Hydrol. Earth Syst. Sci. Discuss., 4, S1857–S1860, 2008
www.hydrol-earth-syst-sci-discuss.net/4/S1857/2008/

Hydrology and Earth System Sciences Discussions

Interactive comment on "Statistical distribution of series of 12 monthly concentration samples for environmental classification of rivers" *by* J. Eliasson and T. Thordarson

J. Eliasson and T. Thordarson

Received and published: 4 January 2008

Anonymous Referee #2 Authors response included.

This paper deals with an important issue: how do we classify the water quality for streams where only limited data are available. However, I have fundamental concerns with the study:

1) The authors want to answer the question which distribution is most suitable to describe concentrations in streams. The data used in this study is not suitable to answer this question! If we want to derive the correct distribution we need a data set with more than 12 concentration values. Especially to investigate the tails of the distributions one

would need a larger sample. It would also be necessary to look on data from different years to support the general conclusions found in the abstract.

Response

No. The authors want to answer the specific question which distribution is most suitable for Statistical distribution of series of 12 monthly concentration samples for environmental classification of rivers as is said in the title. Maybe the DoC can have a more general use, that is for the future to decide. This is the only data available as Icelandic law framework does not demand more. The problem of limited data is discussed on P2569 L10 and onwards. Lowering the sampling frequency to get for data for investigation of the distribution tails does not help. For that there is too much autocorrelation in river flow, both water discharge and concentrations. Even for monthly samples, serious variance defect due to autocorrelation created by storage effect may be suspected in naturally regulated rivers. See P2571 L22 onwards.

2) The authors treat concentration as a purely random variable. The authors claim that there was no seasonal pattern (no results shown). Even if we accept this, I still would expect clear correlations of at least some of the concentrations with runoff. If we want to make progress with classification in the case of limited data we need to consider such correlations! Obviously ignoring runoff at the time of sampling is a severe limitation.

Response

Correlation coefficients to air temperature are given in table 1. Seasonal variation is reflected in the air temperature in the northern hemisphere so geophysical variables with seasonal variation have either a high positive or negative correlation to air temperature. Water temperature (not a constituent) has a strong seasonal correlation (0.64 - 0.83). The constituents themselves have (numerically) much lower correlation to air temperature. There is no general agreement on the limit between strong and moderate correlation. 15 constituents are shown in the figures, 11 are used in the pooling and 4 let out because of cross correlation as discussed on P2567 L15 so correlation is not to-

HESSD

4, S1857–S1860, 2008

tally ignored even though the coefficients are not used directly. It is interesting to note, however, that in data pooling, correlation does not destroy the distribution in general. Two random variables, say f and g where g = a + b f become one and the same when normalized, provided a and b are constants. So if the samples were 100 % correlated to runoff (they never are anywhere) the constituents would certainly all have the same distribution. The victims in data pooling are the IRL (Independent record length) and confidence limits, see e.g. Buisand and Schaefer referred in Eliasson 1997.

3) I do not agree that it is reasonable to compile one distribution from the concentrations of different constitutes like done in Figure 6. Even if these concentrations are normalized before this seems like comparing apples and pears.

Response

Pooling data (62000 hits in Google) is a well known tool in statistic, e.g. meteorology. Why is the reviewer against it ? We find the likeness of the DoC with the well known lognormal distribution in the upper 40% of the data particularly interesting and feel that this is a tangible result of the pooling. This shows that the lognormal can be used, but then the lower 60% of the data has to be discarded and then it takes 30 months to get 12 data points, not 1 year. In view of comment no. 1 about the tails we feel we have really achieved something here.

Minor comments: What is the correlation coefficient in Table 1? May be correlation with air temperature? In that case the correlations actually would indicate some seasonal variations of concentrations.

Response

With air temperature yes (corell = 1). For discussion of correlation see P2567 L15 - 21.

How where probability values (0-1) assigned to the ranks (1-12)?

Response

HESSD

4, S1857–S1860, 2008

This is ranked data, there is no probability value assigned to the ranks 1 - 12 in figs 2 - 5 in the paper. In fig 6 11 constituents are pooled in a pool of 132. Plotting positions for this data set are found using Greengortens formula, but for probabilities below 90 % (Icelandic regulation limit) it makes no visible difference what plotting position formula is used.

A cumulative frequency curve should be monotonic. Some of the curves in figures 2-4, however, show some increases with increasing rank (where the curves should be expected to decrease monotonically). The reason for this must be that the authors used Excels smoothing function when preparing the plots (in this case the figures provide a good example why one NOT should use this function!)

Response

The reason for this choice is clarity of graphical presentation as explained in P2565 L23 and onwards. The method suggested by the reviewer was tried, it produced unusable graphics as too many of the lines clashed into each other. The ranked data figures are mostly for demonstration of systematic deviations from the normal distribution (in linear and log space) that statistical tests accept when the lines are tested one at the time, but clearly show when the lines are plotted together so they can be visually compared. These systematic deviations are avoided when the DoC is used (in linear or log space) instead of the normal distribution.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., 4, 2561, 2007.