EUROPEAN COMMISSION
Executive Agency for Small and Medium-sized Enterprises (EASME)

Department A - COSME, H2020 SME and EMFF
Unit A3 - EMFF

Agreement number: EASME/EMFF/2015/1.2.1.1/SI2.709624

Project Full Name: Monitoring the Ocean Climate Change with Argo

European Maritime and Fisheries Fund (EMFF)

MOCCA

D4.4.5 Report on the update of the CTD reference database for salinity Delayed-Mode Quality Control in the Nordic Seas

Circulation:	PU: Public
Lead partner:	
Contributing partners:	BSH, Ifremer
Authors:	Ingrid M. Angel-Benavides, Birgit Klein, Christine Coatanaon
Quality Controllers:	Romain Cancouët, Sylvie Pouliquen
Version:	1.0
Reference	D4.4.5 Report on the update of the CTD reference database for Salinity DMQC in the Nordic Seas_v1.0.docx
Date:	12.06.2020

Euro-Argo ERIC
European Research Infrastructure
(2014/261/EU)

Consisting of:

Organisation/Natural person	Represented by	Statute	Contributing entities[1]
Euro-Argo ERIC	N/A	Coordinator	N/A
The French Republic	Ifremer	Member	SHOM, INSU/CNRS, Meteo-France, IRD, IPEV
The Federal Republic of Germany	BSH	Member	GEOMAR, University of Hamburg, Alfred-Wegener-Institute for Polar and Marine Research (AWI), Institute for Chemistry and Biology of the Marine Environment (ICBM)
The Hellenic Republic	HCMR	Member	N/A
The Italian Republic	OGS	Member	N/A
The Kingdom of the Netherlands	KNMI	Member	N/A
The Republic of Finland	FMI	Member	N/A
The United Kingdom of Great Britain and Northern Ireland	Met Office	Member	NOCS, BODC
The Kingdom of Norway	IMR	Observer	N/A
The Republic of Poland	IOPAN	Observer	N/A

This document may change without notice.

Document History

Version[2]	Issue Date	Stage	Content and Changes
0.1	07.04.2020	Draft	Initial document creation by Ingrid M. Angel-Benavides
0.2	27.05.2020	Draft	Comments and changes by C. Coatanoan
0.3	29.05.2020	Draft	Comments and changes by B. Klein
0.4	02.06.2020	Draft	Edits and changes by Ingrid M. Angel-Benavides
0.5	07.06.2020	QC	Romain Cancouët
1.0	12.06.2020	Final	Final version

[1] As indicated in the "Technical and Scientific description of the Euro-Argo ERIC" July 2013 attached to the Euro-Argo Statutes.

[2] Integers correspond to submitted versions.

Table of Contents

Table of Figures

Table of Tables

1. INTRODUCTION

This document describes the methodology used for the update of the CTD reference database (CTD-RDB) for salinity Delayed-Mode Quality Control (DMQC) of Argo floats in the Nordic Seas. The resulting data is included in the latest release of the database (2019v01, October 2019). The DMQC procedures are described in the deliverable D.4.3.1 Report on Delayed-Mode processing on the MOCCA fleet, and the partners responsible for their execution are listed in deliverable D4.1.1 Organization of Float Data Management among DAC and DM-operators.

BSH is responsible for the DMQC of the MOCCA fleet and MOCCA cofounded floats operating in the Nordic Seas, which is one of the regions selected for the expansion of the Argo observation system into marginal Seas and high latitudes. Figure 1 shows the Argo profile density in the Nordic Seas, which concentrates on the four deep basins: Greenland Sea (GS), Lofoten Basin (LB), Norwegian Basin (NB), and Iceland Sea (IS).

The report is organized as follows. In Section 2, we present an overview of the purpose and requirements of the CTD-RDB, as well as the procedures currently implemented for its maintenance. In Section 3 we introduce the region of interest and in Section 4 we present the status of the database for the Nordic Seas in the 2018V02 version. In Section 5, we outline the actions performed for the update, followed by a detailed description of the data sources and the preparation procedures in Section 6. In Section 7, we specify the procedures for data merging and post-processing. In Section 8, we present the characteristics of the resulting regionally updated version of the CTD-RDB. The report concludes with a brief outlook on the remaining tasks to further improving the CTD-RDB at both regional and global levels.

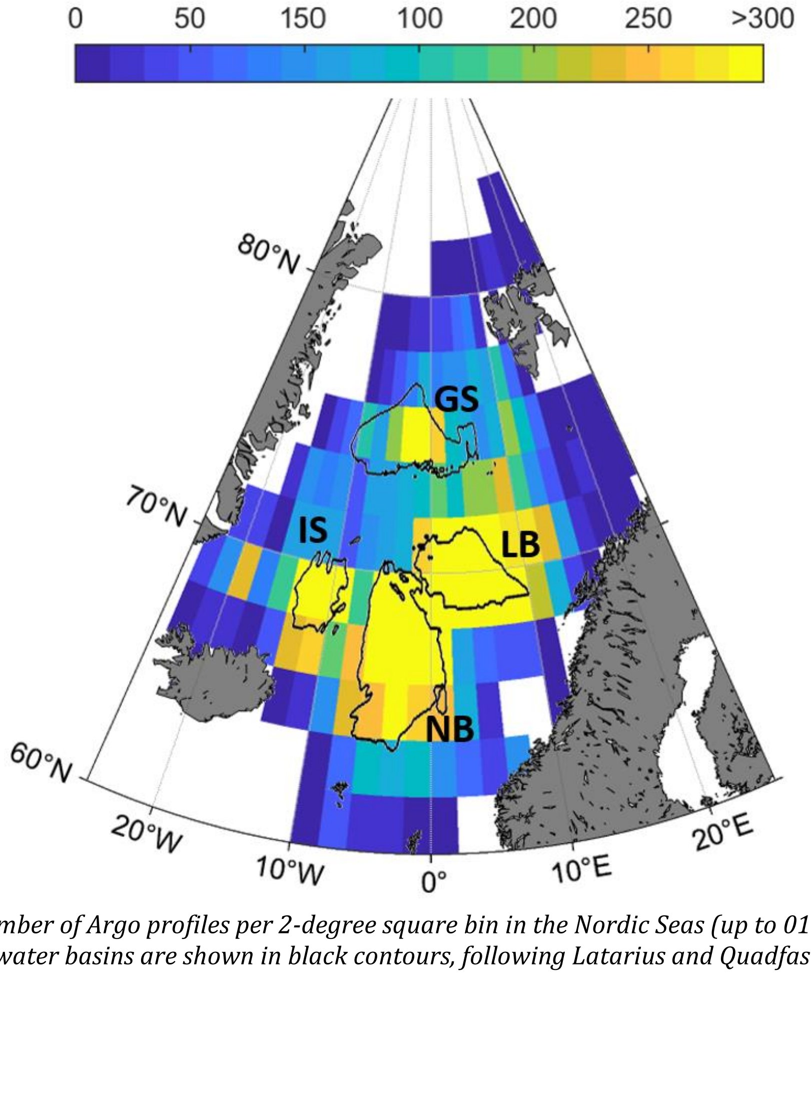

Figure 1 Number of Argo profiles per 2-degree square bin in the Nordic Seas (up to 01 June 2020) The deep-water basins are shown in black contours, following Latarius and Quadfasel (2010).

2. CTD-REFERENCE DATABASE FOR SALINITY DMQC

Data collected by Argo floats undergo a strict DMQC procedure to ensure their scientific quality. The guidelines for the DMQC, as provided by the Argo Data Management Team, are documented in the Argo user's manual v3.2[3] and the Argo quality control manual for CTD and trajectory data v3.1[4].

In particular, the salinity data is carefully screened looking for artificial trends and offsets, which result from instrumental drifts in the conductivity sensor. Using the method described in Owens and Wong (2009) and improved by Cabanes et al. (2016), hereafter referred to as the OWC method, DMQC operators identify such salinity errors and correct them when possible.

The OWC method uses historical hydrographic data to estimate a climatological reference salinity for the float's positions and times using objective mapping. Therefore, an appropriate correction requires reference databases with a temporal and spatial coverage that allows a realistic estimate of such reference, making it possible to distinguish between the signal corresponding to natural variability and sensor drift.

Currently, the Coriolis/Ifremer team for operational oceanography centrally maintains a global CTD-RDB, available to DMQC operators via a password-protected FTP server, which is updated at least once a year using data obtained through downstream services and directly from scientists. The 2018V02 version of the CTD-RDB was the current version at the beginning of the activities covered by this report.

2.1. Data

According to the requirements of the OWC method for Salinity DMQC, the CTD-RDB profiles are delivered as a set of Matlab files. Each one contains CTD profiles inside one of the World Meteorological Organization (WMO) squares/boxes, defined in a 10° latitude x 10° longitude grid (Figure 2), and are named accordingly (e.g. the file ctd_7600.mat contains profiles inside the WMO box 7600). Each mat file contains *n* profiles and the information is stored in vector and matrix variables (Table 1).

The vector variables contain the metadata for each profile: timestamp (*dates*) and geographical position (*lon* and *lat*), plus two internal identifiers: *source*, a profile ID code, and *qclevel*, a code referring to the original database from which the profiles where obtained. The *qclevel* variable is not used by the OWC method but was introduced to provide information about the quality level of the CTD profiles, according to their original data provider, back in CTD-RDB 2016v01. Profiles already present before this update were assigned with a *qclevel* = COR as default value[5]. The codes used are listed in Table 1.

The matrix variables contain the profile data, with samples stored in rows and profiles in columns. As the profiles have a different number of samples, the number of rows (*m*) is determined by the profile with the largest number of samples. For profiles with fewer samples, the extra rows are filled with NaNs.

[3] http://dx.doi.org/10.13155/29825

[4] http://dx.doi.org/10.13155/33951

[5] Personal communication w. Christine Coatanoan.

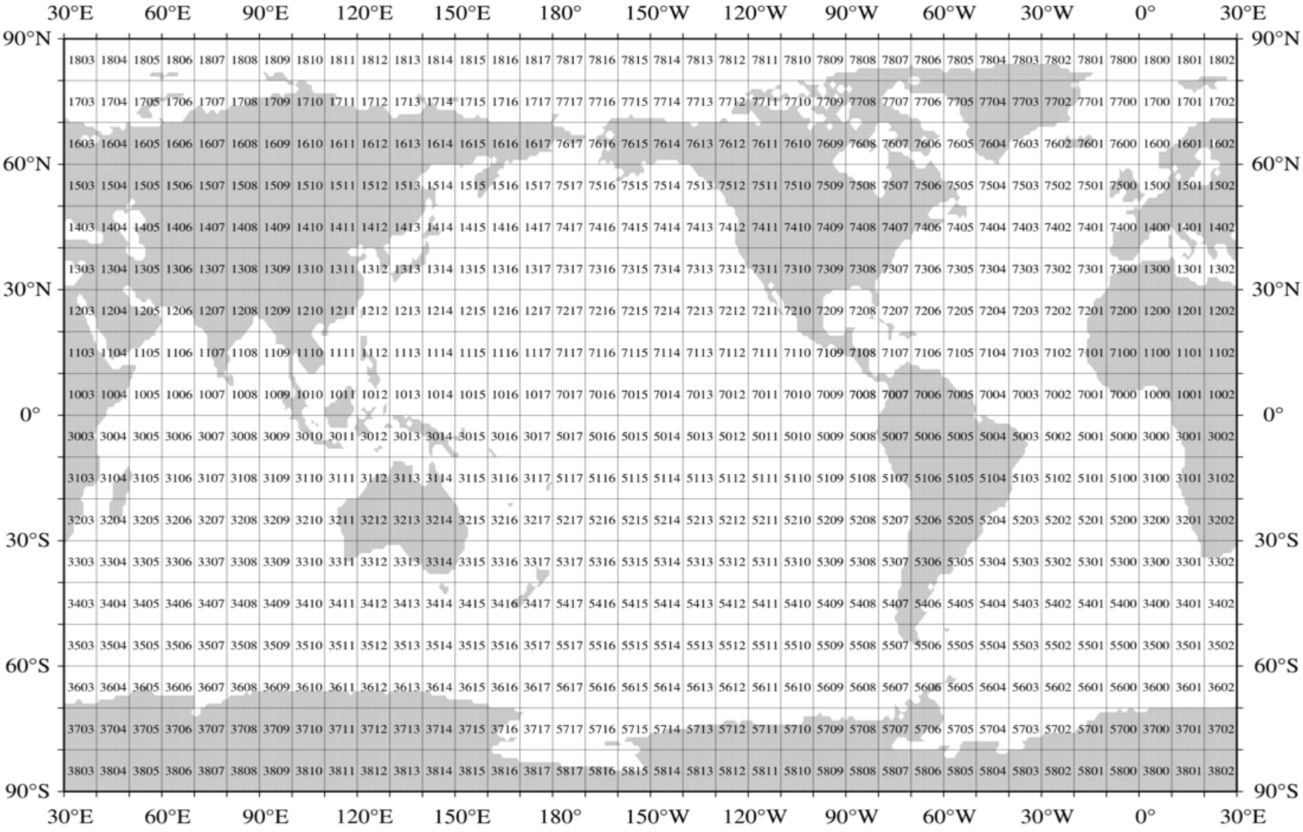

Figure 2 World Meteorological Organization 10-degree boxes.

Table 1 CTD-RDB content

Class	Variable	Size	Type	Format / unit	Details
Metadata	dates	1 x n	double	yyyymmdd HHMMSS	
Metadata	lat	1 x n	double	Degrees	
Metadata	lon	1 x n	double	-180° to 180°	
Metadata ID	qclevel	1 x n	cell	Original database code	Following codes are used in the Nordic seas: COR (Coriolis), OCL (Ocean Climate Library – World Ocean Database), CCH (CLIVAR and Carbon Hydrographic Data Office - CCHDO) and SPI (Scientist, Principal Investigator)
Metadata ID	source	1 x n	cell	Codes	- COR: internal station ID. Ex. 11088883 - CHH: cruise name. Ex. 77DN19910726 - SPI: cruise ID and station number. rr17d0049_001
Data	pres	m x n	double	dbar	
Data	temp	m x n	double	°C ITS90	
Data	ptemp	m x n	double	°C rel to 0 dbar	
Data	sal	m x n	double	PSS-78	

2.2. CTD selection criteria

Section 4.5 of the Argo Quality Control Manual for CTD and Trajectory Data version 3.1 prescribes the CTD-RDB selection, aggregation, and quality control procedures. Table 2 shows the complete list of requirements along with details about how the procedures are currently implemented and possible improvement actions.

Table 2 Profile selection criteria for the CTD-RDB

Argo QC manual criteria	Implemented?	Obs.	Improvement actions
1). Use only data that have passed all NODC[6] quality control tests for observed level data.	YES	Originators flags and other quality controls are used	
2). Use all country codes.	YES		
3). Use only profiles that sampled deeper than 900 dbar.	YES	Invalid samples may be present deeper than 900 dbar, making some profiles useless for OWC.	-Remove invalid samples, defined as those with any Data variable equal to NaN. -Afterwards, check if the profile is still deeper than 900 dbar
4). Weed out all data points outside these ranges: 24 < S < 41, 0.01 < P < 9999, 0°C < T < 40°C, except for WMO boxes with latitudes north of 60°N or south of 50°S, where –2.5°C < T < 40°C.	YES		
5). For WMO boxes that contain more than 10,000 profiles, only select profiles that are post-1995.	YES		
6). Eliminate nearby duplicates.	NO	Only exact duplicates checks are implemented	- Check for and remove near-duplicates and nearby duplicates
7). Do objective residual analysis using previously qc'd reference data to identify anomalies. Then do a visual inspection of anomalies.	Partially	Quality control is made visually using several qc'd reference databases available in-house.	
8). Identify each reference profile with a unique ID, e.g. under the variable *source*.	Partially	IDs are not always unique.	- Make *source* values unique

[6] US National Oceanographic Data Center now National Centers for Environmental Information.

3. THE NORDIC SEAS

Following the definition by Furevik and Nilsen (2005), the Nordic Seas can be found between the Greenland-Scotland Ridge and the Fram Strait-Spitsbergen-northern Norway transect. The Nordic Seas have been monitored with Argo since 2001 and exhibit low natural variability in temperature and salinity in the deeper layers. However, warming and salinification trends have been observed over recent years in the upper 2000 m (Latarius and Quadfasel, 2010; Lauvset et al., 2018). The reported salinification rate for 1000 m and 1500 m depth is of 0.0008 ± 0.0001 PSU year^{-1}, which is near to the order magnitude of the OWC-based salinity corrections applied to Argo floats in the region (10^{-3}). Therefore, to distinguish between artificial and natural trends, the CTD-RDB must include recent profiles.

The region between 60°N and 80°N and 20°E to 20°W comprises most of the Nordic Seas, including WMO boxes 1600, 1601, 1700, 1701, 7600, 7601, 7700 and 7701 (Figure 3). For completeness, the WMO boxes surrounding the Nordic Seas to the West, North, and East were also included, namely WMO boxes 7602, 7702, 7802, 7801, 7800, 1800, 1801, 1802 and 1702.

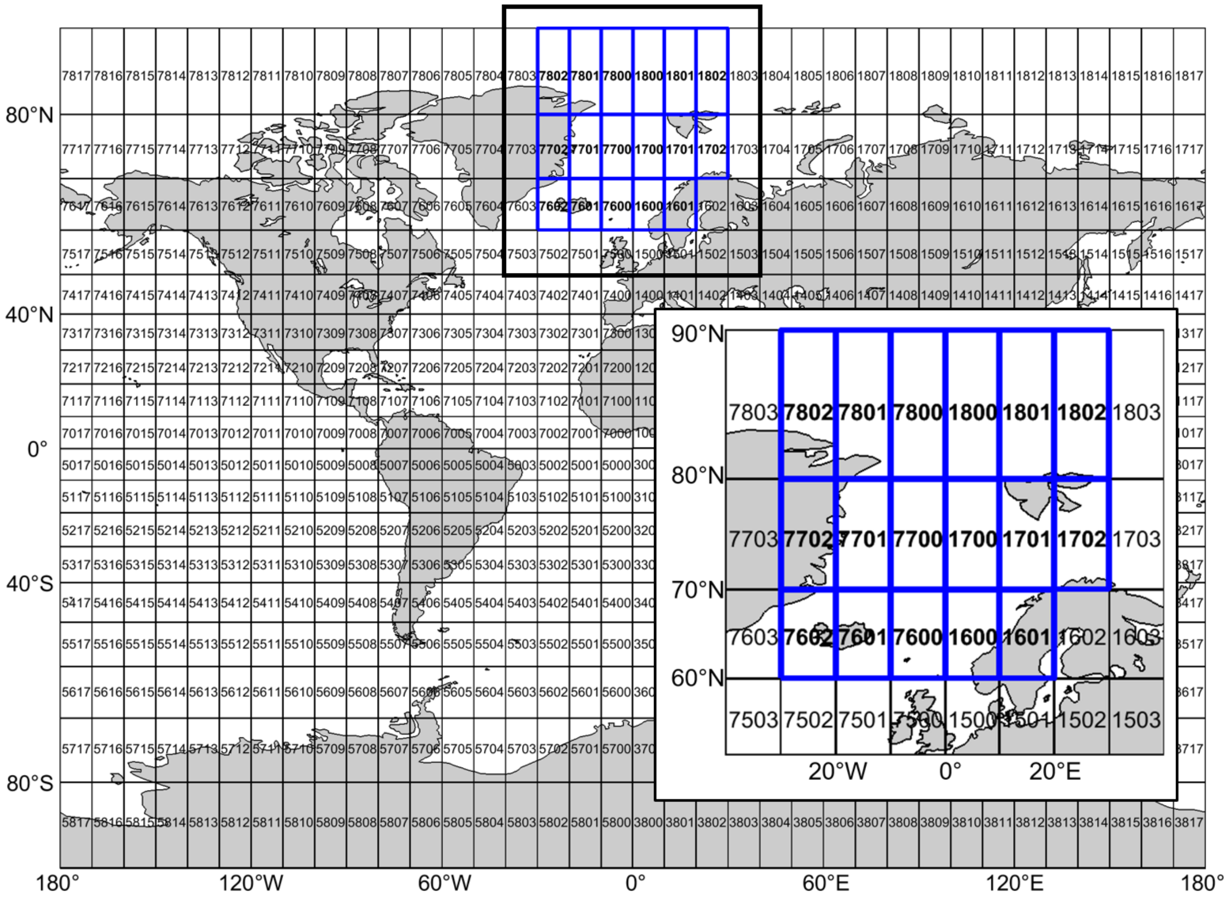

Figure 3 The Nordic Seas WMO boxes (highlighted in blue).

4. CTD-RDB 2018V2

The 17 boxes listed above contain 10509 profiles. However, 1130 profiles are in the North Atlantic Basin (boxes 7601 and 7602) and are masked out for this analysis.

The spatial distribution of the remaining 9460 profiles is shown in Figure 4. Some coastal profiles (142) can be seen in the deep Sognefjorden fjord off the Norwegian coast. Although the maximum recorded pressure selection criterion effectively excludes coastal profiles everywhere else, the fjord is deep enough to pass this criterion and contribute profiles with maximum recorded pressure higher than 900 dbar.

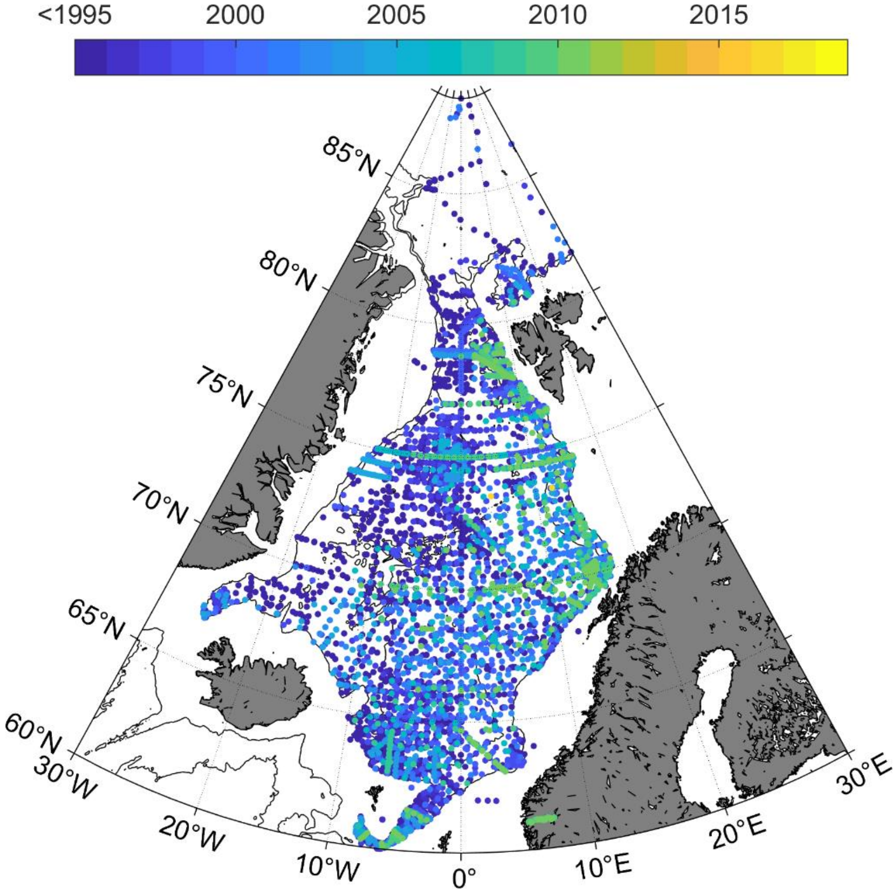

Figure 4 Spatial distribution of the CTD profiles (CTD-RDB 2018v02). The year of sampling is color-coded.

The profile density is shown in Figure 5 using 2-degree square bins. The European Arctic region (north of 82^{o}N) shows a low number of profiles, as expected due to ice presence. From the deep basins, the Iceland Sea and the Norwegian Basin are the ones with the worst and the best coverage, respectively.

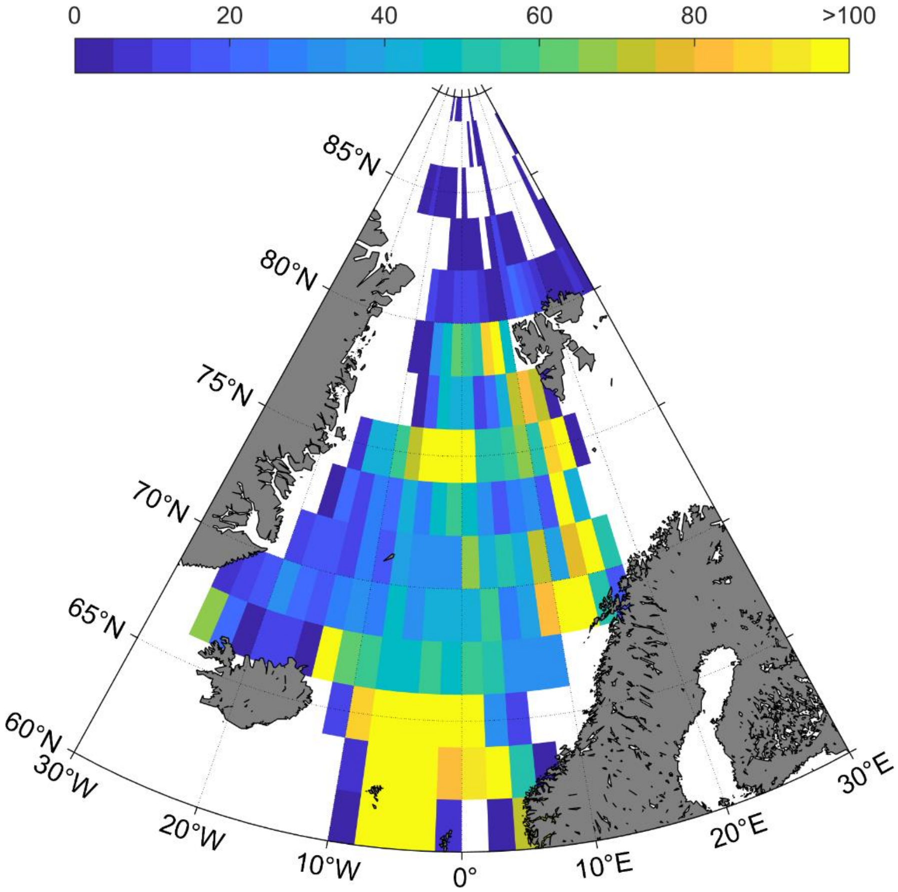

Figure 5 Number of CTD profiles per 2-degree square bin (CTD-RDB 2018v02)

A virtual absence of recent profiles is apparent in Figure 4, where the sampling year is color-coded. This is confirmed in Figure 6, which shows the temporal distribution of the profiles. The oldest profile was collected in 1972 and the number of profiles per year increases until the 90s, reaching a maximum in 2000 (691). Afterwards, the number of profiles decreases considerably. The last year with a relatively large number of profiles is 2011 (252) followed by a gap of 5 years with no data and 2 profiles collected in 2016. Only 15% of the profiles were collected after 2005. The year of the most recent profile for each 2-degree bin is shown in Figure 7.

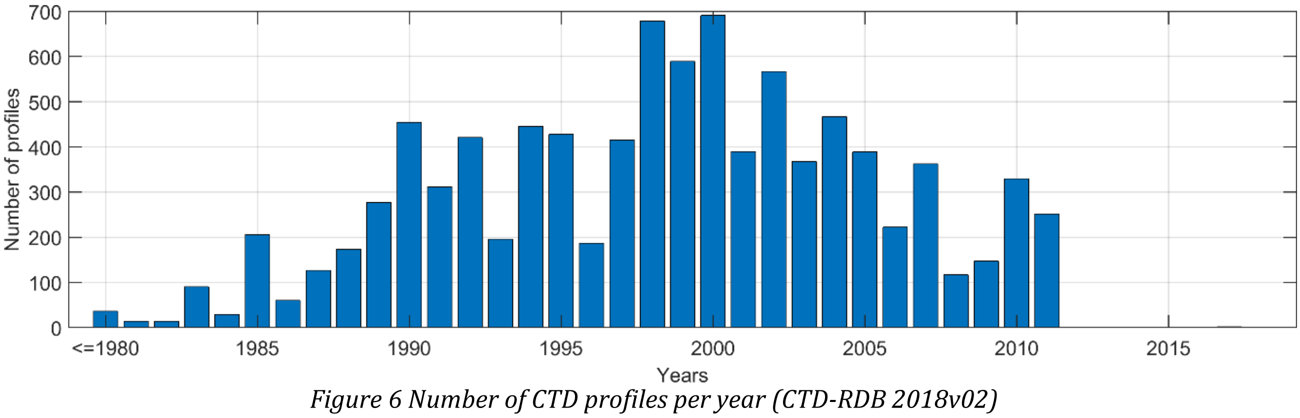

Figure 6 Number of CTD profiles per year (CTD-RDB 2018v02)

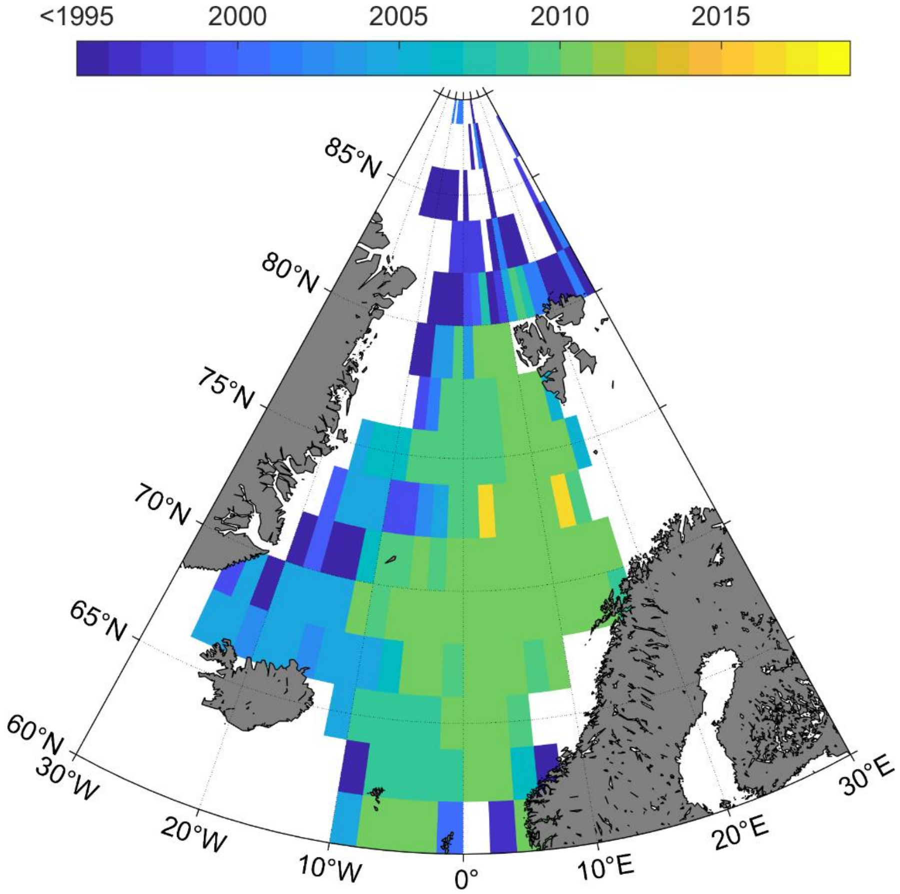

Figure 7 Year of the most recent CTD profile per 2-degree square bin (CTD-RDB 2018v02)

From the 9460 profiles, the *qclevel* of 9445 profiles was COR (Coriolis). Only 13 and 2 profiles had *qclevel* CCH (CCHD) and SPI (obtained directly with scientists), respectively.

The dataset is supposed to include only profiles with maximum recorded pressure larger than 900 dbar. However, 1158 (12%) profiles do not fulfill this requirement and are shown in red in Figure 8. The number increases to 1449 (15%) once samples with incomplete pressure-temperature-salinity triplets are removed. Most of these shallow profiles (1086) correspond to box 7600 and represent 35% of the profiles in that box. The presence of these large numbers of shallow profiles in this box was investigated with C. Coatanoan, and it was traced back to an error that occurred during the preparation of the CTD-RDB 2012v01. This error also implied that many profiles were stored with wrong metadata. Boxes 7601 and 7700 also contain 39 (10.5%) and 33 (2.5%) shallow profiles, respectively.

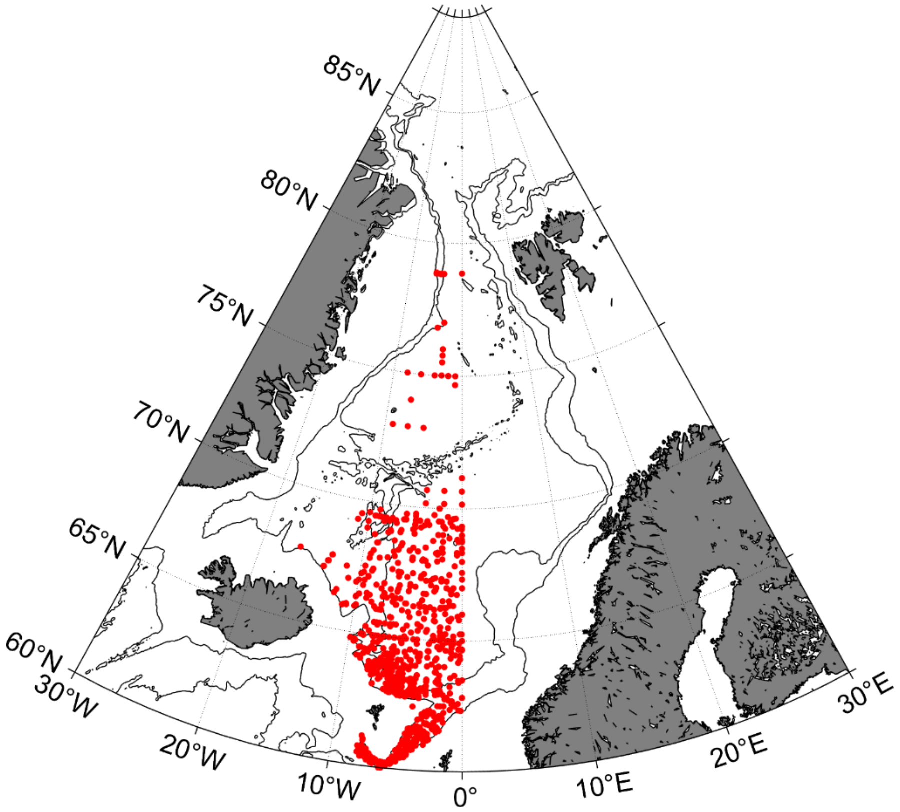

Figure 8 Positions of CTD profiles with maximum recorded pressure < 900 dbar (CTD-RDB 2018v02)

We checked for monotonically increasing pressure in the profiles to detect profiles containing more than one cast. The three profiles that fail this test are shown in Figure 9.

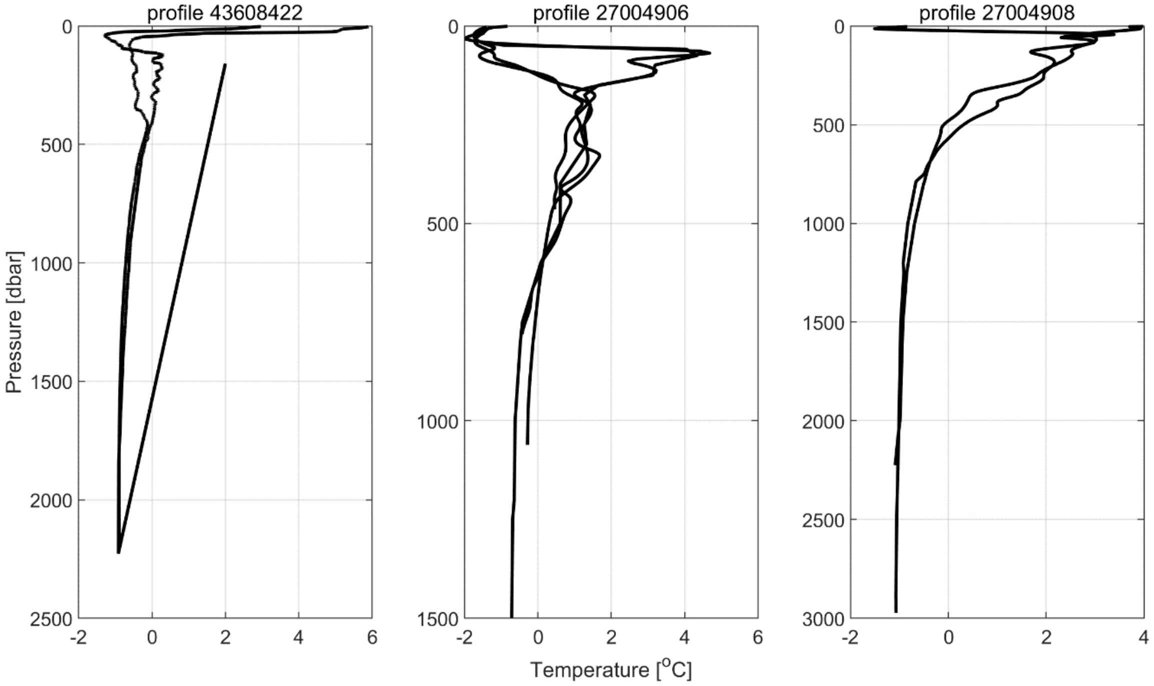

Figure 9 Profiles containing multiple CTD casts (CTD-RDB 2018v02)

To identify duplicates, we compared the profile metadata (*lon, lat, dates,* and *source*). We found seven duplicates.

In five cases one of the pair members was falsely associated with box 7702, which did not correspond to profile positions. For the remaining 2 duplicates, both pair members were found in box 1801.

We further checked if the content of duplicated profiles matched, which was the case for six duplicates. The remaining pair was only a metadata duplicate, having different temperature and salinity profile in each version.

Finally, we checked for exact content duplicates using a test proposed by Gronell and Wijffels (2008). If two profiles are exact content duplicates, their number of pressures, their sum of pressures, their sum of temperatures, and their sum of salinities are identical. Besides the metadata duplicates cited above, one more duplicate was identified. In this case, all the metadata variables were different, and each pair member was found in a different box (1700 and 7700). The *source* codes of these profiles were used by C. Coatanoan to extract the temperature and salinity profiles from the Coriolis. The contents of the CTD-RDB did not match those stored in Coriolis. C. Coatanoan confirmed that the mismatch occurred because the same errors that affected box 7600 during the preparation of the CTD-RDB 2012v01, also affected boxes 7701 and 1700.

5. UPDATE ACTIVITIES

The main issues identified in the CTD-RDB 2018v02 are:

- Some profiles in boxes 1700, 7701, and 7600 had wrong metadata, an error that was traced back to the 2012 update of the database. Also, some profiles are assigned to the wrong box (7702).
- The database is outdated (absence of profiles after 2010).
- Presence of coastal and shallow profiles.
- Presence of duplicated profiles.
- Presence of incomplete samples (either *pres, temp* or *sal* is missing)
- Traceability of the profiles is limited. Most of the profiles are from the Coriolis database and can only be accessed internally at IFREMER.

The following actions were taken to improve and update the CTD-RDB for the Nordic Seas regions.

- Fix the 2018v02 versions of boxes 1700, 7600, and 7701. For this, the boxes were rebuilt by taking their 2011v01 versions as the starting point and adding the updates prepared by C. Coatanoan from 2012 on.
- Improve temporal coverage by adding profiles from alternative data sources:
    - Unified Database for Arctic and Subarctic Hydrography – UDASH (Behrendt at al., 2018): Profiles north of 65^{o}N and collected between 1995 and 2015.
    - International Council for the Exploration of the Sea – ICES: Profiles collected between 2015 and 2019.
- Remove coastal stations and shallow profiles.
- Check for duplicates, including near and content duplicates.
- Remove incomplete samples
- Increase traceability by adding meaningful *source* codes to the profiles added in this update, and when possible, also to profiles already present in CTD-RDB 2018v02.
- Perform final basin-based quality check of salinity values, to remove profiles with suspicious quality.

The flowchart in Figure 10 outlines the actions performed to obtain update the Nordic Seas subset of the CTD-RDB which were incorporated in the current version CTD-RDB 2019v01. The processes and data inside the dotted line where executed/provided by C. Coatanoan at IFREMER. More details about the data preparation and merging procedures are described in the next sections.

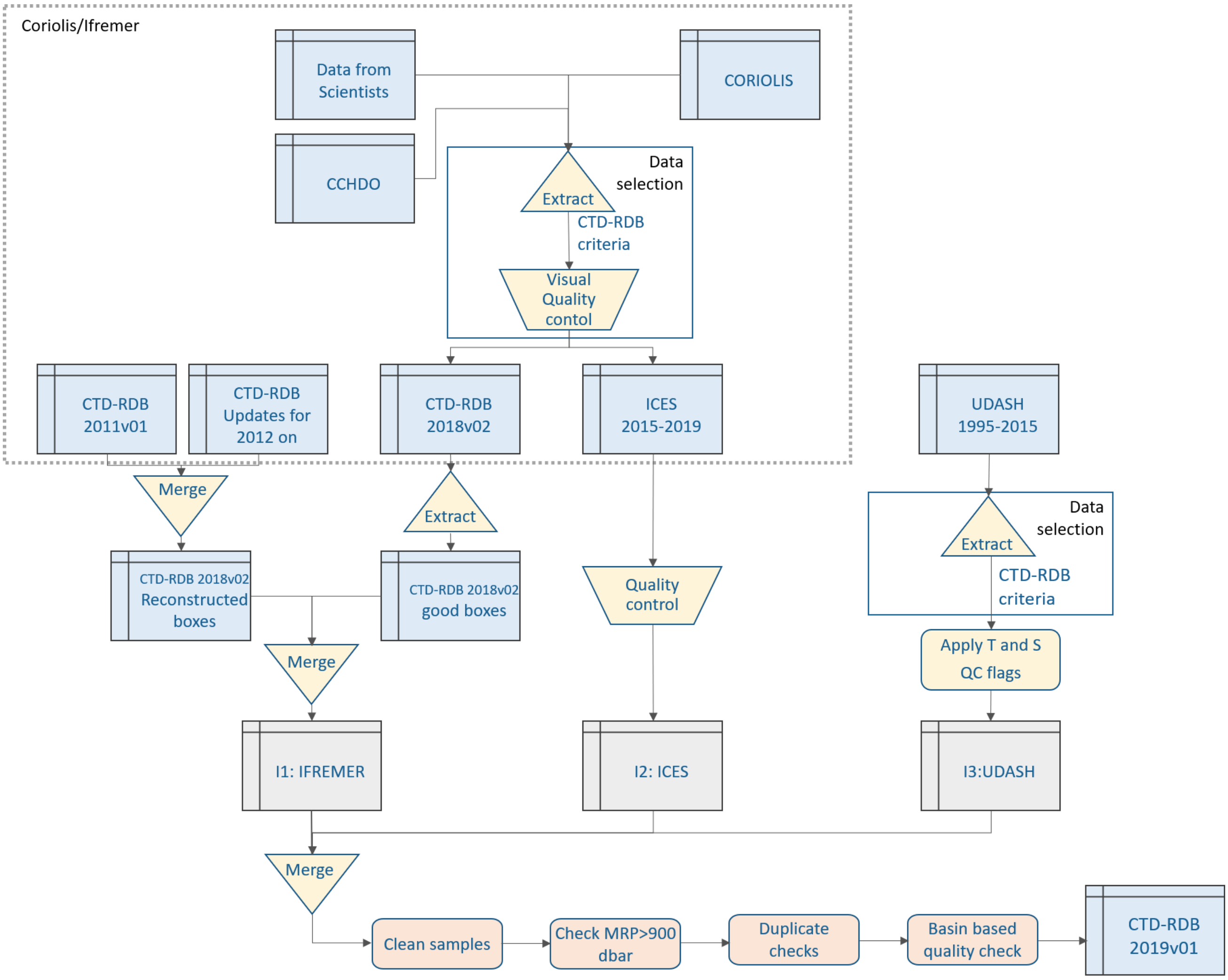

Figure 10 Flowchart of the activities performed to update the Nordic Seas subset of the CTD-RDB.

6. DATA SOURCES AND PREPARATION

As shown in Figure 10, the CTD-RDB 2019v01 was obtained by merging data from three data sources: IFREMER, UDASH, and ICES. Some of the preprocessing steps necessary for each data source are different, due to their different quality levels, and are described below.

6.1. IFREMER

All profiles from the cruise 77DN19910726, originally obtained from CCHDO, had the same *source* value. We corrected this by assigning a unique source value CRUISENAME_ST (ex. 77DN19910726_030), where ST is station number, by comparing the metadata of the profiles in the CTD-RDB and both the NetCDF files and cruise report found in the CCHDO website. In this process, we identified that longitude from station 40 was stored with the wrong sign in both the CTD-RDB and the NetCDF files. This was corrected and communicated to the CCHDO team.

As mentioned in Section 2.1, the default *qclevel* value COR was assigned to all profiles added to the database before the CTD-RDB 2016v01 update, on which the *qclevel* variable was introduced. Since we reconstructed boxes 1700, 7600, and 7701 by updating their 2011v01 versions, all profiles in that version were assigned the default value COR and the updates from the World Ocean Database and CCHDO were labeled appropriately (OCL and CCH).

6.2.UDASH

UDASH dataset combines temperature and salinity profiles from different data sources, instruments, and platforms for the Arctic region (north of 65^{o}N) collected up to 2015 and is publicly available in the PANGAEA data Center. All profiles were subjected to strict quality control tests, aiming to achieve a high and uniform data quality with quality flags assigned to each sample. Duplicate checks were also applied to remove redundant information. All procedures are described in detail in Behrendt et al. (2018).

We selected all ship-borne profiles collected with CTDs from 1995 on. Only samples with temperature and salinity quality flags equal to 0 (best quality) were kept. The potential temperature was calculated using the Gibbs SeaWater (GSW) Oceanographic Toolbox of TEOS-10.

UDASH's main source is the World Ocean Database, accounting for near 80% of the total amount of data, and therefore is expected that many of the profiles selected are already in the IFREMER dataset.

For these profiles, the *source* values are composed by the name of the original database (source variable in the UDASH database, described in Table 1, in Behrendt et al., 2018) and the UDASH profile number.

6.3.ICES

All ICES profiles in the Coriolis Database were selected by C. Coatanoan who performed a quick visual quality control. However, as ICES profiles do not have any sample quality flagging it was necessary to ensure the quality of the added profiles by performing a thorough quality control. Each profile was visualized together with other profiles collected by the same ship in the same WMO box to provide context (Figure 11). Therefore, it was necessary to include the ship identification code in the *source* value of each profile along with the Coriolis internal ID code. In this way, we identified suspicious and bad quality profiles that were deleted and reported to ICES. The most common cause for removal was the presence of large multidirectional spikes in the temperature profiles. The remaining profiles were later examined to flag and remove bad samples (outliers) between 900 and 2000 dbar. The *qclevel* value assigned to these profiles is COR:ICES. These procedures were followed also for those profiles located in the North Atlantic basin of boxes 7601 and 7602.

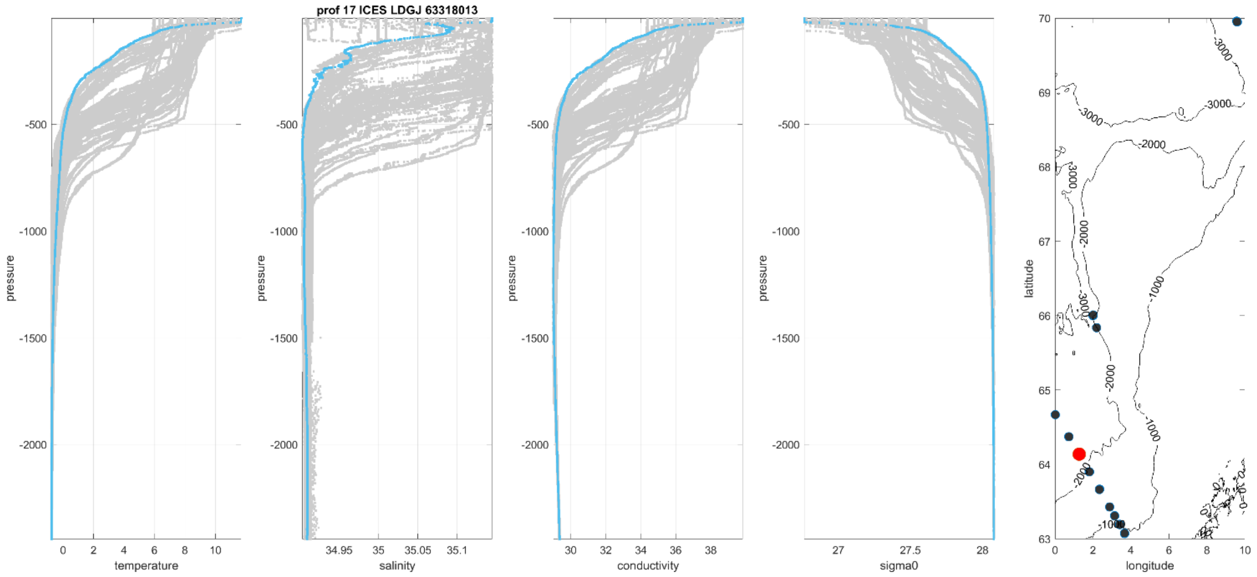

Figure 11 Profile visualization for quality control of ICES data
Example of a profile collected by the ship with code LDGJ in box 1600

7. MERGING AND POST-PROCESSING

7.1. Merging

The data from all databases were merged in the box files. The invalid samples, defined as those for which temperature, salinity or pressure values were missing, were removed from each profile. We removed those profiles that had maximum recorded pressures shallower than 900 dbar after the removal of invalid samples. In the box 1600, the profiles in the Sognefjorden fjord region were deleted.

7.2.Duplicate checks

Duplicate checks are necessary when merging the profiles extracted from the different data sources due to their inherent redundancy. For example, the UDASH database contains data from WOD13 and ICES, which at least partially, were included in the 2018v2 version of the CTD-RDB. Therefore, it can be expected that many profiles are present in more than one data source. It is important to remove these and other duplicates, to avoid data redundancy and hence skewed statistics. The OWC method identifies sensor drifts and anomalies using these profiles in the CTD-RDB, by objectively interpolating the salinity fields into the Argo profiles positions and times over many cycles. Although the OWC software itself does not directly provide statistics about the number of profiles used for the objective interpolation for each cycle, many DMQC operators have developed in house tools to assess the number of profiles available when using different search distances (ellipses) and temporal restrictions. This is necessary because the lower the number of profiles used for the objective interpolation, the more unreliable are the salinity corrections proposed by the OWC method. The presence of duplicated profiles makes this auxiliary information is unreliable. The implications of the presence of duplicated profiles for the objective interpolation itself is negligible because the method accounts for redundant information, except if one of the profiles contains bad quality data (outliers). Thus, it is also important to select the best quality profile when removing duplicated profiles that have gone through different subsampling and quality screenings, to preserve the highest amount of information and avoid the presence of bad quality data. Two types of duplicates must be considered: Metadata and content duplicates.

7.2.1. Metadata duplicates

Two profiles are metadata exact duplicates when they have the exact same position and date, or metadata near-duplicates (or nearby duplicates) when they have almost the same position and time. The detection of exact metadata duplicates is straightforward, by comparing the profiles longitude, latitude, and timestamp in their full precision. However, near metadata duplicates requires the definition of the timespans and distances that are considered “near”.

7.2.2. Content duplicates

Two profiles are content exact duplicates when their temperature and salinity values are the same and correspond to the same pressure levels. If two profiles show only small differences in the pressure, temperature, and salinity values, they are content near duplicates. These content near-duplicates are different versions of the same profile and arise when the profiles were:

a) subsampled or interpolated to different vertical resolutions
b) trimmed to different pressure ranges

c) quality controlled using different criteria, which lead to sample removal (exclusion of bad quality samples)
d) saved with a different number of decimal digits resulting in temperature and salinity values have different precisions.

For the detection of content duplicates, we follow the approach proposed by Gronell and Wijffels (2008). If two profiles are exact content duplicates their number of pressures, the sum of pressures and the sum of temperatures and salinities of all samples are identical. This is further confirmed by a sample-by-sample comparison.

The detection of near content duplicates requires preprocessing of the profiles to account for the differences in the sampling pressure levels between the profiles, and thus make them comparable sample-by-sample. In Gronell and Wijffels (2008) the comparison test uses rounding and truncation in both depth and temperature to obtain profiles with a degraded but common vertical resolution and precision. Here, inspired in the implementation of Gronell and Wijffels (2008) algorithm in Behrendt et al. (2018): a) the profile with the highest vertical resolution is interpolated to the pressure levels of the one with the lowest resolution at the overlapping pressure levels, accounting for the different sampling pressure levels; and b) the precision of the temperature and salinity values is degraded. If more than 95% of the such preprocessed temperature and salinity samples are equal, the profiles are automatically labeled as content duplicates If more than 75% are equal, the operator must confirm the duplicate by examining the profiles visually. The workflow is shown in Figure 12.

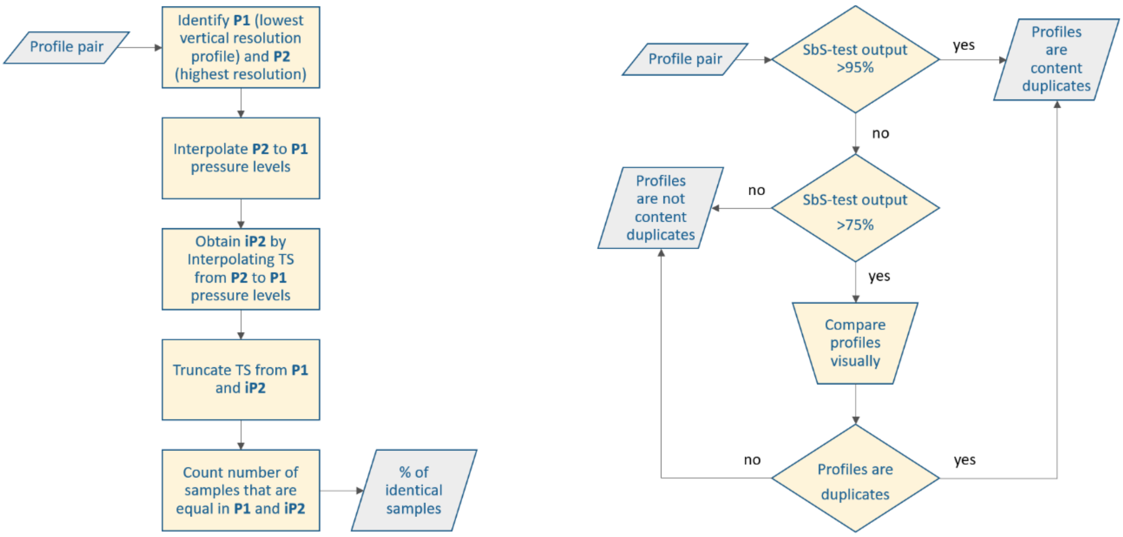
Profile pair
Identify **P1** (lowest vertical resolution profile) and **P2** (highest resolution)
Interpolate **P2** to **P1** pressure levels
Obtain **iP2** by Interpolating TS from **P2** to **P1** pressure levels
Truncate TS from **P1** and **iP2**
Count number of samples that are equal in **P1** and **iP2**
% of identical samples
Profile pair
SbS-test output >95%
yes
Profiles are content duplicates
no
SbS-test output >75%
no
Profiles are not content duplicates
yes
Compare profiles visually
no
Profiles are duplicates
yes

Figure 12 Flowchart for the identification of content duplicates.
Left panel: Sample by sample (SbS-test). Right Panel: Identification of content duplicates

The content near-duplicate test described above is computationally heavy and has been used in Gronell and Wijffels (2008) and Behrendt et al. (2018) only to check for nearby duplicates, those profiles near in time and space. However, a comparison of all profiles is desirable because it allows finding the profiles that have been mislabeled and are therefore in the wrong position. To be able to handle a large number of profiles, we split the near content duplicate check in two parts. First, we identify profile pairs that are likely to be content duplicates (Figure 13), and then these possible duplicates are either confirmed or disproved using the sample-by-sample comparison algorithm describe above. Therefore, the profiles were interpolated to common pressure levels and truncated to 1 and 2 decimal digits for temperature and salinity, respectively. Then, we compare the profiles using the exact content duplicate test. If the sums of temperature and salinity are equal these profiles as marked as possible duplicates.

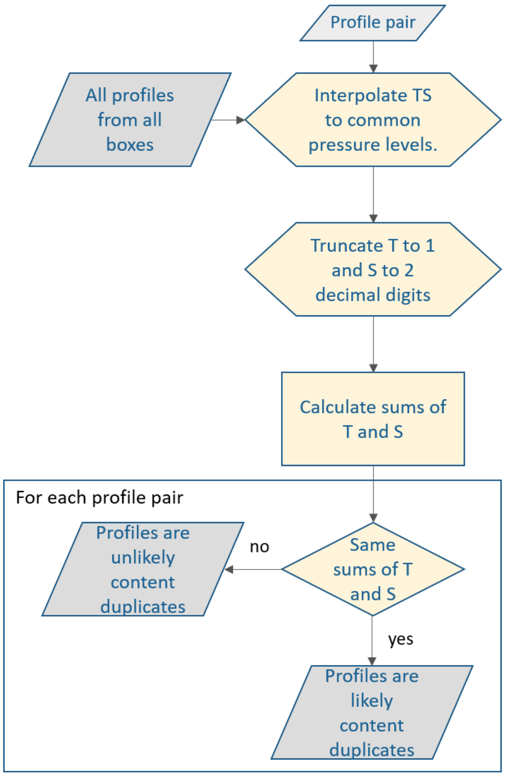
Profile pair
All profiles from all boxes
Interpolate TS to common pressure levels.
Truncate T to 1 and S to 2 decimal digits
Calculate sums of T and S
For each profile pair
Profiles are unlikely content duplicates
no
Same sums of T and S
yes
Profiles are likely content duplicates

Figure 13 Flowchart for the identification of profile pairs that are likely content duplicates

7.2.3. Identifying the best copy of a profile

One must decide which profile in the content duplicate pair to keep/exclude, i.e. the best and worst copy. We use two criteria: the information content and the information about the origin of the profiles, giving priority to the first. If the criteria comparison delivers no decision, meaning that the profiles are exact duplicates, we prefer to keep the profile most recently added (Figure 14)

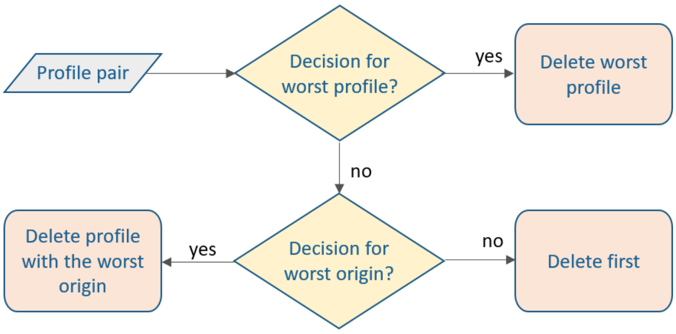
Profile pair
Decision for worst profile?
yes
Delete worst profile
no
Delete profile with the worst origin
yes
Decision for worst origin?
no
Delete first

Figure 14 Flowchart for deciding which profile to delete

We use three metrics to compare the information content of the profiles (Figure 15):

- Maximum recorded pressure. Prevails only if the difference is larger than 50 dbar.
- Salinity precision: Number of decimal digits
- Vertical resolution: Number of samples/pressure range

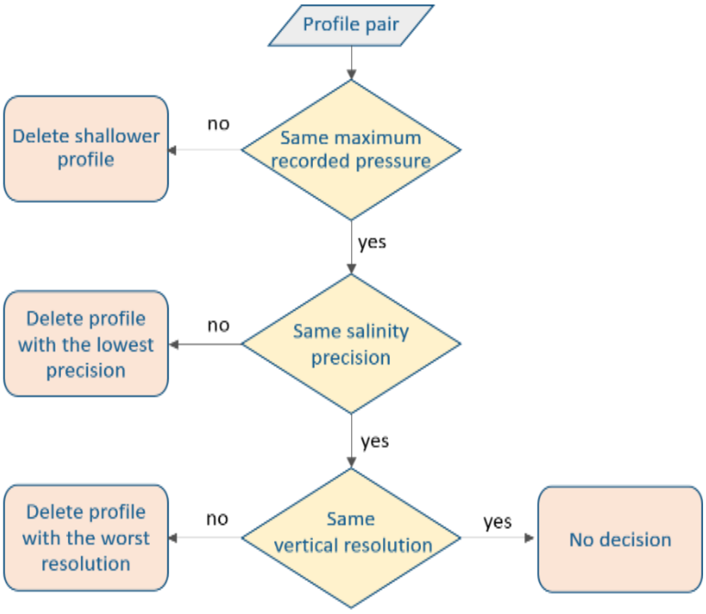
Profile pair
Same maximum recorded pressure
no
Delete shallower profile
yes
Same salinity precision
no
Delete profile with the lowest precision
yes
Same vertical resolution
no
Delete profile with the worst resolution
yes
No decision

Figure 15 Flowchart for deciding which profile to delete according to its information content

For the profile origin, we should keep the profiles with higher quality control. Thus, profiles with *qclevel* UDASH and COR:ICES, which were subjected to mostly automated and manual quality control respectively, should be preferred to those with *qclevel* COR and OCL. However, a code error resulted in the opposite selection. We will fix this bug in the next update of the database. For duplicated profiles with qclevel COR or OCL, which *source* value is the internal Coriolis ID, the profile with the higher source number is preferred, as they entered later in the Coriolis database, likely due to resubmission, which we assume implies an improved quality. Inside the Coriolis database, the older profile is removed, but they were not removed from the CTD-RDB. The workflow is shown in Figure 16.

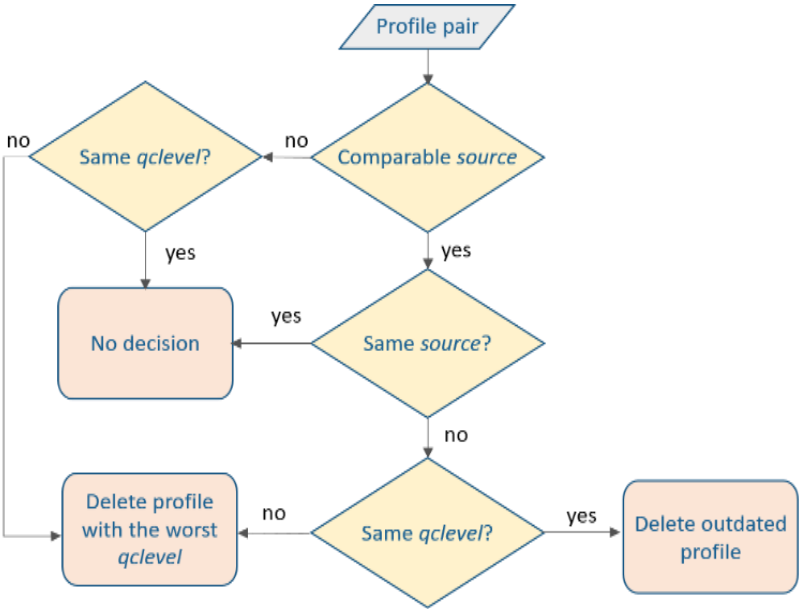

Figure 16 Flowchart for deciding which profile to delete according to its origin

7.2.4. Workflow for duplicate checks

The duplicate check consists of three consecutive steps:

- Check for exact metadata duplicates in each box (Figure 17). If the pair is also a content duplicate delete the worst profile. If the pair is not a content duplicate, both profiles we delete both profiles because their metadata/content is uncertain.

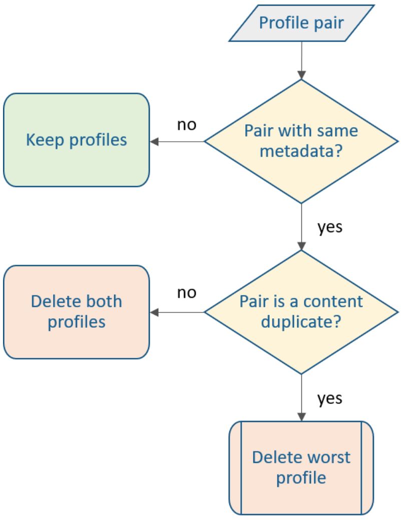

Figure 17 Flowchart for metadata exact duplicate check

- Check for metadata near duplicates in each box (Figure 18). Here we compare the rounded/truncated variables down to one decimal digit for latitude and longitude, and 1 day for the timestamp. If the pair is also a content duplicate, we delete the worst profile. Otherwise, both profiles are kept.

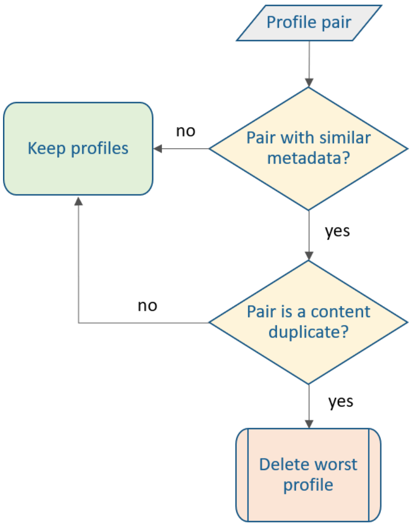

Figure 18 Flowchart for metadata near-duplicate check

- Check for content duplicates in all boxes (Figure 19). Find profile pairs likely to be content duplicates. Confirm content duplicate with a sample-by-sample test. For content duplicates that are near in time and space (distance smaller than 3 km and time difference shorter than 3 days) delete the worst profile. If they are far in time or space, we delete both profiles because their metadata is uncertain.

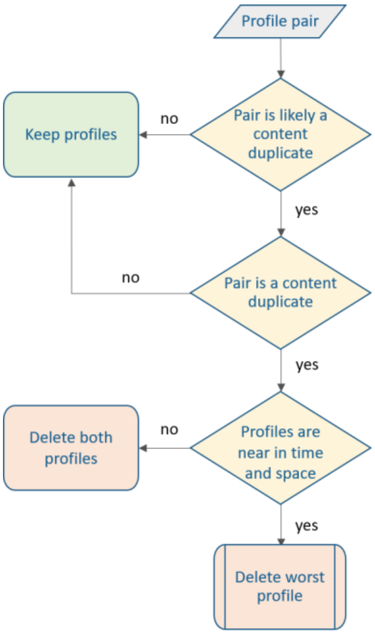
Profile pair
Pair is likely a content duplicate
no
Keep profiles
yes
Pair is a content duplicate
no
yes
Profiles are near in time and space
no
Delete both profiles
yes
Delete worst profile

Figure 19 Flowchart for content duplicate check

7.3. **Final quality control**

To check for outliers, we interpolated the salinity to 900 dbar inside each one of the four deep basins, which limits were defined using a combination of the geographical constraints and their characteristic *f/H* ratio, where *f* is the planetary vorticity and *H* is the water depth, as in Latarius and Quadfasel (2010). Figure 1 shows the *f/H*-characteristic as black contour lines. The *f/H* threshold is 0.079 for the Iceland Sea (IS) and 0.045 for the Greenland Sea (GS), the Lofoten Basin (LB), and the Norwegian Basin (NB).

Figure 20 shows the time series of the interpolated values for the four basins. The data points highlighted with the red circles were considered outliers and the profiles from which they originated were excluded (boxes 1600, 7600, 7601).

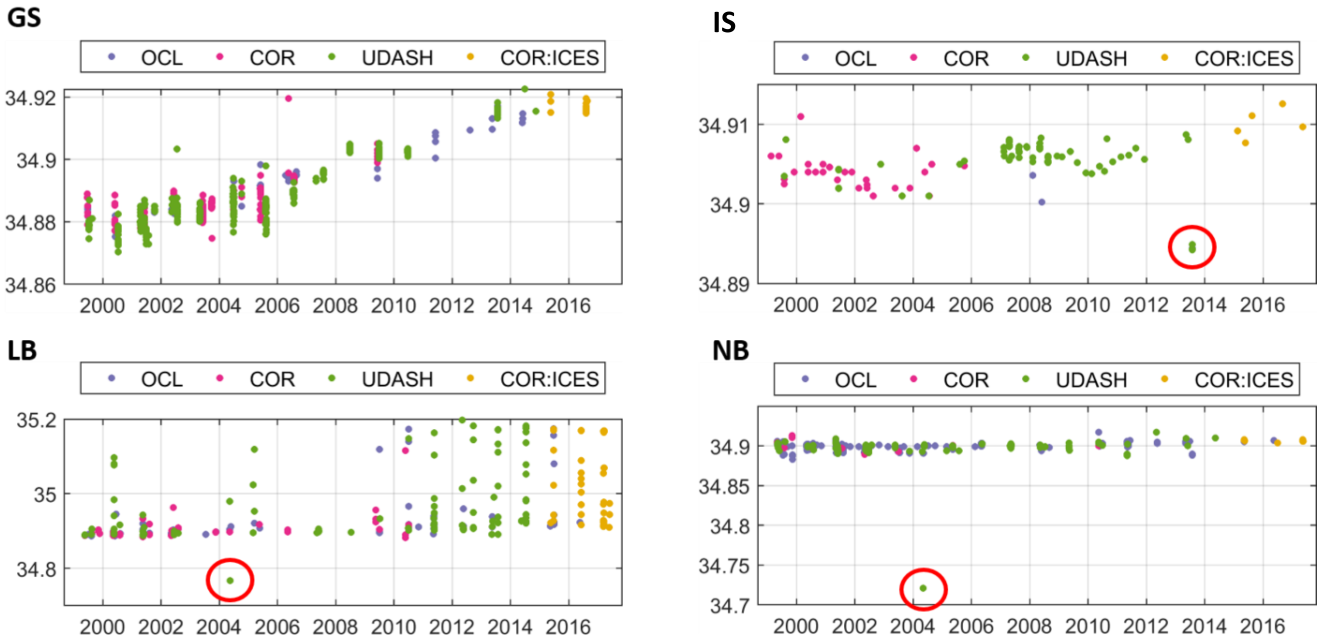

Figure 20 Time series of salinity interpolated to 900 dbar in the deep basins of the Nordic Seas

Given that in all cases the *qclevel* of the profiles is UDASH, we traced back the origin of the profiles. In the Iceland Sea, the four profiles are from the ship Haakon Mosby, one obtained from ICES and three from the WOD13. From the latter, we could identify that they correspond to cruise NO-5474. For the Lofoten and Norwegian Basins, the outliers correspond to one and three profiles respectively. They were all obtained from the WOD13 and collected with the ship DANA and cruise DK-2953. We reported the dubious quality of these profiles to the WOD13 team. In the future, we should consider the deletion of all data from these problematic campaigns from the database, to ensure the quality of the database. Weeding out entire campaigns is only possible if the *source* values of each profile are meaningful, which highlights the importance of increase the traceability of the CTD-RDB in facilitating the assessment and improvement of its quality.

8. CTD-RDB 2019V01

The updated version of the CTD-RDB (2019v01), obtained with the procedures described in this report was released in October 2019. The 17 WMO boxes contain 15319 profiles. The 884 profiles are in the North Atlantic Basin (boxes 7601 and 7602) and are masked out since the report is focused on the Nordic Seas. The spatial distribution of the remaining 14340 profiles is shown in Figure 21. This represents an increase of 4880 profiles when compared with the CTD2018v02.

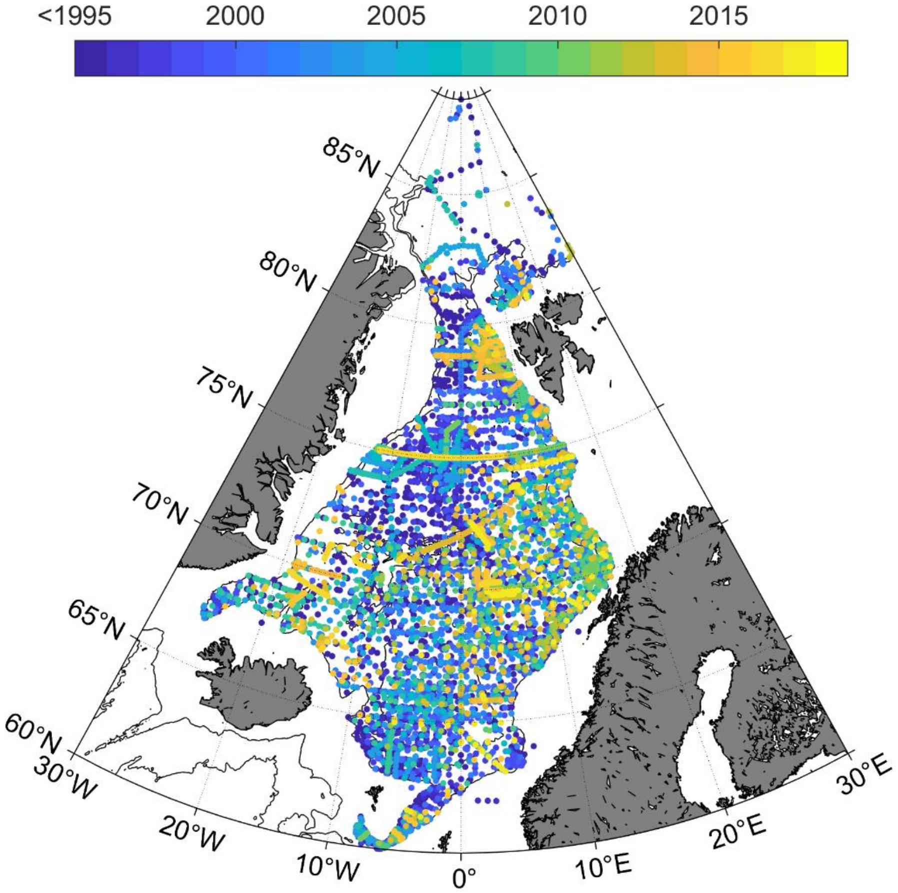

Figure 21 Spatial distribution of the CTD profiles (CTD-RDB 2019v01). The year of sampling is color-coded.

The profile density is shown in Figure 22 using 2-degree square bins. The increase in the number of profiles can be seen in the entire region but the coverage increases particularly in the Lofoten Basin and the Greenland Sea.

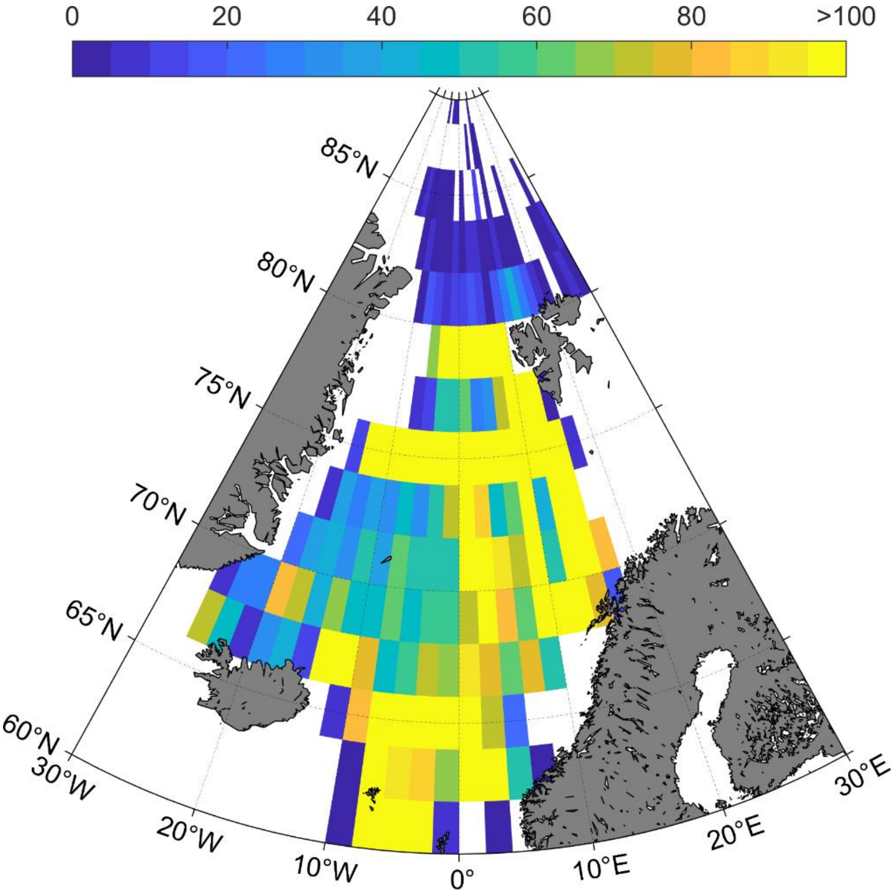

Figure 22 Number of CTD profiles per 2-degree square bin (CTD-RDB 2019v01)

The temporal distribution of the profiles has considerably improved, as seen in Figure 23 that shows the temporal distribution of the profiles. While the CTD-RDB 2018v02 only two profiles were collected after 2012, a total of 1592 profiles are present in the CTD-RDB 2019v01. The absence of profiles after 2017 may reflect the time lag between the time of measurement and the submission of the data to the public databases. A lack of recent profiles in the Arctic (north of 80^{o}N) persist, as shown in Figure 24, were the year of the most recent profile for each 2-degree bin is depicted.

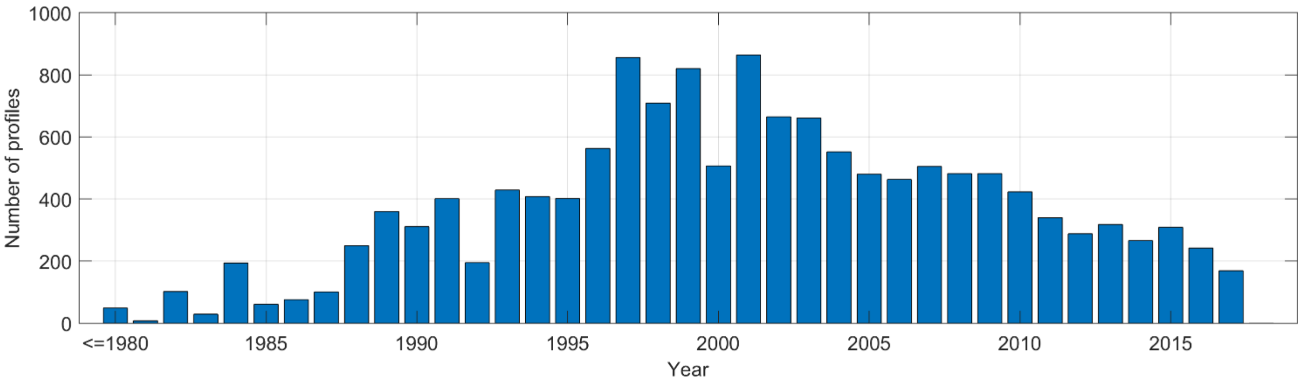

Figure 23 Number of CTD profiles per year (CTD-RDB 2019 v01)

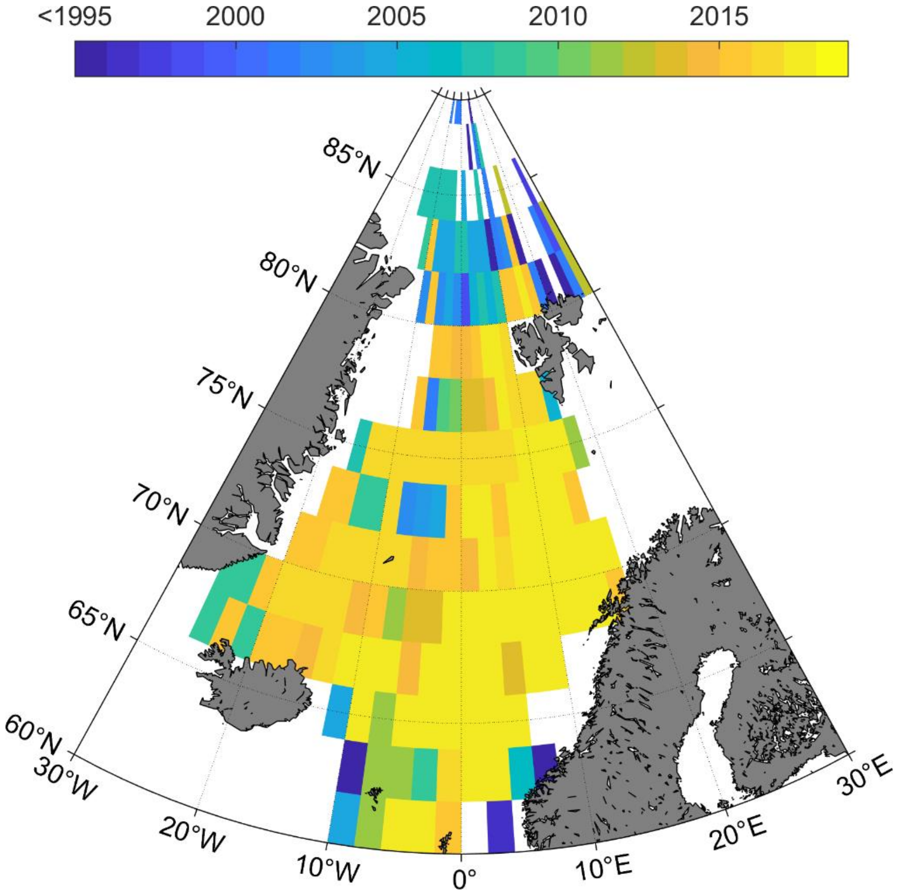

Figure 24 Year of the most recent CTD profile per 2-degree square bin (CTD-RDB 2019v01)

Due to the changes in the *qclevel*, performed while assigning the correct values to the CCHDO cruise 77DN19910726 and the reconstruction of boxes 1700, 7600, and 7701, the number of profiles labeled as CCH and OCL increased from 13 and 0 in CTD-RDB 2018v02 to 31 and 3835 in CTD-RDB 2019v01 respectively. From the newly added profiles, a total of 663 profiles come from ICES (COR:ICES) and 5414 from UDASH.

9. OUTLOOK

Given the importance of the appropriate temporal and spatial coverage of the CTD-RDB in the performance of the OWC method, all DMQC operators should be encouraged to check the status of the database in their regions of interest. A Matlab tool was created for this purpose and can be found in *https://github.com/euroargodev/check_CTD-RDB*. Based on the output of this initial diagnosis, the operators can identify gaps and contribute profiles to the database.

Aiming to further improve the CTD-RDB at the global level, the scripts used for duplicate and other quality checks will be implemented by C. Coatanoan for the next global updates. This work will be part of the EA-RISE project and will be shared with the Argo community via the Euro Argo collaborative framework in Github.

In the regional level, CTD profiles obtained directly from the EA-RISE partners in Poland (Institute of Oceanology of the *Polish* Academy of Sciences - IOPAN) and Norway (Institute of Marine Research - IMR) will be added to the next version of the CTD-RDB. Moreover, given the EA-RISE objective of increase Argo coverage in regions shallower than 900 m, such as the East Greenland Current and the Arctic, the DMQC procedures will need reference profiles with maximum recorded pressure shallower than 900 dbar which are currently not included in the database. A reference database will be built for this purpose.

10. BIBLIOGRAPHY

Behrendt A., H. Sumata, B. Rabe, and U. Schauer. 2017: UDASH - Unified Database for Arctic and Subarctic Hydrography. Earth Syst. Sci. Data, 10, pp 1119–1138. https://doi.org/10.5194/essd-10-1119-2018

Gronell, A., and S. E. Wijffels. 2008: A semiautomated approach for quality controlling large historical ocean temperature archives, J. Atmos. Ocean. Tech., 25, pp. 990–1003. https://doi.org/10.1175/JTECHO539.1

Latarius K. and D. Quadfasel. 2016: Seasonal to inter-annual variability of temperature and salinity in the Greenland Sea Gyre: heat and freshwater budgets, Tellus A: Dynamic Meteorology and Oceanography, 62:4, pp. 497-515. https://doi.org/10.1016/j.dsr.2016.04.012

Lauvset S.K., A. Brakstad, K. Våge, A. Olsen, E. Jeansson, and K.A. Mork. 2018: Continued warming, salinification and oxygenation of the Greenland Sea gyre, Tellus A: Dynamic Meteorology and Oceanography, 70:1, 1-9.10.1080/16000870.2018.1476434.

Cabanes, C., V. Thierry, and C. Lagadec. 2016: Improvement of bias detection in Argo float conductivity sensors and its application in the North Atlantic. Deep Sea Research Part I: Oceanographic Research Papers. 114. 10.1016/j.dsr.2016.05.007.

Furevik, T., and J.E. Nilsen. 2005: Large-Scale Atmospheric Circulation Variability and its Impacts on the Nordic Seas Ocean Climate-A Review. Washington DC American Geophysical Union Geophysical Monograph Series. 158. 105-136. 10.1029/158GM09.

Owens, W., and A. Wong, Annie. 2009: An improved calibration method for the drift of the conductivity sensor on autonomous CTD profiling floats by θ-S climatology. Deep Sea Research Part I: Oceanographic Research Papers. 56. 10.1016/j.dsr.2008.09.008.