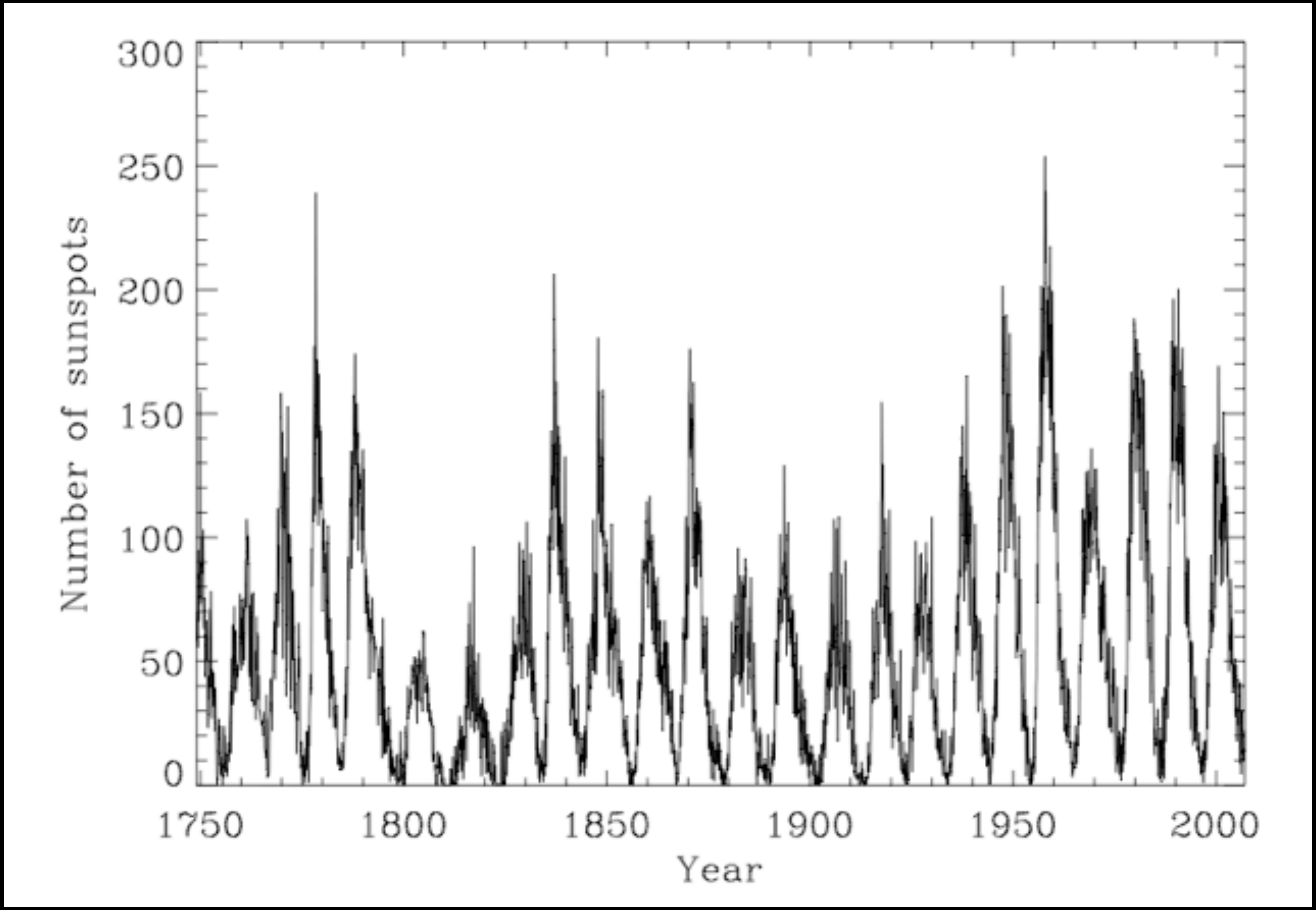
Extreme value statistics in astronomy

David Valls-Gabaud

CNRS, Observatoire de Paris

Atelier Astrostatistiques
Grenoble 2011 Dec 09

Extreme values of a sample



Basics of extreme value theory I

Let $X_1 \ldots X_n$ be independent random variables having a common distribution function F.

Let M_n be the maximum of n observations :

$$M_n = \max\{X_1, \ldots, X_n\}$$

The theoretical distribution of M_n

$$\begin{aligned}\Pr\{M_n \leq z\} &= \Pr\{\mathrm{X}_1 \leq \mathrm{z}, \ldots, \mathrm{X}_\mathrm{n} \leq \mathrm{z}\} \\ &= \Pr\{X_1 \leq z\} \times \cdots \times P\{X_n \leq z\} \\ &= \{F(z)\}^n.\end{aligned}$$

not very useful since usually F is unknown...

Basics of extreme value theory II

Make a simple linear renormalisation of M_n

$$M_n^* = \frac{M_n - b_n}{a_n},$$

for constants $\{a_n > 0\}$ and $\{b_n\}$ chosen appropriately to stabilise the location and scale of M_n^* as n increases.

Basics of extreme value theory III

Theorem (Fréchet 1927): If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$\Pr\left\{\frac{M_n - b_n}{a_n} \leq z\right\} \to G(z) \qquad \text{as } n \to \infty,$$

where G is a non-degenerate distribution function, then G belongs to one of the following families:

$$\begin{aligned}
\text{I}: \quad G(z) &= \exp\left\{-\exp\left[-\left(\frac{z-b}{a}\right)\right]\right\}, \quad -\infty < z < \infty; \\
\text{II}: \quad G(z) &= \begin{cases} 0, & z \leq b, \\ \exp\left\{-\left(\frac{z-b}{a}\right)^{-\alpha}\right\}, & z > b; \end{cases} \\
\text{III}: \quad G(z) &= \begin{cases} \exp\left\{-\left[-\left(\frac{z-b}{a}\right)^{\alpha}\right]\right\}, & z < b, \\ 1, & z \geq b, \end{cases}
\end{aligned}$$

independently of the underlying F ... (REM: central limit theorem for sample means)

- **The Gumbel distribution**

$$G(x) = \exp(-\exp(-x))$$

- **The Fréchet distribution**

$$G(x) = \begin{cases} 0 & x \le 0 \\ \exp\left(-x^{-\alpha}\right) & x > 0, \alpha > 0 \end{cases}$$

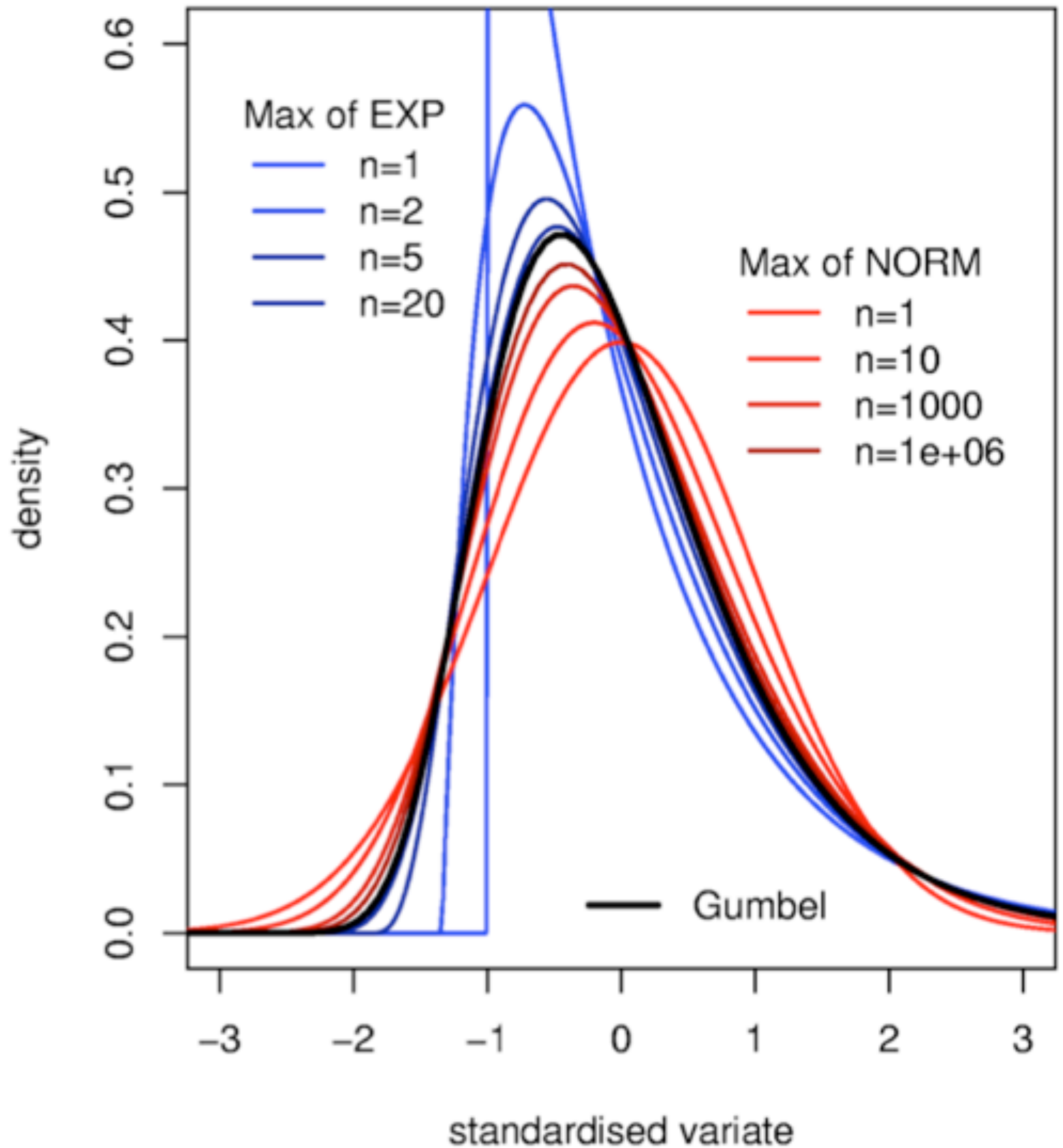
- **The Weibull distribution**

$$G(x) = \begin{cases} \exp\left(-(-x)^{\alpha}\right) & x < 0, \alpha > 0 \\ 1 & x \ge 0 \end{cases}$$

Emil Julius Gumbel
1891-1966

Maurice René
Fréchet
1878-1973

Ernst Hjalmar
Waloddi Weibull
1887-1979


Max of EXP
n=1
n=2
n=5
n=20
Max of NORM
n=1
n=10
n=1000
n=1e+06
Gumbel
density
standardised variate
0.0
0.1
0.2
0.3
0.4
0.5
0.6
−3
−2
−1
0
1
2
3

Corollary: If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$\Pr\left\{\frac{M_n - b_n}{a_n} \leq z\right\} \to G(z) \qquad \text{as } n \to \infty,$$

for a non-degenerate distribution function G, then G is a member of the GEV family

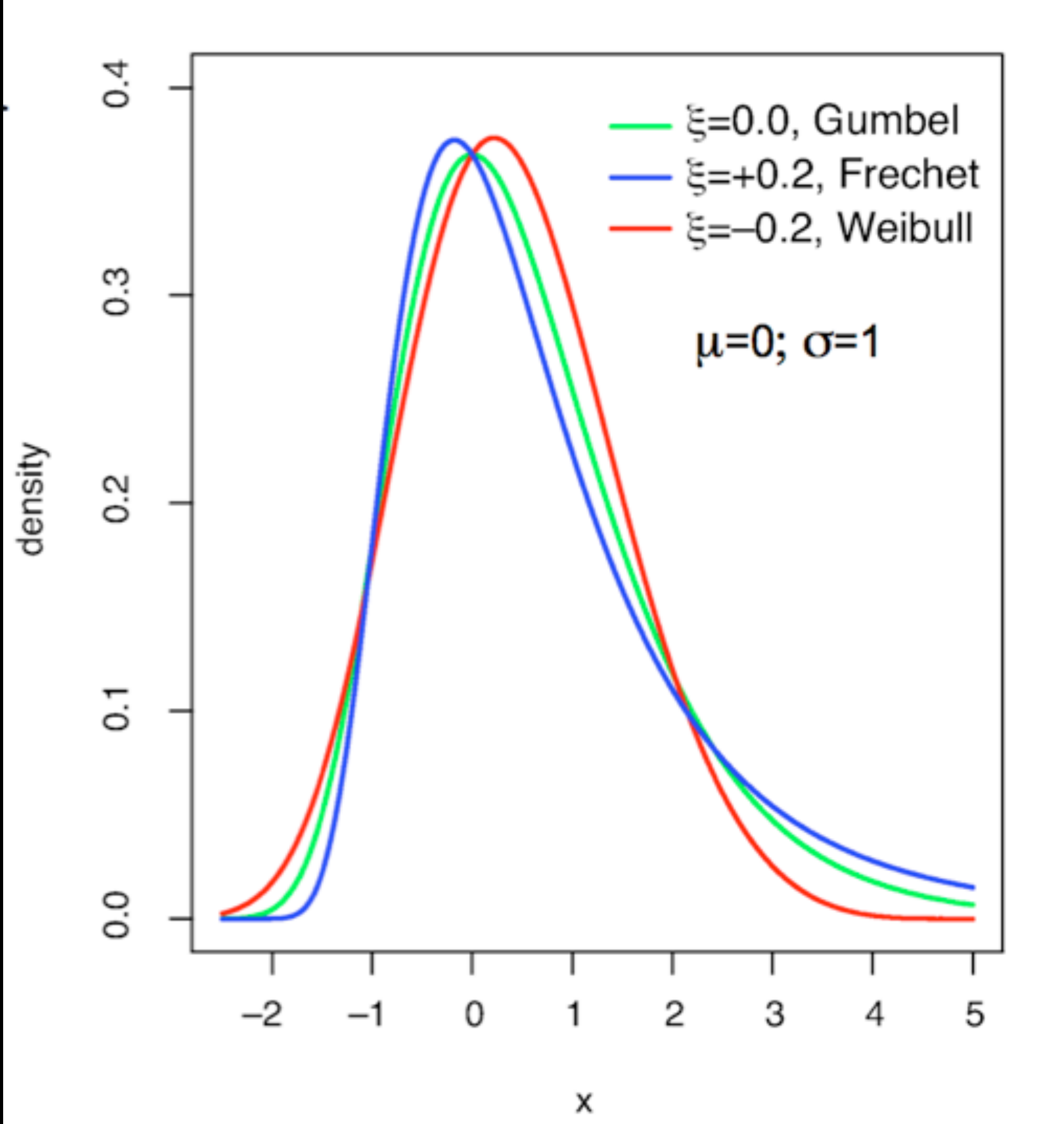
$$G(z) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, \tag{2.2}$$

defined on $\{z : 1+\xi(z-\mu)/\sigma > 0\}$, where $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$.

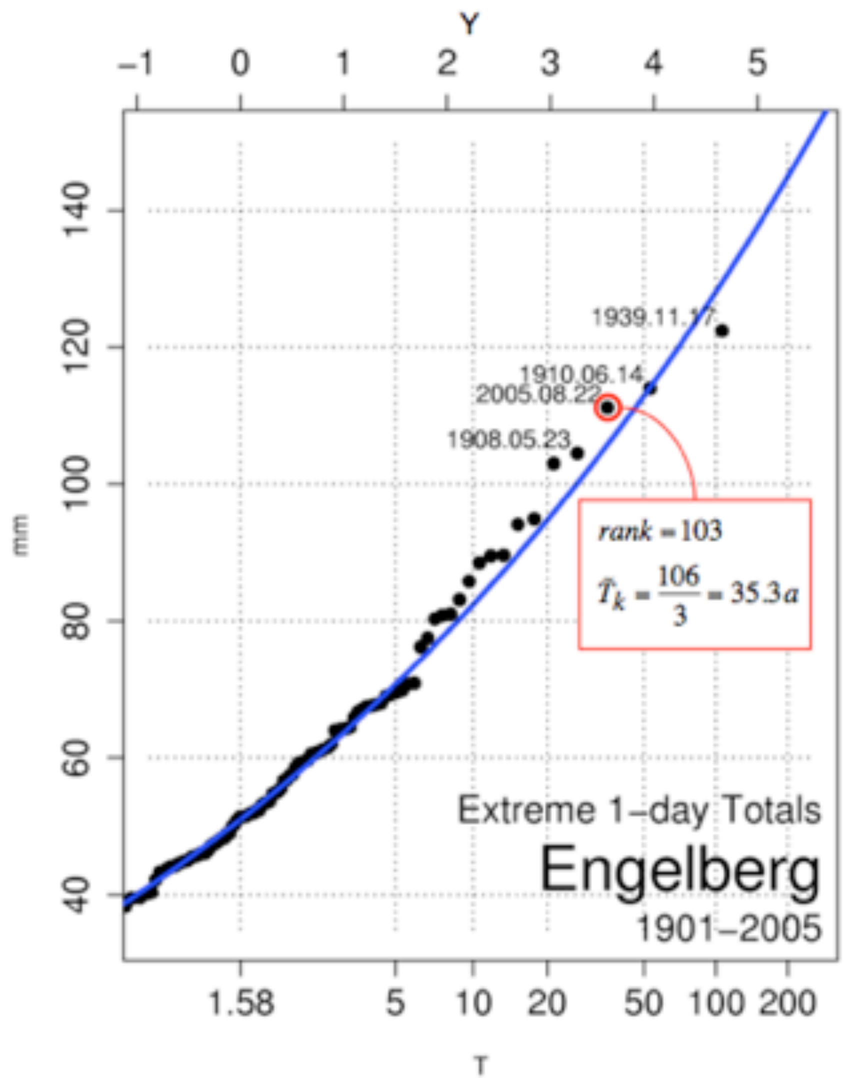
$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1-\{-\log(1-p)\}^{-}\xi\right], & \text{for } \xi \neq 0, \\ \mu - \sigma\log\{-\log(1-p)\}, & \text{for } \xi = 0, \end{cases}$$

where $G(z_p) = 1 - p$.

Definition The extreme quantile $z_p = G^{-1}(1-p)$, where G is the distribution function of M_n, is called the *return level* associated with the *return period* $1/p$.



μ : location
σ : scale
ξ : slope



T-axis is transformed such that Gumbel-Distribution is a straight line.

$F(x) = GEV(x;\mu,\sigma,\xi)$ estimated CDF

$Y(x) = -\log(-\log(F(x)))$ Gumbel Variate

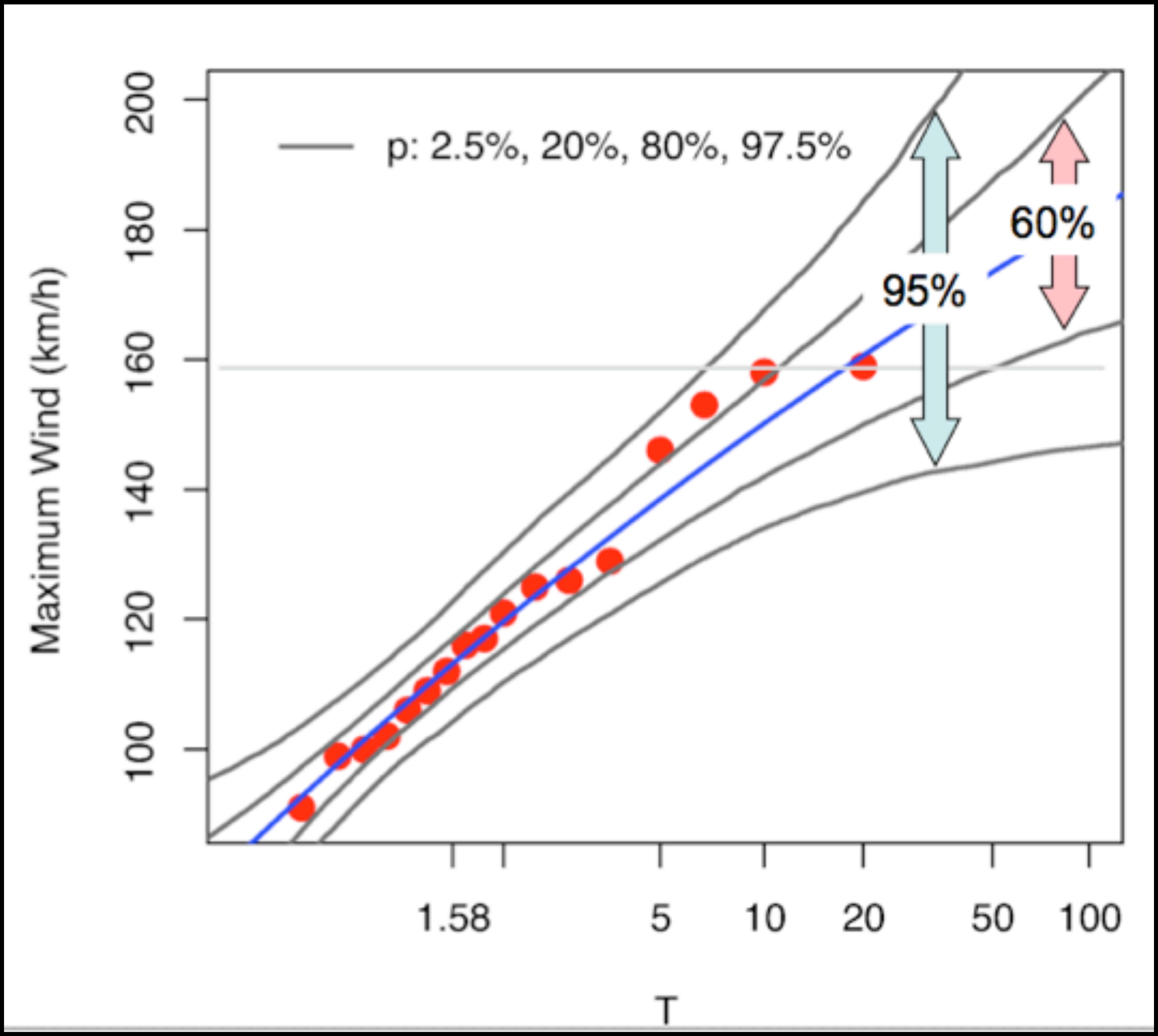
Horizontal axis is linear in Y.

$T(x) = 1/(1 - F(x))$ Return period

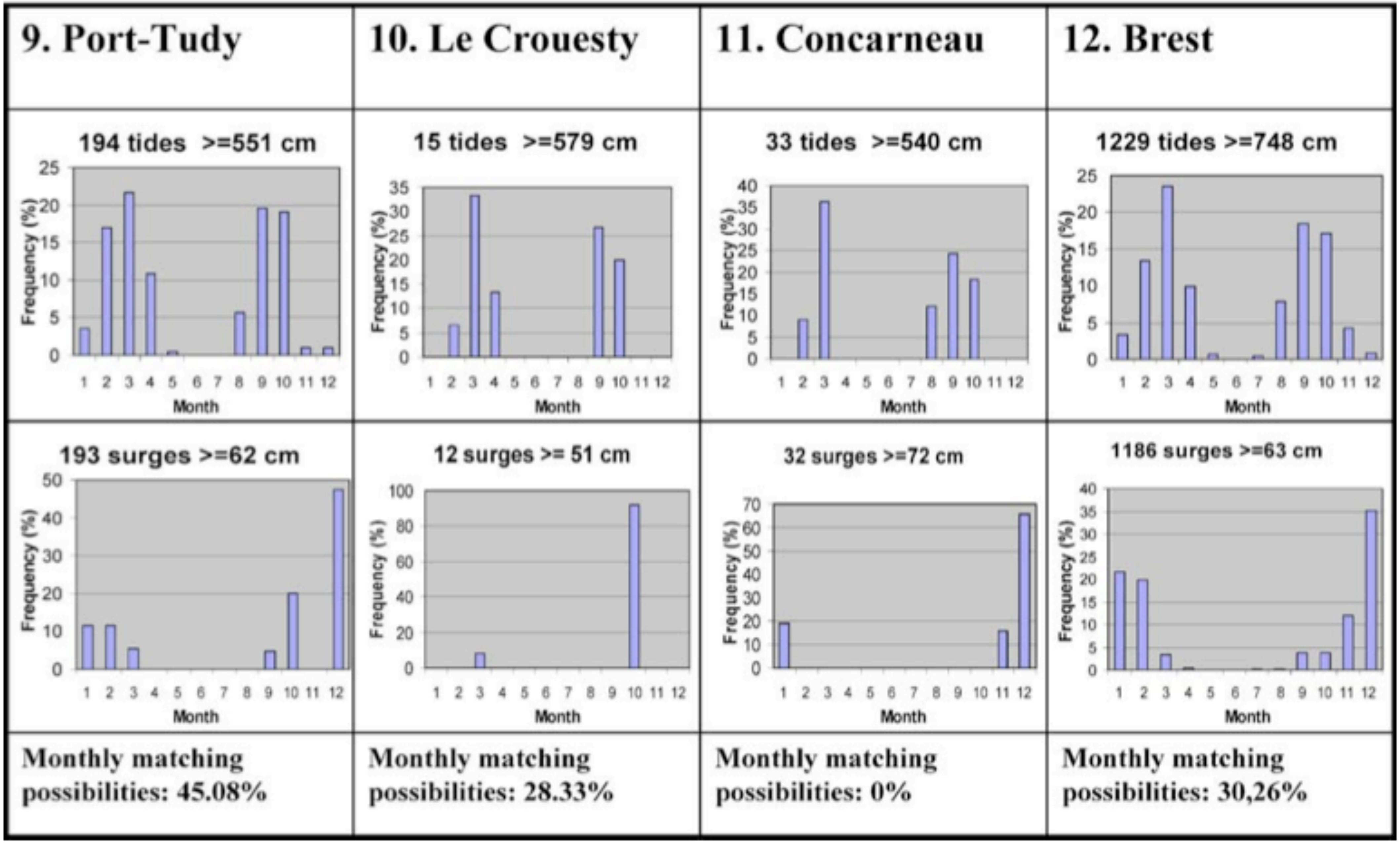
$x_k \quad k = 1,..,N$ Block Maxima

$$\tilde{T}_k = \frac{N+1}{N+1-rank(x_k)}$$ plotting points of block maxima x_k

Parametric resampling and confidence intervals

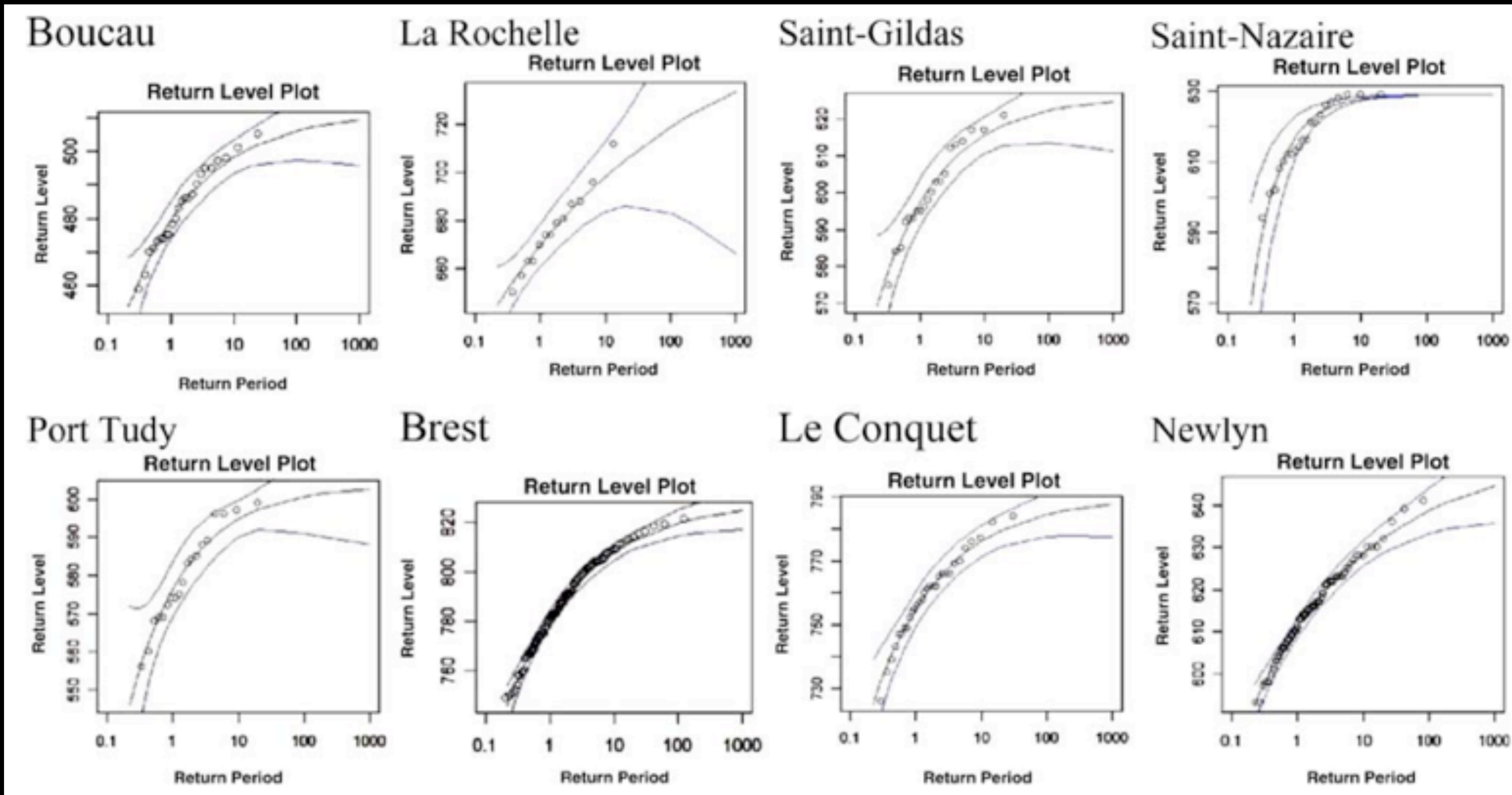


Extreme tides: temporal series



Pirazzoli & Tomasin (2007) Ocean Dynamics 57, 91

Expected return periods

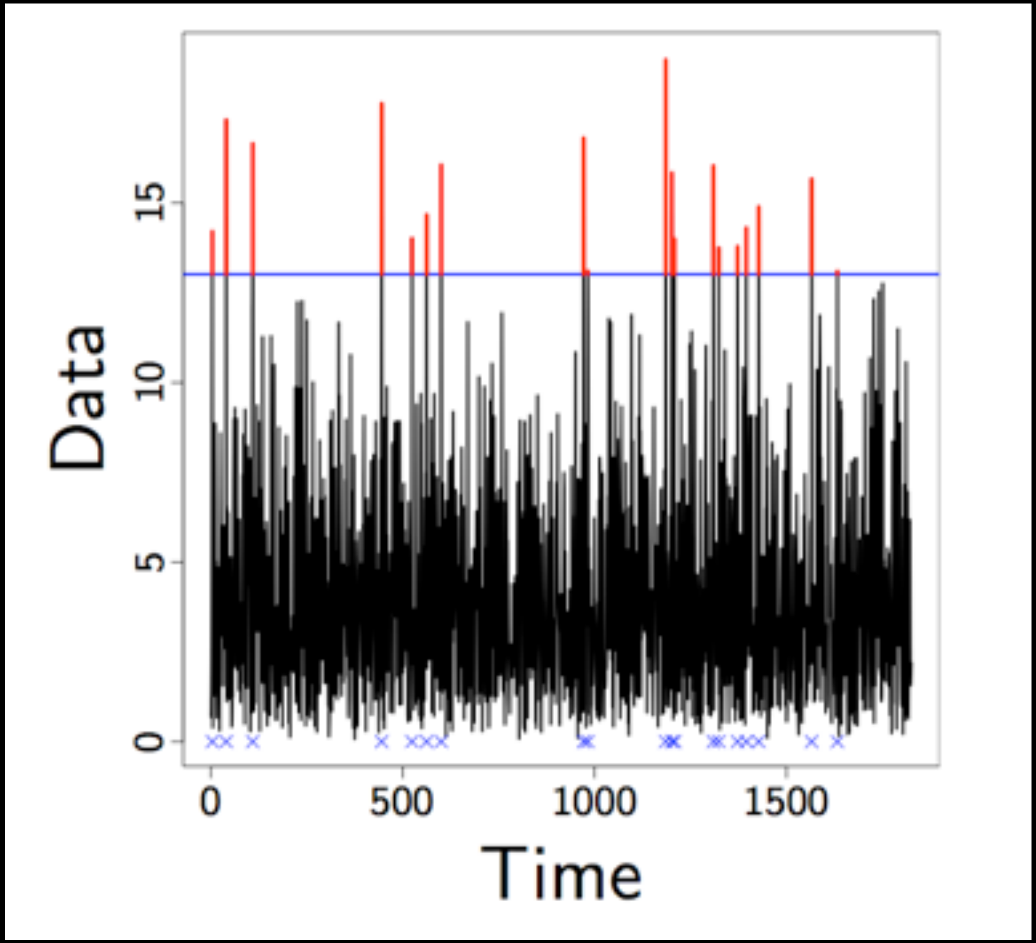


Pirazzoli & Tomasin (2007) Ocean Dynamics 57, 91

Basics of extreme value theory IV

Estimation of GEV parameters:

- ★ Method of moments: not robust and VERY unstable
- ★ Block maxima: extreme events in FIXED intervals
- ★ Peak over thresholds: all extreme events ABOVE a fixed value



Basics of extreme value theory V

Conditional excess distribution function for a threshold level *u* :

$$F_u(y) = P(y \geq X - u \mid X > u), \qquad 0 \geq y < \infty$$

$$F_u(y) = \frac{F(u+y) - F(u)}{1 - F(u)}, \qquad y > 0$$

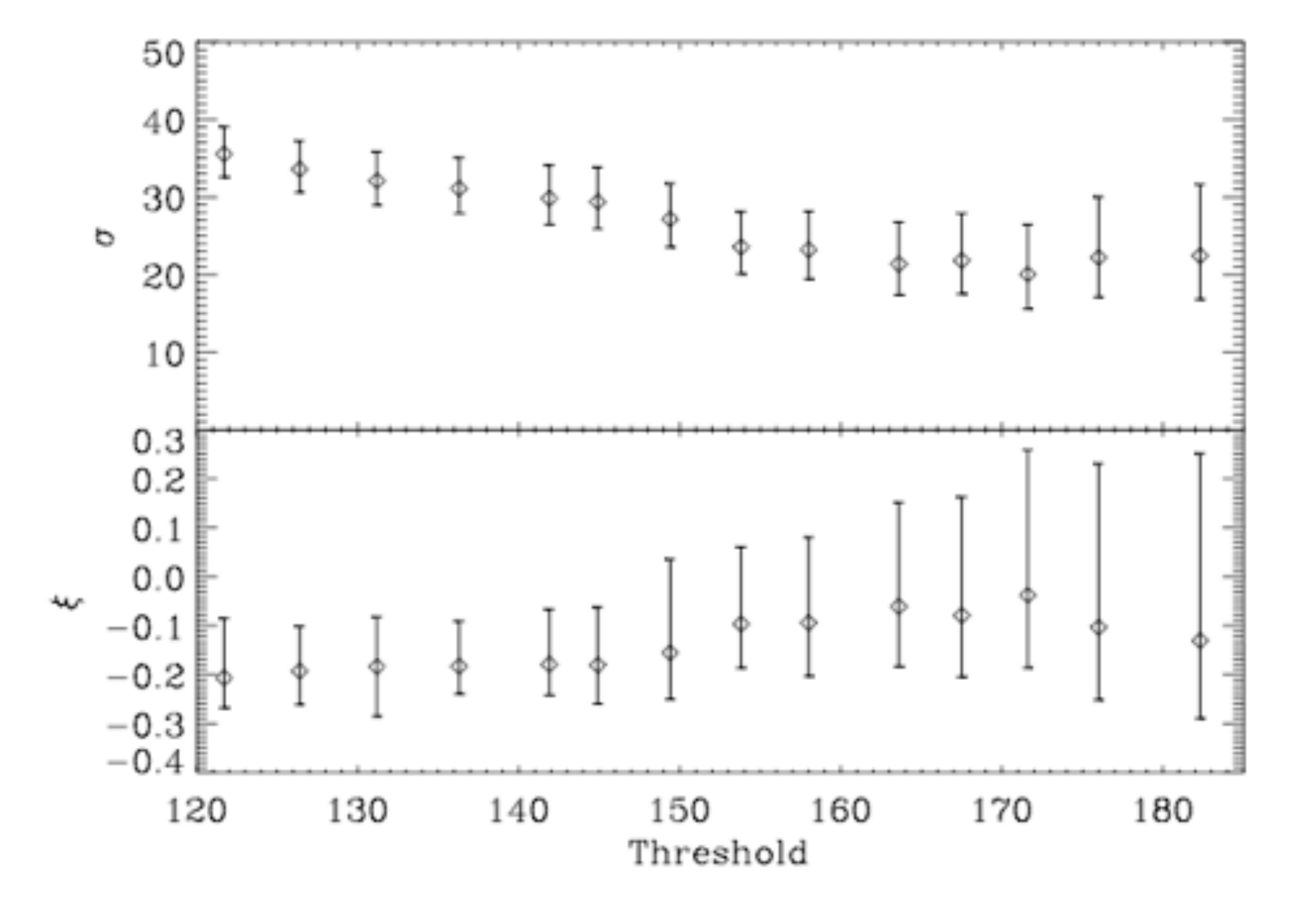
Theorem (Pickands 1975):

$$F_u(y) \approx \begin{cases} 1 - \left(1 + \frac{\xi}{\sigma} y\right)^{-1/\xi} & \text{if } \xi \neq 0 \\ 1 - e^{-y/\sigma} & \text{if } \xi = 0, \end{cases}$$

so that the cumulative DF for events above *u*, taking $x=u+y$, is :

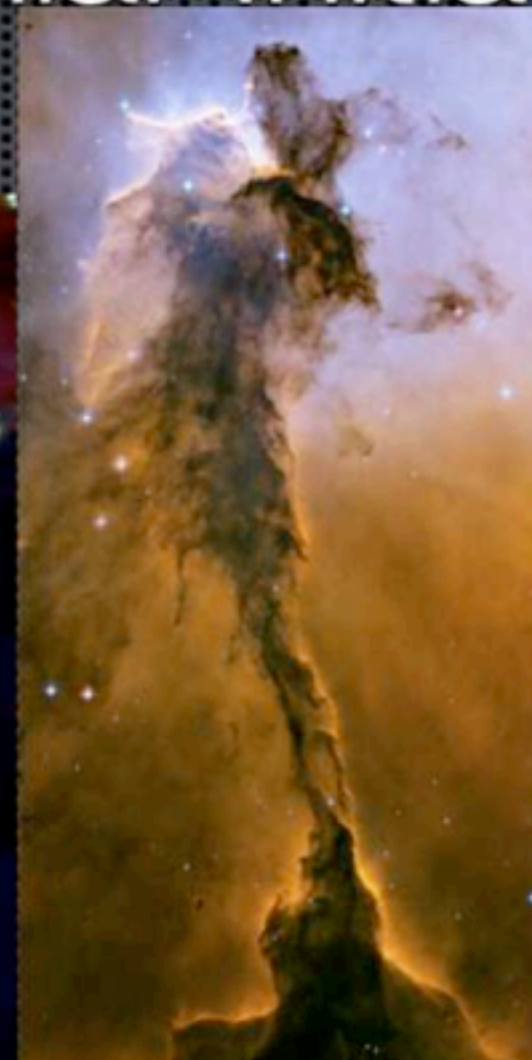
$$F(x) = 1 - \frac{N_u}{n}\left[1 + \frac{\xi}{\sigma}(x - u)\right]^{-1/\xi}$$

For a given sample, calculate the likelihood and apply standard Bayesian estimation :

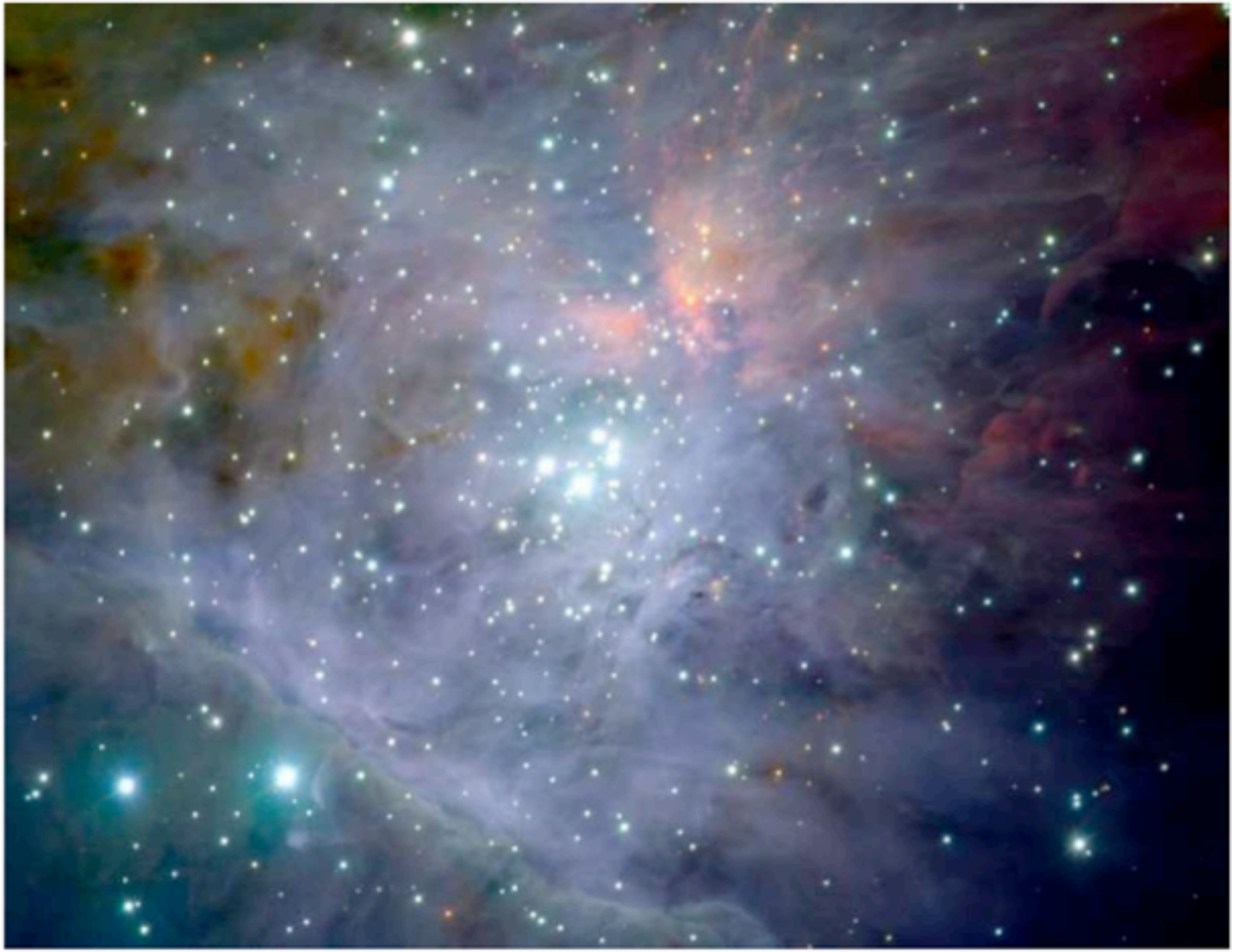


Asensio Ramos (2007) A&A 472, 293

Star Formation and the Origins of the Stellar Initial Mass Function



Why are there fewer massive stars than low-mass stars?



- Typical stellar mass $\sim$ Jeans Mass.

$$M_{\mathrm{Jeans}} \equiv \left[\frac{5}{2}\frac{R_g T}{G\mu}\right]^{\frac{3}{2}} \left[\frac{4\pi}{3}\rho\right]^{-\frac{1}{2}}.$$

- In Molecular clouds:
  - Temperatures $\sim$ 10 K,
  - Lots of structure,
  - Dense cores: $\rho \sim 10^{-19}$ g/cm^3.
- $\Longrightarrow$ masses M $\sim$ 0.7 $\mathrm{M}_{\mathrm{Sol}}$.
  - Agrees well with observations:
    - $\mathrm{M}_{\mathrm{median}} \sim 0.5\ \mathrm{M}_{\mathrm{Sol}}$.
- Not all stars have the same mass $\Longrightarrow$ Distribution?
- The IMF!

- **Salpeter-IMF** (power-law): $\xi(m) = k\ m^{-2.35}$, originally only for 0.4 to 10 $\mathrm{M_{Sol}}$.

- **Miller-Scalo IMF** (log-normal):

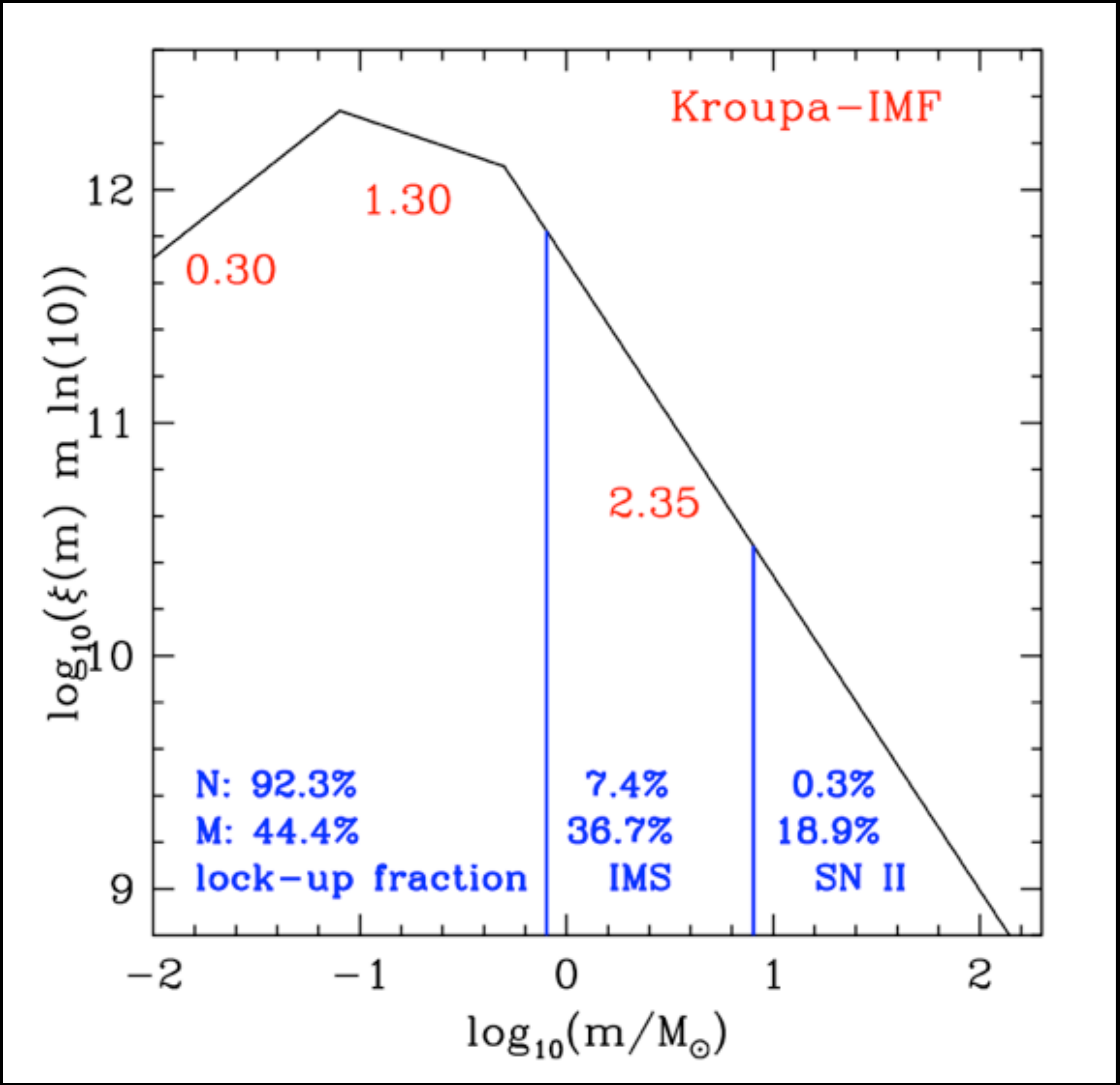
$$\xi(lm) = k\ e^{-\frac{(lm+1.02)^2}{0.9248}},$$

with $lm = \log_{10} m$.

- **Kroupa-IMF** (multi power-law):

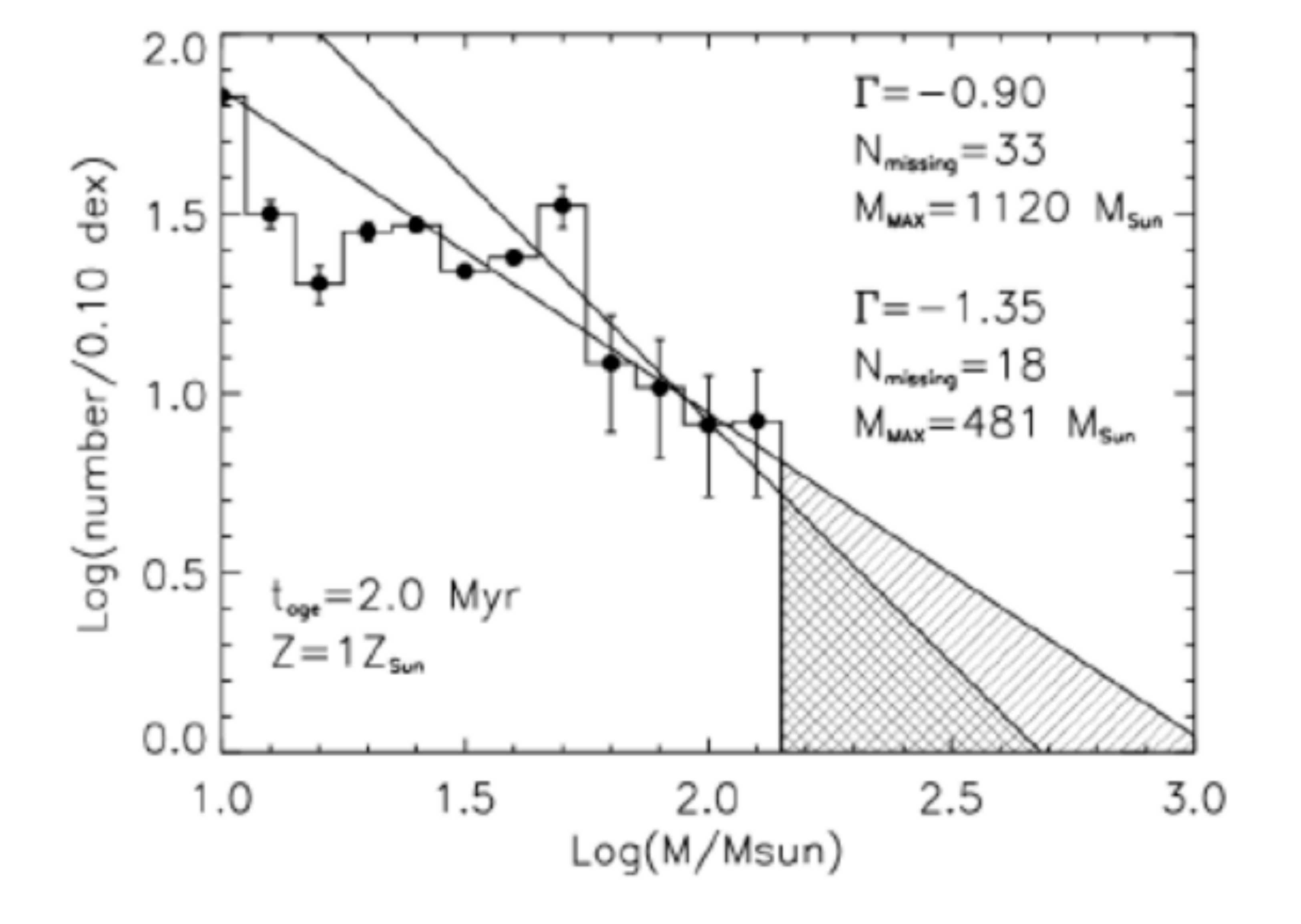
$$\xi(m) = k \begin{cases} \left(\frac{m}{m_{\rm H}}\right)^{-\alpha_0} & , m_{\rm low} \le m < m_{\rm H}, \\ \left(\frac{m}{m_{\rm H}}\right)^{-\alpha_1} & , m_{\rm H} \le m < m_0, \\ \left(\frac{m_0}{m_{\rm H}}\right)^{-\alpha_1} \left(\frac{m}{m_0}\right)^{-\alpha_2} & , m_0 \le m < m_1, \\ \left(\frac{m_0}{m_{\rm H}}\right)^{-\alpha_1} \left(\frac{m_1}{m_0}\right)^{-\alpha_2} \left(\frac{m}{m_1}\right)^{-\alpha_3} & , m_1 \le m < m_{\rm max}, \end{cases}$$

$$\begin{array}{lcl} \alpha_0 = +0.30 & , & 0.01 \le m/M_\odot < 0.08, \\ \alpha_1 = +1.30 & , & 0.08 \le m/M_\odot < 0.50, \\ \alpha_2 = +2.35 & , & 0.50 \le m/M_\odot < 1.00, \\ \alpha_3 = +2.35 & , & 1.00 \le m/M_\odot. \end{array}$$

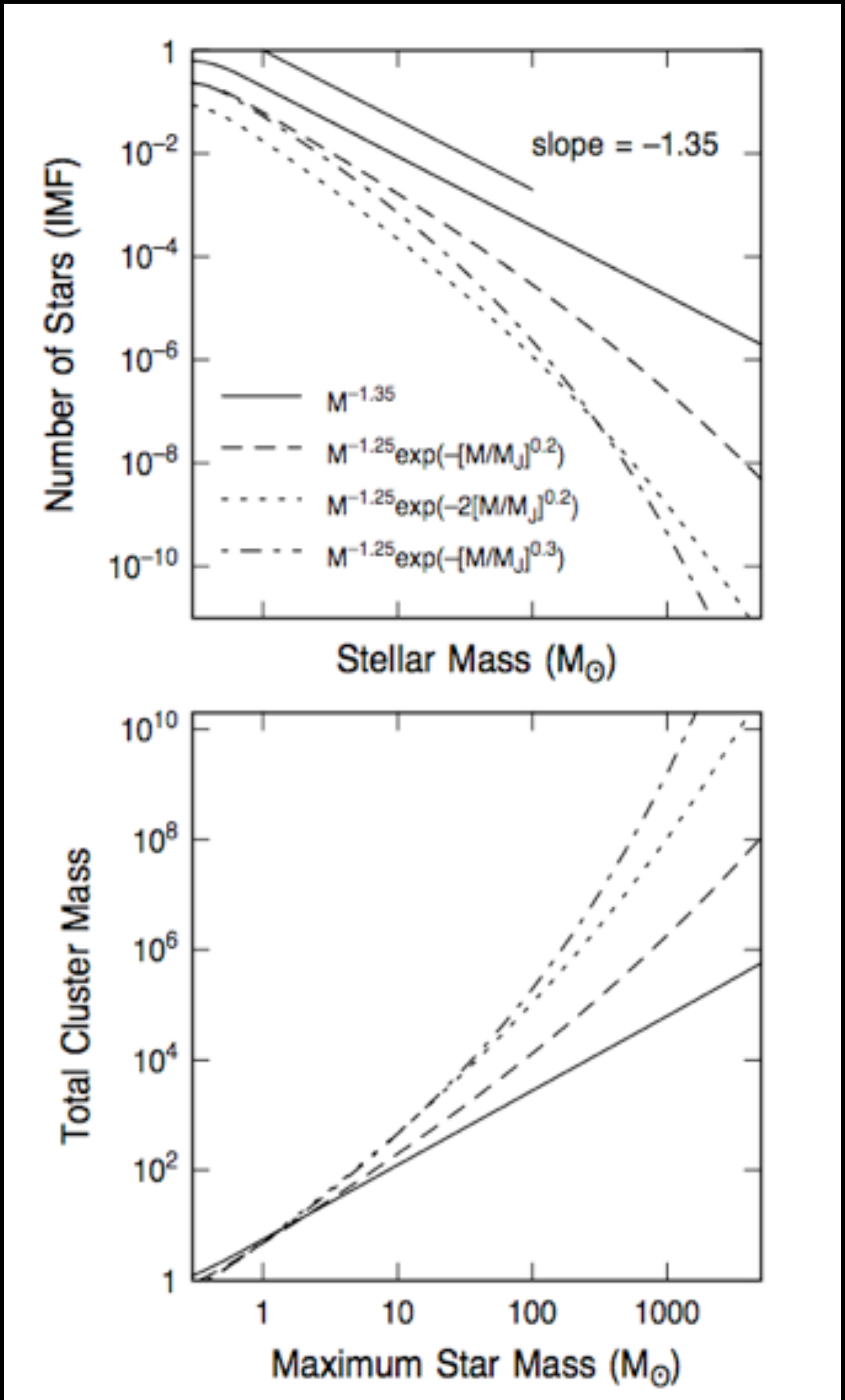


Kroupa−IMF
0.30
1.30
2.35
N: 92.3%
M: 44.4%
lock-up fraction
7.4%
36.7%
IMS
0.3%
18.9%
SN II
12
11
10
9
−2
−1
0
1
2
$\log_{10}(\xi(m)\ m\ \ln(10))$
$\log_{10}(m/M_\odot)$

Evidence for an upper mass cutoff around 150 $M_\odot$?

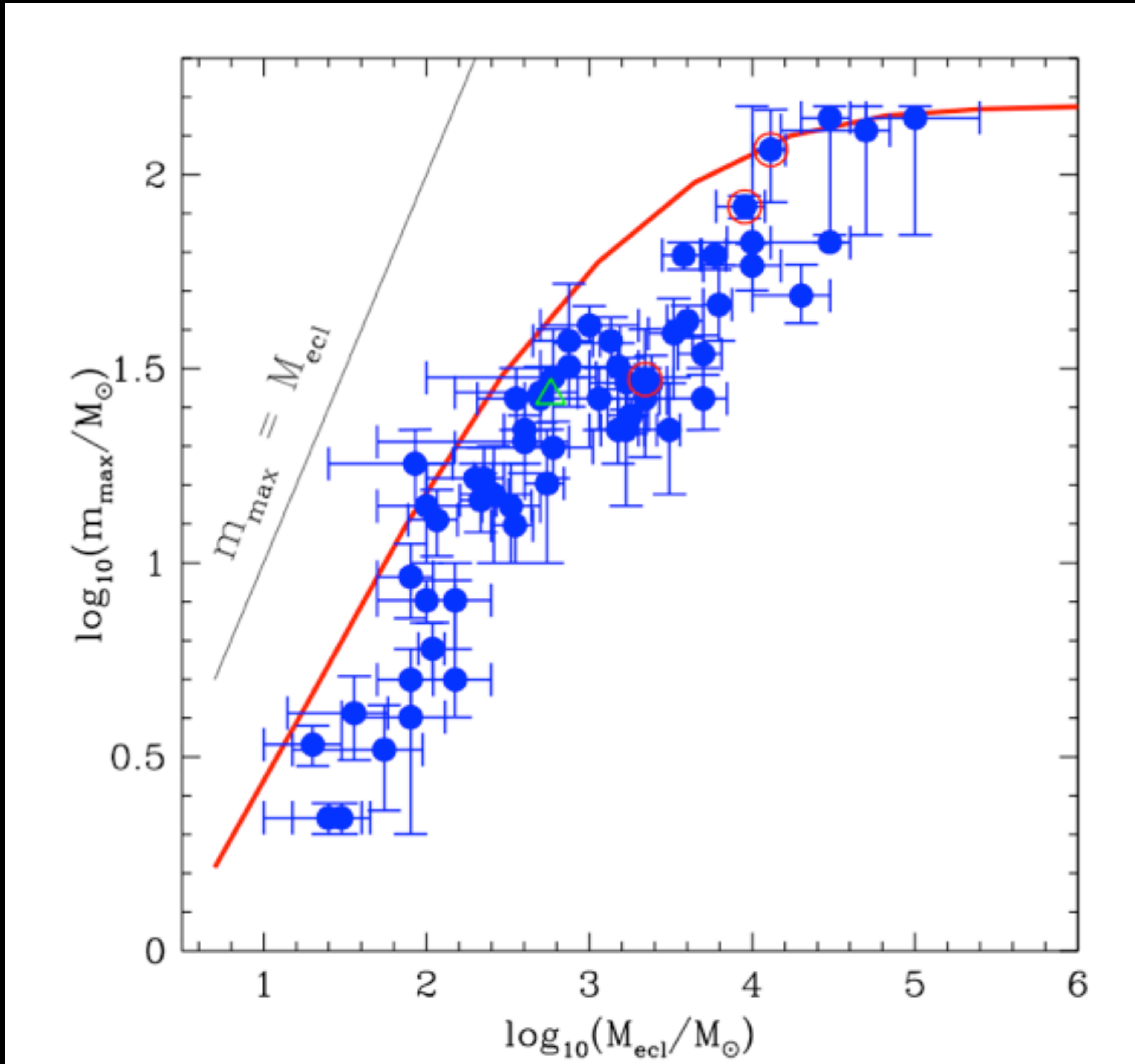


Figer (2005) Nature 434, 592

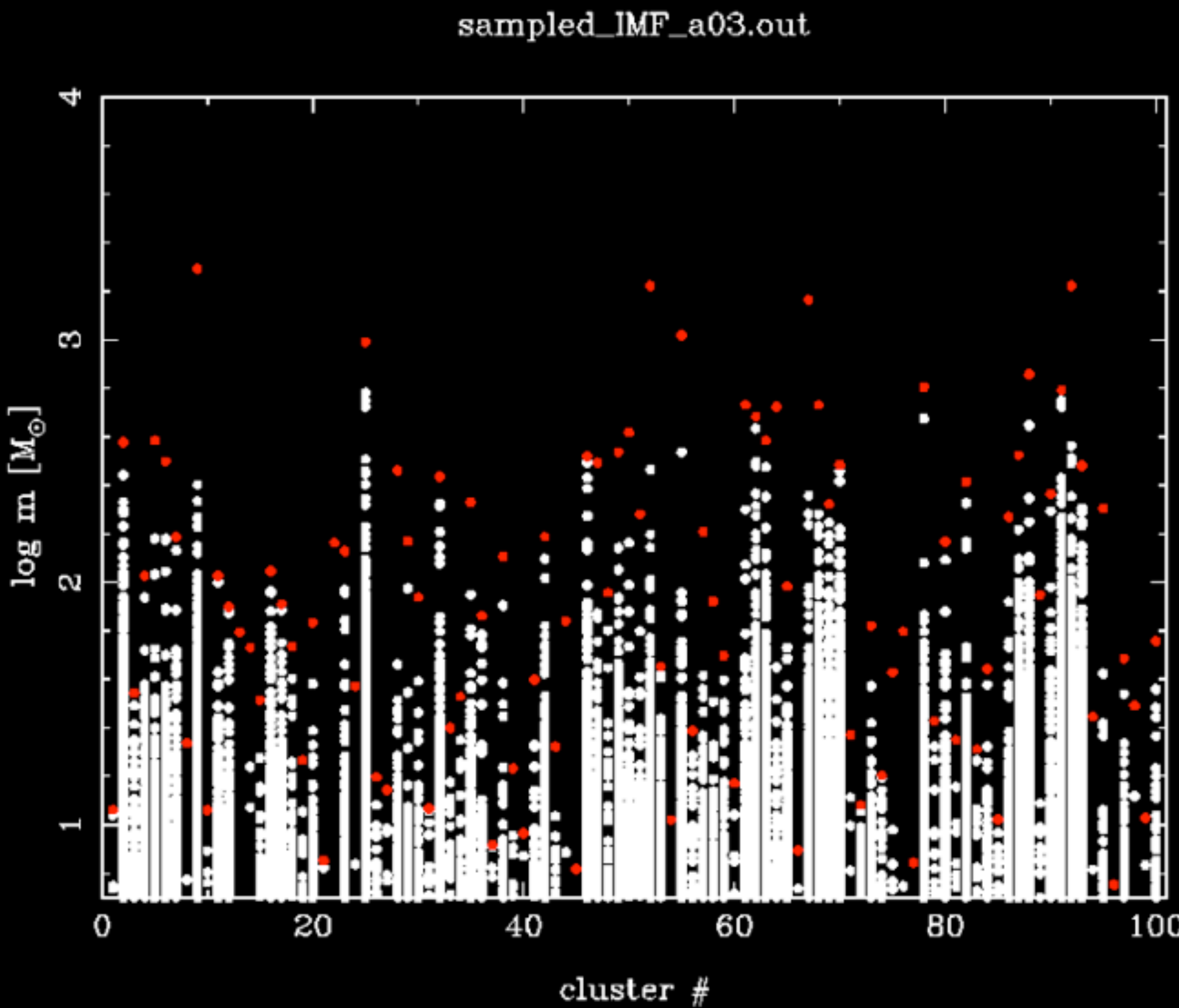


One expects a correlation between the total cluster mass and the maximum observed stellar mass

Elmegreen (2000) ApJ 539, 432

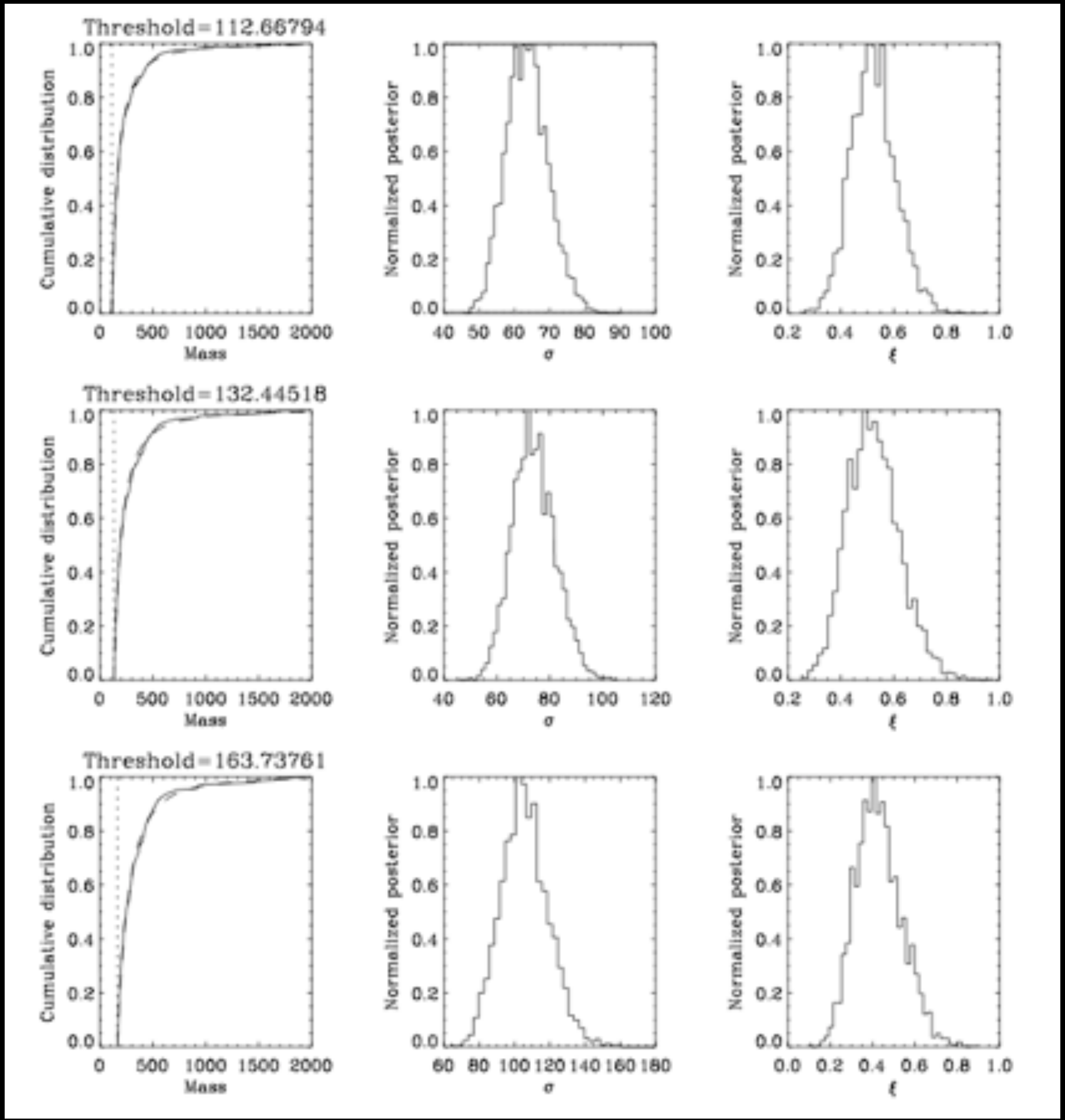


Weidner, Kroupa & Bonnell (2010) MNRAS 401, 275



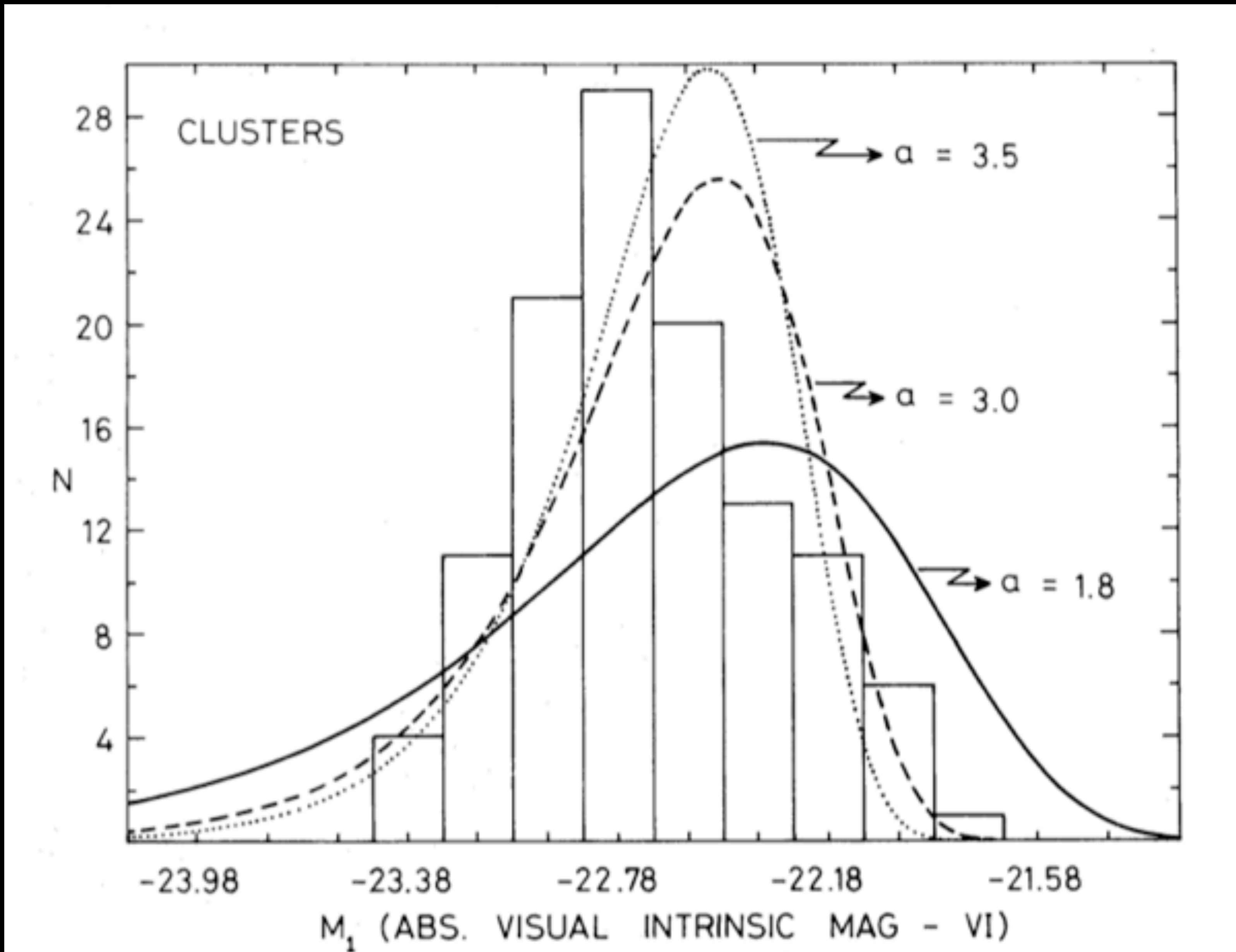
sampled_IMF_a03.out
log m [$M_\odot$]
cluster #
0
20
40
60
80
100
1
2
3
4

Posterior PDFs for scale and slope of IMF for massive stars



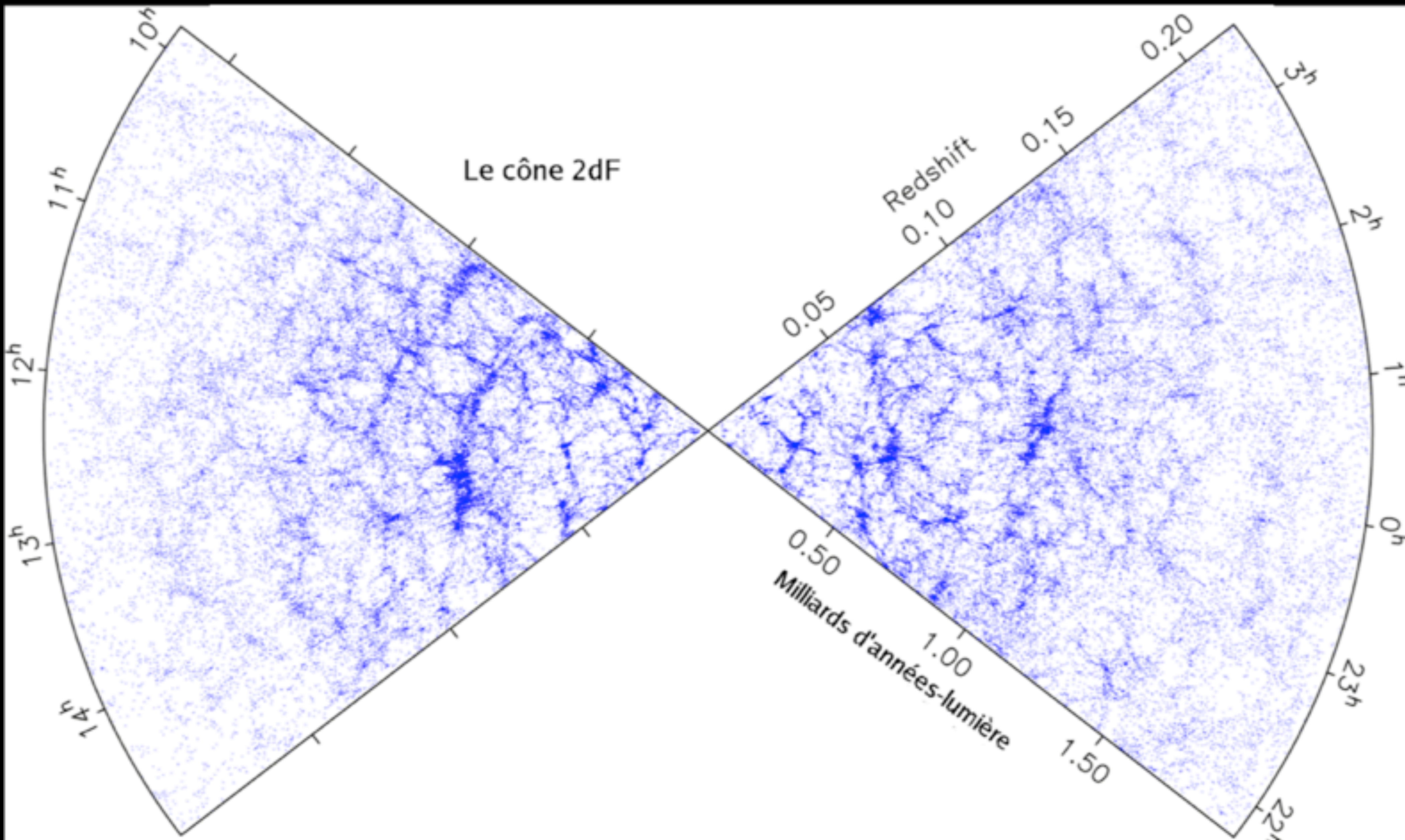
Valls-Gabaud & Asensio Ramos (2012)

Distribution function of brightest galaxies in clusters

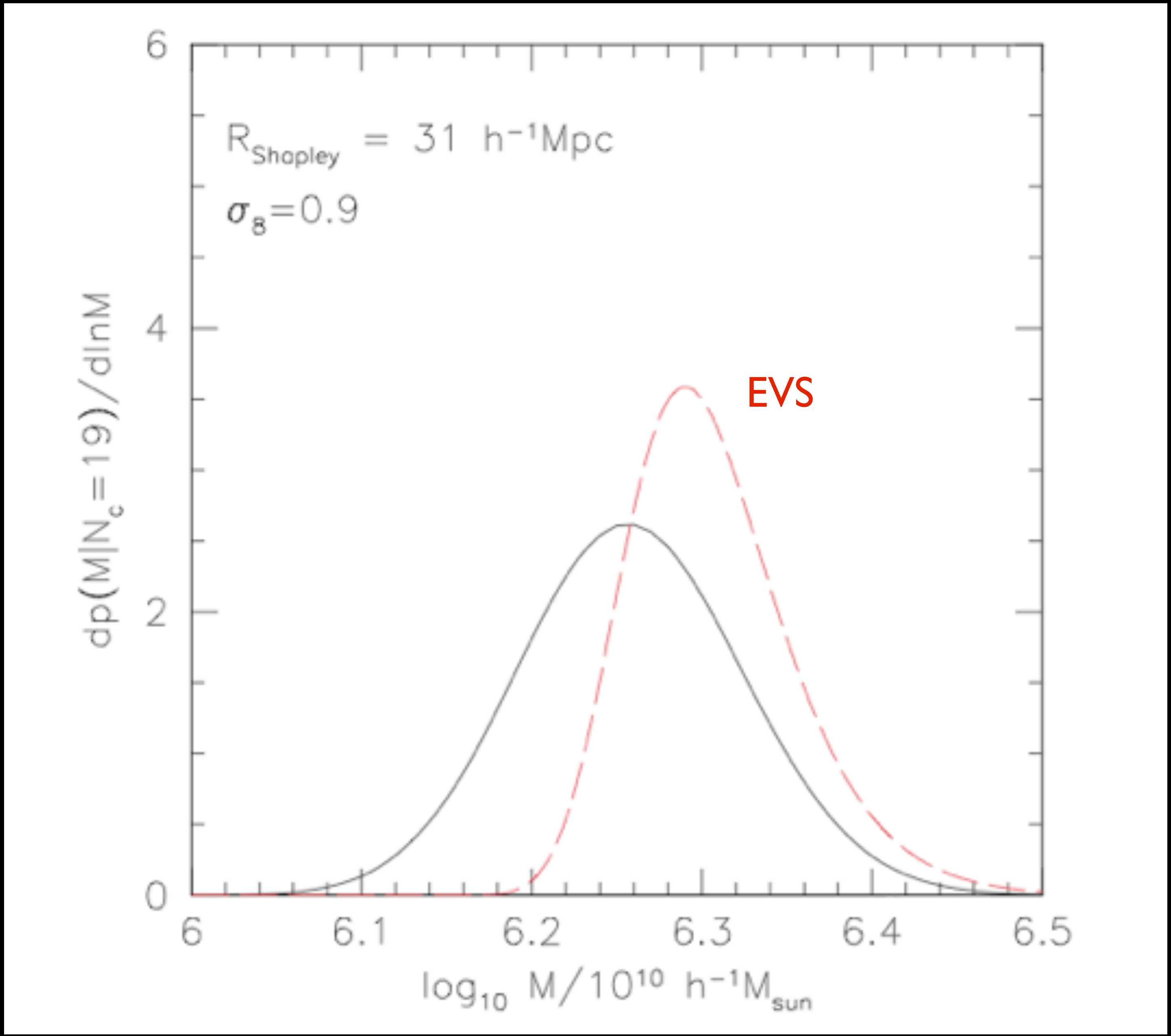


Bhavsar & Barrow (1985) MNRAS 213, 857

Probability of finding the largest structures in a cosmological volume

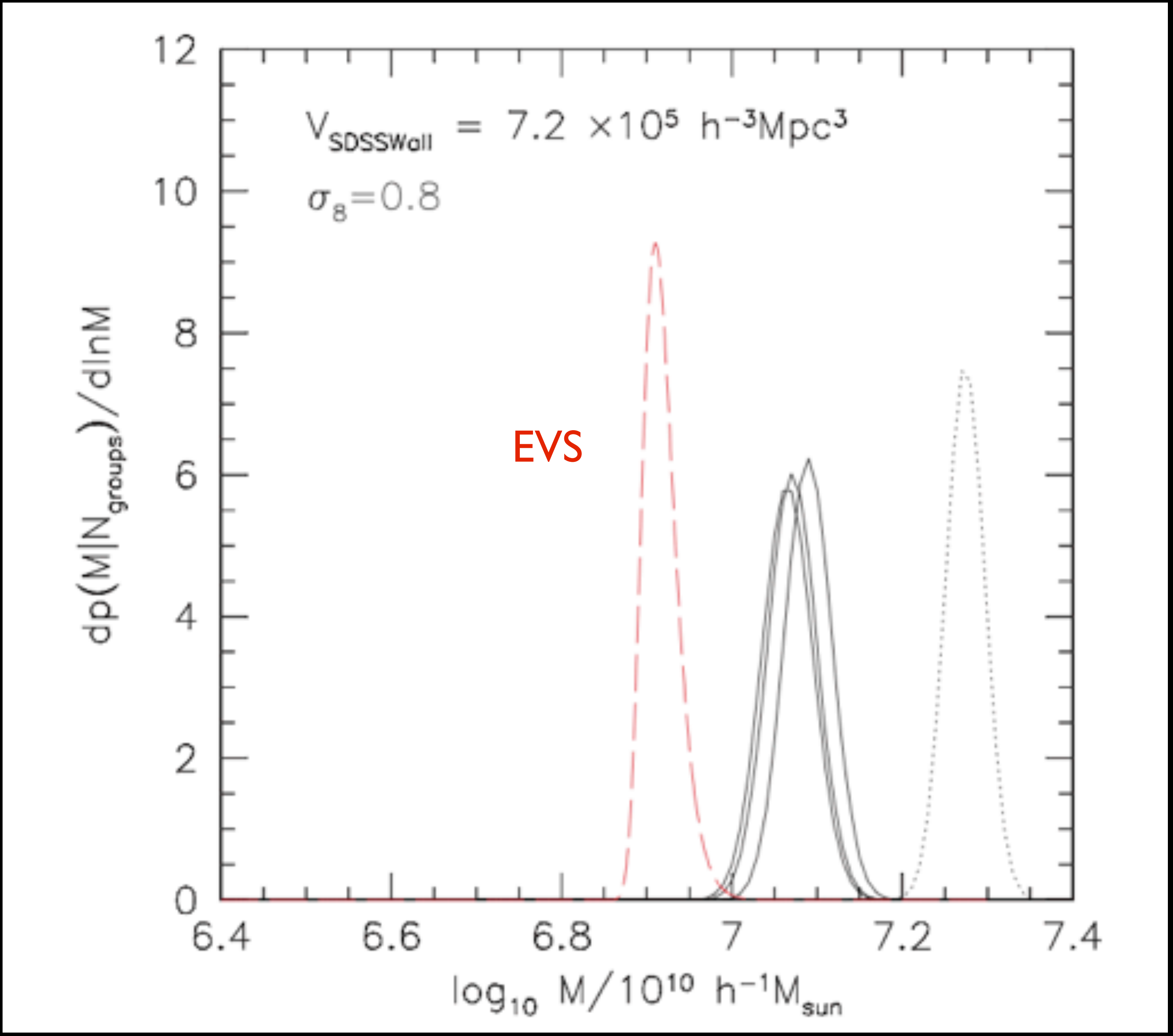


Probability of finding a Shapley supercluster

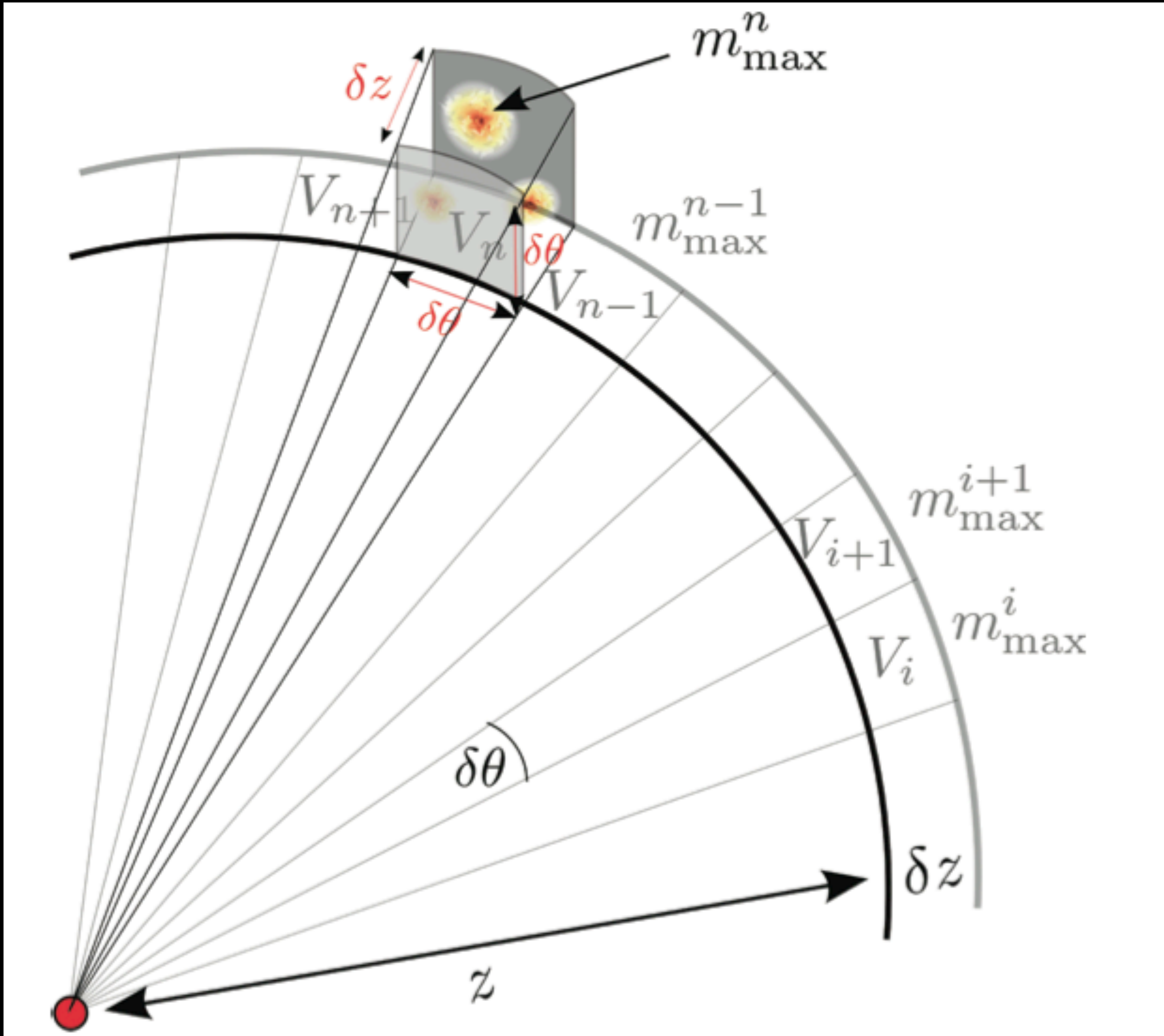


Sheth & Diaferio (2011) MNRAS 417, 2938

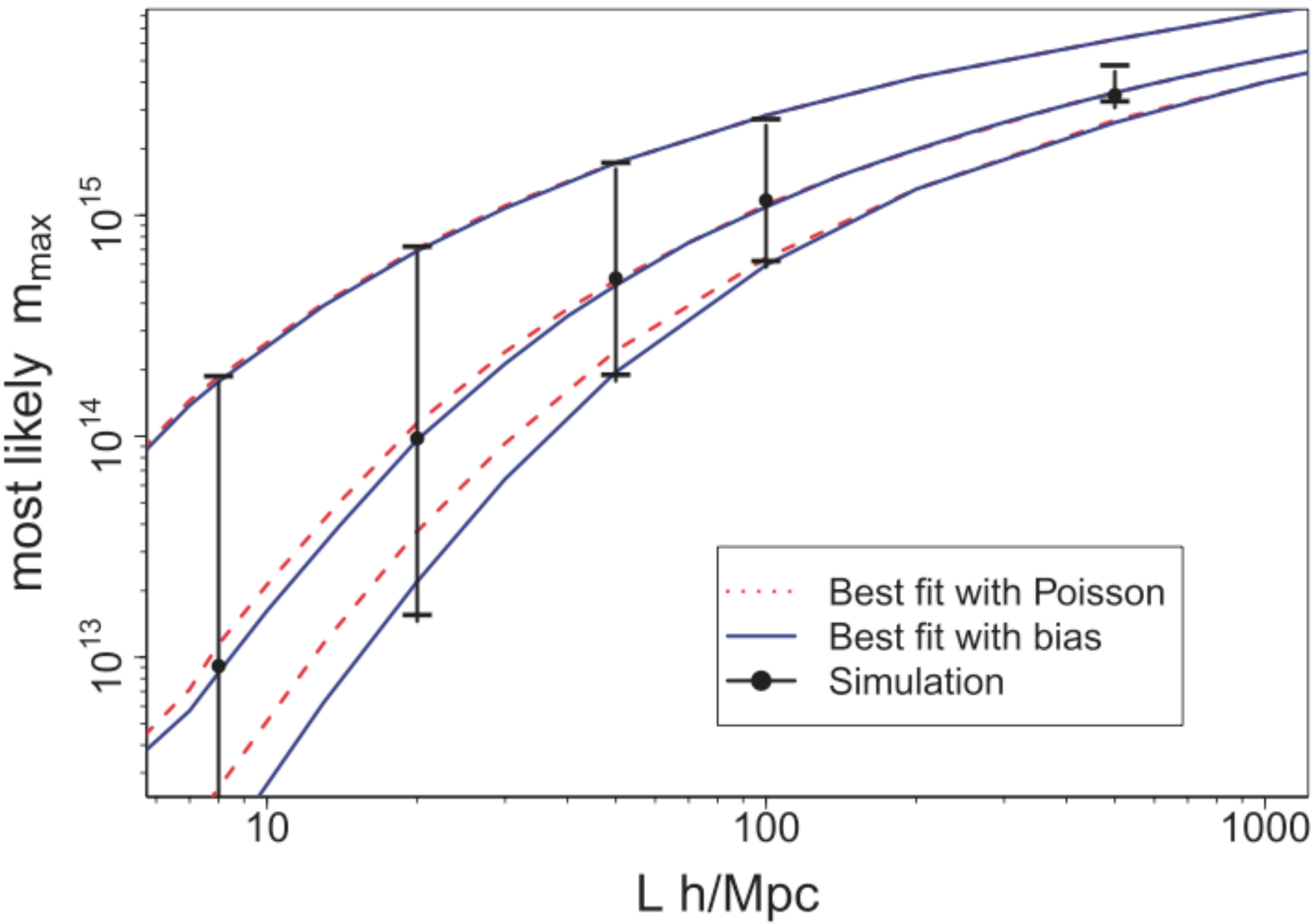
Probability of finding the SDSS Great Wall



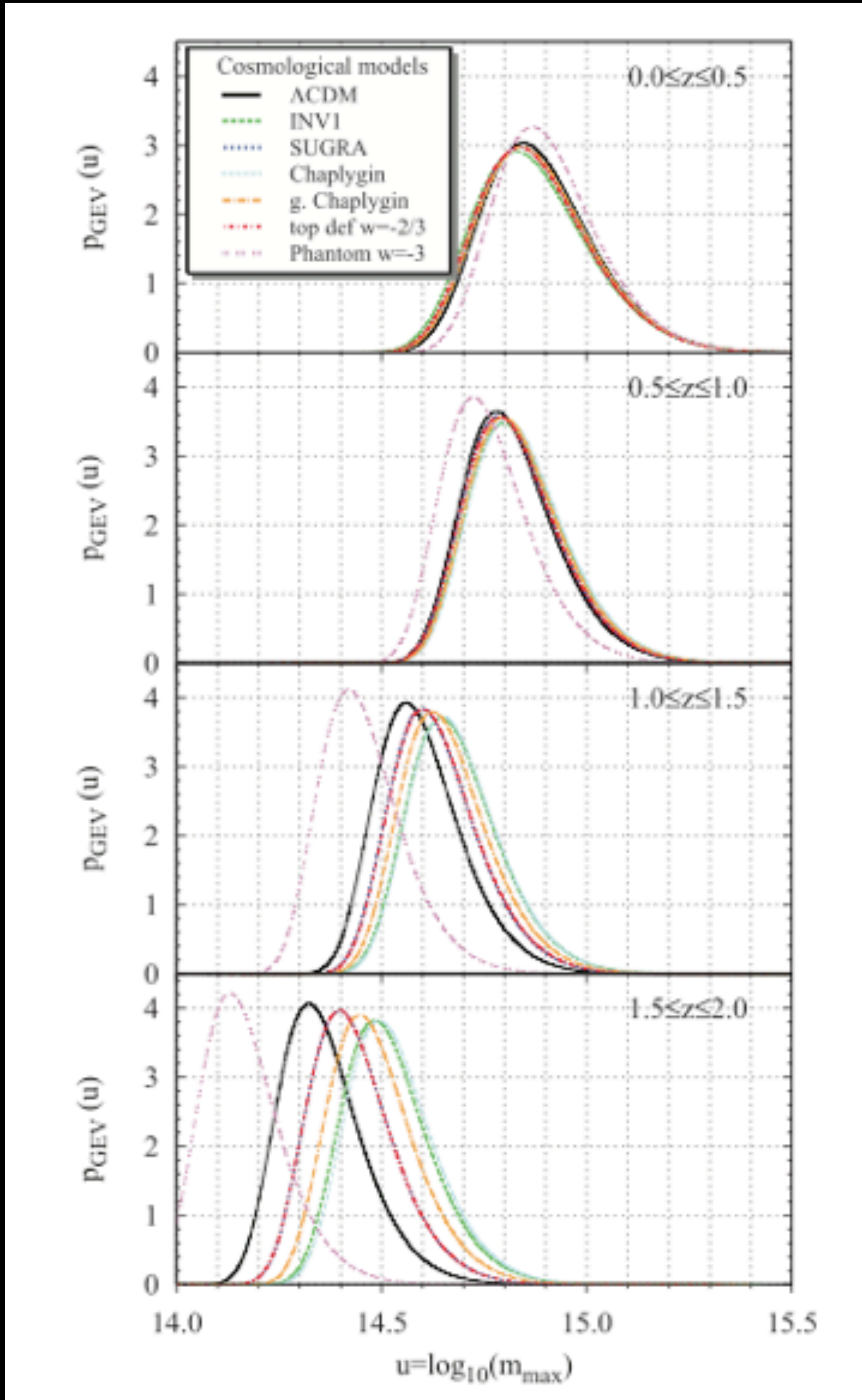
Sheth & Diaferio (2011) MNRAS 417, 2938



$m^n_{\max}$
δz
V_{n+1}
V_n
$\delta\theta$
$m^{n-1}_{\max}$
V_{n-1}
$\delta\theta$
$m^{i+1}_{\max}$
V_{i+1}
$m^i_{\max}$
V_i
$\delta\theta$
δz
z



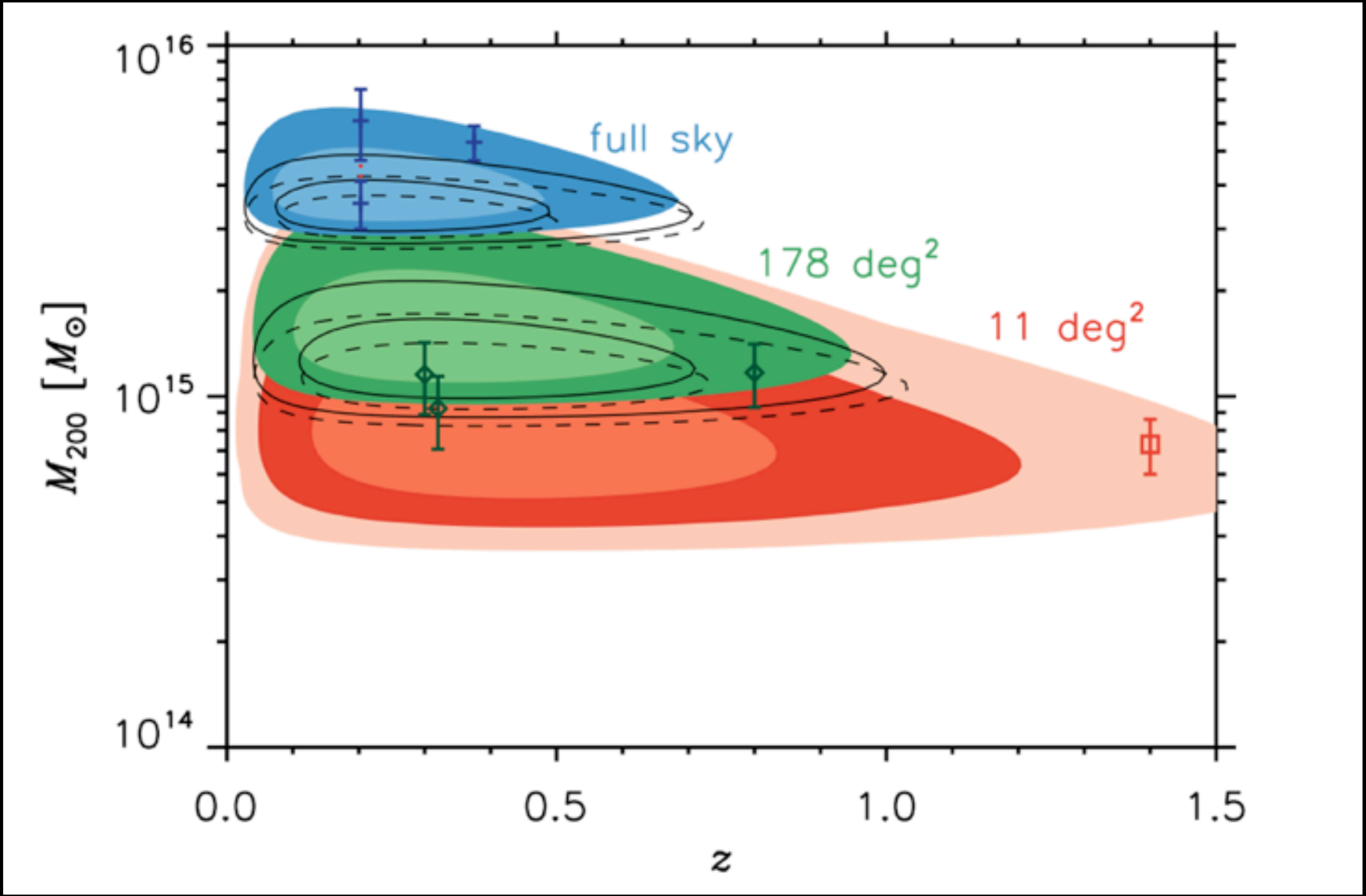
Davis et al. (2011) MNRAS 413, 2087



PDF of the most massive clusters of galaxies as a function of redshift and equation of state of Dark Energy

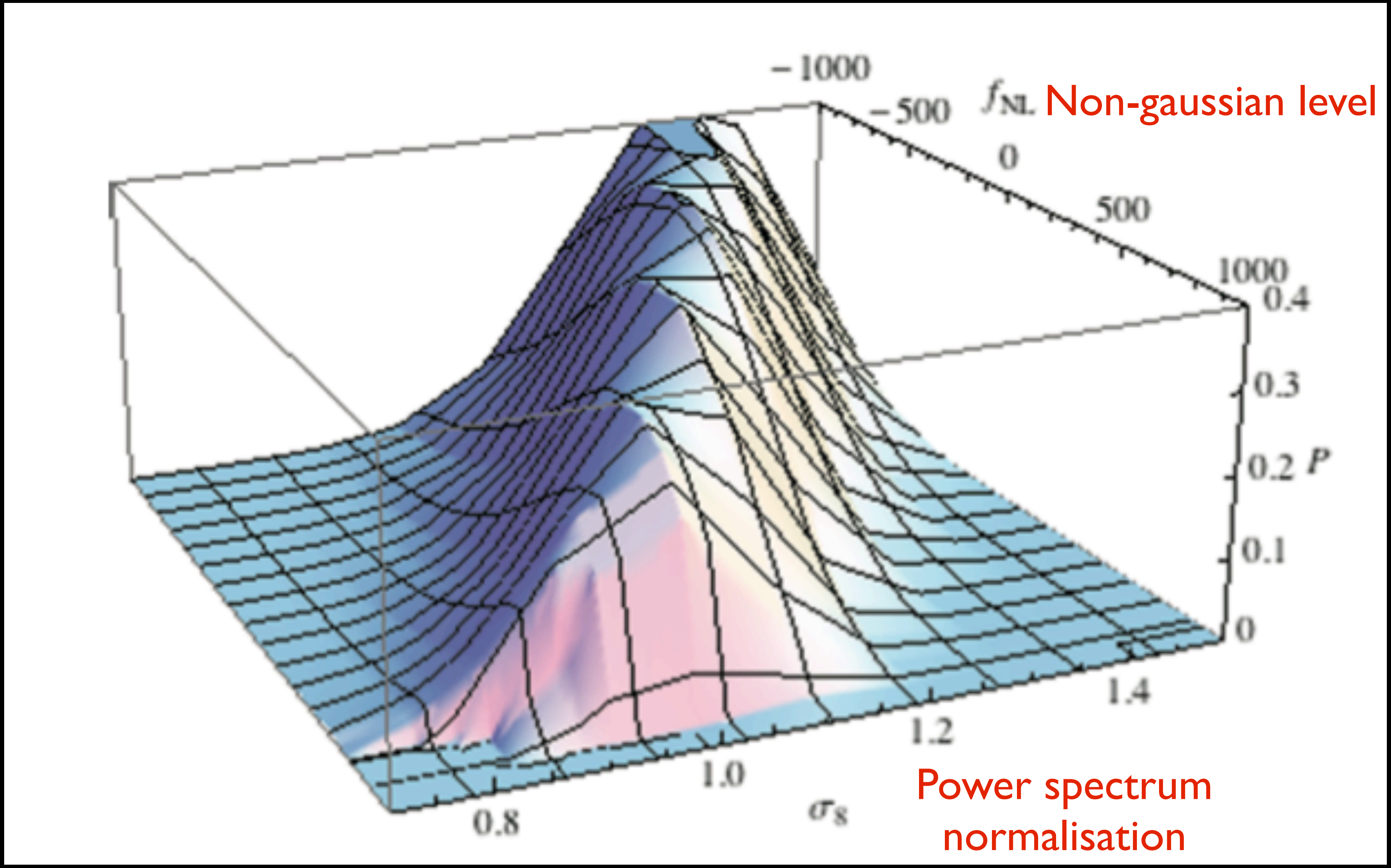
Waizmann, Ettori & Moscardini (2011) MNRAS 418, 456

Most massive galaxy clusters expected in a given survey



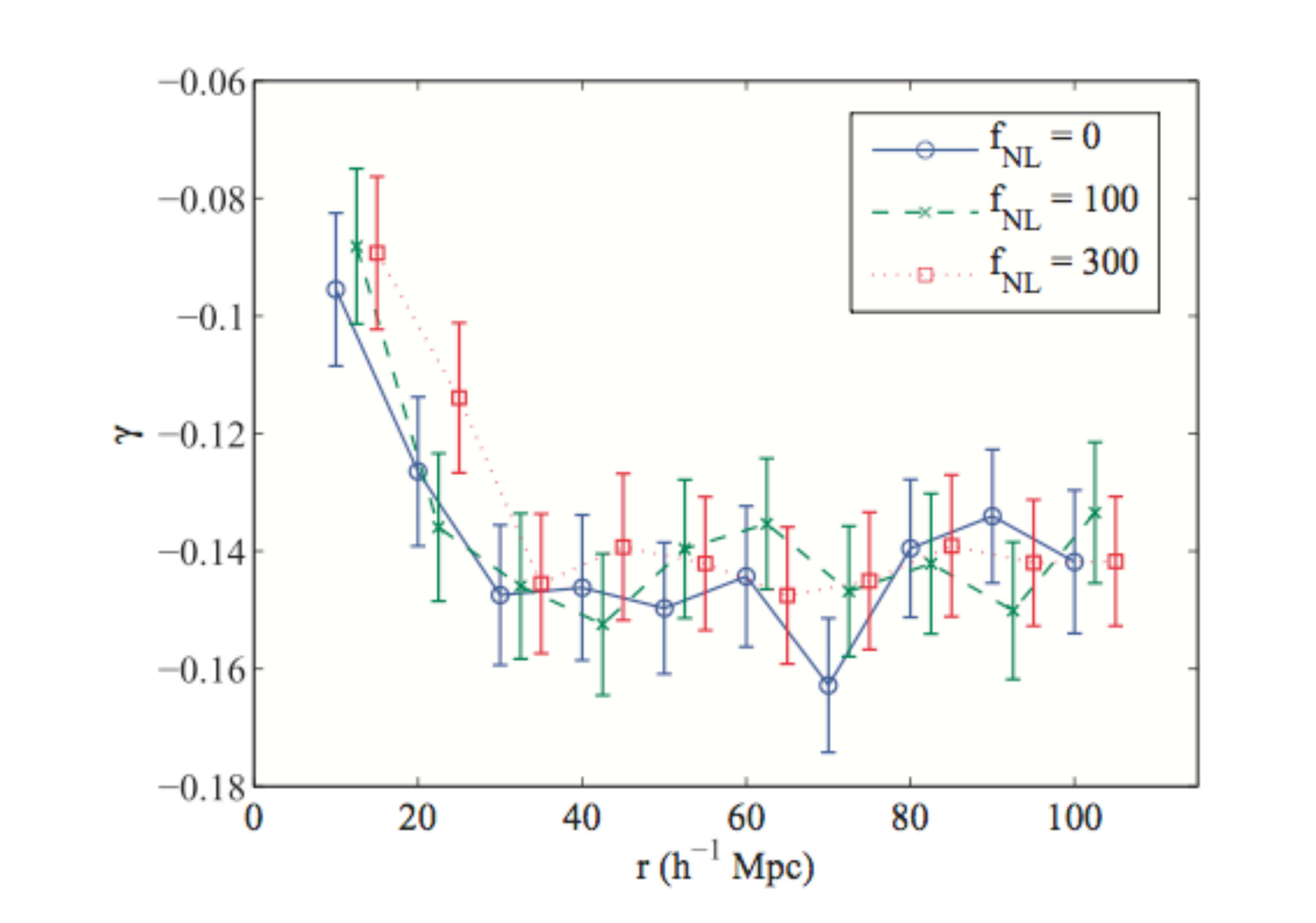
Holz & Perlmutter (2010) arXiv:1004.5349

XMMU J2235.3-2557: a massive cluster
$M_{324} = (6.4 \pm 1.2)\ 10^{14}\ M_\odot$ at $z=1.4$



Cayon, Gordon & Silk (2011) MNRAS 415, 849

.... but the shape parameter of the EVS of the halo mass function does not discriminate non-gaussianity



Harrison & Coles (2011) MNRAS 418, L20

Testing General Relativity

Abundance

$$n(M,z) = \int_0^M f(\sigma)\,\frac{\bar{\rho}_{\rm m}}{M'}\,\frac{d\ln\sigma^{-1}}{dM'}\,dM'$$

$$\sigma^2(M,z) = \frac{1}{2\pi^2}\int_0^\infty k^2 P(k,z)|W_{\rm M}(k)|^2 dk$$

$$P(k,z) \propto k^{n_{\rm s}} T^2(k,z_{\rm t}) D(z)^2$$

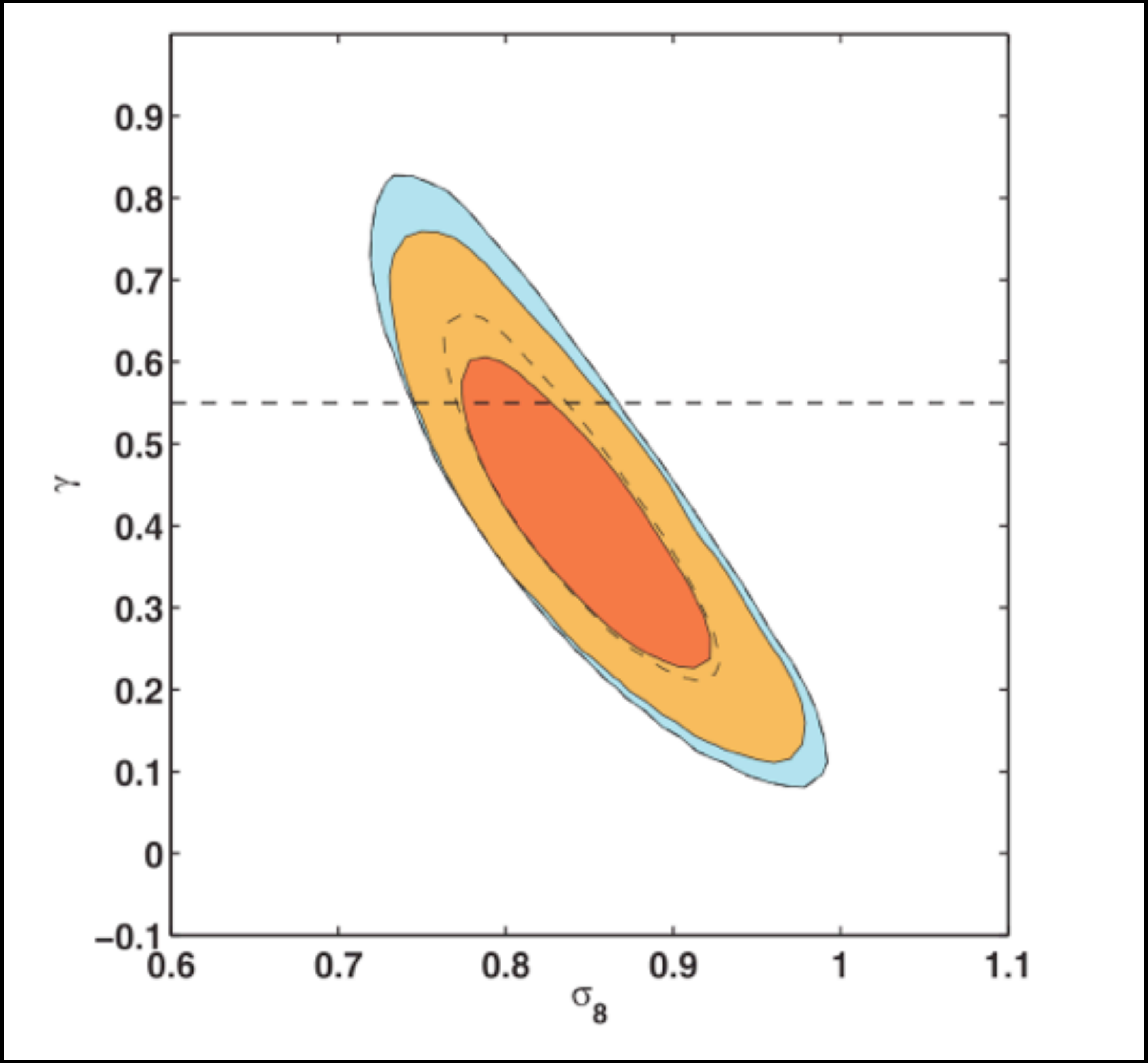
$$D(z) \equiv \frac{\delta(z)}{\delta(z_{\rm t})}$$

$$\frac{d\ln\delta}{d\ln a} = \Omega_{\rm m}(a)^\gamma$$

$$\Omega_{\rm m}(a) = \Omega_{\rm m} a^{-3}/E(a)^2$$

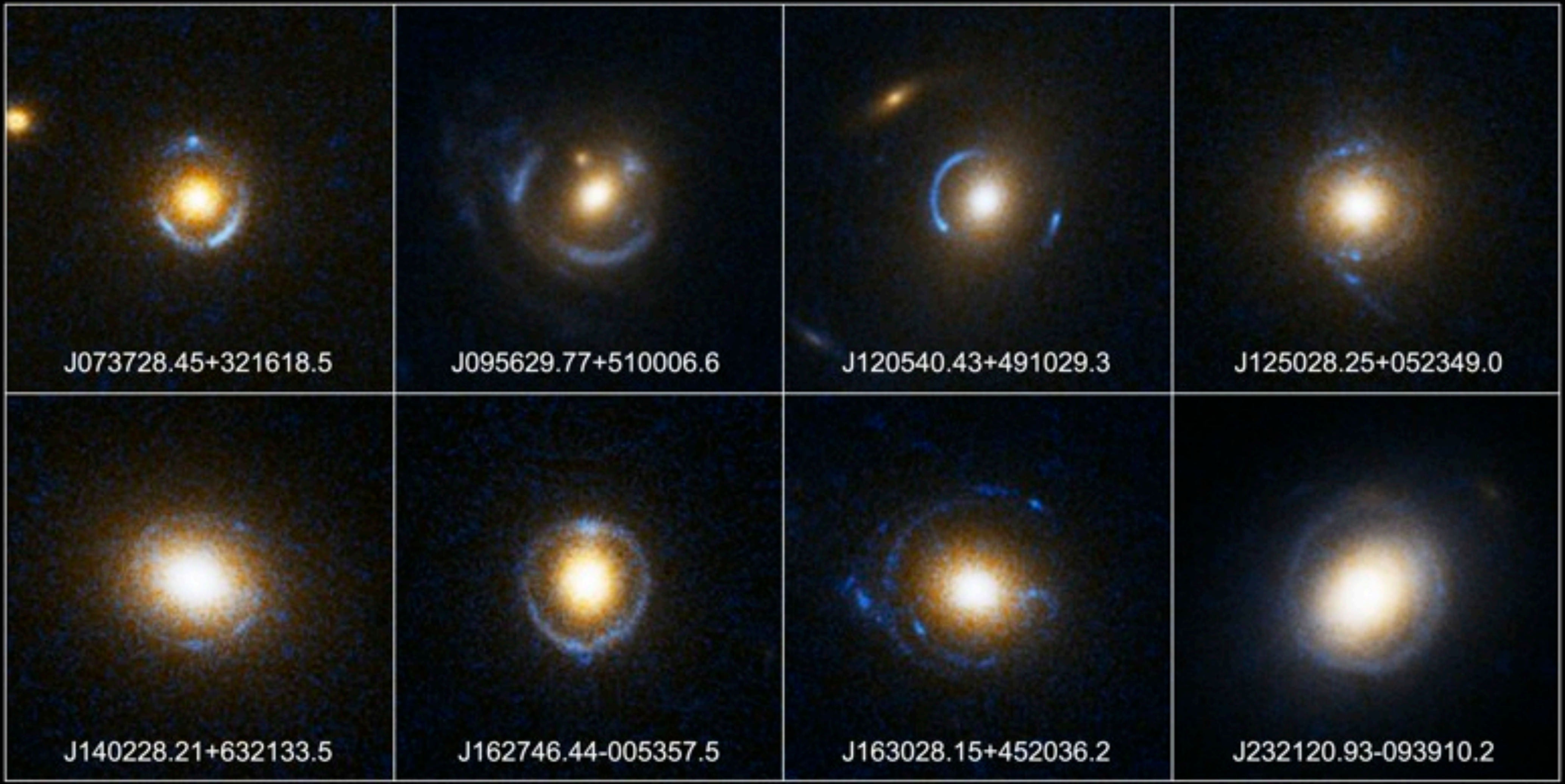
GR $\gamma \sim 0.55$

$$E(a) = \left[\Omega_{\rm m}\,a^{-3} + \Omega_{\rm de}\,a^{-3(1+w)} + \Omega_{\rm k}\,a^{-2}\right]^{1/2}$$



Rapetti et al. (2011) Prog. Theor. Phys. Suppl. 190, 179

Searching for strong gravitational lenses



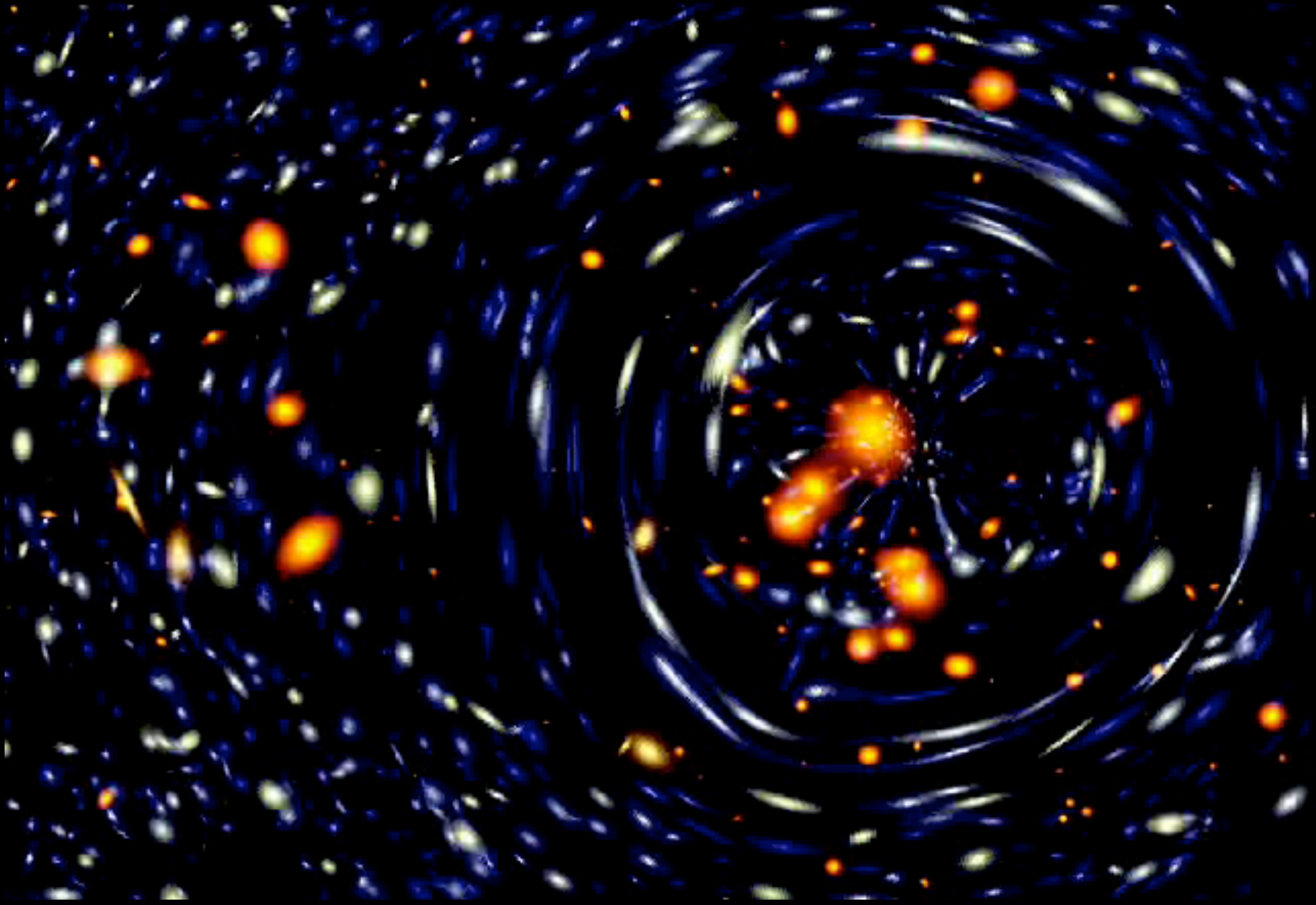
Einstein Ring Gravitational Lenses

Hubble Space Telescope • Advanced Camera for Surveys

NASA, ESA, A. Bolton (Harvard-Smithsonian CfA), and the SLACS Team

STScI-PRC05-32

Gravitational lens effect by a massive cluster



The size of the Einstein radius can be measured directly from the positions of the background galaxies.

It can also provide us with the depth of the potential well (i.e. dark matter content) as well as the cosmological parameters of a given metric if the redshifts of the background galaxies can also be measured:

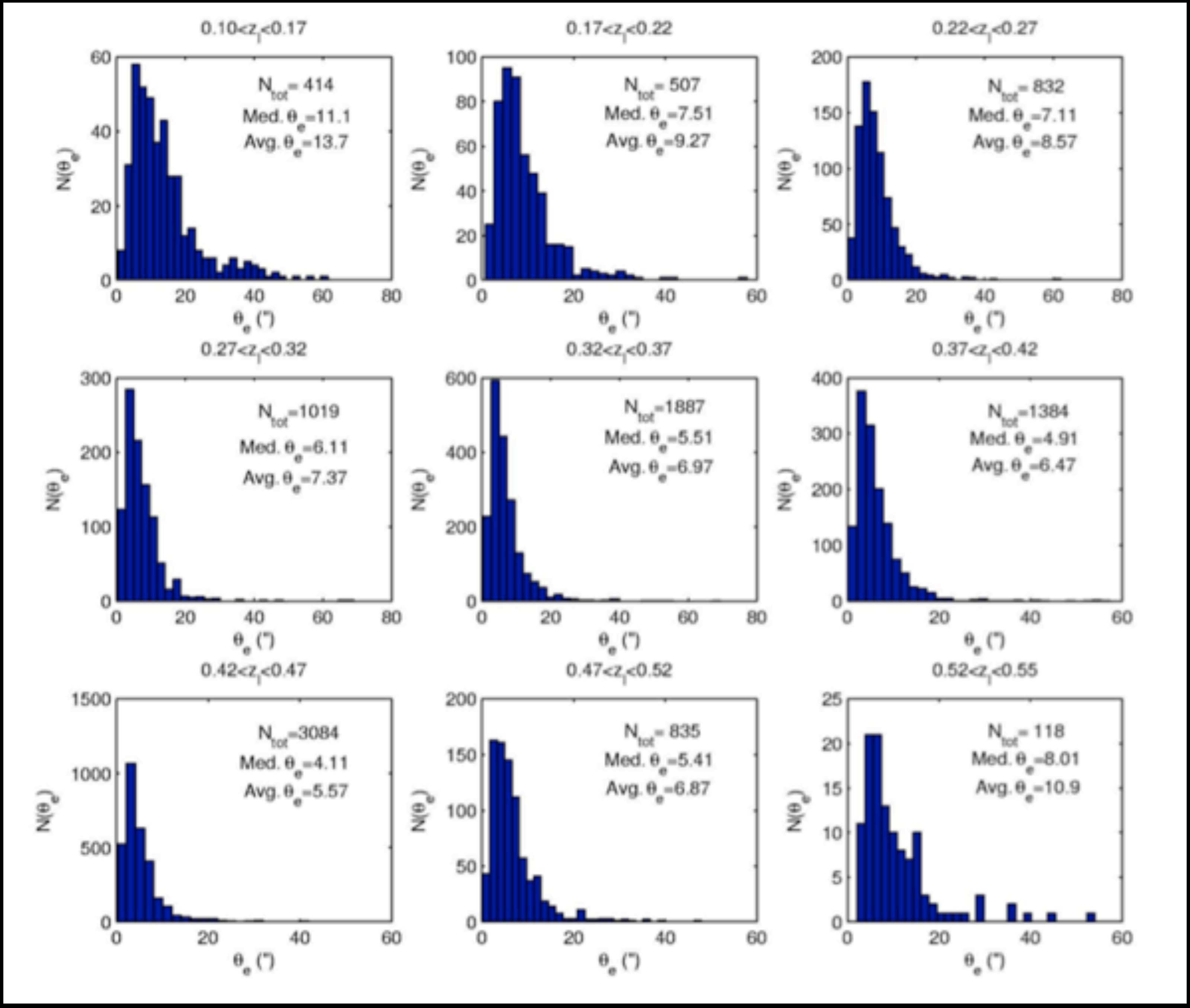
$$\theta_E = 4\pi \left(\frac{\sigma_{SIS}}{c}\right)^2 \frac{D_{LS}}{D_{OS}}$$

positions of the images

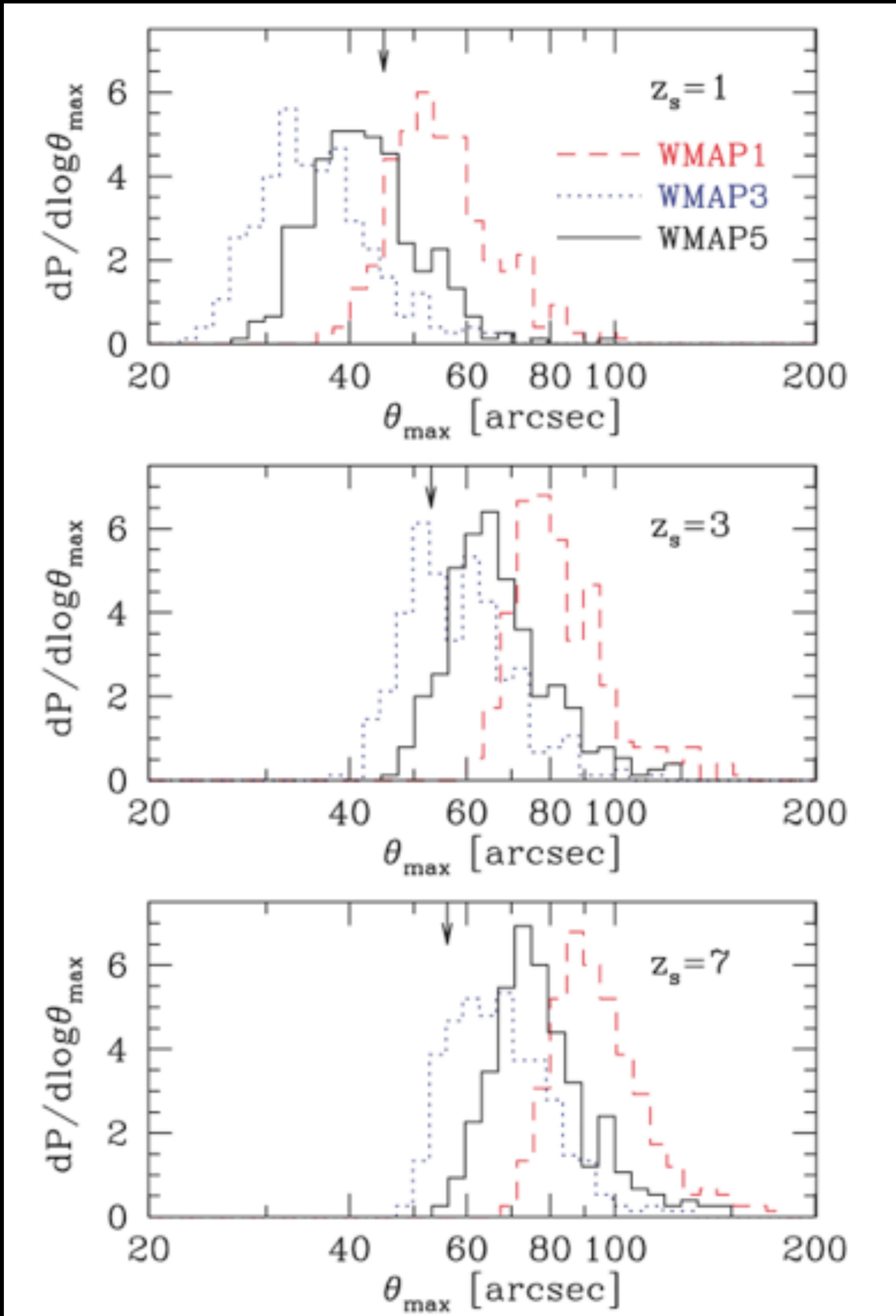
spectrum of the lens

spectrum of the source

Inferring Einstein radii for 10,000 SDSS clusters



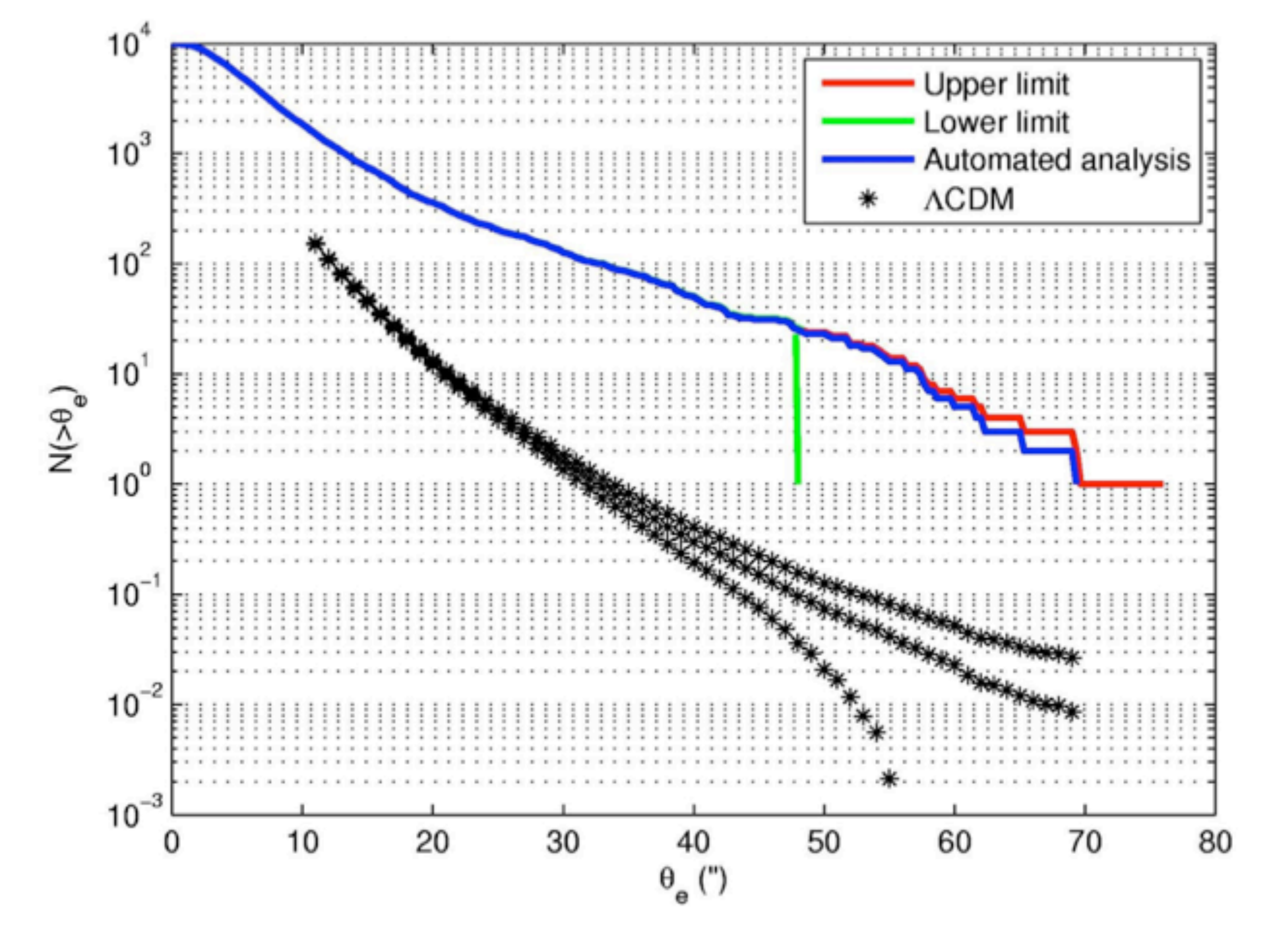
Zitrin et al. (2011) arXiv:1105.2295



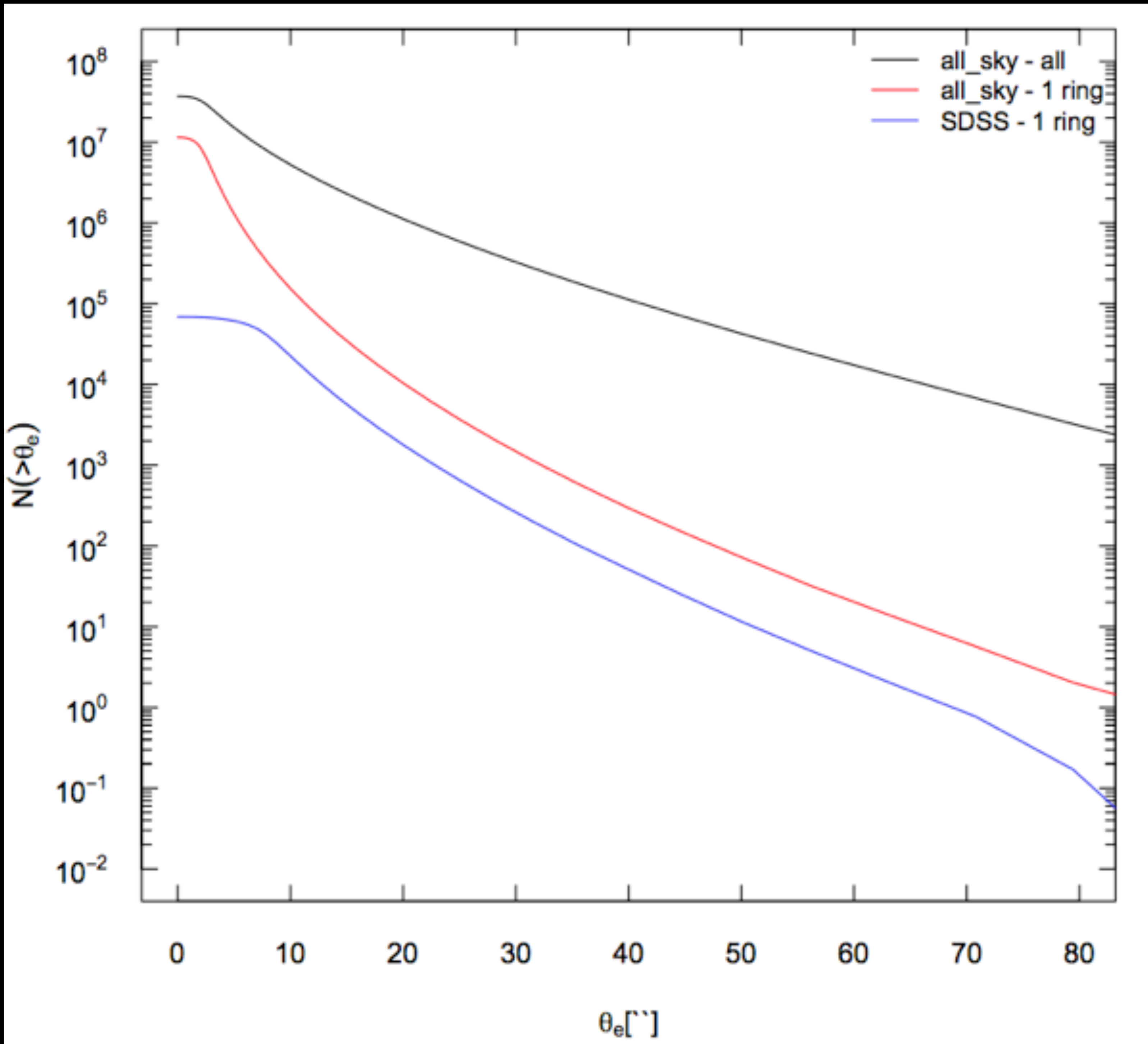
Theoretical expectations

Oguri & Blandford (2009)
MNRAS 392, 930

Tension with ΛCDM scenario?



Zitrin et al. (2011) arXiv:1105.2295



Vega & Valls-Gabaud (2012)

Summary

Methodology

- Proper statistical theory for extreme events
- Bayesian thresholding estimates work best
- Wide applications to astrophysics and cosmology

Limitations

- Assessment of selection biases in samples
- May require extensive testing with simulations