QUALITY INFORMATION DOCUMENT

In Situ TAC
INSITU_GLO_PHY_TS_DISCRETE_MY_013_001

Issue: 1.14

Contributors: T. Szekely

Approval date by the Copernicus Marine Service product quality coordination team: 29/11/2023

CHANGE RECORD

Issue	Date	§	Description of Change	Author	Checked By
1.0	20/06/2013	All	First version of document	Antoine GROUAZEL	S Pouliquen
1.1	20/12/2013		Update for V4.0 or CORA-GLOBAL	Antoine GROUAZEL	S Pouliquen
1.2	14/02/2014		Update after QUARG review	S Pouliquen	
1.3	01/02/2014		Update for CORA-GLOBAL V4.1	T. Szekely	
1.4	10/02/2015	All	Revision after V5 acceptance	T. Szekely	
1.5	May 1 2015	All	Change format to fit CMEMS graphical rules		L. Crosnier
1.6	19/01/2016		Update for CORA-GLOBAL V4.2	T. Szekely	
1.7	14/01/2017	All	Update for CORA 5.0	T.Szekely	
1.8	05/03/2018	All	Update for CORA 5.1	T.Szekely	
1.9	04/04/2019	All	Update for CORA 5.2 and Easy CORA dataset	T.Szekely, P. Roltan	Mercator Ocean
1.10	11/09/2020		Switch to the Netcdf file format	Tanguy Szekely	
1.11	24/08/2021		Major release update	Tanguy Szekely	
1.12	10/02/2022		Minor release update	Tanguy Szekely	S. Tarot
1.13	25/11/2022		New template Product name change Full reprocessing	Tanguy Szekely	S. Tarot
1.14	28/08/2023		Major release update Product extension up to December 2022	Tanguy Szekely	S.Tarot

TABLE OF CONTENTS

I EXECUTIVE SUMMARY

I.1 Products covered by this document

INSITU_GLO_PHY_TS_DISCRETE_MY_013_001 or CORA (Coriolis Ocean Dataset for Reanalysis), also called "Global Ocean-Full CORA- in situ Observations Yearly Delivery in Delayed Mode" is a global dataset of ocean in situ sea temperature and salinity (T/S) measurements.

This document aims to give a detailed picture of the processes and tools used to validate this dataset.

Short Description	Product code	Area	Delivery Time
Global MY	INSITU_GLO_PHY_TS_DISCRETE_MY_013_001	Global	Yearly

"MY" stands for MultiYear product or reprocessed product.

The **INSITU_GLO_PHY_TS_DISCRETE_MY_013_001** product is made up of 2 different datasets:

- The CORA dataset including the full profiles and full metadata covering the global ocean, with distinct data distribution for the In-Situ TAC regions (including global ocean and the European regions). See Fig. 1 for a list of the regions.
- A dataset derived from CORA, the Easy CORA dataset, with a simplified data structure and format.

I.2 Summary of the results

The whole validation and update process for this product is performed twice a year:

- in June for a temporal extension of six months
- November for a temporal extension of six months and a full reprocessing of the complete time series that includes modifications and improvements of the product

The quality of the observations is tested using automatic procedures and a comparison to climatology. Quality flags are set to inform the users about the level of confidence attached to each observation.

The observations are aggregated by the In Situ Thematic Assembly center (In Situ TAC). The In Situ TAC will be described in section II.1. The In Situ TAC relies on observing systems maintained by institutes that are not part of the In Situ TAC. Moreover, the Copernicus marine service is not contributing to the maintenance and setting up of the observing systems it uses. The In Situ TAC have thus a limited impact on the instrument deployments and maintenance.

I.3 Estimated Accuracy Numbers

The accuracy of the in situ observations depends on the platforms and sensors that have been used to acquire them.

- The variety of platforms available to monitor the status of the ocean is very diverse within the different regions.
- In some regions the number of available platforms is at a critically low level to provide an adequate representative overall view of the state of the ocean
- Some data are obtained by regular vessel cruises or dedicated scientific expeditions. The availability of data from these scientific expeditions is often delayed so they are not available for the real time dataset, consequently not available for assimilation of the operational models. When available, these data are added in the multiyear product during the yearly reprocessing.
- This product is based on the delayed time mode validation of a frozen copy of the INSITU_GLO_PHYBGCWAV_DISCRETE_MYNRT_013_030 product. The source product is updated in near real time. Consequently, some profiles distributed in the CORA product may have been updated at the distribution date and may be available with a different update date and data mode and flags in the NRT product.
- The percentage of data flagged as 'good data' varies from region to region.

Table 1 summarizes the accuracy of the T/S measurements that can be expected depending on the platforms and sensors. The definition of the reference values is obtained from different sources. The platform specific references that differ from the common ones are given.

Data-type	**Temperature[1] [°C]**	**Salinity[1] [PSU]**
CTD (Conductivity, Temperature, Depth sensor)	0.005-0.001	0.02-0.003
XBT (Expendable Bathythermograph)	0.1	
XCTD (Expendable CTD sensor)	0.02	0.003
PFL (profiling floats)	0.01[2]	0.01[2]
Moored buoy data:	0.002	0.003
Drifting buoy data	0.01	0.01
Marine mammals	0.005	0.01
Glider	0.005	0.02
Underway (Ferrybox, Research vessel TSG)	0.001-0.1[3]	0.003-0.2[3]

Table 1 Accuracy numbers for temperature and salinity observations for the different platforms. Data is obtained from in the In Situ TAC, NOAA 2009 and ARGO Buoys depending on sensor type.

II PRODUCTION SYSTEM DESCRIPTION

II.1 Production system description

The In Situ TAC is a distributed center organized around 7 oceanographic regions: the global ocean and the 6 EUROGOOS (European Global Ocean Observing System) regional alliances (see Figure 1). The TAC involves 17 partners from 11 countries in Europe. It does not deploy any observing system itself and relies on data that are exclusively funded by sources other than the Copernicus Marine Service.

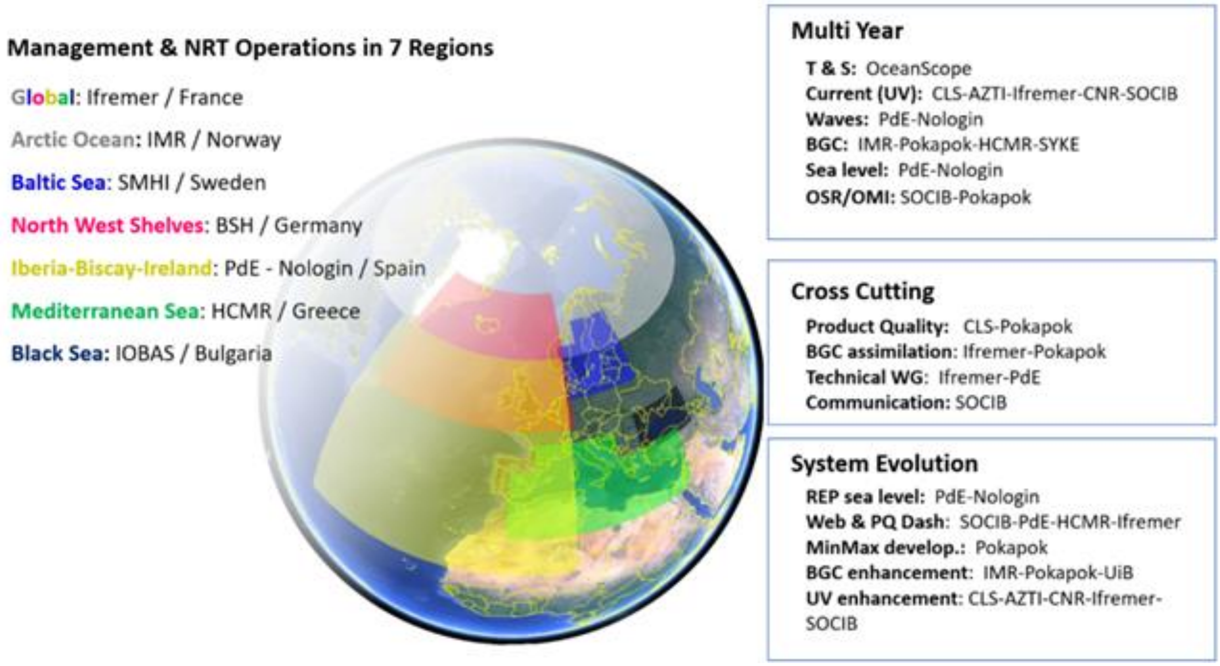


Figure 1: The In Situ TAC (regions map), components and partners. The CORA dataset refers to the Multi year T&S product.

The In Situ TAC architecture is decentralized. However, the quality of the products delivered to users is equivalent irrespective of where the data are processed. The different functions implemented by the global and regional components of the In Situ TAC are summarized in *Figure* 2.

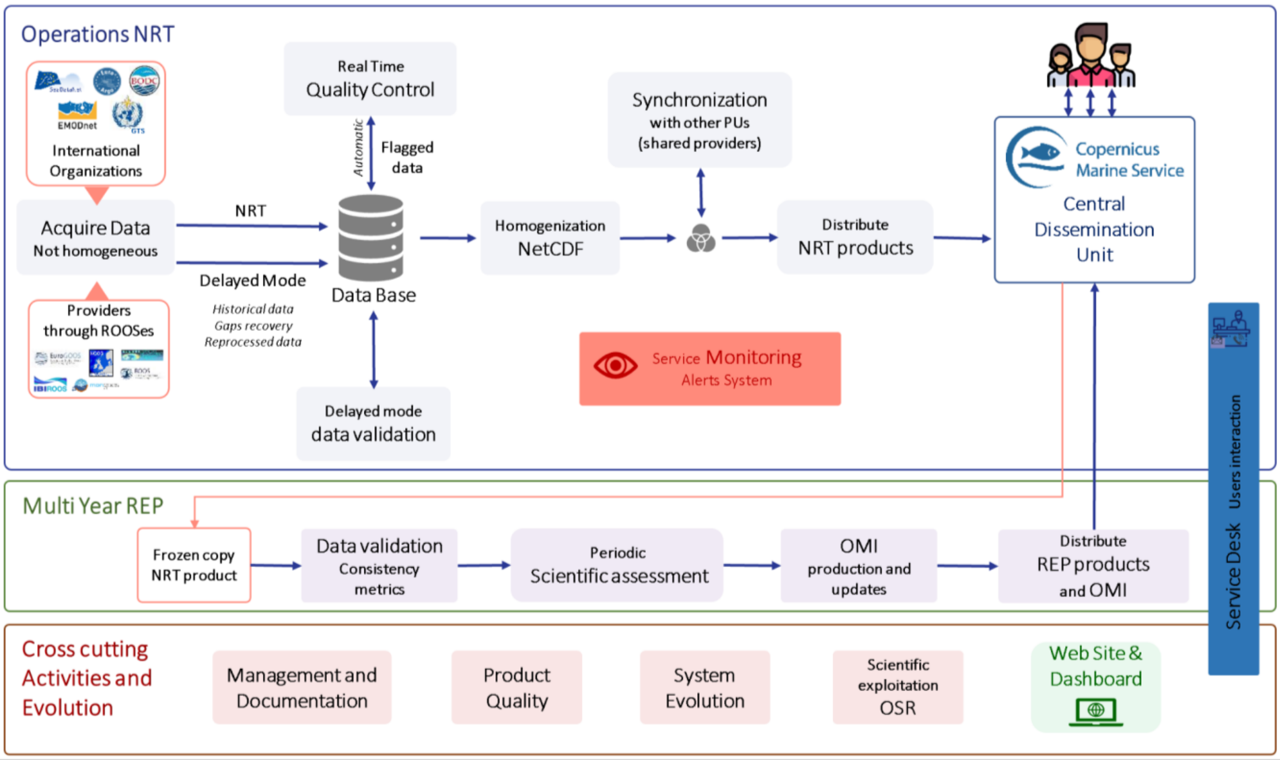


Figure 2: Functions implemented by the In Situ TAC components. For the CORA dataset (T&S multi-year product, the data is acquired by scientific teams and international organizations (ARGO, etc.). The T&S data are validated in near real time and distributed by the Copernicus Marine service central dissemination unit (DU). The present dataset is a frozen copy of the NRT dataset, validated in delayed time mode and distributedvia the DU.

Each of the 7 regions implements 4 core functions:

- Data Acquisition: Gather data available on international networks or through collaboration with regional partners
- Data Quality control: Apply automatic quality controls that have been agreed at the In Situ TAC level. These procedures are defined by parameter, developed in coherence with international agreements, in particular SeaDataNet (https://www.seadatanet.org/).
- Product validation: Assess the consistency of the data over a period of time and an area to detect data that are not coherent with their neighbours but could not be detected by automatic Quality Control (QC)
- Product distribution: Push the data available to the Copernicus Marine Service Distribution Unit (DU). The DU is responsible for data distribution to users.

In any case, the Global component of the In Situ TAC collects the data from the regional components and integrates them into the global product acting as a backup of the regional centers.

II.2 Production subsystem production

CORA dataset production time schedule

The CORA dataset has two yearly release dates. A major release each November, which covers the period from January 1950 to December of the previous year, and a minor update in June, covering January to June of previous year. For instance, the November 2023 release will cover the data from January 1950 to December 2022. The June 2024 release will cover the period from January to June 2023. The minor and major update are based on frozen copies (A copy of the NRT dataset at a given date. See Fig 2) of the INSITU_GLO_PHYBGCWAV_DISCRETE_MYNRT_013_030 product generated in October and March, respectively. Figure 3 gives a schematic overview of the product production and release schedule.

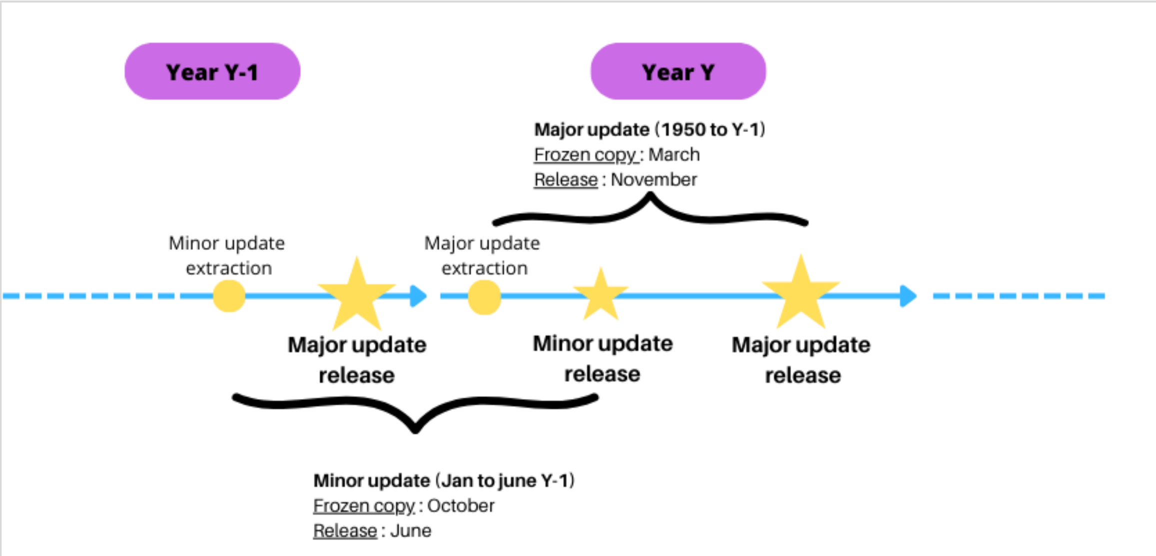


Figure 3: Overview of the CORA dataset extraction and release dates time schedule.

CORA dataset nomenclature

Data in CORA are ordered by date and type of data: nomenclature of the files is **CO_code_YYYYMMDD_PR**or**TS_TT.nc** where:

code is the name of the analysis performed: DMQCGL01.

YYYYMMDD is the date (year, month, day) of the data.

PR stands for vertical PRofile, **TS** stands for Time Series.

TT is the file type (data) which is described below.

File types: The CORA file system is based on 15 daily file types corresponding to the instrument type of the data provider. The file type code and meaning are given in *Table* 2.

CODE	Meaning
BO	Bottles
CT	CTD
DB	Drifter buoys
FB	Ferry boxes
GL	Gliders
ML	Mini loggers for fishery observations
MO	Moorings
PF	Profilers
SD	Sail drones
SF	Scanfish (towed CTDs)
SM	Sea mammals
TS	Ship underway data, thermosalinograph
TX	Thermistor chain data
XB	XBT, XCTD or MBT profiles
XX	Not yet identified

Table 2 : File type code and meaning for the CORA dataset nomenclature

Easy CORA data nomenclature

The Easy CORA files have a slightly different nomenclature, for the file types, from the CORA files, E**CO_code_YYYYMMDD_PR**or**TS_TT.nc** where:

code is the name of the analysis performed: DMQCGL01.

YYYYMMDD is the date (year, month, day) of the data.

PR stands for vertical PRofile, **TS** stands for Time Series.

TT is the file type (data) which is described below.

File types:

- **PR_PF files**: data from Argo floats directly received from the Argo DACS (data centers). Profiles are in real Time or delayed mode if available[1].
- **PR_BT files:** XBT and MBT measurements extracted from the CORA files.
- **PR_CT files:** All the other profiles distributed from the CORA dataset.
- **TS_TM files:** Timeseries from the tropical moorings (TAO/TRITON/RAMA/PIRATA moorings).
- **TS_MO files:** all the other moorings.
- **TS_TS files:** TSG measurements.
- **TS_DR files:** Drifters timeseries measurements.

[1] Principal Investigators have to provide a delayed mode validation on their floats within 6 months after the date of observation

In addition to a different data file nomenclature, the Easy CORA profiles have a different use of the metadata (as explained in the product Product User Manual (PUM)). The Easy CORA profiles have also been vertically subsampled following Table 3 specifications and timely subsampled following Table 4 specifications.

Depth/pressure	**bin thickness**
0-200 m	1 m
200-1000 m	10 m
1000- seafloor	50 m

Table 3: Thickness of the vertical bin for the vertical subsampling scheme. A profile distributed in the Easy CORA dataset only provides one point per vertical bin.

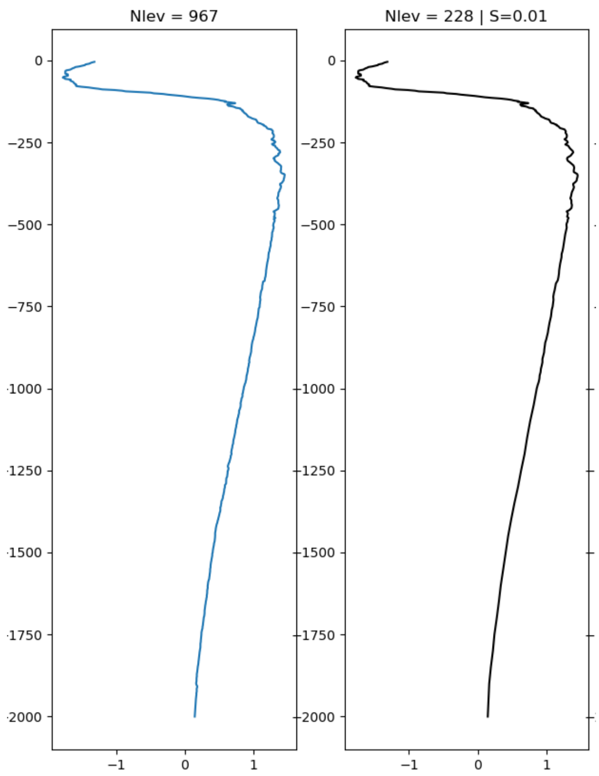


Figure 3: An example of vertical subsampling of a salinity profile (unit: PSU), Nlev is the number of vertical levels of the profile, before and after the subsampling. S in the subsampling representativity.

The subsampling algorithm is selecting one data point in each vertical bin when available. The selection is based on the minimization of the least square distance between the original profile and the subsampled profile interpolated on the original vertical grid. This function is computed and summed for the temperature and the salinity (TEMP and PSAL) parameters, when available, in order to ensure to prefer a data point with TEMP and PSAL parameters to a data point with only one parameter. If one parameter is missing, the function is calculated for the remaining parameter. The integrated least square distance is monotonous, positive, and is null when the two profiles are identical. It is saved in the metadata variable TEMP_SUBSAMPLING_REPRESENTATIVITY (respectively PSAL_SUBSAMPLING_REPRESENTATIVITY). Figure 3 gives an example of a temperature profile vertical subsampling.

Time subsampling is based on time window defined in table 4 and is also suited to minimize an integrated least square distance. Boolean metadata VERTICAL_SUBSAMPLING_STATUS and a TIME SUBSAMPLING_STATUS are set to one if a subsampling has been performed and zero if it was not necessary.

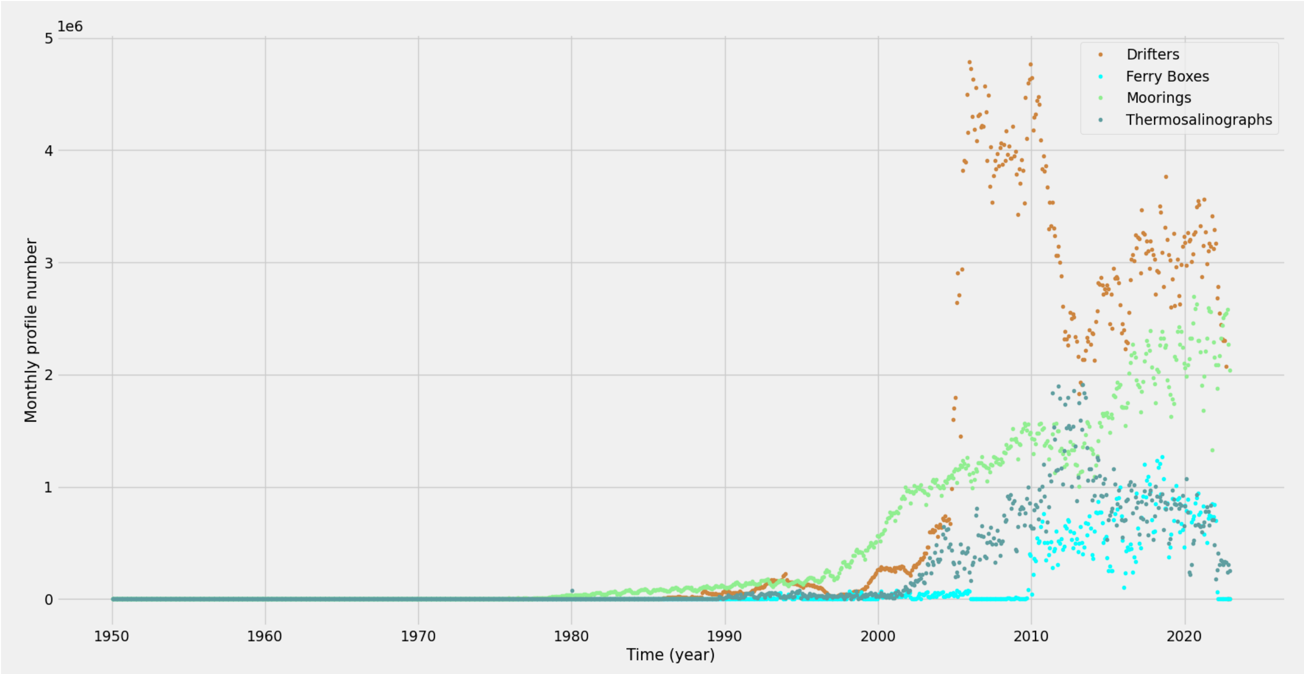
Probe type	**Time window**
Sea Glider	1 hour
Scanfish/Seasoar	1 hour
XBT	6 hours | 10 km
CTD	6 hours
MBT	6 hours
Sea Mammals	6 hours
Profilers	6 hours
Moorings	1 hour
Thermosalinograph	1 hour
drifters	1 hour
Other instruments	6 hours

Table 4: Time window of the profile's time subsampling as a function of the probe type.

II.3 data contained by CORA

CORA contains data from different types of instruments: mainly Argo floats (PF), XBT/XCTD (XB), CTD (CT) and moorings (MO).

Figure 4 gives the yearly number of collected profiles associated to each datatype. Most of the data acquired during the historical period (before 1990) measurements are MBT (XB) profiles. The XBTs and CTD (CT) have been developed during the mid ´60s. The XBTs have been widely deployed by military and scientific programs since the early ´70s. The CTDs deployment rate however increases slowly given the higher cost of CTD probes deployment. The number of collected profiles decreases in the mid-late 1990s with a decreasing number of XBT/CTD casts. The development of worldwide moorings programs (TAO/RAMA/PIRATA) in the mid 1990s is associated with a sharp increase in the number of mooring profiles (MO). Finally, the number of profiles (PF) increases in the early 2000s thanks to the ARGO program.



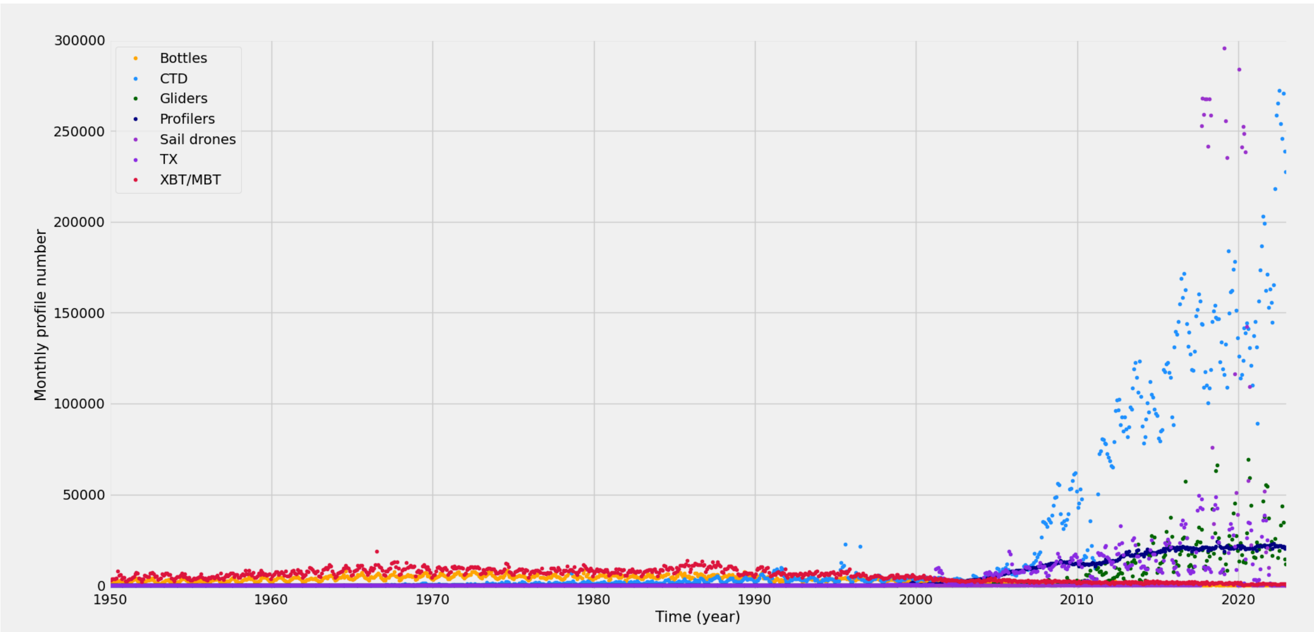


Figure 4: Estimated number of profiles associated to raw instruments types for PR files (top) and TS files (bottom).

In figure 5 the ocean sampling rates are given. These are calculated by counting the monthly sampling rate of an ocean basin gridded at 1°. It shows that most of the North Atlantic Ocean is well sampled in the early 70s. However, the other basins sampling rates are low and increase slowly from 1950 to 1990. The ocean sampling rate decreases in the north Atlantic basin and in the north Pacific basin during the 90s. This decrease is however compensated by a better sampling at depth (see Fig. 6). In the early 2000s, the development of the ARGO program has provided a rapid increase of the ocean sampling rate at a global scale. The ocean sampling scale reaches a maximum over 15% in the Antarctic basin (increasing in the year 2010s thanks to the Marine Mammals Exploring the Oceans Pole to Pole – MEOP - program). In the Arctic basin, the exceeds 20% during the 1970-1990 period and reaches 30 % in the late 80s. It decreases during the 1990s and the early 2000s to reach 20% during the Argo area (2008 – now). This basin coverage also has a very strong seasonal variation for the polar basins (not shown). For the other basins, the coverage also increases from 1950 to 1990, peaks in the late 1980, decreases during the 1990 and stabilized at 40% with the ARGO related coverage after 2008.

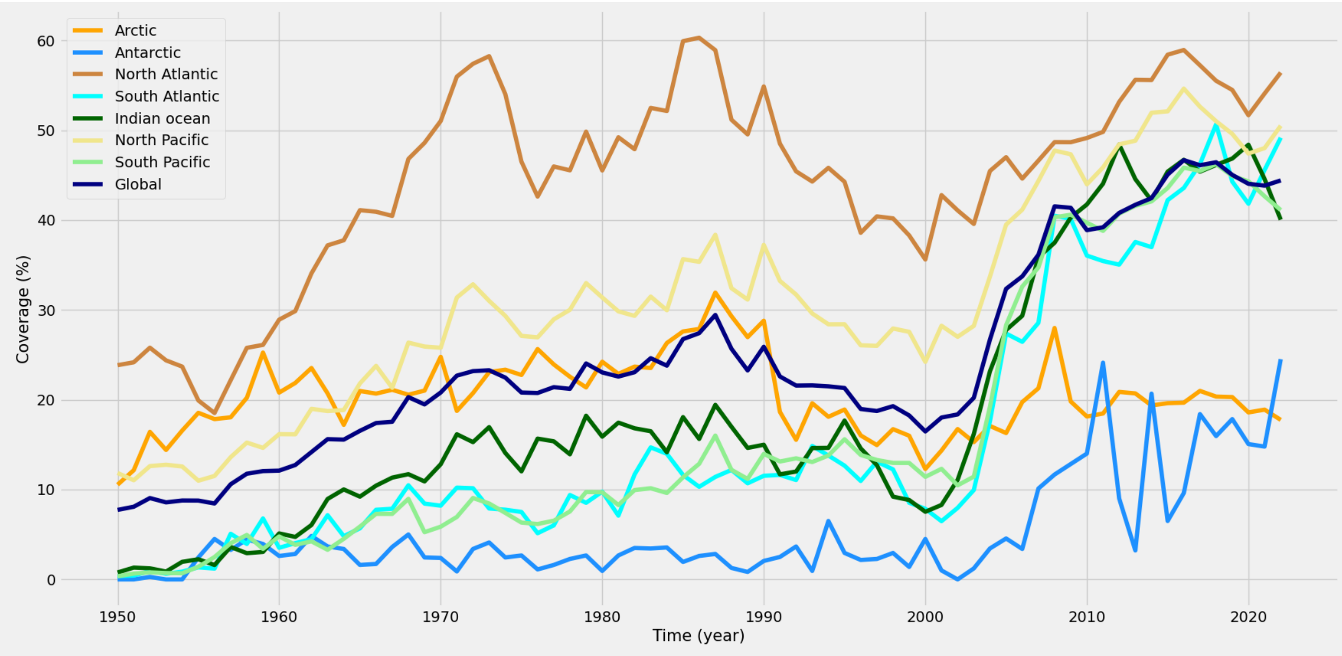


Figure 5: Ocean basins and global ocean sampling rate on the CORA dataset.

Figure 6 gives an overview of the depth distribution of CORA measurements over time. During the 1950-1965 period, most of the observations are MBTs temperature measurements with a few bottle salinity measurements. The increasing number of T4 XBT (250 m depth XBT) after 1965 gives a better ocean coverage at shallow depths for temperature only. The T7 XBTs (1000 m depth XBT) development slowly gives a better sampling at depth. At the same time, the use of CTD sensors provides good temperature and salinity measurements at shallow depth. The number of CTD profiles is however lower. The ocean sampling at depth has gradually improved after the early 1990s with the WOCE program (early 90s), the TAO/RAMA/PIRATA program (late 90s) and the ARGO program (mid 2000s).

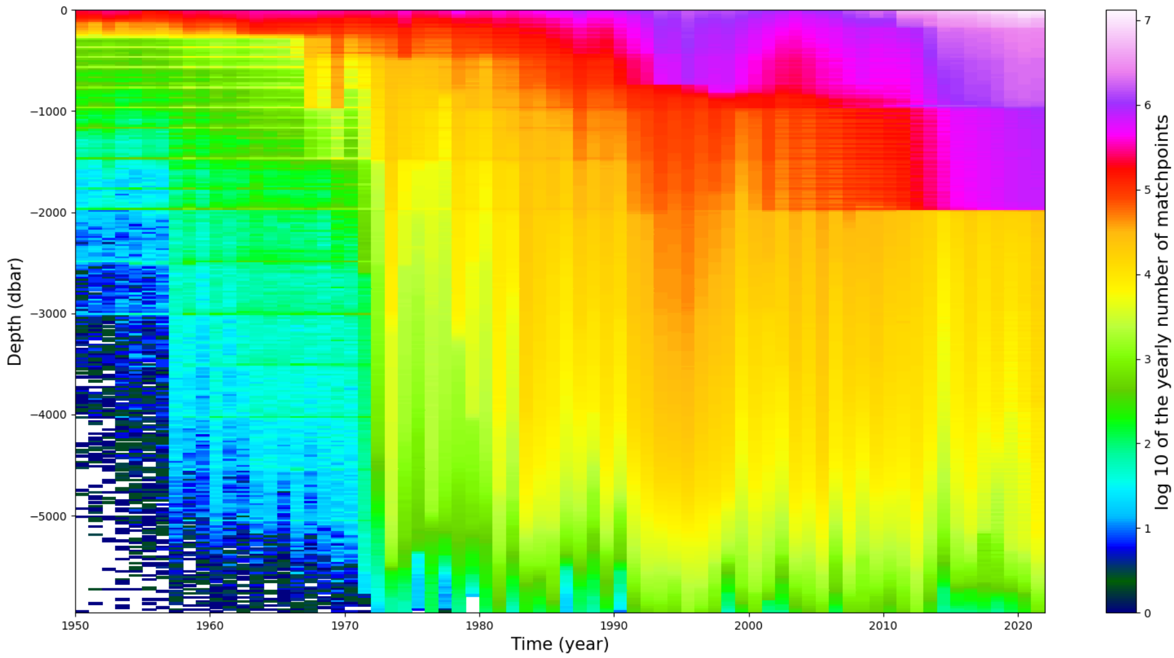


Figure 6: Time depth diagram of number of good profiles per month and per 20 m bins. To ensure the quality of the data, the diagram is based on the sole "good' and "probably good" measurements (QC = 1 or 2).

The CORA dataset is also distributed among regional subsets corresponding to the In Situ TAC regions presented in Figure 1. The time variation of the CORA profiles per instrument type for each regional subset as well as the time diagram of the measurements numbers per depth are given in Figure 7a-f and Figure 8a-f. The In Situ TAC regions are not affected the same way by the evolution of oceanographic instruments. For instance, the contribution of the ARGO program to the Arctic and Back Sea regions coverage happens later than for the Mediterranean region or the Atlantic regions because of the later deployment of the Argo program in the Arctic and Black Sea. On the contrary, the monthly profile number of the North Atlantic, South Atlantic and Mediterranean Sea regions is well correlated with the global region.

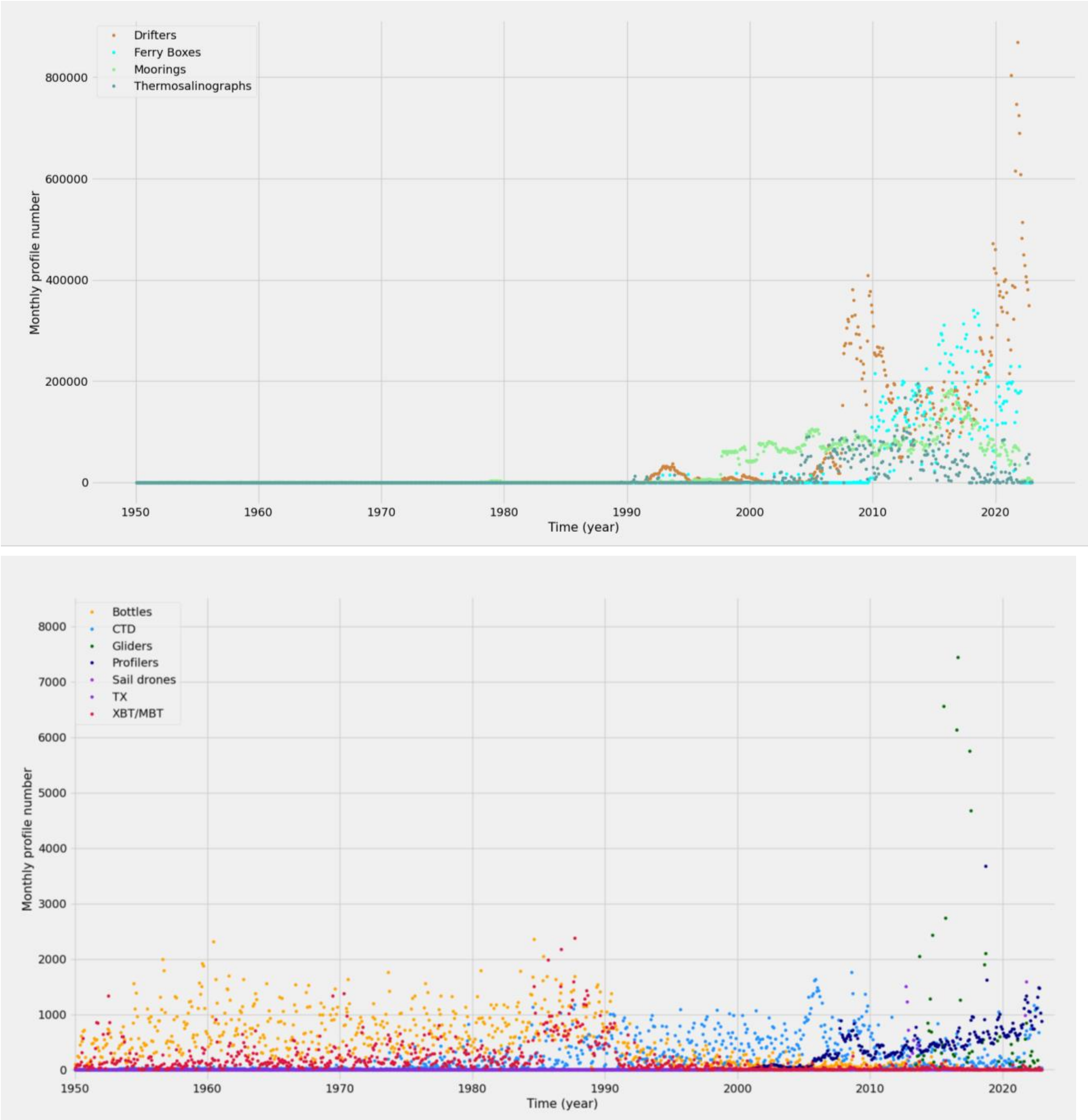


Figure 7a: Distribution of the in situ instruments in the In Situ TAC Arctic region: time evolution of the different platforms (XBT, MBT/bottles, CTD, moorings, profilers, gliders, drifters, TSG, towed CTDs).

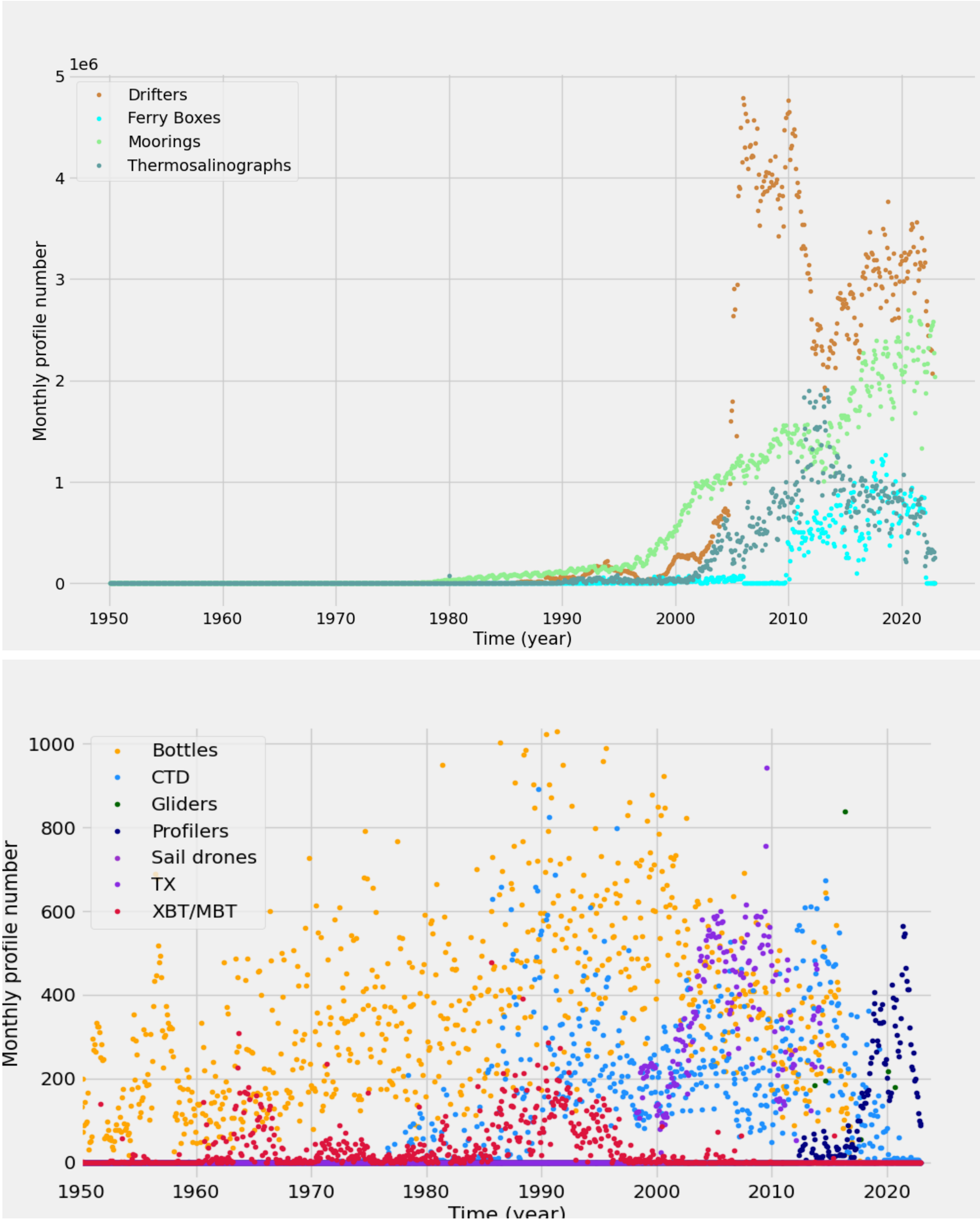
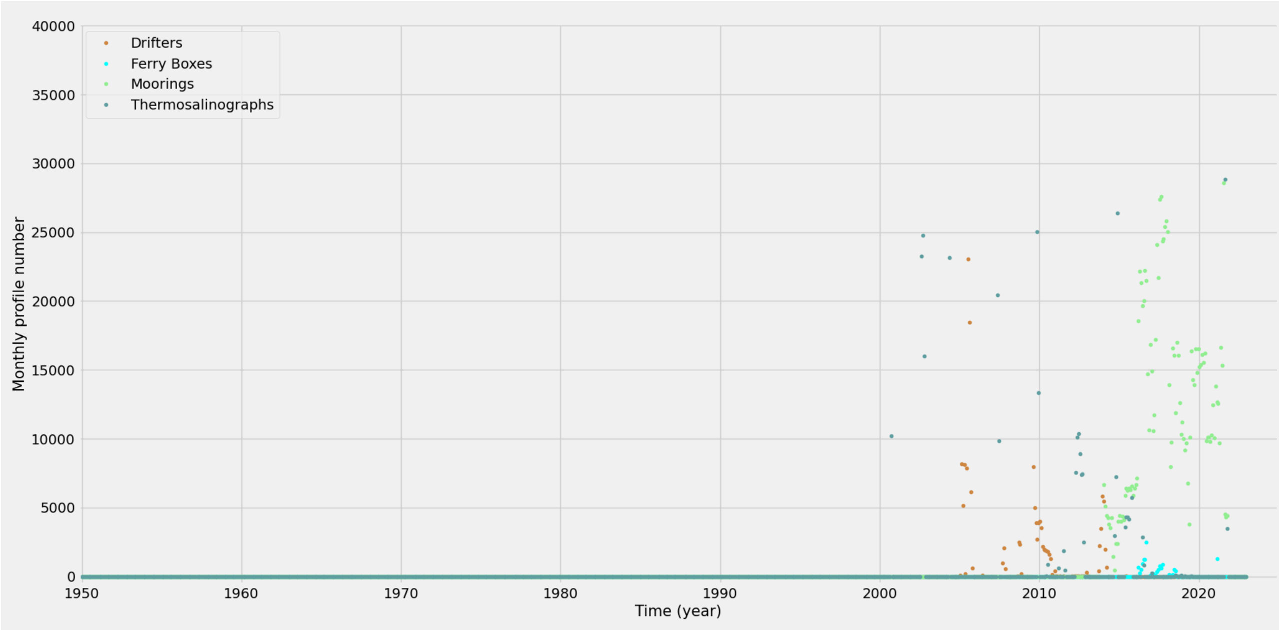


Figure 7b: Distribution of the in situ instruments in the In Situ TAC Baltic region: time evolution of the different platforms (XBT, MBT/bottles, CTD, moorings, profilers, gliders, drifters, TSG, towed CTDs).



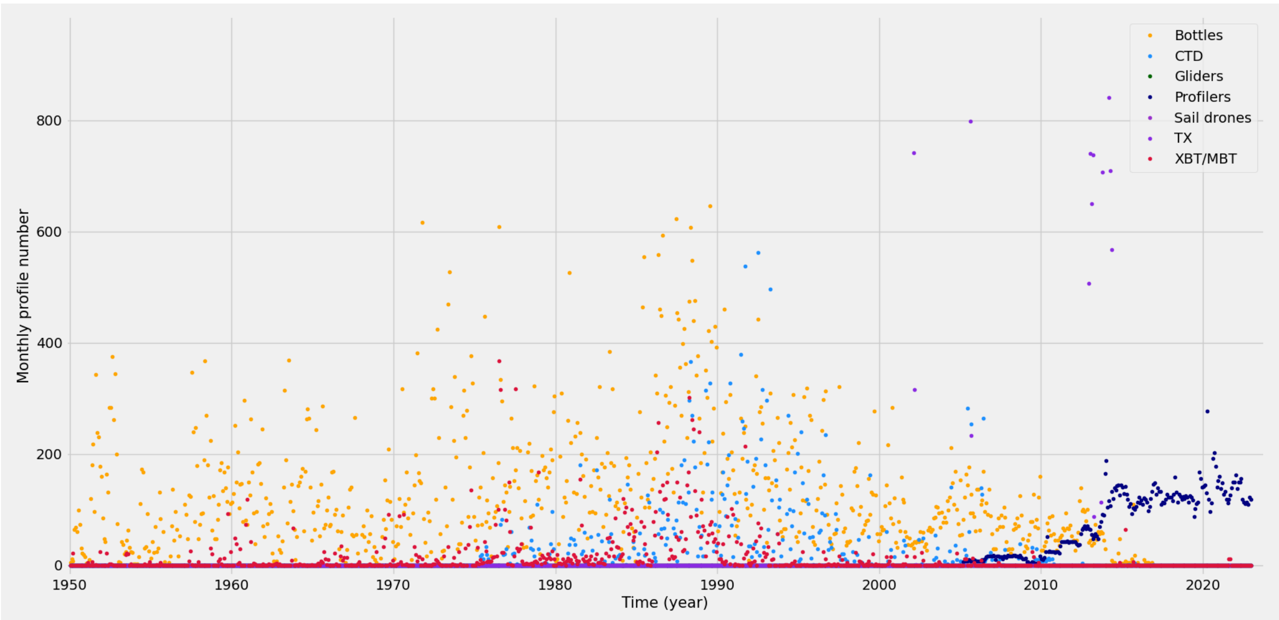


Figure 7c: Distribution of the in situ instruments in the In Situ TAC Black Sea region: time evolution of the different platforms (XBT, MBT/bottles, CTD, moorings, profilers, gliders, drifters, TSG, towed CTDs).

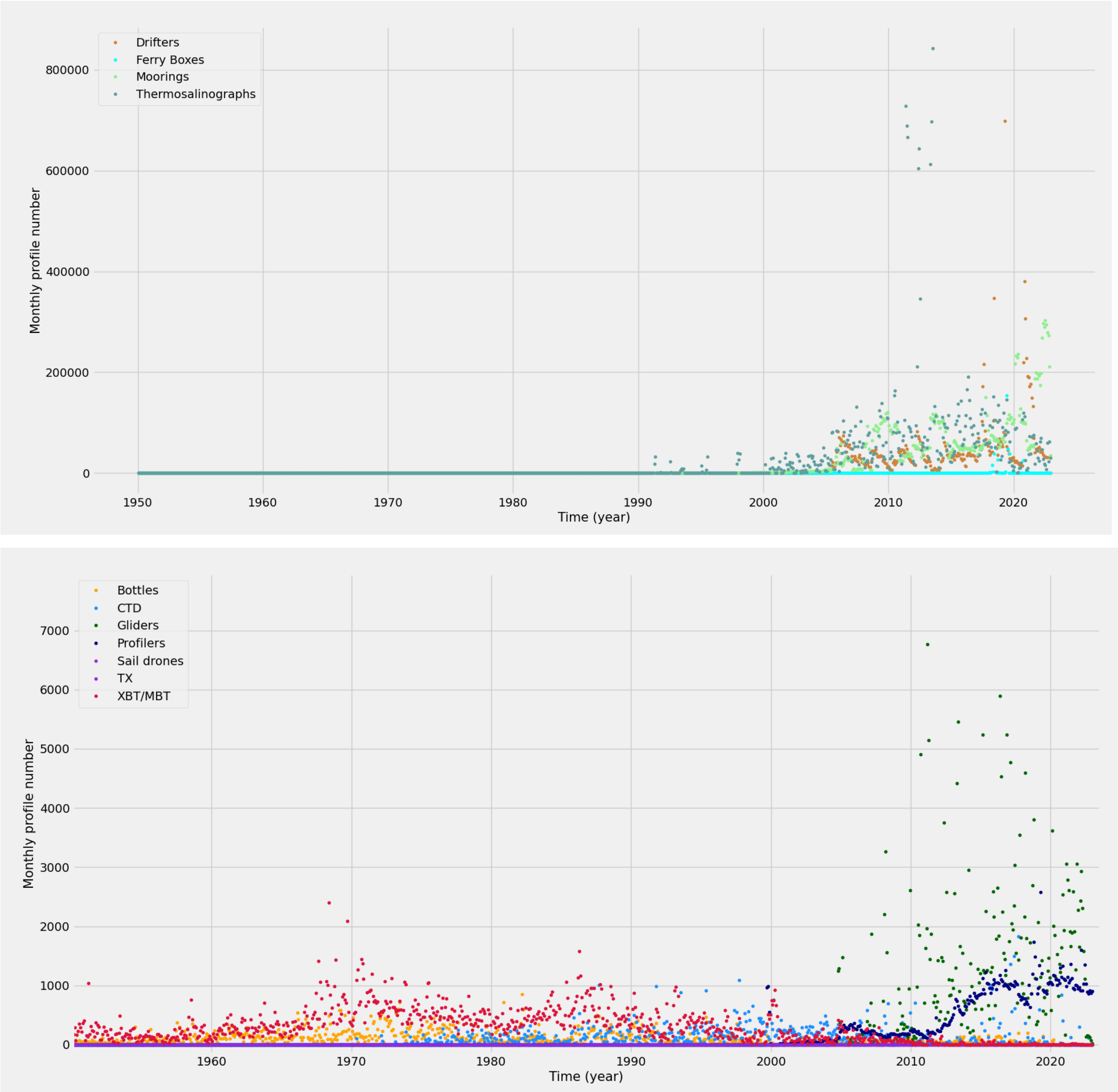


Figure 7d: Distribution of the in situ instruments in the In Situ TAC Mediterranean Sea region: time evolution of the different platforms (XBT, MBT/bottles, CTD, moorings, profilers, gliders, drifters, TSG, towed CTDs).

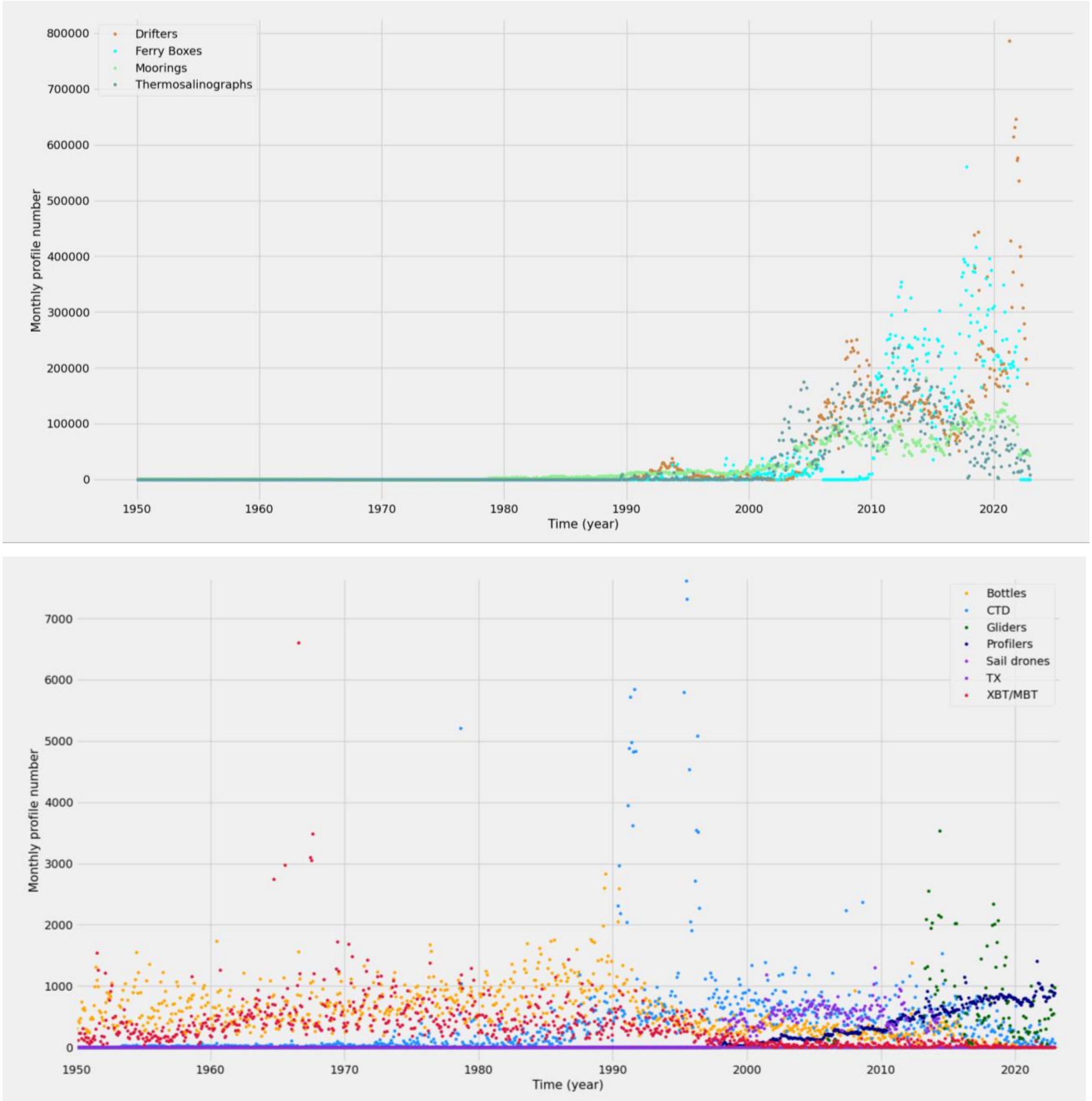
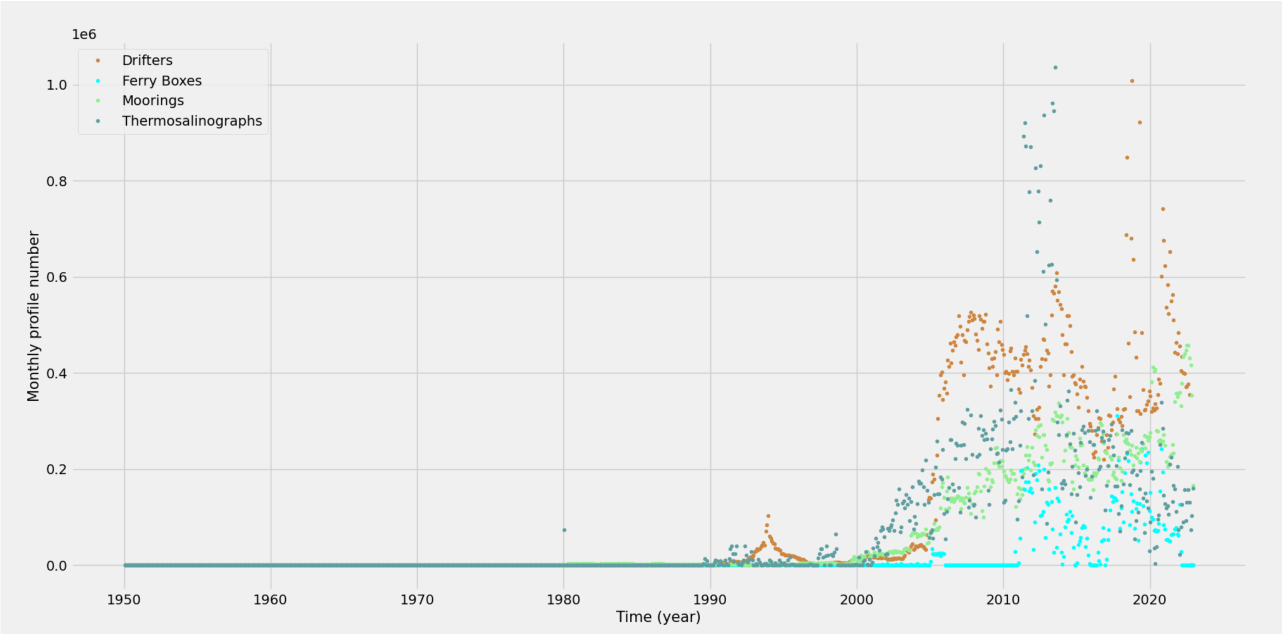


Figure 7e: Distribution of the in situ instruments in the In Situ TAC North West Shelf region: time evolution of the different platforms (XBT, MBT/bottles, CTD, moorings, profilers, gliders, drifters, TSG, towed CTDs).



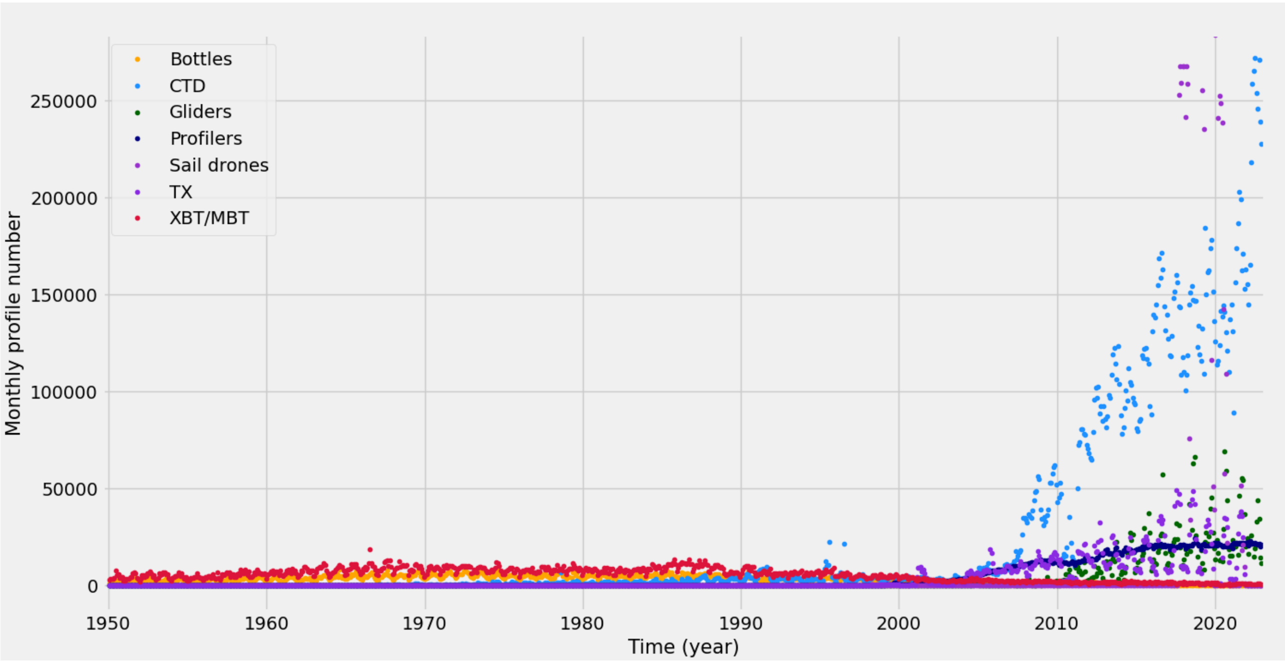


Figure 7f: Distribution of the in situ instruments in the In Situ TAC South West Shelf region: time evolution of the different platforms (XBT, MBT/bottles, CTD, moorings, profilers, gliders, drifters, TSG, towed CTDs).

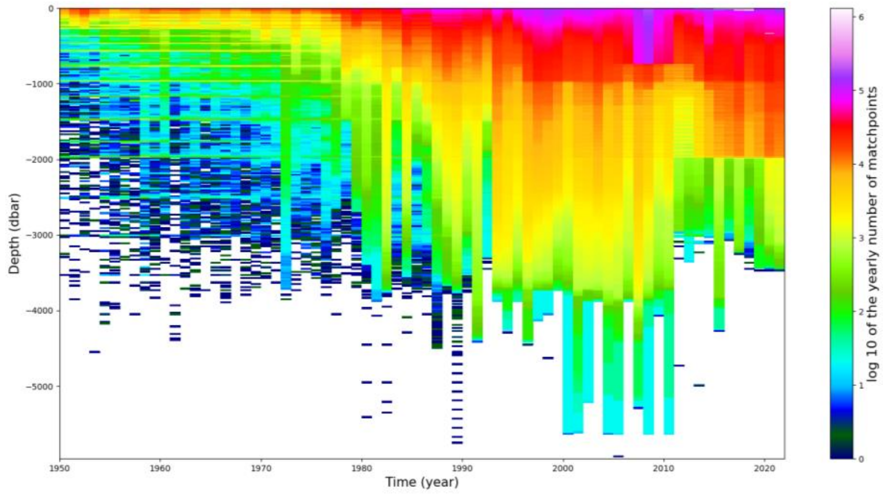
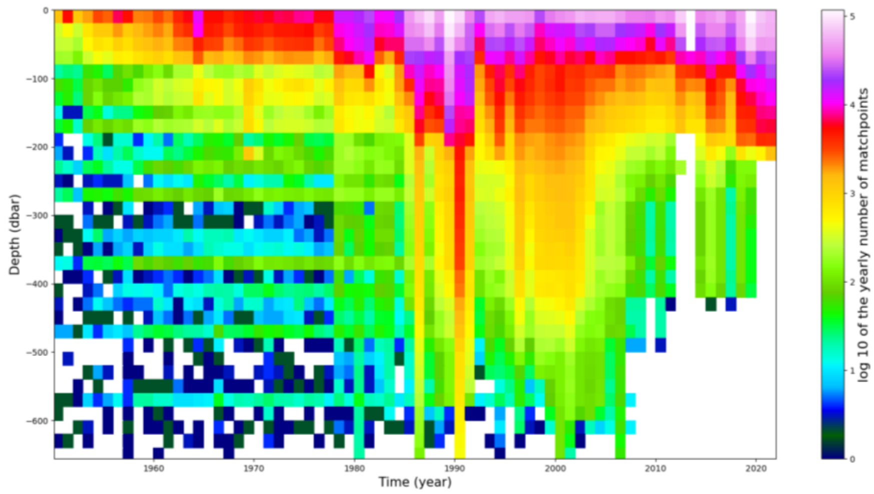
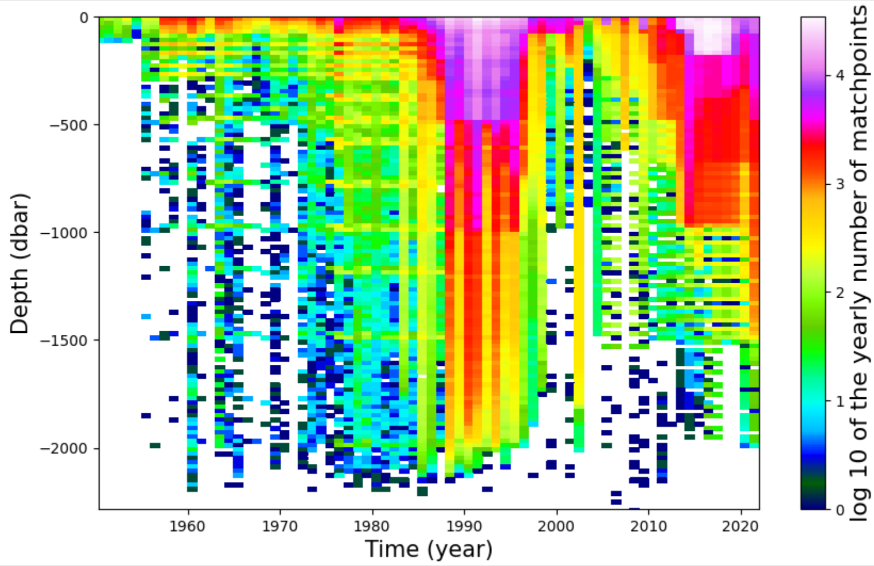
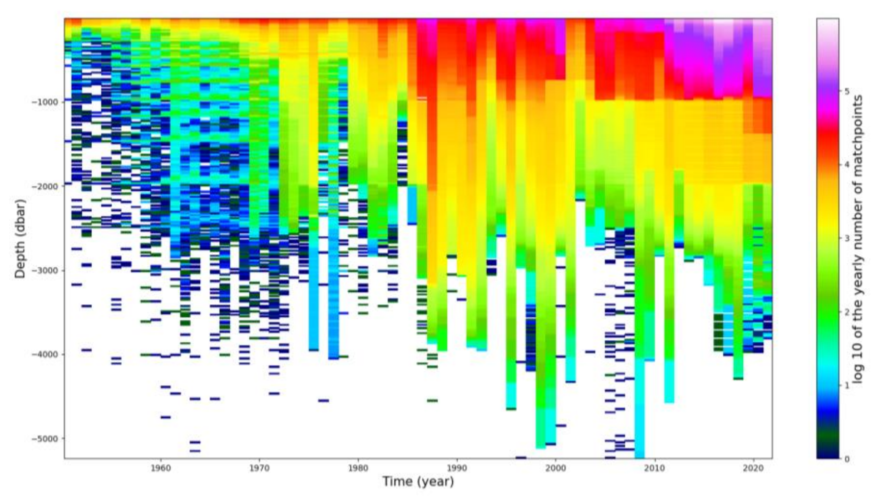
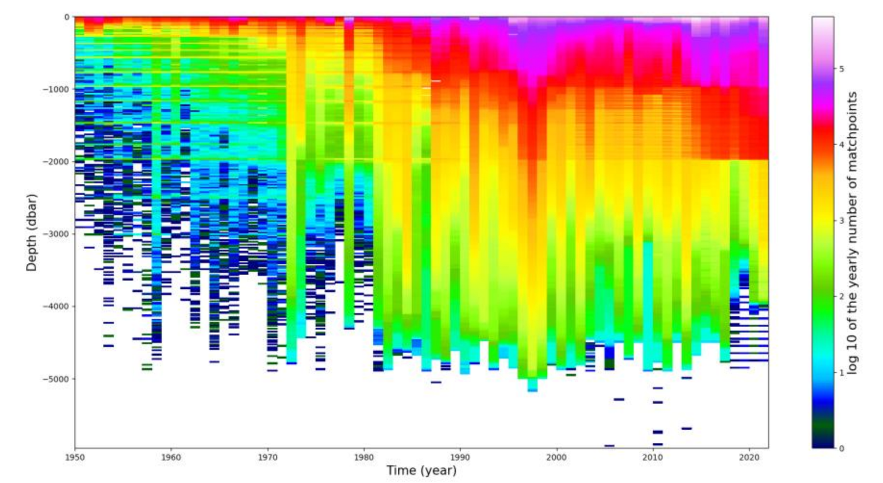
Arctic	
Baltic	
Black Sea	

Figure 8: Distribution of the In-Situ instruments in the In Situ TAC sub regions: depth information of the observations (continues on next page).

Med Sea



North West Shelf



IBI (South West Shelf)

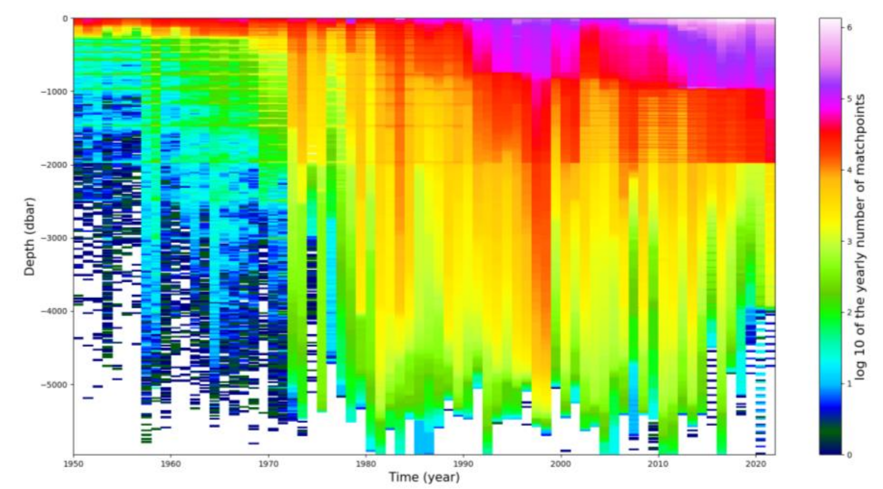


Figure 8: (continued) Distribution of the In-Situ instruments in the In Situ TAC sub regions: depth information of the observations.

I.1 Easy CORA dataset

A side dataset has been developed from the CORA dataset to better fit the reanalysis user's needs. This secondary dataset aims to provide only the "good quality" and "probably good" quality measurements (QC= 1 or 2, respectively). In addition to that, the Easy CORA dataset provides the best estimation of each dataset. For instance, when TEMP and "TEMP_ADJUSTED" parameters are available in the CORA dataset, a single TEMP parameter is distributed by the Easy CORA dataset, with the best available estimation of the temperature measurement. A vertical and a temporal subsampling is also applied to provide a more homogeneous dataset. All the information about these operations is available in the CMEMS-INS-PUM-013_001 document (https://data.marine.copernicus.eu/product/INSITU_GLO_PHY_TS_DISCRETE_MY_013_001/services).

Figure 9 gives the number of profiles by instrument type from 1950 to 2022. It reveals a clear decrease of the number of profiles provided for the timeseries format. This decrease is due to the time sub-setting of the surface drifters, mooring and termosalinograph platforms. The sub-setting however never exceeds one profile per 1 hour time period, leaving enough measurements to provide a well sampled view of the day to day variability. This subsampling drastically decreases the number of thermosalinographs measurements since it concerns not many platforms with a very high frequency sampling. The number of profiles from other instruments are consistent with the CORA product profile distribution.

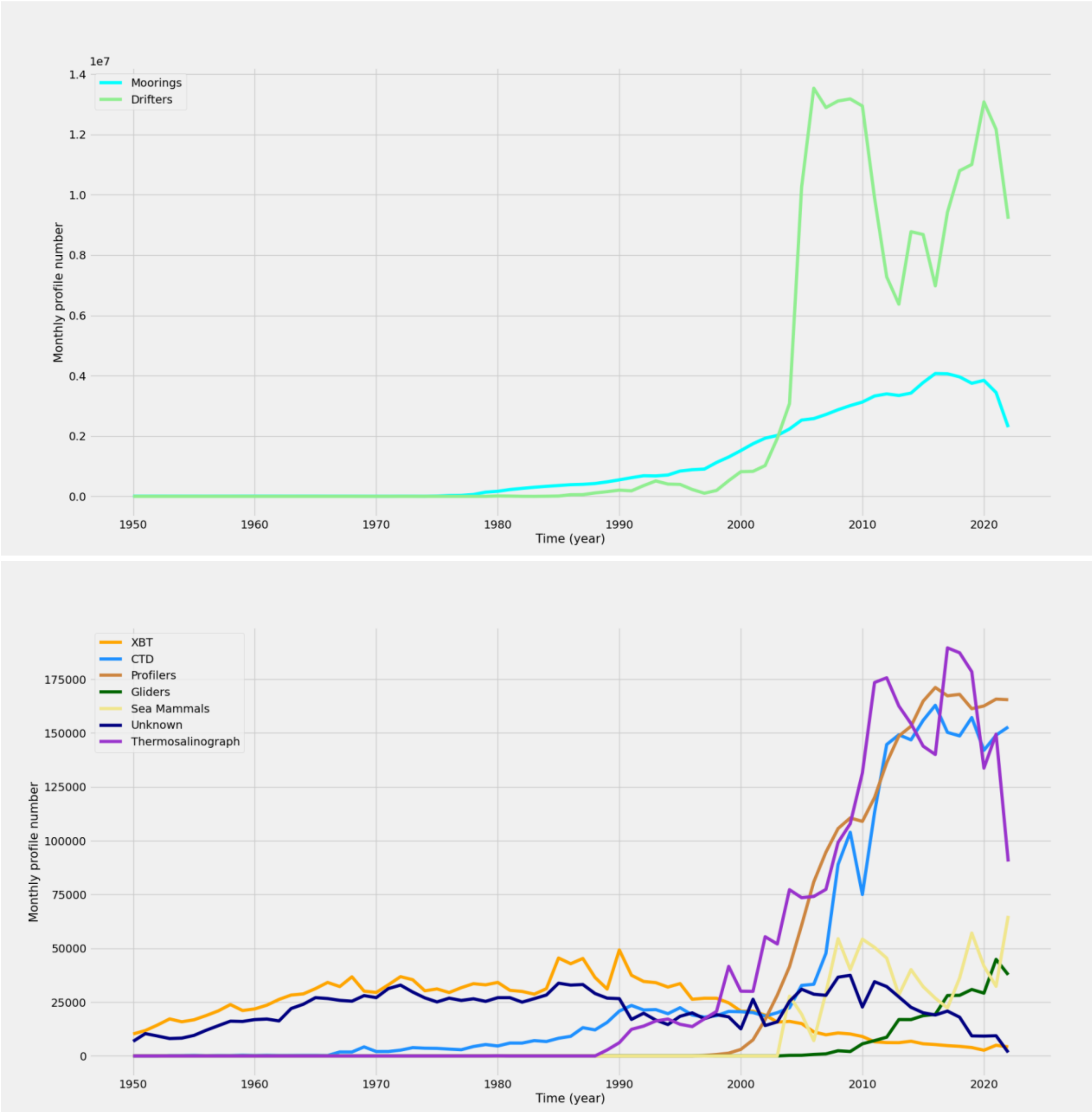


Figure 9: Estimated number of profiles associated with raw instruments types for Easy CORA dataset (top: Moorings and Drifters; bottom: other instruments).

I.2 Difference with the other global Copernicus in-situ products

The product called INSITU_GLO_PHYBGCWAV_DISCRETE_MYNRT_013_030 comes from the GLOBAL In Situ TAC Copernicus Distribution Unit (monthly, latest, or historical repositories). It is also a Copernicus distribution of the Coriolis database but with two main differences with respect to the CORA dataset (Figure 10):

1) the update of Copernicus distribution is continuous in the time and the content of Copernicus /monthly distribution is an image of In Situ TAC database contrarily to CORA which is a hard copy of this database corresponding to a given date.

2) In the Copernicus in-Situ NRT product, the file distribution is organized by platform. Consequently, each file gathers all the profile of the corresponding instrument. On the other hand, the CORA dataset is designed to be assimilated by models and is thus organized into daily files.

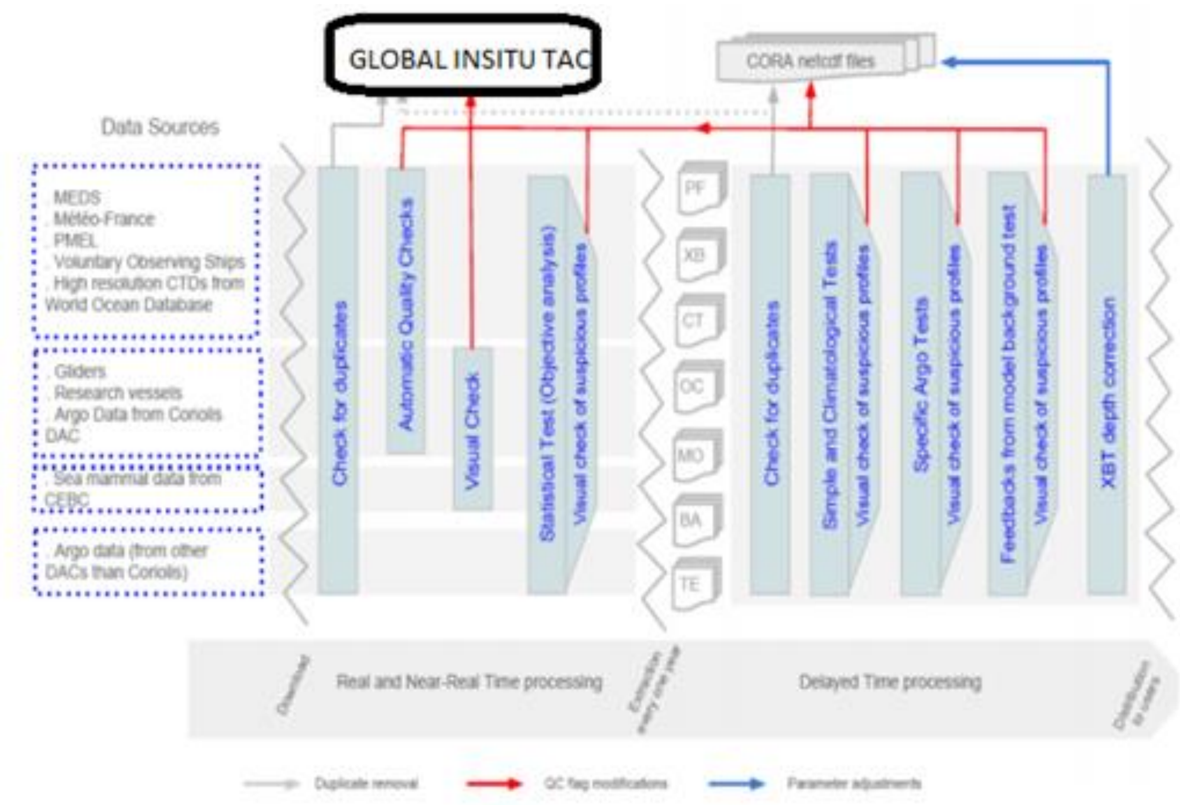


Figure 10: Data source and validation process for CORA GLOBAL.

II VALIDATION FRAMEWORK

In this section we will describe the systems that enable the assessment of the CORA dataset on observed levels. The validation procedure has not evolved since the CORA-GLOBAL-04.2 product, released in 2016.

The production system that handles and manages the validation of CORA dataset involves both the Coriolis Data centre and "Coriolis Research and development team".

The validation chain of CORA begins within the Coriolis database with real time automatic checks. Those tests check that meta-data are consistent with archiving, then objectives analyses are run on the data for real-time and near-real-time checks.

After extraction from the Coriolis database (i.e. In Situ TAC) data are re-analysed through a delayed mode assessment process especially dedicated to CORA dataset. It contains a series of 14 generic tests plus extra specific corrections: duplicates suppression and XBT correction. And at the end of the process, we perform a comparison of the data with a final objective analysis that guarantees the consistency between each measurement in time and space.

Corrections on measurements are avoided as much as possible in the Copernicus In Situ TAC, and when performed they are provided alongside the raw measurements to let the users choose the level of processing they want to use. Results of validation are provided by through the Quality flags associated to the measures (see Table 5).

Quality Code (QC)	**meaning**
0	No QC was performed
1	Good data
2	Probably good data
3	Bad data that is potentially correctable
4	Bad data
5	Value changed (almost never used)
6	Not used
7	Nominal value
8	Interpolated value
9	Missing value (99999)

Table 5: Quality flags and their definition.

II.1 Detection and correction of repeated problems

What we call "*repeated problems*" are issues that occur very frequently in the data and where there is no doubt on the truthfulness of the alert raised. For instance a sensor that does not sink quickly may provide temperature or salinity measurements at constant depth. Those are actually not bad measurement in the sense of physical measurements, but are not suited to the CORA dataset users. Such measurements are automatically flagged to ensure the consistency of the dataset. A log file is then edited and sent to the GLOBAL In Situ TAC in order to validate the alert and correct the data at source.

As "*repeated problems*" are those that occur frequently and the alert itself does not need any confirmation thus the correction applied is automatic. In the CORA dataset, detection and correction are done in delayed mode directly into CORA NetCDF files and then those files are resubmitted to the GLOBAL TAC in order to replace previous measurements.

II.2 Detection and correction of occasional problems

On the contrary some problems are rare and it is necessary that an operator confirms the truth of the alert and changes carefully the values of quality flags. For example, some spikes in the data are rather hard to highlight. In such a case the detection ran on a measure or on metadata in CORA permits the creation of a log file containing the suspicious measurements. Fortunately, the number of "occasional problems" is much lower than "repeated problems" and operators from the TAC can visualize each of them to confirm/dismiss the alert and then correct or leave the quality flags as it.

II.3 Chronology of the tests

II.3.1 Before extraction of CORA data in NetCDF format

The measurements extracted from the Coriolis database are subject to a first set of quality checks.

1. Automatic checks at data loading in GLOBAL In Situ TAC database
2. Comparison with Objective analysis on 30 days frame in real time
3. Comparison with Near Real Time objective analysis once a month on the last 30 days
4. First validation run on In Situ TAC GLOBAL distribution Unit (*monthly* repository) with feedback to the GLOBAL INSITU TAC

The tests listed above are exhaustively described in Coatanoan and Petit de la Villeon, (2005) and discussed by Cabanes, (2013).

II.3.2 After extraction of CORA data in netCDFformat

A second set of tests is set up after the extraction of the measurements from Coriolis.

1. Second validation run on CORA with the DMQC methods. See Table 6 for a list of all the tests.

2. Feedback (as log files or netCDF resubmission) to the GLOBAL In Situ TAC and correction of "repeated problems" in files or re extraction of corrected data from the GLOBAL In Situ TAC database
3. Integration of other feedbacks from different partners: (Mercator-Ocean, CLS, Altran, Meteo-France, etc.)
4. Suppression of duplicated measurements (Table 7 and Table 8).
5. Correction of temperature and depth on XBT measurements
6. Last objective analysis ran on the whole dataset with feedback to the GLOBAL In Situ TAC and correction of "repeated problems" in files
7. CF (File Checker) compliance checks on CORA RAW files and GRIDDED files.

II.4 Content of the second pass of generic tests

Table 6: Content of the second pass of generic tests (continues in next pages).

Name of the validation	**Description**	**Alert Validation**	**Correction applied in CORA**	**Method of Correction applied in GLOBAL TAC**
Measure on earth	Compare position to bathymetry, to reject: - on land positions. - Position more than 5km distant from nearest coastline with elevation above 50m	Operator visualisation	Position QC edited in file	Visualizing and manual QC position edition
Date check	Checks that the dat correspond to the name of the file.	automatic	Quality flag edited in file	Visualizing and manual QC position edition
Parameter Range check	Check that TEMP PSAL PRES and DEPH have acceptable values -2.5<TEMP<45 (units : °C) 0<PSAL<50 (units : PSU) -2.5<PRES<20000 (units : dbar) -2.5<DEPH<20000 (units : meter)	automatic	Quality flag edited in file	Visualizing and manual QC position edition
Constant check	Check that TEMP PSAL PRES and DEPH have identical values along the vertical	automatic	Quality flag edited in file	Visualizing and manual QC position edition
Ascending immersion check	Check that PRES and/or DEPH are monotonous	automatic	Quality flag edited in file	CORA corrected files loaded into In Situ TAC database to replace former one

Name of the validation	Description	Alert Validation	Correction applied in CORA	Method of Correction applied in GLOBAL TAC
Duplicate levels check	Check that PRES or DEPH are not duplicated.	Automatic	Quality flag edited in file	CORA corrected files loaded into GLOBAL In Situ TAC database to replace former one
Minmax test	Comparison of the profiles to a minimum and a maximum reference field.	- Operator visualisation for doubtful profiles - automatic for large errors	Extraction and concatenation of corrected data from GLOBAL In Situ TAC	Visualizing and manual QC position edition
Spike check	spot spikes on profiles when TEMP ΔPRES/ΔLEVELS>1.5 PSAL ΔPRES/ΔLEVELS>0.5	Operator visualisation	Extraction and concatenation of corrected data from GLOBAL In Situ TAC	Visualizing and manual QC position edition
Quality Flag relevance	Control that a given QC is relevant with the associated measure. For Example, if a measure is at fill value then corresponding QC must be have the quality flag fill value	automatic	Quality flag edited in file	CORA corrected files loaded into GLOBAL In Situ TAC database to replace former one
Depth wrote in pressure field	Check that depth measurements are not written in PRES field (especially for XBT and X-CTD)	automatic	Quality flag edited in file	Independent correction in database (to be implemented)
Duplicate profile	Detect duplicate profiles (see Table 7 and Table 8 for the detection criterion and decision chart).	Operator visualisation	Suppression of the duplicated profile	CORA corrected files loaded into GLOBAL In Situ TAC database to replace former one
XBT correction	Empirical correction of depth bias and temperature offset on XBTs. Method by M. Hamon et al. (2012) based on co-localisation with CTD profiles. - Correction of temperature offset Correction of depth bias	Operator visualisation	Measure and quality flag edited in file	No feedback to GLOBAL In Situ TAC for this correction

Name of the validation	**Description**	**Alert Validation**	**Correction applied in CORA**	**Method of Correction applied in GLOBAL TAC**
Assimilation feedback	Alerts on profiles raised by a too strong innovation value (innovation is high when the data-model anomaly is high) when assimilated in a model.	Alerts from Mercator-Ocean validated by operator (fake alert declared by operators are then analysed back at Mercator-Ocean)	Quality flag edited in file	Visualizing and manual QC position edition
Comparison to ENSEMBLE dataset	A comparison to En3v2 (ECMWF in situ T&S dataset) is performed to check the temporal and spatial coverage of data.	N/A	N/A	Missing datasets in CORA are identified and we try to integrate them in the next release.
Ultimate objective analysis	Last step of the process that guarantee a global (spatial and temporal) consistency of the dataset by computing a residual value between each measure and the "background" given by the climatology and the other measurements. two kind of alerts can be raised by the analyse objective: - Standardisation alert (spikes, climatology differences) Residual alert (differences between measure and field computed)	automatic	Quality flag edited in file	Visualizing alert logs and manual QC position edition

	Couple with same type (ex: BA BA or TE TE)	Couple with different type (ex: BA TE, BA OC)
delta date = +/-	0,00001 days (0,864s)	0,042 days (1h28s) - exception for TE PF : 24h -
Delta longitude = +/-	0,0001°	0,1°
delta latitude = +/-	0,0001°	0,1°
Platform number	Can be different	Must be the same

Table 7: Duplicated profiles detection criterion.

types	OC	BA	XB	CT	PF	MO	TE	HF	IC
OC	?	del. BA	no	del. OC	no	no	del. TE	del. HF	del. OC
BA		?	del. BA	del. BA	del. BA	del. BA	del. BA	del. HF	del. BA
XB			?	no	no	no	del. TE	del. HF	no
CT				?	no	no	no	del. HF	del. CT
PF					?	no	del. TE	del. HF	no
MO						?	del. TE	del. HF	no
TE							?	del. HF	del. TE
HF								?	del. HF
IC									?

Table 8: Duplicated profiles decision chart.Fon instance in the case of a profiles of XB type (expndable bathythermograph data) duplicated with a TE profile (data extracted from the TESSAC message from real-time data distribution maintained by the NOAA institute), the TE profile will be deleted.

Minmax method description:

Each of the temperature and salinity profiles of the CORA database are compared to a minimum and maximum measured value reference field. The field is a gridded mesh of 1 degree resolution horizontal hexagonal cells of 20 m depth. The reference fields are the maximum and minimum measured values on a set of 1.2 million ARGO profiles, 10,000 CTD profiles and sea mammals' measurements extracted from the MEOP database, vertically interpolated from the surface to 2,000 m depth.
The CORA 5.0 measurements are compared to the minimum and maximum reference of the corresponding cell and the upper and lower adjacent cells. The profiles containing measurements exceeding the reference values are visually checked by an oceanographer. The minmax method is relaxed on the continental shelf since the minmax sampling is insufficient in the continental shelf zones. As a consequence, the temperature and salinity profiles measured in the continental shelf (depth < 1,800 m) are compared to the ARIVO climatology field (Gaillard et al, 2009) plus or minus 10 times the climatological standard deviation field.

II.5 Data validation results

The assessment of the validation method applied in the delayed time mode validation framework is difficult to perform. Methods are based on the objective analysis of the local temperature and salinity variance before and after the delayed time mode validation process. The results, reproduced after Szekely et al (2019), are shown in Figure 11 to Figure 14. Here it can be seen how the mean variance of the temperature and salinity between 60N and 60S is almost constant in the CORA dataset and very noisy in the dataset before the validation process. This difference in the ocean temperature and salinity variability is due to the efficiency of the delayed time mode validation framework.

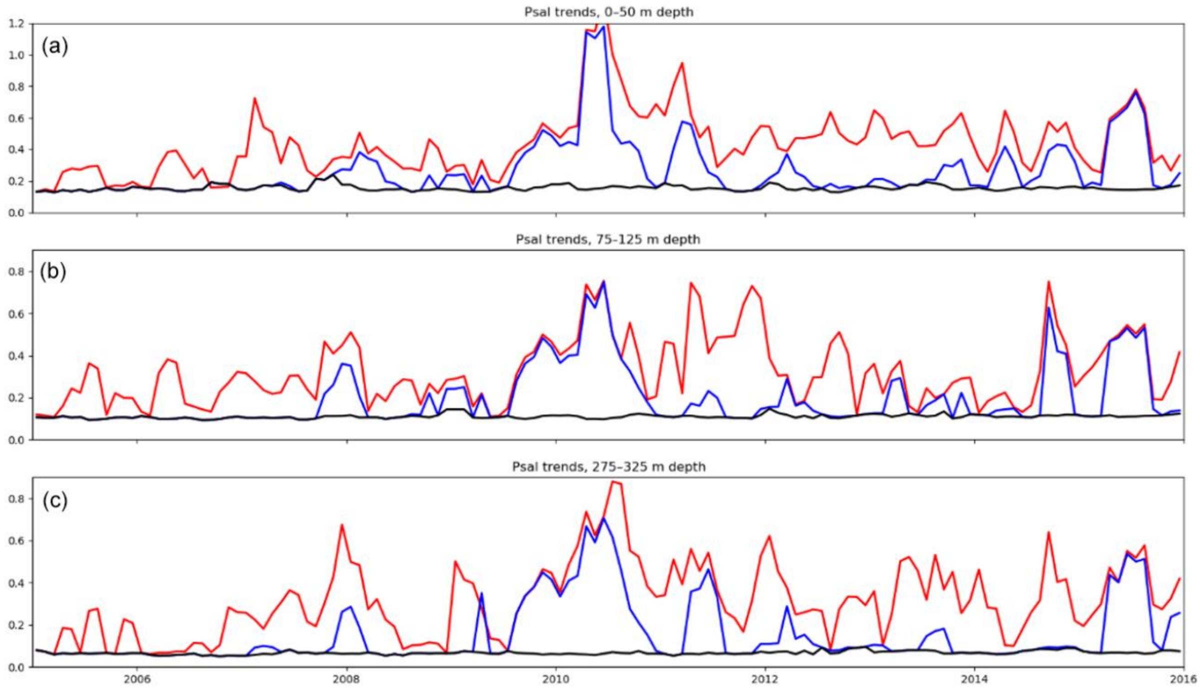


Figure 11: Mean salinity standard deviation in the 0-50m layer (top), 75-125 m depth layer (mid.) and 275-325 m depth layer (bot.). The CORA dataset with QC flags 1,2,3 and 4, CORA dataset with NRT flags 1 and 2 (blue) CORA dataset with CORA flags 1 and 2 (black) are represented.

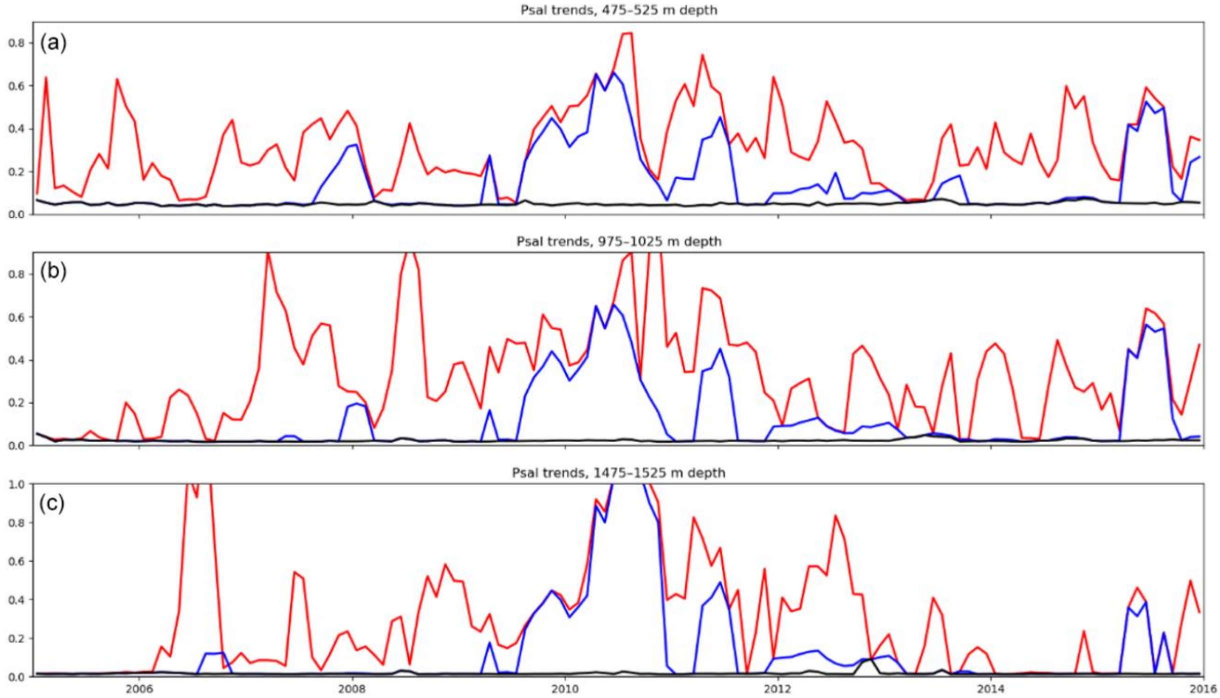


Figure12: Mean salinity standard deviation in the 475-525 m depth layer (top), 975-1025 m depth layer (mid.) and 1475-1525 m depth layer (bot.). The CORA dataset with QC flags 1,2,3 and 4, CORA dataset with NRT flags 1 and 2 (blue) CORA dataset with CORA flags 1 and 2 (black) are represented.

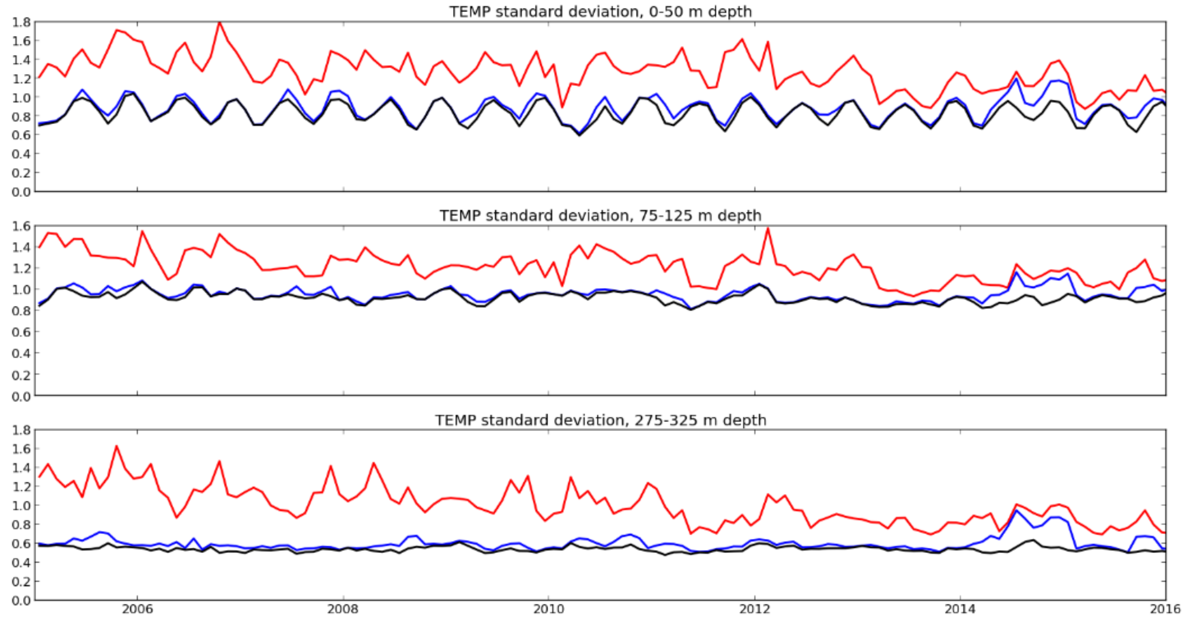


Figure13: Mean temperature standard deviation in the 0-50m layer (top), 75-125 m depth layer (mid.) and 275-325 m depth layer (bot.). The CORA dataset with QC flags 1,2,3 and 4, CORA dataset with NRT flags 1 and 2 (blue) CORA dataset with CORA flags 1 and 2 (black) are represented..

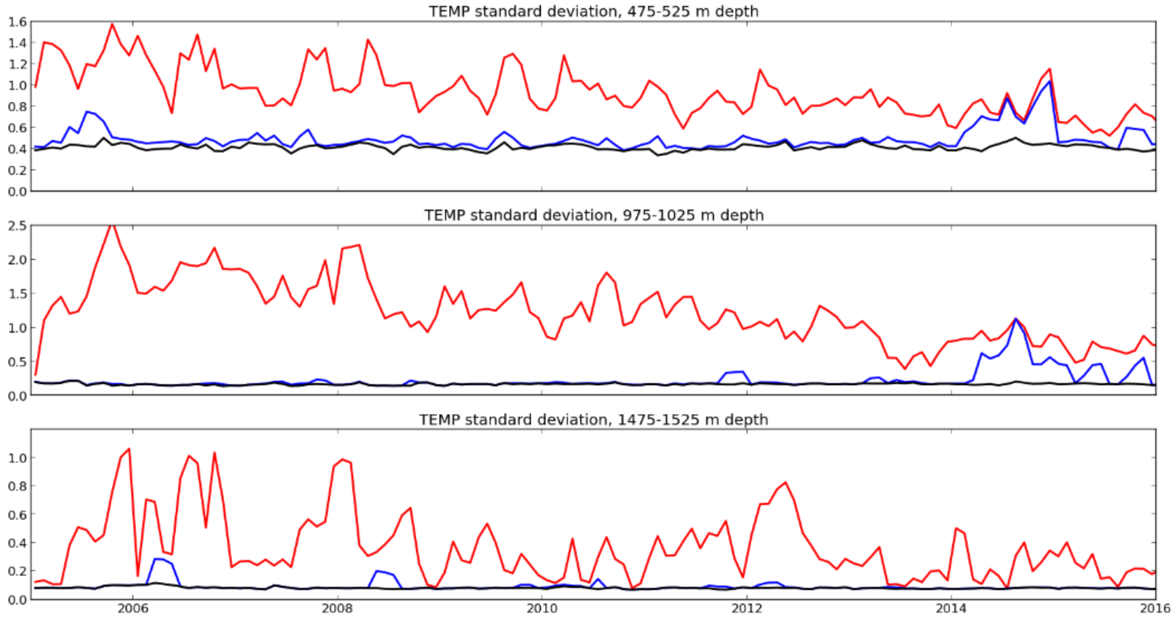


Figure14: Mean temperature standard deviation in the 475-525 m depth layer (top), 975-1025 m depth layer (mid.) and 1475-1525 m depth layer (bot.) The CORA dataset with QC flags 1,2,3 and 4, CORA dataset with NRT flags 1 and 2 (blue) CORA dataset with CORA flags 1 and 2 (black) are represented.

II.6 XBT correction

Attention is given to apply a correction on the expendables bathythermograph (XBT) measurements. As demonstrated by Levitus et al. (2009), the XBT drift may lead to a significant drift in the global ocean heat content. An adjustment is thus necessary.

The XBT is a rocket-shaped instrument dropped into the water from a moving ship without it slowing down. The temperature sensor is linked to the ship by a capillary wire and transfers the measured temperature while falling. It is then necessary to assess the falling rate of the XBT in order to calculate the depth of the measurements. The most common XBT falling rate equation is described in Hanawa et al. (1995) by a quadratic falling rate equation:

$$z = at + 10^{\{-3\}b} t^2 \qquad Eq. (1)$$

Note that equation 1 does not take into account parameters such as the in-situ water viscosity, ^the XBT manufacturer, or the ship speed. There is moreover an uncertainty about the correction method applied to the XBT profiles, since the metadata of XBT profiles are often incomplete in the early period of XBT exploitation. The scientific community has thus developed various methods to adjust the Hanawa coefficients in order to lower the XBT drift. In the previous CORA datasets, the XBT profiles were corrected using the method developed by Hamon et al. (2012). In CORA 5.2, the Cheng et al. (2014) method has been implemented since it is the latest recommendation from the scientific community.

II.6.1 Summary of the method

Different issues with the data of eXpendable BathyThermograph (XBTs) exist and, if not corrected, they are known to contribute to anomalous global heat content variability as shown by Owen and Wong (2009). The XBT system measures the time elapsed since the probe entered the water and thus inaccuracies in the fall rate equation result in depth errors. There are also issues of temperature offset but usually with little dependence on depth.

G1	G2	G3	G4	G5	G6	G7	G8	G9
Sippican T7-DB	DX	Sippican T4-T6	SX	Sippican T10	Sippican T5	TSK T4-T6	TSK T5	TSK T7-DB

Table 9: Groups of XBT types corrected by the Cheng et al. (2014) method. To each groups correspond different adjustment coefficients.

In the method developed by Cheng et al. (2014), the XBT profiles are gathered in 9 groups (*Table* 9) according to the manufacturer and the probe type. Different coefficients are developed according to the year and the group of the corrected XBT.

The original Hanawa et al. (1995) coefficients are modified following the Cheng et al. (2014) framework:

$Z = At - Bt^2 - dz$

$A = A_{H95} + dA_{temp} + dA_{year}$

$B = a_1 A + a_2$

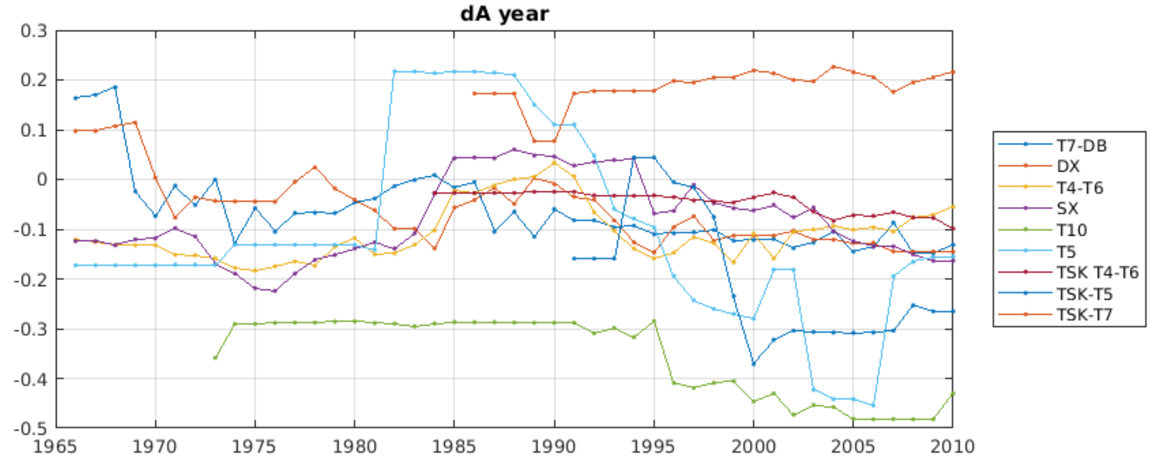
$dz = b_1 A + b_2$.

With A_{H95} the original falling rate equation coefficient A defined by Hanawa et al. (1995), and a_1, a_2, b_1 and b_2 the probe type coefficients defined by Cheng et al. (2014) in Table 10.

	a_1	a_2	b_1	b_2
T4-T6, SX and T10	*0.0069 (±0.0007)*	*-0.435(±0.0045)*	*5.791(±1.015)*	*-37.285(±6.775)*
T7-DB, DX and T5	*0.0070(±0.0002)*	*-0.0440(±0.0016)*	*6.376(±0.657)*	*-40.293(±4.359)*
TSK	*0.0034(±0.0017)*	*-0.0204(±0.0025)*	*8.317(±4.601)*	*-55.746(±30.714)*

Table 10. Values of a_1, a_2, b_1 and b_2 coefficients given by Cheng et al. (2014).

The dA_{temp} and dA_{year} values are given on Figure 15. Each curves of figure 15 corresponds to the evolution of a parameter for an XBTgroup. The time curves are correlated to the variation of the instrument group distribution among the ocean.



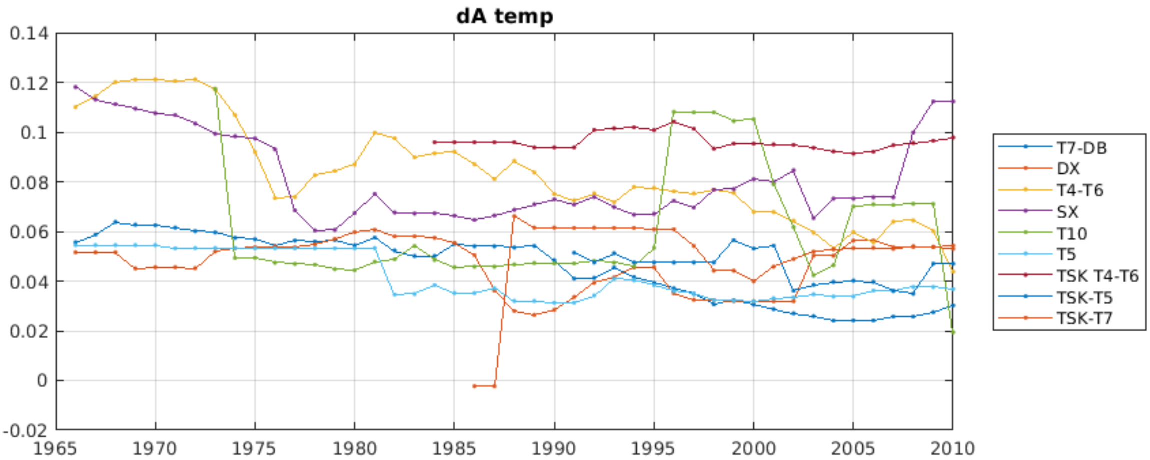


Figure 15: Evolution of dA_{temp} and dA_{year} values through the years for the XBT groups (units m/s).

This correction is divided in two parts: first the computation of the thermal offset then the correction of depth. The temperature offset and the error are calculated by comparing the XBT to reference profiles. The reference profiles are non XBT, co-located profiles (3 km ray and +/-15 days temporal frame), with a maximum average temperature of 1°C and a bathymetric difference inferior to 1000 m. Sole the QC 1 and 2 (good and probably good data) are taken into account for this operation.

II.6.2 Results for CORA-GLOBAL-5.2

Over 1 million XBT profiles have been corrected thanks to the Hamon et al. (2012) correction method over the period 1970-2016. However, for the period 1965-1965, the XBT data are too sparce to perform an accurate correction and are thus not corrected. Figure 16 shows the yearly mean difference between the measured temperature field and the co-located XBT profiles before (top) and after (bot) the XBT correction. The anomaly overreaches 0.5 °C in during the 1975-1981 period and in the vicinity of 1000 m depth from 2004 to 2014, both before and after the correction. However this anomaly is lower after the XBT correction. The early period overstated anomaly is a consequence of the low number of deep XBT profiles (>500 m) between 1975 and 1980. The late period anomaly is also caused by a lack of deep XBT profiles (T-5a and fast deep XBT) in the 2000s.

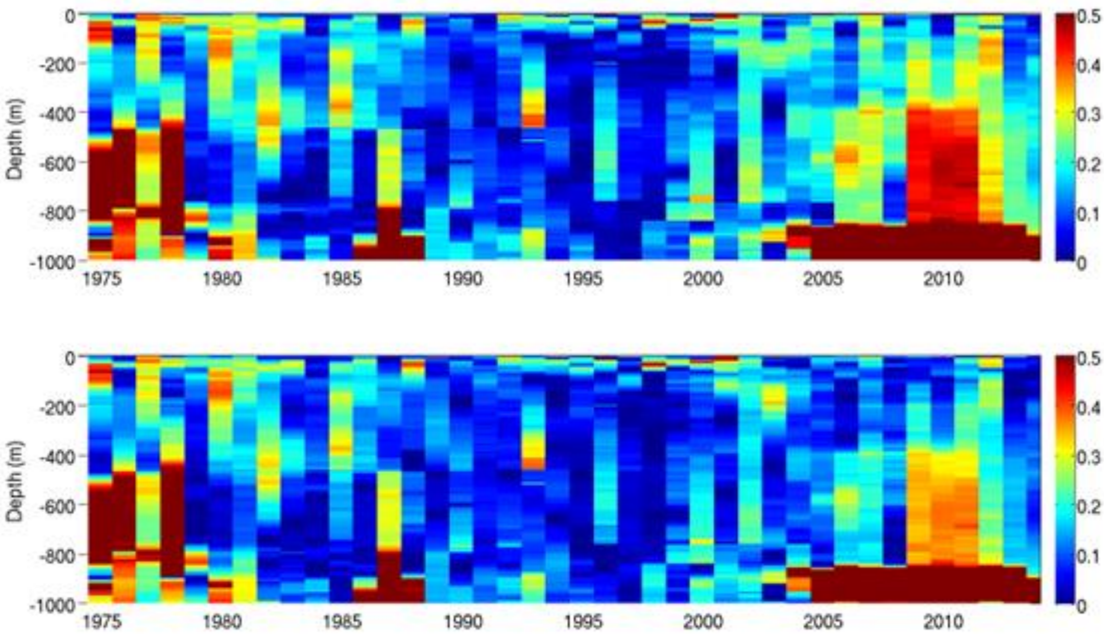


Figure 16: Yearly temperature anomaly between XBT and reference profiles before (A) and after (B) temperature offset and depth corrections in CORA5.2.

IV SYSTEM'S NOTICEABLE EVENTS, OUTAGES OR CHANGES

There have been no changes in the system version since the earlier version.

Date	Change/Event description	System version	other

V QUALITY CHANGES SINCE PREVIOUS VERSION

Date	Data quality change – evolution	System version	other
Nov 2022 release	Update in the Minmax processing field - increased data quality in well sampled coastal zones (mostly north Atlantic and Mediterranean Sea coasts) - Improved bias detection for salinity timeseries in zones covered by satellite SSS data (2010 – present for the global) - Improved drifter stranding detection and flag - Major new data profile uptake	5.2	

III REFERENCES

C. Coatanoan and L. Petit de la Villéon. Coriolis Data Centre, In-situ data quality control procedures. *Ifremer report, 17 pp*, 2005.

F. Gaillard, E. Autret, V. Thierry, P. Galaup, C. Coatanoan, and T. Loubrieu. Quality controls of large Argo datasets. *Ocean. Tech.*, 26:337–351, 2009.

M. Hamon, G. Reverdin, and P.-Y. Le Traon. Empirical correction of XBT data *.J.Geophys. Res*, 2012

K. Hanawa, P. Rual, R. Bailey, A. Si, and M. Szabados. A new depth-time equation for Sippican or TSK T-7, T-6 and T-4 expendable bathythermographs (XBT). *Deep-Sea Res.Part I*, 42:1423–1451,1995.

S. Levitus, J. I. Antonov, T.P. Boyer, R. A. Locarnini, H. E. Garcia, A. V. Mishonov. Global heat content 1955-2008 in light of recently related instrumentation problems. *Geophysical Research Letters*, 2009, 36:L07608