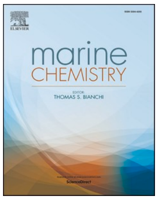
Marine Chemistry 269 (2025) 104498

Contents lists available at ScienceDirect

Marine Chemistry

journal homepage: www.elsevier.com/locate/marchem



Achieving measurement comparability in mercury speciation analysis in seawater: Key requirements and best practices

Igor Živković [a,b,*], Lars-Eric Heimbürger-Boavida [c], Mariia V. Petrova [c], Aurélie Dufour [c], Ermira Begu [a], Milena Horvat [a,b]

[a] *Jožef Stefan Institute, Jamova Cesta 39, Ljubljana, Slovenia*
[b] *Jožef Stefan International Postgraduate School, Jamova cesta 39, Ljubljana, Slovenia*
[c] *Aix Marseille Université, CNRS/INSU, Université de Toulon, IRD, Mediterranean Institute of Oceanography (MIO), Marseille, France*

ARTICLE INFO

Keywords:
Mercury speciation
Measurement uncertainty
Interlaboratory comparison
Data comparability

ABSTRACT

The comparability of measurement results is an important issue in contemporary mercury (Hg) speciation in seawater. Sampling campaigns must be properly designed to determine significant differences on spatial and temporal scales, considering two major parameters: the variability of expected data at a given sampling point/transect and variability in the results due to the intrinsic properties of specific analytical methods, particularly the measurement uncertainty. This study assessed the required sample size, considering several aspects of data variability when determining total Hg, dissolved gaseous Hg, and methylated Hg species in seawater. The required sample sizes were calculated using (1) the measurement uncertainty of a single-laboratory measurement of analytical methods used; (2) performance of the laboratories in interlaboratory comparison exercises; and (3) natural variability in Hg species/fractions in a selected case study in the Central Adriatic Sea. It was shown that the measurement uncertainty of a particular method and interlaboratory variability among laboratories have significant influence on data interpretation in case natural variability of Hg fractions is relatively small, such as for example the open seawater depth profiles. In contrary, in areas with large natural variability of Hg contractions, such as coastal and contaminated sites, their influence on data interpretation is negligible. The present paper introduces the importance of proper estimation of measurement uncertainty in international programs, such as GEOTRACES, where data comparability is of fundamental importance to assess temporal and spatial trends of Hg measurements in the marine environment.

1. Introduction

Mercury (Hg) is a toxic trace element that poses risks to humans, biota, and the environment (Basu et al., 2023; Singh et al., 2023; Wu et al., 2024). In seawater, Hg concentrations are typically very low, measured in the ng L^{-1} range in open ocean waters. However, concentrations tend to be higher in coastal areas, particularly near contaminated sites, where levels can reach several tens of ng L^{-1}. The concentrations of specific Hg fractions or species may account for less than 10 % of the total Hg content (Gworek et al., 2016). Measurements of these low concentrations and corresponding transformation rates underpin regional and global models that aim to realistically demonstrate the distribution, speciation, transport, and transformation of Hg in different environmental compartments. Hg speciation analysis in various environmental compartments is carried out by research groups for different purposes and utilizes different analytical methods. This can create problems for assessment of trends on spatial or temporal scales. Measurement results must be comparable, and harmonized approaches that include sampling design are therefore necessary to guarantee comparable results (Snoj Tratnik et al., 2019).

In the past decade, several projects (MeTra, MercOx, SI-Hg) have aimed to provide common ground for producing comparable results. New analytical procedures for the determination of Hg traces in the atmosphere and seawater have also been improved (Heimbürger et al., 2015; Kotnik et al., 2015; Quétel et al., 2016; Sprovieri et al., 2016; Torres-Rodriguez et al., 2024). The Minamata Convention on Mercury (Article 22: Effectiveness Evaluation) requires signatory parties to evaluate the effectiveness of the Convention's "establishment of arrangements for providing itself with comparable monitoring data on the presence and movement of mercury and mercury compounds in the

[*] Corresponding author at: Jožef Stefan Institute, Jamova Cesta 39, Ljubljana, Slovenia.
E-mail address: igor.zivkovic@ijs.si (I. Živković).

https://doi.org/10.1016/j.marchem.2025.104498
Received 30 September 2024; Received in revised form 10 January 2025; Accepted 3 February 2025
Available online 4 February 2025

environment" (UN, 2013; UNEP, 2021). We need precise, accurate, and comparable measurements to answer the following research questions: 1) exactly how much Hg resides in the ocean; 2) by how much we have increased environmental Hg levels; 3) what is the link between Hg emissions, environmental Hg levels, and biota Hg levels; 4) what is the time lag for decreasing environmental Hg levels following reduction in Hg emissions; 5) why are different ocean basin behaving differently depending on Hg reservoir size, source-sink balance, and trophic web? Our measurements must be able to resolve environmental variability (spatial and temporal), which is often <10 %.

Metrology involves the theoretical and practical aspects of measurement, regardless of measurement uncertainty and the field of application (BIPM et al., 2012). To achieve metrological traceability, it is essential to relate each result to a reference through a documented unbroken chain of calibrations (sequence of comparisons linking a measurement to a reference standard) that contribute to measurement uncertainty (BIPM et al., 2012). Measurement traceability is essential in Hg research to ensure accuracy and comparability of data across studies on environmental contamination. Traceability allows for the consistent calibration of analytical instruments and the reliable quantification of Hg levels in air, water, and soil (Andron et al., 2024; Kleindienst et al., 2023; Živković et al., 2017a). There are two different concepts in metrology: metrological comparability and metrological equivalence (De Bièvre, 2006). Measurement comparability is the ability to compare results from different methods, instruments, or locations within a framework of traceability to a common reference (e.g., the SI units), even if they are obtained using different methods. Measurement equivalence, on the other hand, typically refers to an agreement between results from different measurements that are statistically indistinguishable within the declared measurement uncertainty (De Bièvre, 2006). Strict equivalence is rarely demanded unless in highly controlled interlaboratory comparisons.

The National Institute of Standards and Technology (NIST) provides Standard Reference Material (SRM) 3133, a certified Hg standard that underpins measurement traceability. NIST SRM 3133 is used as a primary calibration standard for the quantitative determination of Hg (NIST, 2016). Using NIST SRM 3133 as a calibration benchmark ensures that Hg measurements are traceable to internationally recognized standards, promoting scientific integrity and regulatory compliance in environmental monitoring. A corresponding CRM for monomethyl Hg (MMHg) primary calibration does not exist. Problems arise due to a lack of appropriate CRMs for quality assurance; for example, NMIA MX014 (the National Measurement Institute of Australia's standard for trace elements in seawater) specifies an extremely large concentration (433 ± 10 ng kg^{-1}), making it inappropriate for Hg trace analysis in seawater, whereas BCR 579 (the European Union [EU] Joint Research Centre's standard for Hg in coastal seawater) certifies a much lower concentration (1.9 ± 0.5 ng kg^{-1}) but with relatively large uncertainty (26 % at the coverage factor k = 2) making it less appropriate for temporal and special trend analysis at low concertation levels typical for open seawater. ERM CA400 (the EU Joint Research Centre's standard for Hg in seawater) certifies a much higher concentration (16.4 ± 1.0 ng kg^{-1}; 95 % confidence interval: 15.4–17.4 ng kg^{-1}) than BCR 579, but with a lower relative uncertainty. According to the ISO Guide 33, the analyst should decide what CRM properties are relevant to the measurement procedure, taking into account the certification approach, the statement on intended use, and the instructions for the correct use (ISO, 2015a). Following this guide, all these CRMs for quality assurance suffer from drawbacks. Therefore, the production of a low-level low-uncertainty CRM for Hg in seawater is required. In the absence of proper reference methods and materials, the comparability of Hg measurement results in proper metrological terms is difficult to demonstrate.

Sample preservation also poses an additional challenge for data equivalence. Acidification can cause Hg species to transform, and many differences are attributable to changes in species composition (Guevara and Horvat, 2013). Acidification of seawater can cause oxidation of dissolved Hg^0 and quantitative decomposition of dimethyl Hg (DMHg) to MMHg (Black et al., 2009); therefore, Hg^0 and DMHg in seawater samples cannot be preserved using acidification. When seawater samples are acidified, methylated Hg (MeHg) is commonly reported as the sum of MMHg and DMHg (Heimbürger et al., 2015, 2010; Kleindienst et al., 2023). However, any information about the MMHg vs DMHg ratio is lost. Dissolved gaseous Hg (DGM)-purged samples have been used to measure the remaining MMHg and to calculate DMHg by a mass balance from a separate MeHg sample (Petrova et al., 2020). As DGM consists of Hg^0 and DMHg, selective purge-and-trap method has been utilized to separate these two species. During purging, DMHg can be trapped on Carbotrap, Tenax, or the equivalent traps, while Hg^0 that mostly passes through these traps can be collected on gold traps (Cutter et al., 2017). However, due to instability of these two species in acidic environment, purging must be conducted immediately upon sampling.

Interlaboratory comparison (ILC) exercises for Hg determination are seldomly performed for Hg speciation in seawater (Cossa and Courau, 1990; Lamborg et al., 2012) as they are difficult to organize primarily due to instability of the Hg species in the seawater. ILCs should be compliant with the commonly agreed and standardized protocols for the production of reference materials (Trapmann et al., 2017), and require labor intensive stability and homogeneity testing prior shipment of the samples to different laboratories (ISO, 2015b). ILCs offer opportunity to assess the comparability of the measurement results. However, they can only effectively demonstrate comparability when there is a proper assessment of measurement uncertainty and establishment of the metrological traceability (De Bièvre, 2006). In practice, this has not yet been demonstrated in any of the ILCs organized for Hg speciation in seawater. The exception is the certification campaign for BCR 579 (Kramer et al., 1998).

Proper sampling planning requires the determination of an appropriate sample size for testing whether new measurements are statistically different from previous measurements or showing that they belong to a specific population (with a pre-determined level of statistical certainty). The purpose of this study was to determine the important factors that can significantly influence the comparability of measurement results. In this paper, we enhance to sampling design by illustrating how various factors influence sample sizes needed for statistically significant discrimination between Hg concentrations in seawater. To achieve this, we calculated sample sizes based on (1) population variability, (2) measurement uncertainty within a single laboratory, and (3) interlaboratory variability based on an ILC exercise. We examined several scenarios regarding the sample size required to observe (1) the difference between mean concentrations and postulated values, (2) differences in mean concentrations between two stations, and (3) the difference in mean concentrations between two stations under conditions of low natural variability. Although our demonstration relies on data collected some time ago, the principles and concepts of metrology applied in this study remain essential and relevant. These foundational principles ensure that our findings are robust and can be adapted to current and future studies. Given the specific population variability in our case study (the Central Adriatic Sea), along with the data distribution and analytical methods used, we propose an appropriate method of data processing and offer recommendations for reducing measurement uncertainty.

2. Materials and methods

2.1. Case study: central Adriatic Sea

The Adriatic Sea is a phosphorus-limited basin located in the northernmost part of the Mediterranean Sea (Šolić et al., 2015). The principal sources of Hg in the Adriatic Sea are discharges from the Hg-rich Soča River and former chlor-alkali plants. These sources contribute to elevated Hg concentrations in seawater and sediments (Živković et al., 2017b). The data used for the statistical analyses were obtained from three sampling stations in the Central Adriatic Sea: the Bay of Kaštela (ST103; 43°31′48”N, 16°27′12″E), the Island of Hvar (CJ008; 43°12′00”N, 16°19′00″E), and the Island of Vis (CJ009; 43°00′00”N, 16°20′00″E) (Fig. 1). Samples were collected mostly on a monthly basis from March 2014 to December 2015. At every station, samples were collected from the surface to the near-bottom water layer. Details of the individual profile depths are presented elsewhere (Živković et al., 2019).

The Bay of Kaštela is a shallow, semi-enclosed bay that was heavily contaminated between 1950 and 1990 by industrial and urban waste water from the former chlor-alkali plant, causing contamination of seawater and sediments (Kwokal et al., 2002; Živković et al., 2017b). The Islands of Vis and Hvar are two inhabited islands in the Central Adriatic Sea, located away from coastal waters. These islands are influenced by the presence of the eastern South Adriatic Current on the surface, while the deeper layers are characterized mostly by the presence of Middle Adriatic deep water, North Adriatic deep water, and modified Levantine intermediate water (Artegiani et al., 1997). Since these water masses are not as susceptible to great oscillations as coastal waters, the stations were suitable for observing long-term trends in Hg concentrations.

2.2. Analytical methods and measurement uncertainties

Measurement uncertainties were estimated for the analytical methods previously used in the Central Adriatic Sea, and the data used for the estimation of measurement uncertainty were previously published (Živković et al., 2019). The focus of that paper was on Hg biogeochemistry in relation to the abundances of different marine microorganisms. However, in this paper, we do not present all previous data, but only provide summary data and use them to determine the sample size required for future sampling. These results were only used as reference data for the novel determination of the required sample sizes based on a metrological approach using both measurement uncertainty and natural sample variability. These results are a convenient dataset for this purpose as they belong to a rare long-term study in the transect from coastal to open seawater. Since these data were obtained during an almost two-year period at a transect between coastal and open seas, they covered large natural variability. This variability was used in this paper only to demonstrate spatial/time trends and differences at high resolution.

Details on the clean sampling methods and analytical procedures used for the determination of Hg fractions in seawater are provided elsewhere (Živković et al., 2019). In short, THg in seawater was determined using cold vapor atomic fluorescence spectrometry (CVAFS—a double amalgamation system) following the United States (US) EPA method 1631 (US EPA, 2002). DGM was determined by purging a nonacidified seawater sample onto gold traps (double amalgamation), followed by thermal desorption and CVAFS detection (Kotnik et al., 2017). MeHg in samples was determined either by a direct method based on derivatization by hydride generation (Živković et al., 2017a) or by a conventional method based on solvent extraction (Horvat et al., 1993) (in some samples from coastal stations close to the former chlor-alkali plant where inorganic Hg concentrations were elevated).

The measurement uncertainty for the determination of THg and DGM in seawater was estimated using the ISO-GUM/Eurachem approach (BIPM et al., 2008; Ellison and Williams, 2012). Following these guidelines, the concentrations of THg and Hg fractions and the corresponding uncertainties in seawater were determined using mathematical models to calculate concentrations based on analytical signals. Additional uncertainty sources included uncertainty due to the recovery and reproducibility of the measurements. Relative combined standard uncertainty ($u_{r,c}$) was calculated by following the general relationship between the combined standard uncertainty and the uncertainty of the individual parameters for the mathematical model expressed in a product (quotient) form (Ellison and Williams, 2012). The expanded relative combined standard uncertainties ($U_{r,c}$), calculated using the coverage factor k = 2, were determined at various concentration levels. The $U_{r,c}$ for the determination of MeHg in seawater was previously published (Živković et al., 2017a) and these values were also included in the results.

The effective degrees of freedom (ν_{eff}) of the combined measurement uncertainty (u_c) were calculated using the Welch-Satterthwaite formula (BIPM et al., 2008):

$$\nu_{\text{eff}} = \frac{u_c^4}{\sum_{i=1}^{N} \frac{u_i^4}{\nu_i}} \tag{1}$$

where u_i is the uncertainty of each individual contribution and ν_i is its respective degree of freedom.

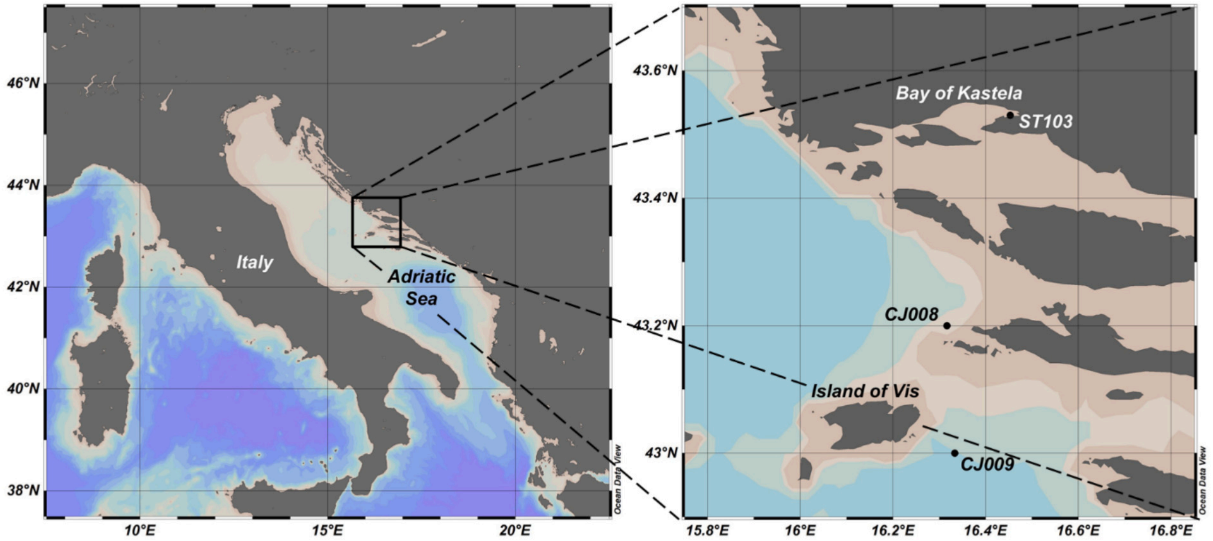


Fig. 1. Sampling stations in the Central Adriatic Sea are indicated with dots and names. Samples were collected mostly on a monthly basis from March 2014 to December 2015. Samples from each station were collected from the surface to the near-bottom water layer. Sampling locations were mapped using Ocean Data View software (Schlitzer, 2024).

2.3. Interlaboratory comparison

2.3.1. The 2014 GEOTRACES international ILC exercise for THg and MeHg

The 2014 GEOTRACES Hg interlaboratory comparison exercise included 10 participating laboratories. The exercise was organized onboard RV *Pourquoi Pas?* on the French-led GEOTRACES GEOVIDE cruise in the North Atlantic. Seawater samples were taken using a trace-clean rosette system (Measures et al., 2008) equipped with 24 × 12 L GOFLO (General Oceanics) bottles. The intercomparison sample for the determination of THg and MeHg was obtained on June 22, 2014, in the Labrador Sea (55°50′31.2”N, 48°5′34.8”W) at a 2365 m depth. The samples were poured directly into individual acid-cleaned 2 L FEP Teflon bottles without filtration, acidified to 0.4 % (*v*/*v*) with double-distilled HCl, and labeled 1–10. All samples were stored in a single box under dark, cold conditions until shipping.

2.3.2. The 2017 GEOTRACES ILC cruise for Hg species in seawater

ILC exercises can only address analytical biases for preserved Hg species, such as THg and MeHg. Some Hg species, especially gaseous species, cannot be preserved over time; therefore, the 2017 GEOTRACES Hg ILC cruise was organized to include all Hg species and procedures, from sampling to analysis. The 2017 GEOTRACES ILC cruise took place in the Northwestern Mediterranean Sea. The Mediterranean Sea was chosen because it is one of the best studied bodies of water regarding Hg speciation (Cossa et al., 2022, 2009, 1997; Heimbürger et al., 2010; Horvat et al., 2003; Kotnik et al., 2017). Thirteen laboratories participated, joined the cruise, and shipped their analytic equipment to the Mediterranean Institute of Oceanography (MIO), Marseille, France. At least one person from each participating scientific group was onboard for the sampling, while the others stayed in the laboratory for the calibration of equipment and sample analyses. Six daily cruises were organized onboard RV *Antédon II* for seawater sampling. Three selected stations covered a coastal–open ocean/shallow–deep gradient: Julio (43°06′N, 5°15′E, 100 m depth), Cassidaigne Canyon K1 (43°06′N, 5°29′E, 500 m depth), and Deep Sea K2 (42°59′N, 5°25′E, > 1000 m depth). During this exercise, THg, MeHg, MMHg, DMHg, and DGM were analyzed. Seawater samples were taken using 6 × 12 L GOFLO sampling bottles mounted on a CTD carousel frame, covered with metal-free epoxy paint. Samples were brought back to the MIO laboratory daily and analyzed using each group's equipment within a couple of hours of sampling.

Although all Hg species were analyzed during the 2017 ILC, only DGM data from this exercise were used for sample size determination based on interlaboratory variability. We opted for this approach to simulate the analytical conditions present in the case study, which measured DGM during field sampling, while THg and MeHg were determined in the home laboratory.

2.3.3. Data analysis and calculations

All data were screened for outliers before calculating the means. MeHg data with unrealistic percentages for the overall mean THg concentrations were discarded as outliers. The modified *Z*-score method was selected for outlier removal because it is robust for both small and large sample sizes. Data points with modified Z-scores greater than 3.5 were labeled outliers (Filliben, 2012) and excluded from the calculation of the means. The values from laboratories that reported repeated measurements were averaged so that the overall mean and its standard deviation (SD_{inter}) reflected variations between individual laboratories, not measurements. Due to the small number of DGM measurements obtained during the 2017 ILC exercise, the interlaboratory variability of DGM was estimated as a variation under intermediate precision conditions between pair-wise laboratories according to the Eurachem approach (Ellison and Williams, 2012).

2.4. Statistical methods

Descriptive statistics (arithmetic mean [$\overline{x}$], standard deviation [SD], geometric mean [GM], geometric standard deviation [GSD], and 95 % confidence interval [95 % CI]) were obtained to provide insight into the variability of THg, DGM, and MeHg concentrations in Central Adriatic seawater. The normality of the data was tested using the Shapiro-Wilk test (SigmaPlot 14; Systat Software, Erkrath, Germany). As the original (linear) data were not normally distributed, they were transformed using the natural logarithm function (ln function). All further statistical analyses were performed on ln-transformed data, and the results were returned to a linear scale using appropriate transformations (Bland and Altman, 1996a).

GM was calculated as an inverse ln function of the mean of ln-transformed data, while GSD was calculated as an inverse ln function of the SD of ln-transformed data; that is, as antiln(mean(ln(x_i)) and antiln(SD(ln(x_i)), respectively (with x_i representing an individual untransformed datum). Lower and upper boundaries of the 95 % CI were calculated as GM / GSD^2 and GM × GSD^2, respectively (Carobbi, 2010; Gao and Martos, 2019).

The sample sizes (n) required to detect significant differences between the mean concentrations of Hg fractions were calculated using STATA 12 software (StataCorp, College Station, Texas, US). Means of ln-transformed data ($\overline{x}_{ln}$) and corresponding sample standard deviations ($SD_{s\text{-}ln}$) were used for sample size determination. A one-sample *t*-test (comparison of mean to hypothesized value) was used to calculate the sample size by comparing the sample mean with the postulated mean. Postulated means were set at 105–130 % of GM (i.e., a difference of 5–30 %) and converted to ln scale prior to sample size determination. We used a one-sided significance level ($\alpha = 0.05$); statistical powers of 0.80, 0.90, and 0.95; and equal predicted sample size ratios (Snoj Tratnik et al., 2019).

Besides sample variability, we used the measurement uncertainty (u) of Hg determinations performed at the JSI laboratory and standard deviations from the interlaboratory comparison exercise (SD_{inter}) to calculate sample sizes. Uncertainty u and SD_{inter} had to be transformed to ln scale so that the units/scale matched those for $SD_{s\text{-}ln}$. Since they were calculated from linear data (normal distribution conserved at linear scale), they were converted to their respective ln-transformed values, u_{ln} and $SD_{inter\text{-}ln}$. An example of the calculation of u_{ln} is shown as Eq. 2:

$$u_{ln} = \sqrt{ln\left(\frac{u^2}{\overline{x}^2} + 1\right)}, \tag{2}$$

The corresponding mean value on the linear scale ($\overline{x}$), for which the uncertainty/SD was estimated, was transformed to the ln scale using Eq. 3 (Carobbi, 2010; Higgins et al., 2008):

$$\overline{x}_{ln} = ln(\overline{x}) - \frac{1}{2} {*} ln\left(\frac{u^2}{\overline{x}^2} + 1\right). \tag{3}$$

The combined variability on the ln scale ($SD_{c1\text{-}ln}$) was calculated as:

$$SD_{c1-ln} = \sqrt{SD^2_{s-ln} + u^2_{ln}}, \tag{4}$$

where u_{ln} is the ln-transformed measurement uncertainty (variability within the JSI laboratory) and $SD_{s\text{-}ln}$ is the standard deviation of the sample. The combined variability on the ln scale, including interlaboratory comparison ($SD_{c2\text{-}ln}$), was calculated as:

$$SD_{c2-ln} = \sqrt{SD^2_{s-ln} + u^2_{ln} + SD^2_{inter-ln}}, \tag{5}$$

where $SD_{inter\text{-}ln}$ is the ln-transformed standard deviation from the interlaboratory comparison exercise (interlaboratory variability).

Two-sample *t*-tests were used to calculate the sample sizes by assessing the differences between two sample means (two stations). The means of the ln-transformed data ($\overline{x}_{ln}$) and corresponding sample standard deviations ($SD_{s\text{-}ln}$) were used for sample size determination. We used a two-sided significance level ($\alpha = 0.05$); statistical powers of 0.80, 0.90, and 0.95; and equal predicted sample size ratios.

The presence of significant differences between sampling stations was tested by applying the Student's t-test (SigmaPlot 14 software) to the ln-transformed data because variances were similar between groups. In contrast, the Welch t-test was used to test whether average sample (spike) recovery significantly differed from the reference value, since the Welch t-test considers differences in sample variances, while the Student's t-test assumes equal variances (Bland and Altman, 1996b).

3. Results and discussion

3.1. Measurement uncertainties for THg, DGM, and MeHg in seawater

We estimated measurement uncertainties for the analytical methods commonly used in the JSI laboratory to determine THg, DGM, and MeHg (Table 1). As the probability distribution was approximately normal and the effective degrees of freedom were sufficient, we assumed that taking the coverage factor k = 2 would provide an approximately 95 % level of confidence for the expanded standard uncertainty (BIPM et al., 2008).

We observed a common trend for all expanded uncertainties: a decrease in the relative standard uncertainty with increasing corresponding concentration level (Table 1). This was due to the increasing difficulty in accurately measuring Hg concentrations as they approached the respective detection limit. Therefore, repeatability and reproducibility were usually the main contributors to the overall uncertainties at the lowest concentration levels. Several adjustments to analytical procedures (i.e., the use of a narrower calibration curve) could result in considerably reduced uncertainty. The details of these modifications are discussed in Section 3.6.

3.2. Case study: presentation and variability of data

Linear values for Hg concentrations in the environment rarely follow normal distribution, except when considering only off-shore data. Hence, the appropriate data transformation might be applied to linear values prior to performing a statistical analysis. The most commonly used transformation is natural logarithmic transformation (ln transformation) (Bland and Altman, 1996a). Fig. 2 shows a histogram of THg concentrations at station CJ009; the left panel presents linear data, while the right panel presents ln-transformed data. Linear data do not follow normal distribution, but ln-transformed data do (Shapiro–Wilk test); therefore, in most cases, Hg statistics should be performed on ln-transformed data.

Table 1
The JSI laboratory expanded relative combined standard uncertainties ($U_{r,c}$) and effective degrees of freedom (d.f.) for the methods of determining THg, DGM, and MeHg in seawater.

Fraction	Concentration level (ng L^{-1})	$U_{r,c}$ (%) (k = 2)	Effective d.f.
THg	0.20–0.30	23.6 (13.0*)	16.3 (35.1*)
	0.40–0.50	16.8 (10.1*)	18.5 (35.7*)
	0.60–0.80	12.0 (8.45*)	25.6 (45.3*)
	1.00–2.00	9.56	34.1
DGM	0.02–0.04	21.9	22.9
	0.08–0.10	15.7	63.1
	0.15–0.30	13.2	139
MeHg (hydride generation) **	< 0.01	21.3	32.2
	0.02–0.03	15.0	29.1
	> 0.08	11.1	26.4
MeHg (ethylation) **	< 0.01	19.3	34.2
	0.02–0.03	18.2	32.4
	> 0.08	15.8	28.1

* Calculation based on a narrow calibration curve (0.10–1.00 ng L^{-1})—see Section 3.6.1.

** Reference: Živković et al. (2017a).

The most commonly used statistics (e.g., *t*-test and analysis of variance) are simple and straightforward when using ln-transformed data, but problems occur when results need to be returned to a linear scale using appropriate transformations to determine the correct interpretation of the results (e.g., $\overline{x}_{ln}$, $SD_{s\text{-}ln}$, and the corresponding 95 % CI_{ln}). The inverse ln function of $\overline{x}_{ln}$ gives GM, whereas $SD_{s\text{-}ln}$ is converted to GSD. The 95 % CI_{ln} are symmetrical around $\overline{x}_{ln}$, but the corresponding 95 % CIs on a linear scale are asymmetric due to the non-linearity of the ln function (Fig. 3) (Carobbi, 2010). Despite knowing that most Hg data in seawater are probably ln-normally distributed, the results are usually presented as the arithmetic means of linear data ($\overline{x}$) and their corresponding SDs. GM is rarely presented, and the corresponding 95 % CI is almost never reported. To avoid inappropriate use of means and SDs that may result in reporting incorrect results (SD much larger than mean value), it is important to first test whether experimental data follow normal distribution using an appropriate statistical test (e.g., the Shapiro–Wilk or Kolmogorov–Smirnov test).

We calculated the sample sizes required to detect significant differences between the mean concentrations of Hg fractions based on the sample variability of THg, MeHg, and DGM concentrations previously determined in the Central Adriatic Sea (case study) (Živković et al., 2019). Three stations illustrate differences between coastal and open sea, and between two open water stations. The descriptive statistics for THg, DGM, and MeHg concentrations in seawater in the Central Adriatic Sea are presented in Table 2. Means and SDs are provided only for assessment of differences with geometric means. The concentration variability in Hg fractions was greatest for THg, since the geometric mean at the ST103 coastal station was fivefold greater than those at open water stations, due to this coastal station being strongly influenced by industrial effluents from the former chlor-alkali plant having been released directly into the sea. By contrast, THg levels did not differ significantly between the two open water stations ($p > 0.05$). Differences between coastal and open-water stations were observed for both DGM and MeHg, but the concentration variability was not as pronounced as in the case of THg (Table 2). For the sample size calculations presented in this paper, we considered the concentrations of Hg fractions at the open water station (CJ009). Since the data were log-normally distributed, geometric means and corresponding variances were considered for the ln-scale sample size calculations.

3.3. Interlaboratory comparisons: THg, MeHg, and DGM in seawater

During the 2014 ILC exercise, unfiltered THg concentrations varied between 102 and 185 pg L^{-1} with a mean value of 139 pg L^{-1} and a standard deviation of 28.9 pg L^{-1} ($n = 9$). One laboratory could not analyze the sample in time and withdrew from the exercise. One value had to be excluded as an outlier. The results of the 2014 ILC exercise showed good agreement between laboratories (RSD = 20.7 %). Of the 10 participating laboratories, 8 submitted results for MeHg. None of the measurements was below the detection limit, although one laboratory reported that the measurement failed. Unfiltered MeHg concentrations varied between 26.1 and 36.1 pg L^{-1} with a mean value of 29.8 pg L^{-1} and a standard deviation of 3.58 pg L^{-1} ($n = 8$). The results showed excellent agreement (RSD = 12.0 %) between the participating laboratories.

During the 2017 ILC cruise for Hg species in seawater, only three participating laboratories provided results for DGM, making it difficult to estimate variability using an interlaboratory comparison. Further difficulties arose from considerable differences in the methods applied. One approach was based on the utilization of two sorption media in series to discriminate between DMHg and Hg^0 (Lamborg et al., 2012), while another approach was based on the more commonly used

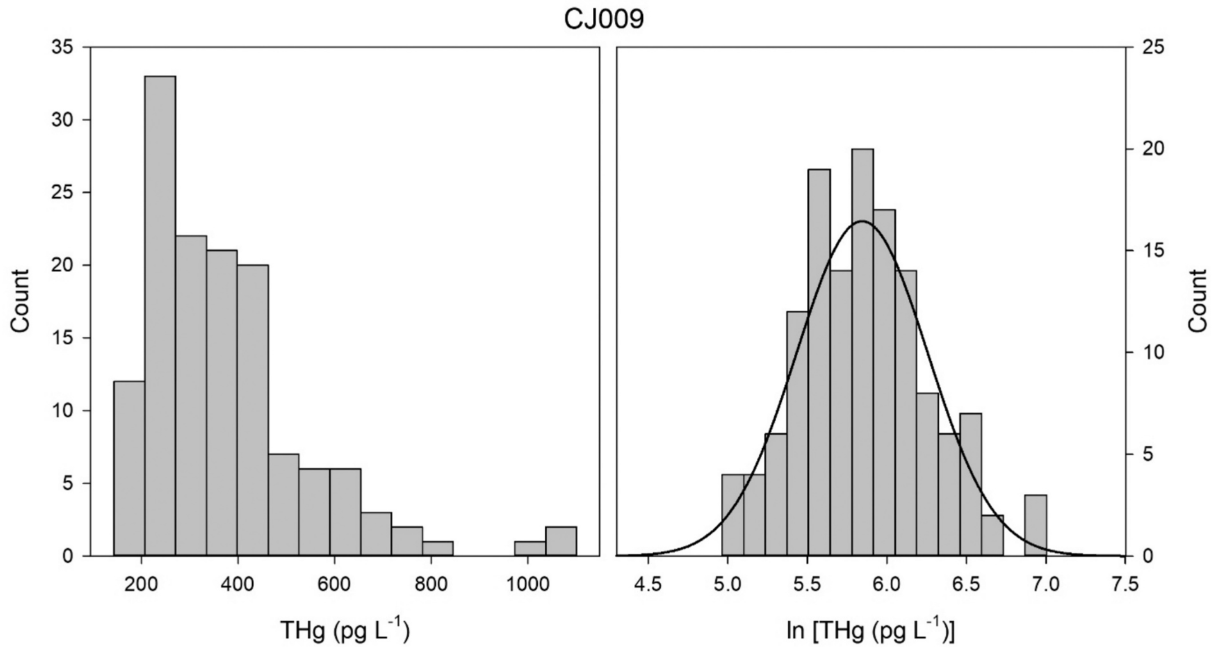


Fig. 2. Histograms of THg concentrations at Station CJ009. The left panel presents linear THg data that are not normally distributed, while the right panel presents normally distributed ln-transformed data (with an indicated theoretical normal distribution curve).

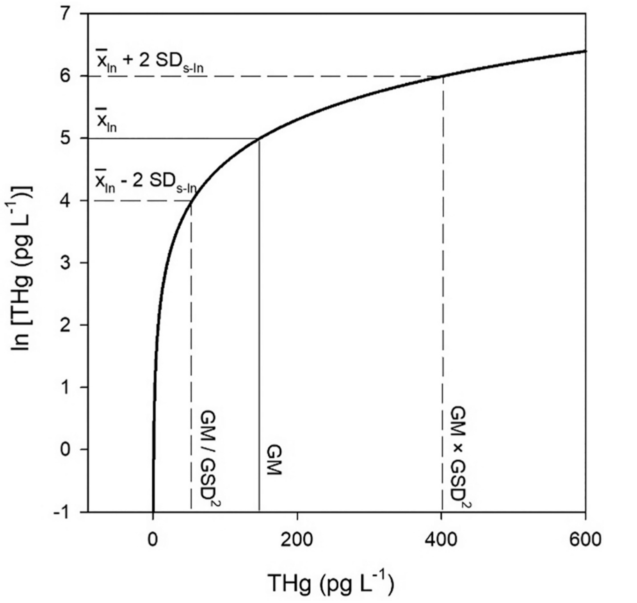


Fig. 3. An example of a conversion of THg concentrations from ln to linear scale. The curved line represents the conversion function from an ln to a linear scale (antiln function). Dashed lines indicate 95 % confidence intervals on respective axes ($\overline{x}_{ln}$, mean of the ln-transformed data; SD_{ln}, standard deviation of the ln-transformed data; GM, geometric mean; GSD, geometric standard deviation).

amalgamation of all volatile Hg species on a gold trap. Furthermore, some laboratories measured DGM in filtered water samples while others measured DGM in unfiltered water samples (or in both). Consequently, we could only use a small amount of data (only data obtained under the same conditions) to estimate variability during the ILC exercise. Based on the intermediate precision approach described in the Materials and Methods section, we estimated the variability in DGM concentrations during the ILC exercise to be 20.3 %.

Table 2
Variability of THg, MeHg, and DGM in seawater (pg L^{-1}) from three sampling stations in the Central Adriatic Sea (n, sample size; Min–Max, range; Mean, arithmetic mean; SD, standard deviation; GM, geometric mean; GSD, geometric standard deviation; 95 % CI, 95 % confidence intervals). Data from each station included all samples collected from the surface to the near-bottom water layer.

Analyte	Station	n	Min–Max	Mean (SD)	GM (GSD[a])	95 % CI
THg	All	323	108–5575	725 (851)	492 (2.20)	102–2377
	ST103	68	693–5575	2025 (1101)	1783 (1.65)	655–4850
	CJ008	119	108–834	379 (141)	354 (1.46)	165–759
	CJ009	136	143–1101	377 (174)	345 (1.51)	151–787
DGM	All	323	19.9–606	116 (90.0)	89.1 (2.10)	20.1–394
	ST103	68	31.8–606	175 (128)	142 (1.92)	38.3–523
	CJ008	119	21.6–393	113 (75.6)	91.7 (1.94)	24.2–347
	CJ009	136	19.9–245	88.9 (60.6)	68.9 (2.09)	15.7–302
MeHg	All	323	1.28–34.3	10.1 (5.00)	8.90 (1.66)	3.21–24.7
	ST103	68	2.21–34.3	13.6 (7.04)	11.8 (1.75)	3.82–36.3
	CJ008	119	1.28–20.7	8.32 (3.74)	7.41 (1.67)	2.65–20.7
	CJ009	136	2.56–22.2	9.80 (3.72)	9.09 (1.50)	4.06–20.4

[a] Dimensionless value.

3.4. Sample size determination (one station)

An example of the sample size determination is given for Hg fractions at Station CJ009. We determined the sample size required to observe significant differences in the concentrations of Hg fractions in seawater based on the observed variability ($SD_{s\text{-}ln}$) within the data at CJ009. We presumed that all samples were measured in one laboratory using the CVAFS. To observe a 20 % difference in THg GM between the two sample groups, a minimum of 56 samples were required at a statistical power of 0.95 and a significance level of 0.05. Smaller differences in THg GM required a greater number of samples (e.g., 203 samples) to observe a 10 % difference, while 773 samples were needed to observe a 5 % difference (at $\alpha = 0.05$ and a statistical power of 0.95). A smaller sample size was required to detect significant differences in the corresponding THg GM at lower statistical power (Fig. 4).

We also determined the sample sizes for DGM and MeHg; for example, the sample variability of DGM in seawater at Station CJ009 required a sample size of 178 to observe a 20 % difference in DGM GM between groups (at $\alpha = 0.05$ and a statistical power of 0.95), assuming all DGM determinations were performed at the JSI laboratory. Under the same statistical conditions, 651 samples were needed to observe a 10 % difference, while 2483 samples were needed to observe a 5 % difference (Fig. 4). Similarly, we determined that 53 samples were required to observe a 20 % difference in MeHg GM between groups at Station CJ009, 194 samples for a 10 % difference, and 739 samples for a 5 % difference (at $\alpha = 0.05$ and a statistical power of 0.95) (Fig. 4).

The relationship between the sample size and the difference (Δ) in GM values followed an inverse second-order function; therefore, the required sample sizes could easily be determined from postulated differences in GM values. The required sample sizes could be calculated using a non-linear regression: $n = \beta_2\Delta^{-2} + \beta_1\Delta^{-1} + \beta_0$ (where β_i is the corresponding correlation coefficient). All inverse second-order correlations had R^2 values >0.9999, and the error in the sample size determination was less than one sample for any Δ between 1 % and 30 %. For example, 578 samples were required to observe a 7 % difference in seawater DGM GM at Station ST103 at $\alpha = 0.05$ and a statistical power of 0.80, calculated using the corresponding inverse second-order regression.

This study's primary limitation, in the context of current environmental challenges, is its inability to account for the effects of climate change on the natural variability of Hg concentrations. The dataset's two-year time frame is too brief to capture shifts in global seawater temperatures, Hg deposition and emissions, or their subsequent impacts on Hg variability and biogeochemical transformations. Future research should address this gap by utilizing data collected over a significantly

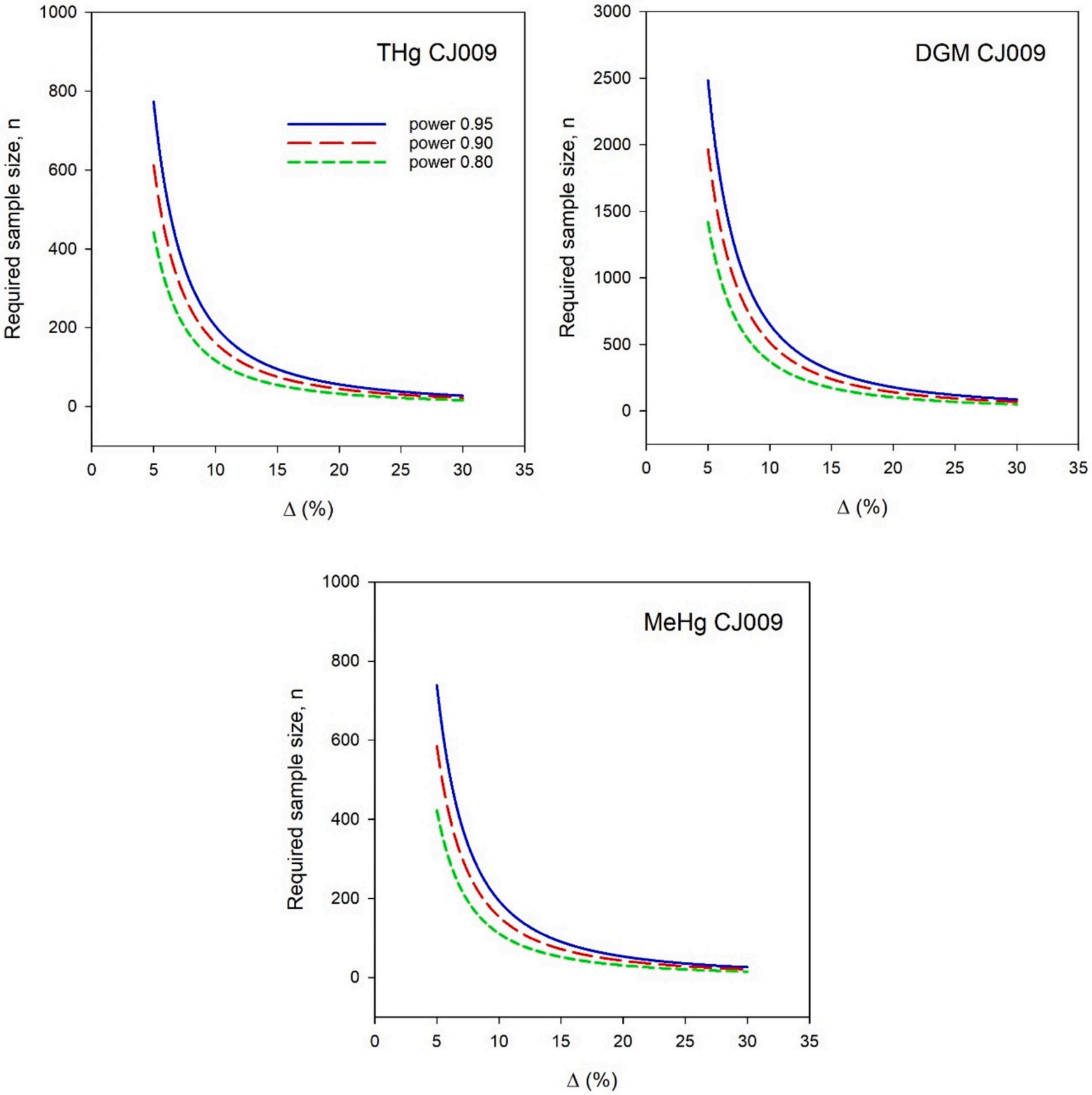


Fig. 4. Sample sizes (n) required to detect significant differences in THg, DGM, and MeHg concentrations in seawater at Station CJ009 between the two groups. Δ represents the expected difference between the postulated mean and the determined geometric mean (%). We assumed that all analyses were performed in a single laboratory. Sample size calculations were determined at three statistical powers (0.95, 0.90, and 0.80) using a one-sided *t*-test and an alpha value of 0.05.

longer period. Currently, this is out of the scope of this study.

3.4.1. *Effect of measurement uncertainty within a single laboratory*

Measurement uncertainty (u) at relevant concentration ranges was calculated for THg, DGM, and MeHg to account for analytical variability within a single laboratory. Prior to determining the required sample sizes, the respective uncertainties were transformed to ln-scale so that they matched the scale for $SD_{s\text{-}ln}$. Considering the measurement uncertainty calculated for the analytical procedures within the JSI laboratory (Section 2.2), the sample sizes required to observe 10 % differences in geometric mean concentrations at Station CJ009 needed to be increased from 203 to 211 (for THg), 651 to 658 (for DGM), and 194 to 205 (for MeHg) (Table 3).

3.4.2. *Effect of measurement variability among different laboratories*

Accounting for the interlaboratory variability (SD_{inter}) observed during the ILC exercises (Section 3.3), the required sample sizes needed to be increased by an additional 8 % for DGM and MeHg, whereas for THg, the needed increase was approximately 24 % (Table 4). The limiting factor in this analysis was that the ILC exercise for THg and MeHg was performed on only one sample (i.e., at only one concentration level). The general observation during the ILC exercises was that the equivalence of results between participating laboratories increased (and SD_{inter} decreased) with greater mean values of Hg species. For the higher concentrations of THg observed in the case study (about 1.8 ng L^{-1} at Station ST103; Table 2), the required sample size would actually be lower than indicated in Table 4.

3.5. *Sample size determination (other examples)*

In the previous section, we determined the sample sizes, assuming only differences between GM and the postulated means. However, it is not always realistic to assume the same variance at different stations, even for those that are relatively close. In the Supplementary Text S1, we presented an example of a sample size calculation based on two-sample comparisons of means (two-sided statistics). Sample size determination for a station with low natural variability was presented in Supplementary Text S2, while the importance of uncertainty in single-sample measurements was described in Supplementary Text S3.

3.6. *How can contributions to measurement uncertainty be reduced?*

The lower the measurement uncertainty, the narrower the 95 % CI will be. The uncertainty may be large for low concentrations of Hg fractions (near the detection limits), especially in samples from open ocean waters. Common errors during analysis can originate from extrapolations of the calibration curves near the detection limit or from using too small amount of sample. Given the fact that the components of measurement uncertainty vary across methods, it is advisable, when possible, to lower the uncertainty component that makes the greatest contribution to overall measurement uncertainty. It is extremely difficult, if not impossible, to influence the reproducibility of an analytical method, but it is relatively easy to influence other uncertainty components. Here, we provide guidelines for reducing measurement uncertainty by slightly modifying the analytical procedure. We focused on the calibration of the method and recovery corrections.

Table 3

Sample sizes (n) required to observe 5 %, 10 %, and 20 % differences in geometric mean concentrations of THg, DGM, and MeHg in seawater at Station CJ009 between the two groups. Sample size was based on $SD_{c1\text{-}ln}$, which considered sample variability ($SD_{s\text{-}ln}$) and measurement uncertainty within a single laboratory (u_{ln}). Sample sizes were determined at a statistical power of 0.95 using a one-sided *t*-test and an alpha value of 0.05.

	$\bar{x}_{ln}$ [a]	$SD_{s\text{-}ln}$ [a]	u_{ln} [a]	$SD_{c1\text{-}ln}$ [a]	n for 5 % [b]	n for 10 % [b]	n for 20 % [b]
THg	5.844	0.412	0.084	0.421	805	211	58
DGM	4.232	0.739	0.078	0.743	2511	658	180
MeHg	2.207	0.403	0.096	0.414	781	205	56

[a] Dimensionless value.

[b] Percentage represents the expected difference between the postulated mean and determined geometric mean of the original data.

Table 4

Sample sizes (n) required to observe 5 %, 10 %, and 20 % differences in geometric mean concentrations of THg, DGM, and MeHg in seawater at Station CJ009 between the two groups. Sample size was based on $SD_{c2\text{-}ln}$, which considered sample variability ($SD_{s\text{-}ln}$), measurement uncertainty within a single laboratory (u_{ln}), and standard deviation in the interlaboratory comparison exercise ($SD_{inter\text{-}ln}$). Sample sizes were determined at a statistical power of 0.95 using a one-sided t-test and an alpha value of 0.05.

	$\bar{x}_{ln}$ [a]	$SD_{s\text{-}ln}$ [a]	u_{ln} [a]	$SD_{inter\text{-}ln}$ [a]	$SD_{c2\text{-}ln}$ [a]	n for 5 % [b]	n for 10 % [b]	n for 20 % [b]
THg	5.844	0.412	0.084	0.205	0.468	996	261	72
DGM	4.232	0.739	0.078	0.201	0.770	2694	706	193
MeHg	2.207	0.403	0.096	0.120	0.431	847	222	61

[a] Dimensionless value.

[b] Percentage represents the expected difference between the postulated mean and determined geometric mean of the original data.

3.6.1. *Uncertainty of the calibration curve*

The estimation of the measurement uncertainty of an analytical result obtained from a calibration curve is often approximated using the so-called error of prediction, which is given in its relative form ($u_{cal,r}$) by the following equation:

$$u_{cal,r} = \frac{1}{x_{pred} * m} * \sqrt{\frac{\sum_j \left(y_j - \hat{y}_j\right)^2}{H-2}} * \sqrt{\frac{1}{G} + \frac{1}{H} + \frac{\left(\overline{y_0} - \overline{y}\right)^2}{m^2 * \sum_j \left(x_j - \overline{x}\right)^2}} \tag{6}$$

where x_{pred} is the predicted (calculated) value for Hg concentration in the sample, y_j is the observed detector's response (y) for a given Hg concentration in a standard solution (x_j), $\overline{x}$ is the mean of the x_j values, $\hat{y}_j$ is the value of y predicted by the equation of the calibration curve for a given x_j, $\overline{y_0}$ is the mean of G repeat measurements of y for the sample, $\overline{y}$ is the calculated mean of the detector's responses for all calibration standards, m is the calculated slope of the calibration curve, and H is the number of calibration points (Hibbert, 2006; Prichard and Barwick, 2003; Theodorou et al., 2012).

According to Eq. 6, there are two possible ways of lowering the $u_{cal,r}$ of the x_{pred}, either by lowering the numerator or by increasing the denominator in Eq. 6. The slope of the calibration curve (m) is a proxy for the sensitivity of the instruments, and the detector's response to Hg should be as great as possible. This can be achieved by increasing the source lamp's luminosity (voltage), by using a low-noise (clean) photomultiplier for an AFS detector, or by using a mirrored cuvette (given that noise is low). The other option is to increase the number of parallel samples (sample repeatability G) and calibration points (H). The term $y_j - \hat{y}_j$ represents the residual of the calibration curve, and it is lowest when the R^2 value is greatest. Finally, the means of Hg concentrations in standard solutions ($\overline{x}$) can be increased, but these values relate to the corresponding means for the detector's responses ($\overline{y}$).

The $\overline{y}$ can be lowered by narrowing the range of Hg concentrations in the calibration standards (consequently also lowering $\overline{x}$). US EPA Method 1631 states that a calibration curve should be created for calibration standards between 0.5 and 100 ng L^{-1} (US EPA, 2002). The JSI laboratory commonly uses calibration standards from 0.5 and 25 ng L^{-1}, which are adequate for determining THg in practically all water samples. Using this approach, we determined a relative combined standard uncertainty of 23.6 % (k = 2) at the 0.3 ng L^{-1} level. However, if the calibration curve was narrowed to 0.1–1.0 ng L^{-1}, the corresponding uncertainty decreased to 13.0 % (k = 2). This single change in the experimental setup achieved the greatest decrease in measurement uncertainty. It is important to note that the corresponding R^2 value was much lower for the 0.1–1.0 ng L^{-1} calibration ($R^2 = 0.99717$) than for the 0.5–25 ng L^{-1} calibration ($R^2 = 0.99997$) (Fig. 5). It is difficult to obtain a good calibration curve at low Hg levels. All reagents must be freshly prepared, and all tubes and vials used for sample preparation must be thoroughly washed. Nevertheless, this calibration curve could not be used because the commonly used certified reference material for THg in seawater (BCR 579) has a reference value of 1.90 ± 0.25 ng kg^{-1} (i.e., 1.94 ± 0.25 ng L^{-1} at 21 °C, k = 1), falling outside the range of calibration standards. As an alternative, a purged Hg-free sample could be spiked with a small amount of the CRM (e.g., ERM-CA400). However, this is not the correct use of a CRM.

3.6.2. *Uncertainty due to recovery*

Uncertainty due to recovery (u_R) can be omitted if it does not substantially contribute to overall measurement uncertainty and if the results are not corrected using recovery factors. However, there is no consensus in the literature about when to apply recovery factors to Hg speciation results. Harmonized guidelines for the use of recovery information in analytical measurement state that measurement results can be corrected using recovery factors if recovery differs significantly from 100 % (Thompson et al., 1999). We propose the same approach for deciding whether recovery factors should be applied to Hg speciation results. Unlike the aforementioned guidelines, we present a slightly different approach for testing statistical differences due to the relatively large uncertainties of certified reference materials for Hg in environmental samples and possible differences in sample variance. Examples are provided for the recovery of THg in BCR 579 and spiking of MeHg in seawater.

Welch *t*-tests were used to test whether the experimental means of THg in BCR 579 significantly differed from the reference value of 1.90 ± 0.25 ng kg^{-1} (i.e., 1.94 ± 0.25 ng L^{-1} at 21 °C, k = 1). The number of experimental determinations of THg concentration required to perform the Welch t-tests was based on QA/QC data from the JSI laboratory, whereas the number of reference determinations was obtained from the CRM certificate ($n = 6$). We assumed that the reference data were normally distributed because the certificate stated that the expanded uncertainty with a coverage factor of k = 2 corresponded to a CI of about 95 %. The Welch t-tests showed no significant differences between the experimental and reference values ($p > 0.05$, two-sided); therefore, the final results for THg in seawater samples were not corrected using the appropriate recovery factors. For the comparison, the theoretical normal distribution curves for the experimental data and reference values (Fig. 6) were calculated from the means and standard deviations (or uncertainty). It is also important to note that data should be transformed to an ln scale if the obtained experimental results do not follow normal distribution (which was not the case here).

A similar approach was used to determine statistical differences between the means for experimental recoveries for 5-pg MeHg spikes and the spike that was used in the method to determine MeHg in seawater using extraction as a preconcentration step (Liang et al., 1996). Extraction recoveries are usually highly variable, which results in a wide spread of recoveries. In contrast, the uncertainty of the spike is narrow (Fig. 6) and mainly depends on the uncertainty of the standard solution. The main reason why Welch t-tests should be used instead of Student's t-tests and the aforementioned guidelines is that the Welch t-test does not assume equal variance between groups. The number of experimental determinations of MeHg spikes required to perform the Welch t-tests was the same as the number of spikes. Welch t-tests showed statistically significant differences between the measured recovery of the spike and the spike itself ($p < 0.05$, two-sided). Low extraction recoveries of MeHg spikes could be attributed to several factors, including both chemical and procedural influences: matrix effects, complexation with dissolved organic compounds or other ligands, insufficient mixing during the extraction, and loss during phase separation. The final results for the determination of MeHg in seawater using extraction were corrected using the recovery factor, and the uncertainty of the recovery was included in the estimation of the overall measurement uncertainty.

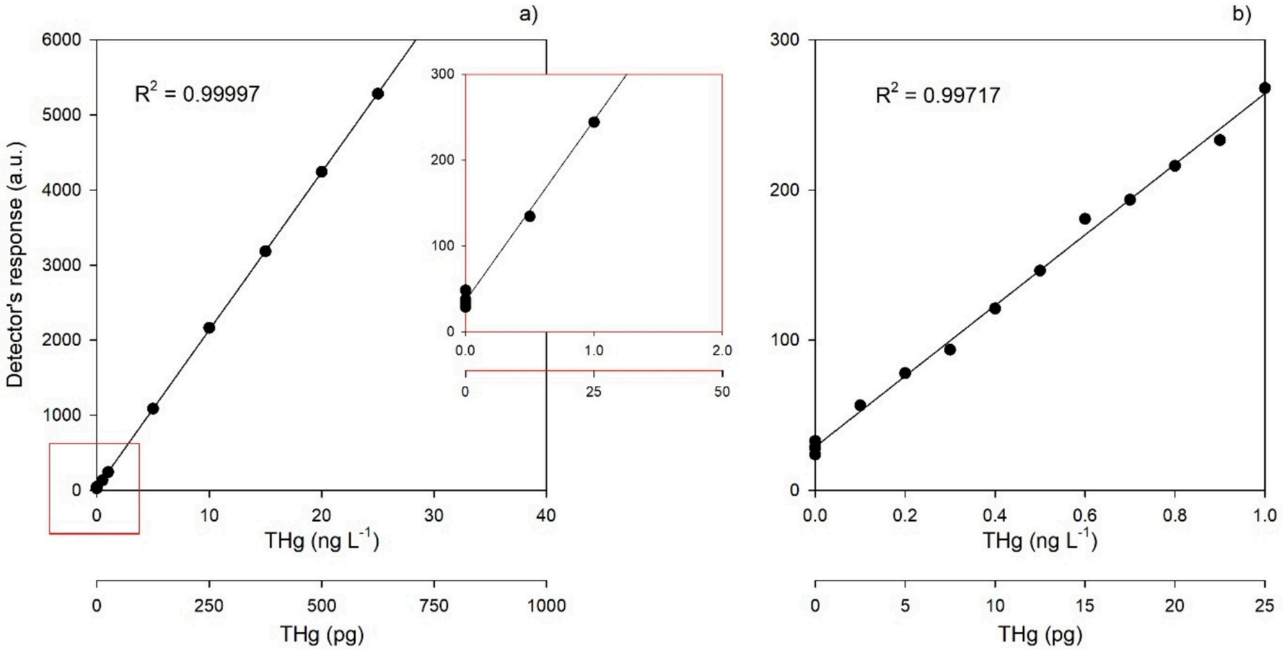


Fig. 5. Comparison of calibration curves for THg in seawater obtained for different ranges of calibration standards. The calibration curve presented in panel a) was obtained for THg concentrations of 0.0–25.0 ng L^{-1} (x axis) (i.e., 0–625 pg of added Hg standard [offset x axis]). The calibration curve presented in panel b) was obtained for THg concentrations of 0.0–1.0 ng L^{-1} (x axis) (i.e., 0–25 pg of added Hg standard [offset x axis]). The red square in panel a) presents the lowest THg standards and is enlarged for clarity (with the same axis titles). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

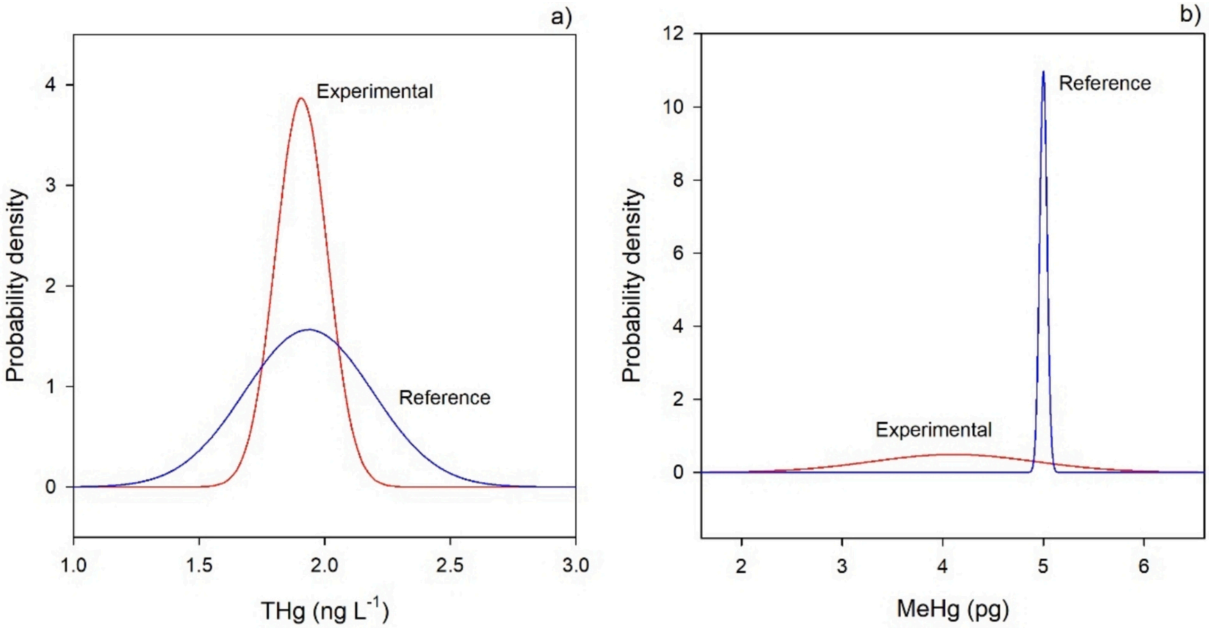


Fig. 6. Theoretical normal distribution curves for experimental data and reference values for a) THg in BCR 579 determined using CVAFS and b) a 5-pg spike of MeHg in seawater determined using CVAFS after sample extraction and ethylation.

4. Conclusion

This study evaluated the significance of natural sample variability, measurement variability within a single laboratory and between multiple laboratories, examining their effect on the required sample sizes to observe spatial and temporal trends of Hg species in seawater. Natural sample variability contributed most to the sample size, as inferred from the data for the Central Adriatic Sea. The largest sample variability in the whole dataset (all vertical profiles from the surface to the near-bottom water layer) was observed for DGM, causing an unrealistically large required sample size, even at the lowest statistical power. Generally, significant variabilities in Hg concentrations arise from differences in Hg sources between coastal and open-water stations. In contrast, lower variability was observed for MeHg because this fraction is natural in origin. While measurement uncertainty and variations due to interlaboratory variability had a minimal impact on the overall required sample size, they could significantly influence the sample size needed for samples with lower natural variability.

Based on the results from this study, we provide the following guidelines/recommendations for the improvement of Hg measurements comparability and harmonization of Hg analyses:

- Standardization of sampling and analysis: Adopt standardized methods for Hg sample collection, preservation, and analysis (e.g., GEOTRACES Cookbook) to minimize interlaboratory variability of results.
- Standardize ILC conditions: Regularly participate in ILCs with agreed standardized protocols to improve laboratory performance. Conduct thorough stability and homogeneity testing of ILC samples before distribution to participating laboratories. Ensure all participating laboratories follow identical sample handling, storage, and analysis protocols during ILCs to minimize variability.
- Implementation of metrological traceability: Link all Hg measurements to a primary reference standard (e.g., NIST SRM 3133) to ensure traceability and comparability across different studies and between laboratories.
- Development of low-level CRMs: CRMs with low Hg concentrations and uncertainties suitable for trace analysis in seawater should be produced. Current limitations in CRM concentration ranges and large uncertainties can hinder their utilization in quality assurance.
- Implement log-transformation for statistical analysis: For datasets that follow a log-normal distribution, standardize the use of log-transformed data for statistical analyses of Hg concentrations. Report geometric means, geometric standard deviations, and appropriate confidence intervals to properly present summary data.
- Determine sample size based on target variability: Tailor sample sizes to detect specific differences in (geometric) mean concentrations. Conduct more frequent and continuous sampling campaigns to better capture temporal variability in Hg speciation data. In areas with high natural variability (e.g., coastal sites), increase sample sizes to distinguish between natural fluctuations and true spatial or temporal trends in Hg concentrations.
- Incorporate measurement uncertainty in sample size calculations: If required, log-transform data before statistical analysis to properly determine sample size estimates. Factor in both intra-laboratory and interlaboratory uncertainties to calculate sample sizes. This ensures that sample size estimates reflect variability introduced by analytical methods and laboratory differences.
- Optimization of calibration curve: Narrow the range of the calibration curve to align with target Hg concentration levels in seawater. Improve calibration by increasing the number of calibration points and replicates, ensuring higher sensitivity and stability of the detector. This reduces uncertainty due to calibration, particularly near detection limits, and improves accuracy for low-level measurements.
- Recovery uncertainty management: Apply recovery factor corrections only when recoveries deviate significantly from 100 %. Use robust statistical methods, such as the Welch *t*-test, to assess whether recovery deviates from unity, considering variability in certified reference materials and analytical procedures. When applying recovery factors, include the corresponding uncertainty contribution in the overall measurement uncertainty budget.

A key takeaway from this exercise is that without accounting for measurement uncertainties and relying solely on a single measurement of Hg concentrations in oceanic waters, it becomes challenging to interpret and discuss results in a broader context. The comparability and compatibility of these results cannot be fully demonstrated, rendering any conclusions speculative. Therefore, it is strongly recommended that the measurement community prioritize these aspects, as demonstrating comparability is essential for global Hg assessments in oceans. Moreover, to fully comply with the requirements of the Minamata Convention, adopting such an approach will enable meaningful assessments of Hg levels.

CRediT authorship contribution statement

Igor Živković: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Lars-Eric Heimbürger-Boavida:** Writing – original draft, Validation, Supervision, Resources, Methodology, Investigation, Formal analysis, Data curation. **Mariia V. Petrova:** Writing – review & editing, Validation, Investigation, Formal analysis, Data curation. **Aurélie Dufour:** Writing – review & editing, Validation, Investigation, Formal analysis, Data curation. **Ermira Begu:** Writing – review & editing, Validation, Methodology, Investigation, Formal analysis. **Milena Horvat:** Writing – original draft, Validation, Supervision, Resources, Project administration, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The work has been funded by ERA-NET co-fund, grant agreement 689443 ERA-PLANET/IGOSP. The research was supported by the Slovenian Research and Innovation Agency (program P1-0143 and projects J1-3033-1 and J7-60123). The authors thank to the Metrology Institute of the Republic of Slovenia (MIRS) under contract No. C3212-10-000071 (6401-5/2009/27) for activities and obligations performed as a Designate Institute. The authors also thank the crew of research vessels and participants of the GEOTRACES intercomparison exercises and cruise. The funders had no role in the study design; in the collection, analysis and interpretation of data; in the writing of the report; and in the decision to submit the article for publication.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.marchem.2025.104498.

Data availability

Data will be made available upon reasonable request.

References

Andron, T.D., Corns, W.T., Živković, I., Ali, S.W., Vijayakumaran Nair, S., Horvat, M., 2024. A traceable and continuous flow calibration method for gaseous elemental mercury at low ambient concentrations. Atmos. Meas. Tech. 17, 1217–1228. https://doi.org/10.5194/amt-17-1217-2024.

Artegiani, A., Paschini, E., Russo, A., Bregant, D., Raicich, F., Pinardi, N., 1997. The Adriatic Sea general circulation. Part II: Baroclinic circulation structure. J. Phys. Oceanogr. 27, 1515–1532. https://doi.org/10.1175/1520-0485(1997)027<1515:TASGCP>2.0.CO;2.

Basu, N., Bastiansz, A., Dórea, J.G., Fujimura, M., Horvat, M., Shroff, E., Weihe, P., Zastenskaya, I., 2023. Our evolved understanding of the human health risks of mercury. Ambio 52, 877–896. https://doi.org/10.1007/s13280-023-01831-6.

BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, OIML, 2008. Evaluation of measurement data — guide to the expression of uncertainty in measurement. Joint Committ. Guid. Metrol. JCGM 100, 2008. https://doi.org/10.59161/JCGM100-2008E.

BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, OIML, 2012. International vocabulary of metrology — basic and general concepts and associated terms (VIM). Joint Committ. Guid. Metrol. JCGM 200. (3rd edition). Doi:10.59161/JCGM200-2012.

Black, F.J., Conaway, C.H., Flegal, A.R., 2009. Stability of dimethyl mercury in seawater and its conversion to monomethyl mercury. Environ. Sci. Technol. 43, 4056–4062. https://doi.org/10.1021/es9001218.

Bland, J.M., Altman, D.G., 1996a. Transformations, means, and confidence intervals. BMJ 312, 1079. https://doi.org/10.1136/bmj.312.7038.1079.

Bland, J.M., Altman, D.G., 1996b. The use of transformation when comparing two means. BMJ 312, 1153. https://doi.org/10.1136/bmj.312.7039.1153.

Carobbi, C.F.M., 2010. The use of logarithmic units in the uncertainty evaluations of EMC measurements. In: O'Neil, J. (Ed.), IEEE EMC Society Newsletter, Winter 2010 – Issue No. 224. IEEE EMC Society, New York, USA, pp. 46–50.

Cossa, D., Courau, P., 1990. An international intercomparison exercise for total mercury in seawater. Appl. Organomet. Chem. 4, 49–54. https://doi.org/10.1002/aoc.590040109.

Cossa, D., Martin, J.-M., Takayanagi, K., Sanjuan, J., 1997. The distribution and cycling of mercury species in the western Mediterranean. Deep Sea Res. Part II Top. Stud. Oceanogr. 44, 721–740. https://doi.org/10.1016/S0967-0645(96)00097-5.

Cossa, D., Averty, B., Pirrone, N., 2009. The origin of methylmercury in open mediterranean waters. Limnol. Oceanogr. 54, 837–844. https://doi.org/10.4319/lo.2009.54.3.0837.

Cossa, D., Knoery, J., Bănaru, D., Harmelin-Vivien, M., Sonke, J.E., Hedgecock, I.M., Bravo, A.G., Rosati, G., Canu, D., Horvat, M., Sprovieri, F., Pirrone, N., Heimbürger-Boavida, L.-E., 2022. Mediterranean mercury assessment 2022: an updated budget, health consequences, and research perspectives. Environ. Sci. Technol. 56, 3840–3862. https://doi.org/10.1021/acs.est.1c03044.

Cutter, G., Casciotti, K., Croot, P., Geibert, W., Heimbürger, L.-E., Lohan, M., Planquette, H., van de Flierdt, T., 2017. Sampling and Sample-Handling Protocols for GEOTRACES Cruises, 3rd ed. GEOTRACES International Project Office, Toulouse, France. https://doi.org/10.25607/OBP-2.

De Bièvre, P., 2006. Comparability vs. degree of equivalence: another terminological headache (in the usage of the English language). Accred. Qual. Assur. 11, 487–488. https://doi.org/10.1007/s00769-006-0182-0.

Ellison, S.L.R., Williams, A., 2012. Eurachem/CITAC Guide: Quantifying Uncertainty in Analytical Measurement, 3rd ed. Eurachem/CITAC, Teddington, UK. https://doi.org/10.25607/OBP-952.

Filliben, J.J., 2012. Exploratory data Analysis. In: Croarkin, C., Tobias, P. (Eds.), NIST/SEMATECH e-Handbook of Statistical Methods. NIST/SEMATECH. https://doi.org/10.18434/M32189.

Gao, A., Martos, P., 2019. Log transformation and the effect on estimation, implication, and interpretation of mean and measurement uncertainty in microbial enumeration. J. AOAC Int. 102, 233–238. https://doi.org/10.5740/jaoacint.18-0161.

Guevara, S.R., Horvat, M., 2013. Stability and behaviour of low level spiked inorganic mercury in natural water samples. Anal. Methods 5, 1996–2006. https://doi.org/10.1039/c3ay26496c.

Gworek, B., Bemowska-Kałabun, O., Kijeńska, M., Wrzosek-Jakubowska, J., 2016. Mercury in marine and oceanic waters—a review. Water Air Soil Pollut. 227, 371. https://doi.org/10.1007/s11270-016-3060-3.

Heimbürger, L.E., Cossa, D., Marty, J.C., Migon, C., Averty, B., Dufour, A., Ras, J., 2010. Methyl mercury distributions in relation to the presence of nano- and picophytoplankton in an oceanic water column (Ligurian Sea, North-Western Mediterranean). Geochim. Cosmochim. Acta 74, 5549–5559. https://doi.org/10.1016/j.gca.2010.06.036.

Heimbürger, L.-E., Sonke, J.E., Cossa, D., Point, D., Lagane, C., Laffont, L., Galfond, B.T., Nicolaus, M., Rabe, B., van der Loeff, M.R., 2015. Shallow methylmercury production in the marginal sea ice zone of the Central Arctic Ocean. Sci. Rep. 5, 10318. https://doi.org/10.1038/srep10318.

Hibbert, D.B., 2006. The uncertainty of a result from a linear calibration. Analyst 131, 1273–1278. https://doi.org/10.1039/b615398d.

Higgins, J.P.T., White, I.R., Anzures-Cabrera, J., 2008. Meta-analysis of skewed data: combining results reported on log-transformed or raw scales. Stat. Med. 27, 6072–6092. https://doi.org/10.1002/sim.3427.

Horvat, M., Liang, L., Bloom, N.S., 1993. Comparison of distillation with other current isolation methods for the determination of methyl mercury compounds in low level environmental samples: part II. Water. Anal. Chim. Acta 282, 153–168. https://doi.org/10.1016/0003-2670(93)80364-Q.

Horvat, M., Kotnik, J., Logar, M., Fajon, V., Zvonarić, T., Pirrone, N., 2003. Speciation of mercury in surface and deep-sea waters in the Mediterranean Sea. Atmos. Environ. 37, 93–108. https://doi.org/10.1016/S1352-2310(03)00249-8.

ISO, 2015a. ISO 33: Reference Materials — Good Practice in Using Reference Materials. ISO, Geneva, Switzerland.

ISO, 2015b. ISO 13528: Statistical Methods for Use in Proficiency Testing by Interlaboratory Comparison. ISO, Geneva, Switzerland.

Kleindienst, A., Živković, I., Tessier, E., Koenig, A., Heimbürger-Boavida, L.E., Horvat, M., Amouroux, D., 2023. Assessing comparability and uncertainty of analytical methods for methylated mercury species in seawater. Anal. Chim. Acta 1278, 341735. https://doi.org/10.1016/j.aca.2023.341735.

Kotnik, J., Horvat, M., Ogrinc, N., Fajon, V., Žagar, D., Cossa, D., Sprovieri, F., Pirrone, N., 2015. Mercury speciation in the Adriatic Sea. Mar. Pollut. Bull. 96, 136–148. https://doi.org/10.1016/j.marpolbul.2015.05.037.

Kotnik, J., Horvat, M., Begu, E., Shlyapnikov, Y., Sprovieri, F., Pirrone, N., 2017. Dissolved gaseous mercury (DGM) in the Mediterranean Sea: spatial and temporal trends. Mar. Chem. 193, 8–19. https://doi.org/10.1016/j.marchem.2017.03.002.

Kramer, K.J.M., Quevauviller, P., Dorten, W.S., van der Vlies, E.M., de Haan, H.P.M., 1998. Certification of total mercury in a sea-water reference material, CRM 579. Analyst 123, 959–963. https://doi.org/10.1039/a705039i.

Kwokal, Ž., Frančišković-Bilinski, S., Bilinski, H., Branica, M., 2002. A comparison of anthropogenic mercury pollution in Kaštela Bay (Croatia) with pristine estuaries in

Öre (Sweden) and Krka (Croatia). Mar. Pollut. Bull. 44, 1152–1157. https://doi.org/10.1016/S0025-326X(02)00134-0.

Lamborg, C.H., Hammerschmidt, C.R., Gill, G.A., Mason, R.P., Gichuki, S., 2012. An intercomparison of procedures for the determination of total mercury in seawater and recommendations regarding mercury speciation during GEOTRACES cruises. Limnol. Oceanogr. Methods 10, 90–100. https://doi.org/10.4319/lom.2012.10.90.

Liang, L., Horvat, M., Cernichiari, E., Gelein, B., Balogh, S., 1996. Simple solvent extraction technique for elimination of matrix interferences in the determination of methylmercury in environmental and biological samples by ethylation-gas chromatography-cold vapor atomic fluorescence spectrometry. Talanta 43, 1883–1888. https://doi.org/10.1016/0039-9140(96)01964-9.

Measures, C.I., Landing, W.M., Brown, M.T., Buck, C.S., 2008. A commercially available rosette system for trace metal-clean sampling. Limnol. Oceanogr. Methods 6, 384–394. https://doi.org/10.4319/lom.2008.6.384.

NIST, 2016. Standard Reference Material 3133. NIST, Gaithersburg, MD, USA.

Petrova, M.V., Krisch, S., Lodeiro, P., Valk, O., Dufour, A., Rijkenberg, M.J.A., Achterberg, E.P., Rabe, B., Rutgers van der Loeff, M., Hamelin, B., Sonke, J.E., Garnier, C., Heimbürger-Boavida, L.-E., 2020. Mercury species export from the Arctic to the Atlantic Ocean. Mar. Chem. 225, 103855. https://doi.org/10.1016/j.marchem.2020.103855.

Prichard, L., Barwick, V., 2003. Preparation of Calibration Curves: A Guide to Best Practice. LGC Limited, Teddington, UK. https://doi.org/10.13140/RG.2.2.36338.76488.

Quétel, C.R., Zampella, M., Brown, R.J.C., 2016. Temperature dependence of Hg vapour mass concentration at saturation in air: new SI traceable results between 15 and 30°C. TrAC Trends Anal. Chem. 85, 81–88. https://doi.org/10.1016/j.trac.2015.12.010.

Schlitzer, R., 2024. Ocean Data View. https://odv.awi.de, 2024.

Singh, A.D., Khanna, K., Kour, J., Dhiman, S., Bhardwaj, T., Devi, K., Sharma, N., Kumar, P., Kapoor, N., Sharma, P., Arora, P., Sharma, A., Bhardwaj, R., 2023. Critical review on biogeochemical dynamics of mercury (Hg) and its abatement strategies. Chemosphere 319, 137917. https://doi.org/10.1016/j.chemosphere.2023.137917.

Snoj Tratnik, J., Mazej, D., Horvat, M., 2019. Analytical quality requirements in human biomonitoring programs: trace elements in human blood. Int. J. Environ. Res. Public Health 16, 2287. https://doi.org/10.3390/ijerph16132287.

Šolić, M., Krstulović, N., Šantić, D., Šestanović, S., Ordulj, M., Bojanić, N., Kušpilić, G., 2015. Structure of microbial communities in phosphorus-limited estuaries along the eastern Adriatic coast. J. Mar. Biol. Assoc. United Kingdom 95, 1565–1578. https://doi.org/10.1017/S0025315415000442.

Sprovieri, F., Pirrone, N., Bencardino, M., D'Amore, F., Carbone, F., Cinnirella, S., Mannarino, V., Landis, M., Ebinghaus, R., Weigelt, A., Brunke, E.-G., Labuschagne, C., Martin, L., Munthe, J., Wängberg, I., Artaxo, P., Morais, F., Barbosa, H. de M.J., Brito, J., Cairns, W., Barbante, C., Diéguez, M. del C., Garcia, P. E., Dommergue, A., Angot, H., Magand, O., Skov, H., Horvat, M., Kotnik, J., Read, K. A., Neves, L.M., Gawlik, B.M., Sena, F., Mashyanov, N., Obolkin, V., Wip, D., Feng, X. Bin, Zhang, H., Fu, X., Ramachandran, R., Cossa, D., Knoery, J., Marusczak, N., Nerentorp, M., Norstrom, C., 2016. Atmospheric mercury concentrations observed at ground-based monitoring sites globally distributed in the framework of the GMOS network. Atmos. Chem. Phys. 16, 11915–11935. https://doi.org/10.5194/acp-16-11915-2016.

Theodorou, D., Zannikou, Y., Zannikos, F., 2012. Estimation of the standard uncertainty of a calibration curve: application to sulfur mass concentration determination in fuels. Accred. Qual. Assur. 17, 275–281. https://doi.org/10.1007/s00769-011-0852-4.

Thompson, M., Ellison, S.L.R., Fajgelj, A., Willetts, P., Wood, R., 1999. Harmonized guidelines for the use of recovery information in analytical measurement. Pure Appl. Chem. 71, 337–348. https://doi.org/10.1351/pac199971020337.

Torres-Rodriguez, N., Yuan, J., Petersen, S., Dufour, A., González-Santana, D., Chavagnac, V., Planquette, H., Horvat, M., Amouroux, D., Cathalot, C., Pelleter, E., Sun, R., Sonke, J.E., Luther, G.W., Heimbürger-Boavida, L.-E., 2024. Mercury fluxes from hydrothermal venting at mid-ocean ridges constrained by measurements. Nat. Geosci. 17, 51–57. https://doi.org/10.1038/s41561-023-01341-w.

Trapmann, S., Botha, A., Linsinger, T.P.J., Mac Curtain, S., Emons, H., 2017. The new international standard ISO 17034: general requirements for the competence of reference material producers. Accred. Qual. Assur. 22, 381–387. https://doi.org/10.1007/s00769-017-1285-5.

UN, 2013. Minamata Convention on Mercury. United Nations, Geneva, Switzerland.

UNEP, 2021. UNEP/MC/COP.4/INF/12 - Guidance on Monitoring Mercury and Mercury Compounds to Support the Effectiveness Evaluation of the Minamata Convention. UNEP, Bali, Indonesia.

US EPA, 2002. Method 1631, Revision E: Mercury in Water by Oxidation, Purge and Trap, and Cold Vapor Atomic Fluorescence Spectrometry. US EPA, Washington, DC, USA.

Wu, Y.-S., Osman, A.I., Hosny, M., Elgarahy, A.M., Eltaweil, A.S., Rooney, D.W., Chen, Z., Rahim, N.S., Sekar, M., Gopinath, S.C.B., Mat Rani, N.N.I., Batumalaie, K., Yap, P.-S., 2024. The toxicity of mercury and its chemical compounds: molecular mechanisms and environmental and human health implications: a comprehensive review. ACS Omega 9, 5100–5126. https://doi.org/10.1021/acsomega.3c07047.

Živković, I., Fajon, V., Tulasi, D., Obu Vazner, K., Horvat, M., 2017a. Optimization and measurement uncertainty estimation of hydride generation–cryogenic trapping–gas chromatography–cold vapor atomic fluorescence spectrometry for the determination of methylmercury in seawater. Mar. Chem. 193, 3–7. https://doi.org/10.1016/j.marchem.2017.03.003.

Živković, I., Kotnik, J., Šolić, M., Horvat, M., 2017b. The abundance, distribution and speciation of mercury in waters and sediments of the Adriatic Sea. Acta Adriat. 58, 165–186. https://doi.org/10.32582/aa.58.1.14.

Živković, I., Fajon, V., Kotnik, J., Shlyapnikov, Y., Obu Vazner, K., Begu, E., Šestanović, S., Šantić, D., Vrdoljak, A., Jozić, S., Šolić, M., Lušić, J., Veža, J., Kušpilić, G., Ordulj, M., Matić, F., Grbec, B., Bojanić, N., Ninčević Gladan, Ž., Horvat, M., 2019. Relations between mercury fractions and microbial community components in seawater under the presence and absence of probable phosphorus limitation conditions. J. Environ. Sci. 75, 145–162. https://doi.org/10.1016/j.jes.2018.03.012.