frontiers | Frontiers in Marine Science

TYPE Original Research
PUBLISHED 15 November 2022
DOI 10.3389/fmars.2022.1002153

OPEN ACCESS

EDITED BY
Juliet Hermes,
South African Environmental Observation Network (SAEON), South Africa

REVIEWED BY
Mark Bushnell,
National Ocean Service (NOAA), United States
Hiroshi Uchida,
Japan Agency for Marine-Earth Science and Technology (JAMSTEC), Japan

*CORRESPONDENCE
Christoph Waldmann
waldmann@uni-bremen.de

SPECIALTY SECTION
This article was submitted to Ocean Observation, a section of the journal Frontiers in Marine Science

RECEIVED 24 July 2022
ACCEPTED 31 October 2022
PUBLISHED 15 November 2022

CITATION
Waldmann C, Fischer P, Seitz S, Köllner M, Fischer J-G, Bergenthal M, Brix H, Weinreben S and Huber R (2022) A methodology to uncertainty quantification of essential ocean variables.
Front. Mar. Sci. 9:1002153.
doi: 10.3389/fmars.2022.1002153

A methodology to uncertainty quantification of essential ocean variables

Christoph Waldmann[1]*, Philipp Fischer[2], Steffen Seitz[3], Manuela Köllner[4], Jens-Georg Fischer[4], Markus Bergenthal[1], Holger Brix[5], Stefan Weinreben[6] and Robert Huber[1]

[1]Center for Marine Environmental Sciences (MARUM), University of Bremen, Bremen, Germany, [2]Biological Station, Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research (AWI), Helgoland, Germany, [3]Electrochemistry Department, Physikalisch-Technische Bundesanstalt, Braunschweig, Germany, [4]Oceanographic Assessment M22, Bundesamt fuer Seeschifffahrt und Hydrographie, Hamburg, Germany, [5]Institute of Carbon Cycles, Helmholtz-Zentrum Hereon, Geesthacht, Germany, [6]Physical Oceanography and Instrumentation, Leibniz-Institute for Baltic Sea Research, Rostock, Germany

The goal of this study is to provide a universally applicable procedure for a systematic evaluation of *in situ* measured data from single sensors regarding quantifying the uncertainty of the measurement results. As determining uncertainty for an environmental parameter also depends on the parameter itself, the focus here will be set on the variable water temperature in the first place. A separate analysis for salinity and other data will follow in later publications. With this first of a series of planned manuscripts on different parameters, we aim at providing a common understanding of how measurement uncertainty on single sensor measurements can be derived. Using an experimental *in situ* set-up with 6 different standard CTD sensors of two different brands, we created a four month-long, high-quality data set to be used to develop a reliable method for quantifying measurement uncertainties. Although the CTDs were deployed in a mooring in a coastal environment the described method can be extended to other deployment configurations as well. The described procedures have evolved as a stepwise process that takes the different perspectives of the involved authors into account, as well as the special conditions for environmental measurements, which are collected while the observed volume/area is undergoing a constant change. By sharing the ideas with other stakeholders, the basic concept can be extended to other observing programs and to other essential ocean variables.

KEYWORDS

uncertainty quantification, essential ocean variables, CTD, coastal observatory, calibration, metrology, quality control, flagging

Introduction

Considering the importance of judging the significance of observations to detect long term trends in earth systems, the current study appears to be timely and relevant for different kinds of observational activities e.g., as part of the UN Decade of Ocean Science for Sustainable Development (UN DECADE, U 2021). Measuring variables in the field is fundamentally different from lab measurements as *in-situ* measurements are unique in space and time and are of transient character. In the ocean sciences, where access to environmental data is often limited due to required ship time or due to weather constraints, single sensor data without the chance of replication and limited information on data quality are the only available source of information. The challenge is to define the concept of "data quality" which is connected to "measurement uncertainty".

Although in other disciplines like atmospheric observations as conducted by the World Meteorological Organization (WMO) the concept of uncertainty has already found entry (WMO, 2008), while ocean sciences have only dealt with uncertainties for specific parameters and often with a limited scope.

Over the past years, various aspects on data quality have been considered (Wong et al., 2022), (Bushnell, 2019) including the "FAIR" data concept (Wilkinson, 2016). The FAIR concept means that scientific data must be "findable", "accessible", "interoperable" and "reusable" including a minimum of associated metadata information which make data transparent with respect to their origin and their processing workflow. However, even FAIR data do not contain an adequate description of data quality, as is classically requested as a minimum standard in natural sciences. Referring to international conventions (UNESCO, 2013) here, data quality refers to the availability of "accuracy" and "precision" information [see Supplementary Material, Appendix 4 on terminology which is based on (BIPM, 2008)]. This will allow the calculation of a statistically robust uncertainty resp. confidence interval for each measured data point spanning the range within which the best estimate value lies with a specified (e.g., 95%) probability. Only when this statistical information for each individual data point is available can the "quality" of this data point be quantitatively assessed in a close context to the respective scientific question. While, for example, in behavioral ecology, temperature measurements used to determine whether an area is habitable for a particular species usually do not need to be more accurate than one degree Celsius, studies of the effects of climate change-induced heat content changes in the deep sea require uncertainties that do not exceed one hundredth of a degree or even less. Thus, while for the behavioral ecologist, the above-mentioned dataset is of sufficient and thus of "high" quality, for the oceanographer the same data set is of insufficient and therefore of "poor" quality. This implies that a data point without information about its uncertainty is neither good nor bad, but in a kind of premature raw data state that is not yet suitable for scientific use and publication without further refinement.

The study that has been carried out strives to perform an analysis of the relevant factors required to calculate a measured data points uncertainty. In a series of follow-up papers, we will address other related topics such as uncertainty of salinity measurements or uncertainty due to sensor drift.

In a first step, a basic concept of uncertainty calculation for time-series data measured by standard oceanographic probes (CTD, Conductivity, Temperature, and Depth) is presented. To keep the analysis as comprehensive as needed, while being as simple as possible, we here concentrate on analyzing the variable temperature as an example. Nevertheless, it should be relatively straightforward to extend this analysis to other, more complex, Essential Ocean Variables [EOVs, (GOOS, 2020)].

Along a data processing chain from the raw sensor output (e.g., T=14.345°C), we evaluate different procedures for a simple but statistically robust numerical uncertainty calculation of the measurement to finally come to an output in the form of

$$T = 14.345 \pm 0.003\,^{\circ}C$$

or verbalized,

Measured Value = Best estimate ± Uncertainty

(Taylor, 1997)

The procedures presented here are not meant to replace existing procedures and frameworks for data quality assurance developed and established in ocean sciences over the last decades. The intention of the manuscript, however, is to complement the information that is provided for environmental data.

A core element of "quality control" procedures are quality flags that assign collected data into different quality categories. The processing and quality control results are stored and published alongside with the data to allow scientists to decide whether data are plausible within a set of mathematical and logic criteria. Flags assigned to data are independent from the later scientific question and provide information if data fulfill simple criteria which make them theoretically valid or invalid. The idea is to exclude obviously or probably wrong data from a dataset.

Quality flags usually consist of a very basic defined code of numbers. A flag categorizes a data point as e.g., "good" or "bad". It can describe if data have been changed, replaced or added to the original raw data set during processing (e.g., "interpolated value") or it can reveal certain events within a data set (e.g., offset detected, spike detected). Usually, a single data point is marked with a unique flag corresponding to a specific interpretation of the "quality". Unfortunately, this marker is a combination of the results from different performed tests highly influenced by specified thresholds defined within each test. So far, there is no international agreement upon standards for flagging, as well as the choice of performed data quality tests and accompanied thresholds. Recommendations for necessary and optional useful

tests vary depending on the scientific party providing the data. Additionally, definitions for the used codes vary, although there are a number of similarities between the used conventions. For example, flagging schemes based on OceanSITES (OceanSITES, 2020), ARGO (Wong et al., 2022), Copernicus (Copernicus, 2020), and SeaDataNet (2010) follow the convention that "no quality test performed" is defined as flag="0", while schemes based on GO-SHIP (Swift, 2010) use flag=1 and (UNESCO, 2013) and IOOS (Bushnell, 2020) use flag=2 for the same. This causes major efforts regarding the mapping of flag information between the data providers. Interpretation of quality flags coming along with data from different data providers can thus be very time consuming for the user. Another issue can occur in the case of a lack of information about the results of individual quality tests when data meant for a specific purpose do not meet these predefined quality criteria and are excluded although they may be useful for other scientific questions under consideration.

Data flagging is therefore highly useful as a plausibility filter to exclude wrong data from datasets without a detailed knowledge of the specific sensor characteristics and functionality as well as without a specific knowledge on the later scientific question. Data flagging however cannot replace a quality assurance procedure providing statistically robust quantitative information on the data's uncertainty at a specified confidence range.

Another contribution to the overall uncertainty budget can be extracted from the sensor specifications determined in the manufacturer laboratory at the time of production and calibration that also should find entry into the metadata description of measured data. Most sensor manufacturers provide initial accuracy and precision values for their sensors and sometimes also information about the stability or drift over time. Even though this information is exactly the type of metadata required to calculate a sensor's uncertainty or confidence, one has to keep in mind that these manufacturer metadata are laboratory values referring to a brand-new or re-calibrated sensor and therefore do not take the sensor lifetime and environmental conditions during storage, transportation and/or deployment into account. Furthermore, it must be considered that manufacturers sometimes provide only information for their sensors describing a typical accuracy and/or precision for a sensor but not for a specific sensor instance. Better qualified sensor specific metadata are only available if the manufacturer provides a sensor specific calibration sheet with detailed information on the serial number of the respective sensor or if a recalibration will be carried out by another calibration laboratory. Therefore, we have to consider different levels of availability and reliability of given sensor accuracies indicating the demand for proper documentation of sensor metadata.

As mentioned before, flags are markers for data plausibility and provide workflow transparency. They do not include detailed information about the significance or robustness of a single data point/measurement. The manufacturers quality parameters of a sensor's data provide this information but cannot be easily applied to the operational phase of a measurement program. From the scientific point of view the knowledge of both information, realistic uncertainty in the operational phase of an experiment as well as flags determined from plausibility tests, would be helpful to prevent scientists from misinterpretation of data. However, while flags can be assigned to data points independent from the operational phase and status of a measurement, a way to assign uncertainty information on sensor measurements in the operational phase seem to be largely unknown and not yet widespread in ocean sciences.

Experimental set-up

The experimental setup was designed to be as close as possible to a normal monitoring program that would be conducted in a coastal area with a duration of several months. A balanced experimental approach with six different multi-parameter probes [CTD: three Sea & Sun Technology (Sea&Sun, 2022) and three Sea-Bird Scientific (Seabird, 2022)] (see also Supplementary Materials, Appendix 1, Table A1 for more details) from four different marine institutes in Germany were chosen. These six probes were deployed in the MarGate underwater test site (Wehkamp et al., 2013) off Helgoland in the southern North Sea from July 20th to November 25th, 2020. This underwater experimental field is jointly operated by the two Helmholtz institutes, Alfred-Wegener-Institute Helmholtz Centre for Polar- and Marine Research (AWI) and the Helmholtz Institute HEREON (formerly HZG) as an international monitoring and test facility for marine observing components. It has been part of the EU project Jerico-Next (Jerico-Next, 2019) where international cooperating partners could apply for the financial cover of time slots to evaluate marine sensors for scientific purposes.

The experimental area has a cable connected underwater node with ten submersed ports for continuous power and high-speed data connection for the remote-controlled operation of underwater sensor systems (Fischer, 2019).

The underwater field is continuously monitored for the main essential ocean variables such as temperature, conductivity, oxygen saturation, chlorophyll-a, turbidity, photosynthetic active radiation (PAR), current and wave height, as well as additional variables as pCO2 and methane concentration. Experiments in the so called MarGate field are supported year-round by specifically trained scientific divers who are responsible for sensor maintenance, repairing and replacing of sensors and new experimental set up. The area provides a highly demanding environment with average wind speed peaks of more than 6 bft (10.8-13.8 m/s) on more than 200 days a year and tidal currents up to 1 m/s.

The six CTD sensors used in this experiment were mounted in 9 m (+/- 1.5m tide) water depth in a metal lander frame (Figure 1). The sensors were mounted such that the sensor heads with the measuring cells were in the same height of 92 cm (+/- 5 cm) above the seafloor, perpendicular to the main current direction and offset at a horizontal distance of 15 cm to each other. This setup ensured that, except for micro turbulences below decimeter scale, all sensors were exposed to the same water body without disturbing each other.

The experimental procedure was designed as follows. Prior to the deployment, all sensors were registered in the AWI Sensor Registry (Registry, 2019) and calibrated in the calibration laboratory at Leibniz Institute for Baltic Sea Research, Warnemuende (IOW) (see Supplementary Material under Appendix 2) to ensure a consistency in the calibration process. On June 19, 2020, all six CTD were returned to Helgoland Centre for Scientific Diving. Five out of the six CTD's were deployed on July 22, 2020 11:00 hours in the experimental field, connected to the node system and the internal data logging mode was started. From this day on, the data from all sensors were downloaded every workday between 10:00 and 13:00 hours, if necessary, converted to ASCII data and transferred to the AWI-O_2A Near-Realtime Database (NRT) (Koppe et al., 2015) to ensure open-access within the group. All measured data and additional auxiliary sensor data from the MarGate test field were automatically displayed in a real-time dashboard at AWI O2A (Dashboard, 2020) so that the sensor and the environmental situation could be monitored in near real time throughout the experiment. Using the O_2A REST API data has been transferred into the PANGAEA ingest format and prepared for long-term archiving and publication at PANGAEA from which the data can be accessed (PANGAEA, Database 2020). To gain as much information as possible on sensor behavior due to different handling like *in situ* cleaning and the calibration and cleaning procedure, the different sensors were individually treated (details in Supplementary Material Table A2). This information will be also used in a further publication on the evaluation of salinity measurements. Sensor 3 and sensor 4 were deployed over the entire experiment without any *in situ* cleaning or maintenance, sensor 1 was regularly cleaned under water with a soft tissue, sensor 2 was recovered for lab calibration in the very beginning of the experiment, sensor 5 was recovered for lab calibration

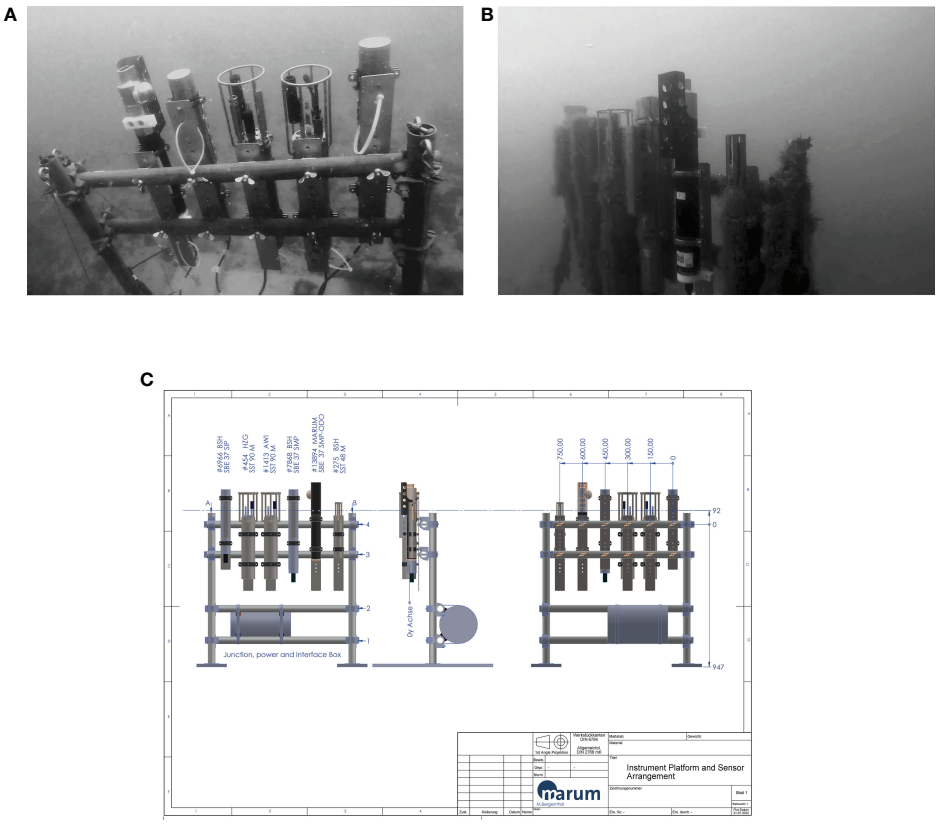

FIGURE 1
In situ images (upper panels **A** and **B**) and CAD drawing (lower panel **C**) with dimensions of the experimental setup of the deployed sensors. Upper left panel: Sensors immediately after deployment, Upper right panel: Sensors three month after deployment with one sensor (second from the right side) freshly lab-calibrated and maintained. For details see text.

after about 1 month and sensor 6 was recovered for laboratory calibration in the middle of the experiment.

As it was the objective of this experiment to develop a common procedure to calculate the uncertainty of sensor measurements, specifics of the design and age of the probes were not considered. For each sensor the maximal sample frequency for data collection has been configured.

Calibration procedure

The calibrations were performed in the calibration laboratory of the Leibniz Institute for Baltic Sea Research (the used instruments are described in appendix 2). This laboratory has been in operation for more than 50 years and it received an accreditation according to the ISO/IEC 17025 (ISO, 2022) by the DAkkS (Accreditation, 2022), the national accreditation body of Germany. The laboratory is accredited for the measurands temperature, pressure, and electrical conductivity.

The calibration of all devices was done for the measurands temperature and electrical conductivity. The temperature calibration is a comparison measurement with Standard Platinum Thermometers (SPRT) in a water bath. It is based on the International Temperature Scale (ITS-90) (Preston-Thomas, 1990) and traceable to the International System of Units (SI)-system.

The temperature probes are calibrated in a bath containing a volume of 80 l of seawater. The bath consists of two compartments: a main volume inside, where the calibration device is mounted and a second volume outside, where a hose that is connected to an external thermostat is installed. The water of the outer volume is pumped through a heating unit for the stabilization of the temperature into the inner volume by means of a vaporizing unit for a uniform distribution. In the inner volume a thermistor sensor is mounted and connected to an external control unit. Three Standard Platinum Thermometers (SPRT) are mounted in the vicinity of the sensors of the device to be calibrated.

The basic calibration schemes that are applied in the ocean science community are very similar. For temperature calibration Negative Temperature Coefficient (NTC) thermistor sensors are often used as the temperature reference. They are very stable and not as sensitive against mechanical stress as SPRTs. The advantage of the SPRTs is that the temperature between the fixed points is defined by the temperature resistance relation according to ITS-90 (Preston-Thomas, 1990). The NTC-sensors must be calibrated in a comparison measurement with SPRTs. This is an additional source of uncertainty which contradicts the asserted lower uncertainties of calibrations with NTC-sensors vs. SPRTs a claim that is hard to explain.

A big difference from most other calibration laboratories is the accreditation correspondent to the ISO/IEC 17025 (ISO, 2022) standard. That means that the process of calibrations and the traceability of the results and the declared uncertainties are ensured. Due to stricter criteria the ascertained uncertainties are typically bigger than in many other unaccredited calibration laboratories.

Uncertainty analysis

Today's ocean sensors typically provide an electrical signal (usually an AC or DC voltage, a frequency or a digital value) representing an ocean parameter measured at a specified time and location. Depending on the type of sensor, this signal is converted into the numerical value of the parameter (e.g., a temperature value) within the sensor with calibration coefficients stored within the instrument or it is necessary to calculate the numerical value subsequently using a calibration file.

The values of the acquired signal are not only a result of the ocean parameter being measured, but it is also influenced by additional, external effects. It is affected by inevitable instabilities and inhomogeneities of the water body near the sensor during the acquisition of the signal. Furthermore, it is affected by the technical properties of the sensor, for instance, by the signal noise of the instrument or by sensor drift or bias. The latter are meant to be determined by a calibration measurement. However, the calibration measurement is likewise affected by the effects mentioned above. Consequently, even under excellent measurement conditions, using calibrated sensors and excluding any instrument or other failures, it is impossible to know to what degree the signal value deviates from the value that truly represent the ocean parameter being measured. Hence, the true value of the sensor signal is uncertain and a method to quantify the uncertainty must be defined to estimate a range around the signal value in which the true is lying with a specified probability. Figure 2 illustrates the complex input of information affecting measured raw data points.

In the following sections, we will propose and evaluate a method to quantify the uncertainty of a measured ocean parameter in a relatively simple and practical manner. To this end, we will use seawater temperature measurements measured with the sensors and measurement setup described in the section 2. Firstly, we will show and discuss the measured data. Then, we will demonstrate how to quantify the uncertainties of the results of an individual sensor and discuss the meaning of the uncertainties. Afterwards, we will compare the results of several sensors. Based on an evaluation of the uncertainties, we will finally discuss to what extent the result of a single sensor is a good representative for the parameter of interest, compared to a multi-sensor measurement.

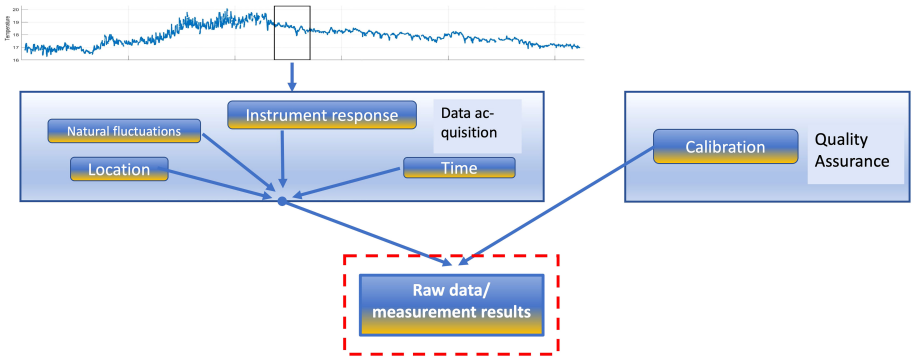

FIGURE 2
Schematic of information (blue) flow and their corresponding uncertainty (yellow) defining a single field measurement. The top part shows the environmental signal which is intended to be measured. The data acquisition box shows the different information and their uncertainty a measurement is facing: natural fluctuations of the environmental signal between the selected sampling interval, the sensor response capabilities, related measurement of time facing clock drift or data storage delays, uncertainty of location because of movement within the water column or if bottom mounted, change of water depth due to tidal effects. The quality assurance box is representing the information an uncertainty from the sensor calibration as detected in the laboratory under stable and well-known conditions.

4.1 Temperature data analysis

Every environmental measurement shows a certain degree of variability that cannot be assigned to any known process and therefore can be seen as a purely statistical phenomena that reflects the continuous transition of the water body into a new state. To estimate this type of variability and derive from that a contribution to the uncertainty of the temperature measurements, data sets from six sensors were evaluated within specific measurement periods. Figure 3 shows the measured temperature data over the entire period. Total data availability of the individual sensors varies between 69.7% and 100% (see Supplementary Materials, Appendix 1, Table A3) The different data availability is due to the temporal removal (for calibration and functional testing) of individual sensors and short telemetry interruptions. However, the data availability is acceptable and sufficient for an evaluation.

The study uses data collected using the O2A NRT data infrastructure which has been published at PANGAEA as described above. The data sets are available in the form of measured values with a maximum sampling frequency depending on the individual sensor configuration (see Appendix 1, Table A1). Sampling intervals vary between 1 s for sensors 1-3 and 10 s for sensors 4-6.

From the entire data set (covering almost four months), selected periods were chosen for further evaluation. The selection of time periods includes the beginning and end of fieldwork as well as periods of high and low variability during

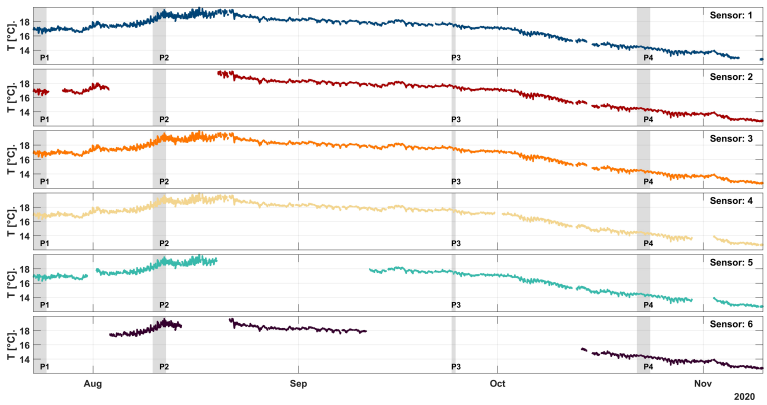

FIGURE 3
Measured temperature data of the six sensors for the whole measurement period (June 22 – Nov 10). Grey shaded areas indicate selected analysis periods. Unlabeled data gaps are due to different download speeds or caused by telemetry interruption.

measurement. Figure 3 shows the full data set and the selected time periods (numbered grey shaded areas P1-P4).

Further information on data availability and number of measurements of the periods are summarized in Table A2 in the Supplementary Material.

The area where the measurements were conducted is located in the south-eastern part of the North Sea. The prevailing ocean condition in this region is mainly influenced by tidal and wind-driven circulation systems as well as the atmospheric boundary. In general, a tidally well-mixed water mass can be expected, characterised by a typical atmospheric annual cycle. Maximum and minimum water temperatures range from 2-20°C over the year.

For this study, temperature measurements were collected over a period of four months. In all individual time series, the characteristic seasonal variation in temperature for the region can be observed (see Figure 3). Since measurements were only taken in summer and autumn, the minimum and maximum temperatures are in the typical range of 12 -20°C. During the summer months, the variability of the results are typically slightly increased, as stronger spatial and temporal temperature fluctuations (heat exchange with atmosphere (diurnal cycle, induced by solar irradiation, variations in the surface layer processes) can appear. In addition, the measurements will be affected by the increased marine fouling (biofouling) during the summer and autumn months. During the autumn months the variability decreased but was more strongly influenced by other environmental factors such as wind and the resulting waves.

Four representative periods from the complete time series were chosen for the determination of the statistical parameters. The rationale behind this is to consider different scenarios to obtain a complete picture of different phases (during a long-term measurement) of the data collection.

- The first selected period P1 is at the beginning of the measurement campaign. The sensors are freshly calibrated and clean (no marine fouling). In addition, the temperature curve shows relatively stable conditions with only minor seasonal fluctuations.
- The second period P2 is in the summer months (August) with relatively strong temperature fluctuations (diurnal cycle). The seasonal effect is also clearly visible (constant temperature increase in the summer months). In addition, the sensors have been in operation for a month, so alterations of the sensors (e.g., sensor drifting) and biofouling effects can have an impact on the data recording.
- The third period P3 had, with very low variability and high data availability, low external influences and stable temperature conditions over the entire measurement period. This provides the possibility to assess calibration uncertainties (in situ) as the data are (nearly) not dominated/influenced by external conditions.
- The fourth period P4 close to the end of the measurements in the autumn months has fairly steady temperature conditions, but high biofouling activity (autumn bloom). Moreover, individual sensors have already been replaced, cleaned or recalibrated.

Statistics of the selected study periods are calculated for an averaging interval of 5 min, with interval size of 300 seconds always starting at the full minute of the interval. The chosen averaging interval correspond to often found measurement intervals in common coastal observing programs and campaigns, but can also be easily adapted to other intervals as needed. Figure 4 shows an example of the results for the 5 min time average (T_{mean}) of one of the sensors.

As already mentioned, the choice of the averaging interval used is individually selectable, but should be adapted to the measurement environment or the measurement objectives. Especially for measurements at sea, there are some limitations in the area of energy and data storage possibilities as well as accessibility and maintenance options. Thus, the scientific focus (highest possible temporal resolution) cannot always be fully addressed, as the mentioned constraints must also be taken into account. An interval of 5 min was chosen for the calculations of variability and measurement uncertainty in this study. This selection based on the intention to resolve prevailing environmental conditions (e.g., tidal influences) of the measuring region in the data. Furthermore, there were no restrictions on the energy supply as the cabled infrastructure of the Helgoland Underwater Observatory (MarGate) was used, so there was relative flexibility in the choice of measurement acquisition settings.

To get a more definite estimate of the variability and uncertainty of the different temperature measurements, statistical parameters (standard deviation and standard error of the sample mean) of each single sensor are determined. The standard deviation (STD) is derived from the temporal variations of the temperature signal and indicates the dispersion of the individual data samples relative to the sample mean over that selected time period. In contrast, the standard error of the mean (SEM) is a measure of the dispersion of the sample mean (further details in the following section). The SEM depends on both the STD and the sample size (N) through the relatively simple relationship

$$\mathbf{SEM} = \frac{\mathbf{STD}}{\sqrt{\mathbf{N}}} \qquad \text{(eq. 1)}$$

and is therefore always smaller than the STD. The SEM is therefore an indicator of the variability of the temperature samples within the period and commonly used to indicate the

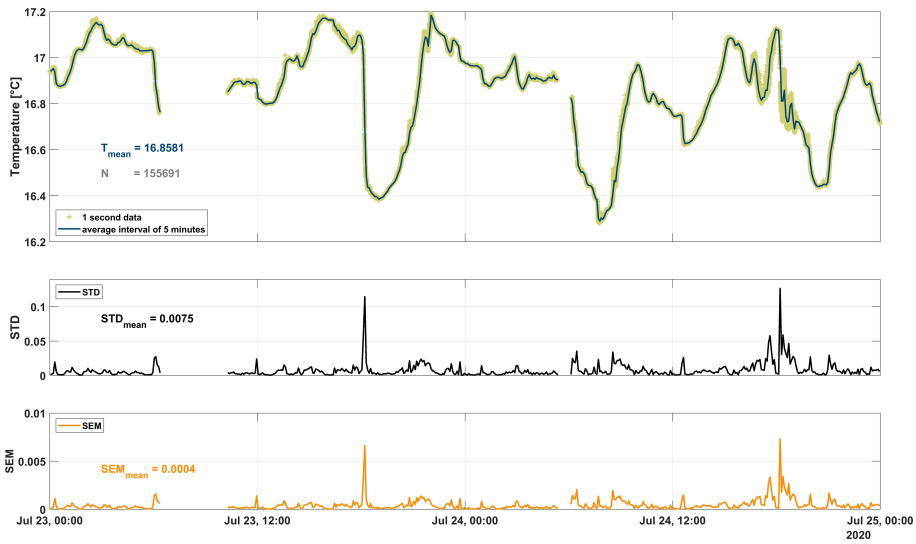

FIGURE 4
Example of measured temperature data from a single sensor (Sensor 1) for period P1. The upper panel shows the temperature data (grey markers indicate the raw data and the blue line indicates the averaged 5 min data sets. The lower panels show the STD (black line) and the SEM (orange line) for the corresponding 5 min averaging intervals). Numbers in the colours of the respective lines indicate the values of the averaged statistical parameters for the complete 5 min intervals.

uncertainty (Lee et al., 2015) (Altmann, 2005).

Results for T_{mean}, STD and SEM of all sensors and periods are summarised in Table A4 in the appendix. For all individual sensors and periods, the results are comparable and in the same range, there are no clear or obvious deviations. The variations are also rather small and in the normal measuring range. As expected, the highest variabilities are observed in the second period and the lowest during the last period. The mean STD for the 5 min mean values of the second period is about five times larger than in the last period, which confirmed the increased variation in the measurement of this second period. In contrast, the variability in the first period is only half that in the second period. Accordingly, the calculated uncertainties (SEM) are also highest in the second period, while in the other periods the uncertainties are lower with lowest values in the third period. In summary, the SEM values are all within a tolerable range and are comparable across all sensors and periods. The values in the tables (see the Supplementary Materials, Appendix 1, Tables A3 and A4) are only the average values for the selected (in this case five minutes) time interval. As shown in Figure 4, the values in the selected interval can vary greatly in variability and uncertainty. This should always be taken into account when particularly temporal fine-scale measurements are necessary or required.

The results also show the influence of the applied size of the sampling interval. Three (of the six) sensors have a longer sampling interval, which results in a larger uncertainty because of the factor $1/\sqrt{N}$ for the calculation of SEM. The difference is low but can be clearly seen. As mentioned before, the selected sampling interval depends also on various boundary conditions (sometimes it is not possible to run a shorter sampling interval due to limitations of the measurement set-up or insufficient energy supply) and measurement targets. The influence on the results of the average values (T_{mean}) is rather insignificant. The results show a good correlation in this case. Again, the choice is up to the user and the specific measurement task.

To look more closely at quantifying measurement uncertainty, in the next subsection we will focus on contribution to the calibration uncertainty, and the uncertainty related to the fluctuations of the individual sensor outputs. Other systematic contributions are the instrument resolution/quantization error that amounts to 0.14 mK for sensors 1-3 while for sensors 4-6 that amounts to 0.03 mK which is negligibly small. The long-term stability that is below 5% of the systematic uncertainty budget is not considered significant for this study.

4.2 Quantification of the uncertainty of single sensor measurements for two 3 h intervals

Four out of the six sensors have been evaluated, since only these have measured temperatures in both selected periods. Figure 5 shows the results of temperature measurements of the four sensors in two three-hour periods. The results shown in the

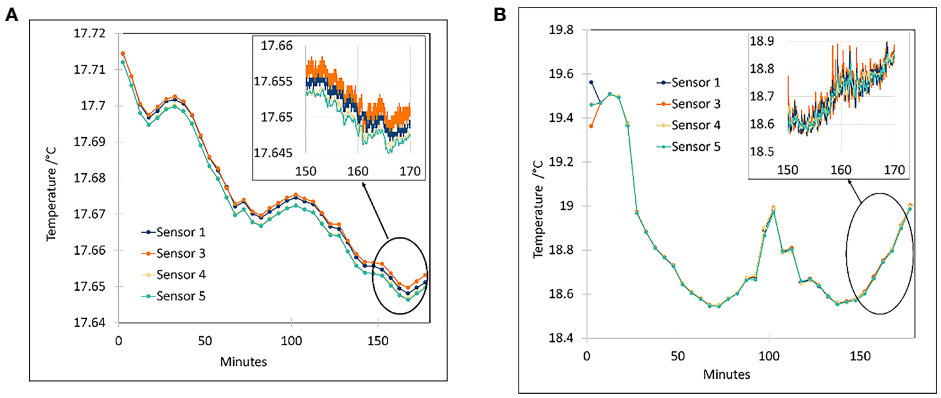

FIGURE 5
Time series of 5 min temperature means for P3 (dots, left panel **A**), while temperature variability was small and (a part of) P2 (right panel **B** with high variability). The insets magnify 20 min time windows and show the spreads of the raw data.

figure on the left-hand side, collected on 23/24 September 2020 between 22:40 and 01:40 (corresponding to time period P3), have been measured during rather stable environmental conditions. The overall change in temperature within that time interval amounts to about 60 mK. The figure shows the arithmetic means of 5 min intervals, indicated by the dots. The inset magnifies a representative 20 min period. There, the original raw data are shown without any separate averaging to illustrate the scattering of the original sensor signals. Two of the sensors had a sampling rate of 60/min or more (green and yellow lines), while the other two had a sampling rate of 6/min (blue and orange).

The results shown in the figure on the right, measured in the time between 11 August 2020 at 19:18 to 22:18 (corresponds to measurements in period P2), have been measured under highly variable environmental conditions. The variations in temperature amounts to almost one degree Celsius within period P2. Again, the inset shows the fluctuations of the unaveraged raw signals. The spread of the temperature signal is in the range of up to 100 mK, compared to a few mK during the calm period. Numerical values for the sensors are shown in the Supplementary Materials, Appendix 1, Table A4.

Based on the "Guide to the expression of uncertainty in measurement" (GUM, 2008) the combined uncertainty $u_c(T)$ of a temperature measurement result T can be calculated by combining the standard uncertainties of individual contributions, here:

$$\mathbf{u_c(T)} = \sqrt{\mathbf{u_{cal}^2 + u_{fluc}^2}} \qquad \text{(eq. 2)}$$

u_{cal} is the standard uncertainty assigned to the calibration and $u_{fluc}$is the standard uncertainty attributed to the variability during the measurement. The standard uncertainty indicates a range ± around the best estimate of the measured parameter value, in which the true value is assumed with a probability around 68%. The expanded uncertainty indicates a respective 95% range, which is usually calculated by multiplying the standard uncertainty with a factor of 2 (see section 6 of (GUM, 2008) Hence,

$$\text{Measured value} = \text{best estimate} \pm \text{uncertainty (68 \%)}$$

$$\text{Measured value} = \text{best estimate} \pm 2 \cdot \text{uncertainty (95 \%)}$$

The numerical value of u_{cal} is provided in the calibration certificate of a sensor. As mentioned, u_{fluc} is the standard uncertainty assigned to the variability of the parameter, which corresponds to the fluctuation of the measured values. The numerical value of u_{fluc} depends on the chosen representation of the parameter, meaning on how the best estimate is determined. Here, we will consider two kinds of representation:

(i) Temperature, at a specific time, is estimated by a single measurement ("raw data")

(ii) Temperature, at a specific time, is estimated by the arithmetic mean of values measured in a 5 min interval around this point in time ("5 min means")

It must be noted that equation 2 is a rather simple, but practical approach, than can be expected to cover the major uncertainty contributions. However, depending on the scientific task, other contributions might become relevant. More details are given in (Bushnell, 2019).

(i) Temperature estimated by a single measurement ("raw data")

If, for whatever reason, the scientific evaluation of a measurement series requires use of the raw data rather than averaged values, the fluctuation uncertainty of a single raw data value must be estimated. Usually, it is determined by quantifying the spread of fluctuating data measured under stable measurement conditions. However, only data measured under

unstable conditions are usually available from environmental measurement series. Therefore, a time interval must be defined, in which the measurement conditions can roughly be considered as being approximately stable. This means, the standard deviation should not exceed the change of the parameter in that interval. An estimate for the change could be the difference of the moving average at the beginning and the end of the interval[1]. Then, assuming a normal distribution, fluctuation uncertainty can be estimated by the standard deviation of the data within this interval (equation 3). The interval should however be large enough to have a minimum number of values included (at least 10) to be statistically meaningful. Otherwise, a factor, a, has to be applied that is given by the student-t distribution (GUM, 2008). Thus, the fluctuation uncertainty of the i^{th} temperature value T_i is estimated by all n values T_{ij} in the chosen interval around the value T_i:

$$\mathbf{u_{fluc}}\ (\mathbf{T_i}) = \mathbf{a}\ \cdot \sqrt{\sum_{j=1}^{n} \frac{(\mathbf{T_{ij}} - \mathbf{T_i})^2}{\mathbf{n} - 1}} = \mathbf{a}\ \cdot\ \mathbf{STD_i} \quad \text{(eq. 3)}$$

For instance, a 2 min interval in our measurement series would include 12 data points for sensors 4, 5, and 6 (having a sample rate of 6/min), which involves a student-t factor of a=1.05 (≈1) for a 68% probability range (see Table G2 in (GUM, 2008)).

Obviously, the sampling rate of a measurement series must be sufficiently large so that a suitable interval can be defined. If the sampling rate is too small to catch the fluctuation of the measurement signal other ways must be found to estimate fluctuation uncertainty. In this case it may be quantified by independent experiments in the lab or simply based on the experience of the scientist evaluating the data [so called type B uncertainty (GUM, 2008)].

It must be emphasized that the time interval mentioned in this subsection is used to calculate an estimate for the fluctuation uncertainty of a single (raw) data point, which reflects the temperature variability seen in the insets of Figure 5. Likewise, the panel in the middle of Figure 4 shows the fluctuation uncertainty of single temperature points in the measurement period P1 (based on a 5 min interval). However, fluctuation uncertainty of single points must be distinguished from that of mean values. Fluctuation uncertainty of mean values, using a 5 min time interval as an example, will be discussed in the next subsection.

(ii) Temperature estimated by the arithmetic mean of values measured in a 5 min interval ("5 min means")

An estimate for fluctuation uncertainty can be calculated with the standard deviation of the mean of n values T_{ij} within the i^{th} 5-minute period:

$$\mathbf{u_{fluc}}\ (\mathbf{T_i})_{\mathbf{5min}} = \mathbf{a} \cdot \sqrt{\sum_{j}^{n}{}_{=1} \frac{(\mathbf{T_{ij}} - \mathbf{T_i})^2}{\mathbf{n} \cdot\ (\mathbf{n} - 1)}} = \mathbf{a} \cdot \mathbf{SEM} \quad \text{(eq. 4)}$$

Here T_i is the arithmetic mean over 5 min in the i^{th} time interval, with n = 30 for the sensors having a sample rate of 10 s, and n = 300 for sensors having a 1 s sample rate. $a \approx 1$ as $n \geq 10$. The fluctuation uncertainty of the mean is obviously smaller than that of a single result (see i) due to the additional factor $1/\sqrt{n}$. The fluctuation uncertainty of the mean is reflected by the smoother behavior seen in the main parts of Figure 5 and the smaller values illustrated in the lower panel of Figure 4.

Figure 6 on the left-hand side shows the combined, expanded uncertainties of four sensors during a period with low variability (P3). Their values are a few mK, they are largely constant over the complete measurement time and are dominated by calibration uncertainty as the temperature has low variability. The sensor corresponding to the orange results has a slightly larger calibration uncertainty. The dashed line indicates the expanded combined uncertainties of individual results, meaning representation (i) for comparison. Note that it corresponds only to those sensors with smaller calibration uncertainties. The uncertainty of the raw data is somewhat larger, so that averaging is also advantageous under low variability conditions.

The figure on right shows expanded uncertainties of the 5 min means during the period with high variability (P2). There, the dashed line indicates the calibration uncertainty of the sensors. The fluctuation uncertainty increases the expanded uncertainty by about a few hundred's Kelvin. It can also be seen that those sensors measuring with higher time resolution (sensor 1 & 3), have smaller fluctuation uncertainties compared to those with lower sample rates (sensor 4 & 5) because of the averaging, despite smaller fluctuations of the raw data of the latter (see inset of Figure 5). Hence, it seems that larger numbers of samples lead to less uncertainty in comparison to longer integration times of the other two sensors. The uncertainty of the individual results is not shown, since it is too large to be shown on that scale.

4.3 Data analysis based on multiple sensor measurements

Figure 7 shows exemplary spreads of the results of four sensors. Each result is the mean of a 5 min interval and is shown as a colored dot. The uncertainty bar of each result indicates its combined uncertainty as described in the previous section. The three groups seen in Figure 7 belong to 3 different, but subsequent 5 min intervals. Note that the results of each

1 A more sophisticated criteria would include a trend analysis. To this end, a linear regression would be applied in the defined interval. If the slope of the regression line is smaller than the expanded uncertainty of the slope, the parameter can be considered stable.

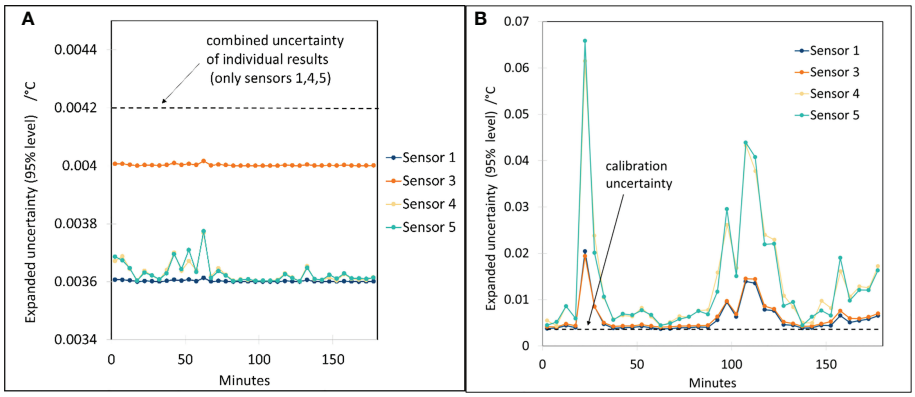

FIGURE 6
Combined uncertainties corresponding to the 5 min means shown in Figure 5. The dashed line in the figure on the left-hand side (panel **A**, low variability period P3), show the approximate combined uncertainties of the individual values (raw data) of sensors 1,4,5 (that of sensor 3 is slightly larger because of its larger calibration uncertainty). The dashed line in the figure on the right-hand side **(B)** approximates the calibration uncertainty of the sensors (high variability P2).

individual group have been measured in the same time interval, however, the values of a group have been slightly shifted along the x-axis to improve visibility of the individual results and their uncertainty bars. The red dot in the middle of each group indicates the average of the four sensor results in that group, simply calculated by the arithmetic mean of the four results. The corresponding uncertainty bar indicates +/- twice the standard deviation of the four results (95% confidence). It is an estimate of the uncertainty of the average in the respective interval. It should be noted that other estimators could be used for the average and its uncertainty (Maronna et al., 2006). The uncertainty weighted mean would for instance consider the uncertainties of the

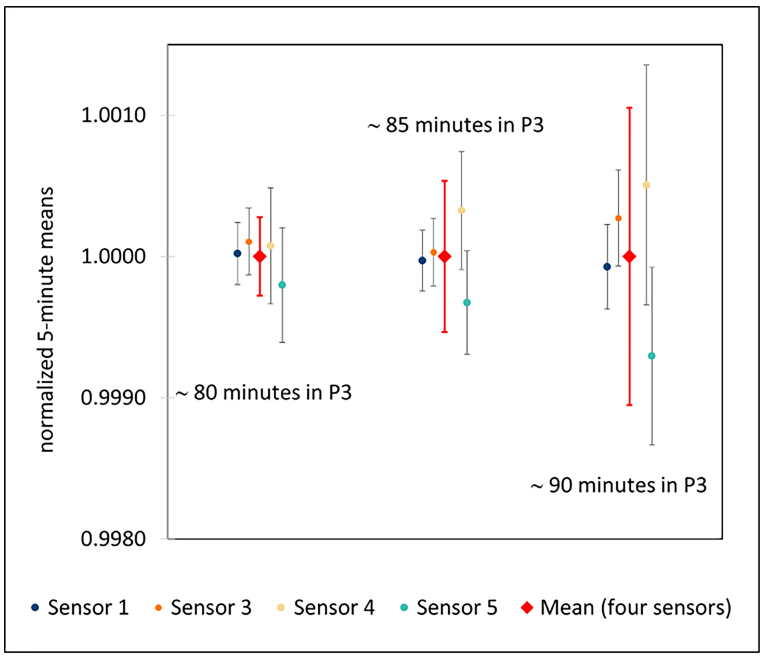

FIGURE 7
Magnified 5 min means corresponding to the Figure 6 (right hand side) with growing variable conditions, illustrating the spread of the best estimate derived from the four different sensors and their expanded uncertainties. The associated times have been slightly shifted to improve the presentation. The red dots indicate the arithmetic mean of all four sensors and the uncertainty bars their expanded uncertainties.

individual results. The median would be less sensitive to potential outliers.

A crucial element of this study is to assess the significance of the measurement result of an individual sensor and its uncertainty in comparison to a multi sensor approach.

The uncertainty will be quantified from all parallel measuring probes and then compared with the measurement uncertainties derived from a single probe. To make both results comparable 5 min averages were calculated and then the standard deviation over the 4 sensors were derived. With the same approach described above a factor based on the student t distribution has to be used to take the low sample number into account (a=1.20, assuming 3 degrees of freedom and a significance of 68%).

$$\mathbf{u_{fluc}}\left(\mathbf{T_i^{5min}}\right) = \mathbf{a}\cdot\sqrt{\sum_{j}^{n}{}_{=1}\frac{\left(\mathbf{T_{ij}}-\mathbf{T_i}\right)^2}{\mathbf{n}\cdot(\mathbf{n}-1)}} \qquad \text{(eq. 5)}$$

Where i is a specified moment in time and u_{fluc} is the value taken as the contribution to the uncertainty based on the variability of the measured parameter across all parallel measuring probes. As above, the combined (equation 2) and expanded uncertainty (equation 3) can be derived from the calibration uncertainty, often confused as the overall measuring uncertainty, and other influencing effects into account.

Small scale mixing process with a scale below the distance of the individual sensors between each other will cause a decorrelation between spatial and temporal variabilities. Those major differences between the sensors typically occur in region of strong temporal/spatial gradients as for instance the thermocline.

As one can see from the comparison between Figures 8 and 6 there appears to be a rather good match between both. The differences can be traced down to the processes that cause strong fluctuations and their related spatio-temporal correlation

Discussion

The focus of this study has been to what extent the measurement result of a single sensor together with the assigned uncertainty as calculated is representative for the observed parameter under consideration. For that purpose, parallel measuring probes had been used to be able to intercompare and judge on temperature measurement results of individual sensors, using the mean of the results of all available sensors as a reference. Only if a single sensor output is consistent with the mean of all sensor output and within the range of the calculated uncertainties can it be considered a reliable representative of the measured parameter. In that case the measurement uncertainty of the individual sensor is also quantifying the uncertainty range within which the consistency is valid. Mathematically, consistency can be expressed by comparing the deviation of the temperature result T_k of an individual senor k from the mean T_M of the results of all sensors with the uncertainty of the deviation (see Figures 7 and 8).

$$\left|\mathbf{T_{k\_i}}-\mathbf{T_{M\_i}}\right| \le \sqrt{\mathbf{u^2}\left(\mathbf{T_{k\_i}}\right)+\mathbf{u^2}\left(\mathbf{T_{M\_i}}\right)} \qquad \text{(eq. 6)}$$

The index i refers to the respective values of the i^{th} measurement interval. According to eq. 6, a temperature measured with sensor k is considered consistent with the mean temperature calculated from results from all sensors, if the deviation from the mean is smaller than its uncertainty. It should be noted that, strictly speaking, a statistically consistent

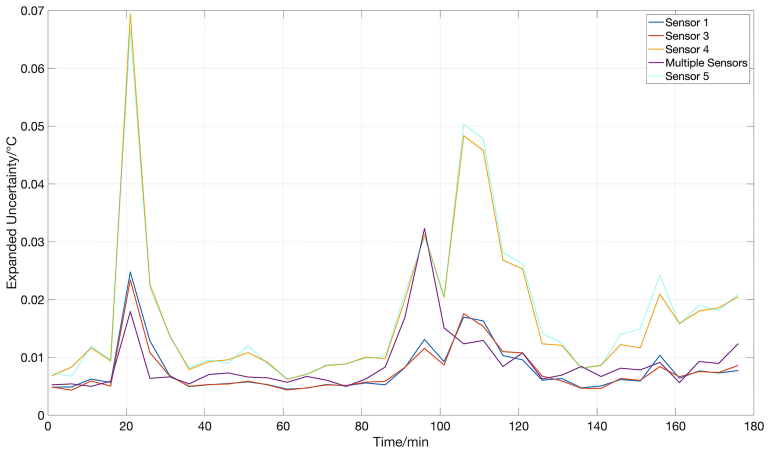

FIGURE 8
The standard uncertainty derived from the standard deviation between the 5 sensors within the same time interval as in Figure 6 on the right-hand side (Period 2, high variability). In violet is the graph for the multiple sensor uncertainty.

approach must consider the correlation between T_{k_i} and T_{M_i} and correlations between the results of the individual sensors, i.e., that all sensors have been calibrated by the same institute in the same way. Using other estimators, e.g., the weighted mean, as a representative for the average of several sensors, also effects the form of eq. 6. However, eq. 6 can be considered as an acceptable approximation for practical use. Intuitively, the uncertainty bars of an individual sensor result must overlap with that of the mean to some reasonable extent. For the sake of simplicity, we will discuss the representative character of a sensor result rather in terms of the overlap of uncertainties considering Figures 6–8. Basically, three cases should be distinguished which have different implications on the representativeness of an individual result.

a) The combined uncertainty value of a result is in the order of the calibration uncertainty. Hence, fluctuation uncertainty is small compared to calibration uncertainty, as can be seen in the left-hand side of Figure 6. Temperature can be assumed stable and homogenous in the vicinity of the sensors. In this case, the uncertainties of all sensors will overlap quite well, as can be seen in the cloud on the left-hand side of Figure 7 (around 80 min). Each sensor measures a good estimate for the mean temperature and the combined uncertainty of the sensor result is a good estimate for the uncertainty of the temperature in the considered time window and measurement volume. The deviation of an individual sensor (e.g., that of sensor 5) from the mean cannot necessarily be considered as systematic measurement bias that should be compensated. In fact, the deviation lies in the range of the calibration uncertainty. Correcting results for deviations that are smaller than calibration uncertainty, cannot be justified.

b) Both calibration and fluctuation contribute to the combined uncertainty of a sensor result in roughly the same order of magnitude. Temperature variation with respect to measurement time and special distribution must be assumed to some extent. The combined uncertainties of the individual sensors do still overlap fairly well with the uncertainty of the mean, as illustrated by the cloud on the right-hand side of Figure 7 (90 min). However, the overlap of some individual sensors with each other is marginal (see sensors 3, 4 and 5). Hence, the uncertainty u_c of an individual sensor does not well represent the actual mean temperature and its uncertainty. An uncertainty factor a_s reflecting the spread of several sensor results should therefore be included in the uncertainty of an individual sensor if only one sensor had been deployed (and temperature variability cannot be neglected):

$$\mathbf{u_{cs}} = \mathbf{a_s} \cdot \mathbf{u_c} \qquad \text{(eq. 7)}$$

u_c is calculated according to eq. 2 from the available measurement data. a_s is an estimated value reflecting the spread of the results of several sensors and u_{cs} denotes the enlarged uncertainty of an individual sensor. Obviously, assigning a number to a_s is somewhat arbitrary if only results of a single sensor are available, as is typical for oceanographic practice. However, GUM (section 4.3 of GUM, 2008) suggests evaluation of a so-called type B standard uncertainty that is based on the available information if repeated observations (here in the sense of several sensors) are not possible. Thus, looking at the results shown in Figure 7, the uncertainty bars of all sensors would reasonably overlap if they were about 50% larger. Therefore, setting a_s=1.5 is an arbitrary, but reasonable choice. If u_{fluc} is smaller than 0.5 u_{cal} its contribution to u_c becomes less relevant. The relative difference between u_{cal} and u_c is then less than 11%. Hence, it is also reasonable to set u u_{fluc}< 0.5 u_{cal} as a limit, below which it is reasonable to assume stable temporal and spatial conditions and, consequently, to set a_s=1 in that case.

c) Figure 8 compares the uncertainties assigned to the means of the sensor results with those of the individual sensors. There are measurement intervals in which the combined uncertainties of the individual sensors are in the order of several tens of mK. Hence, the corresponding fluctuation uncertainties are significantly larger than calibration uncertainties. Due to ongoing mixing processes significant instability in temporal and spatial temperature distribution must be assumed. While temporal averaging still provides a reasonable estimate of temperature and its uncertainty due to temporal variability at the exact position of a sensor, it cannot readily be assumed that the values are also adequate representatives for the entire time range of the measurement. Additional information is needed for instance by averaging the results of several sensors, potentially by weighing the individual results with their uncertainties (Maronna et al., 2006) and assigning uncertainties to the averages as mentioned above.

Cases a) and b) apply to measurement results where the uncertainties indicate no or moderate temperature variability. We propose, as a rule of thumb, that the combined uncertainty of a single sensor measurement can be considered as an adequate representative of the mean temperature in the specified measurement time (here, 5 min) and for the ambient water body next to the sensor within reasonable limits, if the fluctuation uncertainty of an individual sensor is not larger than two times the calibration uncertainty. If fluctuation uncertainty is larger, the uncertainty must be estimated using additional means. For instance, a multi sensor approach could provide reasonable uncertainties, also accounting for spatial inhomogeneity. If no further experimental data is available, the factor a_s can only be quantified based on the experience of the scientist evaluating the data.

A proposed flowchart for processing uncertainty information is presented in the Supplementary Materials, Appendix 3.

Uncertainty quantifications in the QA/QC framework

The sources of uncertainty for *in-situ* collected data are interwoven with instrument effects and the limited knowledge on processes that are influencing the variation of the observations that renders those effects as statistical. Mixing through the action of tides, insolation of surface waters and the related heat transfer, strong weather events, advection of water masses, biofouling, and instrument drift are the processes that determine the bias and fluctuation of the signal. Some of them can be quantified as uncertainties, other factors like biofouling are, practically speaking, impossible to quantify in a reliable manner. Another aspect has to be taken into account regarding the oceanographic assessment of the data before data are quality checked. Oceanographic assessment can thereby mean any influence on the handling/processing of the data due to the usage of other data for validation purposes and the accompanying uncertainty. These can typically be corrections of offsets, instrument drift, interpolations of missing values or reduction of prominent outliers as an additional step to the data processing procedure. By specifying the uncertainty, a statement can be made to what degree the observation had been influenced by systematic and transient, stochastic processes.

Specifying the uncertainty is complementary to data flagging so that uncertainties have to be added to the metadata description. Details on where the uncertainty quantification enters the picture of the standard QA/QC process is indicated in Figure 9.

Independently from the route of data processing, sensor signals are often collected under harsh environmental conditions producing erroneous data sometimes. Likewise, unexpected technical defects or other events might affect data integrity. Quality assurance (QA) measures are applied in the preparation stage of the measurement to improve the quality of the measurement system and therefore data quality. Quality control (QC) measures assess the usability of the measured data whether to use, discard or correct them. To this end, QC in terms of flagging criteria, and oceanographic assessment in terms low-level test of reasonableness and high-level process view are applied. It must be emphasized that QA/QC measures serve to avoid and identify unusable data and to minimize their uncertainty. However, the overall measurement uncertainty as the quantification of the doubt that remains of eventually accepted 'good' data is indispensable.

Typically, measurements at a specific location are used as representatives for much wider areas and results are averaged over time to reduce fluctuations in a time series or simply to reduce the amount of data to be handled. Thus, such results are used as representative for the ocean parameter under investigation over a defined space and over a defined time. Since, the ocean parameter can significantly vary over the defined area and over the averaging time, there is an uncertainty associated with these representative results, so called representation uncertainty, that depends obviously on the definition of the representative. Significant errors and apparent contradictions between representative results coming

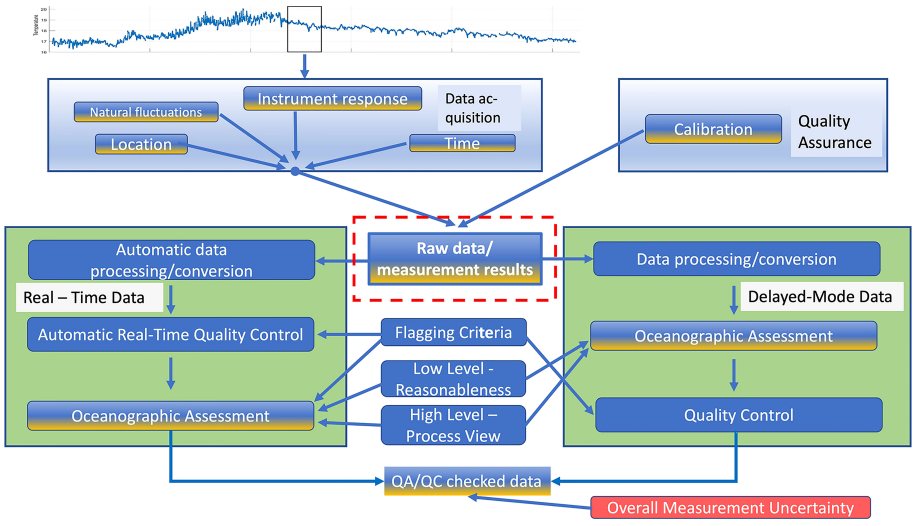

FIGURE 9
Schematic of information (blue) flow and their corresponding uncertainty (yellow) defining a single field measurement. The top part shows the environmental signal which is intended to be measured. The data acquisition box shows the different information and their uncertainty a measurement is facing: natural fluctuations of the environmental signal between the selected sampling interval, the sensor response capabilities, related measurement of time facing clock drift or data storage delays, uncertainty of location because of movement within the water column or if bottom mounted, change of water depth due to tidal effects. The quality assurance box is representing the information an uncertainty from the sensor calibration as detected in the laboratory under stable and well-known conditions.

from different sources can occur because of improper definition and inadequate uncertainty assignment.

Outlook

The presented approach to quantify uncertainties of measured EOVs in a practical manner has been presented on the basis of temperature measurements. The parameter can be directly measured using a single datum that can immediately be calibrated with a temperature standard. Moreover, temperature sensors show good long-term stability. The collected time series have shown that measurement results of all sensors are matching well even after some months. Other EOVs are however more challenging. Their numerical value has to be calculated from different parameters, all having their own uncertainties. Salinity, for instance, is derived from temperature, conductivity and pressure measurements. The respective calibration procedure is relatively elaborate and due to the fact that correlations between involved parameters exist, the uncertainty calculations are not straightforward. Moreover, the stability of the sensors is strongly affected by environmental effects, i.e., biofouling. In a subsequent paper the collected salinity data of the measurement series will discuss the effect of stability issues and multi-parameter measurements on the uncertainty quantification of measured ocean variables.

Other contributions to uncertainty quantifications have to be considered as well. One example would be the pressure sensitivity of temperature in profiling observations, such as shipboard CTD and Deep ARGO observations (Uchida et al., 2015)). Checking the time drift of temperature sensors in profiling float observations (Oka, 2005), would be another important topic, as profiling floats are not usually recovered and therefore post-observation calibrations for the temperature sensors are not possible.

Similar initiatives to enable the quantification of uncertainties had been started like the US CLIVAR Working group on Ocean Uncertainty Quantification (US-CLIVAR, 2020) and the International Quality-Controlled Ocean Database (Cowley, 2021). In the publication of Cowley et al. an additional aspect is mentioned that is described as the "Representativeness Errors". This aspect is putting the uncertainty assessment in the framework of what processes shall be observed and what type of fluctuations can be expected. Because here assumptions have to be made that are related to the used models this contribution to the uncertainty will probably change over time.

With this study a contribution to the UN Decade for Ocean Sciences shall be made. It will be a unique opportunity to use established platforms like the IODE Ocean Best Practice System (IODE, 2022) to disseminate the ideas and methods developed here. A close interaction with expert groups within WMO is already ongoing and will provide an additional impulse bridging the existing gap between ocean and meteorological practices.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: PANGAEA - https://doi.pangaea.de/10.1594/PANGAEA.942643.

Author contributions

CW, PF, and HB conceptualized the experiment, SS was overseeing and contributed to the systematic application of metrological principles, MK and J-GF did the oceanographic evaluation of the complete data set, MB designed the experimental set-up and evaluated the sensor performance during the lifetime of the study, SW carried out all calibration tasks. RH was contributing to the data management description. CW wrote an initial draft of this manuscript. All authors contributed to the article and approved the submitted version.

Funding

The project named DAUNE was carried out using existing funding resources of all involved institutions.

Acknowledgments

The authors very much acknowledge the support by the AWI Diving Team located on the island of Helgoland who deployed and safely retrieved all instruments, did the maintenance and repair and handled all sensor data. We also thank Kai Herklotz from BSH, Germany, for his support and valuable input.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmars.2022.1002153/full#supplementary-material

References

Accreditation (2022) *Deutsche akkreditierungsstelle*. Available at: https://www.dakks.de/de/home.html (Accessed March 2022).

Altmann, D. B. J. (2005). Standard deviation and standasd errors. *BMJ* 331, 903. doi: 10.1136/bmj.331.7521.903

BIPM (2008) *International vocabulary of metrology - basic and general concepts and associated terms*. Available at: https://www.bipm.org/documents/20126/2071204/JCGM_200_2012.pdf/f0e1ad45-d337-bbeb-53a615fe649d0ff1?version=1.15&t=1641292389029&download=true (Accessed March 2022).

Bushnell, M. E. A. (2019). Quality assurance of oceanographic observations: Standards and guidance adopted by an international partnership. *Front. Mar. Sci.* 6. doi: 10.3389/fmars.2019.00706

Bushnell, M. (2020) *QARTOD data flagging*. Available at: https://ioos.noaa.gov/project/qartod/ (Accessed March 2022).

Copernicus (2020) *Copernicus In situ TAC, real time quality control for WAVES CMEMS-INS-WAVES-RTQC* (Accessed March 2022).

Cowley, K. (2021). International quality-controlled ocean database (IQuOD) v0.1: The temperature uncertainty specification. *Front. Mar. Sci.* 11. doi: 10.3389/fmars.2021.689695

Dashboard, A. N. (2020). Available at: https://dashboard.awi.de/?dashboard=3783&resolution=HOUR&from=2020-07-20%252000:00&to=2020-11-25%252018:00).

Fischer, P. (2019). "Intelligent sensor technology: A 'Must-have' for next-century marine science," In F. Kirchner, S Straube, D Kühn and N Hoyer *AI Technology for underwater robots* (Springer Nature Switzerland AG 2020, Springer, Cham), p 19–36. doi: 10.1007/978-3-030-30683-0

GOOS (2020) *EOVs*. Available at: https://www.goosocean.org/index.php?option=com_content&view=article&id=14&Itemid=114.

GUM (2008) *Guide to the expression of uncertainty of measurements*. Available at: https://www.bipm.org/documents/20126/2071204/JCGM_100_2008_E.pdf/cb0ef43f-baa5-11cf-3f85-4dcd86f77bd6?version=1.9&t=1641292658931&download=true (Accessed March 2022).

IODE (2022). Available at: https://www.oceanbestpractices.org/about/ocean-best-practices-system/.

ISO (2022) *Wikipedia*. Available at: https://www.iso.org/standard/66912.html (Accessed 2022).

Jerico-Next (2019). Available at: https://www.jerico-ri.eu/previous-project/jerico-next/.

Koppe, R., Gerchow, P., Macario, A., Haas, A., Schafer-Neth, C. A., and Pfeiffenberger, H. (2015). *O2A: A generic framework for enabling the flow of sensor observations to archives and publications* (Genova: IEEE OCEANS).

Lee, D., In, J., and Lee, S. (2015). Standard deviation and standard error of the mean. *Korean J. Anesthesiology* 68 (3), 220–223. doi: 10.4097/kjae.2015.68.3

Maronna, R., Martin, R., and Yohai, V. (2006). *Robust statistics: Theory and methods* (Chichester, England: John Wiley & Sons Ltd). doi: 10.1002/0470010940

OceanSITES (2020) *OceanSITES*. Available at: http://www.oceansites.org/docs/oceansites_data_format_reference_manual.pdf (Accessed March 2022).

Oka, E. (2005). Long-term sensor drift found in recovered argo profiling floats. *J. Oceanogr* 61, 775–781. doi: 10.1007/s10872-005-0083-6

PANGAEA, D. B (2020). Available at: https://doi.pangaea.de/10.1594/PANGAEA.942643.

Preston-Thomas, H. (1990). The international temperature scale of 1990. *Metrologia* 27, 3–10. doi: 10.1088/0026-1394/27/1/002

Registry, A. S. (2019). Available at: https://hdl.handle.net/10013/sensor.177e9789-d85f-4c68-b6e0-fd496078a6e7.

Sea&Sun (2022) *Seae&Sun technology*. Available at: https://www.sea-sun-tech.com (Accessed March 2022).

Seabird (2022). Available at: https://www.seabird.com (Accessed March 2022).

SeaDataNet (2010) *SeaDataNet data quality control*. Available at: https://www.seadatanet.org/content/download/596/file/SeaDataNet_QC_procedures_V2_%28May_2010%29.pdf (Accessed March 2022).

Swift (2010). Available at: https://www.go-ship.org/Manual/Swift_DataEval.pdf (Accessed September 2022).

Taylor, J. R. (1997). *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements, 2nd Edition* (Sausalito, University Science Books). s.l.

Uchida, E. A., Nakano, T., Tumba, J., Widiatmo, J. V., Yamazawa, K., Ozawa, S., et al. (2015). Deep ocean temperature measurement with an uncertainty of 0.7 mK. *J. Atmospheric Oceanic Technol.* 32, 2199–2210. doi: 10.1175/JTECH-D-15-0013.1

UN DECADE, U (2021) *UN Decade of ocean science /UNESCO*. Available at: https://en.unesco.org/ocean-decade.

UNESCO (2013). "IOcean data standards, Vol.3," in *Recommendation for a quality flag scheme for the exchange of oceanographic and marine meteorological data* (Brest France, Ifremer). s.l.

US-CLIVAR (2020). Available at: https://usclivar.org/working-groups/ocean-uncertainty-quantification-working-group.

Wehkamp, S., Walcher, C., and and Fischer, P. (2013). MarGate - the underwater in situ lab off helgoland. *Abstracts of the 3rd International Workshop. Research in Shallow Marine and Fresh Wate Systems* (Bremen: Berichte, MARUM – Zentrum für Marine Umweltwissenschaften, Fachbereich Geowissenschaften, Universitaet Bremen) 292, 134.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., and Baak, A. (2016). The FAIR guiding principles for scientific data management and stewardship. *Sci. Data - Nat*, 3 (1), 160018. doi: 10.1038/sdata.2016.18

WMO (2008) *WMO report no 8*. Available at: https://community.wmo.int/activity-areas/imop/wmo-no_8 (Accessed March 2022).

Wong, A., Keeley, R., Carval., T.Argo Data Management Team (2022). *Argo Quality Control Manual for CTD and Trajectory Data*. doi: 10.13155/33951