FIRST RELEASE OF THE AGGREGATED DATA SETS PRODUCTS

WP11 - Deliverable D11.2

HORIZON ***2020***

Deliverable number	Short title
D11.2	WP11 V1 Aggregated data sets

Long title
WP11 First release of the Aggregated data sets

Short description
This document describes the SeaDataCloud V1 datasets per sea region (Mediterranean Sea, Black Sea, Arctic Sea, Baltic Sea, North Sea and the North Atlantic Ocean).

Author	Working group
Christine Coatanoan, S. Simoncelli, O. Back, V. Myroshnychenko, S. Scory, H. Sagen	Regional coordinator

Dissemination	Copyright terms
Public	

History

Version	Authors	Date	Comments
1.0	Christine Coatanoan	12/02/2018	Creation and North Atlantic Ocean
	Simona Simoncelli		Med Sea
	Örjan Back		Baltic Sea
	Volodymyr Myroshnychenko		Black Sea
	Serge Scory		North Sea
	Helge Sagen		Artic Sea
2.0	Christine Coatanoan	24/04/2018	Updated version for all the regions with the new dataset

Table of contents

1. Introduction

1.1. Objectives

The main objectives of the WP11 is to improve the quality of the overall infrastructure content, to create the best data products (aggregated dataset and climatologies) and to serve the many user groups (academia, operational oceanography, climate, institutional).

The different steps of this work package are:

- To analyse data distribution and density (space, time, depth) of the regional aggregated datasets;
- To assess data quality making use of the ODV tool and verify the overall improvement of SeaDataNet database content;
- To report to the data providers on the data quality and possible shortcomings;
- To deliver the regional aggregated datasets for dissemination in order to include them on Sextant catalogue, Oceanotron and OceanBrowser.

1.2. Quality Check Strategy

During the SeaDataNet2 activities, the Quality Control strategy QCS, schematized in Figure 1, has been developed and continuously refined by WP11 Regional Coordinators in order to improve the SeaDataNet database content and to create the best product deriving from SeaDataNet data (phase 1). The QC strategy involved NODCs and data providers that, on the base of 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 CMEMS in order to implement a true synergy at regional level, create the best historical datasets to serve operational oceanography and climate change communities.

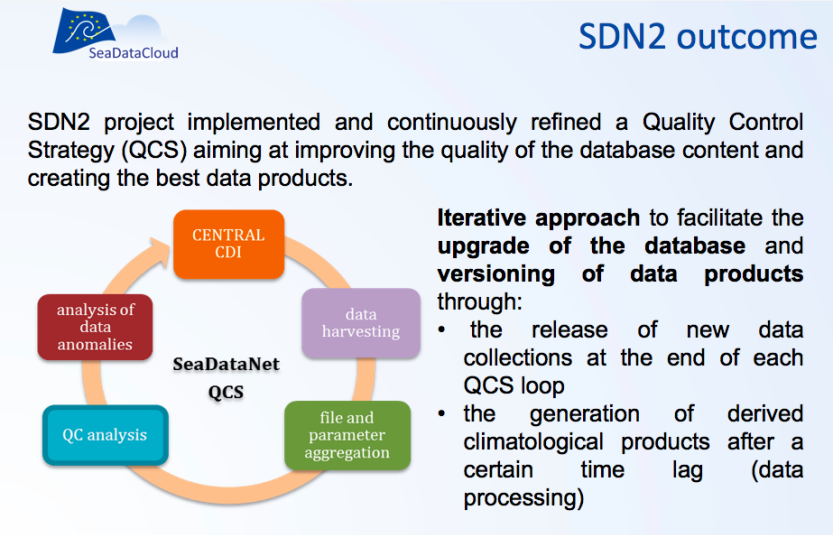


Figure 1. Quality Control Strategy implemented during the SeaDataNet2 Project and adopted during the phase 1 of SeaDataCloud.

The innovation within SeaDataCloud (Figure 2) is to implement a European cloud environment to optimize and automate the QCS at the central level (WP9). The phase 2 will assure a continuous monitoring of the database content and quality, together with the possibility of generating database snapshots on a regular basis, which would undertake the data products versioning.

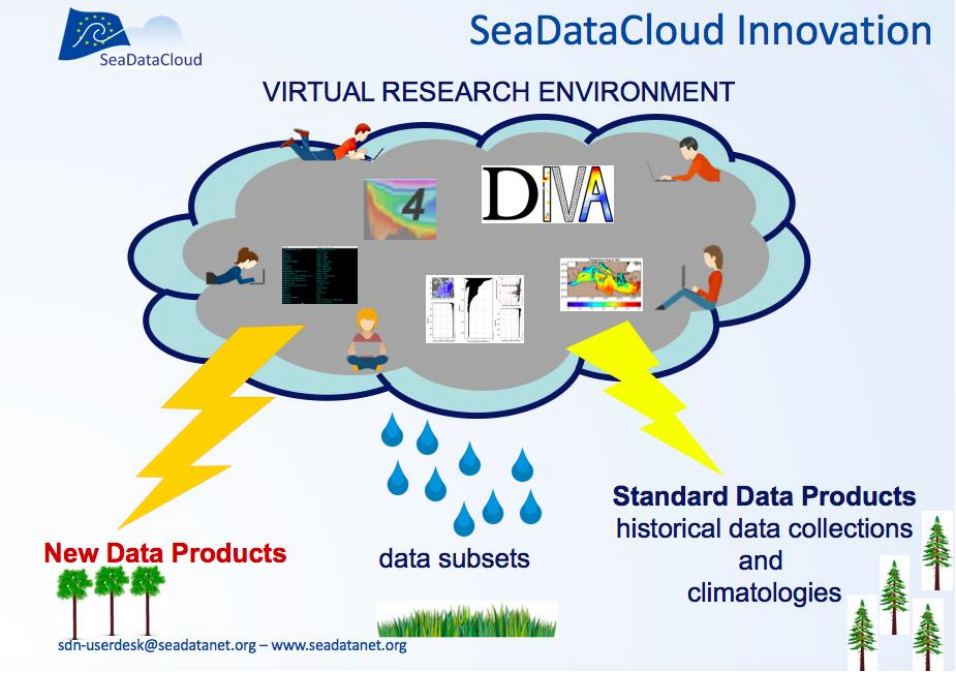


Figure 2. Planned Quality Control Strategy for phase 2 of SeaDataCloud project

1.3. Dataset aggregation

The first step to aggregate the dataset uses the data import from the "Import > SeadataNet Formats" ODV option. All data for all parameters found in the data files are imported into automatically created ODV profile or time series collections. During import, no parameter aggregation or unit conversions take place. In addition to the CDI metadata, the import ODV collections also have a "P01 Codes" meta variable containing for every station/CDI the list of P01 codes found in the given file. The last version of ODV (ODV 5) also performs subsampling of underway data and discards stations with spatial and temporal distances to the previous station of less than about 300 m and 10 min. Typically, this leads to large reductions of the number of stations (only 10 to 20% of stations retained), without significant loss of track accuracy or scientific information content.

Collections created via SDN data import, have the "Export > Station Data > SDN Aggregated ODV Collection" option enabled, which performs the P35 parameter aggregation. As first step, the latest version of the P35 vocab is downloaded from the BODC vocab server, and the latest version of the unit conversion document is downloaded from the ODV site. The original parameters in the ODV import collection are grouped into P35 parameter strictly as defined in the P35 vocab. P35 also specifies the target units for each P35 parameter. For each station/CDI and P35 parameter, ODV calculates value and quality flag by combining all contributing values. In the majority of cases, a given station/CDI only has a single contributing P01 parameter for a given P35 parameter.

Using the P35 aggregated ODV collections, delivery ODV collections are created by combining the various profile and time-series created during the phase of data import and phase of the P35 parameter aggregation, each time excluding parameters that should not be part of the delivery. This

results in separate combined data collections for profile and time-series data, separate for restricted and non-restricted data.

The first delivery of the aggregated dataset was made at the end of November 2017. In this version the underway data sets were imported in full resolution. In February 2018, after some analysis from regional leaders, some cruises were found missing and a new aggregated dataset was made available at beginning of March. In this new version, a sub-sampling of the underway data was taken into account.

1.4. Data Quality Assessment procedure

The QC analysis during SeaDataNet2 project was conducted using ODV software. The idea in SDC is to start with the same approach and develop meantime automated QC procedures.

The basic QC analysis steps applied during SeaDataNet2 Project were:

1. Data coverage;
2. Data distribution maps per Temperature, Salinity and TS couples;
3. Data density maps (domain binning);
4. Histograms with annual, seasonal and monthly data distribution;
5. Statistics about Quality Flags;
6. TS scatter plots of observations with QF=1 (good) and QF=2 (probably good);
7. Gross range check to detect observations with temperature and salinity out of reasonable values;
8. TS Scatter plots after the range check;
9. Scatter plot observations with QF=0 (no quality check) to disclose good data;
10. Visual control of scatter-plots to identify wrong profiles (outliers);
11. Visual check of spikes;
12. Identification of stations falling on land;
13. Identification of wrong or missing data;
14. Stability check on density

Sub-regional checks are advisable per specific areas (areas with similar hydrodynamic characteristics) and per layers (surface layer, intermediate waters, bottom layer).
In view of climatology computation, the QC analysis could be conducted over decades or specific periods, when particular climatic events took place (i.e. Eastern Mediterranean Transient, Western Mediterranean Transition, and Norther Ionian Reversal)

Since some additional steps have been developed by basins, those procedures are described hereafter.

1.4.1. Mediterranean Sea

In the Mediterranean domain the basic QC analysis has been performed following the common QC guidelines in paragraph 1.4.

The first phase was dedicated to correct data anomalies from specific EDMO_codes:

- EDMO_CODE=486 → all measurements flagged as 0 were changed to 1 after consulting the data provider. These observations have been flagged zero by the aggregation procedure but they are good;

- EDMO_CODE=840 → all measurements with temperature and salinity values equal to 0 were wrongly flagged 0. The QF was set to 4.

The second phase was dedicated to the data with QC=0 in order to quality assess all the data within the collection. The adopted strategy was to substitute it with 2 (probably good) following these steps:

- To select all measurements with QF=0 for depth&T&S and assign QF=2;
- To select T measurements with QF=0 assign QF=2;
- To select S measurements with QF=0 assign QF=2.

Then gross range check has been applied:

- → negative depth values have been flagged 4
- → T<2°C and T>33°C assigned to QF=4;
- → S>42 assigned to QF=4.

Law salinity values should be carefully analysed before data usage since a lot of measurements have been sampled at the river mouth. A gross range check for low salinity values has not be applied to preserve these coastal observations, which are crucial for many data analysis and applications.

Many spikes have been identified and flagged 4.

The QC analysis was performed also by the principal instrument type in order to study their monitoring data space-time coverage and their metadata population. Figure 3 show the data distribution maps of measurements sampled by:

1. thermosalinographs (550176 stations)
2. CTD (51537 stations)
3. bathythermographs (56274 stations)

This is phase is fundamental for consistency analysis among data belonging from different instrument types before their usage for long term studies.

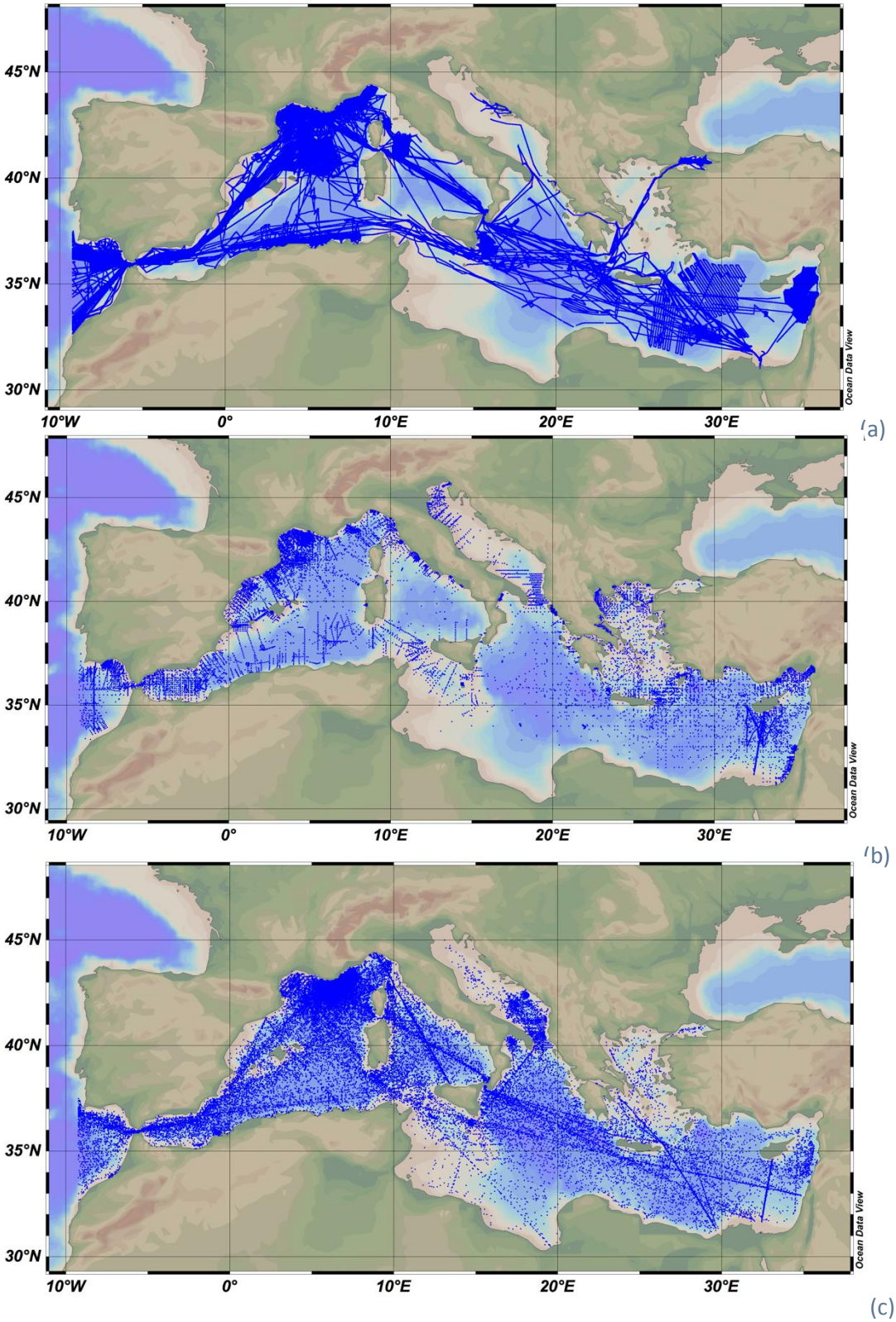


Figure 3. Data distribution maps divided by the principal instrument types: (a) thermosalinograph data (ferry box); (b) CTD data; (c) bathythermographs data.

In order to detect data anomalies more efficiently by visual inspection the data domain has been subdivided in 17 regions according to the definition in Figure 4. The Marmara Sea has been analysed

separately as region 18. This last analysis phase was conducted per sub-regions where specific water types are apparent and data anomalies are easy to find.

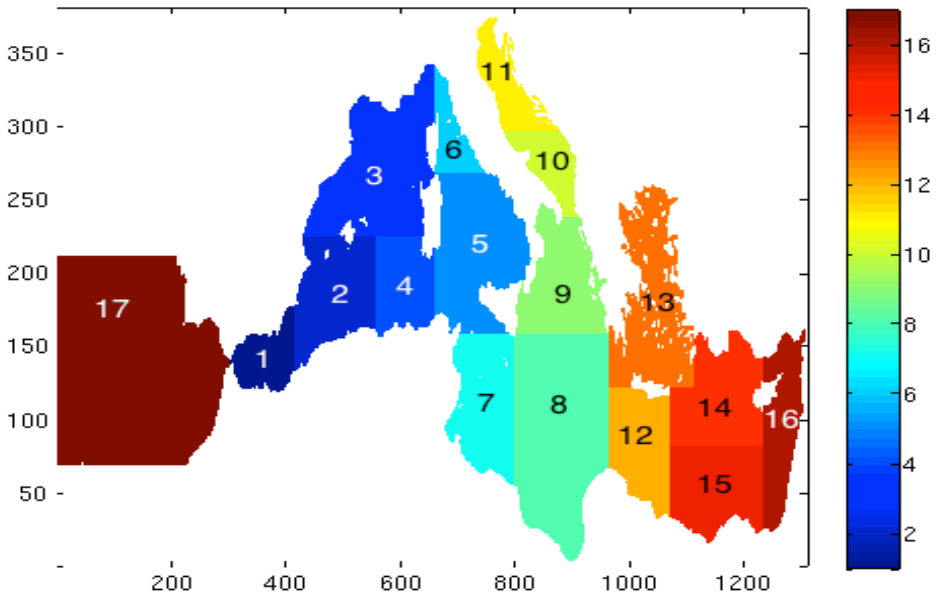


Figure 4 Domain subdivision in 17 sub-regions..

1.4.2. Black Sea

Prior performing quality control of data the collection was checked for presence of **duplicates**. The initial collection contained data at 151530 stations from 2550 cruises (datasets). Analysis revealed presence of more than 12,000 duplicate stations. The duplicates were excluded (deleted) from the collection along with stations having wrong location (on land) or being out of region (e.g. Bosporus stations). **The final collection contains 137723 stations from 2286 cruises**.

The quality control of the Temperature and Salinity data was performed following the guidelines in paragraph 1.4 of this document and taking into account peculiarities of the Black Sea water masses such as permanent halocline and a two-layered structure of the waters, presence of Cold intermediate Layer (CIL), conservativeness of properties of deep water layer (below 200 m), etc. The physical properties of the Black Sea water masses remain the same practically through the whole basin except the North-Western shelf, which is under influence of inflow from large rivers, and area of “Bosporus plume”, where saline Mediterranean waters flow to the Black Sea.

Despite the efforts undertaken in SeaDataNet2 project to improve the overall quality of data within the SeaDataNet infrastructure the initial Black Sea data collection still contained significant amount of not QC-ed data (i.e. having QF=0) as well as wrongly flagged data (i.e. bad data flagged as good and vice versa). Prior performing quality control the QF=1 (good) was assigned to all data having QF=0 (not checked), then the results of data flagging in SeaDataNet2 project were applied to the data in the current collection where possible, and then the quality of data flagged as good was revised with the help of ODV software.

The following simple **range check** procedures were applied to the whole data array, allowing to identify and flag obvious erroneous data:

- Depth < 0,
- Temperature < 0 and QF<>4,

- Temperature > 30 and QF<>4,
- Salinity < 0 and QF<>4,
- Salinity S > 39 and QF<>4,
- Salinity > 23 out of “Bosporus plume” area (28.8<Longitude<29.3, Latitude <41.6),
- Temperature < 6 at depth > 200,
- Temperature > 10 at depth > 200.

Further the profiles were analyzed for **spikes** (using both gradient plots and visual checks), and for **stability** with the help of plots of density derivative. In addition to duplicates and location checks described above the following **metadata checks** were performed for informing data providers about found mistakes and requesting for corrections:

- Identification of stations with Bottom Depth < 0,
- Identification of stations with profile depth > Bottom Depth,
- Identification of stations with wrong date,
- Identification of stations with missing time.

The collection contains some amount of data from stations located in river estuaries. Data values at such stations can vary in wide range depending on season and weather conditions.

1.4.3. Artic Sea

The quality control of the Temperature and Salinity data was performed following the guidelines in paragraph 1.4 of this document. The Artic data collection contained ferry box stations where year 2000 was interpreted as 1900 data in a January cruise. Year was corrected to 2000. Fixed stations along the Norwegian coasts had “No value” data without correct flag QF 9. Spaces were interpreted as Zero values in the production of the Arctic data collection, and these zero values were deleted (replaced by space) in ODV software, and ODV assigned QF 9. 3 cruises had no negative temperatures in the deep water. As the temperature decreased with depth and reached 0, the temperature started to increase with depth. Visually it was obvious that below this depth the sign was wrong and that by correcting the sign to the negative, the profiles appeared normal. These profiles were corrected and flagged QF 1. All data with values flagged QF=0, were revised with ODV. 2 cruises with anomalies in salinity in the deep stable waters were revised to probably bad and bad, in all values in the profiles. Temperature at 1000m is negative and salinity between 34.89-34.92 in most of the Arctic Sea north of the seabed threshold in the Shetland channel between Faroe Islands and Shetland. This range has been used as a QF 1 indicator for the deeper CTD stations. In the Lofoten Basin west of northern parts of Norway, these indicators cannot be used, as temperatures can be 4-5 degrees higher, and salinity close to 35 at 1000m in a small central area.

Following range checks were used for identification and flagging obvious erroneous data:

- Depth < 0 QF <> 4
- Temperature < -1.9 QF <> 4
- Temperature > 25 QF <> 4
- Salinity < 10 QF <> 4
- Salinity > 35.5 QF <> 4

Profiles were inspected for spikes by visual check
Identification of stations with profile depth > Bottom depth
Identification of time errors with year 1999 to 2000 shift.

1.4.4. Baltic Sea

The quality control procedure followed the best practices that were defined during the project SeaDataNet 2 and summarized in paragraph 1.4. The most powerful and useful quality control tool used was visual inspection of subsets of data in ODV to discover spikes, outliers, unstable profiles and stations on land. Errors found during the work in SeaDataNet 2 were also checked to see if old errors/suspicious data were still present.

Salinity field in the Baltic Sea has a large geographical variability, from 0 in the north up to 36 in the southwest (see Figure 5a) thus sub-regions have been defined and used to make the quality control more efficient (Figure 5b).

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.

Considering the large seasonal variability of the temperature field (below 0°C at the surface during winter time, and over 25°C during summer time) it is hard to find suspicious data using range checks or by visually inspecting all data. The solution was to analyze subsets of data for season or month at the time.

Checking bottom water is easier than the surface layer since it is not affected by the seasonal temperature variability and also salinity presents most stable concentration, especially in the Baltic Proper.

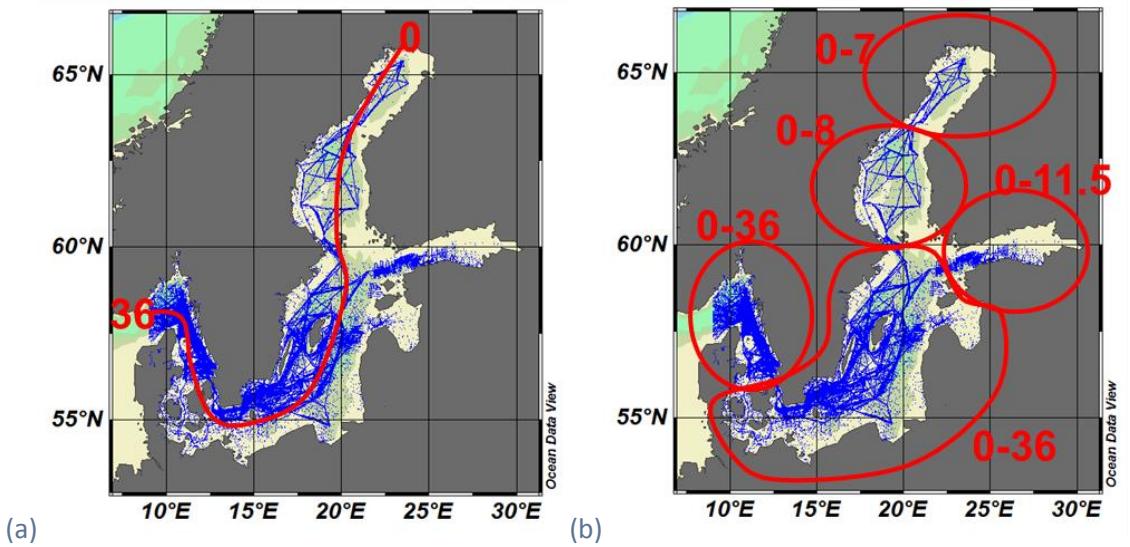


Figure 5. Salinity variability in The Baltic Sea (a), sub-regions used to make the quality control more efficient (b).

1.4.5. North Sea

Errors found during the work in SeaDataNet II were checked in first instance as it was unsure whether our recommendations for improving the data had been implemented. Some systematic errors were still there but most of the time the overall quality had improved.

Using ODV, after handling obvious outliers in the whole data set we took two approaches: the first was to inspect the data by collating centre. Our experience has indeed shown that the routine

applied at each centre for processing the data tends to generate systematic errors. The second approach was to work on sub-regions (Figure 6):

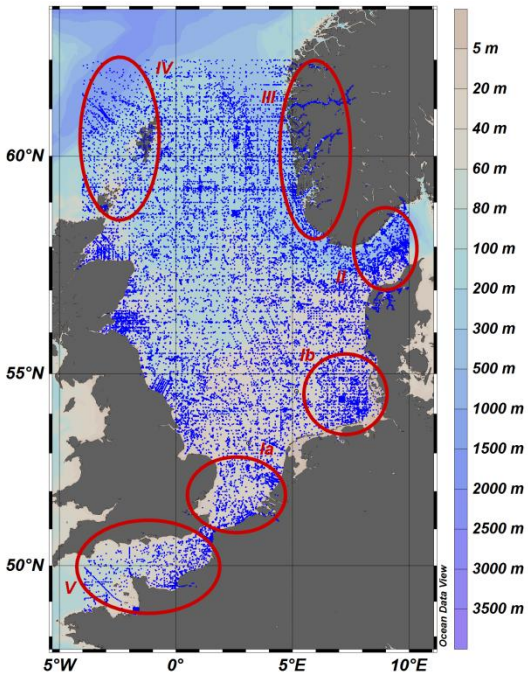


Figure 6. The various sub–regions used for the detailed quality control. I: Channel, IIa & IIb: shallow regions, respectively the Southern Bight and the German Bight; III: Skagerrak; IV: Norwegian fjords; V: Deep oceanic zone.

Whenever needed, data were checked "in context", *i.e.* by looking at all data on smaller geographical or temporal scale or for a given period of the year.

The DIVA analysis tool incorporated in ODV has also been used for spotting anomalies (Figure 7).

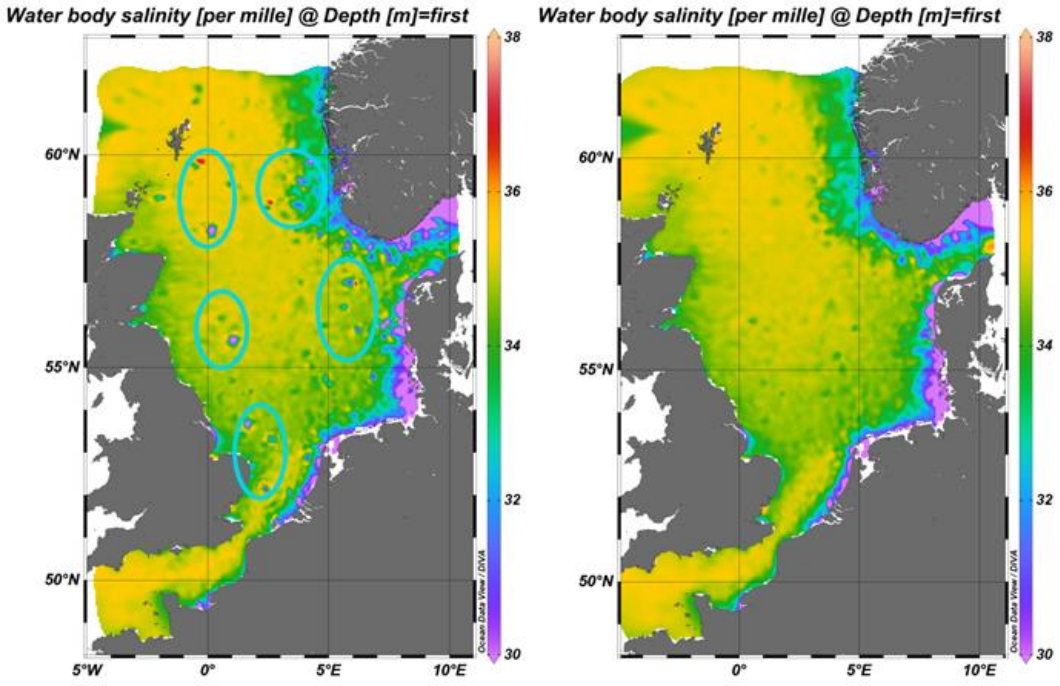


Figure 7. Examples of anomalies detected using DIVA in ODV. Maps show the Salinity field ("discrete" collection) in winter, before and after cleaning of the data.

1.4.6. North Atlantic Ocean

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. 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 data set has been split into 6 sub-regions for the QC visualization (Figure 8) following the water masses characteristics, with similar hydrodynamic. In some sub-region, a new selection has been applied to take into account some time periods to decrease the number of stations to quality check.

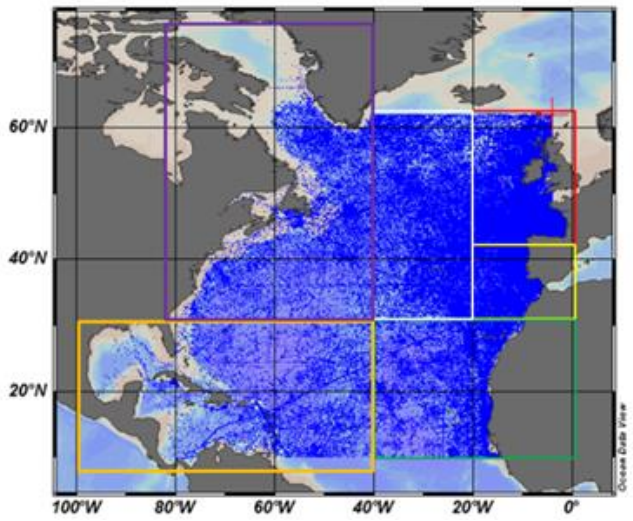


Figure 8. Regional subsets to individually check the QC on the data.

An additional data selection by EDMO_CODE had also been used to focus on some anomalies, thus the same QC procedure has been applied per data centre to detect their eventual systematic errors and data anomalies.

Then potential density was calculated from temperature and salinity profiles and plotted to detect unstable profiles. θS scatter plots with isopycnals helped to further identify data anomalies. Visual inspection was the most used practice to identify the outliers, spike, unstable profiles and stations on land.

The same procedure was applied for all data, considering only the quality flag 0, 1 and 2; since it is well known that quality controlled data still can contain errors.

Checking data in the bottom layer is easier than the surface one because it is not affected by the seasonal temperature variability and it presents stable salinity concentrations.

The quality control analysis followed the best practices that were defined during the project SeaDataNet 2 and summarized in paragraph 1.4, here after reported for the regional specifications:

- Checks of the data coverage, by sub-region when necessary (distribution for T, S, TS couples), by time periods, by layers (distinction between surface, intermediate and bottom layers);
- TS scatter plots of the entire dataset: T versus Z, S versus Z, θS diagram with isopycnal levels for all the QF<3 (check the outliers and change the QF to 4); sometimes the outliers were the missing data values (i.e. T=0.9999) with not appropriate QF (1 instead of 9);
- By sub-region, scatter plot of observations with QF=1 (good) with a secondary plot showing the density;

- By sub-region, scatter plot of observations with QF=2 (probably good) with a secondary plot showing the density;
- Scatter plot observations with QF=0 (no quality check) only to change the bad data with QF4;
- Identification of stations falling on land;
- Identification of stations having unreal depth (depth values<0);

The most useful and powerful quality control used was visual inspection of subsets of data in ODV to discover spikes, outliers, unstable profiles and stations on land.

2. Data statistics per sea basin

Some statistics from regional TS data collections were analyzed in order to:

- identify the progresses on the quality of the overall CDI data content monitoring data Quality Flags (QF) statistics and performing a comparison with SDN2 V2 datasets;
- point out the advancement of number of temperature and salinity data contained in the CDI.

2.1. 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.

SDC_MED_DATA_TS_V1 collection has been obtained harvesting all measurements contained within SeaDataNet infrastructure at the end of November 2017 belonging to 24 data providers (distributors) and 100 data originators.

The spatial distribution and the data density of measurement stations are shown in Figure 9. The spatial distribution of data (Figure 9a) presents a good data coverage in the Western Mediterranean basin and the Atlantic box, while in the Eastern Mediterranean still many areas are characterized by few and sparse data, like the coastal areas of Tunisia, Libya, Egypt and Croatia. Data density map (Figure 9b) 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. Data density is high also in the Marmara Sea.

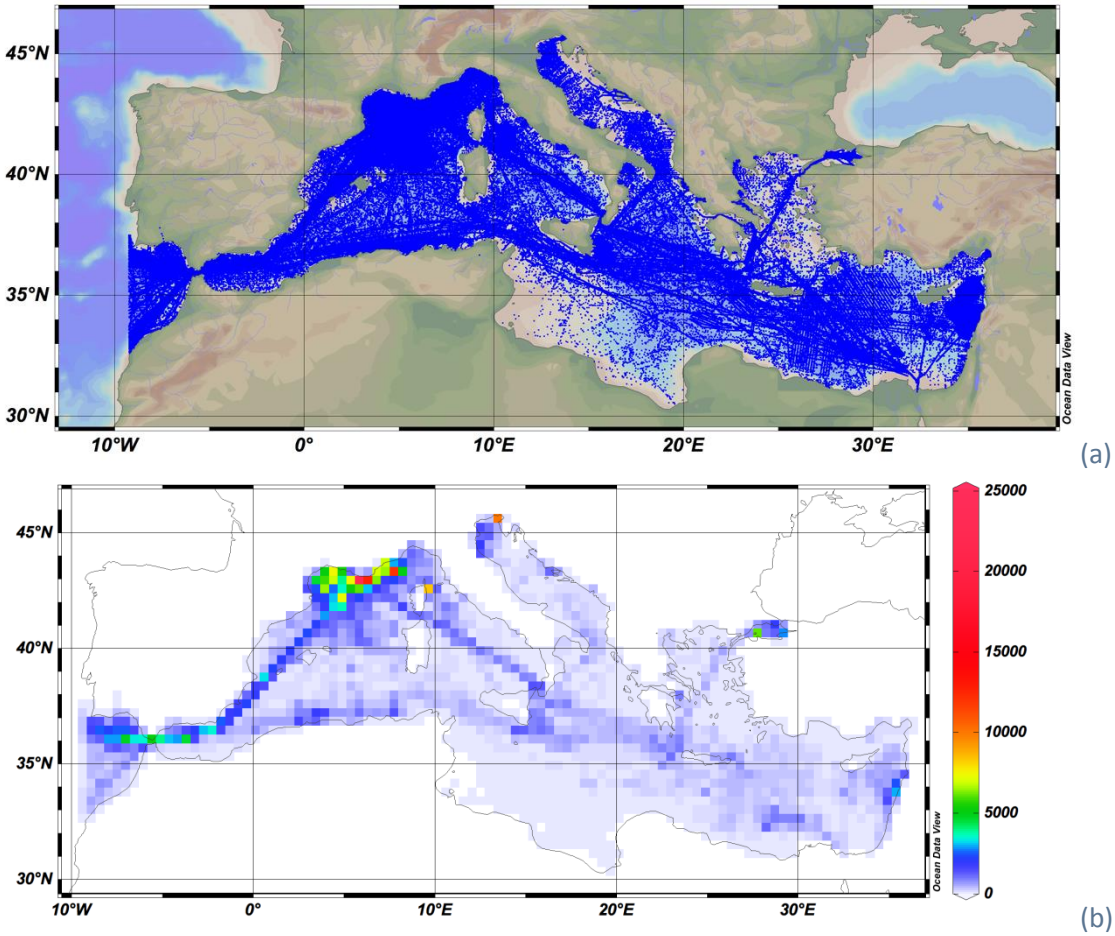


Figure 9. Temperature and Salinity data collection for the Mediterranean Sea in the time period 1900-2017: (a) Data distribution map; (b) Data density map.

Temporal distributions of data are in Figure 10 and Figure 11. Annual distributions (Figure 10a and Figure 11) prove that data are very sparse before 1950 and they start to increase systematically from the sixties and concentrate mostly in the noughties, due to the increase of high resolution ferry box data. This must be taken into consideration during climatological data analysis. Seasonal distribution of data (Figure 10b) 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.

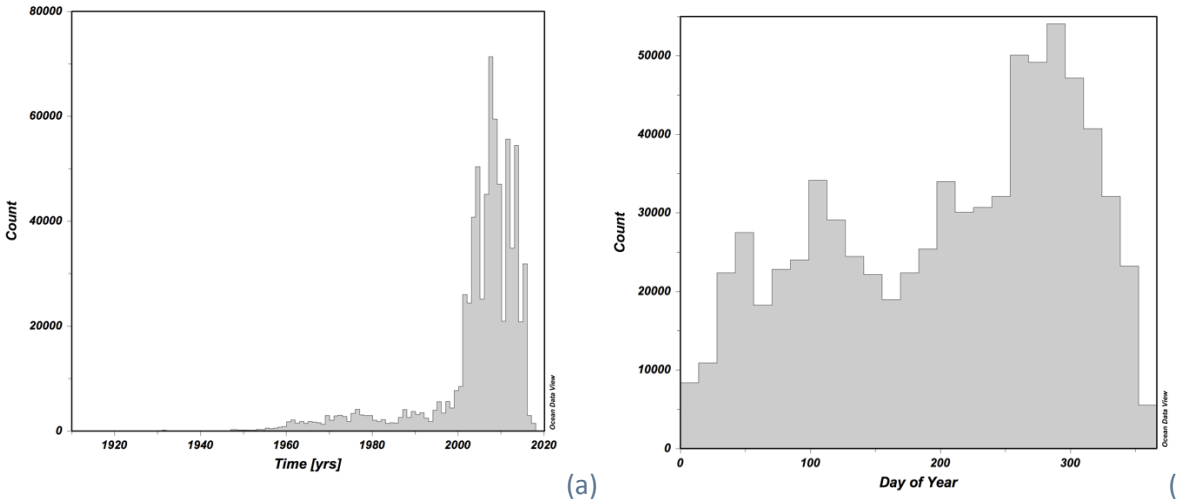


Figure 10. Temporal data distribution over the time period 1900-2017 in the Mediterranean Sea: (a) annual, and (b) seasonal.

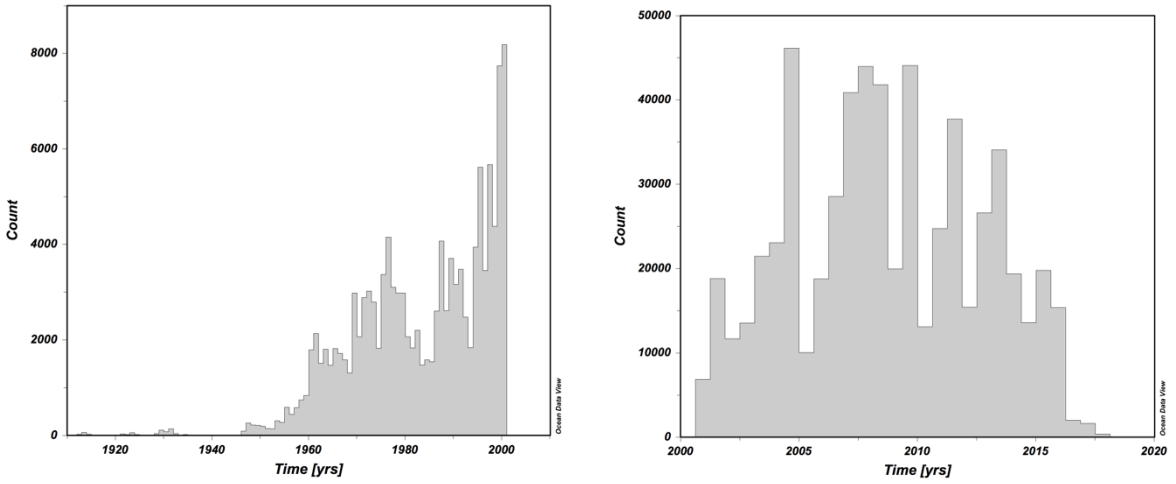


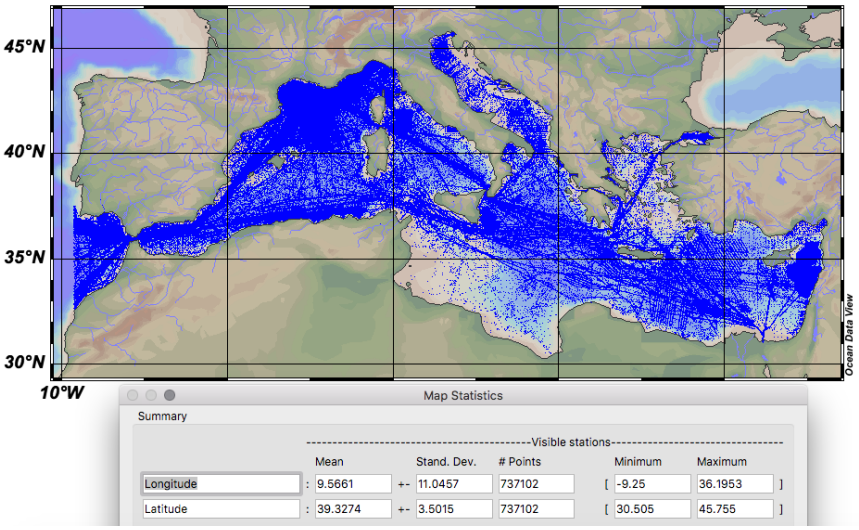
Figure 11 Temporal data distribution over the time periods: (a) 1900-2000 (b) 2001-2017..

par	# stations	%	# points
Total	739784		
T	737102	99,6	41223609
S	671052	90,7	28518660
TS	665388	89,9	28119926

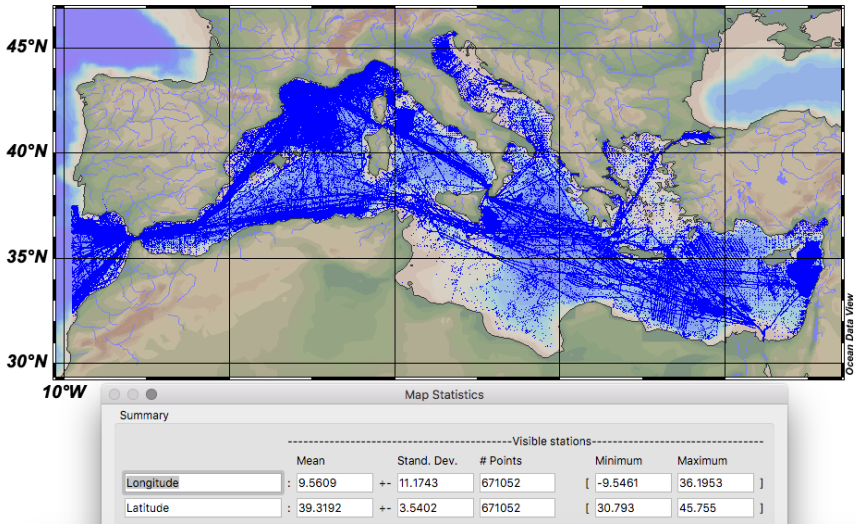
Table 1. SDC_V1 number of stations and samples for Temperature, Salinity and TS couples and relative percentages

Table 1 summarizes the number of observed temperature and salinity stations and stations having both measurements contained in SDC_MED_DATA_TS_V1 collection. Temperature stations represent 99.6% of total stations, salinity stations represent 90.7% of total stations while measurements containing both temperature and salinity are 89.9% In terms of samples it should be noticed the increased difference among temperature and salinity monitoring.

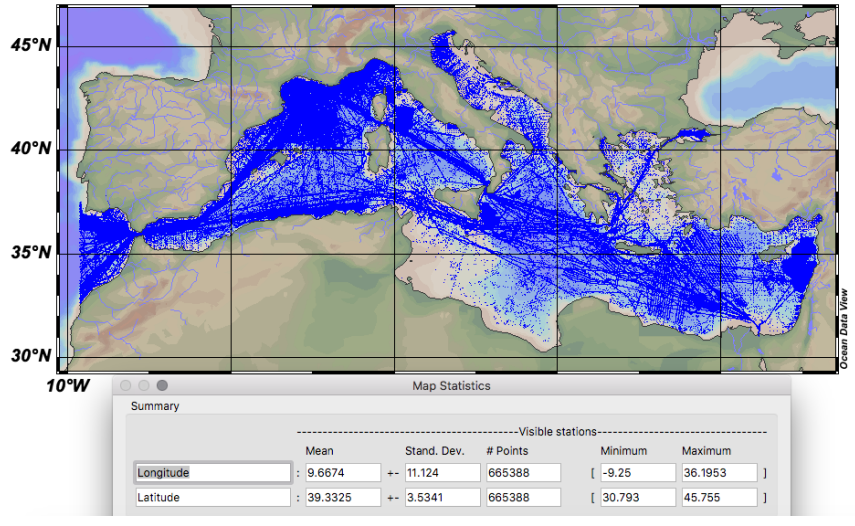
Figure 12 shows Temperature (a), Salinity (b) and TS couples data distribution. Salinity observations are less and sparser than temperature ones. Both maps show the presence of data along ship tracks, along coastal transects, and regular monitoring arrays.



(a)



(b)



(b)

Figure 12. Station distribution map for the Mediterranean Sea 1900-2017: (a) Temperature; (b) Salinity; (c) TS couples.

The statistics related to data quality flags are summarized in Table 2. The percentage of data not checked by data providers (QF=0) ranges from 2.7% of temperature data to 4.5% of salinity data. 97% of temperature and depth records are flagged as good (QF=1) or probably good (QF=2), while for salinity the percentage is equal to 94.6. Bad (QF=4) or probably bad (QF=3) data are less then 1% for the three parameters, and almost absent are samples with flags from 5 to 9.

%	QF=0	QF=1-2	QF=3-4	QF=5-9
Depth	3.0	96.9	0.1	0.0
Temperature	2.7	97.0	0.3	0.0
Salinity	4.5	94.6	0.9	0.0

Table 2. Quality Flag statistics related to depth, Temperature and Salinity parameter expressed in percentages before the QC procedure.

Figure 13 shows the final scatter plots of the good data (QF=1, 2) after QC analysis and Table 3 presented the relative percentages of data quality flags. Data not checked (QF=0) have been all quality assessed and the final quality of the data collection results very high with good data representing more than the 99% of the whole data set.

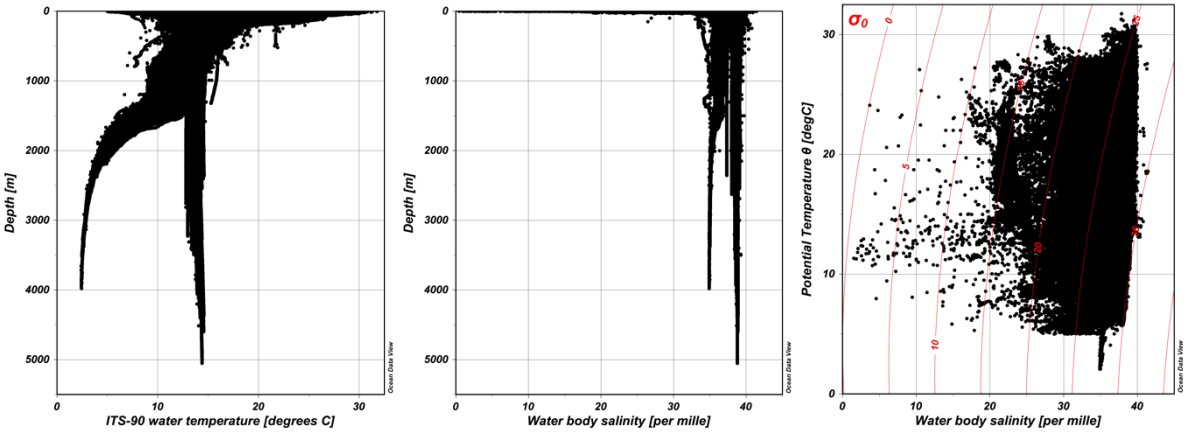


Figure 13. Scatter plots of good (QF=1, 2) observations after QC: (left) temperature versus depth (middle) salinity versus depth and (right) salinity versus potential temperature.

%	QF=0	QF=1-2	QF=3-4	QF=5-9
Depth	0	99.8	0.2	0
Temperature	0	99.8	0.2	0
Salinity	0	99.2	0.8	0

Table 3. Quality Flag statistics related to depth, Temperature and Salinity parameter expressed in percentages after QC analysis.

2.2. Black Sea

Temperature and Salinity Historical Data Collection for the Black Sea contains water body temperature and salinity data (profiles and surface measurements) retrieved from SeaDataNet infrastructure at the end of 2017. It includes non-restricted data belonging to 42 data providers and covers the time period 1868-2017. All data in the collection have been quality controlled according to procedures described in chapter 1.3. The duplicates and bad data (e.g. stations on land, empty depth levels and profiles i.e. those without Temperature and Salinity) were excluded from the collection. Data statistics for the final collection are presented in Table 4.

The collection includes underway data (surface Temperature and Salinity) from two cruises, which trajectories are easily recognized in Figure 14. The underway data were sampled with high frequency (e.g. 1 measurement per minute), therefore the number of data points per cruise is huge. Since every data point has different coordinates and time, in ODV these data points are considered as separate stations. The number of underway stations in the collection is 18563, i.e. stations from 2 datasets represent 13.5% of total. Such large amount of irregularly obtained data introduces bias in data statistics. For better understanding data statistics, they are provided either for the whole collection and excluding underway data.

	Cruises	Stations			Values
		All	Profiles	Underway	
All data	2286	137723	119160	18563	4240346
Temperature	2282	137370	118807	18563	4238207
Salinity	2116	129731	111168	18563	4111531

Table 4. Data statistics for the Black Sea.

Spatial distribution and data density maps of the final dataset are presented at Figure 15.

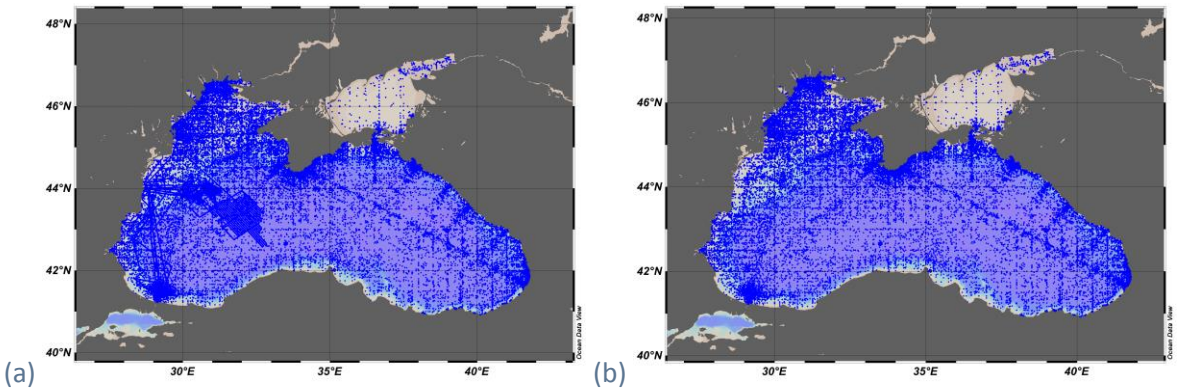
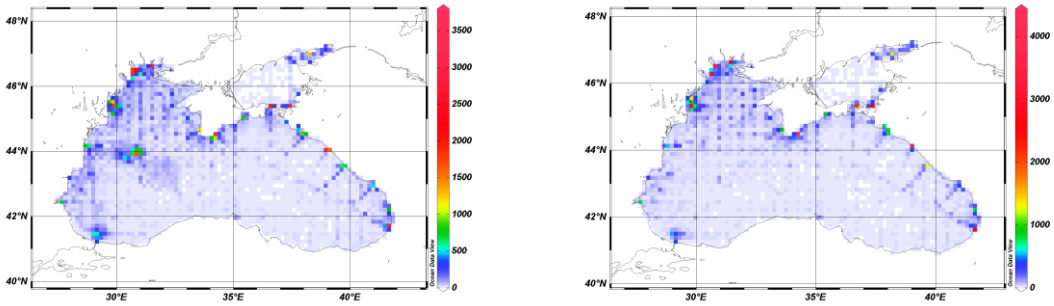


Figure 14. Spatial distribution of observations: all (a) and excluding underway data (b).

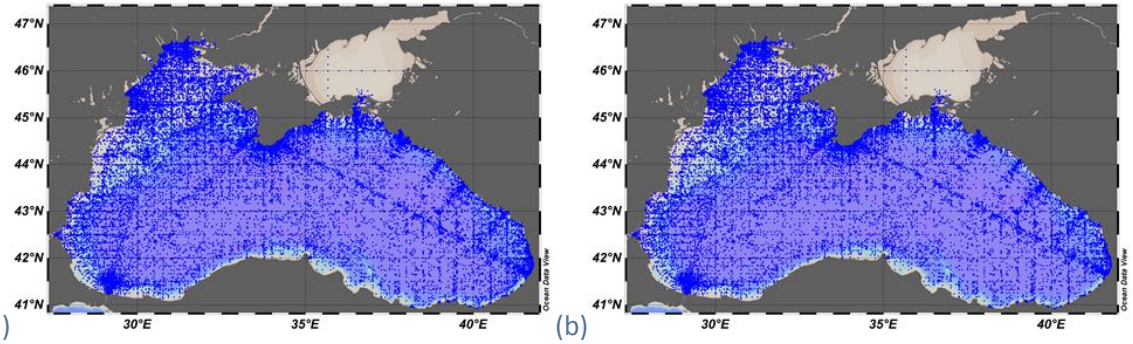


(a) (b)

Figure 15. Data density of observations: all (a) and excluding underway data (b).

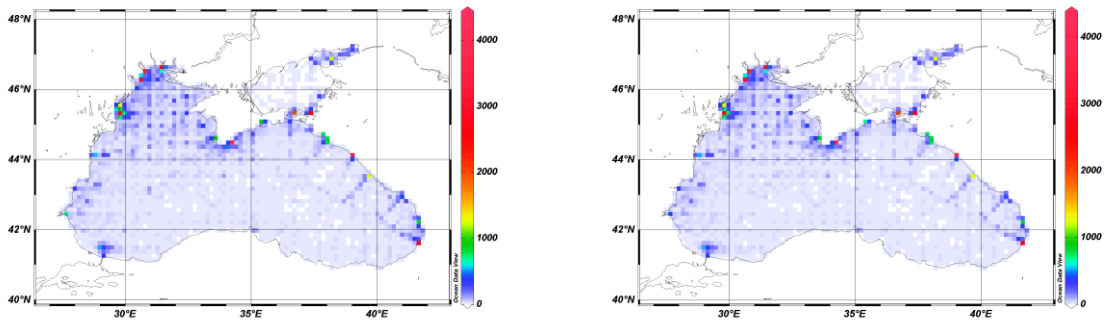
The data distribution in Black Sea is rather uneven (see Figure 14.and Figure 15): the increased concentration of measurements can be observed in areas of intensive navigation and along the standard oceanographic transects, while the interior of the sea, the areas along southern coast and along the central part of the western coast are covered rather poor. About 9% of stations in the collection belong to the Sea of Azov, however most of them come from several coastal stations while the interior of the sea is covered poorly.

The separate maps for spatial distribution and data density of Temperature and Salinity observations are presented at Figure 16 and Figure 17 respectively. The maps for two parameters are practically identical; the small difference in spatial coverage can be noticed only along the eastern coastline and along the north-western shelf.



(a) (b)

Figure 16. Spatial distribution of Temperature (a) and Salinity (b) observations (excluding underway).



(a) (b)

Figure 17. Data density of Temperature (a) and Salinity (b) observations (excluding underway).

Temporal distribution of all observations is presented at Figure 18. Though the first oceanographic measurements in Black Sea date from 1868, the total number of observations before 1955 is rather small, just about 3000. The most intensive oceanographic observations were performed in the Black

Sea during the time period 1970-1995. The peaks in 2002 and in 2015 represent underway data from two cruises: the first one was performed in August – September, 2002, and the second – in August – October, 2015.

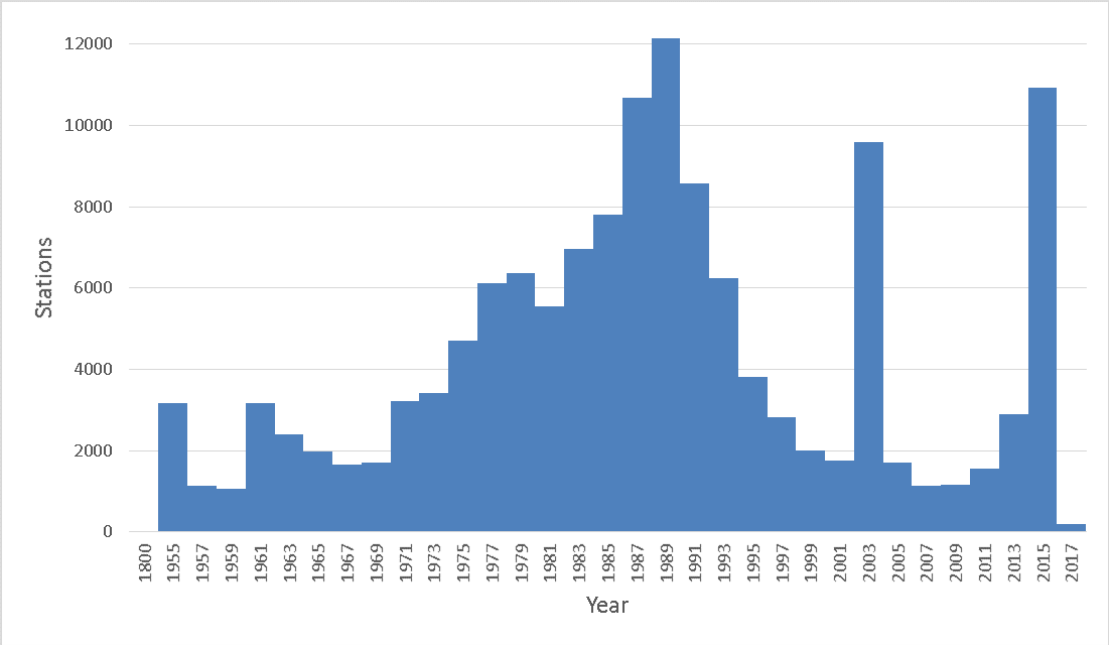


Figure 18 Temporal distribution of observations

Temporal distributions of Temperature and Salinity observations (excluding underway) are presented in Figure 19, Figure 20 and in Figure 21, Figure 22 below. The distributions are practically identical except the small difference in annual distributions in the 1960-s, when there were less Salinity observations compared to Temperature.

The amount of observations from recent years (2015-2017) is very small – just about 300 stations: this might be due to an overall decrease of observations in Black Sea, but mainly to the time lag between sampling and data submission to SeaDataNet. Moreover, recent data might have status “restricted” or “moratorium” and therefore could not be included in the current collection.

Monthly and seasonal distributions, as expected, have dome-like shape having maximum in summer (more observations) and minimum in winter (less observations).

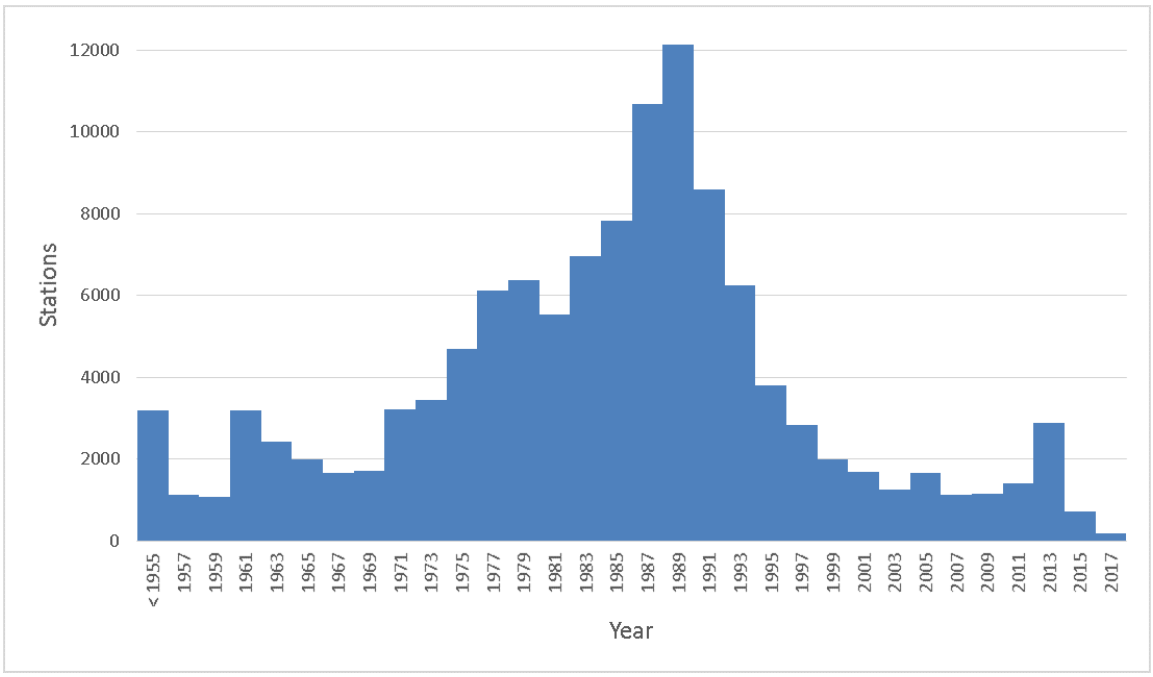
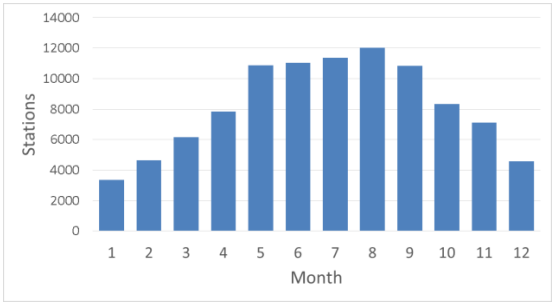


Figure 19. Annual distribution of Temperature observations excluding underway (left bar represents all stations before 1955).

(a)



(b)

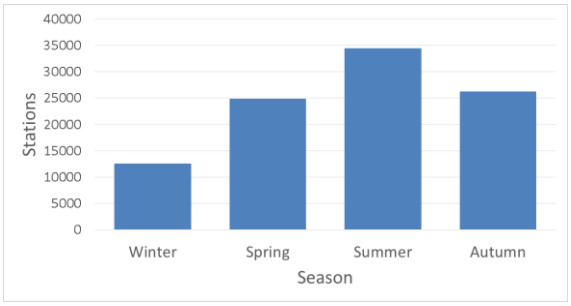


Figure 20. Monthly (a) and seasonnal (b) distribution of Temperature observations (excluding underway).

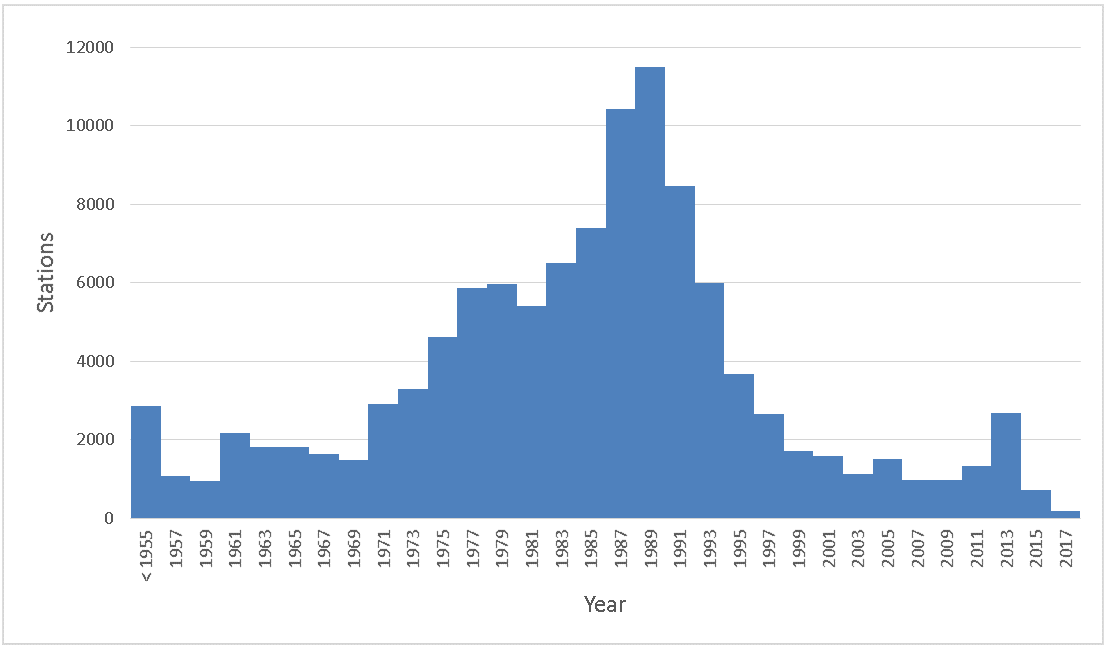
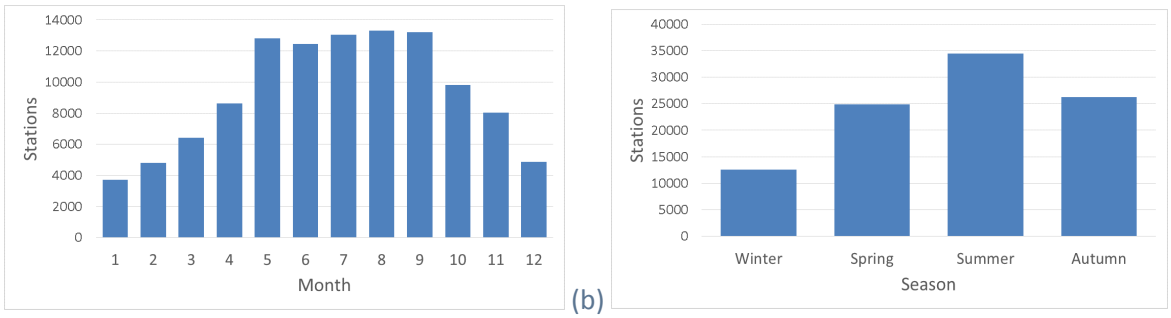


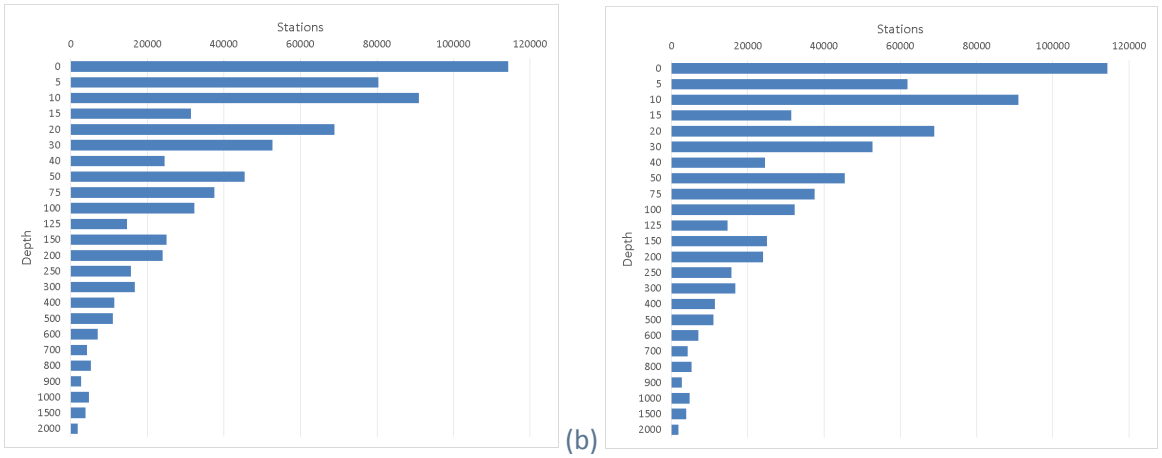
Figure 21. Annual distribution of Salinity observations excluding underway (left bar represents all stations before 1955)



(a) (b)

Figure 22. Monthly (a) and seasonal (b) distribution of Salinity observations (excluding underway)

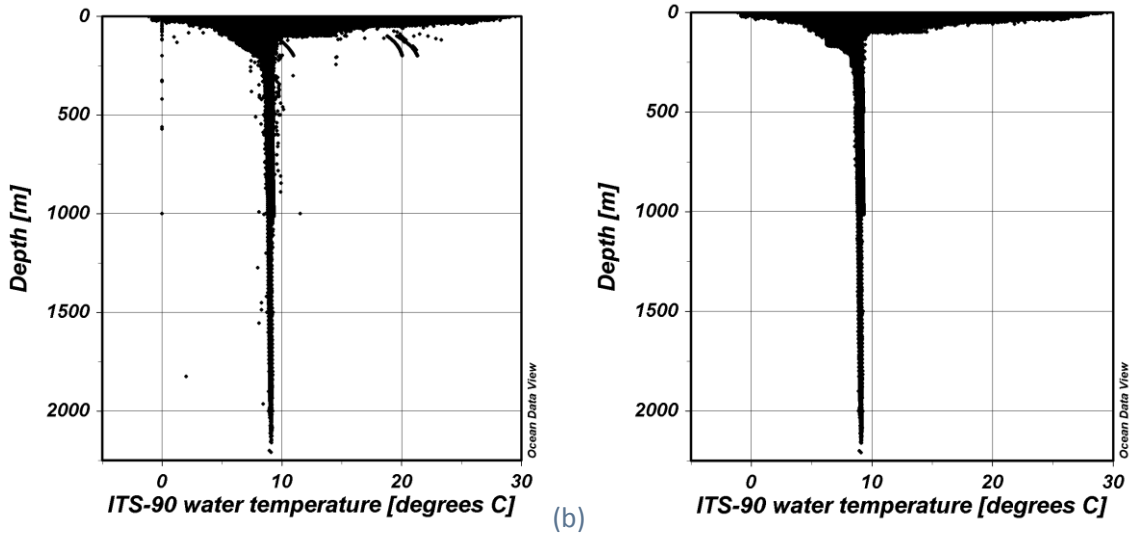
Vertical distribution of observations (Figure 23) shows that data availability drastically decreases with depth. The difference between all data and data without underway appears at 5 m depth level. As the significant part of the collection is bottles data there are also gaps at non-standard depth levels, e.g. at 5, 15, 40, 125 m.



(a) (b)

Figure 23 Vertical distribution of observations: all (a) and excluding underway (b)

The initial data collection contained about 12% of non-controlled data as well as a number of bad data flagged as good that can be seen at scatter plots of Temperature, Salinity and TS diagrams in Figure 24a and Figure 26a). The scatter diagrams for QC-ed data are presented in Figure 24b and Figure 26b respectively for comparison.



(a) (b)

Figure 24. Scatter plots of Temperature before (a) and after (b) quality control.

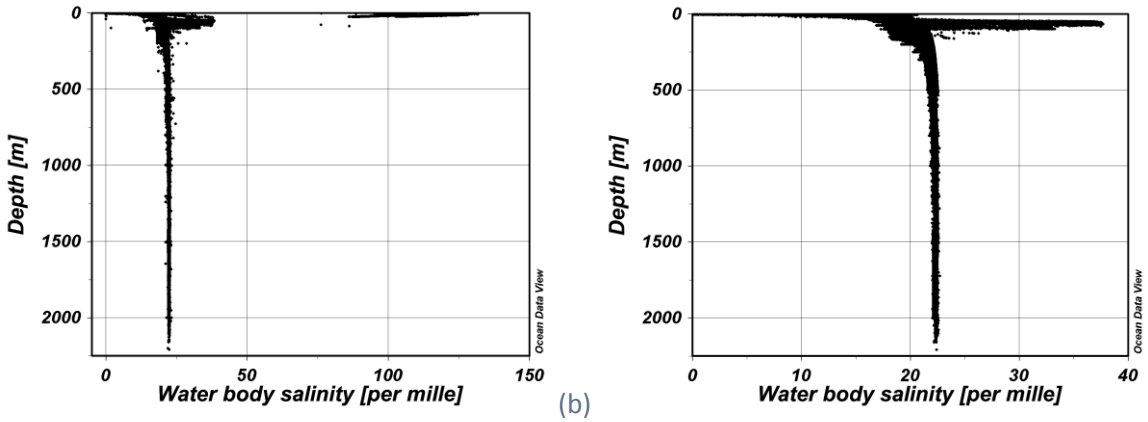


Figure 25. Scatter plots of Salinity before (a) and after (b) quality control.

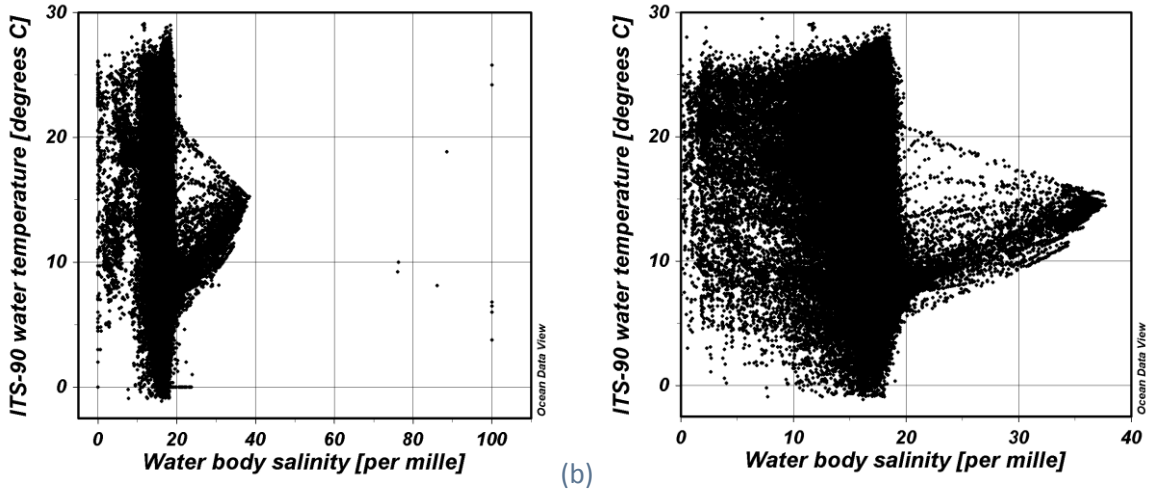


Figure 26. TS diagram plots before (a) and after (b) quality control.

The statistics of the Quality Flaggs (QF) in the initial and final collections are presented in Table 5 and Table 6 respectively. About 96% of data in the final collection have QF=1 ("good"). Another 2-3% of data are flagged as "probably good" (QF=2). Thus in total about 98-99% of the collection data can be considered as valid for usage in different studies and applications.

QF	0	1	2	3	4	Total
	not checked	good	probably good	probably bad	bad	
Depth	500240	3575857	72266	91973	10	4240346
	11.80%	84.33%	1.70%	2.17%	0.00%	
Temperature	491550	3470913	195737	75114	4893	4238207
	11.60%	81.90%	4.62%	1.77%	0.12%	
Salinity	491915	3448470	118380	48207	4559	4111531
	11.96%	83.87%	2.88%	1.17%	0.11%	

Table 5. Data Quality flagging statistics in initial collection.

QF	0	1	2	3	4	Total
	0	4073158	131197	35702	289	
Depth	0.00%	96.06%	3.09%	0.84%	0.01%	
	0	4045723	131358	56030	5096	4238207
Temperature	0.00%	95.46%	3.10%	1.32%	0.12%	
	0	3956660	74707	71960	8204	4111531
Salinity	0.00%	96.23%	1.82%	1.75%	0.20%	
	0	4073158	131197	35702	289	4240346

Table 6. Data quality flagging statistics in final QC-ed collection.

Special cases of QC:

1. The big dataset "AQUALOG Moored Profiler" of P.P.Shirshov Institute of Oceanology, RAS contains **raw data**, i.e. data that did not undergo typical CTD post processing such as filtering, alignment, loop-edit, bin-averaging etc. According to the description of Aqualog, "the profiler makes repeated round trips up and down a taut mooring wire between the subsurface flotation and the anchor". Judging from the data the profiler stays for a while at the end points of its trips generating significant amount of data. When imported to ODV these data are automatically sorted by depth. As a result the data points are getting mixed up, and, consequently, the test of profile for stability fails. Nevertheless, these data were not flagged as bad except the obvious outliers.

2. A number of profiles originating from the Institute of Fishery Resources, Bulgaria also contain raw data for which stability test fails. As in previous case the data were not flagged as bad except the obvious outliers.

The previous version of the product was released in framework of the SeaDataNet II project and it is available at SEXTANT Catalogue (http://sextant.ifremer.fr/en/web/seadatanet) under the name "Black Sea - Temperature and salinity observation collection V2" (SDN_V2, http://doi.org/10.12770/227e9f7b-ddfc-4004-b0e5-f4785d36d43f). Compared to SDN_V2 collection there is an overall significant increase in terms of: cruises, stations and data (Table 7). The total increase in number of stations is about 43%, but excluding the underway data it gives 23.5% of increase. Cruises and values statistics are instead not much affected by underway data. The significant increase of data values (57%) is instead mainly due to the additional CTD profiles with high vertical resolution: 1 m, 0.1 m or even raw data.

Cruises			Stations			Samples (data)		
SDN2_V2	SDC_V1	±%	SDN2_V2	SDC_V1	±%	SDN2_V2	SDC_V1	±%
1723	2284	+32.6%	96487	137723	42.7%	2696215	4240346	+57.2%
Excluding underway data								
1723	2282	+32.5%	96487	119160	23.5%	2696215	4221783	+56.6%

Table 7. Data statistics of previous (SDN2_V2) and current (SDC_V1) versions of collections.

The comparative statistics on Quality Flags of the two collections (SDC V1 versus SDN V2) are very similar, therefore the comparison is not provided hereby - please see the Table 6 instead.

2.3. Artic Sea

The Arctic Seas historical unrestricted data set contains about 7312286 profiles (24203161 values) for both salinity and temperature and covers the years 1903-2017 from 1956 cruises. The collection includes open access data retrieved from the SeaDataNet II infrastructure at the end of 2017.

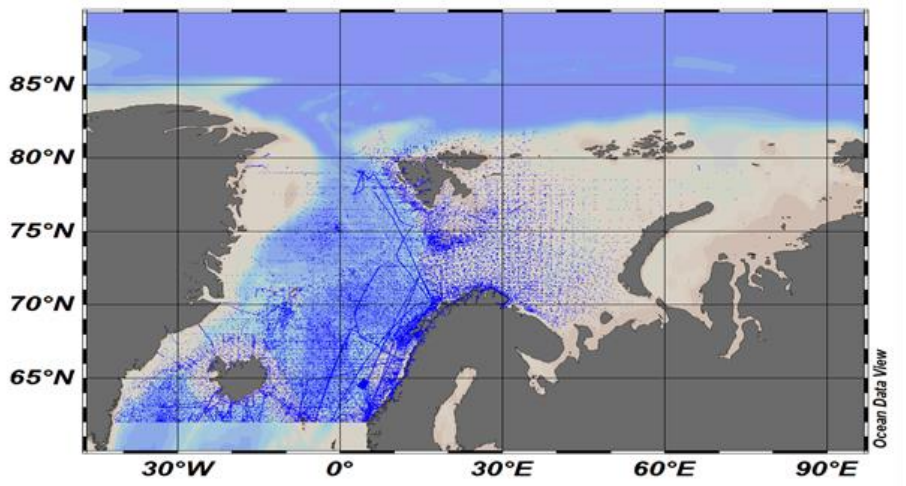


Figure 27. All data points in the Full domain, unrestricted profile data.

Most of the data are from profiles, dots in Figure 27, but there are also some high-resolution data that are from trajectories (ferry box system), which look like solid lines in Figure 27. Figure 28 shows the data density map with data availability. The overall geographical coverage is good over the last 20 years with some exceptions, especially along the Greenland coast, northern Greenland shelf, and coasts of Russia.

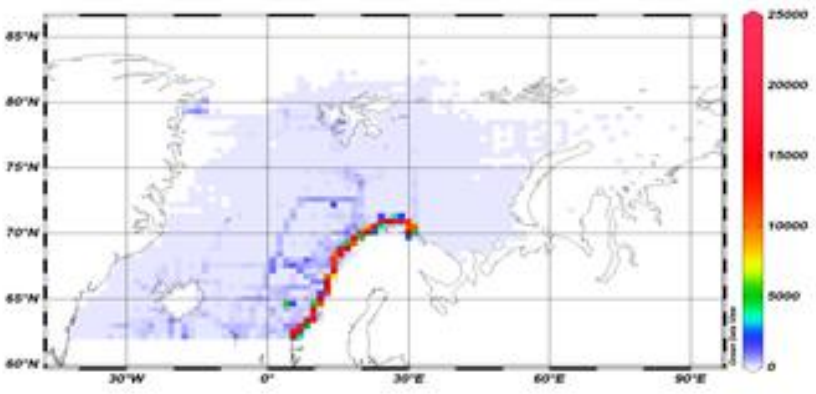


Figure 28. Spatial data density map.

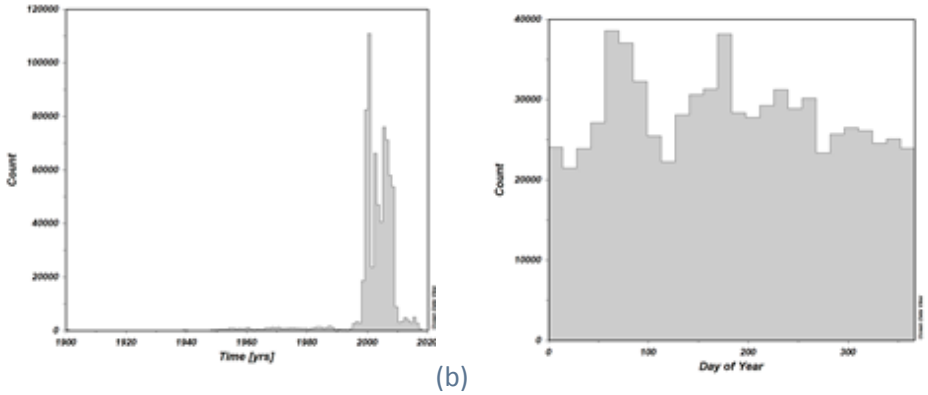
Most of the measurements contain both temperature and salinity data, Table 8. 16.6 % of the stations have Temperature and Salinity, and 94% of the values.

	Number of Stations	Number of values	% of total
Inventory	731286	24203161	
Depths	731286	24202760	99.99

Temperature	731142	23570227	97.4
Salinity	121770	22834692	94.4
T and S	121673	22750100	94.0

Table 8. Data statistics for temperature and salinity in the entire Arctic data set 1935-2017.

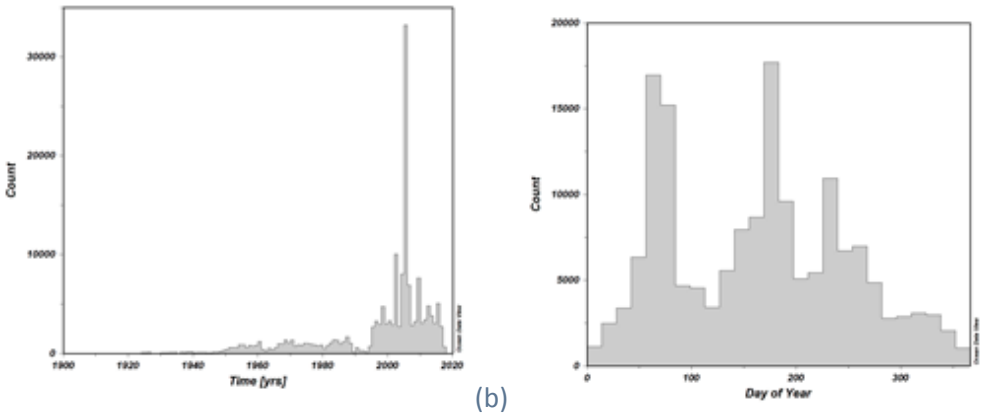
Annual (Figure 29a) and seasonal (Figure 29b) time distributions show low number of data before 2000. The data distribution is somewhat even seasonally, due to the large amount of ferry box data.



(a) (b)

Figure 29. Temporal distribution (a) and seasonal distribution over the year (b) for the entire dataset.

Annual (Figure 30a) and seasonal (Figure 30b) distributions without ferry box data show higher number profile data in spring, summer and autumn.



(a) (b)

Figure 30. Temporal distribution (a) and seasonal distribution (b) without ferry box data.

Ferry box data

Figure 31 shows the geographical distribution of ferry box data and Figure 32 the annual and seasonal data distribution in time. The geographical distribution of the ferry box data is close to the Norwegian coast. The seasonal distribution of the ferry box data is rather good. The ferry box data correspond to 568877 stations and 568877values.

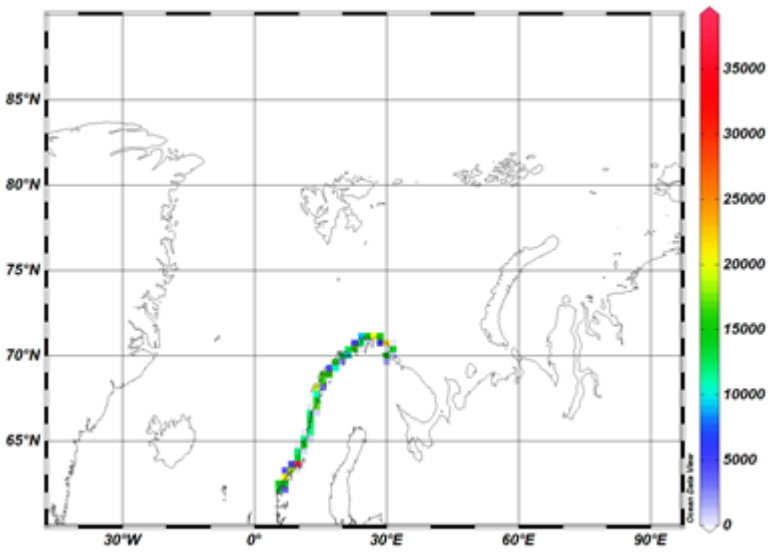
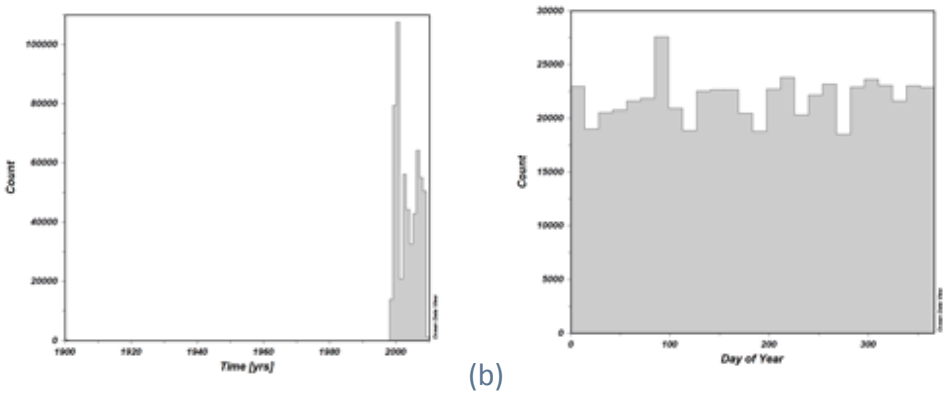


Figure 31. Ferry box data distribution map.



(a) (b)

Figure 32. Annual distribution (a) and seasonal distribution (b) for the ferry box data sub-set.

Temperature and salinity distribution

The dataset has been split in three groups, temperature data (Figure 33) salinity data (Figure 34) and both temperature and salinity data together (Figure 35)

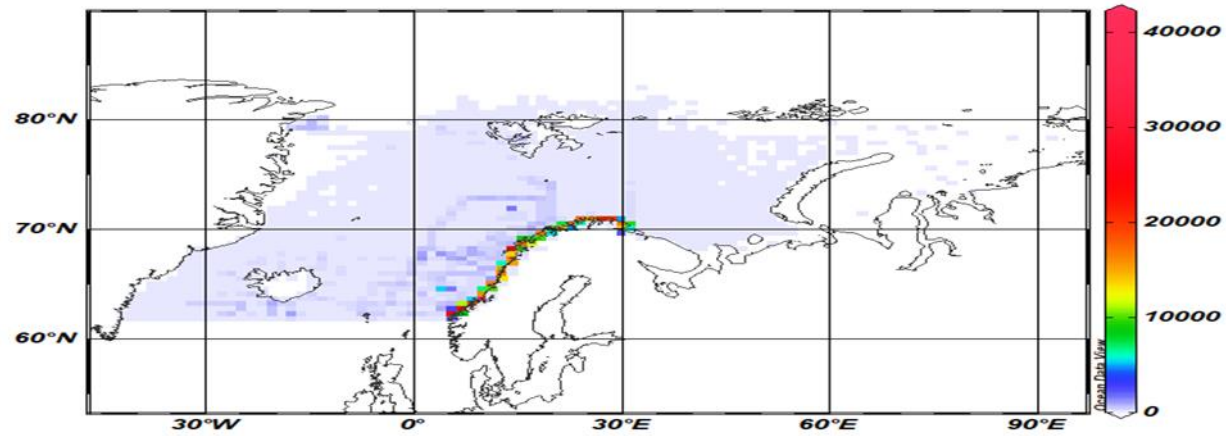


Figure 33. Distribution of temperature data, 731142 stations.

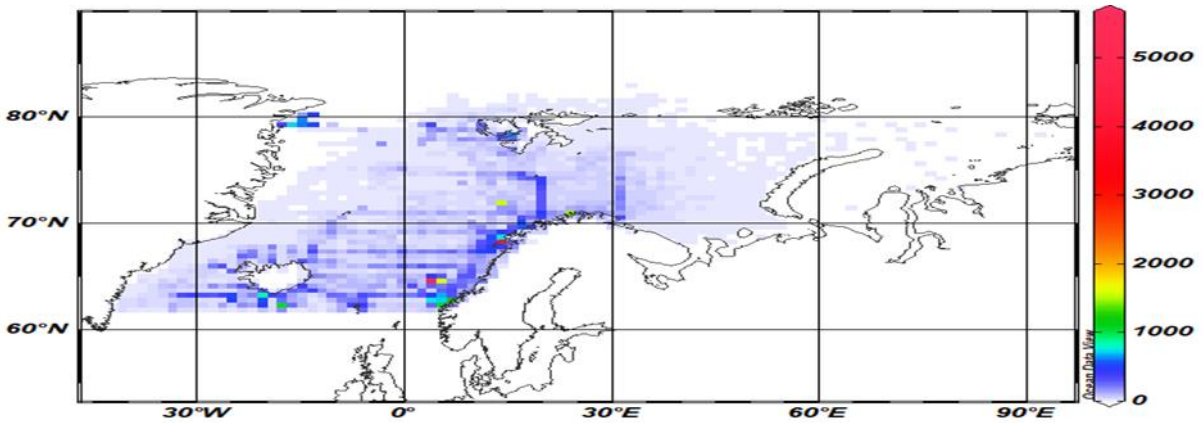


Figure 34. Distribution of salinity data, 121770 stations.

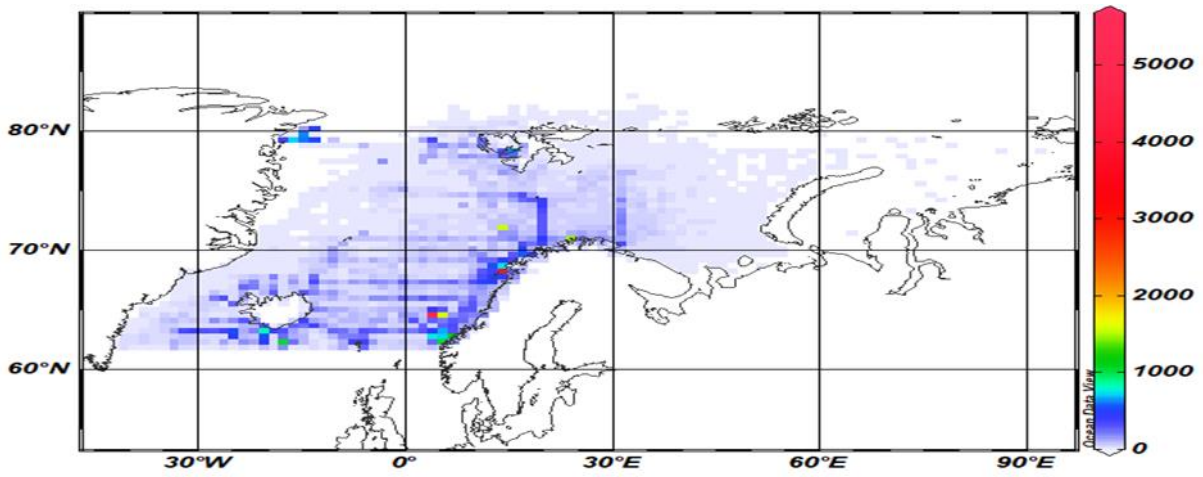
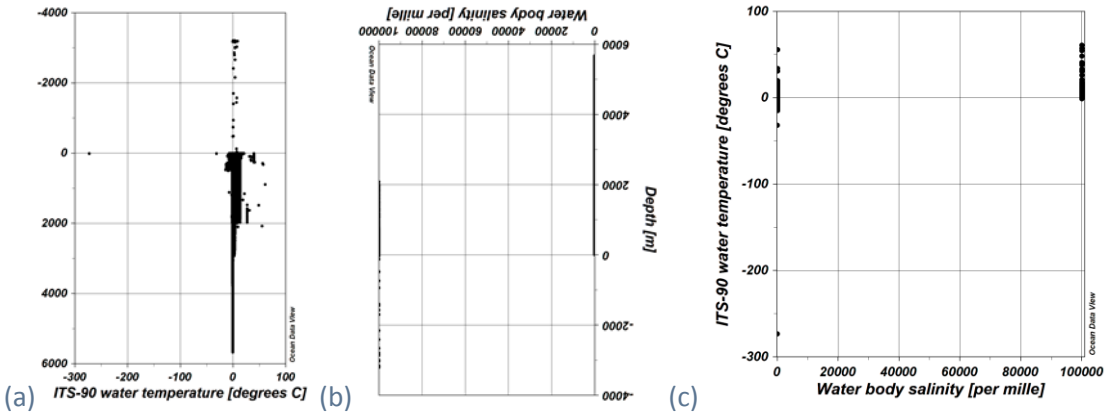


Figure 35. Distribution of temperature and salinity data, 121763 stations.

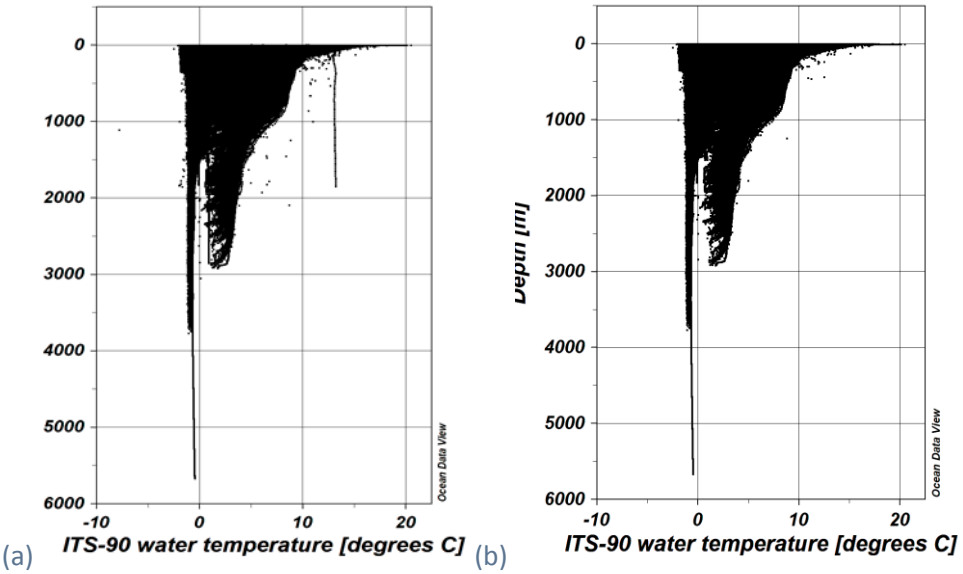
The Figure 36 shows the scatter plots for all the observations before correction.



(a) (b) (c)

Figure 36. Scatter plots distribution for (a) temperature, (b) salinity, and (c) temperature/salinity.

The Figure 37 shows the scatter plot of temperature (QF=1 and QF=2) before (a) and after correction (b).



(a) (b)

Figure 37. QF 1-2 Depth/T scatter (a) before correction (b) after correction.

The Figure 38 shows the scatter plot of salinity (QF=1 and QF=2) before (a) and after correction (b).

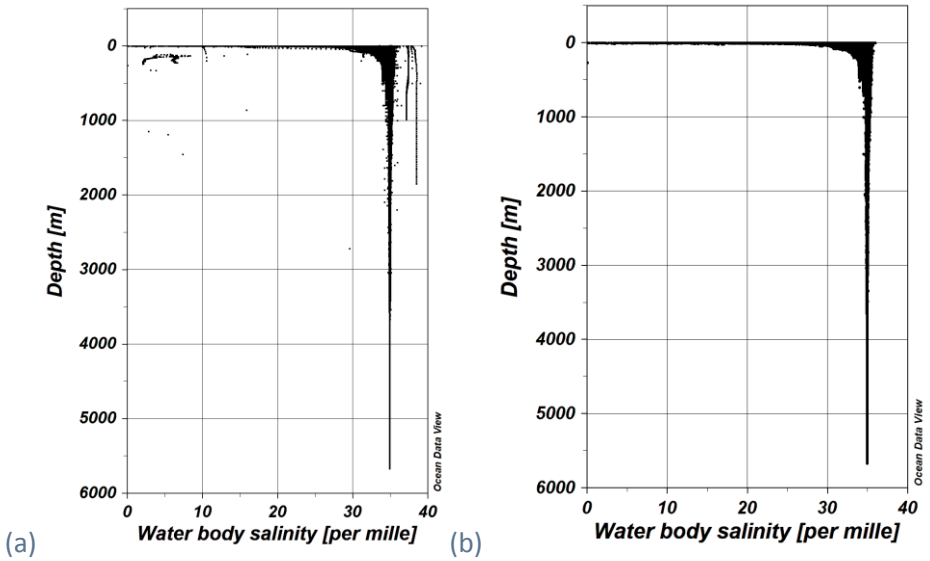


Figure 38. QF 1-2 Depth/S scatter (a) before correction (b) after correction.

The Figure 39 shows the scatter plot of temperature/salinity (QF=1 and QF=2) before (a) and after correction (b).

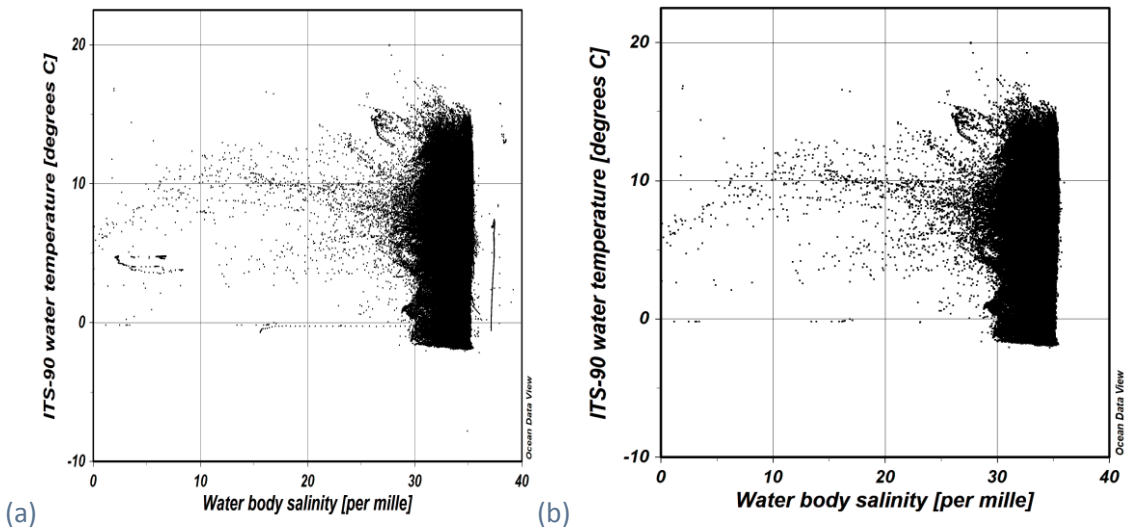
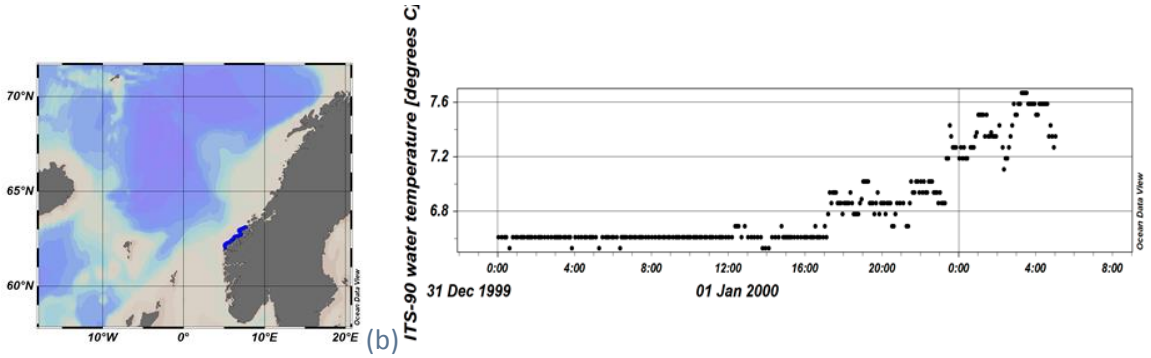


Figure 39. QF 1-2 T/S scatter (a) before correction (b) after correction.

Ferry box data with wrong time

272 stations in collection year 1900 were edited to 2000 after inspection of original data. Subset with 272 stations was written to spreadsheet file, stations were deleted, year 1900 replaced in spreadsheet file with year 2000, and spreadsheet file were read back in collection (Cruise map with track in 1 January 2000 in Figure 40a, and time plot of temperature after QC in Figure 40b).



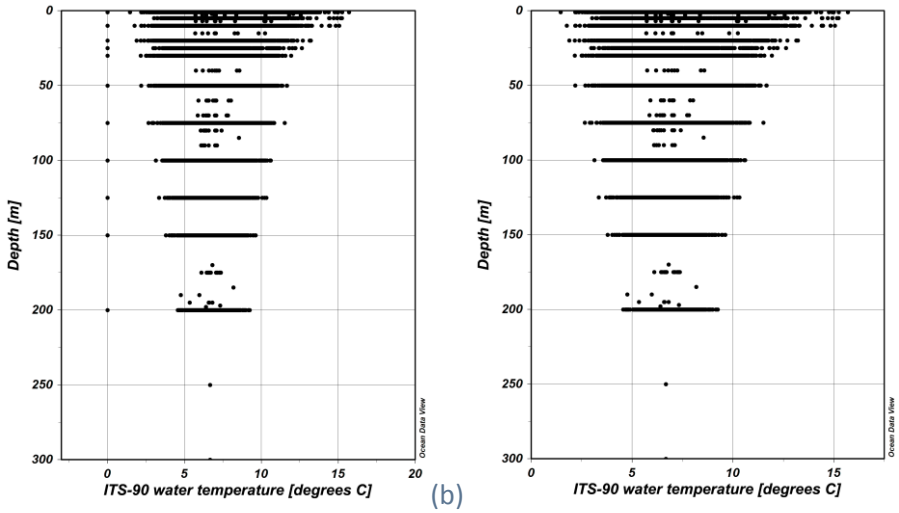
(a) (b)

Figure 40. (a) Map Ferry box 1. Jan 2000 (b) Ferry box temperature time plot after QC.

Stations with all values flagged with QF=0 : 15 cruises were found with 22 stations with all values flagged with QF=0. All values were flagged with QF=1 after stability check, and range check, on original IMR data.

Fixed stations along Norwegian coast where inspection identified 0 degree. Temperature values flagged good instead of original space for no value. 4 fixed stations with 163 Zero temperature values were deleted and ODV flagged QF9.

The Figure 41 shows the fixed station Eggum, the Depth/temperature scatter before and after correction.



(a) (b)

Figure 41. Fixed station Eggum Depth/temperature (a) before QC and (b) after QC.

3 IMR cruises have been found with all deep profiles with only positive temperatures. Profiles were edited in spreadsheet file and read back in collection, and flagged with QF 1.

Stations have been found with negative depths: 123 Temperature values and 66 salinity values, mainly Argo floats at the top of the profiles. All looks good, most flagged as good, while some are flagged probably bad. Depths are negative and flagged good. They are in the collection with initial Quality Flag; they fall out by use of range 0-6000 meters in Depth.

4 stations with deeper observations than bottom depth at given position have been found. Position can be wrong or profiles are remembered from earlier sub bottom depth profiles for some Argo floats in the Barents Sea. All values were flagged bad data, as it is uncertain what caused the problem.

Statistics of the quality flagging in initial and QC edited collection are presented in Table 9 for initial collection, and in Table 10 for QC edited collection.

Quality flag	0	1	2	3	4	5	Total
Depth	156086	24020343	18259	7992	80	0	24202760
% of total	0.64%	99.24%	0.08%	0.03%	0.00%	0.00%	
Temperature	154704	23331579	2695	14614	10755	55756	23570227
% of total	0.66%	98.99%	0.01%	0.06%	0.05%	0.27%	
Salinity	154794	22458390	3226	82258	21958	113879	22834692
% of total	0.68%	98.35%	0.01%	0.36%	0.10%	0.50%	

Table 9 Data Quality flags in initial collection.

Quality flag	0	1	2	3	4	5	Total
Depth	0	24176399	18260	7992	109	0	24202760
% of total	0.00%	99.89%	0.08%	0.03%	0.00%	0.00%	
Temperature	0	23483059	3094	14755	13788	55244	23570064
% of total	0.00%	99.63%	0.01%	0.06%	0.06%	0.23%	
Salinity	0	22492768	2866	194303	31410	113156	22834691
% of total	0.00%	98.50%	0.01%	0.85%	0.14%	0.50%	

Table 10 Data Quality flags in QC edited collection.

The reduction in total number of temperature values is due to replacing 0 degrees Temperature values, were original data, were no values (space) with wrong flags QF=1. ODV puts automatic flag QF 9, when values are replaced by space.

The previous version of the product was released in framework of the SeaDataNet II project and it is available at SEXTANT Catalogue (http://sextant.ifremer.fr/en/web/seadatanet) under the name "Artic Sea - Temperature and salinity observation collection V2" (SDN_V2, http://dx.doi.org/10.12770/f080166b-0632-4de2-85df-97829d56eabf). Compared to SDN_V2 collection, there is an increase of 23% in data samples (Table 11). Period is extended from 2012 (SDN_V2) to 2017 (SDC_V1). New cruises from the period 2013 to 2017 and some older, increase number of cruises by 82%. There is a relative high increase in number of stations due to more ferry box data where each data sample has a unique station. Glider data (Unmanned self-propelled Vehicle) at the Greenland shelf near the Fram strait have been reduced in SDC_V1.

	SDN2_V2	SDC_V1	% of increase
Cruises	1075	1956	82%
Stations	266291	731286	174.6%
Data samples	19681474	24203161	23%

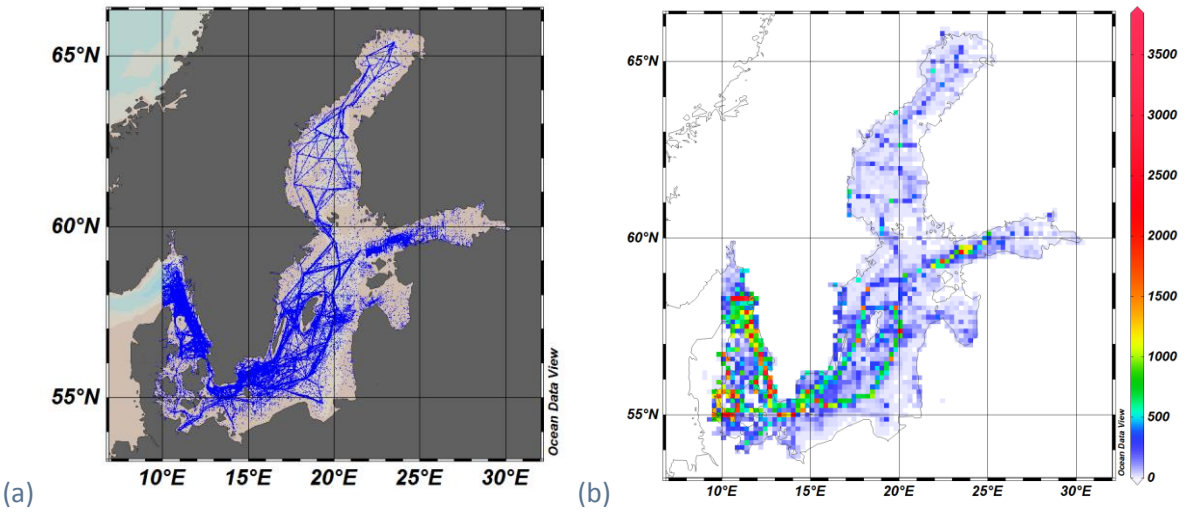
Table 11. Collection SDC_V1 compared to collection SDN_V2.

2.4. Baltic Sea

Baltic Sea historical data set contains about 13700000 values for both salinity and temperature and covers the years 1900-2017. The collection includes open access data retrieved from the SeaDataNet infrastructure at the end of 2017. The majority of all measurements contain both temperature and salinity data, Table 12. Most of the data are from profiles, dots in Figure 42(a), but there are also some high resolution data that are from trajectories (ferrybox system), looks like solid lines in Figure 42(a). Figure 42(b) shows the data density map with data availability. The overall geographical coverage is good with some exceptions, especially along some of the coasts.

	Number of values	% of total
Depths	14069156	
Temperature	13714778	97.5
Salinity	13740926	97.7
T and S	13514087	96.1

Table 12. Number of values for temperature and salinity in the entire data set for the Baltic Sea.



(a) (b)

Figure 42. Map with all data, dots are profiles and what appears to be lines are high resolution ferry box data (a). To the right a data density plot showing where most values have been sampled (b).

Time distribution in Figure 43(a) is low for the first 60 years, it increases somewhat after 1960 until around 1990 or late 1980s where it increases again and is somewhat constant at high levels. In the latest years there is a decrease in data which is caused by a natural time lag between sampling and until data becomes available in the SeaDataNet infrastructure. Seasonal distribution shows a good spread of data during the whole year, Figure 43(b)

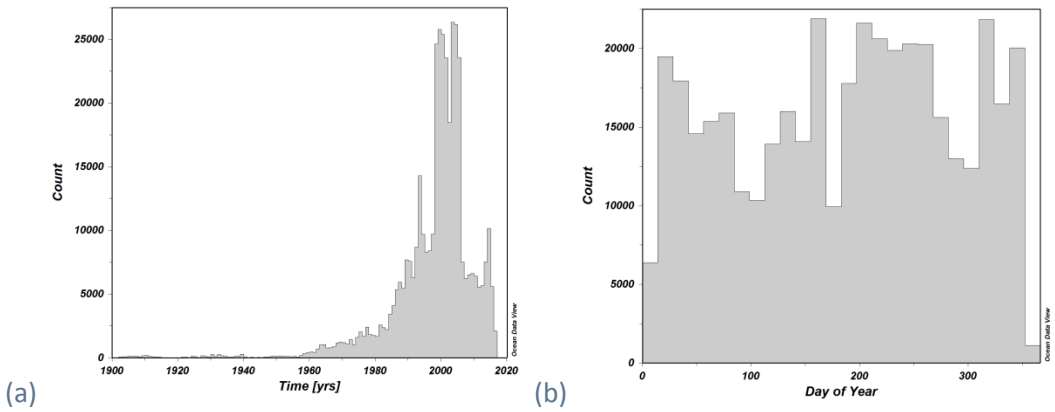


Figure 43. Temporal distribution (a) and seasonal distribution over the year (b) for the entire data set.

Ferry box data with high resolution in space and time makes some statistical figures skewed, a good example of this is the station definition in ODV. In the collection the total number of stations is 407456 of which the ferry box data stands for 145841, 35.8% of all stations. However when comparing number of actual values you get a better picture. The total number of values is 14069156 of which 145841 are from ferry box, 1.0%. If considering the upper 5 metres only (0-5 metres); the ferry box data stands for about 6.4% of all the data.

To get a better understanding of the data distribution, data have been split in two parts; profiles and ferry box data. Distribution plots have been created separately for each part. Figure 44 shows the geographical distribution of ferry box data and Figure 45 the data distribution in time. The geographical and seasonal distribution of the ferry box data is rather good, but the data originates from just a few years so the time coverage is scarce.

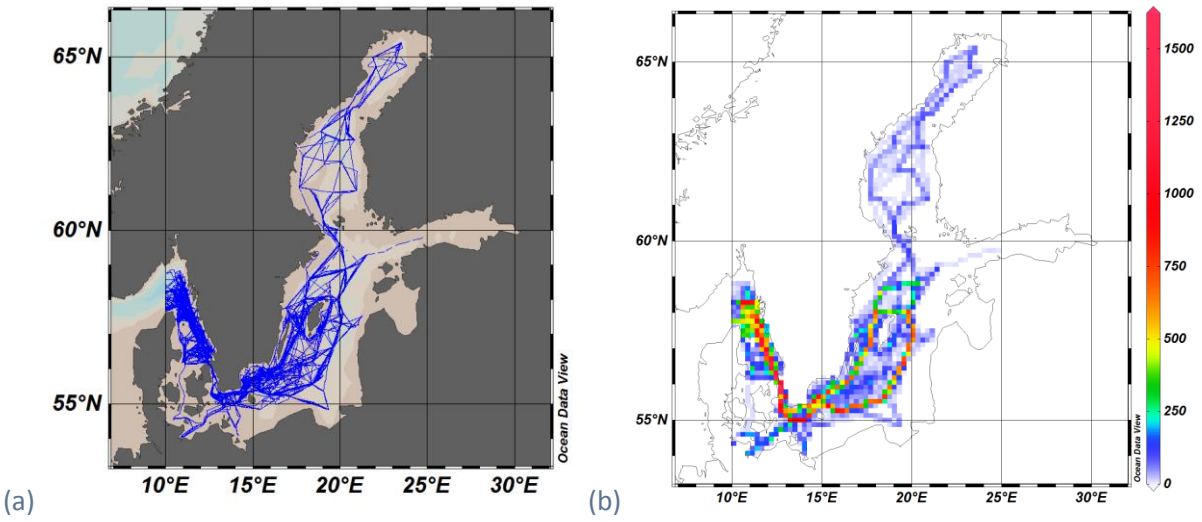


Figure 44. Map of ferry box data (a) and data density map (b) showing where most measurements have been sampled.

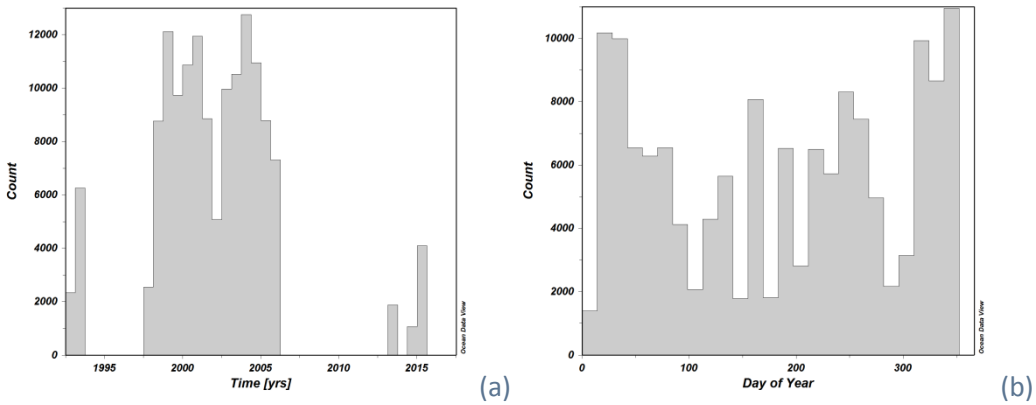


Figure 45. Temporal distribution (a) and seasonal distribution over the year (b) for the ferry box data.

The data distribution of the profile data, both geographical and temporal, can be seen in Figure 46 and Figure 47. The data coverage is best in the Skagerrak and the Kattegat, also decent in the Baltic Proper and the Gulf of Finland. Worst coverage is found in the Gulf of Bothnia, in particular in eastern part and close to the coast.

Comparing to the historical data set constructed in SeaDataNet II (SDN2_V2) which in the Baltic covered the years 1900–2014 there is an increase from around 11 million values to 13.4 million values, almost an **increase by 22% for both temperature and salinity**.

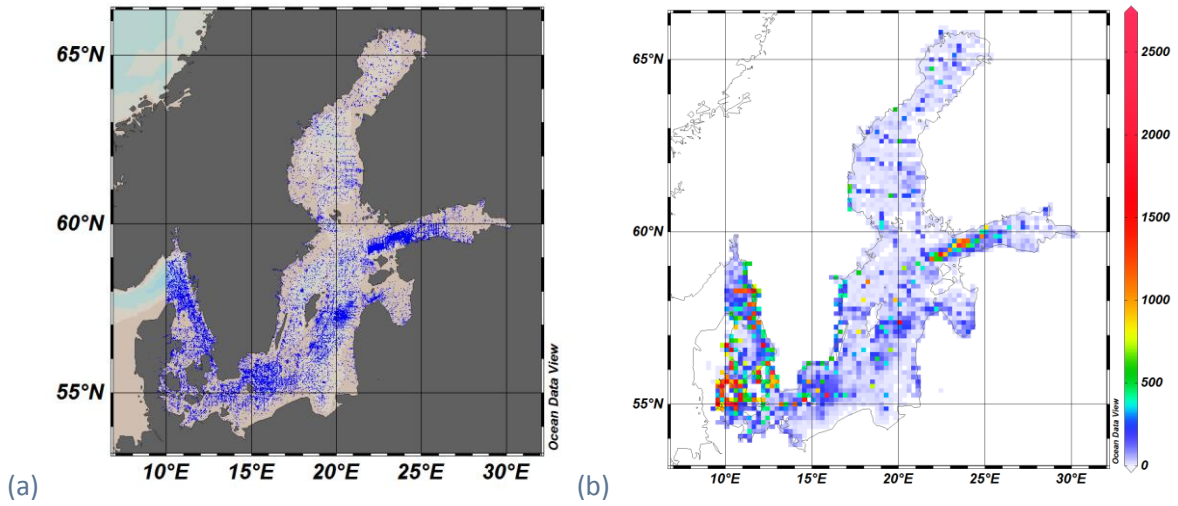


Figure 46. (a) Map with all profile data; (b) data density map showing where most measurements have been sampled.

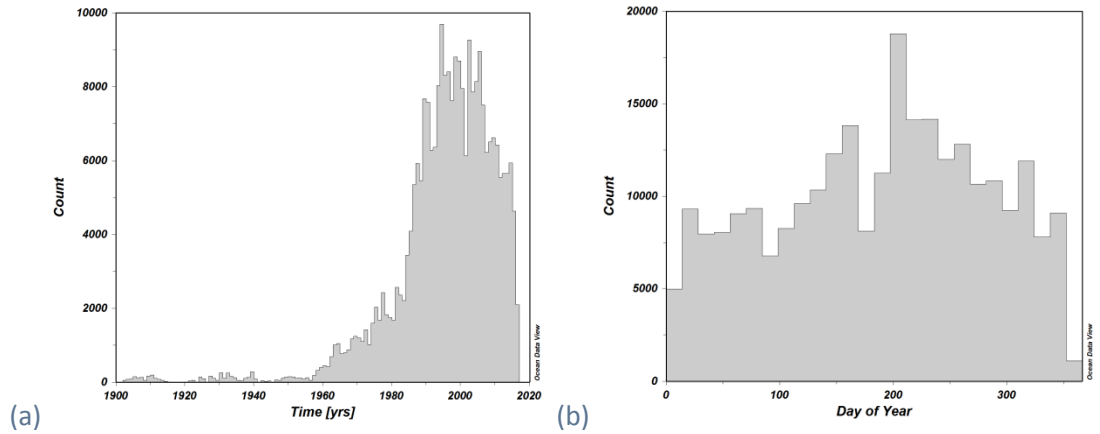


Figure 47. Temporal distribution (a) and seasonal distribution over the year (b) for the profile data.

All data have been quality controlled according to chapter 1.3 (QC best practices). Around 10400 salinity values and 3700 temperature values were flagged as suspicious/bad, flag 3 and 4. This is about 0.076% of the total amount of salinity values and 0.027% of the total amount of temperature values, further details can be seen below in Table 13.

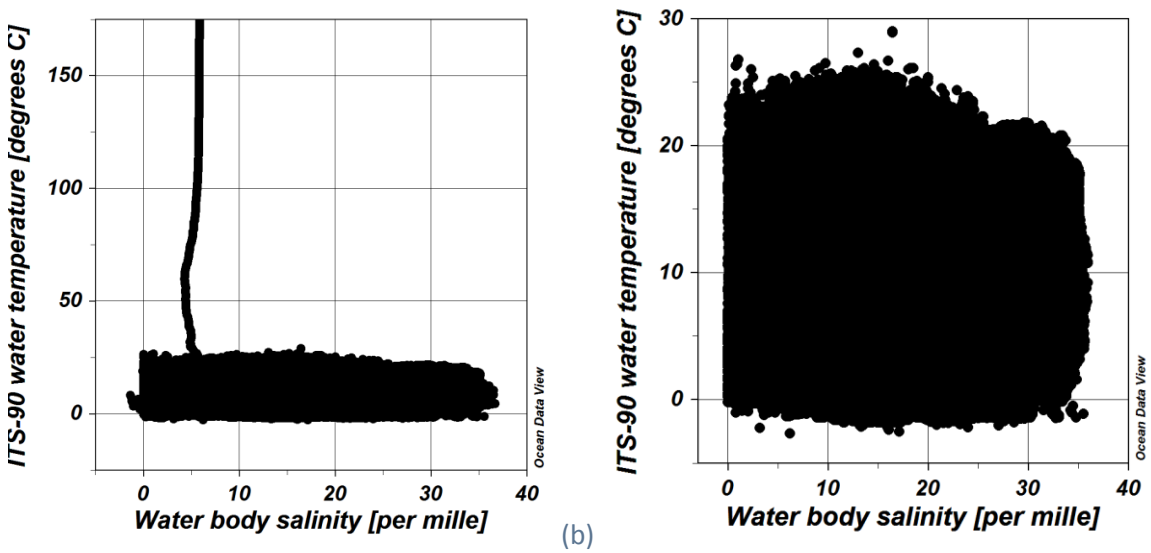
The suspicious values consist mainly of spikes, outliers and unstable density profiles, but there are also some other problems:

- 35 measurements appear to be on land.
- 1 measurement appears to be too deep, probably wrong position.
- 3 measurements have confirmed wrong position.
- 1 measurement has depth as temperature and temperature as salinity, probably a shift of columns in the original data file.
- 17 values have negative depths.
- 2540 measurements lack both temperature and salinity data, probably wrongly having the P02 terms for these parameters in the metadata (CDI file) when there actually are no data present in the data file.
- at least 280 measurements contain several stations/measurements in the same original data file, with different values and different positions.

Errors found in the previous project SeaDataNet 2 were also checked, most of them have been corrected but there are still some measurements that not yet have been corrected. These will be included in the quality control feedback that will be sent to the different SeaDataCloud partners.

Quality flag	0	1	2	3	4	5	6	7	8	A	Total
Salinity	67238	11025622	2642323	10250	126	1105	437	0	4234	0	13740926
% of total	0.5	80.2	19.2	<0.1	<0.1	<0.1	<0.1	0	<0.1	0	
Temperature	65974	11001912	2642906	3599	96	24	0	0	4205	0	13714778
% of total	0.5	80.2	19.3	<0.1	<0.1	<0.1	0	0	<0.1	0	

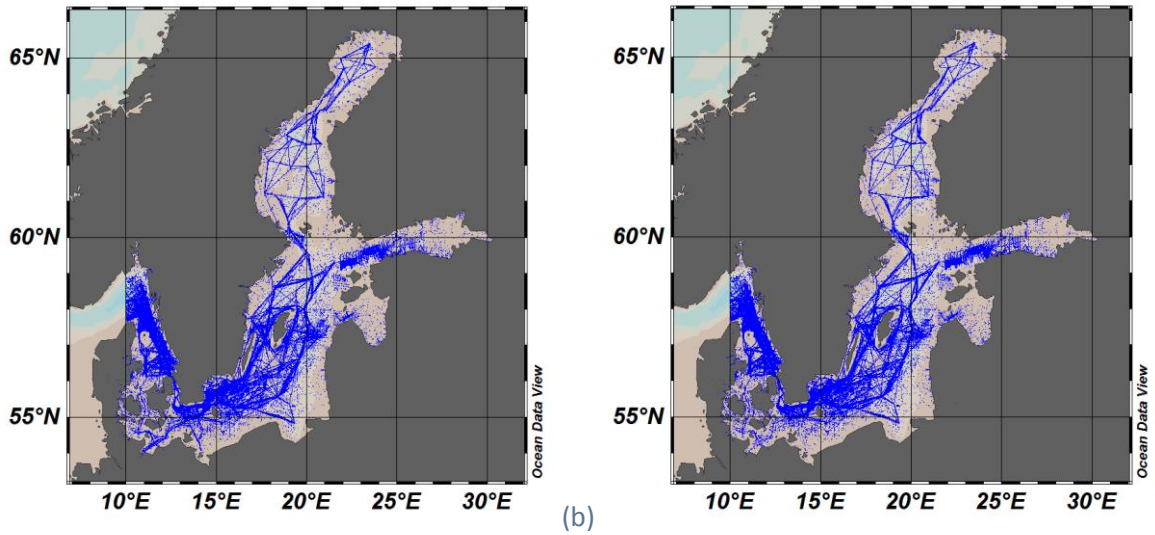
Table 13. Number of values for each quality flag for the entire dataset, 1900-2017.



(a) (b)

Figure 48. TS diagram showing data before (a) and after (b) quality control.

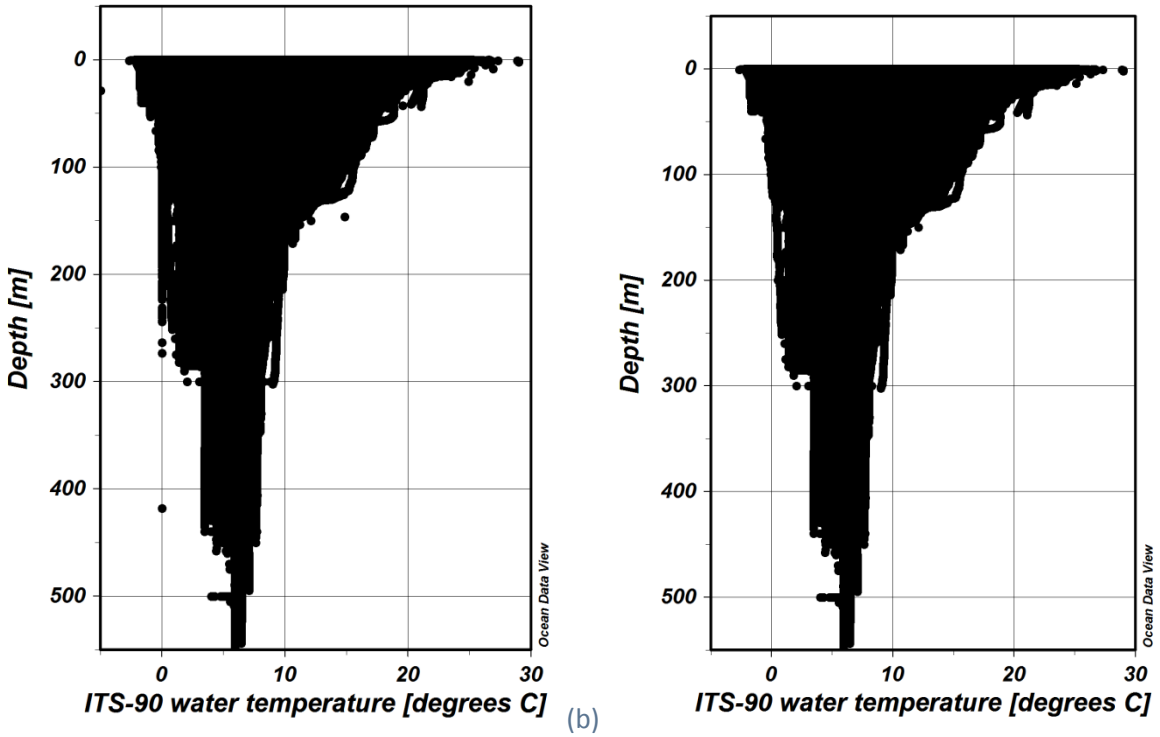
Figure 49presents T and S data distributions after QC where suspicious data have been removed.



(a) (b)

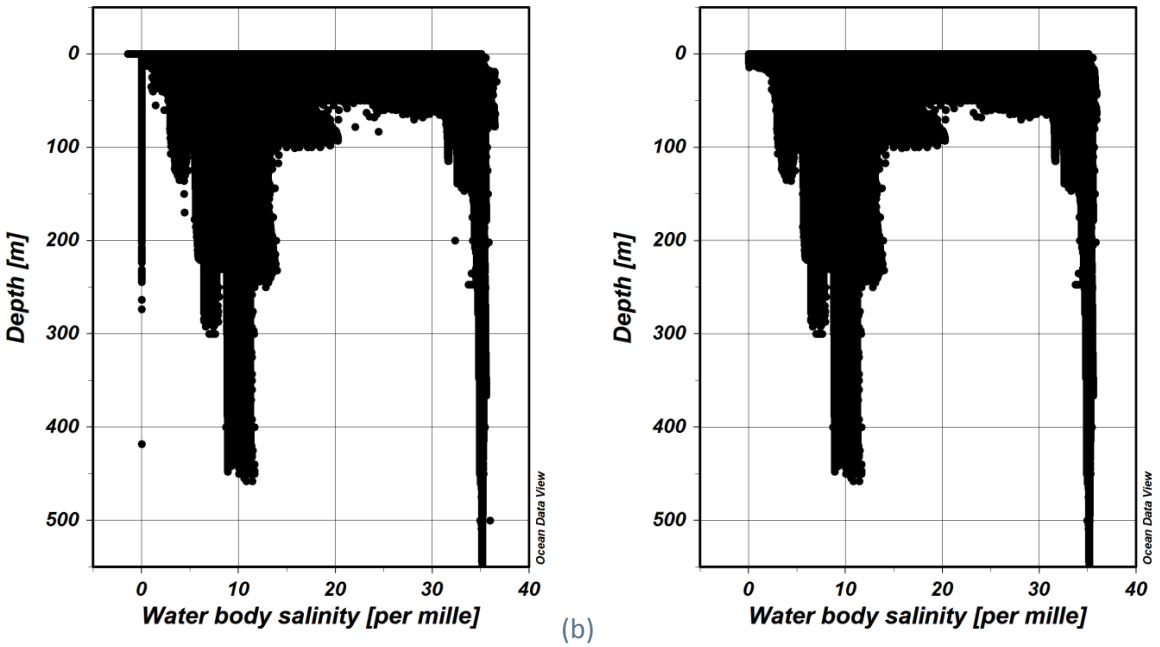
Figure 49. Spatial distribution of temperature (a) and salinity (b).

Figure 50 and Figure 51 show temperature and salinity scatter plots respectively, before (a) and after (b) the quality check assessment.



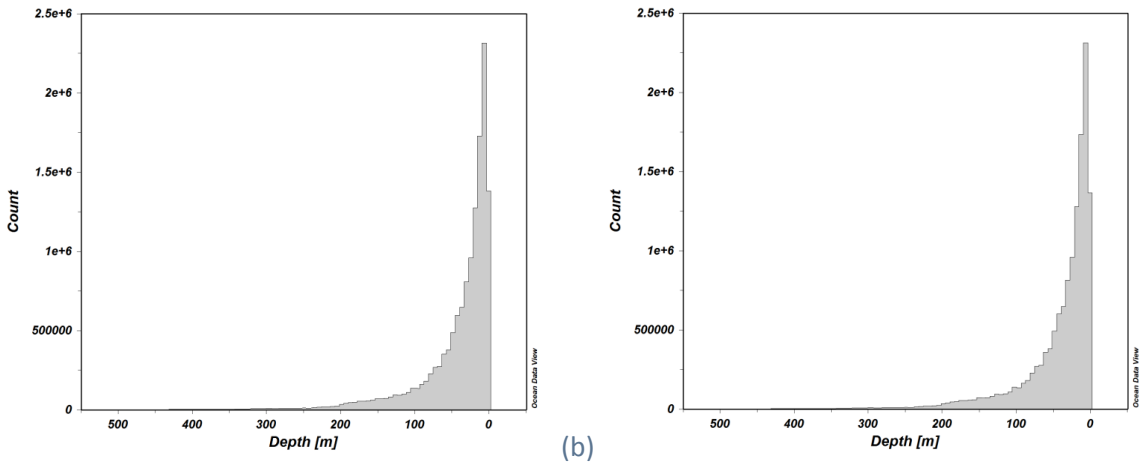
(a) (b)

Figure 50. Temperature observations before QC (a) and after QC (b).



(a) (b)

Figure 51. Salinity observations before QC (a) and after QC (b).



(a) (b)

Figure 52. Temperature (a) and Salinity (b) distribution over depth

The previous version of the product released in framework of the SeaDataNet II project is available at SEXTANT Catalogue (http://sextant.ifremer.fr/en/web/seadatanet) under the name "Baltic Sea - Temperature and salinity observation collection V2" (SDN_V2, http://doi.org/10.12770/1610aa44-0436-4b53-b220-98e10f17a2d4). Compared to SDN_V2 collection there is increase in data in the current collection of SeaDataCloud (SDC_V1) (Table 14). The spatial domain is the same but the time period extended from 1900–2014 to 1900–2017.

	SDN2_V2	% of total	SDC_V1	% of total	% increase
Total	11100238		13780801		24.1
Temperature	11053247	99.6	13434811	97.5	21.5
Salinity	11011231	99.2	13470734	97.8	22.3
Both T and S	10985814	99.0	13234739	96.0	20.5

Table 14. Number of values for Temperature, Salinity and TS couples for the Baltic Sea. The time period 1900-2014 has been considered in order to compare SDN2_V2 and SDC_V1 datasets.

2.5. North Sea

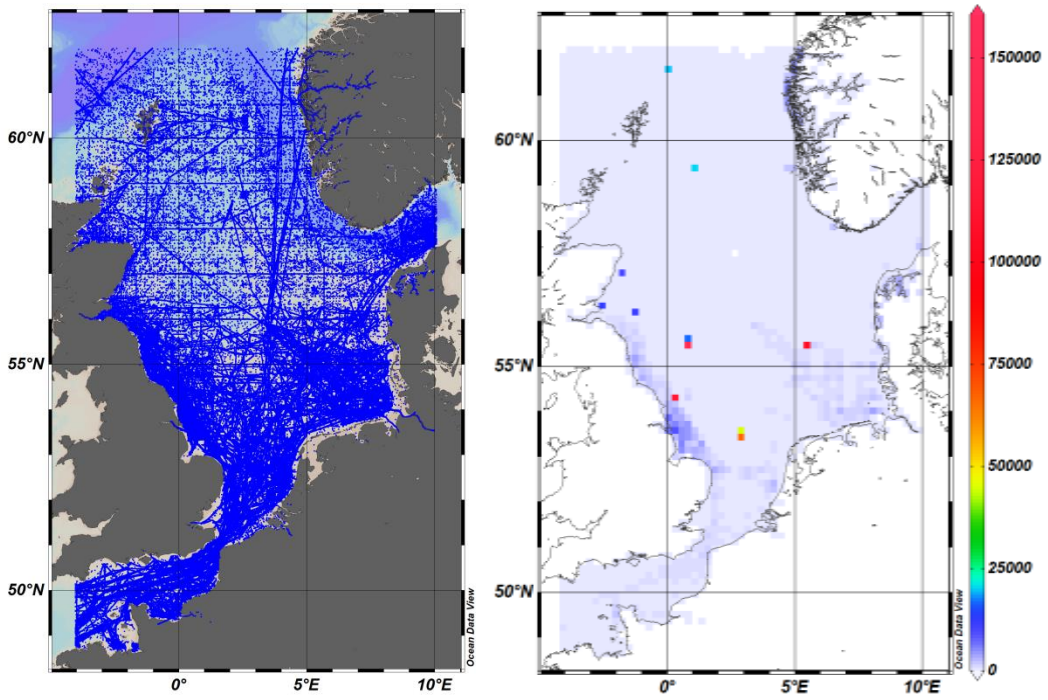
The SeaDataCloud Temperature and Salinity Historical Data Collection for the North Sea contains data on temperature and salinity of water body (profiles and surface measurements) retrieved from the SeaDataNet infrastructure in March 2018. The collection includes non-restricted data obtained from 85 organisations (originators and collating centres).

The collection covers the period 1893 – 2017. All data in the collection have been quality controlled according to procedures described in paragraph 1.4. "En-route" data (Ferry box series, GOSUD series, etc.) were extracted and handled separately. Moorings have not been taken into consideration. Data statistics for the final collections are presented in Table 15.

	Discrete		Trajectories	
	Stations	Values	Stations	Values
TOTAL	162 452		580 376	
Temperature	158 622	7 817 193	576 356	576 356
Salinity	157 545	7 707 384	431 903	431 903
T and S	153 880	7 699 641	431 809	420 867

Table 15. Number of and of values in the final "discrete" and final "trajectories" data set.

The spatial distribution original collection is dominated by data from fixed moorings that haven't been further taken into consideration (they should have been reported as time series). Both "discrete" and "en-route" collections show high densities in certain areas, due to intensive monitoring or research programmes (Figure 53): Belgian coastal zone and Rhine/Meuse Delta, Danish coast, Skagerrak, Firth of Forth and two transects at the Northern boundary.



(a)

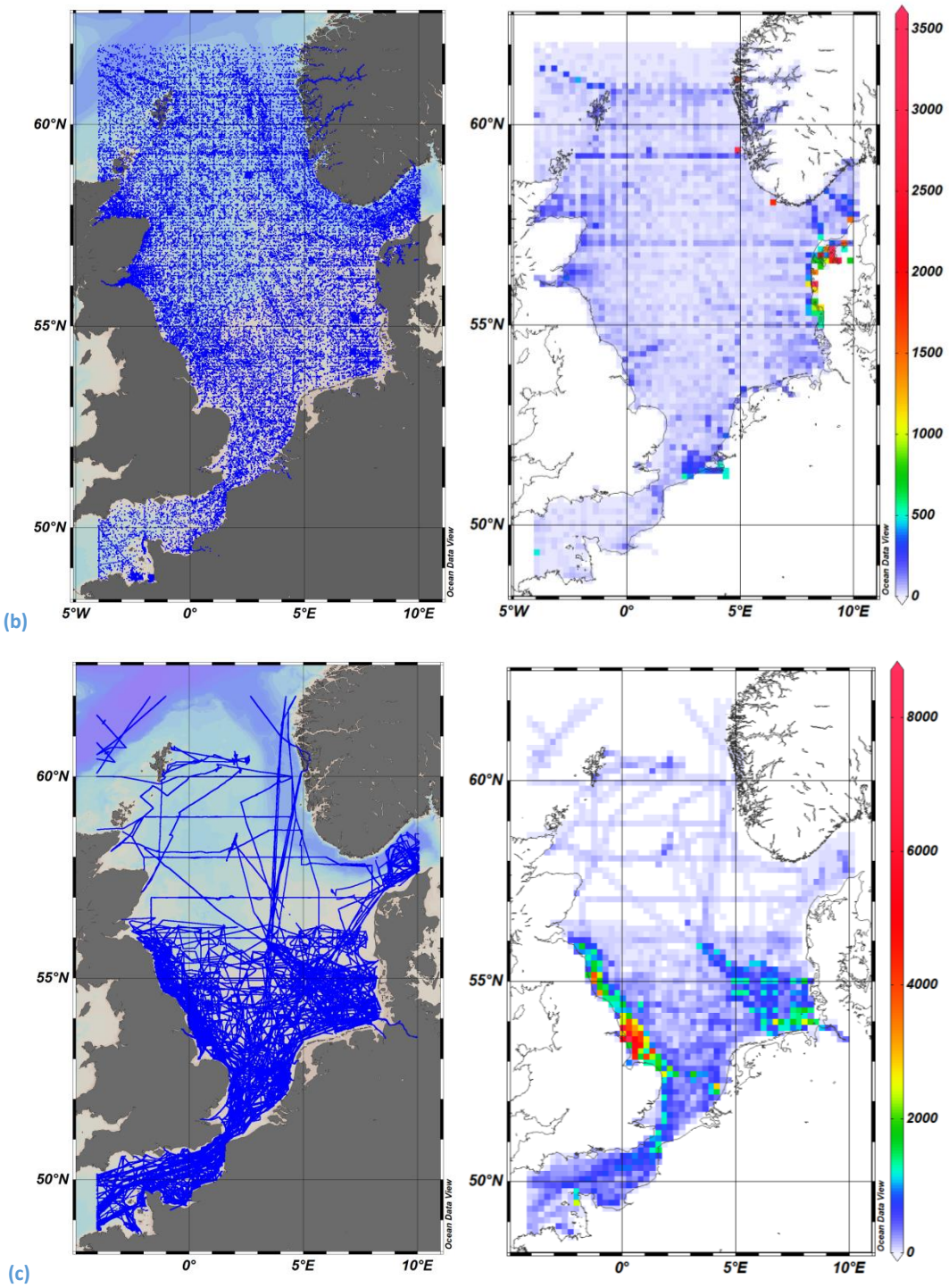


Figure 53 . Spatial distribution of the stations (left) and their density (right). (a) Original collection; (b) Profiles and discrete samplers; (c) "En-route" data.

Although the collection spans from 1893 till 2017, most of the measurements were made during the last 30 years: 115 415 discrete and profile stations (out of 162 452) were performed after 1985 and en-route measurements started in 1985. The coverage over the year is rather uniform (Figure 54).

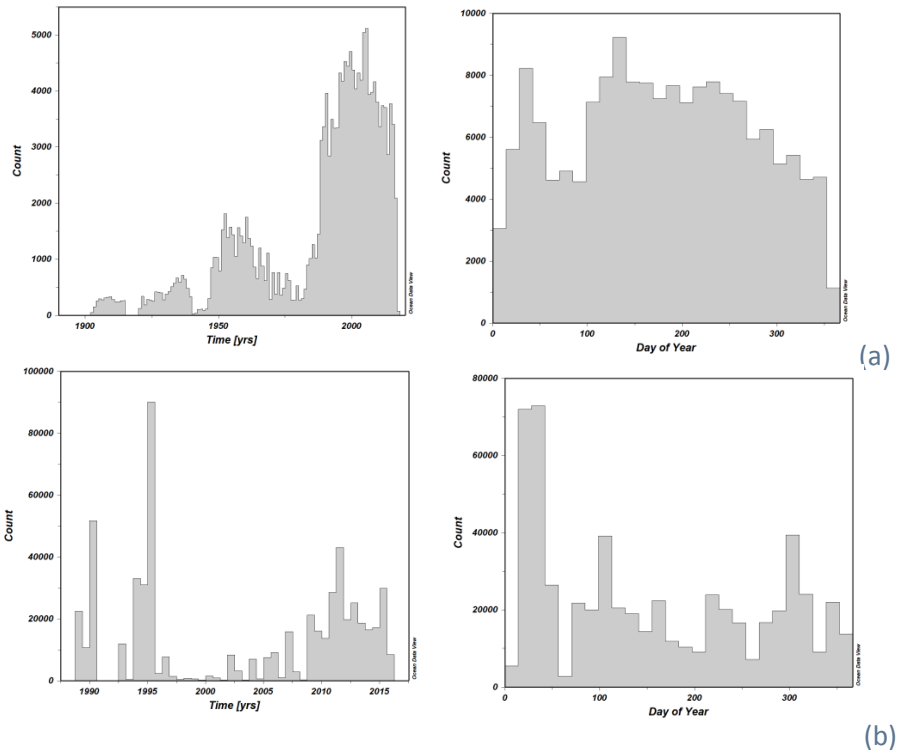


Figure 54. Distribution of the measurements (stations) over time (left) and seasonal distribution (right). (a) Profiles and discrete samplers; (b) en-route measurements.

All data have been quality controlled according to paragraph 1.4. At the end, around 19 000 salinity values and 3 200 temperature values are flagged as suspicious/bad, (flags 3 and 4). This is about 0.3% of the total amount of salinity values and 0.05% of the total amount of temperature values, further details can be seen below in Table 16 and Table 17. Number of S & T values for each quality flag, where the QF Depth is equal to 1 (good) or 2 (probably good) after quality control of the data set.

The majority of the suspicious values come from profiles generated by instruments that were not stabilised at the beginning of the cast. This is mostly visible in profiles published by Belgian and Dutch data centres. There are two apparent causes: a common bad practice that produces (relatively) more bad results in shallow waters and the fact that the operators are interested in parameters that require less or no stabilisation for the instrument to provide rather accurate results.

Some data centres report both down– and upcast in the same file. Data are not per se bad but as the data set only contains the depths of sampling and not the chronology, it is impossible to sort the data. In that case, we flagged the profile with “A”. For the same reason (data might be good), the two station on land were flagged with “A”.

A series of profiles in the Skagerrak consistently shows very high values of salinity. These values were also flagged as “A”. Other outliers were flagged 3.

Three sampling depths are negative. Final sampling depths were checked, when looking suspicious, against the latest (February 2018) data set published on the EMODnet Bathymetry portal, resulting in several of them being flagged as 3 or 4. As these profiles were all coming from the same centre we shall recommend them to check all their profiles; we indeed cannot check if there might be a systematic decoding error. (No conclusion can be drawn when the sampling depth is smaller than the local bottom depth.)

Compared to SDN_V2, the number of samples that have a flag “0” (i.e. not–QCed) inside a profile of which the other values are quality controlled has significantly decreased. But we sometimes found

values flagged as “3” (probably bad) although they are perfectly in agreement with the neighbouring ones, which are flagged “1”.

A feature that would need to be further investigated is the impressive number of profiles with constant values along the depth. Although conditions of well mixing isn’t rare in large parts of the North Sea, perfectly constant values of salinity –for example–, up to the third decimal figure, over a depth of more than 100m, are counter–intuitive.

Quality flag	Number of S values	%	Number of T values	%
0	0	0	0	0
1	7 575 306	98.29	7 722 333	98.79
2	105 593	1.37	90 972	1.16
3	16 533	0.21	1 340	0.02
4	5 282	0.07	813	0.01
5	0	0	0	0
6	0	0	0	0
7	0	0	0	0
8	4 245	0.06	1 209	0.02
9	0	0	0	0
A	0	0	0	0
TOTAL	7 706 959		7 816 667	

Table 16. Comparison of the number of stations in the original SDC_V1 data set and after removal of the high frequency series. “Discrete” collection: Number of S & T values for each quality flag, where the QF Depth is equal to 1 (good) or 2 (probably good) after quality control of the data set.

Quality flag	Number of S values	%	Number of T values	%
0	0	0	0	0
1	419 787	97.19	564 124	97.88
2	1 082	0.25	128	0.02
3	10 453	2.42	11 028	1.91
4	581	0.13	1 076	0.19
5	0	0	0	0
6	0	0	0	0
7	0	0	0	0
8	0	0	0	0
9	0	0	0	0
A	0	0	0	0

TOTAL	431903		576 354	

Table 17. "Trajectories" collection: Number of S & T values for each quality flag, where the QF Depth is equal to 1 (good) or 2 (probably good) after quality control of the data set.

The following Figures show the "discrete" data set scatter plots before and after cleaning: temperature over depth (Figure 55), salinity over depth (Figure 56) and TS diagram (Figure 57 and Figure 58).

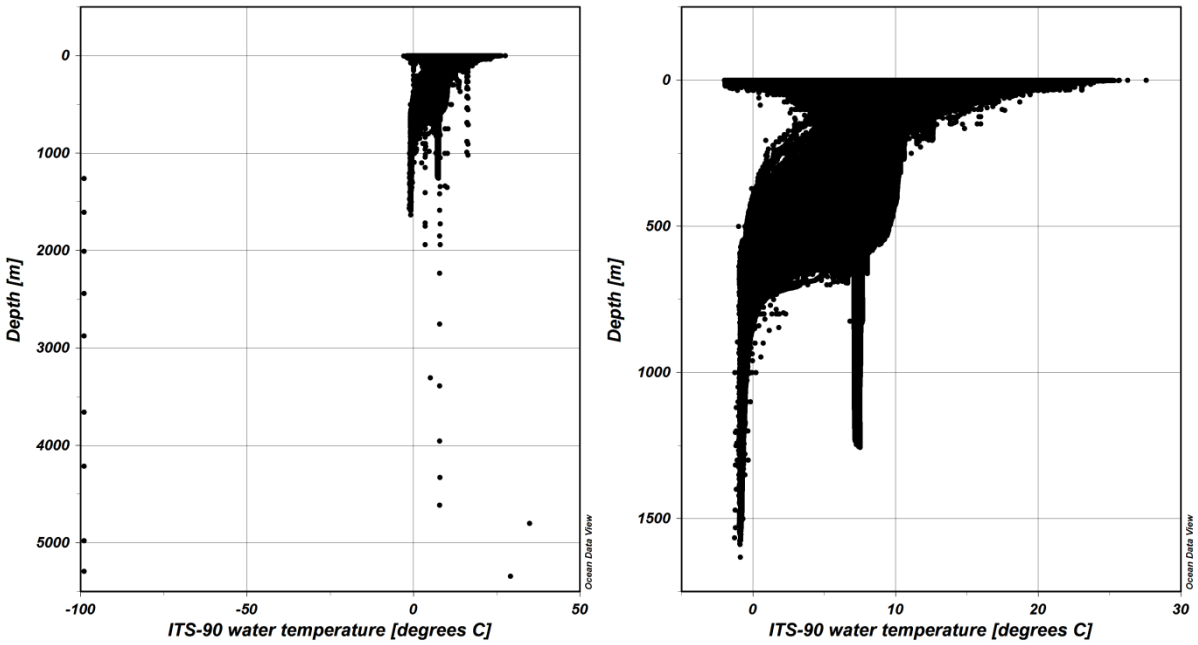


Figure 55. Temperature: data with quality flags set to 0 ("not controlled"), 1 ("good"), 2 ("probably good"), 5 ("changed value") and 8 ("interpolated value") in the original data set (left) and after quality control (right).

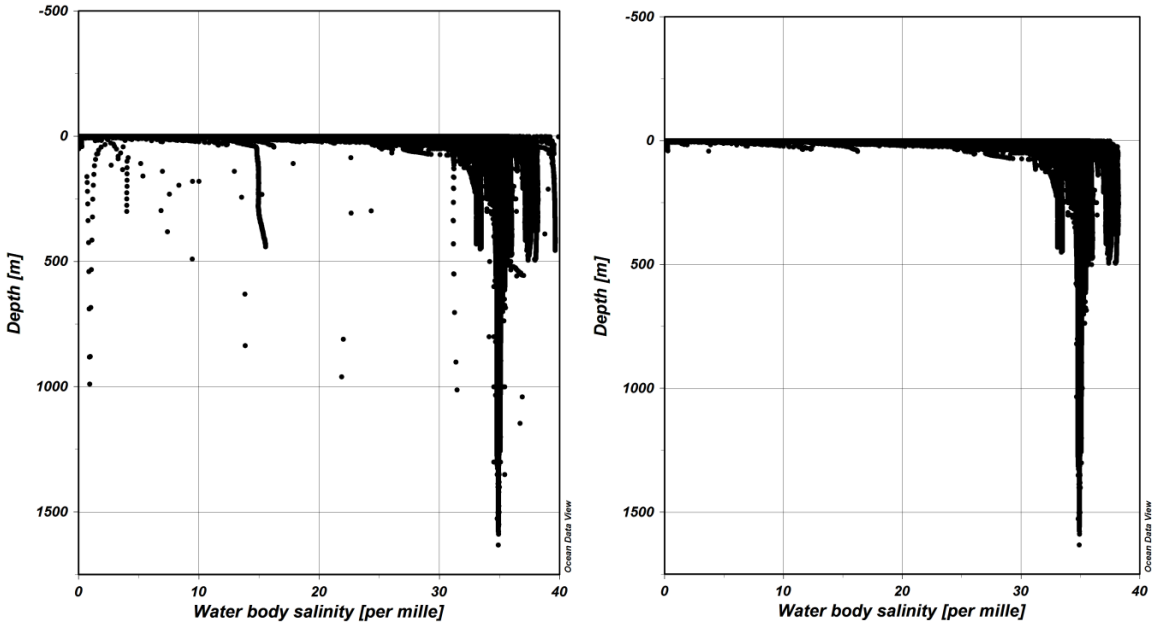


Figure 56. Salinity: data with quality flags set to 0 ("not controlled"), 1 ("good"), 2 ("probably good"), 5 ("changed value") and 8 ("interpolated value") in the original data set (clipped to the 0–40 range, left) and after quality control (right).

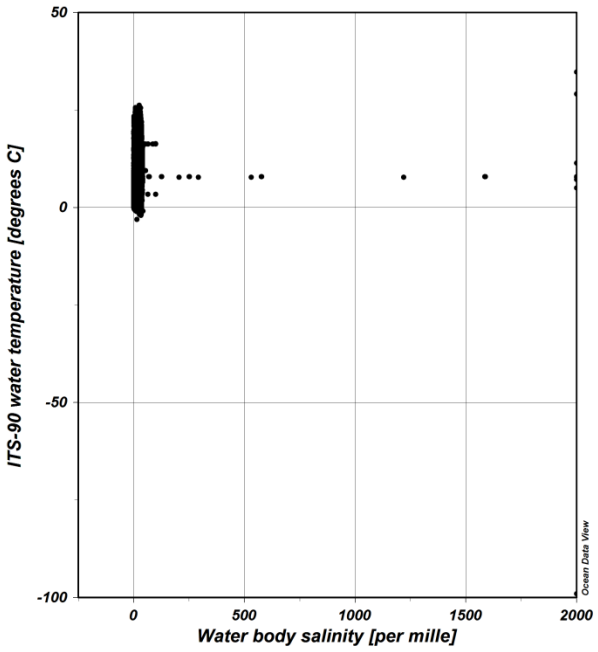


Figure 57. TS-diagram of the original data set.

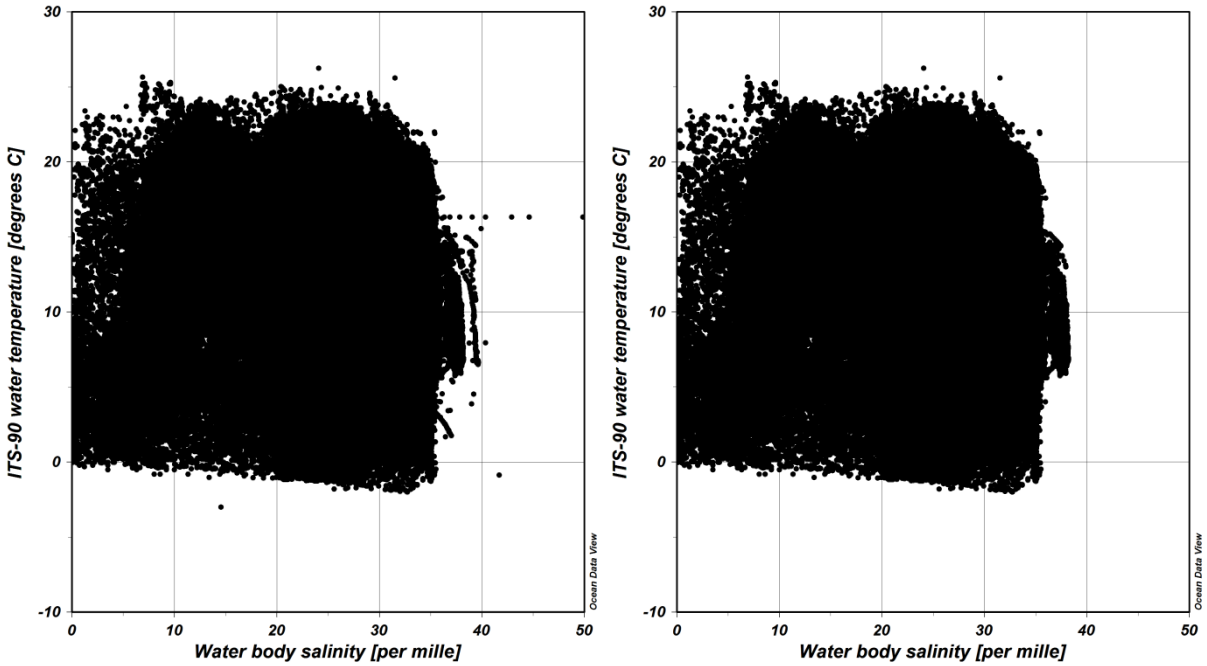


Figure 58. TS-diagram of the original data set (zoomed, left) and of the data set after quality control (right).

The previous version of the product was released in framework of the SeaDataNet II project and available in SEXTANT product catalogue under the name "North Sea - Temperature and salinity observation collection V2" (SDN_V2) (http://dx.doi.org/10.12770/8a51f275-6a8a-4ac2-ba7e-fe491e63a17d). The spatial domain is the same but its time coverage extended from 1900–2014 to 1900–2017.

"Discrete" collection

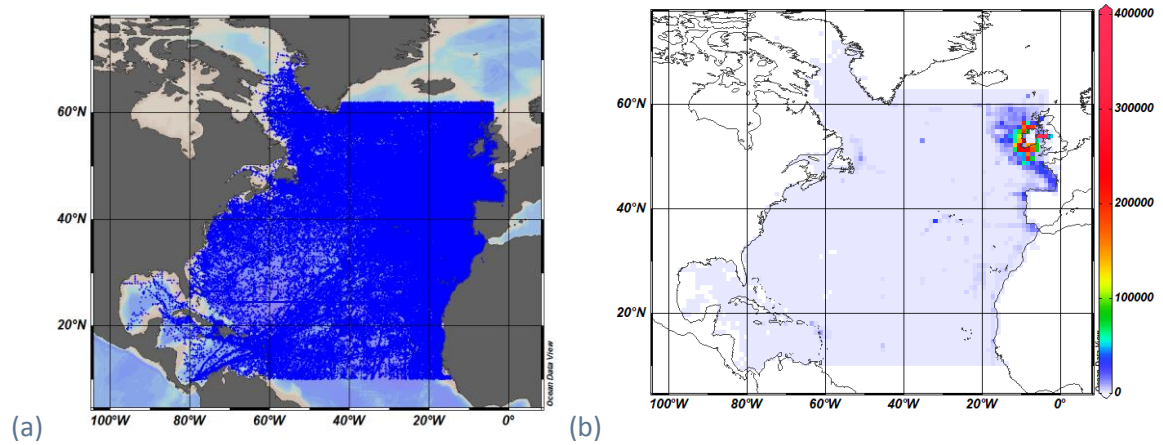
Compared to SDN_V2 collection there is a decrease in data in the current collection of SeaDataCloud (SDC_V1) due to the separate handling of high frequency series (e.g. Ferry box data) (see Table 18).

	SDN2_V2 #stations	% of total	#values	SDC_V1 #stations	% of total	#values
Total	1610854			162452		
Temperature	1590331	98.73%	15943941	158622	97.64%	7817193
Salinity	946724	58.77%	7287495	157545	96.98%	7707384
T and S	944129	58.61%	7344660	153880	94.72%	7699641

Table 18. Comparison of the number of stations in SDN_V2 and SDC_V1, and of the number of those containing respectively T values, S values and T,S-pairs.

2.6. 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 and till gulf of Mexico for the west part. The spatial distribution and the data density maps of T and S observations from the entire data collection are shown in Figure 59(a) and (b). Data distribution maps show a good geographical spread with the best coverage on the eastern part of the domain, mainly close to the areas off Ireland and in the Bay of Biscay (Figure 59 (b)). This higher coverage on the east part is also due to a large number of thermosalinograph measurements (4177186 stations from the MI data centre), which are off the coast of Ireland. The North Atlantic Ocean historical data set contains just over 8108995 stations for the period 1890-2012 and 982778 stations for recent years (2013-2017). The data collection contains 6002 cruises for 9091773 stations.



(a) (b)

Figure 59. TS stations collection for the North Atlantic Ocean: (a) data distribution map; (b) data density plot showing where most values have been sampled.

Table 19 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; 3134 stations have only salinity parameter. Only 39.25% of the observations have the couple TS that means that most of the data have only temperature observations (most of TSG observations).

PAR	#stations	%
All	9091773	
T	9074128	99.80%
S	3572113	39.29%
TS	3568979	39.25%
Z	9091756	

Table 19. Synthesis table with data statistics.

Figure 60 shows the distribution map by parameter (temperature, salinity and couple temperature/salinity). The stations with only temperature data are mainly in the northeast part of the map.

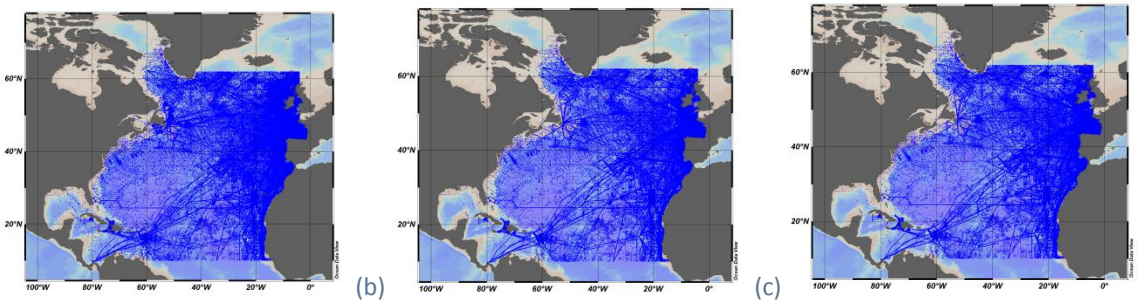


Figure 60. Spatial distribution of: (a) temperature observations; (b) salinity observations; (c) temperature and salinity observations.

Temporal data distribution is shown in Figure 61. The distribution in time is poor for the first 80 years, it increases after 1980 until the end of 1990s where it further increases. In the latest years there is a decrease in data which is caused by a natural time lag between sampling and until data becomes available in the SeaDataCloud system. Figure 4b shows the seasonal distribution of data. Most of the data have been collected during spring, summer and autumn, with a larger peak during summer.

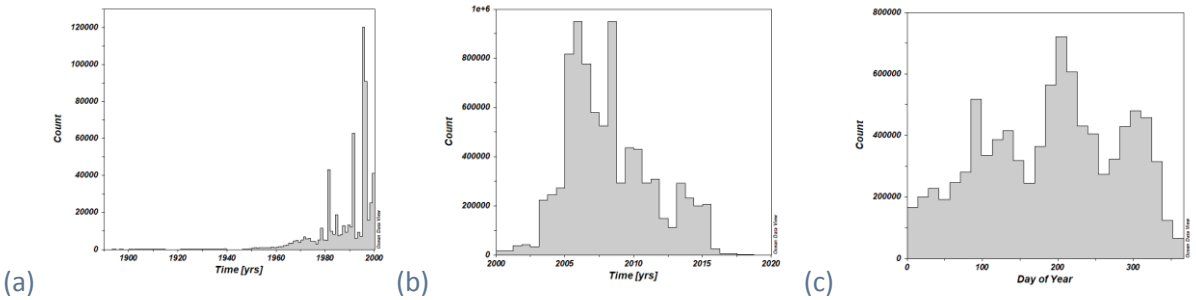
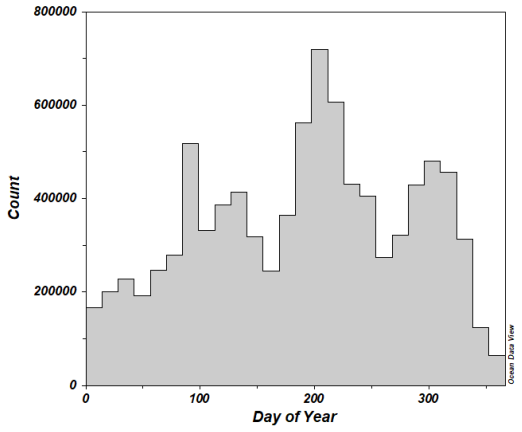
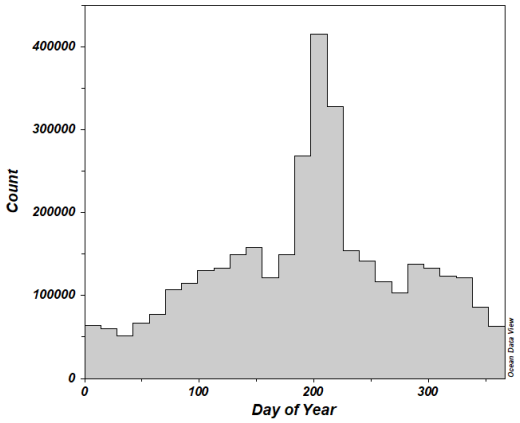


Figure 61. Time distribution for the period 1890-1999 (a) and 2000-2017 (b), and seasonal distribution over the year (c) for the entire data set.

Splitting the temporal distribution by parameter (Figure 62) shows a sampling of salinity data mainly during the summer.



(a)



(b)

Figure 62. Seasonal distribution for temperature (a) and salinity (b).

All data have been quality controlled according to the criteria defined in chapter 1.3. Around 12703 salinity values and 7770 temperature values were flagged as suspicious/bad, flag 3 and 4. This is about 0.024% of the total amount of salinity values and 0.011% of the total amount of temperature values, further details can be seen below in Table 20 and Table 21.

The suspicious values consist mainly of spikes, outliers and unstable density profiles, but there are also some other problems:

- Some profiles appear to be more salty, due to wrong position.
- Few values have negative depths (Argo profiles).
- Some measurements contain values 0; corresponding certainly of missing values (the QC 1 has been updated to 4).
- Some salinity measurements contain QF0 but it is due to a wrong QC mix between psal and ssal measurements during aggregation.
- Few measurements appear to be “on land”, or close to rivers.

The list of the QC changes will be included in the quality control feedback that will be sent to the different SeaDataCloud partners.

PAR	TOT	QF0	QF1	QF2	QF3-9
T	70139531	65237	69642822	69641	361831
%		0.09	99.30	0.09	0.52
S	53549152	1274601	51080353	122079	1072119
%		2.38	95.39	0.23	2.00

Table 20. Quality Flags statistics for the initial data collection (without QC procedure).

PAR	TOT	QF0	QF1	QF2	QF3-9
T	69921467	61640	69421176	69046	369601
		0.09	99.29	0.09	0.53

S	53331088	1274555	50850255	121456	1084822
		2.39	95.34	0.24	2.03

Table 21. Quality Flags statistics QC after the quality check procedure.

The following figures show the distribution of the parameters versus depth. Figure 63 shows the temperature versus depth scatter plots before quality control for all the QF, for the QF1 and QF0.

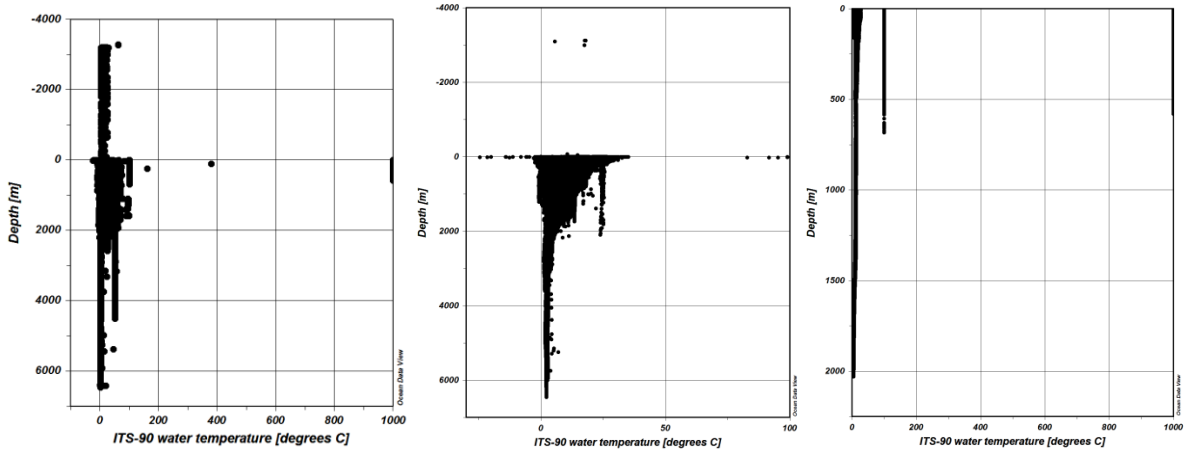


Figure 63. Temperature versus depth scatter plot of the North Atlantic data collection covering the time period 1900-2017: (a) all data Quality Flags; (b) QF=1; (c) QF = 0 (no quality control).

Figure 64 shows the salinity versus scatter plots before quality control for all the QF, for the QF1 and QF0.

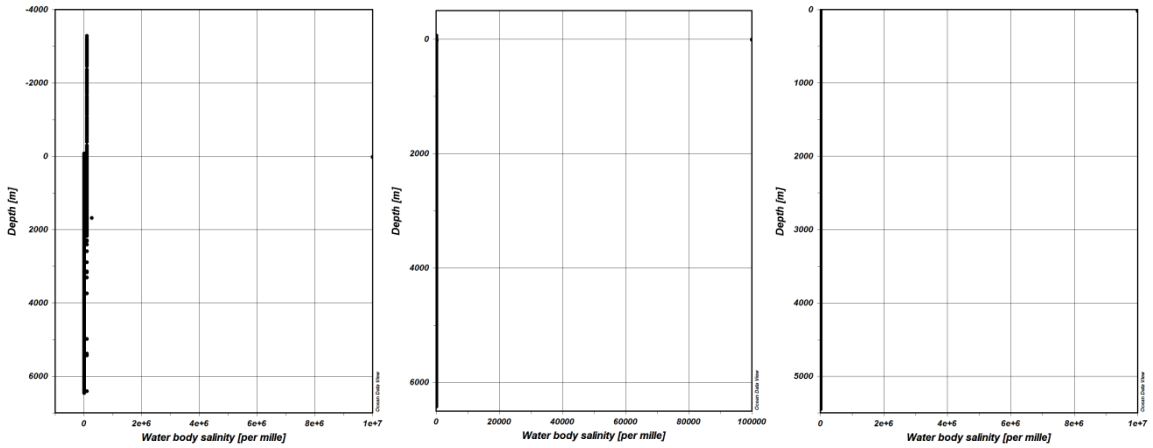
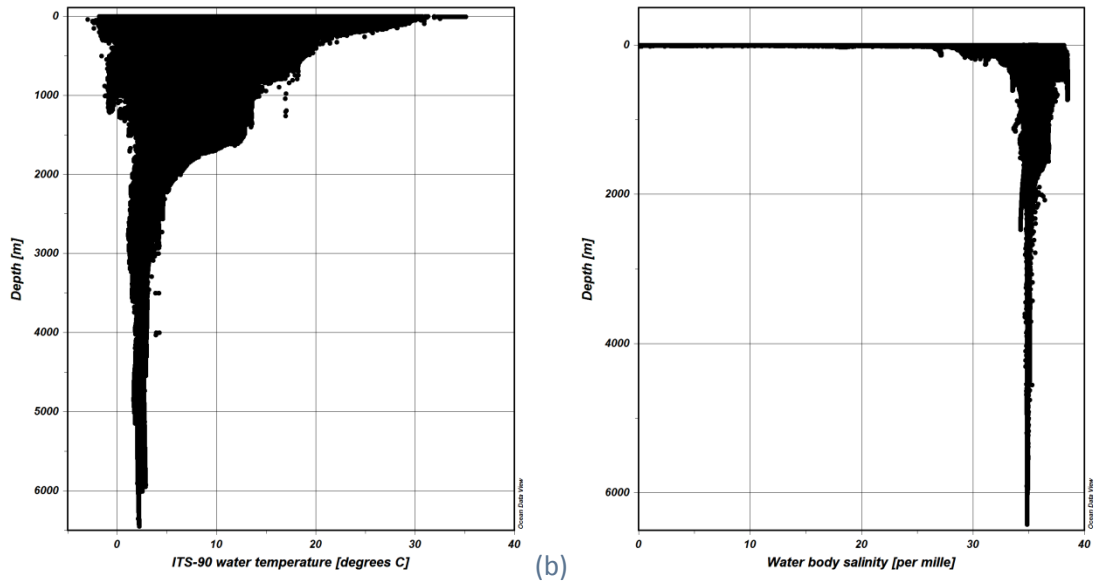


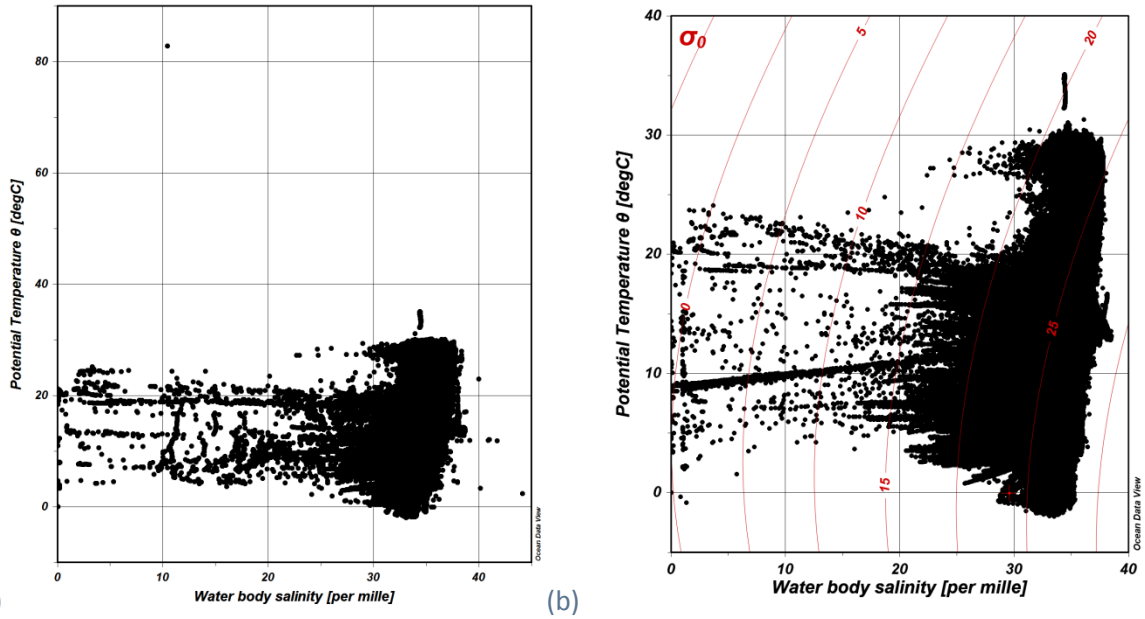
Figure 64. Salinity versus depth scatter plot of the North Atlantic data collection covering the time period 1900-2017: (a) all data Quality Flags; (b) QF=1; (c) QF = 0 (no quality control).

Plots in Figure 63 and Figure 64 show that among data with good quality flag (QF=1) there are still some wrong values that need to be updated with a QF4. Some obvious outliers were easy to detect and remove from the good dataset. Figure 65 displays the parameters versus depth of good quality data after QC analysis. Figure 66 presents the θS diagram plots and Figure 67 shows the vertical distribution of the potential density anomaly before and after the quality control procedure.



(a) (b)

Figure 65. North Atlantic data collection (1890-2017) considering only data with QF = 1 (good): (a) Temperature versus depth; (b) Salinity versus depth.



(a) (b)

Figure 66. North Atlantic data collection: θS diagram (QF=1) showing data before (a) and after (b) quality control.

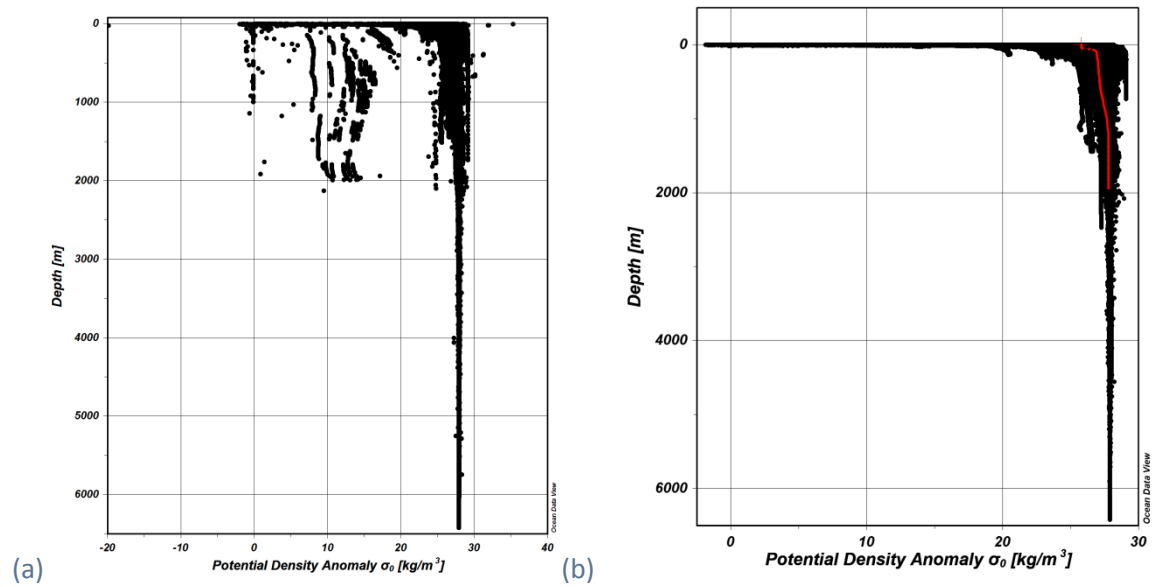


Figure 67. North Atlantic data collection considering only data with QF = 1 (good). Potential Density Anomaly: (a) before correction and (b) after correction.

The previous version of the product (SDN2_V2) was released at the end of 2015 in the framework of the SeaDataNet II project and it is available at SEXTANT Catalogue (http://sextant.ifremer.fr/en/web/seadatanet) under the name "North Atlantic Ocean - Temperature and salinity observation collection V2" (http://doi.org/10.12770/a61129f0-afbc-4bfa-8307-00f37d37d98a).

Comparing the SDC_NAT_DATA_TS_V1 to the SDN2_V2 collection over the same spatial domain but time period from 1900-2014 to 1890-2017 in Table 22 it shows a large data increase (+403%), mainly for temperature measurements (435.7% of increase). This increase is mainly due to the large data ingestion from the Marine Institute (Ireland). Most of their data have only temperature measurements (and most are from thermosalinograph instrument type), explaining the large increase for this parameter between SDN2_V2 and SDC_V1. Moreover, the data aggregation for SDC_V1 contains underway data imported after subsampling.

#stations	Total	T	S	TS
SDN2_V2	1807266	1693840	785476	784015
SDC_V1	9091773	9074128	3572113	3568979
% of increase	+403%	+435.7%	+354.7%	+38.1%

Table 22. Data statistics of previous (SDN2_V2) and current (SDV_V1) version of the North Atlantic Ocean historical data collections.

3. Report on data anomalies

Each regional leader has made a list of anomalies with information on data providers (using EDMO_code), LOCAl_CDI_ID and parameters on which the anomalies have been detected. This list is provided by the ODV logs. The anomalies are sorted by EDMO_code and are sent to each data provider. A system will be implemented at the central CDI level to keep track of the anomalies'

corrections and CDIs updates. Moreover it will be fixed a deadline to perform and submit the necessary corrections and avoid to harvest the same data anomalies in the next historical data collections.

What are those anomalies files? ODV logs and/or other logs files created to detect data anomalies. The thing is to have the apropriate parameters in those files to help the EDMO_code to find the anomalies in their data. See the guideline done in SDN2 when working with the feedback of MyOcean. The central CDI service has to control the resubmission of the corrected data.

In phase II, the VRE will be ready and data anomalies will be managed in an automatic way sending the feedbacks to the data providers and getting their corrective actions. The best thing would be to create/send a message in the VRE provider environment with information (ODV logs and/or other logs files created to detect data anomalies) about the changes that regional product leaders have made on the data QC. With this information, they can easily check the data and correct if necessary. If the correction is not appropriate, a message can be sent in the VRE to the regional leaders.

4. Summary and conclusions

The procedure of the data aggregation is an extensive and constructive exercise to manage a huge volume of data, and involving many people and institutions.

The objectives 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 SDC 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 develop new and refined QC procedures to analyze data completeness and consistency for long term studies;
- To introduce metadata analysis in order to maximize the information we could get from historical data collections and provide recommendations to the Steering Group for including or modifying the metadata description;
- to rise advices about future harvesting and quality assessment procedures;
- to contribute to the development and consolidation of ODV software.

The Quality Assessment procedure permitted to identify and correct lots of data. It seems that there is still needs to improve the aggregation procedure, the quality control from data providers and the data formation at the data center level. All regional leaders expressed the need of actions from data centres to correct bad file formats. Most of the data anomalies are due to bad file formats or bad QC procedures from the data centres. A solution could be to implement automatic check before data submission to avoid the main problems detected.

5. Annex 1. Naming Convention

Table A1 Name convention for SeaDataCloud regional data collections

Project	Region	Product	Parameter	Version	Product Name	Extended Name
SDC	ARC	DATA	T and S	V1	SDC_ARC_DATA_TS_V1	SeaDataCloud Temperature and Salinity Historical Data Collection for the Arctic Ocean (Version 1)
SDC	NWS	DATA	T and S	V1	SDC_NWS_DATA_TS_V1	SeaDataCloud Temperature and Salinity Historical Data Collection for the North-West-Shelf (Version 1)
SDC	BAL	DATA	T and S	V1	SDC_BAL_DATA_TS_V1	SeaDataCloud Temperature and Salinity Historical Data Collection for the Baltic Sea (Version 1)
SDC	NAT	DATA	T and S	V1	SDC_NAT_DATA_TS_V1	SeaDataCloud Temperature and Salinity Historical Data Collection for the North Atlantic Ocean (Version 1)
SDC	MED	DATA	T and S	V1	SDC_MED_DATA_TS_V1	SeaDataCloud Temperature and Salinity Historical Data Collection for the Mediterranean Sea (Version 1)
SDC	BLS	DATA	T and S	V1	SDC_BLS_DATA_TS_V1	SeaDataCloud Temperature and Salinity Historical Data Collection for the Black Sea (Version 1)

6. Annex 2. List of Acronyms

Acronym	Definition
ARC	Arctic ocean
BAL	Baltic Sea
BLS	Black Sea
CDI	Common Data Index
CLIM	Climatology
CMEMS	Copernicus Marine Environment Monitoring Service
DATA	Aggregated Dataset
DIVA	Data-Interpolating Variational Analysis (software)
DOI	Digital Object Identifier
EC	European Commission
EDMO	European Directory of Marine Organisations (SeaDataNet catalogue)
GLO	GLobal Ocean
IOC	Intergovernmental Oceanographic Commission
IODE	International Oceanographic Data and Information Exchange (IOC)
MED	Mediterranean Sea
NAT	North Atlantic Ocean
NWS	North West Shelf

ODV	Ocean Data View Software
QC	Quality Checks
QF	Quality Flags
SDC	SeaDataCloud
SDN	SeaDataNet
TS	Temperature and Salinity
WOA	World Ocean Atlas
WP	Work Package

7. References

- SeaDataCloud project (2016-2020), grant agreement 730960, EU H2020 programme, www.seadatanet.org/About-us/SeaDataCloud
- Seadatanet II project (2011-2015), grant agreement 283607, EU Seventh Framework Programme, https://www.seadatanet.org/About-us/SeaDataNet-2
- Schlitzer, R., Ocean Data View, odv.awi.de, 2017
- Sextant catalogue, Seadatanet products, http://sextant.ifremer.fr/en/web/seadatanet
- Troupin, C., Barth, A., Sirjacobs, D., Ouberdous, M., Brankart, J.-M., Brasseur, P., Rixen, M., Alvera Azcarate, A., Belounis, M., Capet, A., Lenartz, F., Toussaint, M.-E., & Beckers, J.-M. (2012). Generation of analysis and consistent error fields using the Data Interpolating Variational Analysis (Diva). Ocean Modelling, 52-53, 90-101.