# 2013 Freshwater Eionet Workshop - 19/20 September 2013, Copenhagen

## Session 2: Data quality and aggregation

## Document 2d: Quality and representativity of SoE nutrient/ BOD data: length of time series, outliers, inclusion of additional parameters (PO4 versus Ptot, BOD versus COD) by Kari Austnes/ Lidija Globevnik, ETC/ICM

### Background

CSI19 (oxygen consuming substances) and CSI20 (nutrients) are based on data from rivers, lakes and groundwaters (the latter two CSI20 only) stored in Waterbase (see agenda item 1a document). The assessment includes both spatial analysis of the most recent data (WISE maps) and time series/trend analysis. The data are re-analysed every year, after the update of Waterbase. In 2013 a more thorough investigation of the data will be conducted, to evaluate the potential for indicator improvement. This is part of the draft assessment which will be distributed prior to the Eionet meeting (see agenda item 1a document).

### Some suggested data issues to be presented and discussed

* **Data quality**: Some data series reported to the EEA contain suspicious values. Many such values are flagged in the automatic QA procedure (agenda item 2a), but some are only discovered during data analysis (example in figure 1). At this stage it is (too) late to get back to the data providers, so highly suspicious values are excluded from the further analysis (as of this year, though, such exclusions are reported to the data managers so that the data are flagged and checked with the data providers at a later stage). Errors in the data series can affect the data analysis (example in figure 2). The more manual QA procedure is also time consuming, and there is no guarantee that all possible suspicious values will be discovered. So, it would be a great advantage if errors are identified prior to reporting.
* **Length of time series**: For time series analysis it is necessary to use consistent time series, i.e. all time series must cover the same years. Since 2010 up to 3 years of inter/extrapolation has been allowed to be able to use more of the available data series. Still, the number of data series is far lower than the number of stations used for the WISE maps, and the number is decreasing, causing concern about the representativity. Some issues: **1)** The start year is now 1992. This means that no data series starting in 1996 or later will be part of the analysis. One solution is to choose a later start year. Table 1 gives examples of the effects of this solution. Another solution would be to include several parallel analyses with different start year (including, possibly, single, far longer time series), but this would increase the length and complexity of the indicator. **2)** As soon as there is more than three years since data was reported for a station, it is excluded from the analysis. This could be due to a) end of monitoring; b) missing reporting or c) change of station code without informing the ETC data managers (then the station is registered as a new station and the time series is broken). a) may not always be avoided, but b) and c) should be avoided. **3)** In the draft assessment it will be explored whether it is possible to combine comparable parameters to increase the number of time series (e.g. BOD and COD, total phosphorus and orthophosphate).
* **Number of parameters**: Is there added value of including more parameters in the indicators? I.e. parameters that are already reported, but so far not used in this context? Examples will be given (e.g. lake total nitrogen and Secchi depth, river total phosphorus and total nitrogen)
* **Coupling SoE and WFD information**: E.g. aggregating SoE stations by current WFD class was done for some stations in the EEA report on European waters status and pressures published in 2012 (<http://www.eea.europa.eu/publications/european-waters-assessment-2012>, chapter 7), but there was difficulties coupling the data sources due to coding issues. It is important that the water body code as reported to the WFD is supplied in the stations table for the SoE stations.



*Figure 1: Example of possible outlier not discovered in the automatic QA procedure*



*Figure 2: Example for river orthophosphate of the effect of excluding suspicious values (upper panel) as opposed to keeping them in (lower panel). See especially data for the Black Sea.*

Table 1: Effects of changing the start year. Example for river orthophosphate

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Time span | Number of stations | Number of countries | % strong increasing trend | % weak increasing trend | % no trend | % weak decreasing trend | % strong decreasing trend |
| 1992-2010 | 1027 | 22 | 6 | 2 | 38 | 6 | 48 |
| 2002-2010 | 1730 | 29 | 5 | 2 | 64 | 7 | 22 |

### Questions to NRCs

1. **Data quality**: How is the QA procedure in your country before data reporting? Are the most recent data checked towards the whole time series? Which data should be excluded from the analysis?
2. **Length of time series**: Do you have old data that have so far not been reported (from existing SoE stations or other, still monitored stations)? Is it possible that station coding has changed, but that this has not been reported? Do you have other solutions than the suggested ones of how to include shorter time series in the indicator analysis?
3. **Coupling SoE and WFD** (for countries who have adopted the WFD only): Has the water body code for the water body in which the station lies been supplied in the stations table for all SoE stations from your country?