Waterbase – Rivers
Version 9

Quality control documentation

Waterbase – Rivers

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 rivers 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.

Rivers 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 - Rivers 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 - Rivers database provided in January 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=1081

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

Waterbase – Rivers dataset contains four data tables:
- HAZSUB
- PRESSURES
- QUALITY
- STATIONS

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

1. Basic tests

1.1 Summary
1.1.1 Waterbase - Rivers: 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	957	417	438	0	0	102	1	2250	2250	345
AT	586	586	0	0	0	0		22242	22242	1785
BA	476	476	0	0	0	0		2679	2679	164
BE	859	859	0	0	0	0		6705	6705	216
BG	1108	1108	0	0	0	0		10989	10989	521
CH	56	56	0	0	0	0		725	725	
CY	259	176	83	0	0	0		273	273	31
CZ	828	828	0	0	0	0		12660	12660	98
DE	0	0	0	0	0	0		23442	23442	3198
DK	0	0	0	0	0	0		12417	12417	16
EE	779	779	0	0	0	0		10862	10862	32
ES	5545	5545	0	0	0	0		63747	63747	11495
FI	4255	4255	0	0	0	0		139207	139207	389
FR	19532	19532	0	0	0	0		115378	115378	3354
GB	0	0	0	0	0	0		26045	26045	3986
GR	0	0	0	0	0	0		1833	1833	276
HR	1418	642	641	0	0	135	1	5542	5542	340
HU	10600	7419	0	2139	0	3181	1	86992	86992	15285
IE	418	418	0	0	0	0		4291	4264	268
IS	7	7	0	0	0	0		52	52	2
IT	11539	10342	689	11	0	508	1	19793	19793	308
LI	0	0	0	0	0	0		8	8	
LT	642	642	0	0	0	0		20080	20080	115
LU	25	25	0	0	0	0		328	328	
LV	490	490	0	0	0	0		10062	10062	16
ME	?	0	0	0	0	?	3	0	0	
MK	120	120	0	0	0	0		1791	1791	310
NL	0	0	0	0	0	0		4353	4353	157
NO	1909	276	1633	0	0	0		1909	1909	
PL	1810	1681	0	68	0	129	1	25364	25364	1281
PT	399	399	0	0	0	0		826	826	62
RO	825	823	0	0	0	2	2	5886	5886	23
RS	1059	1059	0	0	0	0		4120	4120	295
SE	1228	1228	0	0	0	0		37683	37683	91
SI	133	133	0	0	0	0		4334	4334	401
SK	940	521	0	0	0	419	1	7435	7435	184
TR	90	30	60	0	0	0		90	90	4
Total	68892	60872	3544	2218	0	4476		692393	692366	2341

*
1 – records with non standard determinand
2 – records without mandatory values
3 – delivery was not possible to process

1.1.2   Waterbase - Rivers: Stations

Country Code	Number of records							ETC working database	Waterbase	
	latest (2008) country delivery								Total	with quality issue/s detected
	total	processed by the ETC				excluded				
		total		with quality issue/s detected by the ETC		number	reason*			
		new	redelivered	new	redelivered					
AL	51	22	29	0	0	0		52	52	1
AT	71	3	68	0	0	0		290	290	
BA	44	0	44	0	0	0		52	52	
BE	59	3	56	0	0	0		63	63	
BG	111	1	110	0	0	0		111	111	
CH	8	0	8	0	0	0		8	8	
CY	33	24	9	0	0	0		33	33	
CZ	0	0	0	0	0	0		73	73	
DE	0	0	0	0	0	0		151	151	
DK	0	0	0	0	0	0		42	42	
EE	61	8	53	0	0	0		61	61	
ES	1222	0	1222	0	0	0		1514	1514	
FI	138	0	138	0	0	0		231	231	
FR	1569	0	1569	0	0	0		1939	1939	
GB	0	0	0	0	0	0		204	204	15
GR	0	0	0	0	0	0		94	94	5
HR	45	0	45	0	0	0		45	45	
HU	97	44	53	0	0	0		154	154	
IE	180	0	180	0	0	0		209	209	
IS	1	0	1	0	0	0		1	1	
IT	866	752	114	1	0	0		1403	1403	2
LI	0	0	0	0	0	0		1	1	
LT	53	0	53	0	0	0		99	99	
LU	4	0	4	0	0	0		4	4	
LV	49	21	28	0	0	0		141	141	
ME	0	0	0	0	0	0		0	0	
MK	20	0	20	0	0	0		20	20	
NL	0	0	0	0	0	0		23	23	
NO	46	0	46	0	0	0		46	46	
PL	136	0	136	0	0	0		136	136	
PT	59	0	59	0	0	0		59	59	
RO	118	0	118	0	0	0		126	126	
RS	76	0	76	0	0	0		77	77	
SE	122	11	111	0	0	0		127	127	
SI	0	0	0	0	0	0		30	30	
SK	89	65	24	0	0	0		125	125	
TR	5	5	0	0	0	0		5	5	
Total	5333	959	4374	1	0	0		7749	7749	23

1.1.3   Waterbase - Rivers: Pressures

Country Code	Number of records							ETC working database	Waterbase	
	latest (2008) country delivery									
	total	processed by the ETC				excluded			Total	with quality issue/s detected
		total		with quality issue/s detected by the ETC		number	reason*			
		new	redelivered	new	redelivered					
AL	51	16	35	0	0	0		45	45	
AT	71	3	68	2	0	0		292	292	2
BA	0	0	0	0	0	0		0	0	
BE	59	0	59	0	0	0		59	59	
BG	0	0	0	0	0	0		0	0	
CH	8	0	8	0	0	0		8	8	
CY	32	23	9	0	0	0		32	32	
CZ	0	0	0	0	0	0		72	72	
DE	0	0	0	0	0	0		147	147	
DK	0	0	0	0	0	0		42	42	
EE	61	8	53	0	0	0		61	61	
ES	1222	0	1222	0	0	0		1514	1514	
FI	0	0	0	0	0	0		5	5	
FR	1569	0	1569	0	0	0		1877	1877	
GB	0	0	0	0	0	0		190	190	15
GR	0	0	0	0	0	0		0	0	
HR	0	0	0	0	0	0		0	0	
HU	97	44	53	0	0	0		154	154	
IE	180	0	180	0	0	0		209	209	
IS	1	0	1	0	0	0		1	1	
IT	0	0	0	0	0	0		0	0	
LI	0	0	0	0	0	0		1	1	
LT	53	0	53	0	0	0		99	99	
LU	4	0	4	0	0	0		4	4	
LV	49	21	28	0	0	0		141	141	
ME	0	0	0	0	0	0		0	0	
MK	20	0	20	0	0	0		20	20	
NL	0	0	0	0	0	0		12	12	
NO	46	0	46	0	0	0		46	46	
PL	136	0	136	0	0	0		136	136	
PT	59	0	59	0	0	0		59	59	
RO	0	0	0	0	0	0		124	124	
RS	0	0	0	0	0	0		0	0	
SE	122	11	111	0	0	0		127	127	
SI	0	0	0	0	0	0		24	24	
SK	89	64	25	0	0	0		123	123	
TR	0	0	0	0	0	0		0	0	
Total	3929	190	3739	2	0	0		5624	5624	17

1.1.4   Waterbase - Rivers: 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	0		0	0	
AT	704	704	0	0	0		16011	16011	
BA	1825	1478	0	0	0	1, 3	1443	1443	
BE	16310	16310	0	0	0		7806	7806	217
BG	844	844	0	0	0		1267	1267	11
CH	114	114	0	0	0		152	152	
CY	1688	495	1193	0	0		1110	1110	
CZ	17134	17134	0	0	0		6387	6387	41
DE	0	0	0	0	0		63066	63066	860
DK	0	0	0	0	0		0	0	
EE	4184	4184	0	0	0		1819	1819	
ES	0	0	0	0	0		0	0	
FI	14781	14781	0	0	0		24641	24641	
FR	3081460	3081460	0	0	0		579753	579753	
GB	0	0	0	0	0		42470	42470	35587
GR		0	0	0	0		2668	2668	
HR	8893	4675	4218	0	0		2041	2041	16
HU	43072	43072	0	28	0		17107	17107	47
IE	6556	6469	0	0	0	1	1588	1588	
IS	16	14	0	0	0		48	48	
IT	237653	237653	0	26	0		29610	29610	26
LI	0	0	0	0	0		0	0	
LT	3517	3517	0	0	0		6672	6672	
LU	90	90	0	0	0	2	456	456	
LV	1397	1397	0	94	0		2286	2286	114
ME	0	0	0	0	0		0	0	
MK	1442	1442	0	0	0		1090	1090	
NL	0	0	0	0	0		383	383	
NO	10175	2648	7527	0	0		1852	1852	
PL	898	898	0	0	0	2	1961	1961	
PT	332	332	0	0	0	2	948	948	
RO	286	252	0	0	0	1	452	452	
RS	6072	6072	0	0	0		3348	3348	
SE	13157	13157	0	0	0		6371	6371	
SI	1512	1512	0	0	0		10017	10017	680
SK	33219	33045	0	0	0	1	4675	4675	
TR	0	0	0	0	0		0	0	
Total	3507331	3493749	12938	148	0		839498	839498	680

*
1 – supportive determinands were populated in appropriate fields (e.g. Hardness)
2 – data delivered aggregated
3 – data with no Concetration removed before aggregation

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
HAZSUB: CountryCode, WaterbaseID, Determinand_HazSubs, Year, AggregationPeriod

Result:
No primary key error has been detected.

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
BA	6	11.54
CY	9	27.27
EE	1	1.64
ES	551	36.39
FR	250	12.89
GR	9	9.57
IE	11	5.26
IT	8	0.57
PT	3	5.08
SE	2	1.57
SK	34	27.20
Total	884	11.41

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	7	13.46
BA	52	100.00
BE	4	6.35
BG	111	100.00
CY	1	3.03
CZ	1	1.37
DE	4	2.65

Country Code	No. of stations	Percentage of total no. of stations
FI	226	97.84
FR	62	3.20
GB	14	6.86
GR	94	100.00
HR	45	100.00
IT	1403	100
NL	11	47.83
RO	2	1.59
RS	77	100
SI	6	20
SK	2	1.6
TR	5	100
Total	2127	27.45

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 stations
QUALITY	HU	177	0.2
QUALITY	Total	177	0.03
PRESSURES	AT	2	0.68
PRESSURES	Total	2	0.04
HAZSUB	BE	211	2.7
HAZSUB	BG	3	0.24
HAZSUB	DE	859	1.36
HAZSUB	GB	32006	75.36
HAZSUB	HR	16	0.78
HAZSUB	HU	42	0.25
HAZSUB	LV	114	4.99
HAZSUB	SI	680	6.79
HAZSUB	Total	33931	4.04

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" and "HAZSUB" data tables. These tables contain 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 **(2139 Quality records; 605 HazSub 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" 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.

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 has been confirmed by the respective country as correct **(6 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 **(588 Quality records; 49 HazSub records)**

2 – measurements are probably taken from a highly polluted locations **(124 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 country delivery is considered as problematic because it contains too many quality issues **(7419 Quality records - HU 2007)**

10 – the Mean value = 0 **(6049 Quality records; 882 HazSub records)**. Value is not correct and records should not be used.

99 – the Mean value is empty **(4261 Quality records; 244 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 **(77709 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 – stations coordinates fall slightly outside the country land boundary but country confirmed them as correct **(1 Station)**

1 – monitoring station is located outside the respective country borders – either on the sea or in another country **(15 stations, 3024 Quality records, 15 Pressures records, 2988 HazSub records)**

2 – coordinates are missing **(7 stations, 88 quality records, 26 HazSub records)**

99 – station is not available in the Stations table (**177 Quality records, 2 Pressures records, 33931 HazSub records)** – see result 1.3.3

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