Project

Near Real Time Quality Control and validation of Sea Level *in-situ* data within MyOcean.

Authors:

Begoña Perez, Marta de Alfonso Puertos del Estado, Madrid, Spain

Vibeke Huess, DMI, Denmark

Lesley Rickards, BODC/PSMSL, UK

TABLE OF CONTENTS

I INTRODUCTION

This report describes the recommendations for near real time quality control (QC) and validation of sea level *in situ* observations within the MyOcean project, to be applied in V1. These will be based on the QC procedures and flagging system suggested by the Global Sea Level Observing System (GLOSS) in a report compiled in 2009 as an adaption of the ESEAS Data Quality Manual. The reason for this MyOcean recommendation to follow the GLOSS QC system is explained below together with the suggested QC procedures and flagging system.

I.1 Data flow in Europe

The exchange of *in situ* sea level observations within the European countries has grown significantly during the last decade. This is caused by an increased interest in getting access to real time information of the variations in the sea level on a regional scale. Sea level observations are used in applications with many different purposes and on different time scales such as tsunami warning issues, nowcasting and forecasting the sea level variations in operational oceanography and long term estimates for climate change studies.

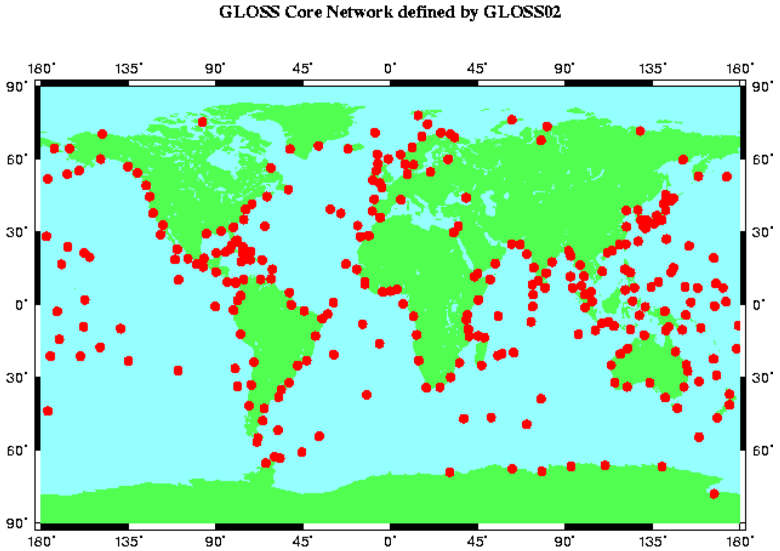

Figure 1: GLOSS Core Network of stations. Main objective: global coverage and long term studies

The global perspective with main focus on monitoring the global mean sea level, and its long term variations related to climate, has been covered by GLOSS. The European countries contribute to the GLOSS Core Network (Figure 1), composed of around 300 stations around the world, with sea level data along the European coastline from a subset of the existing European tide gauges. Within Europe a regional implementation of GLOSS, the European Sea Level Service (ESEAS), has been established to both support the GLOSS data exchange within the European area, and to enhance the exploitation of sea level in situ observations in the region by increasing the density of stations. A recent survey of existing sea level stations in Europe has been published in 2009 (Woodworth et al. 2009), which includes a review of their status about real-time capacities and contribution to international programs of operational oceanography (Figure 2). The number of the latter is increasing from day to day with respect to 10 years ago.

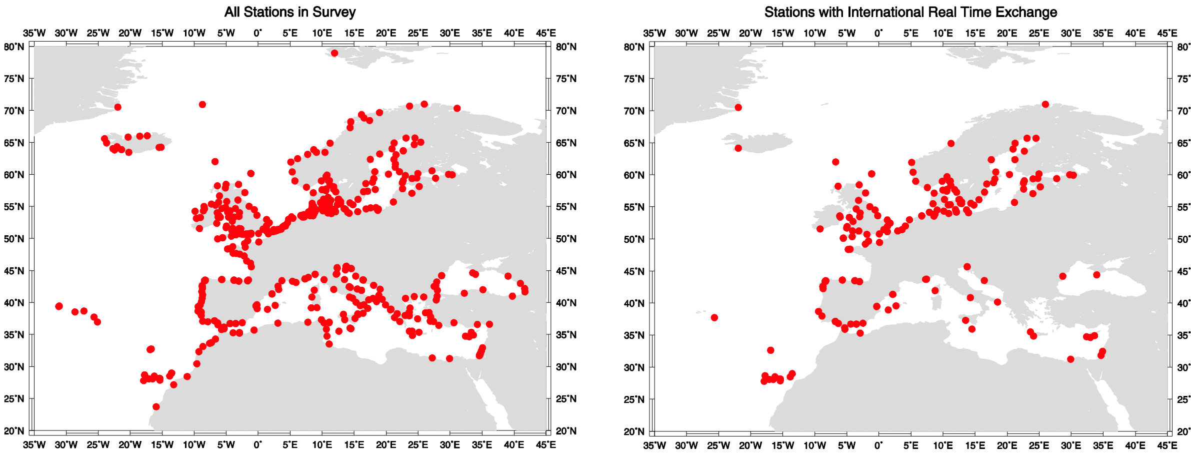

Figure 2: sea level stations in Europe, extracted from: "A survey of European sea level infrastructure", Woodworth, et al. 2009. Left: all stations included in the survey, right: stations actually contributing in real time or near-real time to international activities.

Concerning operational oceanography applications, the field of the MyOcean project, where information and data are needed in real time or near real time, the local and regional perspective within Europe has mainly taken part under the EuroGOOS umbrella within the regional ROOS'es (BOOS, BS-ROOS, IBI-ROOS, MOON, Arctic and NOOS). The exchange of near real time and real time sea level observations within the ROOS'es has grown from a wish of gaining information of the sea level from the neighboring countries. In some of the ROOS'es the system has developed into a full open exchange system with nearly all existing real time tide gauge observations available between all their members. The main initial focus in the ROOS'es has been on establishing the data exchange system, and increasing the data flow with the existing data quality from the data provider.

The latest initiative on data exchange of in situ sea level observations in Europe is the Tsunami Early Warning and Mitigation System in the North-Eastern Atlantic, the Mediterranean and Connected Seas (NEAMTWS) established by the IOC. Compared to the data exchange within the ROOS'es, where for storm surge applications latency required can be of one or several hours, the NEAMTWS data exchange system concerns only on real time data exchange, with a maximum latency of 1 minute, and on data with high sampling frequency (1 minute or less). This application is beyond the scope of MyOcean and this particular task at this moment.

I.2 The need of near-real time quality control. GLOSS/ESEAS QC system.

As the data exchange system is well established now in several of the ROOS'es, a natural step forward is to focus on the QC procedures. One of the most immediate applications of near-real time sea level data is the validation of storm surge models; in this aspect there is a well-established tradition of this use of the data in the NOOS region, where storm surge phenomena reach largest magnitudes and their effects may become more catastrophic. However, the interest on forecasting the meteorological component of sea level, or the total sea level signal, is extending now to other ROOS'es such as IBI-ROOS and MOON, where, in spite of being less prompt to dramatic events, it has become useful for better harbour operations and large vessels docking manoeuvres.

Concerning the objectives of this MyOcean task, near-real time quality control of sea level data is recommended for the main applications related with operational oceanography. This implies the need of implementing automatic software of error detection and flagging. This report will present the recommended procedures and characteristics of this software, in order to guarantee a standardized sea level quality control within MyOcean. This will be based on already existing documentation from GLOSS and ESEAS concerning QC techniques, where three types of delivery timelines can be distinguished, with logically different level of quality control:

- Real Time data (RT)
- Near Real Time data (NRT)
- Delayed Mode data (DM).

Real-time:

For real-time data provided as part of the tsunami monitoring system, with latencies under 1 minute, very little quality control is required. It is of prime importance that the data are provided without delay to the IOC Sea Level Station Monitoring Facility, as an interim solution in Europe, and that quality control does not remove tsunami events by rejecting out of range data. When the final regional tsunami warning centers are in operation, data must be checked by experienced personnel before entering any alert process. Just a few simple checks in real time can be done as detection that the tide gauge stops working – so that it can be fixed as soon as possible.

Near-real-time:

Data are considered to arrive in near-real time for latencies normally between 1 hour and several weeks, and this is normally the situation for storm surge forecasting or altimetry data calibration. This larger latency allows the implementation of some level of automatic quality control (L1 quality control) prior to archiving and use of the data. L1 quality control consists basically of detection of strange characters, wrong assignment of date and hour, spikes, outliers, interpolation of short gaps, stabilization of the series and, depending on the application, even filtering to hourly values and computation of residuals.

Delayed Mode:

This is the case of long time series, which require a more complete checking and analyzing procedure, including computation of all derived sea level products such as harmonic constants, extremes, mean sea levels, tide ranges, etc. One of the critical points in this case, especially for long term mean sea level studies, is datum control and detection of reference changes, with the study of operational history and maintenance incidences at the tide gauge.

Apart from L1 quality control, a second level of data processing can be performed, called L2, that is normally applied to one or more years of data, and that includes: tidal analysis, computation and inspection of residuals, basic statistics (highs and lows, extremes), computation of daily, monthly and annual means, comparison with neighbouring tide gauges, comparison with models or predictions, and detection of reference changes.

The MyOcean quality control for Sea Level observations for V1 will consist of:

* 1) an automatic near real time data QC (called **QC1**) which will be mandatory by all regions to apply and will include some of the L1 modules.

* 2) a second level of automatic QC (called **QC2**) which will be on voluntary level by each region to perform. QC2 includes all the L1 modules.

* 3) Assesment/validation of the products. This consists basically of human inspection of the data, and small changes based on information on reference level adjustments. The assessment will in most of the MyOcean regions rely only on procedures done at data provider level. Assessed data will be included in the MyOcean system, when available in delayed mode. See section III.

II AUTOMATIC QC ALGORITHMS APPLIED WITHIN MYOCEAN FOR NEAR-REAL TIME SEA LEVEL DATA

One of the commitments of the In-Situ TAC within MyOcean Project is to ensure the quality control and validation of the distributed data (see fig. 3). The procedures might be homogeneous and produce similar and uniform quality flags.

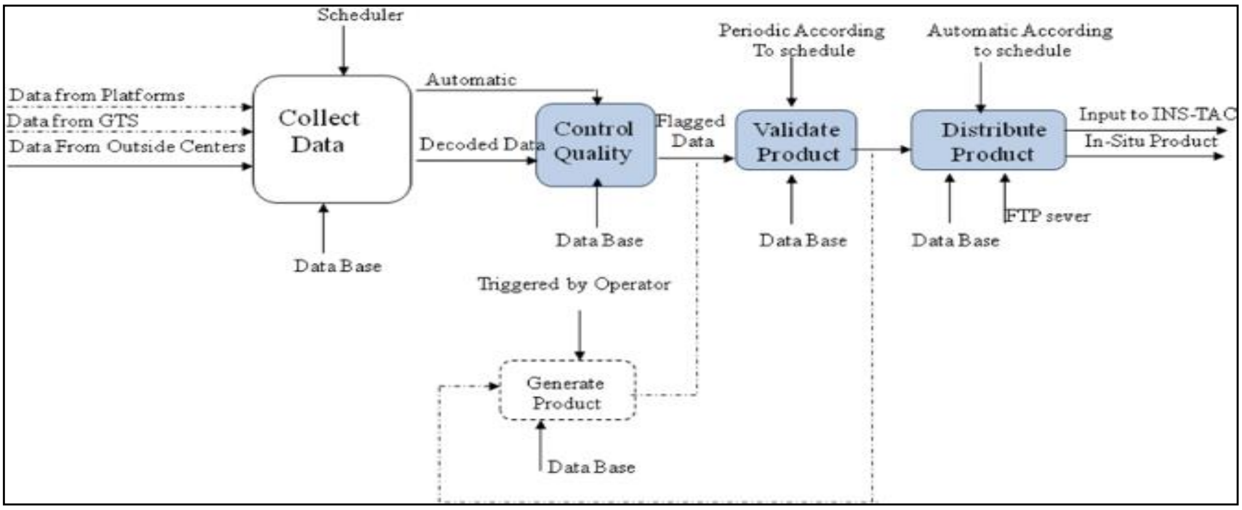

Figure 3. Functions to be implemented by the In-Situ Tac component.

The intrinsic nature of Sea Level data makes the QC procedures to have some special characteristics. Here we show the different quality levels and modules to perform the sea level QC. We will have a set of modules constituting the mandatory QC level (QC1), and we will have a higher level recommended but voluntary including the rest of the modules in the complete QC (Figure 4).

At Puertos del Estado, the software is applied to a moving window of two weeks of raw data with the original data sampling, normally between 1 and 15 minutes. The reason for this length of the window is that the output of this software is immediately sent to the storm surge forecasting system at Puertos del Estado, which needs the last weeks of data for adjustment of the model bias.

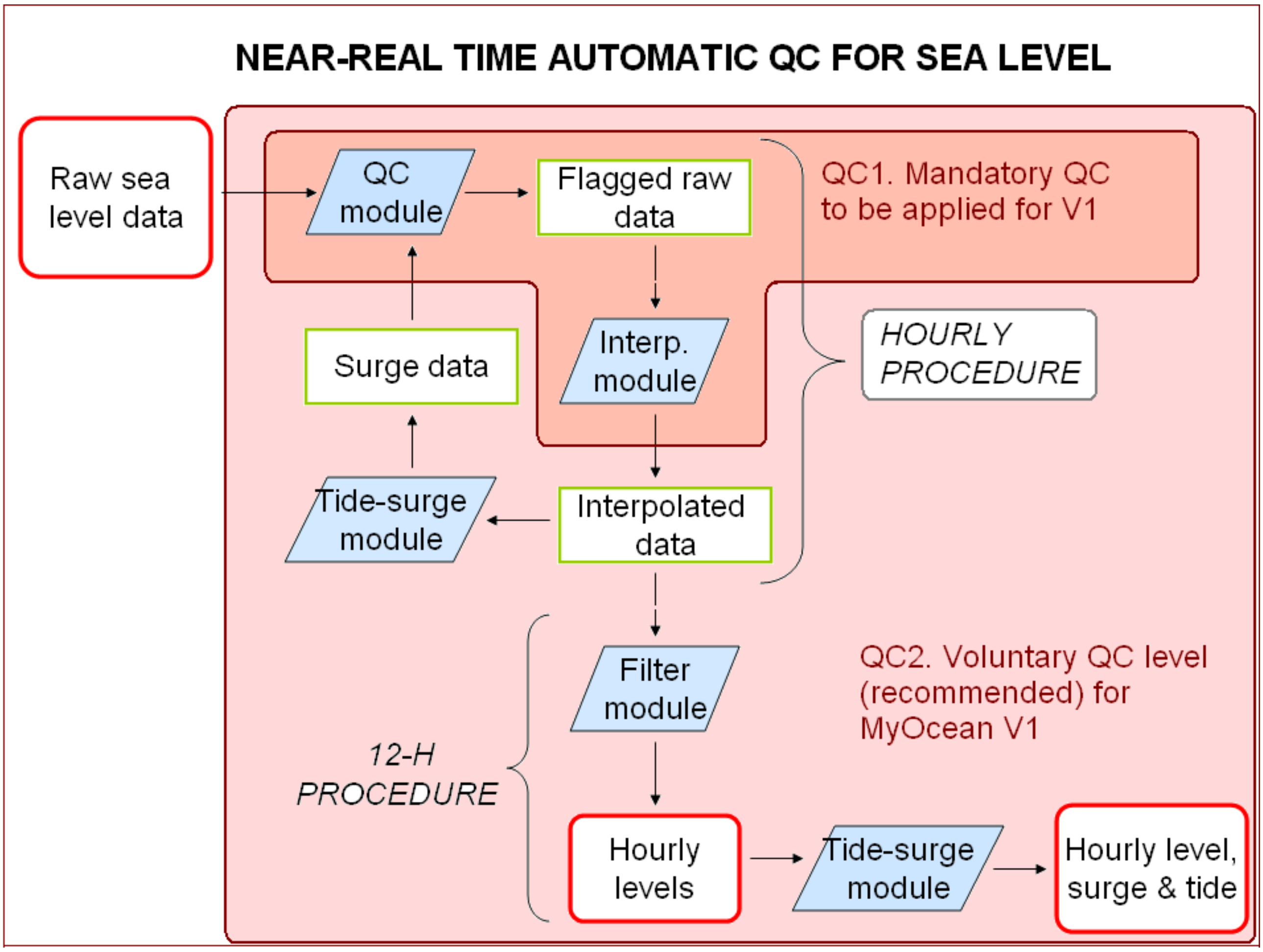

Figure 4: Scheme of the automatic software for QC in near-real time now in place at Puertos del Estado. Mandatory and recommended modules to be implemented in the In-Situ TAC for MyOcean V1.

II.1 QC1 (MANDATORY)

This quality control will be mandatory for all the regions. It will include the following modules:

<u>QC Module:</u>

responsible for:

- Strange characters detection (in which case the record is discarded)
- Out of range values flagging (based on extremes included in the metadata for each station)
- Algorithm for detection of spikes (explained below)
- Stability test: flagging values when there is no change in the magnitude of sea level after a number of time steps. The number of data values or time steps to begin to

flag depends obviously on the time interval. A typical value, for example, is 3 for 5 minute data.

- Date control

The algorithm for detection of spikes is the main component of the QC-module: it is based on the fit of a spline to a moving window of around 12-16 hours. The reason why this can not be applied in real time (latencies of 1 minute) is because it needs this long moving window to be able to detect spikes correctly and not flag real phenomena such as sudden high frequency oscillations due to “seiches” or tsunamis. The degree of the spline (which is normally 2) and the size of the window can be selected and determined depending on the characteristics of the tide, the data sampling, etc. The algorithm flags as spikes the values that differ more than N sigmas from the fit (normally N=3, although this can also be selected in the configuration file). Repeating the process for non-tidal residuals (obtained as total observed sea level minus predicted astronomical tide) is crucial to detect less obvious spikes not detected in the first step; this is why the QC-module is applied again when the residuals are obtained (Figure 4).

This algorithm has proved to be very efficient during the last years at Puertos del Estado, as can be seen in Figure 5, detecting more than the 95% of the wrong values of a very “bad” series.

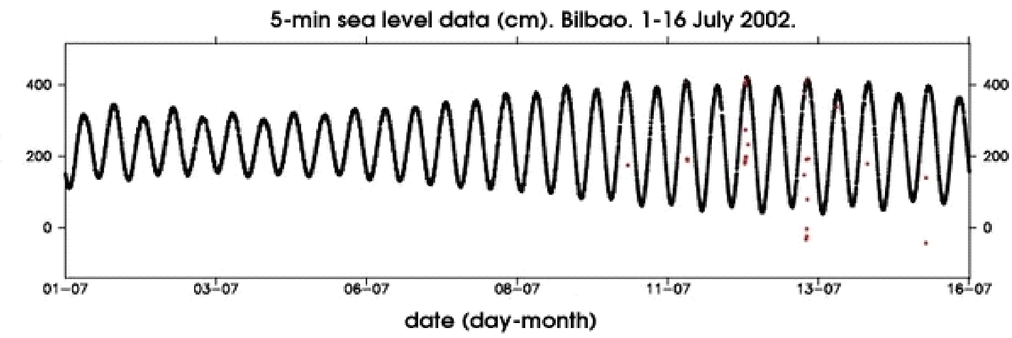

Figure 5: Example of the output of the fit of spline method to Bilbao tide gauge, in Spain. Spikes are plotted in red.

Interpolation Module:

most of raw data from a tide gauge arrive with a data sampling of several minutes, although for many applications in operational oceanography normally 1 hour is considered enough; besides, this data sampling is not always regular and, for example, 5 minute data supposed to arrive at 00,05,10…start arriving at 02,07,12. This is just an example of what can be found on raw data. So the interpolation module has the following objectives:

- checking and adjusting the time interval
- interpolation of wrong values previously flagged on the QC-module
- filling the gaps with new records with the correct date assignment and null-values for the sea level

- interpolation of very short gaps (smaller than 10-25 minutes, depending on the tidal range)

The output is a “clean” time series, called “interpolated series”, ready to enter the filter and harmonic analysis programs, i.e., it will be the one used for the rest of the data processing.

II.2 QC2 (RECOMMENDED)

This quality control level will be voluntary for MyOcean V1 but is highly recommended to guaranty a reliable quality control. It includes the modules in QC1 and also the following ones:

Filter module:

this software performs the computation of hourly values by means of the adequate filter, depending on the original data sampling. In the case of 5-minute data, as is the case of Puertos del Estado REDMAR data, a symmetrical filter of 54 points, following the expression:

$$X_f(t) = F_0 . X(t) + \sum_{m=1}^{M} F_m \left[X(t+m) + X(t-m) \right]$$

Where $X_f(t)$ is the hourly filtered value and F0…m the weights applied to the high frequency values. Details can be found in Pugh, 1987.

The selection of the filter is made taking into account the experience at Puertos del Estado and is also one of the recommended filters found in the ESEAS and GLOSS QC manuals. Figure 6 shows the differences between original and filtered data for Las Palmas station, showing that the algorithm eliminates just the frequencies larger than 0.5 cycles/hour.

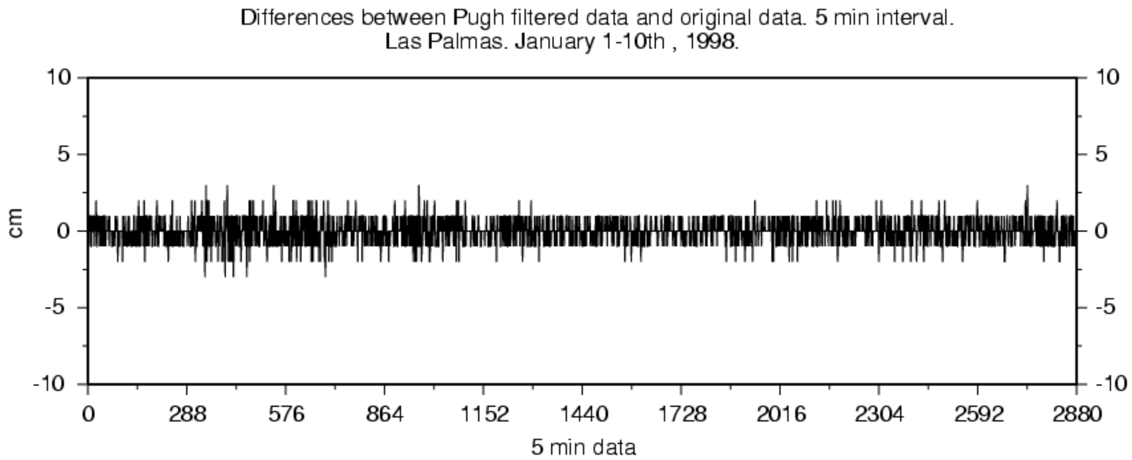

Figure 6: Differences between original data and hourly values for the Pugh filter show clearly that only the high frequency is eliminated, keeping the whole the tidal signal.

Tide-surge module:

this module computes the astronomical tide for the window of data, and then the surge component subtracting the tide to the original sea level. This is performed by means of the Foreman software of tide prediction (Foreman, 1977), and it requires the availability of the main harmonic constituents at each particular station, obtained offline from ideally 1 year of data. This is important because it implies the need of access to these previous data in order to compute a reliable set of harmonic components.

As it has been said, once the first residuals are computed, the QC-module is applied again to surge data (see Fig. 4), in order to detect less obvious spikes. If detected, these new wrong values are flagged again in the total sea level series and the rest of the process repeated to obtain the final products: interpolated series and hourly levels, surge and tide. Then the time series is ready to enter, for example, a storm surge forecasting system.

II.3 METADATA

Some basic additional information (metadata) must be included for each particular tide gauge station, as input for the quality procedures, as well as for archiving and exchange of data.

This metadata must be provided by the data producer when the regional InSitu TAC registers the station

Metadata for QC1 level:

- Data provider
- Country
- Instrument type
- Geographic location (latitude, longitude, coordinate system)
- WMO code of the station or if no WMO code, name of the station to generate MYO code.
- Datum information (chart datum, national datum ?)

Metadata for QC2 level:

The regional In-situ TAC needs also the following information necessary to apply the recommended quality control QC2:

1 year of data or:

- Harmonic constants of one year of data (at least 68 constituents) (this is for Tide-surge module).
- Maximum – minimum expected levels (for out of range detection)
- Maximum – minimum expected surge

II.4 FLAGGING SYSTEM

The GLOSS QC recommendations include a flag system which is in agreement with the SeaDataNet suggested QC flag system. We will adopt the same flags within MyOcean for sea level data.

Flag	Meaning	Definition
0	No quality control	No quality control procedures have been applied to the data value.
1	Good	Good quality data value that is not part of any identified malfunction and has been verified as consistent with real phenomena during the quality control process.
2	Probably good (previously 'correct but extreme')	Data value that is probably consistent with real phenomena but this is unconfirmed.
3	Probably bad (previously 'doubtful')	Data value recognized as unusual during quality control that forms part of a feature that is probably inconsistent with real phenomena.
4	Bad (previously isolate spike or wrong value)	An obviously erroneous data value.
8	Interpolated	This value has been derived by interpolation from other values in the data object.
9	Missing	The data value is missing. Any accompanying value will be a magic number representing absent data.

All the flags except flag=8 can be assigned after the first module (QC module) whereas after the Interpolation module, data can only have flags: 1, 8 or 9.

NOTE: Please note that this QC procedures done within the MyOcean regional centres might result in a data set provided within the MyOcean system that do not always corresponds exactly to the data set that could be provided directly from the data provider, as the different data providers might perform slightly different QC procedures."

III VALIDATION

The validation and assessment activities consist basically of human inspection of data and checks in clock problems and reference changes. To perform these checks, some additional information is needed that only providers can know. The validaton/assessment will, in most of the MyOcean regions, rely only on procedures done at data provider level.

Providers will be asked in a regular basis to report any validation and assessment results obtained and if any update in the delayed mode data is needed and then, assessed data will be included in the MyOcean system, when available in delayed mode.

When these checks are completed, a second level of data processing can be performed, that is normally applied to one or more years of data, and that includes: tidal analysis, computation and inspection of residuals, basic statistics (highs and lows, extremes), computation of daily, monthly and annual means, comparison with neighbouring tide gauges and comparison with models or predictions.

IV REFERENCES

- Manual for tidal heights analysis and prediction. M.G. G. Foreman. Canadian Pacific Marine Science Report No. 77-10, 10pp. 1977.
- Tides, surges and mean sea-level. D.T. Pugh, J. Wiley & Sons, 1987.
- GLOSS report: Quality control of Sea Level Observations. Adapted from the ESEAS Data Quality Manual compiled by Garcia, Pérez Gómez, Raicich, Rickards and Bradshaw. Version 0.5. 2009.
- A survey of European sea level infrastructure. P. L. Woodworth, L. J. Rickards, and B. Pérez. Nat. Hazards Earth Syst. Sci., 9, 1–9, 2009
- García, M.J., Pérez Gómez, B., Raicich, F., Rickards, L., Bradshaw, E., Plag, H.-P., Zhang, X., Lilja Bye, B. and Isaksen, E., 2007. European Sea Level Monitoring: Implementation of ESEAS Quality Control. In: Tregoning P., Rizos C. (eds) Dynamic Planet. International Association of Geodesy Symposia, vol. 130. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-49350-1_11