Waterbase – Lakes
Version 10

Quality control documentation

31 May 2010

Waterbase – Lakes

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.

The dataset contains data on nutrients and organic matter, proxy pressure data on the upstream catchments and physical characteristics of the WISE-SoE lake monitoring stations.

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, 3 and 4).
The Quality control tests have been performed on the Waterbase - Lakes database provided in April 2010 by ETC/WTR. This database is included in the EEA data service as version 10, and is publicly available. The database and metadata are available at the following URL: http://www.eea.europa.eu/data-and-maps/data/waterbase-lakes-6

Waterbase – Lakes dataset contains three data tables:
- QUALITY
- STATIONS
- PRESSURES

Five types of the tests have been performed on the data tables. Basic tests, Logical rules violation test, Outlier detection, Stations tests, and Valid data type and codes tests.

Chemical rules tests, that were first time introduced previous year, were also tested by the ETC/Water but the result were not incorporated due to critical comments obtained from countries as a reaction to number of false positive errors detected by the tests. The rules are presently being revised.

1. Basic tests

1.1 Summary

1.1.1 Waterbase - Lakes: QUALITY

Country	Number of records								
	from the last delivery					in the ETC working database		Waterbase	
	total	inserted into working database							
		new		redelivered old					
		total	QA issue	total	QA issue	total	QA issue	**total**	QA issue
AL	93	0	0	0	0	95	80	95	76
AT	317	317	0	0	0	1457	9	1457	31
BA	50	50	50	0	0	525	477	525	461
BE	64	64	7	0	0	186	13	186	10
BG	180	180	0	0	0	1153	88	1153	47
CH	148	148	0	0	0	2772	309	2772	321
CY	117	117	0	0	0	469	33	469	19
CZ	0	0	0	0	0	0	0		
DE	379	379	379	0	0	2380	380	2380	82
DK	691	691	5	0	0	5705	733	5705	808
EE	285	285	32	0	0	1568	62	1563	58
ES	4127	4127	547	0	0	4127	432	4127	285
FI	9218	9218	8	0	0	199173	22057	182148	22510
FR	0	0	0	0	0	400	182	400	180
GB	3408	3408	1187	0	0	11956	6677	11864	7066
GR	0	0	0	0	0	0	0		
HR	162	162	2	0	0	896	53	896	10
HU	0	0	0	0	0	9149	1408	9149	930
IE	716	716	0	727	0	4386	292	4133	576
IS	10	10	10	0	0	137	10	137	
IT	3726	3726	1660	0	0	9664	2811	9516	2368
LI	0	0	0	0	0	0	0		
LT	177	177	0	0	0	1397	18	1382	28
LU	0	0	0	0	0	0	0		
LV	180	180	0	0	0	2504	100	2503	102
ME	0	0	0	0	0	0	0		
MK	15	15	15	0	0	88	16	88	3
MT	0	0	0	0	0	21	0	21	
NL	400	400	0	0	0	1713	93	1713	51
NO	573	573	415	0	0	7867	415	7867	
PL	367	367	359	0	0	1281	1273	1281	1275
PT	324	324	3	0	0	887	86	867	55
RO	198	198	198	0	0	811	237	811	53
RS	2627	2627	2575	0	0	5202	1190	5202	154
SE	1350	1350	0	0	0	34805	345	34805	343
SI	14	14	0	0	0	1006	74	1006	48
SK	349	349	58	0	0	518	123	518	111
TR	36	36	8	0	0	81	53	81	57
Total	30301	30208	7518	727	0	314379	40129	296820	38118

1.1.2 Waterbase - Lakes: STATIONS

Country	Number of records								
	from the last delivery					in the ETC working database		Waterbase	
	total	inserted into working database							
		new		redelivered old					
		total	QA issue	total	QA issue	total	QA issue	**total**	QA issue
AL	5	0	0	0	0	5	4	5	4
AT	28	0	0	28	0	37	0	37	
BA	6	0	0	6	4	11	8	11	11
BE	5	0	0	5	0	5	5	5	5
BG	15	0	0	15	0	27	0	27	
CH	26	0	0	26	0	26	0	26	
CY	9	0	0	9	0	9	0	9	
CZ	0	0	0	0	0	0	0		
DE	46	27	0	19	0	47	0	47	
DK	19	0	0	19	0	26	0	26	
EE	17	0	0	17	0	17	0	17	
ES	408	352	0	56	0	754	0	754	
FI	246	0	0	246	1	274	1	274	1
FR	198	198	147	0	0	410	296	410	296
GB	117	28	48	89	0	228	160	256	188
GR	0	0	0	0	0	25	0	25	
HR	9	0	0	9	0	29	0	29	
HU	0	0	0	0	0	35	0	35	
IE	73	0	0	73	0	96	1	96	96
IS	1	0	0	1	0	39	0	39	
IT	219	7	0	212	76	376	98	376	97
LI	0	0	0	0	0	0	0		
LT	12	1	0	11	0	57	0	57	
LU	0	0	0	0	0	0	0		
LV	10	3	0	7	0	45	17	45	17
ME	0	0	0	0	0	0	0		
MK	3	0	0	3	0	3	0	3	
MT	0	0	0	0	0	2	2	2	2
NL	17	12	0	5	0	25	6	25	6
NO	148	0	0	0	0	148	0	148	
PL	41	1	0	40	40	47	46	47	46
PT	29	0	0	29	0	32	0	32	
RO	16	16	0	0	0	16	0	16	16
RS	72	1	1	71	0	78	1	78	
SE	119	0	0	119	0	200	0	200	
SI	0	0	0	0	0	12	0	12	
SK	23	23	0	0	0	23	0	23	23
TR	4	4	0	0	0	9	5	9	5
Total	1941	673	196	1115	121	3173	650	3201	813

1.1.3 Waterbase - Lakes: PRESSURES

Country	Number of records								
	from the last delivery					in the ETC working database		Waterbase	
	total	inserted into ETC working database							
		new		redelivered old					
		total	QA issue	total	QA issue	total	QA issue	**total**	QA issue
AL	5	0	0	0	0	0	0		
AT	0	0	0	0	0	32	0		
BA	0	0	0	0	0	0	0		
BE	0	0	0	0	0	0	0		
BG	0	0	0	0	0	15	0	14	
CH	26	26	0	0	0	26	0	26	14
CY	8	8	0	0	0	8	0	8	
CZ	0	0	0	0	0	0	0		
DE	39	25	0	14	0	44	0	40	34
DK	0	0	0	0	0	0	0		
EE	17	0	0	17	4	17	4	17	
ES	408	352	0	56	0	754	0	754	
FI	0	0	0	0	0	0	0		
FR	0	0	0	0	0	0	0		
GB	0	0	0	0	0	0	0		
GR	0	0	0	0	0	0	0		
HR	0	0	0	0	0	0	0		
HU	0	0	0	0	0	18	0	18	
IE	69	0	0	69	0	84	12	72	72
IS	1	1	0	0	0	1	0	1	1
IT	0	0	0	0	0	0	0		
LI	0	0	0	0	0	0	0		
LT	12	1	0	11	0	29	0	29	29
LU	0	0	0	0	0	0	0		
LV	0	0	0	0	0	27	0	27	
ME	0	0	0	0	0	0	0		
MK	0	0	0	0	0	0	0		
MT	0	0	0	0	0	2	0	2	2
NL	0	0	0	0	0	0	0		
NO	0	0	0	0	0	148	0	148	
PL	40	0	0	40	0	40	40	40	40
PT	29	0	0	29	0	29	0	29	
RO	0	0	0	0	0	0	0		
RS	0	0	0	0	0	0	0		
SE	119	0	0	119	0	193	0	193	
SI	0	0	0	0	0	0	0		
SK	23	13	0	10	1	36	14	36	14
TR	0	0	0	0	0	0	0		
Total	796	426	0	365	5	1503	70	1454	206

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 which has to be solved.

List of data tables primary keys:

- QUALITY: CountryCode, Waterbase_ID, Determinand, Year, AggregationPeriod
- STATIONS: CountryCode, Waterbase_ID
- PRESSURES: CountryCode, Waterbase_ID

Result:
No primary key error has been detected.

1.3 Table relations tests

The unique Waterbase identifier (WaterbaseID) is present in each of the data tables. It can be used to link data from one table to another. The table 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 ststions
BG	7	25.93%
DE	8	17.02%
DK	6	23.08%
ES	448	59.42%
FI	28	10.22%
FR	373	90.98%
GR	25	100.00%
HR	20	68.97%
HU	12	34.29%
IE	3	3.13%
IT	35	9.31%
LT	28	49.12%
LV	1	2.22%
MK	1	33.33%
NL	3	12.00%
PT	1	3.13%
SE	8	4.00%
Total	1007	31.46%

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

Country Code	No. of GW bodies	Percentage of total no. of GW bodies
AL	5	100.00%
AT	37	100.00%
BA	11	100.00%
BE	5	100.00%
BG	13	48.15%
CY	1	11.11%
DE	7	14.89%
DK	26	100.00%
FI	274	100.00%
FR	410	100.00%
GB	256	100.00%
GR	25	100.00%
HR	29	100.00%
HU	17	48.57%
IE	24	25.00%
IS	38	97.44%
IT	376	100.00%
LT	28	49.12%
LV	18	40.00%
MK	3	100.00%
NL	25	100.00%
PL	7	14.89%
PT	3	9.38%
RO	16	100.00%
RS	78	100.00%
SE	7	3.50%
SI	12	100.00%
SK	1	4.35%
TR	9	100.00%
Total	1761	55.01%

1.3.3 “QUALITY” and “PRESSURES” table records where none of the stations is present in the “STATIONS” table

Table	Country Code	No of records	Percentage of total no of records
QUALITY	IT	10	0.11%
QUALITY	NL	3	0.18%
QUALITY	SK	12	2.32%
QUALITY	Total	25	0.01%
PRESSURES	SK	14	38.89%
PRESSURES	Total	14	0.96%

All of these records are marked in the dataset (see section 4 for more details)

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
Rule	**Specific logical rules**
18	IF NoOfSubsites > 1 THEN (MethodOfAggregation Is Not Null)

The following exceptions and modifications were been applied:

IF Maximum = 0 AND StandardDeviation = 0 THEN rule 6 is not violated

A special QA field (QA_LRviolations) has been added to the data tables. 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 **(3553 Quality records)**, should not be used in a further analysis or only after a careful consideration. The detected data quality inconsistencies will be tried to be solved in the near future.

3. Outlier detection

Detection of outliers was performed on the “QUALITY” data.

Measurement “Mean” values were tested against limiting values individually defined by an expert for each of the determinands and also statistically 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.

Records where Mean value is not provided are also acknowledged as outliers.

The findings described above have been stored in a special QA field (QA_outlier) added to data table. Following QA flags have been used:

-1 – record has been confirmed by the respective country as being correct **(134 Quality records)**

1 – standard potential outlier - value is either higher/lower than limit value or is suspiciously high/low comparing to the rest of the time series or value change between two consecutive values is suspiciously abrupt or was marked as an potential outlier by a content expert **(278 Quality record)**

2 – measurements are probably taken from a highly polluted locations **(1 Quality records)**. It is recommended not to use them for calculation of average concentrations of nutrients for broader areas like RBD or whole Country.

3 – the whole or a big part of the particular country delivery is considered as problematic because it contains too many quality issues **(910 Quality records: 910 records HU 2007)**

10 – the Mean value = 0 **(1385 Quality records)**. Value is not correct and records should not be used.

99 – the Mean value is empty **(19316 Quality records)**. Record can’t be used.

4. Stations tests

Positions of all reported monitoring stations have been tested using the coordinates provided as well as stations availability. The cases when the station coordinates fall outside the respective country borders, when coordinates are missing or when the monitoring station available in the Quality or Pressures data tables is not available in the Stations table, are documented in a special QA field (QA_station_problem). In addition some other station related issues were tested. Following QA flags have been used:

-1 – station coordinates fall slightly outside the respective country boundary, but were confirmed as correct by country **(1 station, 791 quality records)**

1 – monitoring station is located outside the respective country borders – either on the sea or in another country **(1 station, 49 quality records)**

2 – coordinates are missing **(23 stations, 178 quality records)**

4 – more stations with the same coordinates **(505 stations, 7603 quality records)**

13 – lake area is suspicious **(141 stations, 3872 quality records, 40 pressures records)**

14 – lake depth is suspicious **(5 stations, 250 quality records)**

99 – station is not available in the Stations table **(148 quality records, 14 pressures records)** – see result 1.3.3

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

5. Data type and codes tests

All Lakes dataset values have to follow specifications defined in the respective Data dictionary (DD) definitions. Values, which are of a different data type as requested (e.g. string instead of numeric) or which are not available in a set of allowable values, have been either removed or, if possible, replaced by a correct value.

There is one exception from this rule. Some of these “errors” are only formally wrong. The value is still valid but was not foreseen as possible and was therefore omitted to be included in the current DD definitions of the respective table field. In this case the original code has been left in the field untouched. It is planed that these codes will be added into the code list during the next DD update.

In all the cases the original, incorrect value or value missing in the DD code list, has been stored in a special QA field (QA_datatype_error) in the following format:

Name_of_field: Erroneous_Value; [Name_of_field: Erroneous_Value; …]

Test result summary:

Quality table: 21402 records
Stations table: 179 records
Pressures table: 167 records

The cases where the errors couldn’t be corrected will be tried to be solved in the near future in cooperation with the data providers.