Integration and duplication remove of the marine and meteorology data in NMDIS

LIU Yulong
National Marine Data & Information Service of CHINA
Oct. 04, 2016, TOKYO

OUTLINE

1、About us

2、Quality Control System and Software

3、Eliminate Duplicates of Ocean Data

4、Future work

1、About us

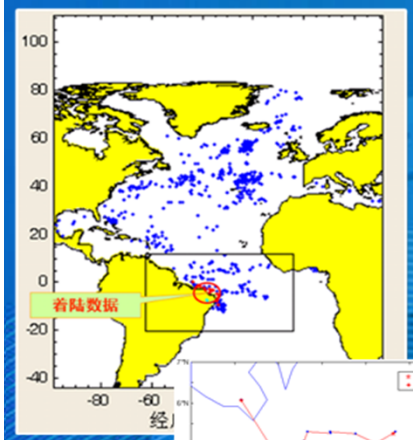
- MDC（marine data center）of NMDIS responsibility;
- National operational ocean observation data source;
- operational procedures for receiving, collecting and processing data; the techniques and methodologies of data processing and application R&D;
- R&D the methodologies and procedures generating statistic and analysis products;
- Integrate marine-meteorological and oceanographic data, actively improve method about Quality Control and Eliminate Duplicates, produce specialized datasets.
- Undertake the international data exchange; establish, maintain and operate the China ocean databases;

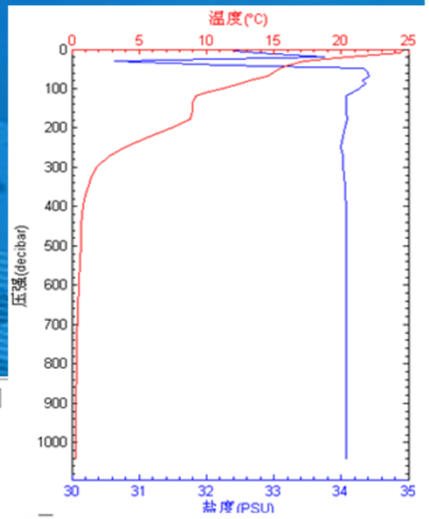
2、Quality Control System and Software

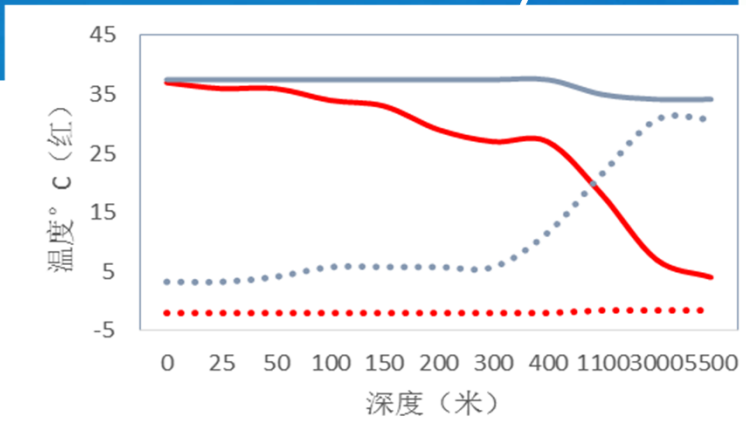
2.1 Quality control of temperature and salinity data

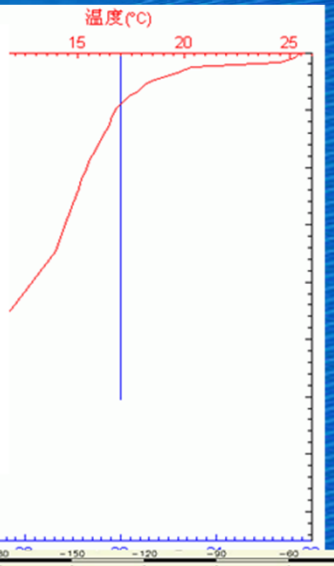
time and space basic test, profile test, climatic characteristics test, profile consistency test , Visual and manual examination. Among them, QC should be specialized by oceanography elements and observation instruments and way

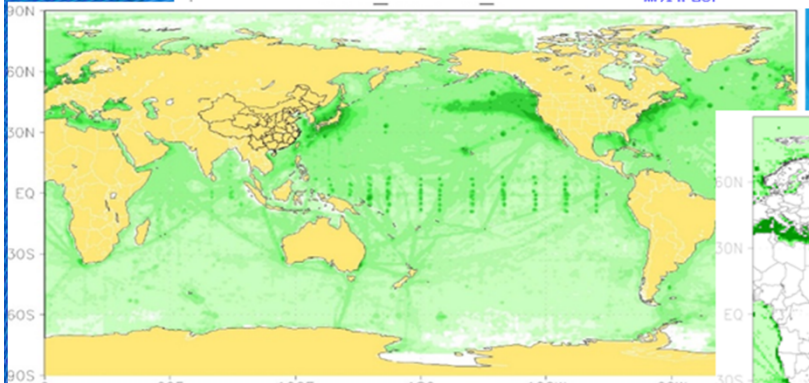
2.1 Quality control of temperature and salinity data

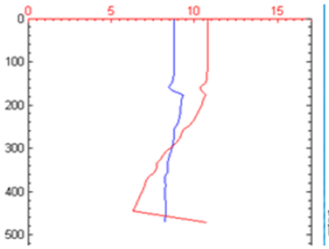


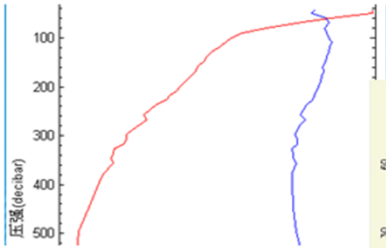


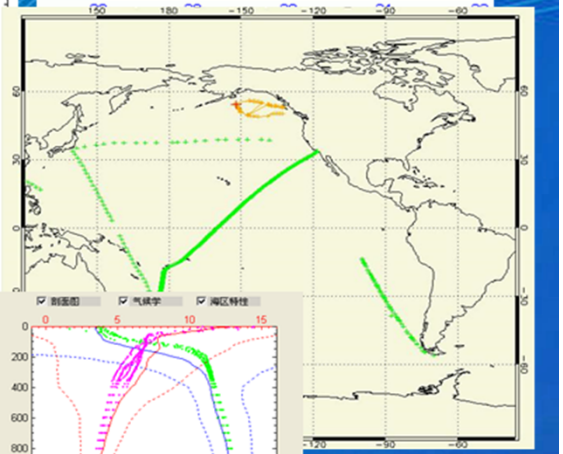


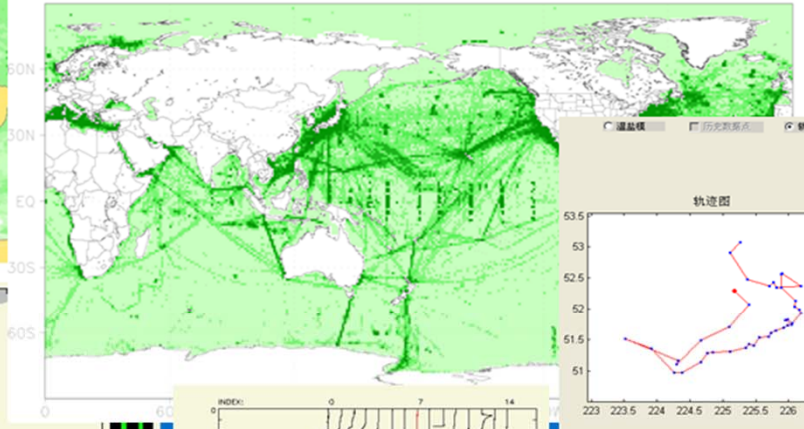


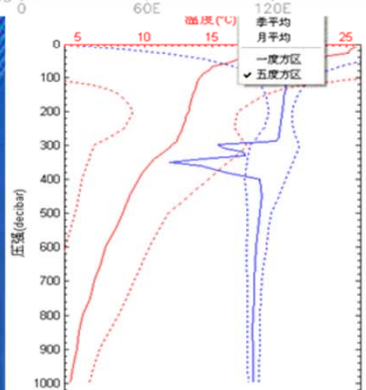


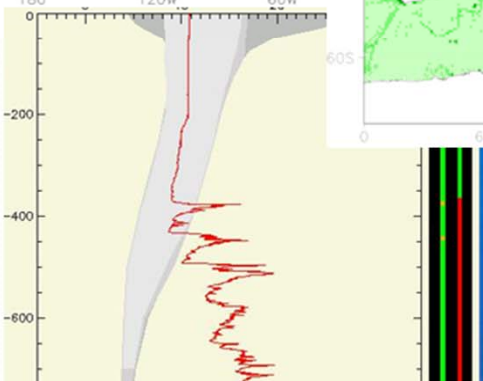


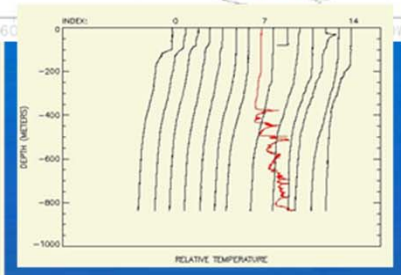


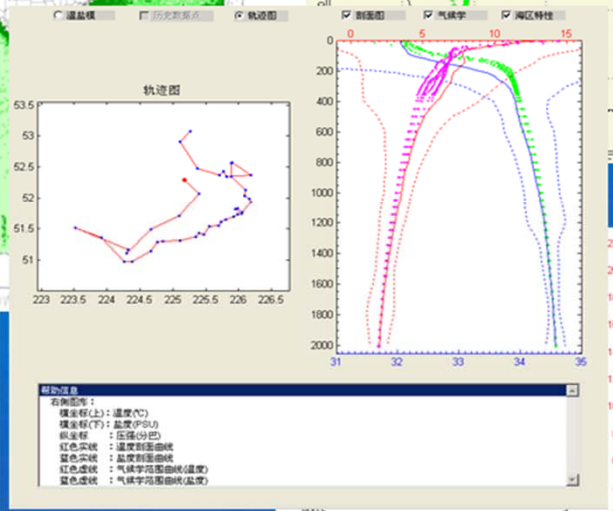


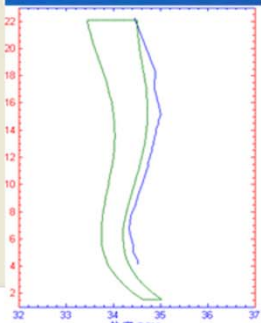






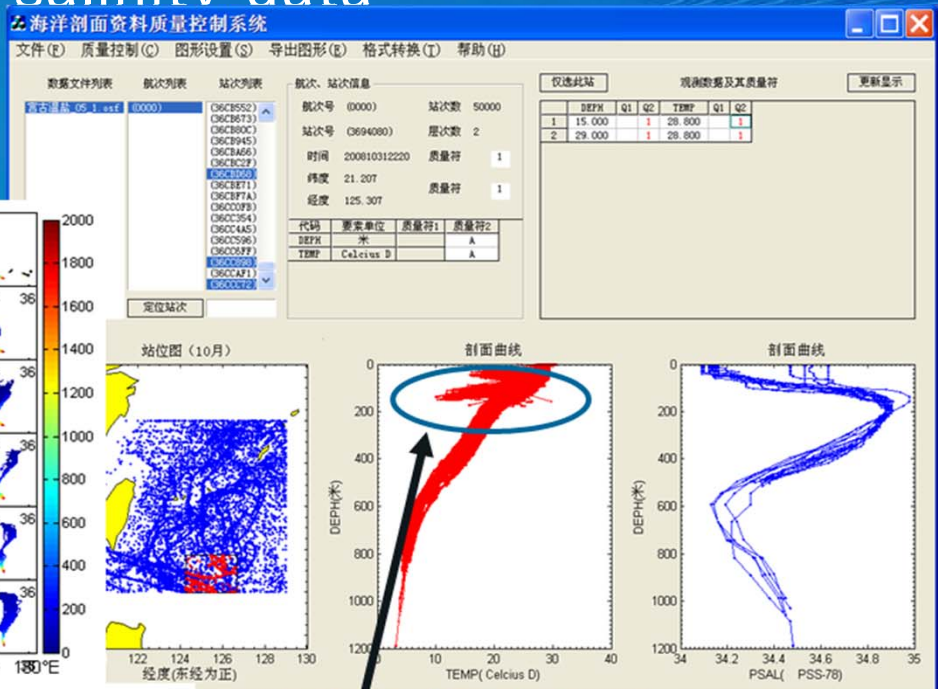






2.1 Quality control of temperature and salinity data

Improve QC and parameters in North-West Pacific



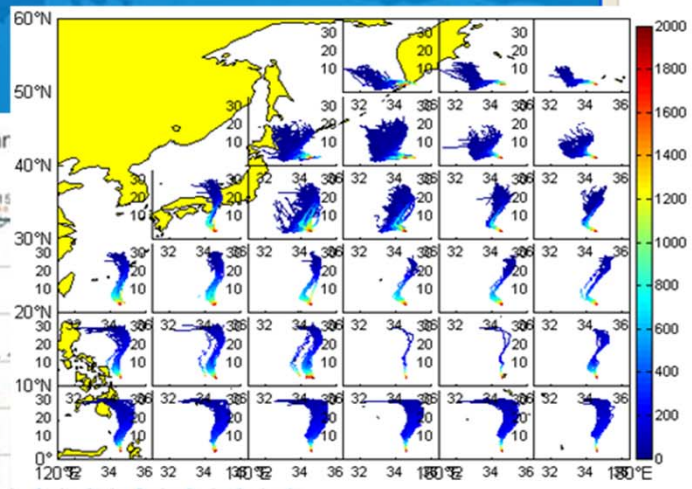
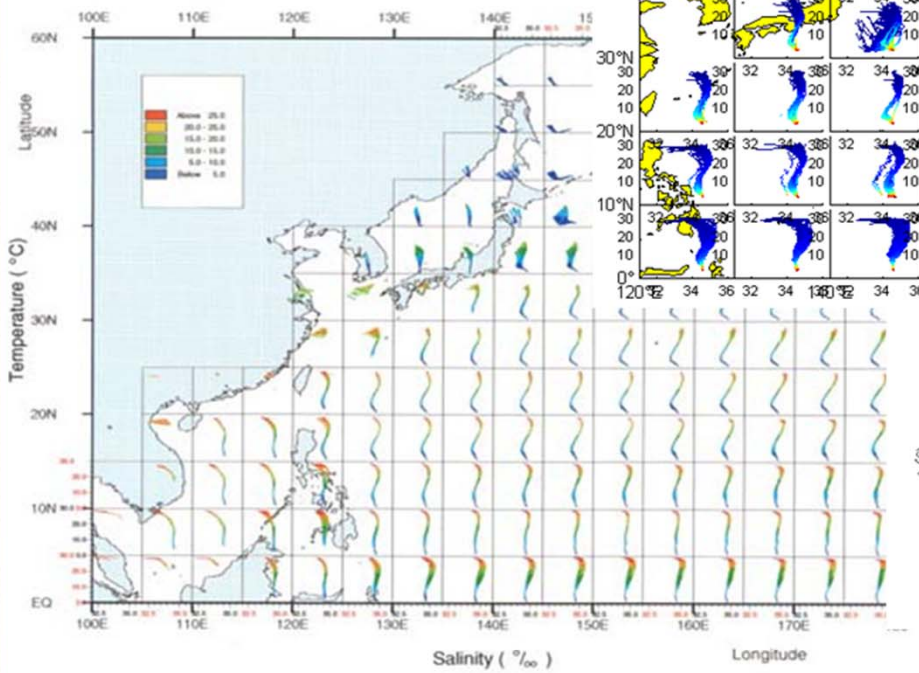
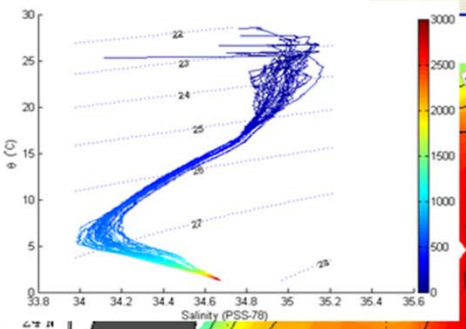


Figure 05f. North-West Pacific Ocean Anr





Objective Analysis Method for QC

- ‘Abnormal data ‘ or ‘real phenomena’ ?

2.2 Data Processing

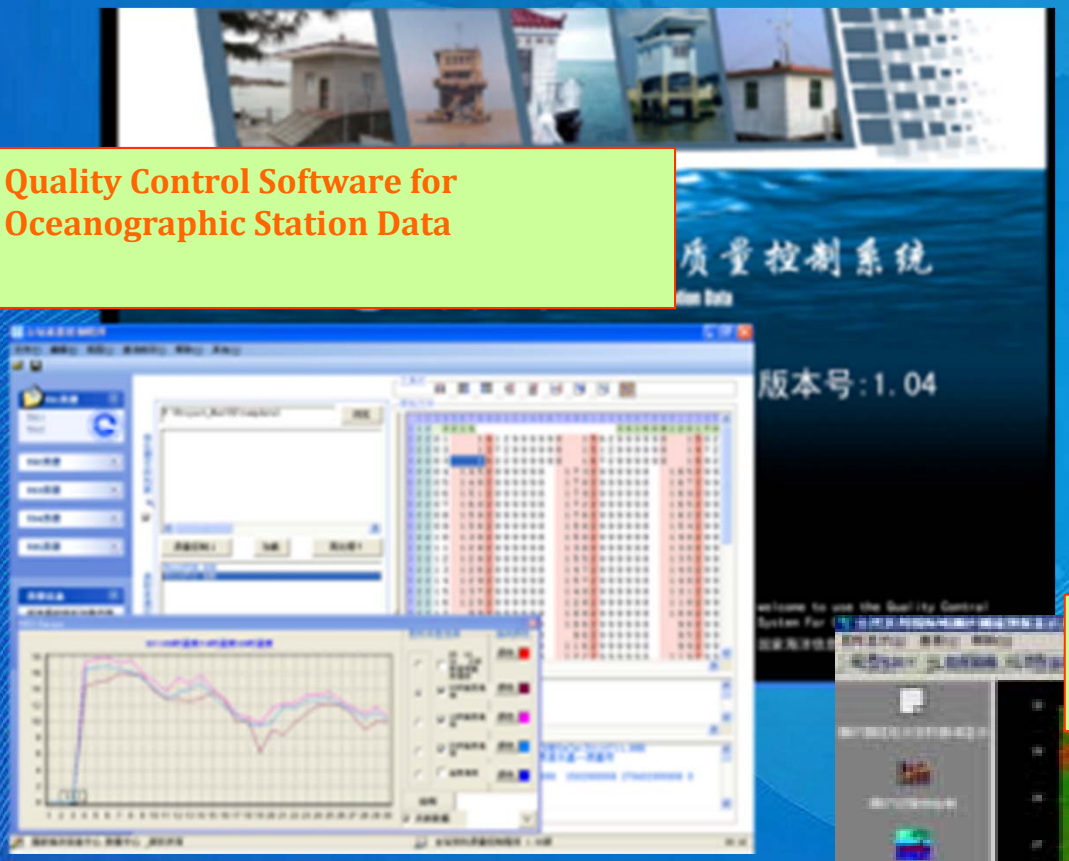
National Ocean Observing System (NOOS-China) including oceanographic stations, buoys, shore-based radars, voluntary observing ships, GPS stations and standard sections etc.

Oceanographic and marine meteorological data by other ocean related agencies, institute, private sectors.

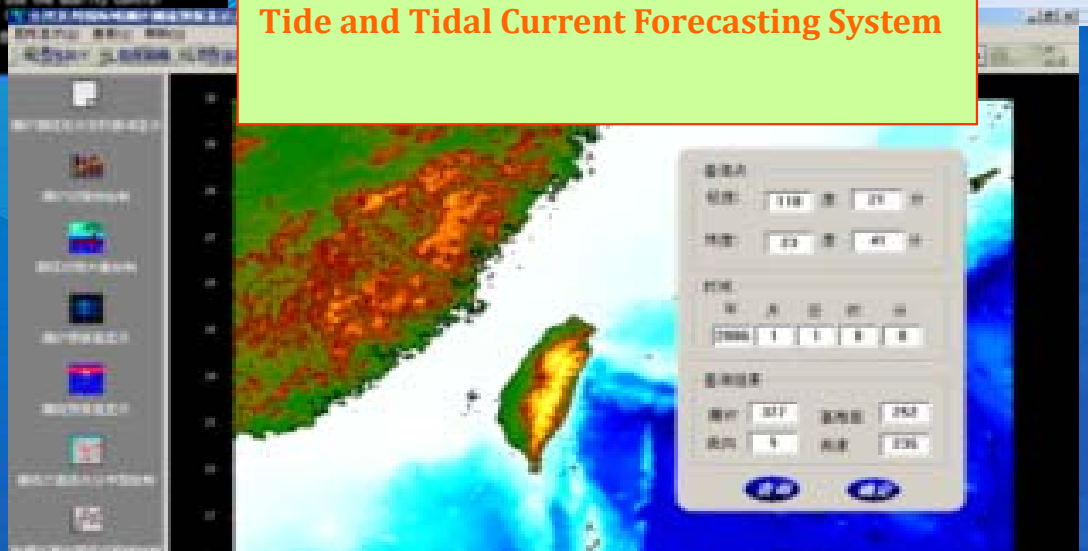
2.2 Data Processing System and Software

Voluntary Observing Ship Data Receiving and Processing System

Quality Control Software for Oceanographic Station Data

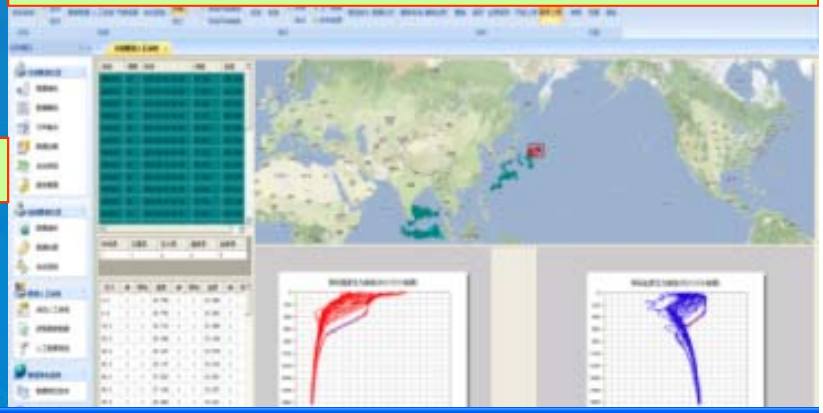


Tide and Tidal Current Forecasting System

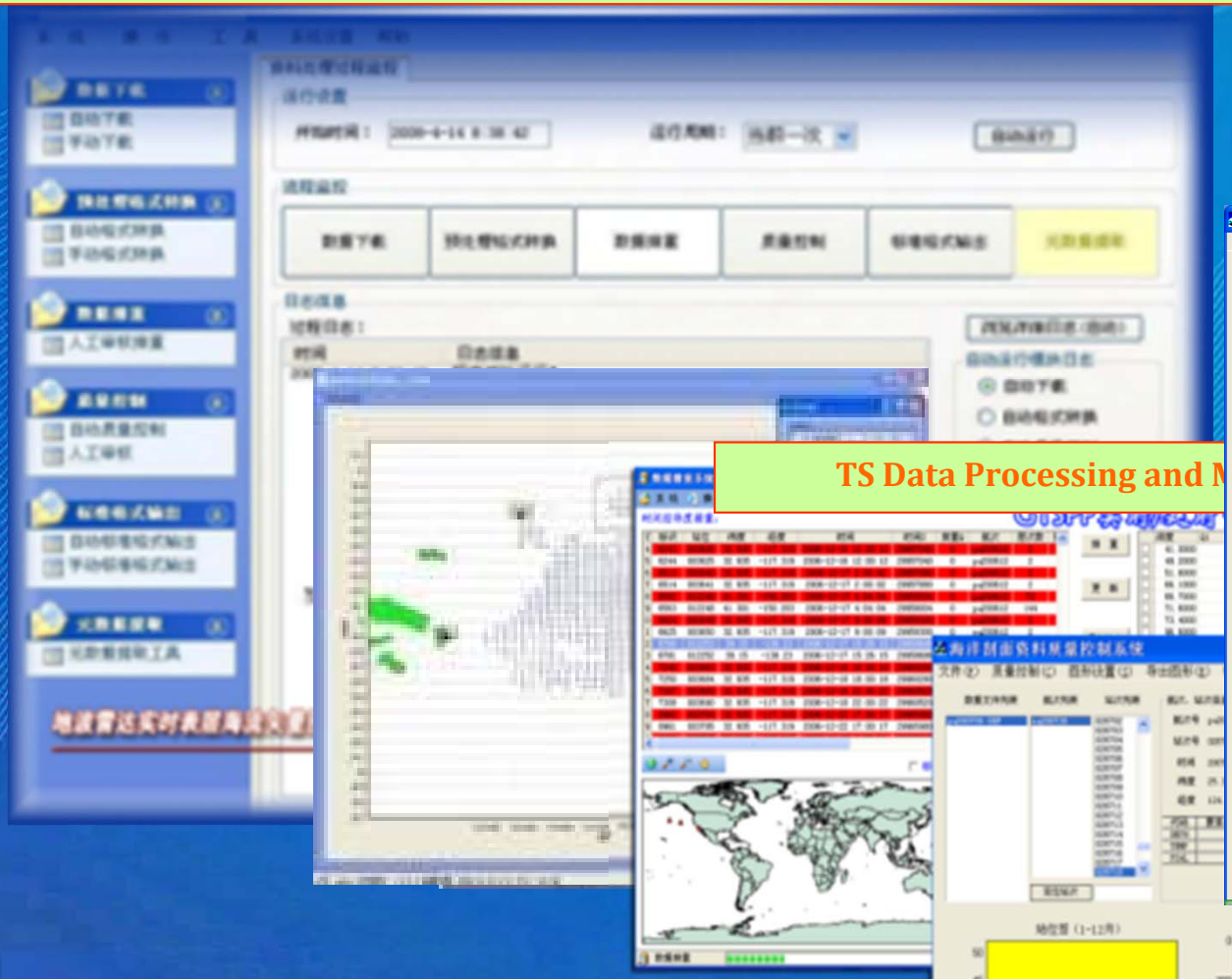


2.2 Data Processing System and Software

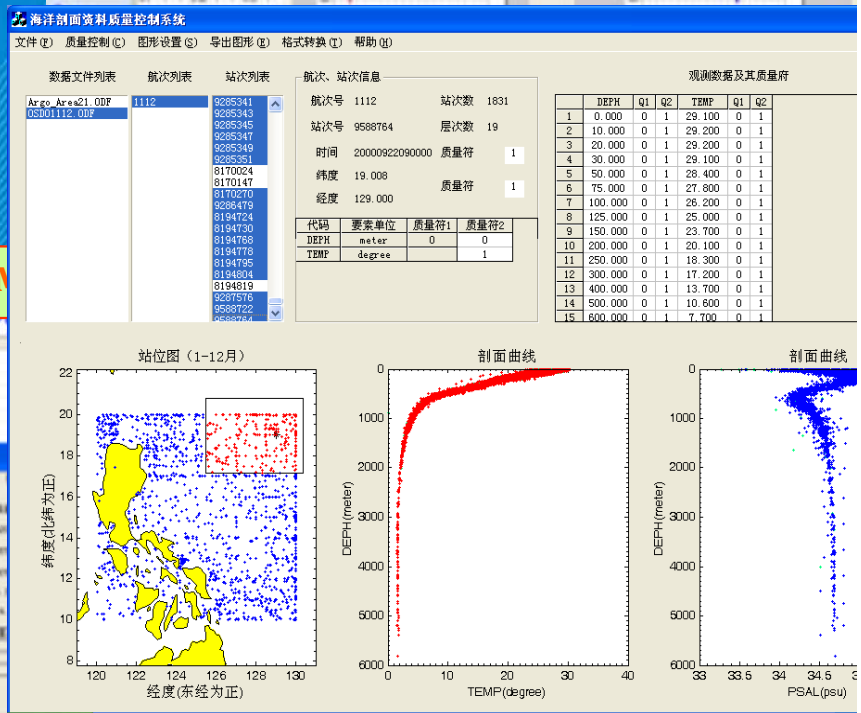
Argo Data Processing and Management System



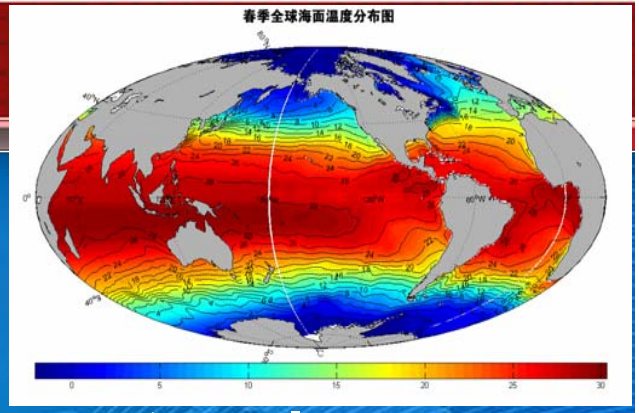
Radar Data Receiving and Processing System



TS Data Processing and M



2.3 Data Product

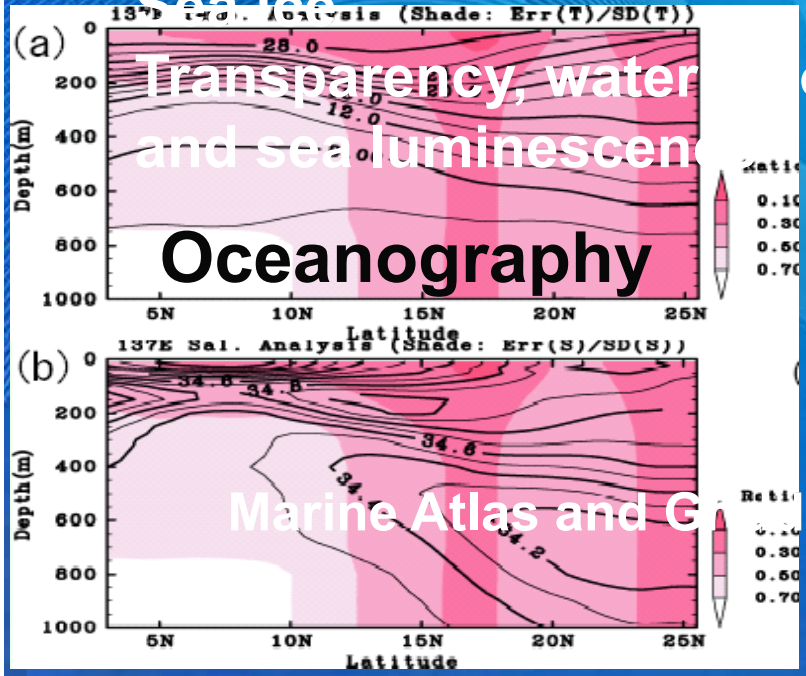


Fixed station
Buoy, Radar
Cruise obs.

Sea surface meteorology
Aerological data

Temperature and Salinity
current
Wave
Sea level
Sea ice

wind
air temperature, air pressure, humidity, precipitation, visibility, sea fog, cloud
Air-sea flux
Sea surface radiation



Transparency, water or and sea luminescen

Oceanography

Marine meteorology

Marine Atlas and G ed Statistical Analysis Product

3、Eliminate Duplicates of Ocean Data

Necessity of Eliminate Duplicates

- Many excellent marine datasets（WOD、ICOADS…）
- Gather information from different sources
- Between excellent datasets cross duplication
- The basic work need many resources
- scientific research requires integrated datasets

2.2 Eliminate Duplicates

Rules to Eliminate Duplicates

- set different levels of datasets producers.
- complete duplicate data, priority reserve high levels.
- inexact duplicate data , more additional information.
  - Keep more depth levels
  - Keep more additional variables
  - Keep higher precision
  - Keep the original observed level data

2.2 Eliminate Duplicates

Methods of Eliminate Duplicates

- **Instrument:** CTD 、UCTD、Argo 、Glider、Drift buoy、Ocean stations、Moored buoy、Autonomous Pinniped data、Undulating Oceanographic Recorder 、XBT and MBT etc.
- **key information items (key items)** : instruments, time, position, temperature profile, salinity profile，etc.
- **Accessorial information:** the country, Institute, the project

2.2 Eliminate Duplicates

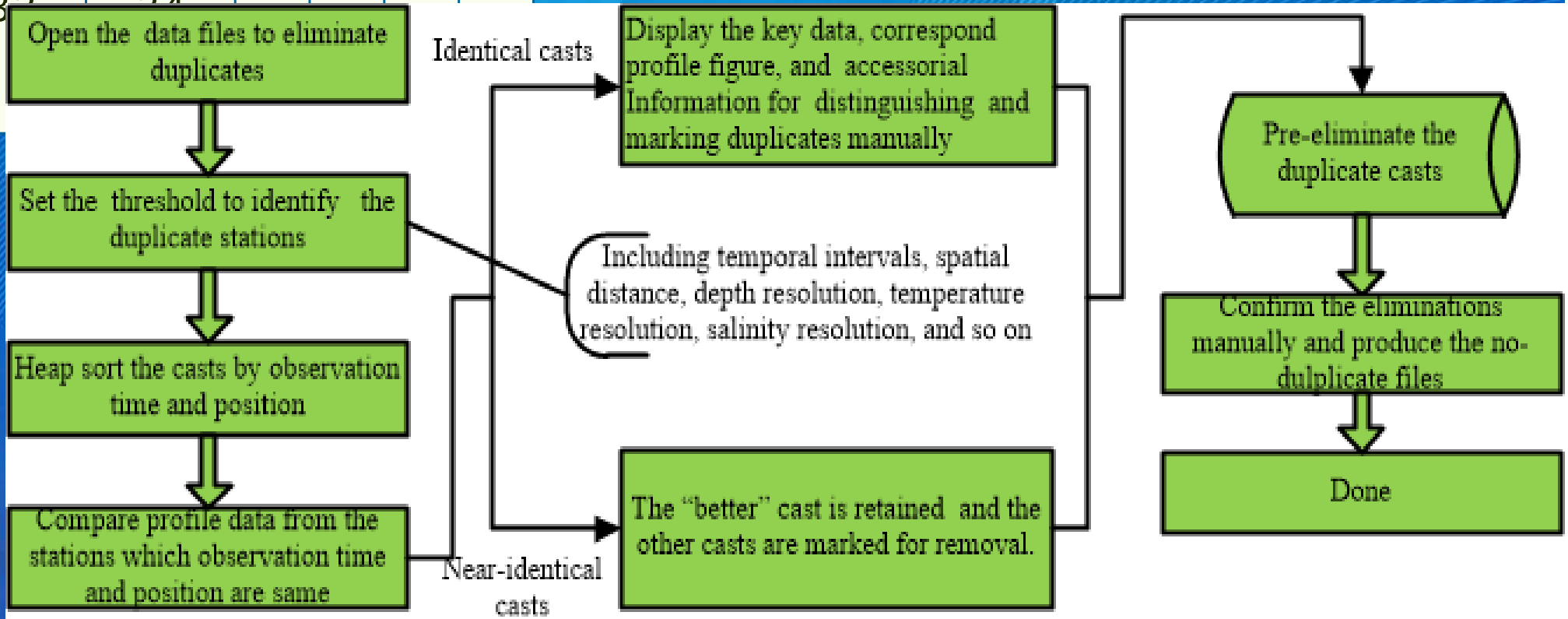
different parameters to identify the duplicate cast for different instruments ,and keep optimizing

Instrument	Temporal interval	Spatial distance(m)	Observed level Temperature difference（℃）	Observed level salinity difference
CTD	1 hours	10	0.01	0.01
UCTD	10 seconds	10	0.01	0.01
Argo	24 hours	10	0.01	0.01
Glider	3 hours	10	0.01	0.01
Drift buoy	1 hours	10	0.01	0.01
Ocean stations	1hours	10	0.01	0.01
Moored buoy	10 minutes	10	0.1	0.1
Autonomous Pinniped data	1 minutes	10	0.1	0.1
Undulating Oceanographic Recorder	10 seconds	10	0.1	0.1
XBT	10 minutes	10	0.1	
MBT	2 hours	10	0.1	

2.2 Eliminate Duplicates

1		2			
31	32	411	412		
		413	414		
33	34				

Flowchart of eliminate duplicates



3、Eliminate Duplicates

国际温盐数据排重系统

操 作 配 置 帮 助

站位信息

igroup	pid	long	lati	strtime	filename
36	338951	35.937	160.001008	2011-02-03 23:06	E:\9-观测业务
36	338952	35.935	159.985992	2011-02-03 23:08	E:\9-观测业务
37	338666	7.985	179.856003	2011-02-05 20:00	E:\9-观测业务
37	338694	7.986	179.845992	2011-02-05 20:00	E:\9-观测业务
38	330243	2.021	164.994003	2011-02-06 20:00	E:\9-观测业务
38	330271	2.021	165.005997	2011-02-06 20:00	E:\9-观测业务
39	338668	7.977	179.863997	2011-02-07 20:00	E:\9-观测业务
39	338696	7.992	179.849	2011-02-07 20:00	E:\9-观测业务
40	346405	41	141.882994	2011-02-08 15:46	E:\9-观测业务
40	355160	41	141.882994	2011-02-08 15:46	E:\9-观测业务
41	330167	8.012	170.046997	2011-02-13 20:00	E:\9-观测业务
41	330194	8.024	170.061006	2011-02-13 20:00	E:\9-观测业务
42	330251	2.008	164.996994	2011-02-14 20:00	E:\9-观测业务
42	330279	2.008	164.990997	2011-02-14 20:00	E:\9-观测业务
43	355197	22.5	160.050003	2011-02-17 21:34	E:\9-观测业务
43	355198	22.5	160.050003	2011-02-17 21:36	E:\9-观测业务
44	330172	7.994	170.039994	2011-02-18 20:00	E:\9-观测业务
44	330199	8.01	170.054992	2011-02-18 20:00	E:\9-观测业务
45	330173	7.997	170.044006	2011-02-19 20:00	E:\9-观测业务
45	330200	8.005	170.057006	2011-02-19 20:00	E:\9-观测业务
46	329995	5.013	170.009994	2011-02-21 20:00	E:\9-观测业务
46	330041	5	170.022994	2011-02-21 20:00	E:\9-观测业务
47	330175	7.999	170.044997	2011-02-21 20:00	E:\9-观测业务
47	330202	8.003	170.054	2011-02-21 20:00	E:\9-观测业务
48	330177	8.007	170.070008	2011-02-23 20:00	E:\9-观测业务
48	330204	8.003	170.056	2011-02-23 20:00	E:\9-观测业务

站位信息

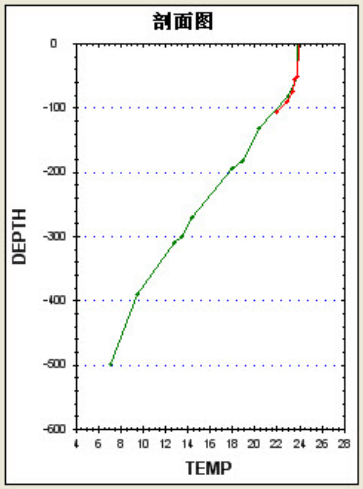
数据一

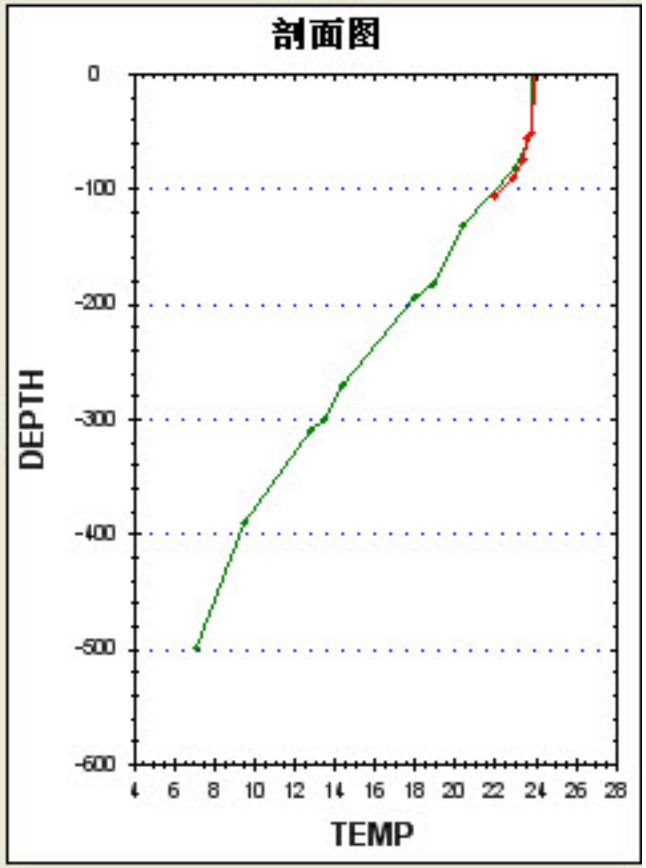
DEPTH	TEMP
0.0	24.400
2.0	24.000
53.0	23.900
57.0	23.600
76.0	23.400
91.0	22.900
107.0	22.000

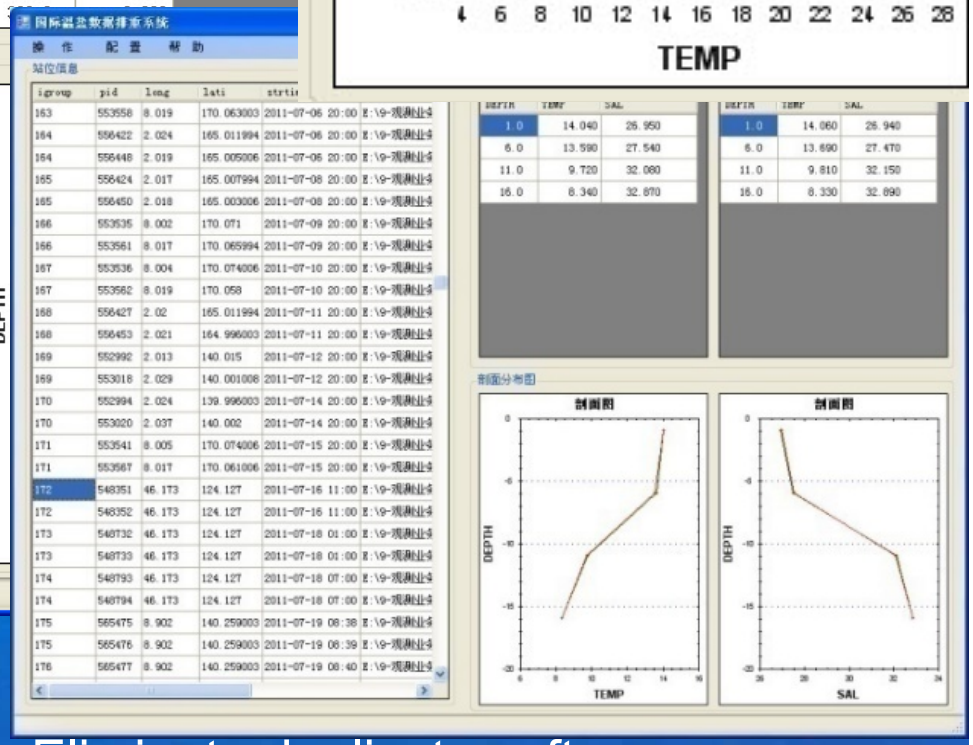
数据二

DEPTH	TEMP
0.0	24.600
2.0	23.900
52.0	23.800
83.0	23.100
132.0	20.400
183.0	18.900
195.0	18.000
272.0	14.400
302.0	13.500
311.0	12.900

剖面分布图



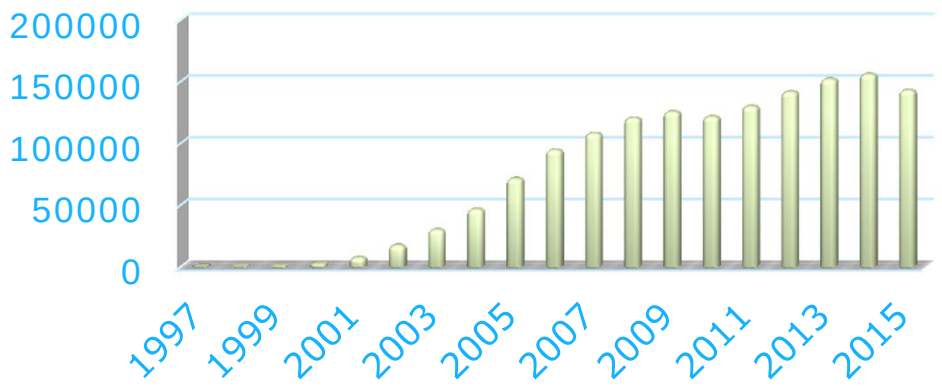


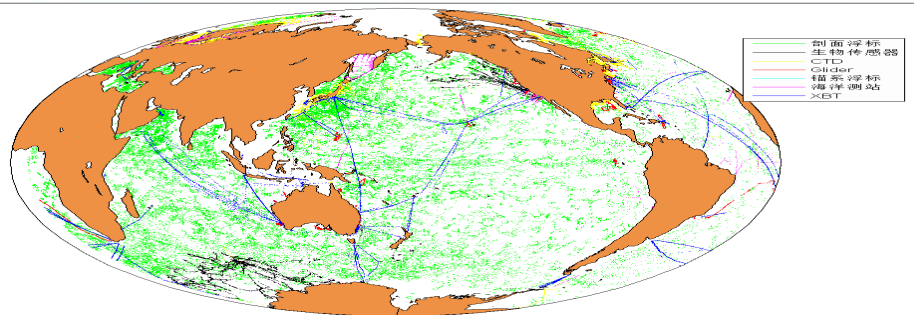


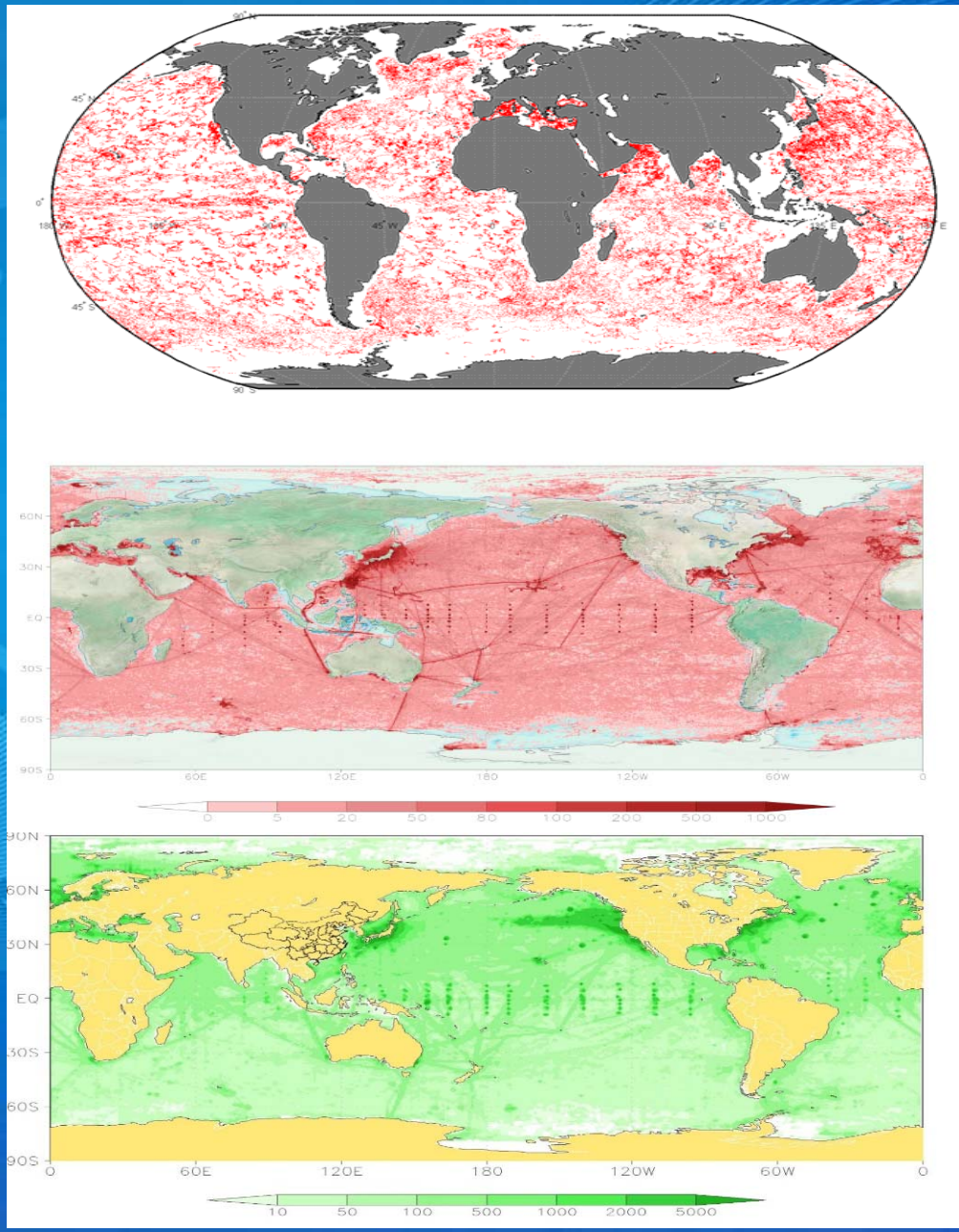
Eliminate duplicate software

Progress of TS data processing---example

Argo+GTSPP+WOD

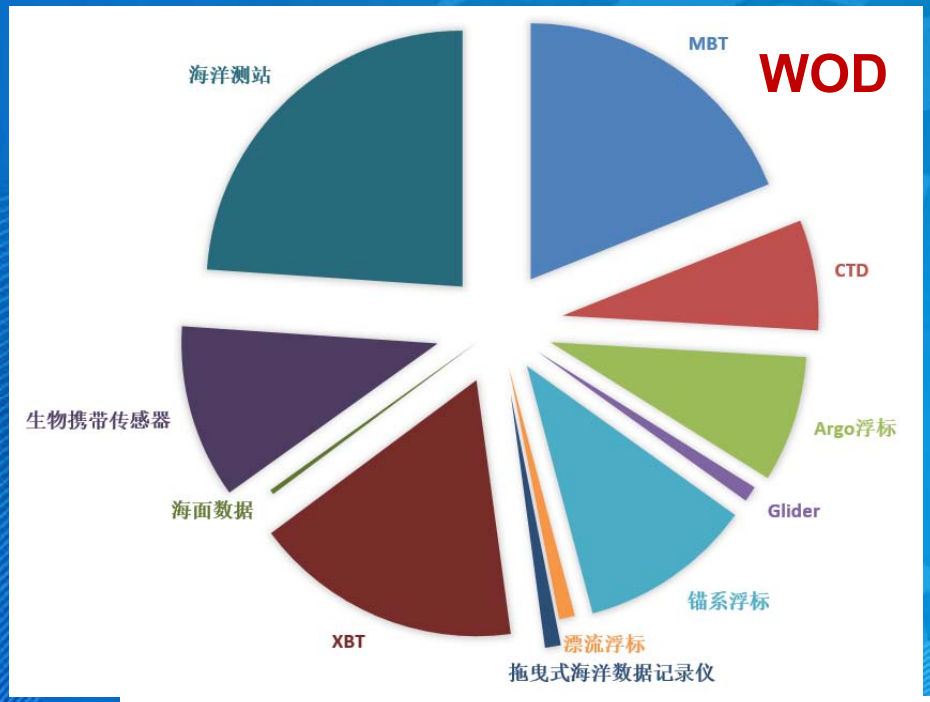


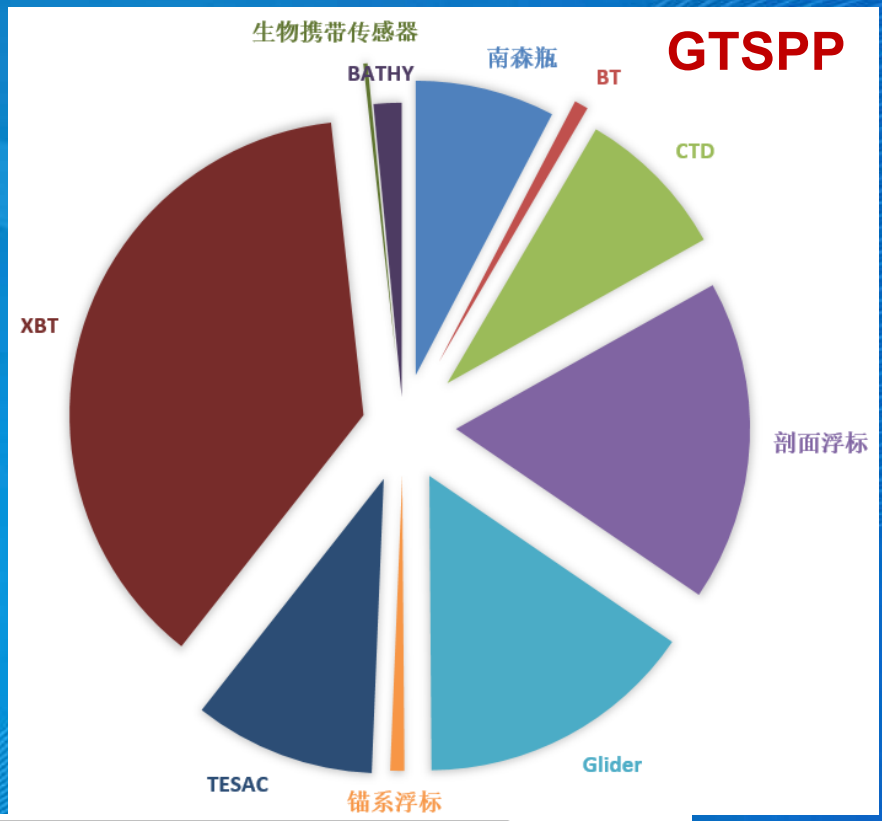


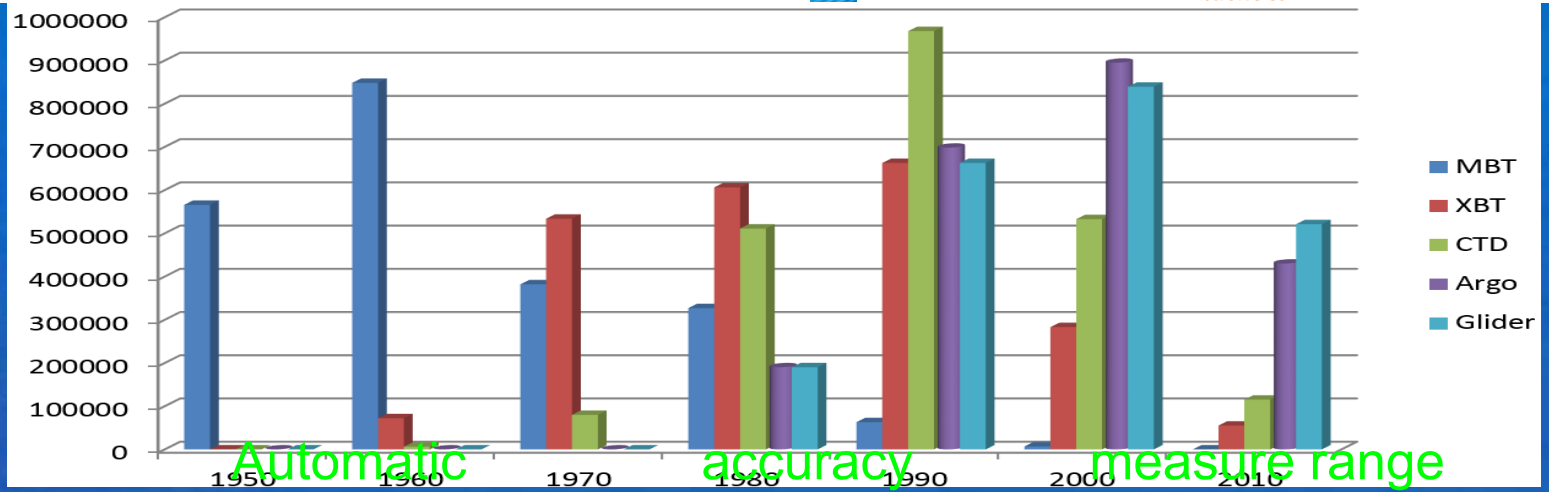


- Argo observational instruments: Argo
- GTSPP observational instruments：XBT、profiling buoy、TESAC and CTD etc.
- WOD observational instruments：CTD 、UCTD、Argo 、Glider、Drift buoy、Ocean stations、Moored buoy、Autonomous Pinniped data、XBT and MBT etc.

consist of instruments







Mechanic Automatic accuracy measure range parameters

Profile buoy data in WOD and Argo

- According to the observational period, the accuracy and the transmission frequency of the ocean observing instruments, the duplicate data are determined.
- the spatial distance threshold is 10 m, temporal interval threshold is 24 hours and the level depth interval threshold is 1m, the same layer temperature is not more than 0.01 degrees Celsius, the same layer salinity is not more than 0.01 PSS-78.

	WOD Profile buoy / Global Argo
number	925295/1086237
duplicates	837071

90% Argo

Profile buoy data in GTSPP and Argo

the spatial distance threshold is 10 m, temporal interval threshold is 24 hours and the level depth interval threshold is 1m, the same layer temperature is not more than 0.01 degrees Celsius, the same layer salinity is not more than 0.01 PSS-78.

	GTSPP Profile buoy / Global Argo
number	821053/925295
duplicates	737710

80% Argo

GTSPP and WOD XBT data

According to the overall composition of the amount of GTSPP and WOD data, XBT data is the very important body of these two kinds of data sets.

Comparison of XBT data in GTSPP and XBT data in WOD

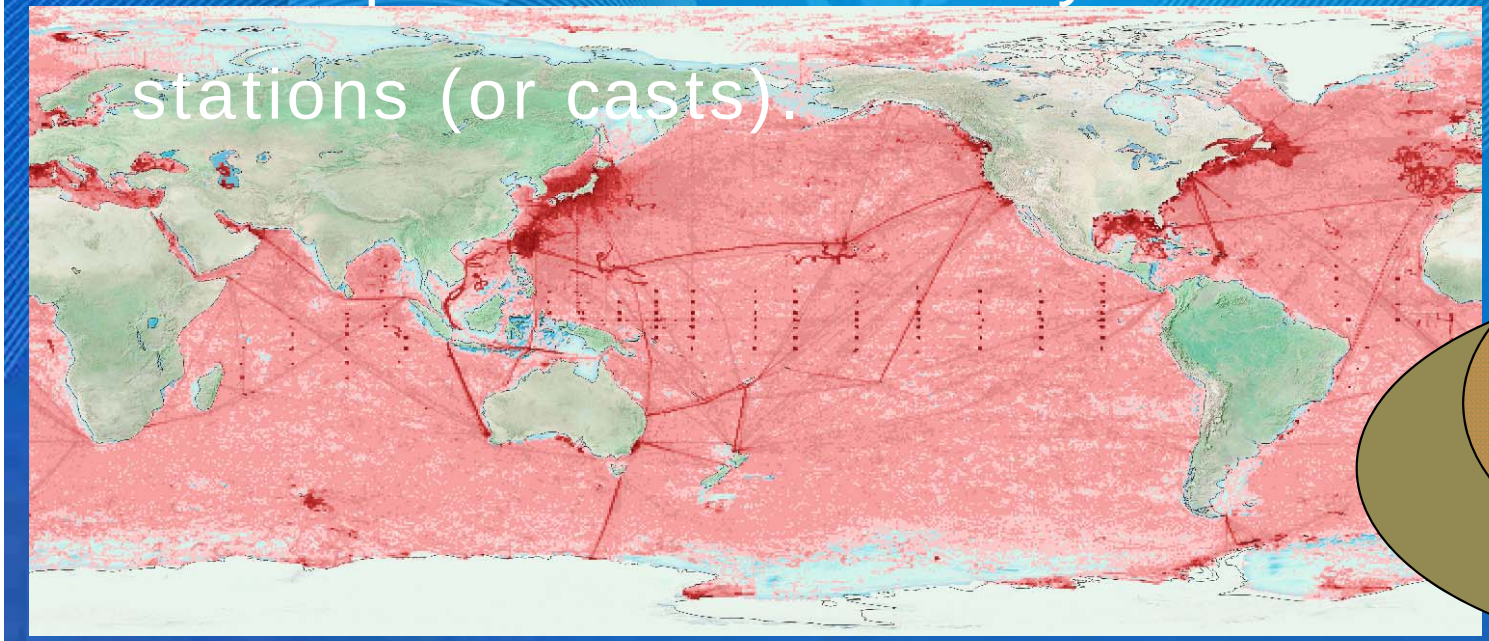
	GTSPP XBT/WOD XBT
number	567837/1000980
duplicates	538777

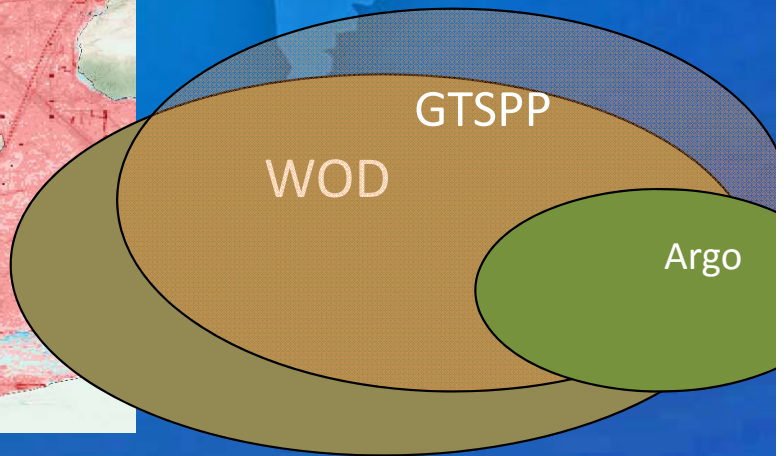
95% GTSPP XBT
WOD 2 times XBT than GTSPP

Progress of TS data processing

WOD, Argo and GTSPP integration

Based on the Argo data released by the global Argo data center with the complete observation information. We eventually completed the integration of temperature and salinity datasets. Total 13 million stations (or casts).

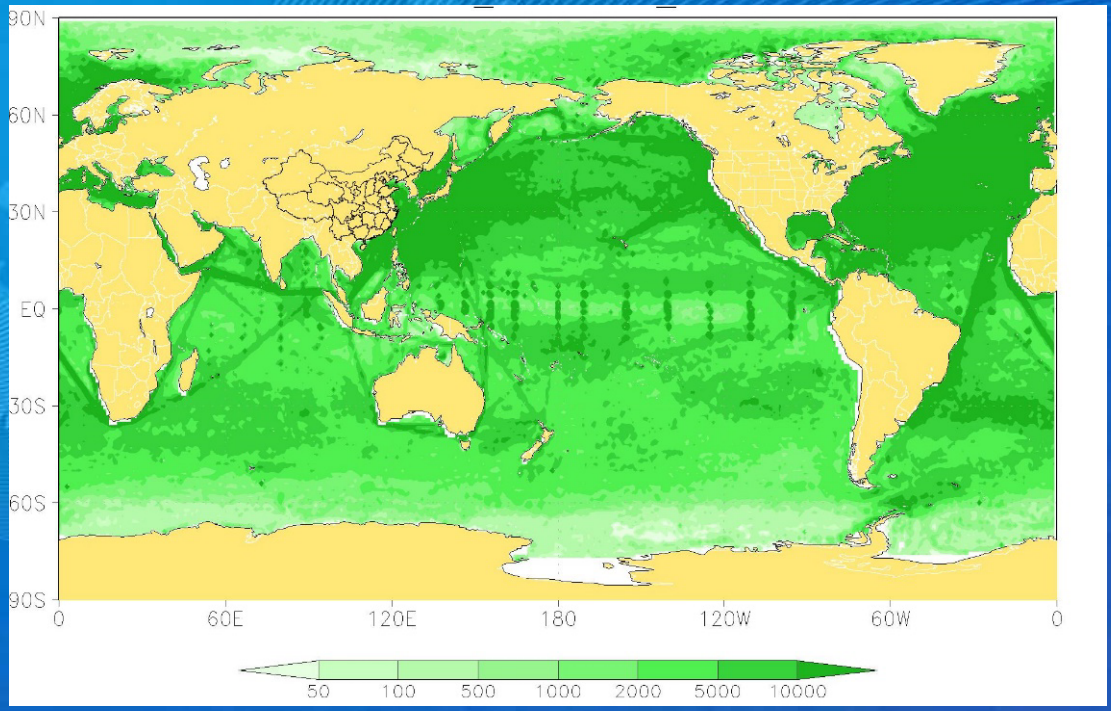




Progress of Marine Meteorological data processing

ICODAS total 402 million times distributed worldwide , Meteorological data includes sea surface air temperature, humidity, air pressure, wind, visibility, cloud, etc.

Quality control including the duplication check is performed on the ICOADS-IMMA data during the period of 1662-2013. With a duplication rate of 0.26%, the quality of these data varies with time and geographic locations. The average ration at 92.12%.



Progress of Marine Meteorological data processing

ICOADS and GTS transmission data (AAXX¥BBXX) eliminate duplicates (April 2014 as an example)

- The eliminate duplicates results of ICOADS and GTS fixed station observation data AAXX, GTS AAXX data is total 1444657 stations which 2742 stations(0.1866%) is the same with ICOADS data.
- The eliminate duplicates results of ICOADS and GTS ship observation data BBXX, GTS BBXX data is total 949604 stations which 906421 stations(95.45%) is the same with ICOADS data. Duplicates in data integration should be removed.

日期	观测次数	重复次数	重复率（%）
0401	49041	120	0.2
0402	76349	222	0.3
0403	85760	117	0.1
0404	71157	150	0.2
0405	82982	128	0.2
0406	71515	132	0.2
0407	73155	153	0.2
0408	69312	110	0.2
0409	67771	128	0.2
0410	85489	131	0.2
0411	69226	160	0.2
0412	68594	144	0.2
0413	67948	140	0.2
0414	69251	158	0.2
0415	75402	148	0.2
0416	35191	63	0.2
0417	23142	53	0.2
0418	23330	49	0.2
0419	23404	46	0.2
0420	23389	64	0.3
0421	23534	71	0.3
0422	23581	45	0.2
0423	23453	35	0.1
0424	23256	30	0.1
0425	23356	27	0.1
0426	23066	41	0.2
0427	19012	23	0.1
0428	14777	22	0.1
0429	27095	32	0.1
0430	32119	49	0.2
	1444657	2742	0.1867

日期	观测次数	重复次数	重复率（%）
0401	19634	18695	95.2
0402	32691	31191	95.4
0403	30965	29520	95.3
0404	31000	29538	95.3
0405	31334	29963	95.6
0406	32134	30106	93.7
0407	32775	30319	92.5
0408	30710	29145	94.9
0409	31103	29748	95.6
0410	32562	31186	95.8
0411	33337	31946	95.8
0412	33720	32257	95.7
0413	33155	31746	95.8
0414	32950	31606	95.9
0415	32769	31391	95.8
0416	36911	35464	96.1
0417	32866	31470	95.8
0418	32959	31541	95.7
0419	34259	32832	95.8
0420	33847	32435	95.8
0421	33766	32392	95.9
0422	33773	32478	96.2
0423	33214	31804	95.8
0424	33491	32161	96.0
0425	33344	32003	96.0
0426	33007	31683	96.0
0427	23518	22614	96.2
0428	19985	19170	95.9
0429	32831	31158	94.9
0430	30994	28859	93.1
合计	949604	906421	95.45

4、Future work

- Continue to upgrade QC and TEST eliminate duplicate data methods.
- Based on the geographical advantage，focus to data fine processing of the Asia Pacific region.
- introduce more data sources to do the eliminate duplicates TEST of the marine data.
- TEST the integration of TS and marine meteorological data operation gradually.


THANK YOU!