Entropy based metrics to evaluate physical models (*and data*)

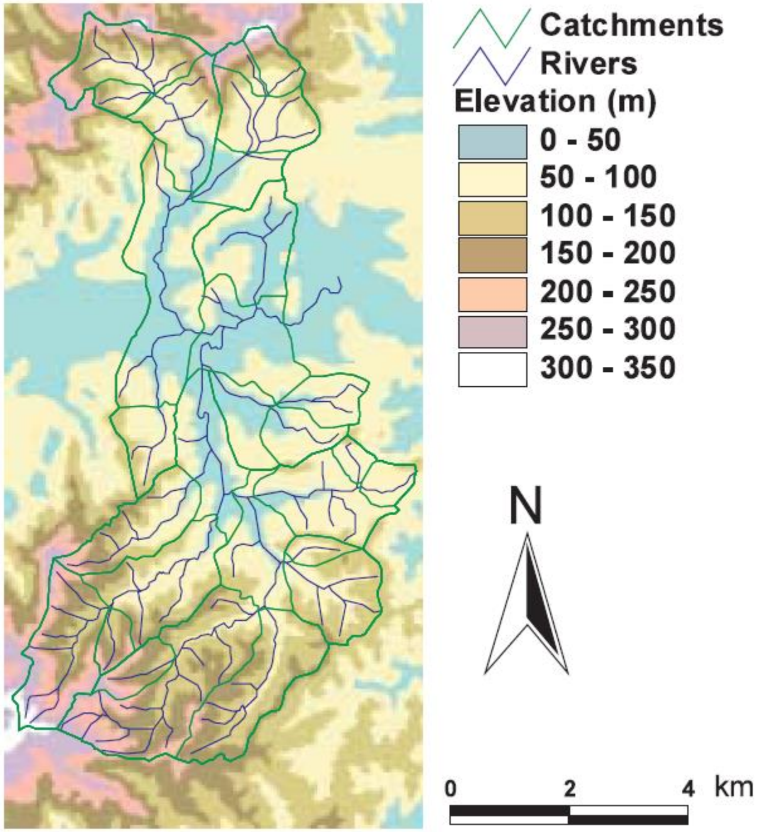
BETHANNA JACKSON, ILIAS PECHLIVANIDIS, HILARY MCMILLAN AND HOSHIN GUPTA

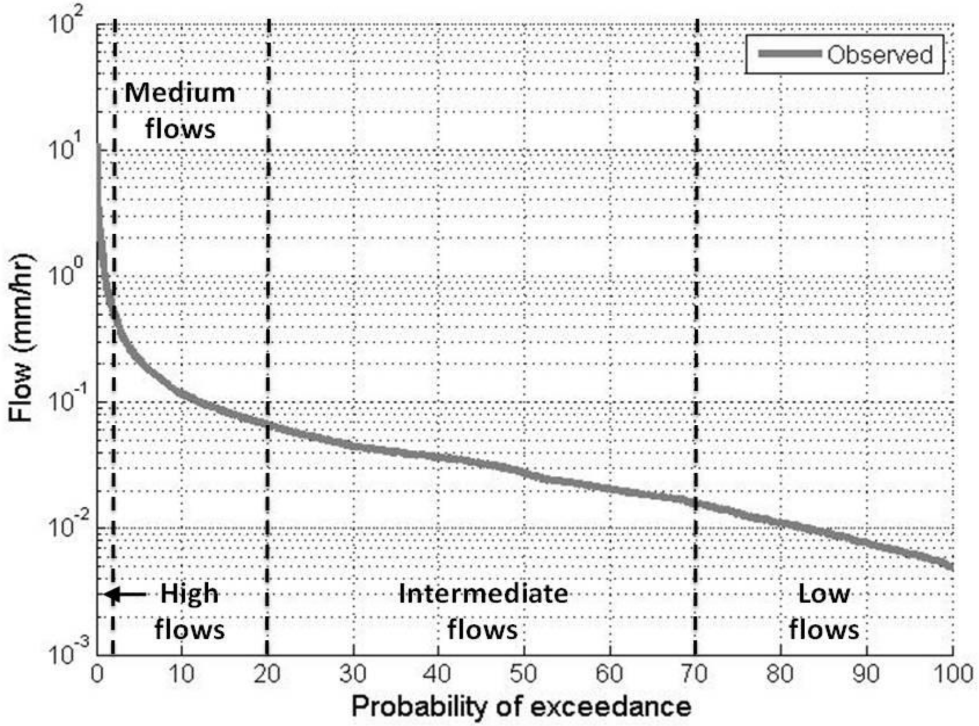
Information theory at the coal face

- Moving from philosophy to practice …
- Or a story of what happens when people new to information theory decide to pick it up and run with it…
- Naïve implementations; artefacts; issues reconciling physical theory with information flow as well as models and data

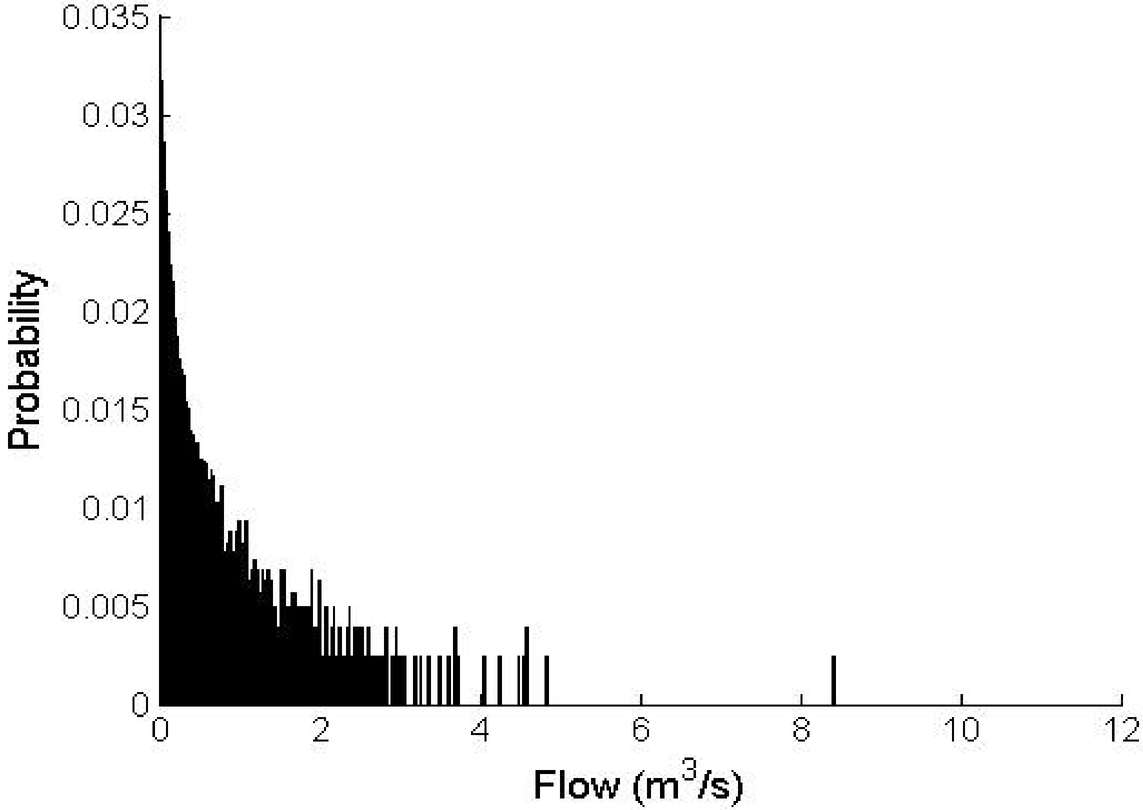
- Pechlivanidis, I.G., Jackson, B., McMillan, H., & Gupta, H.V. (2016) Robust informational entropy-based descriptors of flow in catchment hydrology. *Hydrological Sciences Journal*, 61(1).
- Pechlivanidis, I.G., Jackson, B.M., McMillan, H., & Gupta, H. (2014). Use of an entropy-based metric in multiobjective calibration to improve model performance. *Water Resources Research, 50(10),* 8066-8083.

Application: Mahurangi catchment





Probability distribution of flow



Shannon entropy:

Shannon entropy is defined as: $H_X = E[-\log_{base} p(X)] = -\sum_{i=1}^{M} p(x_i) \cdot \log_{base} p(x_i)$

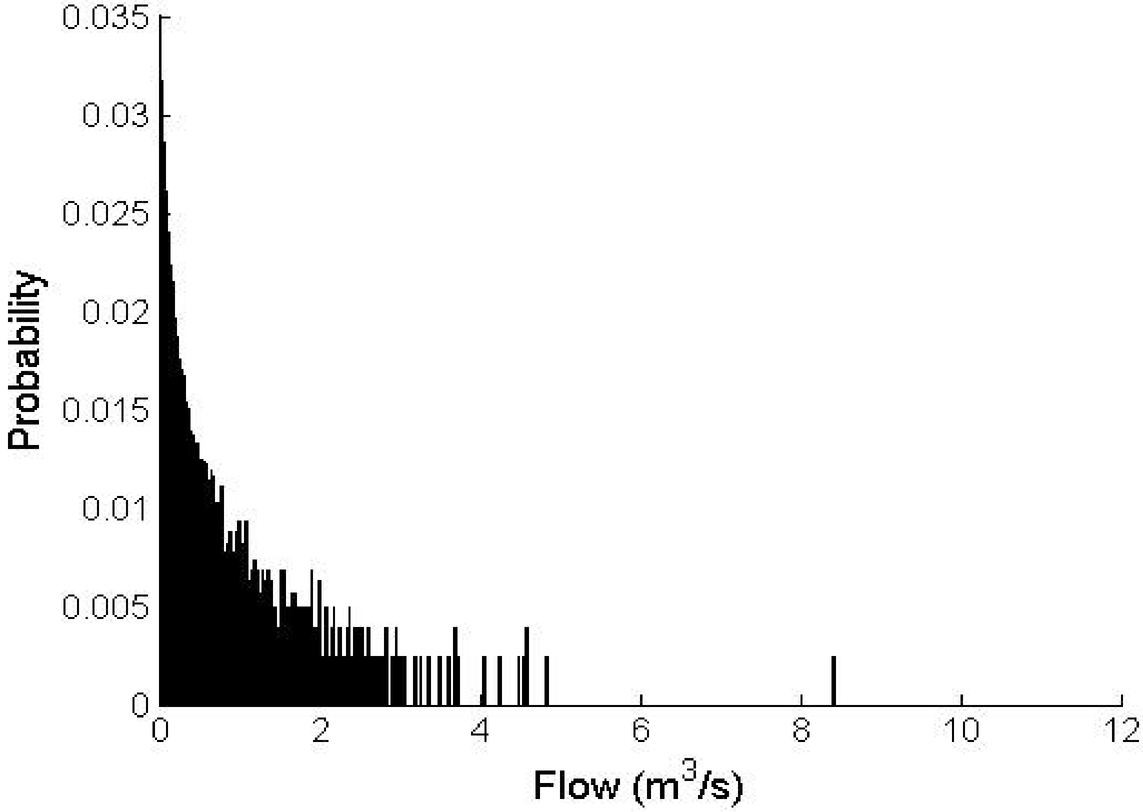
where $p(x_i)$ is the probability of outcome x_i of the discrete, non-negative random variable X such that $\sum_{i=1}^{M} p(x_i) = 1$, *M* is the number of possible outcomes and *E()* denotes expected value. Lies between 0 (complete information) and $\log_{base}(M)$ (no information); commonly normalised to (0, 1).

"Analog" for continuous X with probability density *f*(*x*) satisfying $\int_{-\infty}^{\infty} f(x) = 1$. This continuous Shannon entropy or "differential entropy" is defined as: $H = E[-\log_2 f(X)] = -\int_{-\infty}^{\infty} f(x) \cdot \log_2 f(x) dx$

Not fully consistent with its discrete counterpart:

1. Discrete goes to infinity as number of bins goes to infinity, continuous entropy always finite.
2. Discrete entropy>0, continuous may be positive, zero or negative,
3. Discrete entropy does not (directly) depend on *X*; continuous does.

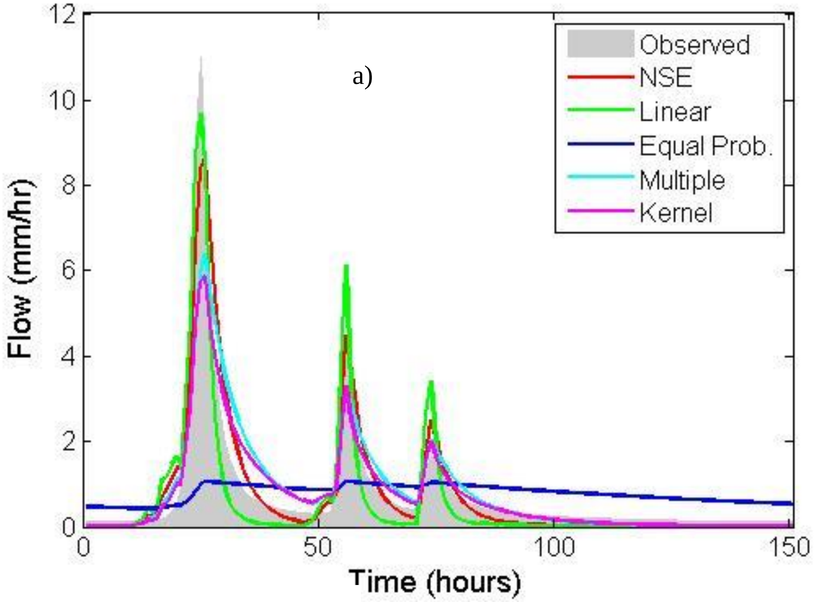
Probability distribution of flow

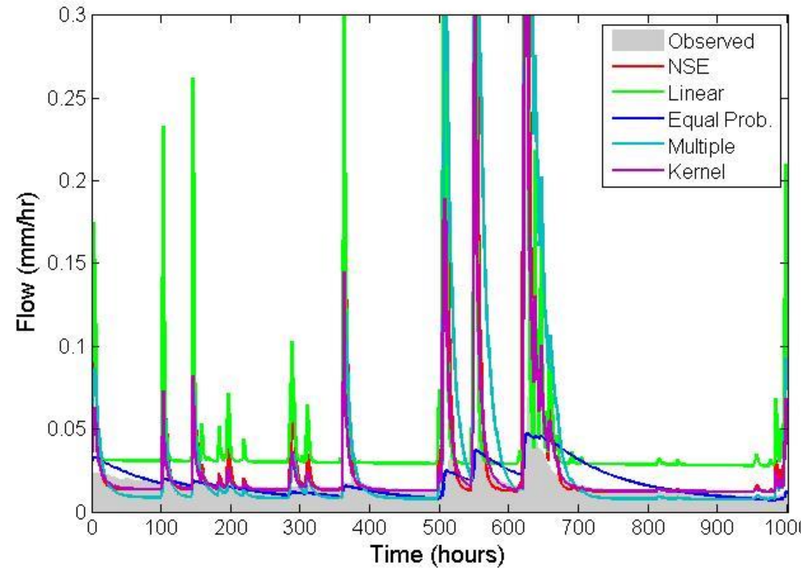


How do we extract meaningful entropy measures from flow?

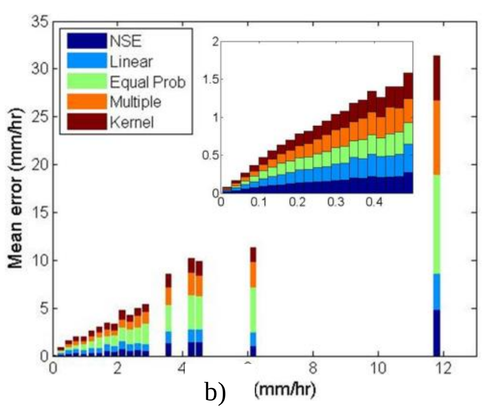
1. What is the impact of binning resolution on calculation of entropy?
2. What artefacts arise due to emphasis of information theoretic (and other) measures towards flow ranges having more data (statistically dominant information)
3. What are the effects of truncation errors on discharge measurements or model simulations?
4. Can we use entropy of flow to help diagnose inconsistencies between models and data; to understand structure of systems (observed or modelled); and/or to understand where new data presents “new” information; and what that might mean?

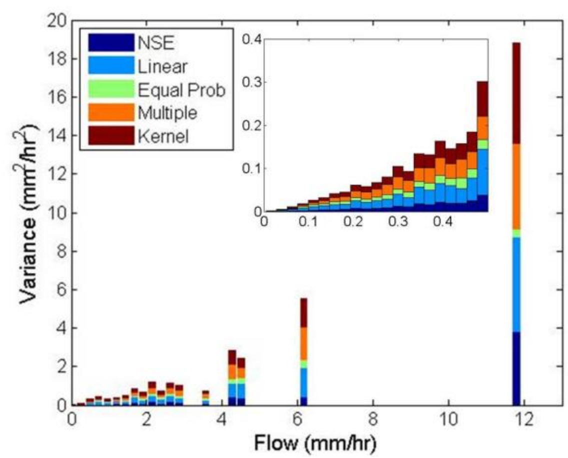
Performance of different binning methods



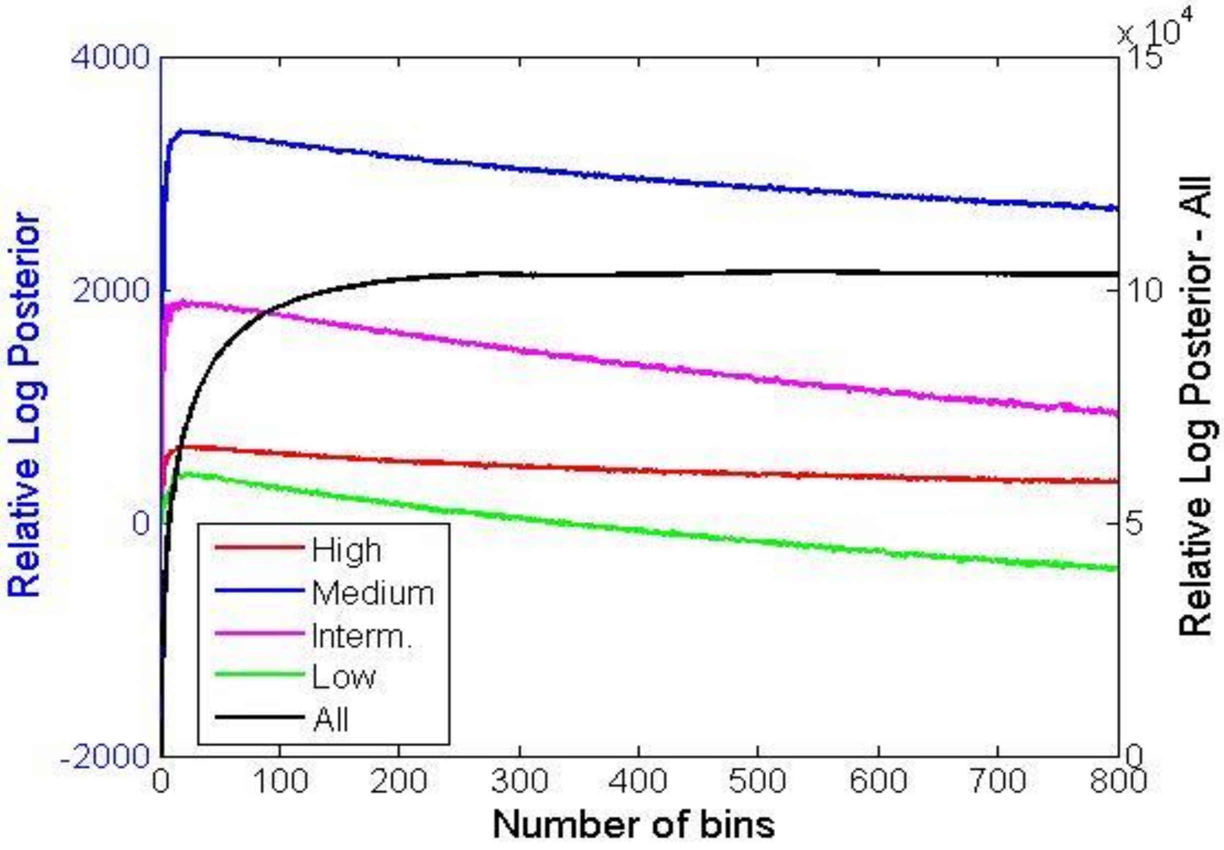


c)





“Optimal” bins



Used Knuth, K.H. 2006.:Optimal data-based binning for histograms. (Physics/0605197).

Also added constraint that bin width >= data precision: Reduced “low flow” bins to 60.

(alternate approach to truncation problem is adding noise with appropriate amplitude)

Conditioned Entropy Metric

$$SUSE = \max\left[\left|H_{sim}^{U} - H_{obs}^{U}\right|,\ \left|H_{sim}^{S} - H_{obs}^{S}\right|\right]$$

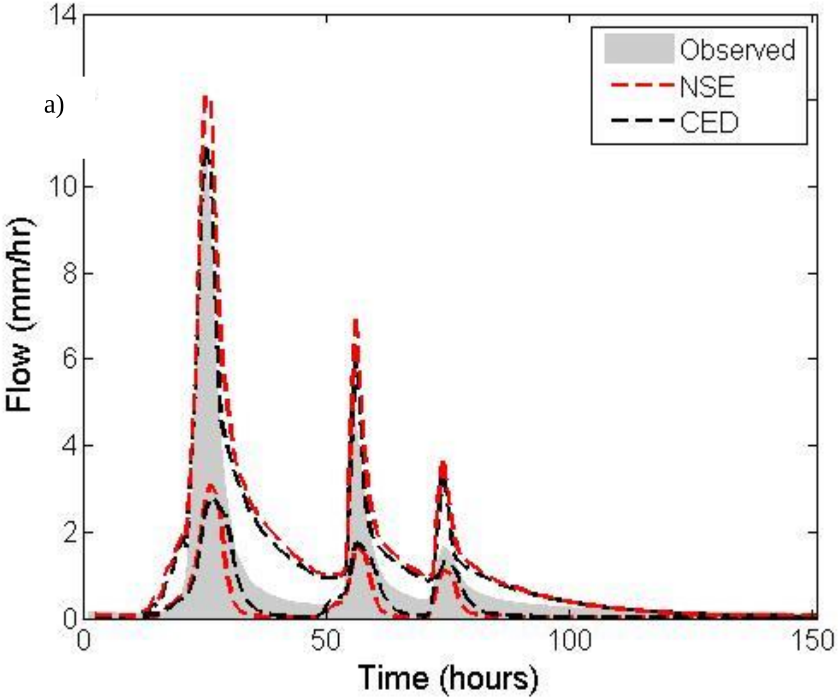
To overcome problems in extracting information from skewed data, The CED metric splits the FDC into multiple segments and evaluates entropy characteristics over all segments.

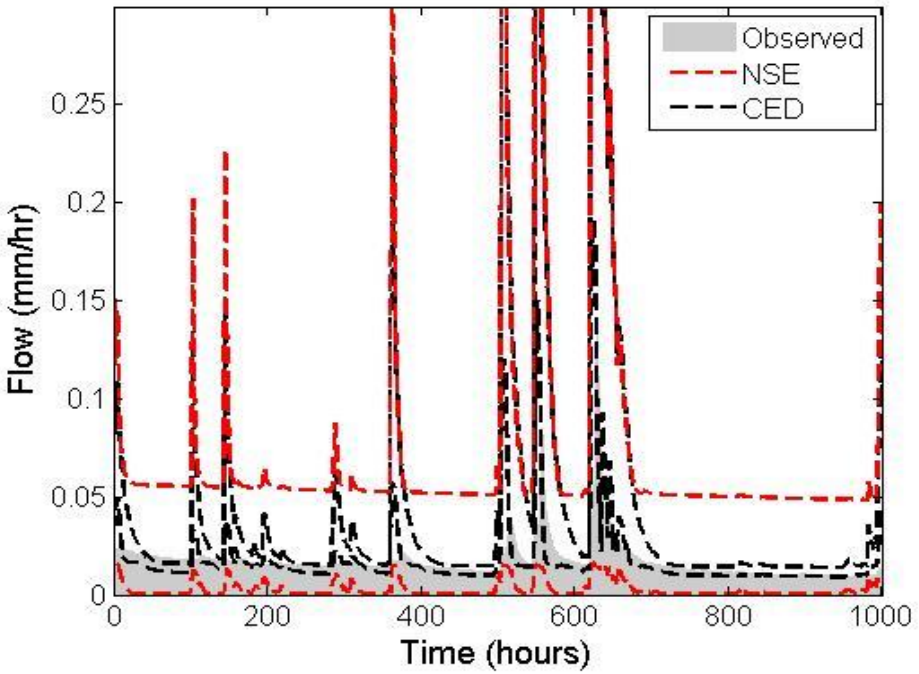
CED is defined as the maximum Scaled/UnScaled Entropy difference present in the different FDC segments

Lies between 0 (perfect agreement) and 1.

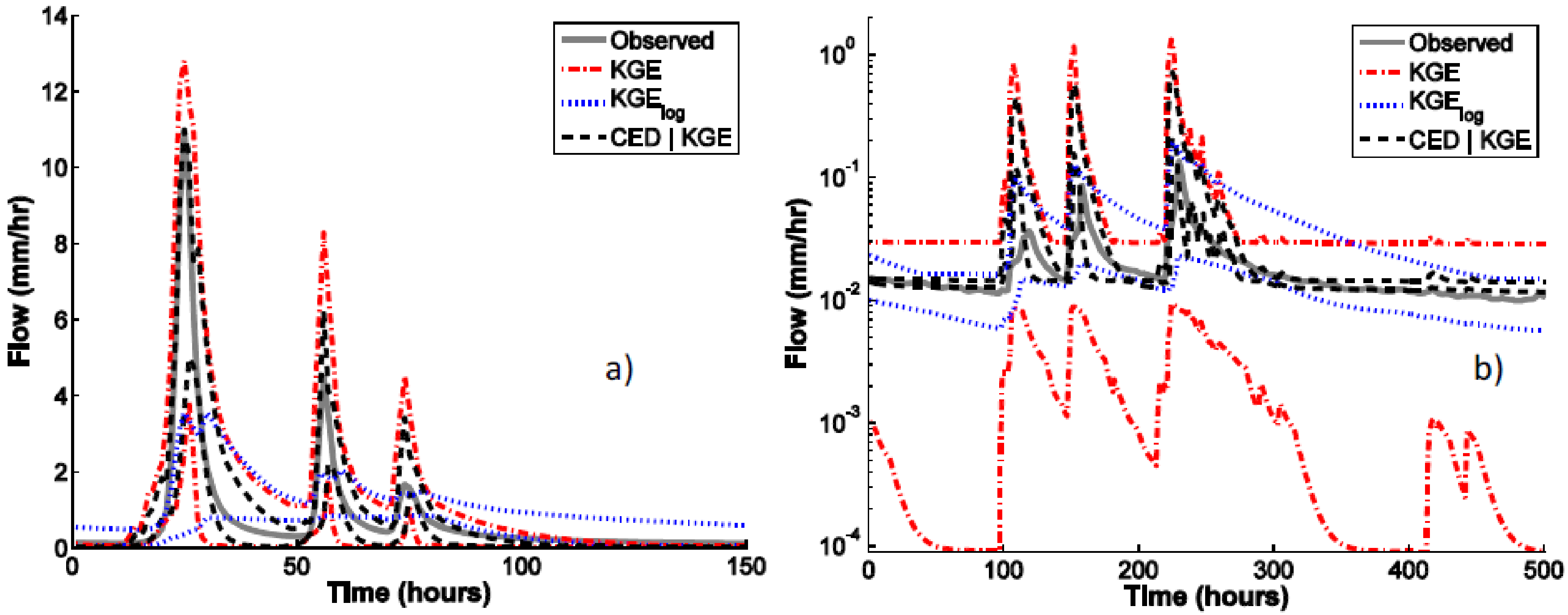
For Mahurangi, partitioned four segments of the FDC; respecting high (0-2%), medium (2-20%), intermediate (20-70%) and low (70-100%) flows.

Simulated streamflow performance

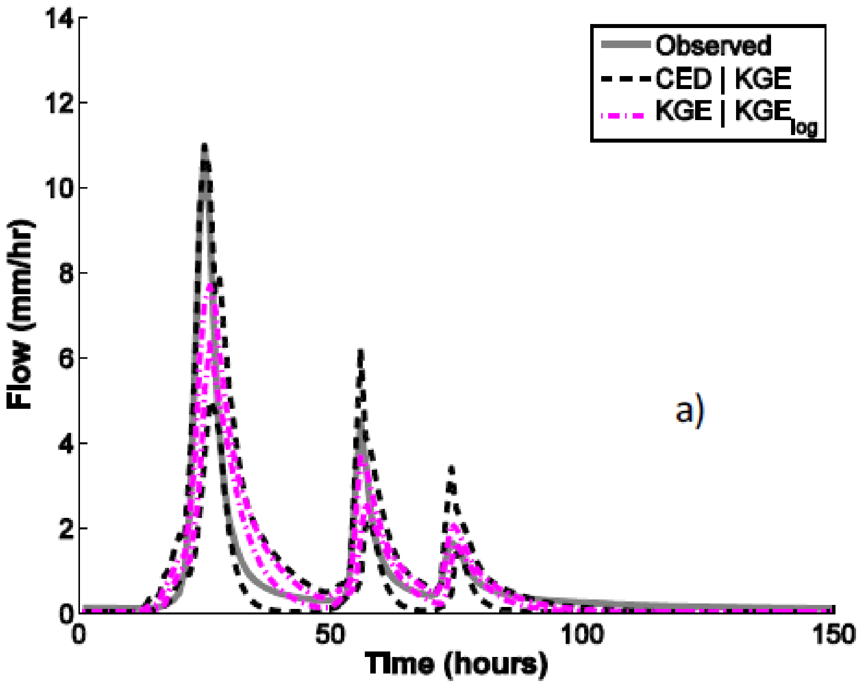


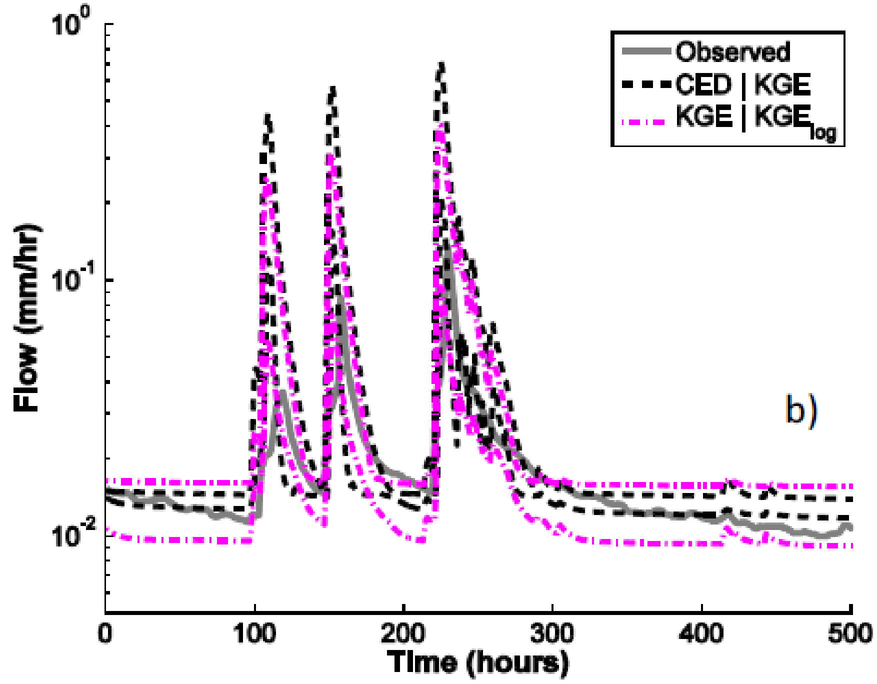


KGE, KGElog, CED|KGE

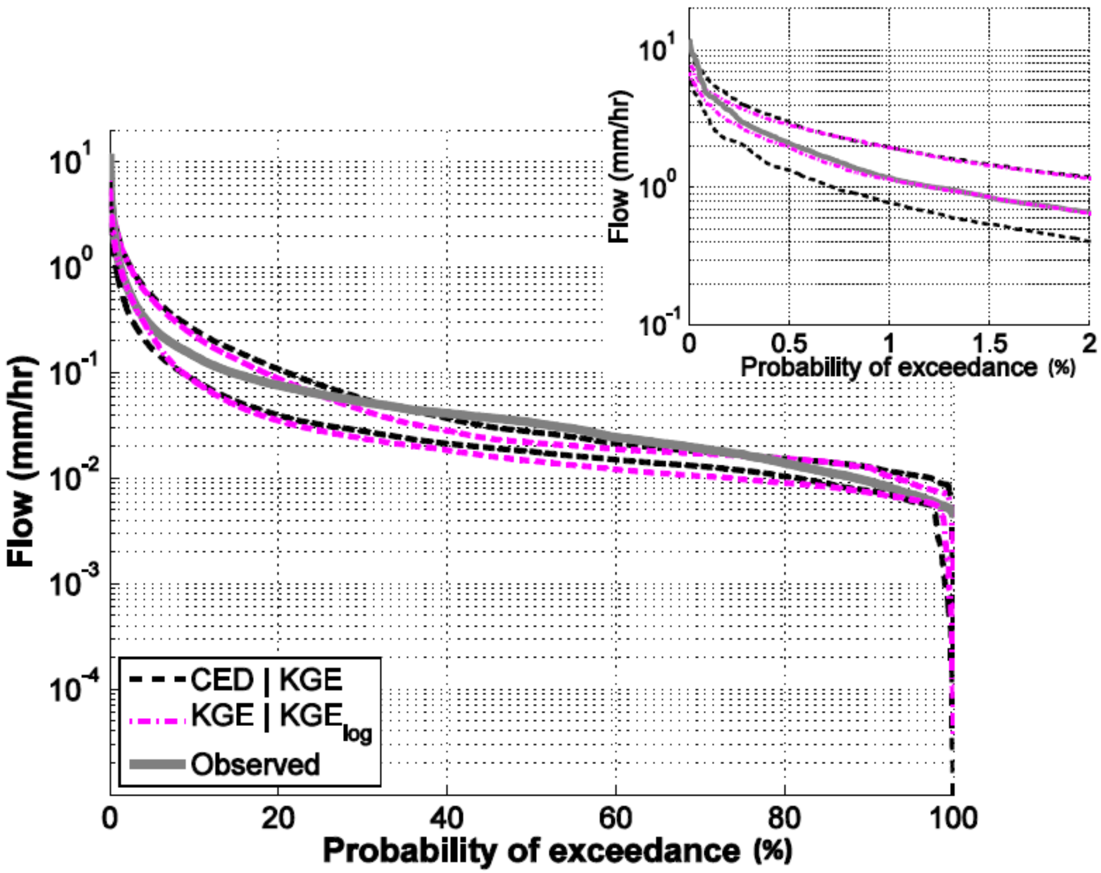


CED|KGE vs CED|KGElog





Envelope of FDC



Conclusions

- Naïve implementations of entropy may contain large artefacts; need for community effort to establish good practice and diagnostics to ensure we are producing meaningful metrics
- We can use entropy to help diagnose models; and in particular it is an excellent representation of "shape" and of the flow duration curve
- Difficulties but also opportunities as distributions of observed (or modelled) data change; opportunity to automatically flag potential errors or obtaining previously unexplored climatic, antecedent, or other information