# Measuring Performance, Skill and Accuracy in Operational Oceanography : Overview of approaches proposed by the GODAE/Ocean Predict Intercomparison and Validation Task Team

Fabrice Hernandez (IRD) and Greg Smith (ECCC)

OceanPredict Intercomparison & Validation Task Team

- Introduction: Operational oceanography developments since the 1990‘s
- Validation and Verification in Operational Oceanography: Concepts and overarching priorities
- International assessment framework: from GODAE to OceanPredict with the Intercomparison and Validation Task Team:
  - ✓ Network, organisation and standardisation
  - ✓ The Copernicus Marine Service: organized with verification/validation at the heart of the system
  - ✓ Scientific outcomes
- New challenges and metrics

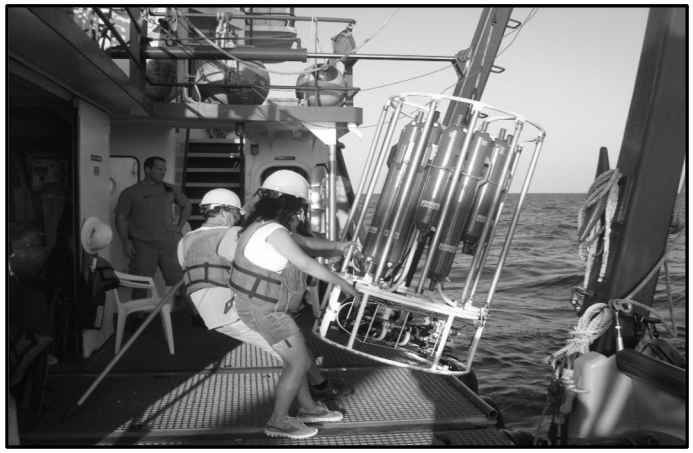
*Hernandez, Smith et al, (2018)*

# Operational Oceanography: Early initiatives in the 1990's

**From early ocean studies to Operational Oceanography**

**PAST**

«Expeditions» and offline sampling



Ocean observing system enhanced with satellite era (satellite altimetry, radiometry)

Improved ocean models

Adapting assimilation technique to ocean

Computers able to run global forecasting systems in real time

**NOW/FUTURE**

Continuous real-time monitoring of ocean and environment

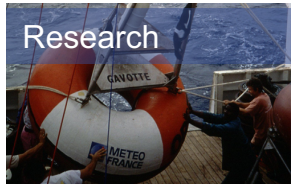
# Operational Oceanography: Users/Services identified in the 1990's – 2000's

**A « core service » to serve a wide range of applications**





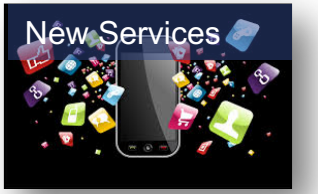














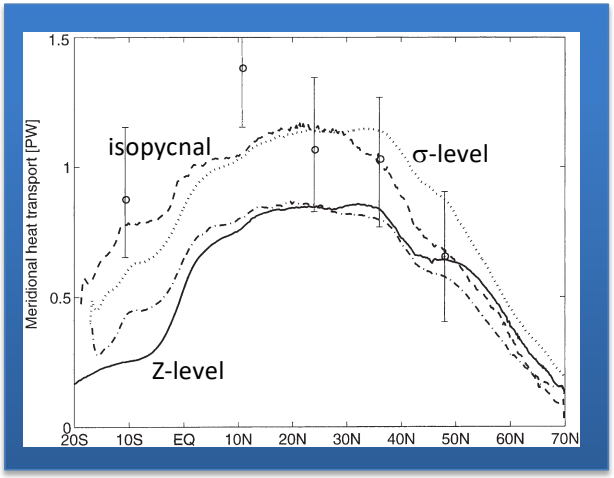

# Operational Oceanography Validation, Verification, Assessment activities:

# Concepts and Priorities

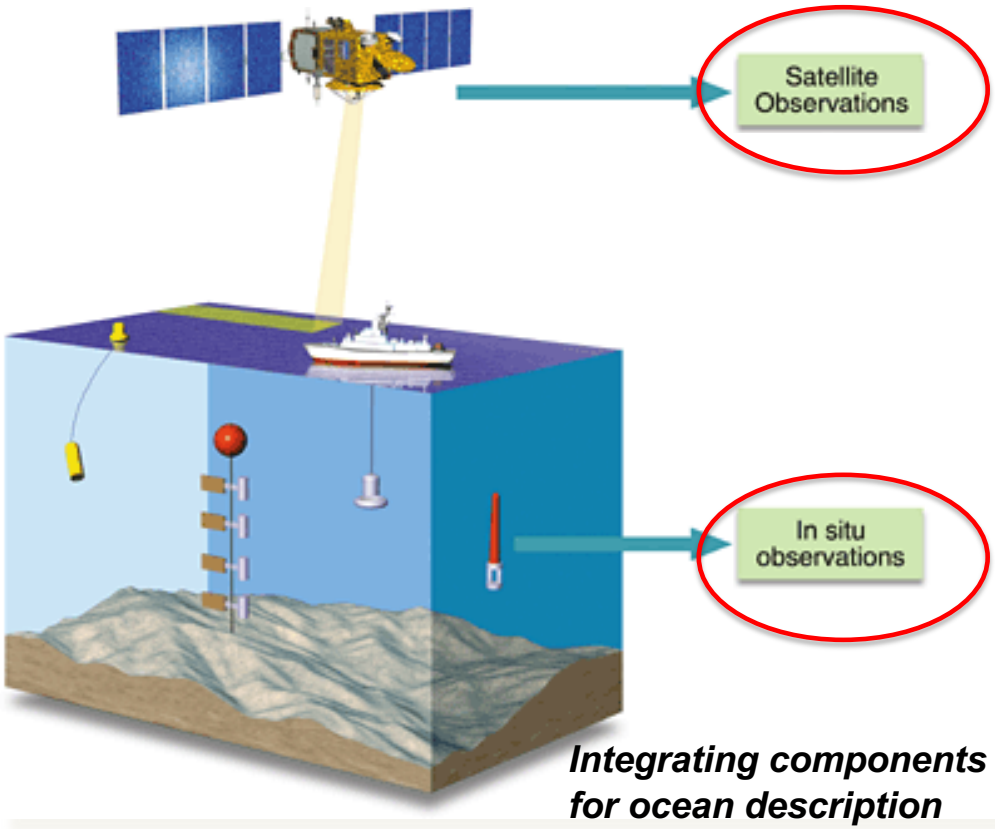
**Challenge: the ocean is poorly observed and undersampled !**

# Evaluation in oceanography: from academic to mature operational system assessment

Ocean models: **tools to describe and understand** ocean processes



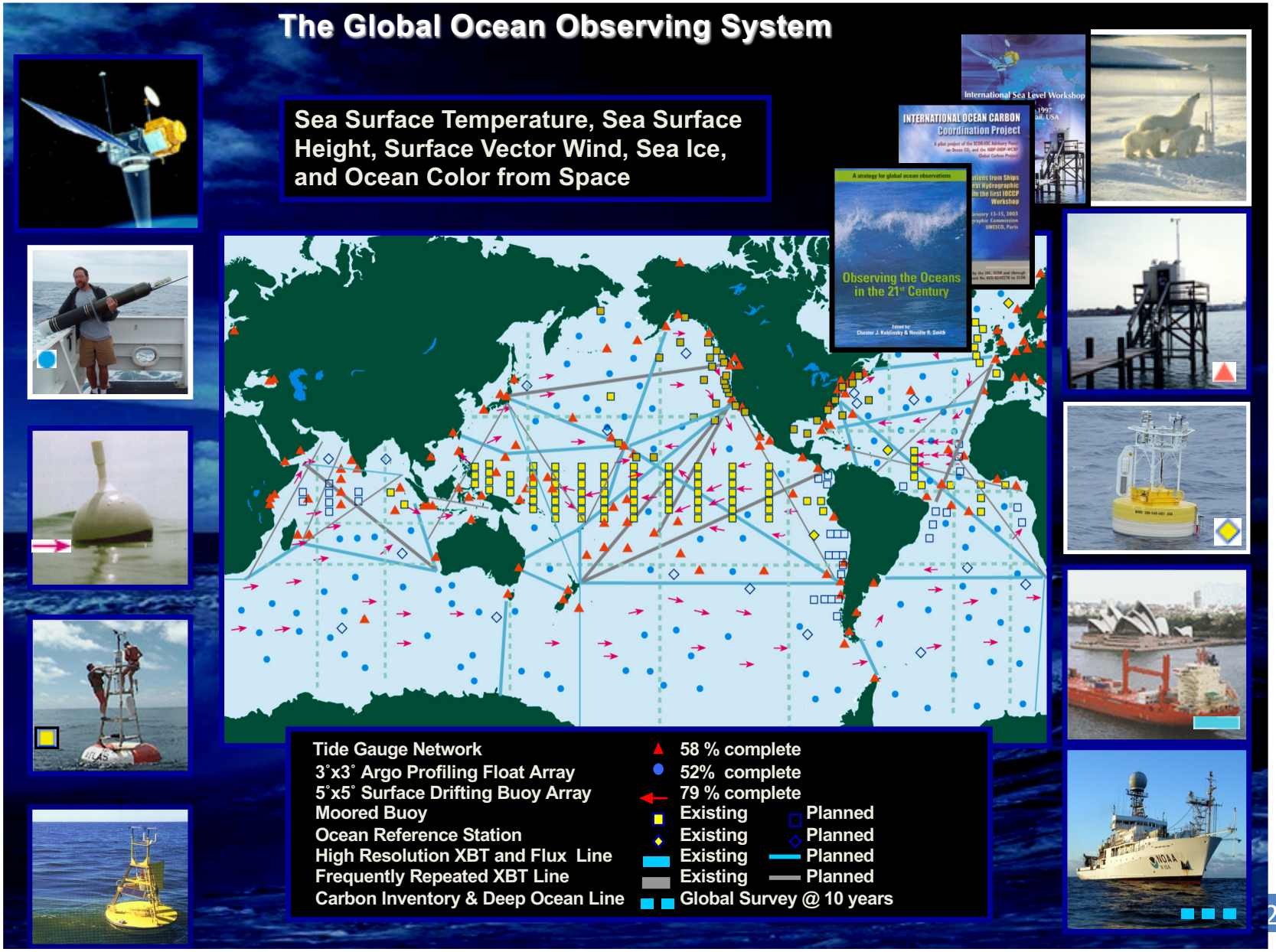

Assessment projects:
DYNAMO: **Meridional Heat Transport** in the North Atlantic Ocean. Values and errors bars given by Macdonald and Wunsch [1996]. Taken from figure 9 of Willebrand et al. (2001).




***Integrating components for ocean description***

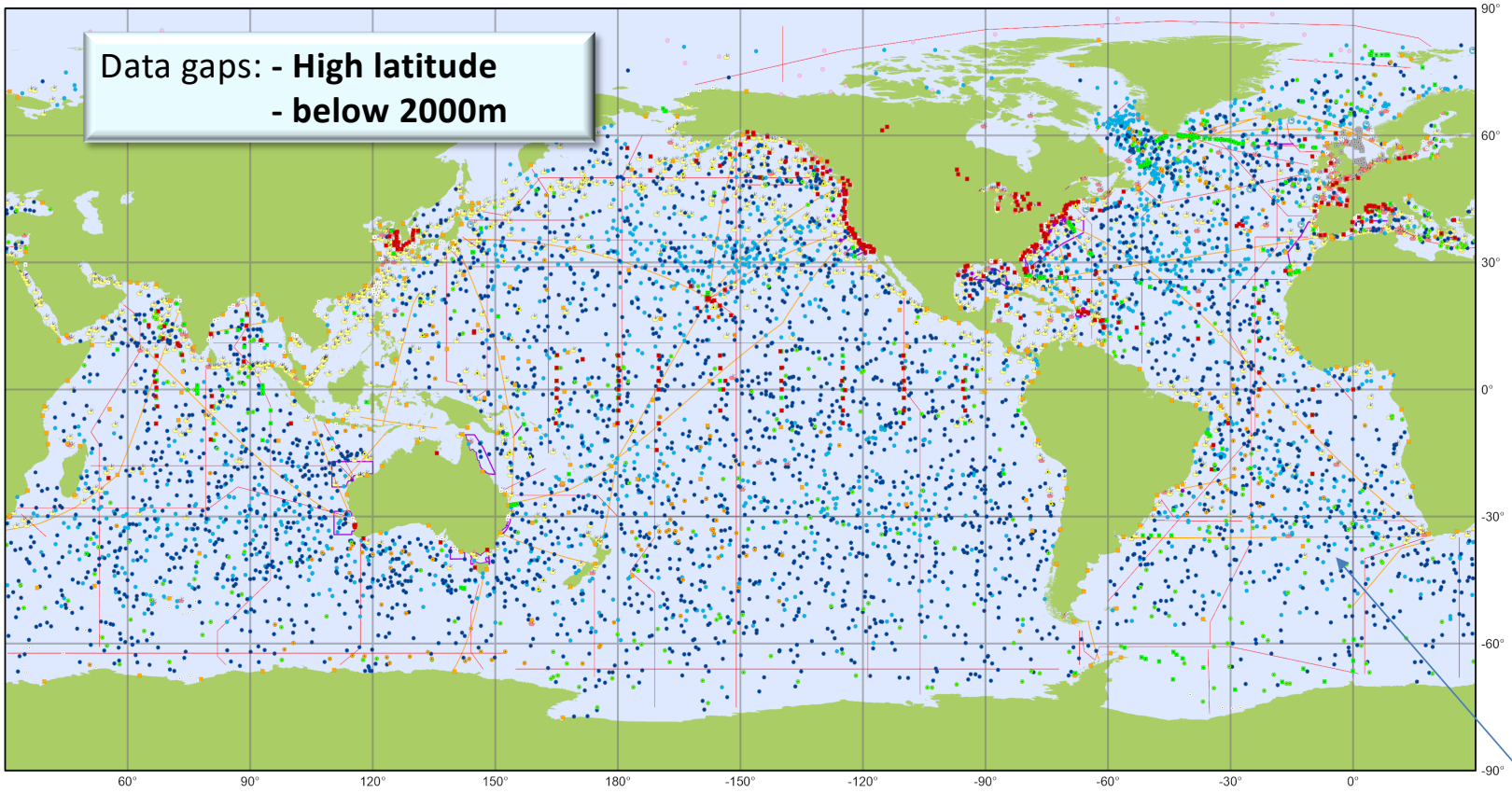
- Observation errors
- Observation representativity (distribution, sparseness)
- Estimation and mapping errors

Observations can not provide a complete description of ocean processes !




Ocean Predict

Ocean Observing System: Sparse compared to the Atmosphere !




Global ocean observing system

In situ operational platforms monitored by OceanOPS

September 2020

**Mobile systems**
- Core floats - Argo
- Deep floats - Argo
- Biogeochemistry floats - Argo
- Underwater gliders - OceanGliders
- Drifting buoys - DBCP
- Polar buoys - DBCP
- Animal borne sensors

**Fixed systems**
- Tsunameters - DBCP
- Offshore platforms - DBCP
- Moored buoys - DBCP
- Ocean reference stations - OceanSITES
- Sea level gauges - GLOSS
- High Frequency radars

**Ship based measurements**
- Manned weather stations - SOT/VOS
- Automated weather stations - SOT/VOS
- Radiosondes - SOT/ASAP

**Reference lines and areas**
- Repeat hydrography - GO-SHIP
- eXpendable BathyThermographs - SOT/SOOP
- Sampled sites - OceanGliders

*Generated by www.ocean-ops.org, 2020-10-23*

Ocean Predict

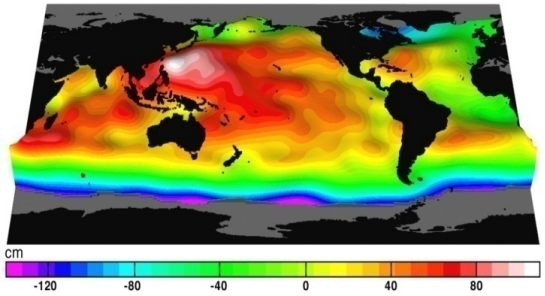
Ocean Observing System: Sparse compared to the Atmosphere !

3000 Argo floats = 1 T/S profile over 300 x 300 km every 10 days- 0-2000m

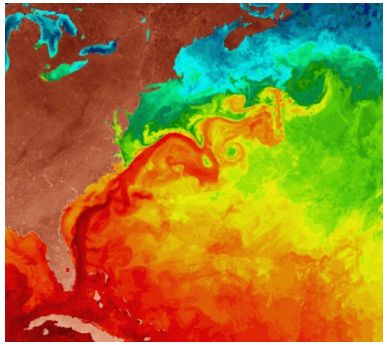
➔ Samples seasonal variability, not the ocean mesoscale

020-IVMW-O

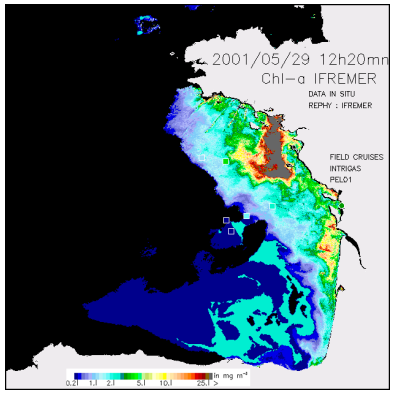
# Ocean Observing System: Sparse compared to the Atmosphere !




Altimetry and gravimetry
(sea level and ocean currents)



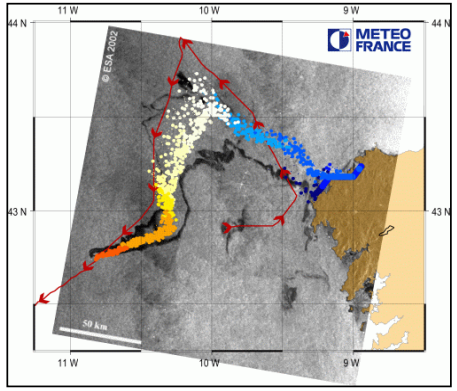
Sea Surface Temperature



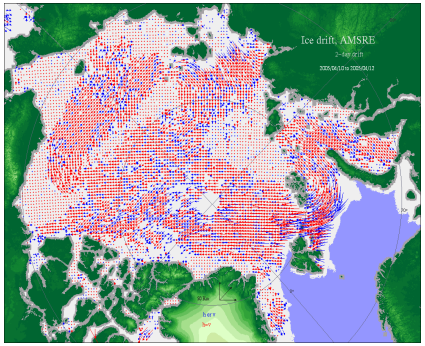

Ocean Colour
(Chl-a, SPM)

**Satellites:**
- **Synopticity**
- **High resolution**
- **Large amount of data**

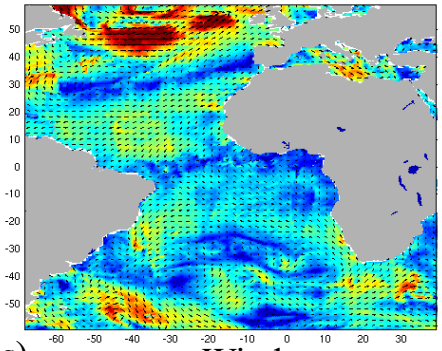
**But only observe the ocean surface !**




Surface roughness from SAR
(e.g. waves, winds, oil slicks)




Sea Ice (concentration, drift, thickness)




Winds
(speed and direction)

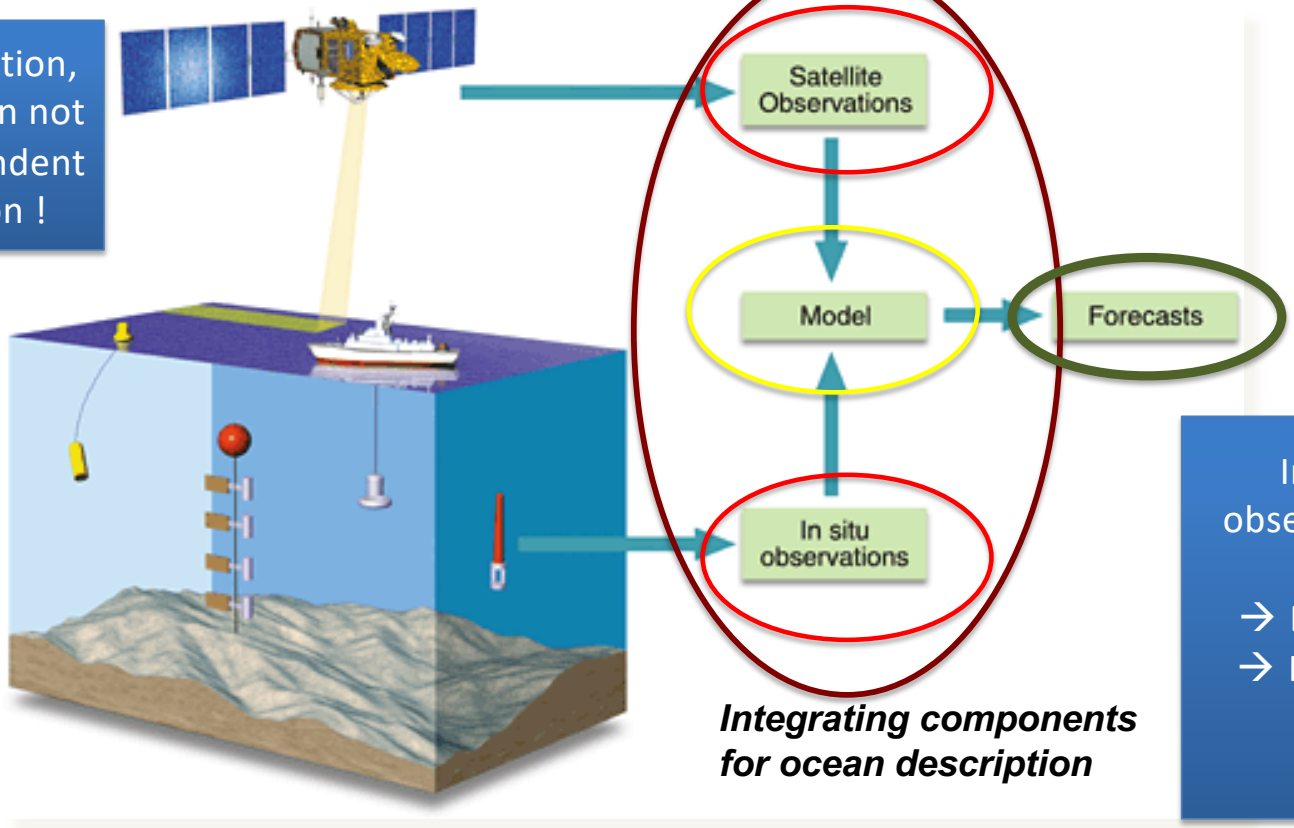
Sub-mesoscales (<10 km) still not observed by satellites

Operational models with resolution reaching 1km cannot be constrained by assimilation

# Evaluation in oceanography: from academic to mature operational system assessment

Ocean Predict

Once used by assimilation, ocean observations can not be considered independent for further evaluation !




***Integrating components for ocean description***

- Performance of the system
- Forecast skill

In real time, few ocean observations for assessment !

→ Delayed mode validation
→ Performance assessment through retrospective experiments

# Quality assessment objectives of operational oceanography

- Evaluate and **monitor performance** of operational system
  - Impact of the observing system
  - Model errors
  - Data assimilation efficiency
- Evaluate **accuracy of products**:
  - Products derived from observation (RT or reprocessed),
  - Routine hindcast and forecast (skill)
  - Reanalyses
- Measure **strength and weaknesses** of operated system for further improvements

***Internal***
***Science driven***

Provide timely robust and reliable products for a useful and cost effective service

- Assess **product's reliability considering user's needs**

***External***
***User driven***

# Quality assessment objectives of operational oceanography

Provide timely robust and reliable products for a useful and cost effective service

**Operational oceanography is now facing same challenges and user's expectation than weather forecast, or climate assessment !**

- From science driven to user driven
  - Challenges in understanding user's requests
- Operational oceanography **continuously evolving toward more complex system**
  - Coupling (atm, wave,ice, biogeochemistry, rivers) & nesting (scales)…
  - Diversity of products (blue, green and white ocean)
  - Focus: climate, seasonal, short-term predictions / open ocean, coastal
- Many applications… and decision makers
- Challenges in communicating product's reliability
  - **General public not really aware** of oceans behaviour (at the opposite of weather !)
  - Ocean intermediate **users** are **experts, with requirements in terms of quantified errors**

# The validation « philosophy » adopted in our community

- **Basic principles. Defined for ocean hindcast and forecast**

  **(Murphy, 1993, adopted in GODAE by Le Provost 2002, MERSEA Strand 1):**

  - **Consistency: verifying that the system outputs are consistent with the current knowledge of the ocean circulation and climatologies**
  - **Quality (or accuracy of the hindcast) quantifying the differences between the system "best results" (analysis) and the sea truth, as estimated from observations, preferably using independent observations (not assimilated).**
  - **Performance (or accuracy of the forecast): quantifying the short term forecast capacity of each system, i.e. Answering the questions "do we perform better than persistency? better than climatology?…**

- **A complementary principle, to verify the interest for the customer (eg, Pinardi and Tonani, 2005, MFS):**

  - **Benefit: end-user assessment of which quality level has to be reached before the product is useful for an application**

# Validation and verification in operational oceanography: The international framework

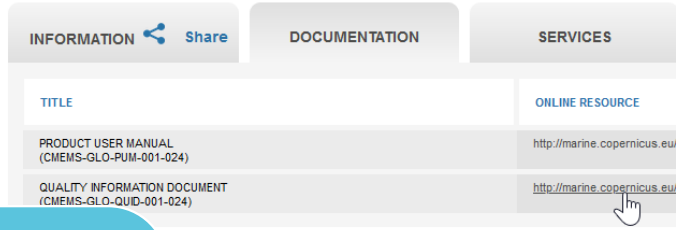
- 1990 → 2010 early stage of operational oceanography, during which we have settled our validation and verification practices
- 1998-2008: The operational oceanography community get organized with the **Global Data Assimilation Experiment (GODAE)** becoming **GODAE OceanView** (GOW 2008-2018), now **OceanPredict** (2018+)
  - First intercomparison (Crosnier and Le Provost 2007; Hernandez et al, 2009)
- Validation and verification scientific issues: one of the first task team raised inside GODAE, still active in OceanPredict
  - The IV-TT, since 2013 is carrying on an **intercomparison project** among international global forecasting centres
  - First outcomes in 2015: **SSH, SST, T(z)/S(z)** (Divakaran et al., 2015; Hernandez et al., 2015: Ryan et al, 2015; Zhu et al., 2016)
  - Addition of **Sea-Ice** intercomparison in 2015 (Smith et al, 2016)
  - Recent addition of **surface velocity** intercomparison in 2018
- GODAE+CLIVAR/GSOP (Global Synthesis and Observations Panel): In parallel, since 2002, the **ocean reanalysis** community start intercomparison projects
  - Ocean State Estimation Project published in 2009: 10 ocean reanalyses (Lee, 2009)
  - ORA-IP published in 2015: 26 ocean reanalysis/reprocessed estimation (Balmaseda et al, 2015)
- Regional initiatives: in Europe the operational community get organized (2002-2004: the MERSEA Std1 project) both for global and coastal systems and, in conjunction with GODAE, develop, test, experiment, implement validation and verification approaches in a common and distributed framework: **Copernicus Marine Service** (Drévillon et al, 2018)

# Scientific quality assurance loop in OO

Marine Monitoring

Collect and answer users' feedback

## Oceanographic information on CMEMS website




Validation in support of service desk

Planned evolution

Pre operational qualification

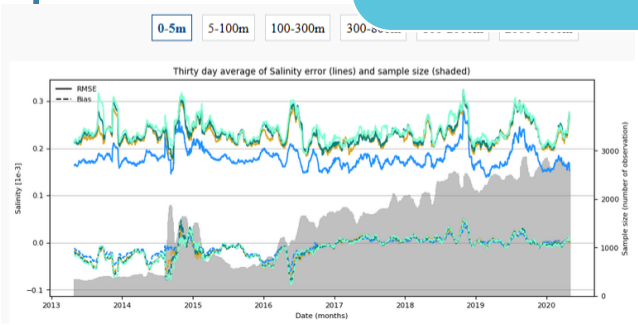
Routine verification

Delayed time verification

Recent & regional values of **EANs** on website

**Operational oceanography Internal process**

Reference QUID document including quality summary, and **estimated accuracy numbers EANs** available for each product on catalogue



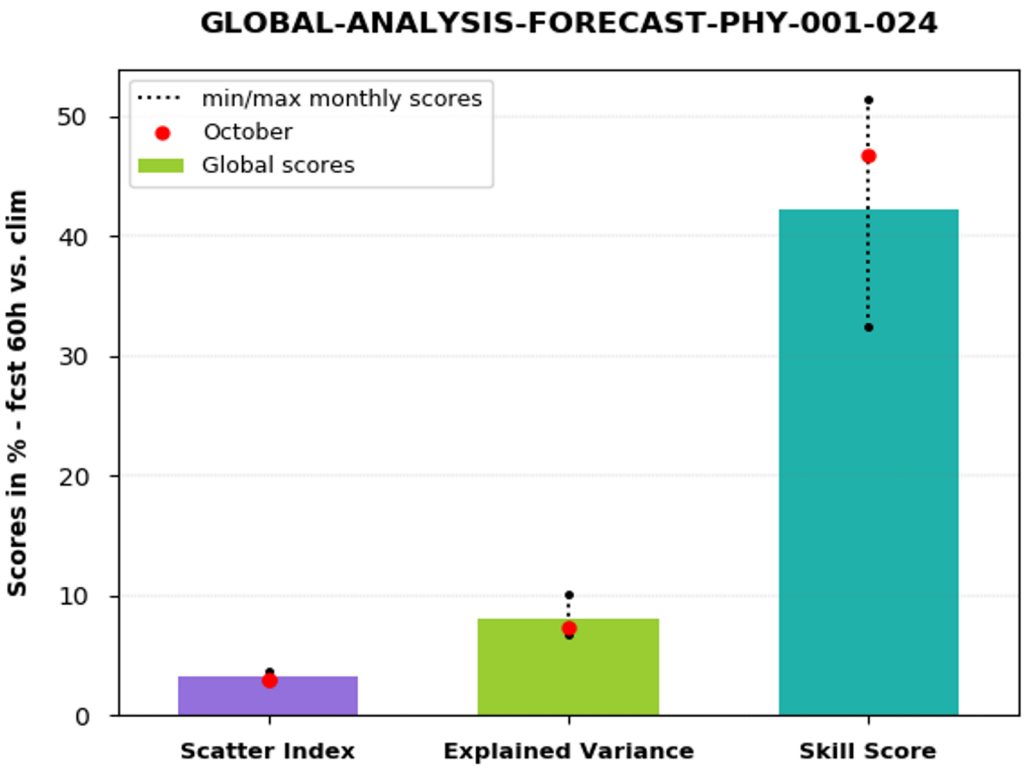

European Commission

Copernicus Europe's eyes on Earth

Implemented by MERCATOR OCEAN INTERNATIONAL

Marine Monitoring

# Product Quality Dashboard




Example for global SST forecast 60h

Website in development:

- Dynamic display of quality information from more than 15 different data producers
- Display of quality scores including forecast skill score
- Starting with climatological values of score (lack of reference observations in real time, maturity of the production chains)

The reference documentation is evolving to include **more variables & more summaries**

European Commission

opernicus Europe's eyes on Earth

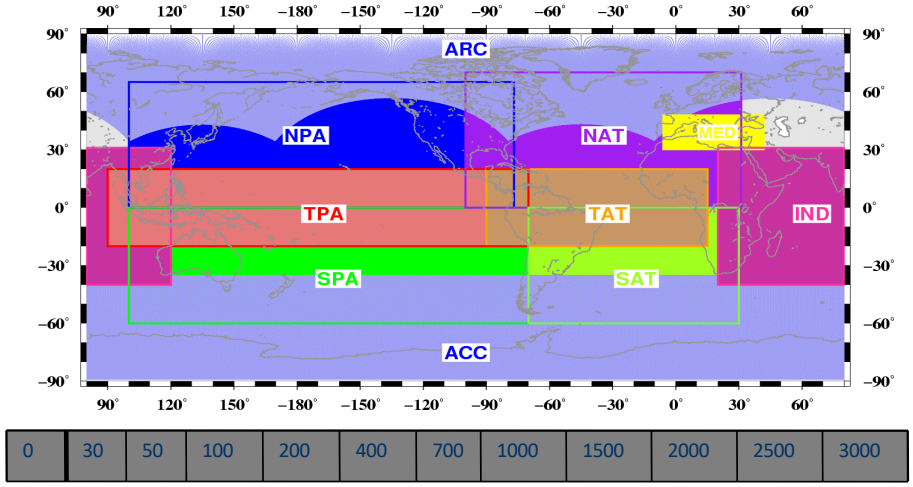
Implemented by MERCATOR OCEAN INTERNATIONAL

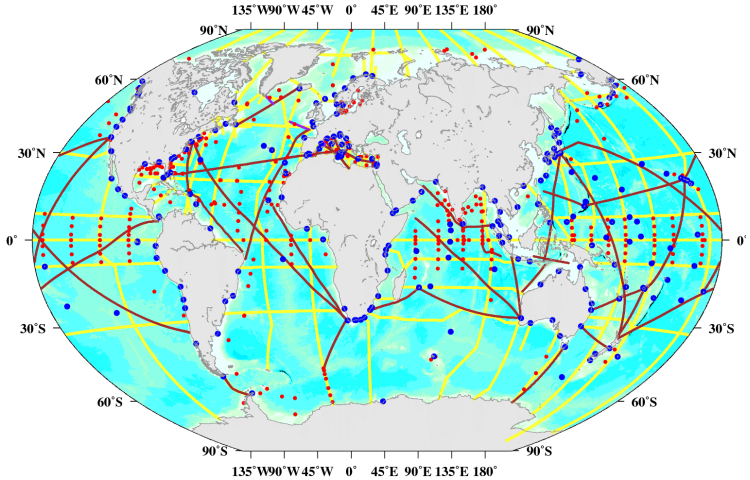
## GODAE, GODAE OceanView, OceanPredict
## Validation and Intercomparison Task Team


- Design metrics in common
- Implement, compute metrics
- Define a technical framework for exchanging information (meta-data, standard, NetCDF file, access procedures for all participants, storage…)
- Perform assessment and intercomparison
- Then publish result

# GODAE metrics definitions: Le Provost 2002, MERSEA Strand 1

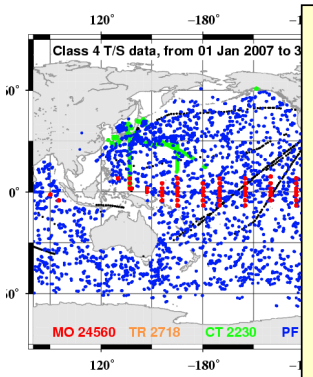
- **Class 1** – daily average **model fields interpolated onto pre-defined grids** (*eddy-permitting view*) on specified levels
- **Class 2** – model fields interpolated to **pre-defined mooring locations** and sections.
- **Class 3** – transports through sections and other **integrated quantities** such as Meridional Overturning Streamfunction and heat transports.
- **Class 4** – assessment **of forecasting capabilities through comparison** of model with assimilated and independent **observations**

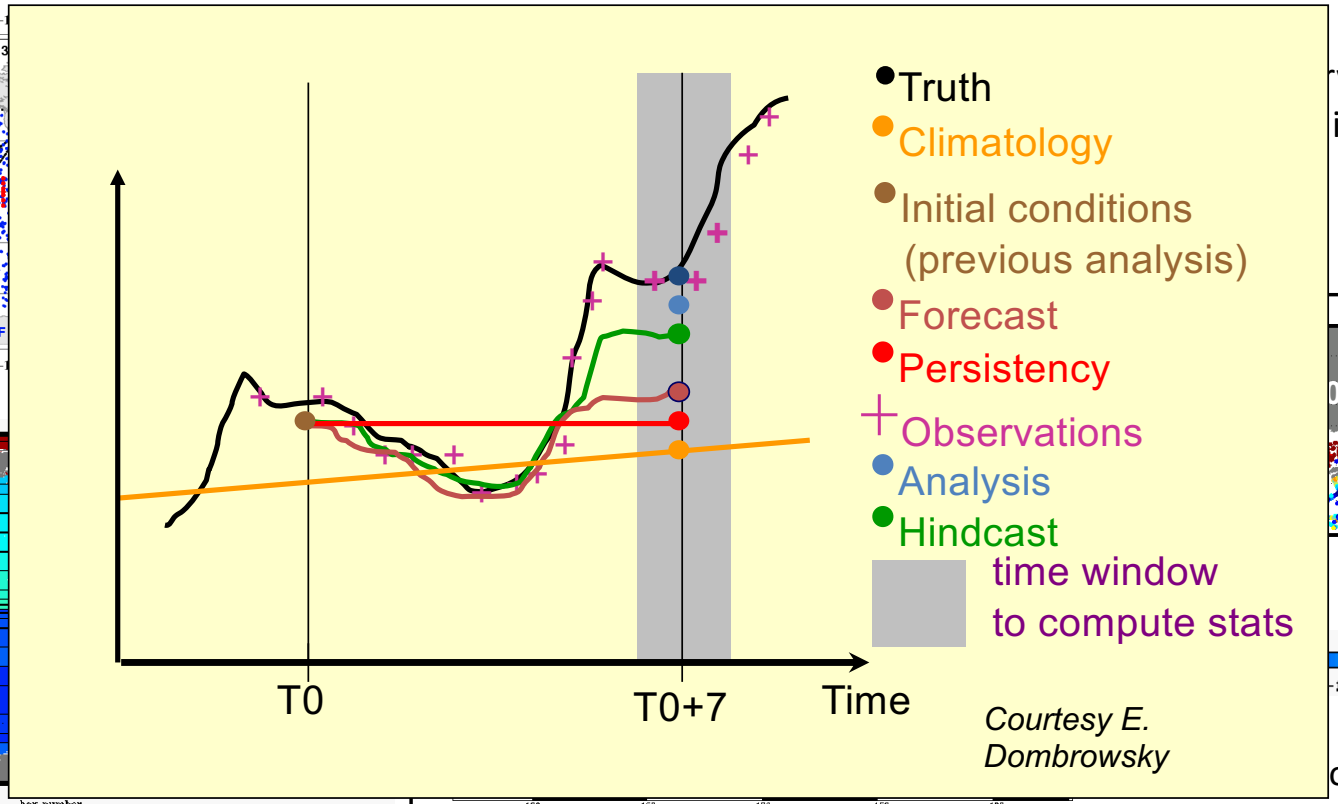


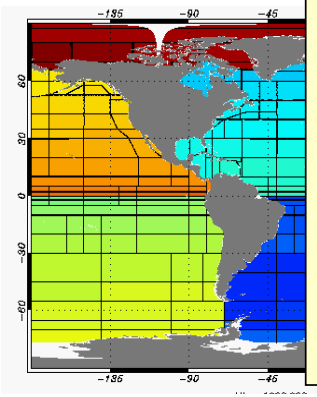
# GODAE metrics definitions: Le Provost 2002, MERSEA Strand 1

- **Class 4** – assessment **of forecasting capabilities through comparison** of model with assimilated and independent **observations**

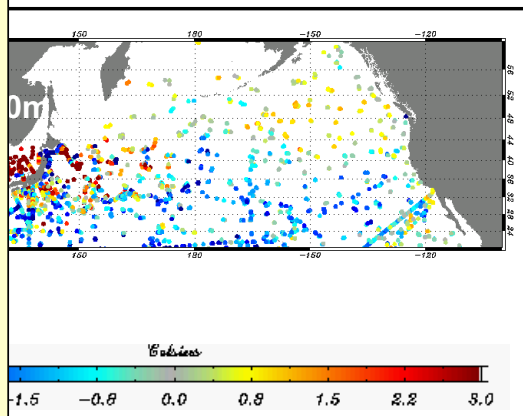



*Courtesy E. Dombrowsky*

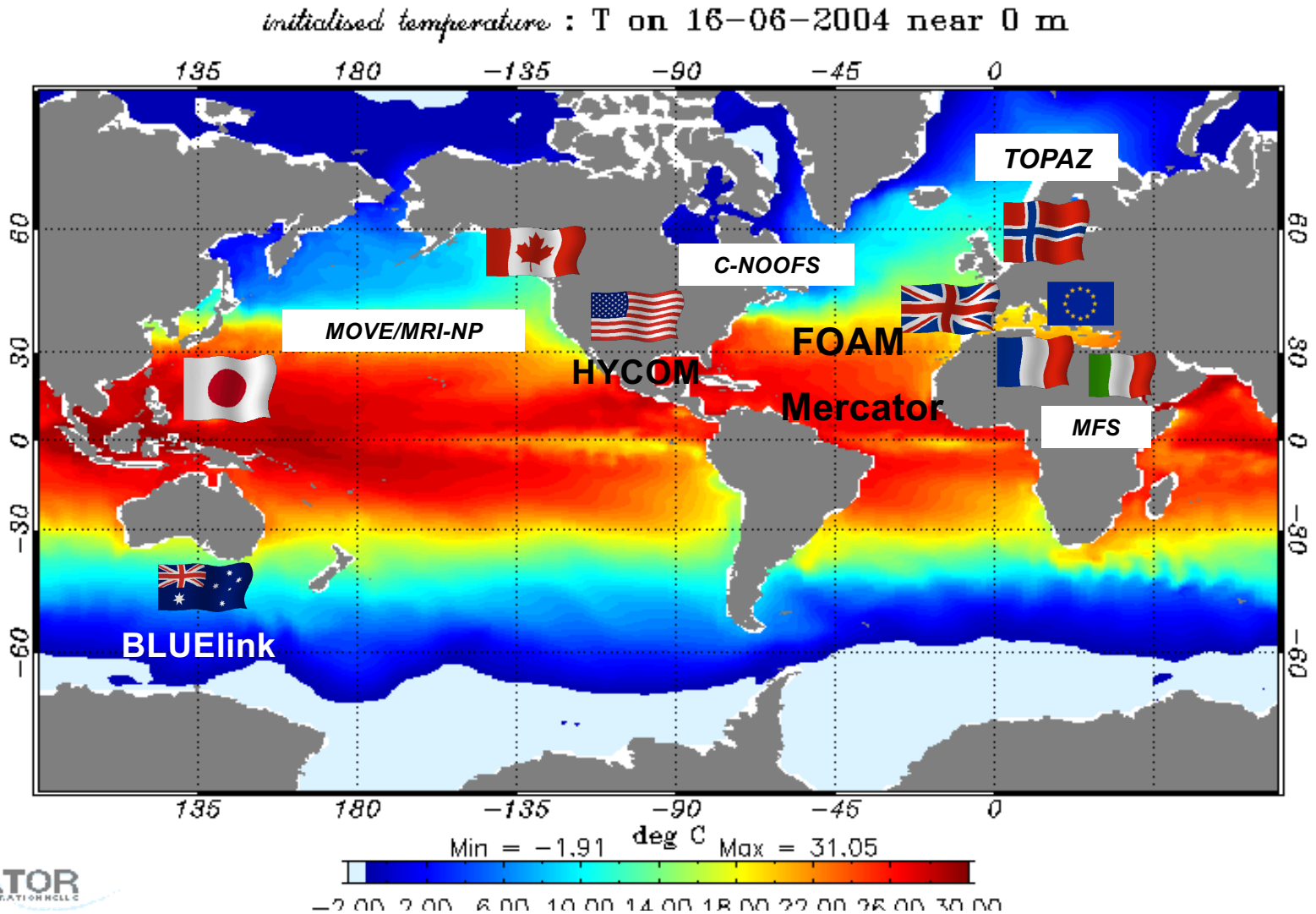


rvation (limitation
istribution)

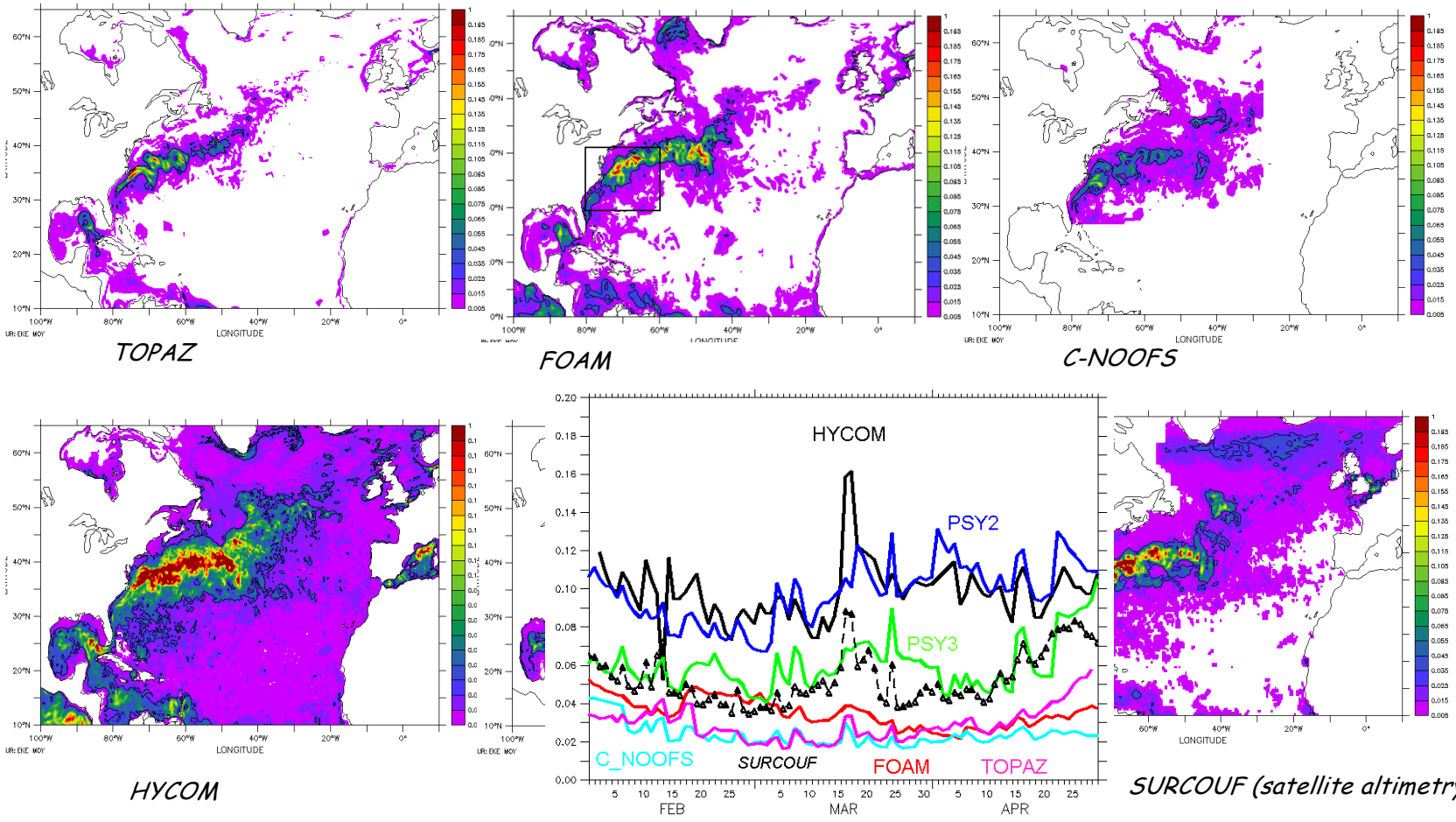


observed temperature over 1 year

# Ocean Forecasting Centres involved in GODAE Intercomparison (2002-2008)



# Assessment of Eddy Kinetic Energy (GODAE intercomparison)



# GODAE OceanView (2013→ ) Real Time Fcst Skill Monitoring (Class 4)

Ocean Predict

Environment Canada

Mercator Ocean
Ocean Forecasters

Met Office

NEMO systems:
**GIOPS**-CONCEPTS
**PSY3** and **PSY4**
**FOAM**

NMEFC (Cn)
*Remo (Br)*
JAMSTEC/MRI (Jp)
MFS (It)
*TOPAZ (Nw)*
*NRL, Rutgers (US)*
*CREG (Ca)*

Global **RTOFS** HYCOM

NOAA
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
U.S. DEPARTMENT OF COMMERCE

NCEP
NATIONAL CENTERS FOR ENVIRONMENTAL PREDICTION

NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

CSIRO

ROYAL AUSTRALIAN NAVY

Australian Government
Bureau of Meteorology

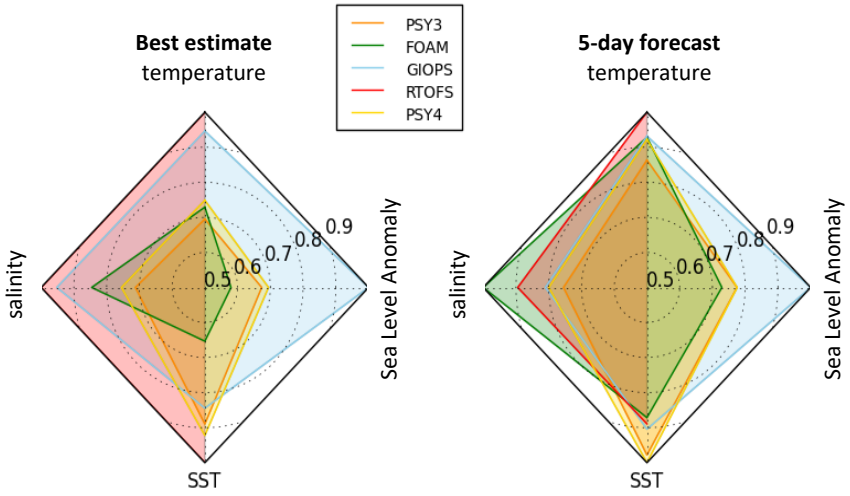
BLUELink project:
**OceanMAPS**
(MOM4)

# Intercomparison and Validation Task Team (**IV-TT**)

Coordinates and promotes the development of a framework for the scientific validation and intercomparison of operational oceanography systems (OOFS) by

- Fostering **scientific discussions** on ocean physical and biogeochemical systems validation, link with CLIVAR and Climate community on common interests
- **defining metrics** to assess the quality of analyses and forecasts ocean products
- Offering **multi-system demonstrations** and visibility to the community, link with JCOMM
- leading to **improvements** of GOV's systems

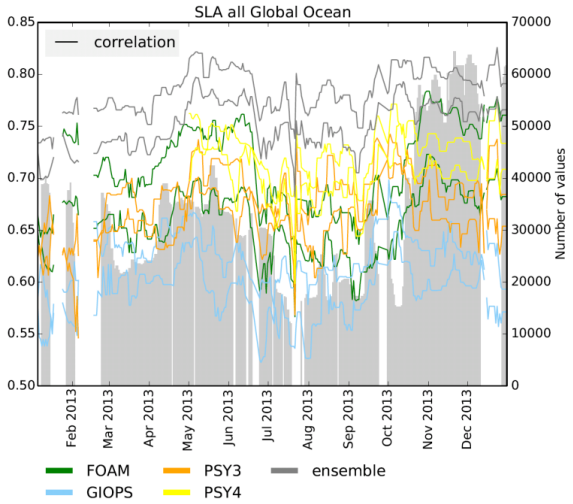
**Forecast accuracy of GODAE systems in 2013**

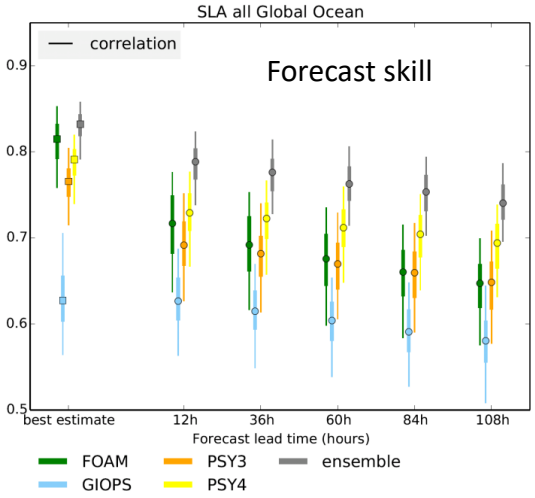



**Real time multi-assessment, and intercomparison started in 2013**

New metrics: Radar chart synthesis from the statistics of validation against observations and intercomparison of 5 operational global forecasting systems during 2013
Scores for 4 Ocean Essential Variables are provided along the four axes, normalized by the largest error

*Hernandez et al, JOO 2015*

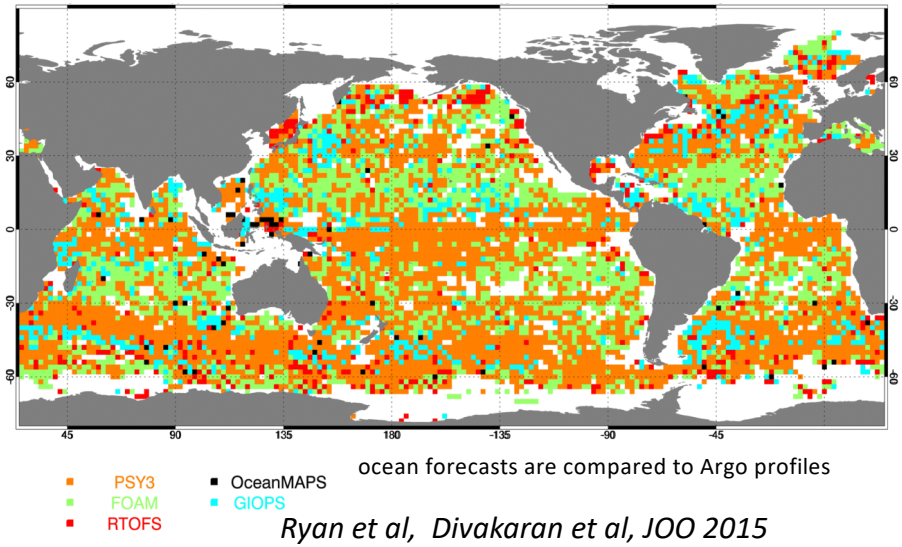



## Real time multi-assessment, and intercomparison started in 2013: demonstration and monitoring

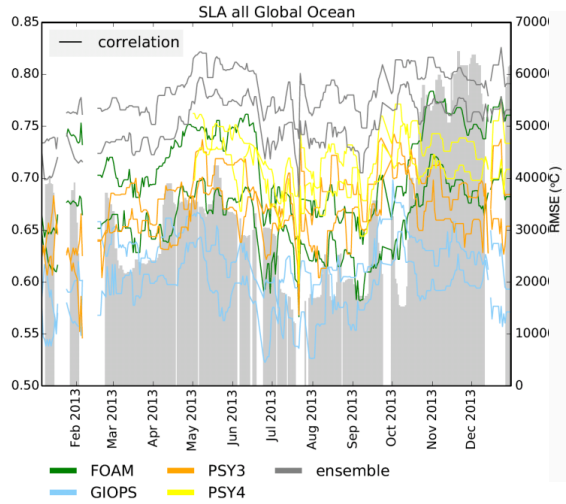
Forecast accuracy with Class 4 metrics: comparison of 4 operational forecast against satellite altimeter sea level anomalies (correlation)
Together with the assessment of the multi-system ensemble estimates (grey)

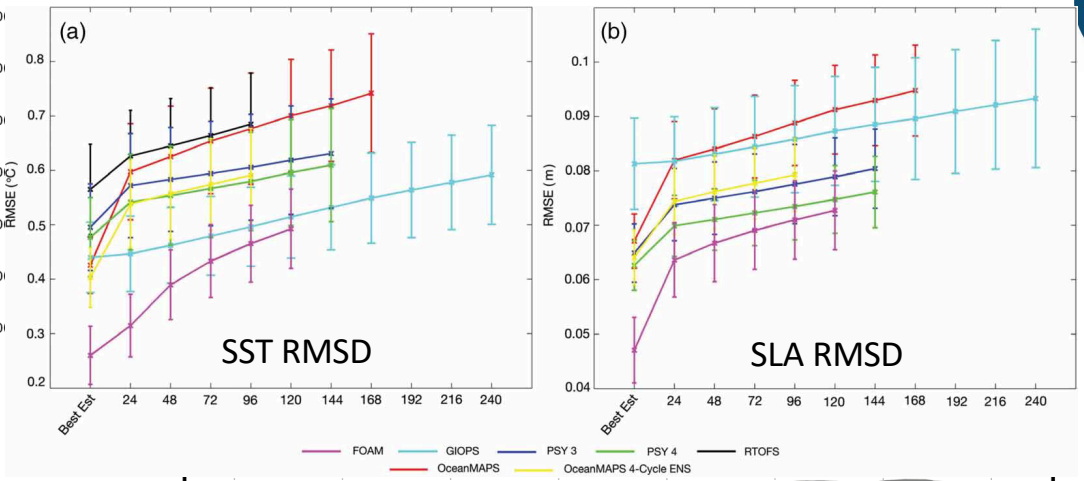
0-100m salinity 1-day forecast: which system performs best in 2013

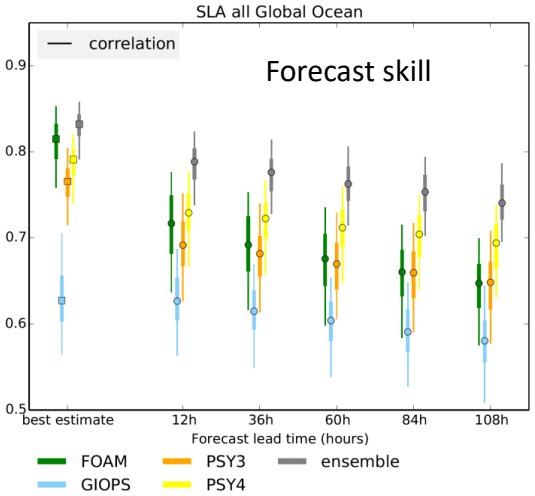


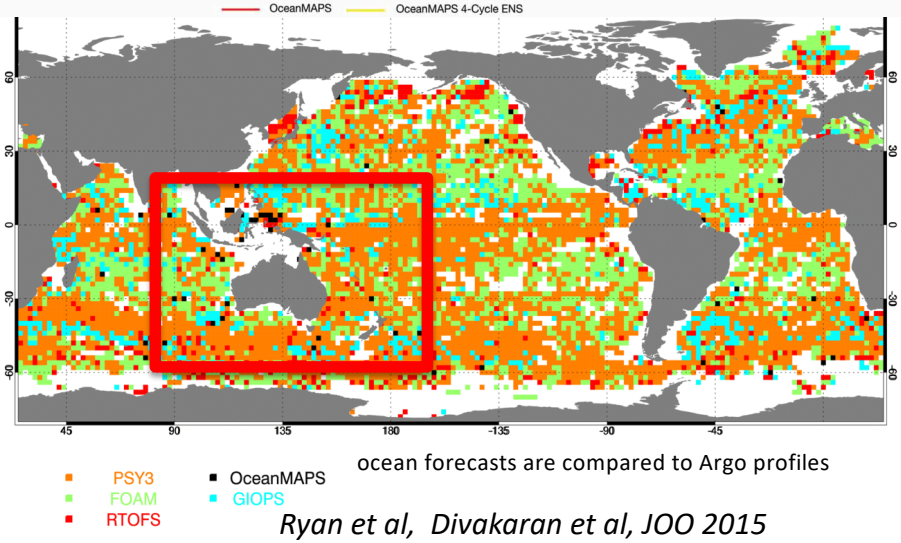

ocean forecasts are compared to Argo profiles

*Ryan et al, Divakaran et al, JOO 2015*



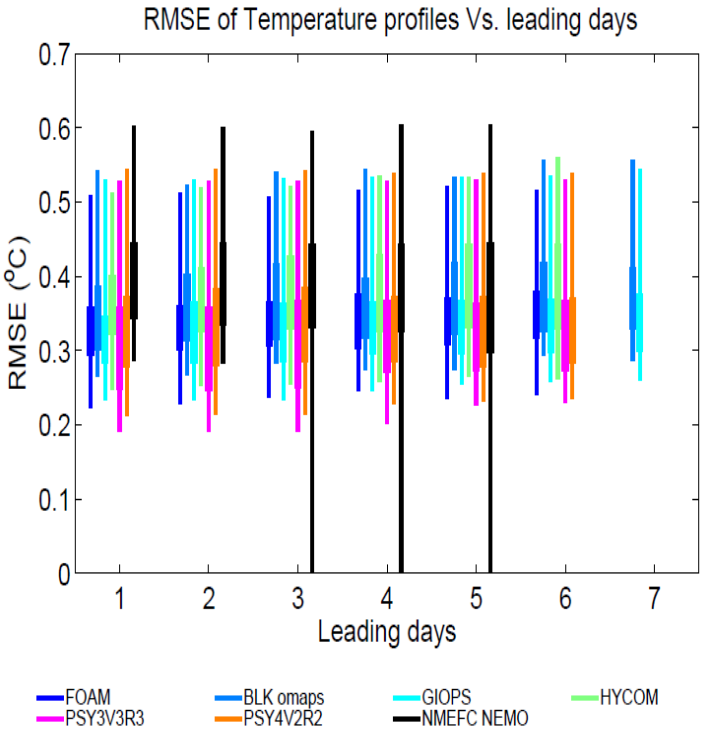



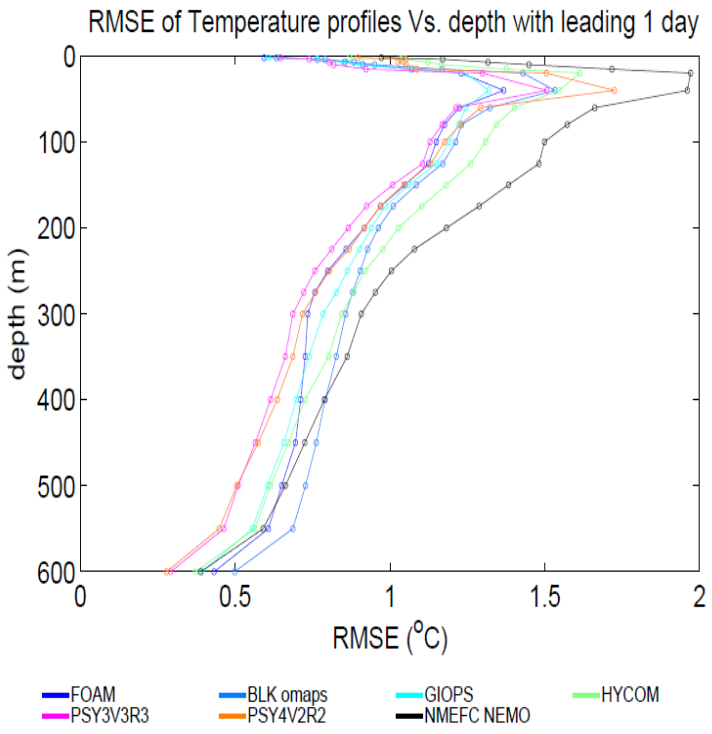




ocean forecasts are compared to Argo profiles

*Ryan et al, Divakaran et al, JOO 2015*

# New contribution: Class 4 Temperature assessment for Chinese global NMEFC system (NEMO +3DVAR/Nudging)



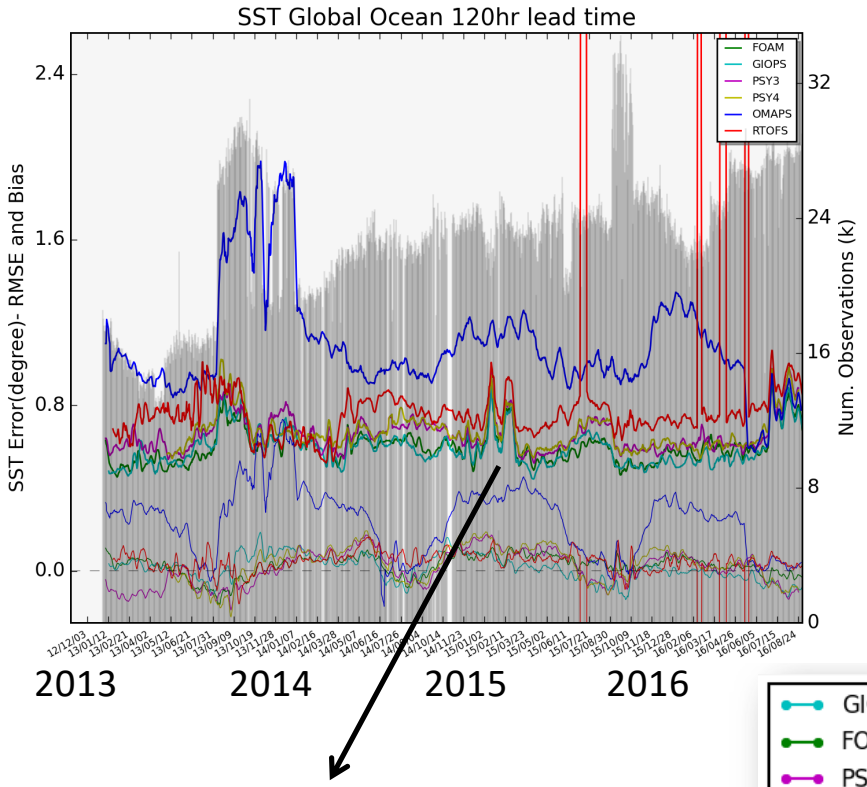




**Huier Mo, Yu Zhang, Zu ziqing and Yinghao Qin**
**National Marine Environmental Forecasting Center,**
***State Oceanic Administration, China***

# Monitoring GOV systems performance

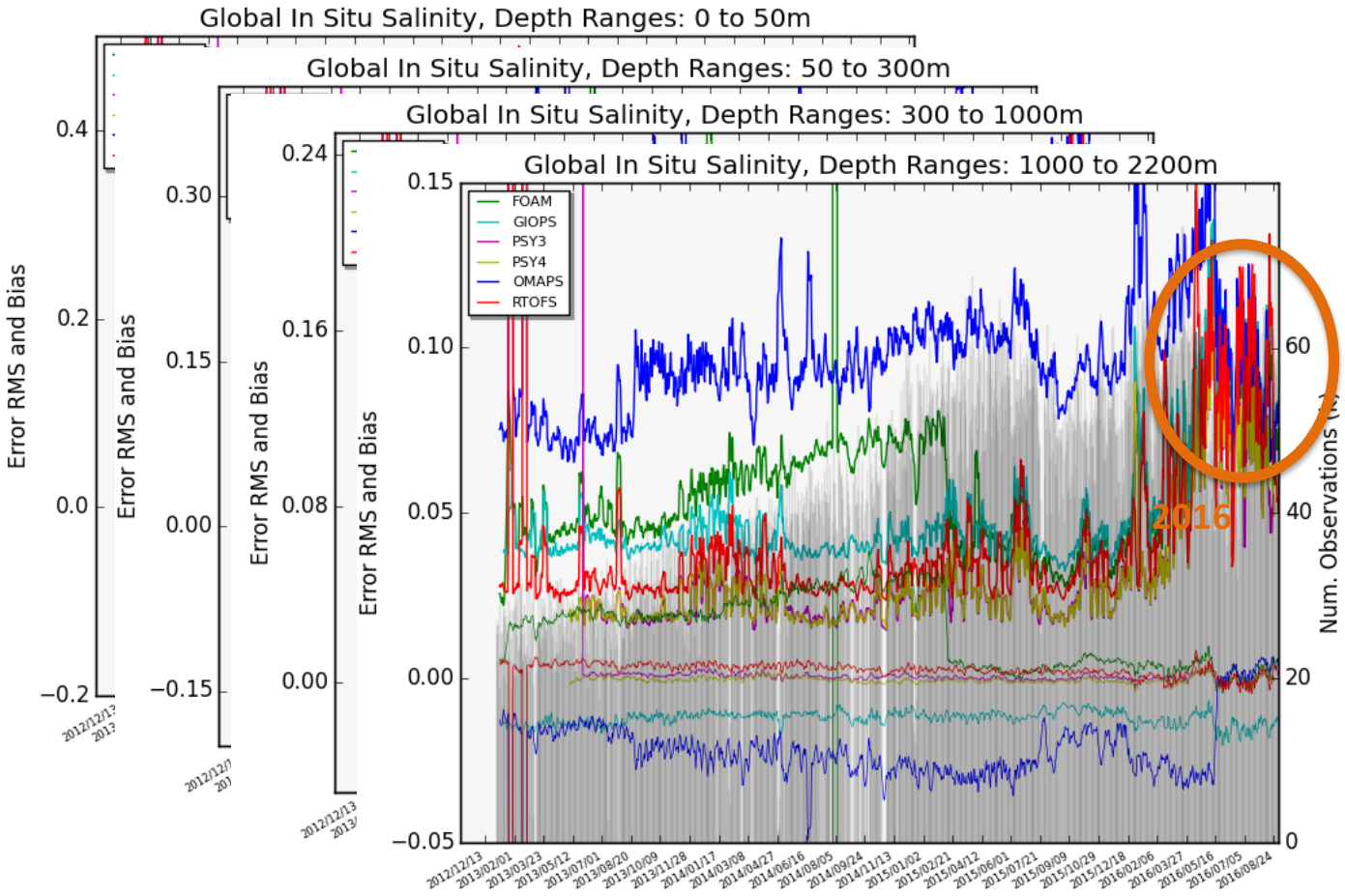
## Global Sea Surface Temperature
## 5 day lead time

- Strong seasonality of bias and RMSD can be seen.
- Strong sensitivity to spatial variability in observation coverage
- Several interesting anomalies with strongly correlated errors
  - E.g. Jan-Feb 2015
- Large error in RTOFS since July 2015 (now corrected) symptomatic of processing issues experienced by various groups.
  - ➢ Improved quality assurance procedures and robustness required for operationalization.




Jan-Feb 2015: Errors strongly correlated between systems.

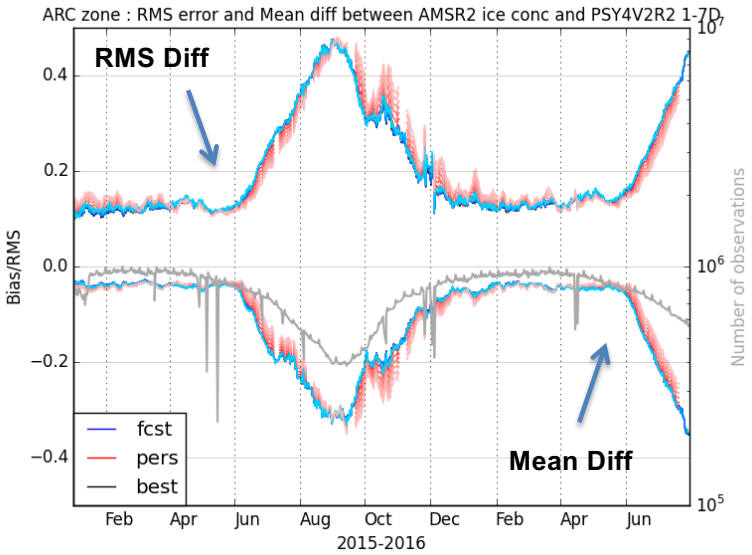
# Monitoring GOV systems performance: Salinity multiyear timeseries

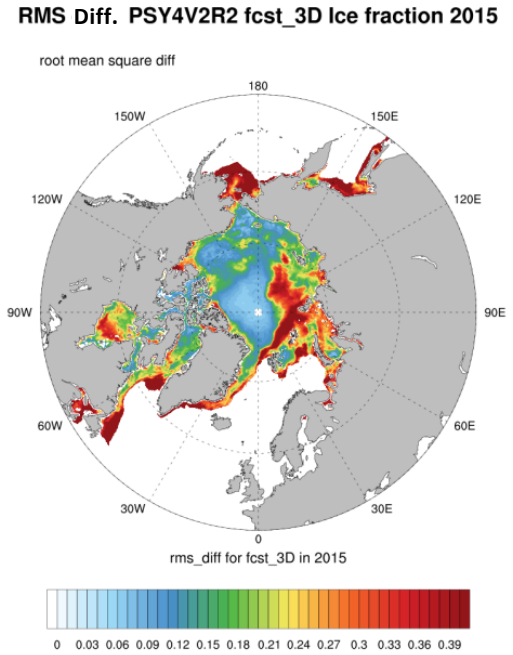



More profiles at depth, more variability in the systems ?

# Global Monitoring of sea ice concentration: Class 4 with AMSR-2

- Compute simple statistics on a stereopolar grid to plot timeseries and 2D map of the error
- *Timeseries* *2D Map*




***C. Regnier, Mercator Ocean***

- Feed in place since mid-2014.
  - Initially only GIOPS and Mercator (PSY4) contributing.
  - UK Met (FOAM) joined in 2016
- Scores calculated using contingency table metrics:
  - Proportion correct total (PCT)
  - Proportion correct Ice (PCI)
  - Proportion correct Water (PCW)
  - Range [0,1]; 1 is perfect score

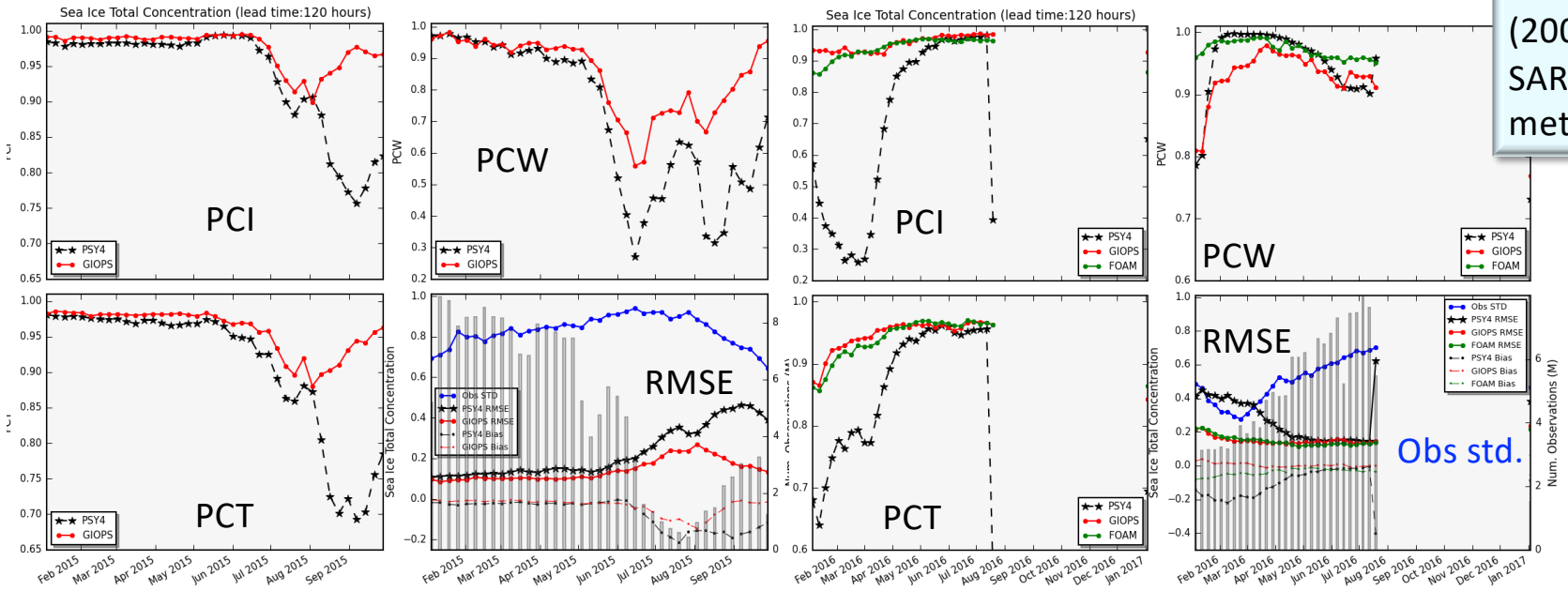
Categorical metrics for sea-ice

| | AMSR2 Ice | AMSR2 Water |
|---|---|---|
| Forecast Ice | Hit ice | False Alarm |
| Forecast Water | Miss | Hit water |

Sea ice introduces challenges as "concentration" no longer relevant at small scales (200m) now observed by SAR requiring binary metrics.

## Northern Hemisphere

## Southern Hemisphere




***G. Smith ,ECCC***

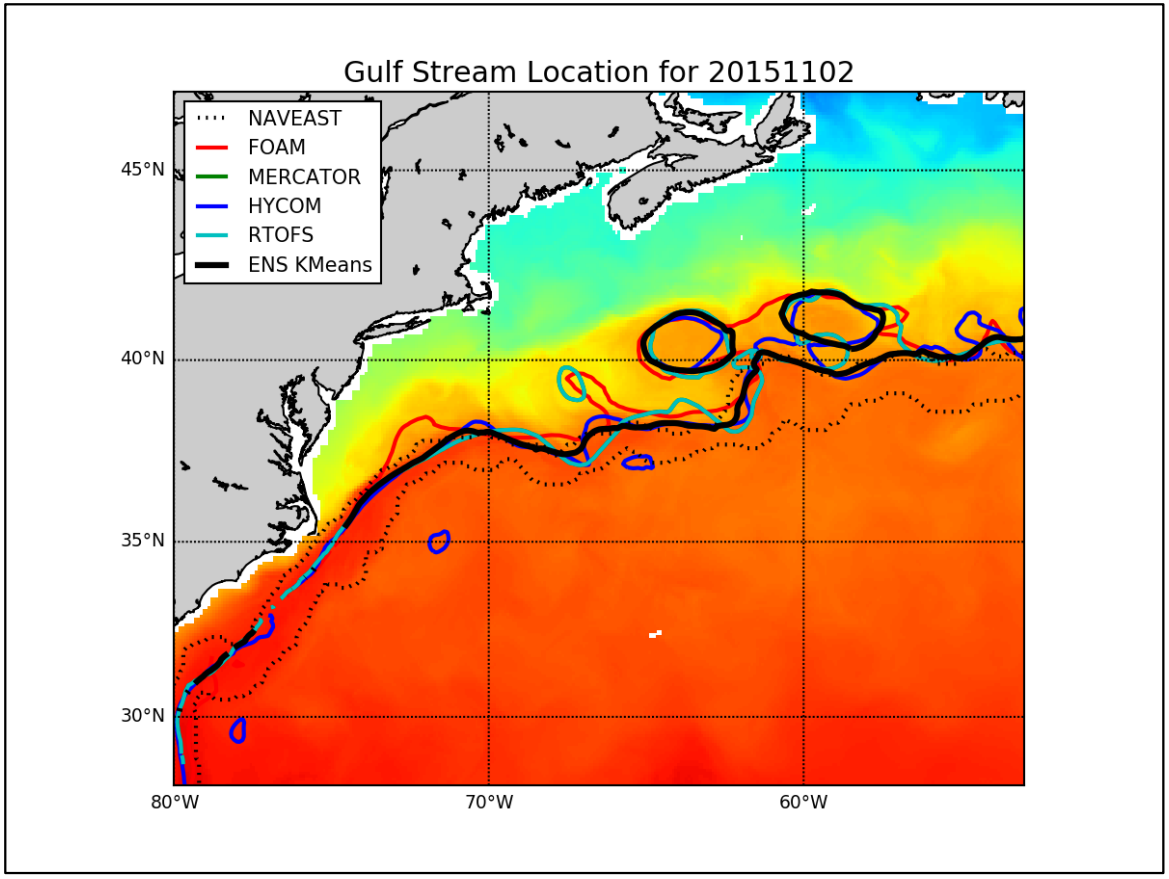
## Comparing surface 2D fcst:

## CLASS 1: Ensemble Approaches
## Consensus Forecast

- Started in 2013, **as a limited effort activity**
- in order to offer **real time opportunities** for participants, able to enhance emergency forecast tasks/decisions with complementary estimated
- Focusing **on surface parameters**, daily forecast
- Every OOFS participant asked to maintain a 2-3 month rolling archive on ftp

- **Model state** assessment, complementary to the Class 4 observation space approaches
- Decision to **standardize on common grids** for at “eddy resolving scales” (1/12°)
- Extend **to full depth parameters**, **short term forecast**
- Repository issues

# Class 1 comparison: Gulf Stream Location




Todd Spindler (NOAA/NCEP)

**Front location defined as the intersection of the 12°C isotherm and the 400m isobath**

**(Halkin and Rossby, 1985)**

**Gulf Stream north and south walls from the Naval Eastern Ocean Center**

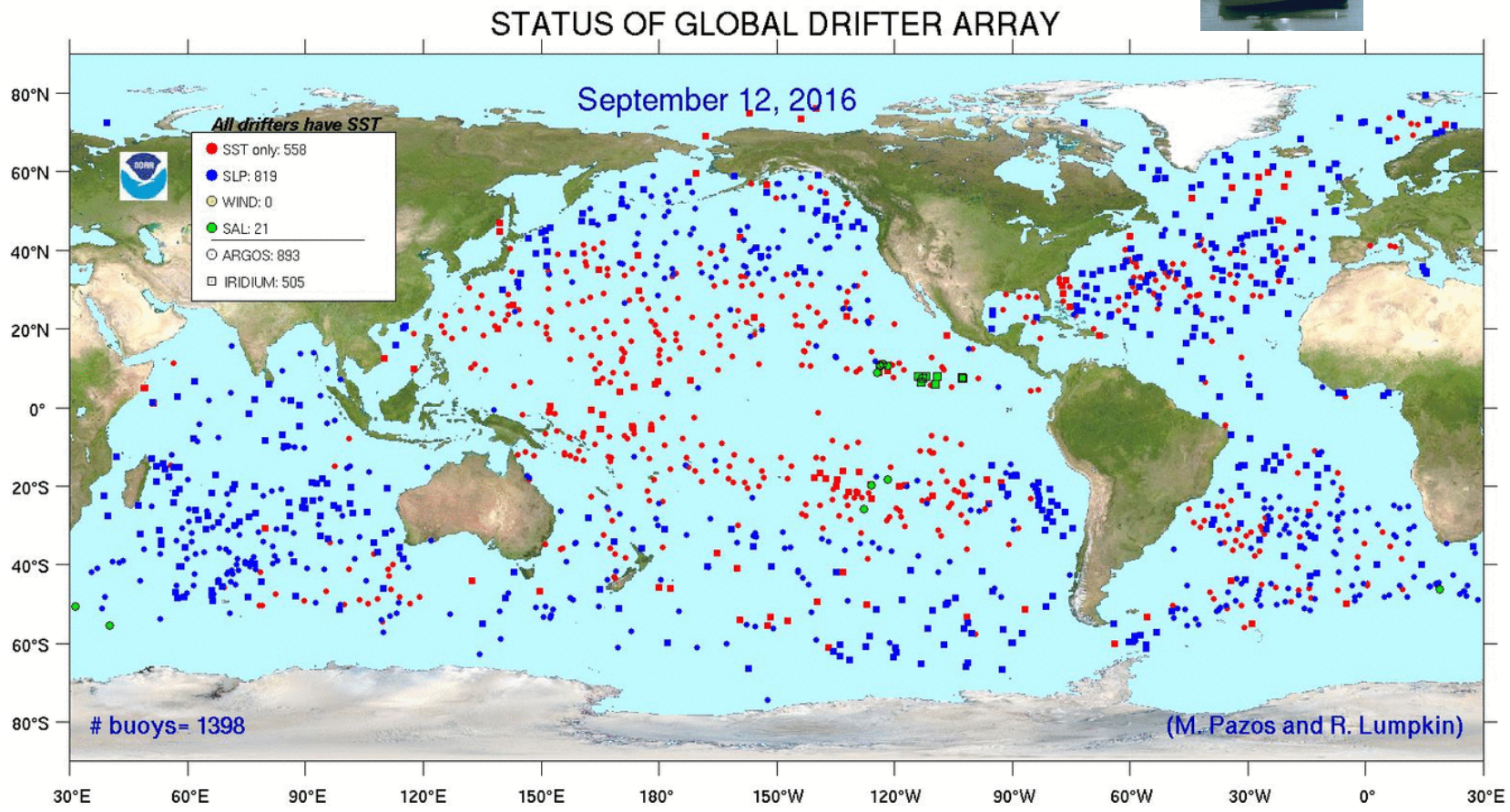
- Testing ensemble mean against individual forecast
- Testing differents ways to produce ensemble mean (K-clustering…)

Ocean Predict

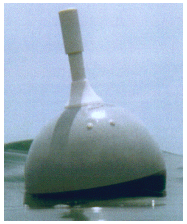
Ongoing task, future metrics….

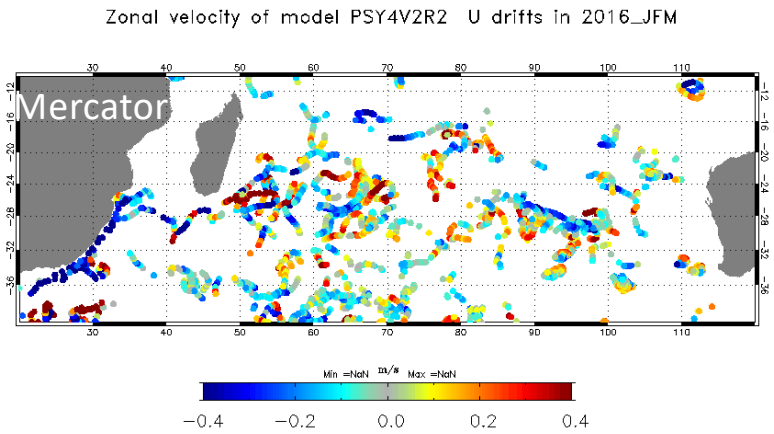
# Global Drifter Program (GDP)

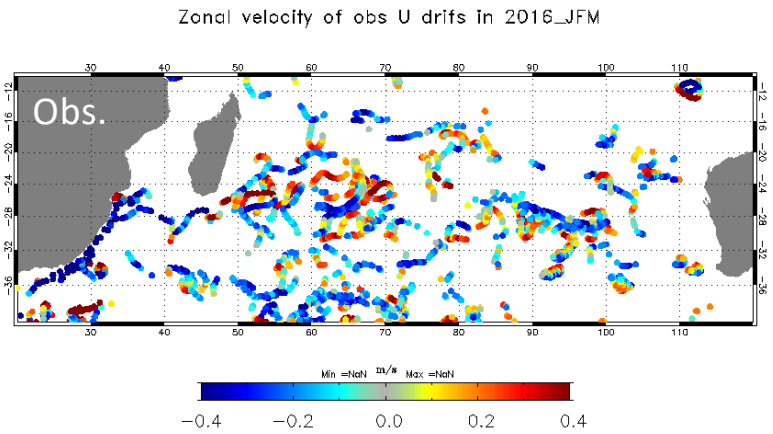
Ocean Predict

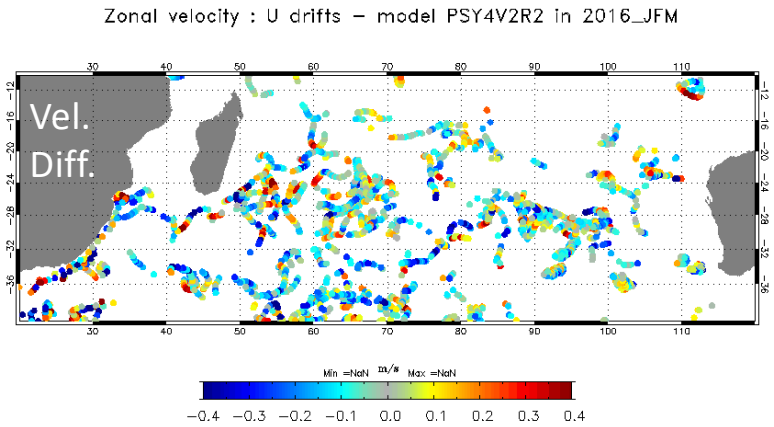


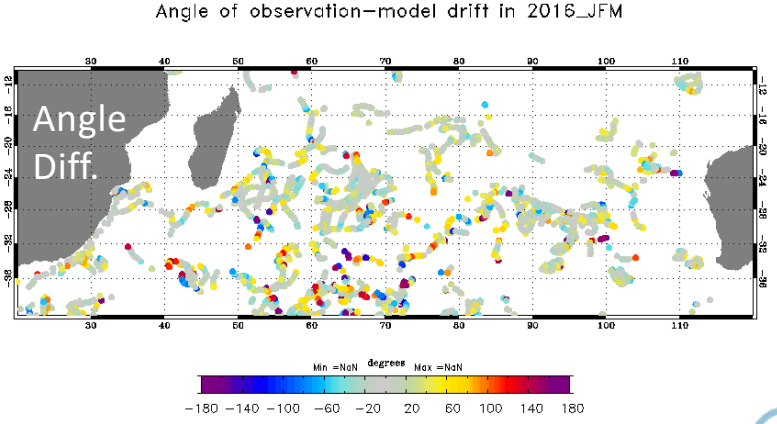
# Class 4 surface velocity assessment against drifters



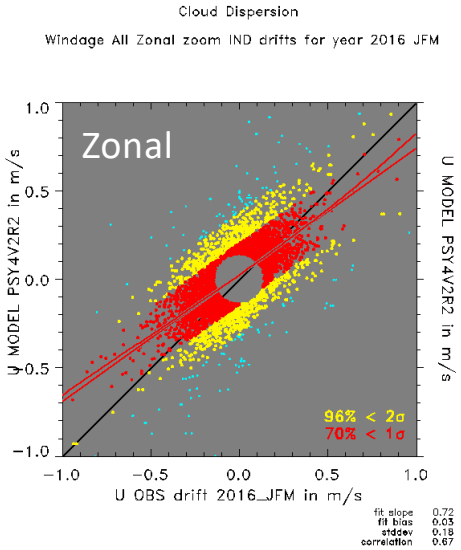


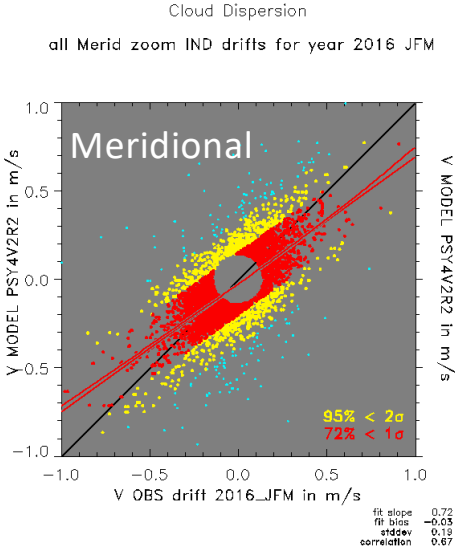



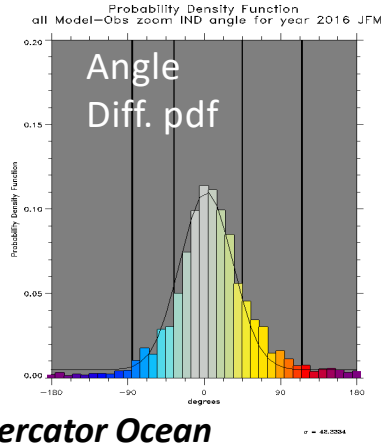



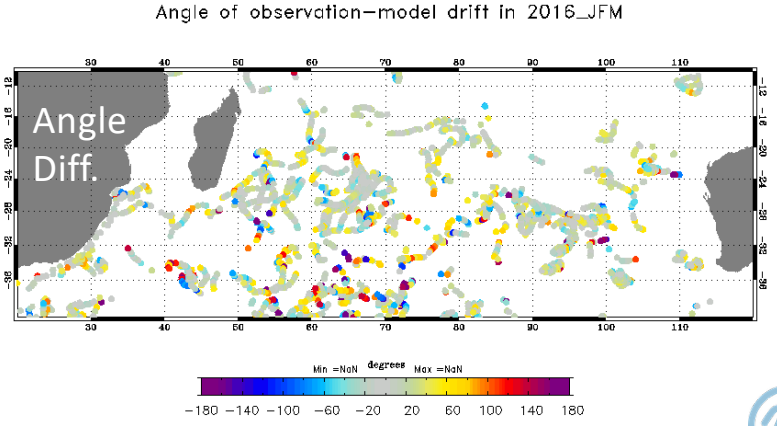




***C. Regnier, Mercator Ocean***



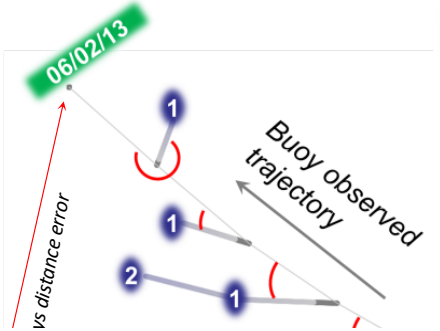



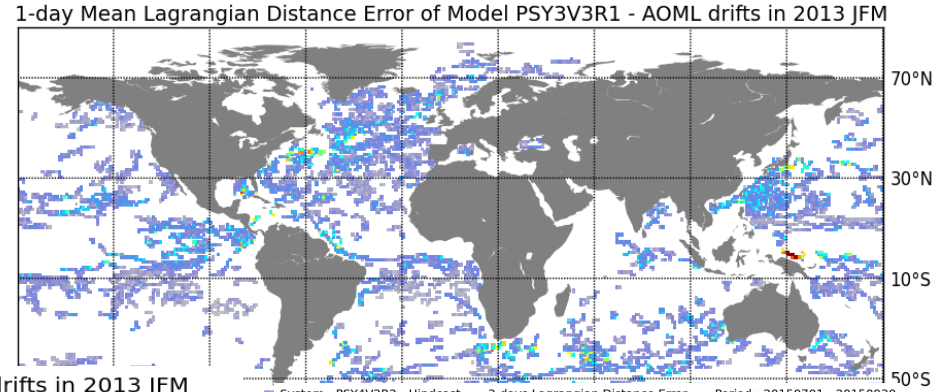



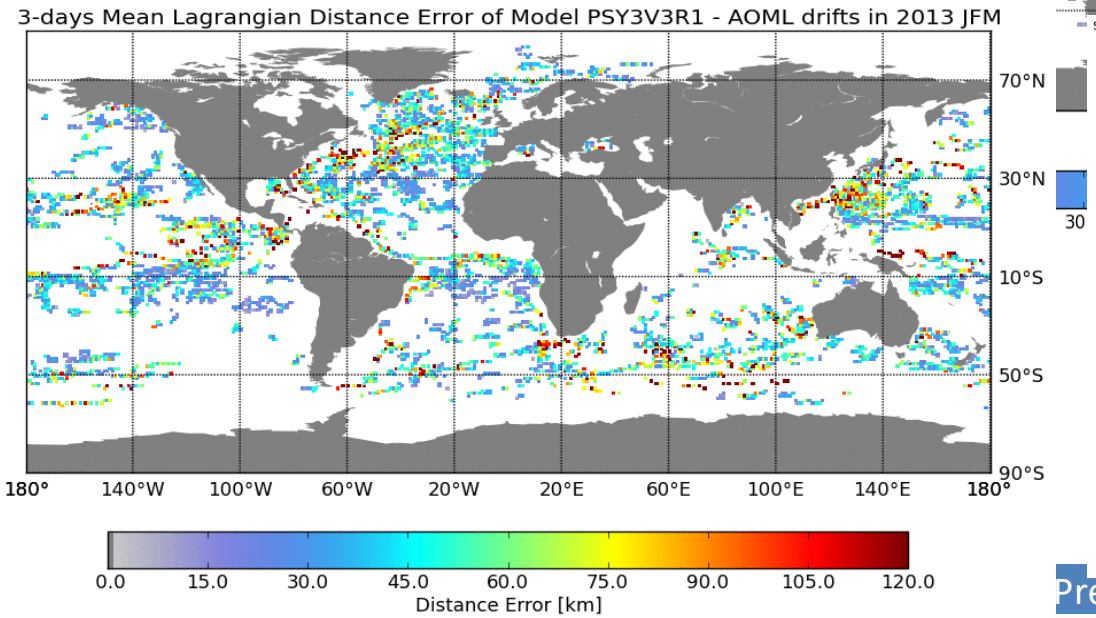

***C. Regnier, Mercator Ocean***

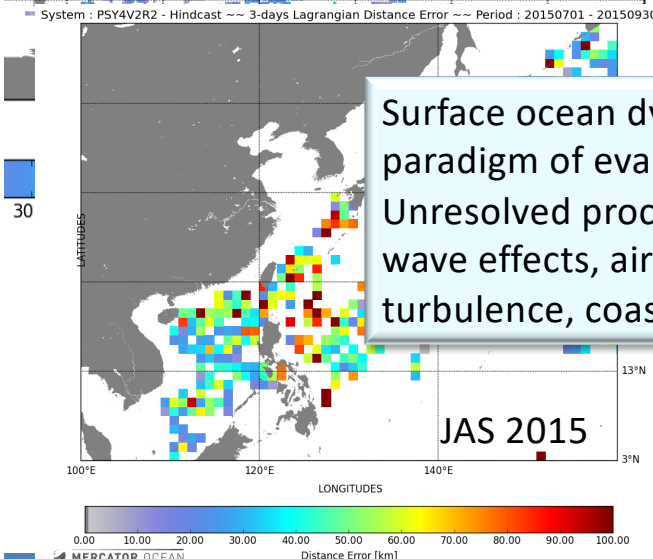
# Derived quantities with drifters :Lagrangian Metrics

*C. Regnier, Mercator Ocean*



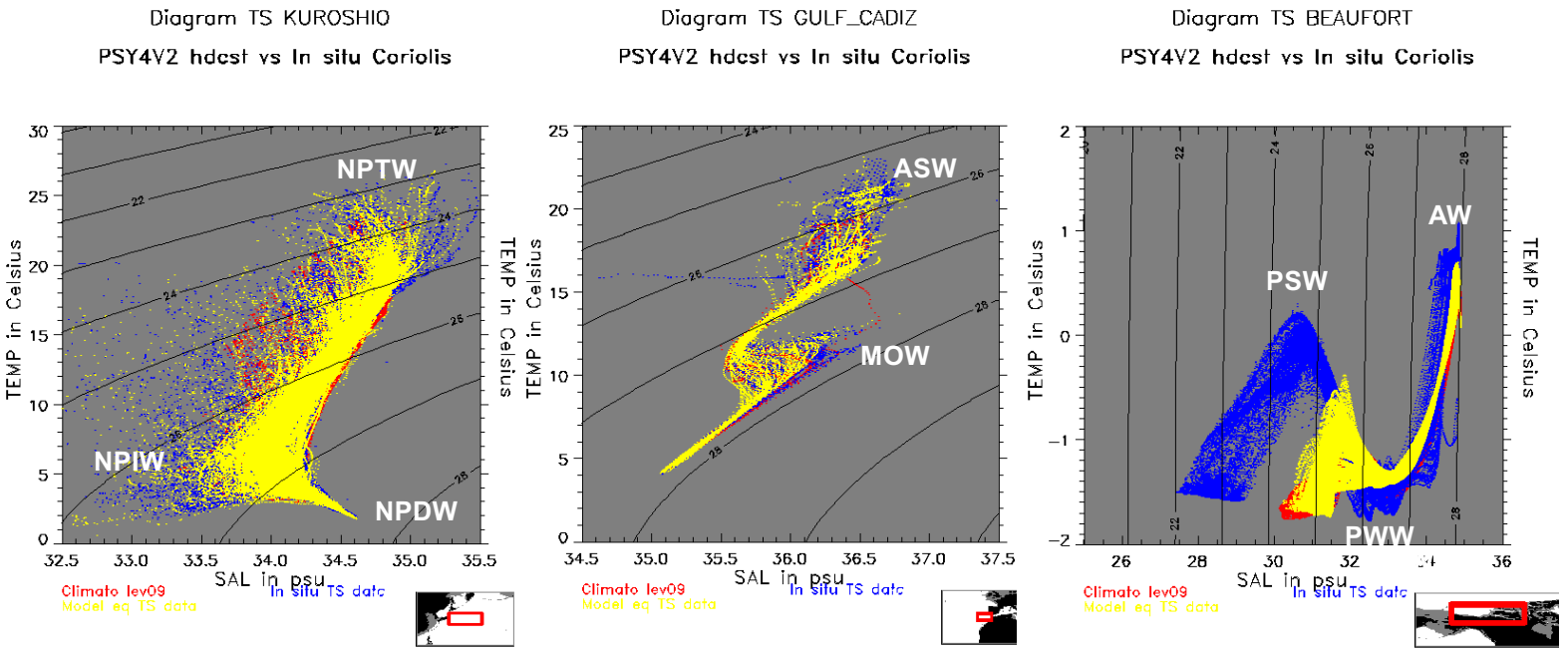




Surface ocean dynamics: the paradigm of evaluation limits: Unresolved processes and coupling : wave effects, air-sea exchanges, turbulence, coastal processes

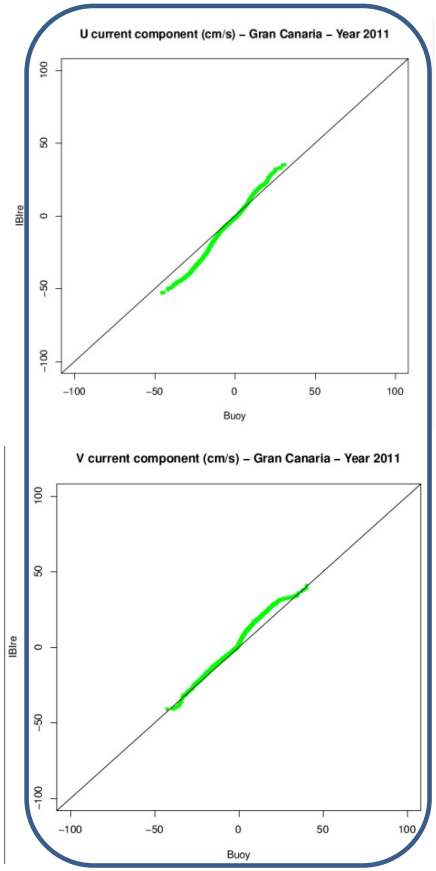
# Next: Class-4 process-oriented metrics
# T/S : control of Water masses



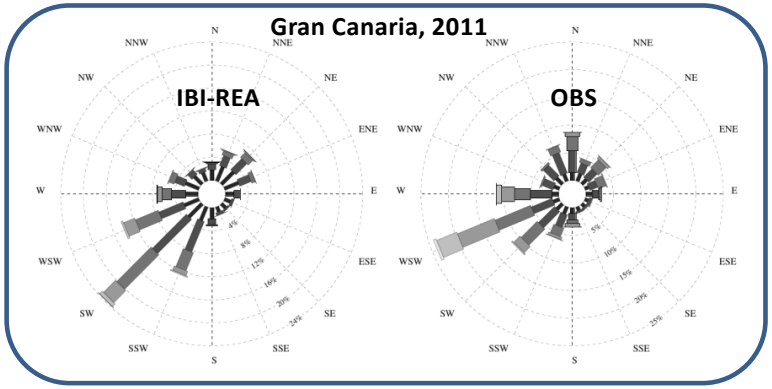

- T/S diagrams of well known water masses
- Identify some missing or not well represented water masses in operational systems

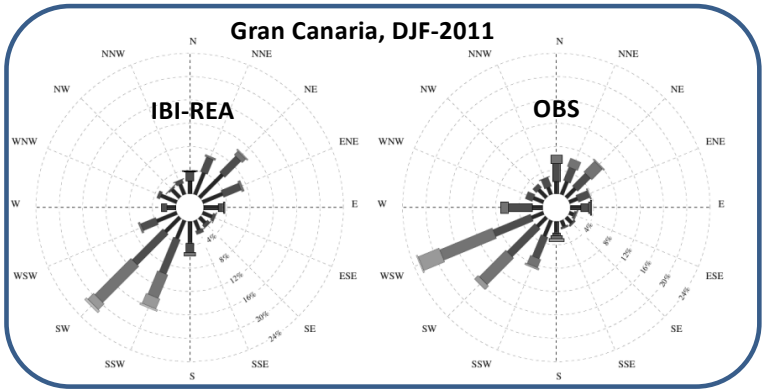
***C. Regnier, Mercator Ocean***

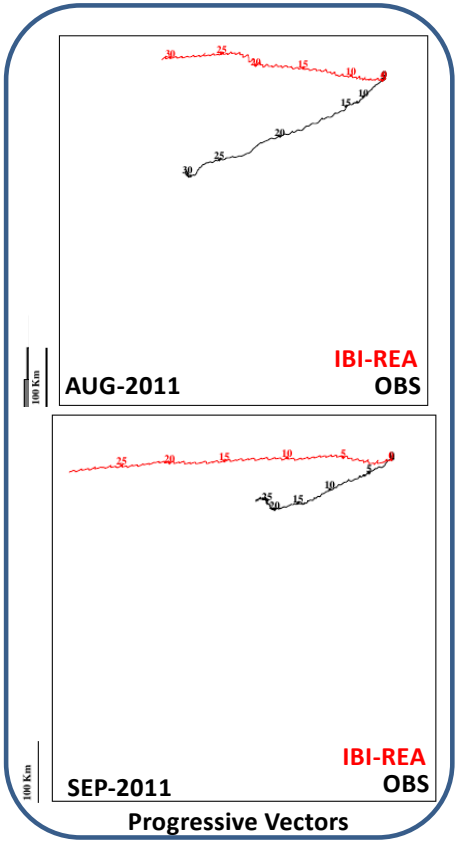
# High freq IBI Reanalysis Currents




***UV-CLASS2-HF-MOORINGS***



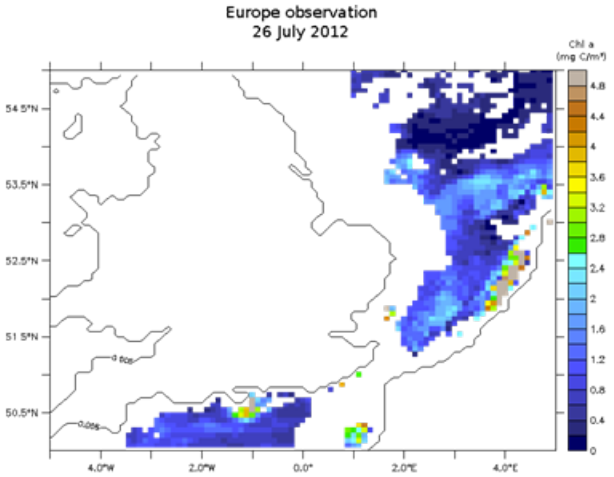





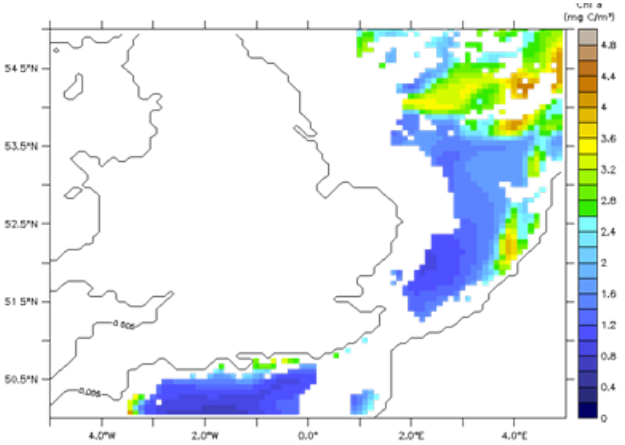
*Courtesy of M. Garcia Sotillo, Puertos del Estado, Spain*

# Characterize Chl skill scores: adapting categorical metrics

## Satellite observation




## CMEMS FOAM model



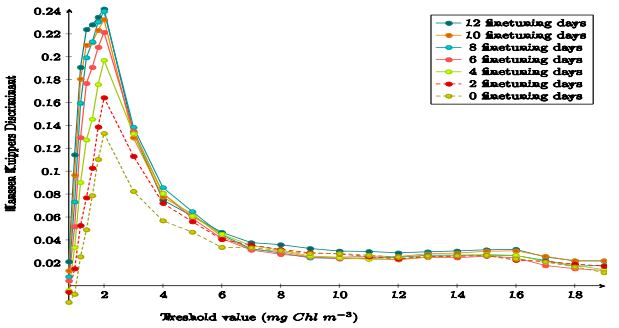
| Event forecast | Event observed | | |
|---|---|---|---|
| | Yes | No | Marginal total |
| Yes | a | b | a + b |
| No | c | d | c + d |
| Marginal total | a + c | b + d | a + b + c + d =n |

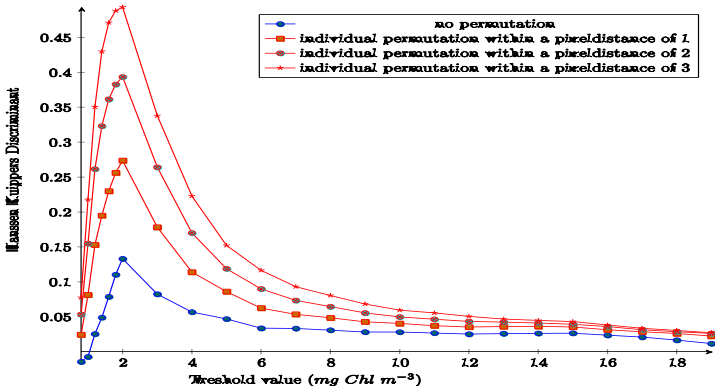
*Courtesy of J. Maksymczuk, MetOffice*

Hanssen-Kuippers Discriminant (1965)
KSS: (ad-bc) / (a+b)(c+d)

Range: -1 to 1, 0 indicates no skill.
Perfect score: 1
(also known as True Skill Statistic)

Varying temporal window



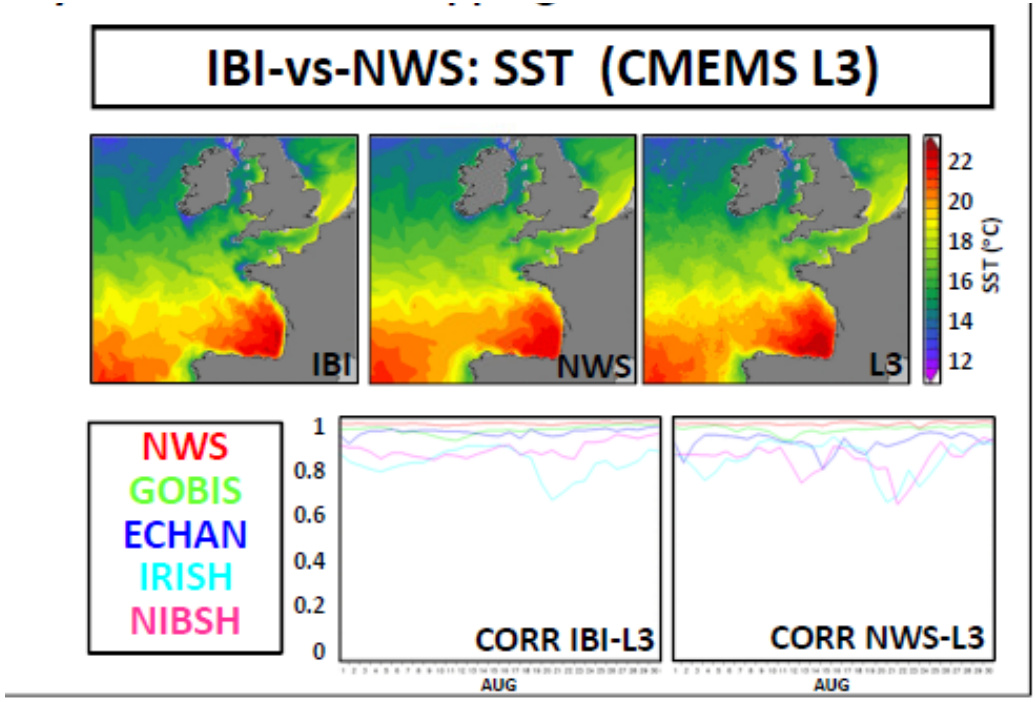

Varying spatial window

## Toward multi model assessment and ensemble approaches

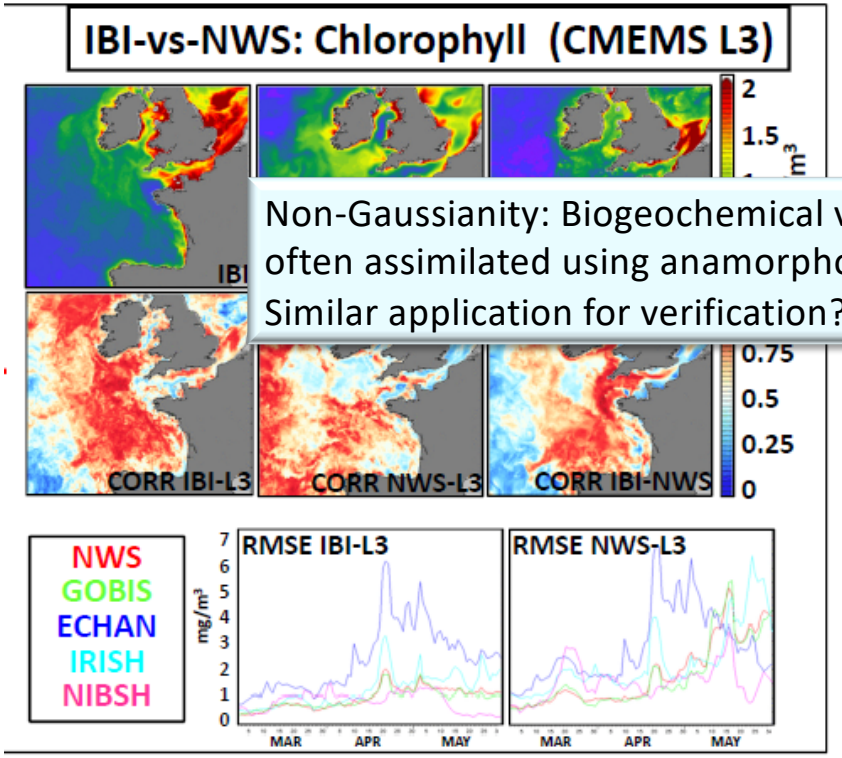
- Why performing MME and Ensembles?
  - Part of the **errors** in observed products and model simulation are **not correlated**
  - Brings to community efforts, exchanges, and faster advances in assessment expertise
  - Many ocean forecasting system are going **probabilistics**: but framework for generating and **evaluating spread not yet established**.
- In practice, MME and Ensemble allow:
  - Evaluate spread as a proxy for overall error, or a probabilistic estimator
  - Compute the ensemble mean, and take benefit of it in many applications
  - Identify outliers of the ensemble

# CMEMS Intercomparison activity between NWS and IBI (NARVAL)

**August 2016**




**Spring 2016**




Non-Gaussianity: Biogeochemical variables often assimilated using anamorphosis. Similar application for verification?

EGU 2017
**Intercomparison of different operational oceanographic forecast products in the CMEMS IBI area**
Pablo Lorente et al.

*Courtesy of M. Tonani, UK-Metoffice*

# Take home messages

- Operational oceanography has been **developing validation and verification approaches for the last 20 years**
- **Strong challenge**: lack of ocean observations, inhomogeneous distribution (high latitude, depth), sparseness, not capable to sample the ocean mesoscale, already represented in most global systems
- **Ongoing intercomparison** experiments at the international level (OceanPredict) and raised inside some community (Europe and the Copernicus Program)
- The community is **looking for user-oriented or process-oriented metrics**, and ways to communicate to users

➢ Reasons why we are glad to **exchange with the weather forecast/climate validation and verification experts**