AIC_i - \min(AIC_1, \dots, AIC_m), i = 1, \dots, m$$

Variance of the aggregated quantile

$$S^2(\bar{x}_F) = S^2(\hat{x}_{F_k}) + \overline{S_k^2}$$

Total variance of aggregated quantile

Variance of quantiles

Mean quantiles variance

The results

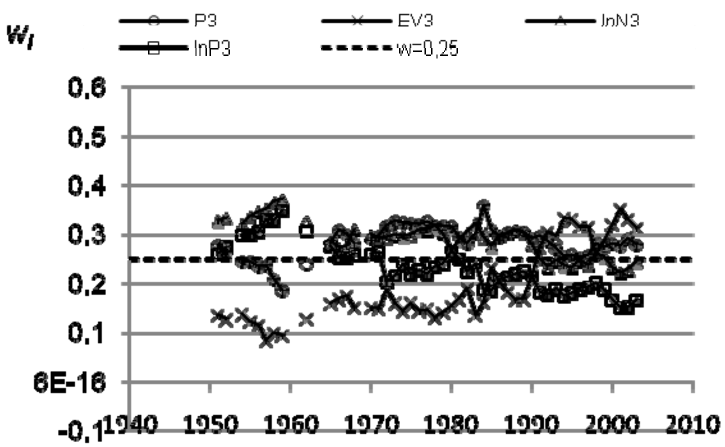
The results for winter maxima at Tczew on the Vistula river (1921-2003)

i	Distribution type	AIC	δ_i	w_i	$x_{0.99}$ ($m^3 \cdot s^{-1}$)	$S(x_{0.99})$ ($m^3 \cdot s^{-1}$)
1	P3	1430.205	0	0.299	7800	549.1
2	EV3	1430.535	0.330	0.254	7570	757.4
3	lnN3	1430.507	0.302	0.257	8270	708.6
4	lnP3	1431.113	0.908	0.190	9310	835.0

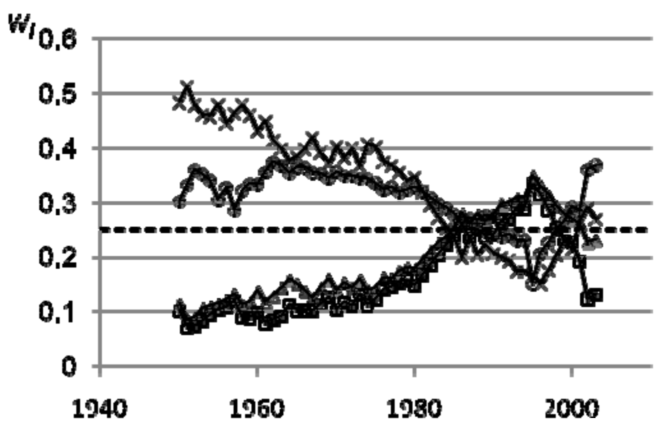
$\bar{x}_{0.99} = 8150\, m^3 \cdot s^{-1}$; $\qquad S(\bar{x}_{0.99}) = 937\, m^3 \cdot s^{-1}$

The weigths versus time (the length of the observation series)

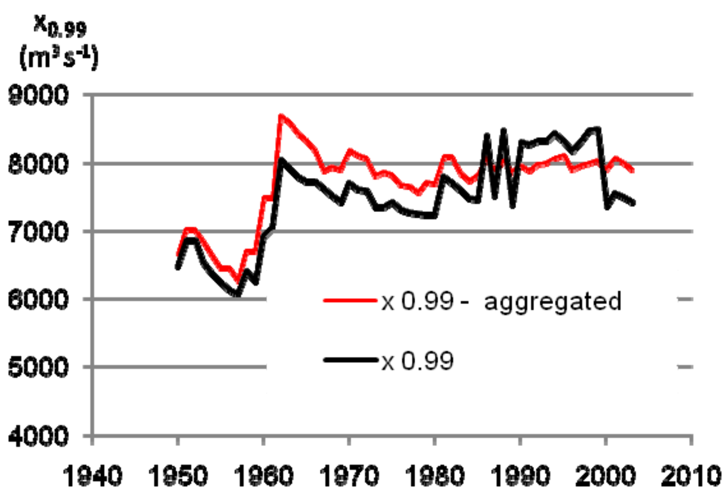
Tczew; winter peak flows

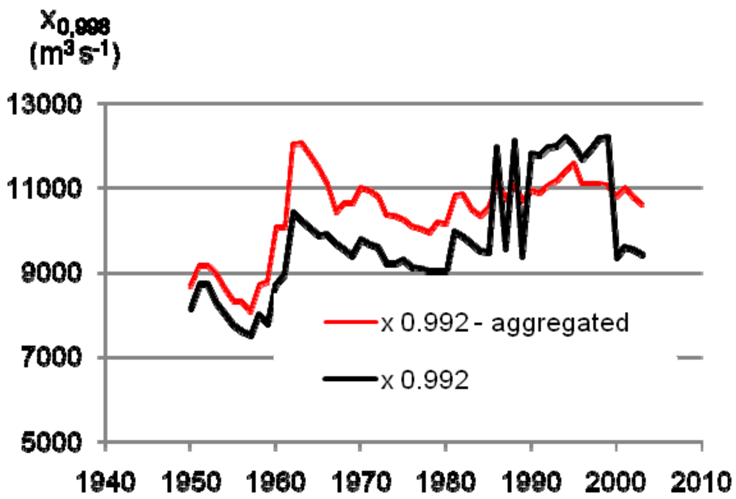


Tczew; summer peak flows



The quantiles versus time (the length of the observetion series)





Conclusions

- ✓ Ranking of estimation methods in respect to upper quantile accuracy depends on:
  - type of distributions, both real and hypothetical
  - number of distribution parameters
  - sample size

- ✓ For two-parameter distributions, in the case of model misspecification, the MLM yields the highest bias of quantile estimates, regardless on the sample size, while the MOM the smallest one

- ✓ Presented analysis can be a source of information about the properties of selected distribution and estimation (D/E) procedures

- ✓ Studies should be extended for other distributions

- ✓ Aggregation method will be regarded as sharpening operation in fuzzy sets theory

THANK YOU