Benefits of interpreted vector programming and Hierarchical Data Format for statistic ocean model evaluation

Paolo Oliveri[1] Alessandro Grandi[2] Emanuela Clementi[2] Simona Simoncelli[1]

paolo.oliveri@ingv.it

Goal

Provide a near real time production and delayed mode flexible evaluation system between:

1. Ocean Model data (e.g. analysis, reanalysis, etc.);
2. Insitu observations data (e.g., moorings, gliders, vessels, etc.)

Insitu observations pros-cons

- Huge and extensive source of information on the real sea conditions;
- Continuous state and quality controlled both from the data provider and the **DAC**;
- Time changing position and depth;
- Different disseminations methods, storage and sampling times;
- Not completely reliable due to the marine environment (electronic problems, durability, continuity of sampling and sensors stability).

Model data pros-cons

- 3D continuously gridded data with regular depth layers;
- Fixed and averaged sampling times;
- Methodical storage of ocean variables (e.g. per-grid or in per-field datasets);
- Completely reliable data;
- Uncertainty of numerical models solutions, even with data assimilation schemes (e.g. analysis of reanalysis).

Problem

Correct and improve data quality and port model data on insitu observations points.

Insitu data post processing

- Input: Horizontal lat, lon limits, standard_names to process, statistic iterations, time range;
- Output: Probes specifications CSV file, post processed, quality checked and time averaged per-field and per-platform datasets.

- Horizontal average and vertical rescaling for not floating devices;
- Original DAC quality control application (v_{qc});
- Gross check quality control:

$$v_{no\_spikes\_min} \le v_{qc} \le v_{no\_spikes\_max};$$

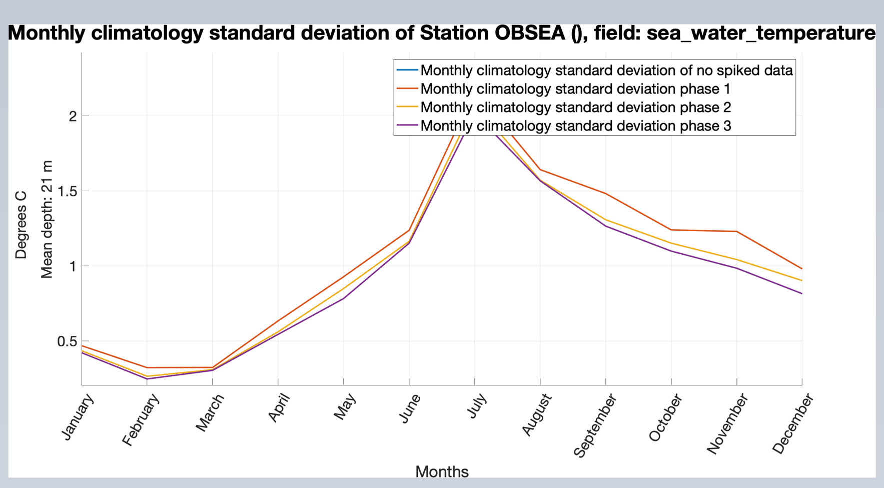
- Redundant statistic quality check: $v_{statistic\_good_{j+1}} = v_{statistic\_good_j}$ when:

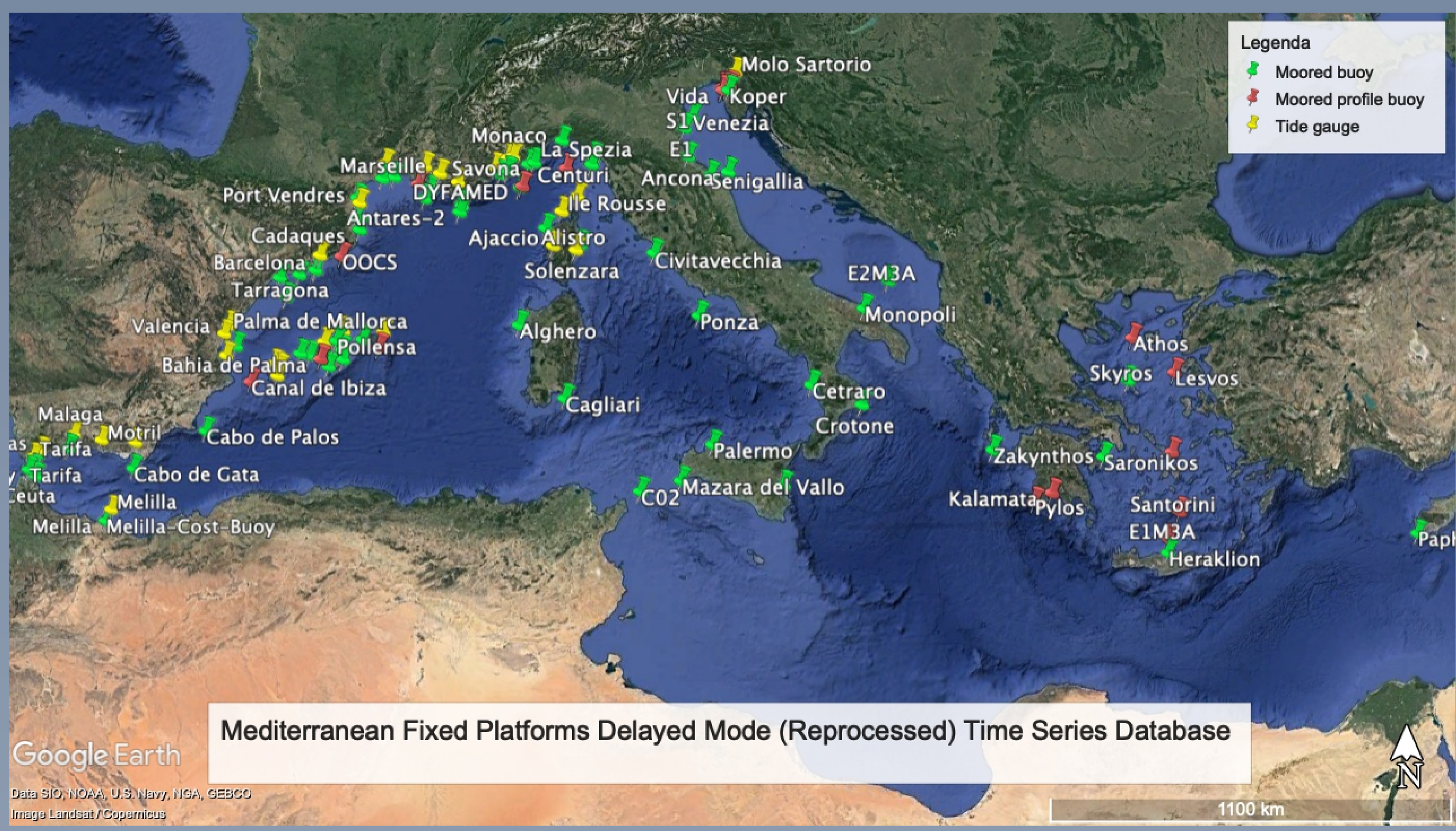
$$|v_{std\_an_j}| <= v_{std\_max} \text{ and } v_{std\_an\_dist_j} \ge 5\%$$

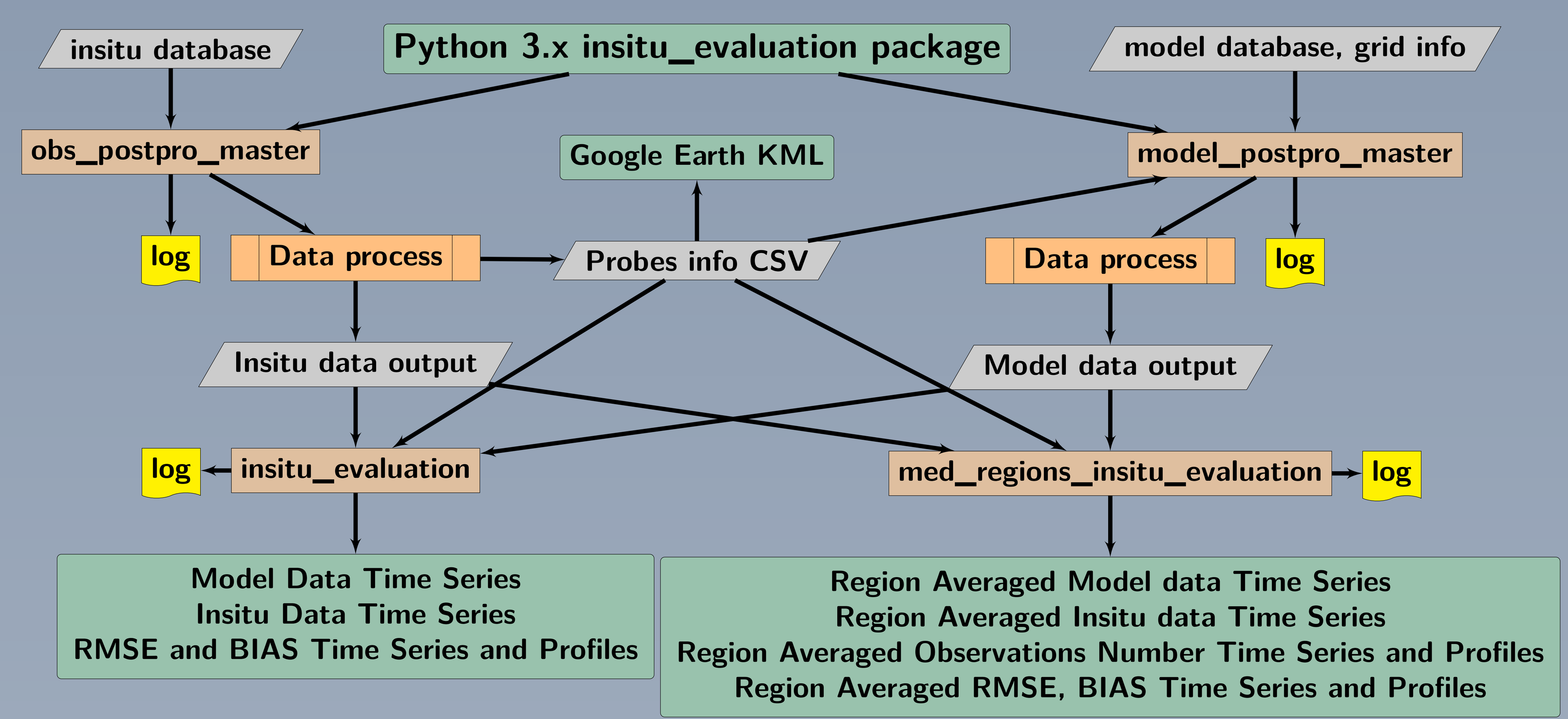
where $v_{std\_an_j} = \frac{v_{statistic\_good_j} - v_{\mu_j}}{v_{\sigma_j}}$ is the standardized anomaly, v_{std_max} is tuned for each field and $v_{std\_an\_dist_j}$ is the probability distribution of $v_{std\_an_j}$ computed by a Kernel Density Estimation.
- Derived fields computation and time averaging of processed datasets.

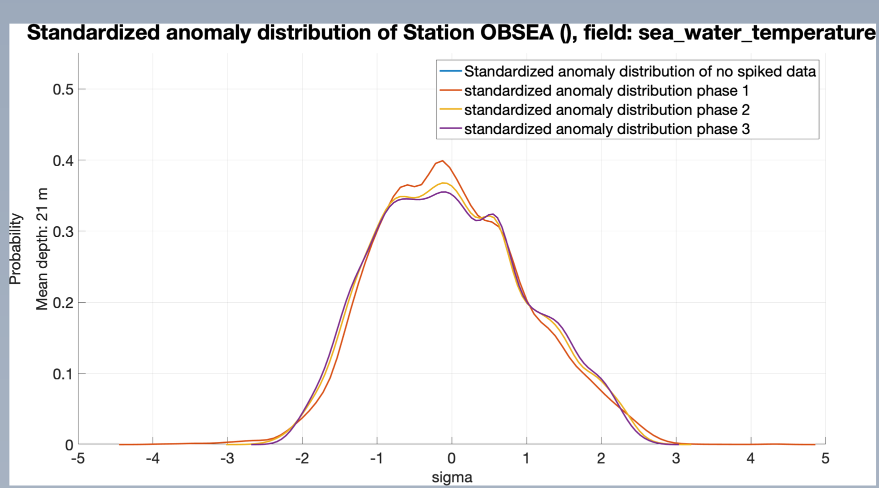
Example

If v is the sea water temperature, then v_{std_max} is set equal to 3 from surface until 10 meters, then 2.5 until 100 meters and 2 for the rest of the water column.

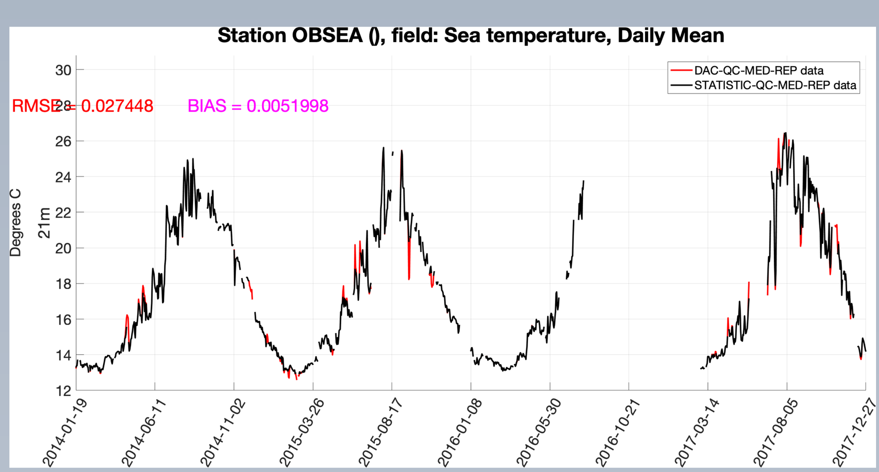








Monthly climatology standard deviation and standardized anomaly distribution for temperature field of the OBSEA platform.



DAC QC vs 3 phases statistic QC for temperature field of OBSEA platform.

Model data post processing

- Input: probes specifications CSV file from insitu part, per-grid or per field daily or hourly datasets, grid information file (e.g. mesh mask file), time range;
- Output: Location ported and vertical interpolated post processed per-field and per-platform hourly and daily mean time series.

- Input file concatenation list generation, high distance of model data from the platform check and variable extraction: $min_{dist_{\mathbb{E}}(model_{lats}, model_{lons}, obs_{lat}, obs_{lon})} V$;
- Computation of the insitu temperature from potential temperature and salinity;
- Linear vertical interpolation on platform depth levels.

Platform insitu evaluation

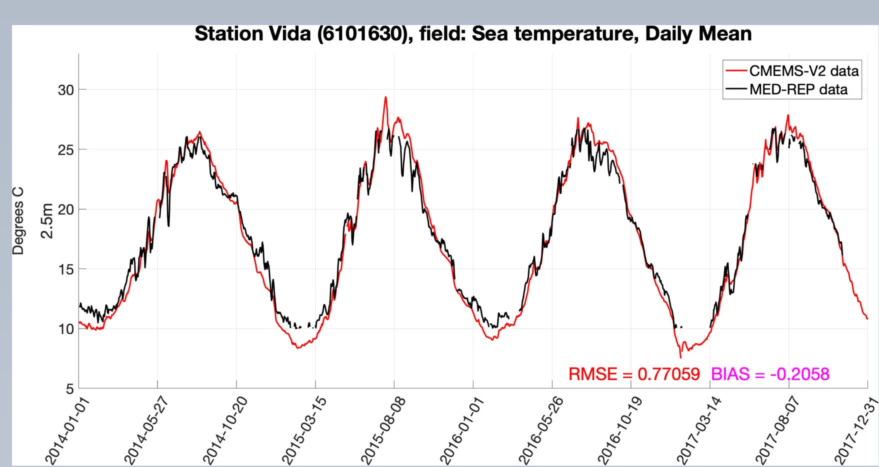
- Input: probes specifications CSV file from insitu part, post processed model datasets directory, post processed insitu datasets directory, time range;
- Output: per-field and per platform evaluation dataset.

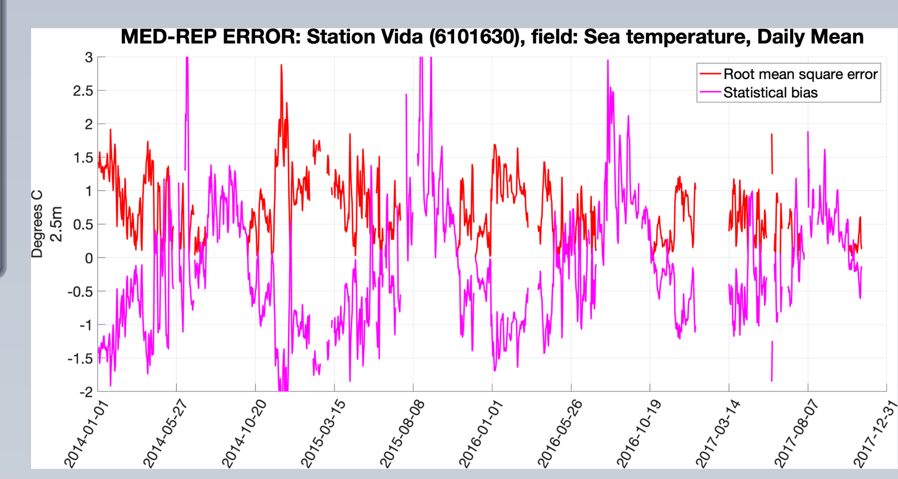
- Insitu and model grid information;
- Model time series;
- Insitu time series;
- **RMSE** time series and profiles:

$$v_{RMSE} = \sqrt{(v_{vert\_interp} - v_{statistic\_good})^2}$$

- **BIAS** time series and profiles:

$$v_{BIAS} = v_{vert\_interp} - v_{statistic\_good}$$





Output of the insitu_evaluation part of platform Vida, showing time series and errors.

Possible solution

Python-3 and NetCDF-4 **insitu_evaluation** package.

HDF and Big Data

Model data and insitu observations databases are **Big Data**. Proposed **Hierarchical Data Format 5**, for:

- High compiled and interpreted programming languages support with bindings and toolkits;
- Not required remote administration;
- Optimal metadata management;
- Speed of accessing, reading and writing datasets.

Interpreted vector programming

The selected programming language is Python 3.x, for:

- Flexibility, portability and platform independence;
- Dynamic typing and scoping;
- Smaller executable program sizes;
- User-friendliness and freedom.

The arrays must be manipulated using vectorization, in order to achieve a similar performance to compiled languages.

Test case subject

- Med. Sea 124 fixed platforms **Reprocessed** data from **CMEMS** in situ **TAC** ftp://my.cmems-du.eu/;
- (Mediterrranean Copernicus Marine ocean model analysis V2 https://doi.org/10.25423/MEDSEA_ANALYSIS_FORECAST_PHYS_006_001)
- **temperature**, **salinity**, **sea level**, **sea water speed**;
- Date range from 2014-01-01 to 2017-12-31.

Med regions insitu evaluation

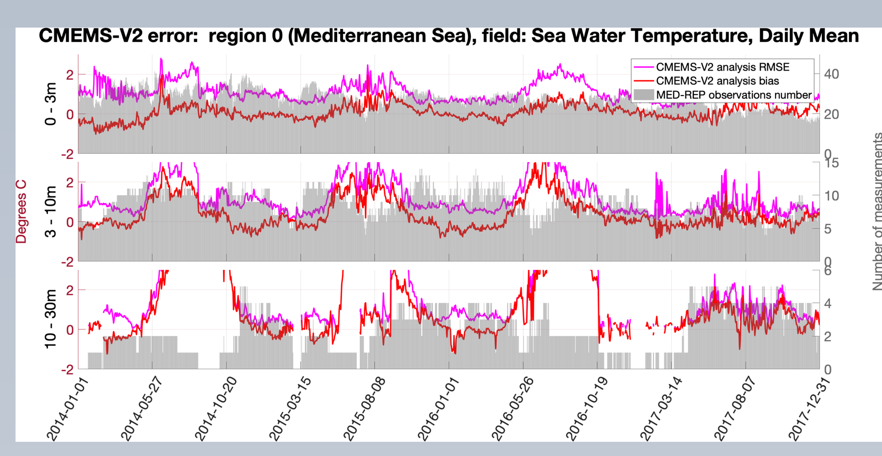
- Input: the same as in platform insitu evaluation;
- Output: per-field and per-med-region evaluation dataset (see **QUID** of the MED-MFC Copernicus products).

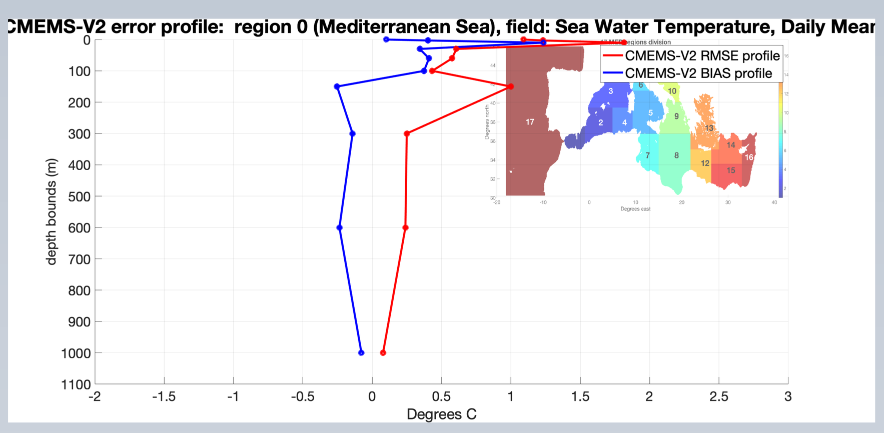
- Computation of averaged **super observation** for each region, linear vertical interpolated on 10 standard depth layers;
- Production of med region grid information, averaged model and insitu locations, depth layers and observations number N time series and profiles;
- Production of **RMSE** time series and profiles:

$$v_{R\_RMSE} = \sqrt{\frac{\sum_{n=1}^{N}(v_{vert\_interp_n} - v_{statistic\_good_n})^2}{N}}$$

- Production of **BIAS** time series and profiles:

$$v_{R\_BIAS} = \frac{\sum_{n=1}^{N}(v_{vert\_interp_n} - v_{statistic\_good_n})}{N}$$





Mediterranean Sea evaluation for temperature field, showing error time series and profile.

[1]Istituto Nazionale di Geofisica e Vulcanologia
[2]Centro Euro-Mediterraneo per i Cambiamenti Climatici