PAN-EUROPEAN INFRASTRUCTURE FOR OCEAN & MARINE DATA MANAGEMENT

SECOND RELEASE OF THE AGGREGATED DATA SETS PRODUCTS

WP10 FOURTH YEAR REPORT - DELIVERABLE D10.4

Project Acronym : SeaDataNet II
Project Full Title : SeaDataNet II: Pan-European infrastructure for ocean and marine data management
Grant Agreement Number : 283607

Deliverable number	**Short Title**
D10.4	V2 Aggregated data sets
Long title	
Second Release of the Aggregated Data Sets Products	
Short description	
This document describes the V2 datasets per sea region (Mediterranean Sea, Black Sea, Arctic Sea, Baltic Sea, North Sea and the North Atlantic Ocean).	
Author	Working group
S. Simoncelli, C. Coatanoan, V. Myroshnychenko, H. Sagen, Ö. Bäck, S. Scory, A. Grandi, R. Schlitzer, M. Fichaut	Regional coordinators

History

Version	Authors	Date	Comments
1.0	S. Simoncelli	01/07/2015	Creation
	O. Back	03/07/2015	Baltic Report
	C.Coatanoan	21/07/2015	Atlantic Report
	V. Myroshnychenko	22/07/2015	Black Sea Report
	S. Scory	23/07/2015	North Sea Report
	H. Sagen	29/07/2015	Provide partial Arctic Report
	V. Myroshnychenko	24/08/2015	Updated Black Sea Report

Table of contents

List of Figures

List of Tables

1. Introduction

Main objectives WP10 during the fourth year of activity were:

1. to analyze data distribution and density (space, time, depth) of the **V2 regional aggregated datasets**;
2. To assess data quality making use of the ODV tool and verify the overall improvement of SDN database content;
3. to report to the SDN national nodes and data providers on the data quality and possible shortcomings;
4. to deliver regional aggregated data sets to specific user communities as MyOcean, now Copernicus Marine Environment Monitoring Service (hereafter CMEMS) In situ TAC as agreed;
5. to deliver the V2 aggregated datasets to the SDN partners responsible of dissemination in order to include them on SEXTANT catalogue, OCEANTRON and OceanBrowser.

A Quality Control strategy, schematized in Figure 1, has been developed and continuously refined by WP10 Regional Coordinators in order to improve the SDN database content and to create the best product deriving from SDN data. The QC strategy involved NODCs and data providers that, on the base of WP10 data quality assessment outcome, checked and eventually corrected anomalies in the original data. The QC procedure has been designed to be iterative and facilitate the update of SDN database content. The QC strategy was implemented in collaboration with MyOcean2 and MyOcean Follow On projects in order to implement a true synergy at regional level, create the best historical datasets to serve operational oceanography and climate change communities.

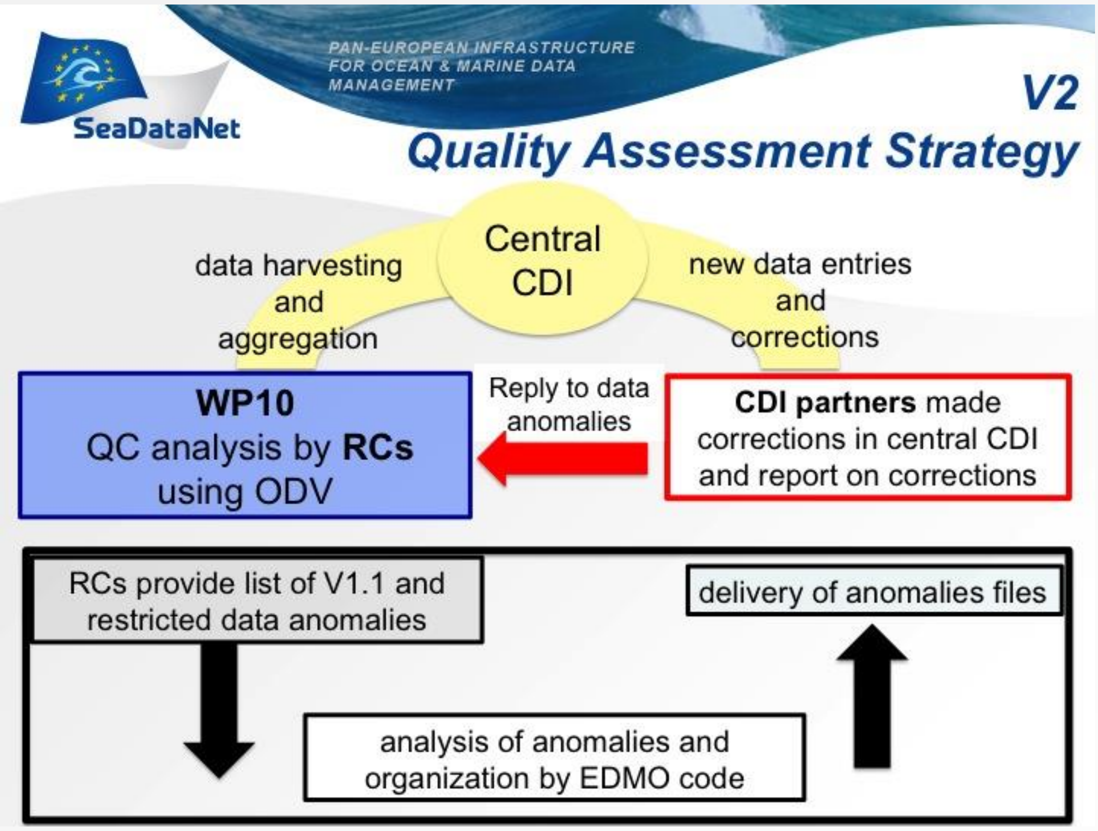
All data unrestricted and under SeaDataNet license were harvested by MARIS from the distributed NODCs and released to AWI for the ODV aggregation procedure at the end of March 2015. The entire data collection contains 1161439 ODV files, which have been divided over 117 zipfiles, each containing ODV files and the relative csv file with metadata. All CDI and ODV were well connected by their Local-CDI_ID-EDMO-code combinations and this allowed creating metadata enriched ODV collections, which represent an important improvement with respect to V1.1 collections.

V2 regional TS data collections (Arctic Sea, Baltic Sea, North Sea, North Atlantic Ocean, Mediterranean Sea, Black Sea) were analysed to assess and report on the quality of these products. The objective were:

- to report to data providers about data anomalies within SDN infrastructure;
- to identify the progresses on the quality of the overall CDI data content;
- to point out the advancement of number of temperature and salinity data contained in the CDI;
- to provide statistics about the data Quality Flags (QF);
- to release SDN qualified temperature and historical data collections to serve the downstream user community;
- to bring about conclusive remarks about the adopted harvesting and quality assessment procedures;
- to bring about conclusive remarks about the adopted harvesting and quality assessment procedures;

- to rise advices about future harvesting and quality assessment procedures;
- to contribute to the development and consolidation of ODV software.

RC after the QC analysis released sub-sets of the aggregated TS collections, covering the time period 1990-2014 to the CMEMS INSTAC. RCs sent the third data anomaly report to NODCs and data providers asking to make the suggested corrections on the data QF according with their expertise.



SeaDataNet

PAN-EUROPEAN INFRASTRUCTURE
FOR OCEAN & MARINE DATA
MANAGEMENT

V2

Quality Assessment Strategy

Central CDI

data harvesting and aggregation

new data entries and corrections

WP10
QC analysis by **RCs** using ODV

Reply to data anomalies

CDI partners made corrections in central CDI and report on corrections

RCs provide list of V1.1 and restricted data anomalies

analysis of anomalies and organization by EDMO code

delivery of anomalies files

Figure 1 Schema of the Quality Check strategy implemented by WP10 in synergy with NODCs.

It follows the description of the regional data collections.

2. Mediterranean Sea

The historical data collection of the Mediterranean Sea contains temperature and salinity observations between -9.25 and 37 degrees of longitude, thus including an Atlantic box and the Marmara Sea that will be treated separately. The spatial distribution and the data density of measurement stations are shown in Figure 2. The spatial distribution of data (Figure 2a) presents a good data coverage in the Western Mediterranean basin and the Atlantic box, while in the Eastern Mediterranean many areas are characterized by few and sparse data, like the coastal areas of Tunisia, Libya, Croatia and Turkey. Data density map (Figure 2b) highlights that observations are more concentrated along the coastal areas of Spain, France and Italy

(Ligurian Sea and Northern Adriatic Sea). In the eastern part of the Basin, maximum data concentration is along the Israeli and the Greek coasts.

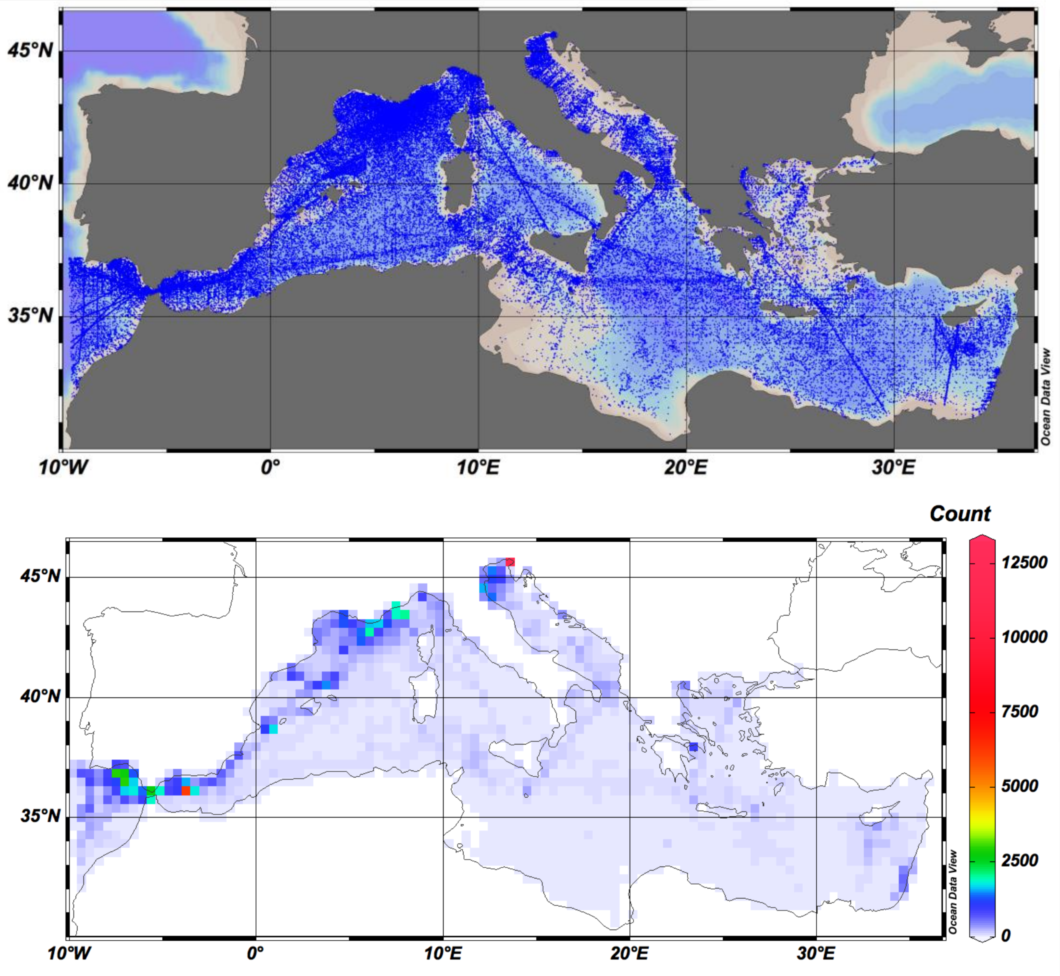
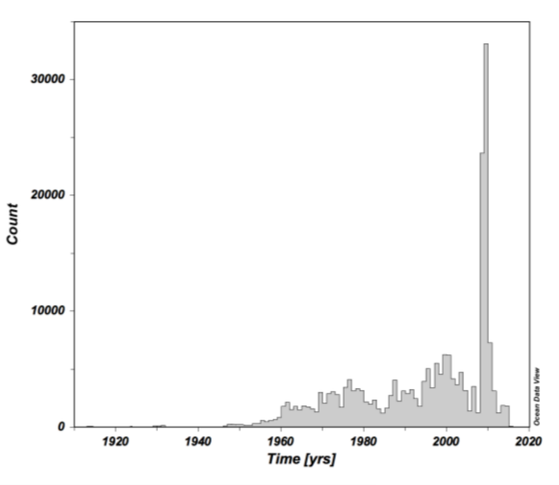
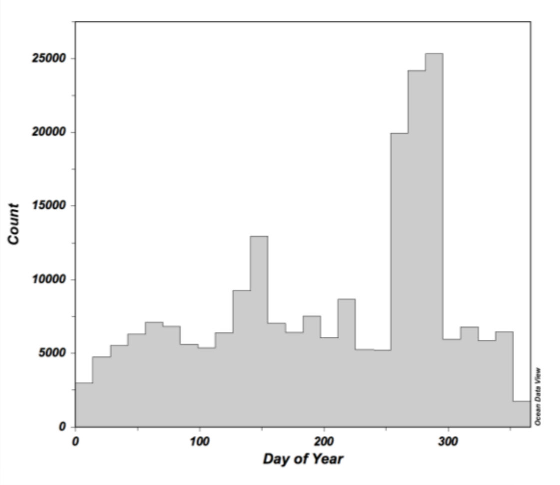


Figure 2 Temperature and Salinity data collection for the Mediterranean Sea in the time period 1900-2014: (a) Data distribution map; (b) Data density map.



(a)



(b)

Figure 3 (a) Annual data distribution and (b) seasonal data distribution for the time period 1900-2014 in the Mediterranean Sea.

Temporal distributions of data are in Figure 3. Annual distribution (a) proves that data are very sparse before 1945 and they start to increase systematically from the sixties and concentrate mostly between the end of the nineties and beginning of noughties. The highest peak corresponds to year 2009, when a lot of sea gliders observations have been sampled along the coasts of France and Spain. This must be taken into consideration during climatological data analysis, like climatology computation. Seasonal distribution of data presents a good coverage all year long. A peak in number of data is present at the end of summer beginning of autumn (September, October) and this might be due to surveys dedicated to monitor particular events. This is another aspect to consider carefully for climatological analysis or other applications.

Figure 4 shows Temperature (a) and Salinity (b) data distribution. Salinity observations are less and sparser than temperature ones. Both maps show the presence of data along ship tracks, along coastal transects, and regular monitoring arrays. Tab. 1 compares V1.1 and V2 number of observed stations and their repartition in temperature and salinity stations and stations having both measurements. The statistics have been computed over the Mediterranean Sea only without considering the Atlantic box and the Marmara Sea, considered separately. The total number of stations increased of about 4% from V1.1 to V2. While the percentage of temperature data over the total amount is more or less invariant, the number of salinity measurements and TS couples incremented of a 5% approximately.

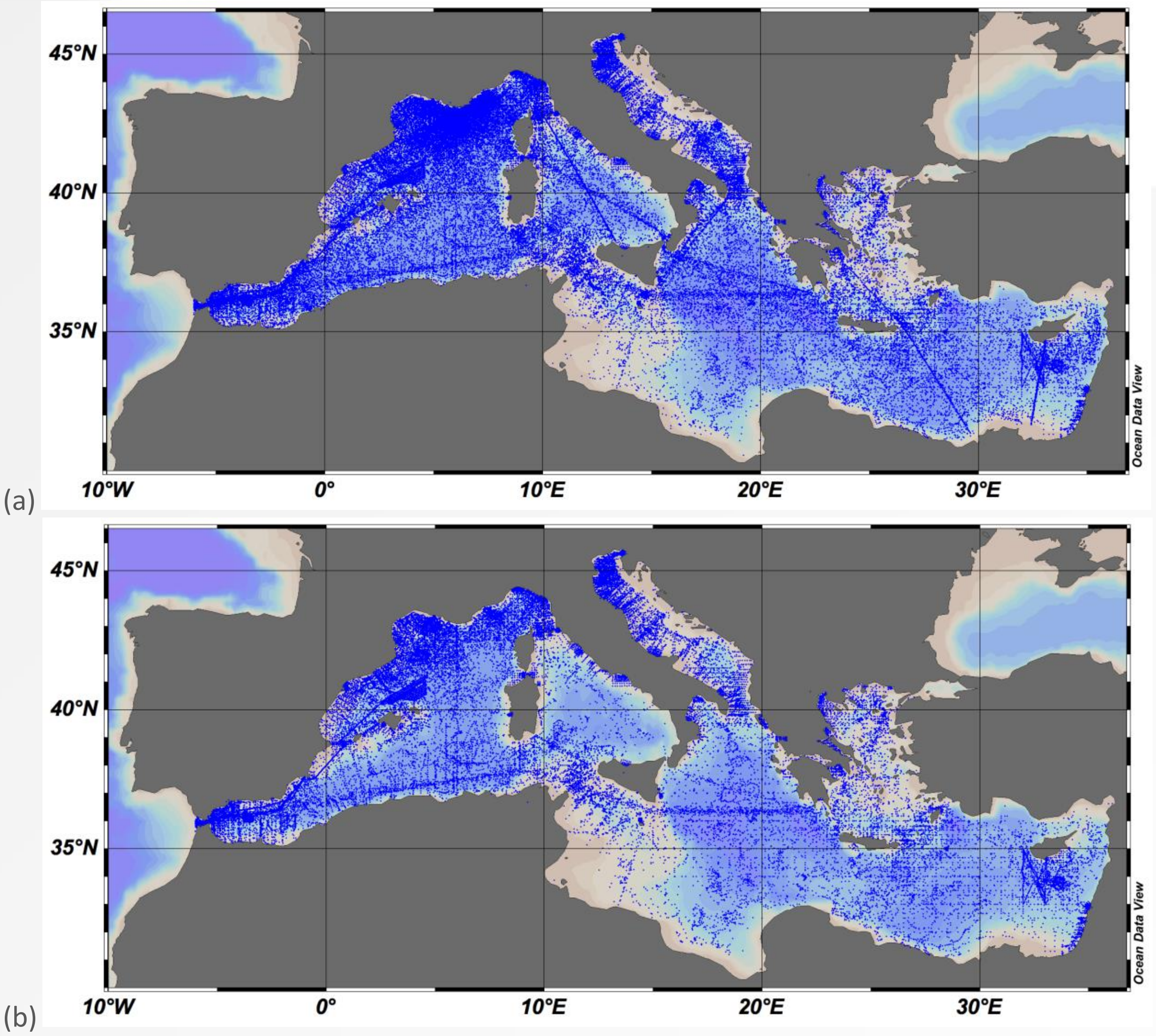


Figure 4 Station distribution map for the Mediterranean Sea 1900-2012: (a) Temperature, (b) Salinity.

par	V1.1	%	V2	%
total	169438		176201	
T	165243	97.5	174246	98.9
S	110670	65.3	123303	70.0
TS	109249	64.5	121927	69.2

Tab. 1 Number of stations for Temperature, Salinity and TS couples for the Mediterranean Sea only. The time period between 1900-2012 has been considered in order to compare V1.1 and V2 datasets.

After a general description of the historical data set a visual control of all observations allowed to assess their quality and to identify the principal anomalies.

Temperature scatter plots are presented in Figure 5. In particular panel (a) represents all T observations with QF equal to 1. Panel (b) represent all temperature observations with QF equal to 0 that did not pass through any QC analysis from data providers. These data seem good data in general and should be QCed by data providers. However they could be considered for further applications according to user needs.

Tab. 2 summarizes the repartition of observations according to their QF and from which it comes out that good temperature data correspond to the 96.8% of the total, while data still not checked are 2.9%.

Salinity scatter plot of good data (Figure 6a) confirms that there are some bad (S>60psu) data flagged as good. There are also a lot of coastal data with salinities close to zero. The user might keep this in mind for applications like climatological map computation that could present strong or unrealistic gradients along the coastline. The scatter plot of salinity data not quality checked (Figure 6b) reveal the presence of good data that could be recovered and analysed. Good data are a 95.3% while data with QC=0 represent a 4.5% (see Tab. 2).

after QC	TOT	QF0	QF1	QF2	QF>3
T	26527320	1129991	25252755	17379	96095
		4.3%	**95.2%**	**0.1%**	**0.4%**
S	19586043	1286013	18170767	50450	78616
		6.6%	**92.8%**	**0.3%**	**0.4%**
TS	19414546	855106	17975659	4034	34764
		4.4%	**92.6%**	**0**	**0.2%**

Tab. 2 Number of samples for Temperature and Salinity and their subdivision according to Quality Flags (QF) 0, 1, 2 and from 3 to 9.

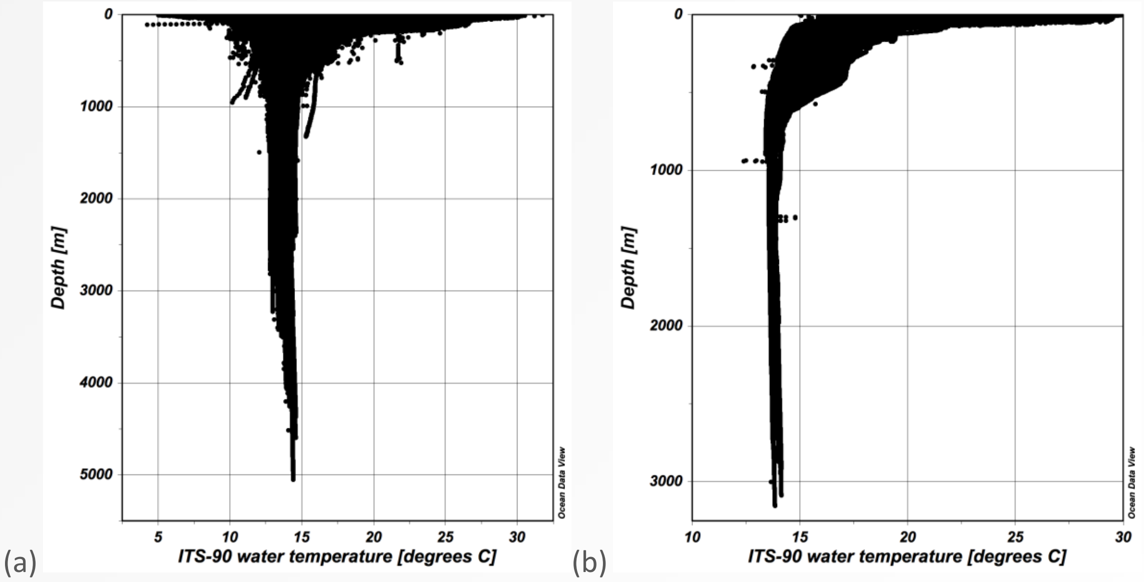


Figure 5 Temperature scatter plots: (a) Data with QF1; (b) data with QF0.

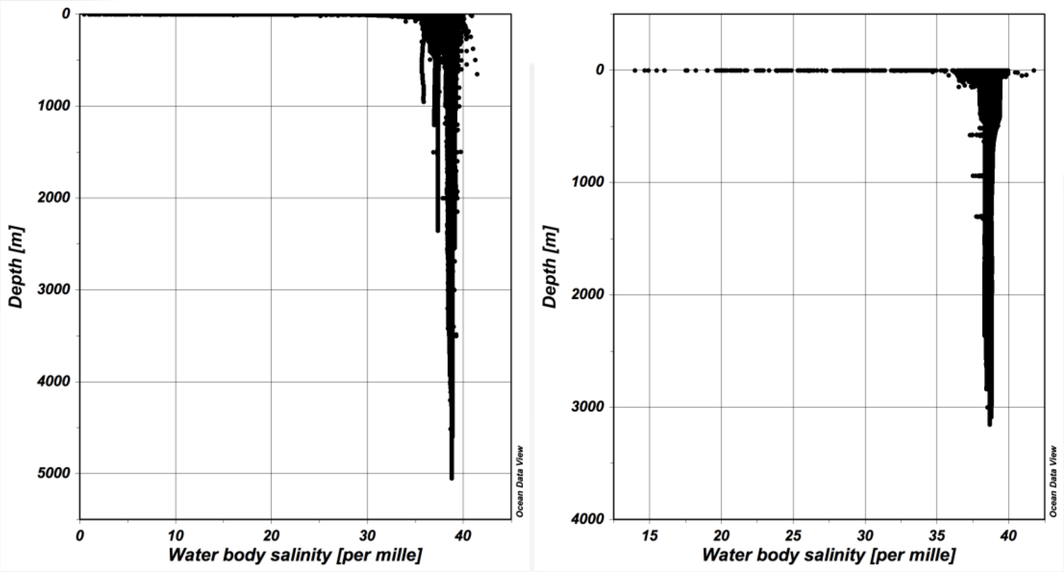


Figure 6 Salinity scatter plots: (a) data with QF1; (b) data with QF0.

Another way to visual check a data collection of temperature and salinity observations is to look at TS scatter plots. Figure 7 present data with QF=1 in (a) while (b) contains data not checked (QF=0). Plot (b) confirms that most of the data not checked could be analysed and recovered for future usage.

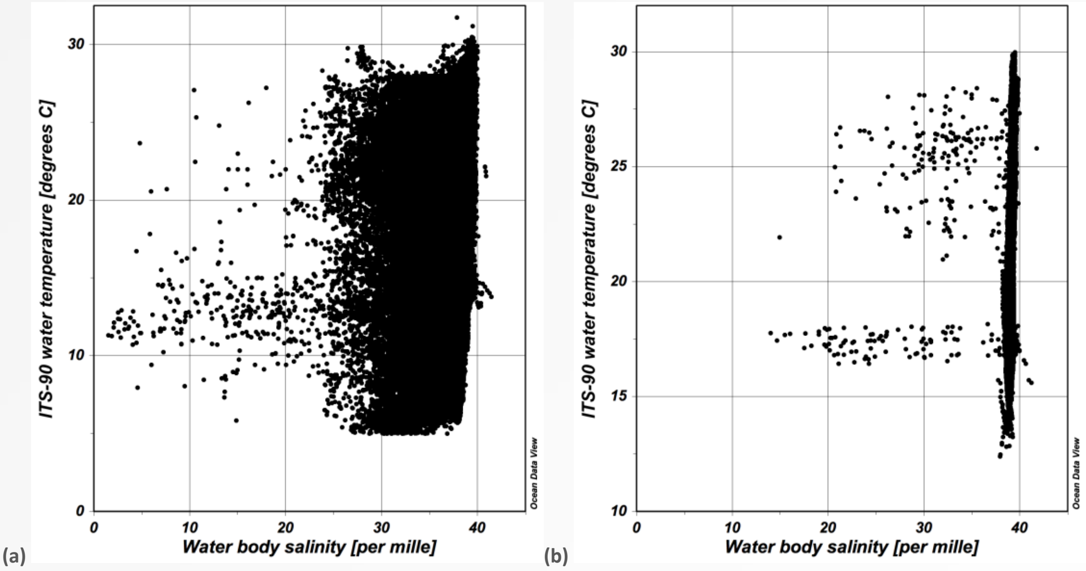


Figure 7 Mediterranean TS data collection for the time period: (a) TS diagram considering only data with QF1 (good); (b) TS diagram considering only data with QF0 (no quality control).

A quality issue concerning the Mediterranean Sea V1.1 data collection was the presence of observation with obvious wrong locations on land, especially in the southern part of Spain, Italy and the coast of Morocco. A list of those observations on land was extracted sent to data providers asking to check the right positioning. Most of these observations were corrected in the central CDI and consequently in V2 aggregated dataset. Some of the observations appearing on land are due to the low resolution land-sea mask applied by ODV.

2.1. The Atlantic Box

The Mediterranean Sea data collection contains also an Atlantic box defined between -9.25 and -6 degrees of longitude. This area has been included in order to have an overlapping zone with the North Atlantic region. The data set consists of 35398 (V1.1 - 31322) stations, mostly concentrated along the Spanish coastal areas (see Figure 8).

The temporal distribution of the Atlantic data in Figure 9 present the same peaks of Figure 3 for the Mediterranean data and this might be due to the presence of a lot of sea gliders observations sampled in 2009.

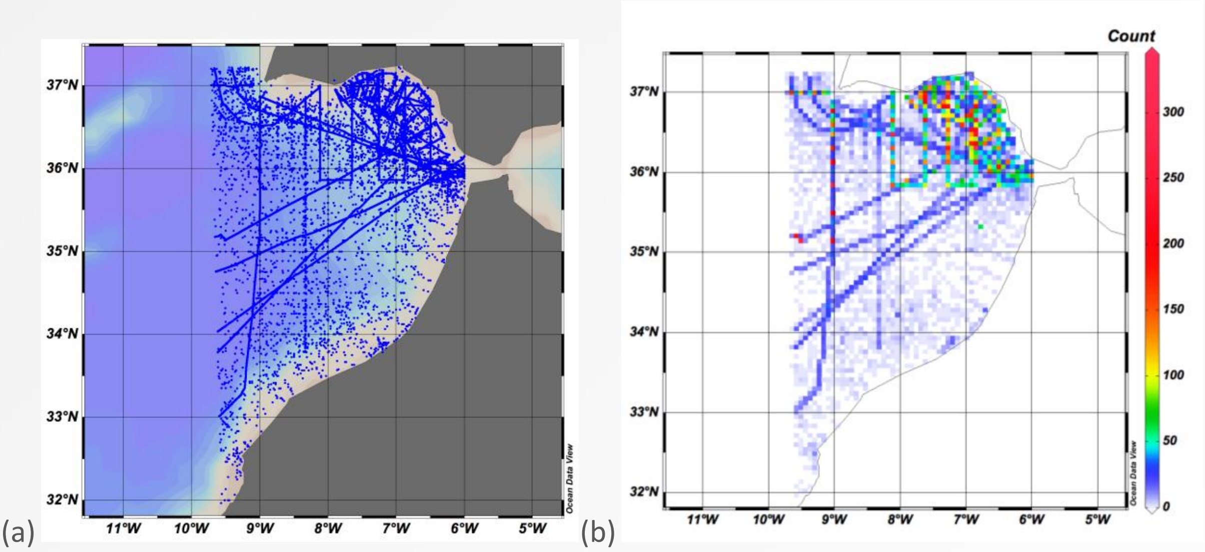


Figure 8 Data distribution map (a) and data density map for the Atlantic box included in the Mediterranean historical data collection. (31322 stations)

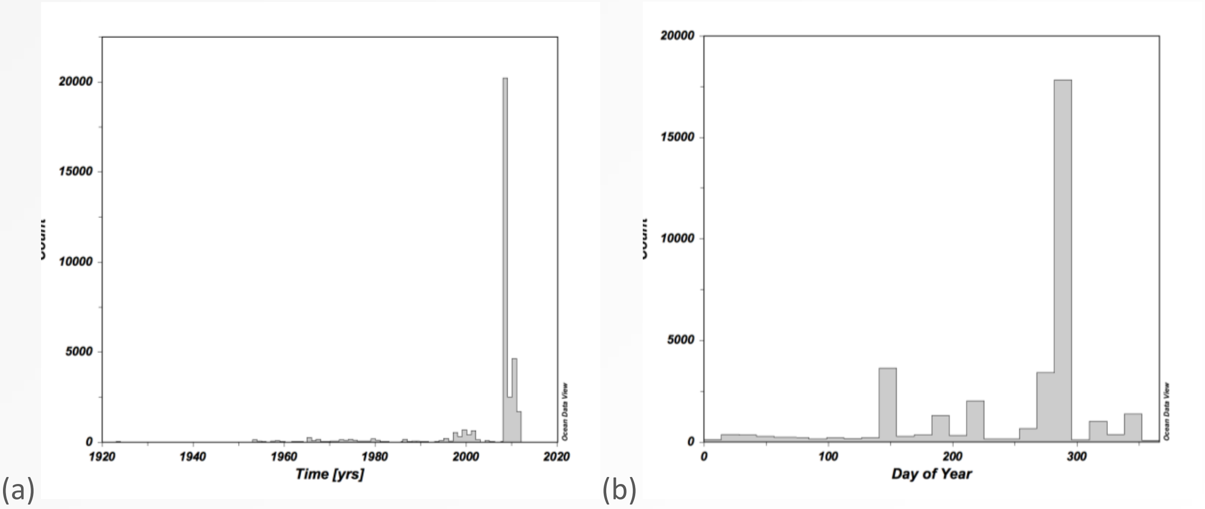


Figure 9 Annual (a) and seasonal (b) data distribution in the Atlantic box.

Figure 10 shows scatter plots of the data flagged as good (QF=1). Visual check did not reveal anomalies since some bad data where filtered out when QF=1 was selected. Salinity scatter plot displays some anomalous measurement with low (<30psu) salinity that could be due to the sampling location close to river mouths.

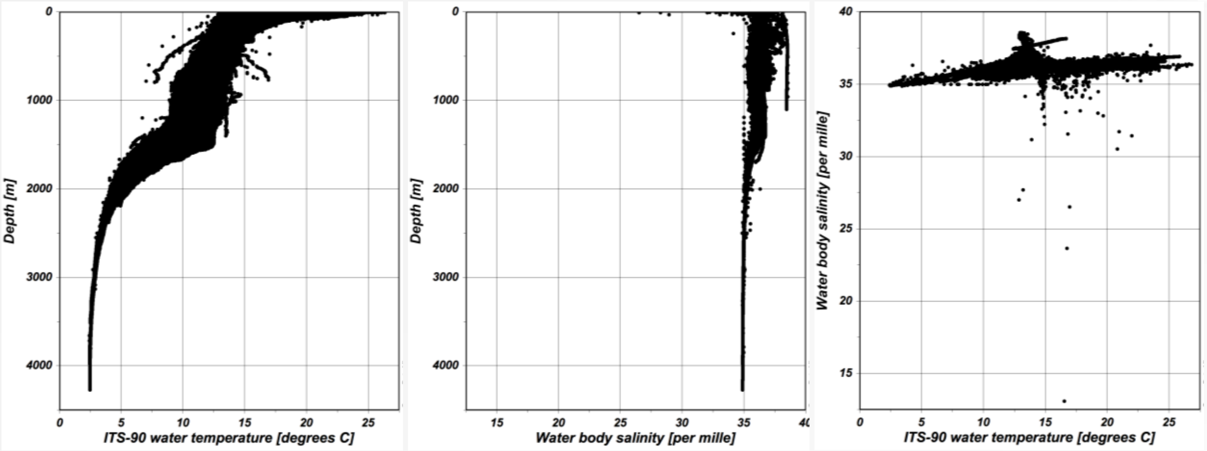


Figure 10 Scatter plots of data with QF=1 in the Atlantic box: (a) temperature; (b) salinity; (c) TS diagram.

Tab. 3 presents data quality flags statistics from which it appears that almost the entire data set has a good quality.

Atlantic Box	TOT	QF0	QF1	QF2	QF>3
T	2676396	10180	2664056	2	2160
S	689606	10317	663104	0	999
TS	689292	10180	677909	0	139

Tab. 3 Number of data points for Temperature and Salinity data within the Atlantic box in Figure 8 and their subdivision according to Quality Flags (QF) 0, 1, 2 and from 3 to 9.

2.2. Marmara Sea

The Marmara Sea has been defined between 26 and 30 longitude degrees and 39.5 and 41.5 of latitude. The data set consists of 190 (156 in V1.1) stations (Figure 11). Observations are few and very sparse both in space and time, as shown in Figure 12. This makes unfeasible the computation of reliable statistical products, thus we will encourage data providers to release more observations in this area.

Figure 13 presents temperature and salinity scatter plot for the entire data set, considering observations with QF=1. Very few points were filtered out from QC as it could be noticed in Tab. 4.

Marmara	TOT	QF0	QF1	QF2	QF>3
T	17267	0	17247	0	20
S	16360	0	16305	0	55
TS	16359	0	16297	0	13

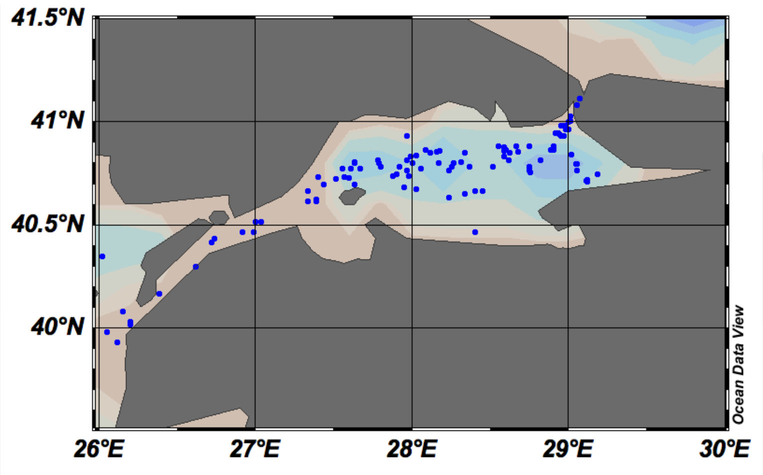


Figure 11 Data distribution map and for the Marmara Sea.

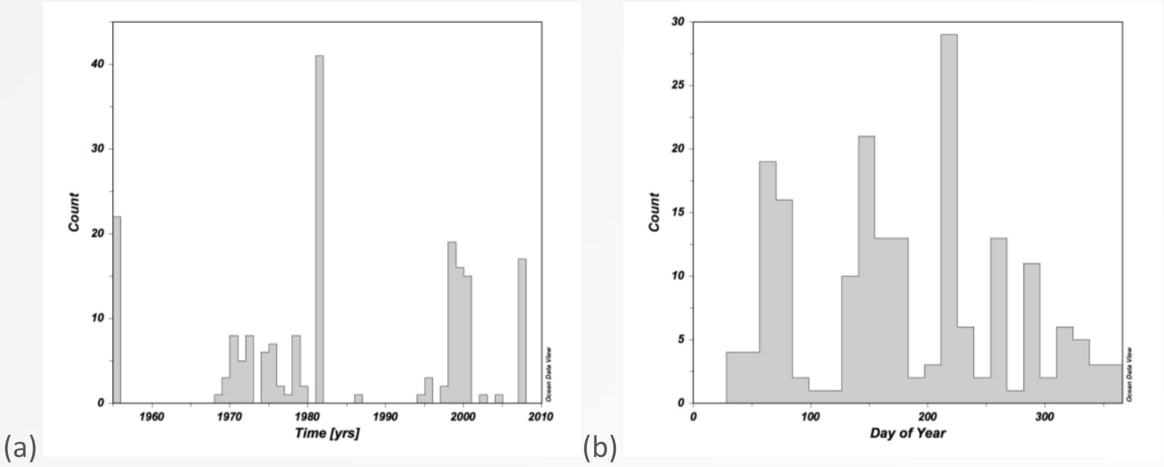


Figure 12 Annual (a) and seasonal (b) data distribution in the Marmara Sea.

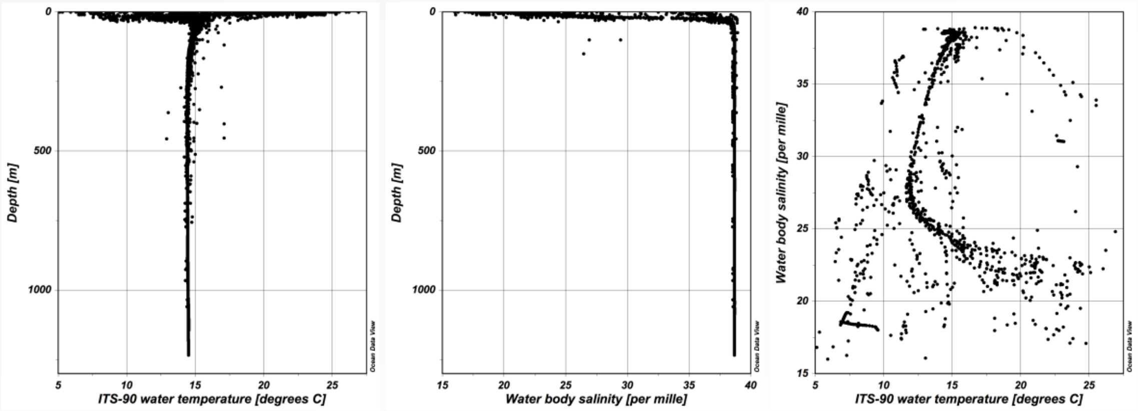


Figure 13 Scatter plots for the Marmara Sea data with QF=1: (left) temperature versus depth; (middle) salinity versus depth; (right) TS diagram.

Marmara	TOT	QF0	QF1	QF2	QF>3
T	17267	0	17247	0	20
S	16360	0	16305	0	55
TS	16359	0	16297	0	13

Tab. 4 Number of data points for Temperature and Salinity data within the Marmara Sea in and their subdivision according to Quality Flags (QF) 0, 1, 2 and from 3 to 9.

2.3. Conclusions about the Mediterranean Sea V2 collection

The QC analysis of V2 highlighted that most of the data anomalies observed in the V1.1 were checked and corrected by the data providers. New data anomalies on 126 stations have been found and the corresponding QF have been changed to 4 (bad data) or 3 (probably bad data). The main anomalies encountered were spikes and outliers. Only one systematic error of salinity values in 2009 around Cyprus Island was found. The list of data anomalies was ordered by EDMO code and send to the corresponding data centre asking for checking and correcting the data.

We might conclude that the implemented QC strategy was very effective in improving the overall content of SDN database in the Mediterranean Sea. Data centres were active and collaborative in responding to our QC results. The quality of V2 Mediterranean Sea aggregated dataset is very high and thus it can be used for a lot of applications.

Restricted data collection for the Mediterranean Sea, retrieved during V1.1 aggregation exercise, represented 17% of the total number of data (28690 stations), which is a high percentage of data. We estimate that about 25% of data were not retrieved from the infrastructure because CDI partners did not approve the request. Figure 14 show the data distribution of the restricted data harvested during V1.1 exercise. Most of the data were located in the Adriatic Sea, the Sicily Channel and the Tunisian coast and are crucial for computing good quality statistical products.

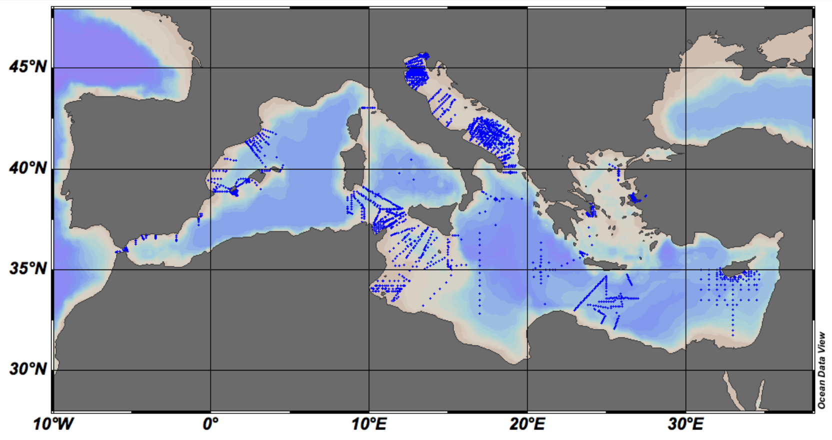


Figure 14 The Mediterranean Sea restricted data collection harvested during V1.1: data distribution map.

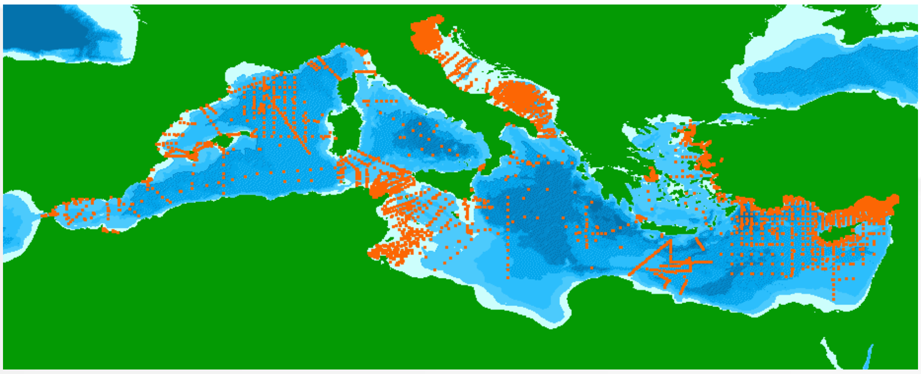


Figure 15 Data distribution map of the restricted temperature and salinity observations within SeaDataNet database at present time.

In date July the 8th 2015 a data search has been performed on the CDI in order to quantify the percentage of restricted data in the Mediterranean Sea and in the Marmara Sea. Tab. 5 summarizes the number of CDIs freely available and those that fall under any kind of restriction. In the Mediterranean Sea there are still 18.2% of temperature and salinity CDIs restricted. Their spatial distribution is displayed in Figure 15, which shows the largest number of observations along the coast of Tunisia, Turkey and the Northern and Southern Adriatic. The availability of these observations is crucial for climatology computation. Comparing Figure 14 and Figure 15 we may notice that many data have not been harvested because partners did not approve the request, even for internal purposes of SDN project.

The situation in the Marmara Sea is noteworthy since 96% of the data are restricted, meaning that it is not possible to compute any climatology. None of those data have been delivered to the Regional Coordinator during the V1.1 aggregation phase, disregarding the indications included in the DoW.

We will encourage data providers to adopt a data policy of free and open access in provision of data to users, in line with international agreements (e.g. WMO, IOC, ICSU, GEO/GEOSS).

Tab. 5 Number of CDIs in the Mediterranean region for water column temperature and salinity measurements.

WEB SEARCH	**Med only**		**Marmara**	
unrestricted+SDN license	163330		281	
restricted	36403	**18.2%**	6709	**96.0%**
TOT	199733		6990	

3. Black Sea

3.1. General Characteristics of the Black Sea aggregated data set

The Black Sea aggregated dataset V2 includes data from the Black Sea and Sea of Azov for period 1868 – 2014. There is decrease of data in V2 compared to V1.1. The reason for decrease is partial elimination of duplicates, which were identified in V1, and absence of CDIs having "SR" – "academic" data access restriction. On the other hand V2 dataset includes a number of new CDIs.

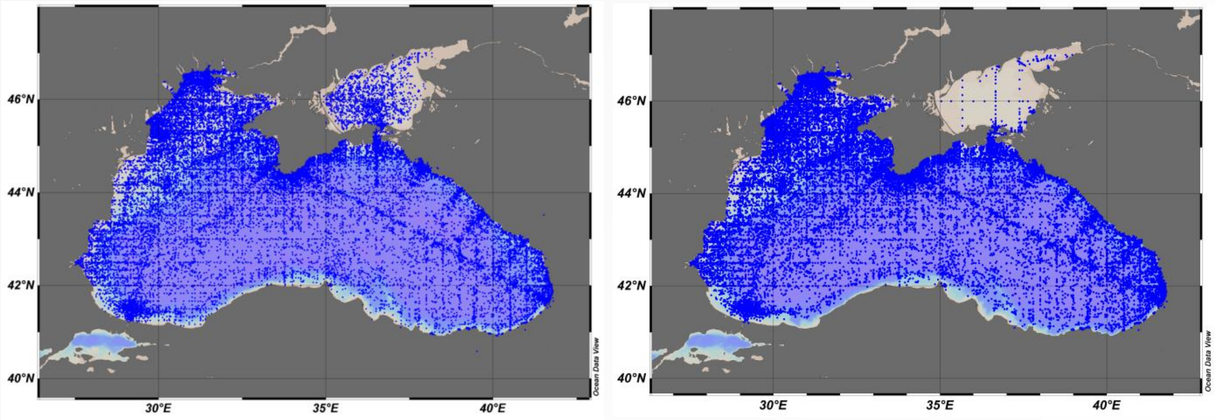
Comparison of two versions is presented in Tab. 6 below. It contains statistics for initial dataset and for final QC-ed dataset. The final V2 dataset was cleaned from duplicates and empty stations, and it also does not contain data of the Southern Scientific Research Institute of Marine Fisheries and Oceanography (Kerch, Crimea, EDMO 688), which was excluded from the SeaDataNet infrastructure earlier in 2015.

Tab. 6 Comparison of V1.1 and V2 aggregated datasets in the Black Sea.

	Initial collection			Final QC-ed collection		
	V1.1	**V2**	**±%**	**V1.1**	**V2**	**±%**
Cruises	2353	2263	-4%	2069	1723	-16%
Stations (CDIs)	125565	122726	-2%	101369	96487	-5%
Samples	3815118	3036865	-20%	3412707	2696215	-21%
Period	1868 - 2013			1868 – 2014		

Excluding of EDMO 688 mostly affected data coverage in the Sea of Azov (Figure 16). Also it can be noticed rather poor data coverage in the southern part of the Black Sea along Turkish coast because all Turkish "SR" CTD data are missing in V2 dataset.

Figure 16 Spatial data distribution in the Black Sea: left – V1.1, right – V2



Upon sorting out the problem with duplicate, empty and split stations (i.e. when Temperature and Salinity on the same stations are submitted as different CDIs) the remaining number of stations (CDIs) in V2 aggregated dataset is 99,171. The 2448[1] (~2.5%) stations appear to be on

[1] Final number to be updated

land or in river mouth as per GEBCO-2014-2min topography. These are mainly stations of the coastal monitoring in ports, which coordinates need to be slightly corrected.

The statistics of the quality flagging in initial V2 aggregated dataset is presented in Tab. 7. The data with QF=1 “good” represent 77% of the total however it does not mean that all they are good. Figure 17 demonstrates numerous outliers on the Temperature and Salinity scatter plots, which are flagged by providers as good data. About 20% of data remains non QC-ed. The QF statistics and plots are similar to those for V1.1 initial dataset, which means that most of data providers did not apply recommended quality flagging based on results of the V1.1 dataset analysis.

Tab. 7 Quality flagging statistics of initial dataset.

QF	0	1	2	3	4	Total
	not checked	good	probably good	probably bad	bad	
Depth	586952	2349163	1368	99377		3036860
	19.33%	77.35%	0.05%	3.27%		100.00%
Temperature	582370	2264869	2	73972	3866	2925079
	19.91%	77.43%	0.00%	2.53%	0.13%	100.00%
Salinity	575947	2222589	3650	46522	2980	2851689
	20.20%	77.94%	0.13%	1.63%	0.10%	100.00%

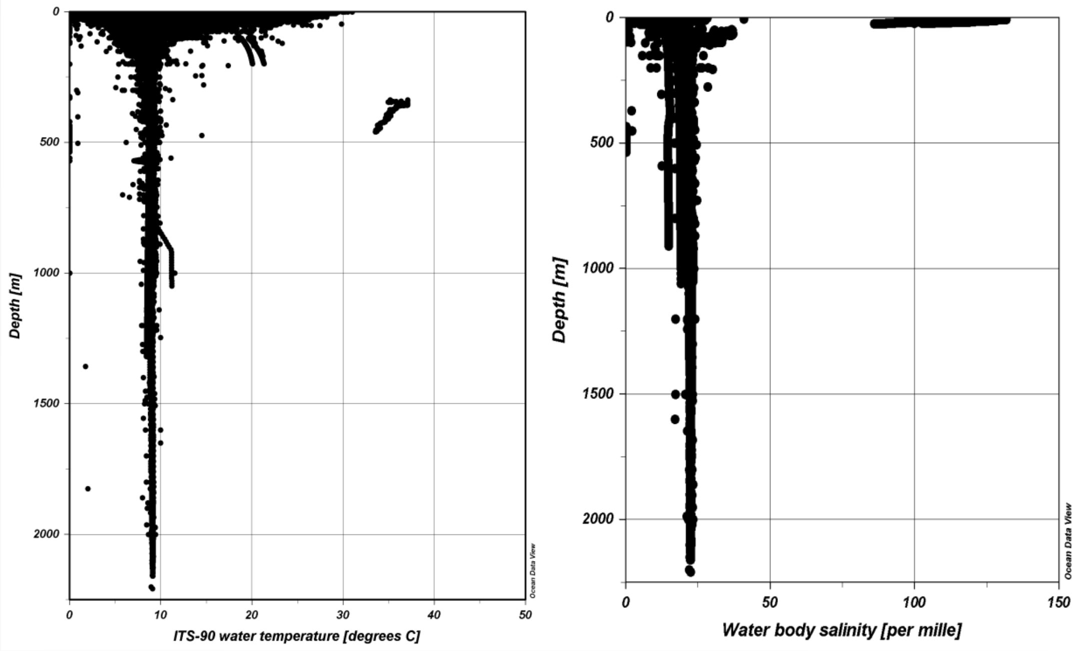


Figure 17 Scatter plots of Temperature (left) and Salinity (right) data in the Black Sea flagged with QF=1 “good” by data providers.

The presented below statistics are obtained for the V2 final dataset.

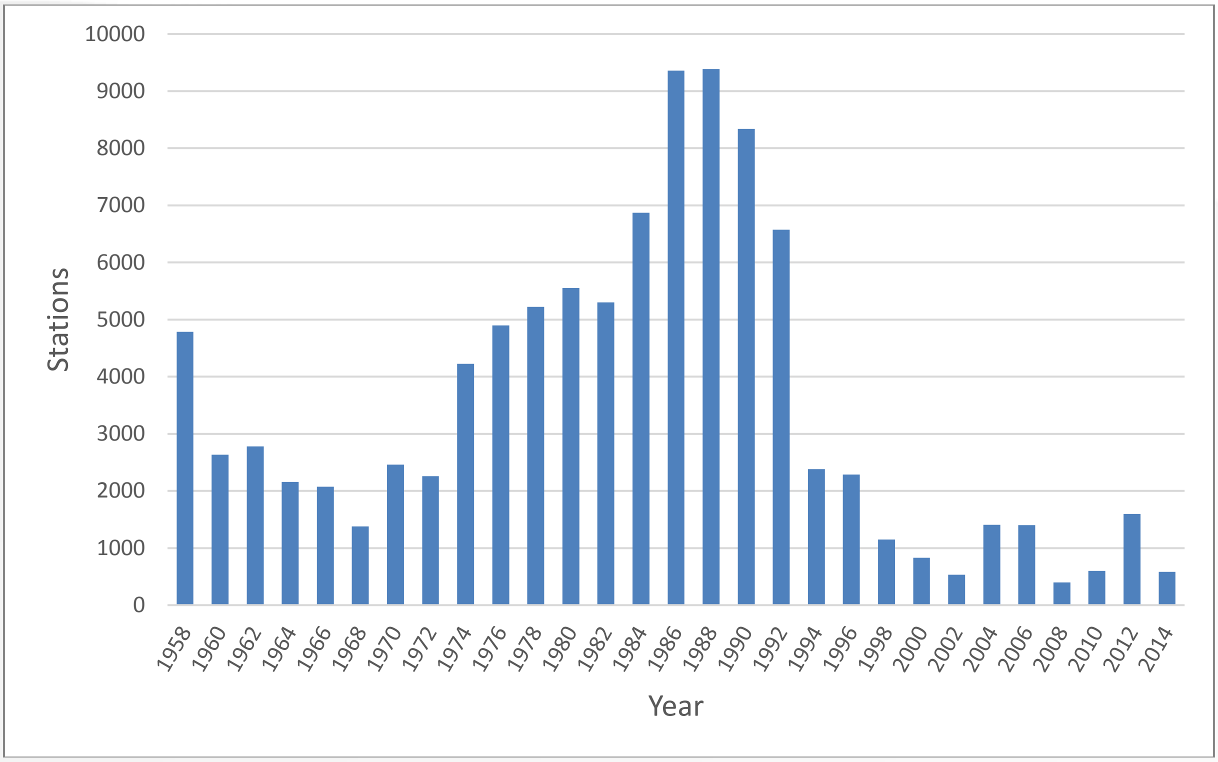


Figure 18 Temporal distribution of stations (left bar represents all stations before 1958) in the Black Sea.

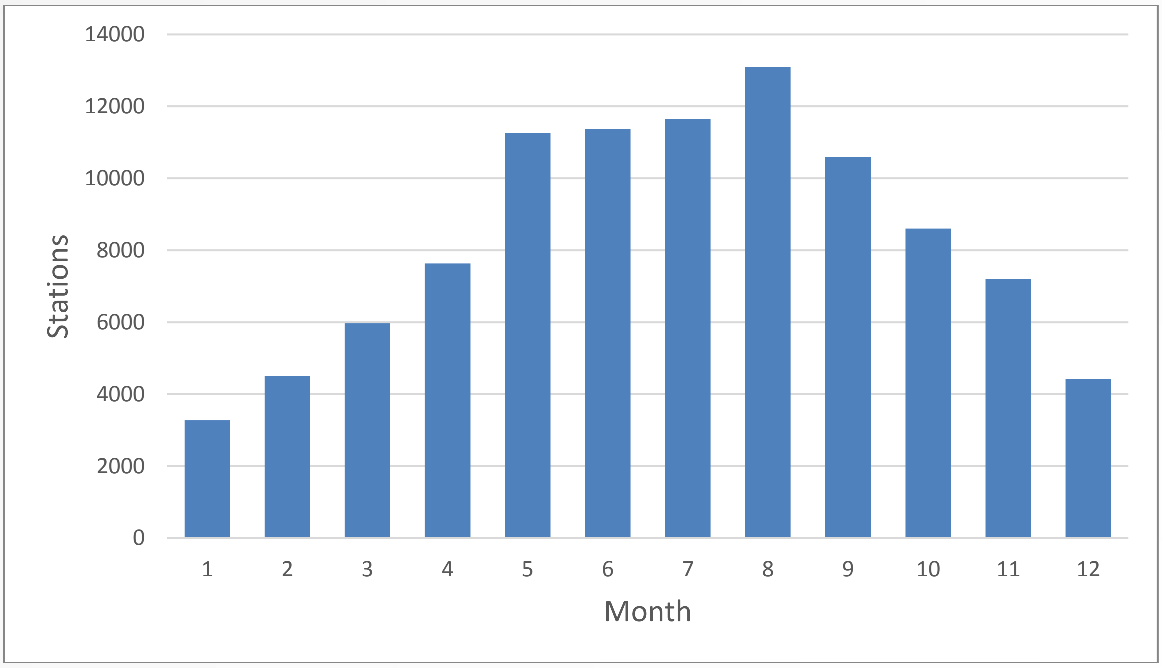


Figure 19 Monthly distribution of stations in the Black Sea.

The temporal distribution of stations (Figure 18 and Figure 19) is similar to the one from V1.1 dataset: the most intensive oceanographic observations were performed in Black Sea in period 1970 – 1995; the amount of observations from recent years (2013-2014) is very small;

most of observations were performed during summer months while in winter period intensity of observations is almost 3 times less than in summer.

3.2. Data Quality assessment procedure for the Black Sea

The results of the quality control performed for V1.1 dataset were applied to V2 dataset. After this procedure approximately 9% of (new) data remained not QC-ed. The same QC procedures as in V1.1 were applied, including range checks (region based), density inversion check, and expert control in ODV with analysis of scatter-plots of parameters , T-S diagrams and Brunt-Viasala frequency plots. Additionally the check of Depth values against GEBCO 30”x30” bathymetry was performed considering that the initial dataset contains rather large amount (99377) of Depth values flagged as “probably bad”. 57600 values appeared to be less than GEBCO interpolated values at the same location, and the respective QF was changed to “probably good”. On the other hand, the maximum depth appears to exceed 1.2 of GEBCO depth on 1091 stations. The respective Depth was flagged as “probably bad”. These changes need to be clarified with data providers because on slope the interpolated GEBCO depth may differ significantly from the real one.

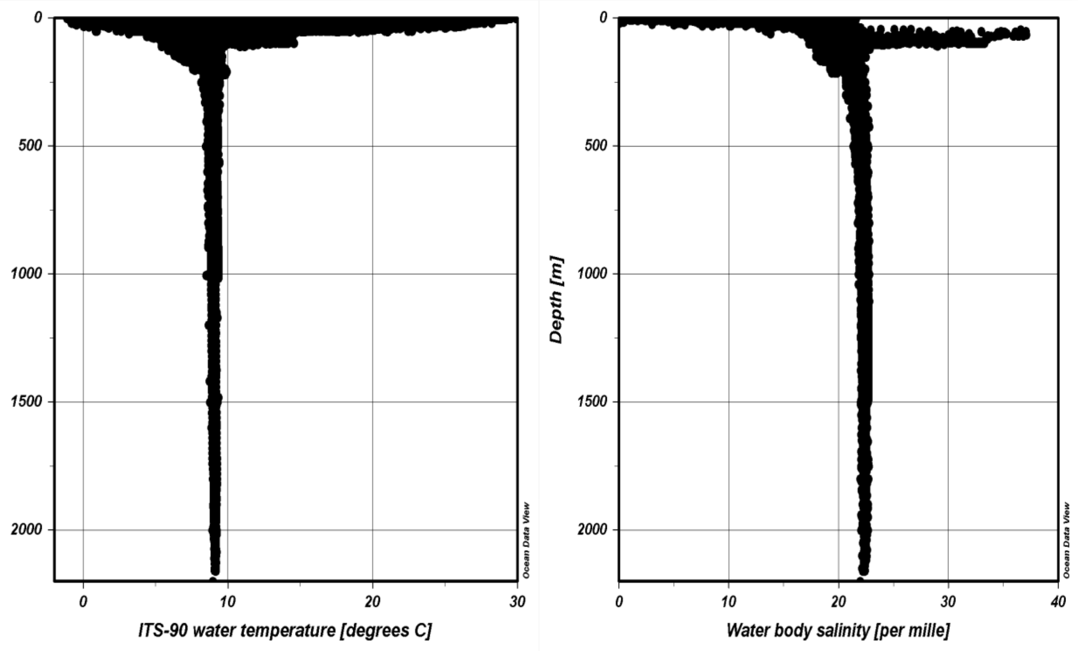


Figure 20 Scatter plots of Temperature (left) and Salinity (right) after QC.

The statistics of the quality flagging for final dataset are presented in Tab. 8 below. Good data represent approximately 96% in the final dataset.

Tab. 8 Quality flagging statistics of the final Black Sea dataset.

QF	0	1	2	3	4	Total

	not checked	good	probably good	probably bad	bad	
Depth	0	2602832	58396	34852	135	2696215
	0.00%	96.5%	2.2%	1.3%	0.0%	100.0%
Temperature	0	2569113	43207	76568	5126	2694014
	0.00%	95.4%	1.6%	2.8%	0.2%	100.0%
Salinity	0	2528715	37416	57186	5426	2628743
	0.00%	96.2%	1.4%	2.2%	0.2%	100.0%

3.3. Conclusions

Amount of data in V2 aggregated dataset decreased compared to V1.1. It happened because the CDIs with “SR” data access flag were not included in the V2 dataset.

Unfortunately most of providers did not apply recommended quality flagging therefore the initial V2 dataset contained significant number of outliers flagged as good.

The problem of duplicates remains to be one of the major issues for the Black Sea aggregated dataset. Approximately 6% of data are duplicates from different providers, i.e. when the same data are submitted by different data centres. Because these data have different cruise labels, different station names and small differences in time and coordinates, identification of such duplicates is not an easy task. Elimination of such duplicates from the SeaDataNet infrastructure is a complicated administrative task to be included in the agenda of further SeaDataNet activities.

The V2 aggregated dataset potentially can be used for computing of the Black Sea Climatologies however to obtain more reliable result it can be complemented with ~1500 CTD profiles having “academic” data access restriction (+1.5%) and ~8,000 restricted CDIs (+9%) - as per the SeaDataNet CDI website.

4. Arctic Sea

The historical data collection of the Arctic Sea V2 covers the time slot 1990 to 2013 and contains temperature and salinity observations in the area 40W to 65 E and 62 N to 83N. In V2, the area was extended a bit further south compared to V1.1, to better fit the other boundaries used. Figure 21 shows the data distribution map.

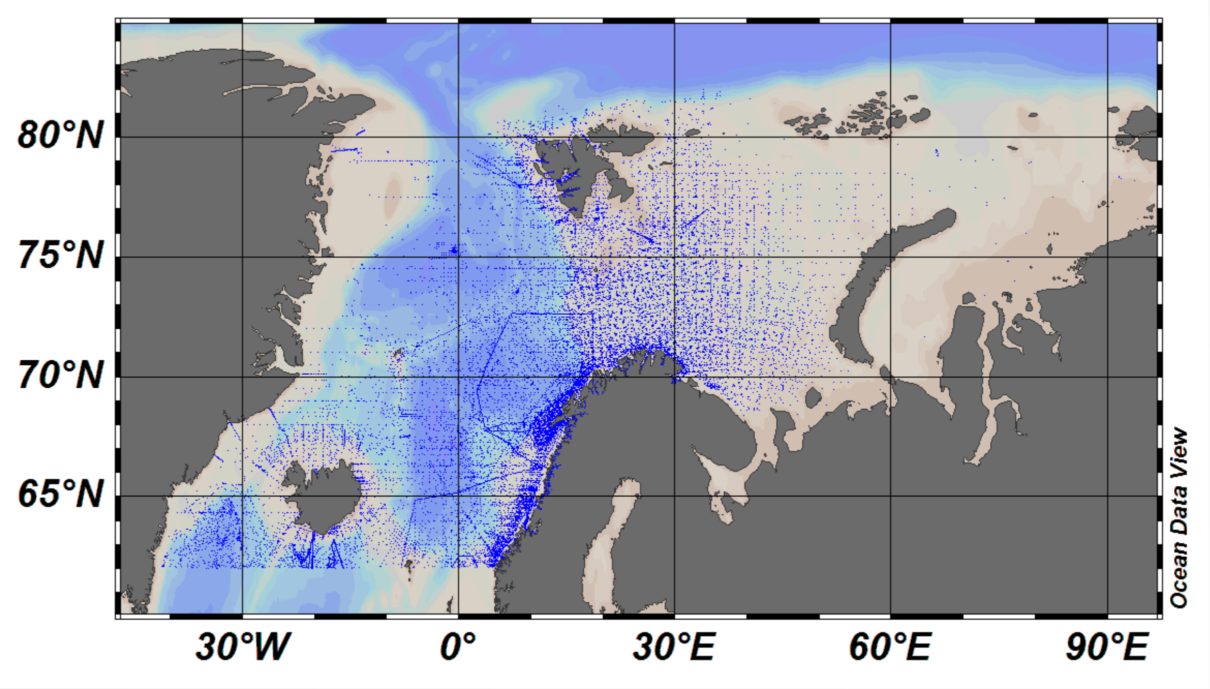


Figure 21 data collection in the Barents Sea in the time period 1990-2013.

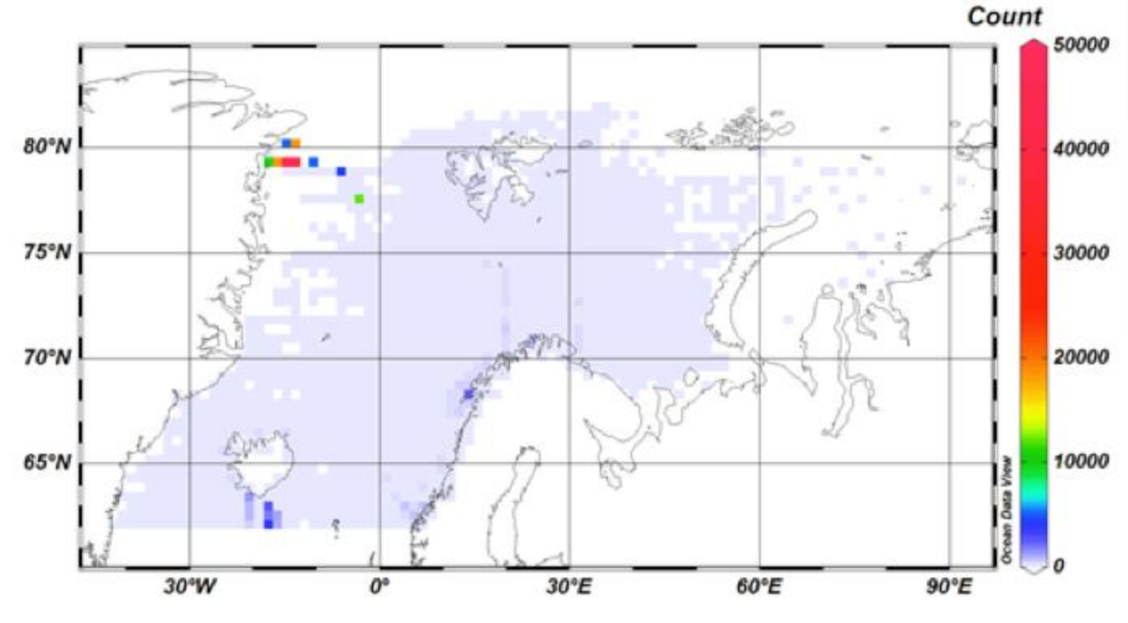


Figure 22 Data density map in the Barents Sea.

The data coverage is good close to the coast (Figure 21 and Figure 22), but in Open Ocean only few data are available because these areas being visited seldom. These areas have better coverage in later years due the introduction of ARGO floats as sampling platform. The east part of the Barents Sea has spares data due to restrictions on Russian data.

The V1.1 dataset contained data from ferry boxes operated along the Norwegian coast. It contained 247.302 data points that are now missing in the V2 collection. The two ferry boxes are mounted on the "Hurtigruten" coastal steamers, MS Lofoten and MS Vesteralen. The time period covered was 1998 to 2009. The reason why they are missing is unknown. The data is available in the SeaDataNet portal for download.

New data (1782 profiles) is introduced in the V2 dataset in 2013 that contains some suspicious data. It is data from Norwegian research cruises operated by hired vessels. The hired vessels are fishing vessels that use lower quality instrumentation (171 profiles).

In the V2 dataset approximate 10 UK glider operations are added. This is data from autumn 2004. In the dataset it represents 153.074 data points out of 266.291 in total. These data is shown in Figure 23a, the annual data distribution, as the dominating pillars. Every measuring point is represented as one profile.

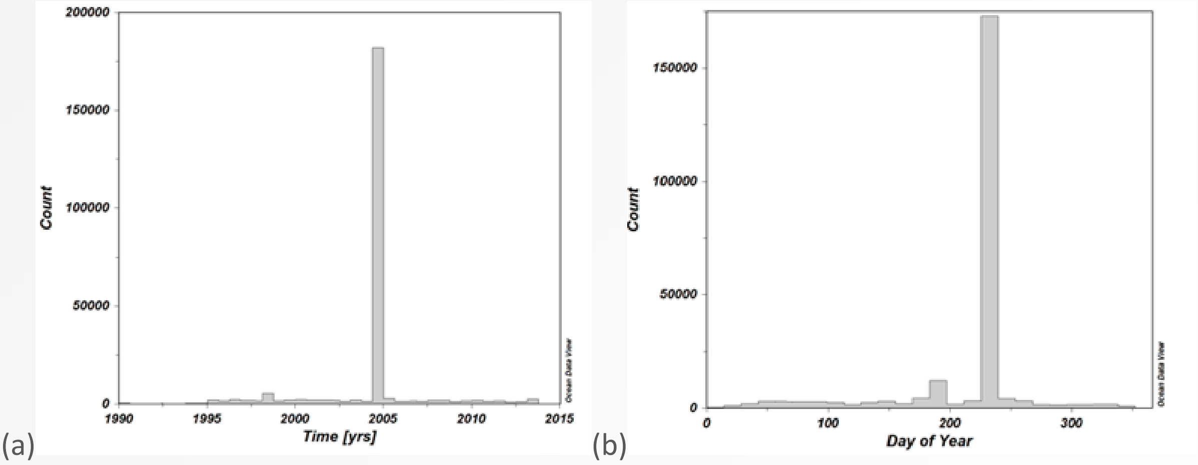


Figure 23 (a) Annual distribution and (b) seasonal distribution of the time period 1990 – 2013 in the Barents Sea

The seasonal distribution Figure 23b shows the same situation in autumn, which implies it to be one dataset with lots of measurements in a short time period in one year. The rest of the data shows the summer period to be the most data intensive period.

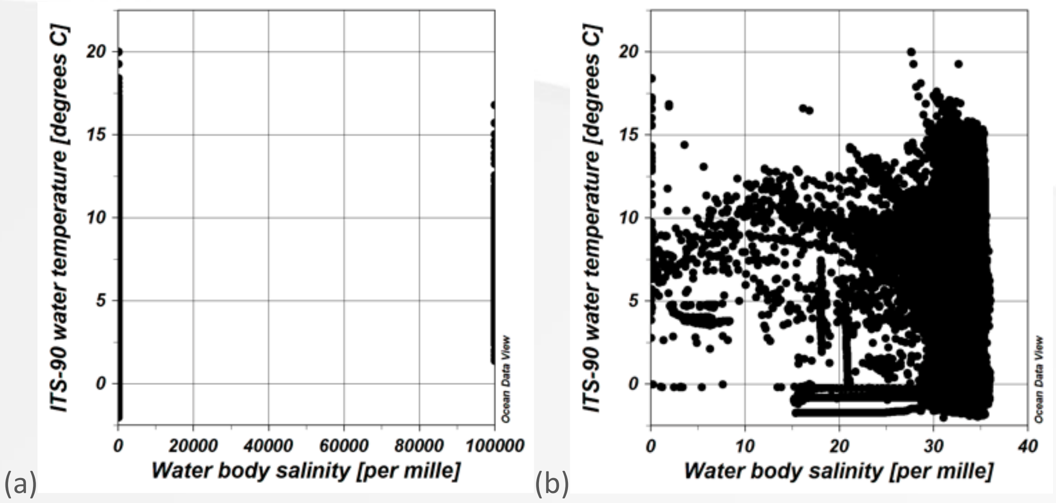


Figure 24 (A) TS diagram before QC (B) TS diagram after QC for entire data collection

TS diagram in Figure 24a before QC analysis shows there are new introduced data that are far out, compared to most data. In the diagram after QC, these data were removed and TS diagram looks better.

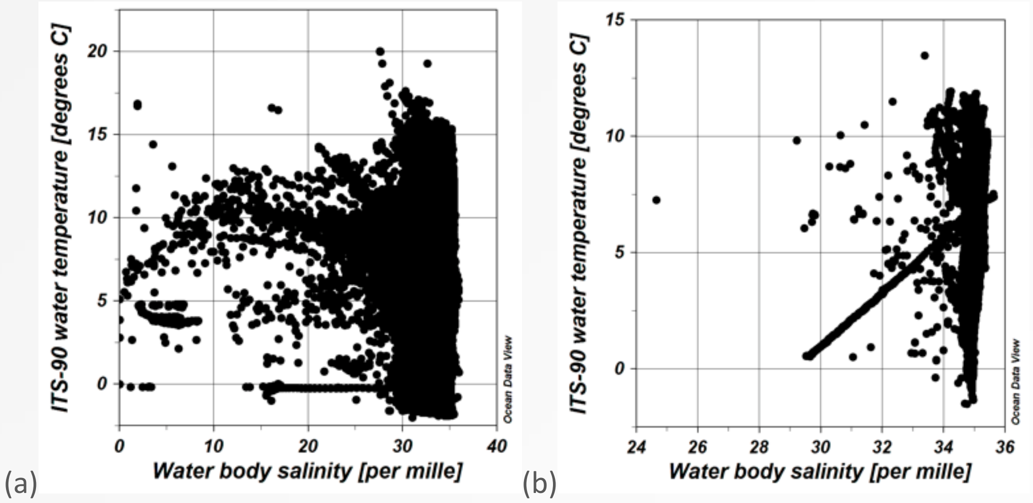


Figure 25 (a) Scatter plot with QF=1 or 2 (b) scatter plot with QF=0

Figure 25b shows that there are still data in the data collection with QF=0 that means no quality control performed. Most of these data are probably good, but have the wrong quality stamp. These data profiles needs to be identified in the collection and marked as good or probably good. The data originator will have to be contacted to avoid these data being wrongly flagged in future data collections.

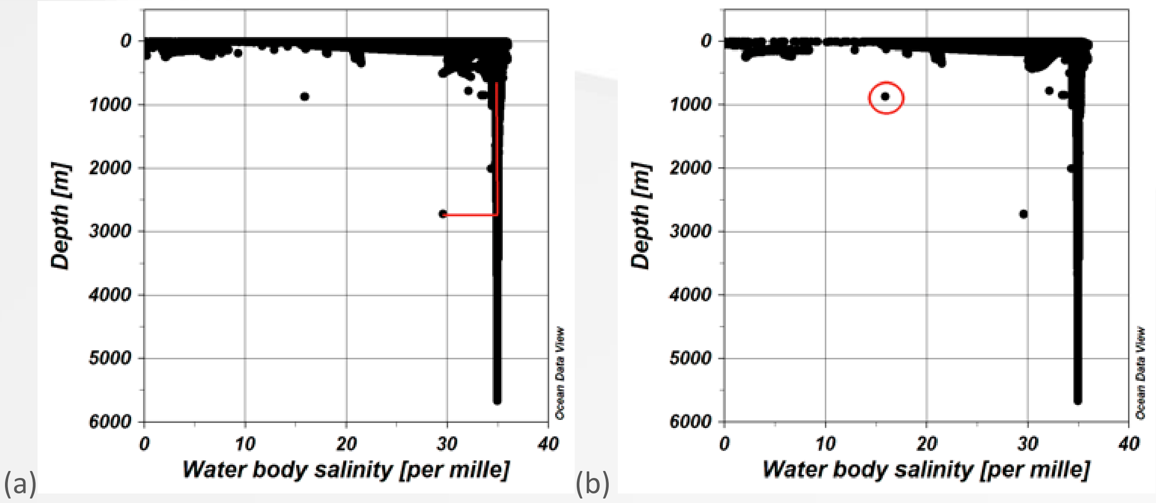


Figure 26 Visual QC (a) identifying spikes (b) identifying bad profiles

Figure 26 a) and b) shows one spike that is far out compared to the rest of the data in the same depth. The profile needs to be marked as bad or erroneous to be excluded from analysis.

QF	0	1	2	3	4	TOT	Out of Range	Range
Depth	13754	18935040	71	89	59	18949013	0	-1,6000m
%	**0.02**	**99.9**						
Temp	70721	18756614	0	7639	1104	18836078	260	-2.5,22°C
%	**0.375**	**99.57**		**0.04**	**0.0058**			
Salinity	117951	18441576	351	64211	12251	18636340	5210	0-36psu
%	**0.632**	**98.95**	**0.0018**	**0.344**	**0.065**			

Tab. 9 Table of QC flag status for temperature and salinity

Tab. 9 summarizes the repartition of observations according to their QF and from which it comes out that good temperature data correspond to the 99.57% of the total, while data still not checked are 0.4%. Good Salinity corresponds to 98.95%, while still 0.6% is not checked. Most of the data flagged as QF=0 look as good and should be reconsidered flagged as QF=1 by the originator.

4.1. Conclusions for the V2 Arctic Sea data collection

The QC analysis of V2 highlighted that most of the data anomalies observed in the V1.1 was checked and corrected by the data providers. New data anomalies have been found. The main anomalies encountered were spikes and outliers. The list of data anomalies was marked by EDMO code and send to the corresponding data centre asking for checking and correcting the data. We conclude that the implemented QC strategy was very effective in improving the overall content of SDN database in the Arctic Sea.

Some salinity values above 35.35 were identified in V2 dataset and these profiles needs to be looked further into to check if the high salinities are real or spikes are occurring in depths where sudden temperature changes creates inconsistency between the sensors behaviour.

The V2 dataset is missing Norwegian ferry box data from the coastal steamer that was present in the V1.1 dataset. A new glider dataset from UK has been introduced in V2 dataset as profiles, but the data are not organised as profiles, rather single points and therefore corrupting the counting of profiles. An on-going QC process is now contacting the respective data owners to improve the overall quality in future data collections.

Data centres were active and collaborative in responding to our QC results. The quality of V2 Arctic Sea aggregated dataset is very high and can be used for many.

5. Baltic Sea

The Baltic Sea historical data set v2 contains about 219000 CDIs with around 12500000 values for both salinity and temperature. Most of the data are from profiles, dots in Figure 27(a), but there are also some data that are from trajectories (ferry box system), solid lines in Figure 27(a).

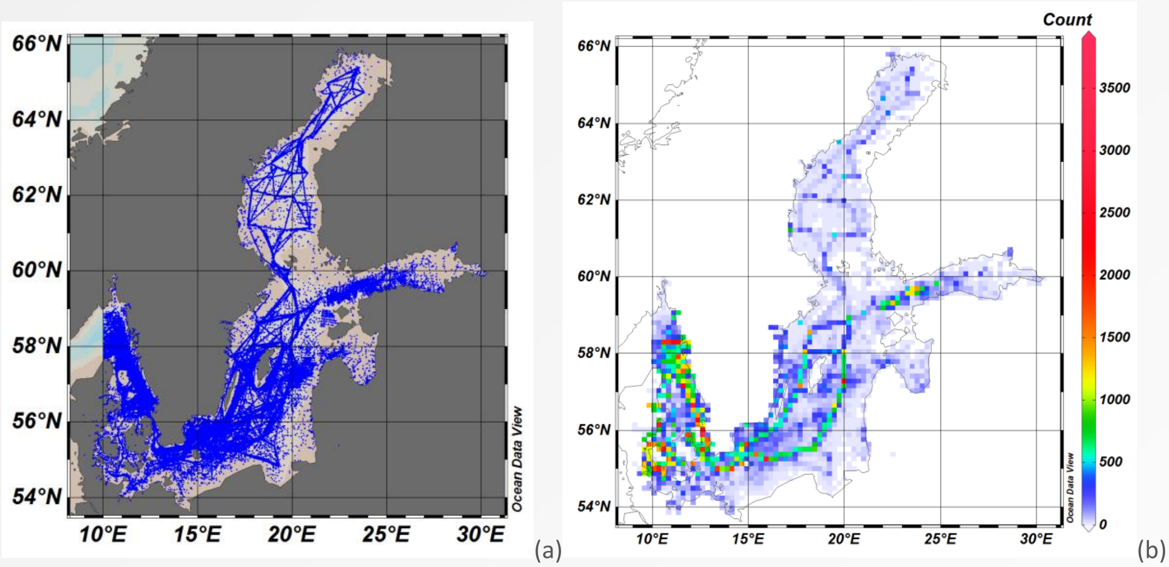


Figure 27 TS data collection for the Baltic Sea in the time period 1990-2013: (a) Data distribution map; (b) Data density map.

Data distribution map (Figure 27a) show a good geographical spread but with a few coastal areas with no or almost no data. Data density map (Figure 27b) is heavily dominated by trajectory data (ferry box system), due to the large number of data points in this type of data. Data density map does not show the vertical data density, this means the trajectory data can create a false illusion of lots of data, when in reality the trajectory only contains data from one depth and can have a small time coverage as well.

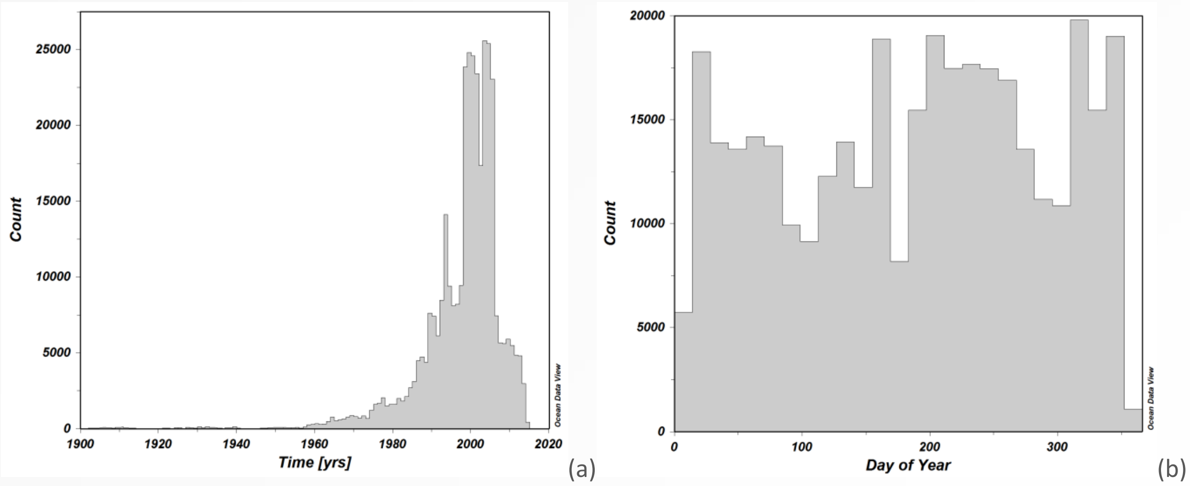


Figure 28 (a) Annual data distribution and (b) seasonal data distribution for the time period 1990-2013 in The Baltic Sea.

Annual data distribution (Figure 28a) shows that there are few measurements up until about 1960. In the 1980s and 1990s the data points increases and stays on a relatively high level up until the latest years where there is a natural time lag between sampling and until data becomes available in the SeaDataNet system. The spikes that can be seen in the late 1990s and early 2000s are from ferry box trajectory data, containing large amounts of data points. Seasonal data distribution (Figure 28b) shows an even spread during the year.

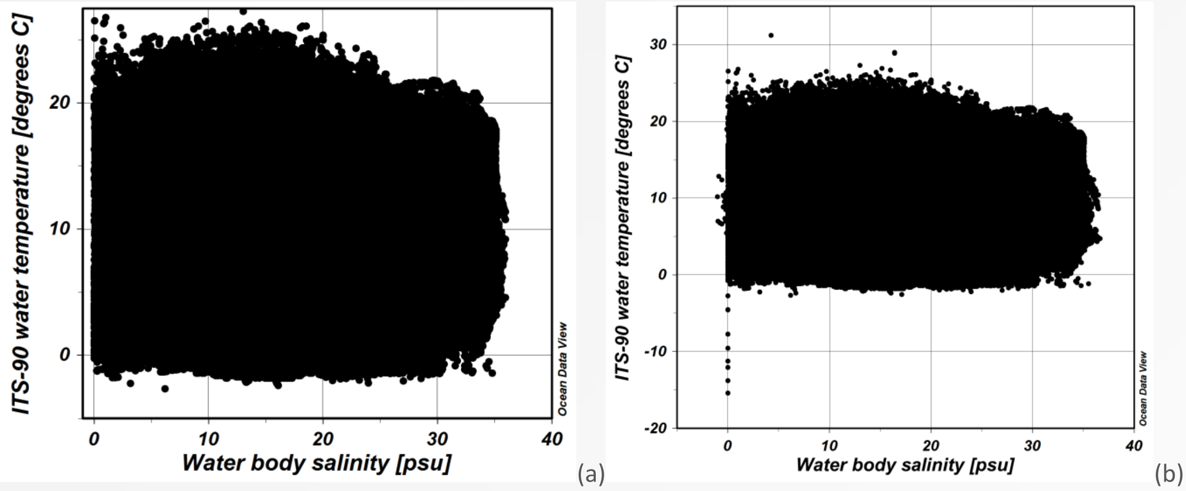


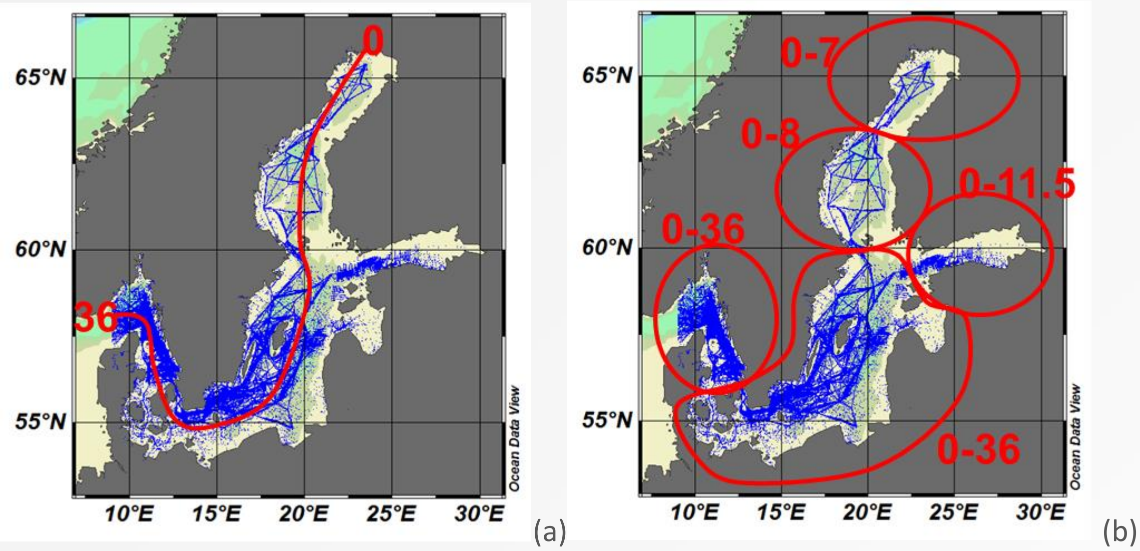
Figure 29 Baltic Sea TS data collection for the time period 1990-2013: (a) TS diagram after QC; (b) TS diagram before QC.

Temperature-Salinity scatter plot before quality control, Figure 29(b), show that are some obvious outliers that are easy to detect and remove. TS scatter plot after quality control show that no obvious outliers are present. It also shows the large range in both salinity, 0 - 36, and temperature, -2°C – 26°C. These large variations make quality control harder, and it makes range checks almost useless. This can be handled by splitting the data into subsets, either by choosing a subset in time or a sub-region with less variation than the whole data set, or both combined.

5.1. Data Quality assessment procedure for The Baltic Sea

Salinity has a large geographical variation, from down to 0 in the north up to 36 in the southwest, Figure 30(a). To handle this, and make QC more manageable, salinity data was divided into sub-regions, Figure 30(b), and visually inspected. Density was calculated and plotted to find unstable profiles. The same procedure was applied for all data, not considering a difference between quality flags 0 and 1; since it is well known that quality controlled data still can contain errors. Obvious bad data were flagged with quality flag 4 (bad), and suspicious data were flagged with flag 3 (probably bad).

In previous version of the historical data collection a more thorough quality control procedure was undertaken; for each region all data were filtered for one year at a time and profiles were visually inspected in ODV to discover spikes, outliers and unstable profiles. For temperature even smaller sub-regions were used to keep the number of profiles in each region lower. Data were then filtered out for one or two months at a time (all years) to handle the large seasonal variation; below 0°C in the surface during winter, and up to above 25°C during summer. Profiles were then visually inspected in ODV to find spikes and outliers.



(a) (b)

Figure 30 (a) Salinity variation in The Baltic Sea, (b) sub-regions used to easier handle the quality control.

All errors found in the previous quality control was checked to see if updates have been made or if the errors were still present. Much of the suspicious or bad data have been corrected or removed from the Seadatanet portal. However in many cases it seems to have happened after the last harvest and the data are still present in the historical data collection, these entire samples have been flagged as bad. Some partners have made corrections in their databases but have not been able to update the data at Seadatanet yet, and some partners have not answered the emails about suspicious data found and done nothing to correct the data.

There are also almost a thousand CDIs with two or more profiles in them containing similar data but different depth resolution. These look to be updated at the Seadatanet portal, the old CDIs have been removed but there are new similar ones available now. These data have been flagged A, value phenomenon uncertain, in the historic collection since the data itself aren’t bad but probably just duplicated.

In the most recent quality control only 40 measurements were found containing suspicious data.

5.2. Conclusions about Baltic Sea V2 data collection

There are still suspicious or bad data that have not been corrected or flagged at Seadatanet, but this is only a very small portion of all data and the overall quality of the data set from the Baltic Sea is high. Less than 1% of all data were flagged as suspicious in the previous quality control, and parts of that are now corrected. The new quality control resulted in only 40 new measurements with suspicious data. Data are well distributed, both spatial and temporal, and works good to use for producing climatologies with DIVA.

6. North Sea

The North Sea aggregated data set V2 contains 1 610 854 stations. Comparing to the V1.1 data set (see Figure 31) it can been seen that most of the data previously of restricted use have in the meanwhile been made public. Some new records have been added and the geographical domain extended to the East (Skagerrak).

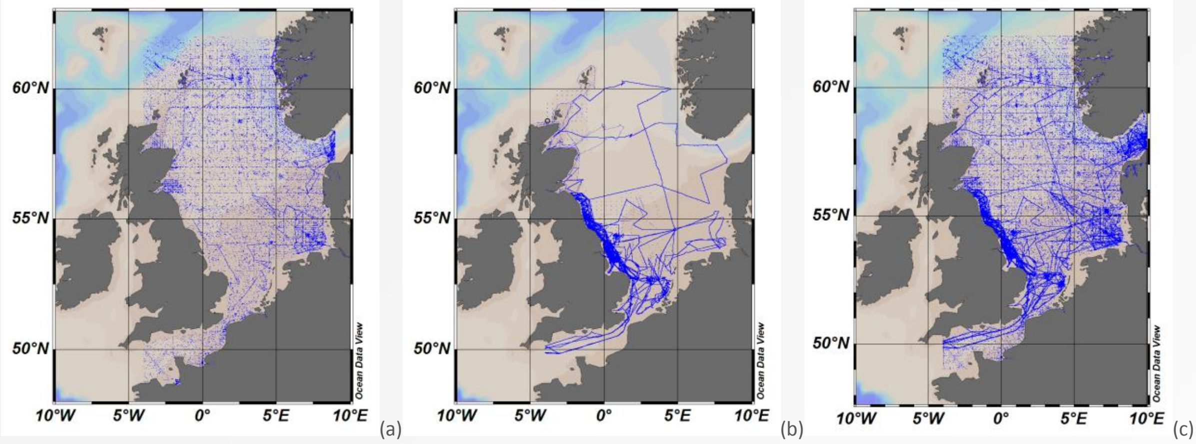


Figure 31 Location of stations in the TS data collections for the North Sea: (a) V1.1 freely accessible data, 751 844 stations, (b) V1.1 restricted dataset, 830 513 stations and (c) V2, 1 610 854 stations.

The geographical coverage of stations is strongly impacted by the two intense measurement programme that took place along the UK coast and at a few stations in the central North Sea in 1988–1990 and 1992-1995.

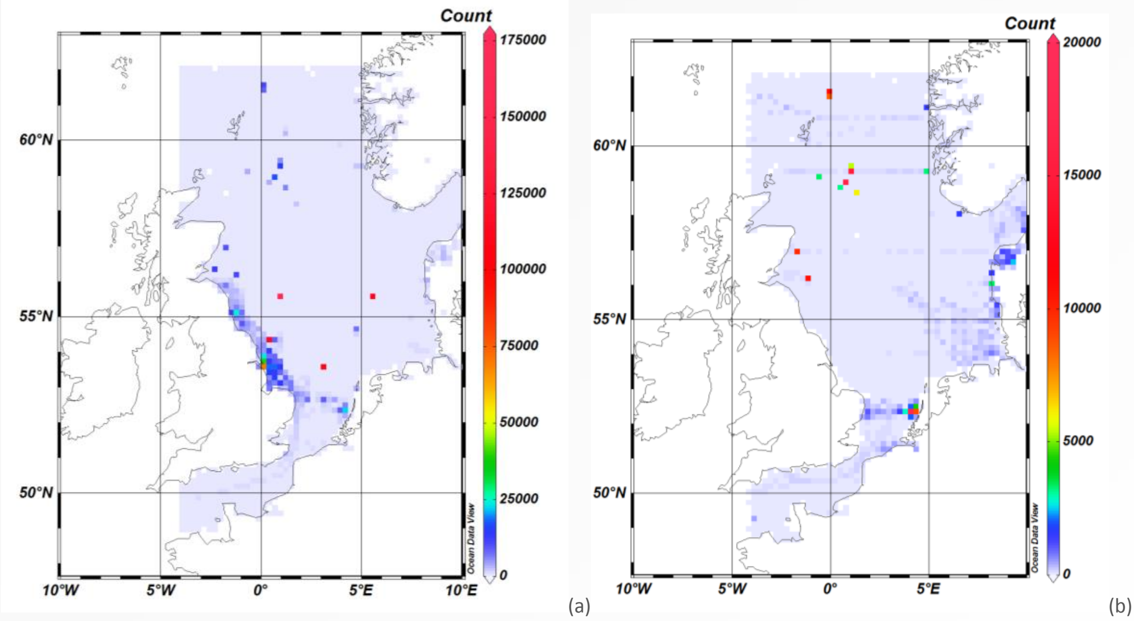


Figure 32 Density map of the TS data for the North Sea. (a) full data set, (b) excluding the period 1988–1995.

The histograms below show the reparation of the data collection effort over the time, for the full data set and with the period 1988–1995 excluded.

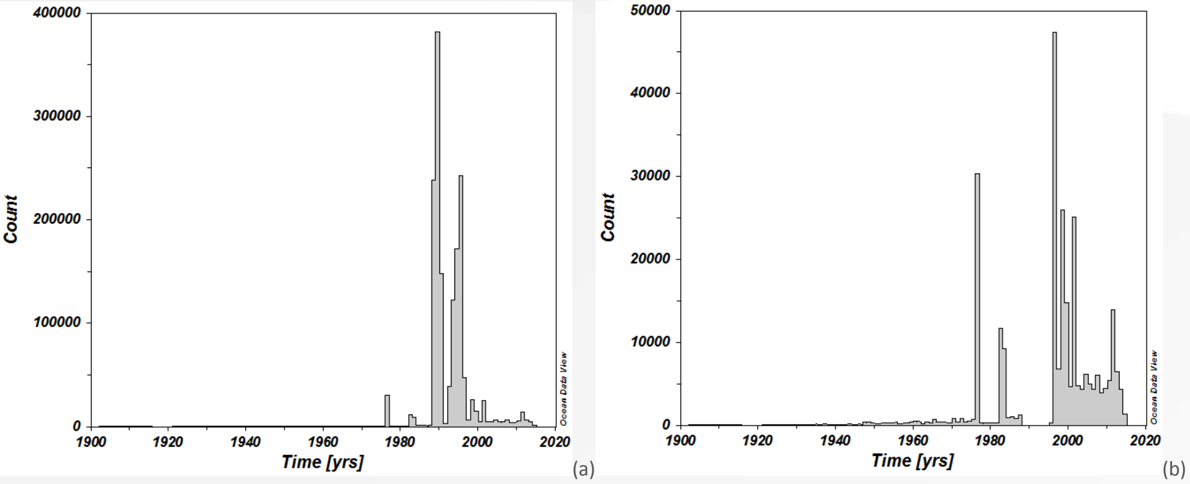


Figure 33 Distribution of the sampling events over the years. (a) whole data set, (b) period 1988–1995 excluded.

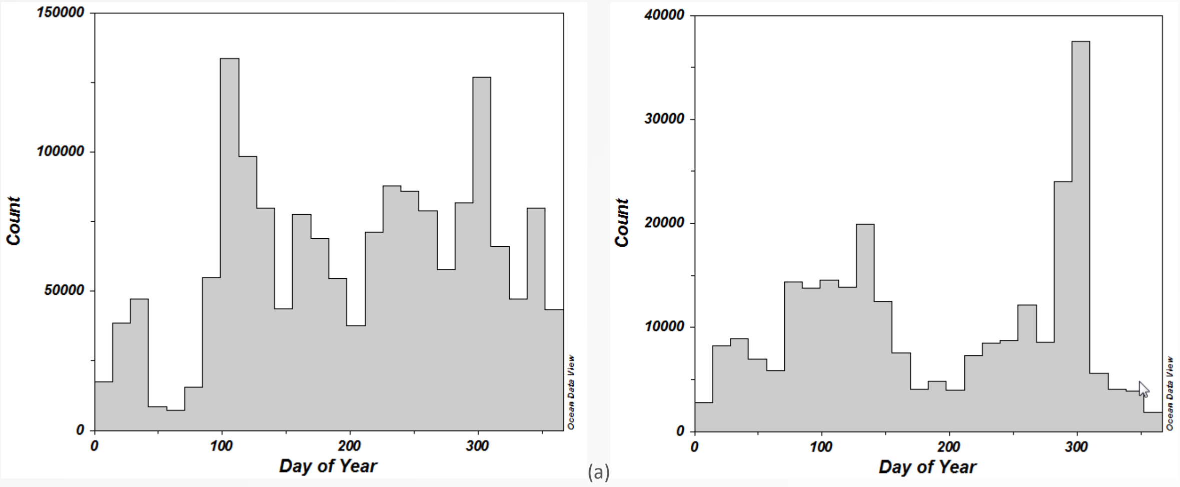


Figure 34 Seasonal distribution of the sampling events. (a) whole data set, (b) period 1988–1995 excluded.

6.1. Characteristics and quality of the historical dataset

The dataset contains 7 344 660 TS-pairs of which 7 256 335 (98.8%) are QC-flagged “1” (good) or “2” (probably good). After quality control, this figure decreases to 7 051 911 (96.0 %).

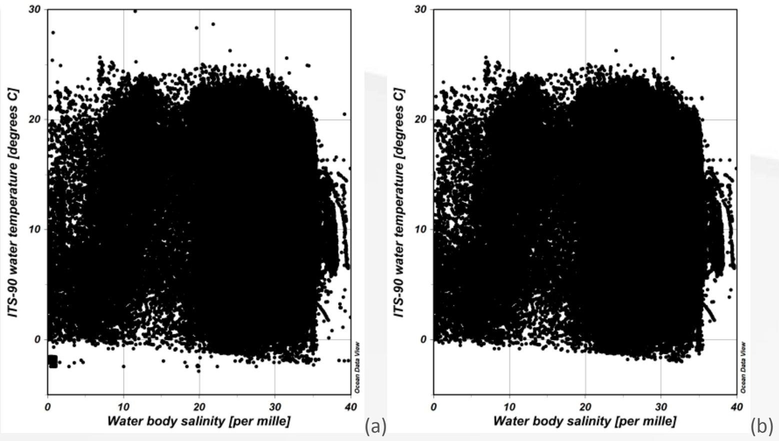


Figure 35 TS diagram of the North Sea data collection before quality control: (a) whole data set; (b) considering only data with QC flags = 1 (good) and 2 (probably good).

The quality control applied follows procedure agreed between the partners.

Data features were examined globally but also more specifically in four regions with their own characteristics (Figure 36):

I. the shallow areas: Southern Bight and German Bight,
II. the Skagerrak
III. the Norwegian coastal region (fjords), and
IV. the area of greater depth around the Shetlands and Orkney.

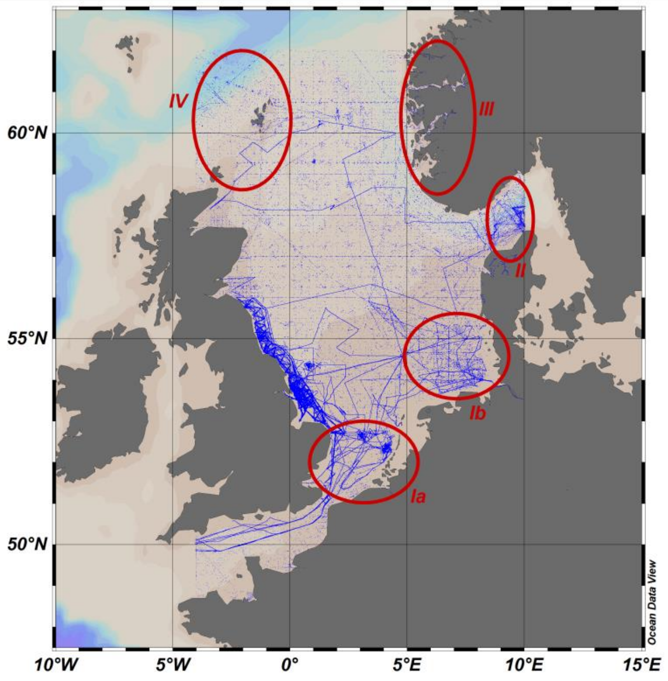
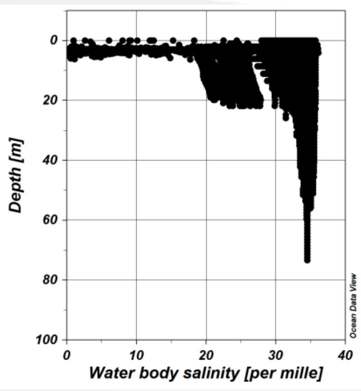
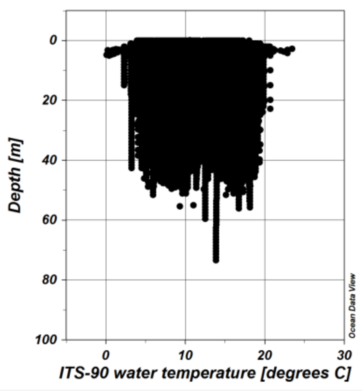


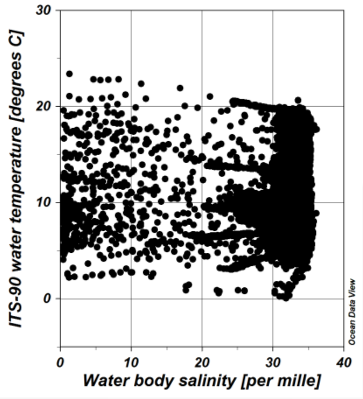
Figure 36 Regions where specific quality control was applied: Ia Southern Bight of the North Sea; Ib German Bight; II Skagerrak; III Norwegian coast; IV deepest region of the basin.



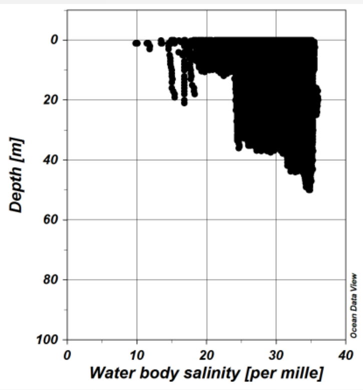
Ia (a)



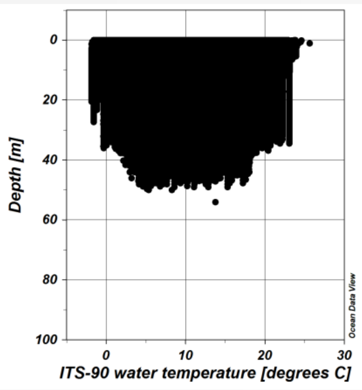
(b)



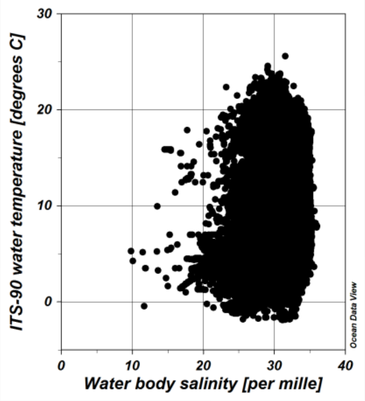
(c)



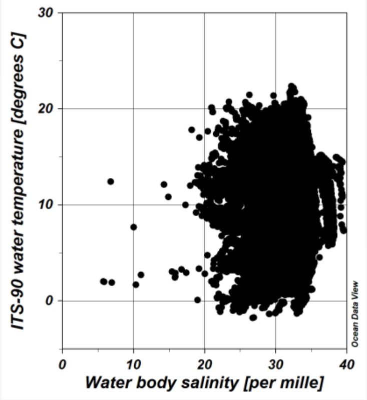
Ib (a)



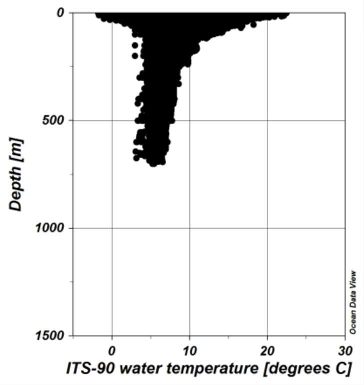
(b)



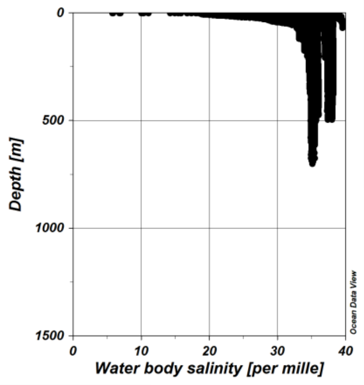
(c)



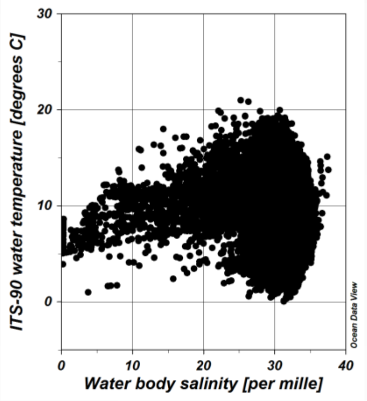
II (a)



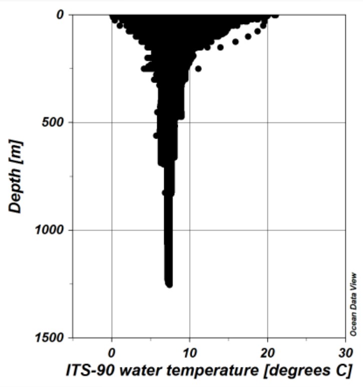
(b)



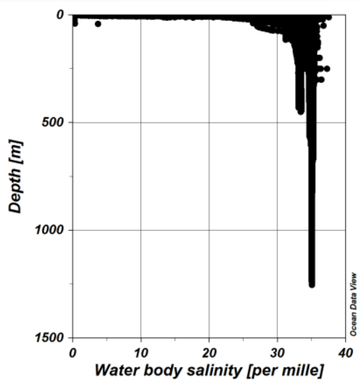
(c)



III (a)



(b)



(c)

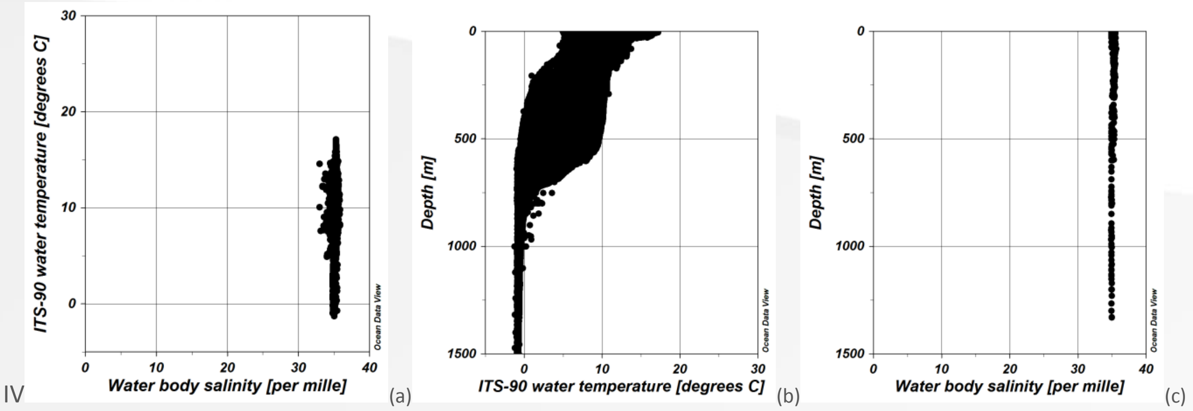


Figure 37 Physical characteristics of the various regions on which quality control focused. From left to right: T,S-diagram, temperature and salinity. From top to bottom, regions Ia to IV.

The main findings of the quality control can be summarized as follows:

- Collating centres apply QC routines that can lead to systematic errors in quality control,
- Data collected when the measuring device was obviously not stabilized are erroneously flagged as "good" (because they pass the range check?). This occurs mainly in shallow waters but not only. We recommend the upper layers not to be published without specific quality control.
- Time series and track data are sometimes published as profiles and had to be rejected (sample position and/or time lacking).
- Oddities are sometimes difficult to identify and to handle, e.g. isolated good data flagged "0" ("not QCed") amongst data flagged "1" ("Good").
- Downcast and upcast both reported.
- Pressure reported as depth.
- QF set to "0" ("not QCed") for obviously interpolated values (should be "8")

These tests lead to the following (Figure 38) diagram of the quality-controlled data sets:

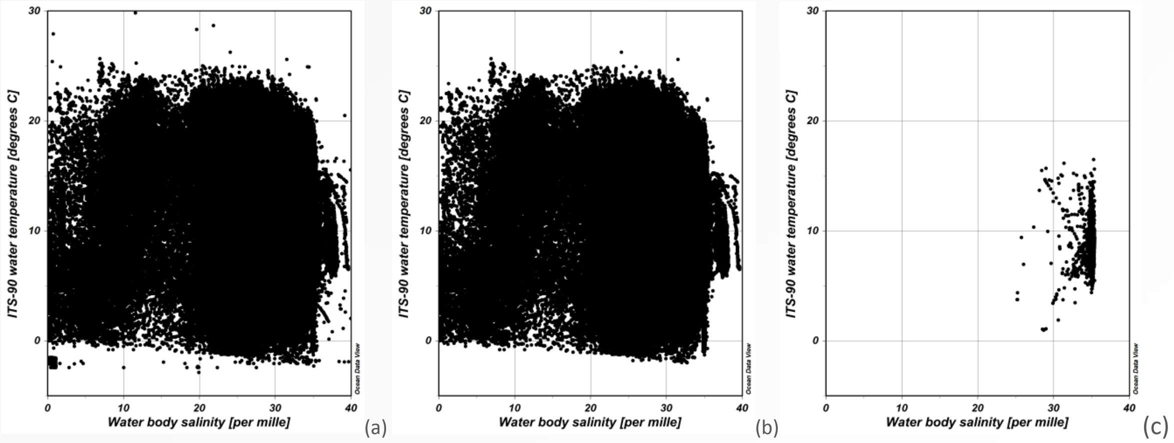


Figure 38 North Sea data collection after quality control: (a) TS diagram after range check analysis; (b) TS diagram considering only data with QC flags = 1 (good) and 2 (probably good) for T and S; (c) TS diagram considering only data with QC flags = 0 (no quality control) for T and S.

After a closer look to the TS-pairs with both QF=0, we decided to promote them to "Probably good data" (QF=2).

	TOTAL	QF0	QF1	QF2	QF≥3
T before QC	16 010 756	286 506 (**1.79**%)	14 907 428 (**93.11%**)	23 818 (**0.15**%)	793 004 (**4.95**%)
T after QC	*Id.*	1 231 (**0.01**%)	14 663 582 (**91.59**%)	27 499 (**0.17**%)	1 318 444 (**8.23**%)
S before QC	7 352 183	19 128 (**0.26**%)	7 238 129 (**98.45**%)	36 269 (**0.49**%)	58 657 (**0.80**%)
S after QC	*Id.*	15 440 (**0.21**%)	7 034 736 (**95.68**%)	39 060 (**0.53**%)	262 947 (**3.58**%)

Tab. 10 Number of Temperature and Salinity data points in the North Sea data collection and their subdivision according to Quality Flags (QF) 0, 1, 2 and from 3 to 9.

6.2. Conclusions

The quality of the North historical dataset is rather good. Adequate handling of track data by some collating centre would add some more valuable data.

The spatial and time distribution of the data (cf. 1988–1995 period) present peculiarities that should be kept in mind when using the datasets.

7. The North Atlantic Ocean

The historical data collection of the North Atlantic Ocean contains Temperature and Salinity observations between 10°N and 62°N of latitude for the east part, and including data into the Labrador Sea till 70°N. The spatial distribution and the data density of Temperature and Salinity observations from the entire data collection are shown in Figure 39(a) and (b). Data distribution maps show a good geographical spread with a best coverage on the east part, mainly close to the coastal areas and in the Bay of Biscay (Figure 39(b)). This higher coverage on the east part is also due to a large number of time-series, which are close to the Irish coast. The North Atlantic Ocean historical data set contains just over 1662303 CDIs for the period 1900-2012 and 79479 CDIs for recent years (2013-2015).

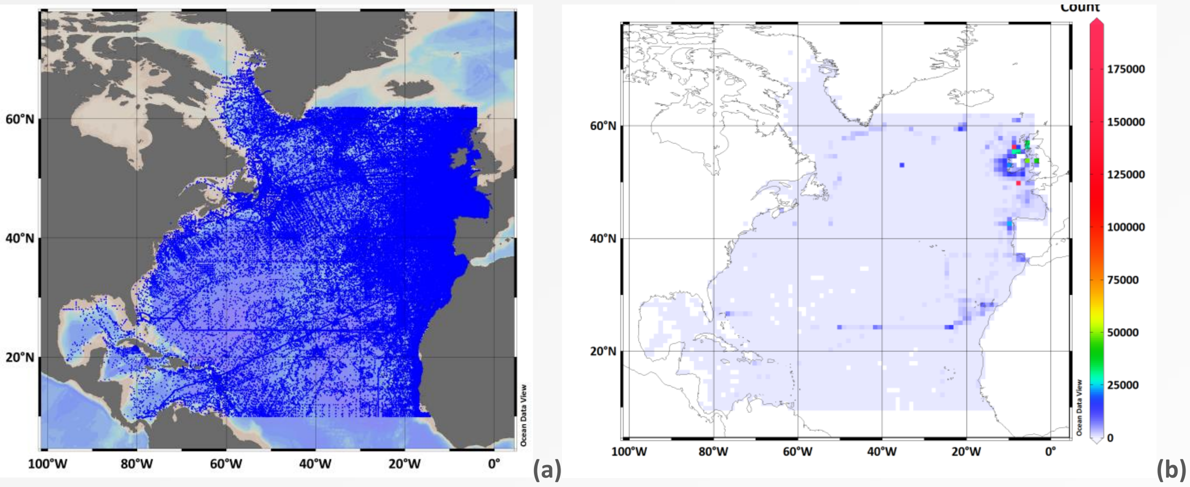


Figure 39 TS stations collection for The North Atlantic Ocean in the time period 1900-2015: (a) Data distribution map; (b) Data density map.

Temporal data distribution is shown on Figure 40. Annual data distribution (Figure 40(a)) shows that there are few measurements up until about 1960. Peaks can be observed in 1996-1997 and some new recent data in 2010-2012, which is good. The number of data increases strongly after the nineties; this might be due to the Argo profilers. Concerning the seasonal distribution Figure 40(b), the peak of the dataset is observed during the summer time.

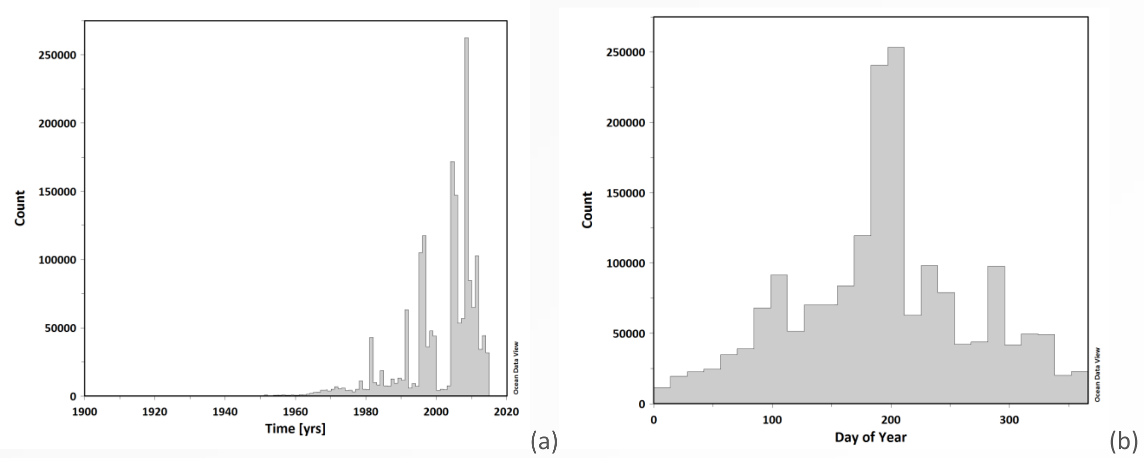


Figure 40 (a) Annual stations distribution and (b) seasonal data distribution for the time period 1900-2015 in The North Atlantic Ocean.

Figure 41 shows Temperature (a) and Salinity (b) data distribution. Salinity observations are less and sparser than temperature ones. Both maps show the largest presence of data to the north, eastwards towards the European coast.

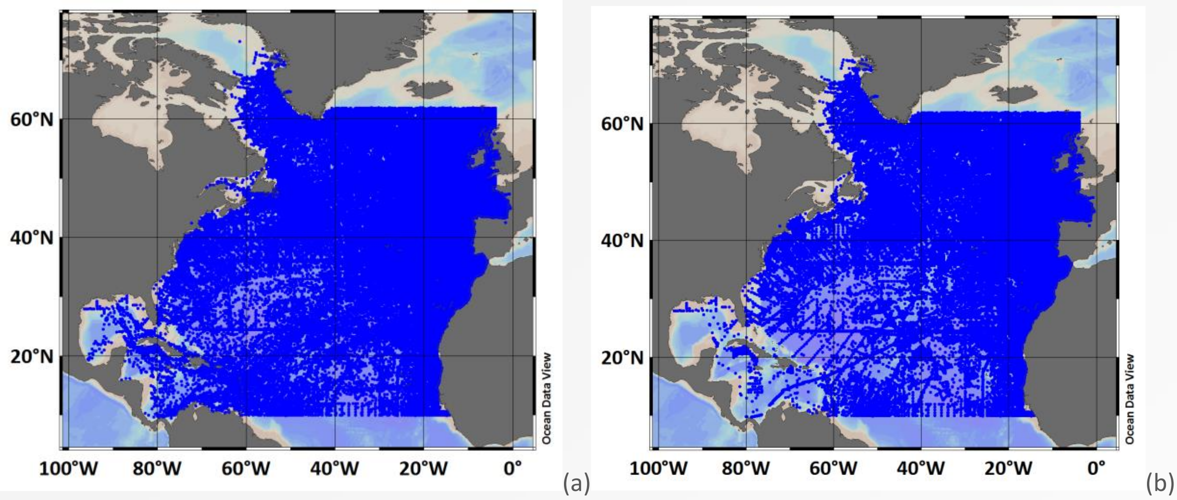
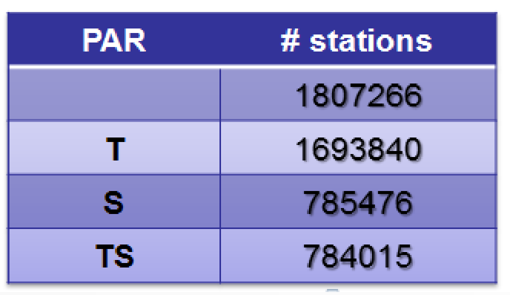


Figure 41 Data distribution map for the North Atlantic Ocean: (a) Temperature, (b) Salinity.

The Tab. 11 shows in details the number of observed stations and its repartition in Temperature stations and Salinity stations and stations that sampled both T and S. Some profiles have Salinity measurements and no Temperature measurements.

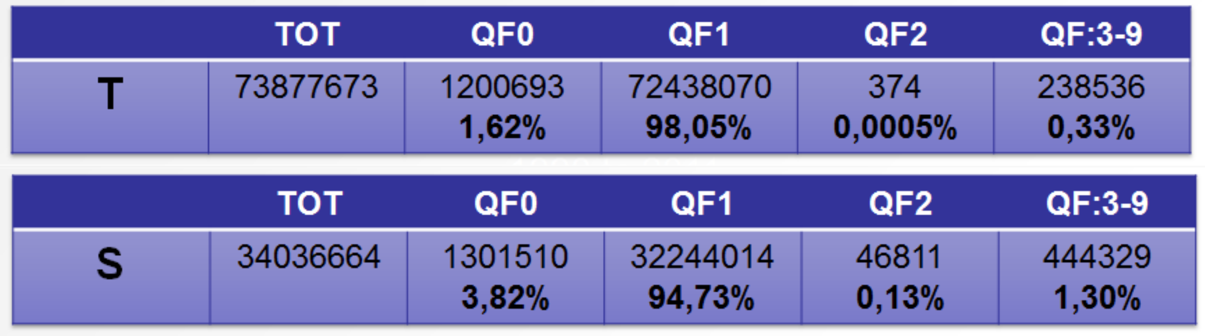


PAR	# stations
	1807266
T	1693840
S	785476
TS	784015

Tab. 11 Number of data points in the North Atlantic Ocean for Temperature, Salinity and TS couples.

After a general description of the historical data set a visual control of all observations allowed to assess their quality and to identify the principal criticalities for possible future applications and users. But the large variability of both salinity and temperature for the North Atlantic Ocean makes the quality control difficult, thus the data set has been split into sub-sets for the QC visualization, either in time or in space (sub-regions) or both combined, with a smaller variation than the whole dataset.

The Tab. 12 shows that the data collection, dividing in groups according to the quality flag (QF) of the measurements, presents a good quality. Nevertheless, some QF0 have been observed (1.62% for T and 3.82% for S) and some bad measurements have still good QF (1). The majority of temperature and salinity measurements (98.05%) have QF1, less than 0.33% of measurements for T and 1.30% of measurements for S are considered as doubtful or bad by the NODCs (see measurements statistics about quality flags in Tab. 12).



	TOT	QF0	QF1	QF2	QF:3-9
T	73877673	1200693 **1,62%**	72438070 **98,05%**	374 **0,0005%**	238536 **0,33%**

	TOT	QF0	QF1	QF2	QF:3-9
S	34036664	1301510 **3,82%**	32244014 **94,73%**	46811 **0,13%**	444329 **1,30%**

Tab. 12 North Atlantic TS measurements collection for the time period 1900-2015: number of measurements (and percent) for temperature and salinity sorted by quality flag: QF0 no control, QF1 good data, QF2 probably good data, QF3-9 others including bad data.

Temperature-Salinity scatter plots before quality control, Figure 42(a)(b)(c), show that there are some obvious outliers (still with QF=1) that are easy to detect and remove from the dataset.

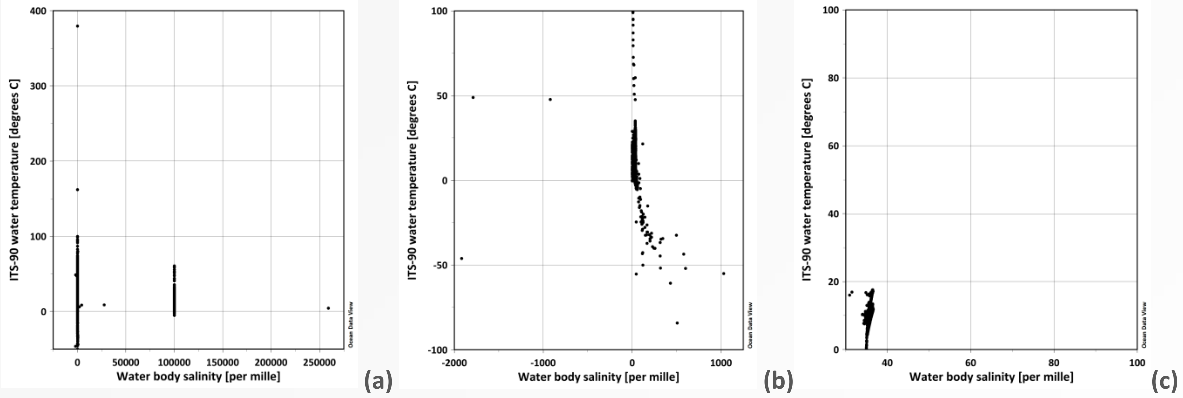


Figure 42 North Atlantic TS data collection for the time period 1900-2015: (a) TS diagram all quality flags; (b) TS diagram with QC flags = 1; (c) TS diagram considering only data with QC flags = 0 (no quality control) for T and S.

TS scatter plots after quality control in Figure 43(a)(b)(c) show that no obvious outliers are present. Few data with QF0 are still present Figure 43(c) and should be updated with a control QF.

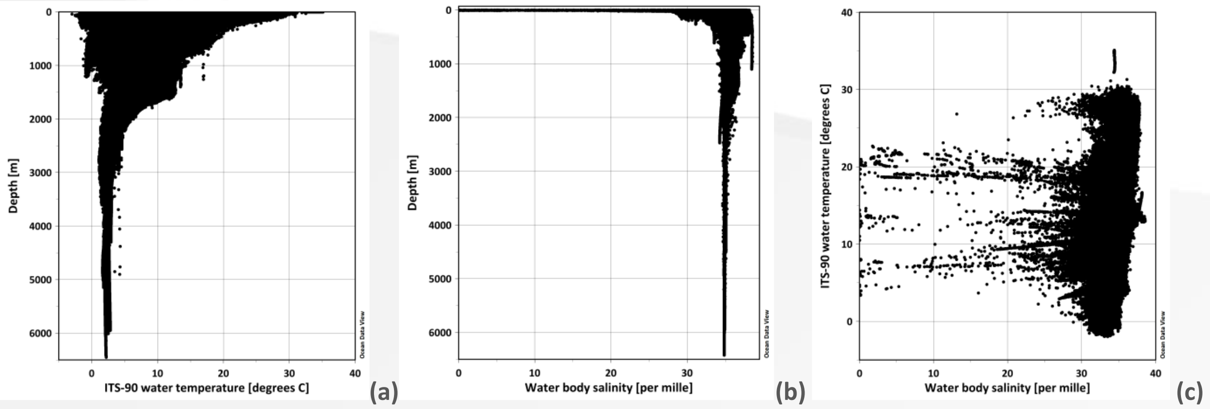


Figure 43 North Atlantic data collection excluding some bad data, considering only data with QC flags = 1 (good): (a) Temperature versus depth for the time period 1900-2015, (b) Salinity versus depth for the time period 1900-2015, (c) TS diagram for the time period 1900-2015.

Those bad data clearly appear in the plots but they represent (after applying QC procedure) less than 1.05 % of the good data local_cdi_ids (10 Edmo_codes are concerned by those anomalies).

7.1. Data Quality assessment procedure for The North Atlantic Ocean

Salinity and temperature have a large geographical variation, since the dataset covers all the North Atlantic Ocean from 62°N (on the east) and 70°N (in the Labrador Sea) to 10°N. To handle this, and make QC more manageable, salinity and temperature data was divided into sub-regions, bands of latitude and longitude; some specifications like time periods and edmo_code numbers have been also used to visually inspect the data. Density was calculated and plotted to find unstable profiles as well as the potential temperature/salinity diagram with isopycnals (see Figure 44 as ex.). The same procedure was applied for all data, taking into account the QF1 of the data, which still contain errors. Obvious bad data were flagged with quality flag 4 (bad), and suspicious data were flagged with flag 3 (probably bad).

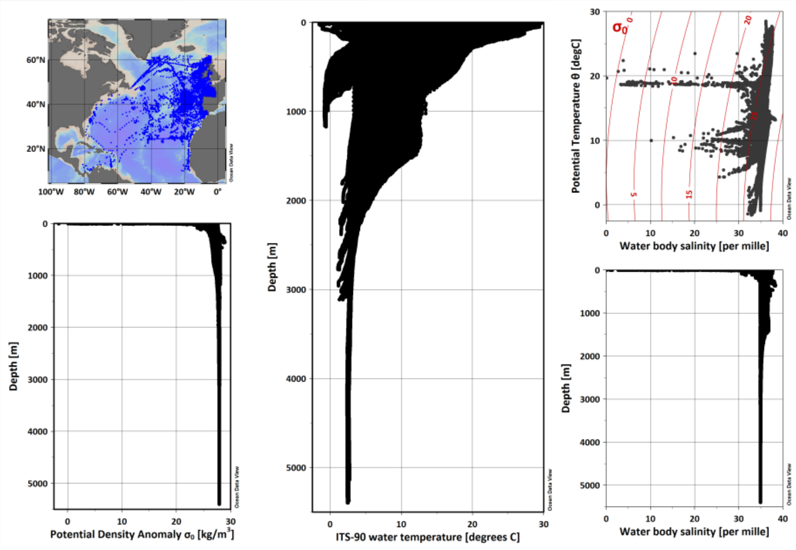


Figure 44 Windows used in ODV to visually inspect the data (map, σ_0 versus depth, temperature versus depth, potential temperature versus salinity and salinity versus depth).

The procedure for the quality assessment analysis is built according to some criteria.

- Outliers: data outside of the regional range defined for temperature and salinity parameters are excluded (QF 4);
- Density inversion, when temperature and salinity measurements are available;
- A visual QF control is applied on all the dataset (with ODV). For all the spikes, density inversions, gradients, doubtful data, that are detected, the QC is changed to 3 (probably bad) or 4 (bad);
- QF0 identified to send feedback to NODC, most of the time QF0 are good data;
- Data on land are also identified.

Some data were found with missing time information, edmo_code and local_cdi_id. Feedbacks have been sent to the CDI partners.

Some time series dataset have also been recorded as profiles, which is not correct since we do not have the time samples for each measurement and cannot be exploited as profiles. Those data have been removed from the V2 dataset and information has been sent to the corresponding CDI partner to update the profiles format to time series format.

The list of anomalies has been sent to the CDI partners.

7.2. Conclusions

The quality of the data set from the North Atlantic Ocean is high. Less than 1.1% of all data were identified as suspicious after a detailed quality control. List of anomalies has to be sent to each concerned CDI partner to correct their QF on data.

The density of the data is higher in the east part of the North Atlantic Ocean than in the west part that could introduce high error for the west part in the upcoming climatologies.

8. General Conclusions

Data Aggregation Procedure was an extensive and constructive exercise to manage more than 1 Million data and involving many WPs, people and institutions. It was huge distributed effort involving 62 data centres and more than 300 data originators.
The **Quality Assessment Procedure for V2 data collections** permitted to identify and correct lots of data. **Aggregation and Quality assessment** allowed to **ameliorate and refine each technical phase** but mainly **to highly improve the quality of SDN infrastructure content.** WP10 promoted collaborations and communication between partners, WPs (WebEX conf) and projects (2 Joint Meetings). All RCs were active and collaborative in QC assessment and reporting activities during the last year of SDN Project.

RCs analysed V2 regional data collections starting from the assessment of SDN data population increase with respect to V1.1. V2 data collections show a data population increase in the Mediterranean Sea and especially in the Atlantic region due to the insertion of new data, as shown Tab. 13. The Black Sea presents a reduction of number of stations due to the removal of data from the central CDI after the Ukrainian Crisis and som duplicate elimination. Statistics from the Baltic and the Arctic could not be computed since the V2 domains do not coincide with V1.1 ones.

REGION	V1	V1.1	V2
Atlantic	431974	1049547 **+143%**	1870266 **+72%**
Baltic Sea (points)	8900000	11700000 **+31%**	NA
Black Sea	93163	101369 **+9%**	96487 **-5%**
Mediterranean Sea	136828	169438 **+24%**	176201 **+4%**
Arctic	407711	445281 **+9%**	NA
North Sea	NA	NA	NA

Tab. 13 Number of data for V1 regional data collections, V1.1 data collections and the relative percentage of data increase.

The **third aggregation phase** allowed retrieving new inserted data but it did not consider the restricted access observations. This third QC phase highlighted that it remains still some bad data flagged as good erroneously and also that there are still data not quality checked with QF=0. RCs sent QC reports to NODCs in order to analyse the data anomalies and make the agreed corrections in the central CDI.

The main conclusions might be summarized in the following point:

- In general V2 data collections contain more data than V1.1 version, with the exception of the Black Sea with an overall reduction of the order of the 5%, due to the removal of duplicates and data from the CDI due to the Ukraine crisis.
- The overall quality of V2 datasets ranges from very good and good, indicating that the QC strategy implemented was successful and that most of the data providers made the suggested corrections in the central CDI.
- Quality assessment highlighted that there are still anomalous data (flagged as good but not good) and data not checked (QF=0).
- QC Strategy was successfully implemented and consolidated, permitting to highly improve the quality of SDN infrastructure content.
- A third feedback on metadata and data anomalies has been sent to data providers asking for corrections in order to continue the monitoring/improvement of SDN database content.
- All RCs participated to WP activities.
- Dissemination through SDN web catalogue (SEXTANT) is on going but it will conclude by the end of the project.

References

S. Simoncelli, *C. Coatanoan, Ö. Bäck,* ***H. Sagen, S. Scory, D. Tezcan, D. M.A. Schaap, R. Schlitzer, S. Iona, M. Fichaut,*** *M.Tonani.* *"TEMPERATURE AND SALINITY HISTORICAL DATA COLLECTIONS FOR THE EUROPEAN MARGINAL SEAS: AGGREGATION AND QUALITY ASSESSMENT PROCEDURES". IMDIS Conference 2013, Lucca, Italy.*
http://imdis2013.seadatanet.org/content/download/73084/949811/file/S1P05_IMDIS2013_SDN2_Products_Simoncelli.pdf