Research Article

Mediterranean Marine Science
Indexed in WoS (Web of Science, ISI Thomson) and SCOPUS
The journal is available on line at http://www.medit-mar-sc.net
DOI: http://dx.doi.org/10.12681/mms.15935

Assessing the quality of biogeochemical coastal data: a step-wise procedure

Luciana SABIA[1], Antonella COSTANZO[1,2], Maurizio RIBERA D'ALCALÀ[1], Vincenzo SAGGIOMO[1], Adriana ZINGONE[1] and Francesca MARGIOTTA[1]

[1] Stazione Zoologica Anton Dohrn, Villa Comunale, 1, 80121 Napoli
[2] Istituto Nazionale per la Valutazione del Sistema di Istruzione (INVALSI)

Corresponding author: francesca.margiotta@szn.it

Handling Editor: Ioanna SIOKOU

Received: 12 February 2018; Accepted: 27 July 2018; Published on line: 23 April 2019

Abstract

Coastal areas host valuable but vulnerable marine ecosystems subjected to increasing anthropogenic pressure and climate change consequences. To assess the impact of these pressures, monitoring programs have proliferated in coastal areas, but most of them follow locally established procedures for quality control (QC). The well-established QC procedure of open ocean data cannot simply be extended to the highly variable coastal area, for which there is the need to develop ad hoc QC approaches. This is particularly crucial for long-term time series, where different instrumentation and analytical methods have been applied over time. This study, based on the biogeochemical dataset collected over 30 years at the LTER MareChiara station (LTER-MC, Gulf of Naples, Mediterranean Sea), addresses potential discrepancies in a long-term dataset, identifying criteria and methods that could also be applied to other coastal datasets. We developed a serial step-wise procedure to characterize the quality of ~ 84,000 data-points, merging statistical tests and expert knowledge. The procedure included nine tests, each addressing potential problems in data generation and management, some of which of general application and others tailored to specific subsets of data. Based on these test, quality flags were assigned to individual data. Critical tests applied to two other independent datasets, showed that the procedure is not dataset dependent. These results contribute to bridge the gap between the need of objective QC criteria and the intrinsic noise of coastal datasets, promoting the discussion on this topic, and improving a proper management and sharing of coastal data.

Keywords: Biogeochemical dataset; Chlorophyll *a*; Coastal oceanography; CTD; Data flags; LTER-MareChiara; Mediterranean Sea; Nutrients; Quality control; Tyrrhenian Sea; Time series data.

Introduction

At present, there is a huge amount of oceanographic data collected all over the world in different environments and with different methodologies. While the overall quality of the measurements has improved with time (Lauvset & Tanhua, 2015), the comparability of the data is still a major issue, irrespective of the laboratory, country of origin or time (Ibe & Kullenberg, 1995). Generating comparable and consistent results in environmental studies is a goal that can be achieved only through intercomparison exercises (e.g., Jakobsen *et al.*, 2015; Aoyama *et al.*, 2016) and, primarily, through well-established quality assurance (QA) and quality control (QC) procedures (Ibe & Kullenberg, 1995), although the great variety of sensor types and different analytical techniques can make this task particularly challenging (Campbell *et al.*, 2013).

A growing body of literature deals with QC of environmental data, mostly addressing large datasets obtained through autonomous sensors that are dispersed all over the oceans (i.e. Argo: Wong *et al.*, 2015; NOAA: Conkright *et al*., 1994; CARINA: Tanhua *et al*., 2010), for which ad hoc QC softwares were developed (e.g., Sheldon, 2008; Horsburgh *et al*., 2015). These QC procedures are generally sufficient to define regional water-mass features and to discriminate anomalous data (de Boyer Montegut *et al.*, 2004). Indeed, QC with open ocean data is relatively straightforward, given the homogeneity of seawater masses that can be considered almost invariant over 1500 m depth, allowing for a prompt identification of outliers and bad data (Lauvset & Tanhua, 2015).

Much more challenging is the task to perform QC on coastal or riverine data (Moatar *et al.*, 2001), which however are more easily obtainable and attract increasing attention, as they concern areas where the major part of anthropogenic activities takes place. There, monitoring programmes regularly produce a great amount of data, used to assess ecosystem health and to support decisions about conservation policies (McQuatters-Gollop *et al.*, 2015). As an example, Long Term Ecological Research

(LTER) activities have become more common in the last decades (Michener, 2016). In Europe, there are currently 38 marine sites in 12 countries participating in the LTER-Europe network (www.lter-europe.net), with most of them located in coastal, shelf and transitional areas and regularly producing new data. Some of these observational activities have taken place for several decades, during which different technologies and methods have been applied, with considerable variations in analytical procedure and changes in personnel.

Although coastal areas are more accessible to sampling than open seas, several forcing mechanisms that are of critical importance are often not easy to grasp or quantify (Jickells, 1998). Agriculture, industrialization and fossil fuel combustion all exert a large influence on ocean chemistry, firstly in local coastal waters and then, on a larger scale, regionally and globally in the open ocean (Doney, 2010). One of the biggest impacts of these activities is exerted by the increasing nutrient loads, which influence productivity, biodiversity and ecosystem health (Cloern, 2001). The main sources of nutrients along the coasts are river inputs, groundwater seepage and wet and dry atmospheric deposition, mostly acting on the continental shelf, where terrestrial and oceanic forces interact, creating dynamics difficult to understand (Burnett *et al*., 2003) and which have high and unpredictable variability at several spatial and temporal scales.

Because of this extreme environmental variability, identifying recording errors and outliers in coastal dataset is not a trivial operation because the risk of false positive results (i.e. good data marked as invalid) is very high (Campbell *et al.,* 2013). To circumvent this problem, quality controls of large datasets in different regions (e.g., Levitus, 1982; Conkright *et al*., 1994; Tanhua *et al*., 2010; Wong *et al*., 2015) allow for relatively wide variability ranges for coastal zones, which however considerably increases the risk of false negative results (erroneous data accepted as valid).

Another source of complication is the mixed origin of data obtained in coastal waters, often collected through both sensors and instrumental analyses. For automatically generated data, possible errors can quite easily be attributed to technical causes, such as fouling of sensors, calibration shifts, or failures of sensors, recorders, and transmission systems, whereas only in some cases unforeseen environmental conditions can have an influence on sensor readings (Wagner *et al*., 2006). In these cases, some common QC practices can be adapted to different conditions merely by setting appropriate tolerance ranges. When dealing with a dataset with mixed origin, there are no universal QC standards applicable to all circumstances. In this case, it is necessary to design QC procedures for each type of data and for each location (Campbell *et al*., 2013). Nonetheless, QC for complex datasets should still follow common principles and criteria, to be applied to coastal data collected over long time spans, so that sampling sites across a wide geographic range can reliably be compared.

This paper proposes a step-by-step approach on how to handle the high variance of coastal data in an effort to identify general criteria for QC that should be suitable for datasets collected in coastal areas. For this purpose, we used the biogeochemical dataset collected in the Gulf of Naples (GoN, Mediterranean Sea) at the Long Term Ecological Research site MareChiara (LTER-MC) as the basis to design and test QC procedures, from the identification of appropriate tests to their application to the different parameters, resulting in the assignment of quality flags to individual data. In addition, we applied some critical steps of the QC procedure developed for the GoN database, with appropriate adjustments, to two different datasets, to test its suitability beyond the dataset for which it was first designed. One dataset was collected within the Italian Coastal Ecosystem Monitoring Project Si.Di.Mar. (http://www.sidimar.tutelamare.it) and consisted of data collected over 5 years and a half in an oligotrophic coastal area of the adjacent Gulf of Salerno (GoS). The other was the result of two sampling campaigns in several coastal areas influenced by rivers in the Mid and North Tyrrhenian Sea and in the South Ligurian Sea (TYR project).

Material and Methods

Datasets

The LTER-MC biogeochemical dataset

The Long Term Ecological Research Program MareChiara (LTER-MC) is one of the longest and most detailed plankton monitoring activities in the Mediterranean Sea, with regular sampling since January 1984. The sampling site LTER-MC (40°48.5’ N, 14°15’ E) is located over a depth of approximately 73 m, two nautical miles from the coastline in the Gulf of Naples (GoN), a relatively deep embayment (average depth ~ 170 m) along the coast of the Mid Tyrrhenian Sea (Fig. 1). Following a series of oceanographic campaigns in the inner GoN in summer 1983 (Zingone *et al*., 1990), the site was selected in an area that receives municipal inputs from one of the most densely populated Mediterranean coastal regions. At the same time, the GoN is also influenced by the dynamics of the offshore oligotrophic Tyrrhenian Sea waters and occasionally by water masses from the adjacent Gulf of Gaeta (Iermano *et al*., 2012).

Sampling at LTER-MC has taken place every fortnight until 1991 and weekly from 1995 to date, with a major interruption from August 1991 to February 1995. In this study, we used the data obtained until December 2014. Over the course of the last 33 years, since the sampling at LTER-MC started, methodologies, instruments, sensors, technologies and individuals handling the data have changed. The dataset encompasses physical and biogeochemical variables, as well as detailed species composition and abundance for phyto- and mesozooplankton. The latter two biological components are not considered in this paper, while QC procedures for phytoplankton

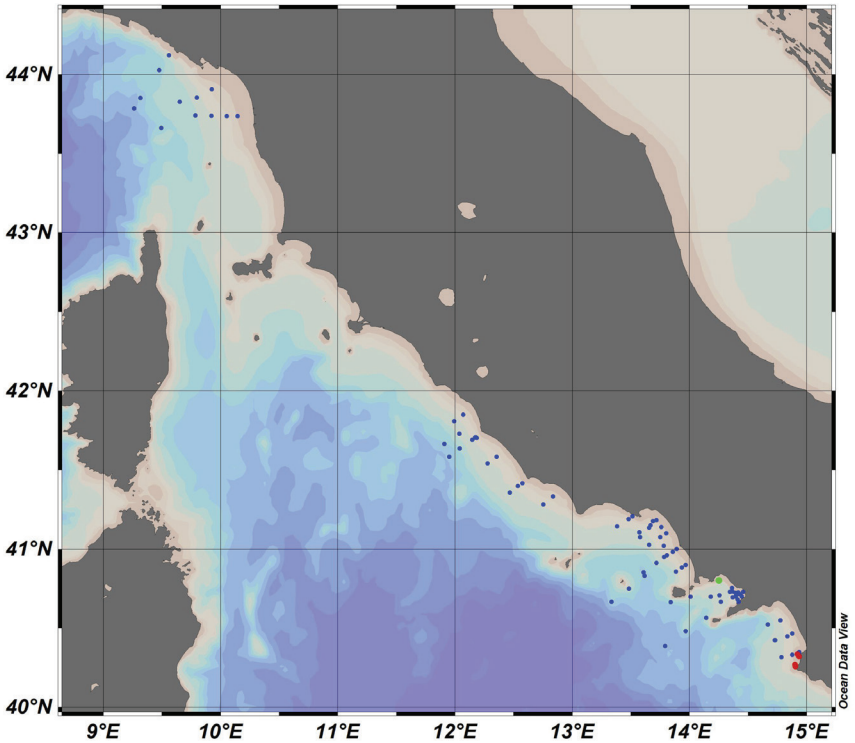

Fig. 1: Geographical position of sampling stations for LTER-MC (●), Si.Di.Mar (●) and TYR (●) datasets.

have been addressed elsewhere (Zingone *et al.,* 2015). Samples for salinity and nutrient analyses were collected at ten fixed depths (0, -2, -5, -10, -20, -30, -40, -50, -60 and -70 m), seven of which (0, -2, -5, -10, -20, -40 and -60 m) were also sampled for chlorophyll *a*. All these data result in a large dataset of 11,100 lines, composed of 1,110 sampling events, each labelled progressively from MC1 (January 26th 1984) to MC1134 (December 22nd 2014) and consisting of eight different variables (temperature, salinity, ammonia, nitrites, nitrates, phosphates, silicates and chlorophyll *a*) at 10 (or 7 for chlorophyll *a*) different depths. The data were obtained through sensors and/or laboratory analyses with procedures and instruments that changed over time, as described in the following.

Temperature (TEMP) was measured with reversing thermometers from January 26th 1984 to September 12th 1995. From October 3rd 1995, different multiparametric profilers were used to acquire temperature and salinity along with pressure. Some missing data of year 2000 were interpolated, on the basis of the salinity and discrete temperature values of adjacent samplings, through an adaptation of the DINEOF (Data Interpolating Empirical Orthogonal Functions), described in Beckers & Rixen (2003).

Salinity (PSAL) in a first period (from January 1984 to December 2001) was determined following the reference method designed by Müller (1983) using initially a salinometer (Beckman mod. RS7C and subsequently AutosalGuildline Instruments).

Since 2002 one multiparametric profiler (Sea-Bird Electronics, 9-11 plus V2) has been measuring physical (pressure, temperature, salinity), biogeochemical (fluorescence and dissolved oxygen) and optical (Photosynthetically Available Radiation, PAR) parameters. This CTD (Conductivity Temperature Density) profiler is periodically calibrated at the Istituto Nazionale di Oceanografia e di Geofisica Sperimentale (OGS) in Trieste.

Samples for the determination of inorganic nutrient concentrations were collected from the Niskin bottles into 20 ml polyethylene vials and immediately frozen. For the whole time series concentrations of the following parameters were determined following Hansen and Grasshoff (1983): ammonia (AMON), nitrates (NTRA), nitrites (NTRI), phosphates (PHOS) and silicates (SLCA). Two different instruments were used for analyses: an Autoanalyzer Technicon II series until 2005 and a FlowSys Autoanalyzer (SYSTEA) thereafter. In order to improve the quality of nutrient data, reference materials were used and the laboratory partook in intercomparison experiments since 2002 (e.g., Aoyama *et al*., 2016).

For the determination of chlorophyll *a* concentrations (CHLT), a variable volume of sea water was filtered under low vacuum on a Whatman glass-fibre filter (GF/F, Ø 25 mm) and then extracted in 10 ml of neutralized acetone 90%. The concentrations were measured initially with a spectrophotometer (Strickland & Parsons, 1972) and afterwards with a SpexFluoromax spectrofluorometer (Neveux & Panouse, 1987) until July 2006, and a SHIMADZU (mod. RF-5301PC) spectrofluorometer (Holm-Hansen *et al*., 1965) from August 2006 onwards.

The Punta Licosa and Punta Tresino Si.Di.Mar. datasets

Data in the Gulf of Salerno (GoS) were obtained in the framework of the Si.Di.Mar. project. Sampling was

performed fortnightly in surface waters at six stations along two coast-offshore transects at two sites (Punta Licosa and Punta Tresino) from June 2001 to December 2006 (804 sampling events) (Fig. 1). The set of biogeochemical data was obtained with the methods described above in the same laboratory.

Information on the GoS is very sparse and only obtained through occasional sampling (Marino *et al*., 1984). Despite the vicinity to the GoN, the GoS differs substantially in physiography and social-economic features, and the two sites of the GoS include protected and pristine areas. Differently from the GoN, the GoS coastal waters are generally restricted to nearshore areas, whereas offshore waters often extend to near the coastline. At the six sampling sites, nutrient concentrations were markedly lower than those recorded in the GoN, and comparable to those of offshore waters.

The TYR dataset

Sampling campaigns took place in October –November 2010 and April 2013 in the Mid and North Tyrrhenian Sea and in the Ligurian Sea next to the major river estuaries (Sarno, Volturno, Garigliano, Tiber, Arno and Magra) from the GoN to the Tuscany coasts (Fig. 1). Also in this case, the set of the same biogeochemical variables was obtained in the same laboratory with the methods described above. Of the entire dataset, we only used the data from the stations closest to the coastline and above the 200 m isobaths. These sampling sites represented those most affected by river plumes and had the highest variability, making them ideally suited to test the efficiency of the QC spike tests for nutrient and chlorophyll *a* data. For our tests, we used a subset of 560 data per each nutrient and of 486 data for chlorophyll *a*, sampled at 100 stations in the upper 100 m of the water column.

Procedures for Quality Control

Errors in large datasets may occur for several reasons and through different routes, ranging from sensor failures or laboratory mistakes to errors introduced while managing the dataset. For this reason, different kinds of tests are needed to deal with all scenarios. In general, when testing the validity of observational results stored in a database, the most effective action is comparing them with the historical knowledge of the features of the area. However, when this knowledge is not available or not accurate enough, the most basic level of QC consists of comparing a dataset against itself (Conkright *et al*., 1994). Our QC procedure was thus based on the dataset itself, relying on the great number of observations.

We performed all tests by designing ad hoc Matlab functions (http://qcbiogeodata.szn.it/). The tests performed on the dataset were:

1. Identification of missing data.
2. Duplicated profiles test: comparison of vertical profiles of individual parameters between different sampling events.
3. Frozen profiles test: comparison of data for individual parameters along profiles.
4. Spike test: detection of anomalies in series of three or four consecutive data in a vertical profile.
5. Range test (upper and lower limits): search for anomalous data based on variability of individual parameters in the whole dataset.
6. Surface / bottom gradient: adapted spike tests for data with no adjacent values.
7. Inter-sampling variability test: check for temperature data variations over consecutive sampling events.
8. Parameter relationship test: preservation of apparently anomalous (bad) data in case of extreme events reflected in more than one variable by comparing chlorophyll *a*, salinity and nutrient data.
9. Detection limit: search for nutrient concentration values lower than method limits.

Some of these tests were performed using individual data (i.e., range test and detection limit), while in other cases it was necessary to take into account all the samples of the water column (profile) or data from at least three consecutive sampling depths (i.e., frozen profile and spike tests).

The choice of tests here presented was inspired by the recommendations of the International Oceanographic Data and Information Exchange (IODE) workshop (IOC, 2010) for minimum QC checks and aimed at defining ranges and thresholds and at proposing criteria to calculate them based on the data themselves.

The ultimate goal of the QC procedure was to assign quality values (i.e., quality flags=QF) to each data-point following the results of these tests. Several quality flag schemes are proposed for oceanographic datasets, e.g. SeaDataNet (https://www.seadatanet.org/), EMODnet Chemistry (http://www.emodnet-chemistry.eu/welcome), IODE, 2013 (IOC, 2013), with different numbers of flag levels. Our list of ten flags was compiled following the suggestions of the IODE workshop (IOC, 2010), which proposed a minimum of five level flags inversely related with the quality. To this minimum scheme we added several other flags in between, the whole list being easily transferable to the 4 quality flag levels used by the Ocean Data View (ODV) software (http://odv.awi.de/en/documentation/), which is widely used for analysis and graphic representation of oceanographic data. While the selection of a flag scheme may reflect different requirements and fit different data handling needs, it is essential to be able to pass from one system to the other based on shared definitions of the flags. To allow for such conversions, in Table 1 our quality flag scheme is compared with the most commonly used ones in European data centres (ODV, SeaDataNet and IODE 2013).

Whereas some tests may allow flagging data immediately, others may not give direct information about the data quality but rather highlight a lack of coherence of some data in the context of the whole dataset. These seemingly incoherent data may represent either low data quality or may

Table 1. List of quality flags (QF) applied in this study, with their meaning and match to other existing QF schemes.

QF	Meaning	ODV	SEADATANET	IODE 2013
0	Good= passed all the applied QC tests	0	1	1
1	Quality not evaluated	1	0	2
2	Probably good data	0	2	1
3	Data below detection limit (DL)	0	6	2
4	Dubious data	4	A	2
5	Reconstructed data	4	8	2
6	Probably bad data	8	3	3
7	Manipulated data	8	-	-
8	Bad data	8	4	4
9	Missing data	1	9	9

result from intense meteorological events or accidental pollution, a scenario that requires verification by further tests.

Setting up the method

Before performing the QC, some preliminary controls were carried out on the metadata and on the dataset structure itself, in order to identify the different sources of the data and recognize possible obvious errors. The cruise numbers and dates of the sampling events were controlled in order to identify any missing events or duplications, referring to the original oceanographic logs in case of doubt. Subsequently we controlled the number of lines (sampling depths) of each event and added a blank line where missing data lines were identified. In case of reconstructed data (data deriving from interpolation between other values/variables in the dataset), such as the temperature in the LTER-MC data-set, the QF 5 was immediately assigned. The dataset thus modified was ready for the QC tests.

Initial analysis of data distributions of the different parameters revealed quite heterogeneous patterns across the dataset. In fact, each parameter was characterized by a different distribution curve, reflecting the complexity of processes involved in the modulation of their variability. Physical parameters showed a skewed normal distribution, while the distributions of the chemical and biological parameters, ranging from 0 to very high values, were non-normal and positively skewed (Fig. 2). The different distributions and characteristics of the data rendered it impossible to apply the same QC criteria to physical, chemical and biological parameters. Consequently, although the tests performed were in principle the same for all the parameters, some tests were adapted to meet precise criteria for the characteristics of specific parameters.

Description of the tests

In the following, we describe the details of the nine tests while an overview of the complete procedure and the QF assigned consequently are reported in Figure 3.

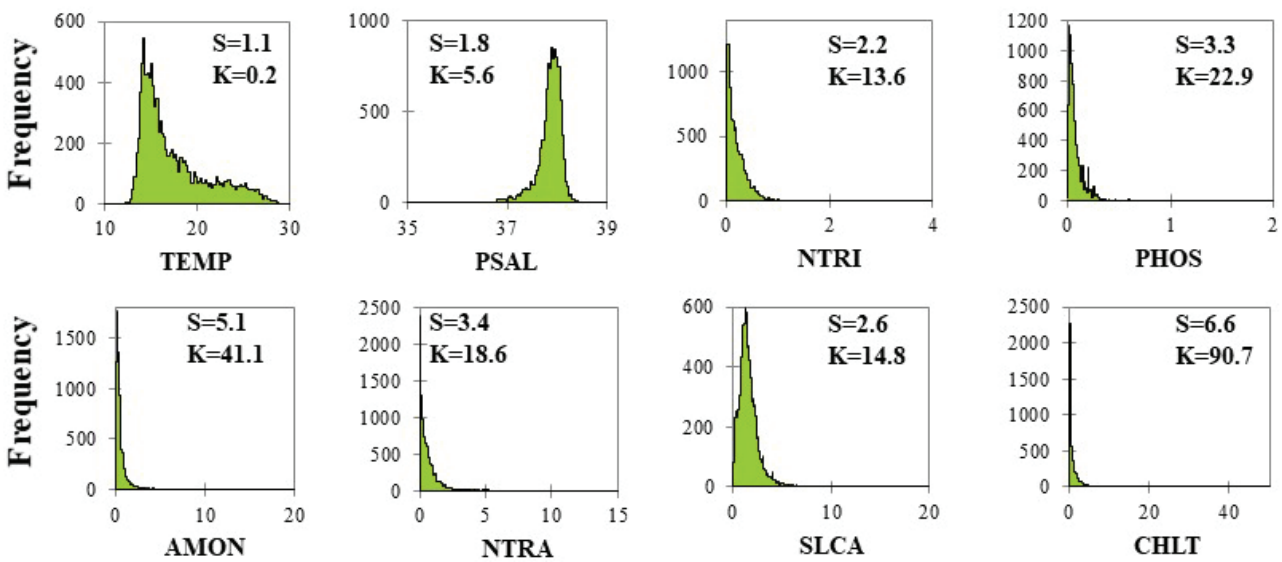

Fig. 2: Frequency distribution of individual parameters in the LTER-MC dataset before applying QC (TEMP: temperature, PSAL: salinity, NTRI: nitrites, PHOS: phosphates, AMON: ammonia, NTRA: nitrates, SLCA: silicates, CHLT: chlorophyll *a*, S: skewness and K: kurtosis).

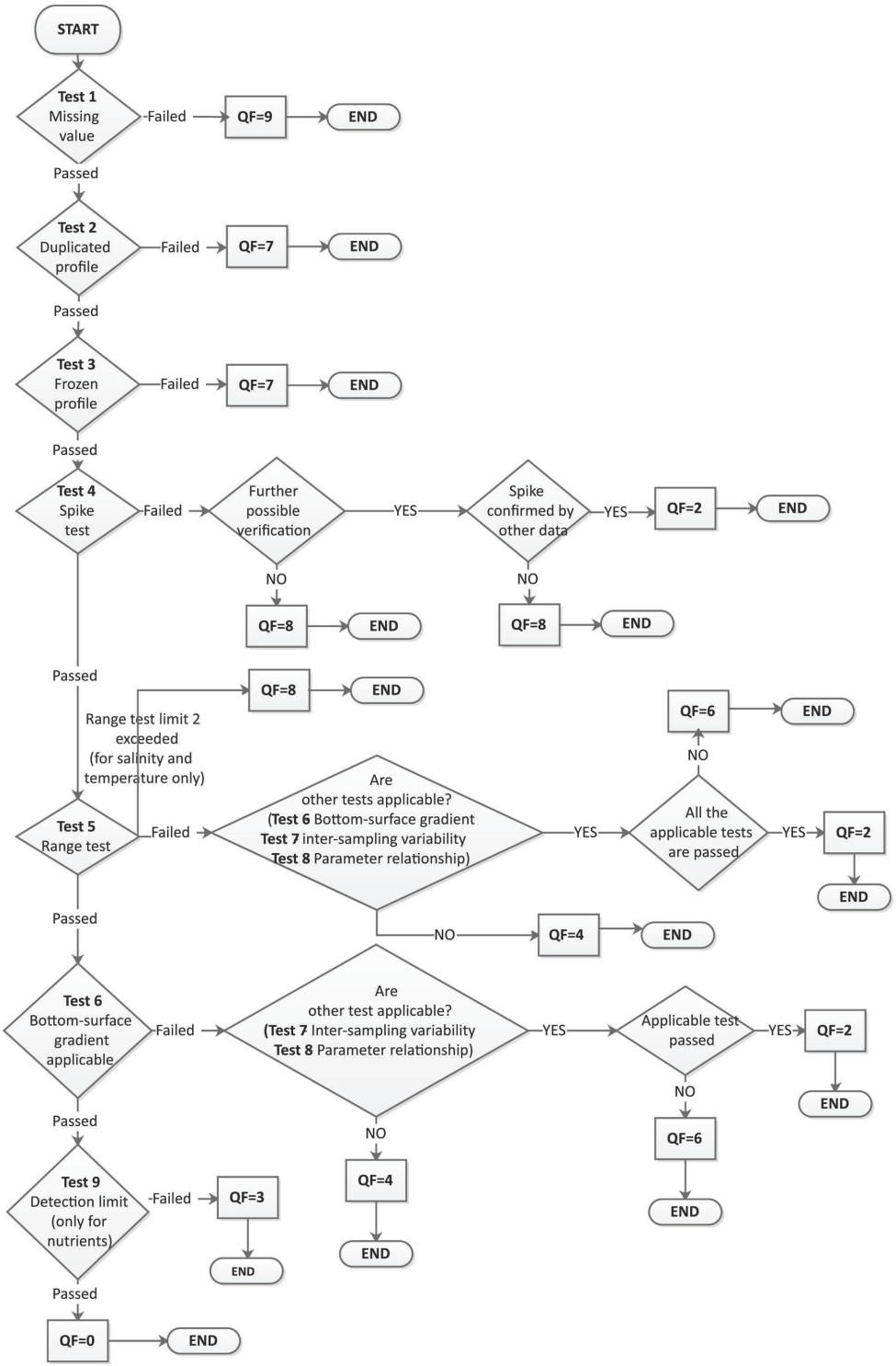

Fig. 3: Workflow of the quality control procedure and quality flag (QF) attribution.

Test 1) Missing values

Problem: In a dataset missing data can exist because of non-performed measurements or lost data.

Flagging method: Empty elements in the data matrix were searched.

Possible correction: When available, data from the oceanographic logs or calculation spreadsheets were inserted.

Action: Empty elements were flagged as missing data (QF 9).

Test 2) Duplicated profiles test

Problem: In a large dataset, errors related to the architecture and management of the database can easily occur, resulting in replicated data strings.

Flagging method: Each vertical profile of a single sampling date was checked against all the other profiles referring to the same parameter. If another (or more than one) profile in the dataset had all ten values equal to the tested one, all the involved profiles were considered as manipulated and identified as bad data. We applied the test in the same way to all parameters.

Possible correction: When available, the original profile data from the oceanographic logs or calculation spreadsheets were used to replace the bad data.

Action: If two or more profiles were identical and it was not possible to identify the correct one, all the profiles were flagged as manipulated data (QF 7), i.e. data that underwent an involuntary modification after their creation.

Test 3) Frozen profiles test

Problem: The occurrence of a high number of data with exactly the same value (three decimal places for laboratory analyses and the full string for CTD data) within the same sampling event is highly improbable and likely caused by an error in the data set handling.

Flagging method: For each sampling date and for each parameter, the number of data-points with the same value was determined, and when this number was higher than the half of the number of data-points in the same profile (>5 for LTER-MC), the whole sampling event was considered manipulated. We applied the test in the same way to all the parameters.

Possible correction: When available, the data from the original profile were used to replace the bad data.

Action: The whole profile was flagged as manipulated data (QF 7), data that underwent an involuntary modification after their creation.

Test 4) Spike test

Problem: When the central value in a sequence of three data in the same profile is markedly different from the two adjacent data, thus representing a spike, it likely results from an error.

Flagging method: This test was based on the assumption that a spike might be present when there is an inversion of the sign of the profile (Fig. 4 A):

$$(V_1-V_2)*(V_2-V_3)<0$$

where V_2 is the data point that is to be checked and V_1 and V_3 are the previous and the subsequent values in the profile, respectively.

When this hypothesis is verified, the spike is flagged if the following inequality is met (Wong *et al.,* 2015):

$$|V_2-((V_3+V_1)/2)|-|(V_3-V_1)/2|>\text{threshold value}$$

where the threshold value represents the maximum variation allowed between the V_2 and the expected V_2 values according to the variation between V_1 and V_3.

We applied this test to all parameters calculating the threshold values to identify spikes according to the variability of each parameter tested. For temperature and salinity, which have a near-normal distribution and a restricted range of variability, we adopted fixed threshold values, calculated as the ratio between the interquartile range and the range of variability of the selected parameters. Based on our dataset, the thresholds were 0.28 for temperature and 0.08 for salinity. Regarding biogeochemical parameters, for which distributions are highly skewed (Fig. 2), with variability ranges at times encompassing several orders of magnitude, a fixed threshold could result in a poor evaluation of spikes, with a sensitivity too high for the highest values and too low for the smallest. For this reason, we defined different threshold values according to the data distribution, with a more tolerant threshold in case of high values and a more restrictive threshold in case of low concentrations. The parameter distributions in terms of skewness and kurtosis were analysed first (data in Fig. 2) to guarantee a high accuracy in the definition of intervals and corresponding threshold values.

For nutrients, we divided the data into three intervals, with boundaries depending on the values of the 25th and

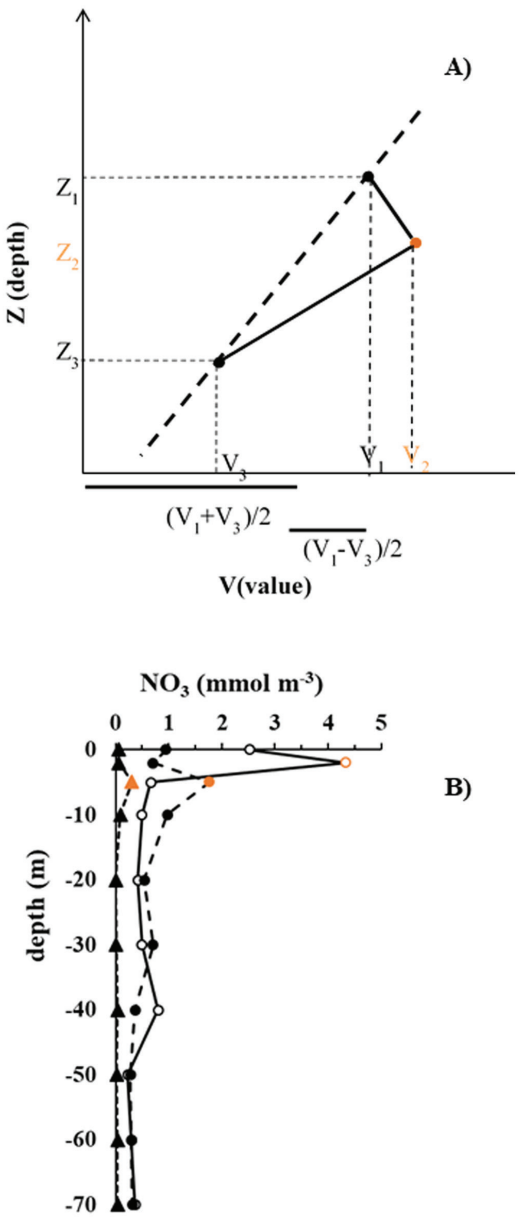

Fig. 4: **A**) Schematic representation of a possible spike (V_2) evaluated based on its previous (V_1) and subsequent (V_3) values of the profile. **B**) Examples of identified spikes in nitrate profiles at low (▲MC 535, -5 m), medium (●M276, -5m) and high (○MC48, -2 m) concentrations. Spikes are in orange.

90th percentiles of their distribution, to give a different weight to the extreme data of the distribution. For this purpose, we chose the 15%, 35% and 75% of the difference between the 90th and 25th percentile as threshold values. This choice allowed us to apply a unique criterion to all parameters for the identification of the threshold values, whilst adapting the threshold to the data for individual parameters (Table 2). If three adjoining values in the vertical profile (V_1, V_2 and V_3) did not fall in the same interval, the threshold value for nutrients was determined according to:

V_2, if $V_2<(V_1+V_3)/2$

$(V_1+V_3)/2$ if $V_2>(V_1+V_3)/2$

This distinction was made in order to be conservative in the application of the spike test, as the lower threshold value is chosen in both cases. Examples of identified spikes in NO_3 profiles at different concentrations are reported in Figure 4 B.

Chlorophyll *a* values instead were characterized by the most skewed distribution and the highest value of kurtosis (~90). Therefore, in order to make the estimation of the spikes comparable with the other parameters, the number of intervals was increased to four (Table 3). In the definition of the threshold, we used the same criterion ensuring different levels of tolerance for low and high values, and selected, respectively, the 7.5%, 20%, 40%, 75% of the difference between the value of the 25th and 90th percentile. If V_1, V_2 and V_3 did not fall in the same interval, the threshold value was set based only on V_2. We made this choice in order to ensure a higher threshold value in the presence of a sub-surface maximum, to avoid erroneously flagging subsurface maxima as spikes.

This test was only possible for measurements that have two contiguous values (above and below in the vertical profile), to which the central data could be compared and was therefore not applicable to the shallowest (surface) and deepest values (70 m, or 60 m for chlorophyll *a* data). In addition, a second test was necessary in case of the occurrence of two consecutive spikes over 4 values. In this case, the presence of one spike may lead to the erroneous assignation of a contiguous spike, due to the presence of an anomalous value in the equation. Therefore, a second test was performed omitting one of two spikes at a time. If both anomalous values still resulted as a spike, they were marked as bad data. If none resulted as a spike, the values were still marked as bad data, as they may have resulted from an inversion in depth labelling. When only one of the two values resulted as a spike after disregarding the other, the latter value was considered correct and hence of good quality.

Possible correction: The original profile data and the correct labelling were used to substitute bad data whenever possible.

Action: Spikes were flagged as bad quality data (QF=8). Temperature and salinity spikes were further verified by analysing density profiles: if density increased with depth, those spikes were flagged as probably good data (QF=2). In the case of chlorophyll *a*, the flagged spikes were further verified through a comparison with fluorescence profiles recorded with the CTD (when available), assigning a spike the probably good quality (QF=2)

Table 2. Ranges (RG) and corresponding threshold values (THV, mmol m^{-3}) for the identification of spikes for nutrient data in the LTER-MC dataset.

	AMON		***NTRI***		***NTRA***		***PHOS***		***SLCA***	
	RG	**THV**	**RG**	**THV**	**RG**	**THV**	**RG**	**THV**	**RG**	**THV**
From 0 to 25th percentile	<0.24	0.17	<0.05	0.06	<0.11	0.18	<0.03	0.02	<0.93	0.28
From 26th to 90th percentile	[0.24,1.4]	0.40	[0.05,0.43]	0.13	[0.11,1.34]	0.43	[0.03, 0.18]	0.05	[0.93, 2.79]	0.56
From 91st to 100th percentile	>1.4	0.87	>0.43	0.29	>1.34	0.92	>0.18	0.11	>2.79	1.40

Table 3. Range (RG) and corresponding threshold values (THV, mg m^{-3}) for the identification of spikes in chlorophyll *a* data in the LTER-MC dataset.

	CHLT	
	RG	**THV**
From 0 to 25th percentile	<0.28	0.16
From 26th percentile to 65th	[0.28, 0.77]	0.42
From 66th to 90th percentile	[0.77, 2.37]	0.84
From 91st to 100th percentile	>2.38	1.57

when the analysed profile corresponded to the fluorescence profile.

Test 5) Range tests

Problem: Anomalous data in a dataset can be identified based on defined ranges of variability for each environmental variable. However, because of the high variability of coastal water properties, values outside these ranges are not necessarily bad data and need further evaluation before being assigned a QF.

Flagging method: The central premise of this test is the definition of ranges to identify possible outliers. To accommodate the strong vertical and seasonal variability of the variables under study (Ribera d'Alcalà *et al.,* 2004), we defined the ranges differently according to depth and time (by month), and depending on the parameter being assessed.

For temperature and salinity, which show a moderately skewed distribution, with a clear range of variability and without extreme measurements, the most important property to evaluate is the consistency of the measurements with the period of the year and the depth. In this time series, measurements obtained with the CTD profiler from 2002 to 2014 provided a basis for comparing the older data and for defining ranges. To define the upper and lower limits of data at a given depth during a given month, we extracted the corresponding maximum and minimum values from the CTD dataset, obtaining ten depth-dependent values per month for the minimum and ten for the maximum. We used these values to generate two series of limits (Range 1 and Range 2) and two different tests.

For temperature:

T_{min} - 2.5%T_{min} < Range1<T_{max}+2.5%T_{max}

T_{min} - 5%T_{min} < Range2<T_{max}+5%T_{max}

For salinity:

S_{min} - 0.1%S_{min} < Range1<S_{max}+0.1%S_{max}

S_{min}-0.25%S_{min} < Range2<S_{max}+0.25%S_{max}

These two tolerance limits take into account the possible difference in measurements, deriving from changes in instrumentation for the parameter determination. Once the limits for each depth and for each month were defined (see Table 1 in http://qcbiogeodata.szn.it/), the whole temperature and salinity dataset were checked and measurements outside the limits for the respective depth and month were flagged as outliers.

For the chemical and biological data, the severe skewness revealed by the data distribution analyses made it necessary to transform the data before applying statistical methods. In addition, the distribution of these data hampered the definition of lower limits, as concentrations can be extremely low under specific conditions. For determining upper limits, a simple boxplot approach could mark as outlier many data that lay in the marginal position of the distribution curve. Therefore, we used a method that allows for an adjustment of the boxplot (Fig. 5) which includes a robust measure of skewness in the determination of the whiskers, resulting in a more accurate representation of the data and a more reliable identification of outliers (Hubert & Vandervieren, 2008). Instead of using the fence [Q1- 1.5 IQR; Q3 + 1.5 IQR] (where Q1 is the first quartile, Q3 the third quartile and IQR the interquartile range) for drawing the standard boxplot, the method defines the intervals as [Q1- h_l(MeC)IQR ; Q3 + h_u(MeC) IQR]. The h_l(MeC) and h_u(MeC) functions allow the fence to be asymmetric around the box, so that adjustment for skewness is indeed possible. The functions are based on a robust measure of skewness obtained using the *medcouple* estimator (Brys *et al*., 2004). This estimator is based on the definition of the quartile skewness (Bowley, 1920) and by using the kernel function, it derives the estimator that will replace the cut off values of outlying observations (Hubert & Vandervieren 2008). The *medcouple* belongs to the class of order statistics, e.g., incomplete generalised L-statistics, like the ordinary median, but it is a nonparametric statistic, thus it can be computed for any distribution.

All the data from a specific depth and month were considered to define the length of corresponding whiskers. Limits obtained by this procedure (Table 1 in http://qcbiogeodata.szn.it/) were used in the same way as the limits defined for the physical parameters: the whole dataset pertaining to chemical and biological parameters was checked and the measurements higher than the defined limits were flagged as outliers. To take into account the motility of plankton, we merged the 0 and 2 m concentration of chlorophyll *a* data, to define just one range for closely spaced depths.

Possible correction: None.

Action: For physical parameters, values exceeding the broadest Range 2, were considered bad data (QF=8),

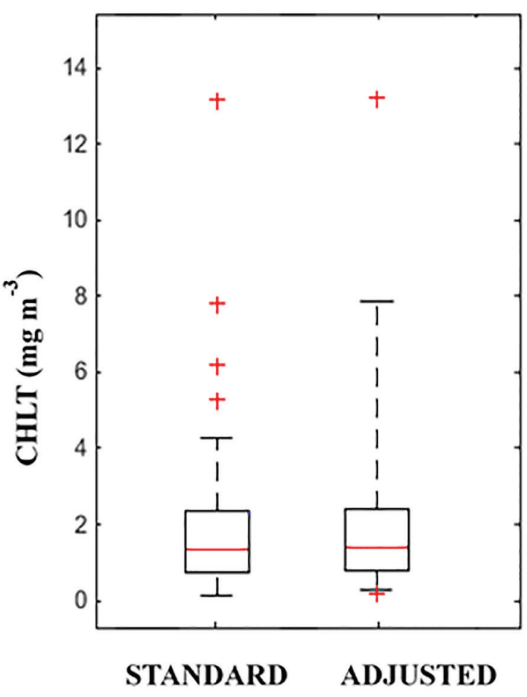

Fig. 5: Chlorophyll *a* data for April at 10 m depth: comparison of a standard (on the left) and an adjusted box plot (Hubert & Vandervieren, 2008) at LTER-MC.

while measurements outside Range 1, were subjected to further investigation by coupling with other tests (tests 6, 7 and 8). For chemical and biological parameters, only one limit was defined, and the measurements marked as outliers were subjected to further investigation and coupling with other tests (test 6 and 8).

Test 6) Bottom-surface gradient test

Problem: Bottom and surface water data cannot be tested by a spike analysis and, in the case of nutrients and chlorophyll *a*, their ranges only have an upper limit. If these data are markedly different from the adjacent ones, they could be considered outliers.

Flagging method: For temperature, we did not perform the test for 0 m data since a high variability is possible for surface waters. For the deepest data, 70 m values that were more than 1% higher than 60 m ones were considered outliers. Regarding salinity, surface values were tested against 2 m values and were marked as outliers when variation was higher than 0.2 %. For deep values, instead, a fixed limit was imposed and 70 m values that were 0.2 times 60 m values were marked as outliers.

Surface nutrient data were tested against 2 m values and were marked as outliers when they were lower than the 2 m values by more than 50%. Nutrient data at 70 m depth were marked as outliers when they were lower than the 60 m data by more than 35%. For chlorophyll *a* data, we tested only surface data against data from 2 m depth and marked as outliers those values that were lower than the 2 m values by more than 25%.

Possible correction: None

Action: Outliers were further investigated with other tests (tests 7 and 8). Temperature and salinity outliers were also compared with the density profile and marked as probably bad data (QF=6) when salinity values below the potential outlier were lower than above.

Test 7) Inter-sampling variability range test

Problem As temperature is a conservative parameter, a severe variation between two consecutive sampling events is not possible and a too high daily variation could indicate bad quality data.

Flagging method: The maximum rate of variation per day was determined using the 2002-2014 CTD dataset by comparing each temperature value with the value from the previous sampling event and dividing the difference by the days between the two sampling dates. Each value was also compared with that of the subsequent sampling date, often obtaining another rate of daily variation. Limits obtained by this procedure (Table 2 in http://qcbio-geodata.szn.it/) were used in the same way as the limits defined for test 5. The maximum daily variation rate was hence used to test the 1984-2001 dataset, in which the daily variation was calculated by comparing each measurement with that at the same depth of the previous sampling event. If this value was higher than the defined daily variation, it was flagged as outlier. The same procedure was applied to compare each temperature value to the value measured in the following sampling event.

Possible correction: None.

Action: This test was coupled with the variability range test or with the bottom-surface test: data marked as outliers in both the variability range test (range 1, the narrowest) and in this test were considered probably bad data (QF=6), otherwise as probably good data (QF=2).

Test 8) Parameter relationship test

Problem: In coastal areas, extreme events that lead to measurements outside the expected ranges are common and the risk of false positive values is high. However, extreme events are usually recorded by more than one variable. This test is aimed at validating data erroneously designated as outliers through the comparison with other variables. Most extreme events reflected in chemical and biological variability in the GoN are driven by terrigenous inputs. As these inputs most strongly affect surface waters, we performed this test only on data from 0-10 m depth.

Flagging method: In the first step, we analysed the strength and the statistical significance of linear correlations amongst the selected variables. We considered only those variables with a significant correlation coefficient ($p<0.01$) and used the outcomes to perform an in depth analysis to detect potential outliers. In the second step, we estimated the linear relationship between the selected variables by performing a classical regression analysis on each of the parameters conditional to the value of the predictor(s). Consistently with the first step of the procedure, we considered only those estimates of the regression model that were statistically significant ($p<0.01$). Then we focused on the distribution of the residuals (R) after performing model diagnostics to assess the assumptions of normality. In reference to the selected parameter, conditional to the predictor(s), data were flagged as outliers if the residual term $|r_i|$ was higher than a specified cut-off value, meaning that the relation with other parameters could not justify the extreme value of those measurements.

Hence if:

$$|r_i| > 2\sigma_R \qquad (1)$$

then the observed i^{th} value of the analysed parameter, conditional to the value of the selected predictors, is marked as an outlier, with σ indicating the standard deviation of the R distribution.

Before performing the test, we checked the statistical significance of the correlations described below.

Salinity: we analysed two regression models with salinity as dependent variable conditional to both chlorophyll *a* and silicates, one at a time. The correlation coefficients of salinity with either parameters were statistically significant (Pearson correlation coefficient for PSAL and CHLT: 0.506; and for PSAL and SLCA:-0.226, $p<0.01$). This is consistent with the hypothesis that freshwater inputs from land are coupled with strong nutrients inputs and are able to trigger phytoplankton blooms in surface waters (0-10 m). As nutrients may be consumed, or the bloom may have a time lag phase, anomalously low salin-

ity values can be supported by either high nutrient loads and/or by high biomass concentrations. Only those values identified as outlier in both regression models (salinity with chlorophyll *a* and salinity with silicates) were flagged as outliers after this test.

Nutrients: we checked for significant correlations among nutrients and with salinity. From the resulting correlation patterns, we performed regression for each nutrient with salinity and with another nutrient parameter, selecting the one with the highest correlation ($p<0.01$). We regressed nitrites with nitrates (Pearson correlation coefficients= 0.655), nitrates, ammonia and phosphate with silicates (Pearson correlation coefficients respectively=0.739; 0.602; 0.380). Also, silicates were compared with DIN (Dissolved Inorganic Nitrogen, obtained by the sum of nitrates, nitrites and ammonia) (Pearson correlation coefficients=0.755). In all these cases, the detection of outlying observations was carried out following the residual analysis procedure described in the flagging method, following equation 1.

Chlorophyll *a:* we analysed the relation of chlorophyll *a* with salinity and nitrates, separately, obtaining statistically significant correlations. In this case we decided to use a less stringent cut-off value for evaluating residuals because of the highly skewed distribution and the more variable range of chlorophyll *a*. In particular, chlorophyll *a* data whose residuals (in absolute terms) exceeded the standard deviation (σ) of the residual distribution (R) by more than four times can be considered outliers.

Possible correction: None.

Action: This test was used for samples that failed the range test, and allowed to flag the data as probably good (at least one correlation with a residual value in the identified range), or probably bad (residual outliers).

Test 9) Detection limit test

Problem: Nutrient concentrations below the method detection limit (MDL) are not necessarily bad data but cannot be considered accurate enough.

Flagging method: For each nutrient, every data point was compared with the method detection limit (0.1 mmol m^{-3} for silicates and 0.01 mmol m^{-3} for all other nutrients) and data below these values were flagged.

Possible correction: None.

Action: Data that fail the test are flagged as data below detection limit (QF=3).

Quality flags

After performing all tests, we assigned the following flags:

Data with **QF =0** passed all the applicable tests.

Data with **QF=2** failed the range test (test 5), or the bottom/surface gradient test (test 6), but passed the inter-sampling variability test (only for temperature, test 7) or at least one of the correlation tests (test 8).

Data with **QF=3** were below the detection limit.

Data with **QF=4** failed the range test (test 5), or the bottom/surface gradient test (test 6) and did not undergo further testing.

Data with **QF=5** were reconstructed data (from metadata analysis).

Data with **QF=6** failed the range (test 5), and did not pass the bottom surface test (test 6), the inter-sampling variability test (test 7), nor a minimum of one of the correlation tests.

Data with **QF=7** failed the duplicated profile test (test 2) or the frozen profile test (test 3).

Data with **QF=8** failed the range test 2 for temperature and salinity, or failed the spike test (test 8).

QF=9 indicated missing data (test 1).

This quality flags scheme provides a comprehensive description of the quality of the data and is implemented for expert users of the dataset, such as data maintainers and producers. In a simplified version, the scheme proposed in Table 1 can be summarized into a four flag scheme, partially coinciding with the ODV quality flag scheme (http://odv.awi.de/en/documentation/) and hence ensuring an automatic recognition of the quality flags by the ODV software. The simplified scheme also provides a more immediate comprehension of the quality of the dataset and allows for a quicker elimination of the bad quality data.

Good= passed all the applied QC tests. This combines our flags from 0 to 3, since data that are considered probably good (flag 2) can be included in the good quality data. The same happens for data below detection limit (flag 3) that, although not reliable in absolute values, are anyway representative of low nutrient concentrations.

Questionable/suspect = inconclusive – failed non-critical metric or subjective QC test(s). This encompasses the flags 4 and 5.

Bad = failed critical metric QC tests, which includes our flag from 6 to 8.

Missing data, indicated as flag 9 (not included in the ODV flag scheme).

Results

Quality Control of the LTER-MC dataset

An example of detailed flagging is presented for temperature by months in Figure 6.

The general results for the whole dataset (Fig. 7 and Table 4) show a very high percentage of good quality data (89%), of which 86% good, 1% probably good and 2% below the detection limit. Among the nutrients, phosphates are more often below the detection limit than any of the other nutrients, whereas ammonia and silicates are rarely below the detection limit. The second most abundant flag is 9, missing data, ranging around 5% and 9.5% of the total amount of data for all the parameters, except for salinity for which about 2% of data were missing. Bad data only constitute approximately 2% of the total amount of data, while 1% resulted as manipulated data. Dubious data (~ 1% of the total) are not present in temperature data, which were verified with the inter-sam-

Table 4. Percentage of quality flags for each parameter in the LTER-MC dataset.

	GOOD			QUESTIONABLE			BAD		MISSING
	Good	Probably good	Below DL	Dubious	Reconstructed	Probably bad	Manipulated	Bad	Missing
	0	2	3	4	5	6	7	8	9
TEMPERATURE	88.7	0.5	0.0	0.0	1.8	0.0	0.7	0.8	7.4
SALINITY	94.8	0.7	0.0	0.7	0.0	0.0	0.5	1.1	2.2
AMMONIA	84.9	0.8	0.1	1.9	0.0	0.3	1.6	2.4	8.0
NITRATES	84.2	0.4	2.9	1.3	0.0	0.1	1.8	1.5	7.8
NITRITES	86.0	0.5	3.3	1.3	0.0	0.2	1.6	2.0	5.1
PHOSPHATES	74.5	0.4	4.4	1.0	0.0	0.3	2.2	7.7	9.5
SILICATES	86.0	0.7	0.7	1.0	0.0	0.3	0.3	2.0	9.0
CHLOROPHYLL a	87.5	1.5	0.0	0.9	0.0	0.1	0.0	0.6	9.4

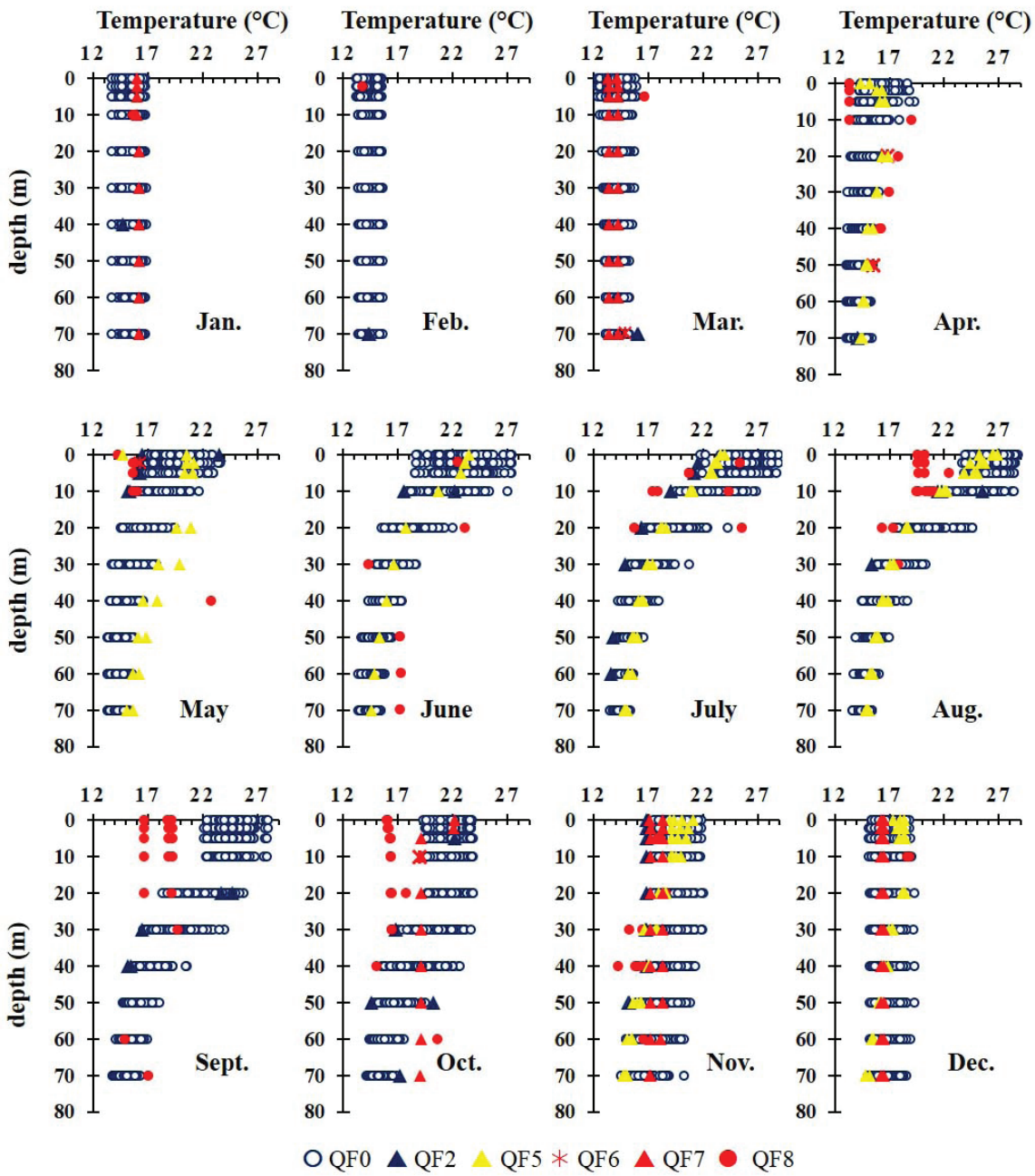

Fig. 6: Colour-coded flags for temperature data in monthly vertical temperature profiles in the LTER-MC dataset.

Table 5. Percentage of quality flags in the LTER-MC dataset according to the simplified QF scheme.

LAYER 0-10 m				
	GOOD	QUESTIONABLE	BAD	MISSING
TEMPERATURE	84.39	1.69	2.03	11.89
SALINITY	95.41	0.00	1.35	3.24
AMMONIA	88.22	0.00	4.66	7.12
NITRATES	89.75	0.00	3.36	6.89
NITRITES	92.43	0.00	3.06	4.50
PHOSPHATES	80.79	0.00	10.23	8.99
SILICATES	88.72	0.00	2.70	8.58
CHLOROPHYLL *a*	90.27	0.00	0.92	8.81
LAYER 20-40 m				
	GOOD	QUESTIONABLE	BAD	MISSING
TEMPERATURE	91.02	1.74	1.62	5.62
SALINITY	96.22	1.11	1.44	1.23
AMMONIA	84.98	1.44	5.05	8.53
NITRATES	86.16	1.74	3.66	8.44
NITRITES	89.10	1.20	3.87	5.83
PHOSPHATES	78.23	1.29	10.75	9.73
SILICATES	86.16	1.50	2.76	9.58
CHLOROPHYLL *a*	87.50	2.38	0.70	9.42
LAYER 50-70 m				
	GOOD	QUESTIONABLE	BAD	MISSING
TEMPERATURE	91.29	1.80	1.02	5.89
SALINITY	95.02	1.32	2.01	1.65
AMMONIA	83.48	4.74	3.18	8.59
NITRATES	85.95	2.58	3.15	8.32
NITRITES	86.82	3.18	4.71	5.29
PHOSPHATES	78.59	2.07	9.49	9.85
SILICATES	86.79	1.92	2.25	9.04
CHLOROPHYLL *a*	86.93	1.52	0.00	11.55

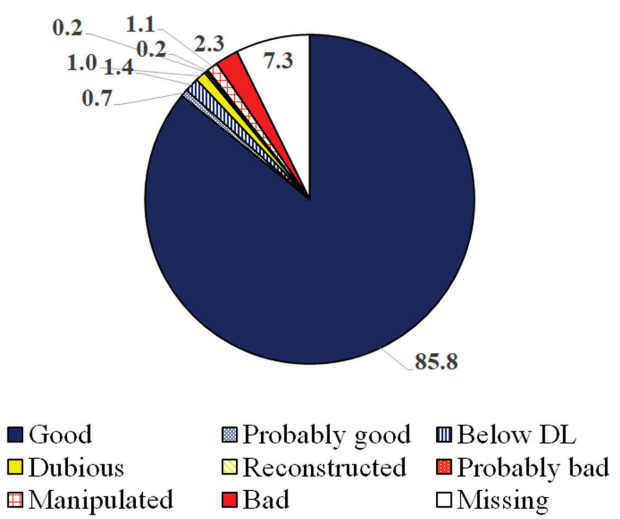

Fig. 7: Percentages of quality flags attributed through the QC procedure to the whole LTER-MC dataset (~84,000 data-points).

pling variability test. Data were flagged as probably bad when this test failed, and as probably good otherwise. Of the other parameters, ammonia and nitrates have the highest percentage of out of range data for which it was not possible to check the validity with other tests. Flag 5 (reconstructed data) is only present for temperature data in the year 2001. Phosphates have the highest percentage of bad and probably bad data (see Table 4). Together with 10% of missing and questionable data, 20% of observations should not be considered for further analyses. Salinity shows the highest percentage of good data, while all the other parameters have comparable percentages of good data.

Looking at the vertical distribution, the highest percentage of bad data is observed in the intermediate layer (Table 5). This result may reflect the complex dynamics of this layer, which is alternatively affected by surface and bottom waters and thus shows a high variability that is impossible to distinguish statistically from the errors. Similarly, the maximum percentage of bad data is found for the summer months (July and August, data not shown). This period is characterised by sudden alternations of coastal and offshore conditions, resulting in abrupt changes in hydrographic features (Zingone *et al*., 1990; D'Alelio *et al.,* 2015). Such high natural variability is difficult to be taken into account in tests without affecting the effectiveness of the QC.

The percentage of questionable data is negligible for surface data because only reconstructed data belong to this class and the correlation test allows flagging these

Table 6. Ranges (RG) and corresponding threshold values (THV) for the determination of the spike for nutrients (mmol m^{-3}) and chlorophyll *a* (mg m^{-3}) in the TYR dataset.

	AMON		*NTRI*		*NTRA*		*PHOS*		*SLCA*		*CHLT*	
	RG	**THV**	**RG**	**THV**	**GR**	**THV**	**RG**	**THV**	**RG**	**THV**	**RG**	**THV**
From 0 to 25th percentile	<0.19	0.12	<0.02	0.04	<0.03	0.32	<0.014	0.01	<0.85	0.37	<0.19	0.15
From 26th to 90th percentile	[0.19, 0.98]	0.28	[0.02, 0.26]	0.08	[0.03, 2.18]	0.75	[0.014, 0.085]	0.02	[0.85, 3.32]	0.86	[0.19, 1.17]	0.34
From 91st to 100th percentile	>0.98	0.59	>0.26	0.18	>2.18	1.71	>0.085	0.05	>3.32	1.85	>1.17	0.73

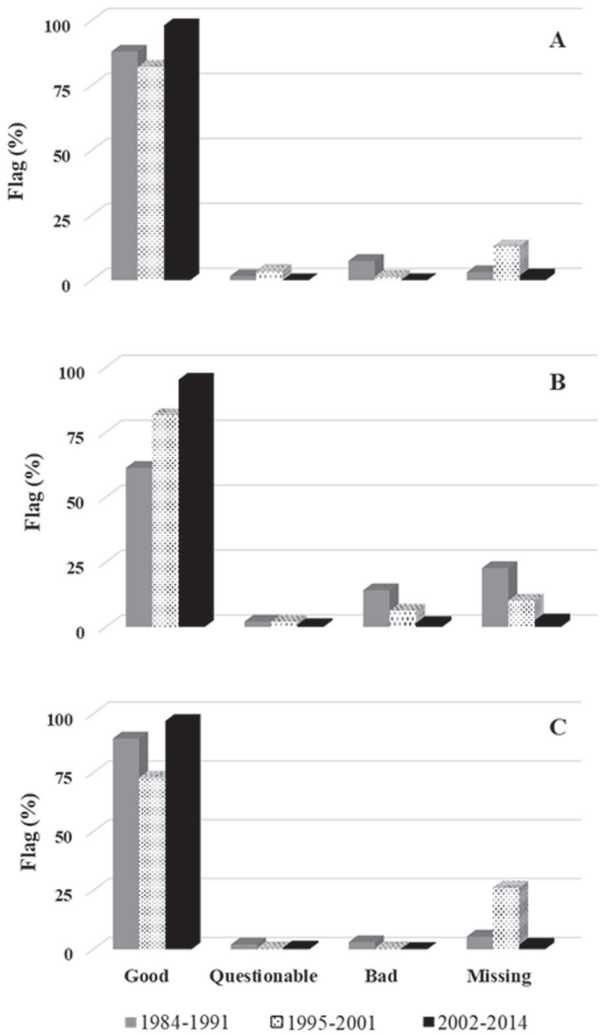

Fig. 8: Percentage of assigned quality flags for **a**) temperature and salinity, **b**) nutrients and **c**) chlorophyll *a* in the LTER-MC dataset grouped by period, following the simplified QF scheme.

data as probably good or probably bad. On the other hand, the percentage of missing data is minimal in the intermediate layer, as bad weather conditions limit CTD data acquisition mostly in the surface and bottom layers.

Over the years, the number of bad data has diminished for all the parameters. Figure 8 shows the percentage of good, questionable, bad and missing data for three groups of variables in three different periods. For physical variables and chlorophyll *a* (Fig. 8 a, c), the percentage of good data is lower during the second period with respect to the first period, coupled with an increase of missing data, but increases again afterwards, and the number of bad data decreases from the beginning of the series to 2014. For nutrients, instead (Fig. 8 b), the quality has improved monotonically and markedly during the years. The percentage of good data during the first period is around 70% for all the nutrients. In the second and in the third period the percentage of good data increases to more than 80% and 95% respectively, while the percentage of questionable, bad and missing data decreases, indicating a greater accuracy in both sampling and analyses.

Validation tests on different datasets

In order to evaluate the general applicability of the most critical tests designed for the LTER-MC dataset, we applied the range and the correlation tests to the Si.Di. Mar. dataset and the spike test to the TYR dataset.

Range test and correlation test: Punta Licosa and Punta Tresino Si.Di.Mar. datasets

The more oligotrophic conditions in the GoS compared to the LTER-MC area in the GoN resulted in upper nutrient concentration limits that were an order of magnitude lower than those obtained for surface data of the LTER-MC dataset, using the same criteria described in the previous section for the range test (test 5) (see Table 3 at http://qcbiogeodata.szn.it). The correlation tests were able to justify, on an average, two thirds of the outliers identified in the test 5, leading to very few data being flagged as probably bad, as reported for nitrates in Figure 9. For most nutrients, correlations with salinity and with silicates or nitrates were found. The only exception was ammonia, which was not significantly correlated to salinity. This is reasonable, because in such oligotrophic environments the presence of ammonia is influenced by regenerated production rather than by anthropogenic input. For ammonia, we used the significant correlations with silicates and DIN to perform the quality control test.

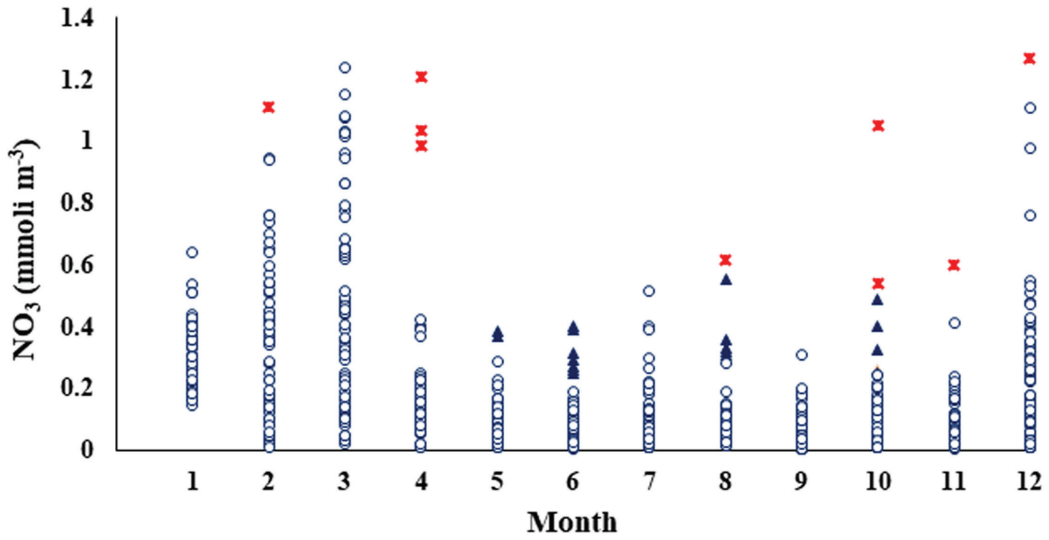

Fig. 9: Flags (○ QF0, ▲QF2, and ✳QF6) for surface nitrates (NO_3) concentrations in the Si.Di.Mar. dataset.

Spike test: the TYR dataset

After checking the distribution, asymmetry and kurtosis of each parameter, we chose range and thresholds to define the spikes according to the distribution following the criteria described above (test 4). In the TYR dataset, nutrient data distributions were comparable to those seen in the LTER-MC dataset. In contrast, chlorophyll *a* data displayed a less skewed distribution than in the LTER-MC dataset and asymmetry and kurtosis values were comparable to those of nutrients. For this reason, we identified only three ranges and, consequently, three thresholds to define the spikes (Table 6). The second difference was the frequent presence of high values within the data profile that were initially classified as spikes but could indicate a Deep Chlorophyll Maximum (DCM), generally absent at LTER-MC. In this case, the comparison with fluorescence data conclusively supported their assignment to DCM data, thus allowing the preservation of a big percentage of data initially considered erroneous.

Discussion

The recommendations concerning data management that resulted from the 2010 IODE meeting on QC (IOC, 2010) were intended as guidelines that could be agreed upon and used by all oceanographic data centres. A minimum set of numerically defined QC tests were proposed, based on quantifiable and generally objective tests, which included the data range, excessive gradient, excessive spike and "no gradient" checks, but no indications were provided to identify precise procedures that could be of general application, while accounting for site specificity. The use of basin climatology as a reference does not set the proper ranges for many coastal areas, even within the same basin. Unfortunately, as already outlined by Campbell *et al.* (2013), there are no universal standards that are applicable in all circumstances, since QC procedures must be designed specifically for the type of data and for the location in which they are collected. This led us to develop additional criteria for an accurate QC of coastal biogeochemical datasets. Our results, built on a quite large dataset, integrate the general criteria proposed by previous workshops and symposia (e.g., IOC, 2010) by providing more refined criteria and proposing a possible approach for QC procedures in an environment characterised by a hardly constrainable variability.

The first consideration stemming from our approach was that QC procedures for coastal datasets require a deep knowledge of the functioning of the area and a preliminary study of the large amount of data accumulated, as in the case of the LTER datasets. The rationale of our study was: 1) to check the validity of each single measurement based on the properties of the whole bulk of the dataset; 2) to verify whether the proposed criteria can reliably be used for coastal sites in general. What clearly emerged from this effort is that the process of implementing a QC procedure tuned on a specific dataset is an iterative, trial and error process that cannot be completely free of subjective, expert knowledge-based decisions. As an example, in some cases the high number of frozen and duplicated profiles, observed for some parameters in selected years, highlighted the need to exclude those data from the definition of ranges for the whole LTER dataset and from parameter relationship tests. Furthermore, we found that adequate statistical tools able to capture the peculiarities of the examined data can strongly improve the QC procedure. This is exemplified by the use of the adjusted boxplot (Hubert & Vandervieren, 2008) which allowed taking into account the skewed nature of a parameters' distribution, thereby rescuing data that would have resulted as outliers with a standard boxplot.

In temperate sites, seasonality plays a pivotal role in shaping the variability of biogeochemical properties. This is particularly true at LTER-MC where the seasonal signal is among the strongest factors in determining the annual variability of biogeochemical parameters (Cloern

& Jassby, 2010). Moreover, the distribution of chemical and biological parameters displays strong vertical gradients (Ribera d'Alcalà *et al*., 2004). For example, at a given depth salinity can be substantially different in March with respect to October (see Table 1 at http://qcbiogeodata.szn.it/). For this reason, the range test was adapted for each depth in each month. However, these criteria may only be suitable for dynamics at temperate latitudes, and different criteria and adjustments are needed when considering datasets from tropical or polar sites, with completely different dynamics.

A crucial result of our study is the use of the parameter relationship test (test 8): the high variability in surface coastal waters or intense meteorological events can lead to seemingly anomalous data that are nonetheless representative of environmental conditions and can be detected in more than one parameter, as observed also in other LTER sites (e.g., Gulf of Trieste, Lipizer *at al.,* 2012a). In our study, the reliability of some data was tested comparing them with the values of the two variables displaying the highest covariance. This allowed us to complement the test of the range, which is hard to constrain in coastal systems, with a test based on the observed long-term dynamics of the site. Assessing the reliability of high values will in turn allow setting more consistently the specific ranges for the site.

Among the tests suggested by IODE (IOC, 2010), we disregarded the excessive gradient test applied to vertical profiles. Although being one of the most common QC tests (Wong *et al*., 2015), it cannot be applied at a coastal site, where terrestrial inputs can lead to events of strong stratification and therefore very steep gradients. In our QC, we used the excessive gradient test only on a time-basis (test 7), as a criterion to test the reliability of consecutive sampling events for temperature, whose variations are dominantly driven by heat fluxes and depend very weakly from horizontal transport.

The results obtained through the application of the QC tests to the LTER-MC dataset demonstrate the value of their use for assessing the reliability of those data. Overall, the quality of the LTER-MC dataset is quite good, with only 2% of bad data in a dataset amounting to more than 84,000 data. Such percentage does not affect significantly the interpretation of the LTER-MC dataset. On the other hand, the QC highlighted several problems occurring over the first years of the sampling and/or some inaccuracies in data storage, probably due to the participation of several people to the process and to the different technologies used to handle the same dataset. Results of the first period from 1984-1991, despite their contribution to a first characterization of the site (Scotto di Carlo *et al.*, 1985), mainly acted as a trial stage, after which the sampling was interrupted for 4 years. In the first part of the second period, from 1995 to 2001, the quality slowly improved. Finally, after 2002 the technological improvement, including automated equipment for sampling and profiling, allowed to obtain much more reliable data. As expected, during the last period, when the group handling the dataset was small and composed by the same people, the percentage of bad data was significantly lower. This suggests that there is an intrinsic variability in generating data added by the operators, despite the use of the same equipment and protocols. This should be taken into account when characterizing trends in time series (Whiltshire & Durselen, 2004). Lastly, the gradual improvement of informatics systems for data management strongly enhanced the quality of the data stored, minimizing the errors due to the step-wise procedure to go from sample acquisition to parameter computation.

The tests performed on the Si.Di.Mar. and TYR dataset confirmed that the criteria and tests designed for the LTER-MC dataset could be extended to other datasets. They also demonstrated that some criteria must be tuned to the specific characteristic of different areas, which again shows the value of using both expert knowledge and preliminary analysis of the dataset in the QC of coastal data.

The design of a procedure for QC of coastal biogeochemical data and the results of its application allow drawing some general considerations and recommendations. First, in the building-up phase it emerges as essential to make good use of complementary parts of a dataset in the evaluation of the quality of selected data, for example, by further expanding the consistency test to the highest number of relevant variables (e.g., pigments vs. cell counts, inorganic nitrogen vs. total nitrogen) or by considering the intrinsic limits of variability between results of consecutive sampling events (e.g., temperature or pigment variations over time cannot exceed realistic values within the same water mass). Second, our results show the value of flagging dubious data rather than eliminating them, which still allows removing those data for specific analyses and rather use interpolated data, at the same time preserving the information that extreme values may contain for future studies and re-examination. Lastly, building up a QC procedure proves to be an iterative process needing reconsiderations and retuning in accordance with the results of long-term trend analysis, eventually leading to retune criteria of data quality assessment and reconsider data initially flagged as wrong.

Besides testing the quality of our data set, the main motivation of our exercise was to attract the attention of the community to the need of developing consensus procedures to make coastal data inter-comparable in time and space, despite changes in methodologies, operators, and programs. We are aware that our approach is affected by subjective decisions, but these decisions were based on expert knowledge and data distribution curves. Likewise, in other sites research groups have developed their own routine procedure to assess the quality of their data (e.g. Raabe & Whiltshire, 2009; Segura-Noguera *et al.*, 2011; Lipizer *et al.*, 2012b). Beyond the local interest, this first effort to identify QC procedures that may be applicable to other heterogeneous coastal oceanographic datasets would greatly benefit from integrations and improvements through the comparison with 'in house', unofficial methodologies used in other sites, to reach a consensus protocol, similarly to what has been developed for open ocean data sets. Indeed, sharing and comparing multiple procedures for quality control is the most ad-

vantageous strategy for the development of an inter-site data validation, which may help in minimizing the subjective component in QC procedures. The application of standardised tests with site-specific criteria will greatly assist in assessing the reliability and allow the comparability of ecological data from coastal regions. Our final remark is on the importance of promoting a discussion on the biogeochemical data quality issue, the solution of which is a prerequisite for a proper coastal management in these marine environments that are most strongly affected by global change. This is particularly crucial in the Anthropocene, when the human impact on the environment is both strong and fast.

Acknowledgements

The basic structure of this work was developed in 2013 in the framework of the project RITMARE. This work was supported partly by a research contract [to L. S.] within the framework of the TERNA project and partly with funds provided by the Stazione Zoologica Anton Dohrn. The authors wish to thank Sietske J. Batenburg for a critical lecture and revision of the English version of this paper and Florian Kokoszka for a final check of the Matlab functions and for useful comments and advice for the readability of the procedures. The authors are grateful to M. Cannavacciuolo, C. Chiaese, F. Conversano, F. Corato, A. Passarelli, F. Tramontano, G. Zazo and all the other people who, since January 26, 1984 sampled, analysed and preserved the LTER-MC dataset, collecting such valuable information.

References

Aoyama, M., Abad, M., Anstey, C., Ashraf, P.M., Bakir, A. *et al.*, 2016. *IOCCP-JAMSTEC 2015 Inter-Laboratory calibration exercise of a certified reference material for nutrients in seawater.* Japan Agency for Marine-Earth Science and Technology, Report Number 1/2016, 176 pp.

Beckers, J.M, Rixen, M., 2003. EOF Calculations and data filling from incomplete oceanographic datasets. *Journal of Atmospheric and Oceanic Technology,* 20, 1839-1856.

Bowley, A.L., 1920. *Elements of Statistics*. Charles Scribner's Sons, New York, 330 pp.

Brys, G., Hubert, M., Struyf, A., 2004. A robust measure of skewness. *Journal of Computational and Graphical Statistics,* 13, 996-1017.

Burnett, W.C., Bokuniewicz, H., Huettel, M., Moore, W.S., Taniguchi, M., 2003. Groundwater and pore water inputs to the coastal zone. *Biogeochemistry,* 66, 3-33.

Campbell, J.L., Rustad, L.E., Porter, J.H., Taylor, J.R., Dereszynski, E.W. *et al.,* 2013. Quantity is nothing without quality: automated QA/QC for streaming environmental sensor data. *Bioscience,* 63 (7), 574-585.

Cloern, J.E., 2001. Our evolving conceptual model of coastal eutrophication problem. *Marine Ecology Progress Series,* 210, 223-253.

Cloern, J.E., Jassby, A.D., 2010. Patterns and scales of phytoplankton variability in estuarine-coastal ecosystems. *Estuaries and Coasts,* 33, 230-241.

Conkright, M.E., Boyer, T.P., Levitus, S., 1994. *Quality control and processing of historical oceanographic nutrient data.* NESDIS 79 NOAA Technical Report, 85 pp.

D'Alelio, D., Mazzocchi, M.G., Montresor, M., Sarno, D., Zingone, A. *et al*., 2015. The green–blue swing: plasticity of plankton food-webs in response to coastal oceanographic dynamics. *Marine Ecology - An Evolutionary Perspective,* 36, 1155-1170.

de Boyer Montegut, C., Madec, G., Fischer, A.S., Lazar, A., Iudicone, D., 2004. Mixed layer depth over the global ocean: an examination of profile data and a profile-based climatology. *Journal of Geophysical Research,* 109, C12003.

Doney, S.C., 2010. The growing human footprint on coastal and open ocean biogeochemistry. *Science*, 328 (5985), 1512-151.

EMODnet Chemistry, 2018. *European Marine Observation and Data Network.* http://www.emodnet-chemistry.eu/welcome (Accessed 26 June 2018).

Hansen, H.P., Grasshoff, K., 1983. Automated chemical analysis, p. 347-379. In: *Methods of Seawater Analysis*. Grasshoff, K., Ehrhardt, M., Kremung, K. (Eds). Verlag Chemie, Weinheim.

Holm-Hansen, O., Lorenzen, C.J., Holmes, R.W., Strickland, J.D.H., 1965. Fluorometric determination of chlorophyll. *Journal du Conseil Permanent International pour l'Exploration de la Mer,* 30, 3-15.

Horsburgh, J.S., Reeder, S.L., Jones, A.S., Meline, J., 2015. Open sourced software for visualization and quality control of continuous hydrologic and water quality sensor data. *Environmental Modelling & Software,* 70, 32-44.

Hubert, M., Vandervieren, E., 2008. An adjusted boxplot for skewed distributions. *Computational Statistics and Data Analysis,* 52 (12), 5186-5201.

Iermano, I., Liguori, G., Iudicone, D., Buongiorno Nardelli, B., Colella, S. *et al*., 2012. Filament formation and evolution in buoyant coastal waters: observation and modelling. *Progress in Oceanography,* 106, 118-137.

Ibe, A.C., Kullenberg, G., 1995. Quality Assurance/Quality Control (QA/QC) regime in marine pollution monitoring programmes: The GIPME perspective. *Marine Pollution Bulletin,* 31 (4-12), 209-213.

IOC (Intergovernmental Oceanographic Commission) of UNESCO. 2010. *First IODE workshop on quality control of chemical oceanographic data collections*. IOC Project Office for IODE, Oostende, Belgium, 8-11 February 2010 Paris, UNESCO, IOC Workshop Report No. 228, 18 pp.

IOC (Intergovernmental Oceanographic Commission) of UNESCO. 2013. Ocean Data Standards, Vol.3: *Recommendation for a Quality Flag Scheme for the Exchange of Oceanographic and Marine Meteorological Data*. Paris, IOC Manuals and Guides, 54: 3, 12 pp.

Jakobsen, H.H., Carstensen, J., Harrison, P.J., Zingone, A., 2015. Estimating time series phytoplankton carbon biomass: Inter-lab comparison of species identification and comparison of volume-to carbon scaling ratios. *Estuarine, Coastal and Shelf Science,* 162, 143-150.

Jickells, T.D., 1998. Nutrient Biogeochemistry of the Coastal Zone. *Science,* 281, 217-222.

Lauvset, S.K., Tanhua, T., 2015. A toolbox for secondary quality control on ocean chemistry and hydrographic data. *Limnology and Oceanography: Methods,* 13, 601-608.

Levitus, S., 1982. *Climatological Atlas of the World Ocean*. NOAA Professional Paper 13, Rockville, Md, 173 pp.

Lipizer M., De Vittor C., Falconi C., Comici C., Tamberlich F. *et al.*, 2012a. Effects of intense physical and biological forcing factors on CNP pools in coastal waters (Gulf of Trieste, Northern Adriatic Sea). *Estuarine, Coastal and Shelf Science,* 115, 40-50.

Lipizer, M., De Vittor, C., Falconi, C. F., Comici, C., Kralj M. *et al.*, 2012b. *Long Term Ecological Research (LTER) site in the Gulf of Trieste – C1 station. Inventory of sampling and analytical methods and quality control of biogeochemical data.* OGS, Technical Report 2012/86 OCE 3 BIPPA, 15 pp.

LTER Europe, 2018. *Long Term Ecosystem Research in Europe.* http://www.lter-europe.net/lter-europe/about/ep-tf/tf-marine (Accessed 13 May 2018).

Marino, M., Modigh, M., Zingone, A., 1984. General features of phytoplankton communities and primary production in the Gulf of Naples and adjacent waters, p. 89-100. In: *Marine Phytoplankton and Productivity*, O. Holm-Hansen, L. Bolis and R. Gilles, Springer-Verlag, Berlin.

McQuatters-Gollop, A., Edwards, M., Helaouet, P., Johns, D.G., Owens, N.J.P. *et al*., 2015. The continuous plankton recorder survey: how can long term phytoplankton datasets contribute to the assessment of Good environmental status? *Estuarine, Coastal and Shelf Science,* 162, 88-97.

Michener, W.K., 2016. Advances in managing Long Term Ecological Research Data. *Ecological Informatics,* 36, 199-200.

Moatar, F., Miquel, J., Poirel, A., 2001. A quality control method for physical and chemical monitoring data. Application to dissolved oxygen levels in the River Loire France. *Journal of Hydrology,* 252, 25-36.

Müller, T.J., 1983. Determination of salinity, p. 41-73. In: *Methods of Seawater Analysis*. Grasshoff, K., Ehrhardt, M., Kremung, K. (Eds). Verlag Chemie, Weinheim.

Neveux, J., Panouse, M., 1987. Spectrofluorometric determination of chlorophyll and pheophytins. *Archiv fur Hydrobiologie*, 109, 567-581.

ODV, 2018. *Ocean Data View*. http://odv.awi.de/ (Accessed 13 May 2018).

Raabe, T., Whiltshire, K.H., 2009. Quality control and analyses of the long-term nutrient data from Helgoland Roads, North Sea. *Journal of Sea Research,* 61, 3-16.

Ribera d'Alcalà, M., Conversano, F., Corato F., Licandro, P., Mangoni, O. *et al*., 2004. Seasonal patterns in plankton communities: an attempt to discern recurrences and trends. *Scientia Marina,* 68 (1), 65-83.

Scotto di Carlo, B., Tomas, C.R., Ianora, A., Marino, D., Mazzocchi, M.G. *et al*., 1985. Uno studio integrato dell'ecosistema pelagico costiero del Golfo di Napoli. *Nova Thalassia*, 7, 99-128.

Gra, 2018. *Pan-European infrastructure for ocean and marine data management*. https://www.seadatanet.org/ (Accessed 26 June 2018).

Segura-Noguera, M., Cruzado, A., Blasco, D., 2011. Nutrient preservation, analysis precision and quality control of an oceanographic database of inorganic nutrients, dissolved oxygen and chlorophyll a from the NW Mediterranean Sea. *Scientia Marina,* 75 (2), 321-339.

Sheldon, Jr. W.M., 2008. Dynamic, rule-based quality control framework for real time sensor data. p. 145- 150. In: *Proceedings of the Environmental Information Management Conference 2008: Sensor Networks*. Albuquerque, NM, 10-11September 2008.

Strickland, J.D.H., Parsons, T.R., 1972. A practical handbook of sea water analysis. *Bulletin of the Fisheries Research Board Of Canada,* 167, 1-310.

Tanhua, T., van Heuven, S., Key, R. M., Velo, A., Olsen, A. *et al.*, 2010. Quality control procedures and methods of the CARINA database. *Earth System Science Data*, 2, 35-49.

Wagner, R.J., Boulger, R.W., Jr., Oblinger, C.J., Smith, B.A., 2006. *Guidelines and standard procedures for continuous water-quality monitors-Station operation, record computation, and data reporting: U.S. Geological Survey Techniques and Methods* 1–D3, 51 pp. + 8 attachments. http://pubs.water.usgs.gov/tm1d3 (Accessed 13 May 2018).

Whiltshire, K., Dürselen, C.D., 2004. Revision and quality analyses of the Helgoland Reede long-term phytoplankton data archive. *Helgoland Marine Research,* 58 (4), 252-268.

Wong, A., Keeley, R., Carval, T., Argo Data Management Team, 2015. *Argo Quality Control Manual for CTD and Trajectory Data*. IFREMER Report, France, 56 pp.

Zingone, A., Harrison, P.J., Kraberg, A., Lehtinen, S., McQuatters-Gollop, A. *et al*., 2015. Increasing the quality, comparability and accessibility of phytoplankton species composition time-series data. *Estuarine, Coastal and Shelf Science,* 162,151-160.

Zingone, A., Montresor, M. Marino, D., 1990. Summer phytoplankton physiognomy in coastal waters of the Gulf of Naples. *Marine Ecology - Pubblicazioni della Stazione Zoologica di Napoli I,* 11 (2), 157-172.