Determination of discrete sampling locations minimizing both the number of samples and the maximum interpolation error: Application to measurements of carbonate chemistry in surface ocean

V. Guglielmi, Touratier F., Goyet C.

Journal of Sea Research 191 (2023) 102336

Contents lists available at ScienceDirect

Journal of Sea Research

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

Determination of discrete sampling locations minimizing both the number of samples and the maximum interpolation error: Application to measurements of carbonate chemistry in surface ocean

Véronique Guglielmi [a,b], Franck Touratier [a,b], Catherine Goyet [a,b,*]

[a] Espace-Dev, UPVD, Perpignan, France
[b] Espace-Dev, Univ. Montpellier, UPVD, IRD, Montpellier, France

ARTICLE INFO

Keywords:
Underway measurements
Sampling strategy
Interpolation error

ABSTRACT

Over the past three decades, a variety of programs have conducted extensive measurements of ocean properties at fixed stations throughout the water column, as well as in the surface ocean via oceanographic ships and ships of opportunity. Ships of opportunity were particularly used to determine the air-sea CO_2 fluxes from automated measurements of sea-surface temperature, salinity, and CO_2 fugacity. These underway measurements, often recorded at a frequency of every minute, generate large data files that need to be quality controlled, stored and analyzed. For practical use these data are often binned by 1° latitude x 1° longitude. Unfortunately, by doing so, there is a consequential loss of accuracy for these data sets.

Here, using the original 2010 underway data sets of sea-surface temperature, sea-surface salinity, total alkalinity and total inorganic carbon, along the cruise track from Hobart (Tasmania) to Dumont D'Urville (Antarctica), we show what would had been a more appropriate sampling strategy for each of these properties, maintaining their full measurement accuracy, while improving their interpolation accuracy. Furthermore, this analysis illustrates a general methodology for objectively determining, under suitable conditions, the appropriate locations for each property measurement according to a required accuracy. These results should greatly facilitate future cruise preparation and reduce the cost of measurements, while improving their scientific value.

1. Introduction

In the actual context of global warming and increasing anthropogenic carbon dioxide into the atmosphere (Komhyr et al., 1989; Kirk et al., 1989; Keeling et al., 1996; Tans et al., 1996, Stephens et al., 2000; Hall et al., 2021; https://gml.noaa.gov/ccgg/trends/), there is a growing interest in quantifying the role of the ocean in the absorption of part of this atmospheric anthropogenic carbon (DeVries, 2014; Friedlingstein et al., 2020). Consequently, over a few decades, time-series stations and repeated transects of underway measurements (ICOS https://www.icos-cp.eu/; Dyfamed; http://www.obs-vlfr.fr/cd_rom_dmtt/sodyf_main.htm; HOT, BATS; https://scrippsco2.ucsd.edu/data/ seawater_carbon/ocean_time_series.html), were designed to quantify the penetration of anthropogenic carbon in the ocean.

In the surface ocean (from the air-sea interface down to the depth of the wintertime mixed layer), many processes (such as air-sea exchanges [heat, gases, nutrients, etc.], mixing of water masses [fresh waters from rivers, surface currents, etc.], and seasonal biological activity), are at play. Thus, it is extremely difficult to disentangle the anthropogenic signal from the natural variations of total CO_2 concentrations (C_T).

At present, the only way to attempt to quantify the penetration of anthropogenic carbon in the surface ocean and to determine CO_2 sink and source areas of the ocean, is to assume the ocean is in quasi-steady-state (with negligible ocean circulation variation) and to perform repeated underway measurements. Thus, with the automation of measuring systems (such as thermosalinographs for the Temperature and Salinity, or Infra-Red based instruments for CO_2 fugacity), it is possible to design programs based upon Ships of Opportunity (SOOP, https://community.wmo.int/ship-opportunity-programme), in addition to those based upon oceanographic research vessels.

In France, ships that supply the bases of the Terres Australes et Antarctique Française (TAAF), also provide an excellent opportunity to acquire such valuable data sets. Thus, several programs such as

* Corresponding author **at:** Espace-Dev, UPVD, Perpignan, France.
E-mail address: cgoyet@univ-perp.fr (C. Goyet).

https://doi.org/10.1016/j.seares.2023.102336
Received 27 September 2022; Received in revised form 3 January 2023; Accepted 3 January 2023
Available online 4 January 2023

SURVOSTRAL (https://www.legos.omp.eu/survostral), and MINERVE (https://campagnes.flotteoceanogra-phique.fr/series/128/fr/), were designed to perform measurements of sea-surface temperature (SST), salinity (SSS), CO_2 fugacity (fCO_2), total CO_2 (C_T), and total alkalinity (A_T) while the supply ship "L'Astrolabe" is underway between Hobart (Tasmania) and Dumont D'Urville (Terre Adélie, Antarctica).

Yet, after more than a few decades of sea-surface measurements, it is still very difficult to disentangle the anthropogenic signal from the natural signal. The seasonal and inter-annual variations are still large compared with the anthropogenic perturbations (Laika et al., 2009; Morrow and Kestenare, 2014; Brandon et al., 2022). Thus, it is essential to intensify both the frequency of the sampling cruises, and the accuracy of measured and interpolated SSS, SST, C_T and A_T properties.

In order to optimize the number of measurements during all cruises, sampling strategies are essentially designed via Observing System Simulation Experiments (OSSE). For instance, Valsala et al. (2021), and Ford (2021) performed such experiments based on dynamical and biogeochemically consistent systems to design optimal sampling strategies for surface ocean CO_2 properties on both global and regional scales.

Here, the objective of this work is to show how to use another method first presented by Davis and Goyet (2021), which is significantly different from OSSE, to determine an appropriate sea-surface sampling strategy adapted to each measurable property. The two main advantages of the design of a sampling strategy based on this method using novel mathematical relationships, are to minimize the number of data while increasing their interpolation accuracy, and to appropriately and precisely determine sample locations in high variability ocean areas.

2. Materials and methods

2.1. Data sets

Over the past few decades, the French Antarctic supply ship "*l'Astrolabe*" provided the opportunity to scientists to perform (mainly in the austral summer), sea-surface measurements and sampling from Hobart, Tasmania (43°S 147°E) to the French Antarctic base Dumont D'Urville (66°S, 140°E).

As part of the SURVOSTRAL program (https://www.legos.omp.eu/survostral), continuous underway temperature and salinity "surface" seawater (at around 5 m), were measured since 1993 from R/V "*l'Astrolabe*" via a thermosalinograph (TSG). The raw data were recorded every minute. These raw data were then corrected for any bias (compared with discrete sample measurements), and by a median filter (over ±12 **min**) to reduce the noise of the measurements. These raw and corrected data sets are freely available (https://sss.sedoo.fr; Alory et al., 2015). Below, we used the corrected data set.

Similarly, as part of the MINERVE program (https://campagnes.flotteoceanogra-phique.fr/series/128/fr/), designed to quantify the interannual variability of the CO_2 properties in the Southern Ocean south of Tasmania, total alkalinity (A_T) and total CO_2 (C_T) were also sampled and measured from R/V "*l'Astrolabe*". These data are freely available (https://data.ifremer.fr/SISMER).

For the purpose of this work, we are focusing only on the transect Hobart – Dumont D'Urville which occurred in February 19–23, 2010. The choice of this transect was randomly picked among the transects where Total alkalinity (A_T) and Total CO_2 (C_T) were measured.

The measurement accuracy of sea-surface salinity (SSS), during this 2010 cruise is estimated to be ±0.005 (Morrow and Kestenare, 2014). The measurement accuracy of sea-surface temperature (SST), is estimated to be ±0.001 °C (from the manufacturer). The measurement accuracy of total alkalinity (A_T) and total CO_2 (C_T) measurements are estimated to be ±3.5 $\mu mol.kg^{-1}$ and ± 2.7 $\mu mol.kg^{-1}$, respectively (similar to the accuracies of these measurements performed on board previous MINERVE cruises [Laika et al., 2009]).

Fig. 1 shows the result of the measurements of these four properties (SST, SSS, A_T, C_T) along the cruise track from Hobart (Tasmania) to Dumont D'Urville (Antarctica), in February 2010. These graphs clearly show the disparity in the frequency of the measurements. There are 7815 data points for SST and SSS (one every minute; $N = 7815$ for SST and SSS), while there are only 238 data points for A_T and C_T (due to the difficulty and time of measurements; $N = 238$ for A_T, C_T).

2.2. Method

Based upon the work of Davis and Goyet (2021), who showed for example, how to determine appropriate Total CO_2 (C_T) sampling

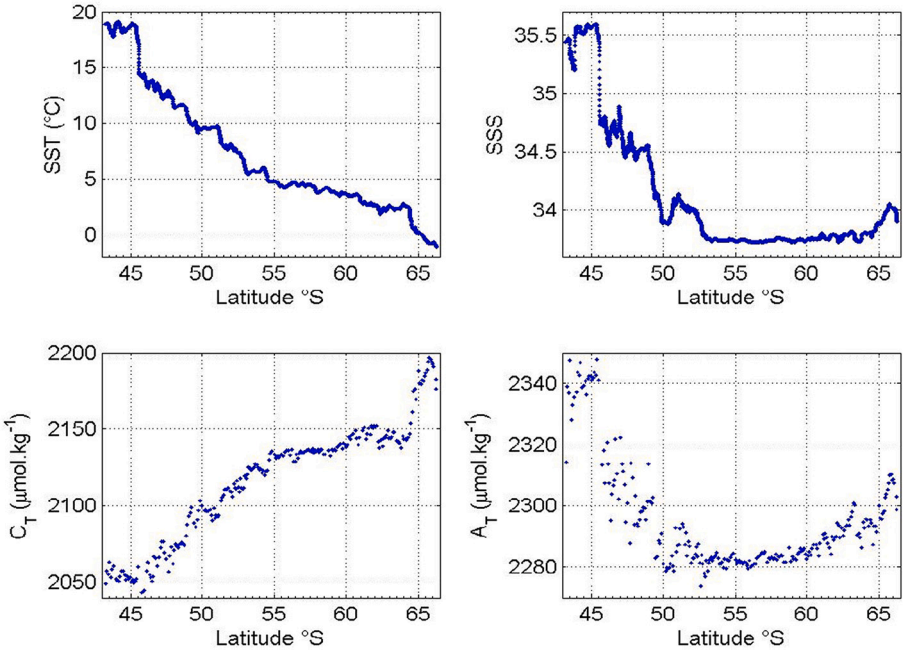


Fig. 1. Measured property as a function of latitude; a) SST and b) SSS measured every minute (7815 measurements), c) C_T and d) A_T measured roughly every 20**min** (238 measurements).

patterns throughout a water column from the surface to the bottom, we will use the same methodology to show how to determine appropriate sampling patterns for underway surface ocean measurements of SST, SSS, A_T and C_T.

In practice, Davis and Goyet (2021) demonstrated that the common intuitive notion, that where an environmental variable is highly variable it should be sampled more frequently than where it is less variable, has a sound basis in an analysis based on specific properties of the data itself. This analysis leads to a new, and practical, methodology for analyzing and improving the sampling of data.

The following is a very brief summary of the methods and results of this rigorous, analytic approach. For further details refer to (Davis and Goyet, 2021).

Assume that the set $\{(X_i, Y_i)\}$, $i = 1, \ldots,N$ represents a sample of a data field where the data signal Y is a function of the variable X, where X represents a position along some one-dimensional path through the data field. This path could be depths of a CTD cast, or surface sample positions along a ship track, or the track of an autonomous vehicle, etc. Assume X_{val} is the value of a position in the sample interval $[X_k, X_{k+1}]$, that is,

$X_k \le X_{val} \le X_{k+1}$, and let Y_{val} represent the value of the signal at the position X_{val}. Define the values Δ^+X_{val}, Δ^-X_{val}, Δ^+Y_{val}, and Δ^-Y_{val} by:

$$\Delta^-X_{val} = X_{val} - X_k \tag{1}$$

$$\Delta^+X_{val} = X_{k+1} - X_{val} \tag{2}$$

$$\Delta^-Y_{val} = Y_{val} - Y_k \tag{3}$$

and

$$\Delta^+Y_{val} = Y_{k+1} - Y_{val} \tag{4}$$

These values clearly represent differences in the spacing and the value of the data point (X_{val}, Y_{val}) to the data points in the sample data nearest to it.

Define three functions of the variable X as follows:

1) The ***sample error*** (or interpolation error), at the data point (X_{val}, Y_{val}) based on the sample data $\{(X_i, Y_i)\}$ is given by:

$$Err(X_{val}, Y_{val}, \{(X_i, Y_i)\}) = Y_{val} - L(X_{val}, \{(X_i, Y_i)\}) \tag{5}$$

where $L(X_{val}, \{(X_i, Y_i)\})$ is the Lagrange (linear) *interpolated value* of Y at X_{val}, given by:

$$L(X_{val}, \{(X_i, Y_i)\}) = (Y_k\, \Delta^+X_{val} + Y_{k+1}\, \Delta^-X_{val})/(\Delta^-X_{val} + \Delta^+X_{val}) \tag{6}$$

(Note that $\Delta^-X_{val} + \Delta^+X_{val} = X_{k+1} - X_k$, is the *length* of the sample interval $[X_k, X_{k+1}]$.)

2) The ***sample variability*** at the data point (X_{val}, Y_{val}) based on the sample data $\{(X_i,Y_i)\}$ is given by:

If Δ^-X_{val} or Δ^+X_{val} equals 0, $Var(X_{val}, Y_{val}, \{(X_i, Y_i)\}) = 0$, otherwise,

$$Var(X_{val}, Y_{val}, \{(X_i, Y_i)\})2 \bullet ([\Delta^+Y_{val}/\Delta^+X_{val}] - [\Delta^-Y_{val}/\Delta^-X_{val}])/(\Delta^+X_{val} + \Delta^-X_{val}) \tag{7}$$

The sample variability is a novel function, but it can be shown that the value of this function at any point (X_{val}, Y_{val}) equals the value of the *second derivative* of the signal function Y(X) at *some point* along the sample path and in the sample interval $[X_k, X_{k+1}]$, that contains X_{val}.

3) The ***sample spacing*** at the data position X_{val} based on the sample positions $\{X_i\}$ is given by:

$$Space(Xval, \{Xi\}) = \sqrt{\Delta^-X_{val} \bullet \Delta^+X_{val}} \tag{8}$$

Thus the sample spacing of a position X_{val} relative to a sample pattern equals the *geometric mean* of its distances from the closest positions of the pattern.

The spacing function is purely a function of *positions* along this sample path and does not involve the data values at these positions. The spacing function squared is a positive parabolic function which equals 0 at any sample position X_i. Its maximum value over any sample interval $[X_k, X_{k+1}]$ equals one half the length of this interval and occurs at the midpoint of the interval.

The first important result is called the Sample Error Theorem:

Sample Error Theorem – Given the above definitions, then:

$$Err(X_{val}, Y_{val}, \{(X_i, Y_i)\}) = -(Var(X_{val}, Y_{val}, \{(X_i, Y_i)\})/2) \bullet (Space(X_{val}, \{X_i\}))^2 \tag{9}$$

Since the values of $Var(X_{val}, Y_{val}, \{(X_i, Y_i)\})$ are *inherently* values of the second derivative of the function that describes the data along the path through the data field, the primary way to reduce the maximum error of any sample interval is to reduce the values of the sample spacing function, or in other words to decrease the spacing between sample points. This is the same as increasing the sampling frequency. It is also to be noted that according to the above result, the sample error varies with the *square* of the sample spacing. Thus, for example, if the sample pattern is evenly spaced, doubling the number of samples reduces the error to one fourth of its current error.

In fact, the Sample Error Theorem is a rigorous analytical relationship between sample error, sample variability and sample spacing that partially explains the common intuitive notion described above. It also provides a practical methodology for improving the sample (interpolation) error accuracy of a sample pattern, by adjusting the lengths and positions of sample intervals to reduce the maximum sample error in each interval.

Define $MaxAbsErr(\{(X_i, Y_i)\}, k)$ to be the *maximum absolute value* of $Err(X_{val},Y_{val},\{(X_i, Y_i)\})$ over all the data points (X_{val},Y_{val}) in the sample interval $[X_k,X_{k+1}]$ that contains X_{val}.

Assume that B(X) is a positive function such that for any sample data on the same path through the same data field, B(X) has the property that

$$|Var(X_{val}, Y_{val}, \{(X_i, Y_i)\})| \le B(X_{val}) \tag{10}$$

Then B(X) is called an ***absolute variability bound*** for sample data along this path.

Consider any sample pattern of positions $\{X_i\}$ along the same path and for each sample interval $[X_k, X_{k+1}]$ of this pattern define

$$MaxErrBnd(B, \{Xi\}, k) = max_{X_k \le X \le X_{k+1}} \left(\frac{\mathbf{B}(\mathbf{X})}{2} \bullet (Space(X, \{X_i\}))^2 \right) \tag{11}$$

By the Sample Error Theorem, it then follows that for any value k, or sample interval $[X_k,X_{k+1}]$,

$$|MaxAbsErr(\{(Xi, Yi)\}, k)| \le MaxErrBnd(B, \{Xi\}, k) \tag{12}$$

In summary, any positive variability bound B(X) for data samples along the path leads to bounds on the maximum absolute sample error of that data over each sample interval $[X_k,X_{k+1}]$.

A general goal of efficient sample pattern design is to find a pattern of a given size such that the *maximum* of the values $|MaxAbsErr(\{(Xi, Yi)\}, k)|$ on each sample interval $[X_k,X_{k+1}]$ is a *minimum*.

The following question then arises: Given a variability bounding function B(X), is there a sample pattern $\{X_i\}$ of a given size, with the property that the values $MaxErrBnd(B, \{X_i\}, k)$ for all k *are equal*, and thus define a *uniform sample error bound* for any sample data $\{X_i, Y_i\}$ using this sample pattern?

The basic idea is that if the MaxErrBnd function is *uniform* on each sample interval, it is near, or equal to, its *minimum maximum* value overall.

Given a bound on variability, a sample pattern $\{X_i\}$ for which the right side of the above inequality (12) is uniform is called a *balanced error pattern*. Such patterns always exist, but can be very difficult to

calculate algorithmically. However, there is a very efficient, practical, algorithm shown in (Davis and Goyet, **2021**) for calculating *semi-balanced error sample patterns* which have many of the same properties of balanced error patterns, and it is these efficient patterns which form the basis of the methodology to be illustrated here. Since the values MaxErrBnd(B, $\{X_i\}$, k) are equal for each interval $[X_k, X_{k+1}]$ of the sample pattern it can be denoted by the expression MaxErrBnd(B, $\{X_i\}$). That is,

$$\text{MaxErrBnd}(B, \{X_i\}) = \text{MaxErrBnd}(B, \{X_i\}, k) \text{ for all } k \tag{13}$$

(Note that an *evenly spaced* sample pattern is a balanced error sample pattern for a variability bounding function B(X) that is a *constant*.)

The second important result is that for a balanced error sample pattern, there is a relationship between the sample size and the maximum sample error bound.

2.2.1. Relation between maximum sample error bound and sample size (N)

If B(x) is any positive, non-zero, function and $\{X_i\}$ is a balanced error sample pattern of size N for B(X), then given the definitions above, it follows that

$$\text{MaxErrBnd}(B, \{Xi\}) \approx \left(\frac{1}{8 \bullet N^2}\right) \bullet \left(\int_{X_1}^{X_N} \sqrt{B(X)}\right)^2 \tag{14}$$

Or, equivalently,

$$N \approx \frac{\left(\int_{X_1}^{X_N} \sqrt{B(X)}\right)}{\sqrt{8 \bullet \text{MaxErrBnd}(B, \{ Xi \})}} + 1 \tag{15}$$

(These are estimates whose accuracy is shown in (Davis and Goyet, **2021**)).

Given a variability bound B(x), the sample positions of a semi-balanced error sampling pattern can be determined as follows (Davis and Goyet, **2021**):

First define the strictly increasing function $A(X) = \int_{X1}^{X} \sqrt{B(t)}\, dt$ in the interval $[X_1, X_N]$. Then the following function calculates the X positions of a sample pattern $\{X_i\}$ such that the Y values of A(X) are *regularly* spaced:

$$\text{Define } \textit{Distribute}(N, A(X)) = \{X_i\}\ i = 1, \ldots, N, X_1 \le X_i \le X_N\} \tag{16}$$

such that $A(X_{i+1}) - A(X_i) = (A(X_N) - A(X_1))/(N)$, with i = 1, …, N,where N represents the number of points to be distributed within the interval $[X_1, X_N]$.

The sample positions $\{X_i\}$ then define a *semi-balanced error* pattern of size N for the variability bound B(X).

Note that if B(X) is a *constant*, then $\{X_i\}$ is *evenly spaced*, and *balanced*. In this case only, Eqs. (14) and (15) are essentially exact (Davis and Goyet, **2021**).

Thus, the Eqs. (14) and (15) are valid estimates whenever the sampling pattern is even, balanced or semi-balanced, (Davis and Goyet, **2021**).

In summary, the basic steps of this methodology are:

1) **Estimate a positive bound on the absolute variability of the signal to be sampled from calculations on existing sample data.**

Estimates of variability can be calculated from any set of sample data $\{(X_i, Y_i)\}$ i = 1,2,..,N, by using *any three* sample data points (X_{k-1}, Y_{k-1}), (X_k, Y_k), (X_{k+1}, Y_{k+1}), for k = 2,3, …,N-1, in place of (X_i, Y_i), (X_{val}, Y_{val}), (X_{i+1}, Y_{i+1}) respectively, in the definition of Var(X_{val}, Y_{val}, $\{X_i, Y_i\}$). Then take the absolute values of these values. Do this for as many sample data sets as possible to get a good estimate for the absolute variability bounding function B(X).

2) **Determine the scientific requirements of the sampling methodology. What is the desired maximum sample error between sample points?**

Using B(X) from 1) and the above relation between a maximum sample error bound and sample size, calculate the number of samples N (Eq. (15)) required to achieve the desired maximum sample error.

3) **Once a bound B(X) is known and the required number of samples is known use the algorithm illustrated above (Eq.(16)) for calculating a semi-balanced error pattern of the required size to be used to sample the data. The resulting pattern is then known to approximate the desired maximum sample error.**

Note that a feature of this approach to sampling design is that it is *data driven*. In other words, everything is based on the specific properties of sample data along a path that has already been sampled, and for which some persistent variability bound has been determined from sampled data along this path, or a similar path.

Also it should be clear that this method is narrowly focused on exploiting the persistent variability properties of much environmental data. It is entirely different in concept and principle from other approaches to sample design such as OSSE. It is also very simple and can be used by anyone (a hand calculator can be sufficient to perform the calculations). In addition, it provides a precise knowledge of the maximum sample error (Eq. (14)) throughout the studied path.

3. Results for underway SST and SSS measurements

During the 2010 cruise, the SST and SSS properties were measured and recorded along with the ship's position (latitude and longitude) every minute. Since the ship was mainly sailing southward, and at a more or less constant speed, we will consider that the"X" axis is only the latitude (L).

Since the TSG instrument measurement error for the temperature and salinity can be estimated as ±0.001 °C and ± 0.005, respectively, we would desire an interpolation maximum error of half that of the measurements. Thus, we define MaxErrT = ± 0.0005 °C and MaxErrS = ± 0.0025, for SST and SSS, respectively. Consequently, the overall uncertainty of the interpolated data along the whole cruise track would be less than (0.001 + 0.0005) ± 0.0015 °C for temperature and (0.005 + 0.0025) ± 0.0075 for salinity.

In practice the temperature and salinity errors can be higher (Morrow and Kestenare, **2014**), if one takes into account the environmental errors in addition to the instrument errors. Thus, below, the results will be shown for the three desired MaxErrT (± 0.0005 °C; ± 0.005 °C; ± 0.05 °C) and three MaxErrS (± 0.0025; ± 0.005; ± 0.01).

3.1. Determination of the temperature and salinity variabilities and their bounds

Fig. 2 illustrates each SST absolute variability (VarT(L)) and SSS absolute variability (VarS(L)), respectively, as calculated according to Eq. (7), as well as their respective bounds (*Bnd*T(L) and *Bnd*S(L)). For instance here, the bounding function of temperature variability is the suite of straight lines (solid red lines in Fig. 2a) using the selected points shown below(*cpBnd*T(L); red stars in Fig. 2a):

*cpBnd*T(L) = {(43.2262, 16,600), (46, 16,600), (52, 10,545), (53.5, 4500), (63, 4500), (64.1, 6940), (66.2598, 6940)}.

(Note that these points were chosen to be close to, but always above the variability data values to form a kind of broad envelope over the variability values.)

Similarly, the bounding function of salinity variability is the suite of straight lines (solid red lines in Fig. 2b) using the selected points shown (*cpBnd*S(L); red stars in Fig. 2b):

*cpBnd*S(L) = {(43.2262, 3260), (44, 3260), (45.3, 1000), (45.5,

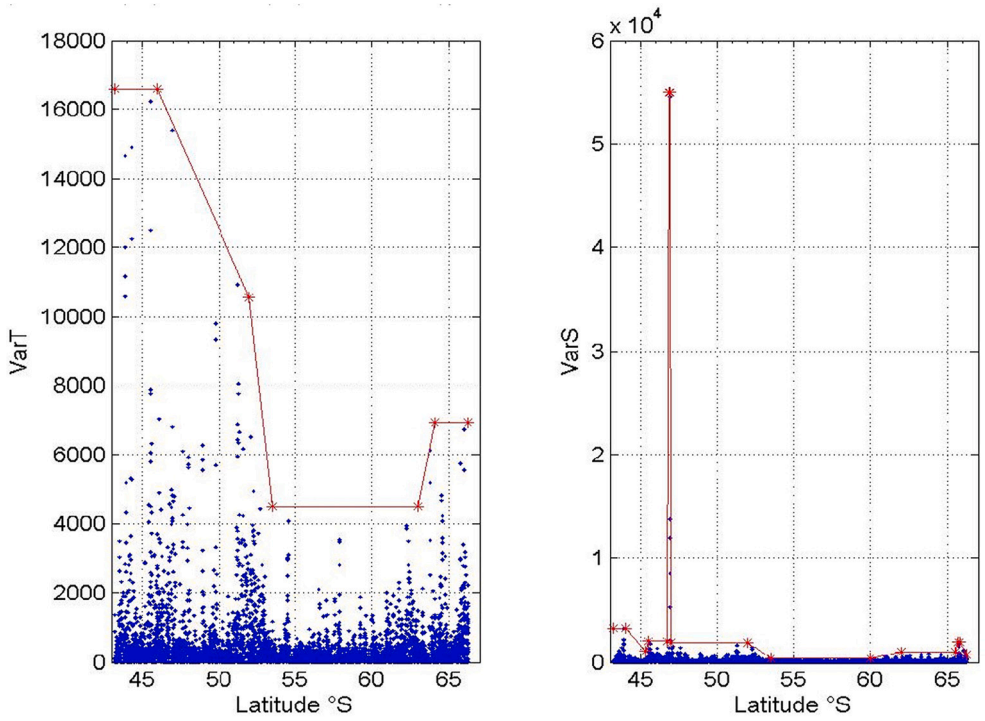


Fig. 2. Variability of a) sea-surface temperature, and b) sea-surface salinity, as a function of latitude (in decimal degree). The solid (red) lines on each graph represents the variability bounds. **(For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)**

2000), (46.8, 2000), (46.91, 55,000), (46.98, 5500), (47, 1800), (52, 1800), (53.5, 400), (60, 400), (62, 900), (65.5, 900), (65.7, 1900), (65.8, 1900), (66.2598, 700)}.

Fig. 2 shows that the temperature and salinity variability bounds have different shapes and thus, it may not be appropriate to measure them simultaneously. This observation is also in good agreement with the results based upon vertical (through a water column), temperature and salinity data, presented in Davis and Goyet (2021). Consequently, in order to minimize the number of measurements in the ocean (from the surface seawater throughout the bottom waters), while insuring the highest accuracy of each property, SST and SSS measurements should be performed at different locations.

For instance, here, as illustrated in Fig. 2, there is a large difference between the SST and SSS variabilities at latitudes near 47°S. There is a huge salinity variability while the temperature variability is almost constant. As shown in Fig. 3, these differences in variabilities reflect the differences in the SST and SSS signals.

Thus Fig. 3, which is a zoom of Fig. 1 within the latitude interval

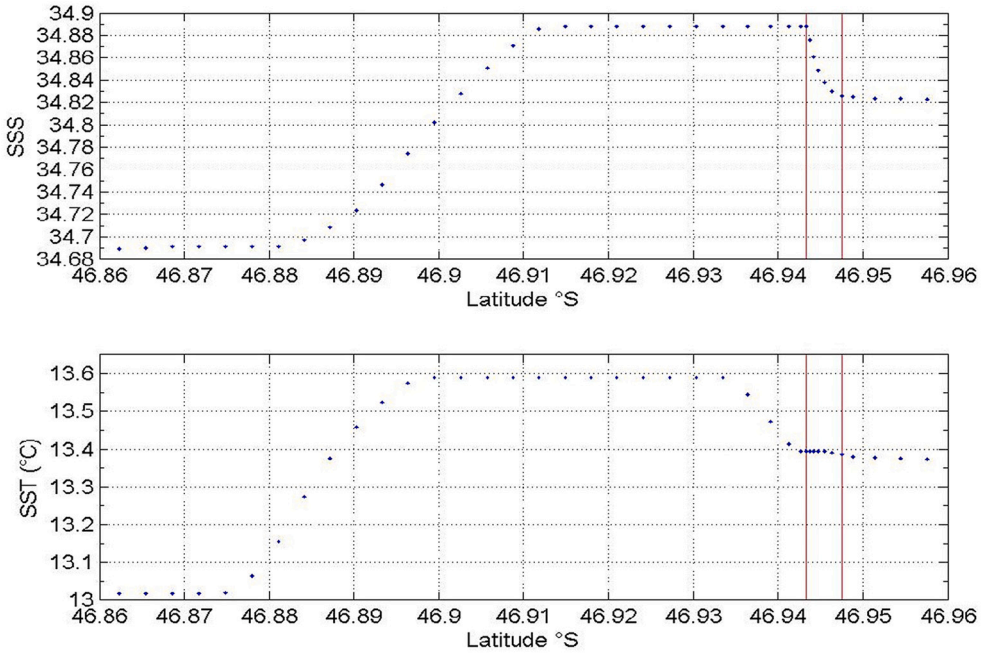


Fig. 3. Zoom of the SSS and SST data within the latitude interval [46,86°S; 46,96°S].

[46,86°S; 46,96°S], clearly shows that both the increase and the decrease in temperature and salinity occur at different latitudes, with a latitude shifted around 0.01° (about 1 km), and at a very different rate. Such features are typical of fine surface ocean structures. Thus, they are expected to occur within this ocean area, especially between the Sub-Tropical Front (STF), and the SubAntarctic Front (SAF) where there are many eddies, cold cores, and filaments with important SST and SSS small scale variations.

Inside such fine structures (below meso-scale), SSS and SST vary on different time scales. In general SST varies much faster than SSS due to air-sea interactions. This was clearly illustrated in a previous study by Morrow et al. (**2004**). They showed that the front signatures of the minimum of SSS and SST in cold core eddies that can last one to two months after the ring detachment, are significantly sharper for SSS than for SST.

In another time-series study, Morrow and Kestenare (**2014**) further illustrated the recurrent high SSS variability in this ocean area, in particular, near 47°S and South of the STF.

3.2. Determination of the temperature and salinity maximum interpolation errors

For temperature measurements along the cruise track from 43.2262°S to 66.2598°S, the result of Eq. (14) for an even sampling pattern is $\text{MaxErrBnd}(SST)_{\text{even}} = 0.018$ °C and that of the same Eq. (14) for a semi-balanced error sampling pattern is $\text{MaxErrBnd}(SST)_{\text{sb}} = 0.009$ °C. Thus, these results indicate that:

1) Either it is unnecessary to use a temperature probe as accurate as 0.001 °C since the interpolation accuracy cannot be better than 0.018 °C. (Thus, a temperature probe with an accuracy of 0.036 °C would suffice for an even sampling pattern for an overall (measurement + interpolation) accuracy of 0.054 °C, or a temperature probe with an accuracy of 0.018 °C would suffice for a balanced error sampling pattern for an overall (measurement + interpolation) accuracy of 0.027 °C.)
2) Or it is necessary to greatly increase the frequency of the measurements to keep the overall uncertainty below 0.0015 °C.
3) Or it may be appropriate to use a higher order interpolation (Davis and Goyet, 2021).

For salinity measurements along the cruise track from 43.2262°S to 66.2598°S, the result of Eq. (14) for an even sampling pattern is $\text{MaxErrBnd}(SSS)_{\text{even}} = 0.0597$ and that of Eq. (14) for a semi-balanced error sampling pattern is $\text{MaxErrBnd}(SSS)_{\text{sb}} = 0.0013$. Thus, these results indicate that:

1) Either it is unnecessary to use a salinity probe as accurate as 0.005 since the interpolation accuracy cannot be better than 0.0597. (A salinity probe with a measurement accuracy of 0.12 would suffice for an even sampling pattern for an overall (measurement + interpolation) accuracy of 0.18.)
2) Or it is necessary to increase the frequency of the measurements to keep the overall uncertainty below 0.0075.
3) Or it is necessary to use a semi-balanced error sampling strategy which will provide an interpolation error of only 0.0013 (below the desired maximum interpolation error of 0.0025), for an overall (measurement + interpolation) accuracy of 0.0063. (Thus, in this case, it would be possible to reduce the number of measurements performed.)
4) Or in order to further reduce the number of measurements it may be appropriate to use a higher order interpolation (Davis and Goyet, 2021).

In other words, these results indicate that for SST, a more efficient sampling pattern would be a semi-balanced error sampling pattern. Yet, it will be necessary to considerably increase the number of measurements to reach a desired maximum interpolation error below 0.005 °C. In any case, SST sampling would be regularly spaced along the latitudinal axis when the variability remains constant in the three latitude intervals [43.2262°S – 46°S], [53,5°S – 63°S], [64,1°S – 66,2598°S]. And sampling would be irregularly spaced along the latitudinal axis when the variability varies in the three intervals [46°S – 52°S], [52°S – 53,5°S], [63°S – 64,1°S].

Similarly, these results show that for SSS, a more efficient sampling pattern would be a semi-balanced error sampling pattern. But contrary to SST, it would be possible to reduce the number of measurements to reach the desired maximum interpolation error below 0.0025, if they were appropriately (irregularly) spread along the cruise track.

In order to determine more appropriate sampling patterns for SST and SSS, it is necessary to calculate the number of samples needed prior to determining their locations.

3.3. Determination of the number of samples needed to reach a desired accuracy using even and balanced error sampling patterns

Since it is appropriate to use an even sampling pattern in areas where the bounds of the variability signal are constant, and to use a balanced error sampling pattern in areas where the bounds of the variability signal varies, the calculated number of samples needed for SST and SSS along the cruise track (depending upon the desired accuracy), between Hobart and Dumont D'Urville can be calculated using Eq. (15). The results are summarized in Table 1 for measurements of SST and in Table 2 for measurements of SSS.

Remark: In these tables the number of samples calculated over the whole latitudinal interval (last line in the tables) are less than the sum of samples within the sub-intervals. This is due to rounding of the result (since a fraction of a sample would be meaningless), and of the limits (ex: one sample at 52°S would be counted twice; once in the interval [46°S; 52°S] and once in the interval [52°S; 53.5°S]).

These results (Table 1) indicate that the desired SST maximum interpolation error of 0.0005 °C is far from being reached with "only" 7815 measurements (quasi-evenly spaced), since it would require a minimum of 32,958 measurements (more than 4 times 7815 points) judiciously located along the cruise track. The target of a maximum interpolation error of 0.005 °C could not even be reached with the 7815 measurements (which represent a measurement every minute), since it would need at least 10,423 data.

On the other hand, if the maximum interpolation error needed were only 0.05 °C, then 7815 measurements would be more than twice too many since only 3297 measurements would suffice.

As expected, it is in the latitude interval [53.5°S; 63°S], where the SST variability is the lowest, that the number of samples measured (3109) is the closest (per degree of latitude), to the one calculated

Table 1
Numbers of samples needed within each latitudinal interval to reach the desired maximum interpolation error for SST measurements. The gray boxes indicate that the number of samples are calculated (Eq. (15)) for a semi-balanced sampling pattern. The white boxes indicate that the number of samples are calculated (Eq. (15)) for an even sampling pattern.

Latitudinal interval	N measured	N for a desired SST Maximum interpolation error		
		0.0005°C	0.005°C	0.05°C
43.2262°S - 46°S	905	5652	1788	566
46°S - 52°S	2015	11030	3489	1104
52°S - 53.5°S	566	2044	647	205
53.5°S - 63°S	3109	10077	3187	1009
63°S - 64.1°S	357	1314	416	132
64.1°S - 66.2598°S	863	2846	901	285
43.2262°S - 66.2598°S	7815	32958	10423	3297

Table 2
Numbers of samples needed within each latitudinal interval to reach the desired maximum interpolation error for SSS measurements. The gray boxes indicate that the number of samples are calculated (Eq. (15)) for a semi-balanced sampling pattern. The white boxes indicate that the number of samples are calculated (Eq. (15)) for an even sampling pattern.

Latitudinal interval	N measured	N for a desired SSS Maximum interpolation error		
		0.0025	0.005	0.01
43.2262°S - 44°S	253	313	222	157
44°S - 45.3°S	432	420	297	210
45.3°S - 45.5°S	60	56	40	28
45.5°S - 46.8°S	432	412	292	207
46.8°S - 46.91°S	36	126	90	64
46.91°S - 46.98°S	30	117	83	59
46.98°S - 47°S	7	24	17	12
47°S - 52°S	1670	1501	1062	751
52°S - 53.5°S	566	346	245	174
53.5°S - 60°S	2137	920	651	461
60°S - 62°S	646	359	254	180
62°S - 65.5°S	1155	743	526	372
65.5°S 65.7°S	71	54	38	27
65.7°S -65.8°S	63	32	23	16
65.8°S - 66.2598°S	257	117	83	59
43.2262°S - 66.2598°S	7815	5527	3908	2764

(3187) for an desired maximum interpolation error of 0.005 °C.

Table 2 illustrates that to reach the desired maximum interpolation error of 0.0025 for SSS, the number of measurements could be significantly reduced if they were performed at key locations rather than evenly spaced. Using a semi-balanced pattern strategy, only 5527 measurements would suffice while the 7815 measurements evenly spaced are not enough to reach the desired maximum interpolation error of 0.0025. Furthermore, if the desired maximum interpolation error is set to only 0.05, the number of SSS measurements needed would drop down to 3908 (a reduction by a factor close to 2 of the 7815 measurements).

Overall, these results indicate that currently, temperature and salinity data recorded every minute along a cruise track do not guarantee linear interpolation errors less than half that of these accurate measurements. They also emphasize the fact that the SSS and SST variabilities may significantly differ in time and space. Thus, it would be best if SSS and SST were measured at different rates to preserve their respective measurement accuracies. This would further avoid over sampling.

These results allow us to quantify the required frequency of the measurements as the expected accuracy decreases. Thus, it would be judicious to ensure that the objectives in terms of interpolation errors could be reached given the accuracy of the measuring systems.

In practice, it may not be possible (or desirable if scientists do not wish to study cold cores, eddies, or filaments), to considerably increase the frequency of measurements when there is a very sharp variation of the variability property (such as that observed for salinity near 47°S). In such case, it may be appropriate to choose to ignore this very high localized variability to determine reasonable variability bounds over the whole signal.

For example, assume the bounding function of salinity variability is the suite of straight lines (solid red lines in Fig. 4) using the following selected points (*cpBndS*(L); red stars in Fig. 4):

cp*BndS*(L) = {(43.2262, 3260), (44, 3260), (45.3, 1800), (52, 1800), (53.5, 400), (60, 400), (62, 900), (65.5, 900), (65.7, 1900), (65.8, 1900), (66.2598, 700)}.

With these new bounds, using Eq. (14), MaxErrBnd(SSS)$_{even}$ = 0.0354 and MaxErrBnd(SSS)$_{sb}$ = 0.0012. Compared with the ones above (MaxErrBnd(SSS)$_{even}$ = 0.0597 and MaxErrBnd(SSS)$_{sb}$ = 0.0013), these results illustrate the importance of the choice of a variability bound. The closer such a bound is to a variability signal, the lower is the number of samples needed to recover the full signal (with a minimum interpolation

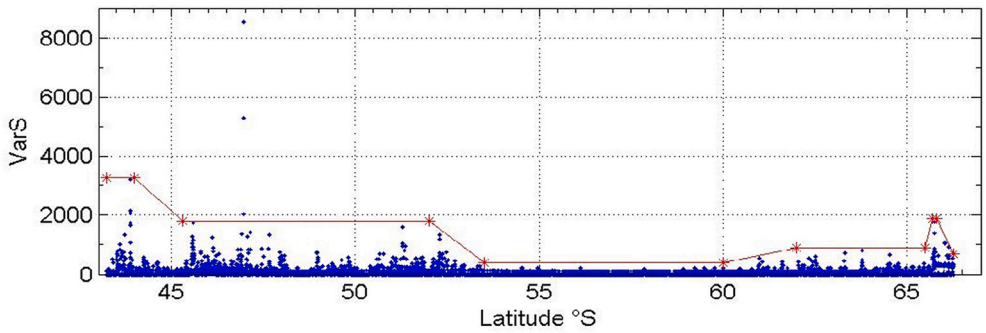


Fig. 4. Variability of sea-surface salinity as a function of latitude. The red lines represent the variability bounds if the highest SSS variability is ignored. **(For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)**

error). As expected, these results further show that the largest difference is with an even sampling pattern. Thus, there is always a significant advantage to using a semi-balanced error sampling pattern, to minimize the number of measurements while ensuring the lowest interpolation error.

Using these new bounds, the results of the sample size needed in each latitudinal interval are summarized below in Table 3.

Table 3 illustrates that without taking into account the high variability near 47°S, to reach the SSS desired maximum interpolation error of 0.0025 the number of measurements could be reduced to 5349. This represents a reduction of 178 measurements compared with the number of measurements needed (5527, Table 2) to take into account the full SSS variability.

Furthermore, as expected, if the desired maximum interpolation error is set to only 0.05, the number of SSS measurements needed would drop down to only 3783 (a reduction of only 125 measurements compared with result [3908] in Table 2). As the expected maximum interpolation error increases, the number of measurements decreases and this reduces the effect of increasing the variability bounds.

Then, knowing the number of measurements needed in each latitude interval, it is easy to determine their positions. In areas where the variability bound is constant, the measurements would be regularly spaced, while in areas where the variability bound is variable, the position of measurements would be simply determined by using Eq. (16).

Remark: here, since the function BndY(t) is a simple linear function (a•t + b), the integral of its square-root can be calculated exactly;

$\int_{t1}^{t2} \sqrt{(a \bullet t + b)}\, dt = (a \bullet t2 + b)^{3/2} \bullet 2/(3\bullet a) - (a \bullet t1 + b)^{3/2} \bullet 2/(3\bullet a)$. However, if the function BndY(t) were more complex, the integral could simply be done numerically.

Given the relatively high number of data points, a figure of the results of Eq. (16) would not help to visualize them. Thus, we choose to show the results of Eq. (16) only for the A_T and C_T sampling (below) which are considerably less numerous.

4. Results for underway A_T and C_T measurements

Since the accuracy of the total alkalinity and total CO_2 concentrations are 3.5 μmol.kg^{-1} and 2.7 μmol.kg^{-1}, respectively, for both properties, we would desire an interpolation accuracy of half the accuracy of the measurements. Thus, we define MaxErrA_T = 1.75 μmol.kg^{-1} and MaxErrC_T = 1.35 μmol.kg^{-1}.

4.1. Determination of the A_T and C_T variability bounds

Fig. 5 illustrates the variabilities and their bounds for total alkalinity and total CO_2 data.

Consequently, the bounding function of A_T variability is the suite of straight lines (solid red lines in Fig. 5a) using the following points (*cpBnd*A_T(L); red stars in Fig. 5a):

cpBndA_T(L) = {(43.25, 3900), (48.8, 3900), (52.8, 2800), (54, 1000), (63, 1000), (64.4, 1750), (66.25, 1750)},

and similarly, the bounding function of C_T variability is the suite of straight lines (solid red lines in Fig. 5b) using the following points (*cpBnd*C_T(L); red stars in Fig. 5b):

cpBndC_T(L) = {(43.25, 2700), (52.8, 2700), (54, 1000), (60, 1000), (61.2, 1520), (64.4, 1520), (66.25, 2320)}.

Fig. 5 further shows that the A_T and C_T bounds do not have the same shape. Thus, knowing the maximum of variability (*MaxBnd*A_T = 3900 for A_T, and *MaxBnd*C_T = 2700 for C_T), the maximum errors of interpolation of these data sets with 238 points, can be easily calculated using Eq. (14). The results along the cruise track (43.25°S - 66.25°S) are:

MaxErrBnd$(A_T)_{even}$ = 4.59 μmol.kg^{-1}; MaxErrBnd$(C_T)_{even}$ = 3.18 μmol.kg^{-1};

MaxErrBnd$(A_T)_{sb}$ = 2.42 μmol.kg^{-1}; MaxErrBnd$(C_T)_{sb}$ = 2.16 μmol.kg^{-1}.

Since the results for both (an even pattern sampling or an irregular sampling pattern), indicate a maximum interpolation error larger than the desired interpolation error (±1.75 μmol.kg^{-1} for A_T and ± 1.35 μmol.kg^{-1} for C_T, as mentioned above), it is clear that 238 samples evenly spaced along the latitude axis between Hobart and Dumont D'Urville are not enough. These results clearly show that the position of the samples has a significant impact on the interpolation accuracy.

Given the acquired knowledge on the A_T and C_T variability bounds, it is now possible to design an appropriate sampling strategy with a minimum of samples. Thus, where the bound is constant, samples will be evenly spaced along the latitudinal axis, and in areas where the bound varies, samples will be unevenly spaced along the latitudinal axis.

In order to determine the exact position of samples to be measured throughout the cruise track, it is necessary to first determine the number of samples to be taken (depending upon the desired maximum interpolation error), within each latitudinal area defined by the variability bound, and then to calculate the sample positions in the areas where the variability bound varies.

Tables 4 and 5 illustrate the results of the sample size needed (calculated using Eq. (15)), in each latitudinal interval for A_T and C_T,

Table 3
Numbers of samples needed within each latitudinal interval to reach the desired maximum interpolation error for SSS measurements assuming the high SSS variability near 47°S does not exist. The gray boxes indicate that the number of samples are calculated (Eq. (15)) for a semi-balanced sampling pattern. The white boxes indicate that the number of samples are calculated (Eq. (15)) for an even sampling pattern.

Latitudinal interval	N measured	N for a desired SSS Maximum interpolation error		
		0.0025	0.005	0.01
43.2262°S - 44°S	253	313	222	157
44°S - 45.3°S	432	462	327	231
45.3°S - 52°S	2235	2011	1422	1006
52°S - 53.5°S	566	346	245	174
53.5°S - 60°S	2137	920	651	461
60°S - 62°S	646	359	254	180
62°S - 65.5°S	1155	743	526	372
65.5°S 65.7°S	71	54	38	27
65.7°S -65.8°S	63	32	23	16
65.8°S - 66.2598°S	257	117	83	59
43.2262°S - 66.2598°S	7815	5349	3783	2675

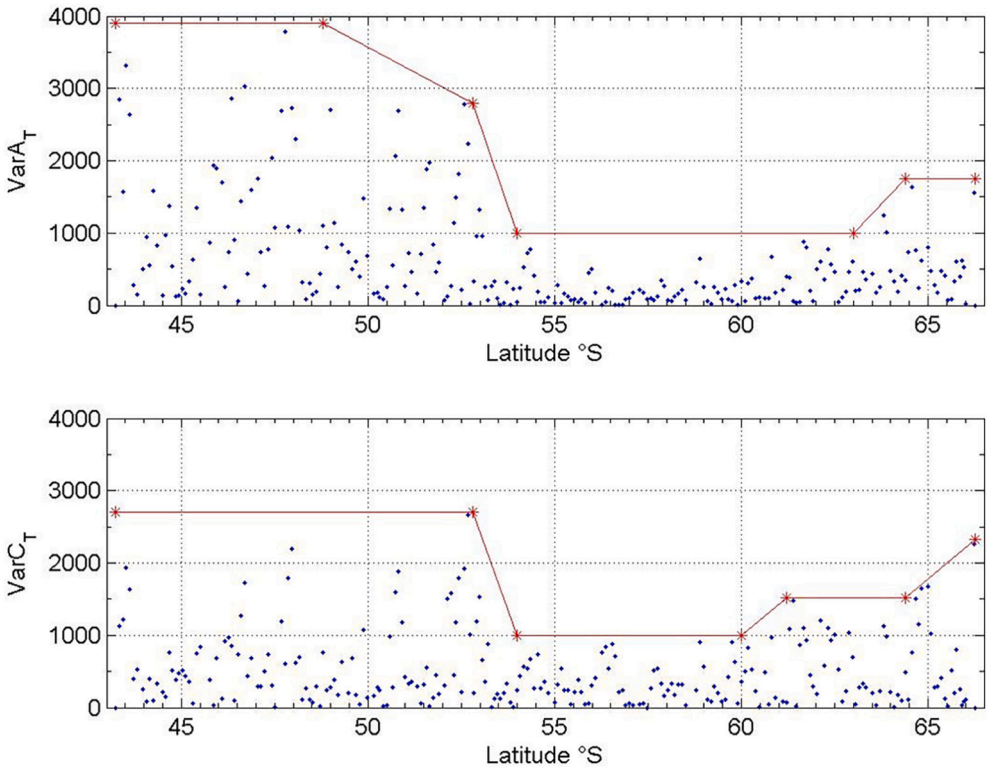


Fig. 5. Variability of a) total alkalinity as a function of latitude, and b) total inorganic carbon as a function of latitude. The solid (red) lines on each graph represent the variability bounds. **(For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)**

Table 4
Numbers of samples needed within each latitudinal interval to reach the desired maximum interpolation error for A_T measurements. The gray boxes indicate that the number of samples are calculated (Eq. (15)) for a semi-balanced sampling pattern. The white boxes indicate that the number of samples are calculated (Eq. (15)) for an even sampling pattern.

Latitudinal interval	N measured	N for a desired A_T maximum interpolation error (μmol.kg^{-1})			
		0.5	1.0	1.35	1.75
43.25°S - 48.8°S	55	174	124	106	94
48.8°S - 52.8°S	43	117	83	71	63
m52.8°S - 54°S	15	27	19	17	15
54°S - 63°S	90	143	102	88	77
63°S - 64.4°S	14	27	19	17	15
64.4°S - 66.25°S	21	40	28	25	22
43.25°S - 66.25°S	238	523	370	318	280

respectively. These Tables further show the results of the calculation performed for three desired maximum interpolation errors.

The first one is guided by our effective cruise measurement accuracies. The second and third ones assume the measurements are performed with an improved accuracy of 2 μmol.kg^{-1} and 1 μmol.kg^{-1}, respectively. Note that these improved targeted accuracies are reasonable and reachable.

For total alkalinity, the results (Table 4) indicate that the desired maximum interpolation error of 1.75 μmol.kg^{-1} was reached within the two latitudinal intervals [54°S, 63°S] and [64.4°S, 66.25°S] since in these intervals 77 and 22 samples respectively, are needed while during the cruise 90 and 21 samples were effectively measured in these intervals, respectively. In the interval [52.8°S, 54°S], the desired maximum interpolation error could have been reached only if, the 15 measured samples would have been measured according to a semi-balanced error sampling pattern (unevenly spaced along the latitudinal axis). Within the remaining three intervals ([43.25°S, 48.8°S], [48.8°S, 52.8°S], [63°S, 64.4°S]), they were not enough samples to reach the desired maximum interpolation error of 1.75 μmol.kg^{-1}.

These results further show the significant increase in the number of samples needed as the desired maximum interpolation error decreases.

For total inorganic carbon, the results (Table 5) indicate that the desired maximum interpolation error of 1.35 μmol.kg^{-1} can be reached only within the latitudinal interval [54°S, 60°S]. All the other latitudinal intervals need to be sampled at a higher rate. This is logical since the desired maximum interpolation error is reduced compared with that of A_T. Here too, the results reported in Tables 4 and 5, show the significant differences in the number of samples needed as the desired maximum interpolation error decreases.

Since the shape of the A_T variability bound is different from that of the C_T variability bound, ideally their sampling patterns would be

Table 5
Numbers of samples needed within each latitudinal interval to reach the desired maximum interpolation error for C_T measurements. The gray boxes indicate that the number of samples are calculated (Eq. (15)) for a semi-balanced sampling pattern. The white boxes indicate that the number of samples are calculated (Eq. (15)) for an even sampling pattern.

Latitudinal interval	N measured	N for a desired C_T maximum interpolation error (μmol.kg^{-1})			
		0.5	1.0	1.35	1.75
43.25°S - 52.8°S	98	249	177	152	134
52.8°S - 54°S	15	26	19	16	15
54°S - 60°S	59	96	68	59	52
60°S - 61.2°S	11	22	16	13	12
61.2°S - 64.4°S	34	63	45	39	34
64.4°S - 66.25°S	21	41	29	25	23
43.25°S - 66.25°S	238	494	350	301	264

different. However, as mentioned above, for various reasons (technique of measurement, convenience, …), it may be required or desirable, to collect samples simultaneously for A_T and C_T measurements. In this case, according to the various areas it will be necessary to oversample one property or the other. For instance, here, assuming the desired interpolation accuracy is 1.75 μmol.kg^{-1} for A_T and 1.35 μmol.kg^{-1} for C_T, a common sampling strategy would require the sample, within each latitudinal interval, to have the maximum of the calculated number of samples required for A_T and that for C_T. The result is summarized in Table 6.

These results show that in order to reach the desired maximum interpolation error of 1.35 μmol.kg^{-1} for C_T and 1.75 μmol.kg^{-1} for A_T measurements along this cruise track between 43.25°S and 66.25°S, it would be necessary to collect at least 324 samples at specific locations, while only 238 samples were measured at quasi-regularly spaced locations along the latitude axis. Thus, in some areas, such as within the interval [54°S; 60°S], the location and number of samples measured were almost sufficient, while in other areas, such as within the interval [43.25°S; 48.8°S], the number of samples measured is clearly not enough (by almost a factor 2). Yet in other areas, such as within the intervals [52.8°S; 54°S] and [60°S; 61.2°S], the number of samples measured is close to sufficient but only if, they would have been measured at appropriate (as calculated below) irregular spacing locations.

Fig. 6 below, illustrates the results of the computation of the sample positions within the latitudinal interval [52.8°S – 54°S]. In this example, using Eq. (16) in which the function $A(x) = \int_a^x \sqrt{BndY(t)}\,dt$ is calculated with "x" the latitude within the interval [52.8°S – 54°S], and *Bnd*Y(t) is *Bnd*C_T(L) within this same interval [52.8°S – 54°S]. The 16 samples needed in this latitudinal interval are then regularly distributed along the A(x) axis to find the position of the samples on the latitude ("x") axis.

This figure shows that within the 0.2° latitudinal interval [52.8°S – 53°S] four samples would be needed, while within the other 0.2° latitudinal interval [53.8°S – 54°S], only three samples would be needed. Thus, with this function, an appropriate irregular spacing between samples can be easily calculated.

Fig. 7 illustrates the result (of Eq. (16)) for an appropriate number of samples and their positions within the interval [43.25°S; 66.25°S], to reach a desired maximum interpolation error of 1.35 μmol.kg^{-1} for C_T and 1.75 μmol.kg^{-1} for A_T.

This **Fig. 7** illustrates that spacing between samples should vary from very small in the area [61.2°S; 63°S], to much larger in the area [54°S; 60°S]. The number of samples required within each degree of latitude is shown below in **Fig. 8**.

Thus, these two figs. (7 and 8) illustrate that the property variability bound determines the relative proportion of samples spread over the latitudinal interval ('x' axis). These results confirm not only what is intuitively expected; where the property variability is high, the number of samples should be high. But in addition they indicate both the exact number of samples needed and their position. Where the property variability bound is constant, the samples should be regularly spaced, as illustrated in figure8, by the same number of samples per degree of latitude at the beginning and at the end of the cruise track. Elsewhere, the number of samples per degree varies (see Figs. 7 and 8, within the latitudinal areas [50°S; 55°S] and [58°S; 62°S]), and the samples should be collected at specific locations as determined by Eq. (16).

It should be emphasized that this methodology cannot only be applied directly to data sets of properties to be sampled but also to data derived from other properties (such as C_T and A_T functions of SST and SSS (Guglielmi et al., **2022**)).

Table 6
Numbers of samples needed for common sampling within each latitudinal interval to reach the desired maximum interpolation error of 1.75 μmol.kg^{-1} for A_T and 1.35 μmol.kg^{-1} C_T measurements. The bold italic numbers are from C_T (Table 5), while the others are from A_T (Table 4).

Latitudinal interval	N measured	N for common A_T and C_T sampling with 1.75 μmol.kg^{-1} and 1.35 μmol.kg^{-1} maximum interpolation error, respectively
43.25°S - 48.8°S	55	94
48.8°S - 52.8°S	43	63
52.8°S - 54°S	15	***16***
54°S - 60°S	59	***59***
60°S - 61.2°S	11	***13***
61.2°S - 63°S	20	***39***
63°S - 64.4°S	14	15
64.4°S - 66.25°S	21	***25***
43.25°S - 66.25°S	238	317

5. Discussion

It is generally the case that little is known about the interpolation properties of sample data along the path where it is collected. The Sample Error Theorem shows that if data is collected along a path using an evenly spaced sample pattern then the variation of sample (interpolation) error over this path follows the variation of the variability of the sampled data. If the variability of the sample data is steady and even then the interpolation error will also be steady and even. However if the data has a high variability along some portions of the path but not over other portions, the interpolation error will exhibit the same variation. It follows that if a sample pattern with the same number of points could be found that 'equalizes' the interpolation error by increasing the sample frequency (decreasing the spacing) along the portions of the path where the variability is higher, and decreasing the sample frequency

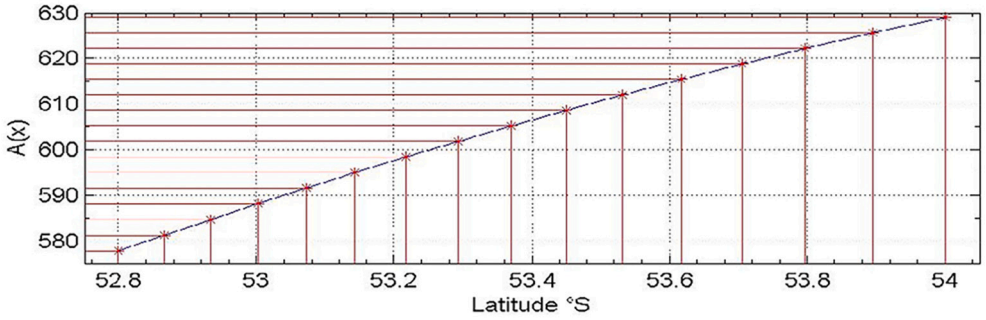


Fig. 6. Function $A(x) = \int_a^x \sqrt{BndCT(t)}\, dt$ within the latitude interval [52.8°S – 54°S].

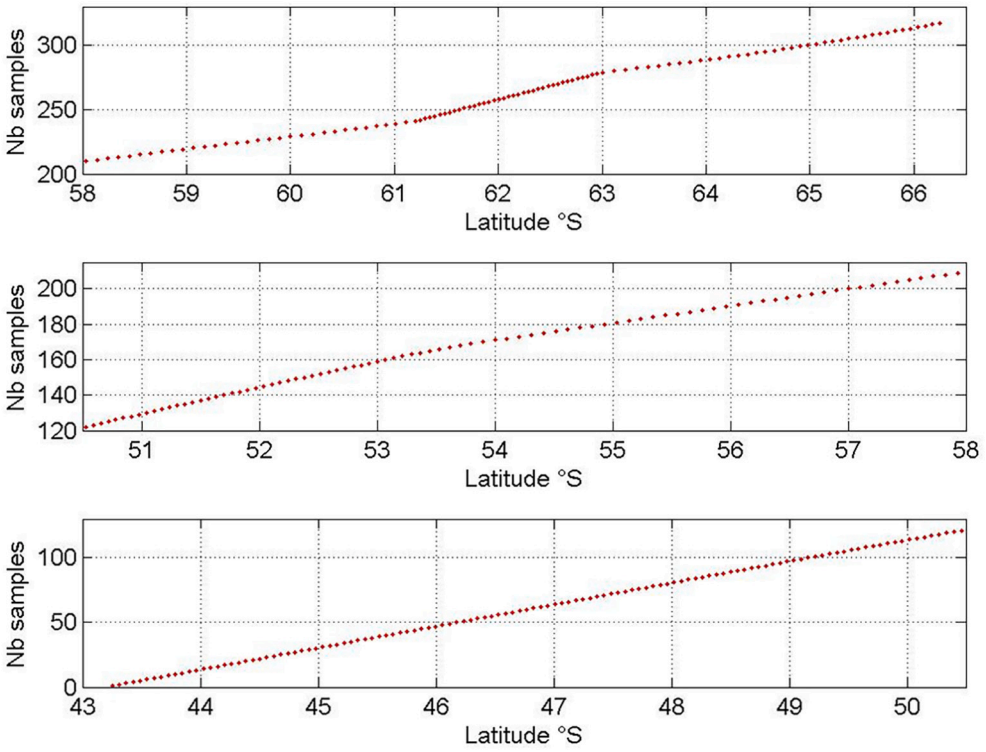


Fig. 7. Sample location for each of the 317 samples within the latitudinal interval [43.25°S; 66.25°S]. In order to best visualize the position of the samples, the figure is split in three latitudinal areas; [43.25°S; 50.5°S], [50.5°S; 58°S], and [58°S; 66.25°S].

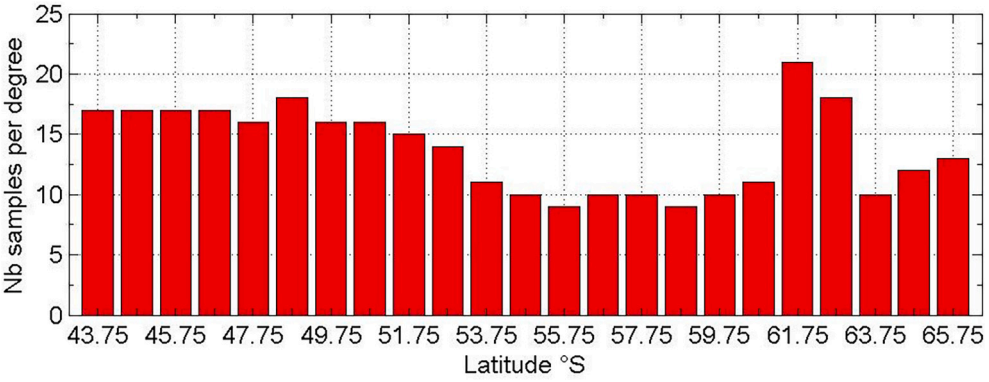
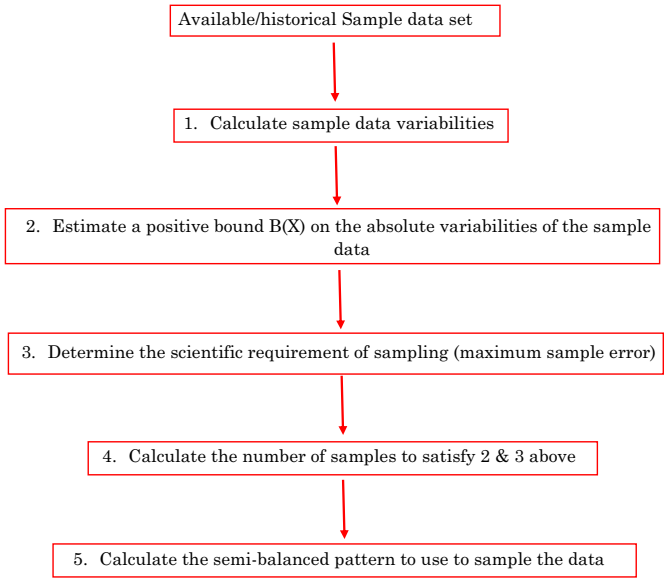


Fig. 8. Number of samples per degree of latitude within the interval [43.25°S; 66.25°S].

(increasing the spacing) along the portions of the path where the variability is lower, then the interpolation error would be more uniform between sample points and the overall maximum interpolation error would be reduced. The primary steps of the new protocol proposed here are shown in the following flowchart:



In particular, this study shows that with simple calculations, it is possible not only to know the maximum linear interpolation error of any measured property, but also to precisely determine the positions of these measurements along a cruise track (based upon previous data sets) while minimizing both the number of these measurements and the maximum interpolation error.

Since the accuracy of each property measurement depends upon the measuring system, each property would be ideally measured using its proper sampling pattern. However, if for any (practical) reason, two (or more) properties should be sampled simultaneously, then it is possible to determine a common sampling strategy. Such common strategy will increase the number of measurements required to preserve the desired maximum interpolation error of each of these properties.

All these calculations are based upon the variability of the signal or more exactly upon bounds of the variability property. In areas where the variability bounds of properties (such as here, SST, SSS, A_T, or C_T), are constant, sampling would ideally be regularly spaced along the 'x' axis (here, the "latitude" axis). However, in areas where the variability bounds vary, sampling would ideally be irregularly spaced along the "x" axis.

The choice of a variability bound of a property is particularly important since all the calculations are based upon these bounds. In order to capture the complete signal variability, it is good to define somewhat larger variability bounds than those calculated from one or a few examples of sample data, but that also means that the property could be oversampled. On the other hand, if the variability bounds are chosen too small, there will be fewer measurement samples but with the risk of missing the highest signal variability. Thus, depending upon the objectives, priorities and various practical constraints, one would have to find an equilibrium between sample size and accuracy. A challenge of this method is to determine and estimate persistent and precise variability bounds in the particular data environment that is being sampled.

A key factor to designing an appropriate sampling strategy (with a minimum of samples/measurements), is to know both the accuracy of the measurements and the desired maximum interpolation error.

In any case, all the results as presented above for SSS, SST, A_T, and C_T, show that the current sampling strategy can be significantly improved by using a semi-balanced interpolation error sampling strategy.

In particular, this study emphasizes the large difference in the number of SSS and SST measurements needed (5527 and 32,958, respectively), along a cruise track between Hobart and Dumont D'Urville to ensure that the maximum interpolation errors remain below half the property measurement accuracy. Concerning A_T, and C_T, "only" 317 measurements could be sufficient. Even if it can be assumed that A_T, and C_T can be measured with an improved accuracy of 1 $\mu mol.kg^{-1}$, "only" 380 measurements could be sufficient. Yet, this would still represent a significant effort since a measurement of A_T and/or C_T usually takes much more time and is much more expensive than a SSS or SST measurement.

Why is there such a large difference between the A_T and/or C_T and SSS number of measurements? As mentioned above, all the calculations are based upon the estimate of the property variability bounds. Since the number of SSS data is large (7815), all short-scale SSS variabilities (such as those near 47°S), are likely to be detected, thus raising the maximum of the bound. On the other hand, since the number of A_T and/or C_T data is relatively small (238), some small-scale variabilities may remain undetected, thus lowering the maximum variability bound.

Furthermore, this work shows that some ocean areas require special attention when processing the underway data, such as the data from eddies as well as convergence and divergence zones, and polar fronts (such in the Barents Sea, or the Antarctic and Sub-Antarctic fronts). These zones can be characteristic of five general processes; 1) the crossing of any boundary between a cold and a warm current (such as the Gulf-Stream and its eddies, or the Sargasso Sea front, etc.), as well as gyres (such as the cyclonic sub-polar gyres in the Wedell and Ross Seas), 2) the crossing of any boundary between coastal and oceanic waters (such as coastal front of the Iroise Sea, etc.), 3) off estuaries (such at the meeting of the Amazon river with the Atlantic waters, etc.), 4) Arounds banks, reefs, shoals, and an island shelf, 5) along the margin of upwelling areas (such as the equatorial front between Peru and Galapagos Islands, or the Benguela upwelling, or the Mediterranean upwelling near the Sicilian coast, etc.).

6. Conclusions

Overall, this study is based upon a methodology (in Section 2.2), that significantly facilitates the determination of appropriate sampling and measurement locations, even when the number of measurements is limited by various constraints (time, cost, technology, accuracy, etc.). The methodology illustrated above is a rigorously based, practical, data driven approach to the design of sampling patterns that achieves these goals. As illustrated here, the somewhat novel concepts of sample variability, sample spacing and the Sample Error Theorem enable scientists to not only estimate sample error (interpolation error) from the sampled data itself, but also can be used to establish the existence of persistent bounds on variability and from this to determine unique sample patterns that meet the scientific requirement of a given maximum interpolation error. Moreover, this methodology is general and can be applied to any kind of environmental study (geology, meteorology, etc.), with any kind of data (in situ, remotely sensed, etc.), and provides a path to significant improvements to the overall scientific value of sampled data.

In summary, the results presented above raise further questions about the scientific need for improvement of both measurement and interpolation accuracy.

Given the present technology, how should sample measurements be distributed to achieve the minimum maximum interpolation error? Which scientific questions could be precisely answered with this present technology? Is there a need to create new technologies with better accuracy (and/or measurement frequency) to make significant scientific progress?

For instance, here, one could conclude that there is an urgent need to develop new reliable technologies for much faster (and cheaper)

measurements of A_T and C_T in seawater.

Thus, this study opens the route to more efficient sampling strategies (and cruise designs), which could significantly enhance scientific progress.

Funding

This research did not receive any grant from funding agencies in the public, commercial, or not-for-profit sectors.

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.

Data availability

The link to the data is in the references

Acknowledgments

We genuinely thank Alain Poisson for the initiation of the MINERVE program, as well as Elodie Kestenare and Rosemary Morrow for the SURVOSTRAL data and helpful comments.

We are grateful to Daniel Davis for helpful discussions, thoughtful suggestions, and providing language help.

We thank the Terres Australes et Antactique Françaises (TAAF) and the Institut Paul Emile Victor (IPEV) for logistic support during the 2010 cruise.

References

Alory, G., Delcroix, T., Téchiné, P., Diverrès, D., Varillon, D., Cravatte, S., Gouriou, Y., Grelet, J., Jacquin, S., Kestenare, É., 2015. The French contribution to the voluntary observing ships network of sea surface salinity. Deep-Sea Res. I Oceanogr. Res. Pap. 105, 1–18.

Brandon, M., Goyet, C., Touratier, F., Lefèvre, N., Kestenare, E., Morrow, R., 2022. Spatio-temporal variability of the CO2 properties and anthropogenic carbon penetration, in the Southern Ocean surface waters. Deep Sea Research Part I. Oceanograp. Res. Papers.103836. https://doi.org/10.1016/j.dsr.2022.103836.

Davis, D., Goyet, C., 2021. Balanced Error Sampling: With **A**pplication to **O**cean **B**iogeochemical **S**ampling. Presses Universitaires de Perpignan (224pp).

DeVries, T., 2014. The oceanic anthropogenic CO $_2$ sink: **s**torage, air-sea fluxes, and transports over the industrial era. Glob. Biogeochem. Cycles 28, 631–647. https://doi.org/10.1002/2013GB004739.

Ford, D., 2021. Assimilating synthetic **b**iogeochemical-Argo and ocean colour observations into a global ocean model to inform observing system design. Biogeosciences 18, 509–534. https://doi.org/10.5194/bg-18-509-2021.

Friedlingstein, P., O'Sullivan, M., Jones, M.W., Andrew, R.M., Hauck, J., Olsen, A., Peters, G.P., Peters, W., Pongratz, J., Sitch, S., Le Quéré, C., Canadell, J.G., Ciais, P., Jackson, R.B., Alin, S., Aragão, L.E.O.C., Arneth, A., Arora, V., Bates, N.R., Becker, M., Benoit-Cattin, A., Bittig, H.C., Bopp, L., Bultan, S., Chandra, N., Chevallier, F., Chini, L.P., Evans, W., Florentie, L., Forster, P.M., Gasser, T., Gehlen, M., Gilfillan, D., Gkritzalis, T., Gregor, L., Gruber, N., Harris, I., Hartung, K., Haverd, V., Houghton, R.A., Ilyina, T., Jain, A.K., Joetzjer, E., Kadono, K., Kato, E., Kitidis, V., Korsbakken, J.I., Landschützer, P., Lefèvre, N., Lenton, A., Lienert, S., Liu, Z., Lombardozzi, D., Marland, G., Metzl, N., Munro, D.R., Nabel, J.E.M.S., Nakaoka, S.-I., Niwa, Y., O'Brien, K., Ono, T., Palmer, P.I., Pierrot, D., Poulter, B., Resplandy, L., Robertson, E., Rödenbeck, C., Schwinger, J., Séférian, R., Skjelvan, I., Smith, A.J.P., Sutton, A.J., Tanhua, T., Tans, P.P., Tian, H., Tilbrook, B., van der Werf, G., Vuichard, N., Walker, A.P., Wanninkhof, R., Watson, A.J., Willis, D., Wiltshire, A.J., Yuan, W., Yue, X., Zaehle, S., 2020. Global **c**arbon **b**udget 2020. Earth Syst. Sci. Data 12, 3269–3340. https://doi.org/10.5194/essd-12-3269-2020.

Guglielmi, V., Touratier, F., Goyet, C., 2022. Design of sampling strategy measurements of CO_2/carbonate properties. Journal of Oceanography and Aquaculture. https://doi.org/10.23880/ijoac-16000227. ISSN:2577-4050.

Hall, Bradley D., Crotwell, Andrew M., Kitzis, Duane R., Mefford, Thomas, Miller, Benjamin R., Schibig, Michael F., Tans, Pieter P., 2021. Revision of the World Meteorological Organization Global Atmosphere Watch (WMO/GAW) CO2 calibration scale. Atmospheric Measurement Techniques 14 (4), 3015–3032. https://doi.org/10.5194/amt-14-3015-2021.

Keeling, C., Chin, J., Whorf, T., 1996. Increased activity of northern vegetation inferred from atmospheric CO2 measurements. Nature 382 (6587), 146–149.

Thoning, Kirk W., Tans, Pieter P., Komhyr, Walter D., 1989. Atmospheric carbon dioxide at Mauna Loa observatory: 2. Analysis of the NOAA GMCC data 1974-1985. J. Geophys. Res. 94, 8549–8565.

Komhyr, W.D., Harris, T.B., Waterman, L.S., Chin, J.F.S., Thoning, K.W., 1989. Atmospheric carbon dioxide at Mauna Loa Observatory: 1. NOAA GMCC measurements with a non-dispersive infrared analyzer. J. Geophys. Res. 94, 8533–8547.

Laika, H.E., Goyet, C., Vouve, F., Poisson, A., Touratier, F., 2009. Interannual properties of the CO_2 system in the Southern Ocean south of Australia. Antarct. Sci. 21, 663–680. https://doi.org/10.1017/S0954102009990319.

Morrow, R., Kestenare, E., 2014. Nineteen-year changes in surface salinity in the Southern Ocean south of Australia. J. Mar. Syst. 129, 472–483. https://doi.org/10.1016/j.jmarsys.2013.09.011.

Morrow, R., Donguy, J.R., Chaigneau, A., Rintoul, S., 2004. Cold core anomalies at the Subantarctic Front, south of Tasmania. Deep-Sea Research I 51, 1417–1440.

Stephens, B.B., Wofsy, S.C., Keeling, R.F., Tans, P.P., 2000. The CO2 budget and rectification airborne study: **s**trategies for measuring rectifiers and regional fluxes. Inverse Methods in Global Biogeochemical Cycles Geophy 114, 311–321.

Tans, P.P., Bakwin, P.S., Guenther, D., 1996. A feasible global carbon cycle observing system: **a** plan to decipher today's carbon cycle based on observations. Glob. Chang. Biol. 2 (3), 309–318.

Valsala, V., Sreeush, M.G., Anju, M., Sreenivas, Pentakota, Tiwari, Yogesh K., Chakraborty, Kunal, Sijikumar, S., 2021. An observing system simulation experiment for Indian Ocean surface pCO_2 measurements. Prog. Oceanogr. 194 https://doi.org/10.1016/j.pocean.2021.102570.

Web references

https://gml.noaa.gov/ccgg/trends/.
https://www.icos-cp.eu/.
http://www.obs-vlfr.fr/cd_rom_dmtt/sodyf_main.htm.
https://scrippsco2.ucsd.edu/data/seawater_carbon/ocean_time_series.html.
https://community.wmo.int/ship-opportunity-programme.
https://www.legos.omp.eu/survostral.
https://campagnes.flotteoceanogra-phique.fr/series/128/fr/.

Data references

https://data.ifremer.fr/SISMER.
https://www.seanoe.org.
https://sss.sedoo.fr.