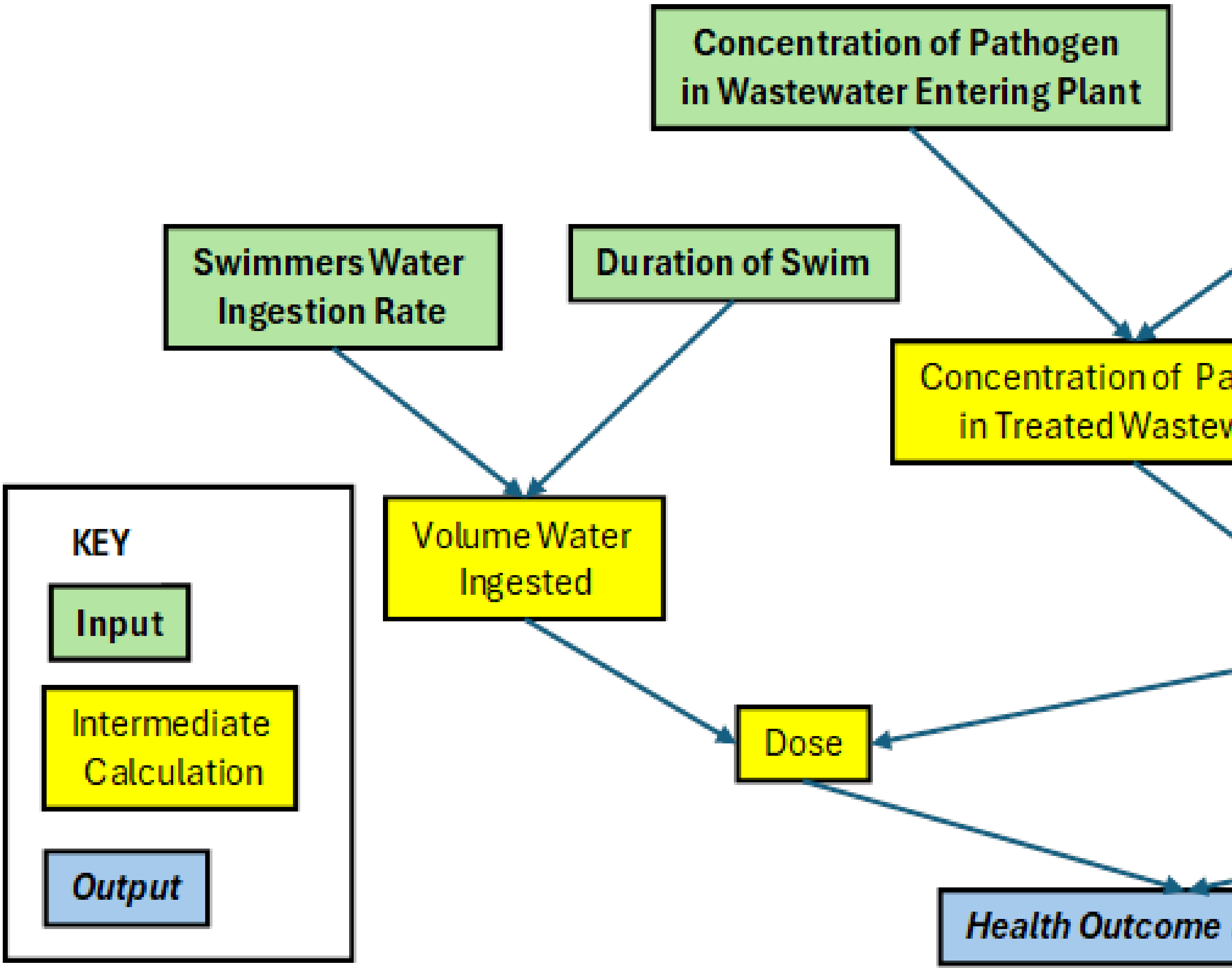
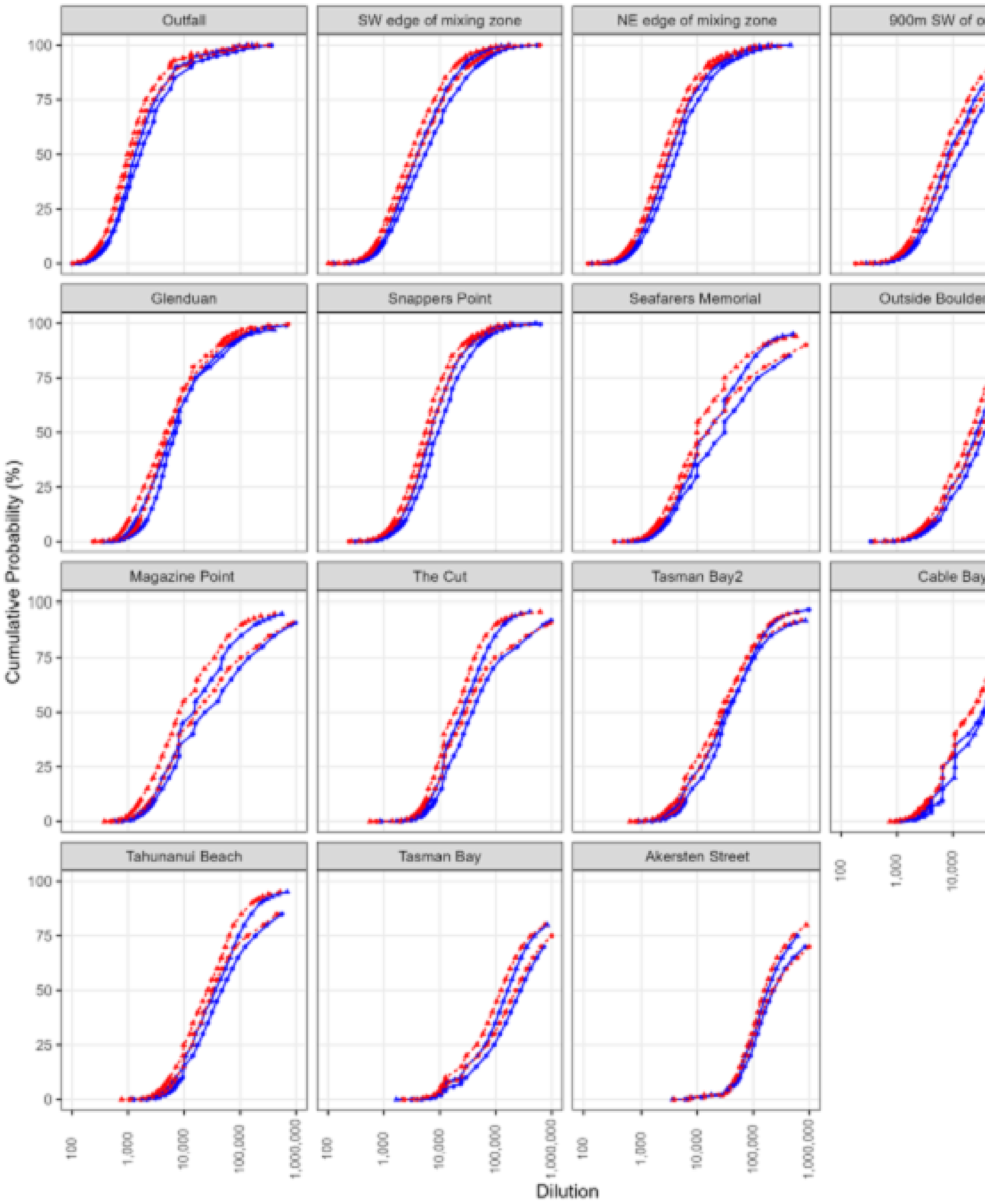
Empirical dilution function

Many of the QMRA problems we are given dilution data and the QMRA model samples this dilution data as part of estimating the concentration of pathogens in the environment.



You have seen the data we get. The usual way to handle it is to estimate the empirical cumulative distribution function (ECDF)and then sample form the ECDF. See some examples of ECDFs from Nelson North shown below.



Outfall
SW edge of mixing zone
NE edge of mixing zone
900m SW of o
Glenduan
Snappers Point
Seafarers Memorial
Outside Boulde
Magazine Point
The Cut
Tasman Bay2
Cable Bay
Tahunanui Beach
Tasman Bay
Akersten Street
Cumulative Probability (%)
0
25
50
75
100
Dilution
100
1,000
10,000
100,000
1,000,000

Some ECDFs are smooth other are not and they depend on tides and other factors. I will send you a spreadsheet and some code I put together to sample from an ECDF the function is called estDistribution.

The Hockey Stick function.

We often estimate the concentration of pathogens using the hockey stick distribution, yet another empirical distributions. See pages copied from McBride:

9.3.2 Hockey stick distribution

Influent virus data can be expected to be right-skewed, along with most environmental microbiological data. Given estimates of the minimum, median and maximum of that distribution, it would be folly to use a single triangular distribution—the right tail would be "too fat" and the median would not be preserved (unless maximum – median = median – minimum). A simple solution is to join the median and maximum by a "hockey stick"—that is, the line **BCD** in Figure 9.4.

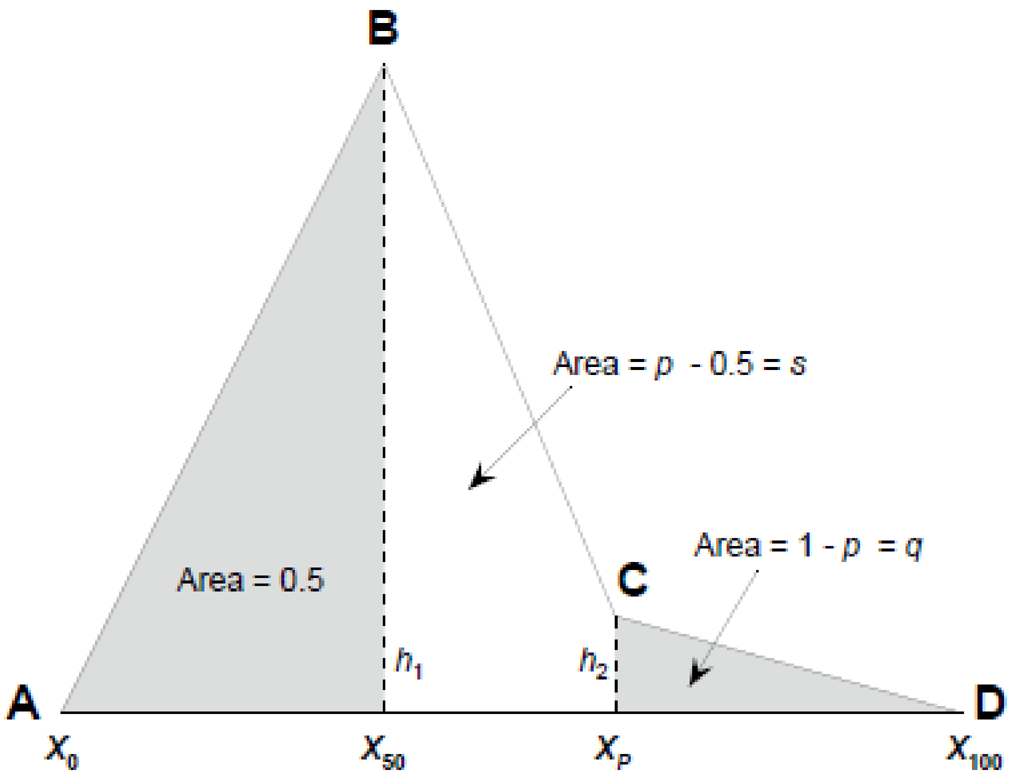


Fig. 9.4 "Hockey stick" empirical influent virus distribution (X_0 is the distribution's minimum, X_{50} is its median, X_{100} is its maximum and X_P is the Pth percentile: $p = P/100$). The value of P (>50) is to be supplied by the risk modeler.

To determine the position of the point **C** the we use the constraint that the total area under the distribution should be unity. To satisfy this requirement we find the value of the Pth percentile (X_P) given a user-supplied value of P. We also use linear interpolation between the distribution's breakpoints.[7] The algebra involves solving

[6]In @RISK this would be coded as c*RiskDiscrete({0,1},Pnone:Psome) (using named cells for c, P_{none} and P_{some}).

[7]Were we to join **BD** by a straight line, X_{50} would be the distribution's mode (not its median) and we would over-estimate the median and the frequency of most of the high values.

To determine the position of the point C the we use the constraint that the total

area under the distribution should be unity. To satisfy this requirement we find the

value of the Pth percentile (XP) given a user-supplied value of P. We also use linear

interpolation between the distribution's breakpoints.7 The algebra involves solving

a quadratic, in the following calculation sequence (referring to the nomenclature on the Figure—see Problem 9.6):

$$h_1 = \frac{1}{X_{50} - X_0}, \tag{9.9}$$

$$X_P = \frac{1}{2}\left\{X_{50} + X_{100} + \frac{1}{h_1} - \sqrt{(X_{100} - X_{50})^2 + \frac{X_{50}(2 - 8q) + X_{100}(2 - 8s)}{h_1} + \frac{1}{h_1^2}}\right\}, \tag{9.10}$$

$$h_2 = \frac{2q}{X_{100} - X_{50}}. \tag{9.11}$$

A pragmatic approach is to set $P = 95$—that is, the toe of the hockey stick is at the 95%ile. Random samples can be taken from this distribution (e.g., using @RISK's RiskGeneral function).

I think Charlote used it in here code and it is also in my code.

I hope those notes help