PROBABILITY MODELS AND UNCERTAINTY IN HYDROLOGY: SOME PERSONAL REFLECTIONS

Presentation to the New Zealand Hydrological Society Conference
November 28, 2012

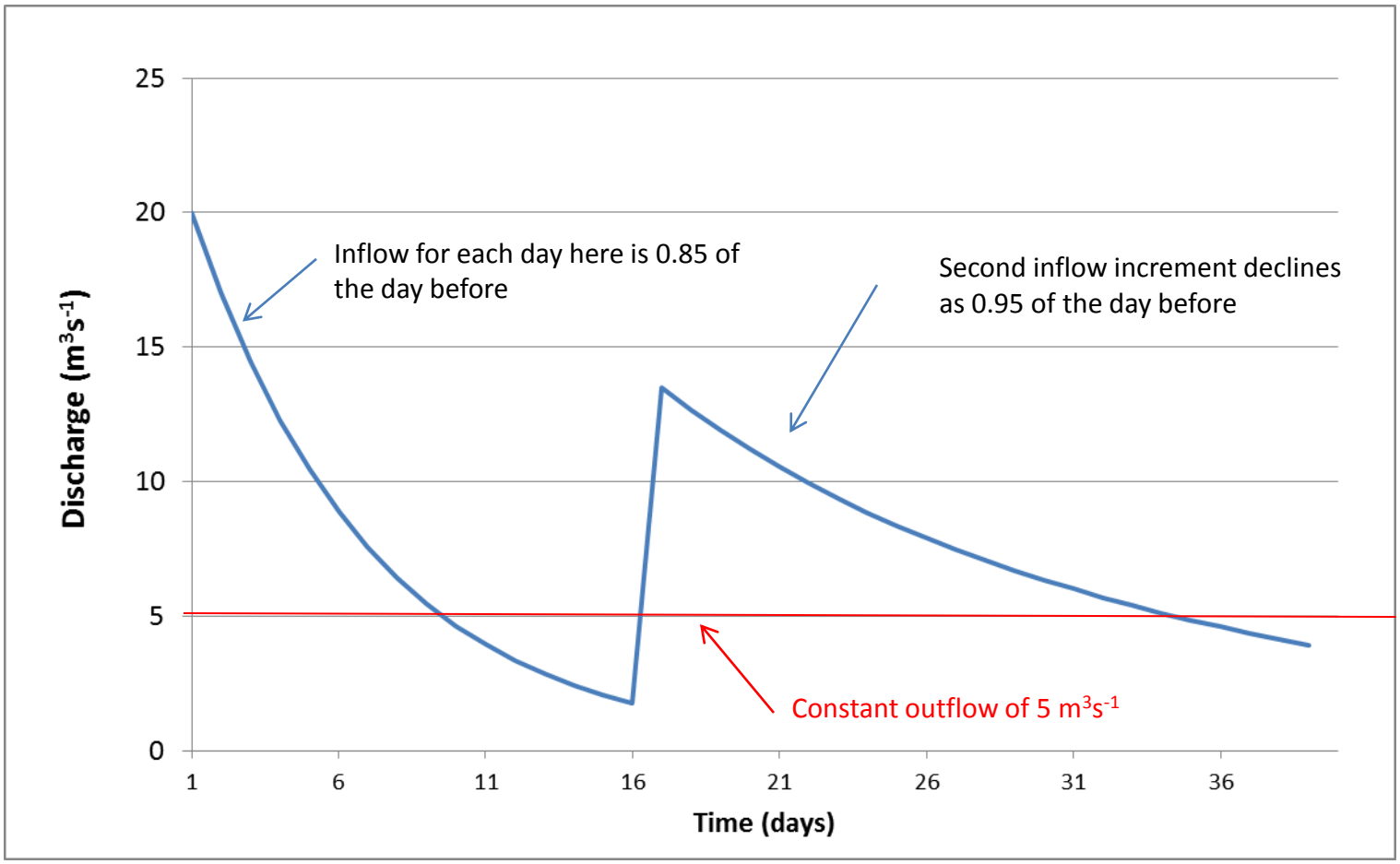
Earl Bardsley
Earth & Ocean Sciences
University of Waikato

New Zealand Hydrological Society Conference, Nelson, 27-30 November

Part 1: Uncertainty

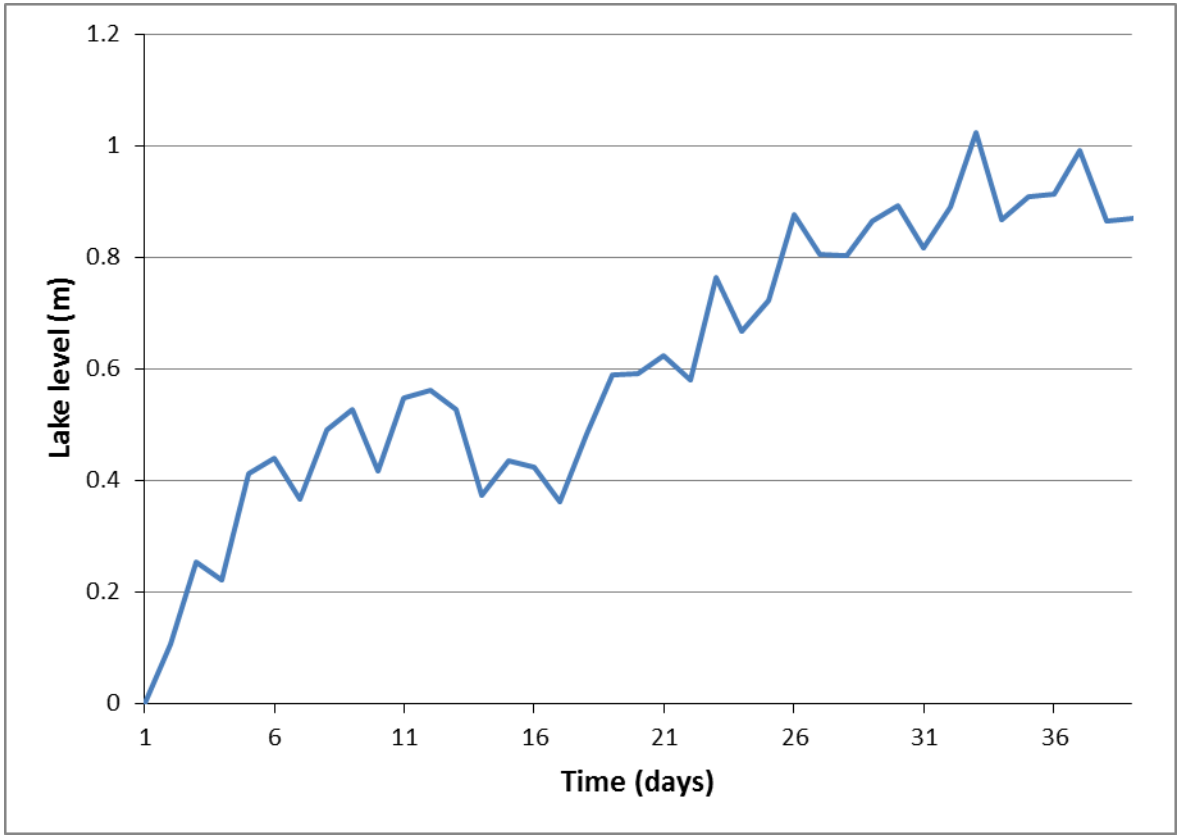
- Reduction of uncertainty with hydrological information (estimating a lake inflow hydrograph and confidence bound for a return period)
- Quantification of model validation fits – introducing a new fit measure

Simulated lake inflow hydrograph with a constant outflow



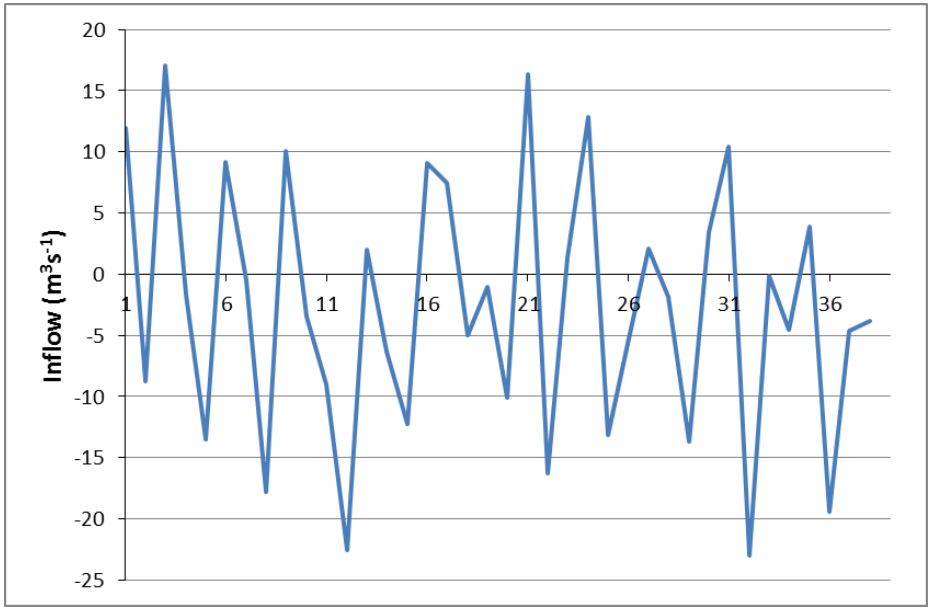
Assume lake area = 10 km^2 , next slide shows resulting lake level changes with random noise (zero mean) added…

Resulting simulated lake level time series (from arbitrary zero)



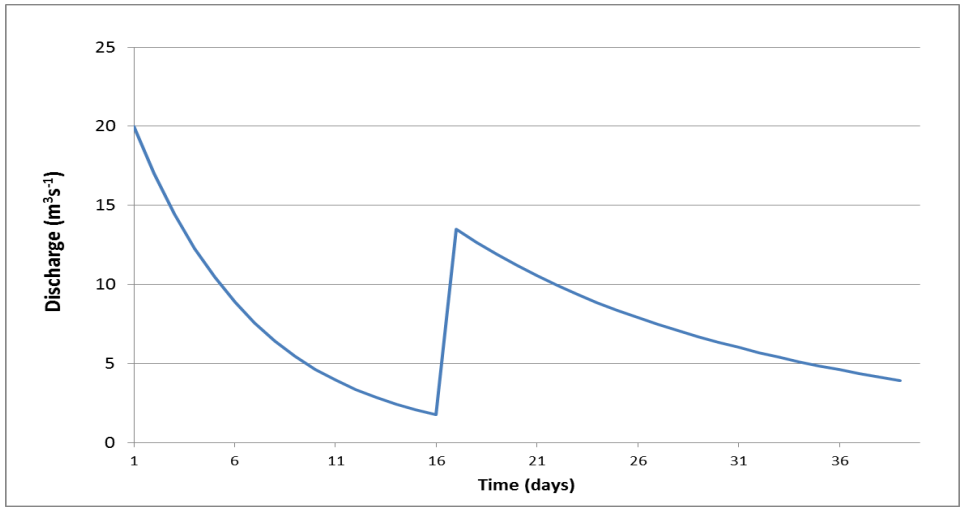
We wish to these lake levels to estimate the daily values of the inflow hydrograph.

Firstly using the water balance from the known outflow and lake level differencing..



Estimated lake inflow hydrograph from lake level differencing

Estimated mean inflow = 3.08 m^3s^{-1} (with negatives assumed to be zero).



Actual lake inflow hydrograph (as shown earlier)

Mean inflow = 7.68 m^3s^{-1}

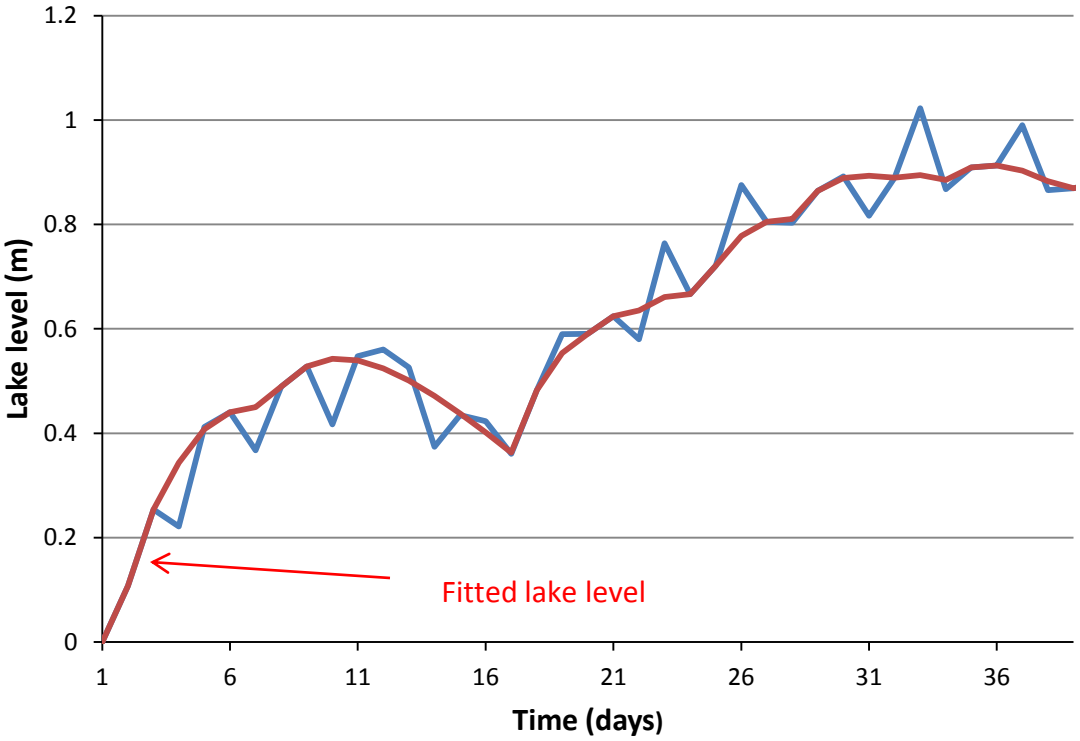
Now let's add one bit of hydrological information as a very simple constraint on the inflow hydrograph:

Inflow (current day) > 0.7 inflow (previous day)

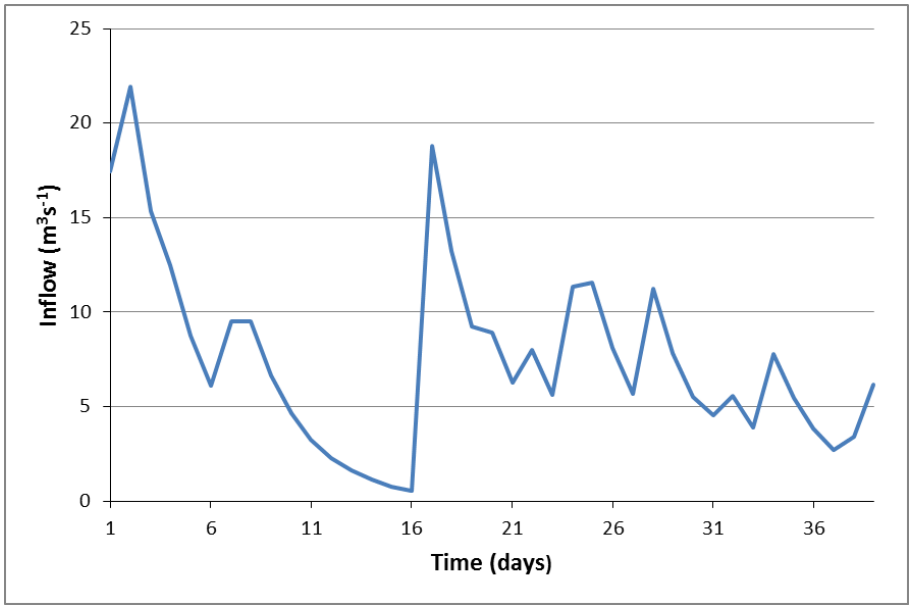
This constrains the rate of hydrograph declines, but not rises.

Linear programming specification: subject to the above constraint, create an inflow time series which matches the lake level time series as closely as possible.

(I used Excel Solver but would recommend a general package like *What's Best*.)



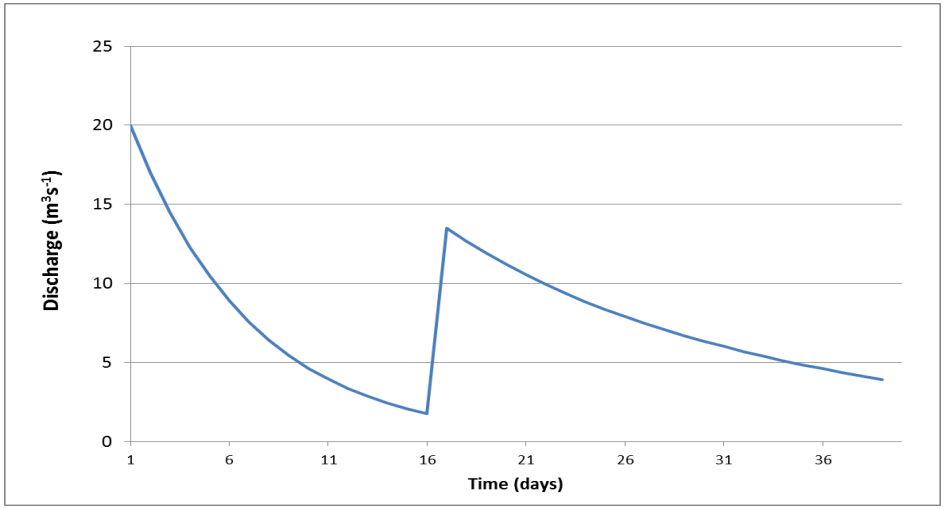
Original and fitted lake levels



Estimated lake inflow hydrograph from linearly constrained lake level fitting

Estimated mean inflow = 7.61 m^3s^{-1}.

Further improvements possible with more constraints.



Actual lake inflow hydrograph

Mean inflow = 7.68 m^3s^{-1}

* Running a standard smoother through the lake levels would not have detected the sharp inflow increase from the second inflow event

A second (brief) example of the value of further hydrological information:

Suppose we have a situation where we only know that nothing (eg extreme flood) has happened at a location in the available record of *R* years. What can we say about the return period?

"Return period" implies constant probability of .. whatever. Given this information we have:

95% lower confidence bound to return period =

$0.33R$

(Derived from a standard χ^2 relation (*J of Hydrol* 49 395-999.)

Quantifying uncertainty in model capability: a goodness of fit measure with validation data

The standard hydrological model goodness of fit measure: the Nash-Sutcliffe efficiency *E*

$$E = 1 - \frac{\sum (O_i - P_i)^2}{\sum (O_i - \bar{O})^2} \qquad -\infty < E \leq 1$$

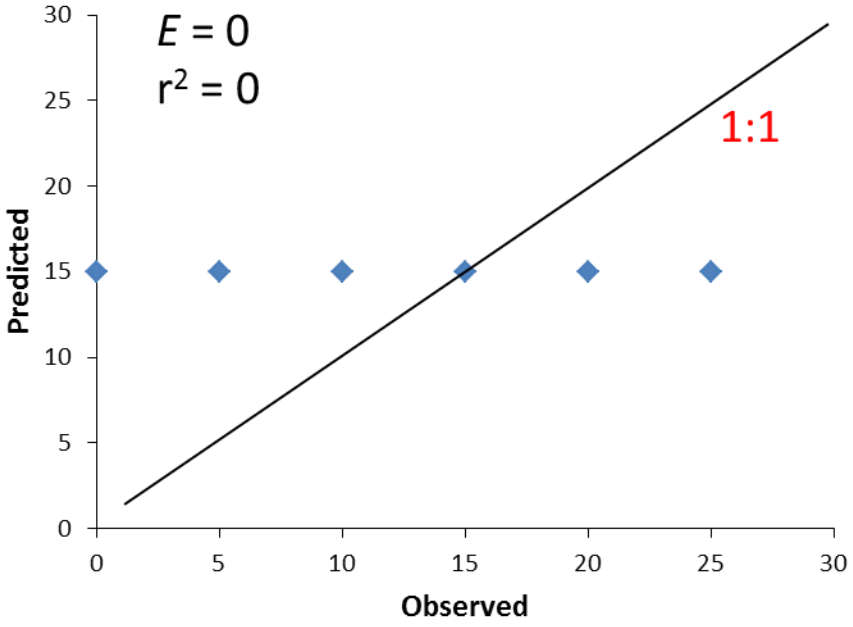
$$= 2 - 1/r^2 \quad \text{(for an unbiased model)}$$

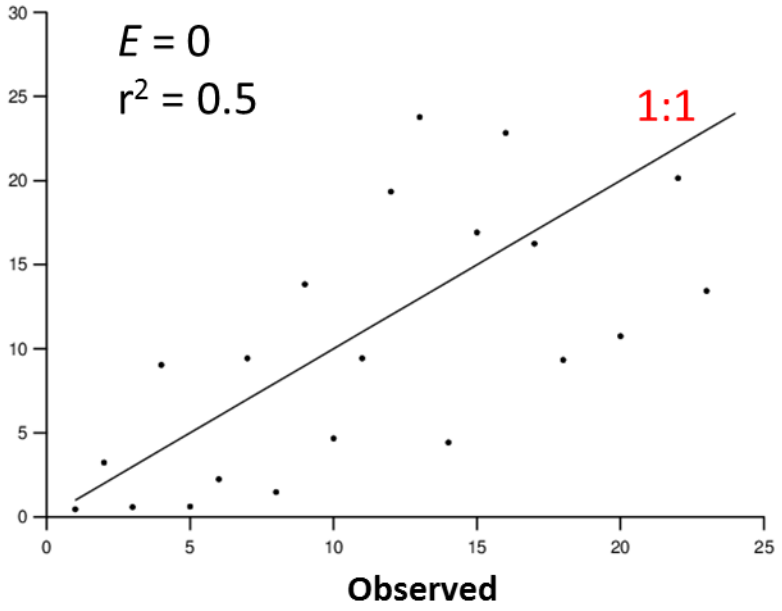
O_i = observed data
P_i = predicted from the model

E = 0.0 = “model is no better than the data mean”

But *E* has some issues…

- Unheard of outside hydrology (not easy to communicate)
- ‘Overly-optimistic’ for good fits
- Somewhat inconsistent for poor fits
- Unusual relation to r^2 in the unbiased case





Anomalous behaviour of Nash-Sutcliffe efficiency for $E = 0.0$

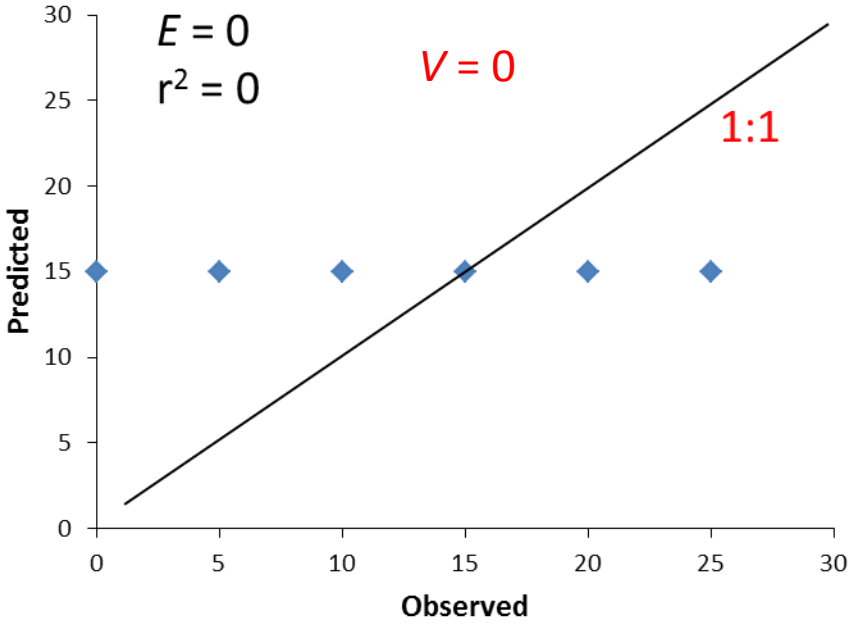
Proposed new goodness of fit index *V*

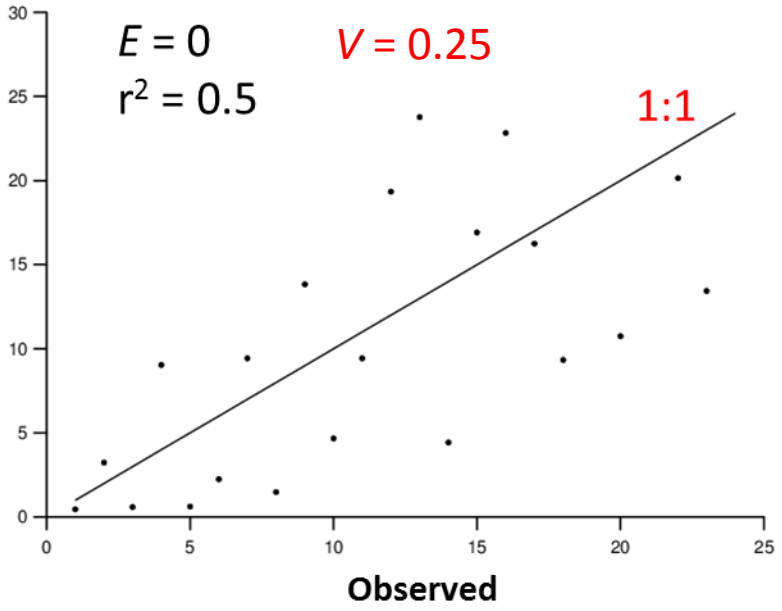
$$V = r^2 / (2 - E) \qquad 0 \le V \le 1$$

$$= r^4 \text{ (for an unbiased model)}$$

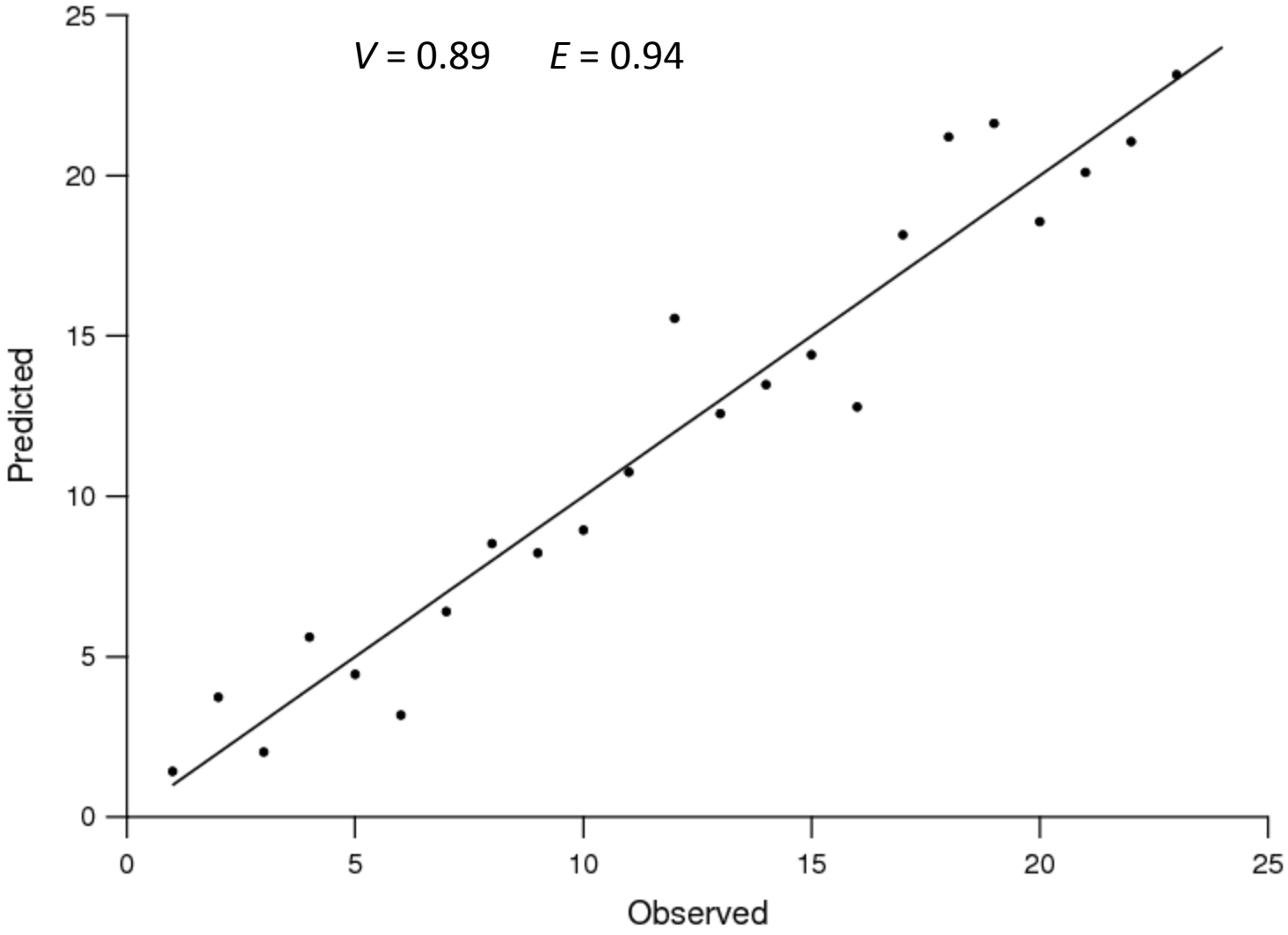
r = correlation coefficient between observed and model-predicted values

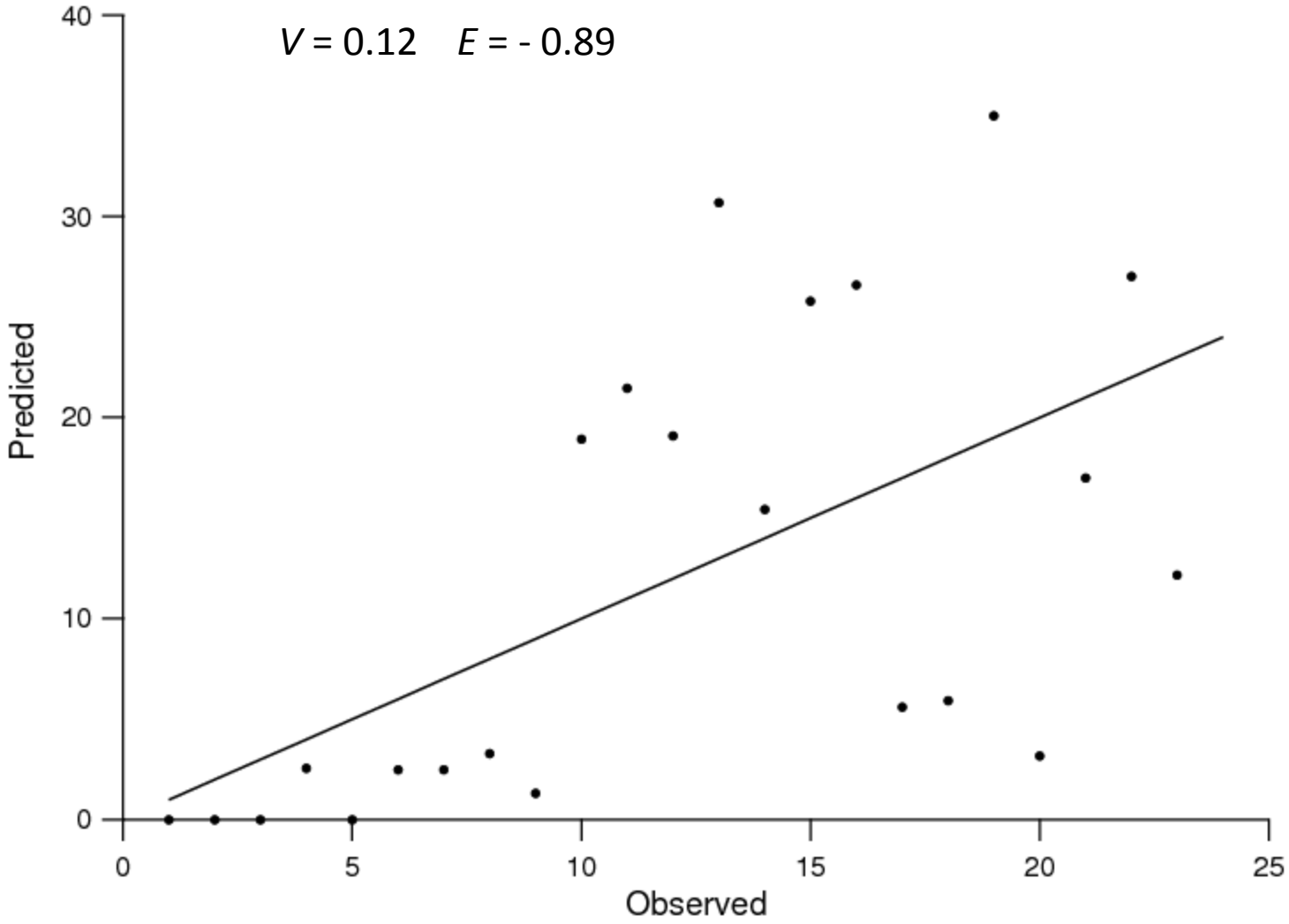
Example application of *V* to the previous data sets

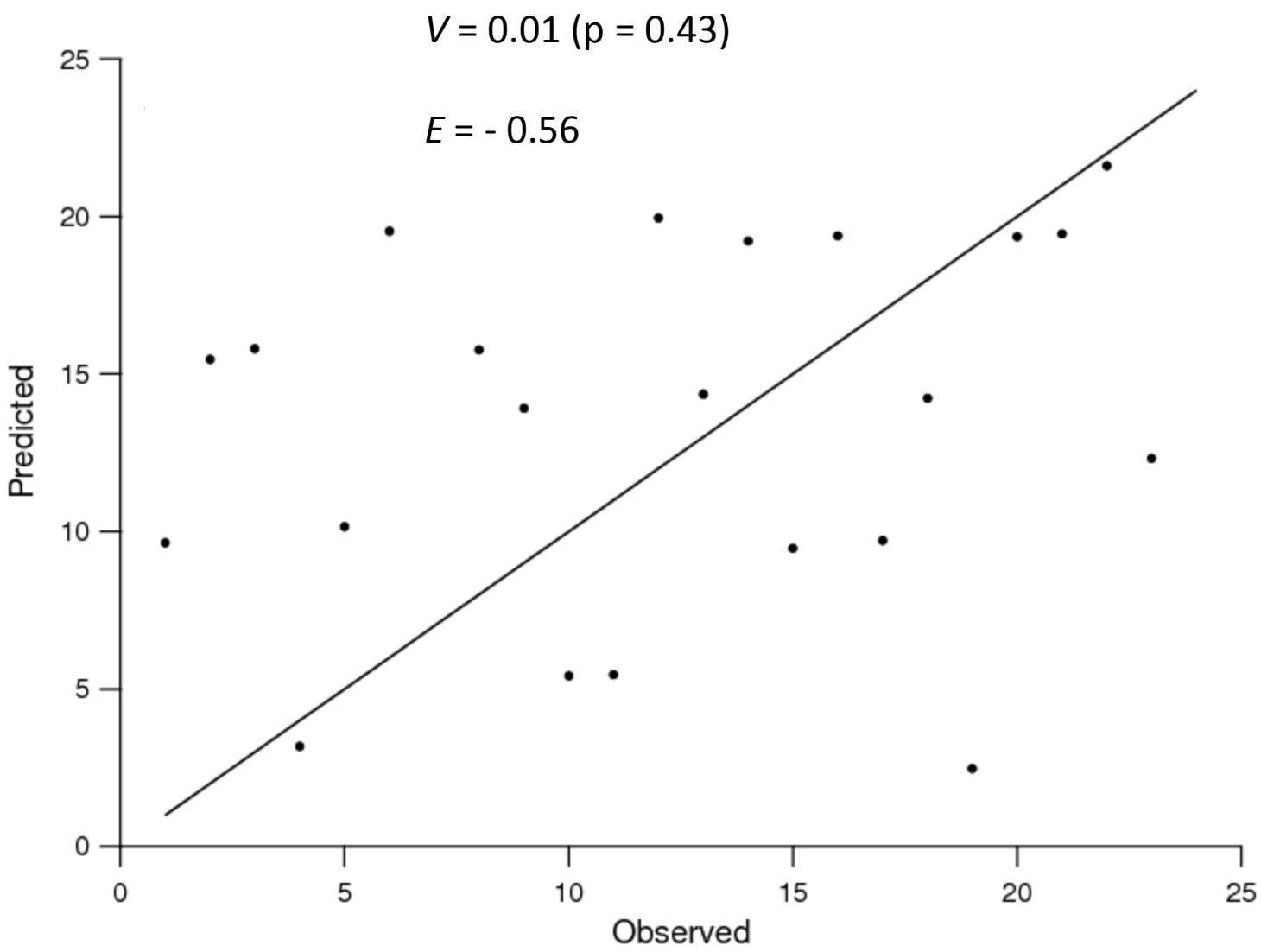


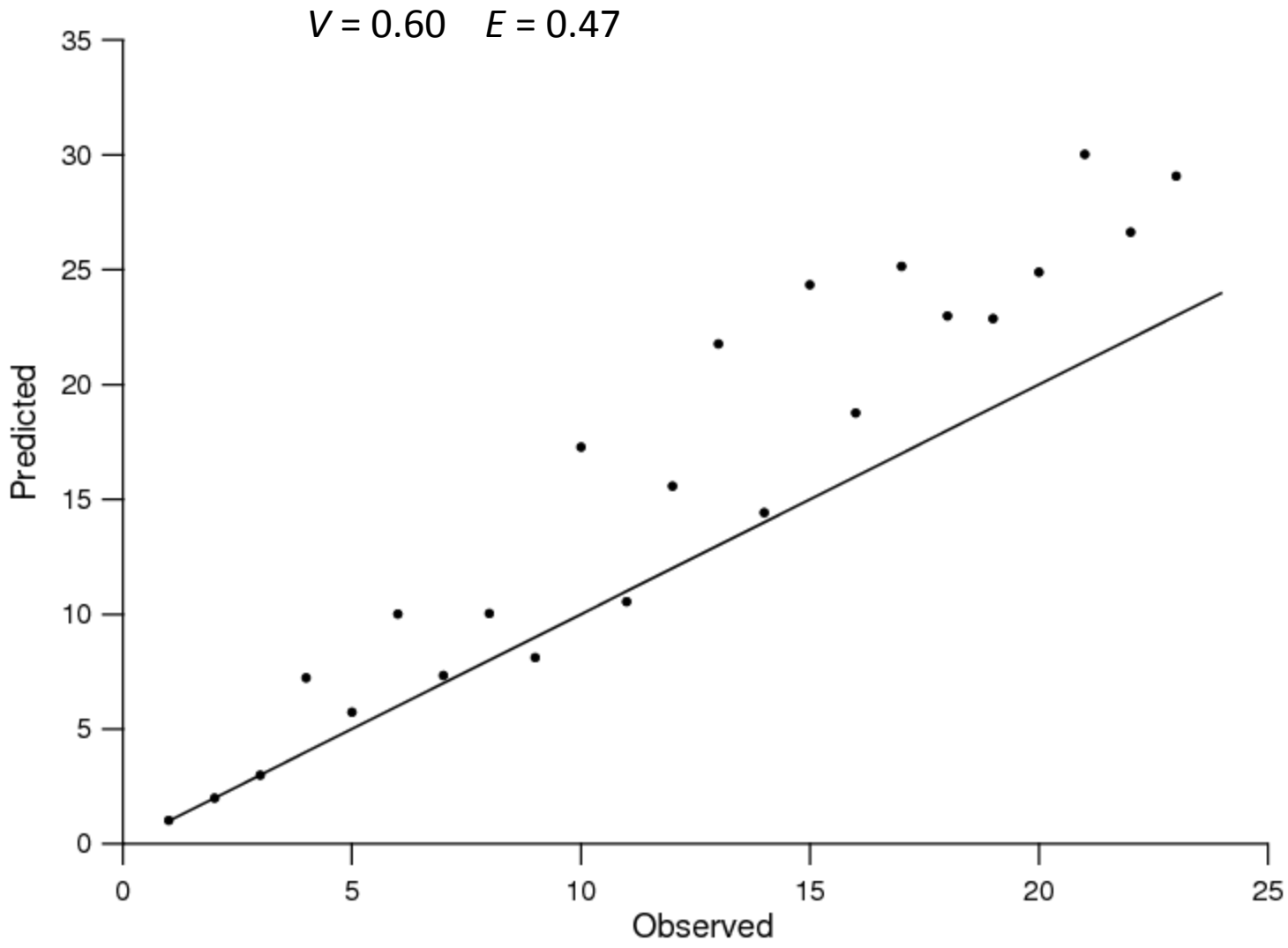


Some further example data sets follow..


V = 0.89
E = 0.94
Predicted
Observed
0
5
10
15
20
25

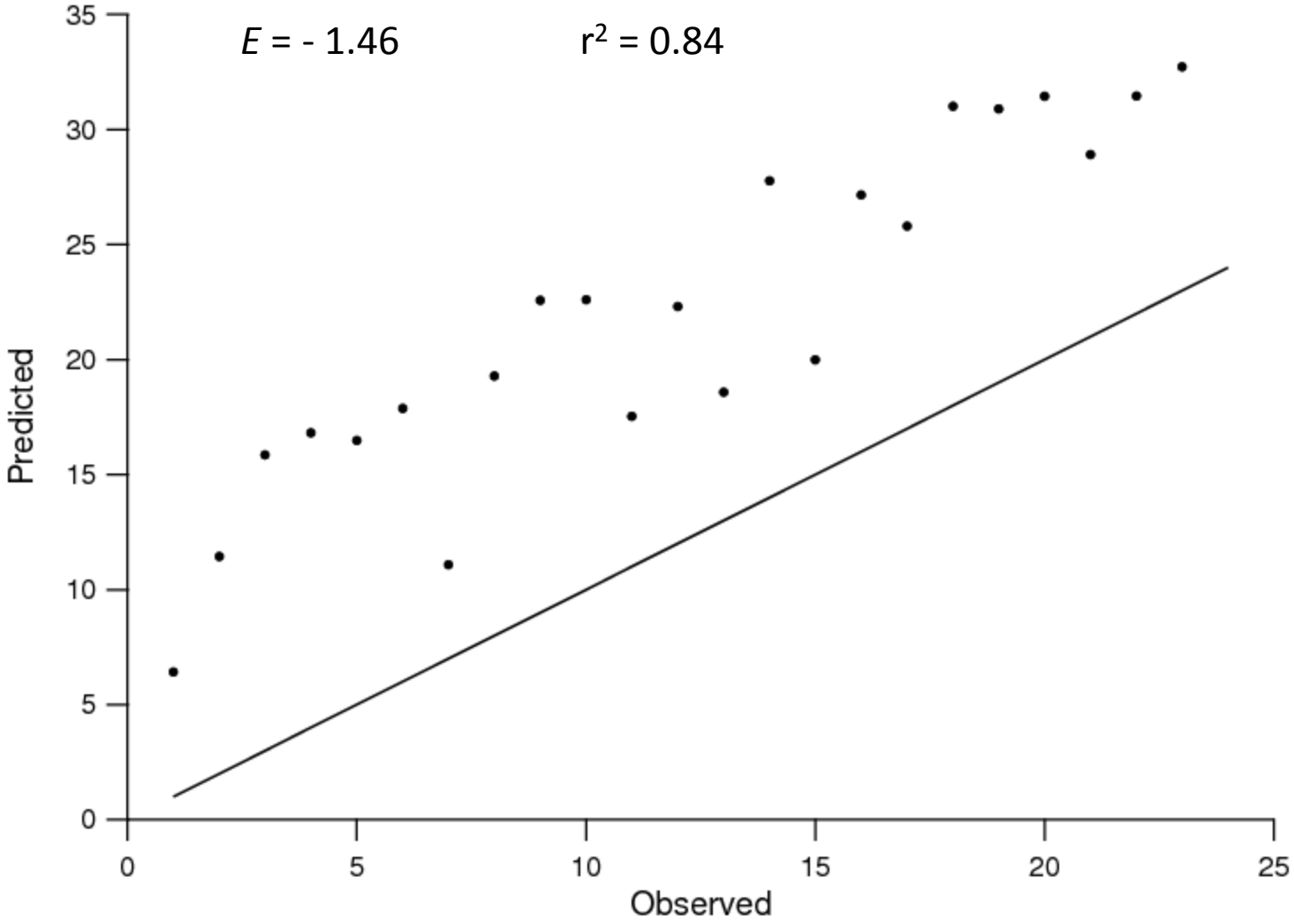

V = 0.12 E = - 0.89
Predicted
Observed
0
10
20
30
40
0
5
10
15
20
25


V = 0.01 (p = 0.43)
E = - 0.56
Predicted
Observed
0
5
10
15
20
25


V = 0.60 E = 0.47
Predicted
Observed
0
5
10
15
20
25
30
35

$V = 0.24$ ($p < 0.01$)

$E = -1.46$ $r^2 = 0.84$

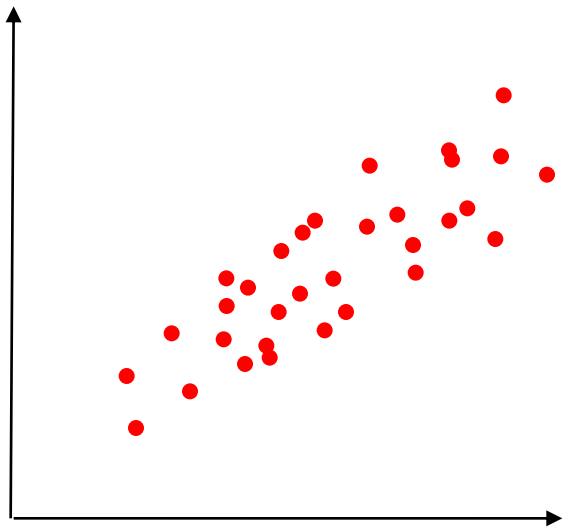


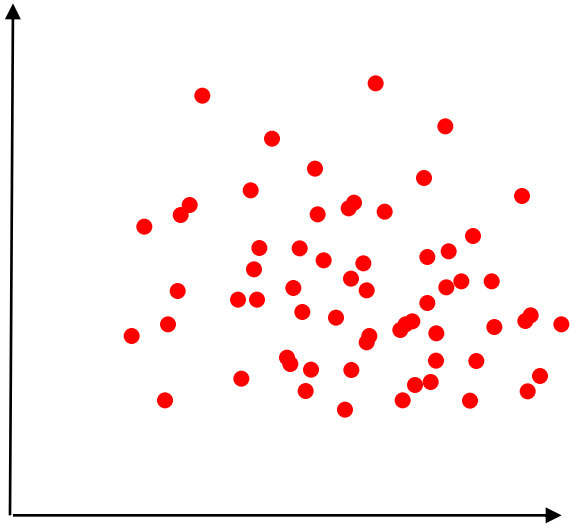
* For *any* fit index, remember to remove from data the obvious variations (eg seasonal or regional means) before calculating the goodness of fit measure.

Part 2: Selected hydrology-applicable probability models

- Permutation distributions
- Quantification of model validation fits – introducing a new fit measure
- Gumbel return period
- Weibull distribution and the hazard function for baseflow recessions
- Lognormal distributions for environmental work rates
- A Beta distribution alternative

Permutation distributions of test statistics (from randomly rearranging data). Permits assumption-free testing of non-standard hypotheses.



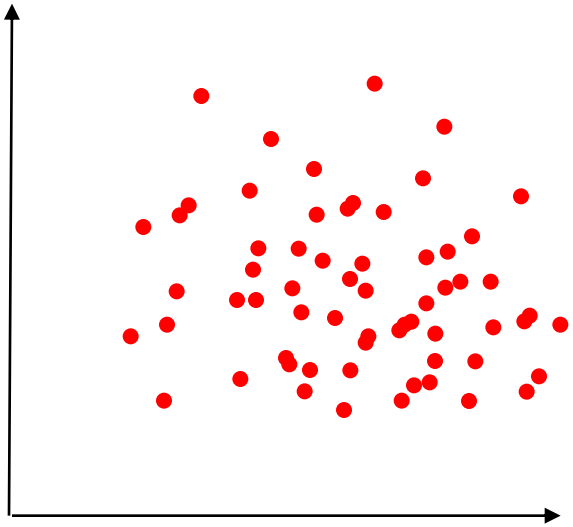


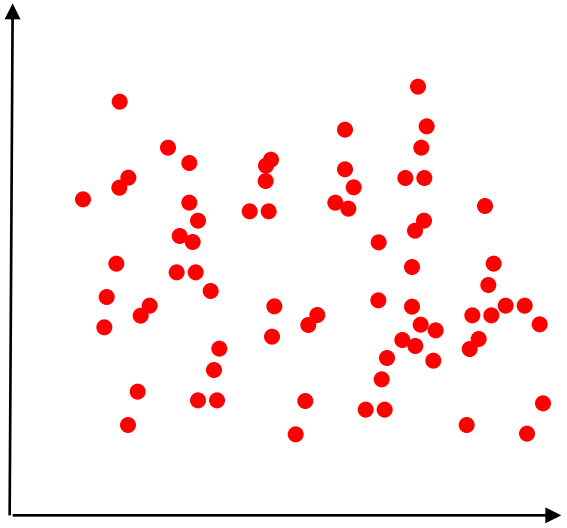
Generally – a good relationship becomes....

.... a mess under any one randomisation

So any statistic calculated here (eg regression gradient, empty space in the upper left corner..

... is likely to be far removed in the distribution of the test statistics calculated from all the randomisations





But if the original data is already a mess ...

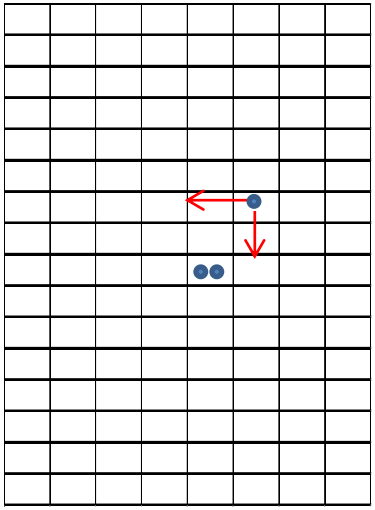
... each randomisation will just give a different mess each time

So any test statistic calculated here ....

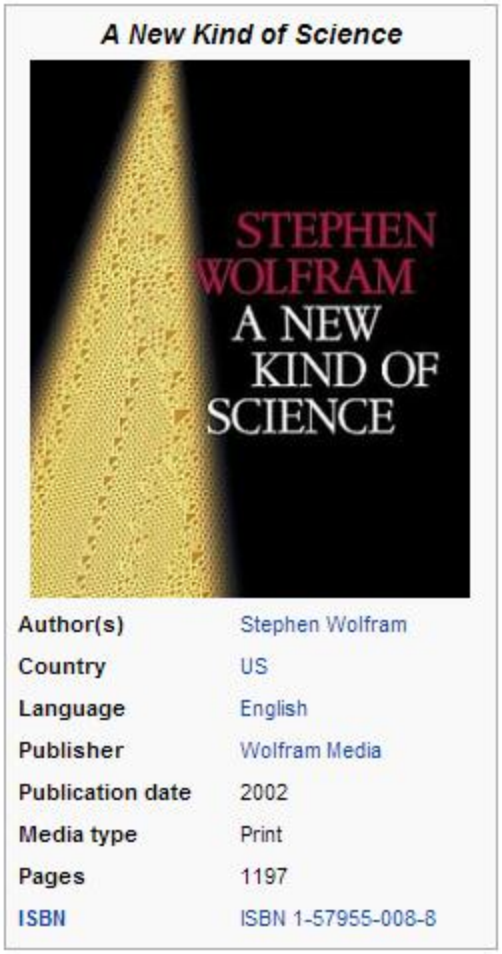
... is likely to end up in the middle of the test statistic distribution – never anywhere near significant.

Vetrova & Bardsley, W.E. *Hydrology and Earth System Sciences*, 16, p.1255-1257

The probability distributions of agent-based finite difference groundwater modelling



Also useful for coupled groundwater microbug /reactive product modelling

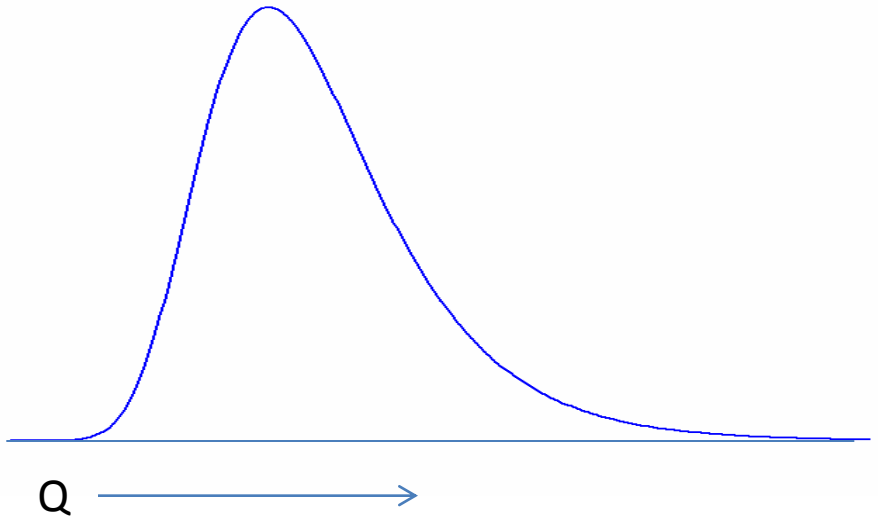


For each “particle”, calculate all positive values of $K\Delta h/\Delta L$ and express as probabilities with probability proportional to $K\Delta h/\Delta L$.
Select one “particle” from the probability distribution and move it (the origin cell loses 1 pressure value and the destination cell gains the pressure increment).
Recalculate the probability distribution and repeat.

The Gumbel distribution and the strange case of the return period time scale

$F(Q_n) = \{\exp(-\exp(-(Q - \xi)/\alpha))\}^n$

= probability that the largest flood in n years will be less than Q



Hypothetical conversation:

Farmer: "You say you have constructed my upstream dam to withstand anything up to the 100-year flood. What actually does that mean?"

Engineer: "It means that if my assumptions are correct and I can ignore estimation error, your dam has been carefully constructed so that there is an 0.63 probability that it will be destroyed within 100 years."

Farmer: "Ummm … run that by me again please."

We can't improve accuracy, but we can improve communicability. If would be more logical to have a time scale such that there was (say) a 1% chance that the flood magnitude concerned would happen in the specified time period.

$$Y_{RP} = -\ln\left\{-\ln\left[1-\frac{1}{RP}\right]\right\}$$

Return period tick-point on the Gumbel y-scale

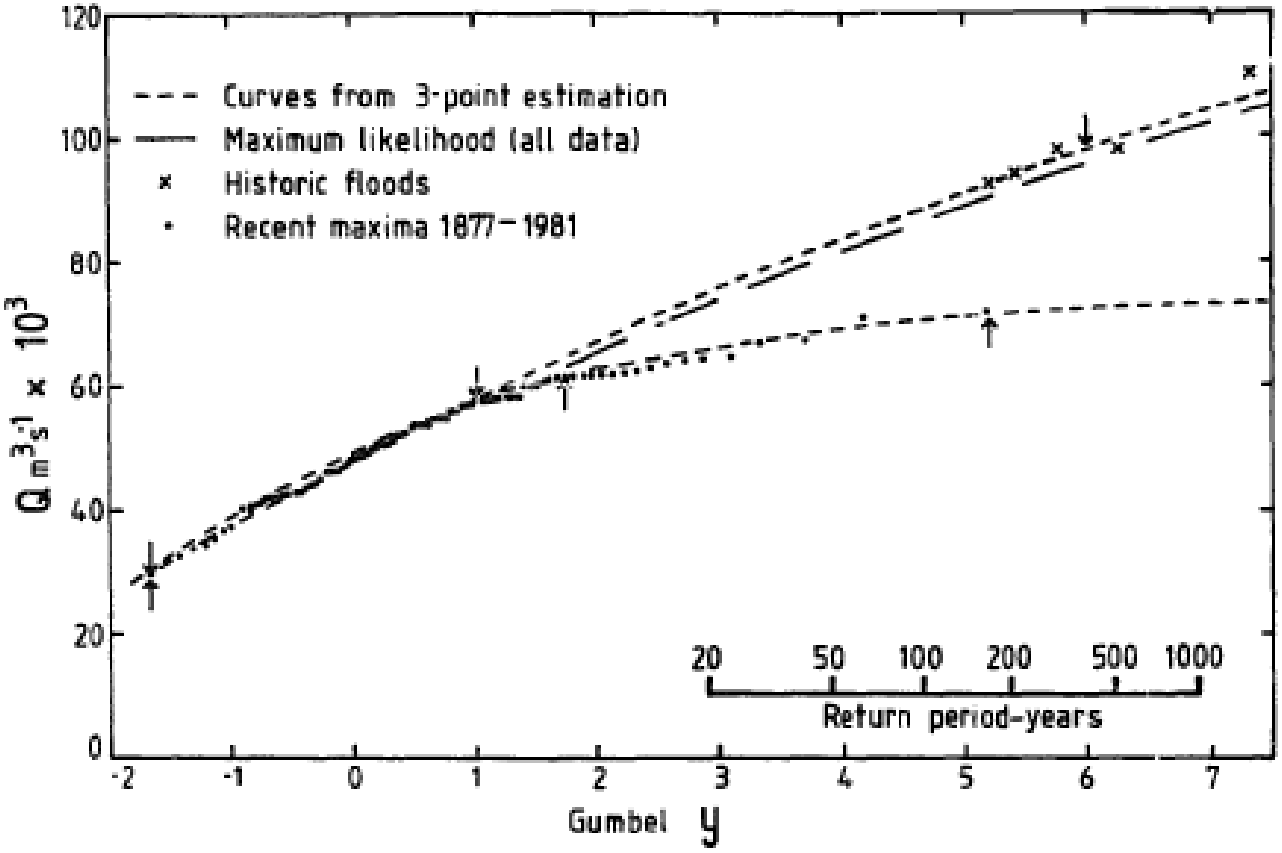
$$Y_T = -\{\ln -[\ln (0.99^{1/T})]\}$$

Time tick-point for T on the Gumbel y-scale for the new scale

Time	Y(RP)	Y(0.01)
10	2.25	6.90
30	3.38	8.00
50	3.90	8.51
100	4.60	9.21
200	5.30	9.90

Journal of Hydrology 119 389-391

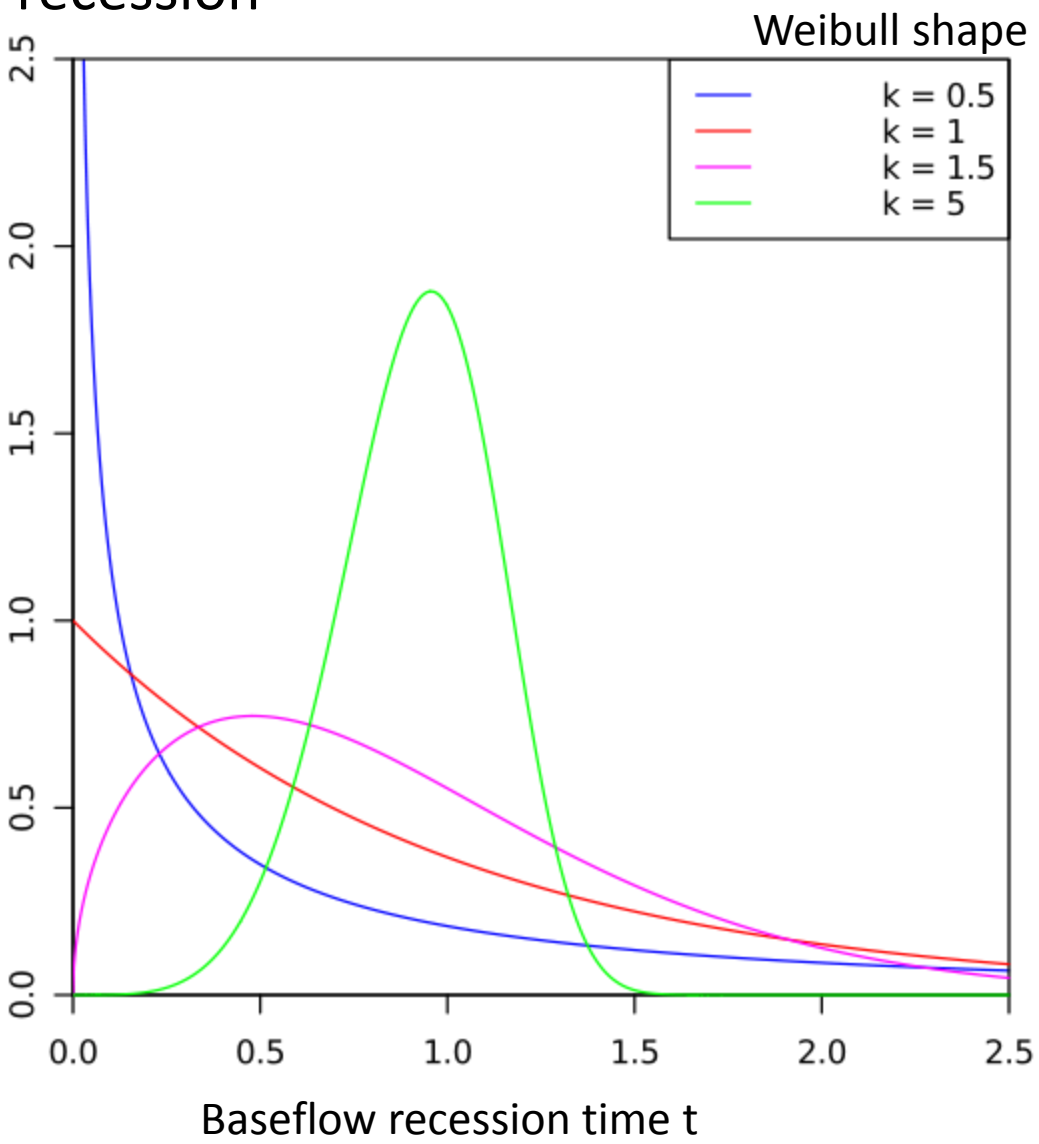
And for the generalised extreme value distribution we need not be confined by maximum likelihood estimation



Maximum likelihood and subjective 3-point fitting of Yangtze River annual maxima, recorded at Yichang.

Journal of Hydrology 110, 315-321

The Weibull distribution and the hazard function for baseflow recession



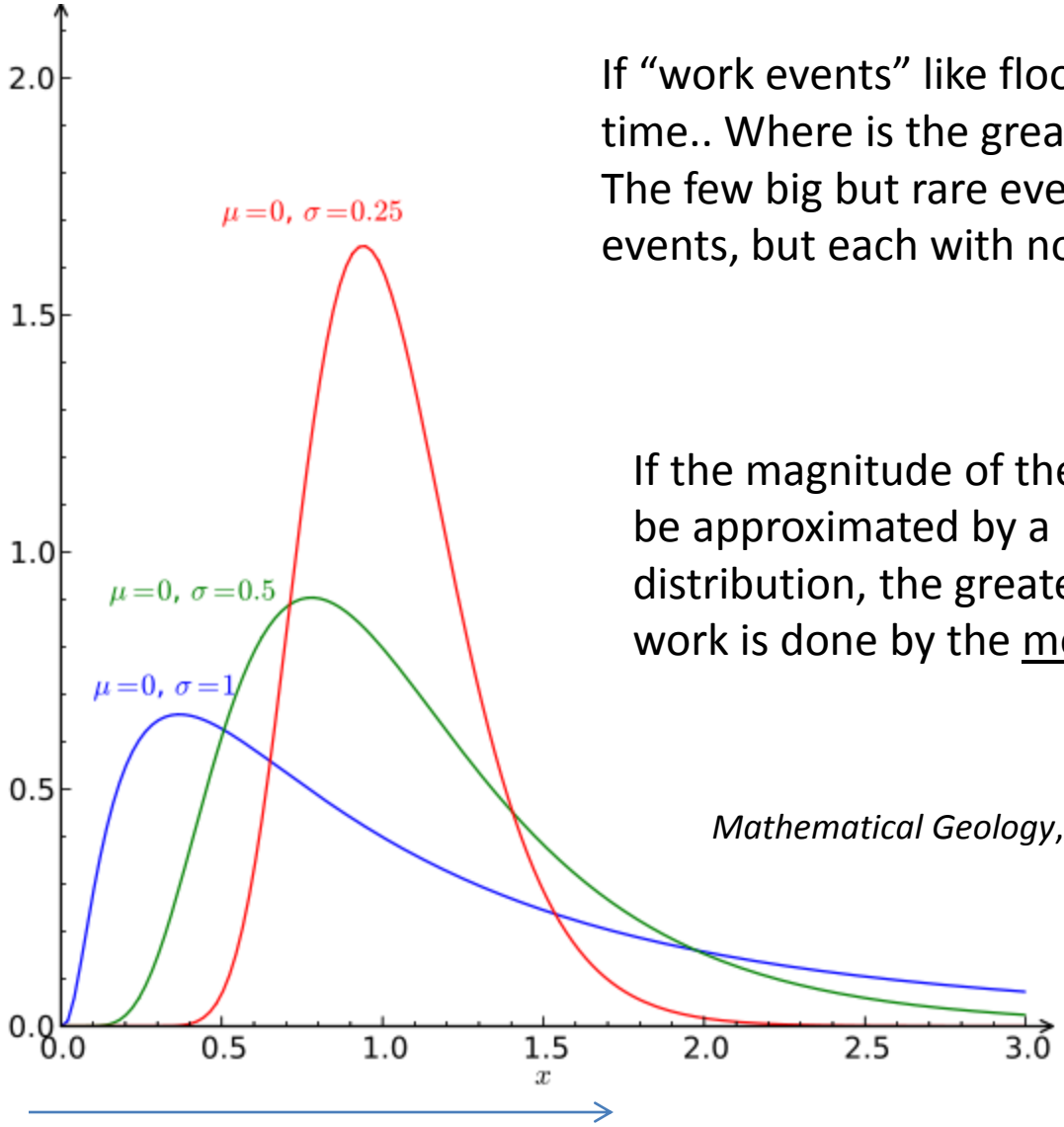
Hazard function (in a store loss sense) = discharge at a given time / (amount of water still in the groundwater store)

= f(t) / [1 – F(t)]

For shape parameter < 1, the Weibull hazard functions decrease with time, so could be suitable as a general model for baseflow recession

K = 1 = exponential distribution = constant hazard function.

The lognormal distribution and natural work rate intensity



If "work events" like flood erosion are random in time.. Where is the greatest cumulative work? The few big but rare events? The many little events, but each with not much work?

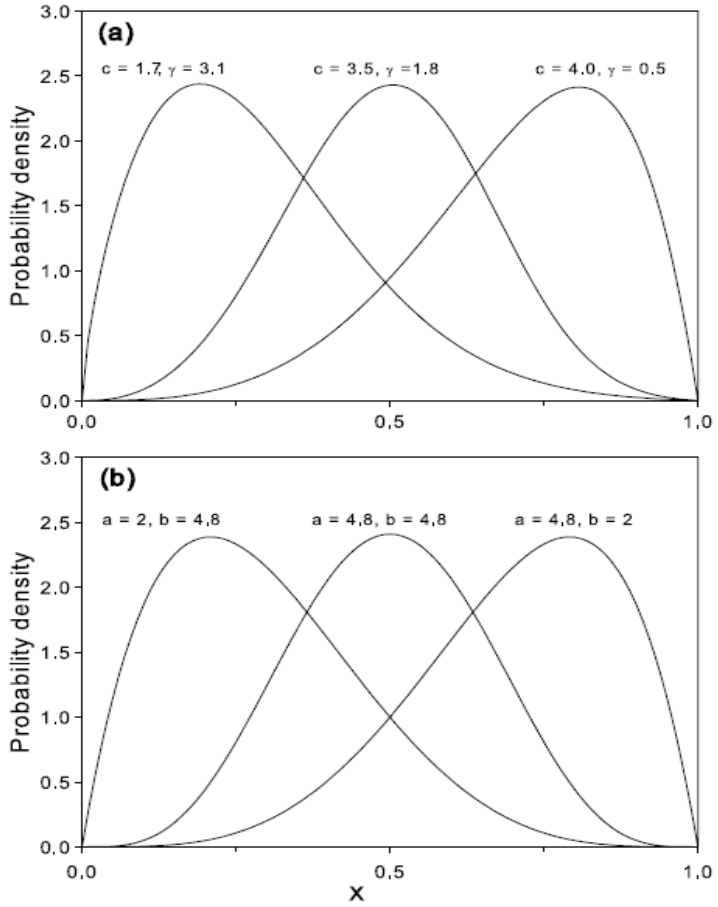
If the magnitude of the work events can be approximated by a lognormal distribution, the greatest cumulative work is done by the median events.

Mathematical Geology, v.23, p.591-608

Magnitude of work per event (erosion event, flood cost..)

The annoying Beta distribution

The Beta distribution is useful for describing events over a bounded range (eg raindrop size distributions) but use if the incomplete beta function makes integration not so easy in a standard spreadsheet. An alternative is:



Earl's distribution
(with simple integration)

Beta distribution

Advances in Water Resources, v.30, p.113-117.

Conclusions

- Even minimal hydrological information can be of value with uncertainty
- Probability distribution models provide a unified approach to a range of hydrological analyses (apologies for not mentioning multivariate models)