Waterbase – Lakes
Version 9

Quality control documentation

16 October 2009

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 lake monitoring stations, proxy pressures on the upstream catchments areas, as well as chemical quality data on nutrients and organic matter and hazardous substances in lakes.

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 March 2009 by ETC/WTR. This database is included in the EEA data service as version 9, and is publicly available. The database and metadata are available at the following URL: http://dataservice.eea.europa.eu/dataservice/metadetails.asp?id=1088

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

Waterbase – Lakes dataset contains four data tables:

- HAZSUB
- PRESSURES
- QUALITY
- STATIONS

Six types of the tests have been performed on the data tables. Basic tests, Logical rules violation test, Chemical rules violation test, Outlier detection, Station data tests and Data type tests.

1. Basic tests

1.1 Summary

1.1.1 Waterbase - Lakes: Quality

Country Code	Number of records									
	latest (2008) country delivery							ETC working database	Waterbase	
	total	processed				excluded			Total	with quality issue/s detected
		total		with quality issue/s detected by the ETC		number	reason*			
		new	redelivered	new	redelivered					
AL	94	94		94		0		189	95	80
AT	216	216				0		1140	1140	7
BA	112	112				0		475	475	424
BE	55	55				0		122	122	
BG	149	149				0		973	962	63
CH	41	41				0		2624	2315	159
CY	352	183	169			0		413	413	32
DE	117	117		117		0		2163	2046	41
DK	0	0				0		5014	4300	14
EE	473	431	42			0		1750	1696	4
ES	0	0				0		0	0	
FI	11510	6225	5285			0		189980	155310	3340
FR	0	0				0		400	400	4
GB	0	0				0		8548	8548	5702
GR	0	0				0		0	0	
HR	142	142				0		734	734	50
HU	1163	1163				0		9397	9149	1406
IE	489	489				0		3432	2960	70
IS	10	10				0		127	127	1
IT	2523	2273				250	1	7425	7159	658
LT	270	270				0		1315	1270	1
LV	200	200				0		2324	2309	85
ME	308	0				308	2	616	0	
MK	5	5				0		73	73	1
MT	2	2				0		21	21	
NL	0	0				0		1313	1313	456
NO	7294	603	6691			0		7294	7294	
PL	401	401				0		953	904	904
PT	377	287				90	3	563	518	30
RO	107	107				0		613	613	38
RS	973	973				0		4234	4161	231
SE	1109	1109				0		33455	33455	156
SI	14	14				0		1098	1054	30
SK	277	229				48	4	229	169	169
TR	125	125				0		125	125	125
Total	28908	16025	12187	211	0	696		289132	251230	749

*

1 – data were aggregated
2 – delivery was not possible to process
3 – data of insufficient quality removed
4 – supportive determinands were populated in appropriate fields

1.1.2 Waterbase - Lakes: Stations

Country Code	Number of records									
	latest (2008) country delivery							ETC working database	Waterbase	
	total	processed by the ETC				excluded				
		total		with quality issue/s detected by the ETC		number	reason*		**Total**	with quality issue/s detected
		new	redelivered	new	redelivered					
AL	5	0	5			0		5	5	4
AT	28	1	27			0		37	37	
BA	9	0	9			0		11	11	8
BE	5	0	5			0		5	5	5
BG	15	0	15			0		27	27	
CH	26	0	26			0		26	26	2
CY	9	3	6			0		9	9	
DE	20	0	20			0		20	20	
DK	0	0	0			0		26	26	23
EE	17	8	9			0		17	17	17
ES	402	0	402			0		402	402	
FI	246	0	246			0		274	274	1
FR	193	0	193			0		212	212	2
GB	0	0	0			0		228	228	197
GR	0	0	0			0		25	25	
HR	9	0	9			0		29	29	
HU	18	13	5			0		35	35	18
IE	76	63	13			0		97	96	
IS	1	0	1			0		39	39	
IT	249	147	102			0		216	369	23
LT	23	13	10			0		56	56	
LV	17	11	6			0		42	42	17
ME	11	0	11	11		0		11	0	
MK	3	0	3			0		3	3	
MT	2	0	2			0		2	2	2
NL	0	0	0			0		13	13	7
NO	148	0	148			0		148	148	
PL	40	36	4			0		46	46	46
PT	32	0	32			0		32	32	
RO	16	0	16			0		16	16	16
RS	70	0	70			0		77	77	
SE	119	12	107			0		200	200	
SI	2	0	2			0		12	12	
SK	0	0	0			0		0	0	
TR	5	5	0			0		5	5	5
Total	1816	312	1504	11	0	0		2403	2544	393

1.1.3 Waterbase - Lakes: Pressures

Country Code	Number of records									
	latest (2008) country delivery							ETC working database	Waterbase	
	total	processed by the ETC				excluded				
		total		with quality issue/s detected by the ETC		number	reason*		**Total**	with quality issue/s detected
		new	redelivered	new	redelivered					
AL	5	0				5	1	0	0	
AT	28	28				0		32	32	
BA	0	0				0		0	0	
BE	0	0				0		0	0	
BG	0	0				0		15	15	
CH	26	26				0		26	26	2
CY	8	8				0		8	8	
DE	20	0	20	20		0		20	19	
DK	0	0				0		0	0	
EE	17	8	9			0		18	17	4
ES	402	402				0		402	402	
FI	0	0				0		0	0	
FR	0	0				0		0	0	
GB	0	0				0		0	0	
GR	0	0				0		0	0	
HR	0	0				0		0	0	
HU	18	18				0		18	18	
IE	72	72				0		84	84	12
IS	1	0				1	1	0	0	
IT	0	0				0		0	0	
LT	23	23				0		28	28	
LV	17	13	4			0		27	27	
ME	0	0				0		0	0	
MK	0	0				0		0	0	
MT	2	2				0		4	4	
NL	0	0				0		0	0	
NO	148	0	148			0		148	148	
PL	40	40				0		40	40	40
PT	29	29				0		29	29	
RO	0	0				0		0	0	
RS	0	0				0		0	0	
SE	119	12	107			0		193	193	
SI	0	0				0		0	0	
SK	23	23				0		23	23	23
TR	0	0				0		0	0	
Total	998	704	288	20	0	6		1115	1113	81

*

1 – data of insufficient quality removed

1.1.4 Waterbase - Lakes: HazSub

Country Code	Number of records								
	latest (2008) country delivery						ETC working database	Waterbase	
	total	processed by the ETC				note*		Total	with quality issue/s detected
		total		with quality issue/s detected by the ETC					
		new	redelivered	new	redelivered				
AL	0	0					0	0	
AT	0	0					0	0	
BA	315	315					656	656	656
BE	577	577					1259	1259	339
BG	46	46					141	141	26
CH	107	107					1703	1703	601
CY	1959	584	1375				2047	2047	1
DE	295	295		295			1759	1464	1
DK	0	0					0	0	
EE	174	159	15				179	179	
ES	0	0					0	0	
FI	4696	4456	240				40559	40559	219
FR	0	0					29	29	
GB	0	0					10497	10497	5867
GR	0	0					0	0	
HR	204	204					986	986	5
HU	2430	2316				1	12383	12383	348
IE	3877	2508				2	2508	2508	2146
IS	21	21					21	21	
IT	11886	8516				1	10480	10480	965
LT	0	0					0	0	
LV	62	62					191	191	1
ME	0	0					0	0	
MK	0	0					0	0	
MT	0	0					0	0	
NL	0	0					0	0	
NO	6324	271	6053				6324	6324	
PL	0	0					0	0	
PT	598	258				3	635	635	2
RO	47	47					47	47	
RS	1897	1897					4219	4219	92
SE	1812	1812					29146	29146	78
SI	0	0					0	0	
SK	237	237					285	285	285
TR	0	0					0	0	
Total	37564	24688	7683	295	0		126054	125759	455

*
1 – data were aggregated
2 – duplicated values removed
3 – data of insufficient quality removed

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, SampleDepth, MethodOfAggregation
HAZSUB: CountryCode CountryCode, WaterbaseID, Determinand_hazsubs, Year, Month, Day, SampleDepth, SampleAnalysis

Result:
3 QUALITY records from FI duplicated after a nonstandard determinand “BOD7 with adding of allylthiourea” changed to official determinand “BOD7”. All these records are flagged in the QA_datatype_error field (see section 6)

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" or “HAZSUB” table by country

Country Code	No. of stations	Percentage of total no. of stations
BG	7	25.93
DK	6	23.08
ES	402	100.00
FI	28	10.22
FR	175	82.55
GR	25	100.00
HR	20	68.97
HU	11	31.43
IE	1	1.04
IT	37	10.03
LT	28	50.00
LV	1	2.38
MK	1	33.3333
NL	3	23.0769
RS	2	2.5974
SE	8	4.00
Total	755	29.68

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	100.00
AT	5	13.51
BA	11	100.00
BE	5	100.00
BG	12	44.44
CY	1	11.11
DE	1	5.00
DK	26	100.00
FI	274	100.00
FR	212	100.00
GB	228	100.00
GR	25	100.00
HR	29	100
HU	17	48.57
IE	12	12.5
IS	39	100
IT	369	100
LT	28	50
LV	15	35.71
MK	3	100
NL	13	100
PL	6	13.04
PT	3	9.38
RO	16	100
RS	77	100
SE	7	3.5
SI	12	100
TR	5	100
Total	1456	57.23

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

Table	Country Code	No. of records	Percentage of total no. of records
QUALITY	IT	36	0.5
QUALITY	NL	18	1.37
QUALITY	SK	169	100
QUALITY	Total	223	0.09
PRESSURES	SK	23	100
PRESSURES	Total	23	2.07
HAZSUB	SK	285	100
HAZSUB	Total	285	0.23

All of these records are marked in the dataset (see section 5 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

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 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 **(3080 Quality records)**, should not be used in a further analysis. 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" and "HAZSUB" data tables. Following values were analyzed:
Measurement values: mean (QUALITY), concentration (HAZSUB)
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.

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 – 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 **(203 Quality records; 1050 HazSub records)**

3 – the whole country delivery is considered as problematic because it contains too many quality issues **(7419 Quality records - HU 2007)**

10 – the Mean/Concentration value = 0 **(1404 Quality records; 128 HazSub records)**. Value is not correct and records should not be used.

99 – the Mean/Concentration value is empty **(121 Quality records; 434 HazSub records)**. Record can't be used.

4. Chemical rule violation tests

Chemical rules were tested in the "QUALITY" data table. Following chemical rules were defined between concentrations of certain related determinands from the same monitoring station, year and aggregation period:

Rule	Definition
1	BOD5 < COD
2	BOD5 < TotalOrganicCarbon
3	Orthophosphate < TotalPhosphorus
4	Total Nitrogen = Kjeldahl Nitrogen + Nitrate + Nitrite
5	Kjeldahl Nitrogen = Organic Nitrogen + Total Ammonium
6	Total Oxidised Nitrogen = Nitrate + Nitrite

The following modifications have been applied to the rules 4, 5 and 6:

If concentration values of all "right side" determinands are provided and are > 0, the +-5% tolerance was applied to the formula to avoid false violations detected because of rounding:

*Rule 4: 0.95*Total Nitrogen < (Kjeldahl Nitrogen + Nitrate + Nitrite) < 1.05*Total Nitrogen*
*Rule 5: 0.95*Kjeldahl Nitrogen < (Organic Nitrogen + Total Ammonium) < 1.05* Kjeldahl Nitrogen*
Rule 6: 0.95 Total Oxidised Nitrogen < (Nitrate + Nitrite) < 1.05 Total Oxidised Nitrogen

If concentration of some of the "right side" determinand is missing or is = 0, it was tested whether the sum of remaining "right side" concentrations is not higher than concentration of the "left side" determinand:

Rule 4: Total Nitrogen > [Kjeldahl Nitrogen] + [Nitrate] + [Nitrite]
Rule 5: Kjeldahl Nitrogen > [Organic Nitrogen] + [Total Ammonium]
Rule 6: Total Oxidised Nitrogen > [Nitrate] + [Nitrite]

A special QA field (QA_CRviolations) has been added to the data table. Information of the rules violated in the respective records (all records of each of the determinands from both sides of formula) 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_CRviolation field is not empty **(3251 records)**, should not be used in a further analysis. The detected data quality inconsistencies will be tried to be solved in the near future.

5. Station coordinates and availability tests

Positions of all reported monitoring stations have been tested using the coordinates provided as well as stations availability. If the station coordinates fall outside the respective country borders or if coordinates are missing or if the monitoring station available in the Quality, Pressures or HazSub data tables is not available in the Stations table, this information is stored in a special QA field (QA_station_problem). Following QA flags have been used:

1 – monitoring station is located outside the respective country borders – either on the sea or in another country **(27 stations, 1011 Quality records, 2 Pressures records, 864 HazSub records)**

4 – more stations with the same coordinates **(137 stations, 5132 Quality records, 6475 HazSub records)**

13 – lake area is suspicious **(236 stations, 6128 Quality records, 40 Pressures records, 5459 HazSub records)**

14 – lake depth is suspicious **(5 stations, 189 Quality records, 120 HazSub records)**

99 – station is not available in the Stations table (**223 Quality records, 23 Pressures records, 285 HazSub records)** – see result 1.3.3

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

6. Data type tests

All Lakes dataset values have to follow specifications defined in the respective Data dictionary. 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. The original, incorrect value 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: 34 records
Stations table: 148 records
Pressures table: 16 records
HazSub table: 2515 records