Waterbase – Lakes
Version 8

Quality control documentation

18 June 2008

Waterbase – Lakes

In the context of the implementation of the Water Framework Directive (WFD), the European Environment Agency (EEA) EIONET-Water annual data flow for waters is in the process of being transferred into the WISE 'State of the Environment' (SoE) voluntary data flow. With this it remains one of the EIONET Priority Data Flows, but gains full integration into the reporting under WISE as the single entry point of water information in Europe and complementarily with data collected under the WFD. Most information that is used for European level 'state of environment' assessments needs to be provided by member countries and there it usually comes from monitoring networks that are to meet several assessment purposes, SOE, as well as different legal requirements..

Data on lakes are collected annually through the WISE-SoE data collection process. Data and information obtained through the WISE-SoE data collection process are primarily used to compile indicator factsheets, associated with the EEA's Core Set Indicators, upon which EEA assessment reports are based. Data collected through the WISE-SoE data collection process are also published in WISE map viewer, Waterbase, a series of water topic-specific databases and web pages, publicly accessible via the EEA Data Service's web site.

Lakes dataset include physical characteristics of the river monitoring stations, proxy pressures on the upstream catchments areas, as well as chemical quality data on nutrients and organic matter in rivers.

QA/QC activities

This document briefly presents the ETC/Water and the EEA activities focused on quality of the Waterbase - Lakes dataset and the results of these activities. In addition a warning is given on the use of certain records for analytical purposes (see section 2 and 3).
The Quality control tests have been performed on the Waterbase - Lakes database provided in April 2008 by ETC/WTR. This database is included in the EEA data service as version 8, and is publicly available. The database and metadata are available at the following URL:
http://dataservice.eea.europa.eu/dataservice/metadetails.asp?id=1039

A subset of the dataset is also used in the WISE (http://water.europa.eu/).

Waterbase – Rivers dataset contains three data tables:

- PRESSURES
- QUALITY
- STATIONS

Four types of the tests have been performed on the data tables. Basic tests, Logical rules violation test, Outlier detection and Station coordinates tests.

1. Basic tests

1.1 Summary

The summary presents number of records subdivided by country for each table of the dataset:

a) which were delivered by the country in the last delivery (very late deliveries are not included)
b) which was possible to process (reasons why some of the records was not possible to process are very various and comprehensive summarization is difficult)
c) total in the working database
d) which are present in the Waterbase - Lakes v8

Numbers of records excluded from the Waterbase subdivided by the reasons of the exclusion are also present.

1.1.1 Waterbase - Lakes: Quality

Country Code	Numbers of records				
	in the latest delivery		total in the working database	in the Waterbase	excluded from the Waterbase - reasons*
	total	processed			1
AL	0	0	95	95	
AT	73	73	924	924	
BA	123	123	363	363	
BE	59	59	67	67	
BG	74	74	824	824	
CH	113	113	2583	2583	
CY	20	20	230	230	
CZ					
DE	91	91	2046	2046	
DK	278	278	5014	5014	
EE	420	420	1646	1646	
ES					
FI	7416	7416	274651	274651	
FR			400	400	
GB			8548	8548	
GR					
HR	349	349	592	592	
HU	419	419	8234	8234	
IE	244	244	2956	2956	
IS			117	117	
IT	1371	1371	5153	5018	135
LI					
LT	298	298	1045	1045	
LU					
LV	209	209	2124	2124	
ME	309	0			
MK	5	5	68	68	

MT	9	9	19	19	
NL			1313	1313	
NO	560	560	6749	6749	
PL	100	100	552	552	
PT	276	276	276	276	
RO	125	125	506	506	
RS	1071	1071	3261	3261	
SE	1818	1818	32346	32346	
SI	120	120	1084	1084	
SK					
TR					
Total	15950	15641	363786	363651	135

*
1 – stations are not present in the stations table

1.1.2 Waterbase - Lakes: Stations

Country Code	Numbers of records				
	in the latest delivery		total in the working database	in the Waterbase	excluded from the Waterbase - reasons*
	total	processed			1
AL			5	5	
AT	36	36	36	36	
BA	11	11	11	11	
BE	5	5	5	5	
BG	15	15	27	27	
CH	26	26	26	26	
CY	6	6	6	6	
CZ					
DE	20	20	20	20	
DK	23	23	26	26	
EE	13	13	13	10	3
ES	402	402	402	402	
FI	246	246	274	274	
FR	193	193	212	212	
GB			228	228	
GR			25	25	
HR	9	9	29	29	
HU	11	11	22	22	
IE	31	31	32	32	
IS	39	39	39	39	
IT	173	173	216	216	
LI					
LT	15	15	43	43	
LU					
LV	14	14	29	29	
ME	11	0			

MK	3	3	3	3	
MT	2	2	2	2	
NL			13	13	
NO	148	148	148	148	
PL	10	10	10	10	
PT	31	31	32	32	
RO	16	16	16	16	
RS	77	77	77	77	
SE	181	181	191	191	
SI			12	12	
SK					
TR					
Total	1767	1756	2230	2227	3

*
1 – stations with the same coordinates have been merged

1.1.3 Waterbase - Lakes: Pressures

Country Code	Numbers of records				
	in the latest delivery		total in the working database	in the Waterbase	excluded from the Waterbase - reasons*
	total	processed			1
AL					
AT	36	36	36	31	5
BA					
BE					
BG	15	15	15	15	
CH					
CY	6	6	6	6	
CZ					
DE	19	19	19	19	
DK					
EE	10	10	10	10	
ES					
FI					
FR					
GB					
GR					
HR					
HU					
IE	32	32	32	32	
IS					
IT					
LI					
LT	15	15	15	15	
LU					

LV	14	14	14	14	
ME					
MK					
MT			2	2	
NL					
NO	148	148	148	148	
PL					
PT	32	32	32	29	3
RO					
RS					
SE	181	181	181	181	
SI					
SK					
TR					
Total	508	508	510	502	8

*
1 – all of the pressure fields are empty

1.2 Primary key tests

Primary key is a field or combination of fields with values which have to be unique in the data table. If primary key is duplicated it is an error.

List of data tables primary keys:
STATIONS: CountryCode, WaterbaseID
PRESSURES: CountryCode, WaterbaseID
QUALITY: CountryCode, WaterbaseID, Determinand, Year, AggregationPeriod

Result:
No primary key error has been detected in the STATIONS or PRESSURES tables.

Concerning the QUALITY table the SampleDepth was also intended to be a part of the primary key. However this information has not been provided for a lot of records and couldn't be used as planed. The proper vertical aggregation of the data was also not possible because of the same reason. Therefore it was decided to leave the data as originally delivered. The duplication of the records has been flagged in a special QA field (QA_duplication) showing the number of records with the same primary key combination. It is recommended to consider this issue when using the data in further analysis.

For the purposes of the WISE map viewer the duplicated records were aggregated and simple average was calculated. The records where SampleDepth value was > 3 m were excluded from the use.

1.3 Table relations tests

The unique Waterbase identifier (WaterbaseID) is contained in each of the data tables. It can be used to link data from one table to another. The able relations tests detect identifiers which are not present in some of the tables.

1.3.1 Number of stations without any data in the "QUALITY" table by country*

Country code	No. of stations	Percentage of total no. of stations
BG	7	25.93
DK	6	23.08
ES	402	100
FI	32	11.68
FR	175	82.55
GR	25	100
HR	20	68.97
HU	11	50
IT	20	9.26
LT	28	65.12
LV	1	3.45
MK	1	33.33
PT	6	18.75
RS	4	5.19
SE	10	5.24
Total	748	33.59

*Some of the detected stations might be used for collecting information about hazardous substances in the water only. These data are not included in the Waterbase yet.

1.3.2 Number of stations without any data in the "PRESSURES" table by country

Country code	No. of stations	Percentage of total no. of stations
AL	5	100
AT	5	13.89
BA	11	100
BE	5	100
BG	12	44.44
CH	26	100
DE	1	5
DK	26	100
ES	402	100
FI	274	100
FR	212	100
GB	228	100
GR	25	100
HR	29	100
HU	22	100
IS	39	100
IT	216	100
LT	28	65.12
LV	15	51.72
MK	3	100
NL	13	100
PL	10	100

Country code	No. of stations	Percentage of total no. of stations
PT	3	9.38
RO	16	100
RS	77	100
SE	10	5.24
SI	12	100
Total	1725	77.46

1.3.3 "QUALITY" and "PRESSURES" table records where "WaterbaseID" is not present in the "STATIONS" table

Quality and Pressures records missing connection in the Stations table were been removed from the Waterbase. They will be included after the country will provide such information.

2. Logical rule violation tests

Logical rules were tested in the "QUALITY" data table. This table contains several measurement value fields, calculated in the aggregation process. Logical relations can be detected between them and mathematically transformed in a set of rules. Following rules have been detected and tested:

Rule	Basic validation rules
1	Mean >= Minimum
2	Mean <= Maximum
3	Median >= Minimum
4	Median <= Maximum
5	Minimum <= Maximum
6	StandardDeviation < Maximum
Rule	**Combined validation rules**
13	IF Minimum < Maximum THEN (StandardDeviation > 0)
14	IF NumberOfSamples = 1 THEN (Mean = Minimum = Maximum = Median)
15	IF NumberOfSamples = 1 THEN (StandardDeviation = 0)
16	IF NumberOfSamples = 0 THEN (AllValueType Is Null)
Rule	**Negative value validation rule**
17	All Values >= 0

The following exceptions and modifications were been applied:

IF Maximum = 0 AND StandardDeviation = 0 THEN rule 6 is not violated
IF Determinand = Temperature the values can be negative (exception of the rule 17)
IF Rule 13 is violated THEN change StandardDeviation to Null

A special QA field (QA_LRviolations) has been added to the data table. Information of the rules violated in the respective record are kept there as a coma separated list of those rules numbers (the numbers are the same as in the table above). It is recommended that the records where QA_LRviolation field is not empty (314 records), should not be used in a further analysis. The detected data quality inconsistencies will be tried to be solved in the near future.
The records where the rules 1 and 2 are violated have not been used for the WISE. The records where the Median value was intended to be used instead of the missing Mean but where the rule 4 is violated have been also excluded.

3. Outlier detection

Detection of outliers was performed on the "QUALITY" data table. Following values were analyzed:
Measurement values: mean
Determinands: all
Aggregation periods: all
Years: all

Measurement values were compared with other values from the same time series. If the value was detected as an outlier it was analyzed whether it can be a possible error or whether it was caused by natural conditions.
Some of previously detected errors have been already corrected by countries or were approved as natural high/low values (noted in the Remarks field).

Some whole time series where the measurement values are naturally very high (e.g. because of the positioning of the monitoring station close to the source of the pollution) have been also detected. These time series have not been included in the subset used for the WISE update.

Last part of the outlier detection process was detection of records where Mean value is not provided.

All types of the information mentioned above have been stored in a special QA field (QA_outlier) added to data table. Following QA flags have been used:

1 – record is a potential outlier (48 records). It is recommended not to use these records in the further analysis until the issue is solved by the data suppliers.

2 – measurements are probably taken from a highly polluted locations (0 records). It is recommended not to use them for calculation of average concentrations of nutrients for broader areas like RBD or whole Country. The representativeness of the result can be negatively affected. Records have not been used for the WISE.

3 – mean value is missing (149 records)

4. Station coordinates tests

Positions of all reported monitoring stations have been tested using the coordinates provided. If the coordinates locate the station outside the respective country borders or if coordinates are missing this information is stored in a special QA field (QA_coordinates_err for the “Stations” table, QA_station_err for the “Quality” an “Pressures” table). Following QA flags have been used:

1 – monitoring station is located outside the respective country borders – either on the sea or in another country (2 stations, 207 quality records, 0 pressures records)

2 – coordinates are missing (30 stations, 340 quality records, 0 pressures records)

These data quality inconsistencies will be tried to be solved in the near future.

In addition if more then one station has the same coordinates, the flag number 3 has been used to mark the records (140 stations, 5218 quality records, 1 pressures records). This is not considered as an error. The coordinates provided are probably not the coordinates of the station but of the lake. It is accepted at the present.