SeaDataCloud Temperature and Salinity Data Collection for the Black Sea: analysis of data quality problems

Volodymyr Myroshnychenko[1], *Reiner Schlitzer*[2], *Michèle Fichaut*[3], *Dick Schaap*[4]

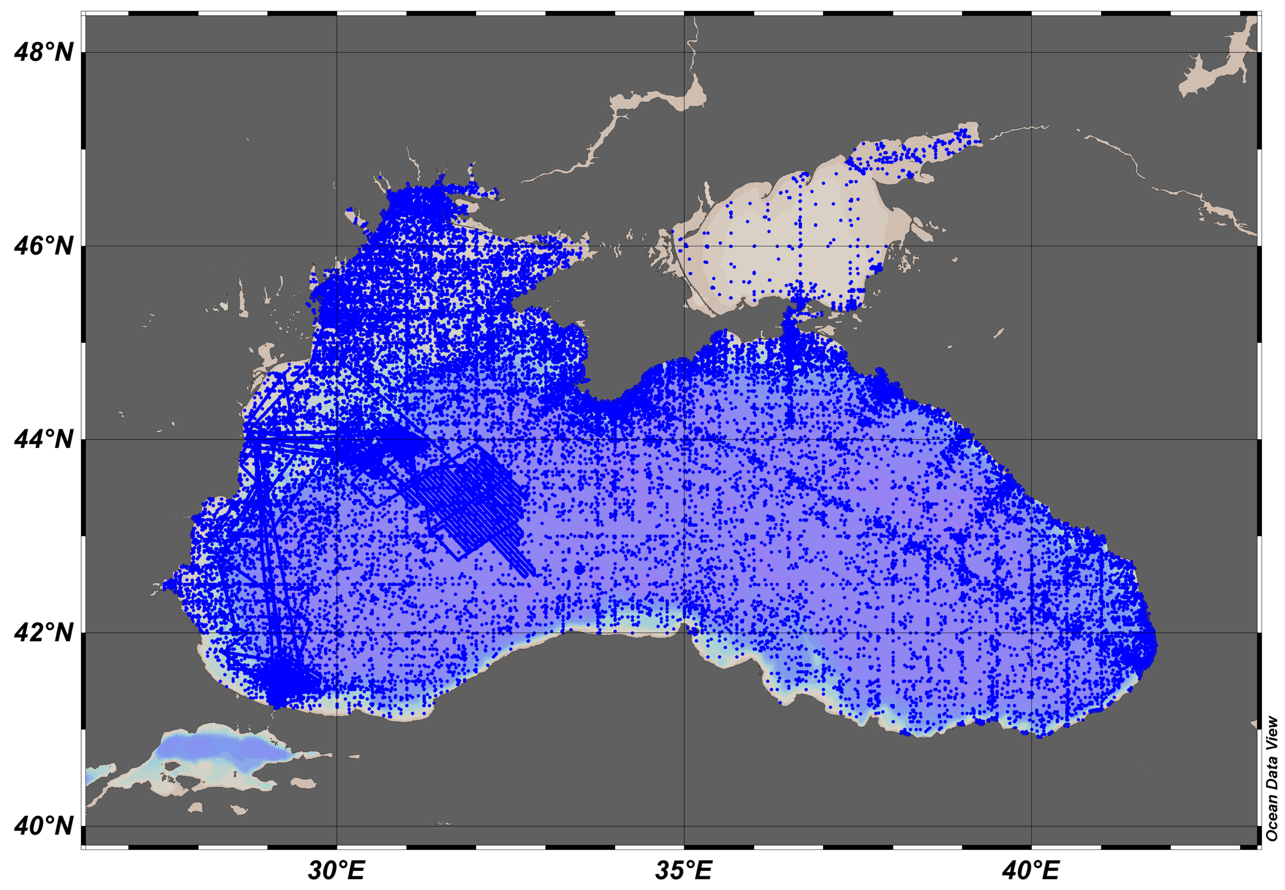
(1) METU, Turkey, volodymyr@ims.metu.edu.tr (2) AWI, Germany (3) IFREMER, France (4) MARIS, Netherlands

General Information about collection

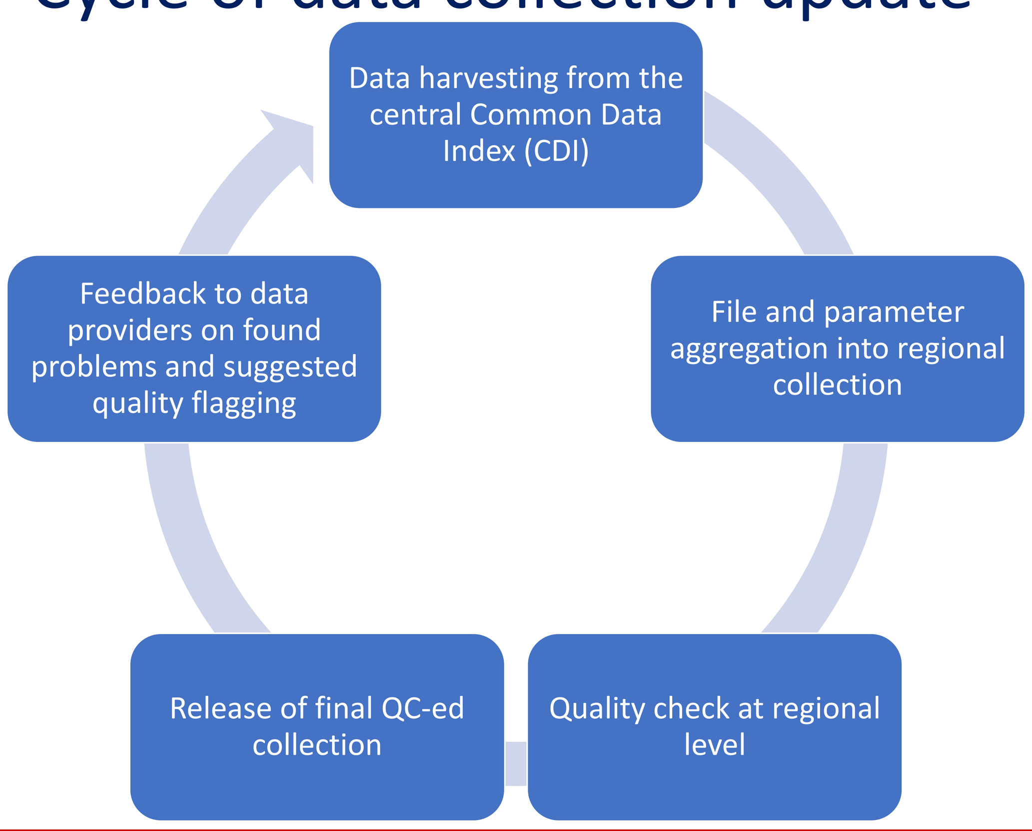
The new qualified **Temperature and Salinity Historical Data Collection for Black Sea SDC_BLS_DATA_TS_V1** was produced within framework of the SeaDataCloud (SDC) project in 2018.

Time period	Cruises	Stations			Values
		All	Profiles	Underway measurements	
1868-2017	2,286	137,723	119,160	18,563	4,240,346

This is the 3rd release of the regional data collection. Two previous were published within framework of the SeaDataNet II project. Each release represents a snapshot of the SeaDataNet (SDN) distributed database content at different time: Jan, 2014 (SDN V1.1), Mar, 2015 (SDN V2), and Nov, 2017 (SDC V1). The collection is available for download at https://www.seadatanet.org/Products/.



Cycle of data collection update



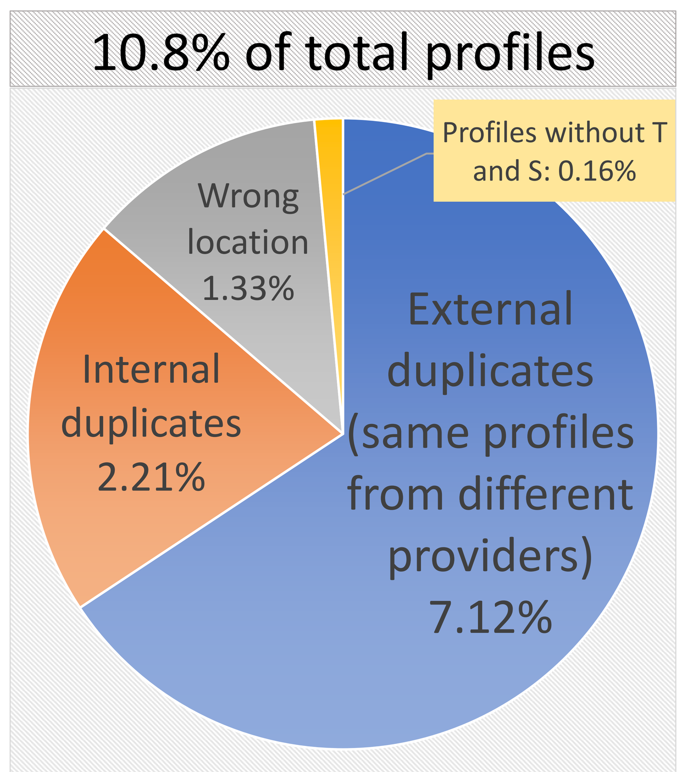
Quality Control and feedback to providers

The cycle of the data collection update consists of 5 steps. It is presumed that harvested Temperature and Salinity have undergone Quality Control (QC) at provider's side, however in reality the initial aggregated dataset contained a number of metadata problems as well as not QC-ed data or data anomalies that needed to be corrected. Therefore the quality check of metadata and data is one of the most important and labour-consuming parts of the cycle. The QC was been performed with the help of Ocean Data View software (ODV http://odv.awi.de/).

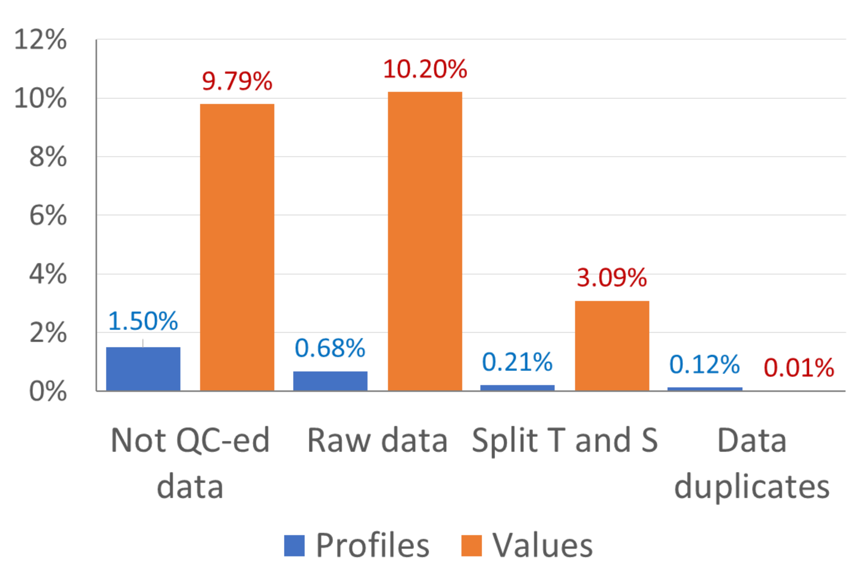
Since the SeaDataNet infrastructure is essentially a distributed database, the suggested corrections should be applied to original data at data sources. Therefore each cycle includes the step of feedback to data originators and distributors on found problems with metadata and on suggested quality flagging of data to be applied. This will ensure supplying the qualified data to future user's requests within the SeaDataNet infrastructure.

Metadata-related problems

- Duplicates. This is most significant problem in the SDC Black Sea Temperature and Salinity Data Collection. The duplicates can introduce bias into derived data products, e.g. in climatologies, therefore they should be eliminated from the collection. In total the 86% of found metadata problems are related to duplicates.
- Wrong location (on land). This error is more typical for old data that were acquired before the reliable navigation systems were introduced.
- Mismatch between CDI and real data with respect to parameters, i.e. when CDI record indicates presence of temperature or salinity while the respective dataset (profile) does not contain nor temperature nor salinity. This kind of error can mislead users who are searching for data via CDI interface.
- Other problems: missing time, missing date, wrong sea depth etc.



Data-related problems



Data-related problems were revealed in 2.5% profiles affecting, however, 22% (!) of total data, because these are data mainly from high-resolution CTD profilers.

- Non QC-ed data. This kind of problem is more common for newly acquired data.
- Raw data. Although the raw data are accepted, they are noisy, the profiles may contain density inversions.
- Profiles with split temperature and salinity. In such profiles one data row contain just temperature , next – just salinity and so on. The temperature and salinity for the same depth should be merged, otherwise it is not possible to calculate derived physical properties.
- Data duplicates within profiles.

Conclusions

The initial aggregated dataset contained metadata- and data-related problems that affected quality of more than 10% profiles and more than 20% data values. The problems were resolved in the final product – the SDC Temperature and Salinity Historical Data Collection for Black Sea: e.g., duplicates were eliminated, quality flags were assigned to non QC-ed data, split data were merged, etc. The same corrections now have to be applied at data sources of the SeaDataNet infrastructure. The recommendations on actions to be performed have been elaborated and send to data providers. The top priority actions:

- Elimination of duplicates. This action requires close cooperation of involved data providers under coordination of SDN managers and readiness of data providers to withdraw duplicates notwithstanding that it will decrease their scores in SDN.
- Mandatory QC of data and processing of raw data followed by resubmission of final CTD profiles to SDN.

Implementing these actions will bring the SeaDataNet infrastructure closer to its ultimate goal - providing the best qualified data on marine environment to end users.