SeaDataCloud

Data Quality: why it is so important?

The experience of the regional products managers and recommendations

S. Scory, S. Simoncelli, C. Coatanoan, Ö. Bäck, S. Iona, V. Myroshnychenko

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

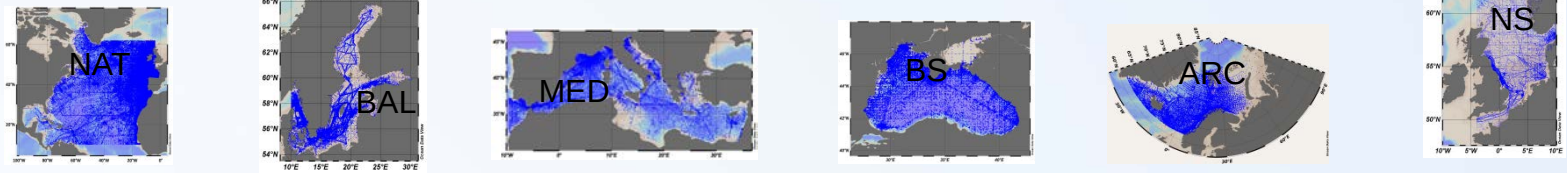
OUTLINE

- SeaDataCloud data products and objectives
- SeaDataCloud Quality Check Strategy
- SeaDataCloud products' timeline
- Quality Control procedure
- Data anomalies and data providers' response
- The importance of full metadata record
- Unlock your data and set them free
- PIDocs and acknowledgment of data providers
- SeaDataCloud innovation

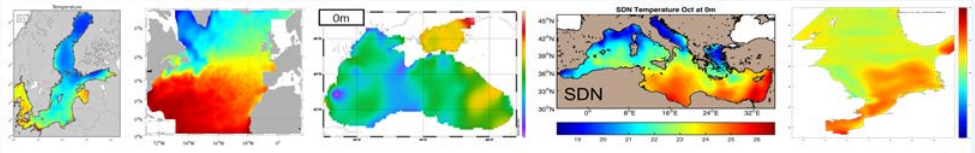
SeaDataCloud Data Products

SeaDataCloud aims at providing **data products** deriving from SeaDataNet infrastructure at **regional and global scale** to serve a diverse user community:

1. **Aggregated data sets for all the European marginal seas**→ all historical temperature and salinity (1900 onwards) data harvested from the central CDI and validated by regional leaders



2. **Climatologies** → gridded fields obtained through a mapping technique (DIVA) and representing the climate of the ocean at both regional and global scale



3. **New data products** → multi-platform and multi-disciplinary approach combining both in situ (e.g. gliders, Argo, ships, drifters, fixed platforms) and remote sensed observations, Ocean Monitoring Indicators for tracking ocean mechanisms and/or climate modes and trends

Objectives

- **Improve the quality of the** overall **infrastructure content** through systematic quality assessment (every 2 years)
- **develop new methods to ensure quality, homogeneity and robust uncertainty measures in long-term time-series of data**
- **Integrate external datasets** (Copernicus Marine Environment Monitoring Service, World Ocean Database) to increase temporal and spatial resolution and further improve products' quality
- **Generate the best data products** to serve different user groups (operational oceanography, climate, marine environment, institutional, academia) adopting the most advanced methodologies
- **Increase user uptake** providing **timely and reliable information** of the full product generation process and its quality

Example application

Ocean Prediction

Atm Forcing

Initial Condition

Ocean general circulation models (OGCMs)

Forecast/analyses reanalyses

Data Assimilation

Example application

Ocean Prediction

Atm Forcing

climatologies

Initial Condition

Ocean General Circulation Models (OGCMs)

Forecast/analyses reanalyses

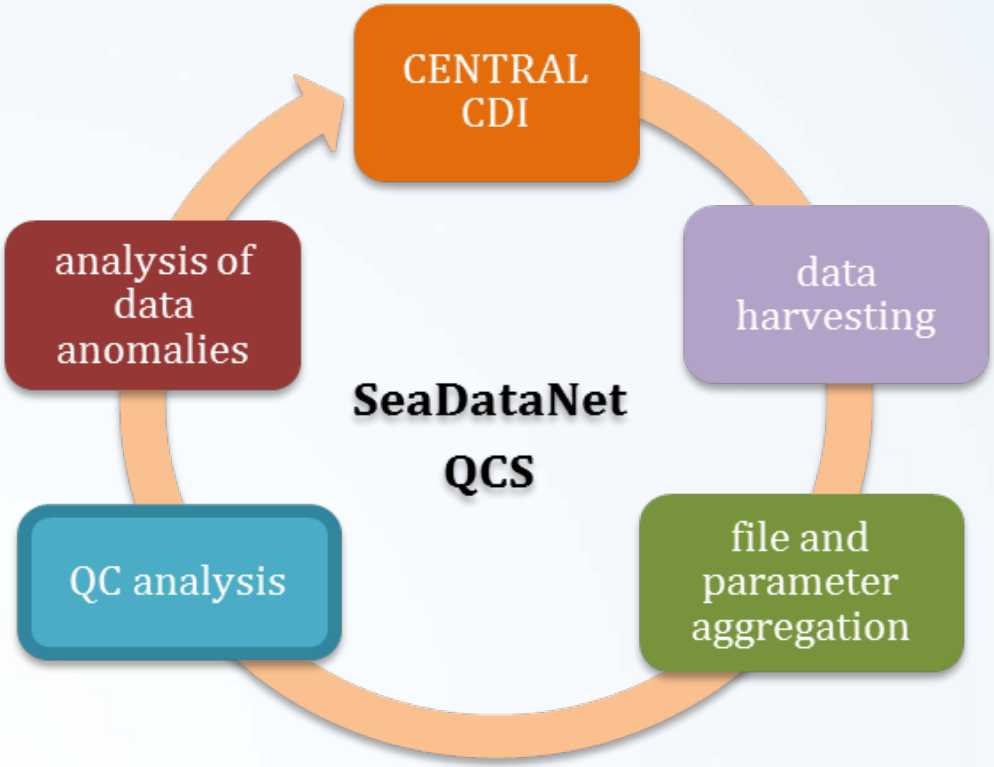
Data Assimilation

Validation procedures rely on climatologies, gridded reconstructions, reprocessed time series of data

Reanalyses→ harmonized historical data collections

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



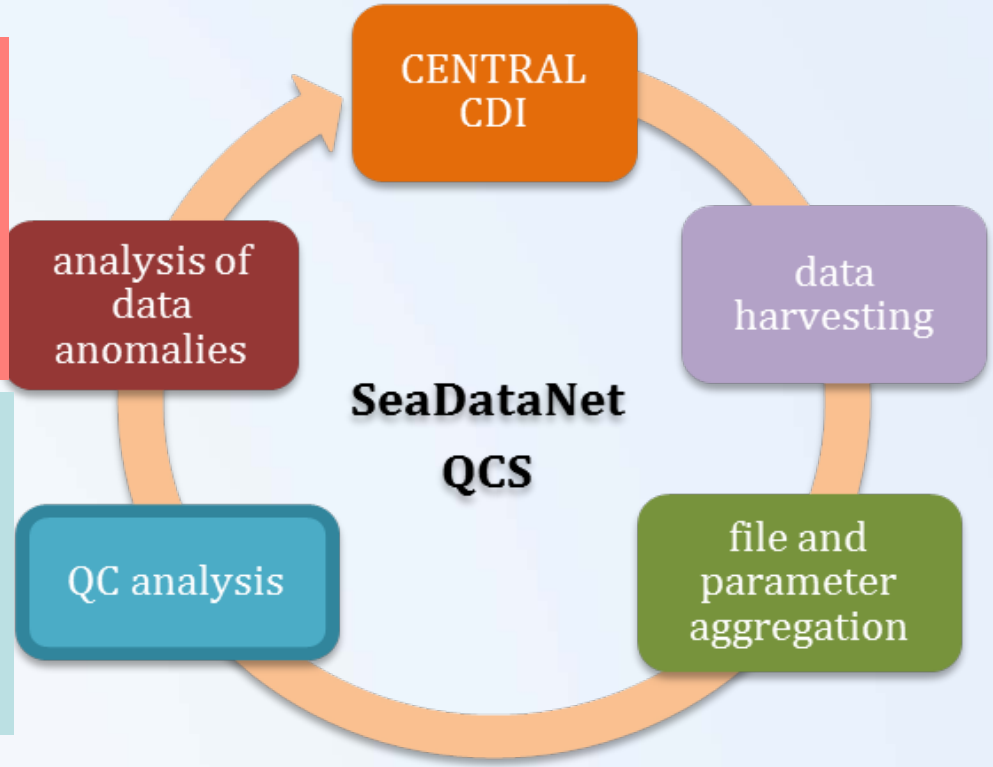
Iterative approach to facilitate the **upgrade** of the database and **versioning** of data products through:

- 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

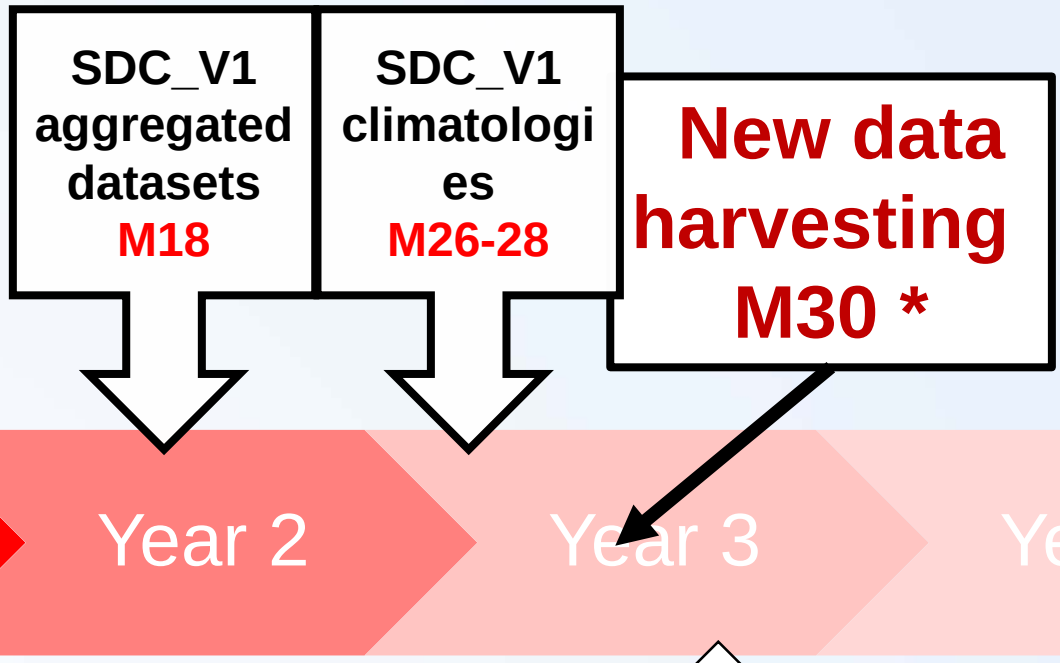
Regional products leaders compile a list of data anomalies and organize it per EDMO code. The list of anomalies is sent to the data providers



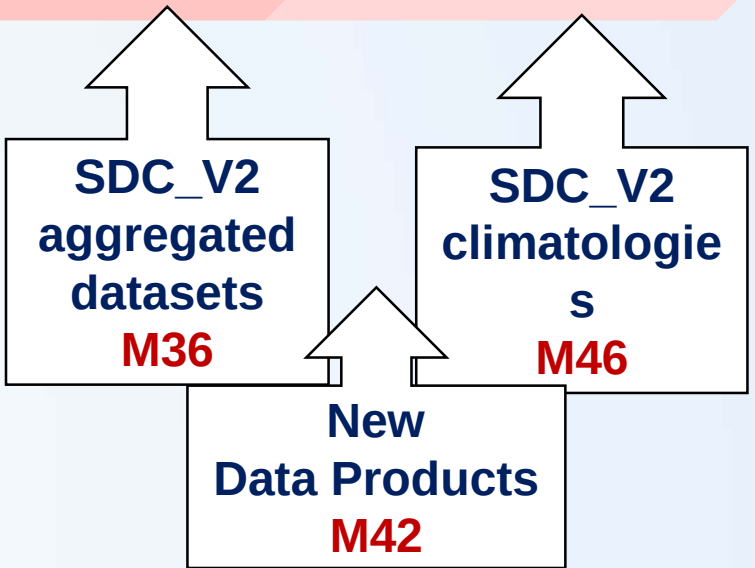
A rapid feedback from the data providers guarantees:

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

WP11 New Timetable



* It is crucial to correct data anomalies (M26→Dec2018) and ingest all available data before next data harvesting (M30→Apr2019)



Quality Control procedure

The quality control work follows the best practices that were defined during the project SeaDataNet 2:

- **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 plot of observations with QF=1 (good) with a secondary plot showing the density;
- By sub-region, scatter plot of observations with 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);
- 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.

Quality Control procedure

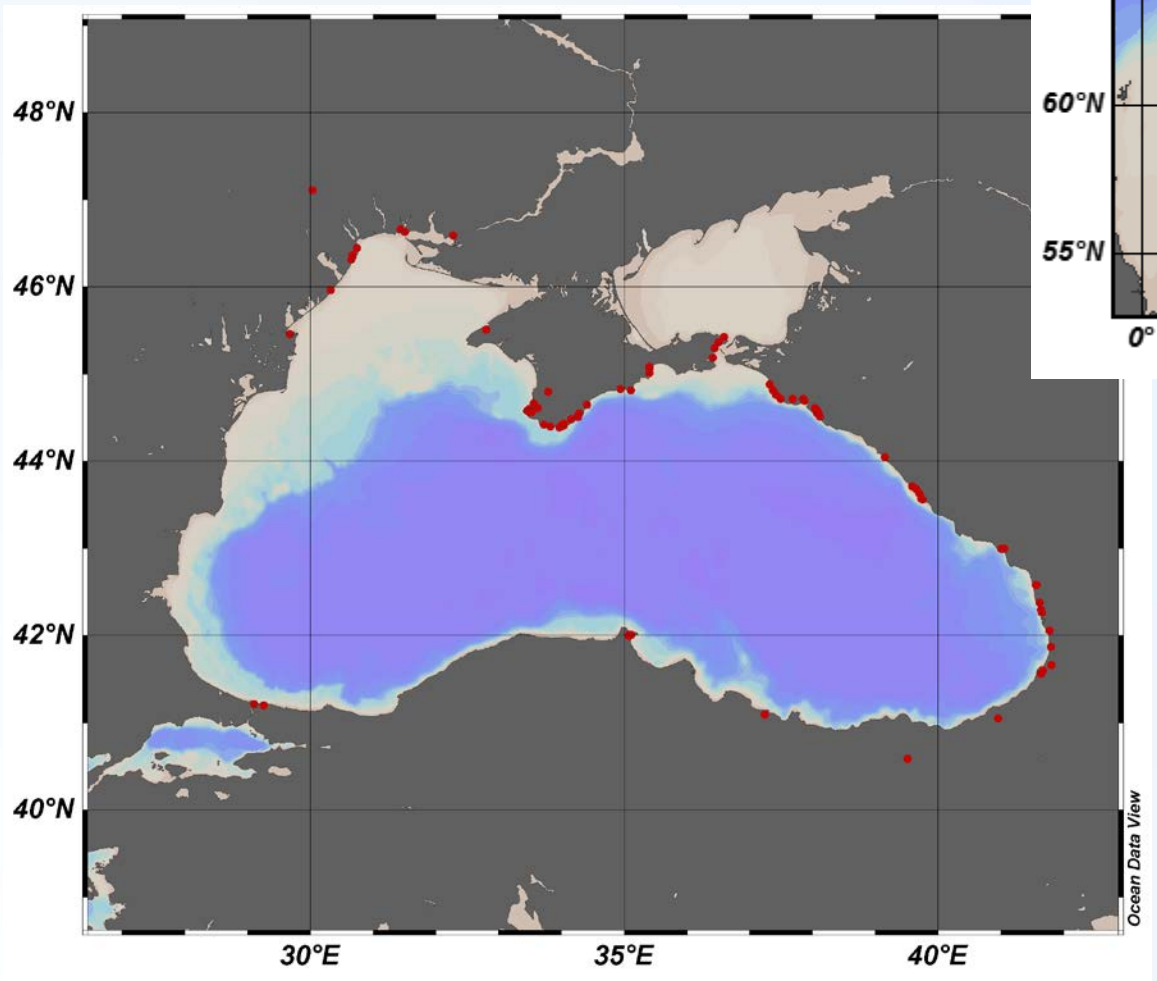
- QF statistics on correction

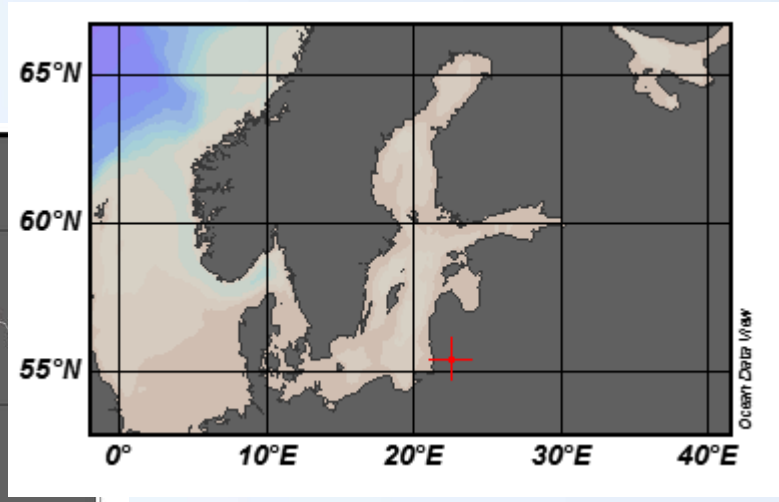
	BEFORE CORRECTION (%)				AFTER CORRECTION (%)			
	QC0	QC1	QC2	QC3-4 (5-9)	QC0	QC1	QC2	QC3-4 (5-9)
ARTIC SEA								
TEMPERATURE	0.66	98.99	0.01	0.34	0	99.63	0.01	0.36
SALINITY	0.68	98.35	0.01	0.96	0	98.5	0.01	1.48
BALTIC SEA								
TEMPERATURE					0.5	80.2	19.3	<0.1
SALINITY					0.5	80.2	19.2	<0.1
BLACK SEA								
TEMPERATURE	11.6	91.9	4.62	1.89	0	95.46	3.1	1.44
SALINITY	11.96	83.87	2.88	1.28	0	96.23	1.82	1.95
NORTH ATLANTIC								
TEMPERATURE	9	99.3	0.09	0.52	0.09	99.29	0.09	0.53
SALINITY	2.38	95.39	0.23	2	2.39	95.34	0.24	2.03
NORTH SEA (discrete)								
TEMPERATURE					0	98.79	1.16	0.04
SALINITY					0	98.29	1.37	0.34
MED SEA								
TEMPERATURE	2.7	96.9		0.3	0	99.8		0.2
SALINITY	4.5	94.6		0.9	0	99.2		0.8

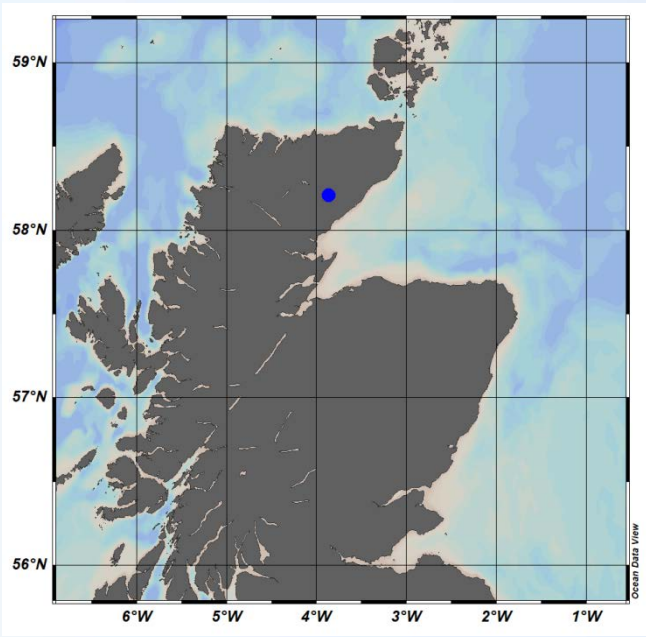
Issues with metadata

- Minimal set of metadata → almost useless
- Wrong “measuring area type” (one single measurement defined as “curve”)
- Moorings defined as “profiles” instead of time series” (particular case of Thermistor chains in one CDI → “Time series of times series”?)
- Vertical resolution given in “minutes”
- Station on land

On land

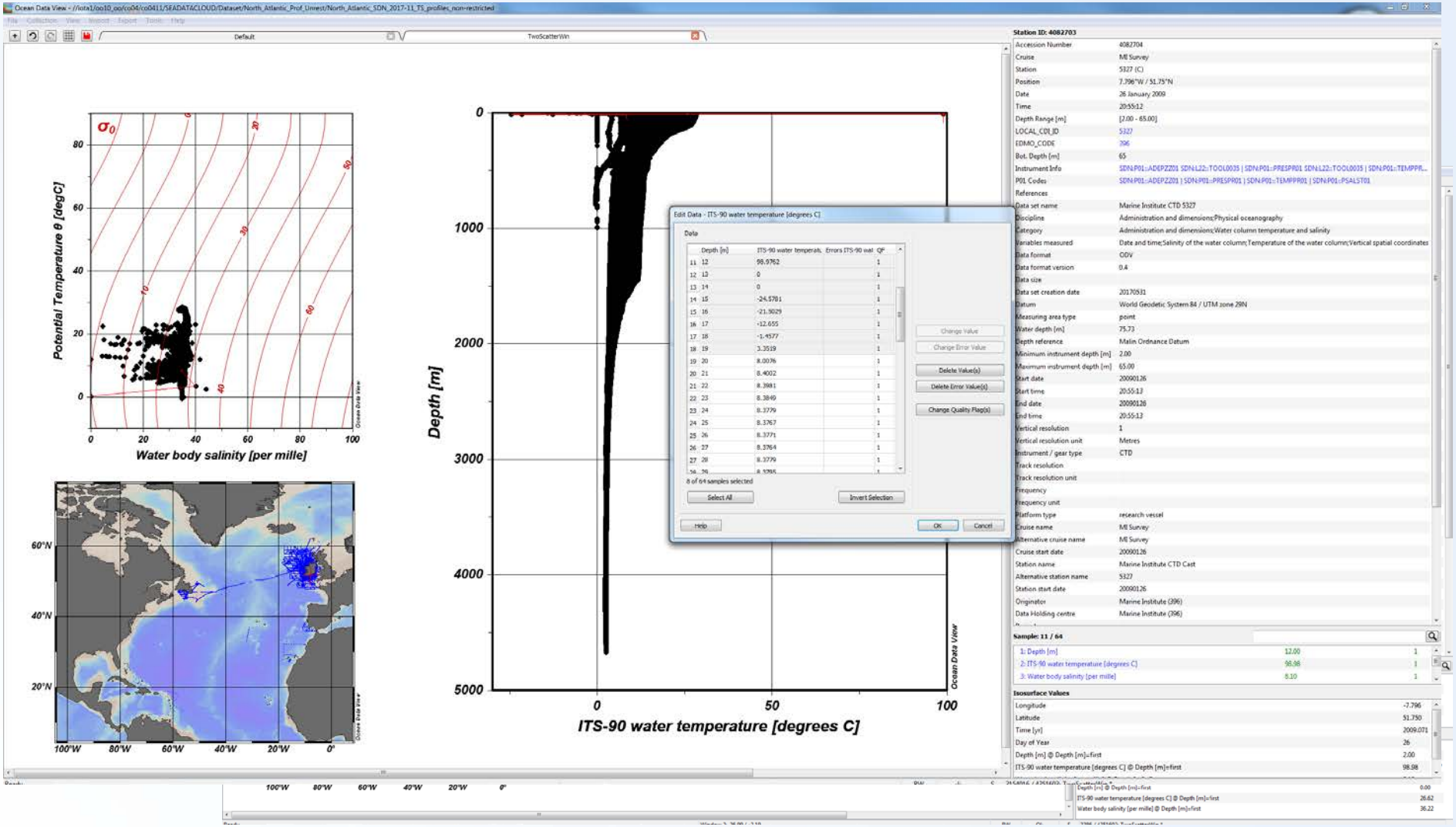

48°N
46°N
44°N
42°N
40°N
30°E
35°E
40°E
Ocean Data View


65°N
60°N
55°N
0°
10°E
20°E
30°E
40°E
Ocean Data View

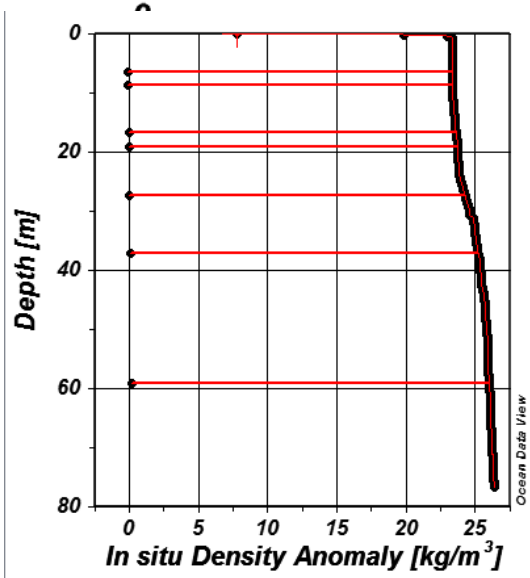

59°N
58°N
57°N
56°N
6°W
5°W
4°W
3°W
2°W
1°W
Ocean Data View

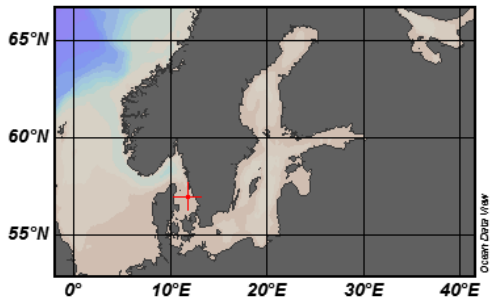
Data anomalies

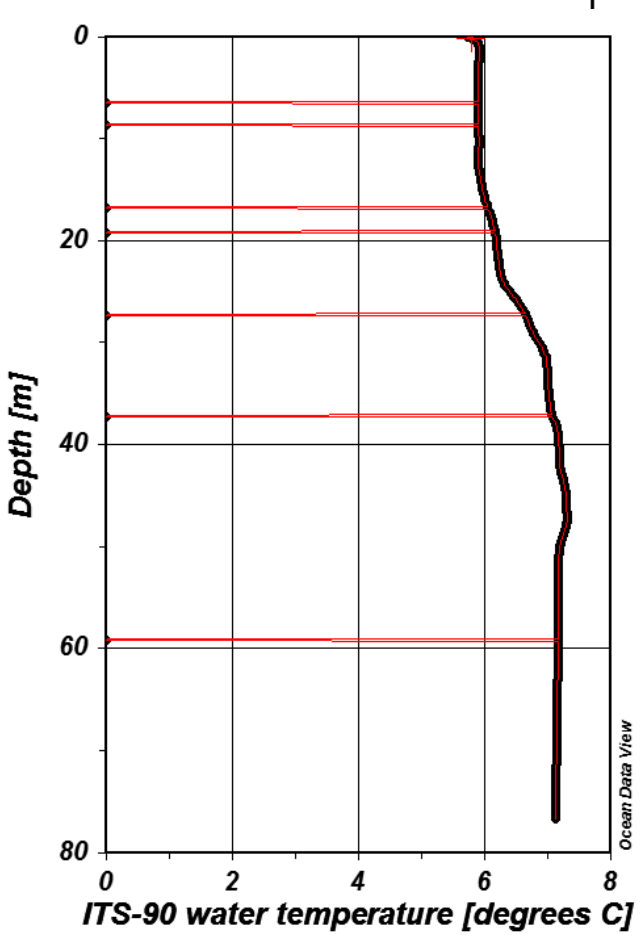
Out of range with QF1

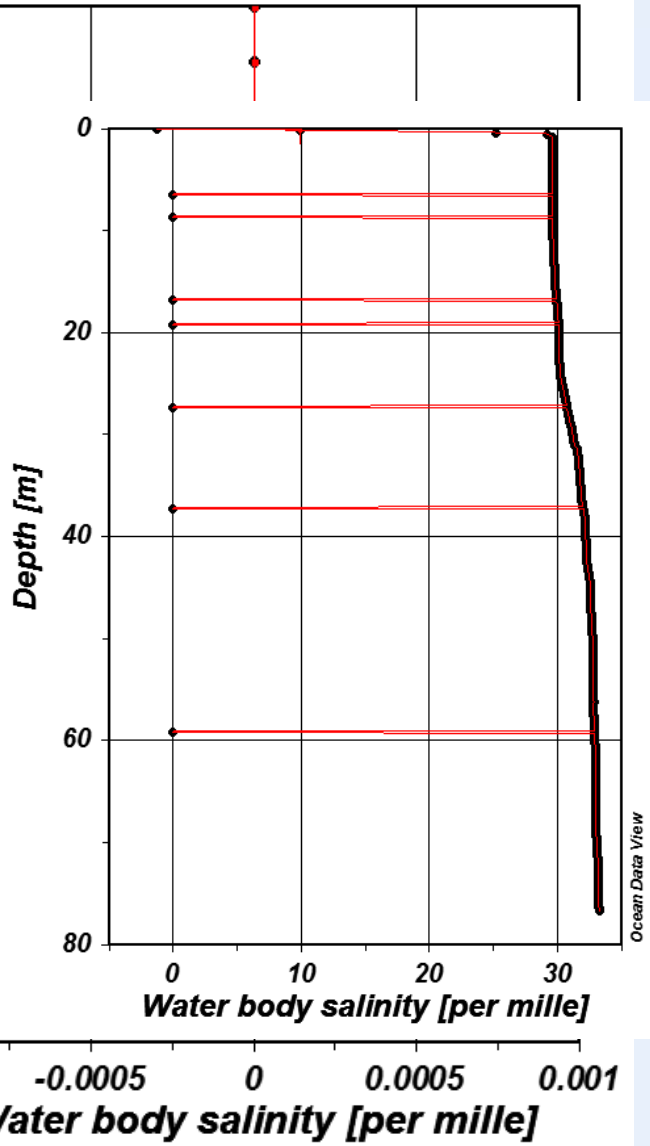


Bad values – Zero reported instead of default values

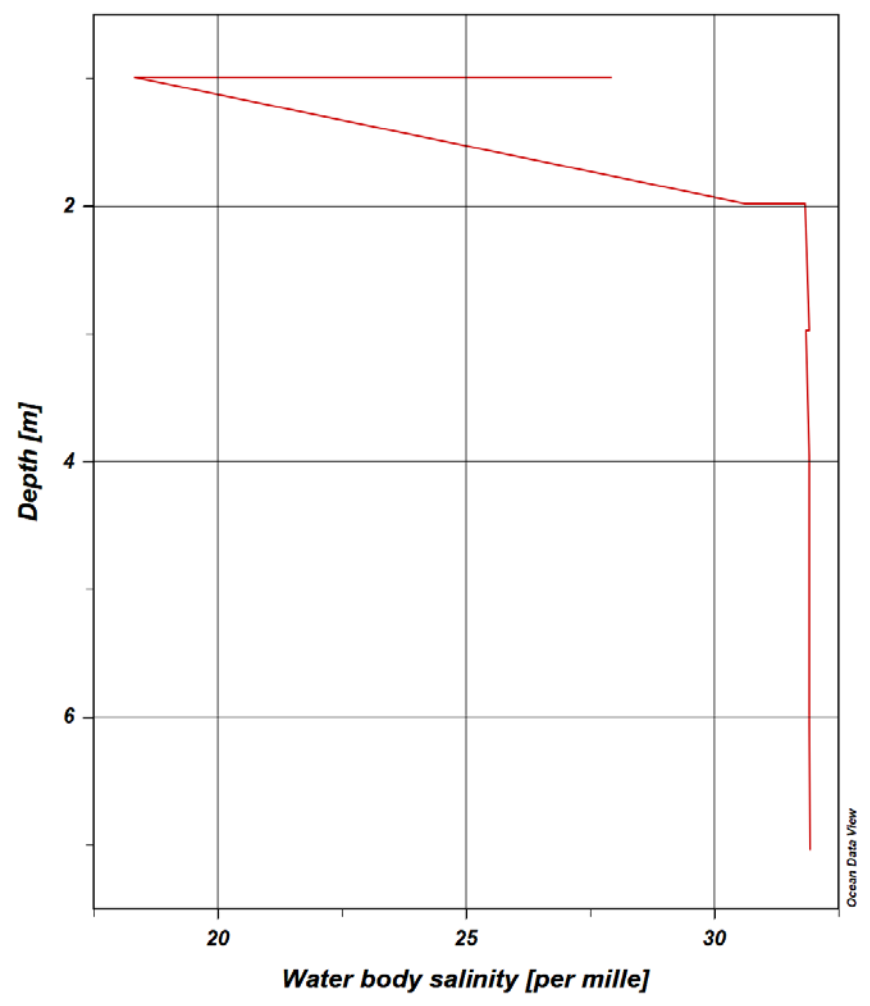




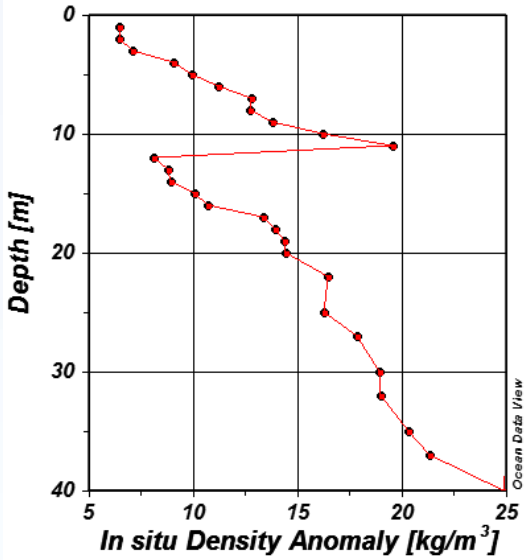


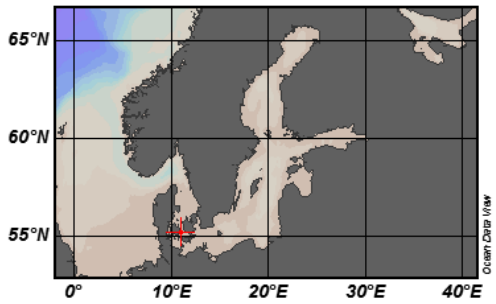


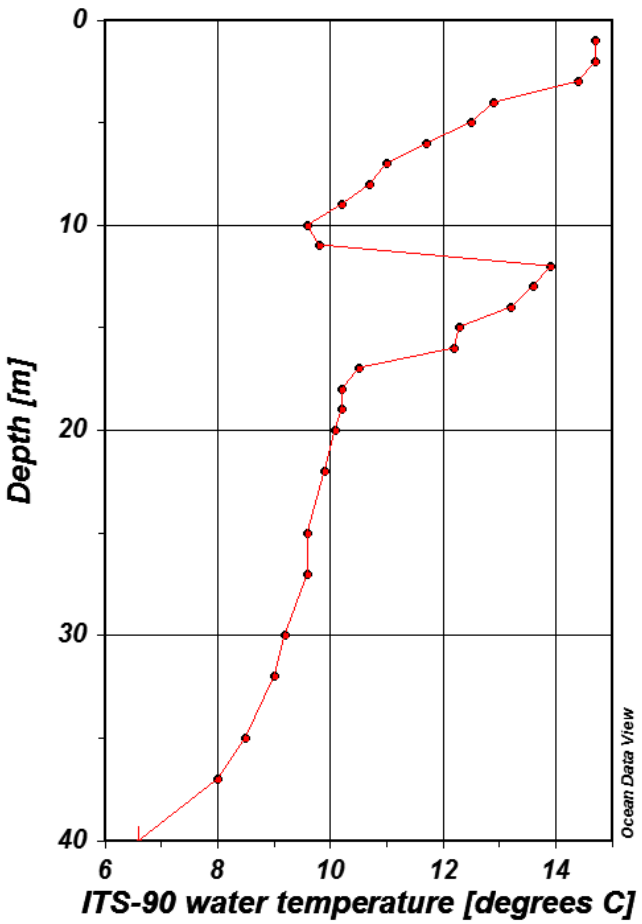
Downcasts and upcasts together
Sensor not stabilized

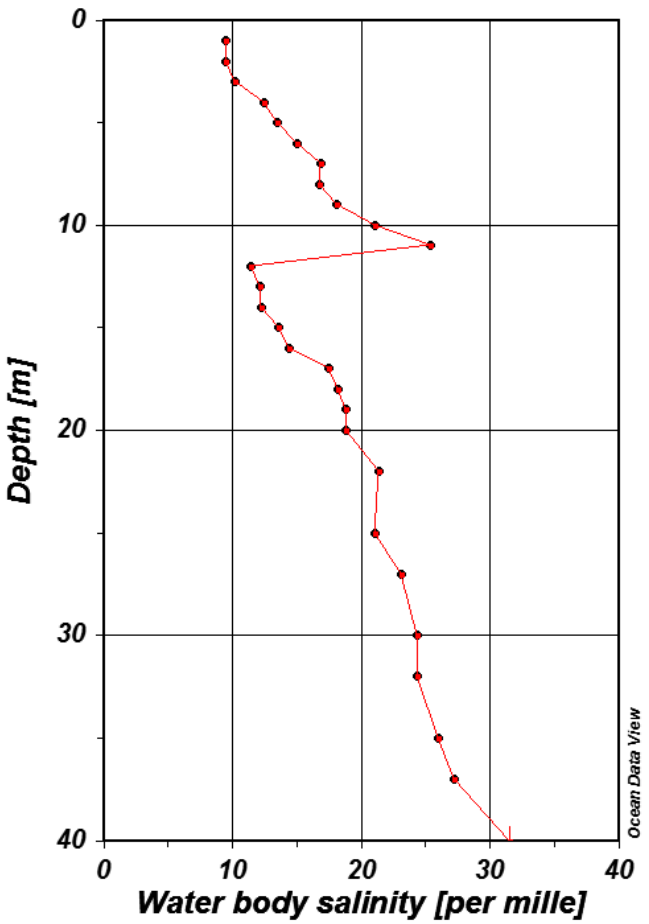


Unstable profiles

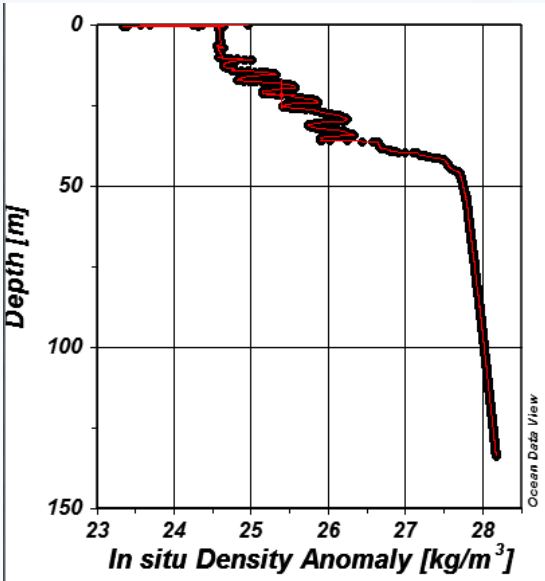


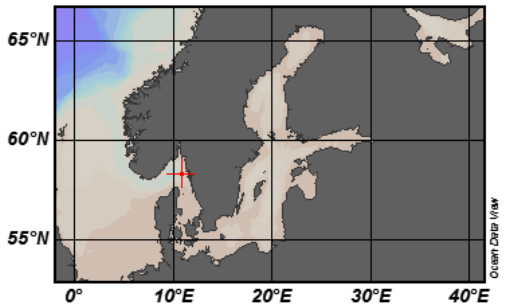


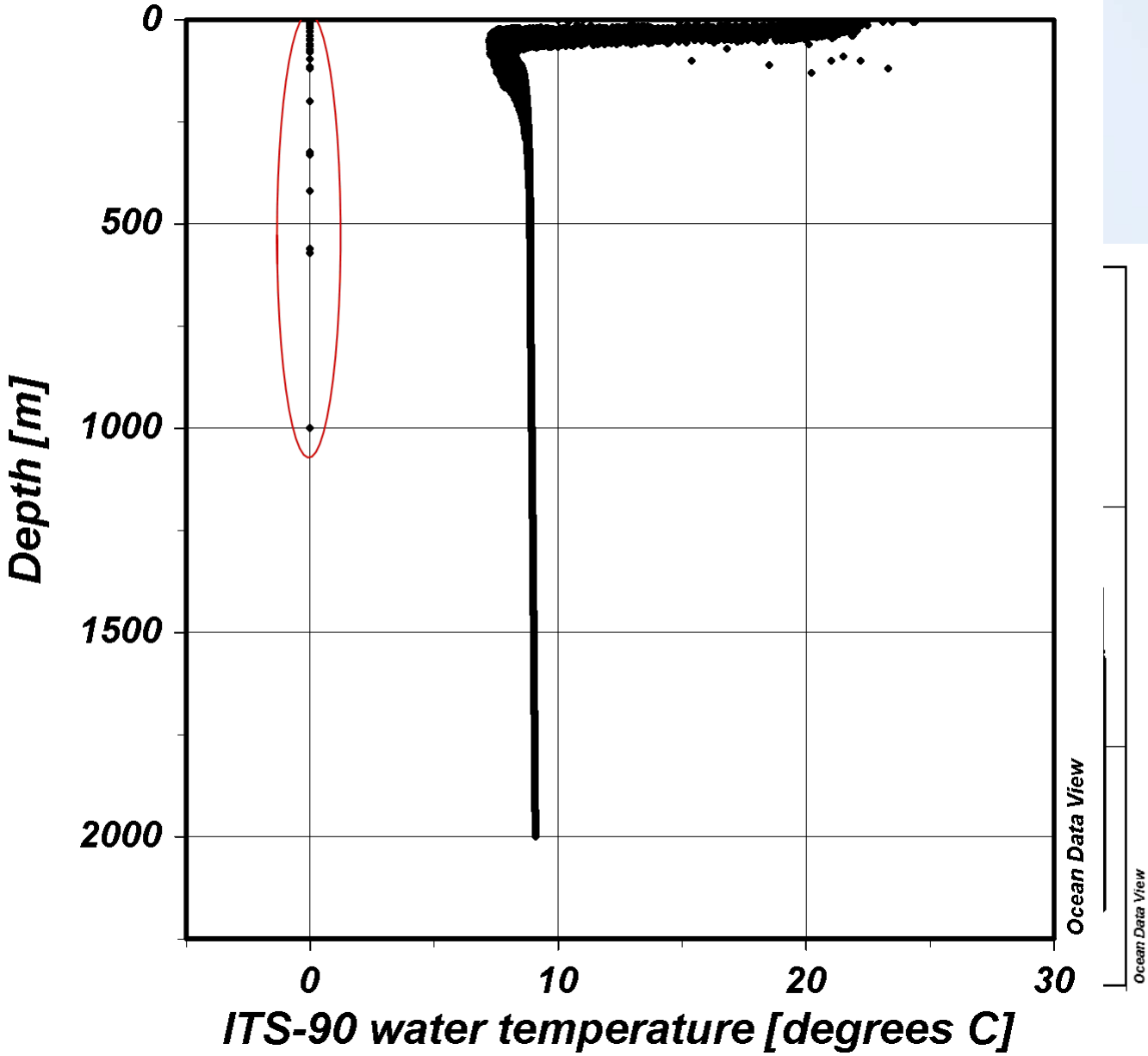




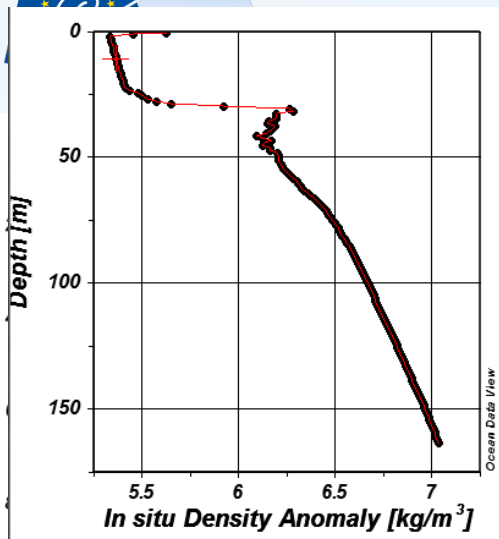
Sensor issues

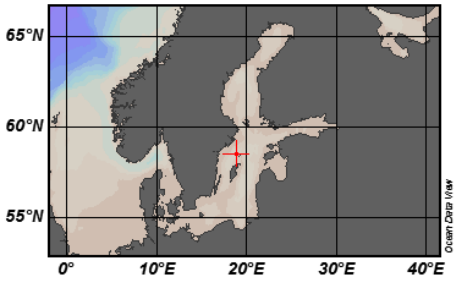


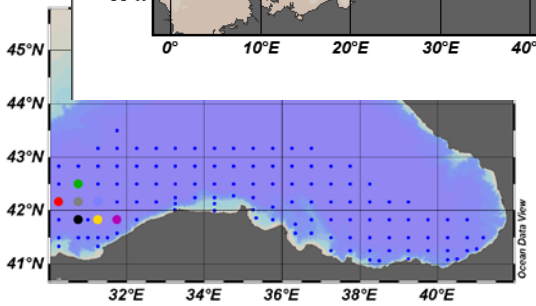


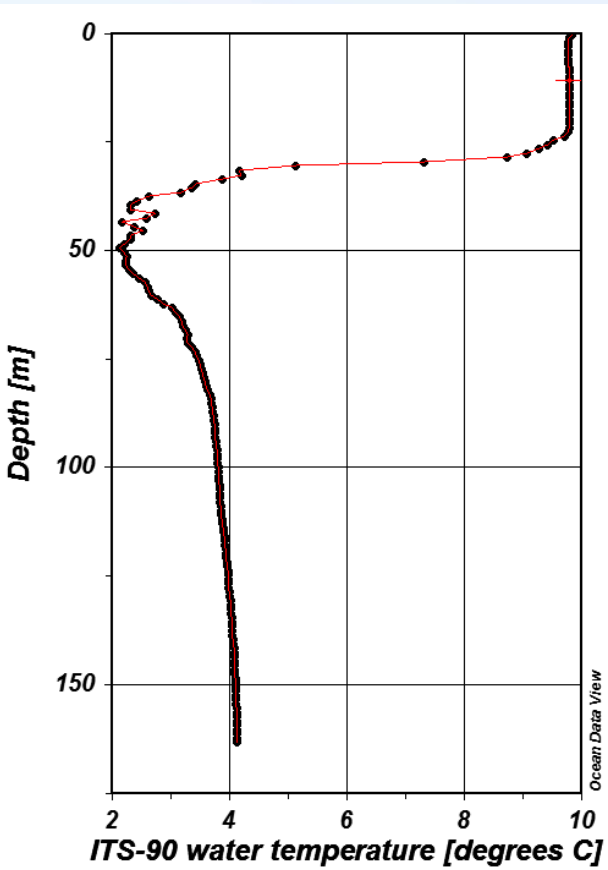


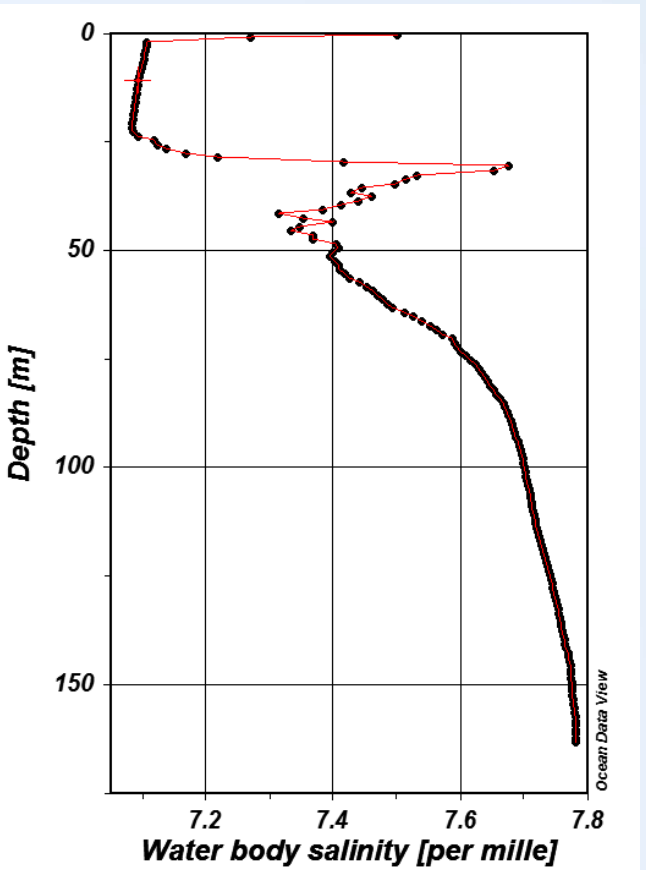
Spike (big and small)

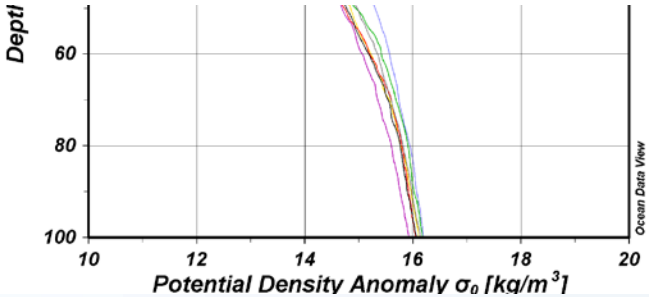


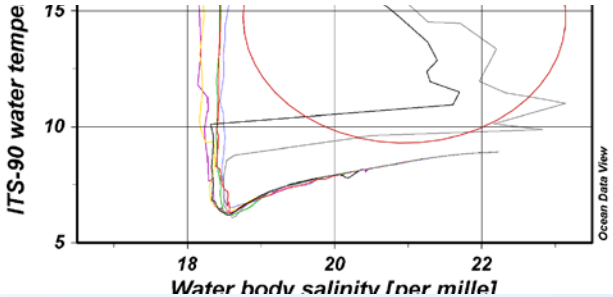






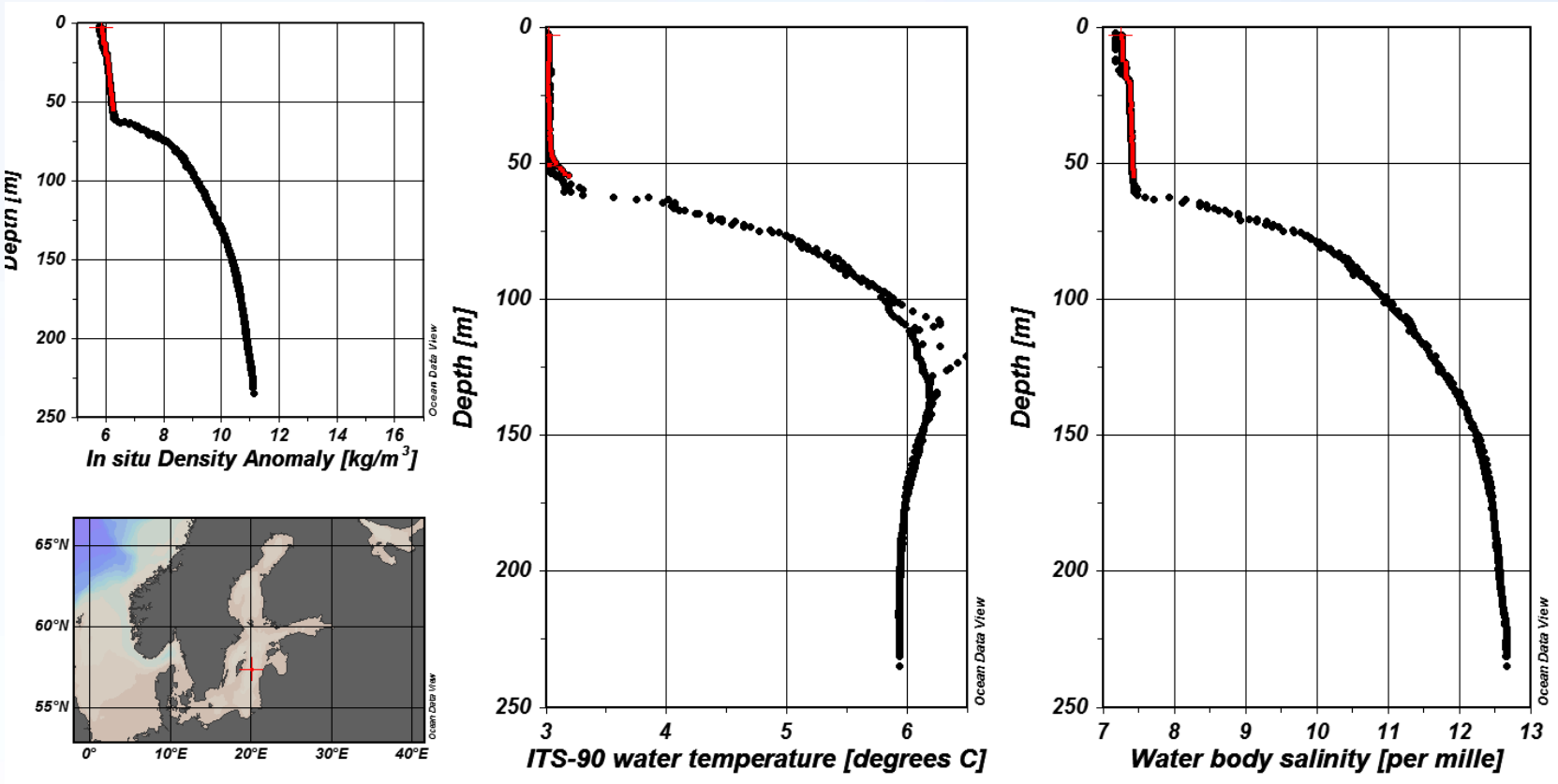






Artefact of ODV aggregation

One CDI_ID with 17 stations



SeaDataCloud

Aggregation procedure with P35

Step 1 ➔ SDN QC flags

PRES		PSAL		SSAL	
1774.0	1	35.086	1	35.087	0
1775.8	1	35.083	1	35.084	0
1777.0	1	35.081	1	35.081	0
1778.0	1	32.123	4	32.123	0

Step 2 ➔ ODV QC flags

PRES		PSAL		SSAL	
1774.0	0	35.086	0	35.087	1
1775.8	0	35.083	0	35.084	1
1777.0	0	35.081	0	35.081	1
1778.0	0	32.123	8	32.123	1

Step 3 ➔ Median value for aggregation

PRES	Salinity
1774.0	35.0865
1775.8	35.0835
1777.0	35.081
1778.0	32.123

Step 4 ➔ Keep the worst ODV flag

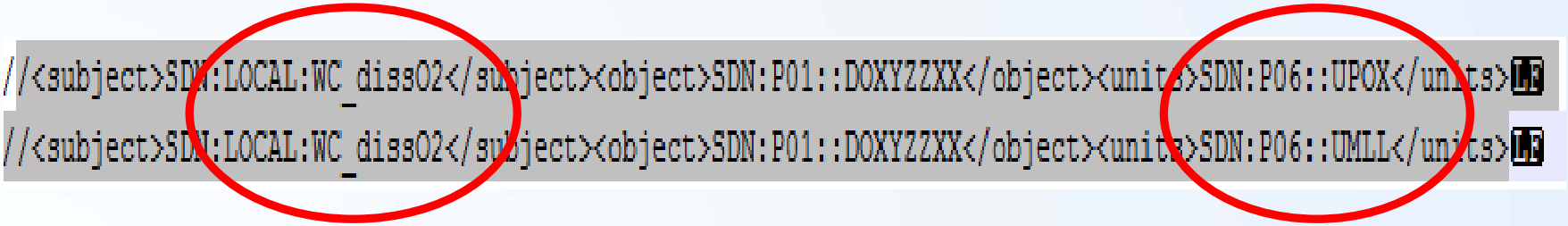
PRES		Salinity	
1774.0	0	35.0865	1
1775.8	0	35.0835	1
1777.0	0	35.081	1
1778.0	0	32.123	8

Step 5 ➔ Back to SDN flag

PRES		Salinity	
1774.0	1	35.0865	0
1775.8	1	35.0835	0
1777.0	1	35.081	0
1778.0	1	32.123	4

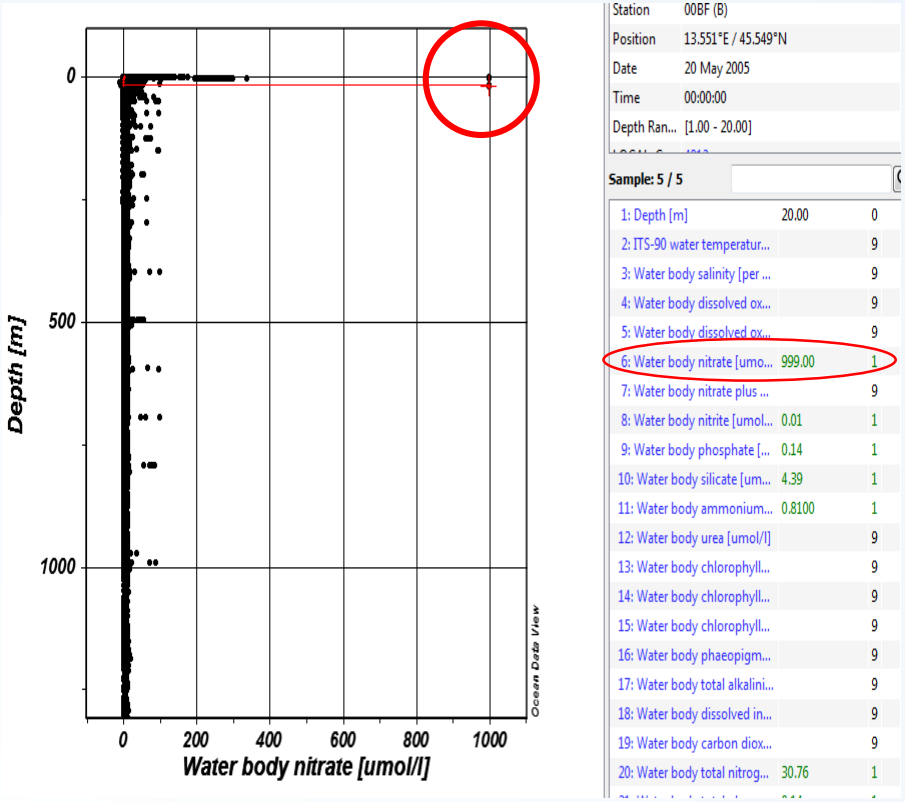
From EMODnet-Chemistry

- Two parameters with the same user label (name) and different units
  - labels were renamed



SeaDataCloud

Values 999 with flag 1



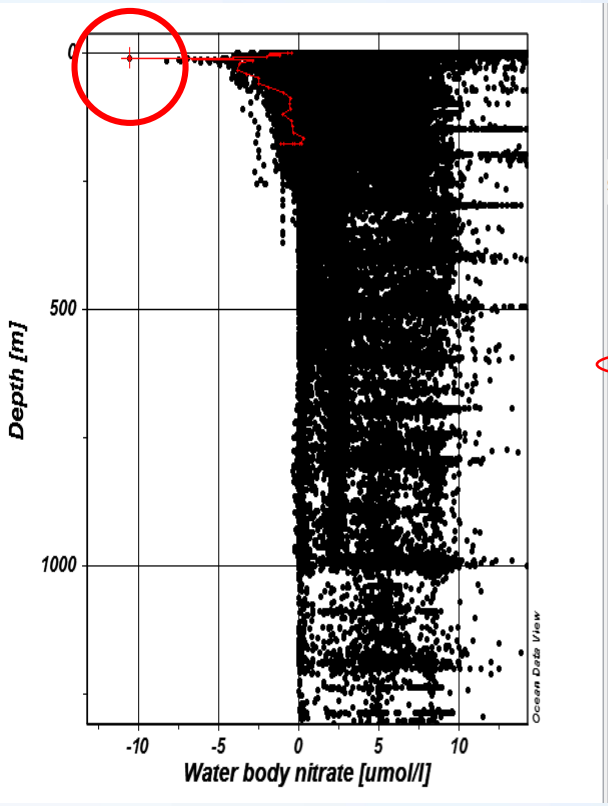
Station	00BF (B)
Position	13.551°E / 45.549°N
Date	20 May 2005
Time	00:00:00
Depth Ran...	[1.00 - 20.00]

Sample: 5 / 5

1: Depth [m]	20.00	0
2: ITS-90 water temperatur...		9
3: Water body salinity [per ...		9
4: Water body dissolved ox...		9
5: Water body dissolved ox...		9
6: Water body nitrate [umo...	999.00	1
7: Water body nitrate plus ...		9
8: Water body nitrite [umol...	0.01	1
9: Water body phosphate [...	0.14	1
10: Water body silicate [um...	4.39	1
11: Water body ammonium...	0.8100	1
12: Water body urea [umol/l]		9
13: Water body chlorophyll...		9
14: Water body chlorophyll...		9
15: Water body chlorophyll...		9
16: Water body phaeopigm...		9
17: Water body total alkalini...		9
18: Water body dissolved in...		9
19: Water body carbon diox...		9
20: Water body total nitrog...	30.76	1

Flag changed to 4

Negative values



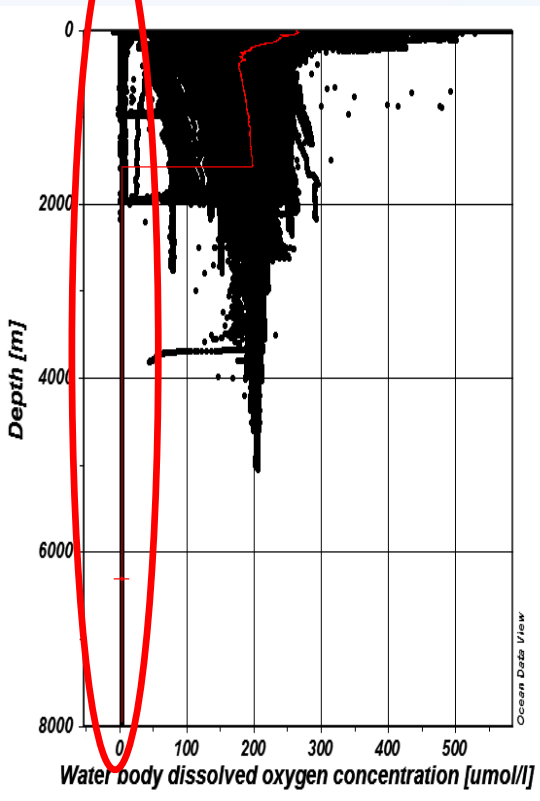
Station	00060 (B)
Position	17.07°E / 41.448°N
Date	30 August 2014
Time	11:09:00
Depth Ran...	[0.00 - 178.59]

Sample: 7 / 31

1: Depth [m]	12.01	1
2: ITS-90 water temperatur...		9
3: Water body salinity [per ...		9
4: Water body dissolved ox...		9
5: Water body dissolved ox...		9
6: Water body nitrate [umo...	-10.56	0
7: Water body nitrate plus ...		9
8: Water body nitrite [umol...		9
9: Water body phosphate [...		9
10: Water body silicate [um...		9
11: Water body ammonium...		9
12: Water body urea [umol/l]		9
13: Water body chlorophyll...		9
14: Water body chlorophyll...		9
15: Water body chlorophyll...		9
16: Water body phaeopigm...		9
17: Water body total alkalini...		9
18: Water body dissolved in...		9
19: Water body carbon diox...		9
20: Water body total nitrog...		9

Flag changed to 4

Wrong values with flag 1



Station	NB026 (C)
Position	5.476°E / 40.502°N
Date	31 March 2000
Time	08:30:31
Depth Range [...	[4.96 - 7959.87]
LOCAL_CDI_ID	NB1_NB026_H10
EDMO_code	149
Bot. Depth [m]	2710
Instrument Info	
P01 Codes	SDN:P01::PRESPR01 | SDN:P01::TEMPS601 | SDN:P0..
References	http://seadatanet.maris2.nl/v_cdi_v3/print_xml.asp..
Data set name	NORBAL 1 CRUISE CTD
Discipline	Administration and dimensions;Biological oceano...
Category	Administration and dimensions;Dissolved gases;O...

Sample: 6453 / 8167

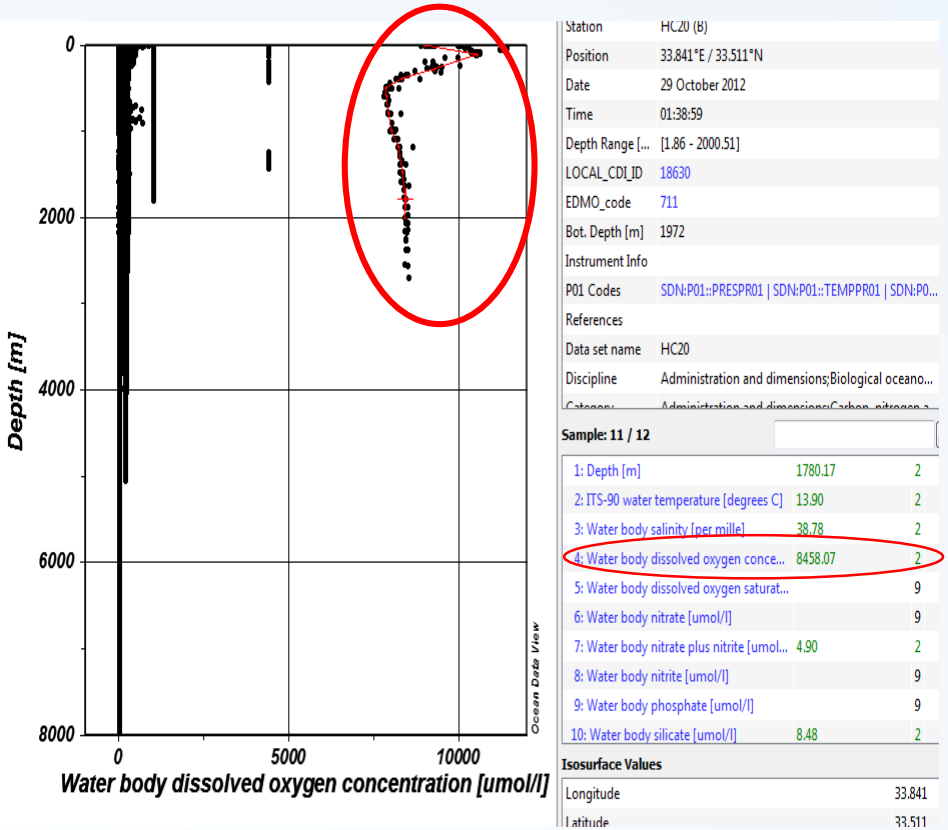
1: Depth [m]	6313.14	1
2: ITS-90 water temperature [degrees C]		9
3: Water body salinity [per mille]		9
4: Water body dissolved oxygen conce...	2.68	1
5: Water body dissolved oxygen saturat...		9
6: Water body nitrate [umol/l]		9
7: Water body nitrate plus nitrite [umol...		9
8: Water body nitrite [umol/l]		9
9: Water body phosphate [umol/l]		9
10: Water body silicate [umol/l]		9

Isosurface Values

Mediter. max depth = 5121 m

- depth flag changed to 4
- Parameter flag changed to 4

Wrong units



Station	HC20 (B)
Position	33.841°E / 33.511°N
Date	29 October 2012
Time	01:38:59
Depth Range [...	[1.86 - 2000.51]
LOCAL_CDI_ID	18630
EDMO_code	711
Bot. Depth [m]	1972
Instrument Info	
P01 Codes	SDN:P01::PRESPR01 | SDN:P01::TEMPPR01 | SDN:P0...
References	
Data set name	HC20
Discipline	Administration and dimensions;Biological oceano...

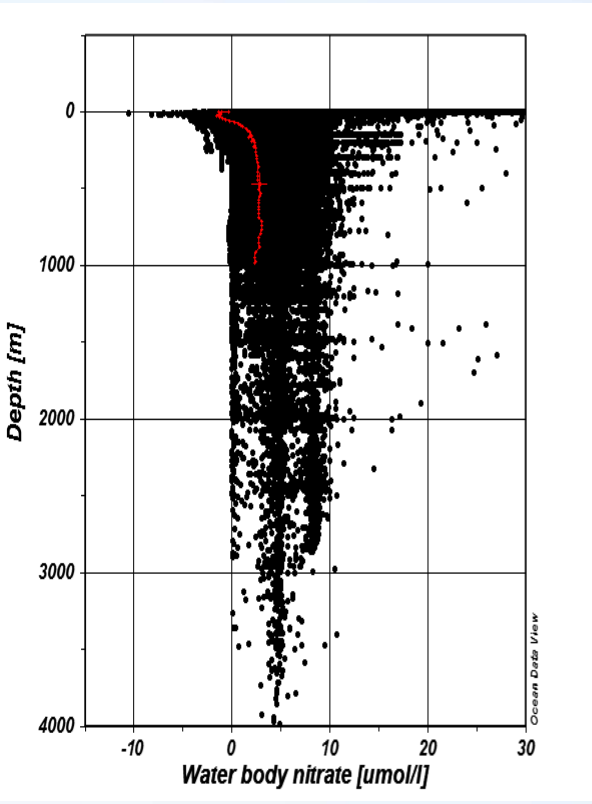
Sample: 11 / 12

1: Depth [m]	1780.17	2
2: ITS-90 water temperature [degrees C]	13.90	2
3: Water body salinity [per mille]	38.78	2
4: Water body dissolved oxygen conce...	8458.07	2
5: Water body dissolved oxygen saturat...		9
6: Water body nitrate [umol/l]		9
7: Water body nitrate plus nitrite [umol...	4.90	2
8: Water body nitrite [umol/l]		9
9: Water body phosphate [umol/l]		9
10: Water body silicate [umol/l]	8.48	2

Isosurface Values

Longitude	33.841
Latitude	33.511

No Qced data



Station	00128 (B)
Position	15.699°E / 37.215°N
Date	14 May 2015
Time	10:35:00
Depth Ran...	[0.15 - 987.43]
LOCAL_C...	429840
EDMO_code	120

Sample: 36 / 56

1: Depth [m]	472.98	1
2: ITS-90 water temperatur...		9
3: Water body salinity [per ...		9
4: Water body dissolved ox...		9
5: Water body dissolved ox...		9
6: Water body nitrate [umo...	2.87	0
7: Water body nitrate plus ...		9
8: Water body nitrite [umol...		9
9: Water body phosphate [...		9
10: Water body silicate [um...		9
11: Water body ammonium...		9
12: Water body urea [umol/l]		9
13: Water body chlorophyll...		9
14: Water body chlorophyll...		9
15: Water body chlorophyll...		9
16: Water body phaeopigm...		9
17: Water body total alkalini...		9
18: Water body dissolved in...		9
19: Water body carbon diox...		9

Flag changed to 4

Flags changed to 1

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!
- Raw CTDs
- Down- & upcasts together, non-stabilized sensors
- Missing data badly flagged
- Out of range
- artefacts generated by ODV aggregation

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, sometimes with more explanation if necessary (doc)

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

What is expected from each CDI partner ?

- The anomalies list updated with NODC comments following this table
- The list of updated CDI :

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

- A report with some informations :

- List of errors and number

QC_Action	Number of anomalies detected by MyOcean	Number of true anomalies	%
Climatology	20	0	0.0
Gradient	328	1	0.3
IncreasingPressure	544	25	4.6
RegionalRange	42	0	0.0
Spike	148	42	18.3
StuckValue	28	0	0.0
VisualInspection	1	1	100.0
Total	**1111**	**69**	**6.2**

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

Column	Description	Comment
LOCAL_CDI_ID	*cdi_identifier*	Partner local CDI identifier, Information from CDI
EDMO_CODE	EDMO_CODE of the organization distributing the data	Information from CDI
PLATFORM_CODE=CRUISE	CDI cruise_name	
STATION_DATE_START	Date at which the station starts	
STATION_DATE_STOP	Date at which the station ends	
UPDATE_DATE	Date of the control done by MyOcean partners	
PARAMETER	PARAMETER exported from ODV (TEMP, PSAL, DEPH or DEPTH [sometimes PRES when MyOcean partners have changed name])	
QC_ACTION	As described in Introduction, to define the type of anomalies (spike, gradient, missing value, etc)	
OLD_QC	QC from original dataset	
NEW_QC	QC suggested by MyOcean (see Annex I)	
VERTICAL_REFERENCE_START	Level at which starts the anomaly in the profile	
VERTICAL_REFERENCE_STOP	Level at which stops the anomaly in the profile	
AGREE WITH THE SUGGESTED CORRECTION (YES/NO)	Fill with Yes/No if you agree/disagree with the corrections suggested by MyOcean	
NODC COMMENT	Column to be added to your file in order to put some information about your our opinion about the suggested correction (agreement, disagreement, explanation if necessary)	
DETAILS	Column to be added to your file in order to put more information about suggested correction	

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

Unlock your data and set them free!

- Importance of sharing data for knowledge advancement
- Make your restricted data → unrestricted

Product Information Document

Goal: to associate to each product a **PIDoc** containing all the specifications about its:

- General characteristics (format, space-time coverage, resolution)
- Quality (validation methodology and results)
- Usability

PIDoc will have a DOI as well as the data products and both will be available through the SDC product catalogue

→ This would increase user confidence and uptake of SDC products

→ It would also provide details on how to reproduce the products in the VRE where data and tools will be available

Product Information Document

Goal: to associate to each product a **PIDoc** containing all the specifications about its:

- General characteristics (format, space-time coverage, resolution)
- Quality (validation methodology and results)
- Usability

PIDoc will have a **DOI** as well as the data products and both will be available through the SDC product catalogue

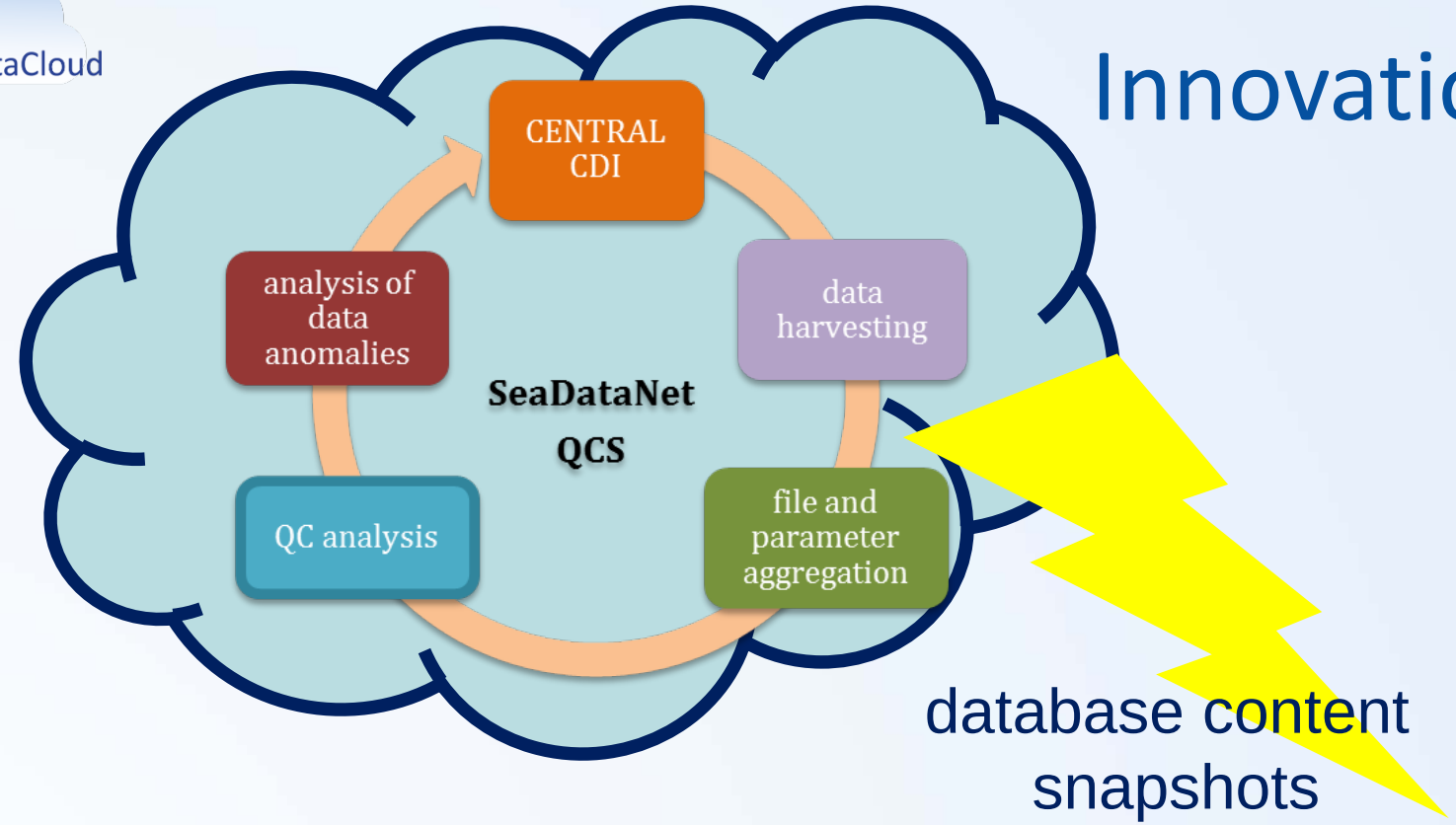
→ This would increase user confidence and uptake of SDC products

→ It would also provide details on how to reproduce the products in the VRE where data and tools will be available

Each PIDoc contains the full list of data distributors and originators

→ acknowledgment of your effort!

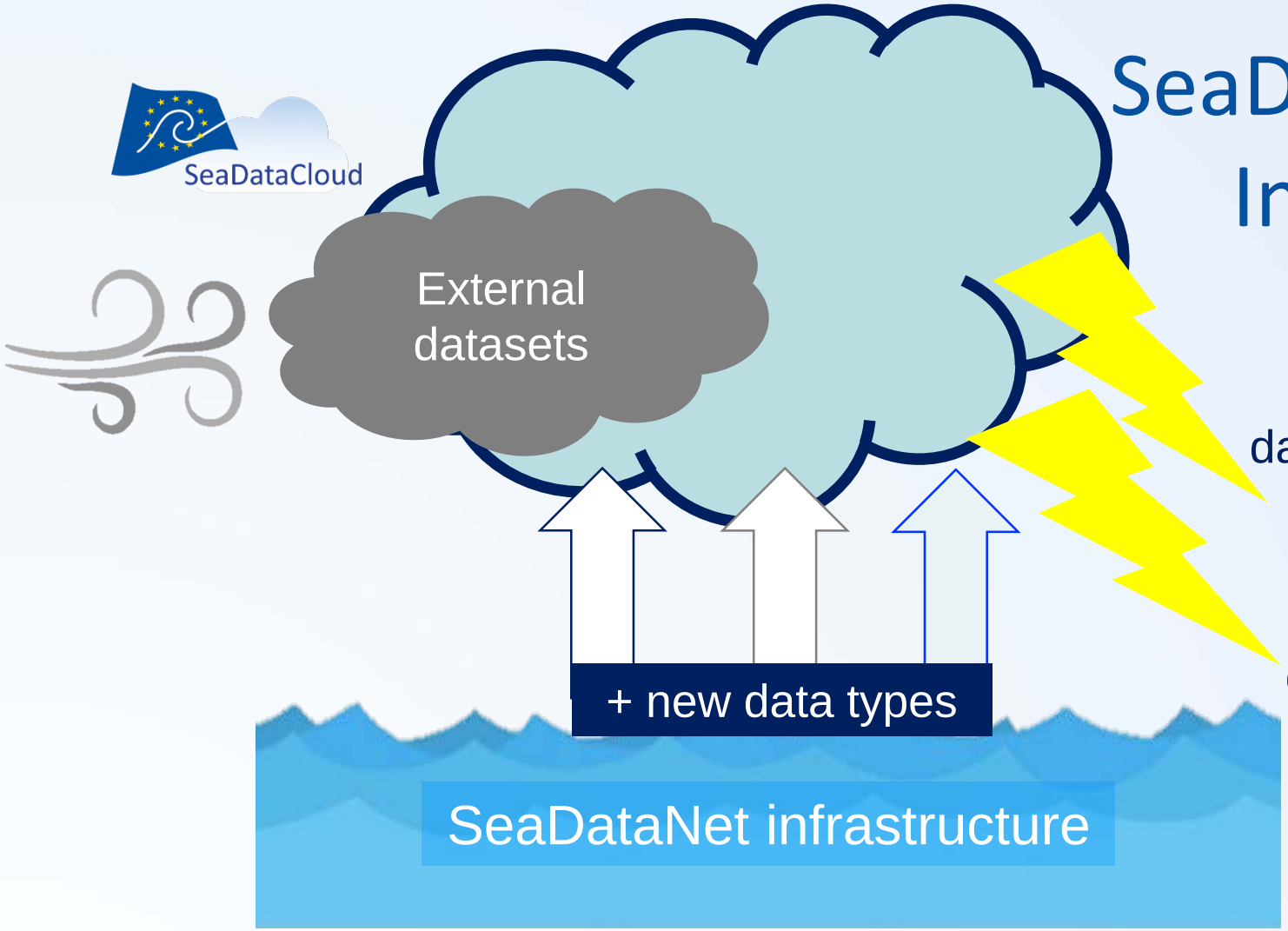
SeaDataCloud Innovation



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

Reactions?

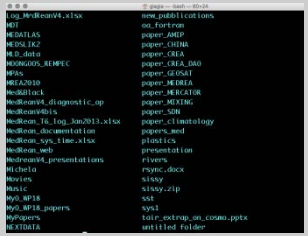
SeaDataCloud Innovation

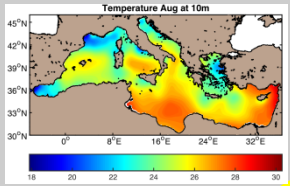


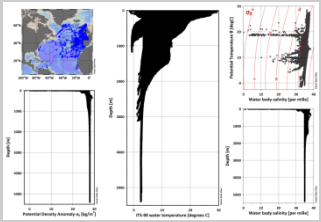
The **ingestion of new data types** (HF radar, glider data) and the **integration of external data sets** are fundamental actions for the creation of appropriate observational data products as demanded by the user community

Virtual Research Environment

SeaDataCloud







New Data Products

data subsets

Standard Data Products
historical data collections
climatologies

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

Virtual Research Environment

- The positive impact of VRE on data providers (i.e. automatic management of data anomalies)

- See following slide (to get a message each time a QF is defined as doubtful or bad, correct and charge automatically the corrected data)

How to improve QCS ?
to an automated way

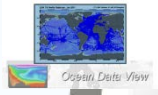
SeaDataCloud

Central CDI catalogue

Data Harvesting

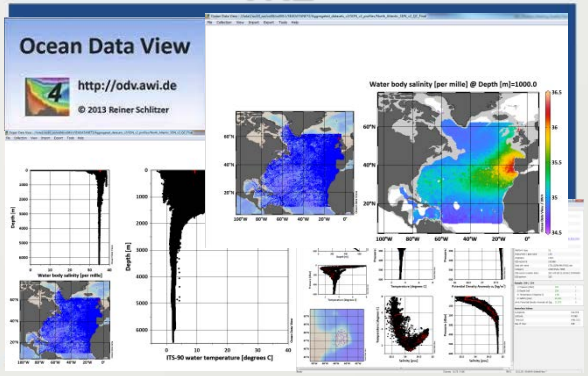
AWI

File and Parameter Aggregation



2nd release : possibility to get only new, updated profiles ? List of removed profiles

VRE



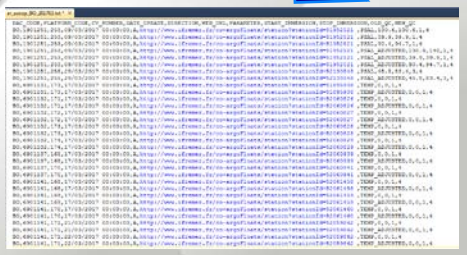
Work on a unique subset

DATA

Message for NODC, in the VRE message window ?

CDI EDMO_CODE ...

CDI EDMO_CODE ...



LIST OF STATIONS WITH CORRECTION ON QC

CONCLUSIONS

SeaDataCloud work plan on data products is very ambitious and our success is dependent from **data availability** and **technical developments** related to the cloud virtual research environment

- More data → highest product quality and increased knowledge
- VRE will allow a fastest access to the data and the tools that will be shared