SeaDataCloud

Improvements of the QC loop
Report on feedback to data centres

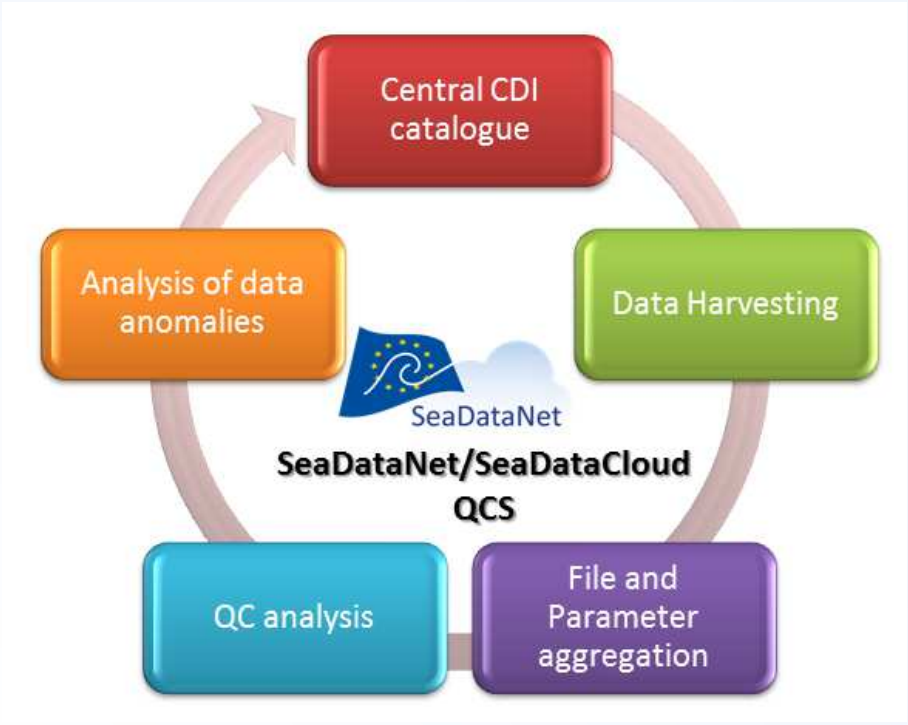
C. Coatanoan & regional leaders

sdn-userdesk@seadatanet.org – www.seadatanet.org

SeaDataCloud Plenary Meeting (November 2018), Barcelona

QC Loop & Quality Check Strategy

SDN2 project implemented and continuously refined a **Quality Control Strategy (QCS)** aiming at improving the quality of the database content and creating the best data products.



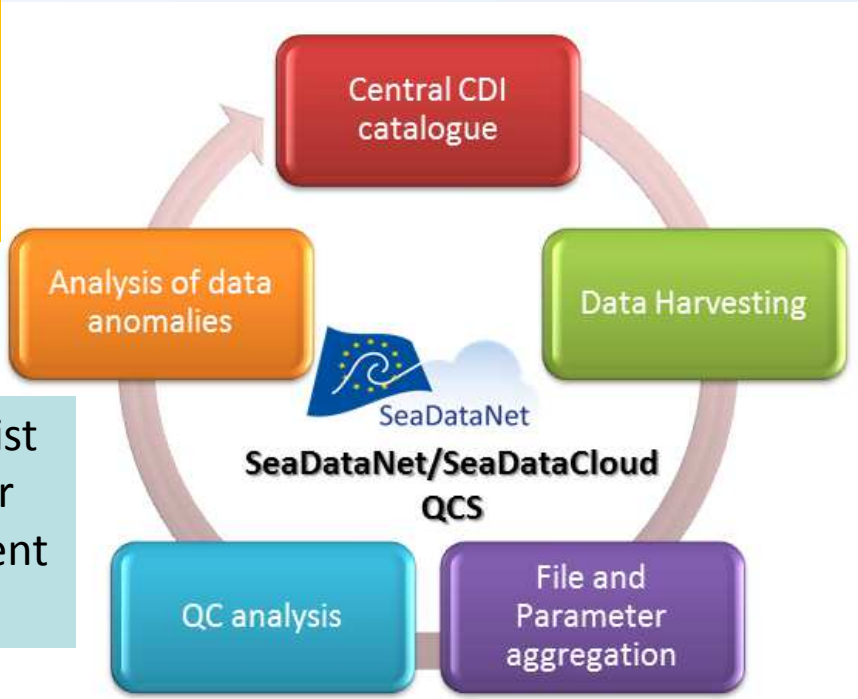
Iterative approach to facilitate the **upgrade** of the database and **versioning** of data products through:

- the release of new data collections at the end of each QCS loop
- the generation of derived climatological products after a certain time lag dedicated to data processing

Quality Check Strategy

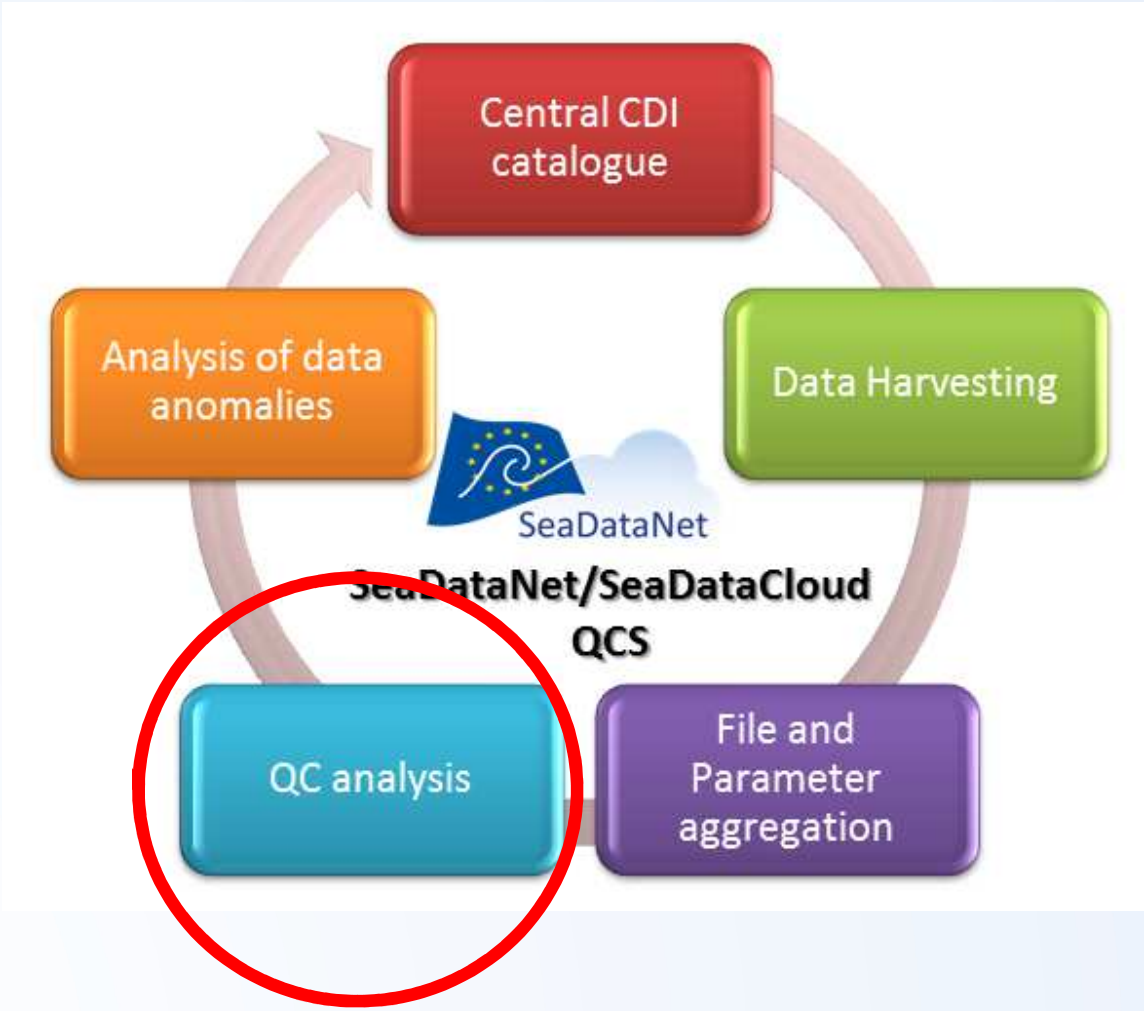
Data providers have to timely analyze the list of anomalies and make the necessary corrections on the quality flags or the data format and update the CDI

Regional products leaders compile a list of data anomalies and organize it per EDMO code. The list of anomalies is sent to the data providers

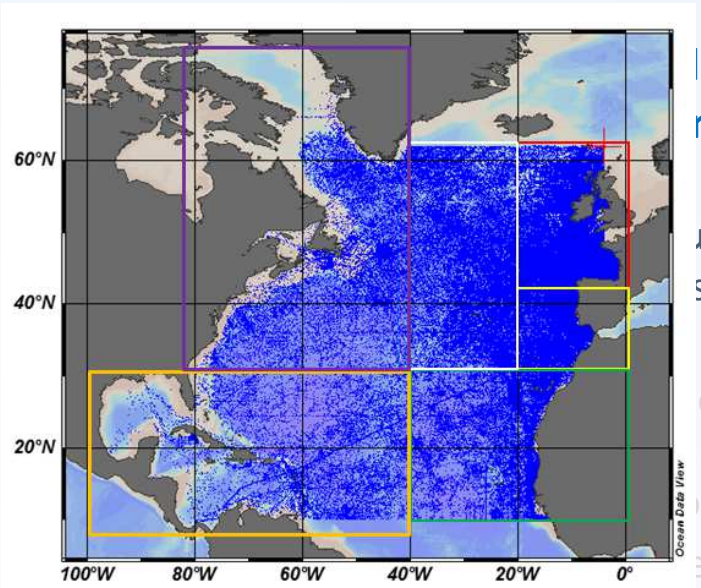


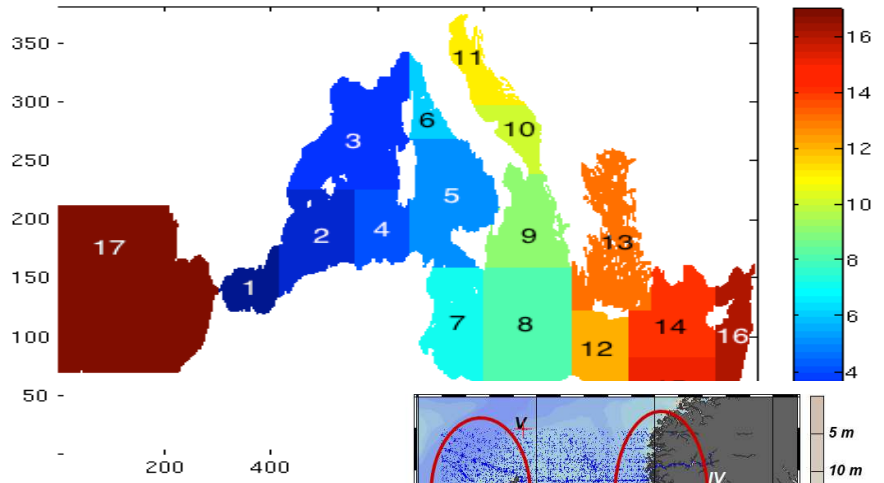
A rapid feedback from the data providers guarantees:

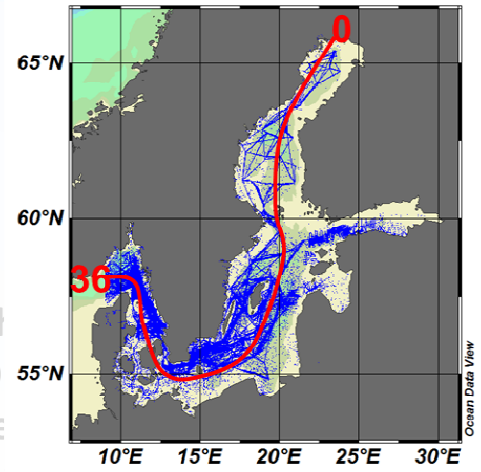
- **the timely generation of data products → increasing user confidence and awareness**
- **the upgrade of the database content → no mismatch among products and CDI service**

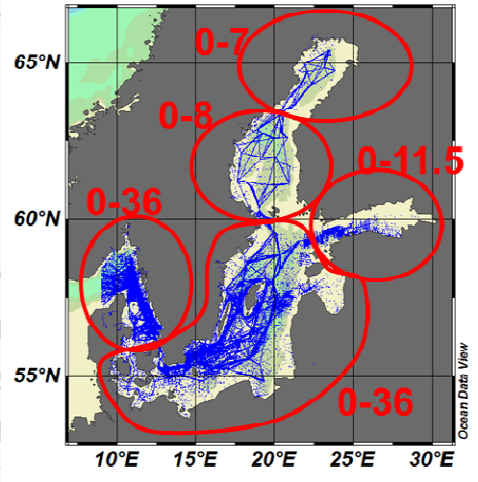

Central CDI catalogue
Data Harvesting
File and Parameter aggregation
QC analysis
Analysis of data anomalies
SeaDataNet
QCS

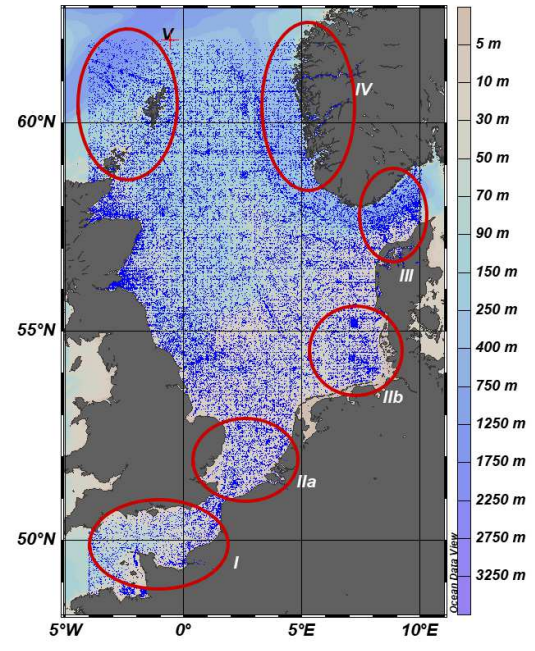
Quality Control procedure









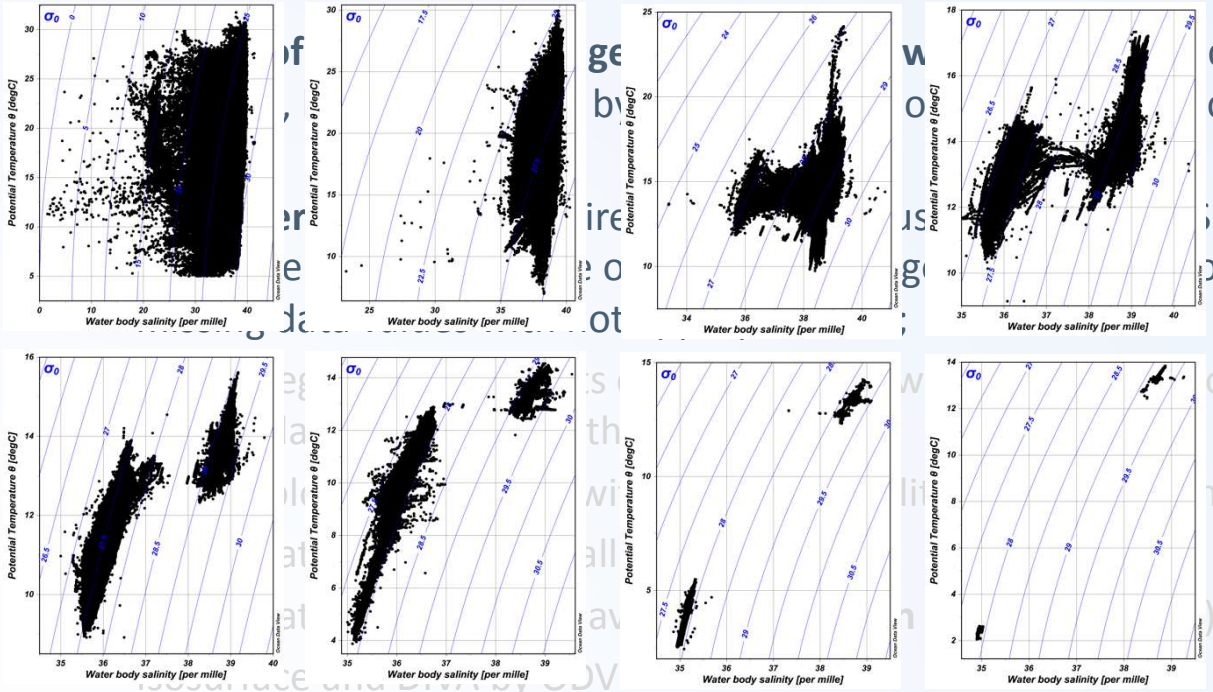


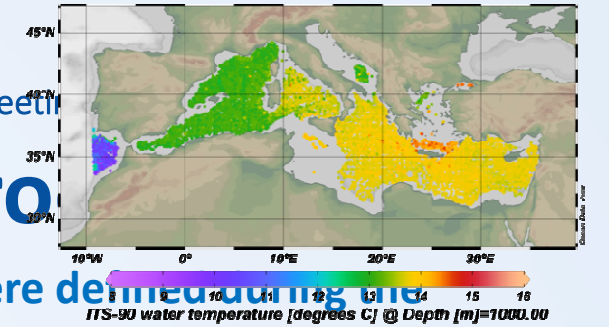
SeaDataCloud

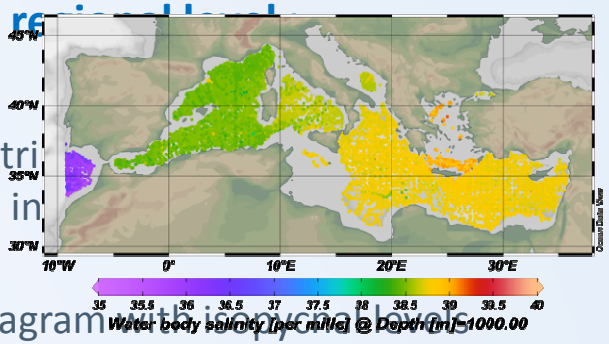
SeaDataCloud Plenary Meeti

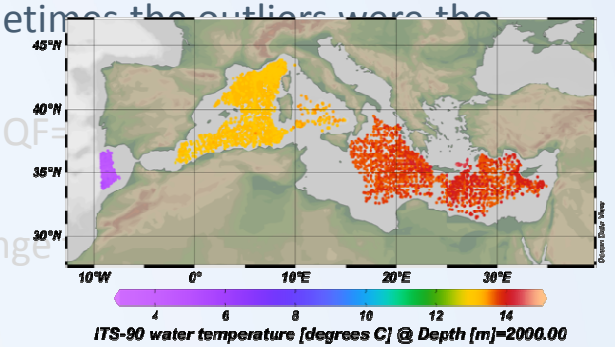
Quality Control pro

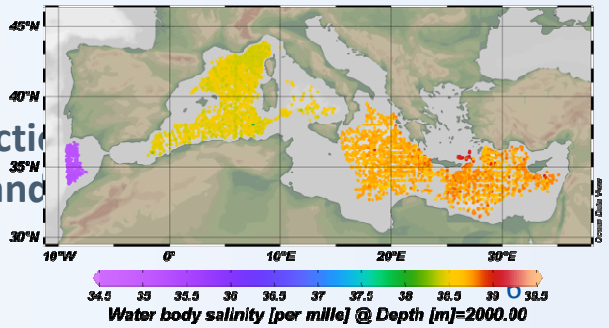
The quality control work followed the best practices that were defined during the **project SeaDataNet 2 and additional steps introduced at the re**gional level











The most useful and powerful quality control used was visual inspection using ODV to discover spikes, outliers, unstable profiles and stations on land, and selection by criteria (instrument type, by data centre ….).

sdn-userdesk@seadatanet.org – www.seadatanet.org

Quality Control procedure

The quality control work followed the best practices that were defined during the project SeaDataNet 2 and additional steps introduced at the regional level :

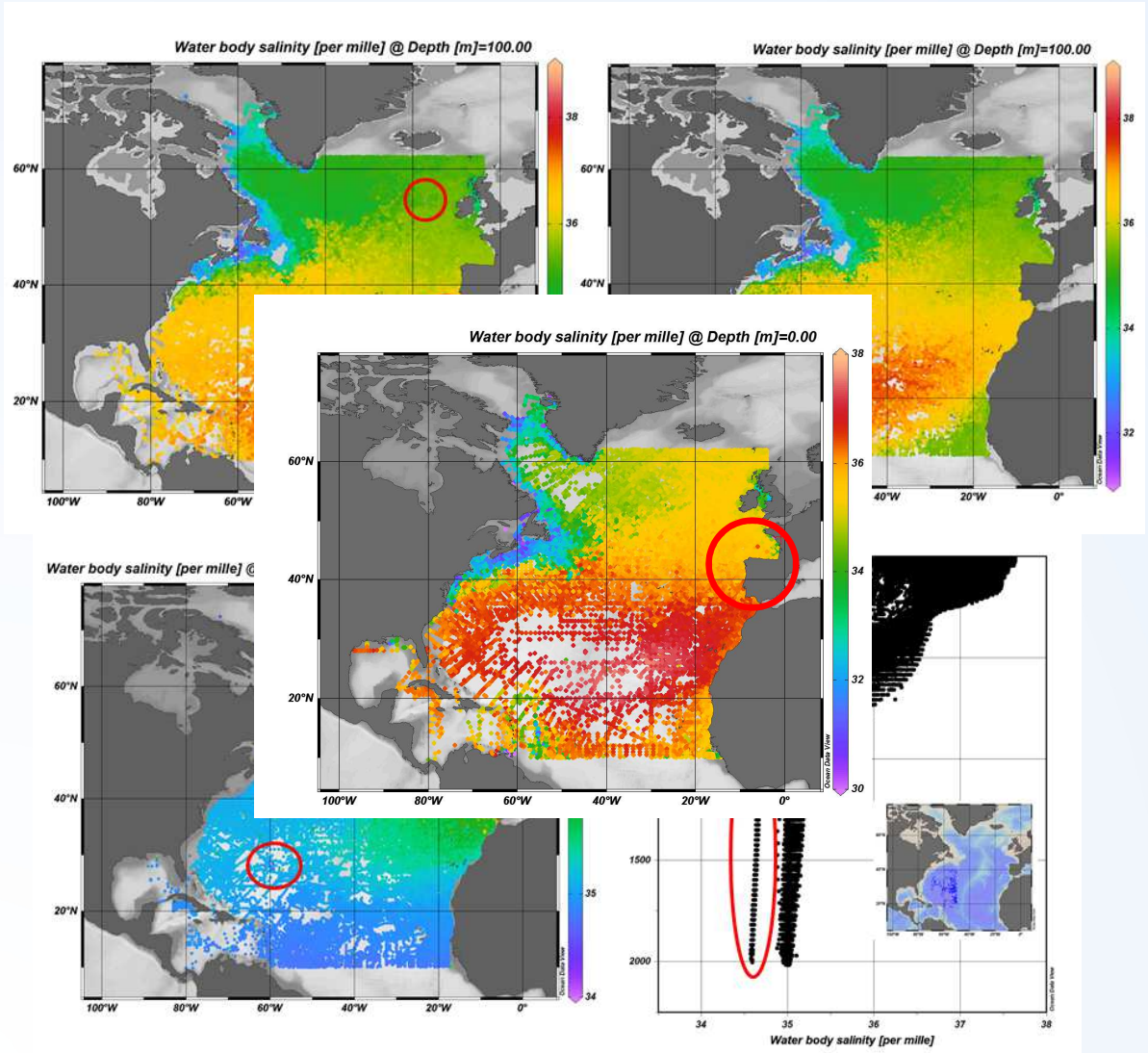
- **Checks of the data coverage, by sub-region when necessary** (distribution for T, S, TS couples), by time periods, by layers (distinction between surface, intermediate and bottom layers);
- **TS scatter plots** of the entire dataset: T versus Z, S versus Z, θS diagram with isopycnal levels for all the QF<3 (check the outliers and change the QF to 4); sometimes the outliers were the missing data values with not appropriate QF;
- By sub-region, scatter plots of observations with QF=1 (good) or QF=2 (probably good) , with a secondary plot showing the **density**;
- Scatter plot observations with **QF=0** (no quality check): only change the bad data with QF4;
- Identification of stations falling **on land**;
- Identification of stations having **unreal depth** (depth values<0);
- Isosurface and DIVA by ODV

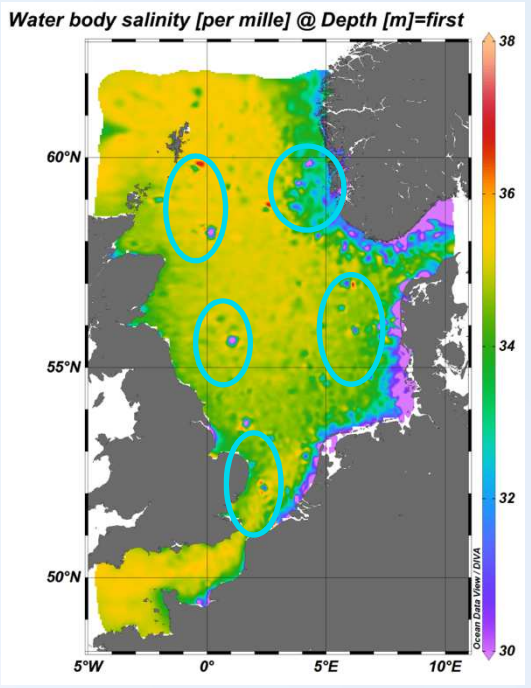
The most useful and powerful quality control used was visual inspection of subsets of data in ODV to discover spikes, outliers, unstable profiles and stations on land. Other use is the selection by criteria (instrument type, by data centre ….).

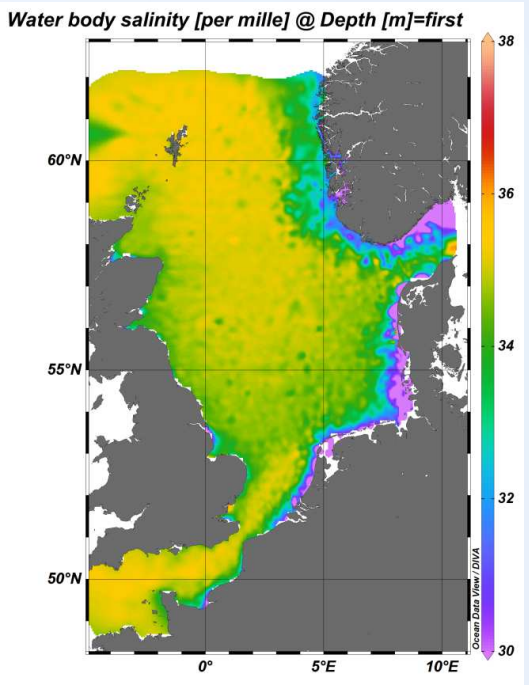
Isosurfaces and DIVA by ODV

SeaDataCloud

SeaDataCloud Plenary Meeti





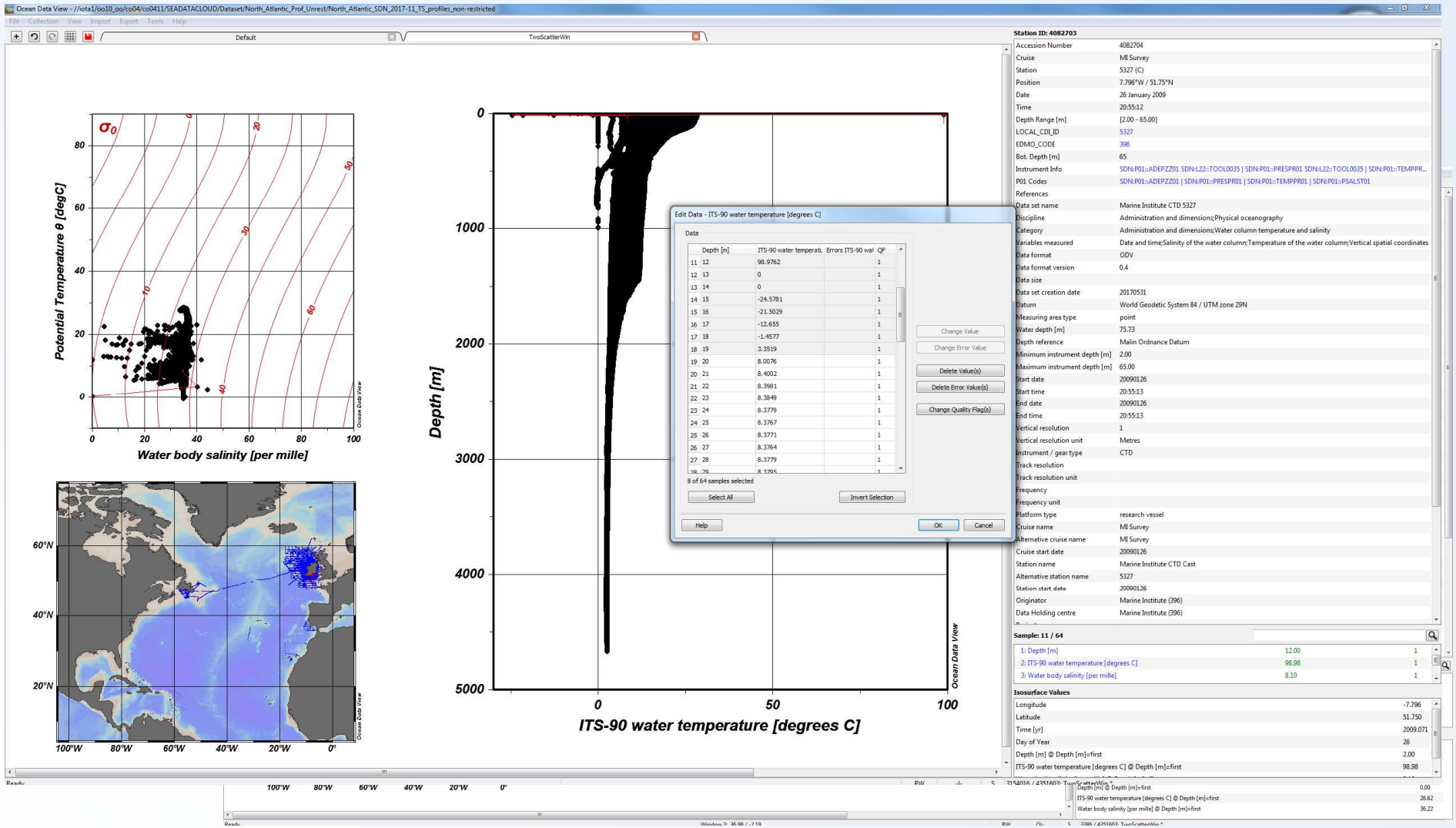


Visual Inspection with ODV

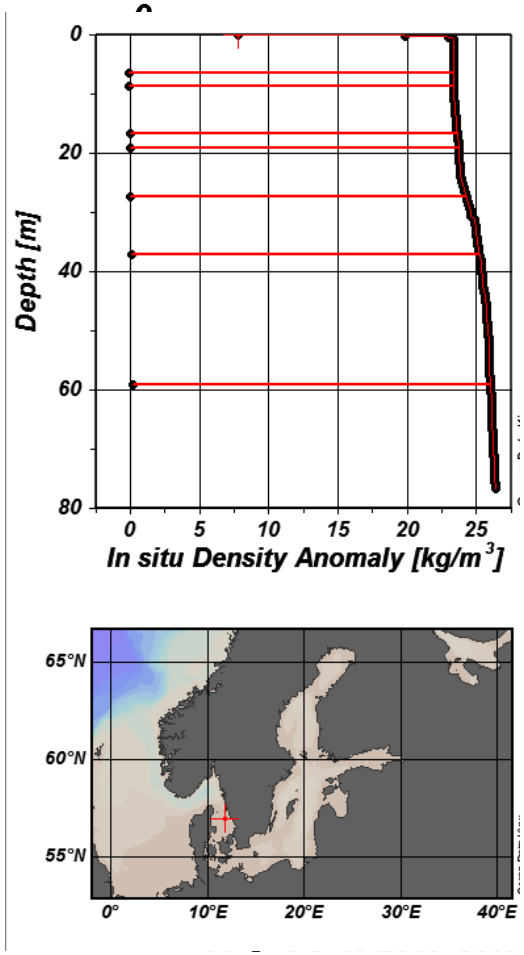
- Ex. of anomalies for some regions (should have been detected at the data center level)

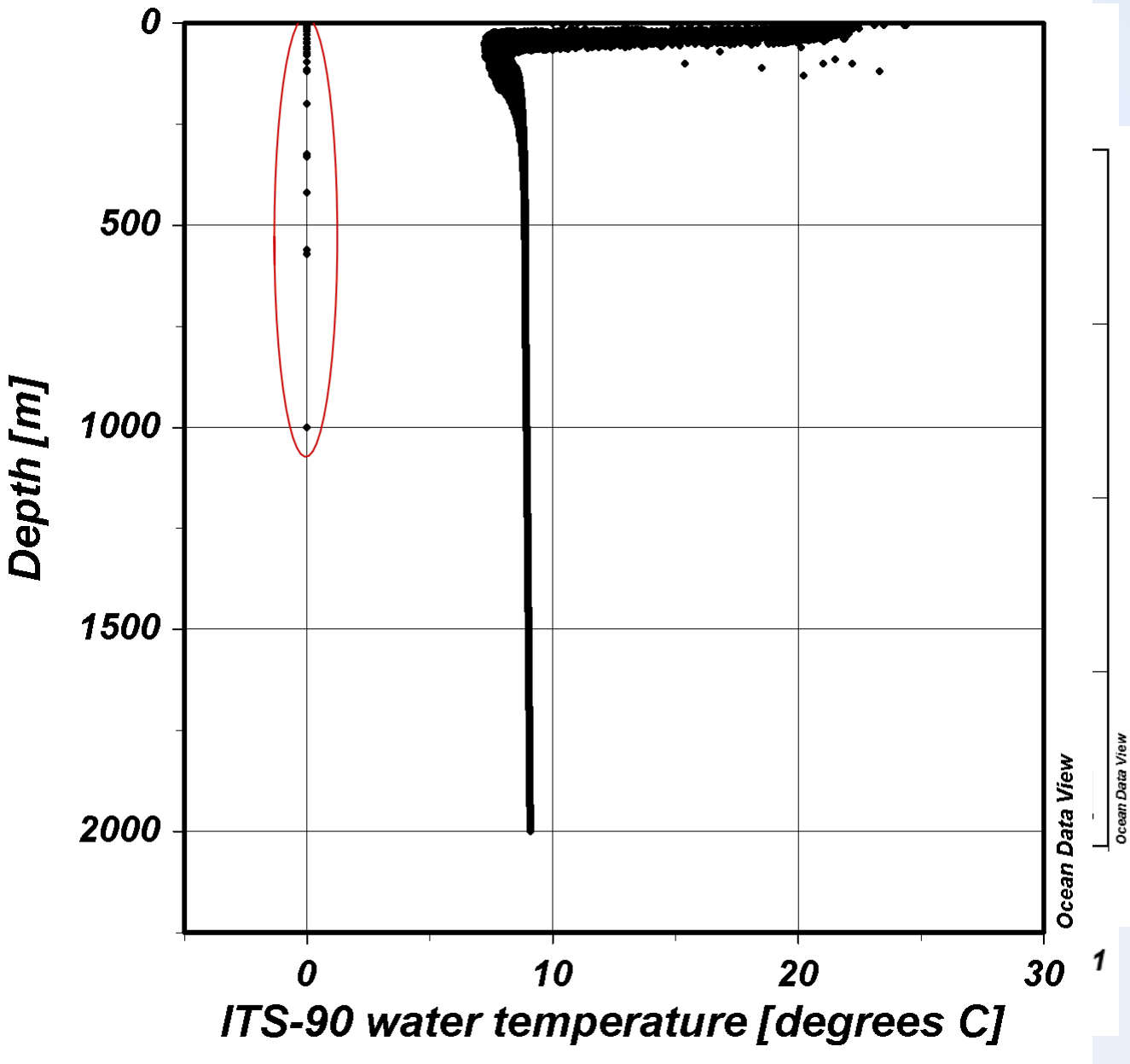
Out of range with QF1

SeaDataCloud Plenary Meeting (November 2018), Barcelona

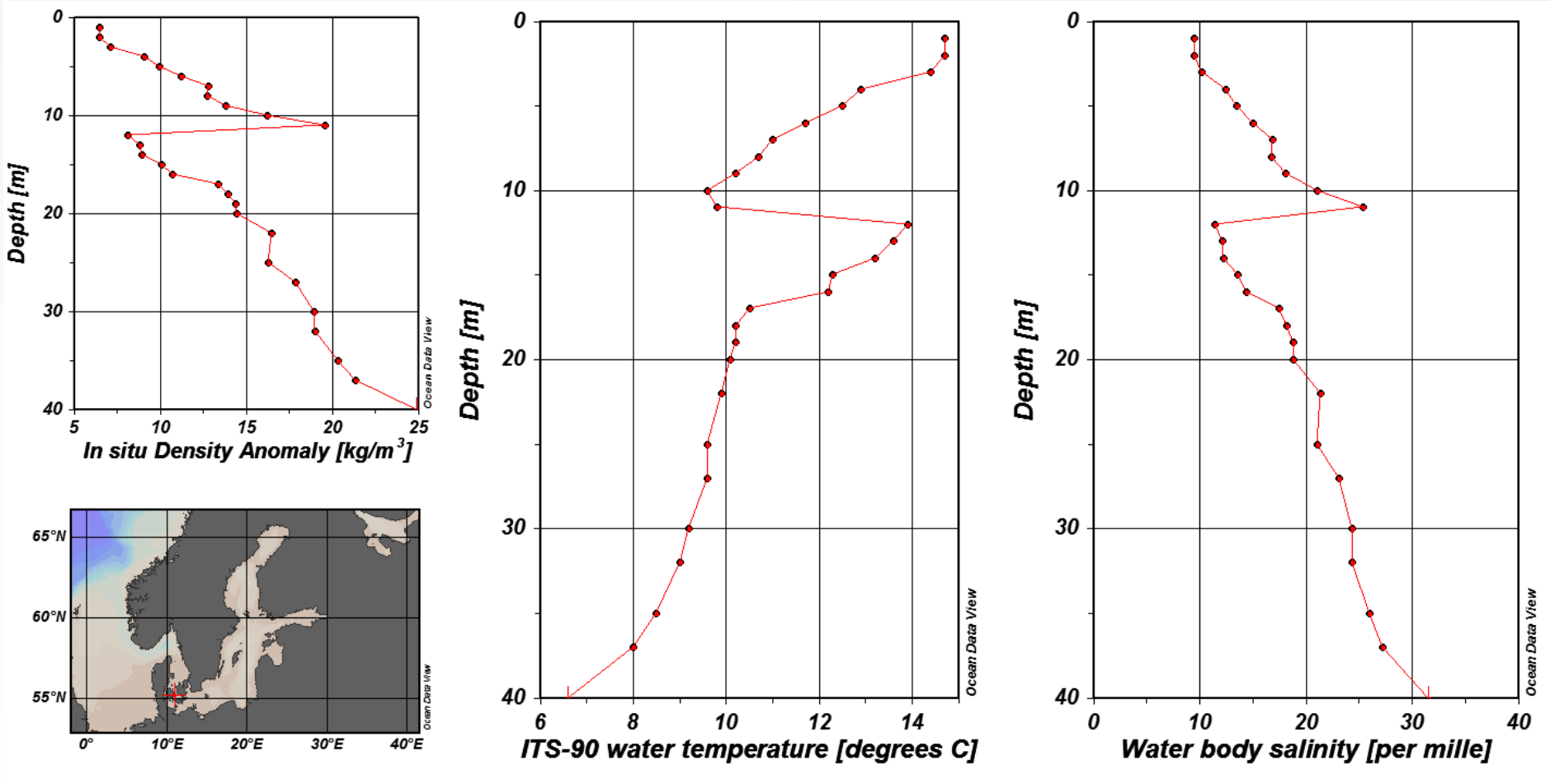


Bad values – Zero reported instead of default values

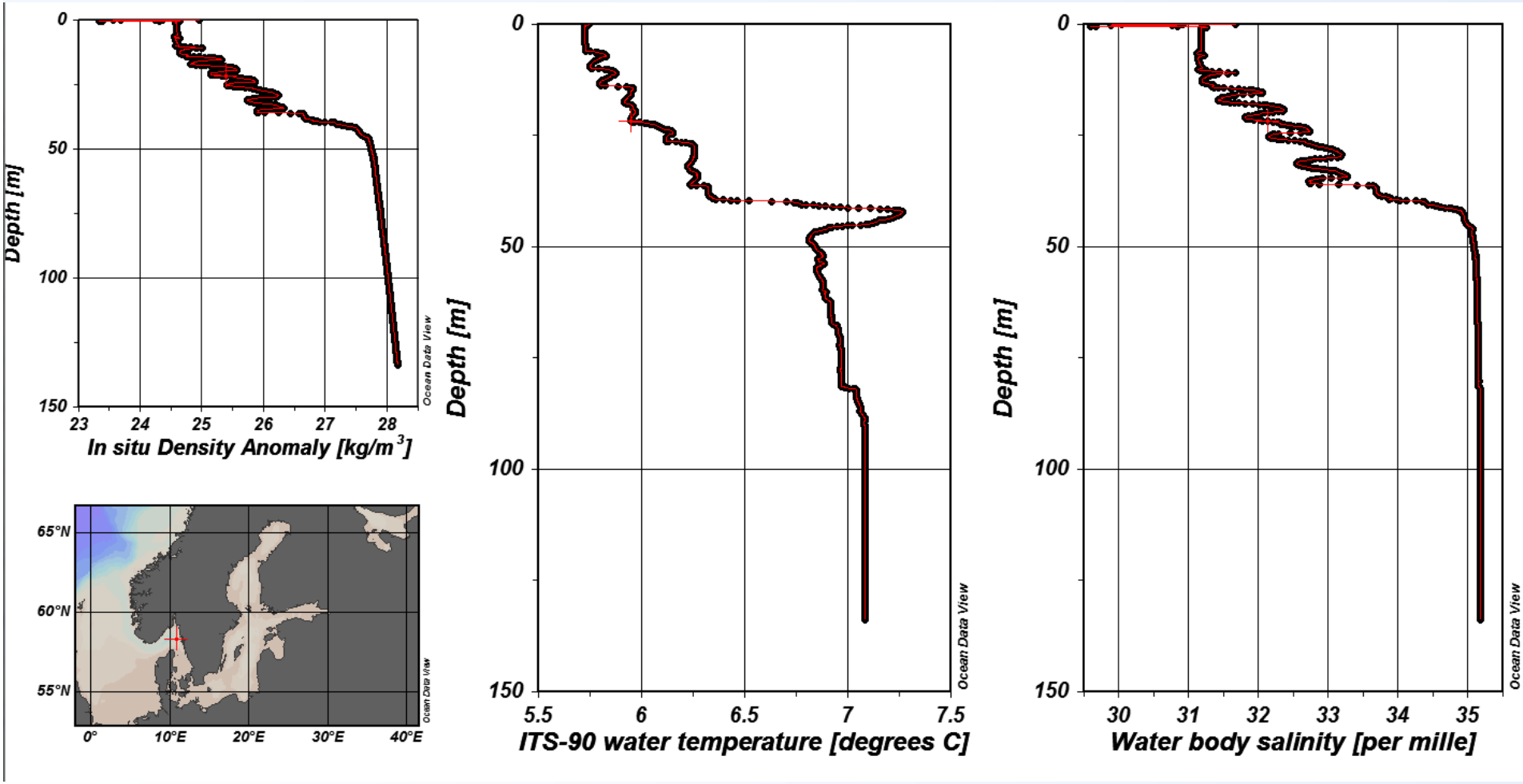




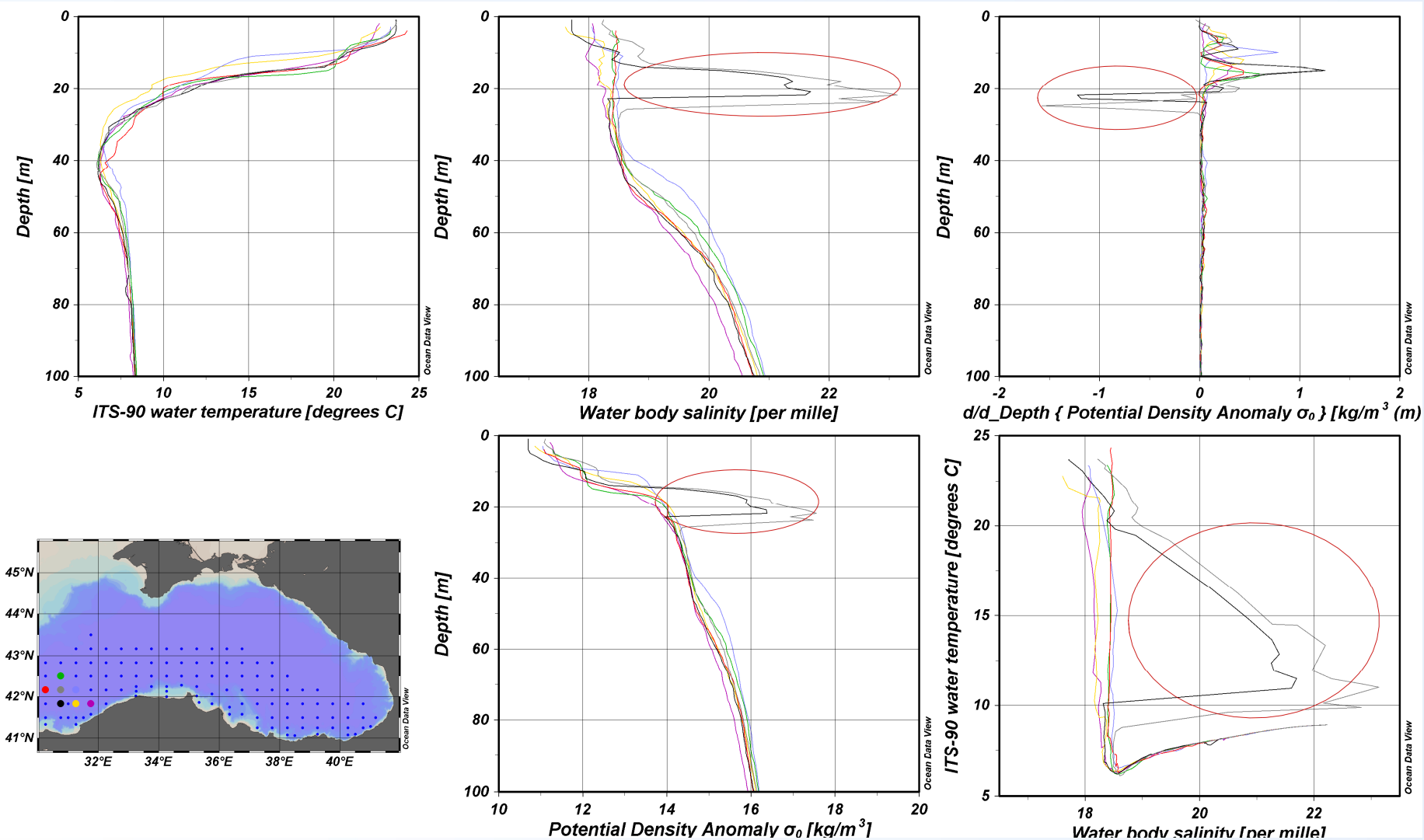
Unstable profiles



Sensor issues



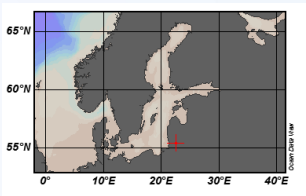
Spike (big and small)



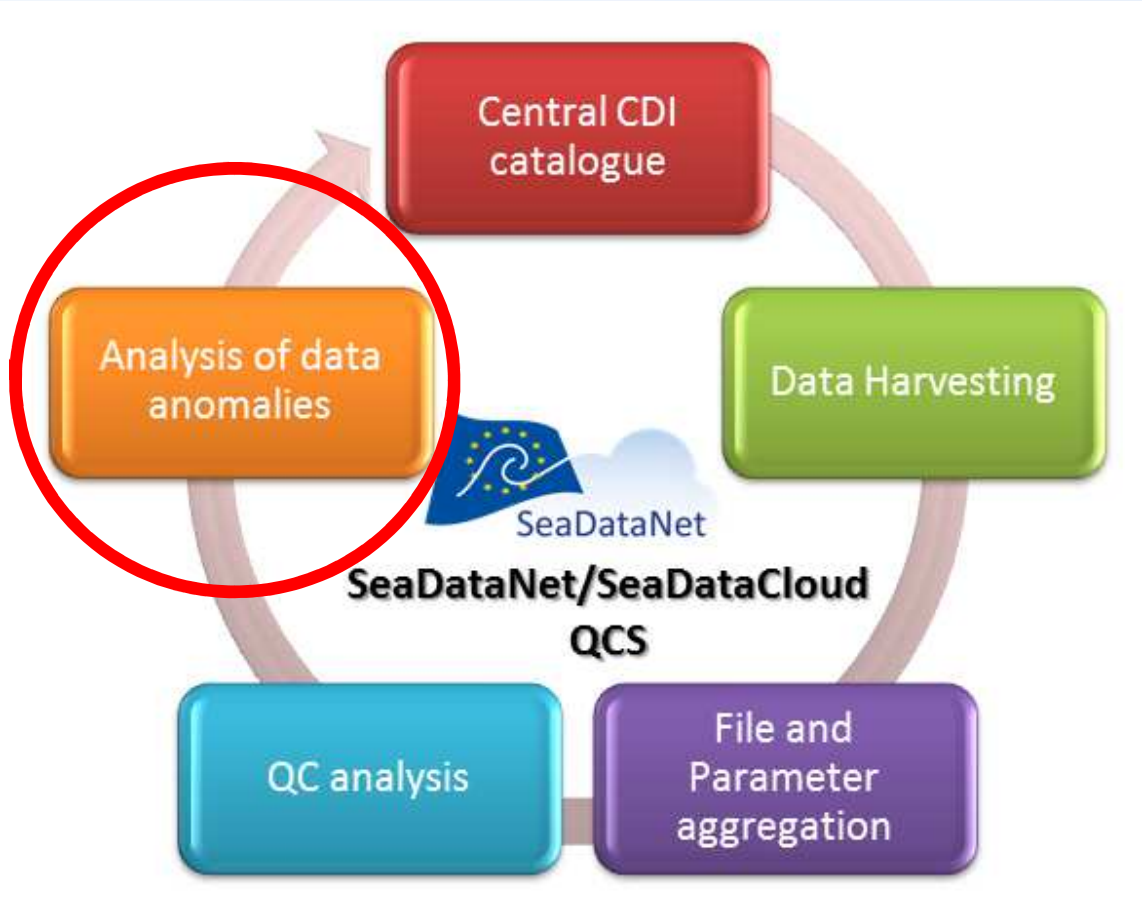
Data anomalies

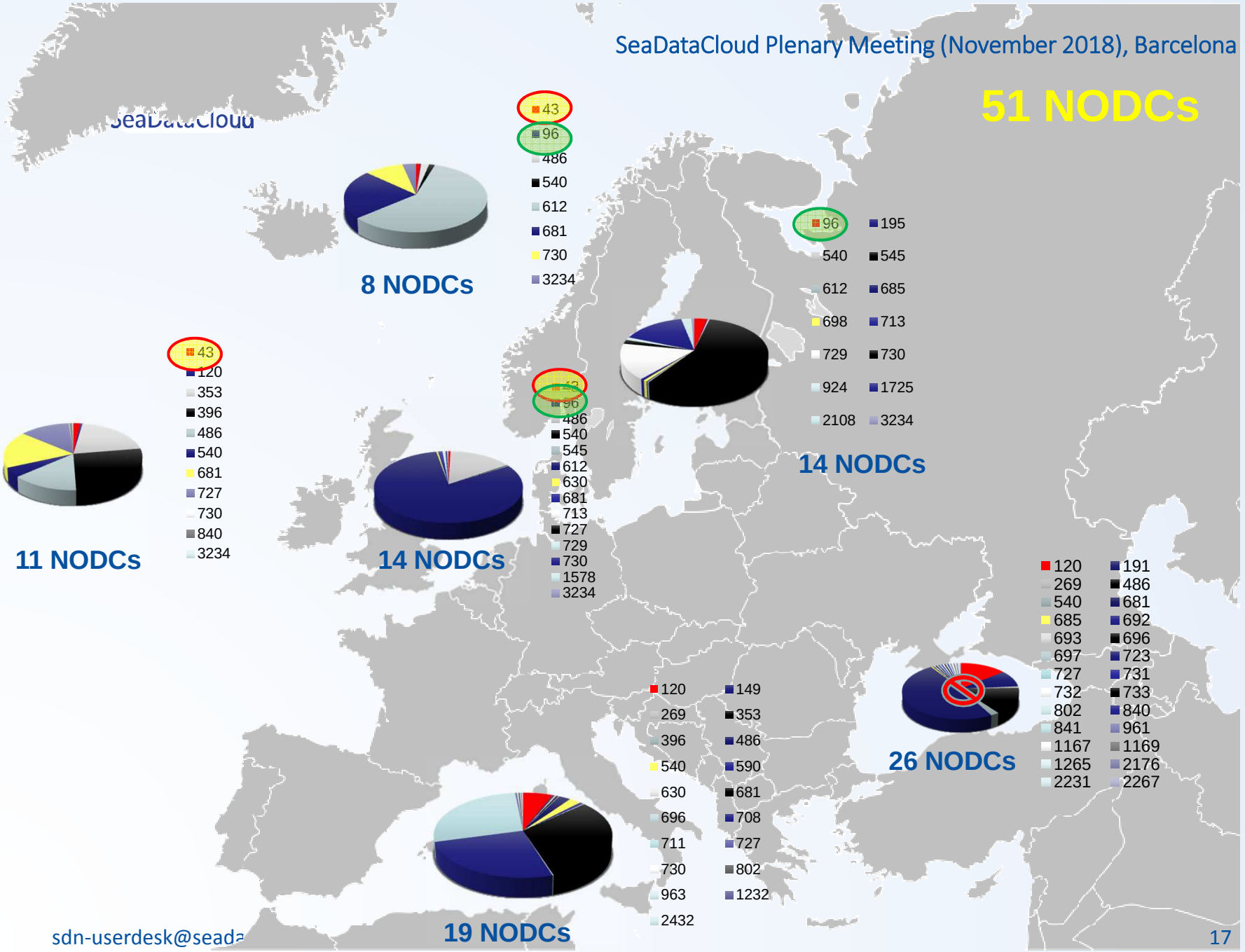
Examples of the various anomalies:

- Format issues : Several missing data values 999.999 or 99.999 or 9.999 or -999.00 or 99.00 and QF0 => missing data can have several values (rules of NODC) but the QF should be 9 (then in ODV : values will be empty and QF9)
- QF 0 (→ must disappear!) [problem identified during aggregation –mix PSAL SSAL bad choice of the QC]
- Missing data badly flagged
- Out of range
- Land points →



- Upgrade QC procedure → refine it, reprocessing is advisable to increase data usage and confidence, but sometimes also to insert some new metadata (i.e. new data types could need some additional metadata)


Central CDI catalogue
Data Harvesting
File and Parameter aggregation
QC analysis
Analysis of data anomalies
SeaDataNet
SeaDataNet/SeaDataCloud QCS


51 NODCs
8 NODCs
43
96
486
540
612
681
730
3234
14 NODCs
96
195
540
545
612
685
698
713
729
730
924
1725
2108
3234
11 NODCs
43
120
353
396
486
540
681
727
730
840
3234
14 NODCs
43
96
486
540
545
612
630
681
713
727
729
730
1578
3234
19 NODCs
120
149
269
353
396
486
540
590
630
681
696
708
711
727
730
802
963
1232
2432
26 NODCs
120
191
269
486
540
681
685
692
693
696
697
723
727
731
732
733
802
840
841
961
1167
1169
1265
2176
2231
2267

Feedback to data providers

What is sent to each CDI partner ?

- The anomalies list with LOCAL_CDI_ID EDMO_CODE PARAMETER_LEVEL and OLDQC NEWQC, and a guideline (explanations of what they have to do and what it is expected)

LOCAL_CDI_ID EDMO_CODE PARAMETER LEVEL_OLDQC_NEWQC

3234_49156 3234 Water body salinity [per mille] @ Depth [m] = {14.897:1 19.863:1 24.828:1 29.794:1 34.759:1 39.724:1 44.689:1 49.654:1 54.618:1 59.583:1 64.547:1 69.512:1 74.476:1 79.44:1 84.404:1 89.368:1 94.332:1 99.295:1 104.259:1 109.222:1} -> 3 3234_49156

3234 Water body salinity [per mille] @ Depth [m] = {114.186:1 119.149:1 124.112:1 129.075:1 134.037:1 139:1 143.963:1 148.925:1 153.887:1 158.849:1 163.812:1 168.774:1 173.735:1 178.697:1 183.659:1 188.62:1 193.581:1 198.543:1 203.504:1 208.465:1} -> 3 3234_49156

3234 Water body salinity [per mille] @ Depth [m] = {213.426:1 218.386:1 223.347:1 228.308:1 233.268:1 238.228:1 243.188:1 248.148:1 253.108:1 258.068:1 263.028:1 267.987:1 272.947:1 277.906:1 282.865:1 287.824:1 292.783:1 297.742:1 302.701:1 307.659:1} -> 3

……

Feedback to data providers

SeaDataCloud Plenary Meeting (November 2018), Barcelona

What is expected from each CDI partner ?

- The anomalies list updated with NODC comments following this table

	A	B	C	D	E
1	Accession number	Local_cdi_id	edmo_code	parameter	Comments
2	58722	Thalassa_CANIGO_0997_Leg_2_9579	396	Water body salinity [per mille] @ Depth [m] = {3.9708:1} -> 3	Doubtful data, correction
3	382688	812	396	ITS-90 water temperature [degrees C] @ Depth [m] = {249:1} -> 4	Bad data, correction

- The list of updated CDI XML :

LOCAL_CDI_ID	EDMO_CODE	PLATFORM_CODE=CRUISE
FI35199101301_00050_H10	486	PRIMO-0 21/03
FI35199443005_25900_H10	486	MBP-FRONT 1994
FI35199502002_00870_H10	486	EUROMARGE
FI35199706005_0K010_H10	486	PELMED 97
FI35199845001_00260_H10	486	BIODYPAR 1

NODCs have to send to MARIS the updated CDI XML

- A report with some informations :

- List of parameters in errors and number

Parameters	Number
Depth	20
Temperature	328
Salinity	544
Duplicates	3606
Others	1
Total	**4499**

- Details on why corrections have not been taken into consideration, etc……

Status of the feedback

- **51 NODCS (First message sent on 10th July – Reminder on 10th October)**

DATELINE for feedback/updated CDI files with corrections : fixed at end of November

(a reminder has been sent on 31th October for some who gave no answer and for which the contact points were correct)

- 36 answers ->> Answers : meaning acknowledged receipt of the message, saying that they are working on, those who need more explanation to understand the message or work done

14/36 real answers (with updated report and CDI XML Files sent to Maris)

1/36 with feedback and report, CDI XML files will be updated in November

70% of answers with ~39% of work already processed (on answers)

Status of the corrections

- Some NODCs agree with our corrections

=> CDI updated – improved data quality in the dataflow and for the next release

- Some NODCs not agree with our corrections

❑ ex. Close to river, specific ice condition showing untypical stratification of water masses, thermal mixing of the water masses

⇒ CDI not updated – (take into account feedback at regional level)

- No answer from NODCs

=> CDI not updated – same data for the next release (will be outliers for our products)

Conclusion at the Regional Product level : to get only the new data for the next integration in the product ("*to not waste our time for checking same 'bad' data*")

- **Highlight importance for NODCs to answer and correct the data and update the missing metadata**

➔ **importance of this step for the data quality, not only for the products but for all the users**

No feedback for those NODCs

christine.coatanoan@ifremer.fr

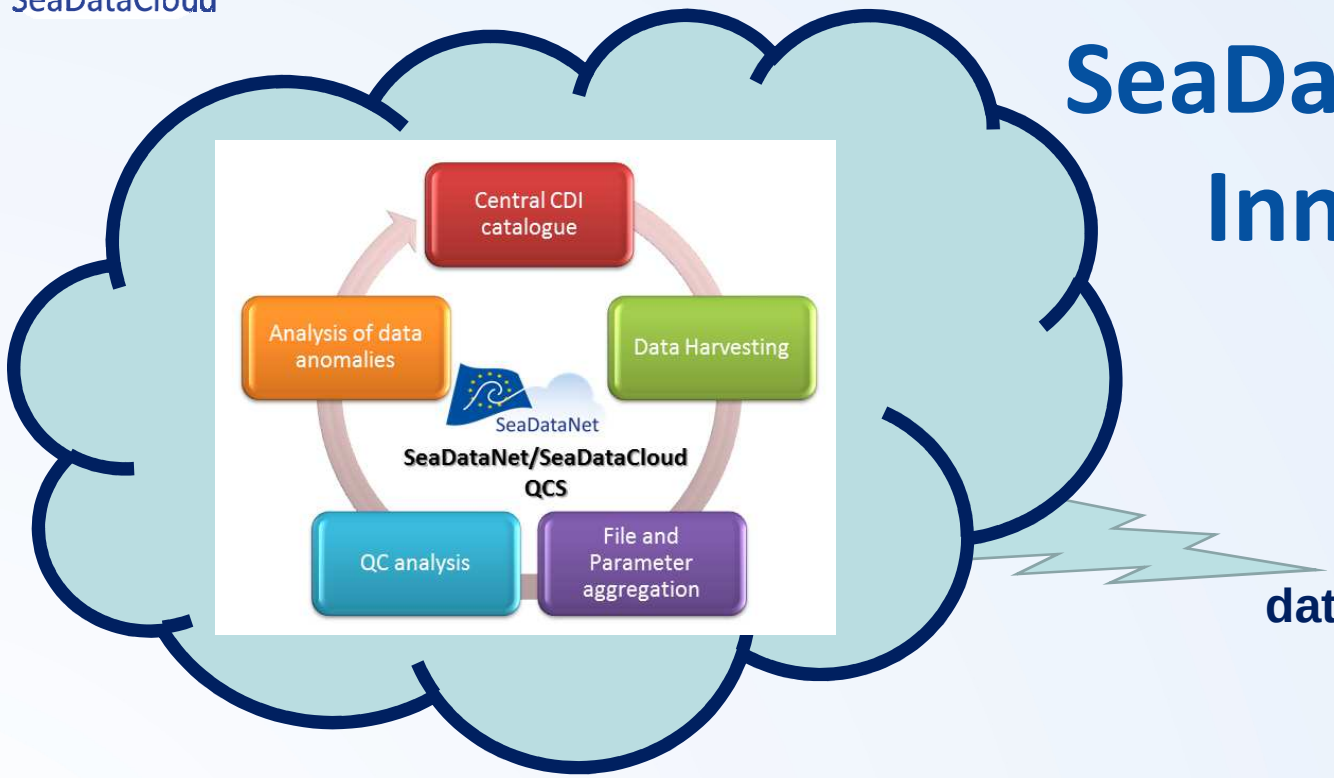
Barcelona

EDMO_CODE			
692	Bulgarian National Oceanographic Data Centre(BGODC), Institute of Oceanology	Bulgaria	Bulgarian National Oceanographic Data Centre(BGODC), Institute of Oceanology
693	Iv.Javakhishvili Tbilisi State University, Centre of Relations with UNESCO Oceanological Research Centre and GeoDNA (UNESCO) (TSU-DNA)	Georgia	Iv.Javakhishvili Tbilisi State University, Centre of Relations with UNESCO Oceanological Research Centre and GeoDNA (UNESCO) (TSU-DNA)
723	State Oceanographic Institute (SOI) Moscow	Russian Federation	All-Russia Research Institute of Hydrometeorological Information - World Data Centre (RIHMI-WDC) National Oceanographic Data Centre (NODC) (RIHMI-WDC)
727	Marine Hydrophysical Institute	Ukraine	Marine Hydrophysical Institute
731	Department of Navigation and Hydrography and Oceanography, Turkish Navy	Turkey	Institute of Marine Sciences, Middle East Technical University
733	Sinop University, Fisheries Faculty	Turkey	Institute of Marine Sciences, Middle East Technical University
802	Istanbul University, Institute of Marine Science and Management	Turkey	Institute of Marine Sciences, Middle East Technical University
840	Institute of Biology of the Southern Seas, NAS of Ukraine	Ukraine	Marine Hydrophysical Institute
924	Russian State Hydrometeorological University, St-Petersburg	Russian Federation	All-Russia Research Institute of Hydrometeorological Information - World Data Centre (RIHMI-WDC) National Oceanographic Data Centre (NODC) (RIHMI-WDC)
961	National Institute of Meteorology and Hydrology, Bulgarian Academy of Sciences	Bulgaria	Bulgarian National Oceanographic Data Centre(BGODC), Institute of Oceanology
1169	Odessa National I.I.Mechnikov University	Ukraine	Marine Hydrophysical Institute
2108	Environmental Protection Agency (EPA)	Lithuania	Department of Marine Research of the Environmental Protection Agency
2176	Ankara University	Turkey	Institute of Marine Sciences, Middle East Technical University
2231	Danube Hydro-meteorological Observatory	Ukraine	Marine Hydrophysical Institute
2267	National Environmental Agency of the Ministry of Environment Protection and Natural Resources (NEA)	Georgia	Iv.Javakhishvili Tbilisi State University, Centre of Relations with UNESCO Oceanological Research Centre and GeoDNA (UNESCO) (TSU-DNA)

From a visual inspection to an automatic approach……

SeaDataCloud Innovation
VRE

SeaDataCloud Innovation



The **implementation of the cloud environment** will optimize and automate the QCS at the central level assuring a continuous monitoring of the database content and quality, together with the possibility of generating database snapshots on a regular basis and allowing data products versioning.

Virtual Research Environment

- How to manage anomalies on data ?

to get the information each time a QF is defined as doubtful or bad at the regional level with WebODV, correct and charge automatically the corrected data by the NODC

- The positive impact of VRE on data providers

automatic management of data anomalies : to increase quality and optimize/ensure timeliness of the exchanges between NODCs and users


THANK YOU
FOR YOUR
ATTENTION