Benefits of interpreted vector programming and hierarchical data format for statistic ocean model evaluation

Paolo Oliveri, Istituto Nazionale di Geofisica e Vulcanologia (Italy), paolo.oliveri@ingv.it
Simona Simoncelli, Istituto Nazionale di Geofisica e Vulcanologia (Italy), simona.simoncelli@ingv.it
Alessandro Grandi, Centro Euro-Mediterraneo sui Cambiamenti Climatici (Italy), alessandro.grandi@cmcc.it
Emanuela Clementi, Centro Euro-Mediterraneo sui Cambiamenti Climatici (Italy), emanuela.clementi@cmcc.it

Big Data world is increasing incredibly fast, as High Performance Computing is near to overcome the Exascale and Artificial Intelligence techniques are invading our lives in new and different manners every day.

From the ocean models point of view, this means that now we have the capacity to build and run newer, more resolute and accurate models, using more or less the same time that we used in the past with less resources; however, this causes the production of more complicated and big datasets that anyway have to be analysed and evaluated.

From the observations point of view, this means that we can install, use, maintain and store very huge amounts of devices, with the same resources that we used in the past; unfortunately, the increasing number of recorded devices data and the temporal dilatation of the time series is producing very huge datasets, even if the single device data is pointwise.

Further, ocean recorded data are not completely reliable yet, because marine environment is not the ideal condition where electronic can work long and optimally, causing sampling discontinuities and needing frequent maintenance and sensors' calibration.

In order to make these two different worlds interact we took advantage of:

1) Interpreted language (for ease of use and portability) -> Python 3.x;
2) Vectorized numerical data analysis (for flexibility and speed) -> NumPy;
3) Output data storage (for efficiency and standardization) -> NetCDF-4 (HDF-5 extension);
4) Software maintenance and update (for security and speed) -> Git,

All of the selected software is free of charge and open source.

A Python module has been set up to automatically do:

1) Data quality assessment of a set of sea observations, using original quality controls (if available), spike removal (gross check) and redundant statistic quality check, the last one by computing standardized anomaly and the probability distribution (kernel density estimation), finally returning the best possible hourly and daily time series for the analysts' purposes;

2) Extraction and aggregation in time of ocean model time series at the horizontal nearest point to the observations, and then interpolating the model depth levels to the observations depth levels;

3) Running evaluation methods in terms of RMSE (root mean square error) and statistical bias, both for each device and for the Mediterranean standard Copernicus regions.

In order to create and set up the system for the first time we have chosen:

1) Water temperature, salinity, sea level, and 3D speed as test variables;
2) Fixed observatories data (moorings) located in the Mediterranean Sea downloaded from the European Copernicus Marine Environment Monitoring Service (CMEMS) in situ TAC (Thematic Assembly Center) Med FTP server (the data is accessible after free registration to Copernicus marine service portal);
3) Two model data from the RITMARE Italian project and from CMEMS.

The results are promising, both from the scientific and the technic point of view:

- An historical evaluation can be processed in acceptable times and it can be easily repeated on any machine equipped by the necessary portable software;
- A real time - production evaluation can be as easily set up as an “one off” one;
- Taking into account of the internal quality data processing of the observations, the evaluation results can provide more reliable and model helpful skill scores.

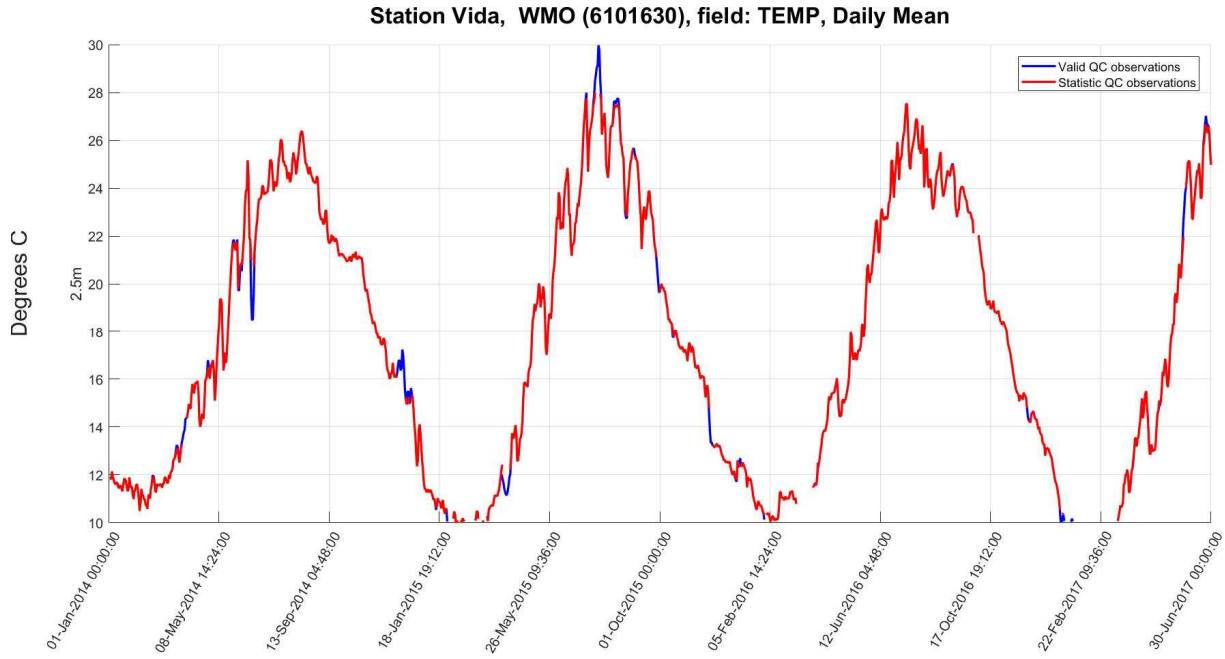

Figure 1. Difference of daily mean post processed insitu data of station Vida using original (blue line) versus 3 iterations statistic quality flag (red line) application.