frontiers | Frontiers in Marine Science

TYPE Technology and Code
PUBLISHED 02 October 2024
DOI 10.3389/fmars.2024.1403175

OPEN ACCESS

EDITED BY
Pierre Yves Le Traon,
Mercator Ocean, France

REVIEWED BY
Antonio Novellino,
ETT SpA, Italy
Noir Primadona Purba,
Padjadjaran University, Indonesia

*CORRESPONDENCE
Lijing Cheng
chenglij@mail.iap.ac.cn

†These authors have contributed equally to this work and share first authorship

RECEIVED 19 March 2024
ACCEPTED 10 September 2024
PUBLISHED 02 October 2024

CITATION
Song X, Tan Z, Locarnini R, Simoncelli S, Cowley R, Kizu S, Boyer T, Reseghetti F, Castelao G, Gouretski V and Cheng L (2024) DC_OCEAN: an open-source algorithm for identification of duplicates in ocean databases. *Front. Mar. Sci.* 11:1403175. doi: 10.3389/fmars.2024.1403175

DC_OCEAN: an open-source algorithm for identification of duplicates in ocean databases

Xinyi Song[1,2†], Zhetao Tan[1,2†], Ricardo Locarnini[3], Simona Simoncelli[4], Rebecca Cowley[5], Shoichi Kizu[6], Tim Boyer[3], Franco Reseghetti[4,7], Guilherme Castelao[8], Viktor Gouretski[1] and Lijing Cheng[1,2*]

[1]International Center for Climate and Environment Sciences, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China, [2]University of Chinese Academy of Sciences, Beijing, China, [3]National Centers for Environmental Information, National Oceanic and Atmospheric Administration, Silver Spring, MD, United States, [4]Istituto Nazionale di Geofisica e Vulcanologia (INGV), Bologna, Italy, [5]Climate Science Centre, Environment, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Hobart, TAS, Australia, [6]Physical Oceanography Laboratory, Department of Geophysics, Tohoku University, Sendai, Japan, [7]Italian National Agency for New Technologies, Energy and Sustainable Economic Development (ENEA), Santa Teresa Research Centre, Lerici, Italy, [8]Scripps Institution of Oceanography (SIO), University of California, San Diego, La Jolla, CA, United States

A high-quality hydrographic observational database is essential for ocean and climate studies and operational applications. Because there are numerous global and regional ocean databases, duplicate data continues to be an issue in data management, data processing and database merging, posing a challenge on effectively and accurately using oceanographic data to derive robust statistics and reliable data products. This study aims to provide algorithms to identify the duplicates and assign labels to them. We propose first a set of criteria to define the duplicate data; and second, an open-source and semi-automatic system to detect duplicate data and erroneous metadata. This system includes several algorithms for automatic checks using statistical methods (such as Principal Component Analysis and entropy weighting) and an additional expert (manual) check. The robustness of the system is then evaluated with a subset of the World Ocean Database (WOD18) with over 600,000 *in-situ* temperature and salinity profiles. This system is an open-source Python package (named DC_OCEAN) allowing users to effectively use the software. Users can customize their settings. The application result from the WOD18 subset also forms a benchmark dataset, which is available to support future studies on duplicate checks, metadata error identification, and machine learning applications. This duplicate checking system will be incorporated into the International Quality-controlled Ocean Database (IQuOD) data quality control system to guarantee the uniqueness of ocean observation data in this product.

KEYWORDS

duplicate checking, ocean data infrastructure, ocean in-situ observations, ocean data quality improvement, temperature and salinity

1 Introduction

Ocean *in-situ* observational data, such as ocean temperature and salinity profiles, are essential for understanding changes in the ocean and climate. Within the ocean science community, researchers deploy various instruments and sensors, such as CTD (Conductivity, Temperature, Depth), XBT (Expendable Bathythermographs), MBT (Mechanical Bathythermograph), Argo, APB (Autonomous Pinniped Bathythermographs), and moored buoys, to gather data from the surface to the deep ocean (Boyer et al., 2018). These profiles are crucial for research in monitoring ocean warming, ocean stratification, vertical mixing, circulation, etc., and also invaluable for policy-makers and science outreach (Mackenzie et al., 2019). For this purpose, the international ocean data exchange centers, such as the World Meteorological Organization (WMO) and the National Oceanic and Atmospheric Administration (NOAA), play crucial roles in data collecting, integration, standardizing, formatting, duplicates removal, quality control and distribution of oceanographic data from various institutions around the world (Boyer et al., 2018; Goni et al., 2019; Abraham et al., 2013; Tan et al., 2023). These actions ensure the oceanographic data are reliable, comparable, and accessible, therefore, supporting multi-disciplines from ocean-related climate research, ocean and weather forecasting to marine ecosystem management (IPCC, 2021).

However, a major challenge in data integration is duplicate checking and removal, which is a major part of quality control of ocean data. The duplicates can occur from the start of data ingestion into a database to data distribution at the end. For example, during the data assembly stage, integrating the same data from multiple data centers in slightly different forms is a frequently encountered problem in ocean data management (Locarnini et al., 2019). During the data transmission, the same records can be transmitted to different decimals, leading to two "different" profiles (Boyer et al., 2018). How to identify the duplicates and then remove duplicates is still a major challenge in the operational oceanography community (Cowley et al., 2023), as the need for duplicate checking becomes especially evident when integrating data from different infrastructures into a more comprehensive database, particularly when consolidating various data sources.

For example, merging data from the Global Temperature-Salinity Profile Program (GTSPP) and data from other sources into the World Ocean Database (WOD) (a project of the International Oceanographic Data Exchange - IODE), will inevitably result in the generation of data with the same observations in the same location at the same time. These duplicates especially have a non-negligible impact on the ocean state estimates (Levitus, 1982; Boyer and Levitus, 1994; Ishii et al., 2017; Simoncelli et al., 2021; 2022; Good et al., 2023; Cheng et al., 2024) and the data assimilation in ocean reanalyzes (Escudier et al., 2021; Carton and Giese, 2008; Balmaseda et al., 2015). One example is the ocean temperature field reconstruction and ocean heat content estimates, which will be illustrated in the following section. At present, some activities in different countries exist for duplicate checking work, for instance, Copernicus Marine Service (Szekely et al., 2024), EN4 from Met Office of the United Kingdom (Good et al., 2013), and Chinese Academy of Sciences Ocean Database (Zhang et al., 2024), etc. Furthermore, analyzing the duplicate checking result serves as a crucial process in identifying problematic data or errors in metadata (Cowley et al., 2023).

To identify duplicated data, some automatic or manual duplicate-checking algorithms or systems (a collection of various algorithms) have been proposed to keep the best data version and to remove or label the duplicates. For example, Gronell and Wijffels (2008) defined "exact duplicated data" and "near duplicated data" and proposed a semi-automatic approach that combines expert manual quality control techniques and automatic statistical checks to identify these two kinds of duplicates. This approach has been used to construct the fourth version of the "EN" series of datasets by Good et al. (2013). Cabanes et al. (2021) also applied this method to Delayed-Mode Quality Control (DMQC) analysis for Argo data. The National Oceanic and Atmospheric Administration (NOAA)/National Centers for Environmental Information (NCEI) data centers designed a system to identify duplicates in various data sources, which is used in the real-time update stream of the WOD database (Garcia et al., 2018). Durack and Wijffels (2010) attempted to identify duplicate salinity profiles by finding matches in time and location within 1 day and 0.02 degrees (in both latitude and longitude). Ji et al. (2022) used a set of specific location, time, and depth thresholds to identify and remove duplicate data for different instruments. Schmidtko et al. (2017) defined duplicated profiles as data pairs within a 5 km distance and 25 hours. However, previous approaches have limitations:

1. Algorithms that rely only on limited metadata (like geographical coordinates) might not be applicable to errors in that metadata, for instance, comparing each profile to the co-located profiles within a 0.1° latitude and longitude box (Gronell and Wijffels, 2008). However, in many cases, duplicate data are widely separated in geographical locations, i.e. the longitude and latitude of two profile data may be opposite, while the rest of the metadata information remains the same. Such profiles have long been neglected as duplicate data.
2. Some algorithms only consider certain metadata (e.g., time, latitude, longitude, and depth) as key information to identify duplicates (Ji et al., 2022), thus are only capable of identifying a subset of duplicate.
3. Previous research has primarily focused on identifying and deleting duplicate data (Gronell and Wijffels, 2008) but there is no available benchmark dataset for duplicate-checking. Detailed analyses of the underlying reasons for the occurrence of duplicates are not always documented.
4. Identifying potential duplicate pairs in previous studies was typically viewed as a time-consuming task due to the necessity of conducting one-by-one comparisons, where each profile had to be individually compared against every other, leading to a substantial number of comparisons in large amounts of data (Ji et al., 2022).
5. The lack of open-source algorithms limits duplicate checking for further use and broader applications.

The goal of this paper is to present some criteria for identifying duplicates and then develop a new duplicate checking algorithm for ocean *in-situ* profiles (Section 2). The method consists of a semi-

automatic procedure, based on crude screening and target screening, which is followed by a manual expert check to review the identified duplicates. This method is developed in an open-source Python package (named DC_OCEAN (https://pypi.org/project/DC-OCEAN/). We incorporate two steps for the automatic checks: 1) Crude screening: it aims to identify as many possible duplicates as feasible by using the Profile Summary Score metric; 2) Targeted screening: it aims to refine the analysis by further selecting possible duplicates based on the results of the crude screening and then by classifying them into various categories. In addition to the automatic checks, outputs from manual duplicate checking (expert screening) are included in the package to validate results of the automatic duplicate checking process. The proposed method is validated (Section 3) by utilizing data sourced from WOD (Boyer et al., 2018; downloaded in February 2022), and a benchmark dataset containing the resulting duplicate flags is released (Section 4). Section 5 encompasses the conclusion and discussion, followed by Section 6 which provides a summary of the data and code developed within this study.

2 Methods

2.1 Definitions

The following two types of duplicates are defined and used in this study:

1. Possible duplicates (the results of automatic algorithms in section 2.3): refer to profiles in which not all the metadata information is identical (Gronell and Wijffels, 2008) or to profiles with erroneous data/metadata.
2. Exact duplicates: refer to profiles with identical measurements and metadata (Gronell and Wijffels, 2008) or "possible duplicates" that have been checked by experts and confirmed to be "exact duplicates".

Because the data accompanied by metadata issues or data problems do not qualify as duplicates, once the identified problems have been resolved or corrected, the data will be either confirmed as exact duplicates or determined to be non-duplicates by experts. This step involves manual duplicate checking (Section 2.2.3). We categorize data with identified data/metadata errors during manual duplicate checking as non-duplicates, while the other data confirmed by experts are considered to be "exact duplicates". For example, profiles close in time and space may be due to metadata errors (our algorithm will find them). If manual duplicate checking reveals that there is indeed an error in the time or location information of the profiles, these profiles will be considered non-duplicates; otherwise, they will be considered exact duplicates.

2.2 Criteria to identify duplicates

We establish seven criteria to identify duplicates based on expert experience and oceanographic knowledge.

2.2.1 Criteria 1

If all the metadata and measurements of depth, temperature, salinity, dissolved oxygen, etc., in two or more profiles are identical, they are directly classified as exact duplicates, and no manual check is required.

2.2.2 Criteria 2

Profiles observed at nearly the same location (within 1 km, considering the resolution of old instruments), at nearly the same time (less than one hour), by the same ship, with the same instrument, are classified as possible duplicates. One of the typical cases is when a vessel has repeated observations at a given spot within a short period for calibration purposes; these data will be identified as possible duplicates.

2.2.3 Criteria 3

Profiles observed by the same ship, at the same time (difference less than one hour, empirically), but at different locations (with a distance threshold of >=30km, which is an empirical choice) are classified as possible duplicates that require further expert validation as a vessel cannot be in multiple places simultaneously.

2.2.4 Criteria 4

Some data originators are known to store and/or distribute data after applying a numerical scaling (i.e., multiplying by a constant factor) or an offset (e.g., adding a constant). These data are measured near the same locations (<=500m, empirically chosen) with the same instrument. Such modified data are identified as possible duplicates here but will be further assessed with expert check.

2.2.5 Criteria 5

Profiles are collected at nearly the same time (<=1 hour) and at nearly the same location/station (within 1 km), but their records are rounded off or truncated. This arises mainly because of the data processing methods by different data originators. These profiles are classified as possible duplicates.

2.2.6 Criteria 6

Measurements are identical, but some parts of the metadata are different. In this case, the profiles are classified as having metadata error requiring further expert review. These profiles are considered as possible duplicates.

2.2.7 Criteria 7

Some measurements are missing, or the values have been interpolated. Because of storage limitations or data processing policies (e.g., practice applied back in time), some data may have been resampled or interpolated before being submitted to data centers. If 85% (an empirical choice) of the measurements are identical in two profiles, these profiles are classified as possible duplicates.

2.2.8 Instrument specific processes

In addition, the position accuracy of the coordinate information for modern Conductivity-Temperature-Depth (CTD), Autonomous

Pinniped (APB), Profiling Float (PFL), and Glider (GLD) is much higher than other instruments due to the use of GPS systems. In such cases, if the longitude or latitude of the profiles differ by four decimal places (empirical chosen), they cannot be classified as possible duplicates (i.e., not passed to the expert check if none of the previously listed criteria is fulfilled).

2.3 Duplicate checking workflow

The semi-automatic algorithm, schematized in Figure 1, consists of two parts: (1) automatic duplicate checking (N00 processes, N01 crude screen and M00 targeted screen) and (2) manual (expert) duplicate checking. Criteria 1 to Criteria 7 are applied in M00.

2.3.1 Automatic checking

Based on the criteria proposed above, we developed an automatic duplicate checking system that calculates a "Profile Summary Score (PSS)" for each profile by integrating measurements and various metadata such as country, time, location, instrument type, etc. These metrics allow us to efficiently compare profiles. This study assumes that the profiles with similar PSS have a large probability to be duplicats.

Program N00 in Figure 1 is used for pre-processing metadata and secondary processing data (e.g., the sum of temperature, sum of salinity, if available). Table 1 shows all the available metadata and secondary processing data used to calculate the Profile Summary Score. During this stage, numerical metadata such as time, longitude, and latitude are retained, while string metadata information (such as country and platform) is converted into numerical values by using the ASCII code table (e.g., letter "A" is 65) and then summing these ASCII code values of each string to derive final numerical values. For example, the sum of ASCII code for string "NODC" is 292 (78 + 79 + 68 + 67).

Numerical values outside the range set in Table 1 are set to NaN (i.e., 'np.isnan' in Python). The missing string values are set to empty values (i.e., '' in Python; empty). These two types of missing values will be ignored during the calculation of the PSS.

Program N01 includes multiple screening processes to detect possible duplicates. This program consists of three independent strategies: 1) Arithmetic Mean; 2) Entropy Weight Method; 3) Principal Component Analysis (PCA) method. The three independent strategies are summed up to a final score because they leverage different statistical techniques to synthesize the Profile Summary Score from complex and varied metadata and measurements. From this program, information on each profile can be reduced to a single numerical value as a flag so that the comparison between profiles is no longer a rigid comparison of one-by-one correspondence between metadata. This approach enhances the robustness of our duplicate detection by ensuring that multiple facets of the profiles are considered and thereby catching multi-types of duplicates as much as possible, with minimizing the likelihood of overlooking duplicates that only one strategy might miss. The order to perform these three independent strategies does not change the result because they are summed up to give a final score.

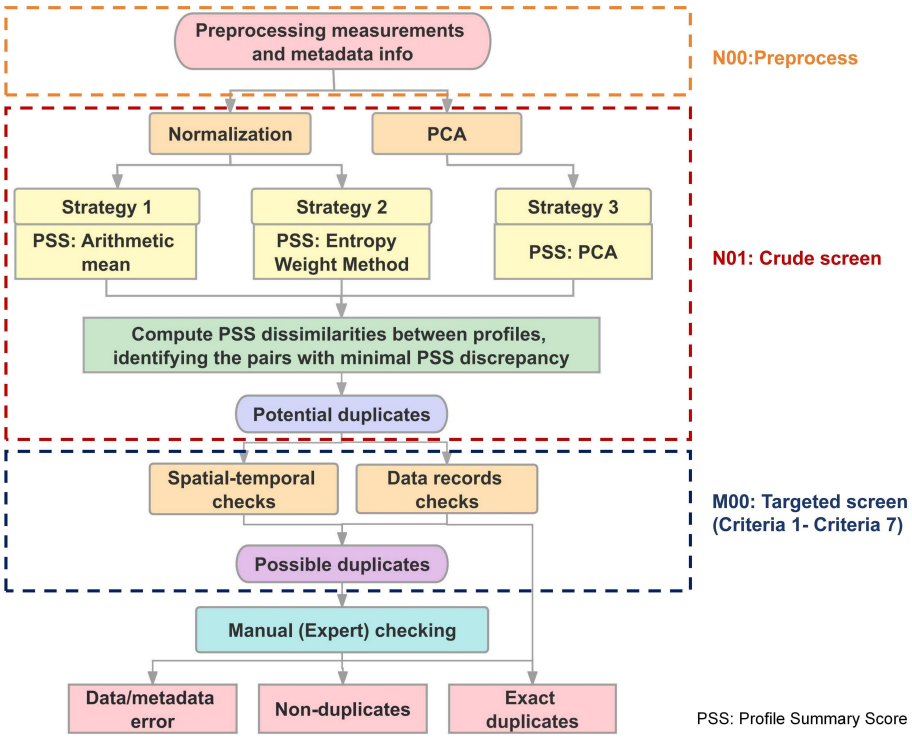

FIGURE 1
Flow chart of the DC_OCEAN duplicate checking.

10.3389/fmars.2024.1403175

TABLE 1 Metadata and secondary processing data extracted from the WOD profiles and used to calculate the Profile Summary Score in the duplicate checking system (DC_OCEAN).

Name	Data Type	Long_name	Dimension	Range
z	float	depth	1D	0 m~12000 m
Temperature	float	Sea water temperature	1D	-2 °C ~40 °C
Salinity	float	Sea water salinity	1D	0 psu~50 psu
Oxygen	float	Sea water dissolved oxygen	1D	–
Chlorophyll	float	Chlorophyll	1D	–
lat	float	latitude	–	-90°~90°
lon	float	longitude	–	-180°~180°
time	float	time	–	–
country	string	country name	–	–
Temperature_Instrument	string	Temperature Instrument	–	–
need_z_fix	string	Instruction for fixing depths for XBT bias correction	–	–
recorder	string	Recorder	–	–
GMT_time	float	GMT_time	–	–
WMO_ID	integer	WMO identification code	–	–
dbase_orig	string	original database	–	–
project_name	string	Project name	–	–
Platform	string	Platform name	–	–
ocean_vehicle	string	Ocean vehicle name	–	–
accession_number	integer	NODC accession number	–	–
Institute	string	Institute name	–	–
WOD_cruise_identifier	string	WOD cruise identifier	–	–
dataset_id	string	dataset name	–	–
sum_temp	float	Sum of temperature	–	–
sum_salinity	float	Sum of salinity	–	–
sum_depth	float	Sum of depth	–	–
std_temp	float	Standard deviation of temperature	–	–
std_salinity	float	Standard deviation of salinity	–	–
std_depth	float	Standard deviation of depth	–	–
depth_number	float	Number of depth	–	–
maximum_depth	float	Maximum depth	–	–
cor_temp_depth	float	Correlation coefficient between temperature and depth	–	–
cor_sal_depth	float	Correlation coefficient between salinity and depth	–	–

The details to calculate the Profile Summary Score are shown as follows:

1. Strategy 1 (Arithmetic Mean): we calculate the simple average of metadata numerical values (converted in the Program N00) for each profile for comparison purposes. Here, before calculating the mean, the algorithm includes a normalization process for all numerical metadata to better homogenize the information and reduce the influence of differences in dimensions and units in profiles and their metadata. For example, the accession number of the profile is "9700235", which is much larger than other metadata values (e.g., latitude – 2.5°N). Without normalization, the accession number information will dominate the results.

2. Strategy 2 (Entropy Weight Method): here we calculated the weighted average by using the entropy weight method (Zeleny, 1998) to determine the weight. This method provides an objective perspective in calculating weights by leveraging entropy values to gauge the discreteness level of variables. A lower entropy value indicates heightened discreteness in variable measurements, resulting in a proportionately larger weight. The weight calculation process is shown in Appendix 1. Similarly, a normalization process for all numerical metadata is done before deploying the Entropy Weight Method. The relevant part of Strategy 2 is the "math_util_functions.entropy_weight" function.
3. Strategy 3 (Principal Component Analysis; PCA): PCA serves as a technique to decrease the dimensionality of data while retaining those that contribute most to variance (Jolliffe, 2002). The PCA initially calculates the eigenvalues and eigenvectors of the data matrix (Z described in Appendix 1) containing the metadata information (variables shown in Table 1) for all profiles. Then, we rank the variables based on the eigenvalues. Variables with larger eigenvalues are considered more significant because they account for a greater amount of variance. We then select the first variables that can explain 95% of the total variance (95% is an empirical choice), which are defined as "key variables". These "key variables" are then used in the entropy weight method to compute PSS. Applying PCA to each profile helps to identify metadata information that more effectively captures variations between profiles. The relevant part of Strategy 3 is the call the "math_util_functions.PCA_PSS_profiles" function.

With the Profile Summary Score value calculated by the above three strategies separately, we then utilize the neighborhood ordering method (Elmagarmid et al., 2007) to identify "possible duplicate pairs" as follows: firstly, sorting the Profile Summary Score of all profiles in ascending order, and then comparing each score with the following scores in turn. This action enhances screening efficiency. When the difference between the two scores is less than 0.0001% (an empirical chosen value), we consider these two corresponding profiles as "potential duplicates". Compared with the algorithm that analyzes profiles one by one, this method changes the time complexity from $O(n^2)$ to $O(n)$. The "potential duplicate pairs" obtained by the three strategies separately are then merged together, as a single "possible duplicate pairs" list (Figure 1).

With the "potential duplicates" list, Programs M00 are targeted screening by performing manual one-by-one duplicates checks or automatic checks on the list of "potential duplicates" created during the crude screening (i.e., N01). These checks can help us determine whether the "potential duplicates" identified in the crude screening are "exact duplicates" or fall into other categories of duplicates. This code checks for duplicates, triplicates, quadruplicates, etc. According to the seven criteria proposed in the Section 2.2, seven corresponding checks have been implemented here (adapted from Gronell and Wijffels (2008) and expertise within IQuOD). The checks can be categorized into two groups: "spatial-temporal checks" and "data record checks", as illustrated in Figure 2. During spatial-temporal checks, we assess whether the identified data is measured simultaneously but at different locations or if it is measured simultaneously and is co-located. The data record checks include correlation, truncation, layer-by-layer, exact duplicates, and interpolation (missing data) checks. The correlation check determines whether the correlation coefficients of temperature (or salinity) and depth of the profiles are consistent. If they are identical, it indicates that data is contaminated by numerical scaling or translation.

The output of this step is the classified list of possible duplicates and non-duplicates. This list is then used as input for the expert manual check (see section 2.3.2).

2.3.2 Manual checks

Manual (expert) checks are aimed at assessing the possible duplicates identified by the automatic algorithm. This step could also be used to analyze the reasons for duplication. Special attention is required for date, time, and location, as they are essential information, which is very important to identify errors. If any of these three variables is missing or wrong, the measurements of the profiles are of no use. Here, based on some potential reasons for duplication and metadata/data issues, the International Quality-controlled Ocean Database (IQuOD) task team members (Simoncelli et al., 2024) recommend additional criteria for all the identified pairs output from the automatic checking. For instance, for duplicate profiles in the Mediterranean Sea, the regional experts among the IQuOD members will review the original database for potential errors. If there is a high confidence that certain cruises or vessels have submitted duplicate data, the member associated with those cruises will check the original details about the vessel and its instruments. This includes gaining insights into the operational status of instruments during data collection by operators.

This manual checking process determines if they are true duplicates or if there are errors in the data and/or metadata. The criteria recommended by IQuOD include:

1. Repeated submissions to the data center, either submitted at different times or through repeated submissions by different organizations (i.e. data providers participating to the same cruise or project), lead to data repetition (Lawrimore et al., 2011; Simoncelli et al., 2022).
2. Some organizations in some countries may have submitted data with incorrect country codes. For example, data from Japanese fishery sectors may have been assigned non-Japanese country codes.
3. Submitting raw data first and the post-processed data later.
4. Data adjustments for confidentiality purposes by the military, such as modifications of latitude, longitude, year, or time. The adjusted data is then re-submitted by different organizations [personal communications from the Institute of Oceanology, Chinese Academy of Sciences (IOCAS)].
5. Constraints related to the submission timeframe for the Global Telecommunications System (GTS). Due to limited

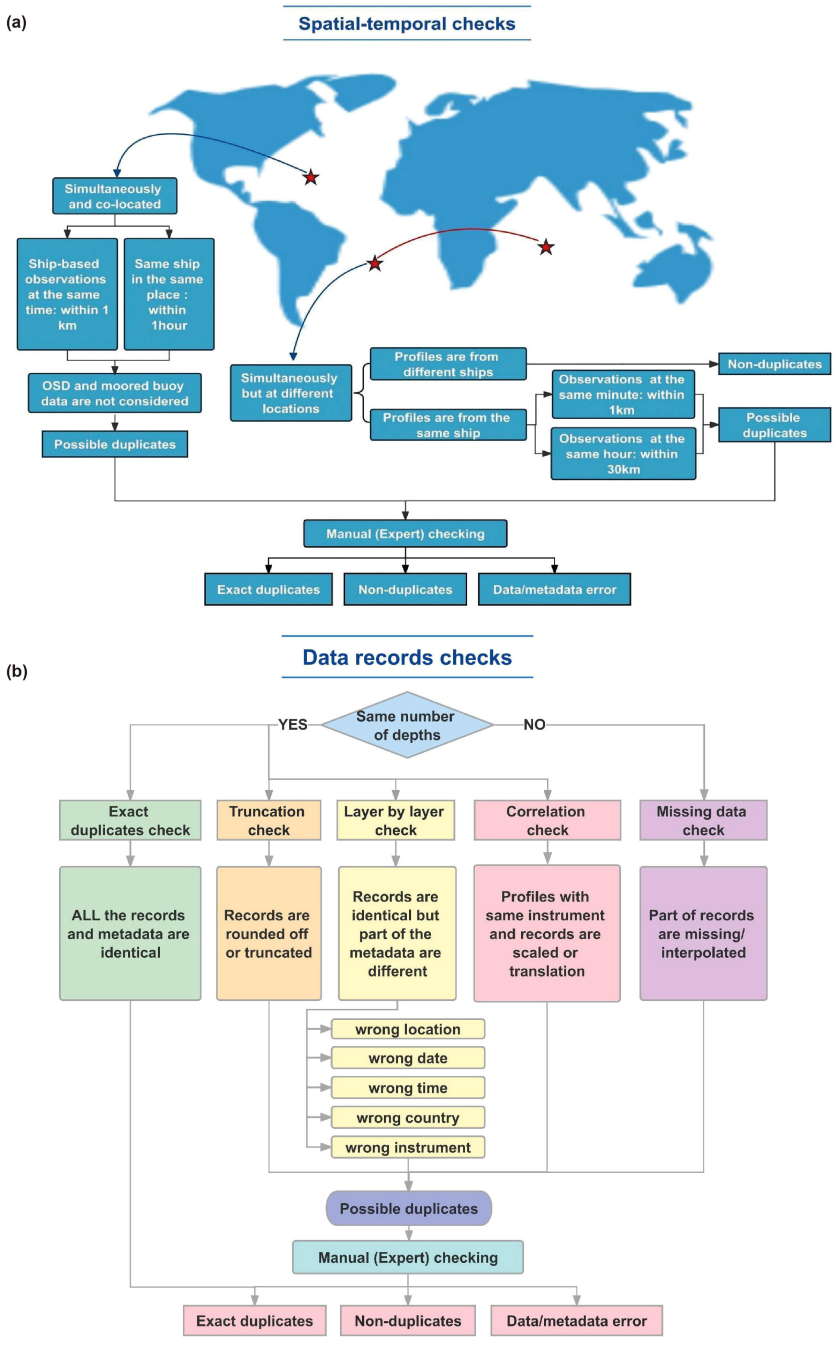

FIGURE 2
Flow chart of targeted screening. **(A)** Spatial-temporal checks. **(B)** Data records checks.

transmission capability, some historical data might have been heavily subsampled. For instance, real-time data from SOOP (Ship of Opportunity) XBT released within 12 hours of collection, albeit at the expense of substantial vertical sub-sampling of the profiles (Manzella et al., 2003).

6. Data delivered to the GTS (Global Transmission System) resulted in missing data or even wrong metadata information in these real-time profiles. From the 1990s, GTS data was reduced in size by creating inflection-point ASCII formats. More recently, full-resolution data has been delivered to the GTS in BUFR (Binary Universal Form for the Representation of meteorological data) format which includes more metadata and data, however, mismatches in GTS data and delayed-mode delivered data still occur.

Here, we noted that some parts of the above criteria also served as a supplement or references to the physical reason for the manual duplicate checking.

Beyond confirming the exactness of possible duplicates and analyzing their causes, the results of the manual (expert) checks are

also invaluable for further refining and optimizing the automatic algorithm. For instance, analysis of the duplicates in 1992-1993 XBT, as manually identified by experts, revealed a novel type of duplicates arising from interpolation or missing data (see Figure 3F). By incorporating the distinctive characteristics of these duplicates, the capability to detect similar instances of duplicates is enhanced.

3 Results

3.1 Validation

World Ocean Database (WOD) (Boyer et al., 2018) is an international effort to collect and archive *in-situ* oceanographic data in a unified format. The WOD converts data, received at the National

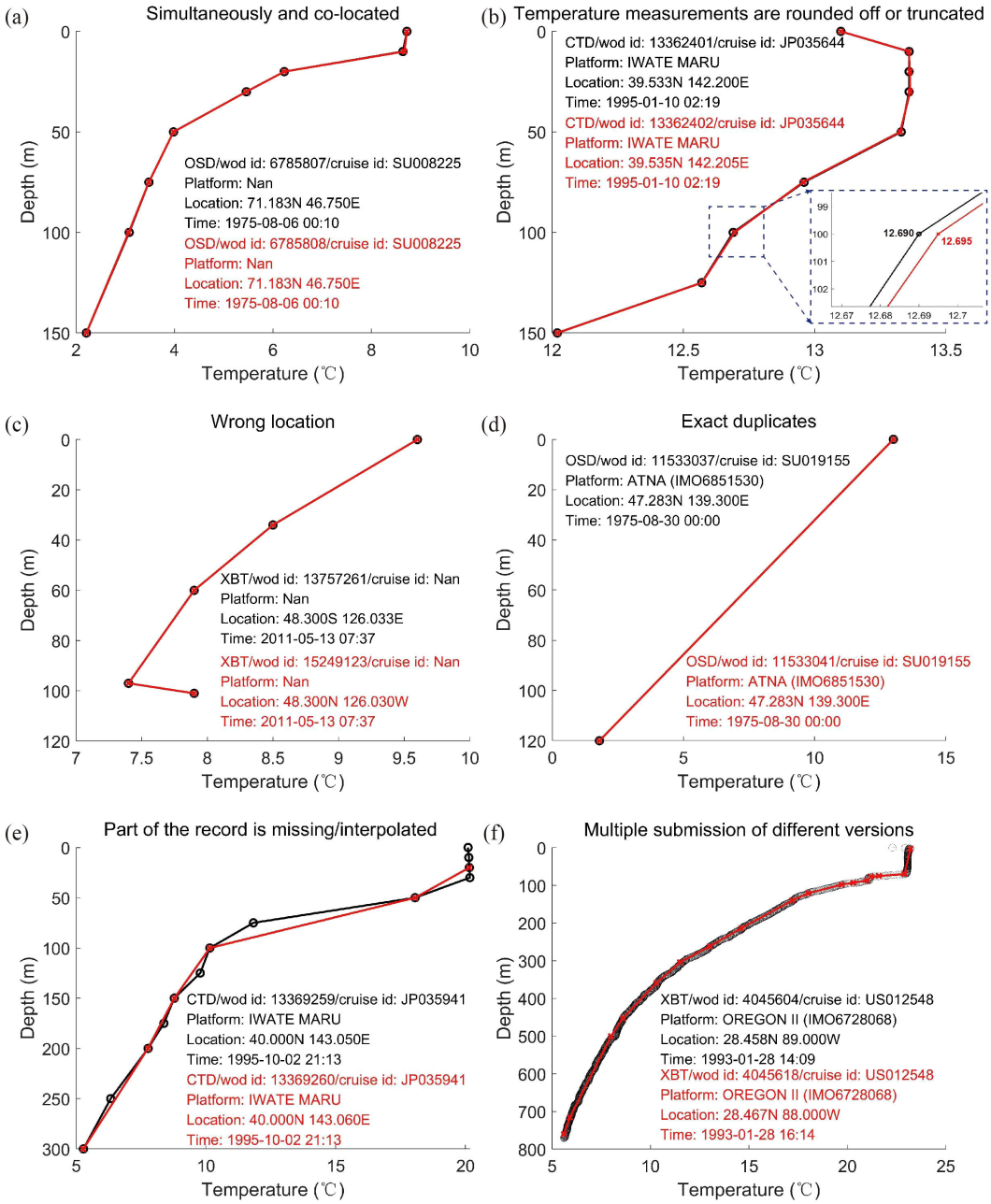

FIGURE 3
Example of some duplicated pairs. **(A)** Simultaneously and co-located **(B)** Observations are rounded off or truncated **(C)** Wrong location **(D)** Exact duplicates **(E)** Part of the profile data is missing/interpolated **(F)** Multiple submissions of different versions. Metadata information inserted into the subfigures are instrument types, WOD unique identifier, cruise identifier, platform name, latitude, longitude, year, month, day, hour, and minute, respectively.

Centers for Environmental Information (NCEI) from different sources, into the internal format and quality controlled for inclusion in the database. In the process of integrating data from different sources, platforms and instruments, there might be a large amount of duplicate data. In the WOD, these duplicate data has not been fully resolved. To validate the proposed duplicate checking system, we used some *in-situ* temperature and salinity profiles from the WOD18 (downloaded in February 2022) for the years 1975, 1995, and 2011. We choose these three years of data to represent different periods of the global ocean observing system. We also incorporated 1992-1993 XBT (eXpendable BathyThermograph) from the Gulf of Mexico with duplicated data (the total amount is 28 duplicated pairs), which has been under the rigorous expert-validated manual check, to validate the proposed system's robustness. In these data, the duplicates are known because they have already been screened by experts.

The results of duplicate checking showed that there were 95, 28, 494 and 831 duplicated groups for 1975, 1992-1993, 1995 and 2011, respectively (Table 2). Here, a duplicated group refers to either a double pair (i.e., two profiles), or triplicates, quadruplicates, etc. In the 1975 subset, exact duplicates (Figure 3D) accounted for approximately 2.8% of all duplicated groups, while possible duplicates comprised 97.2% (Table 3). Among the duplicated groups, a significant proportion fell into categories such as simultaneously and co-located profiles with identical measured values but different locations or dates. For example, Figure 3A displays two profiles measured simultaneously at the same place. In comparison, Figure 3C shows two profiles with identical metadata except for a mismatch in the sign and rounding precision in their recorded latitude/longitude metadata.

In the 1995 subset, the exact duplicates comprised 8.0%, while the remaining 92.0% were possible duplicates (Table 2). In the possible duplicated group, duplications due to rounding off or truncation dominate (38.4%), while profiles with interpolation also account for a large proportion (32.7%). Figure 3B shows an example of possible duplicates resulting from rounding or truncating temperature measurements. Additionally, Figure 3E represents possible duplicates caused by missing data or interpolation.

In the 2011 subset, nearly all duplicates were classified as possible duplicates (99.8%), with simultaneously and co-located duplicates being the predominant types of duplicates. For the Gulf of Mexico XBT data in 1992-1993, all duplicates are detected as possible duplicates, which serves as an indication of the robustness of the Profile Summary Score algorithm with the expert-validated data. In Figure 3F, we observe a possible duplicate resulting from multiple submissions of different versions. In those cases, one of this pair has more temperature records than the other one.

In summary, our algorithms identified 1,448 duplicated groups in total within a dataset comprising over a total of 600,000 profiles investigated in this study.

TABLE 2 Duplicate & metadata checking results of 1975, 1995, and 2011 data and 1992-1993 XBT data in the Gulf of Mexico obtained from WOD18.

	1975	1992-1993 (Gulf of Mexico)	1995	2011
Total number of downloaded profiles	142537	–	142473	337651
Duplicated groups (results of the targeted screen)	95	28	494	831
Duplicate rate (%)	0.067	–	0.347	0.246

TABLE 3 Proportion of exact duplicates and possible duplicates per year.

		1975	1992-1993 (Gulf of Mexico)	1995	2011
Duplicated groups (results of target screen)	Exact duplicates rate (%)	2.8	0.0	8.0	0.2
	Possible duplicates rate (%)	97.2	100.0	92.0	99.8

3.2 Benchmark dataset

A benchmark dataset is constructed by combining the results of checks for the 1975, 1995, and 2011 WOD data and the data from the Gulf of Mexico (Boyer et al., 2018). This dataset is available to the community as a benchmark for new methods that are going to resolve duplicates, such as training of machine learning models. Two additional variables have been incorporated into the WOD standard netCDF file format, while the metadata and observations are all preserved as in WOD (see Table 4).

The benchmark dataset (Song et al., 2023) contains 1,448 groups of duplicate data (each group comprising two or more profiles). 2,956 profiles were labeled, with 542 being labeled as 1 (constituting 18.336%), 1,222 as 2 (41.373%), and 1,192 as 0 (40.291%). The file *duplicate_list_pair.xlsx* contains information about the *Duplicate_flag* and *Duplicate_pair_id* related to the labeled profile data.

4 Conclusion and perspective

This study provides some criteria for defining the duplicates and an open-access tool for identifying duplicate data. Currently, our code can only detect duplicates for data with WOD format. If users possess data in alternative formats, a format conversion is needed. We have provided instructions in the user manual for data formats requirements. The DC_OCEAN code base will become more versatile and applicable to other input data formats in the future, thanks to the continuous IQuOD community effort and the users feedback and collaboration through the GitHub development environment. In fact, although our program is currently targeted at temperature and salinity data, with some modifications it can be applied to other variables (e.g. dissolved oxygen, partial pressure of carbon dioxide, etc.). Three years of data from WOD18 were chosen to test and validate the newly developed algorithm with outcomes published in the benchmark dataset, which will support future activities. Possible reasons for duplication require further analysis

TABLE 4 Details of the two additional variables in the benchmark dataset. Note the corresponding metadata and measurements are the same as WOD18.

Variable 1: Duplicate_flag			Variable 2: Duplicate_pair_id	
Duplicate_flag=0	Duplicate_flag=1	Duplicate_flag=2	Purpose	Pair ID
Duplicate profiles, have been checked by experts. Need to be kept in the dataset by the user.	Duplicate profiles, have been checked by experts, but uncertain which one to keep.	Duplicate profiles, have been checked by experts. Need removal from the dataset by the user.	Used to indicate the duplicate groups (which netCDF files are duplicated).	Order from 1 to 10000.

by the IQuOD experts. The proposed approach is to flag the identified duplicate data without removing it from the database, according with the latest data management best practices, which allows the user to trace back the detecting process with full transparency. Users can decide whether to remove these duplicates based on their research requirements. We highlight that the new duplicate checking system (DC_OCEAN, Tan et al., 2024) could enhance the uniqueness of *in-situ* ocean profiles and facilitate further marine and climate science data applications. The obtained benchmark dataset can serve as a valuable reference for testing and improving other duplicate checking systems.

The duplicate checking system can potentially be useful in various scientific applications, such as reconstructing historical ocean temperature gridded fields. An example is, in the Northwest Pacific region, after removing duplicate data in 1995 from the abovementioned benchmark dataset, the difference in the gridded averaged temperature can be as large as 0.1°C for 0-100 m and 0.06°C for 0-300 m. After applying a gap-filling method by Cheng et al. (2017) to reconstruct gridded fields with full ocean coverage, the impact of the duplicate check can spread to a large area because the reconstruction used nearby observations. The maximum temperature difference in 1995 can reach 0.06°C (upper 100 m) and 0.05°C (upper 300 m). The major difference is mainly distributed along the east coast of Japan, extending the difference to the Kuroshio extension due to the spatial interpolation (mapping), which is characterized by abundant eddy activities (Figure 4). This test, although simple, indicates that duplicates may have a potential non-negligible impact on regional ocean temperature (and ocean heat content) estimates. Therefore, our duplicate checking software can support the improvement of the data analysis, which will be fully investigated in the future.

The system can also support international activities such as IODE-project WOD, where there should be no duplicate data by design. However, there were duplicates in WOD (as shown in this paper), therefore the Duplicate checking system for ocean profile (named DC_OCEAN) can fill the gap between the complete and the incomplete duplicate checking in the WOD. The investigations (as done in this study) are also valuable for improving the data quality

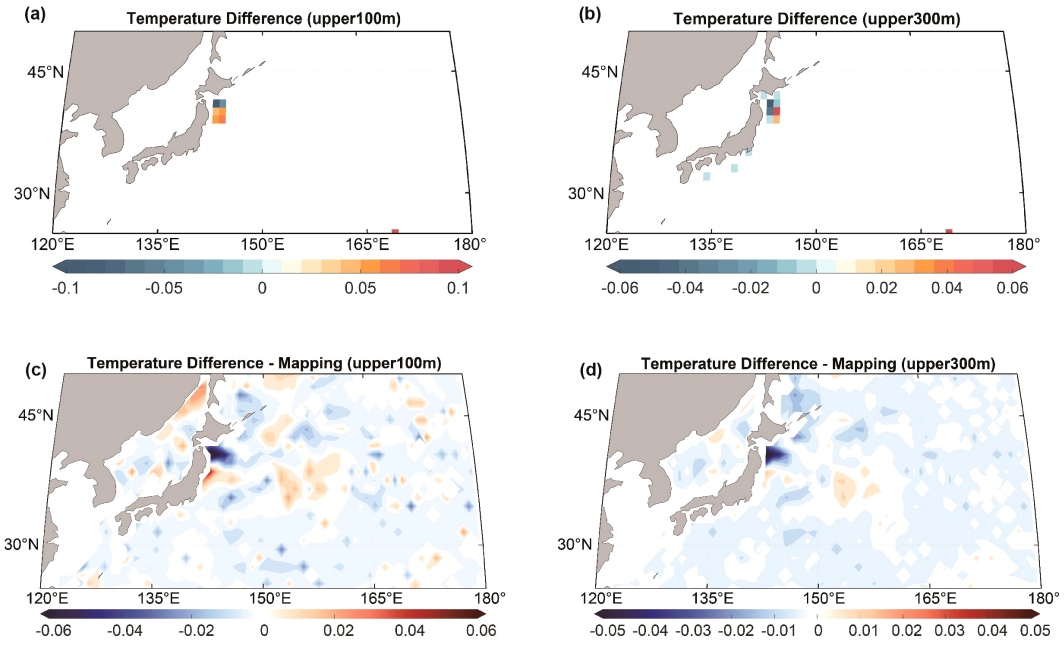

FIGURE 4
The estimated differences in gridded averaged temperature and reconstructed temperature after applying a gap-filling approach in Cheng et al. (2017), before and after excluding duplicate data in 1995. **(A, B)** The gridded averaged temperature at 100 and 300 m, respectively. **(C, D)** The reconstructed temperature field at 100 m and 300 m after gap-filling approach, respectively. The unit is °C.

and data management practices, which is a task of another IODE-endorsed Project: the International Quality-controlled Ocean Database. Furthermore, for some uses, some closely co-located data (e.g., up + down casts of the same CTD station) identified by the DC_OCEAN might not be considered a duplicate for WOD, but might still skew the statistics (e.g., for gridded averages). So, the software can identify and flag these cases for specific users who would like to flag these co-located data in time and space.

The system developed in this study is part of the IQuOD project for identifying duplicates, correcting metadata errors, and eventually improving ocean data quality (Cowley et al., 2023). Because fully implementing the duplicate checking for all historical data requires substantial time and effort, our open-source duplicate checking software is used as a tool for further activities. In addition, the system could also support various efforts in merging or in integrating data from different data sources or data infrastructures, for instance, WOD, the Blue Cloud 2026 European Project (Schaap et al., 2022), the Chinese Academy of Sciences Ocean Data Center (CODC) database, EN4, etc (Zhang et al., 2024; Boyer et al., 2018; Good et al., 2013). For this purpose, for example, the Blue Cloud project has set up a synergy to cooperate with IQuOD by using several data quality improvement tools (including the duplicate checking algorithm purposed in this study) to provide access to multi-disciplinary datasets from observations, and finally generate qualified data collections by merging data from various databases.

5 Future challenges

During the investigation, we found that many profiles lack parts of metadata (such as country, time, location, instrument type, and platform), posing challenges for duplicate checking because an exhaustive metadata description is a requirement to identify the best data version. Therefore, to facilitate the follow-on examination of metadata, we have assigned the value of 99999 to the *Duplicate_pair_id* variable for these data and included them in our benchmark dataset. Tackling metadata errors has always been a central task of ocean data centers, including IQuOD.

Additionally, identifying the duplicates within different data infrastructures is more complicated due to the adoption of different metadata format standards and vocabularies. Currently, the duplicate checking system can only support detecting duplicates within a unified metadata format framework (Here, we are based on WOD format). Therefore, the unification and standardization of metadata are very important for duplicate checking, and the DC_OCEAN will adapt to more kinds of metadata formats in the future (e.g., GTSPP or WMO format). Unifying metadata format has also always been a central task of ocean data centers including IQuOD.

Currently, our code can identify eight types of duplicates (see Section 2.2). However, there may be other potential/new types of duplicates that we may have missed. Therefore, considering the feedback from manual (expert) checks and including the characteristics of newly discovered duplicates will be the future direction of improvement for DC_OCEAN.

Moreover, many of our threshold settings are practical and based on a limited amount of data. Despite achieving satisfactory results when applied DC_OCEAN to data from WOD at 1975, 1995, and 2011, further verification of its performance in larger database is needed.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s. The duplicate checking algorithm (DC_OCEAN; https://github.com/IQuOD/duplicated_checking_IQuOD) is available as an open-source Python package under the Apache-2.0 license (https://doi.org/10.5281/zenodo.12662531; Tan et al., 2024; https://pypi.org/project/DC-OCEAN/). A full.cdl file of the algorithm is https://github.com/IQuOD/duplicated_checking_IQuOD/blob/main/ocean_data_netCDF_format.cdl. The DOI of the benchmark dataset presented in this study is http://dx.doi.org/10.12157/IOCAS.20230821.001.

Author contributions

XS: Writing – original draft, Writing – review & editing, Methodology, Software, Visualization. ZT: Conceptualization, Writing – original draft, Writing – review & editing, Data curation, Methodology, Software. RL: Conceptualization, Writing – review & editing, Data curation, Methodology. SS: Software, Writing – review & editing. RC: Writing – review & editing, Software, Project administration. SK: Writing – review & editing. TB: Writing – review & editing. FR: Writing – review & editing, Data curation. GC: Writing – review & editing, Software, Project administration. VG: Writing – review & editing. LC: Conceptualization, Writing – review & editing, Project administration, Methodology.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This study is supported by the National Natural Science Foundation of China (Grant no. 42122046, 42076202), the Scientific Committee on Oceanic Research (SCOR) Working Group 148, funded by national SCOR committees and a grant to SCOR from the U.S. National Science Foundation (Grant OCE-1546580). We thank the International Oceanographic Data and Information Exchange (IODE) program of the Intergovernmental Oceanographic Commission (IOC) for their financial support.

Acknowledgments

We thank all the IQuOD members who manually checked for possible duplicates. IQuOD is a Programme Activity of the

International Oceanographic Data and Information Exchange (IODE) (https://www.iode.org) of IOC/UNESCO. We would also thank Huifeng Yuan of Computer Network Information Center, Chinese Academy of Sciences for his technical expertise and innovative insights in refining and reviewing the code and software. We also thank Huayi Zheng from the Institute of Atmospheric Physics (IAP/CAS) for his efforts on the independent review of the software. We would also like to thank Edward King and Ann Thresher from CSIRO for their efforts/contributions regarding the duplicate checking code for our references. We would also like to express our gratitude to the editor and reviewers of this manuscript for their valuable insights and patience during the submission process.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Abraham, J., Baringer, M., Bindoff, N. L., Boyer, T., Cheng, L. J., Church, J. A., et al. (2013). A review of global ocean temperature observations: Implications for ocean heat content estimates and climate change. *Rev. Geophys.* 51, 450–483. doi: 10.1002/rog.20022

Balmaseda, M. A., Hernandez, F., Storto, A., Palmer, M., Alves, O., Shi, L., et al. (2015). The ocean reanalyses intercomparison project (ORA-IP). *J. Operational. Oceanogr.* 8, s80–s97. doi: 10.1080/1755876X.2015.1022329

Boyer, T. P., Baranova, O. K., Coleman, C., Garcia, H. E., Grodsky, A., Locarnini, R. A., et al. (2018). *World ocean Database 2018*. Eds. A. V. Mishonov and Technical, (NOAA Atlas NESDIS), 87.

Boyer, T. P., and Levitus, S. (1994). *Quality control and processing of historical oceanographic temperature, salinity, and oxygen data* (US Department of Commerce, National Oceanic and Atmospheric Administration).

Cabanes, C., Angel-Benavides, I., Buck, J., Coatanoan, C., Dobler, D., Herbert, G., et al. (2021). *DMQC cookbook for core Argo parameters* (France: IFREMER Brest).

Carton, J. A., and Giese, B. S. (2008). A reanalysis of ocean climate using Simple Ocean Data Assimilation (SODA). *Monthly. Weather. Rev.* 136, 2999–3017. doi: 10.1175/2007MWR1978.1

Cheng, L., Pan, Y., Tan, Z., Zheng, H., Zhu, Y., Wei, W., et al. (2024). IAPv4 ocean temperature and ocean heat content gridded dataset. *Earth Syst. Sci. Data Discussions.* 2024, 1–56. doi: 10.5194/essd-16-3517-2024

Cheng, L., Trenberth, K. E., Fasullo, J. T., Boyer, T., Abraham, J., and Zhu, J. (2017). Improved estimates of ocean heat content from 1960 to 2015. *Sci. Adv.* 3, e1601545. doi: 10.1126/sciadv.1601545

Cowley, R., Macdonald, A., Good, S., Killick, R., Cheng, L., Tan, Z., et al. (2023). *IQuOD 7th Annual Workshop Report, 10-11 July 2023 Potsdam Institute for Climate Impact Research, Potsdam, Germany 2023 International Quality-Controlled Ocean Database (IQuOD) – 7th IQuOD Annual Workshop 8th IODE SG-IQuOD 4th SCOR WG 148 10-11 July 2023 Potsdam Institute for Climate Impact Research, Potsdam, Germany* (AquaDocs: International Quality-Controlled Ocean Database (IQuOD).

Durack, P. J., and Wijffels, S. E. (2010). Fifty-year trends in global ocean salinities and their relationship to broad-scale warming. *J. Climate* 23, 4342–4362. doi: 10.1175/2010JCLI3377.1

Elmagarmid, A. K., Ipeirotis, P. G., and Verykios, V. S. (2007). Duplicate record detection: A survey. *IEEE Trans. Knowledge. Data Eng.* 19, 1–16. doi: 10.1109/TKDE.2007.250581

Escudier, R., Clementi, E., Cipollone, A., Pistoia, J., Drudi, M., Grandi, A., et al. (2021). A high resolution reanalysis for the mediterranean sea. *Front. Earth Sci.* 9, 9. doi: 10.3389/feart.2021.702285

Garcia, H. E., Boyer, T. P., Locarnini, R. A., Baranova, O. K., and Zweng, M. M. (2018). *World Ocean Database 2018: User's Manual*. Eds. A. V. Mishonov and Technical, (Silver Spring, MD: NOAA).

Goni, G., Sprintall, J., Bringas, F., Cheng, L., Cirano, M., Dong, S., et al. (2019). More than 50 years of successful continuous temperature section measurements by the global expendable bathythermograph network, its integrability, societal benefits, and future. *Front. Mar. Sci.* 6, 452. doi: 10.3389/fmars.2019.00452

Good, S. A., Martin, M. J., and Rayner, N. A. (2013). EN4: Quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates. *J. Geophys. Res.: Oceans.* 118, 6704–6716. doi: 10.1002/2013JC009067

Good, S., Mills, B., Boyer, T., Bringas, F., Castelão, G., Cowley, R., et al. (2023). Benchmarking of automatic quality control checks for ocean temperature profiles and recommendations for optimal sets. *Front. Mar. Sci.* 9. doi: 10.3389/fmars.2022.1075510

Gronell, A., and Wijffels, S. E. (2008). A semiautomated approach for quality controlling large historical ocean temperature archives. *J. Atmospheric. Oceanic. Technol.* 25, 990–1003. doi: 10.1175/JTECHO539.1

IPCC (2021). *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge University Press).

Ishii, M., Fukuda, Y., Hirahara, S., Yasui, S., Suzuki, T., and Sato, K. (2017). Accuracy of global upper ocean heat content estimation expected from present observational data sets. *Sola* 13, 163–167. doi: 10.2151/sola.2017-030

Ji, F., Dong, M., Liu, Y., Xu, S., Wan, F., Shi, X., et al. (2022). "A study on the method of eliminating duplication of ocean temperature and salinity data," in *AIIPCC 2022; The third international conference on artificial intelligence, information processing and cloud computing*. IEEE. 1–7.

Jolliffe, I. T. (2002). *Principal component analysis for special types of data* (New York: Springer), 338–372.

Lawrimore, J. H., Menne, M. J., Gleason, B. E., Williams, C. N., Wuertz, D. B., Vose, R. S., et al. (2011). An overview of the Global Historical Climatology Network monthly mean temperature data set, version 3. *J. Geophys. Res.* 116. doi: 10.1029/2011JD016187

Levitus, S. (1982). *Climatological atlas of the world ocean* (US Department of Commerce, National Oceanic and Atmospheric Administration).

Locarnini, R. A., Mishonov, A. V., Baranova, O. K., Boyer, T. P., Zweng, M. M., Garcia, H. E., et al. (2019). *World Ocean Atlas 2018, Volume 1: Temperature*. Eds. A. Mishonov and Technical, (NOAA Atlas NESDIS 81), 52pp.

Mackenzie, B., Celliers, L., Assad, L. P. D. F., Heymans, J. J., Rome, N., Thomas, J., et al. (2019). The role of stakeholders in creating societal value from coastal and ocean observations. *Front. Mar. Sci.* 6, 137. doi: 10.3389/fmars.2019.00137

Manzella, G. M. R., Scoccimarro, E., Pinardi, N., and Tonani, M. (2003). Improved near real-time data management procedures for the Mediterranean ocean Forecasting System-Voluntary Observing Ship program. *Ann. Geophys.* 21, 49–62. doi: 10.5194/angeo-21-49-2003

Schaap, D., Assante, M., Pagano, P., and Candela, L. (2022). *Blue-Cloud: Exploring and demonstrating the potential of Open Science for ocean sustainability 2022 IEEE International Workshop on Metrology for the Sea; Learning to Measure Sea Health Parameters (MetroSea)* (IEEE), 198–202.

Schmidtko, S., Stramma, L., and Visbeck, M. (2017). Decline in global oceanic oxygen content during the past five decades. *Nature* 542, 335–339. doi: 10.1038/nature21399

Simoncelli, S., Coatanoan, C., Myroshnychenko, V., Bäck, Ö., Sagen, H., Scory, S., et al. (2021). "SeaDataCloud data products for the european marginal seas and the global ocean," in *9th EuroGOOS International conference* (Brest, France).

Simoncelli, S., Cowley, R., Tan, Z., Killick, R., Castelão, G., Cheng, L., et al. (2024). *The International Quality-controlled Ocean Database (IQuOD)* Vol. 80 (Miscellanea INGV), 139–140. doi: 10.13127/MISC/80/50

Simoncelli, S., Manzella, G. M. R., Storto, A., Pisano, A., Lipizer, M., Barth, A., et al. (2022). "A collaborative framework among data producers, managers, and users," in *Ocean Science Data*. Eds. G. Manzella and A. Novellino (Elsevier), 197–280.

Song, X., Tan, Z., Locarnini, R., Simoncelli, S., Cowley, R., Kizu, S., et al. (2023). A benchmark dataset for ocean profiles duplicate checking. *Marine Science Data Center of the Chinese Academy of Sciences*. doi: 10.12157/IOCAS.20230821.001

Szekely, T., Gourrion, J., Pouliquen, S., and Reverdin, G. (2024). *CORA, Coriolis Ocean Dataset for Reanalysis* (SEANOE).

Tan, Z., Cheng, L., Gouretski, V., Zhang, B., Wang, Y., Li, F., et al. (2023). A new automatic quality control system for ocean profile observations and impact on ocean

warming estimate. *Deep. Sea. Res. Part I.: Oceanogr. Res. Papers.* 194, 103961. doi: 10.1016/j.dsr.2022.103961

Tan, Z., Song, X., Yuan, H., Cowley, R., Cheng, L., and Castelao, G. (2024). *IQuOD/duplicated_checking_IQuOD: DC_OCEAN: v1.3.3 (v1.3.3).* (Zenodo). doi: 10.5281/zenodo.13819929

Zeleny, M. (1998). Multiple criteria decision making: Eight concepts of optimality. *Hum. Syst. Manage.* 17, 97–107. doi: 10.3233/HSM-1998-17203

Zhang, B., Cheng, L., Tan, Z., Gouretski, V., Li, F., Pan, Y., et al. (2024). CODC-v1: a quality-controlled and bias-corrected ocean temperature profile database from 1940–2023. *Sci. Data* 11, 666. doi: 10.1038/s41597-024-03494-8

10.3389/fmars.2024.1403175

Appendix 1: The entropy weight method

The entropy weight calculation process is introduced here. The input metadata can be regarded as a data matrix Z,

$$Z = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1m} \\ z_{21} & z_{22} & \cdots & z_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{12} & \cdots & z_{nm} \end{bmatrix},$$

where m is the number of variables (they are used to calculate the Profile Summary Score for a profile) and n is the total amount of input profiles.

The entropy value of each variable is defined as e_j,

$$e_j = -\frac{1}{\ln(n)} \sum_{i=1}^{n} p_{ij} \times \ln p_{ij}, (j = 1, \ldots\ldots, m)$$

,

where $p_{ij} = \frac{z_{ij}}{\sum_{i=1}^{n} z_{ij}}, (i = 1, \ldots\ldots, n, j = 1, \ldots\ldots, m)$.

And the weight of the j_{th} variable $weight_j (j = 1, \ldots\ldots, m)$ is:

$$weight_j = \frac{d_j}{\sum_{j=1}^{m} d_j}, (j = 1, \ldots\ldots, m)$$

,

where $d_j = 1 - e_j$.