Improved Statistical Method for Quality Control of Hydrographic Observations

Jérôme Gourrion, Tanguy Szekely, Rachel Killick, Breck Owens, Gilles Reverdin, Bertrand Chapron

Improved Statistical Method for Quality Control of Hydrographic Observations

JÉRÔME GOURRION AND TANGUY SZEKELY

OceanScope, Plouzané, France

RACHEL KILLICK

Met Office, Exeter, United Kingdom

BRECK OWENS

Woods Hole Oceanographic Institution, Woods Hole, Massachusetts

GILLES REVERDIN

Sorbonne-Université, CNRS/IRD/MNHN (LOCEAN UMR 7159), Paris, France

BERTRAND CHAPRON

LOPS, Ifremer, Plouzané, France

(Manuscript received 26 January 2019, in final form 12 December 2019)

ABSTRACT

Realistic ocean state prediction and its validation rely on the availability of high quality in situ observations. To detect data errors, adequate quality check procedures must be designed. This paper presents procedures that take advantage of the ever-growing observation databases that provide climatological knowledge of the ocean variability in the neighborhood of an observation location. Local validity intervals are used to estimate binarily whether the observed values are considered as good or erroneous. Whereas a classical approach estimates validity bounds from first- and second-order moments of the climatological parameter distribution, that is, mean and variance, this work proposes to infer them directly from minimum and maximum observed values. Such an approach avoids any assumption of the parameter distribution such as unimodality, symmetry around the mean, peakedness, or homogeneous distribution tail height relative to distribution peak. To reach adequate statistical robustness, an extensive manual quality control of the reference dataset is critical. Once the data have been quality checked, the local minima and maxima reference fields are derived and the method is compared with the classical mean/variance-based approach. Performance is assessed in terms of statistics of good and bad detections. It is shown that the present size of the reference datasets allows the parameter estimates to reach a satisfactory robustness level to always make the method more efficient than the classical one. As expected, insufficient robustness persists in areas with an especially low number of samples and high variability.

1. Introduction

Monitoring and predicting the climate evolution at short and longer time scales has been, is, and will be for years to come the main challenge for the Earth sciences community. For the atmospheric, continental and oceanic domains, this challenging task benefits from the information provided by the ever-increasing observation networks and from an increased understanding of the physical and chemical mechanisms contributing to the dynamics of these coupled systems. In practice, these mechanisms are simulated under both physical and mathematical assumptions and technical constraints. Thus, these climate analyses and predictions cannot avoid inherent uncertainties that may make them unrealistic.

In the last few decades, taking advantage of advances in the atmospheric analyses, the ocean community has focused on atmospherically forced ocean models to

Corresponding author: Jérôme Gourrion, jerome.gourrion@ocean-scope.com

DOI: 10.1175/JTECH-D-18-0244.1

better reproduce the available observations, see Le Sommer et al. (2018). Model simulations can provide large and homogeneous sampling, but these models are not a complete version of reality, due to their imperfect representation of the full dynamics and numerical imprecision. How can these predictions be improved using observations that sample an unfiltered reality but with heterogeneous and incomplete sampling? To answer this question, two main activities have received the most attention: 1) development and maintenance of in situ observation networks and 2) design of adequate numerical strategies for data assimilation. In Europe, within the Copernicus Marine Environment Monitoring Service, these activities are conducted at global and regional scales, both in real time and delayed time. At the global scale, Mercator-Océan carries out the modeling and assimilation activities while Coriolis is involved in the observational ones. For these complementary activities to succeed, an essential and critical activity is the data quality control (QC). This paper focuses on QC procedures.

For meteorological data, Gandin (1988) distinguishes three categories of errors: random, systematic and gross errors. Random errors are due to instrument behavior itself and unresolved environmental variability influencing the instrument; they are intermittent and cannot be eliminated, but it is often reasonable to describe them as white noise using a Gaussian probability distribution, zero mean, and specified variance. Systematic errors are usually asymmetrically distributed, and their mean value is called bias. They are usually caused by an unaccounted for, persistent shift in the measurement. These biases usually persist in time so that they can be estimated from time-averaged data. If a priori information about them is available, they may be corrected, otherwise they must be treated as random errors with bias and correlated noise. Gross errors are caused by the malfunctioning of the device and by mistakes during data processing, transmission, reception or decoding, which usually affect only a very small fraction of the data. However, such errors may be very large and severely affect the downstream user. Small errors of this type are usually neither dangerous nor detectable, and can be incorporated into the estimated random errors. In the past, manual QC has been used to detect them, but the increasing data volume makes it excessively time consuming, necessitating automatic, computerized QC procedures.

Here, our attention focuses on gross errors that may have a dramatic impact on the model analysis. Analysis of the results obtained with the proposed approach indicates some interesting ability in the early detection of systematic errors. For most oceanographic observations, random errors are usually at least one order of magnitude smaller than gross errors.

Basic QC procedures usually check for errors in platform identification, date, location, value, digital encoding or stuck values, based on global criteria. Other test categories focus on the temporal and spatial consistency of data subsets. Typical tests on temperature/salinity data detect frozen values or sudden changes in time series, as well as spikes, unrealistic gradients or density inversions in vertical profiles. Possible horizontal inconsistencies are often addressed through comparison with local statistics from a climatological reference dataset, checking that a given value lies within a validity range:

$$X_{\min} \le X \le X_{\max}, \tag{1}$$

where X stands for the relevant variable and $[X_{\min}, X_{\max}]$ defines the range of valid values. A common practice defines the validity range from the climatological mean and standard deviation (std) in the neighborhood of the observation:

$$X_{\text{mean}} - N \times X_{\text{std}} \le X \le X_{\text{mean}} + N \times X_{\text{std}}, \tag{2}$$

where N is an adjustable parameter; see Gandin (1988), Boyer and Levitus (1994), Carton et al. (2000), Delcroix et al. (2005), Reverdin et al. (2007), Ingleby and Huddleston (2007), and Cabanes et al. (2013). Under such an assumption, Eq. (2) is the statistical equivalent of Eq. (1). Gandin (1988) implements this test for his meteorological assimilation but does not give details about the value assigned to N. Boyer and Levitus (1994) used such a strategy when building their *World Ocean Atlas*. They used such intervals to select the observations that enter the computation of their $5° \times 5°$ climatological mean and standard deviation; the N value is set to 3, except for coastal boxes where it may reach 5 and measurements close to strongly varying topography where it is set to 4.

Using the Simple Ocean Data Assimilation (SODA) package for assimilating temperature and salinity profiles in the Geophysical Fluid Dynamics Laboratory MOM2 model, Carton et al. (2000) also use such a strategy to select the data to be assimilated; they choose an N value of 4, which discards 10% of the total data. In their analysis of sea surface salinity in the Pacific Ocean, Delcroix et al. (2005) discard data using validity ranges based on standard deviations and N values between 3.5 and 5.

During QC, the main objective is to both maximize the detection of bad data and minimize erroneous rejection of good data (false alarms). In the following, ''good'' detections should be understood as errors that the QC method is able to detect while ''bad'' detections

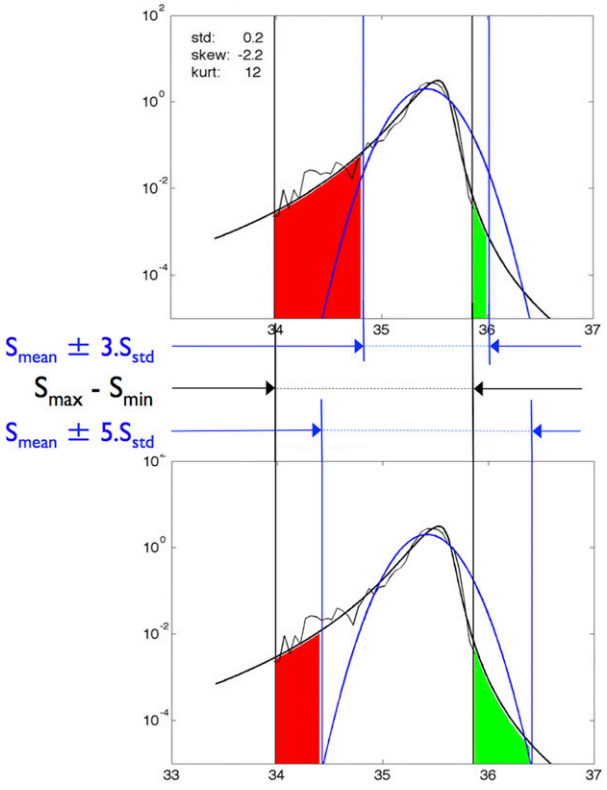

FIG. 1. Scheme describing the impact of Gaussian assumptions on the quality control of a realistic salinity distribution. The thin black curve is an example asymmetric salinity distribution. The thick black line is a skewed Student pdf model with same mean, variance, skewness, and kurtosis. The blue line is a Gaussian model with the same mean and variance and location of the validity interval boundaries with (top) $N = 3$ and (bottom) $N = 5$. The vertical black lines give the validity range based on minimum and maximum values. The vertical blue lines give the validity range based on 3 (top plot) and 5 (bottom plot) standard deviations. The red patches indicate ranges of values for which the classical approach erroneously detects good data, and the green patches correspond to ranges for which erroneous data are not detectable.

refers to good observations that are erroneously detected. Despite its wide use in QC procedures for hydrographical datasets, the approach described by Eq. (2) implicitly assumes that the data are symmetrically distributed around a single modal value. While tuning the N value may account for nonzero kurtosis, it cannot properly represent skewness. Figure 1 schematically illustrates this point. It is clear that this approach has a single degree of freedom, the N parameter, that is insufficient to achieve the two objectives: improving the number of false alarms degrades the number of good detections, and vice versa. As pointed out by Ingleby and Huddleston (2007) in their discussion, "in any QC system, there is a balance between trying to reject all 'bad' observations and retain all 'good' ones, different users might require a different balance."

Thus, it is clear that the classical approach does not account for asymmetry or skewness $\mathscr{S}$ in the local data distribution. Further, there is no a priori choice for N in Eq. (2). Whenever a constant value for N is defined, it implies constant kurtosis $\mathscr{K}$ and it assumes that the detection procedure has a constant statistical significance or performance level, which is a reasonable objective. This classical model-based approach assumes that the data distribution 1) is unimodal, 2) is symmetric, and 3) has a constant kurtosis. Therefore, this approach intrinsically lacks the flexibility to account for the probability distribution in terms of peak enhancement (or flatness) and relative amplitude of its tails. It is not possible to simultaneously optimize the number of good and bad detections.

Uncommon events may be labeled as erroneous if they are too far from the mean, that is, at a distance larger than $N \times$ std, even if they are realistic and included in the climatological dataset used to build the reference mean and standard deviation values.

In this paper, a different statistical estimator is used to define the boundaries of the validity interval in Eq. (1). The main characteristics of the method are as follows:

- Beyond *global-range* or *basin-range* QC procedures, the objective is to detect those gross errors that lie outside a *local* validity interval.
- The validity interval bounds are inferred from the minimum and maximum values found in a reference climatological dataset.
- Random measurement errors are assumed to be much smaller than the observed variability (high signal-to-noise ratio).
- Results do not depend on the probability distribution shape.
- Both good and bad detections are optimized.
- Minimum and maximum reference fields are easily updated with a posteriori verification of uncommon events that are outside the minimum–maximum interval. Such an update will always reduce the amount of rejected data (i.e., reduce—and never increase—the number of detections).
- The detection efficiency only depends on the quality and representativity of the reference minima and maxima estimates and not on the choice of some parameter value.
- The efficiency increases or decreases with low or high variability, respectively, of the parameter (better at depth than at surface; near the surface, better for parameters with weaker seasonal cycle).
- If used as a strategy to identify data for manual QC, the proposed approach should significantly reduce the operator time spent on unnecessary visual data inspection.

The paper is organized as follows. Section 2 presents the strategy to derive the minimum and maximum reference fields, the climatological datasets used as inputs and the statistical parameters of interest. Examples of the resulting fields are described, focusing on the innovation relative to the classical approach and their statistical robustness. In section 3, examples of improved detection using the new approach are presented. Section 4 presents the assessment of the detection method. Concluding remarks are in section 5.

2. Building the minimum and maximum reference fields

a. *Strategy for local extrema estimation*

The extreme values of a climatological dataset can be used to define an efficient validity interval for a given parameter. Indeed, it guarantees that an uncommon, but realistic, event, even if observed only once, will not be discarded when observed again in the future. Nevertheless, estimation of extreme values turns out to be a nontrivial challenge for in situ observations subject to errors with various origins and magnitudes. As minimum and maximum values are extremely sensitive to measurement errors, an adequate strategy must be set up: it should be manual, iterative and based on the spatial consistency of the resulting minimum and maximum fields. The following steps are adopted:

1) All prior quality flags are discarded. Depending on the dataset, some nonlocal quality checks may be applied (see section 2b).
2) Preliminary minimum and maximum values are computed for bins of longitude, latitude and pressure. As expected, these values are very noisy. For all minimum and maximum values, the associated measurement information is stored.
3) Within each geographic bin and pressure level, $T/S/\sigma$ fields are displayed and visually scrutinized. For all minimum (maximum) values judged as significantly smaller (larger) than their immediate neighbors, the corresponding $T/S/\sigma$ vertical profiles are displayed together with 1) all profiles from the same geographical box and 2) all profiles from the same platform. A decision is then made to accept these data as a realistic uncommon event or to reject (flag) them.
4) Minimum and maximum values at all pressures are recomputed. Field inspection, flag activation and minimum and maximum update are repeated iteratively until all extremes are estimated to be realistic.

For the first version of the minimum and maximum fields, several undesirable statistical artifacts appeared and are detailed herein. A specific solution was designed for each of them.

First, with a standard regular longitude–latitude grid, extrema estimates (and statistical moments) are systematically noisier with increasing latitude. This is due to the reduction of the cell surface and the resulting decrease in the number of samples. An unstructured grid, having the remarkable property of homogeneous cell surface (see https://www.discreteglobalgrids.org or Sahr 2011), was used to eliminate this problem. Hexagonal cells with a 110-km distance between two opposite vertices were selected. Since the statistical robustness of the extrema estimates depends on the data coverage, the results for this grid no longer systematically decrease with increasing latitude. The resulting detection efficiency is much more homogeneous.

Second, the vertical sampling scheme of the various instruments used in the reference databases differs significantly. As a result, the vertical data profiles may not have a value within every pressure bin so that the vertical profiles of minimum and maximum estimates may be highly discontinuous. To avoid such a discontinuity, the measured values are propagated to the missing levels from one level above and below to fill the data gap, except for those gaps due to multiple erroneous observations; a maximum of one missing level is authorized. In this case, and after the iterative QC procedure described above, the profiles are linearly interpolated to the center of the pressure bins located between two consecutive valid measurements, and the iterative estimation procedure is repeated. This procedure removes most irregularities in the vertical extrema profiles.

Third, real uncommon events are often present in the database, but probably not at all locations where they may actually occur. As a result, in some cases, horizontal discontinuities may still appear in these fields, even after iterative and manual QC. It is then reasonable to consider that such an uncommon event might be observed in a near neighbor. Optionally, assuming that the statistics are locally ergodic, a smoothed version of the reference fields is computed a posteriori by replacing each minimum (maximum) value by the smallest (largest) value for the cell itself and its immediate neighbors. In the following, all reported statistics are estimated on this basis, that is, provided on the 100-km hexagonal grid but, for each cell, actually computed from the distribution of data corresponding to that cell together with its immediate neighbors. The spatial resolution of the reference fields comes closer to 300 km rather than 100 km. Such a resolution is satisfactory for the present study at global scale and essentially based on observations from the Argo network, but should probably be revisited when focusing at interior seas, marginal seas or

continental shelves when both variability and sampling scales might be somehow different. Spatially extending these uncommon events reduces the number of false-positive detections and improves the overall detection quality (see the assessment section).

b. Datasets

1) ARGO

A snapshot of the Argo dataset was first downloaded in September 2015 (http://doi.org/10.17882/42182#42342) from the Global Data Assembly Center (GDAC) (see Argo 2014). This dataset contained more than 1.4 million profiles, from nearly 10 000 platforms. Only ascending profiles with delayed-time parameter values are used in this analysis.

While Argo provides a relatively homogeneous coverage of the global ocean, it is still sparse near continental shelves and, for lack of deployment opportunities, in the Southern Ocean. Consequently, the dataset is extended using the following datasets.

2) CLIMATOLOGICAL CTD DATASETS

The International Council for the Exploration of the Sea (ICES) also provides a high quality CTD database focused on the Northern Atlantic and Arctic Oceans (http://ices.dk/marine-data/data-portals/Pages/ocean.aspx; 13 000 profiles). Ifremer also maintains a database of all CTD acquired onboard its research vessels (http://donnees-campagnes.flotteoceanographique.fr; 7000 profiles). The Ocean Climate Laboratory (OCL; see https://www.nodc.noaa.gov/about/oceanclimate.html) updates regularly its World Ocean Database (WOD); we collected their historical dataset of CTD profiles at observed depth levels. These three important climatological CTD datasets are included. Practically, we accessed these data through the Coriolis interface. At the time that the Coriolis website was accessed, 43 000 profiles were available from the WOD.

3) CTD MOUNTED ON SEA MAMMALS

Unprecedented sampling of the Southern Ocean is provided by an observation network of CTDs mounted on sea mammals. The data are available through the Marine Mammals Exploring the Oceans Pole to Pole (MEOP) portal (http://www.meop.net; 78 000 profiles). Only data having passed delayed-time QC are retained.

c. Salinity statistical parameters

Once the datasets have been iteratively quality controlled as described above, temperature, salinity and potential density distributions are assembled over all oceanic grid cells and 20-m-thick layers, as described in section 2a. From these distributions, minimum and maximum values, as well as standard statistical moments, are determined. In this section, all of these statistical parameters are presented and intercompared to investigate the consistency of the minimum and maximum estimates. First, some statistical background is recalled to provide some basics and help understand the latter comparison. Then the spatial distributions of data at the surface and 1000 m are presented. Following this, the minimum and maximum fields are introduced and interpreted in terms of validity range; a similarity with third- and fourth-order statistical moments is presented. The robustness of the parameters is investigated through Monte Carlo simulations to characterize systematic errors due to insufficient sampling. Last, a consistent statistical model is proposed and the minimum and maximum values are interpreted in terms of equivalent percentile to illustrate the degree of accuracy of their distribution tail description.

For the sake of brevity, results are only presented for salinity S. All salinity values are given as practical salinity in the pss-78 scale and will be labeled ''psu.''

1) STATISTICAL BACKGROUND

Starting with Pearson (1895), statisticians have studied the properties of various higher-order statistics, and have discussed their utility and limitations. Visual displays (e.g., histograms) often show asymmetry and/or heavy-tailed characteristics. Skewness and kurtosis can be used to characterize these features. Skewness is a measure of lack of symmetry of the data distribution, and for a normal distribution is zero. Kurtosis is a measure of whether the data, relative to a normal distribution, accumulate near the peak and the tails (high kurtosis) or at an intermediate distance sometimes referred to as the shoulders (low kurtosis). The kurtosis for a normal distribution is 3. More precisely, kurtosis characterizes the dispersion of a random variable around its (positive or negative) standard deviation. Data distributions with high kurtosis present enhanced peakedness and heavy tails, or large interval width (maximum − minimum) values. Data distributions with low kurtosis have light tails or heavy shoulders, and the maximum–minimum range value is small.

In this context, it is interesting to recall the origin of the Student's t distribution. The t distribution arises when a normally distributed process is assessed using the sample variance rather than its true value. If the sample variance is normally distributed around the true one, the process distribution will thus depart from the true Gaussian shape, and the resulting t distribution has an increased kurtosis. Furthermore, the composition of processes with different means will impact the kurtosis, even if the processes have the same variance. As addressed by Darlington (1970), Hildebrand (1971), Moors (1986), and

Knapp (2007), a bivalued mean may lead to negative kurtosis anomalies. More generally, sample mean variability will tend to smooth out the distribution peak, reducing the kurtosis, while sample variance variability acts toward increasing the kurtosis.

For asymmetrical distributions, the skewness is nonzero and, generally, kurtosis also increases. For a bimodality case, with two modes having different peak locations and levels, the kurtosis reduction will be accompanied by a nonzero skewness. As such, skewness and kurtosis are partially correlated, something that can be written in a first-order description as follows:

$$\mathcal{K} = \mathcal{K}_{\text{Scor}} + \mathcal{K}_S(\mathcal{S}). \tag{3}$$

To help to derive a kurtosis-type parameter independent of skewness, Blest (2003) proposed to define a kurtosis adjusted by skewness estimates. Rosco et al. (2015) improved Blest's definition, and Jones et al. (2011) provided an analytical expression of the skewness dependence factor. The kurtosis is thus adjusted by the skewness parameter to help interpret the minimum- and maximum-derived parameters.

Note that statistical estimates of sample skewness and kurtosis are often not robust. Various authors have proposed more robust estimators, either from quartile, octile or more generally quantile-based estimates [for kurtosis, see Moors (1988), Kim and White (2004), and Kotz and Seier (2009); for skewness, see Bowley (1920), Hinkley (1975), Groeneveld and Meeden (1984), Mac Gillivray (1992), and Johnson et al. (1994)]. The resulting parameters are then usually found to be well correlated with sample estimates but with a significantly improved signal-to-noise ratio. In the following, the minimum and maximum parameters are combined with other statistical parameters to be interpreted in terms of such robust asymmetry and peakedness characteristics so as to then be compared with sample skewness and kurtosis.

Pearson diagram, a tool for distribution classification

In $(\mathcal{S}^2, \mathcal{K})$ space, Pearson (1905) defined distribution families. In this diagram, reference analytical laws are identified, either by points (Gauss and Rayleigh), single curves (Student, gamma, inverse gamma, and Weibull) or partially bounded domains (generalized beta and beta prime). Figure 2 displays such a diagram with colored lines corresponding to the location of the Weibull, gamma and inverse-gamma distributions. The distribution of sample skewness and kurtosis is shown with light gray contour lines. In section 2c(6), the Pearson diagram is used to identify an analytical distribution family that represents reasonably well our dataset on the basis of the $(\mathcal{S}^2, \mathcal{K})$ distribution.

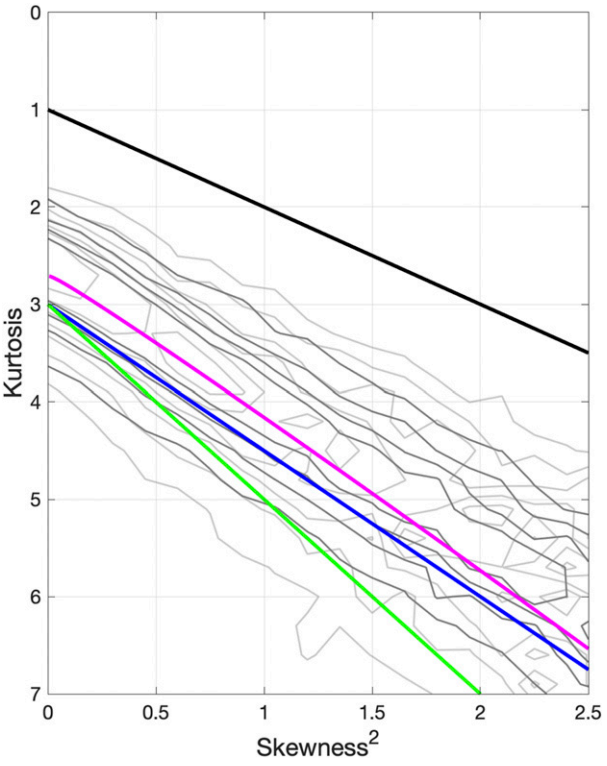

FIG. 2. Pearson diagram: The light-gray lines are contours of the normalized empirical $(\mathcal{S}^2, \mathcal{K})$ histogram. Dark-gray lines refer to the $(\mathcal{S}^2, \mathcal{K})$ parameters as estimated using Eqs. (5)–(7). The pink, blue, and green lines correspond respectively to Weibull, gamma, and inverse gamma distributions. The domain bounded by the blue and black lines corresponds to the generalized beta distribution, and the domain below the green line corresponds to the beta prime distribution.

2) NUMBER OF SAMPLES

The number of vertical profiles with observations in the surface and in the 1000–1020-m layers are shown in Fig. 3. Due to the Argo network sampling, the spatial coverage is rather uniform over the global ocean. Nevertheless, heterogeneities are present. Higher spatial density appears in the vicinity of the Kuroshio and Gulf Stream regions as they are the closing branch of the subtropical circulation; lower density occurs in areas where the platforms either have difficulty entering due to the large part of their life spent at depth (continental shelves, marginal seas) or are less deployed due to scarce ship routes crossing them (South Atlantic and Southern Oceans). The ICES dataset specifically contributes to increased density in the North Atlantic, north of 50°N. The MEOP dataset improves the overall low sampling of the Southern Ocean. The OCL dataset contribution is particularly obvious through the zonal and meridional high density lines, particularly in the central equatorial Pacific Ocean. It also has an important contribution over the continental shelves. For the 1000-m layer, the spatial

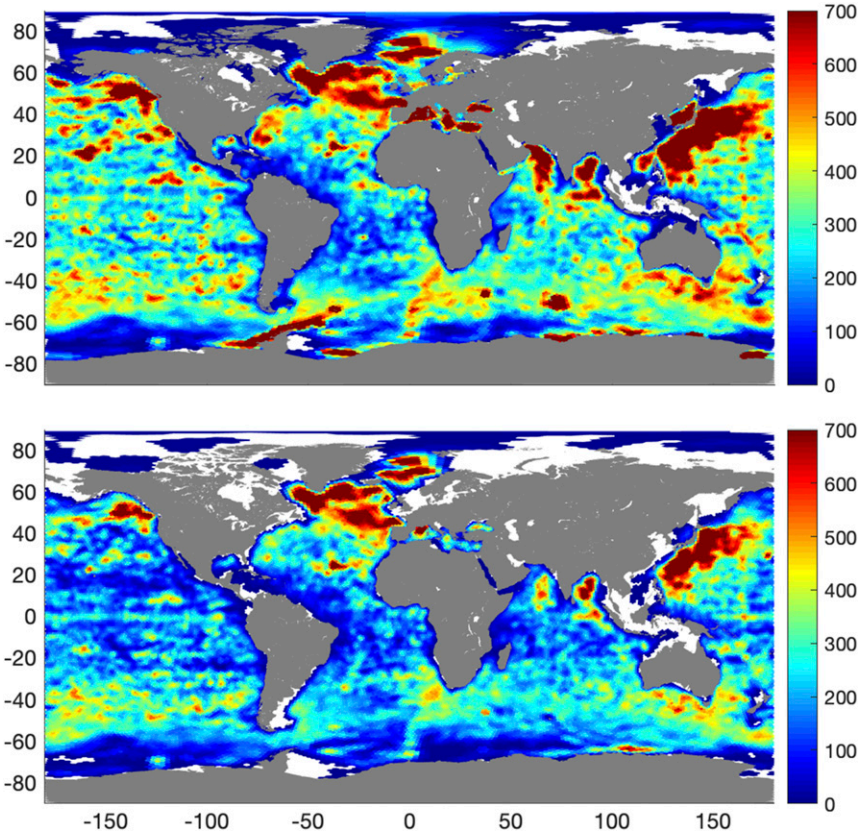

FIG. 3. Number of profiles per grid cell: (top) the 0–20-m layer and (bottom) the 1000–1020-m layer.

distribution is similar to the one at the surface; differences come from vertical profiles that do not reach such a depth, especially in the intertropical domain (shallower Argo sampling) and in the Southern Ocean (sea mammals do not reach such a depth).

As mentioned in section 2a, a particular spatial filtering is applied a posteriori. To increase the statistical robustness, all statistics provided for a given grid cell are computed using data from the cell itself as well as its immediate neighbors. From a theoretical point of view, this is equivalent to a local ergodicity assumption under which the poor description of the temporal variability is improved assuming that it can be estimated from the spatial variability in the near neighborhood.

3) MINIMA AND MAXIMA

The left and right columns of Fig. 4 display minimum and maximum salinity fields, respectively. The top and bottom panels show the surface and 1000-m layers, respectively. In the surface layer, the classical large-scale structure consisting of salinity maxima in the desertic subtropics is a feature in both minimum and maximum fields. In the 1000-m layer, the presence of outflows from evaporation basins (Mediterranean Sea, Red Sea, and Persian Gulf) is a striking feature in both field types, with the Mediterranean outflow being associated with a 2-psu difference between North Pacific and North Atlantic waters. As an example, the minimum surface salinity field displays signatures of seasonal (in an Eulerian way) freshwater inputs such as rain in the Pacific intertropical convergence zone (ITCZ), or runoffs from the Amazon, Niger, Congo or Ganges Rivers.

Changes in the structure between the minimum and maximum fields occur in zones of increased mesoscale eddy activity; the Gulf stream front is clearly displaced northward when shifting from the minimum to the maximum fields; the westward return branch of the Southern Hemisphere supergyre appears in the deep maximum field with high salinities near 40°S in the Atlantic (eddies generated in the Agulhas retroflection area) or oriented northwestward from the southwestern tip of Australia (water coming from the Tasman leakage; see Rosell-Fieschi et al. (2013).

Variability interval amplitude

Variability intervals can be used in validity range check for QC purposes. Here, the interval width obtained from the minimum and maximum fields is compared

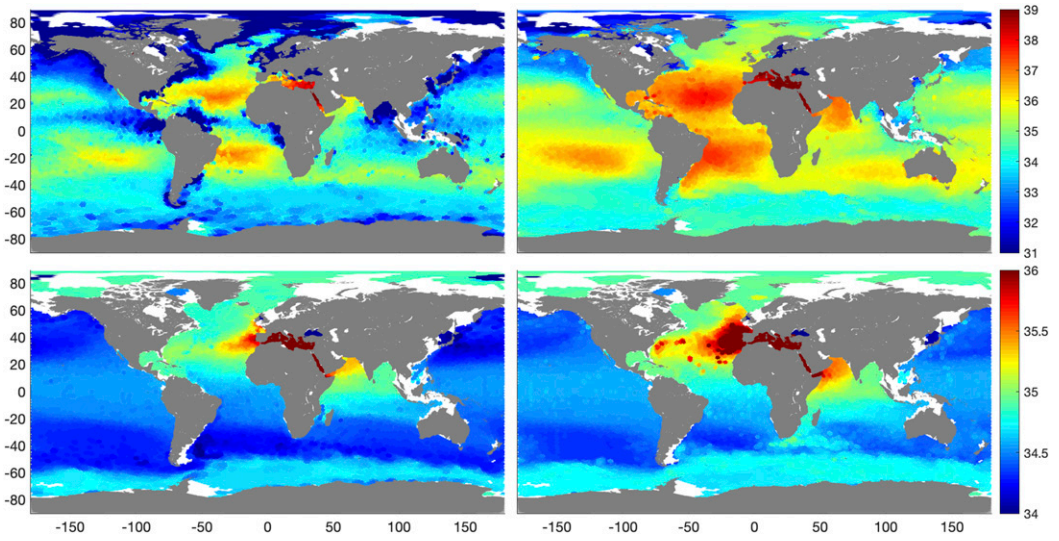

FIG. 4. (left) Minimum and (right) maximum salinity values for the (top) 0–20-m and (bottom) 1000–1020-m layers.

with the classical approach based on standard deviation estimates.

The upper panels in Fig. 5 display the amplitude of the validity interval as computed from the maximum − minimum difference (left panel) or 2*N* times the standard deviation (right panel), as defined in section 1 for the classical approach. Under Gaussian assumption and with *N* equal to 3, the "classical" interval contains 99.7% of the data. The overall similarity between both interval width estimates suggests that, at first order, the Gaussian assumption is reasonable and the interval widths compare well when *N* is equal to 3. This allows us to write

$$S_{\max} - S_{\min} \sim 6S_{\text{std}}. \tag{4}$$

Nevertheless, local differences appear at second order. In the next section, we propose to describe such differences in terms of specific statistical parameters characterizing the distribution shape.

4) INTERVAL CENTER SHIFT AND WIDTH RATIO

Rather than comparing the intervals from both approaches through their lower and upper bounds, we propose to shift to a different framework more focused on the distribution's shape. New parameters are introduced, the interval width ratio (IWR) and the normalized interval center shift (NICS):

$$\text{IWR} = \frac{S_{\max} - S_{\min}}{2 \times S_{\text{std}}} \quad \text{and} \tag{5}$$

$$\text{NICS} = 6 \times \frac{[(S_{\min} + S_{\max})/2] - S_{\text{median}}}{S_{\max} - S_{\min}}. \tag{6}$$

In Eq. (5), IWR represents the ratio of their widths; it can be considered as a robust kurtosis estimate based on quantiles (see section 1) that characterizes distribution tail height relative to height at 1 standard deviation from the distribution mean. In Eq. (6), NICS represents the difference in their center location; it vanishes when the local distribution is symmetric and is normalized so as to be interpretable as a deformation parameter; it can be considered as a robust skewness estimate based on quantiles. The factor 6 in Eq. (6) allows us to scale NICS similarly to skewness, that is, as a ratio to 1 standard deviation; see Eq. (4).

The middle panels in Fig. 5 display surface NICS (left panel) and IWR (right panel) from the quality-controlled dataset. In the surface intertropical Pacific and Atlantic Oceans, the signatures of precipitations and runoff identified in the minimum field (see section 3) are clearly present in the NICS field: relative to the mean value, the validity interval is shifted toward negative values indicating that large fresh anomalies are more likely to occur than salty ones. A striking feature is visible in the Southern Ocean, especially east of the Greenwich Meridian. The subtropical front is a boundary with intense mixing between warm and salty South Indian Central Water (SICW) and fresher and colder sub-Antarctic Surface Water (SASW). This produces an asymmetric NICS structure. On the northern side, salty SICW is dominant, increasing the mean salinity value, while meandering of the front and the presence of SASW eddies produce intermittent fresher anomalies, resulting in a negative shift of the validity interval center. On the southern side, SASW is dominant, the intermittent anomalies are saltier, inducing a positive shift in the location of the validity interval center.

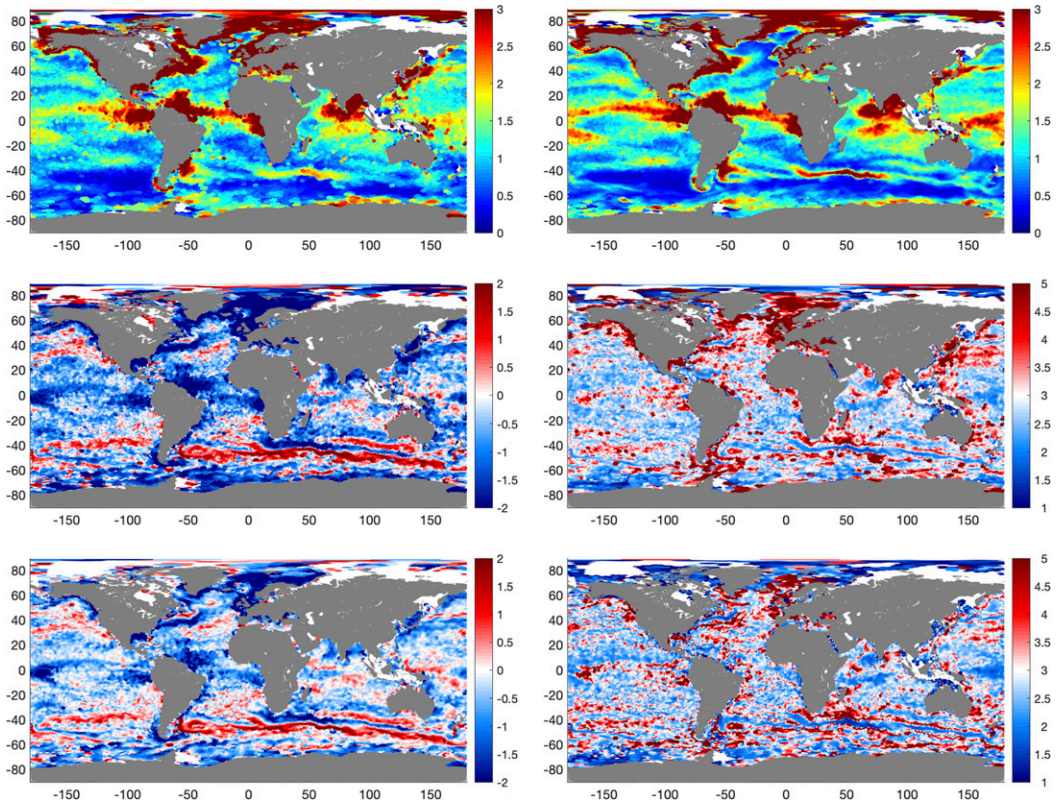

FIG. 5. Variability amplitude as estimated from (top left) $S_{max} - S_{min}$ and (top right) $6S_{std}$. (middle left) Normalized interval center shift as defined in Eq. (6). (bottom left) Skewness. (middle right) Interval width ratio as defined in Eq. (5). (bottom right) Kurtosis corrected from skewness. All plots refer to the 0–20-m layer.

The IWR field shown in the middle-right panel of Fig. 5 is the ratio between the actual validity interval width and 2 times the standard deviation, that is, the effective N value (number of standard deviations) that the classical approach should apply to avoid misestimates of the validity interval width. As already mentioned, global estimates of this ratio provide an average value of approximately 3. A systematic value larger than 3 would overestimate the validity interval width, reducing the QC efficiency by accepting erroneous data. Focusing again on the Southern Ocean east of the Greenwich Meridian, a symmetric structure is observed in the cross-front direction, that can be interpreted in terms of shape departure from the Gaussian one (for which IWR = 3). At the center, values lower than 3 are observed, indicating an excessive standard deviation value relative to the tail height (sometimes referred as ''heavy shoulders''), characteristic of a flattened or even bimodal distribution, that is a combination of processes with similar variance but different means. On both sides of the front, IWR values are larger than 3, reflecting the impact of the large NICS values described above for IWR.

NICS and IWR similarity with $\mathscr{S}$ and $\mathscr{K}$

In this section, the comparison between moments-derived and minimum- and maximum-derived shape parameters requires adjustment of kurtosis with a skewness correction, see Eq. (3). To derive such a correction, a first criterion is based on Fig. 2 and aims to align the principal axis of both ($\mathscr{S}^2$, $\mathscr{K}$) distributions (light and dark gray contour lines). A second criterion aims to match at best the color scale of middle and bottom-right panels of Figs. 5 and 6. As a good trade-off between such criteria, the following correction is proposed:

$$\mathscr{K}_S(\mathscr{S}) = 1.15\mathscr{S}^2. \tag{7}$$

For this qualitative comparison, such an ad hoc correction at first order seems reasonable, even if certainly imperfect.

Figure 5, in the bottom panels, presents the spatial distribution of skewness and kurtosis [adjusted for skewness; see Eqs. (3) and (7)] in the surface layer. Similar comparison at the 1000-m level is shown in Fig. 6. The NICS and IWR fields are very similar to skewness and modified kurtosis fields. This is an

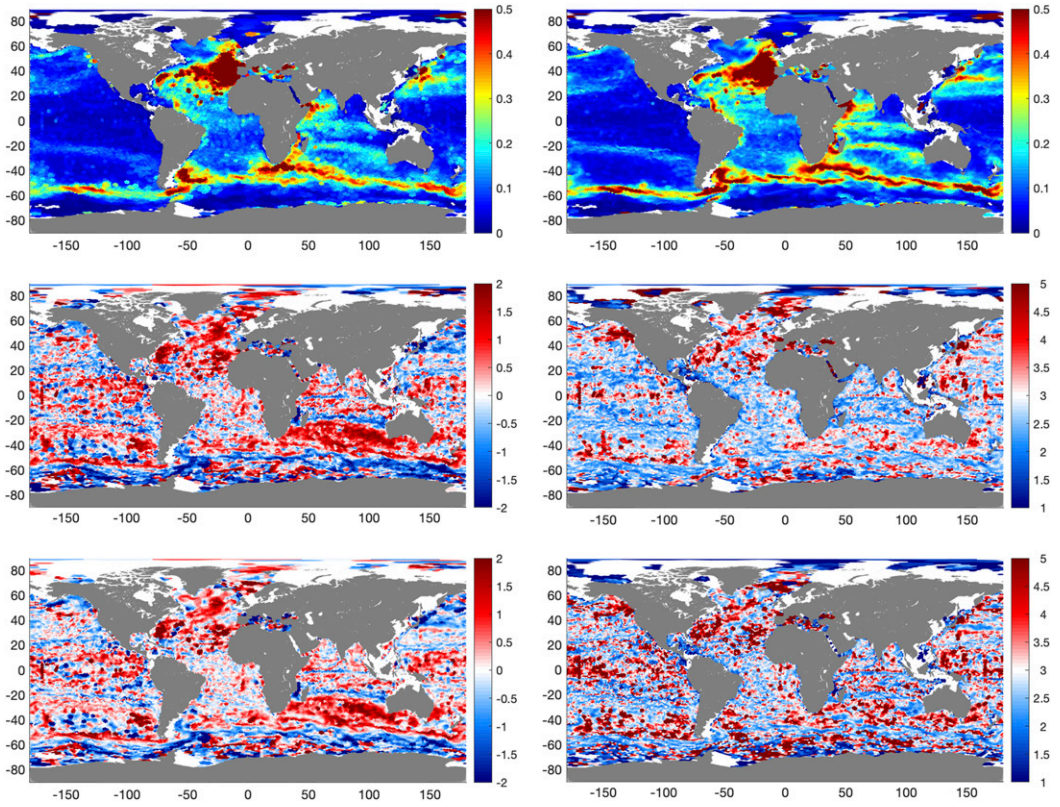

FIG. 6. As in Fig. 5, but for the 1000–1020-m layer.

expected result as NICS and IWR have definitions similar to robust quantile-based estimates of skewness (see discussion in section 1). But such a similarity between the minimum- and maximum-derived and the moments-derived estimates of asymmetry and peakedness suggests that the estimated minimum and maximum values globally reach a significant degree of robustness, even if they can still be improved locally. The middle and bottom panels of Figs. 5 and 6 illustrate the third- and fourth-order variability in the distribution shape that the present approach allows us to account for, which is an improvement relative to the classical one. This is also different from the approach by Gouretski (2018) who proposes an ad hoc solution to account for third-order statistics, while it does not address the contribution of the fourth-order ones.

5) STATISTICAL ROBUSTNESS

Even after gathering the data from one cell together with its neighbors (see section 2), spatial variations of data density are still significant, see Fig. 3. In this section, we focus on potential parameter errors associated with insufficient sampling.

Grid cells with a total number of profiles $n \geq 500$ are selected. A Monte Carlo approach is then used to examine the effects of insufficient sampling. For each selected grid cell, the full distribution is randomly split in n/p subdistributions of size p, ranging from 5 to 500. Sample parameters for the mean, variance, skewness, kurtosis, NICS and IWR are estimated for all subdistributions and normalized by the value for that cell obtained with the full distribution. For each p value, a distribution of normalized parameter values is then obtained as the total average over all the selected cells. Because $\mathscr{S}$ and NICS may take either positive or negative values, the Monte Carlo procedure and the normalization are applied to the square of these quantities, $\mathscr{S}^2$ and NICS^2, but results are expressed in terms of $\mathscr{S}$ and NICS.

Results are presented in Fig. 7. First and as expected, mean, standard deviation, skewness, and kurtosis have increasing random errors with increasing norm associated with their definition (from L^1 to L^4). Second, the spread of the relative error distributions systematically decreases with increasing number of samples.

For each p value, a systematic bias of the normalized distribution median is shown as black lines. No systematic mean bias is evident; variance is only biased by a few percent in the most extreme case; $\mathscr{S}$ and $\mathscr{K}$ appear biased below 200 samples by up to 20%–30%.

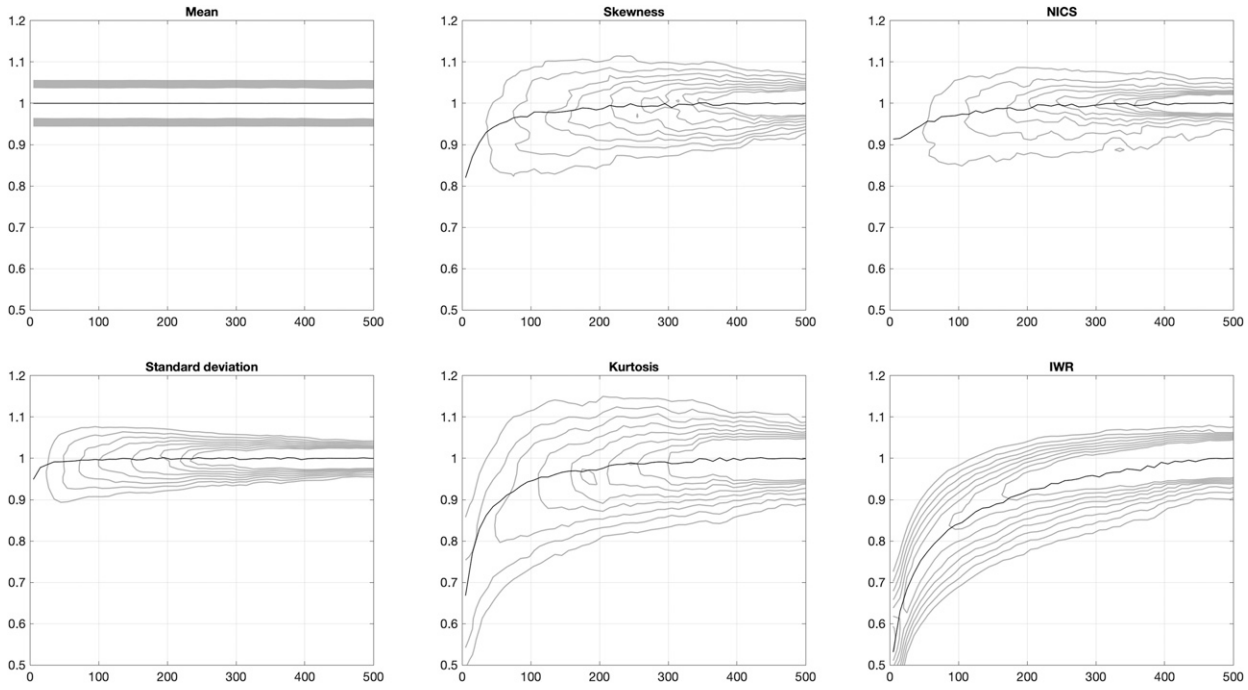

FIG. 7. Contours of the relative error distribution as a function of the number of samples. For all parameters, the contour levels are set from 15% to 50% (by steps of 5%) of the absolute maximum of the bidimensional pdf. (top left) Mean and (bottom left) variance; (center top) skewness and (center bottom) kurtosis; (right top) NICS and (right bottom) IWR. The black lines correspond to the median of the distribution for given number of samples.

For sufficiently large sampling (the right part of each panel corresponding to more than 300 samples), the minimum and maximum-derived parameters (NICS and IWR) have less spread than the sample skewness and kurtosis parameters. Minimum- and maximum-derived estimates of asymmetry and peakedness thus appear more robust than those estimated from high-order moment. With decreasing number of samples, this lower error dispersion is counterbalanced by a much larger bias for IWR. This is an intrinsic consequence of the extreme value-based approach. Minimum and maximum values can provide a more accurate estimate of the validity interval, but are directly related to the parameter distribution tails; their robustness depends on the amount of QC work (see section 2) but, more essentially, on the amount of variability sampled. Minimum and maximum values are critical samples within the distribution; in the Monte Carlo simulation, all parameter estimates from subdistributions that do not include such critical samples dramatically diverge from the true solution, leading to thinner validity intervals, that is, lower IWR estimates. Comparatively, the fourth-order moment includes information closer to the distribution peak and does reach robustness faster; the corresponding validity intervals are somehow less dependent on the distribution tails. It is interesting to note that NICS does not show a larger bias than the skewness; although NICS is estimated directly from minimum and maximum values, its definition [normalized by the maximum − minimum width rather than the std; see Eq. (6)] allows us to derive a weakly biased asymmetry parameter from minimum and maximum estimates with stronger bias.

6) MINIMUM AND MAXIMUM EQUIVALENT PERCENTILE

To further statistically characterize the minimum and maximum estimates and their robustness, we attempt to evaluate minimum and maximum equivalent percentiles; their consistency is evaluated through the expected decrease (increase) of the minimum (maximum) equivalent percentile when the empirical distribution is built from an increasing number of independent samples. With this aim, it is first necessary to select an analytical distribution law that adequately describes our empirical distributions, especially focusing on the distribution tails, which minimum and maximum values are associated with. Clearly, such a distribution should well reproduce the empirical skewness and kurtosis values closely linked to the shape of the distribution tails. As such, we use the $(\mathscr{S}^2, \mathscr{K})$ space of the Pearson diagram to identify an adequate family of distributions. Following, using the cumulative formulation of the selected distribution, the

minimum and maximum values are characterized in terms of equivalent percentiles.

(i) Pearson diagram

We propose the use of the Pearson diagram tool presented in section 1. In the ($\mathcal{S}^2$, $\mathcal{K}$) space of Fig. 2, the data distribution is presented in light gray contour lines using the sample skewness and kurtosis adjusted from the systematic biases identified in the previous section. The dark-gray contour lines correspond to the robust quantile-based estimates of $\mathcal{S}^2$ and $\mathcal{K}$, where $\mathcal{S}^2 = \text{NICS}$, $\mathcal{K} = \text{IWR} + \mathcal{K}_S(\mathcal{S})$, and $\mathcal{K}_S(\mathcal{S})$ is taken from Eq. (7). To account for the bias evidenced in section 5, the sample $\mathcal{S}$ and $\mathcal{K}$ are empirically adjusted for their low sampling bias using the median of the normalized distribution shown in Fig. 7, middle panels. Similarly, NICS and IWR are adjusted using the median of the normalized distribution shown in Fig. 7, right panels, prior to their interpretation in terms of $\mathcal{S}$ and $\mathcal{K}$.

The ($\mathcal{S}^2$, $\mathcal{K}$) distribution is fairly well spread around the curve corresponding to the Weibull law, especially for the lowest $\mathcal{S}^2$ values. The distribution of the minimum- and maximum-derived parameters has a significantly lower spread. Correspondingly, grid cells with $\mathcal{S}^2 < 1$ and $\mathcal{K}$ away from the Weibull law by less than 0.05 are selected. In the following, the two sets of ($\mathcal{S}$,$\mathcal{K}$) parameters are used to estimate minimum and maximum equivalent percentiles.

(ii) Minimum and maximum equivalent percentile

Here we use the selected grid cells for which the distribution should be reasonably approximated by a Weibull law. For each cell and its corresponding empirical distribution, Weibull parameters are adjusted to match the salinity distributions, especially to mimic the skewness and kurtosis values. From these adjusted laws, the inverse Weibull cumulative distribution is used, and minimum and maximum values are interpreted in terms of percentiles. Figure 8 shows the percentile estimates associated with the salinity minimum for cases with negative skewness in the surface layer. The percentile values are obtained both using the direct empirical ($\mathcal{S}$,$\mathcal{K}$) values and the minimum- and maximum-derived ones.

For the largest number of samples, the equivalent percentile does converge toward a value close to 0.2; such a value is in reasonable agreement with the 99.7% of data included inside a 6-std interval under the Gaussian assumption (see section 3). With decreasing number of samples, the minimum value has not fully converged to such a value and the percentile rapidly increases. The minimum estimate is less robust and

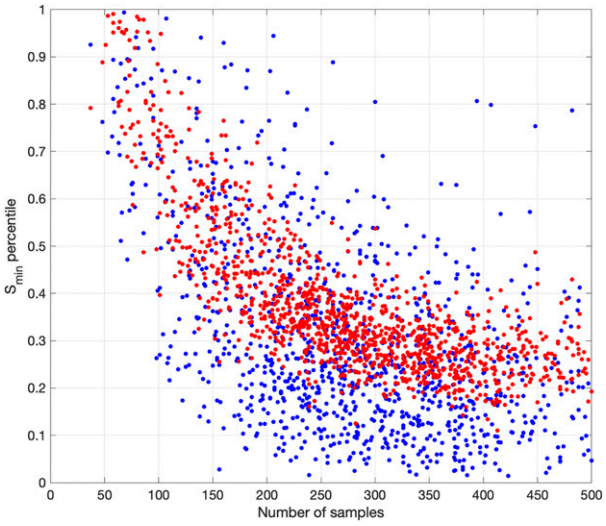

FIG. 8. Evolution of the S_{min}-equivalent percentile as a function of the number of samples, using the empirical S_{min}, $\mathcal{S}$, and $\mathcal{K}$ values (blue dots) or using the empirical S_{min} values and the estimated $\mathcal{S}$ and $\mathcal{K}$ values as defined in Eqs. (5) and (6) (red dots).

describes a location in the distribution tail progressively closer to the interior. A number of 300 to 400 samples can be considered as a threshold value to ensure a sufficient robustness. This value is in good agreement with the results obtained in section 5 for IWR.

Percentiles derived from minimum and maximum estimates have lower spread than those obtained from sample $\mathcal{S}$ and $\mathcal{K}$, suggesting again a higher robustness and quality. Indeed, minimum and maximum values are individual samples with robustness primarily depending on the performed QC work focusing specifically on the extreme values of the dataset, see section 2.

These results conclude the characterization of the minimum and maximum estimates themselves. In the following, an illustration of a practical usage of such estimates for QC purposes is presented, followed by a more academic statistical assessment.

3. Illustration of minimum and maximum usage for QC purposes

Raw noisy data from a CTD sensor mounted in the bilge of the *One Planet One Ocean* (*OPOO*) sailing boat that participated in the Barcelona World Race (BWR) were kindly provided by J. Salat and J. Salvador, from the Institute of Marine Sciences (ICM), Spanish National Research Council (CSIC). *OPOO* covered a round-the-world track starting on 31 December 2014 in Barcelona, Spain. The data are now used to illustrate the potential of the minimum and maximum estimates for QC purposes.

Figure 9 shows the trajectory of the boat and the entire salinity time series, as well as an expanded time period of particular interest here. Under high-hull-speed conditions, the boat rises above the water and air enters the sensor, resulting in negative conductivity and salinity errors. Such errors need to be filtered out from the time series.

Thus, in the middle and bottom panels of Fig. 9, the observed surface salinity time series are shown together with the local minimum and maximum validity interval, as well as the one derived from the classical approach estimated with $N = 5$ (i.e., the value used at the Coriolis data center for delayed-time quality control). From early January through late February, the minimum and maximum interval is systematically thinner than its classical equivalent, suggesting that the minimum and maximum approach has, in general, a more restrictive error detection capability; erroneous measurements have a larger probability of detection, that is, the approach identifies a larger number of good detections. Of course, such a larger relative capability would be reduced by using a lower N value in the classical approach; but the price would be an increased number of bad detections.

To further intercompare the two approaches, we now focus on a couple of salinity anomalies occurring with the boat crossing the ITCZ on 11 January and 24 March or sailing through a freshwater pool by the southern tip of South America on 7 March. For the 7 March event, the low-salinity anomaly encountered near Cape Horn similarly impacts the lower bound of the validity interval in both approaches. Nevertheless, for the upper bound, while the maximum value is not sensitive to that fresher water, the classical upper bound is symmetrically shifted up. This is because the method assumes a symmetrical data distribution while the mean value is not affected by the fresh anomaly, which likely has a low occurrence in the reference dataset. The consequence is that the ability of detecting measurement errors associated to positive errors up to 1 psu is severely degraded locally. In the second case, a similar analysis can be conducted, except that the classical upper bound of the validity interval is impacted less relatively, that is, shifted toward higher values; the fresh anomaly is partly carried by the standard deviation but also by the mean value; in this case, near 5°N, the probability of occurrence of such a fresh event is higher as the ITCZ crosses the area twice a year, during its northward and southward migration respectively in May and November. As intuited in section 1, the performance of the classical approach depends on the occurrence of uncommon events in the reference dataset, while the minimum and maximum approach is independent of such occurrence as long as it is larger than zero.

4. Statistical assessment

In the two previous sections, some examples described the added value of the minimum and maximum approach relative to the classical one. Here, we provide a statistical assessment of this added value in terms of number of good detections (GD) and bad detections (BD).

A robust validation approach should use independent datasets to derive the reference fields and to validate them using Monte Carlo experiments. We derived a set of reference minimum and maximum fields from a randomly selected fraction of the available dataset (typically 70%–90%), and use the remaining fraction (30%–10%) to assess the procedure, repeating this random split a large number of times to reach the necessary statistical confidence.

First, note that, as shown in earlier sections, the robustness of the reference minimum and maximum fields is highly sensitive to the size of the dataset used to build them, and the entire dataset presently available is still insufficient to reach full statistical robustness. Thus, the split of the available dataset into development and validation subsets degrades the suboptimal version of the minimum and maximum approach due to the degraded reference fields, which is the price of a Monte Carlo validation procedure. This degradation has a much weaker effect on the classical approach as the first- and second-order statistical moments converge much more rapidly, see Fig. 7. As a consequence, the method accuracy as obtained from the validation results should be considered as a lower bound estimate of the actual accuracy.

Second, the number of members in the Monte Carlo experiment is restricted by the high computational cost of each of them, that includes 1) building the reference minimum and maximum fields from the first data subset and 2) running the qualification method for the second subset. In the present study, we chose to compute 10 members for each of three different splitting-ratio values (70–30, 80–20, and 90–10). A posteriori, the dispersion between all 10 members has been checked, ensuring that the limited number of members should not restrict the validation conclusions.

a. Good and bad detections

To validate the procedure, the Argo dataset as provided in the global Copernicus Marine Environment Monitoring Service (CMEMS) In Situ dataset (http://marine.copernicus.eu/services-portfolio/access-to-products/?option=com_csw&view=details&product_id=INSITU_GLO_TS_REP_OBSERVATIONS_013_001_b; credit

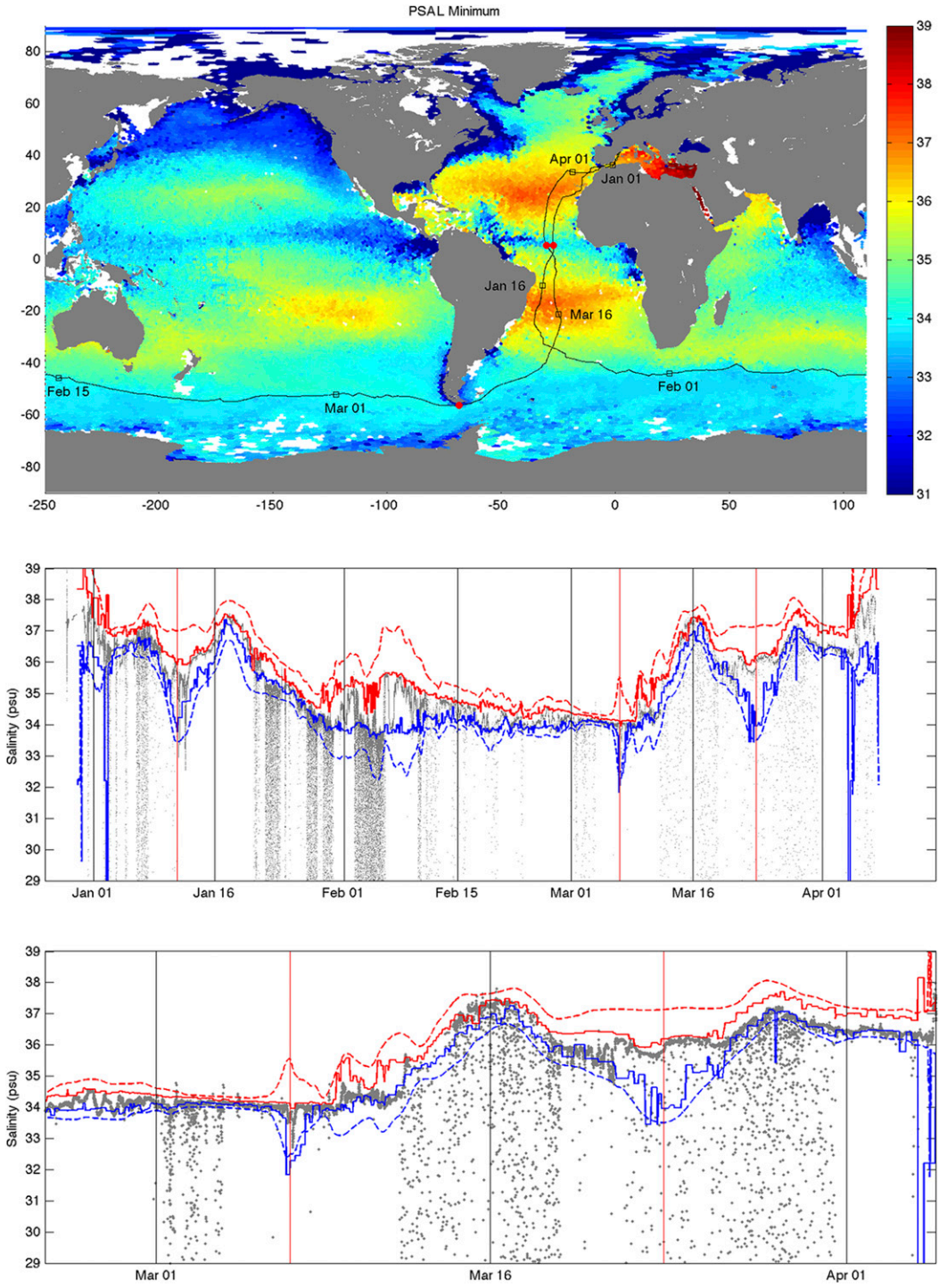

FIG. 9. (top) Sailing ship route during the BWR. The background color corresponds to the surface $S_{\min}$ estimate. (middle) Entire salinity time series during the cruise; blue or red lines indicate the lower or higher bound of the validity interval, respectively, with the full lines corresponding to the minimum and maximum approach and the dashed lines corresponding to the classical approach. Vertical black lines help to locate the beginning and middle of each month; vertical red lines correspond to the events highlighted in the text. (bottom) Same as right half of the middle panel but with a zoom onto the March–April period. The figure is provided through the courtesy of J. Salat and J. Salvador of ICM, CSIC.

to EU Copernicus Marine Service Information) is used for validation. The delayed-time QC procedures include both automatic detection and human inspection, leading to a high-quality set of flags. Assuming that the quality of such flags is perfect, they are used, in this section, to evaluate any automatic qualification procedure. The classical approach for different N values and the minimum and maximum procedure are run and the corresponding flags obtained. These flags are categorized as either GD or BD on the basis of their agreement with the CMEMS ones. Note that, keeping the naming practice of good and bad detections implicitly assumes that CMEMS flags are perfect and can be taken as a ground truth. Given that visualization and confirmation of all these detections was not manageable by the authors, this is a pragmatic assumption. Supported by the consistency of the results, the authors believe that it is also a reasonable assumption as a major part of such detections have already been visualized by the CMEMS delayed-time operator. The optimal detection method should maximize GD and minimize BD. Note that, in order to simplify the presentation and synthesis of the statistics from more than 1 million profiles with typically 100 observations each, we choose to group the observations by profile, or piece of a profile. A ''profile flag'' is activated whenever any of its corresponding individual observation flags is activated. Such a definition leads to some imperfection in the sense that an error in the surface layer may be associated with a good detection while the reference individual observation flag responsible of the profile flag activation is located in a much deeper layer. To account for this, it is preferred to compute flag statistics for different ocean layers, limiting the potential impact of the problem. The four selected layers are 0–200, 200–500, 500–1000 and 1000–2000 m.

b. Results

The left and middle columns of Fig. 10 display monthly GD and BD percent time series. The rows are for different ocean layers, see figure caption. The right column displays efficiency for the different approaches in terms of normalized relative variations of good (horizontally) and bad (vertically) detection statistics. GD reference values are defined as the classical-approach time series for $N = 4$ filtered with an 11-points rectangular window. The BD reference value is defined as the temporal average of the classical-approach time series for $N = 6$. The presented statistics are first computed as the difference to the above defined reference levels and, second, normalized so that the classical-approach results for $N = 6$ and $N = 4$ have respective coordinates (0, 0) and (1, 1) in the diagram frame. The GD statistics are independent of the method, except for the classical approach with $N = 5$ or 6 for which the GD number may be reduced by up to 20%, leading to a reduced overall quality of the dataset under such an automatic QC approach. For reference, the percent of all profiles with real-time QC (at Argo program level) is displayed in the same panel. In delayed-time QC, Argo observations may be not only flagged but also corrected to reduce offsets or biases. As such, a significant reduction of the number of errors is expected at this step. Its steady increase since 2012 is fully consistent with the fact that delayed time QC is made available within a delay varying from several months to a few years. For all approaches, the overall agreement with the GD time series, especially in the last years, suggests that, as expected, the overall quality of the dataset degrades slowly from 1% erroneous data to 2% in the last years before the present with decreasing amount of observations controlled by delayed time QC procedures, confirming the overall robustness of our GD statistics.

For the classical approach, it can be seen that the choice between lower or higher N value is a trade-off between 1) maximizing the number of detected erroneous data and 2) minimizing the number of false alarms, which, in fully automatic mode, does discard a significant amount of valid observations, or, when combined with human control, does imply a poor efficiency of the operator work. The validation results for the minimum and maximum approach show a significant impact of the dataset splitting ratio, particularly for bad detections. In the right panels of Fig. 10, the colored circles indicate that the 90/10 splitting ratio provides the best BD number reduction while the GD number is only marginally reduced. This indicates that the amount of data used in building the minimum and maximum reference fields is more important than the amount used in the assessment procedure. As such, in the following, we use only the 90/10 splitting-ratio value. The results that should be obtained with minimum and maximum reference fields built from the full dataset can be anticipated from the right panels of Fig. 10 through extrapolation of the results obtained for the 3 values of the dataset splitting ratio.

The minimum and maximum approach allows us to somehow uncouple GD and BD statistics: as shown in Fig. 11, within the 200–500-m and the 500–1000-m layers, the GD numbers are close to those from the classical approach with $N = 4$ while the BD ones correspond better to $N = 5$; see the right column of Fig. 11. The results are somehow degraded in the 0–200-m and the 1000–2000-m layers. In the 0–200 m, a lower reduction in BD is observed. Given that there are roughly the

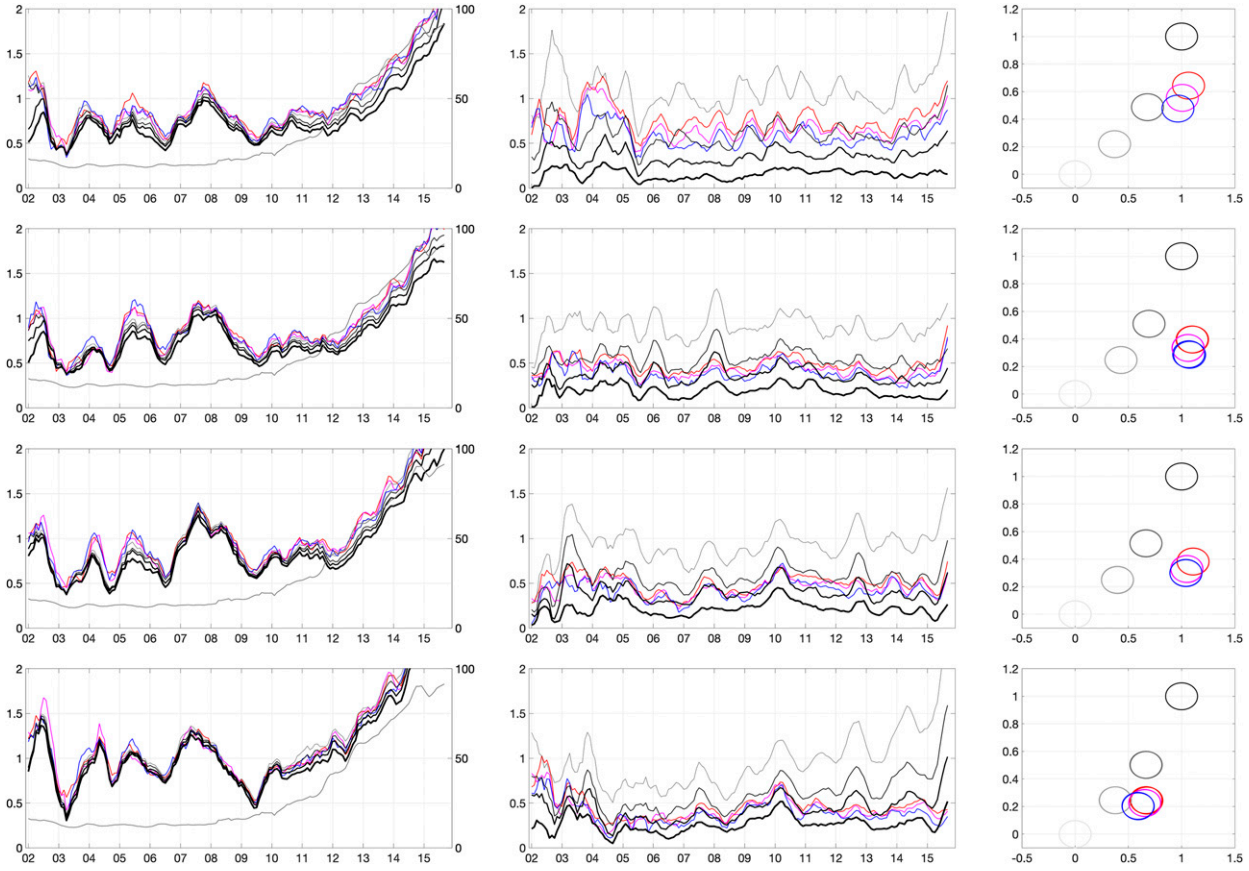

FIG. 10. (left) The left scale corresponds to the monthly percentage of good detections for a set of approaches and configurations: black lines refer to the classical approach for $N = 4, 4.5, 5$, and 6 with increasing line thickness; color curves refer to the minimum and maximum approach based on different dataset splitting ratios: 70% (red), 80% (pink), and 90% (blue); results are presented as computed for different ocean layers: (top) 0–200, (top middle) 200–500 m, (bottom middle) 500–1000 m, and (bottom) 1000–2000 m. For the right scale, the gray curve shows the Argo overall percent of profiles with real-time quality control. (center) As in the left panels, but for bad detections. (right) Efficiency diagram for the different approaches in terms of normalized relative variations of good (horizontally) and bad (vertically) detection statistics for the above-defined ocean layers. Good detection reference values are defined as the classical-approach time series for $N = 4$ filtered with a 5-point rectangular window. Bad detection reference values are defined as the temporal average of the classical-approach time series for $N = 6$. The ellipse semiaxes are scaled with the corresponding standard deviations. The presented statistics are first computed as the difference from the above-defined reference levels and second are normalized so that the classical-approach results for $N = 6$ and $N = 4$ have respective coordinates (0, 0) and (1, 1) in the diagram frame.

same number of observations in all layers, the significantly higher variability in the shallow layers implies a reduced statistical robustness of the minimum and maximum estimates.

In the 1000–2000-m layer, the natural variability is smaller; all validity intervals are narrower and more errors can be detected. As a first consequence, all GD numbers are much larger than in the upper layers (see Fig. 10, lower-left panel). In the lower-right panel, we further observe that the minimum and maximum increase of GD numbers is weaker than in upper layers. This is primarily due to the reduction of the ratio of two numbers when they both increase by the same amount. Second, at these depths where the distributions are usually more symmetrical, the QC statistics do not take advantage of an ability to account for asymmetrical distributions; the minimum and maximum benefit is reduced in comparison with upper layers where asymmetrical distributions are more likely to occur.

5. Conclusions

In the general context of automatic QC procedures for temperature and salinity observations, and beyond ''global range'' or ''basin range'' tests, this paper revisits the idea that validity intervals might be defined locally from the historical knowledge of the local variability. A classical approach estimates the validity interval from the mean and standard deviation of the historical local

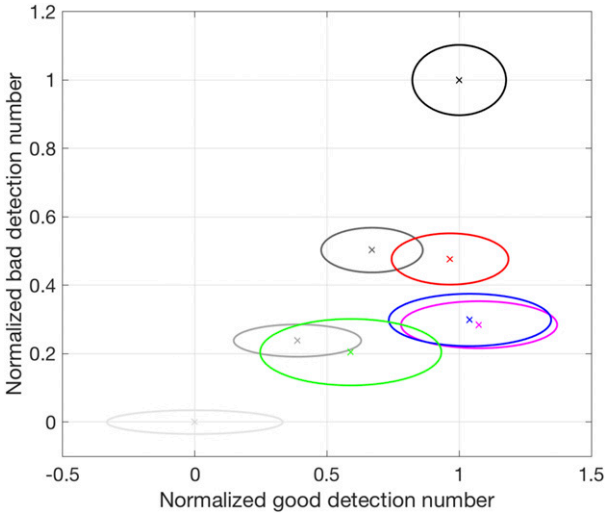

FIG. 11. As in the right panels of Fig. 10 except that all results by depth layer are synthesized in a single panel. Only results for the 90%–10% dataset splitting-ratio value are presented. Colors stand for the depth layers 0–200 m (red), 200–500 m (pink), 500–1000 m (blue), and 1000–2000 m (green).

distribution. In the present study, we propose to directly estimating the validity interval bounds as the local minimum and maximum values from the historical dataset, after dedicated manual QC work. This is a refinement of the statistical description beyond second-order statistics that allows accounting for spatial variations of the historical distribution shape, for example, asymmetry and peak enhancement, see Figs. 5 and 6. Gouretski (2018) proposed an ad hoc modification of the classical approach based on robust skewness estimates, which only accounts for third-order corrections but does not address the case of fourth-order variations of the distribution shape. The consistency and robustness of the minimum and maximum parameters is assessed through different strategies. Using quantile-based statistics, they are interpreted in terms of robust skewness and kurtosis; the comparison with sample skewness and kurtosis demonstrates the consistency of the minimum and maximum parameters. Further, Monte Carlo simulations are realized to characterize the impact of an insufficient number of samples. It is shown that minimum- and maximum-derived parameters are significantly less noisy; the interval width requires more samples to reduce its bias than other parameters, but, beyond 300–400 samples, all biases are highly reduced. Residual biases will potentially result in erroneous detections by the local range QC test. This is quantified in the assessment section.

It is demonstrated that, for a similar number of good detections, the new approach allows an important reduction of the number of bad detections. If used as an alert-raising tool combined with human QC, the number of bad detections can be seen as unnecessary use of human time so that its reduction leads to a significant saving of human resources. The success is attributed to the increased accuracy of the minimum and maximum statistical estimates in accounting for previously observed uncommon events when defining a validity interval. On the one hand, such an increased accuracy comes from the fact that a specific uncommon event introduced in the reference dataset will never raise a detection while, in the classical approach, it might be detected depending on its occurrence in the reference dataset and its weight in the estimates of the first- and second-order statistical moments. On the other hand, this increased sensitivity to rare events requires an extensive and specific manual QC step. It is also evident that, being more selective, the approach may fail more rapidly in areas where the variability is poorly sampled in the reference dataset.

Nevertheless, if the method allows significant reduction in the number of erroneous detections to be checked by the operator in delayed-mode QC, this is a good result. The number of erroneous detections is still too high for an implementation in a operational near-real-time system. This is to be related with the spatial distribution of observations, see Fig. 3, and the number of grid cells that do not reach a threshold number of samples (300–500).

It is clear that statistics will improve at updating the reference dataset with the latest observations; this will help to improve the method performance, that is, reduce the number of detections, progressively with time. Inside an operational data production system (such as the Coriolis facility), it will be pertinent that all the observations with an alert raised by this automated QC procedure but cancelled by an operator be included regularly in the reference dataset.

Future work should aim to further improve the method performance. It will require building some model of the nonsampled variability in order to artificially widen the validity interval. Such extrapolation of the empirical distribution might either be based on an adequate analytical model of the distribution, or on an ad hoc prediction from available moments and quantiles.

Acknowledgments. This study has been conducted using EU Copernicus Marine Service Information and was supported by the European Union within the EU Copernicus Marine Service In Situ phase-I and phase-II contracts led by Ifremer. The publication was also supported by SOERE CTDO2 in France. The Argo data were collected and made freely available by the International

Argo Program and the national programs that contribute to it (see http://www.argo.ucsd.edu, http://argo.jcommops.org). The Argo Program is part of the Global Ocean Observing System (http://doi.org/10.17882/42182). The marine mammal data were collected and made freely available by the International MEOP Consortium and the national programs that contribute to it (see http://www.meop.net; https://doi.org/10.17882/45461). Aleix Gelabert and Dídac Costa were the skippers of the *OPOO*, sponsored by the Intergovernmental Oceanographic Commission (UNESCO) and Pharmaton. The BWR is a periodic oceanic race organized by the Fundació Navegació Oceànica de Barcelona (FNOB). Reviewer D. Briand provided some useful comments on the final version of the draft paper before submission.

REFERENCES

Argo, 2014: Argo float data and metadata from Global Data Assembly Centre (Argo GDAC). Sea Scientific Open Data Edition, accessed 8 September 2015, http://doi.org/10.17882/42182#42264.

Blest, D. C., 2003: A new measure of kurtosis adjusted for skewness. *Aust. N. Z. J. Stat.*, **45**, 175–179, https://doi.org/10.1111/1467-842X.00273.

Bowley, A. L., 1920: *Elements of Statistics*. Vol. 2. King and Son, 289 pp.

Boyer, T. P., and S. Levitus, 1994: Quality control and processing of historical oceanographic temperature, salinity, and oxygen data. NOAA Tech. Rep. 81, 65 pp.

Cabanes, C., and Coauthors, 2013: The CORA dataset: Validation and diagnostics of in-situ ocean temperature and salinity measurements. *Ocean Sci.*, **9**, 1–18, https://doi.org/10.5194/os-9-1-2013.

Carton, J. A., G. Chepurin, X. Cao, and B. Giese, 2000: A simple ocean data assimilation analysis of the global upper ocean 1950–95. Part I: Methodology. *J. Phys. Oceanogr.*, **30**, 294–309, https://doi.org/10.1175/1520-0485(2000)030<0294:ASODAA>2.0.CO;2.

Darlington, R. B., 1970: Is kurtosis really ''peakedness?'' *Amer. Stat.*, **24**, 19–20, https://doi.org/10.2307/2681925.

Delcroix, T., M. J. McPhaden, A. Dessier, and Y. Gouriou, 2005: Time and space scales for sea surface salinity in the tropical oceans. *Deep-Sea Res. I*, **52**, 787–813, https://doi.org/10.1016/j.dsr.2004.11.012.

Gandin, L. S., 1988: Complex quality control of meteorological observations. *Mon. Wea. Rev.*, **116**, 1137–1156, https://doi.org/10.1175/1520-0493(1988)116<1137:CQCOMO>2.0.CO;2.

Gouretski, V., 2018: World Ocean Circulation Experiment—Argo Global Hydrographic Climatology. *Ocean Sci.*, **14**, 1127–1146, https://doi.org/10.5194/os-14-1127-2018.

Groeneveld, R. A., and G. Meeden, 1984: Measuring skewness and kurtosis. *Statistician*, **33**, 391–399, https://doi.org/10.2307/2987742.

Hildebrand, D. K., 1971: Kurtosis measures bimodality? *Amer. Stat.*, **25**, 42–43, https://doi.org/10.1080/00031305.1971.10477241.

Hinkley, D. V., 1975: On power transformations to symmetry. *Biometrika*, **62**, 101–111, https://doi.org/10.1093/biomet/62.1.101.

Ingleby, B., and M. Huddleston, 2007: Quality control of ocean temperature and salinity profiles—Historical and real-time data. *J. Mar. Syst.*, **65**, 158–175, https://doi.org/10.1016/j.jmarsys.2005.11.019.

Johnson, N., S. Kotz, and N. Balakrishnan, 1994: *Continuous Univariate Probability Distributions*. Vol. 1. John Wiley and Sons, 784 pp.

Jones, M., J. Rosco, and A. Pewsey, 2011: Skewness-invariant measures of kurtosis. *Amer. Stat.*, **65**, 89–95, https://doi.org/10.1198/tast.2011.10194.

Kim, T.-H., and H. White, 2004: On more robust estimation of skewness and kurtosis. *Finance Res. Lett.*, **1**, 56–73, https://doi.org/10.1016/S1544-6123(03)00003-5.

Knapp, T. R., 2007: Bimodality revisited. *J. Mod. Appl. Stat. Methods*, **6**, 8–20, https://doi.org/10.22237/jmasm/1177992120.

Kotz, S., and E. Seier, 2009: An analysis of quantile measures of kurtosis: Center and tails. *Stat. Pap.*, **50**, 553–568, https://doi.org/10.1007/S00362-007-0101-4.

Le Sommer, J., E. Chassignet, and A. Wallcraft, 2018: Ocean circulation modeling for operational oceanography: Current status and future challenges. *New Frontiers in Operational Oceanography*, E. Chassignet, A. Pascual, and J. Verron, Eds., CreateSpace, 289–306.

Mac Gillivray, H., 1992: Shape properties of the g-and-h and Johnson families. *Commun. Stat. Theory Methods*, **21**, 1233–1250, https://doi.org/10.1080/03610929208830842.

Moors, J. J. A., 1986: The meaning of kurtosis: Darlington reexamined. *Amer. Stat.*, **40**, 283–284, https://doi.org/10.2307/2684603.

——, 1988: A quantile alternative for kurtosis. *J. Roy. Stat. Soc.*, **37D**, 25–32, https://doi.org/10.2307/2348376.

Pearson, K., 1895: Contributions to the mathematical theory of evolution.—II. Skew variation in homogeneous material. *Philos. Trans. Roy. Soc. London*, **186A**, 343–414, https://doi.org/10.1098/rsta.1895.0010.

——, 1905: ''Das fehlergesetz und seine verallgemeinerungen durch fechner und Pearson.'' A rejoinder. *Biometrika*, **4**, 169–212, https://doi.org/10.2307/2331536.

Reverdin, G., E. Kestenare, C. Frankignoul, and T. Delcroix, 2007: Surface salinity in the Atlantic Ocean (30°S–50°N). *Prog. Oceanogr.*, **73**, 311–340, https://doi.org/10.1016/j.pocean.2006.11.004.

Rosco, J. F., A. Pewsey, and M. C. Jones, 2015: On Blest's measure of kurtosis adjusted for skewness. *Commun. Stat. Theory Methods*, **44**, 3628–3638, https://doi.org/10.1080/03610926.2013.771747.

Rosell-Fieschi, M., S. R. Rintoul, J. Gourrion, and J. L. Pelegrí, 2013: Tasman leakage of intermediate waters as inferred from Argo floats. *Geophys. Res. Lett.*, **40**, 5456–5460, https://doi.org/10.1002/2013GL057797.

Sahr, K., 2011: Hexagonal discrete global grid systems for geospatial computing. *Arch. Fotogrametrii Kartogr. Teledetekcji*, **22**, 363–376.