Delayed time mode validation Minimum / maximum temperature and salinity reference field

Tanguy Szekely[1]

[1]OceanScope, 38 rue Jim Sevellec, 29200, Brest

13-09-2024

Abstract

The twice-yearly update of the Copernicus marine product INSITU_GLO_PHY_TS_DISCRETE_MY_013_001 (hereafter referred to as the CORA dataset) implies performing a delayed-time mode validation of the temperature and salinity *in-situ* profiles. One of the most crucial tests in this validation process involves comparing the new profiles to a reference field of minimum and maximum observed temperature and salinity (DM MinMax field). This test has proven to be highly accurate for delayed-mode validation. The DM MinMax field is refined after each major release of the dataset.

1. Introduction

Improving the quality of ocean datasets by identifying and rectifying erroneous measurements has been a pivotal focus within the scientific community in recent years (Conkright et al., 2002; Cowley et al., 2021; Good et al., 2013). This aspect of dataset management has been a central objective within the CORA dataset production framework (Cabanes et al., 2012; Szekely et al., 2019). In 2017, the validation process received a significant enhancement with the introduction of the MinMax field (Gourrion et al., 2020). This method relies on generating outlier fields representing the minimum and maximum values of sea water temperature and salinity measured by in-situ instruments. Comparing new measurements to these fields efficiently aids in identifying erroneous measurements. Given its basis in known ocean variability, any measurement that deviates from the reference field but is confirmed as valid must be integrated into the next version of the field.

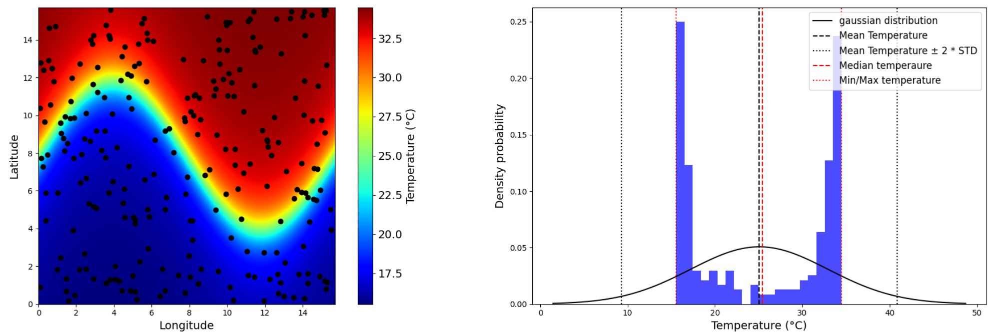
Since then, two versions of the MinMax field have been developed. The first one is tailored for application on the INSITU_GLO_PHYBGCWAV_DISCRETE_MYNRT _013_030 dataset (Gourrion et al., 2023). This dataset receives daily updates with new profiles, necessitating a swift validation process. The NRT MinMax is therefore expanded to emulate unmeasured ocean variability. When deployed, this reference field aims to efficiently detect significant errors while minimizing false detection. A second version is applied for the delayed-time mode (DM) validation of the CORA dataset. This dataset undergoes updates twice a year, and the validation process is geared towards identifying subtle errors overlooked during the NRT validation. Consequently, the DM MinMax field is sharper than its NRT counterpart, but it might be unfit for near real time quality control. This article presents the latest version of the Delayed mode MinMax field deployed for performing delayed-time mode quality control of the CORA dataset.

Section 2 gives a concise description of the MinMax detection method and grid construction, while section 3 provides a concise description of the grid. Section 4 conducts an analysis of the DM MinMax field, with a focus on comparing the Near Real Time (NRT) and delayed-time (DM) validation processes. Finally, section 5 delves into the results and discusses the utilization of the DM MinMax field.

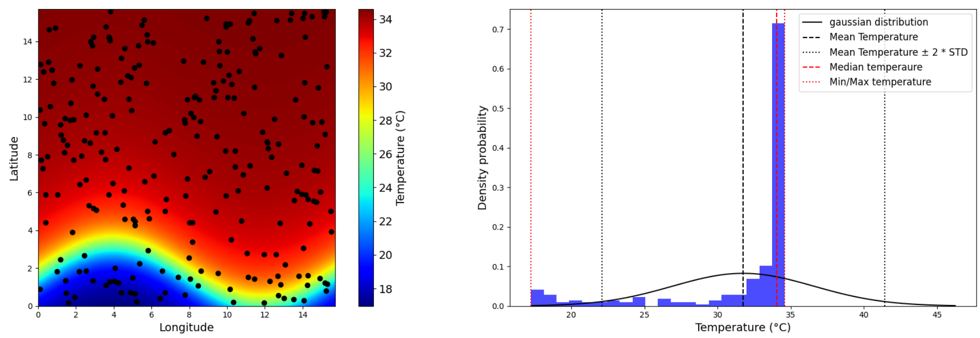
2. Method description

The DM MinMax data validation method is based on the previous work by Gourrion et al. (2020). The method strategy is to provide reference fields of minimums and maximums of temperature and salinity, and to consider as suspect any of the profiles that depart from those limits. The reference field is built by applying a dataset of carefully validated profiles on a regular grid, and to consider the minimum and the maximum of the measured temperature and salinity in each grid cell as the local minimum or maximum reference.

This method has been proven to be efficient at detecting erroneous profiles compared to methods based on climatology. Indeed, climatology fail to represent the actual ocean variability in the regions where the density probability of ocean parameters are not Gaussian. For instance, if you consider an oceanic frontal region between a warm and a cold water mass (See Fig.1, left panels). Performing repeated in-situ measurement in this zone will provide a cluster of warm profiles, and another of cold profiles. This is emulated by 250 random samples (Fig.1, left panels for sample posiitons and Fig.1, right panels, for samples density probability function. A climatology based on these two clusters will give an intermediate temperature solution, which is an improbable case. It is figured by the black dashed line of Fig.1, right panels. In addition to that, combining this climatology with an estimation of the standard deviation will give a high level of standard deviation in this zone, assuming that the warm profile and cold profiles are equally probable. Fig.1, right panels shows an example of this interval with the mean of the samples plus or minus two times the standard deviation (black dotted lines). This hypothesis is plausible on Fig.1, upper panel. On the other hand, outliers based on the detected profiles in the zone provides an more accurate estimation of the local ocean variability that will not be biased by the shape of the profiles repartition function. This case is highlighted on Fig.1, lower panels with an asymmetric probability density function. In this case the use of a climatology and standard deviation underestimates the lower outlier and overestimates the upper outlier. It has also proven to be effective at detecting small drifts in measurement instruments, as the measured variability may be very subtle at depth, enabling early detection of sensor drifts.



(a)



(b)

Figure 1: Left. Example of a temperature with 250 random samples. Right : Probability density function of the samples, with minimum and maximum outliers (red dotted line), median (red dashed line), mean (black dashed line), 2 standard deviations interval (black dotted line) and gaussian probability density function based on the samples mean and probability density function.

The DM MinMax field is produced after each major release of the CORA dataset (see table A1, in Annexe A). It is used to perform the data validation on the next version of the dataset. At the end of the validation phase, the detected profiles that have proven not to be measurement errors are included in the new field calculation to enlarge the outliers.

3. Grid definition

This validation method relies on the comparison of in-situ temperature and salinity profiles to a reference field based on the outliers among all the historical measurements taken within a grid cell. Consequently, it is crucial to select a grid cell size that is sufficiently large to include an ample number of data points, ensuring a reliable assessment of ocean variability. However, the grid cell should not be overly expansive to avoid merging distinct variabilities.

In order to test this hypothesis, it is important to compare the sampling of each grid cell without bias towards latitude. To accomplish this, an unstructured grid of hexagonal-like cells with nearly constant surface area, known as the Icosahedral Snyder Equal Area (ISEA) grid with an aperture of 4, was employed for this study.

The grid utilized for this study is the 4H6 hexagonal grid[1]. The primary criterion in grid selection was the mean inter cell distance, as a 120 km inter cell distance enables a basin-scale ocean resolution while a 30 km inter cell distance enables resolution of a 200 km wide continental shelf (details on table 1).

Table 1: ISEA grids shape parameters

Resolution	Number of Grid Cells	Inter cell Distance (km)	Standard Deviation (km)
4H6	40962	119.913	7.641
4H8	655362	29.978	1.908

The cells whose centers are positioned above the 2000 m depth bathymetry line have been excluded. This adjustment was made because it was found that this grid cell size was too large for coastal zones.

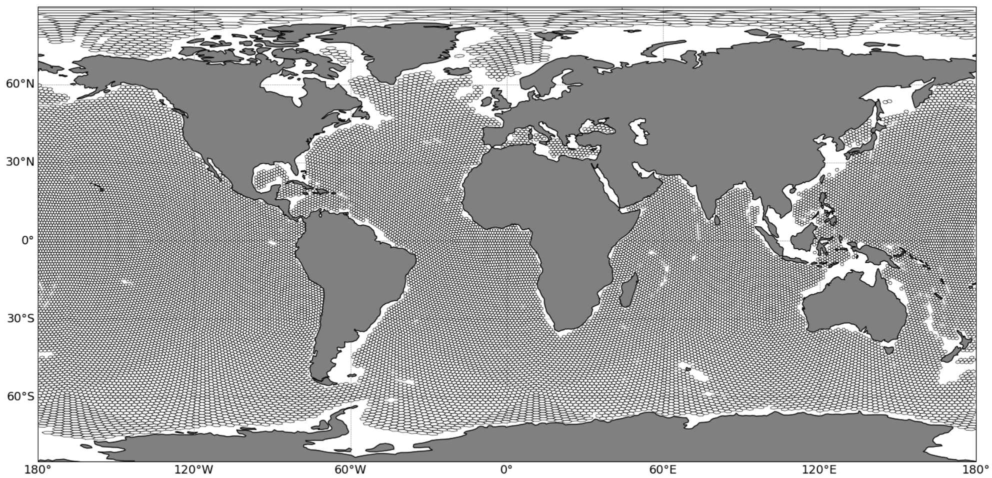


Figure 2: Position of the grid cells at surface.

Figure 2 gives the position and the limits of the grid cells. The grid vertical resolution is 20 dbar, starting from surface to 2000 dbar. This resolution is a compromise between the need for a fine resolution in regions where a sharp thermocline occurs and the necessity to ensure an adequate amount of data in each grid cell to be representative of ocean variability.

[1]https://discreteglobal.wpengine.com/

4. Field description

For the purpose of data validation, four distinct fields are generated. A TEMP MIN field that gives the minimum reference temperature. A TEMP MAX field for maximum temperature, A PSAL MIN for minimum salinity and a PSAL MAX for maximum salinity.

Figures 3 and 4 provide a comparison between the DM MinMax field and the NRT MinMax field. Notably, the NRT MinMax field exhibits variations when compared to the DM MinMax field. One key factor contributing to these differences is the use of distinct reference datasets.

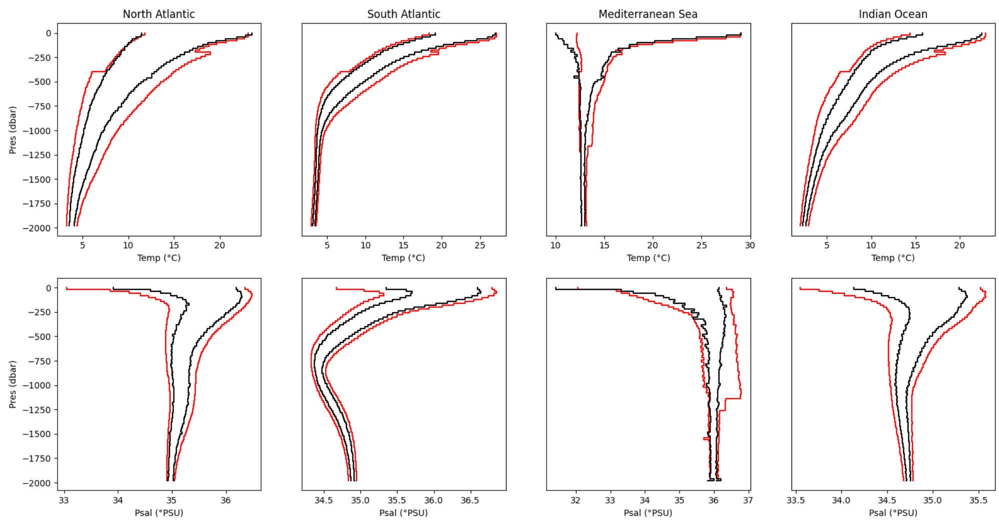


Figure 3: Mean reference fields for temperature and salinity for North Atlantic, South Atlantic, Mediterranean Sea and Indian Ocean for DM MinMax (black) and NRT MinMax (red)

To address data gaps, a widening parameter is introduced, which proportionally increases either the minimum or maximum field values relative to the median. The selection of this widening parameter is a critical decision aimed at maintaining a detection rate that aligns with the requirements of the NRT validation process. This parameter choice is pivotal in achieving a balance between data accuracy and the ability to effectively identify relevant information in real-time scenarios.

The global shape of the DM and NRT MinMax temperature fields are quite close, except two stalls at 200 dbar for temperature max and 400 dbar for temperature min. They are produced by the NRT MinMax enlargement function. In all cases, the temperature and salinity DM outliers are embodied inside the NRT reference field.

Table 2 and 3 give an overview of the performances of the DM MinMax field of the validation of temperature (TEMP) and salinity (PSAL) profiles. The difference in TEMP and PSAL performances mostly lies at the instrument level. First, there is more TEMP profiles than PSAL profiles since most of the XBT measurements are TEMP only profiles. Second, the majority of PSAL measurements are based on a conductivity sensor. The

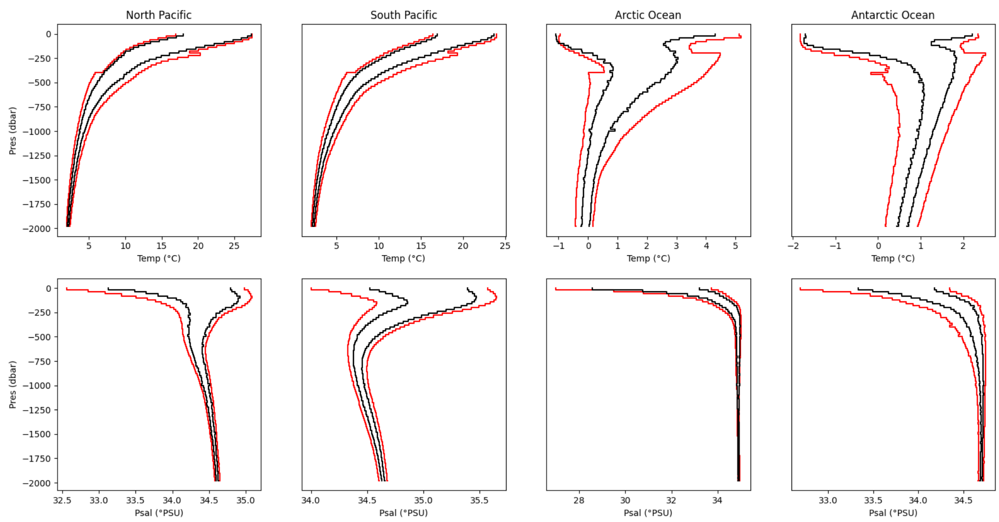


Figure 4: Mean reference fields for temperature and salinity for North pacific, South Pacific, Arctic Ocean and Antarctic Ocean for DM MinMax (black) and NRT MinMax (red)

relation between salinity and conductivity strongly depends on temperature so a small measurement error on temperature induces a large measurement error on salinity. A consequence is that when TEMP and PSAL are measured by the same instrument, an error on the TEMP measurement is almost for sure associated to and error on PSAL on the same level, but an error on PSAL may not be related to an error en TEMP.

There are two different approaches in the validation statistics. First, the validation approach in use for the CORA dataset is a profile based approach. The DM MinMax field is used as a reference field. If any measurement from a tested profile is located below the minimum or above the maximum, then the profile is visually checked and its measurements may be flagged. Another approach is a data point based approach. In this approach, all data points departing from a reference field are flagged as bad. While using this approach, the user should consider suspicious the profiles with all the data points at depth flagged as bad. Most of these profiles are indeed subject to a sensor drift, detected at depth where the ocean variability is low, but undetected in subsurface where ocean variability is higher.

The columns "Profiles detected'" and "Points detected" in Table 2 and 3 help to have an overview of the efficiency of the NRT and DM MinMax fields for a profile based and a data point based validation approach. It appears that the difference between the NRT MinMax and the DM MinMax are often higher when using the pointwise approach rather than using the profile centered approach.

The Statistical difference between basins mostly depends on the repartition of the measurements and the accuracy of the instruments. For instance the numerous flags in the Antarctic basin is related to the wide deployment of sea mammals born instruments in this region. These instruments are subject to sensor drift. This drift is adjusted in the

MEOP database, but a large part of the sea mammals profiles distributed in the CORA database are still not adjusted.

Basin	Profiles detected by NRT MinMax	Profiles detected by DM MinMax	% increase	Points detected by NRT MinMax	Points detected by DM MinMax	% increase
North Atlantic	2788	6748	142%	59842	210346	252%
South Atlantic	1790	2941	64%	138822	834544	501%
North Pacific	14254	34595	143%	509476	4133094	711%
South Pacific	4796	6134	27%	306027	783402	156%
Mediterranean Sea	284	325	14%	2292	3683	61%
Indian Ocean	6579	9131	39%	387085	868961	124%
Antarctic Ocean	905	1942	114%	96507	251586	161%
Arctic Ocean	317	445	40%	15885	32437	104%

Table 2: Profiles and data points detected by TEMP NRT MinMax and DM MinMax

Basin	Profiles detected by NRT MinMax	Profiles detected by DM MinMax	% increase	Points detected by NRT MinMax	Points detected by DM MinMax	% increase
North Atlantic	4253	6363	50%	635681	1115517	75%
South Atlantic	2125	2385	12%	117725	282950	140%
North Pacific	16938	23775	40%	2222490	4390167	98%
South Pacific	5961	8882	49%	496227	1039142	109%
Mediterranean Sea	2152	2273	6%	304354	419724	38%
Indian Ocean	11955	17173	44%	348679	1355997	288%
Antarctic Ocean	9639	26176	172%	411781	1825101	343%
Arctic Ocean	704	1243	77%	57069	146638	157%

Table 3: Profiles and data points detected by PSAL NRT MinMax and DM MinMax

5. Discussion

The DM MinMax field complements the NRT MinMax field by providing sharper detection of outliers, albeit at the cost of increased operational effort. Notably, it has proven to be more effective in identifying sensors that may have drifted, potentially affecting global ocean properties such as heat content and mean halosteric height.

However, it's important to note that this method relies on a fixed reference field and doesn't account for the effects of climate change. Consequently, regular updates are essential to ensure its effectiveness with new measurements. Users must exercise caution when processing data dated after the final date of the reference dataset to maintain accuracy and relevance.

Last, this reference field is linked to the current version of the CORA dataset. It aims to be improved at every major CORA update, and to be updated on a yearly basis.

References

Cabanes, C., Grouazel, A., von Schuckmann, K., Hamon, M., Turpin, V., Coatanoan, C., & Traon, P.-H. L. (2012). The cora dataset: Validation and diagnostics of ocean temperature and salinity in situ measurements. *Ocean Science Discussions*, 9(2), 1273–1312.

Conkright, M., Levitus, S., Antonov, J. I., Baranova, O., Boyer, T., H.E. & Garcia, C. S. (2002). World ocean database 2001 and world ocean atlas 2001. *AGU Fall Meeting Abstracts*, OS52B–0220.

Cowley, R., Killick, R., Boyer, T., Gouretski, V., Reseghetti, F., Kizu, S., & Domingues, C. (2021). International quality-controlled ocean database (iquod) v0. 1: The temperature uncertainty specification. *Frontiers in Marine Science*, 8, 689–695.

et al., A. W. (2020). Argo data 1999–2019: Two million temperature-salinity profiles and subsurface velocity observations from a global array of profiling floats. *Frontiers in Marine Science*, 7(700).

Good, S., Martin, M., & Rayner, N. (2013). En4: Quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates. *Journal of Geophysical Research: Oceans*, 118(12), 6704–6716.

Gourrion, J., leroy, D., & Coatanoan, C. (2023). Minmax qc approach: Characterization of alert statistics and configuration for cms nov. 2023 eis case of argo t/s dataset. *Technical Note 202312-MinMax-01*. https://doi.org/https://doi.org/10.13155/88904

Gourrion, J., Szekely, T., Killick, R., Owens, B., Reverdin, G., & Chapron, B. (2020). Improved statistical method for quality control of hydrographic observations. *J. Atmos. Oceanic. Technol.*, 37, 789–806.

Szekely, T., Gourrion, J., pouliquen, S., & Reverdin, J. (2019). The cora 5.2 dataset for global in situ temperature and salinity measurements: Data description and validation. *Ocean Science*, 15(6), 1601–1614.

A. Dataset description

This section describes the dataset used to build the current version of the DM MinMax field. It aims to be updated together with every field update, as any difference in the dataset may modify the known ocean variability outliers and thus change the reference field shape.

The current version of the reference field (MinMax V1.0) is based on the In-Situ November 2023 release of the CORA dataset, distributed by the Copernicus marine Service[2]. The details of the release scope are shown on Table 2.

Table A1: Dataset description

Name	INSITU_GLO_PHY_TS_DISCRETE_MY_013_001 (CORA dataset)
Release date	November 2023
Parameters	Temperature, practical salinity
Coverage	1950-2022, Global Ocean
Dataset doi	https://doi.org/10.17882/46219

The profiles have been validated at first in near real time, by the Ifremer team[3] and then a delayed time mode validation by the OceanScope team. This procedure ensures a near flawless dataset with the minimum of over flagged data. It is necessary to provide a picture of the ocean variability as accurately as possible, and thus to improve the quality and the accuracy of the DM MinMax field. Figure A1 gives an overview of the yearly number of profiles used for the DM MinMax definition.

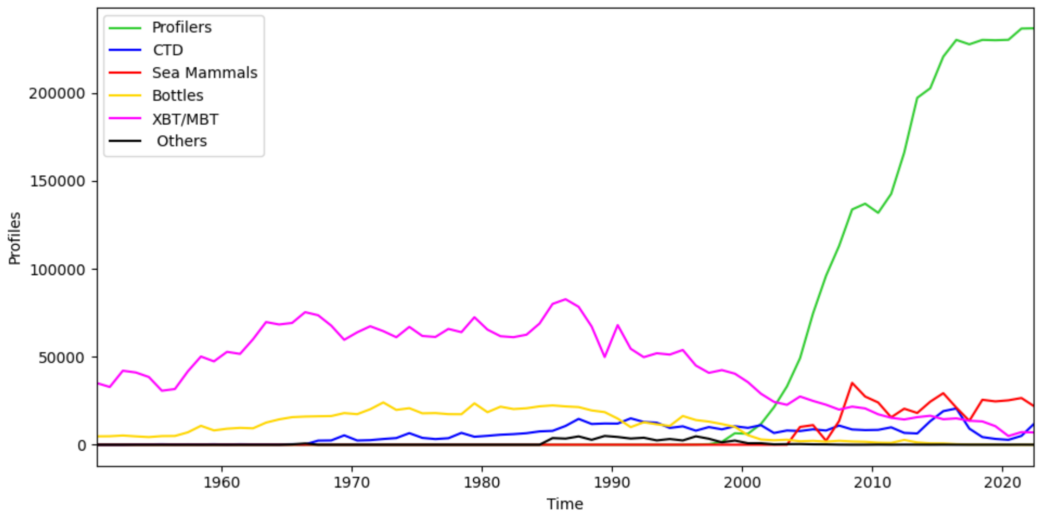


Figure A1: yearly number of profiles for each data types.

[2]https://data.marine.copernicus.eu/product/INSITU_GLO_PHY_TS_DISCRETE_MY_013_001/services
[3]https://marineinsitu.eu/partners/ifremer/

The data types selected for this field are listed below.

A.1 ARGO Profiles

The ARGO dataset is probably the dataset with the best data in terms of data quality and space repartition available for the scientific community (et al., 2020). ARGO floats are autonomous profilers equipped with a CTD sensor (Conductivity, Temperature, Depth). They are deployed in the deep ocean and often drift at 1000 dbar. Once a week they dive up to 2000 dbar or 6000 dbar level and perform a measurement profile from this level to the surface and transmit the data afterward. With a life expectancy of 4 years, each profiler can provide hundreds of profiles. The floats are deployed by more than 30 nations and reached global coverage in 2008. The yearly number of profiles however kept increasing up to over 20000 yearly profiles in 2014 (Figure A1). The space coverage is very high in the northern hemisphere, especially in the North Atlantic, Mediterranean Sea and the North-West pacific ocean. The coverage is however lower in the subtropical gyres and in the Arctic and Antarctic seas (A2).

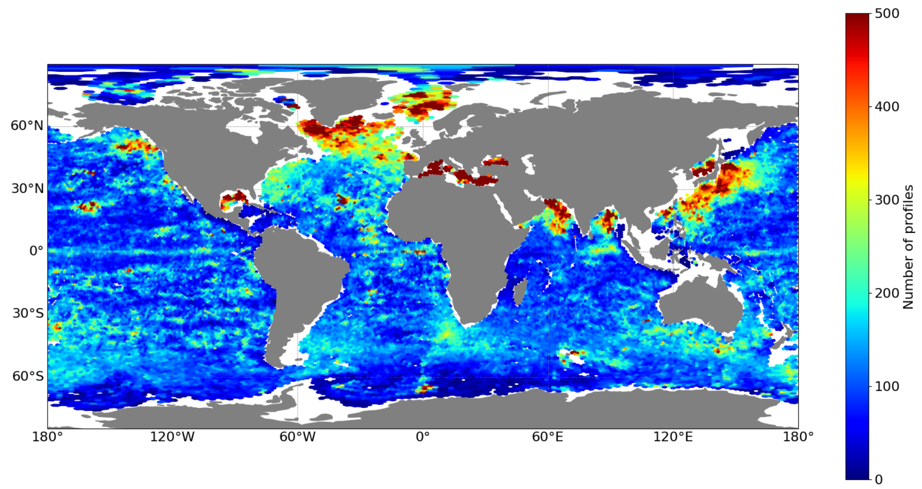


Figure A2: Number of ARGO profiles in selected DM MinMax grid cells.

The ARGO profiles also benefit from the data processing and adjustment procedures of the ARGO data community. Its regular coverage during over 15 years ensures to provide an accurate estimation of the minimum and maximum temperature and salinity, at least at depth where ocean variability is the lowest.

A.2 Expandable bathythermograph/mechanical bathythermograph (XBT/MBT)

The XBT/MBT dataset is the second dataset in terms of number of profiles after the ARGO dataset. The MBT is an instrument that measures sea water temperature thanks to a thermistor plunged in the ocean. The sea water pressure is estimated thanks to the immersion level. The XBT is an expandable version of the MBT where the thermistor is free falling beneath the sea surface. The immersion level is calculated thanks to an estimated falling rate. The XBT/MBTs were widely used during the 1970s and the 1980s. They are still in use today, particularly along commercial lines equipped with automatic launchers. The XBTs are the most common instruments between the 1970s and the 2000s. There is only a few models of XBTs, with a fixed maximum depth. There is consequently a gap in the vertical coverage at 460 m (T4 type XBTs max depth), 760 m (T7 and Deep Blue max depth), 1000 m (Fast Deep max depth) and 1850 m (T11 max depth).

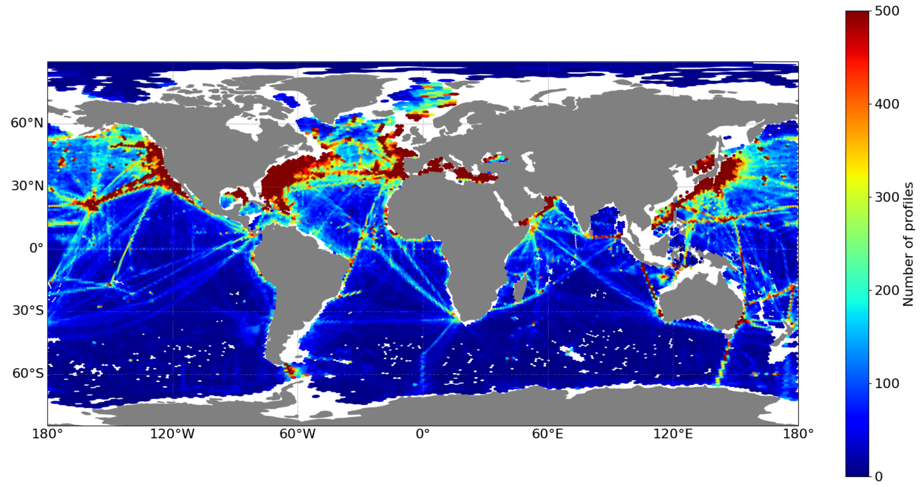


Figure A3: Number of XBT profiles in selected DM MinMax grid cells

The XBT/MBT data coverage is very high in the northern hemisphere, but also along the main commercial lines (Fig. A3). It is however very low in the Antarctic and beside the commercial shipping lines.

The only XBT type instruments able to measure the practical salinity are the XCTSDs. They have been produced in low numbers compared to the XBT, so most on the XBT/MBT type profiles selected here have no practical salinity measurements.

A.3 Conductivity, temperature, depth instruments (CTD)

The CTDs are oceanographic instruments designed to measure the sea water conductivity, the temperature and sea water pressure. It can thus calculate the practical salinity. These instruments have been widely used by the scientific community for the accuracy of the temperature and salinity measurements. Most of the profiles used for the definition of the DM MinMax field are thus located along scientific cruise transects and measurement campaigns. They are mostly located in the northern hemisphere.

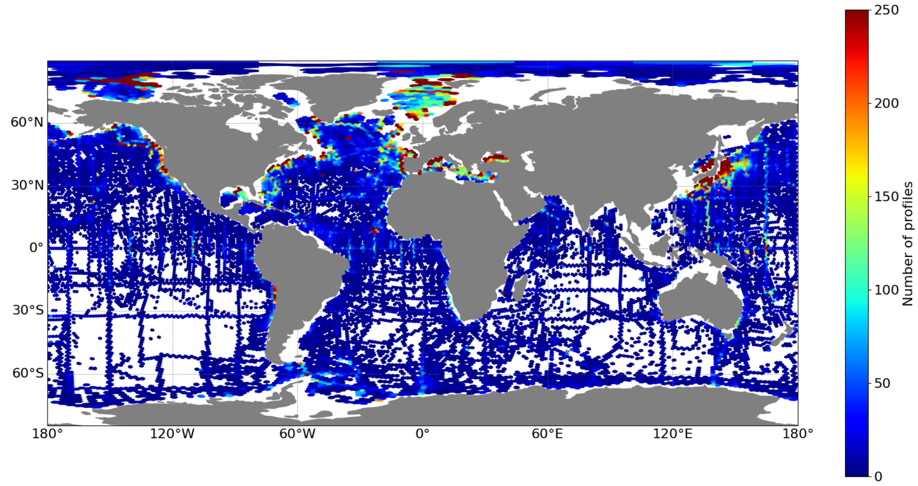


Figure A4: Number of CTD profiles in selected DM MinMax grid cells

A.4 Bottles measurements

The profiles labelled "bottle measurements" are temperature and salinity measurements of the water trapped in a bottle. The bottles instruments have mostly been deployed in a "Rosette" configuration, a cage with a limited number of bottles fixed, trolled along a cable. The bottles are hermetically closed at the measurement level. The salinity is measured from the sample collected in the bottles and the temperature from an internal mercury thermometer on an extra CTD device.

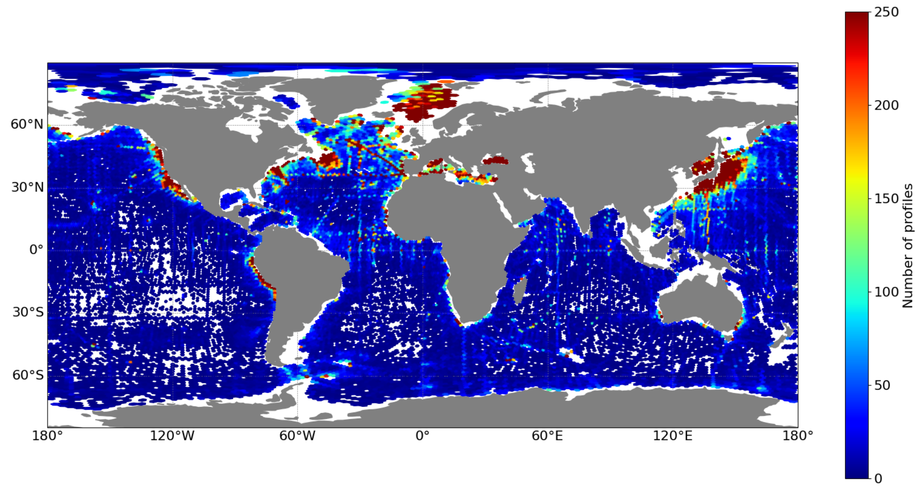


Figure A5: Number of bottles profiles in selected DM MinMax grid cells

The bottles were already in use in the 1950 (Figure A1), and are still in use now, mostly for calibration purposes, or for chemical samplings along with salinity. The Rosette device has however a limited number of bottles, and thus a limited number of points on the vertical.

A.5 Sea Mammals mounted CTD.

The MEOP[4] project has gathered measurements from tags attached to marine mammals (Mostly elephant seals). These tags provide temperature and salinity profiles along the marine mammal migration routes. A low frequency version of the profile is transmitted after all dives of the animal, but some sensors are recovered after the seasonal moult, giving the high frequency version of the profiles. The major deployment zones are the West coast of the USA, the Kerguelen Islands and Spitzberg. Giving a unique vision of the ocean hydrology in these regions. These datasets compensate for the lack of data of the ARGO floats in zones with seasonal sea ice.

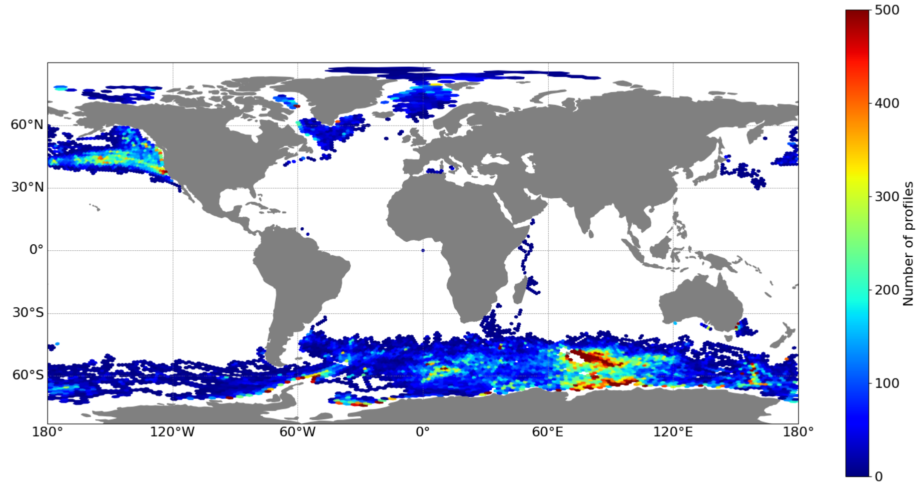


Figure A6: Number of sea mammals profiles in selected DM MinMax grid cells

[4]https://www.meop.net/