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

Volodymyr Myroshnychenko, METU (Turkey), volodymyr@ims.metu.edu.tr
Reiner Schlitzer, AWI (Germany), reiner.schlitzer@awi.de
Michèle Fichaut, IFREMER (France), michele.fichaut@ifremer.fr
Dick Schaap, MARIS (Netherlands), dick@maris.nl

Created for the first time in 2014 within the SeaDataNet II (SDN) EU-project the regional temperature and salinity data collection for Black Sea then underwent 2 updates: one was done in course of the SDN II project while the latest version was released within the SeaDataCloud (SDC) project in 2018. The data collection is managed with ODV software (Schlitzer, R., Ocean Data View, odv.awi.de, 2017) which provides various possibilities for performing quality control (QC), including flagging of wrong and doubtful data, identification of duplicates and data anomalies, etc.

The periodical update intends to supply researches and other end users with the most complete and qualitative data products based on the up to date information retrieved from the expanding SDN infrastructure. Each cycle of the data collection update consists of the following steps:

1. Data harvesting from the central Common Data Index (CDI).
2. File and parameter aggregation into regional collection.
3. Quality check at regional level.
4. Release of final QC-ed collection.
5. Feedback to data providers on found problems and suggested quality flagging.

It is expected that each update cycle will contribute to improvement of the overall quality of data and information in the SDN infrastructure assuming that data originators and distributors will apply the recommended corrections. However the experience earned from the two cycles of collection update suggests that this process is going rather slowly and that some problems continue to persist through update cycles.

Hereby we provide analysis of the problems and factors that affect data quality of the latest SDC temperature and salinity data collection for Black Sea. They can be divided into 2 groups: metadata-related and data-related. The problems from the first group are more serious since they affect whole temperature-salinity profiles and, usually, can’t be corrected during QC of the aggregated dataset because require communication and actions from data providers that takes time. As a result, the affected profiles should be eliminated from the data collection.

The major metadata-related problems were identified as follow:

- Duplicates. This is most significant problem in the SDN Black Sea dataset. The duplicates can introduce bias into derived data products, e.g. in climatologies, therefore they should be eliminated from the collection.
- Wrong location (on land).
- Mismatch between CDIs and datasets 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 data via CDI interface.

The effect of metadata-related problems on the data quality of the SDC temperature and salinity data collection for Black Sea is presented at *Figure* **1**. More than 10% of profiles are affected, all of them were eliminated from the collection.

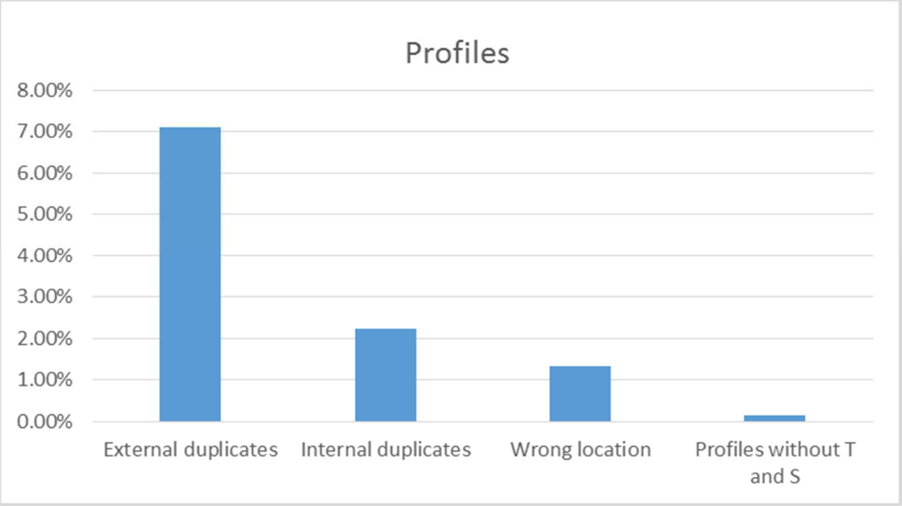

Figure 1 Percentage of profiles with metadata-related problems

The data-related problems include:

- Not QC-ed data.
- Raw data. Although the raw data are accepted, they are noisy, the profiles may contain density inversions.
- Profiles with split temperature and salinity, i.e. profiles that consist of pairs of data rows: one row contain temperature, other – salinity. Such pairs should be merged, otherwise it is not possible to calculate derived physical properties.
- Data duplicates within profiles.

In terms of profiles the effect of the data-related problems on data quality in the collection is not large – just ~2.5%, however in terms of data values it is significant - >22% - because the data are coming from the high-resolution CTD profiles (*Figure **2***).

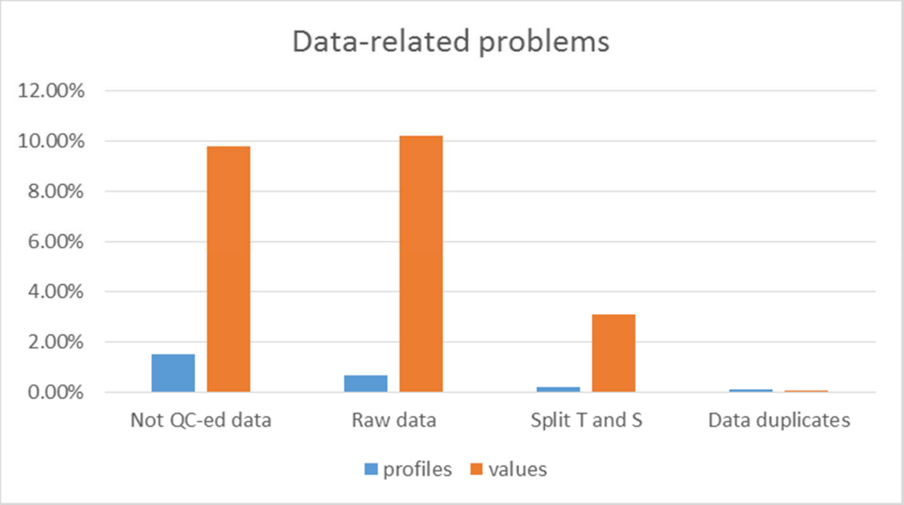

Figure 2 Effect of data-related problems

The performed analysis allowed to identify most serious problems that affect quality of more than 10% profiles and more than 20% data values in the SDC temperature and salinity data collection for Black Sea. The first problem to be resolved is elimination of duplicates, however it 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. The mandatory QC of data and processing of raw data followed by resubmission of final CTD profiles are the other tasks to be performed with high priority by corresponding data providers. The recommendations on actions to be performed have been elaborated and send to data providers.