TEMPERATURE AND SALINITY HISTORICAL DATA COLLECTIONS for the European marginal seas: AGGREGATION and QUALITY ASSESSMENT procedures

S.Simoncelli *(INGV), C. Coatanoan (Ifremer) , O. Bäck (SMHI), H. Sagen (IMR), S. Scory (MUMM), Devrim Tezcan (METU), Dick M.A. Schaap (MARIS), Reiner Schlitzer (AWI), Sissy Iona (HCMR), Michèle Fichaut (Ifremer), Marina Tonani (INGV)*

SeaDataNet

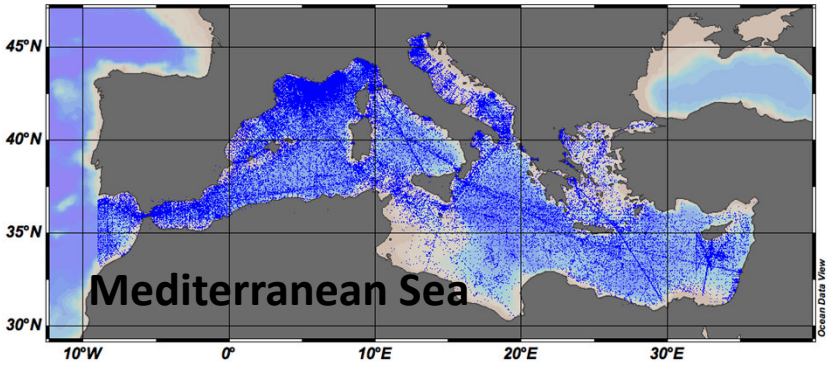
IMDIS Conference
Lucca, Monday 23 September 2013

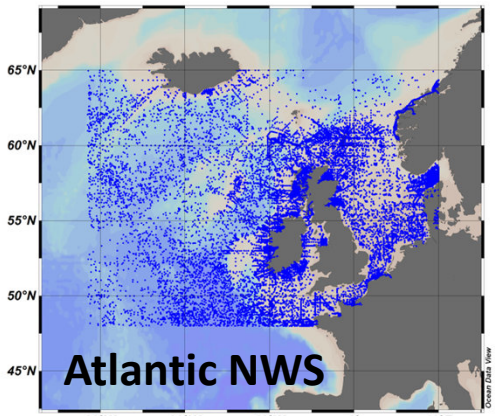
OUTLINE

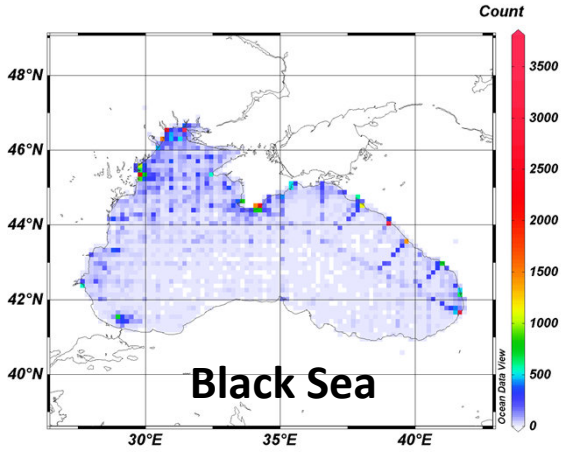
- Introduction
- Duplicates Issue (HCMR)
- Data harvesting of T&S files by CDI Robot (MARIS)
- Building SDN Aggregated Datasets (AWI)
- Quality Control (QC) analysis by Regional Coordinators (WP10)
- SDN-MyOcean In situ TAC feedback on QC
- Results
- Conclusions

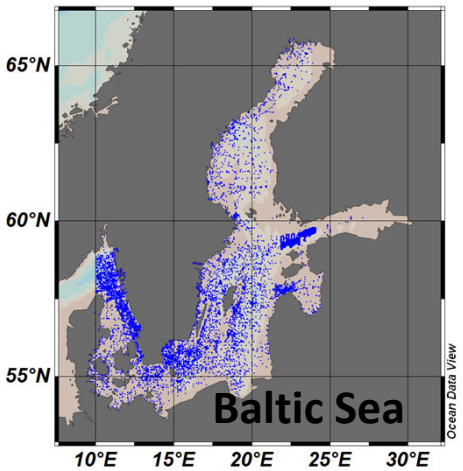
INTRODUCTION

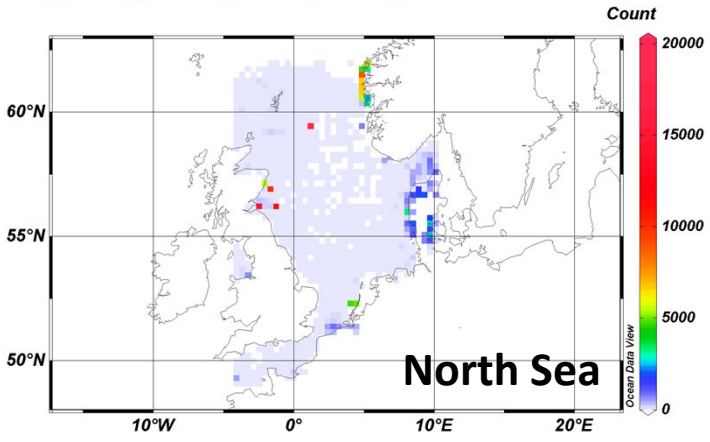
Temperature and Salinity (TS) historical data collections from 1900 were created for each European marginal sea within the Framework of SeaDataNet2 (SDN) EU-Project.







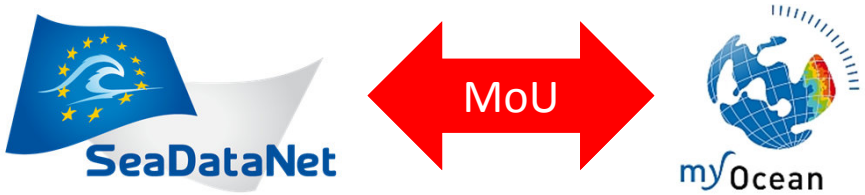




INTRODUCTION

- **Motivation**: TS collections of data were created to meet *operational oceanography* and *climate change community* requirements that need longer and longer time series of in situ observations to study long term ocean phenomena and their implications in the surrounding environment.

- **SDN-MyO collaboration**: this work has been developed in synergy with MyOcean In-Situ Thematic Assemble Centre (INS-TAC) to support and promote monitoring, modeling or downstream service development.



TIME SCHEDULE and ACTIONS

SeaDataNet

myOcean

SDN2 **MyO2**

Dupplicate implementation Plan	Oct2012
Data Harvesting and Aggregation	Dec2012
RCs received TS collections: basic QC	Jan2013

Feb2013
Release 1990-2010 aggregated data sets to MyO →

Apr2013
2° SDN-MyO Joint Meeting on QC

May2013 **← QC feedback to SDN RCs**

Organization of QC feedback to NODCs	Jun2013
QC Guidelines for NODCs	Jul2013
NODCs QC actions dateline	15 Sept2013

Dec2013
Release of TS Historical Data Colletions

Duplicates Implementation Plan

SeaDataNet

- An **implementation plan** (based on ODV duplicates checks was prepared (Oct2012) asking to data providers to:

✓ identify duplicates

✓ clean the data (delete, update, replace, etc)

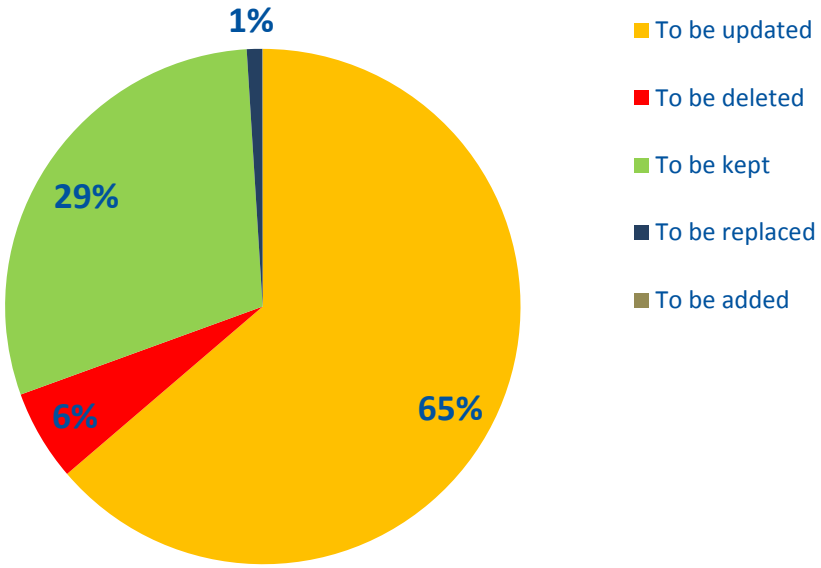
✓ explain their actions in details

- After evaluation of all data modifications both the CDI (Common Data Index) central catalogue and local archives were updated
- **Guidelines** were sent to all partners to avoid similar cases in the future
- **White list** of the cleaned and checked CDIs → new entries in the CDI central catalogue will be checked against this list to avoid future duplicates

Duplicates Implementation Plan: Results

Number of duplicates and actions undertaken:

21 Partners	Potential duplicates	To be updated	To be deleted	To be kept	To be replaced	To be added
Total	60866	38793	3475	17989	596	13



Potential duplicates

- **6%** were real duplicates
- **65%** needing correction
- **29%** not duplicates

Data harvesting of T&S files by CDI Robot

- **MARIS developed a robot user that uses the CDI Data Discovery and Access Service to query, shop and retrieve data sets from the SDN distributed data centres in automatic way**
- Query: search for all data sets with T&S, whose access is unrestricted or falling under SeaDataNet License → ~860.000 CDIs
- Robot was triggered to start harvesting the related ODV files from the distributed data centres through the general CDI shopping mechanism (RSM–DM)
- This procedure was essential to test/tune the performance of the RSM–DM process and find the optimum data requests management procedure

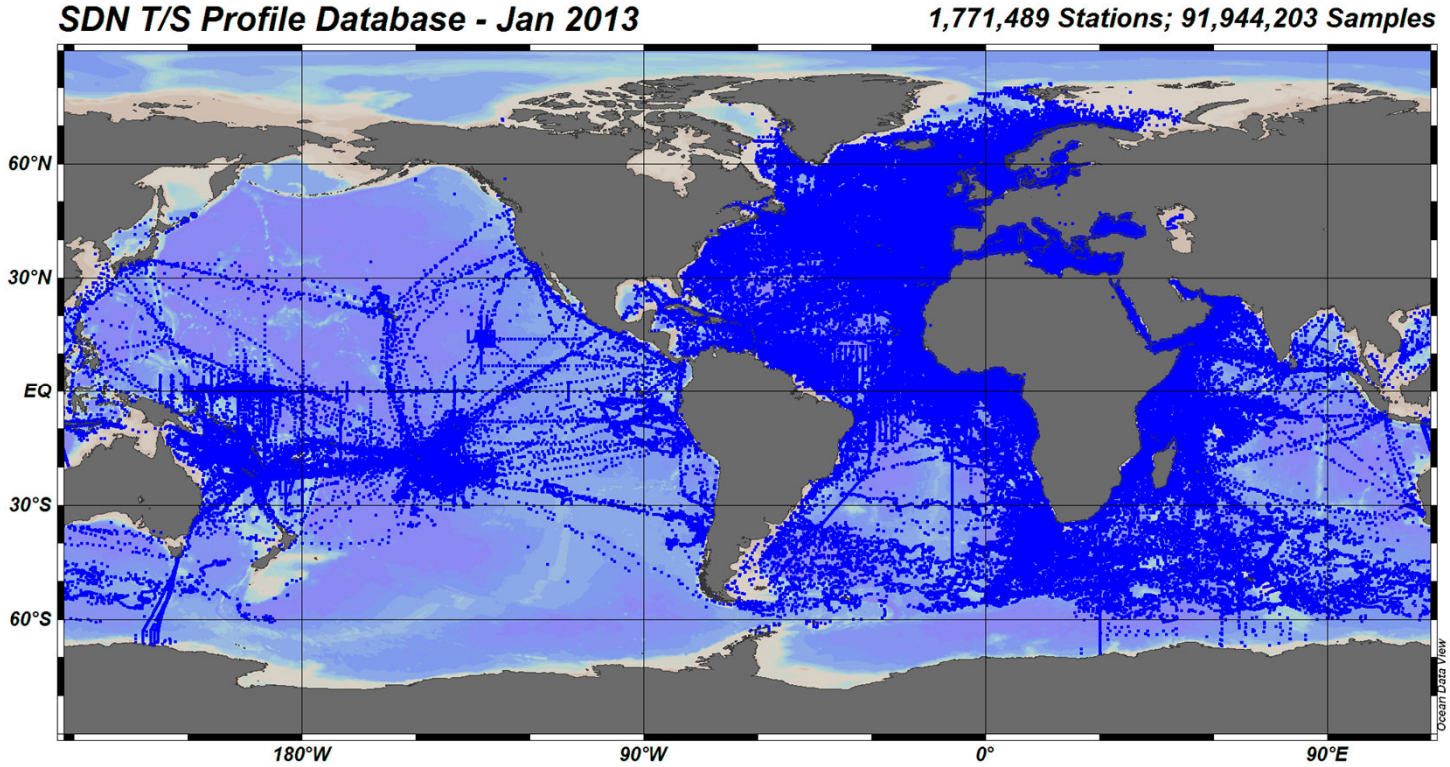
Data harvesting of T&S files by CDI Robot

- RSM (Request Status Manager) is fault proof → it keeps track of all data requests and repeats them in case of disturbances at DM (Download Manager) level
- Robot harvesting and tuning of the shopping system took into account also the **duplicates issue**
- A DVD was delivered to AWI with all the ODV files in a storage structure with the full CDI metadata as CSV file
- ODV files contained in most cases not only T&S but also additional observations

Building SDN Aggregated Datasets (AWI)

- **> 2 Mio SDN data files** in ODV format and metadata file containing CDI information were aggregated
- Aggregation of all data files into a single TS Data Collection was done using **SDN Importer of ODV 4.5.3**. This has been done in 9 pieces of about 250,000 files each, then combined
- original temperature and salinity variables were aggregated into single T and S variables using *Aggregated Derived Variables* function
- Logs files were sent to the coordinator and data centers for fixing problems
- **Regional sub-sets of TS data from 1900 to 2012** were createrd and distributed to SDN Regional Coordinators (RCs) for QC (Quality Control)

Correction of ODV files



More than 14000 files were rejected because ODV was not standard or not SDN standard. The list of errors was sent to 33 data centres → most of the data have been corrected

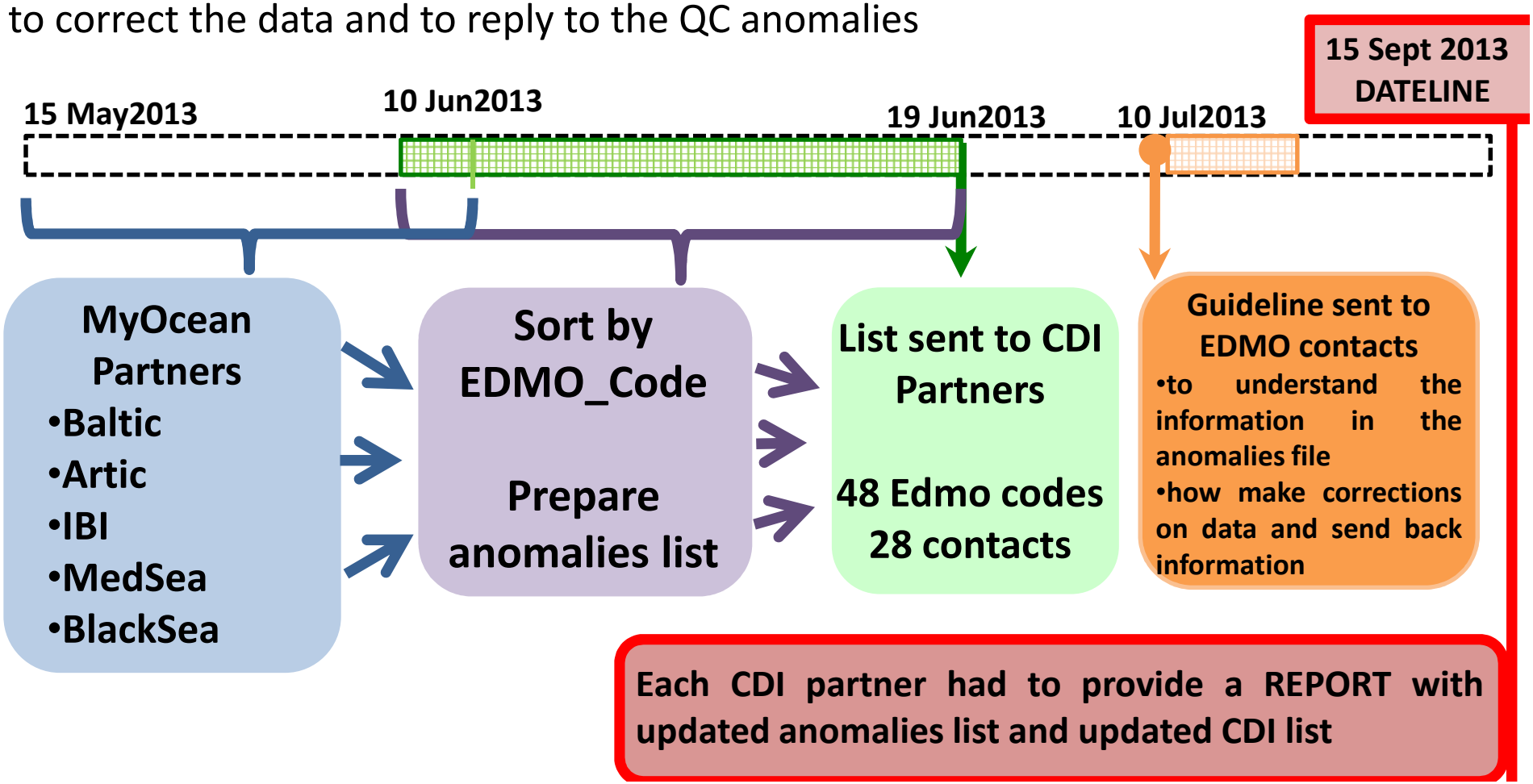
Quality Control Process

- **Jan2013** RCs (SDN WP10) received the TS collections
- **Guidelines for a first basic QC analysis in ODV and a common template for the QC report have been defined**
- **Feb2013** 1990-2012 sub sets of data have been extracted and released to MyO In-situ TAC in order to collaborate on the QC process
- **Mar2013** reports on the entire data collection have been prepared and presented to SDN StComm
- **Apr2013** 2nd SDN-MyO Joint Meeting on data quality assessment
- **May2013** MyO sent to SDN feedback on the quality of regional TS collections

Exchanges SDN-MyO:

Jun2013: results from SDN and MyO QC analysis were collected, organized and sent to the NODCs

Jul-Aug2013: NODCs were asked to check QC data anomalies, to take actions in order to correct the data and to reply to the QC anomalies



QC analysis by RCs

QC analysis considered all Qflags in order to identify anomalies and possible solutions and results were included in short reports

1.Data distribution and data density map

2.Histograms of annual and seasonal data distribution

3.TS scatter plots of the entire dataset→ highlighted the necessity of applying a gross range check

4.TS scatter plot after the **range check**

5.TS scatter plots of: Qflags=1(good),2(probably good) and Qflags=0(no check)

6.Statistics about Qflags

7.Visual control of scatter-plots to identify wrong profiles (outliers)

8.Visual check of spikes

9.Identification of stations falling on land, of wrong or missing data

Outliers have been saved in text files in order to report to both MyO and the NODCs

SDN QC Results

- Some data center probably inverted Qflag 1 and 0 → obs flagged as good (1,2) presented values out of range while most of the obs flagged 0 were reasonable

RCs did not modify any data or Qflag but they defined procedures and priority actions to report on the quality assesment to data providers in order to facilitate the update process and to progressively improve the quality of the infrastructure.

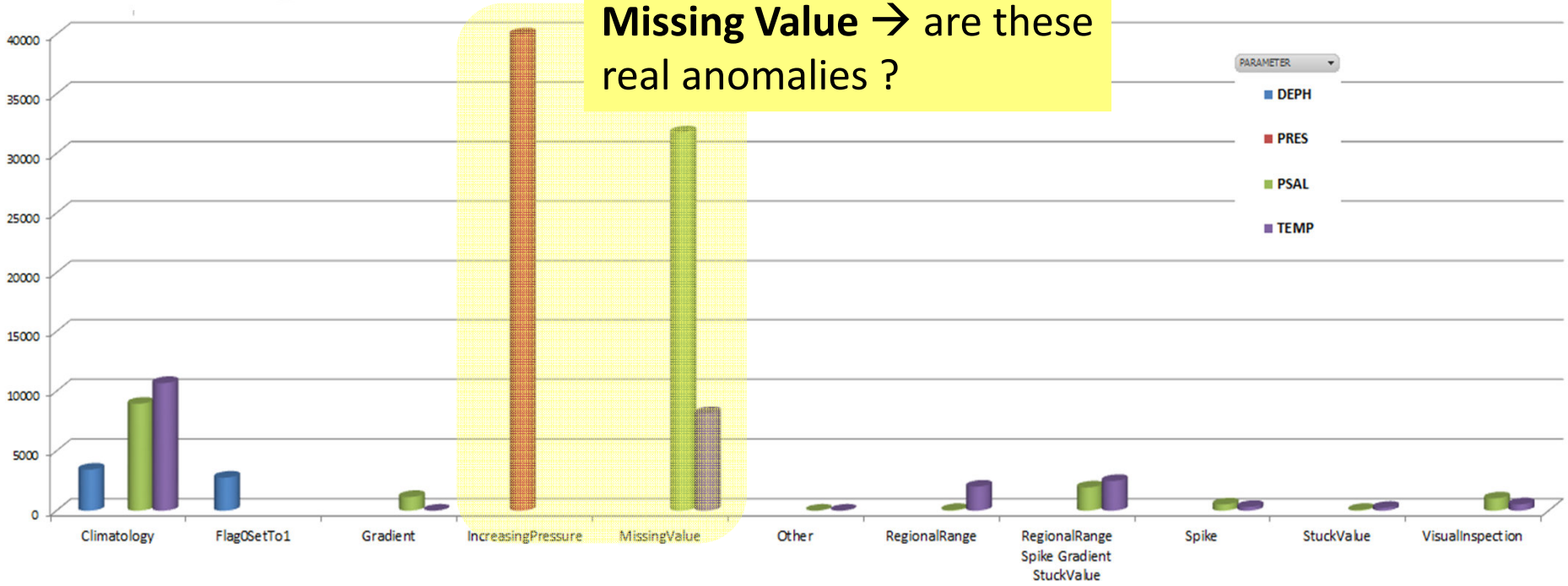
RCs nominated a responsible person to coordinate the comunication between NODC – RCs - MyO INSTAC→ *Christine Coatanon (Ifremer)*

RCs improved the common strategy for future QC analysis:

- sub-regional QC (per areas & per depth)
- stability check on density

MyO Type of anomalies

Increasing Pressure and **Missing Value** → are these real anomalies ?



Climatology: out of the climatology envelope

Gradient: gradient anomaly

Increasing pressure: non increasing depth

Missing Value: missing value with a Qflag≠9

Regional Range: value out of regional range

Spike: spike

Stuck Value: constant profile

Visual Inspection: finding an anomaly visually

CONCLUSIONS

- Data aggregation and quality assessment procedure was an extensive and fruitful exercise involving many people. It allowed to ameliorate and refine each technical phase of the procedure, but mainly to highly improve the quality of SDN infrastructure content.
- The collaboration between SDN and MyOcean was crucial during the data quality assessment and allowed to identify and correct a lot of data anomalies.

Future Work

The quality of historical data collections would be highly improved by an update before the official release thus we decided to repeat the aggregation procedure harvesting all data in the CDI and repeat the QC assessment.

A new aggregation exercise would allow to

1.to deliver the best aggregated data sets which mirror the true infrastructure content (otherwise the TS collection would have a lower quality than in reality)

2.the future SDN products based (climatologies) on these data collections will be higher quality

Historical TS collection of data will be delivered in December 2013

Thanks to

C. Coatanoan (Ifremer)

O. Bäck (SMHI)

H. Sagen (IMR)

S. Scory (MUMM)

Devrim Tezcan(METU)

Dick M.A. Shaap (MARIS)

Reiner Schlitzer (AWI)

Sissy Iona (HCMR)

Michèle Fichaut (Ifremer)

Marina Tonani (INGV)