

COMPARATIVE AND INTERCALIBRATION STUDIES

The ICOS OTC $p\text{CO}_2$ instrument intercomparison

Tobias Steinhoff^{1,2,3*}, Thanos Gkritzalis,⁴ Steve Jones,^{4,5} Vlad A. Macovei⁶, Craig Neill,⁷ Ute Schuster,⁸ John Akl,⁷ Ricardo Arruda,⁹ Dariia Atamanchuk,⁹ Mark Barry,¹⁰ Laurence Beaumont,¹¹ Carolina Cantoni,¹² Andrew Dickson,¹³ Jana Fahning,¹⁴ Jac Fought,¹⁵ Constantin Frangoulis,¹⁶ Lucía Gutiérrez-Loza^{1,3}, Clinton Hagan,¹⁷ Martti Honkanen,¹⁸ Sami Kielosto,¹⁹ Nadja Kinski,²⁰ Arne Körtzinger,² Peter Landschützer,⁴ Siv K. Lauvset^{1,3}, Noah Lawrence-Slavas,²¹ Quanlong Li,^{22,23} Anna Luchetta,¹² Damien Malarde,²⁴ Melf Paulsen,² Markus Ritschel,²⁵ Anna Rutgersson,²⁶ Richard Sanders,^{1,3} Kiminori Shitashima,²⁷ Reggie Spaulding,²⁸ Natalia Stamataki,^{16,29} Ken Stenbäck,¹⁹ Adrienne Sutton,²¹ Witold Tatkiwicz,⁸ Maciej Telszewski,³⁰ Hannelore Theetaert,⁴ Bronte Tilbrook,⁷ Rik Wanninkhof³¹

¹NORCE Norwegian Research AS, Bergen, Norway; ²GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany; ³Bjerknes Centre for Climate Research, Bergen, Norway; ⁴Flanders Marine Institute (VLIZ), Oostende, Belgium; ⁵University of Bergen, Bergen, Norway; ⁶Institute of Carbon Cycles, Helmholtz-Zentrum Hereon, Geesthacht, Germany; ⁷CSIRO, Hobart, Australia; ⁸University of Exeter, Exeter, UK; ⁹Department of Oceanography, Dalhousie University, Halifax, Nova Scotia, Canada; ¹⁰Pro-Oceanus Systems, Bridgewater, Nova Scotia, Canada; ¹¹CNRS, Gif sur Yvette, France; ¹²CNR-ISMAR Institute of Marine Sciences, Trieste, Italy; ¹³Scripps Institution of Oceanography/UCSD, San Diego, California, USA; ¹⁴SubCtech, Kiel, Germany; ¹⁵Sensors & Systems Solutions, LLC, Lexington, South Carolina, USA; ¹⁶Institute of Oceanography, Hellenic Centre for Marine Research (HCMR), Heraklion, Greece; ¹⁷Hagan Technologies, Cape Town, South Africa; ¹⁸Finnish Meteorological Institute, Helsinki, Finland; ¹⁹Finnish Environment Institute, Helsinki, Finland; ²⁰4H-JENA engineering GmbH, Kiel, Germany; ²¹NOAA Pacific Marine Environmental Laboratory, Seattle, Washington, USA; ²²State Key Laboratory of Marine Environmental Science, Xiamen University, Xiamen, China; ²³College of the Environment and Ecology, Xiamen University, Xiamen, China; ²⁴NKE instrumentation, Hennebont, France; ²⁵University of Hamburg, Hamburg, Germany; ²⁶Uppsala University, Uppsala, Sweden; ²⁷Tokyo University of Marine Science and Technology, Tokyo, Japan; ²⁸Sunburst Sensors, Missoula, Montana, USA; ²⁹Department of Physics, Section of Environmental Physics and Meteorology, National and Kapodistrian University of Athens, Athens, Greece; ³⁰IOCCP, Sopot, Poland; ³¹NOAA Atlantic Oceanographic and Meteorological Laboratory, Miami, Florida, USA

Abstract

In 2021, the Ocean Thematic Centre of the European Research Infrastructure “Integrated Carbon Observation System” conducted an international partial pressure of carbon dioxide ($p\text{CO}_2$) instrument intercomparison. The goal was to understand how different types of instrumentation for the measurement of ocean $p\text{CO}_2$ compare to each other. During the two-week long experiment, we installed various instruments in a tank facility using natural sea water (North Sea). These included direct air–water equilibration systems and membrane-based flow-through instruments along with submersible sensors and instruments that are normally installed on buoys and autonomous surface vehicles. In situ instruments were installed inside the tank and the flow-through instruments were fed the same water using a pumping system. We changed the temperature (between 10°C and 28°C)

*Correspondence: tsteinhoff@geomar.de

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Associate editor: Isaac Santos

and the seawater $p\text{CO}_2$ (between 250 and 800 μatm) to observe instrument responses over a wide range. Since there is no reference for surface ocean $p\text{CO}_2$ measurements, we agreed on a set of instruments serving as intercomparison reference. All data from the different instruments were then compared against the intercomparison reference during periods of stable temperature and $p\text{CO}_2$. The study provides important information to enhance future ocean carbon monitoring networks, but makes no direct recommendation for the use of any specific sensor. A major finding is that equilibration through direct air–water contact appears to be more consistent and independent of external factors than equilibration through a membrane or photometric detection. We found several instruments with no temperature measurements at the location of equilibration or with uncalibrated temperature sensors introducing significant uncertainty in the results.

The release of carbon dioxide (CO_2) to the atmosphere due to anthropogenic activities continues, resulting in a net annual growth rate in atmospheric CO_2 content of ca. 5.2 Gt yr^{-1} (Friedlingstein et al. 2023; IPCC 2022). The Global Carbon Project (Friedlingstein et al. 2023) presents independent estimates for CO_2 sources (e.g., fossil fuel burning, cement production, land-use change) and sinks (e.g., ocean and land uptake). Over the last 60 yr, the ocean's contribution as a CO_2 sink has remained at around 25% of global CO_2 emissions, increasing from $1.1 \pm 0.4 \text{ Gt C yr}^{-1}$ in 1960 to $2.9 \pm 0.6 \text{ Gt C yr}^{-1}$ in 2023. Superimposed on the background uptake due to increasing atmospheric CO_2 is considerable interannual and decadal change in the ocean sink for different regions (Gruber et al. 2023; Mayot et al. 2023), which shows the importance of observations to resolve how CO_2 uptake is evolving (Gruber et al. 2023, Mayot et al. 2023). However, Friedlingstein et al. (2023) point out that there is a significant divergence between two independent types of estimates used to calculate the best-guess ocean sink, that is, estimates based on ocean hindcast simulations (models) and those based on surface ocean observations, leading to a difference of ca. $0.6 \text{ PgC yr}^{-1} \text{ decade}^{-1}$ for 2023. Hereby it appears that models tend to underestimate the ocean CO_2 sink compared to observation-based products. This is also supported by independent methods like atmospheric oxygen budgets. However, the uncertainty in the observation-based results, relying on surface water partial pressure of carbon dioxide ($p\text{CO}_2$) measurements of $\approx 0.5 \text{ PgC yr}^{-1}$ precludes a definitive attribution of cause of the discrepancy.

More high-quality observations of surface ocean $p\text{CO}_2$ are required to resolve this model-data mismatch and improve the ocean sink estimate (e.g., Gloege et al. 2022; Hauck et al. 2023). In this context, new observations, especially in the data-poor regions, are required to increase fidelity of gap-filling techniques used to extrapolate sparse observations onto a global grid (Heimdal et al., 2024). Globally, however, Dong et al. (2024) suggest uncertainty in data-based estimates of the ocean CO_2 sink has increased by 65% due to a global decline in data availability since 2017. Recently, the World Meteorological Organization initiated the “Global Greenhouse Gas Watch” aiming to provide “useful and actionable climate information for mitigation, adaptation and early warning systems” (<https://library.wmo.int/idurl/4/68532>, last accessed

01 November 2024). The lack of oceanic observations of greenhouse gas fluxes is identified as a limitation for global estimates. Bridging the gap between the model-based and observation-based estimates requires not only an increase in the number of observations but also improved accuracy and thorough understanding of their uncertainty: the better we know the uncertainty of each data point that is used for global estimates the better constrained these estimates have the potential to be. Observations of the CO_2 flux between the ocean and the atmosphere are commonly calculated using a bulk flux Eq. (1):

$$F = k \times s \times \Delta p\text{CO}_2 \quad (1)$$

where $\Delta p\text{CO}_2$ is the difference between the bulk surface oceanic and lower atmospheric $p\text{CO}_2$ ($p\text{CO}_2 [\text{ocean}] - p\text{CO}_2 [\text{atmosphere}]$), k is the gas transfer velocity, and s is the solubility of CO_2 . Uncertainties in each of the components will impact the uncertainty of the flux estimate F . The gas transfer velocity is often estimated from wind speed (e.g., Wanninkhof and Triñanes 2017) and accounts for 15–20% of the uncertainty in flux estimates (e.g., Takahashi et al. 2009). The uncertainty of the calculated solubility is less than 1% (Weiss 1974). The major uncertainty in $\Delta p\text{CO}_2$ for many ocean regions stems from missing surface ocean $p\text{CO}_2$ observations (Jersild and Landschützer 2024) as ocean $p\text{CO}_2$ is more difficult to measure and is more variable in space and time than atmospheric CO_2 . In this study, we address ocean $p\text{CO}_2$ observations and their associated uncertainties through comparing the results of the largest number of different instruments operated side by side to date.

To track changes in the open ocean, $p\text{CO}_2$ needs to be measured with high accuracy due to the small average global disequilibrium of sea–air CO_2 ($-4.1 \mu\text{atm}$ for 1980–2021; Fay et al. 2024) and the small observed spatiotemporal variability relative to coastal regions. In the coastal ocean, with impacts from land, river systems, and sediments, the criteria can be less stringent since the observed variations can be an order of magnitude greater than in the open ocean (e.g., Dai et al. 2009; Macovei et al. 2022), and the signal is easier to resolve.

Surface ocean $p\text{CO}_2$ is a subvariable of the Essential Ocean Variable “Inorganic Carbon” defined by the Global Ocean

Observing System (<https://goosocean.org/document/17475>). The Global Ocean Observing System Essential Ocean Variable aimed at resolving fluxes to 0.2 PgC yr^{-1} (10%) and the present increased sink makes this about 0.3 PgC yr^{-1} . Global Ocean Observing System defines the accuracy/uncertainty for $p\text{CO}_2$ measurements for different platforms with $\pm 2 \mu\text{atm}$ for ship-based underway observations, $\pm 5 \mu\text{atm}$ for autonomous surface vehicles (AUVs), and $\pm 10 \mu\text{atm}$ for moored observatories. The classification of uncertainty goals for different platforms is also dictated by available instrumentation. Different instruments are designed for different applications. While classical air–water equilibration systems with multiple calibration gases and precise gas analyzers can be installed on ships, they are not suitable for installation on small platforms like AUVs or underwater. The needed accuracy in $p\text{CO}_2$ also depends on the research question, and the user has to decide which instrument is adequate for their application (e.g., when studying regions with large $p\text{CO}_2$ gradients, climate quality data [uncertainty of $p\text{CO}_2$ better than $2 \mu\text{atm}$] are not necessarily needed). The lower quality measurements can be useful in coastal regions where there is larger variability, and this also applies in the subsurface. However, coastal regions are a small percentage of the global ocean, and the exercise was about the performance of sensors for measurements at the surface, not the subsurface. In addition to the Global Ocean Observing System definitions, Newton et al. (2014) defined two criteria for work within the Global Ocean Acidification Observing Network: (1) climate goal: measurements of quality sufficient to assess long-term trends and (2) weather goal: measurements of quality sufficient to identify relative spatial patterns. They also define that the climate goal requires an uncertainty of $p\text{CO}_2$ of less than $\pm 2 \mu\text{atm}$. But even with these definitions, measurements with uncertainties greater than $\pm 2 \mu\text{atm}$ taken at locations where no other measurements are available can support the climate goal. However, knowing the uncertainty of any $p\text{CO}_2$ measurement is essential for making estimates of the uncertainty of global ocean CO_2 uptake. As an approximation, $\pm 2 \mu\text{atm}$ translates to an approximate uncertainty of $\pm 0.4 \text{ Pg C yr}^{-1}$ in the total global ocean CO_2 flux (Bender et al. 2002; Wanninkhof et al. 2019).

One difficulty of calculating the uncertainty of seawater $p\text{CO}_2$ measurements is that there is no reference standard or material available for this seawater parameter, only for CO_2 in air. This is because the seawater $p\text{CO}_2$ is defined as “the partial pressure of CO_2 in air that is in equilibrium with a sample of seawater” (Dickson et al. 2007). As a consequence of this definition, one can separate the measurement of seawater $p\text{CO}_2$ into two steps: (1) equilibration of a volume of air with the seawater sample/flow and (2) calculate the $p\text{CO}_2$ in the equilibrated air (which is the product of the measured mole fraction of CO_2 and the total gas phase pressure). Both steps can be performed in different ways: equilibration can be reached via direct air–water equilibration or via a gas-permeable membrane. The $x\text{CO}_2$ in air can be measured using a variety of

techniques (most commonly nondispersive infrared [IR] or laser-based analyzers). To accurately measure (or estimate) the seawater $p\text{CO}_2$ with such systems, one needs to control several parameters, that is, degree of equilibration (ideally 100%), $x\text{CO}_2$ measurement, temperature and pressure of equilibration and in situ temperature and salinity, which explains why $p\text{CO}_2$ is still not an “off-the-shelf” parameter: while a CO_2 analyzer can be calibrated using reference gases, all the other factors cannot be similarly verified (Pierrot et al. 2009). In addition to instruments equilibrating and measuring CO_2 in the gas phase, others have been developed that equilibrate the CO_2 dissolved in seawater via a gas-permeable membrane with chemical solutions that change the pH depending on the amount of CO_2 (DeGrandpre et al. 1995). In recent years, the variety of instruments that measure $p\text{CO}_2$ in seawater has increased: new instrumentation has been developed, existing instruments have been improved, and new platforms have been introduced to the market (e.g., AUVs). The available instrumentation can be roughly separated in four groups: surface, underway (air–water equilibration), underway (membrane equilibration) and submersible application (a more detailed specification is given in “Methods” section). Instruments in the “surface group” are typically installed on buoys or AUVs, instruments in the “underway group” are used on ship-board installations and instruments in the “submersible group” can be installed under water at several meters’ depth.

Critical questions for all these approaches are: how well do these instruments measure seawater $p\text{CO}_2$, and how can we determine this? In the absence of a reference for $p\text{CO}_2$ measurements, intercomparison exercises are the only way to evaluate the performance of available instrumentation under identical conditions. These exercises should be carried out regularly to include recent technologies, document improvements that are made to existing instrumentation, identify potential errors and biases, the solutions to them, and update the community about the available equipment. Several multi-instrument intercomparison exercises of different sizes and in different environments (at sea, close to shore or on land) (Arruda et al. 2020) have been undertaken in past decades. Table 1 lists some intercomparisons and their main results and challenges encountered.

Results from the exercises at Scripps in 1994 and in Japan in 2003 and 2009 only provided limited information in the reports linked in Table 1. The R/V Meteor 10-d trans-Atlantic at-sea intercomparison of $p\text{CO}_2$ systems in 1996 (Körtzinger et al. 2000) highlighted, among typical sources of error, the possibility of performing $p\text{CO}_2$ measurements with different systems at an agreement to within $1 \mu\text{atm}$ when systems were maintained well and potential sources of error were considered. This confirmed earlier findings from a smaller at-sea intercomparison between two independent $p\text{CO}_2$ systems (Körtzinger et al. 1996). The 2003 exercise included early

Table 1. Previous multi-instrument $p\text{CO}_2$ intercomparisons.

Year	Details	Publication/report (if available)
1994	Andrew Dickson—laboratory intercomparison at Scripps/La Jolla/USA	
1996	Arne Körtzinger—at sea intercomparison in the North Atlantic Ocean during German R/V Meteor Cruise 36/1 Temperature readings are a significant source of error No best choice of equilibrator type Regular calibration of IR sensor is needed Dryer CO_2 measurement should be recommended (i.e., water vapor correction can be a significant source of error) Agreement to within 1 μatm is possible for well-designed and well-maintained systems across a wide range of conditions	Körtzinger et al. (2000)
2003	NIES, Japan—large indoor pool. Compared $x\text{CO}_2$ and used calibrated temperature sensors provided by NIES, excluding temperature and pressure measurements as possible sources of error. Agreement to 5 ppm in $x\text{CO}_2$ among underway systems Issues identified include contamination of equilibrator headspace by ambient air, pressure changes in equilibrator and fouling by bacterial respiration	http://www.ioccp.org/images/01Underway/WR211_eo.pdf (limited information)
2009	Alliance for coastal technologies—protocols for demonstrating the performance of in situ $p\text{CO}_2$ analyzers Difficulty separating instrument error from natural variability at coastal test sites	https://www.act-us.info/evaluations.php#pCO2
2009	NIES, Japan—Same comparison methodology as NIES 2003 Very good agreement among all but one underway system, which demonstrated the importance of knowing equilibration pressure	http://www.ioccp.org/images/D3meetingReports/SOCATPac_final.pdf (limited information)
2018	Lauri Laakso—INTERCARBO Several $p\text{CO}_2$, pH and total alkalinity instruments were compared. The results highlighted that differences between instruments existed, especially in conditions that simulated real-life installations. The importance of repeating similar intercomparison experiments was highlighted by the reported authors. Moreover, the challenges of creating suitable experimental conditions were emphasized, especially relating to water inlet and outlet pressures	Laakso et al. (2019)

prototypes of what would become the General Oceanics (GO) underway system and the “Moored Autonomous $p\text{CO}_2$ ” (MApCO₂) buoy system (both listed in the report as “National Oceanographic and Atmospheric Administration (NOAA)”), as well as the “Submersible Autonomous Moored Instrument” and “Carbon Interface Ocean Atmosphere” (Carioca), that is, systems now operating as standard equipment in the global oceans. The exercise was useful in identifying a small bias in the prototype GO system caused by entrainment of ambient air into the equilibrator, which was corrected by the addition of the small 2nd equilibrator on the vent of the main equilibrator. The 2009 exercise had early production versions of the GO system (listed as “NOAA” and “NIO”) and MApCO₂ (NOAA/MBARI). The agreement between the two NIES reference systems and the two GO systems was better than ± 1 ppm in $x\text{CO}_2$ and agreement with MApCO₂ was better than ± 2 ppm.

These past intercomparisons have shown their importance to evaluate the performance of different instruments under

certain conditions. However, the performance of an instrument also depends on users following proper protocols for deployment.

The Integrated Carbon Observation System (ICOS) Oceans network consists of a wide range of stations including ship lines and moorings (Steinhoff et al. 2019) that are operated by numerous groups. The network needs to assure that the quality across the different stations is comparable; the need of an intercomparison of instruments being used by ICOS-Oceans groups led to the 2021 $p\text{CO}_2$ intercomparison in the Marine Station Oostende at the Flanders Marine Institute.

The goal of this exercise was to operate various $p\text{CO}_2$ instruments under controlled conditions, covering different $p\text{CO}_2$ levels and water temperatures, to compare the results. During and after the intercomparison exercise, we examined instruments and data in an objective way. The intercomparison was conducted in close communication with instruments’ principal investigators (PIs) and suppliers

to ensure optimal operation of the equipment and optimal data quality. The data processing of the instrument's data was done in a transparent and consistent way. The intercomparison of the data was done by the organizers of the intercomparison.

Material and procedure

During the 2021 $p\text{CO}_2$ intercomparison (IC2021), we compared 27 instruments representing 16 different instrument models. We placed the instruments into four different groups according to their intended application and equilibration principle (Table 2). Note that the underway group was divided into air–water equilibration and membrane-based equilibration. The application for both types might be the same, but the measurement principle is different. More detailed descriptions of the instruments are provided below and in the Supporting Information:

Surface

Instruments that are designed to autonomously measure directly at the water surface (mounted on a floating platform e.g., as buoy, AUV) and are not meant to be submerged. All instruments, except the Electrode- CO_2 , used during the intercomparison measure CO_2 using an IR detector and use a zero gas and at least one span gas for referencing the IR cell. The measurement principle is based on air–water equilibration. The Electrode- CO_2 sensor is based on ion-sensitive field-effect transistor-pH (ISFET-pH) technology (see sensor description) and is not based on air–water equilibration.

Underway (air–water and membrane equilibration)

Bench-top flow-through instruments that are designed for ship-board installation (e.g., in laboratories or engine rooms). Seawater is pumped from a seawater intake (typical at 2–11 m depth) to an instrument with an air–water equilibrator or using a membrane equilibrator. In these systems, a defined, relatively small volume of air is equilibrated with a continuously flowing seawater supply that has a practically endless volume such that the CO_2 mass transfer required to establish chemical equilibrium does not measurably affect the seawater $p\text{CO}_2$. This application can be extended from surface water to depths of approximately 50 m using a hose and a submersible pump.

Air–water equilibration

The equilibration takes place at atmospheric pressure using different types of equilibrators (spray head, marbles, bubble...) and currently CO_2 is measured using an IR detector; although much of the open ocean observations in the 1980's utilized gas chromatography (Weiss 1981). The water drains from the equilibrator by gravity, sometimes requiring another pump to discharge the waste water. These systems often use a zero gas and several nonzero reference gases to track the detector drift.

Membrane equilibration

Equilibration of the internal sample gas or liquid volume with the ambient water is achieved via a hydrophobic semi-permeable membrane while the actual detection of CO_2 can be done via measurement in the gas phase or by chemical reaction (and subsequent pH detection) within the liquid phase. In the case of the gas phase, CO_2 is measured using an IR detector, and a zero gas and several nonzero reference gases can be connected to track the IR detector drift.

Submersible

Instruments that can be fully submerged in seawater and allow the user to measure below the surface. Equilibration of CO_2 is achieved by different principles: (a) CO_2 equilibrates across a membrane between the water and air inside the instrument; CO_2 is measured using an IR detector. These systems can generate their zero gas for drift control of the IR cell internally by passing ambient air through an alkaline CO_2 absorbing agent such as sodium hydroxide or soda lime. (b) CO_2 passes through a membrane into a dye. The pH changes in the dye due to CO_2 addition or removal are detected by spectrophotometry and related to the seawater $p\text{CO}_2$.

Different $p\text{CO}_2$ instruments are calibrated in different ways. While in bench-top systems with a separate gas analyzer, the analyzer can be calibrated using reference gases with known CO_2 content, this is not possible for submersible systems. Often, these types of instruments use zero CO_2 gas for drift control, which can be generated inside the sensor. Detailed descriptions for each instrument type and its calibration are given in the Supporting Information.

In the context of this work, we define the PI for the instruments as owners and/or suppliers of the instruments. During the planning phase of the exercise, it was agreed with all participating PIs that results will be published with instruments' names; furthermore, all data will be shared among PIs. An open data sharing policy were important for such a large data set and helped to improve data processing routines for several instruments.

To optimize the data quality from each instrument, we asked for protocols describing the correct handling of each instrument. We defined optimal data quality as the highest quality of data that can be obtained with a certain instrument. This does not necessarily mean that these data have the highest quality compared to other instruments. Some, but not all groups/manufacturers provided documentation on how to use and deploy the systems used during this intercomparison to ensure optimal functioning of each instrument. These "Quick Start Guides" (see data archive, Steinhoff et al. 2025) should enable a trained user to set up the instrument and retrieve data of known quality. The Quick Start Guide does not replace the manufacturer's manual, but documents the necessary steps for a successful deployment and optimal performance of the instrument. The Quick Start Guides were an integral part of the

Table 2. List of instrument types with individual instruments within identified groups. The abbreviation for each instrument is used throughout the document. Instruments that did not produce sensible data (either from the beginning of the experiment or during the experiment) are in parentheses. Detailed instrument descriptions are given in the Supporting Information. Instruments in bold were used as internal references during the experiment (see “Determination of an ‘Intercomparison Reference’” section). Note that the stated uncertainty could not always be reproduced during this study.

Group	Instrument	Uncertainty given by manufacturer	PI	Abbreviation
Surface	ASVCO ₂ (NOAA PMEL Engineering, USA)	< 2 μ atm	Noah Lawrence Slavas	ASVCO ₂
	CaPASOS (Prototype, University of Exeter, UK)	N/A	Ute Schuster	(CaPASOS)*
	Electrode prototype (Tokyo University of Marine Science and Technology, Japan)	N/A	Kiminori Shitashima	Electrode-CO ₂ †
	Hagan-GENx (Hagan Technologies, South Africa)—formerly called VeGAS	< 2 μ atm	Clinton Hagan	Hagan-GENx1 Hagan-GENx2
	MApCO ₂ (Sensors Systems Solutions, USA)	< 2 μ atm	Jac Fought	MApCO ₂
Underway air–water	GO 8050 pCO₂ system using LI-COR LI7000 (General Oceanics, USA)	< 2 μ atm	John Akl	GO-Li7000
	GO 8050 pCO₂ system using LI-COR LI7815 (General Oceanics, USA)	< 2 μ atm	Craig Neill	GO-Li7815
	SuperCO ₂ (Sunburst Sensors, USA)	< 1.5% xCO ₂	Martii Honkanen	SuperCO ₂
	Prototype (VLIZ, Belgium)	< 2 μ atm	Thanos Gkritzalis	VLIZ
	Prototype (Xiamen University, China)	< 2 μ atm	Quanlong Li	XMU-CO ₂ -UW
Underway membrane	AFT-CO ₂ (Sunburst Sensors, USA)	< 5 μ atm	Reggie Spaulding	AFT-CO ₂ ‡
	Contros HydroC-CO ₂ FT (Contros, 4H-Jena Engineering, Germany)	< 0.5%	Yoana Voynova	ControsFT1
			Thanos Gkritzalis	ControsFT2
			Thanos Gkritzalis	ControsFT3
	SOOGuard System (Aanderaa) using a prototype in-chamber Pro-Oceanus CO ₂ equilibration. (Dalhousie University, Canada)	< 0.5% xCO ₂	Dariia Atamanchuk	Dal-SOOP
	SubCtech Cube (1,2) and SubCtech race (3) (SubCtech, Germany)	< 0.5%	Jana Fahning	SCT1
			Peter Landschützer	SCT2
			Jana Fahning	SCT3
Submersible	Carioca system (CNRS—INSU Division Technique, France)	< 0.5% xCO ₂	Laurence Beaumont	Carioca§
	Contros HydroC-CO ₂ (Contros, 4H-Jena Engineering, Germany)	< 0.5%	Nadja Kinski	ControsSUB1
			Martii Honkanen	ControsSUB2
			Carolina Cantoni	ControsSUB3
	NKE (NKE instruments, France)	< 3 μ atm	Damien Malardé	NKE§
	CO ₂ -Pro CV and CO ₂ -Pro ATM (Pro-Oceanus, Canada)	< 0.5% xCO ₂	Constantin Frangoulis	PO-ATM1
			Mark Barry	PO-ATM2
			Mark Barry	PO-CV
	SAMI CO ₂ (Sunburst Sensors, USA)	< 3 μ atm	Anna Rutgersson	SAMI1
			Martii Honkanen	(SAMI2)¶

*Instrument did not function properly; it was damaged during transport.

†Instruments did not function properly and data were not used in the comparison of different instruments.

‡Instrument worked in the beginning but then drifted off scale.

§Instrument's data were off scale and not used for comparison.

||During the intercomparison, the PI was Jack Triest.

¶Instrument was not calibrated before the experiment and data were not used for comparison.

intercomparison and were systematically reviewed and improved during the exercise.

Because of the restrictions due to the COVID-19 pandemic, only a small number of personnel were allowed on site at the

Vlaams Instituut voor de Zee (VLIZ) to perform the physical setup and operate the instruments (i.e., 4 non-VLIZ personnel +3 VLIZ employees). However, there was frequent communication with all PIs, in order for the setup to be optimal and

reduce the risk of failure. The manual, documents, and Quick Start Guides that accompanied all systems proved very helpful both during the setup and the main experiment.

All PIs and users had access to the data from their instruments and the auxiliary data (water temperature and salinity as well as environmental conditions in the experiment room) via online file sharing. Data were updated daily, which allowed the PIs to monitor their instruments' performance throughout the experiment (see "Daily data discussion" section).

Description of participating $p\text{CO}_2$ instruments

A detailed description of the instruments is given in the Supporting Information. Here, we give a short overview of the instruments that were used during the IC2021. The focus of this study was the comparison of the final processed $p\text{CO}_2$ data, that is, the $p\text{CO}_2$ in the tank at the tank's water temperature as measured by each instrument. Practical aspects, such as deployment and duration in the field, were not part of the evaluation. The reader is asked to refer to the manufacturer and/or manual to obtain this information. Note that some instruments malfunctioned and occasionally delivered erroneous data. The parts of compromised instruments' data were excluded from further analysis.

Surface

(Instruments that are often deployed on buoys and AUVs and are not meant to be submerged.)

ASVCO₂: The ASVCO₂ is a climate quality $p\text{CO}_2$ instrument using in situ air–water equilibration. It also measures atmospheric CO₂. It uses a Nafion dryer for drying the measurement gas. CO₂ is measured using a LI-COR® gas analyzer that uses a zero and span gas for calibration in the field.

CaPASOS: The CaPASOS instrument uses in situ air–water equilibration and can measure atmospheric CO₂. It is equipped with three reference gases for calibration and uses a LI-COR gas analyzer for CO₂ detection.

Electrode-CO₂: It is an underwater $p\text{CO}_2$ sensor based on ISFET-pH sensor. The pH sensor can detect $p\text{CO}_2$ change as inner solution pH change caused by permeation of CO₂ gas through the membrane.

Hagan-GENx (formerly known as VeGAS): The Hagan-GENx CO₂ system functions similarly to the ASVCO₂ using a LI-COR gas analyzer for CO₂ detection. It uses a Nafion dryer for drying the measurement gas and a thermally regulated optical bench. It also uses a zero and a span gas for calibration.

MAPCO₂: The MAPCO₂ system is using the same method as described for the ASVCO₂ using in situ air–water equilibration with a different equilibrator design. It also measures atmospheric CO₂. It uses a Nafion dryer for drying the measurement gas. CO₂ is measured using a LI-

COR gas analyzer that uses a zero and span gas for calibration in the field.

Underway, air–water equilibration

(Instruments using air–water equilibration that are normally deployed on ships where water is pumped from a seawater inlet through the instrument.)

General Oceanics 8050 with LI-7000 (GO-LI7000): The GO model 8050 is designed for unattended operation vessels and requires a free-flowing gravity drain. The system uses a shower equilibrator and a LI-7000 IR detector. The system uses several reference gases for calibration.

General Oceanics 8050 with LI-7815 (GO-LI7815): A 2nd GO 8050 system was modified to use a LI-COR LI-7815, a laser-based gas analyzer. This system was a prototype but now is commercially available as the GO 8060 system.

SuperCO₂: The SuperCO₂ system uses a double showerhead equilibrator chamber together with a LI-COR gas analyzer. It also uses several reference gases for calibration.

VLIZ: The VLIZ system is a custom made (noncommercial) system using a marble equilibrator, which is coupled to a Picarro cavity ring-down spectroscopy analyzer. A six-port selector valve is used to switch between sample or a calibration standard gas.

XMU-CO₂-UW: The XMU-CO₂-UW system is a custom-made underway system using a novel water–gas equilibrator based on a jet aerator. It uses a LI-COR for CO₂ detection and reference gases for calibration.

Underway, membrane equilibration

(Instruments using a membrane for equilibration that are normally deployed on ships where water is pumped from a seawater inlet through the instrument.)

AFT-CO₂: The AFT-CO₂ is a commercially available flow-through, autonomous instrument that measures $p\text{CO}_2$ spectrophotometrically. CO₂ in seawater equilibrates across a silicone membrane to an internal solution containing a pH-sensitive dye. The pH change is sensed and translated into $p\text{CO}_2$. The instrument is factory calibrated.

4H-Jena, Contros HydroC CO₂ FlowThrough (ControsFT): The HydroC CO₂-FT is a flow-through version of the below-described HydroC CO₂.

Dal-SOOP: The Dalhousie underway system is composed of an SOOGuard FerryBox System with two flow-through chambers. CO₂ equilibrates via a membrane and is measured through an IR sensor.

SubCTech (SCT): The OceanPack™ measurement principle is based on the diffusion of dissolved CO₂ in water through a semi-permeable membrane across which the CO₂ is equilibrated in a small air volume and then analyzed by a LI-COR sensor. It uses zero and span gas for calibration. The measurement principle is the same for both versions: the Cube and the Race system.

Submersible

(Instruments that can be deployed under water, for example, at moorings.)

Carioca: The Carioca $p\text{CO}_2$ sensor measures $p\text{CO}_2$ spectrophotometrically using a pH indicator. Inside an equilibration cell, dissolved CO_2 equilibrates with the pH indicator solution across a silicon membrane.

4H-Jena, Contros HydroC CO_2 (ControsSUB): The HydroC CO_2 is an optical, headspace-based underwater $p\text{CO}_2$ sensor. The measurement principle is the diffusion of dissolved gases from the ambient liquid through a membrane into an internal gas circuit. The gas concentrations are measured by an IR system. It uses a zero gas for drift correction.

NKE: The NKE is the commercially available sensor of the CARIOCA system.

Pro Oceanus (PO-ATM, PO-CV): Pro Oceanus participated in this activity with three units: two CO_2 -PRO™ ATM and one CO_2 -PRO™ CV. The principle behind both systems is similar and is based on the diffusion of dissolved gas through a hydrophobic membrane into a gas-tight internal gas circulating loop that is connected to an NDIR detector. The main differences between the two systems are the type of membrane and the fact that the ATM version is able to switch between water and atmospheric sampling modes.

SAMI: The SAMI2- CO_2 is an in situ, autonomous instrument that measures $p\text{CO}_2$ spectrophotometrically. CO_2 in seawater equilibrates across silicone membrane tubing to an internal solution containing a pH-sensitive dye. The pH change is sensed and translated into $p\text{CO}_2$. The instrument is factory calibrated.

Tank facility

One of VLIZ's warehouses was modified to host a microcosm seawater facility, equipped with three 5-m^3 (diameter 2 m, height 1.5 m) cylindrical tanks, which are mainly used for small-scale fisheries experiments. The warehouse was temperature regulated ($10\text{--}25^\circ\text{C}$, via an air conditioning/heating unit) and each tank had a dedicated heat exchanger for basic temperature control ($4\text{--}30^\circ\text{C}$, Aqua Medic Titan Professional 8000).

Each tank was connected to additional water treatment and preparation equipment (e.g., protein and fat skimmers, bio-scrubbers). During the preparation phase, we performed a short experiment investigating the suitability of the venue in October/November 2019. During the experiment, we deployed a submersible pump in one of the tanks and connected a GO system to measure the $p\text{CO}_2$ of the tank. In addition, two temperature sensors were deployed and moved around in the tank to test for homogeneity. The temperature of the tank as well as the $p\text{CO}_2$ was altered ($p\text{CO}_2$ changes were induced by adding acid or base) to observe the response and mixing. The conclusion was that the facility was suitable for the experiment, but some minor

modifications were required to optimize it for the needs of the experiment and, most importantly, for the water conditioning and circulation. One point that was identified during this experiment was the temperature hysteresis of the heat exchanger. The heat exchanger could not be regulated to better than $\pm 0.5^\circ\text{C}$, which resulted in a temperature fluctuation around the set point temperature of around 1°C . Since the temperature change was slow (less than 0.4°C h^{-1}) the impact on the intercomparison was considered to be low. During the preparation phase, it was also decided to use only one tank, as it would have been challenging to achieve identical conditions in more than one tank.

An advantage of the facility and setup was the ability to adjust the water conditions within a short period of time. In the preparation test, it was established that a 10°C temperature change can be achieved within 24 h, while CO_2 adjustments of hundreds of μatm (usually $200\text{--}300 \mu\text{atm}$) could be achieved very rapidly with the addition of acid (HCl) or base (NaOH) and allow for a 24-h relaxation period for the water to return to atmospheric equilibrium.

Additional modifications included the isolation of the tank from the rest of the equipment (skimmers, buffer tank, filters), the addition of a small Polyvinylchlorid (PVC) pipe extension on the return of the heat exchanger, which improved the water circulation and prevented the introduction of large air bubbles to the water. A platform, temporary wooden panels, and a wooden A-frame were built to host all instruments and sensors (Fig. 1).

This setup made it possible to have all submersible instruments in the same tank and have a central centrifugal submersible pump (Lowara 8SC3/09/5 C L20 DE 22–24) that supplied the underway instruments. The IC2021 water distribution system was designed in such a way that all instruments received the water flow as specified by manufacturers (Fig. 1B) while warming between the tank and the instrument was minimized. The average warming between the tank and the connected underway instruments was less than 0.3°C . Where necessary, the system allowed for gravity/free drainage (return to the tank) for the systems. The flow rate to every system was regulated via a metering ball valve (GF type 523). The water distribution system was built with flexible PVC hoses (diameter $10\text{--}25 \text{ mm}$) as this allowed for quick changes and modifications. All hoses and connections were continuously checked for leaks and pressure losses. A pressure-retaining valve (GF, type 586) was attached to the water system to maintain constant back pressure in the main flow path and avoid changes in flow rate when single instruments were temporarily disconnected (e.g., maintenance).

The submersible sensors were distributed within the tank and located as close to the submersible pump as possible (Fig. 2). The surface instruments were installed close to the wall of the tank since they needed to be adjusted to the water level.

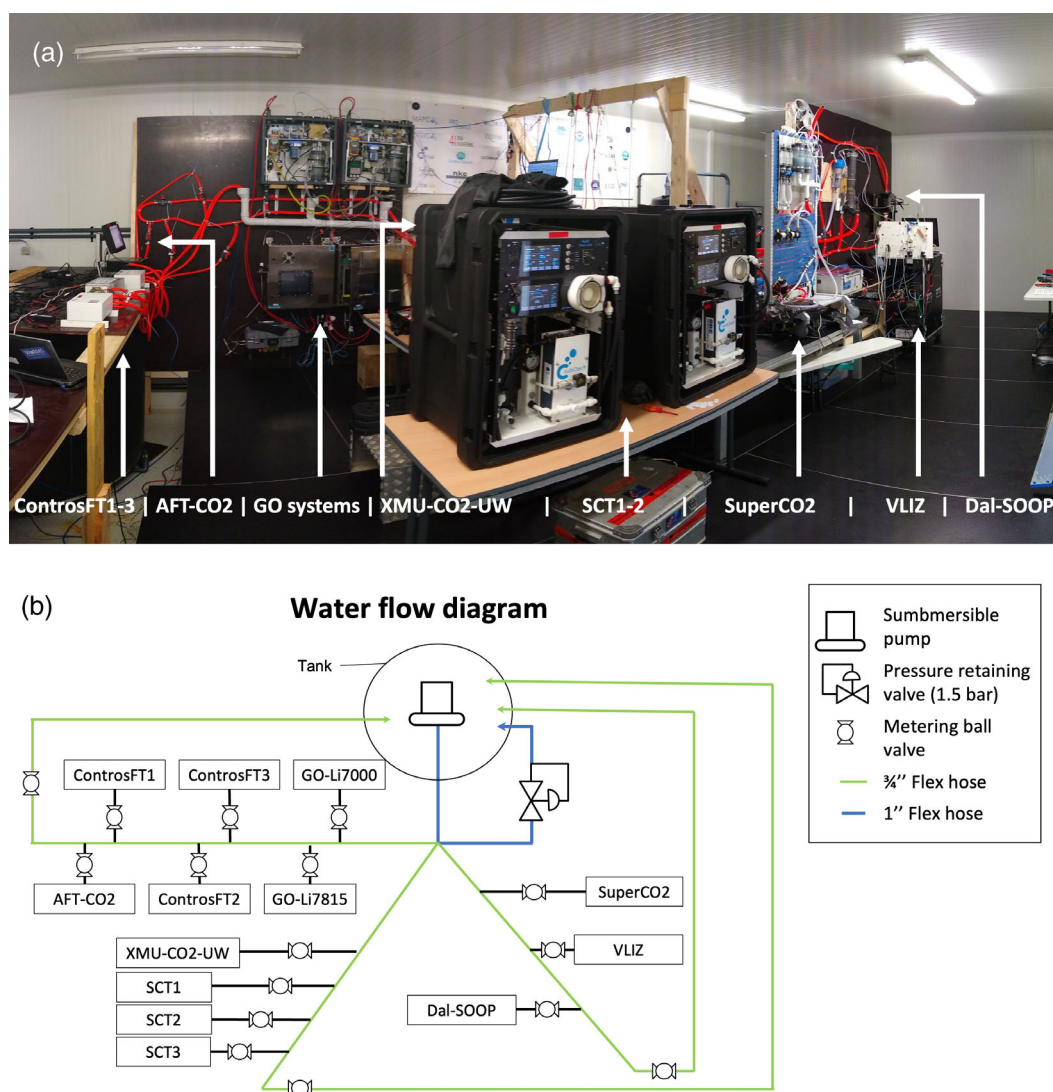


Fig. 1. Setup of instruments during the IC2021 exercise. **(A)** Panoramic view with all underway instruments. In the background, the wooden A-frame for the submersible instruments can be seen. The A-frame is above the water tank. **(B)** Water flow diagram of underway system.

We provided a set of high-quality reference gases (referenced to World Meteorological Organization X2007 scale, uncertainty better than 0.05 ppm CO_2) for all systems that used reference gases for calibration. Two sets of 6 reference gases (nominal mixing ratio/mol fraction: 0, 150, 300, 420, 500, 800 ppm CO_2) were provided by the ICOS Flask and Calibration Laboratory at the Max Planck Institute for Biogeochemistry in Jena/Germany (<https://www.icos-cal.eu/fcl>). While one set of gases was connected, the other set was kept as backup. Supporting Information Table S1 shows the certified values and uncertainty of all reference gases. The same type of pressure regulator was used for all cylinders (SCOTT 20-05-114CD14). The PI of each system decided beforehand how many and which reference gases their system would use, which was necessary to design the gas distribution system (Supporting Information Fig. S1). Stainless steel 1/8" tubing and

appropriate Swagelok® fittings, connectors, and valves were used for this distribution system. After the assemblage, the gas distribution system was tested for leaks, which were addressed before starting the main experiment. The system was tested periodically, and during the experiment, no gas leaks were identified.

Auxiliary data

All groups were provided with additional data that some instruments needed for their $p\text{CO}_2$ calculations or to monitor the environmental conditions. In total, there were five water temperature sensors (see Fig. 2): one Seabird SBE37 (measuring temperature and salinity [conductivity] inside the tank) was placed next to the submersible pump close to the middle of the tank; two Seabird SBE38 temperature sensors were placed near the surface of the tank, and two further SBE38

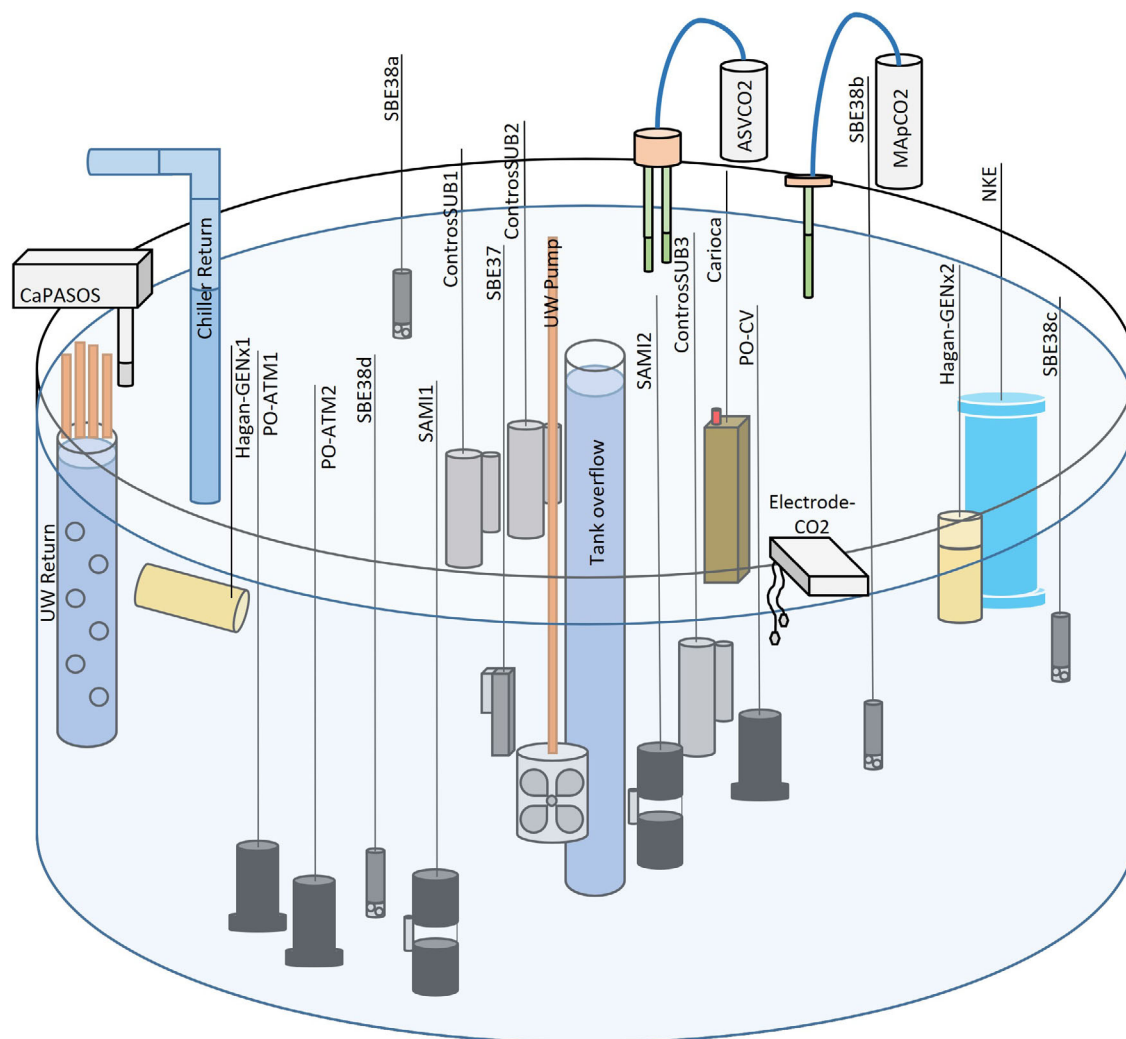


Fig. 2. Schematic of the distribution of surface and submersible sensors in the tank. The water flow path between the underway (UW) pump and the UW return is shown in Fig. 1.

temperature sensors were placed approximately 50 cm above the bottom of the tank. All temperature sensors were calibrated and tested for internal consistency before the exercise. The SBE37 was calibrated for temperature and conductivity before the intercomparison. In addition to the water variables, we provided the room temperature and the ambient pressure using a Vaisala PTB-200 pressure sensor that was calibrated before the exercise (measurement uncertainty 0.02 kPa). All auxiliary data were recorded in 10 s intervals.

The temperature recorded by the SBE37 was used by all instruments as “in situ temperature” and all CO₂ measurements were corrected to and reported at this temperature. The data from the four SBE38 sensors were used to monitor the heterogeneity of the water in the tank. Missing temperature data from periods of up to 1 min were interpolated by linear time-dependent interpolation between the nearest temperature measurements. We identified six intervals where

temperature readings were missing for longer than a few minutes (ranging from 4 min to 15 h) due to a computer failure during night time. For these time intervals, data from a thermosalinograph (Seabird SBE45) attached to the SubCtech system SCT1 were used to reconstruct the tank temperature using a multiple linear regression. The difference between the SBE45 and SBE37 measurements was ca. 0.1°C. For all time intervals, an multiple linear regression was done using the temperature and water flow recordings from SCT1 and the temperature recordings from the SBE37:

$$T_{\text{SBE37}^*} = a \times T_{\text{SCT1}} + b \times WF_{\text{SCT1}} + c \times T_{\text{SCT1}} \times WF_{\text{SCT1}} + d \quad (2)$$

where T_{SBE37^*} is the estimated tank temperature, T_{SCT1} the temperature readings from the SBE45 included in SCT1, and WF_{SCT1} is the water flow at SCT1. The mean difference between SBE37* and measured data from SBE37 and the

standard deviation were calculated for 30 min before and after the interval of missing data where data from both instruments (SBE37 and SB45) were available. The estimated tank temperature has an uncertainty of better than $\pm 0.02^\circ\text{C}$, which translates into a $p\text{CO}_2$ uncertainty of $0.4 \mu\text{atm}$ for a $p\text{CO}_2$ of $420 \mu\text{atm}$.

During these six time intervals, the salinity readings from the SBE37 were also missing. Compared to temperature, salinity has a minor effect on $p\text{CO}_2$. The SBE45 data from SCT were available for the whole time of the experiment; therefore, the times where data for both instruments (SBE37 and SBE45) were available were used to correct the salinity data from SBE45 with the calibrated SBE37. The corrected salinity data from the SBE45 were then used as the tank salinity. The accuracy is estimated to be better than 0.1 PSU. Missing salinity data from intervals of up to 1 min were interpolated by linear time-dependent interpolation between the nearest salinity measurements.

During longer time intervals without pressure recording, the data gaps were filled using the atmospheric pressure recorded by one of the GO systems. The difference between the Vaisala pressure sensor in the room and the Druck pressure sensor in the GO system was less than 0.01 kPa. Missing pressure data from intervals of up to 1 min were interpolated by linear time-dependent interpolation between the nearest pressure measurements.

All auxiliary data were linearly interpolated on a 1 s grid. Thus, temperature, salinity, and pressure data were available at the timestamp of every $p\text{CO}_2$ measurement from all

instruments, and avoided the necessity for data interpolation by individual PIs. We assigned flags to all auxiliary data depending on whether they were measured or estimated. A detailed description of the correction methods and the intervals with missing data is stored together with the data set at the ICOS Carbon Portal (*see* below for data availability) and in the Supporting Information.

During the experiment, not only did the water temperature vary, but also the room temperature. Figure 3 illustrates data from the auxiliary sensors, showing the different conditions the $p\text{CO}_2$ sensors were exposed to. During the setup phase and until 01 July, the water temperature was set at 19°C . The variability of approximately 1°C seen during that interval is related to the heat exchanger's temperature regulation controller (as explained earlier). After 01 July, the temperature was set at different levels within a 20°C range ($10\text{--}30^\circ\text{C}$). Within that interval, we tried to maintain intervals of stable temperature; times when the temperature changed as fast as possible, and instances/intervals with slow temperature changes.

All temperature sensors (Fig. 2) agreed within 0.01°C , which showed that the tank was well mixed, and we could assume homogeneity. The salinity was stable at approx. 31.5 PSU until 04 July. On 04 July, the water was replaced with water of approx. 34.5 practical salinity unit (PSU). The increase up to 34.8 seen in the last days of the experiment might be associated with minor evaporation of the seawater in the tank related to the higher water temperature during that interval. This is also supported by the changes that were made

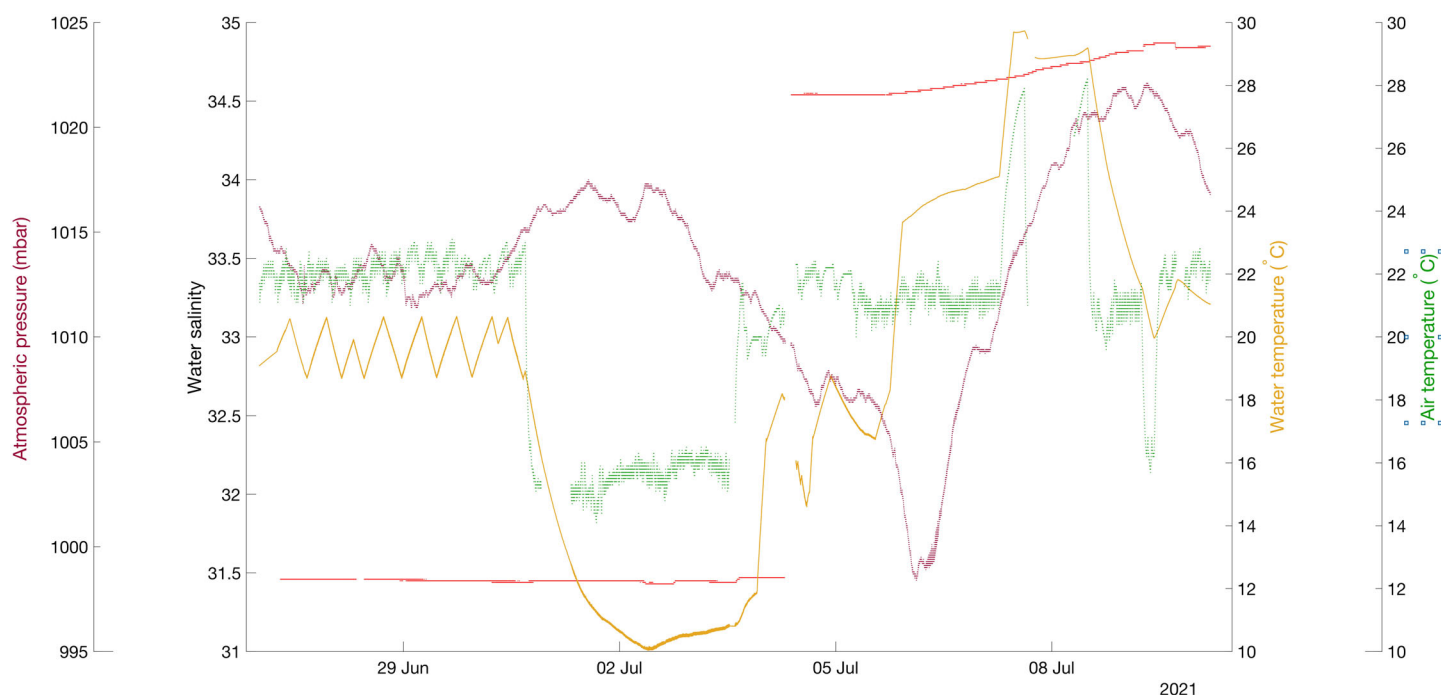


Fig. 3. Auxiliary measurements throughout the experiment: seawater salinity in the tank (black), seawater temperature in the tank (gold), air temperature inside the experiment facility (green) and atmospheric pressure (purple). The x-axis shows the date in 2021.

in the room temperature. As seen in Fig. 3, until 01 July, the ambient room temperature was kept at approx. 22°C. Between 01 July and 03 July, the room temperature was set at 15°C. Between 04 July and 07 July, the room temperature was set back to 22°C, and on 07 July, it was attempted to increase the temperature to 28°C, but this was proven too challenging for the air conditioning/heating of the room to maintain. The variability of the atmospheric pressure is due to local meteorological conditions.

Determination of an “Intercomparison Reference”

One objective of the IC2021 was to assess the performance of state-of-the-art seawater $p\text{CO}_2$ measuring instruments. As Certified Reference Materials for seawater $p\text{CO}_2$ measurements do not exist, we resorted to a different method of estimating the “reference” $p\text{CO}_2$ in the tank based on the measurements from selected participating instruments, which we called the “Intercomparison Reference” (hereby IC-Reference). These methods are not appropriate to get the true value of $p\text{CO}_2$ but it was the community agreed consensus on the best estimate of the true value. When choosing instruments that could help us calculate the IC-Reference, we used criteria described in the Standard Operating Procedures (SOP) in Dickson et al. (2007) and Pierrot et al. (2009): We only considered instruments that (a) were correctly functioning for at least 50% of the time of the experiment, (b) used at least two nonzero gas standards for internal calibration, and (c) had a measurement frequency of at least once every 2 min. We calculated the difference between all the instruments that fulfilled (a), (b), and (c) and chose the ones that agreed for 80% of the data set better than $\pm 2 \mu\text{atm}$ to establish the IC-Reference. Following all the selection criteria, four of the underway air–water equilibrator instruments (Table 2) fulfilled the requirements. Out of these, the Super CO_2 had unstable readings due to the intermittent blocking of its pressure sensor, so it was excluded, leaving the two GO instruments and the custom-built VLIZ instrument as the three instruments that were used for the IC-Reference calculation. This choice was agreed upon by the community participating in the daily meetings (see below). We linearly interpolated the data from these instruments over every second for the duration of the experiment to ensure having a reference value for every data point from all participating instruments. Data in the interpolated products where the time separation between two consecutive measurements by an instrument was greater than 2 min were removed. Such measurement gaps were caused by the instrument performing an internal calibration or by the instrument being turned off for troubleshooting. The IC-Reference was calculated as the mean of any of the three interpolated products present at any one point in time, with a flag given according to the number of instruments included in creating the mean. We believe the choice of instruments for the IC-Reference calculation was appropriate, especially since the standard deviation of the mean when all three interpolated products were available was

never higher than $2 \mu\text{atm}$. The choice of IC-Reference was supported by the measurements of the other “surface” systems ASV CO_2 , Map CO_2 , and Hagan-GENx (Table 2). These instruments were not included in the IC-Reference creation as their measurement interval is higher than 2 min, but their data fall within the $\pm 2 \mu\text{atm}$ band around the IC-Reference. The IC-Reference was therefore used as the closest possible estimation of the real tank conditions, and the instrument performance of all participating instruments was assessed as the difference from it.

Daily data discussion

To monitor the progress and guarantee the success of the IC2021, every interested participant was invited to a daily meeting where the newest data were presented and the plan for the next days was discussed. Before each meeting, the data for every sensor was processed using QuinCe (<https://quince.science>), a web-based tool developed by ICOS Ocean Thematic Centre for standardized processing and quality control of data from a variety of ocean CO_2 sensors, and converting the results into a standard format. This produced a complete set of data which could easily be analyzed and plotted for review. The data processing routines for different sensors were of differing levels of maturity: some well-established sensors had code audited by the community and/or sensor manufacturers to produce publication-quality data, while several of the newer sensors required the development of experimental routines specifically for the intercomparison. Consequently, the quality of the daily data was not the same for all sensors, and it was used only to inform the discussions and allow PIs to see that their sensors were producing usable data and identify possible issues as quickly as possible. Each PI could also access the data from their sensor to perform their own analysis of their sensor’s ongoing performance. To ensure fairness in the comparison of tank $p\text{CO}_2$ data, we used data processed by the PIs using their own routines or improved and agreed upon processing routines.

Time interval used for intercomparisons

In order to find suitable time intervals where the $p\text{CO}_2$ was stable enough to be able to compare the readings of different instruments, we chose intervals where the $p\text{CO}_2$ of the IC-Reference did not change by more than $4 \mu\text{atm h}^{-1}$, and starting at least 1 h after acid or base additions. This resulted in 11 intervals at different $p\text{CO}_2$ levels and temperatures. We chose 6 intervals out of these 11 to represent combinations of high, medium, and low $p\text{CO}_2$ levels and high and low temperatures. Table 3 and Fig. 4 show the chosen time intervals that are used in the remainder of the study to compare the instruments. The last column also shows the color that was used to differentiate the six intervals in the following plots.

Table 3. Time intervals that were chosen for detailed comparison of the instruments. The intervals cover the observed range in $p\text{CO}_2$ and temperature (cf. Figure 4). For the temperature, we choose the tank temperature as described above. The values for $p\text{CO}_2$ are the IC-Reference $p\text{CO}_2$.

Interval	Start time	End time	Duration	Mean temperature (min/max), °C	IC-reference mean $p\text{CO}_2$ (min/max), μatm	
1	01 Jul	02 Jul	24 h	12	425	Mid $p\text{CO}_2$
	07:50	08:00		(10.1/12.8)	(421/429)	Low SST (light blue)
2	02 Jul	02 Jul	7 h	10	260	Low $p\text{CO}_2$
	10:15	17:15		(10.0/10.4)	(254/276)	Low SST (light green)
3	03 Jul	03 Jul	13 h	11	630	High $p\text{CO}_2$
	08:16	21:30		(10.7/11.9)	(621/677)	Low SST (light red)
4	06 Jul	06 Jul	7 h	25	470	Mid $p\text{CO}_2$
	10:30	17:00		(24.5/24.7)	(462/479)	High SST (dark blue)
5	06 Jul	07 Jul	9 h	25	280	Low $p\text{CO}_2$
	20:45	06:00		(24.8/25.1)	(277/310)	High SST (dark green)
6	07 Jul	08 Jul	8 h	29	730	High $p\text{CO}_2$
	20:30	04:30		(28.9/28.9)	(726/744)	High SST (dark red)

Assessment (results)

The whole experiment lasted only 2 weeks, which limits the conclusions that can be drawn. The presented dataset does not allow for quantifying instrument drift that might be

observed during regular deployments that can last up to several months. Also, it was not the scope of this study to examine response times of instruments.

In the following assessment, not all instruments are included. Some instruments did not function properly, and

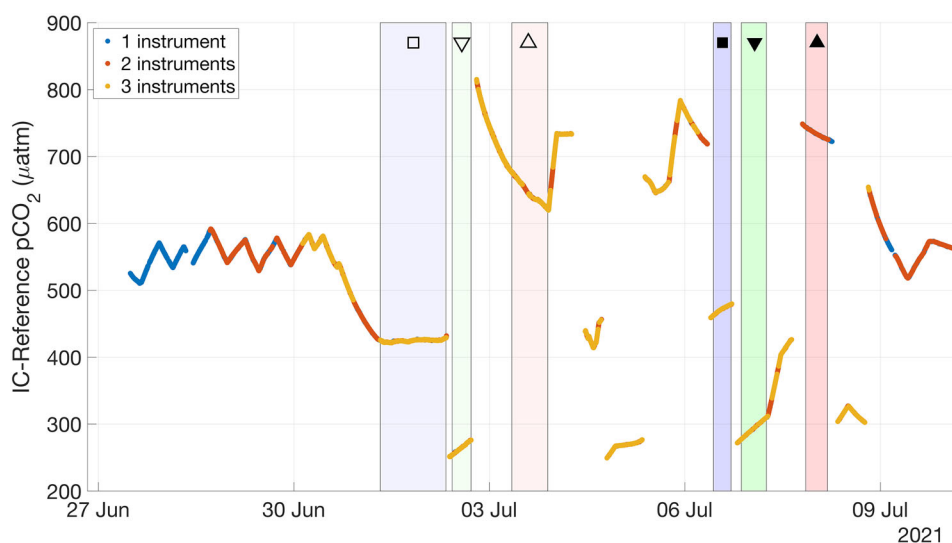


Fig. 4. The IC-Reference $p\text{CO}_2$ throughout the experiment, color-coded by the number of instruments that were used for the averaging at that particular point in time. Availability of < 3 instruments was due to either instruments not being ready, temporarily malfunctioning, or performing a calibration. We chose relatively stable intervals of low (green, downwards triangle symbol), medium (blue, square symbol) and high (red, upwards triangle symbol) $p\text{CO}_2$ to use as comparison intervals. Each of these had an instance when the tank seawater temperature was low (lighter shades, open symbols) or high (darker shades, filled symbols). The x-axis shows that date in 2021.

we decided to exclude instruments when a user could identify them as obviously erroneous without a 2nd dataset for comparison (e.g., $p\text{CO}_2$ values higher/lower than 500 μatm compared to what was expected in this environment):

1. CaPASOS: The system was damaged during transport. We tried to fix the broken parts during the exercise, but the equilibrator was irreparably damaged.
2. AFT: During the 1st week of the exercise, the AFT CO_2 instrument produced reasonable data. Around 02 July, the instrument drifted significantly (more than 60 μatm per day) and the data after that were not used for comparison. On 03 July, the indicator outflow from the AFT was directed outside of the main waterflow to avoid any effects on other instruments. This seems to have allowed air bubbles to form within the AFT optical cell, causing poor precision and accuracy in the measurements. The outflow was placed in a flask of water on 05 July to produce back pressure and prevent the formation of air bubbles. This improved measurement precision, but the accuracy did not improve.
3. SAMI2: The instrument was not calibrated before it was sent to participate at the IC2021. The observed $p\text{CO}_2$ values were obviously too high.
4. CARIOCA: The data were consistently several hundreds of μatm too high. After discussion with the PI, we concluded that the instrument might not be suitable for the water conditions during the experiment (i.e., coastal North Sea water). We speculated that organic acids might diffuse through the membrane and alter the indicator pH

independently from the seawater $p\text{CO}_2$ values. This needs to be confirmed in further experiments.

5. NKE: The instrument showed similar behavior as the CARIOCA instrument; and since they use the same measurement principle, we assume that the same is true for the NKE instrument.
6. Electrode- CO_2 : The instrument showed unrealistic values (e.g., negative $p\text{CO}_2$) from the beginning of the experiment. We assume that the pH calibration on site was not sufficient.

Results of intercomparison $p\text{CO}_2$ IC-Reference

The $p\text{CO}_2$ record during the experiment is presented in Fig. 4. This record shows $p\text{CO}_2$ data from the IC-Reference. As described above, the IC-Reference was created from three instruments. However, data were not available from the three instruments at all times, in which case we used the average of only two instruments. This is indicated in Fig. 4, and for five out of the six chosen intervals, all three instruments were available for the majority of the time. In the beginning of the experiment, the $p\text{CO}_2$ inside the tank was between 500 and 600 μatm which corresponds to the CO_2 concentration in the room. The observed zig-zag variation in the beginning is due to warming and cooling of the tank water. During this time, the $p\text{CO}_2$ was not altered by chemical additions. The decrease on 30 June corresponds to the cooling of the tank (Fig. 3) and the abrupt changes in the observed $p\text{CO}_2$ correspond to the addition of NaOH (decreasing $p\text{CO}_2$) or HCl (increasing $p\text{CO}_2$). The observed $p\text{CO}_2$ ranged from 250 to

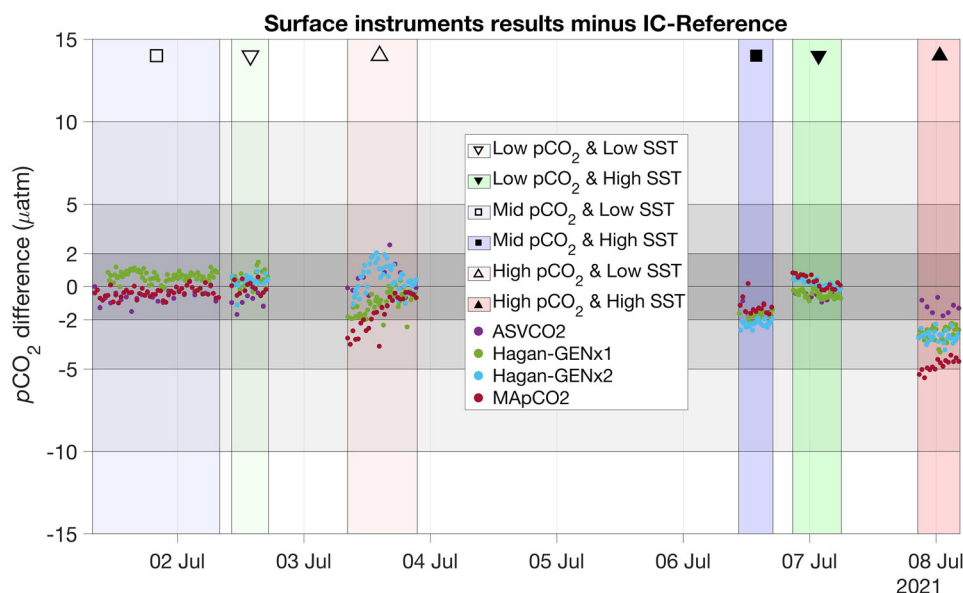


Fig. 5. The difference from the IC-Reference $p\text{CO}_2$ for the surface instruments measurements at the six chosen comparison intervals (vertical shading and symbols). Positive (negative) values represent $p\text{CO}_2$ overestimation (underestimation) compared to the IC-Reference. The ± 2 , ± 5 , and ± 10 intervals are also shown (horizontal shading). The CAPASOS and Electrode- CO_2 results are not shown. Note the y-scale being -15 to $+15$ μatm . The x-axis shows that date in 2021.

slightly over $800 \mu\text{atm}$ covering nearly the whole range of $p\text{CO}_2$ that can be observed in marine surface waters.

As described in “Methods” section, we estimate the uncertainty of the IC-Reference to be less than or equal to $2 \mu\text{atm}$. This means that differences smaller than $2 \mu\text{atm}$ between the IC-Reference and a specific instrument cannot be resolved in the intercomparison.

Instruments' responses in relation to the IC-Reference

Surface instruments

For the comparison of the surface instruments, CaPASOS and Electrode- CO_2 were removed from the analysis since they were not working properly (see above), resulting either in noisy data with differences of several tens of μatm between two consecutive data points or in negative $p\text{CO}_2$ data. The differences in $p\text{CO}_2$ of the remaining four instruments from the IC-Reference are presented in Fig. 5. For the selected comparison intervals, the difference is within $\pm 5 \mu\text{atm}$ and most of the time all four systems fall within $\pm 2 \mu\text{atm}$ of the IC-Reference. These results show that these systems used as autonomous sensors on moorings and AUVs can deliver $p\text{CO}_2$ data of climate quality under varying conditions. The main deviations from the IC-Reference occur in intervals with high $p\text{CO}_2$, especially in combination with high sea surface temperature (SST). This could be explained by the systems using only one nonzero gas standard, which for IR detectors can be a limiting factor as their response is nonlinear. Some variability at high SST may also be an issue, which might be due to elevated water vapor interfering with the CO_2 detection during times with high SST. Furthermore, at high $p\text{CO}_2$, the difference from the IC-Reference for the duration of an individual comparison

interval is more variable and often shows drift compared to the lower $p\text{CO}_2$ intervals.

Underway air–water equilibration instruments

All air–water equilibration systems show good agreement with the IC-Reference and show a deviation of less than $5 \mu\text{atm}$ during most of the time (Fig. 6). Three of the five underway equilibrators instruments available during the experiment were used in the IC-Reference calculation. The mean of the two GO instruments and the custom VLIZ instrument always had a standard deviation $< 2 \mu\text{atm}$. It is therefore expected that these three instruments have a difference from the IC-Reference of $< 2 \mu\text{atm}$ at all times. The XMU- CO_2 -UW fell within the ± 2 difference band for three of the comparison intervals and within ± 5 at all other times. The Super CO_2 measurements were mostly within $\pm 5 \mu\text{atm}$ of the IC-Reference and displayed high-frequency noise of about $4 \mu\text{atm}$ peak to peak during the 1st three intervals. The high-frequency noise was likely due to water droplets in the tubing connecting the equilibrator to the pressure sensor, and replacing the tubing resulted in less noisy data, although the higher SST intercomparisons toward the end of the experiment were typically different by about ± 2 to $\pm 5 \text{ ppm}$ from the IC-Reference.

Underway membrane instruments

Of the 8 underway membrane sensors deployed in the exercise, only the three SCT sensors used nonzero reference gases to characterize the response of the NDIR sensor. All membrane-based sensors showed greater deviation from the IC-Reference than any of the air–water equilibrator-based systems (Fig. 7). During the 1st comparison interval, the data of

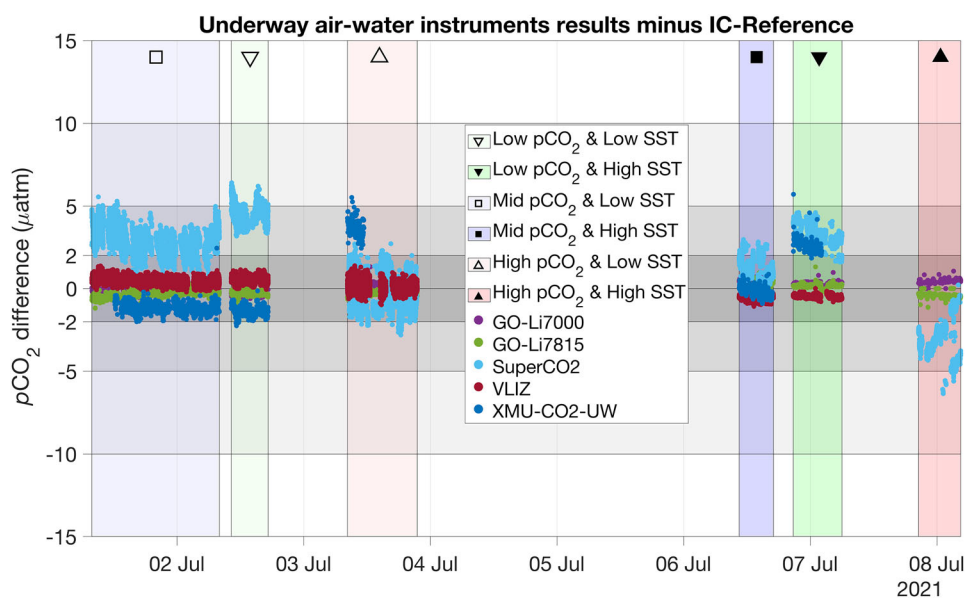


Fig. 6. The difference from the IC-Reference $p\text{CO}_2$ for the underway air–water equilibration instruments measurements at the six chosen comparison intervals (vertical shading and symbols). Positive (negative) values represent $p\text{CO}_2$ overestimation (underestimation) compared to the IC-Reference. The ± 2 , ± 5 , and ± 10 intervals are also shown (horizontal shading). Note the y-scale is -15 to $+15 \mu\text{atm}$. The x-axis shows the date in 2021.

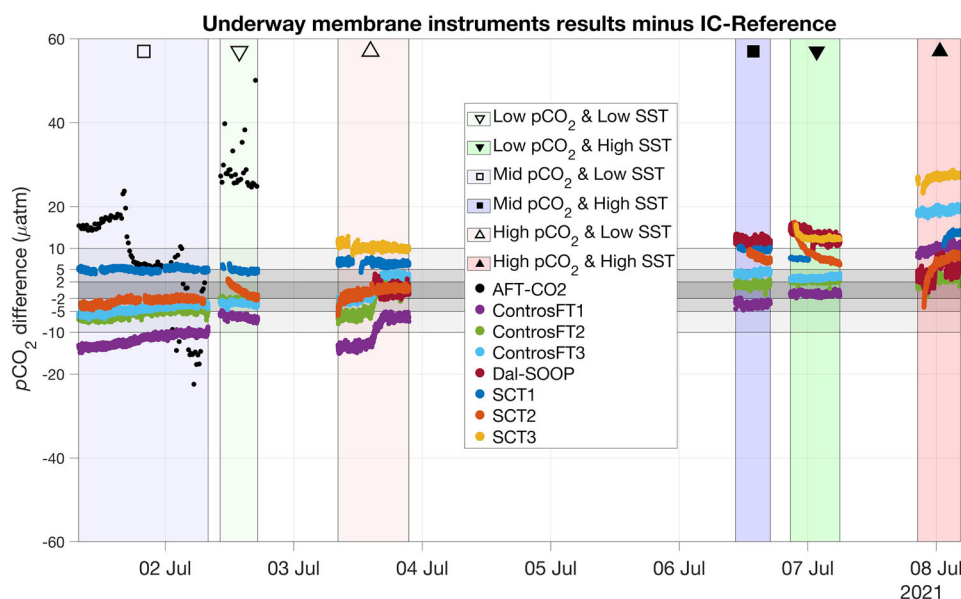


Fig. 7. The difference from the IC-Reference $p\text{CO}_2$ for the underway membrane-equilibration instruments measurements at the six chosen comparison intervals (vertical shading and symbols). Positive (negative) values represent $p\text{CO}_2$ overestimation (underestimation) compared to the IC-Reference. The ± 2 , ± 5 , and ± 10 intervals are also shown (horizontal shading). The AFT- CO_2 results are not shown after 03 July due to potential contamination of tank seawater by the dye. Note the y-scale being -60 to $+60$ μatm . The x-axis shows that date in 2021.

the AFT- CO_2 instrument were within ± 20 μatm of the IC-Reference; but after the 2nd comparison interval, the difference between the sensor and IC-Reference has increased, and the performance of the sensor was unstable, likely due to the change in location of the indicator outlet, as described above. The DAL-SOOP was not fully installed until the 3rd

comparison interval, after which it was always within 20 μatm of the IC-Reference and within 10 μatm during the two high- $p\text{CO}_2$ intervals.

The SCT1 and SCT2 differences were for most of the exercise within 20 μatm from the IC-Reference, with many intervals being within 10 μatm . The SCT3 was included in

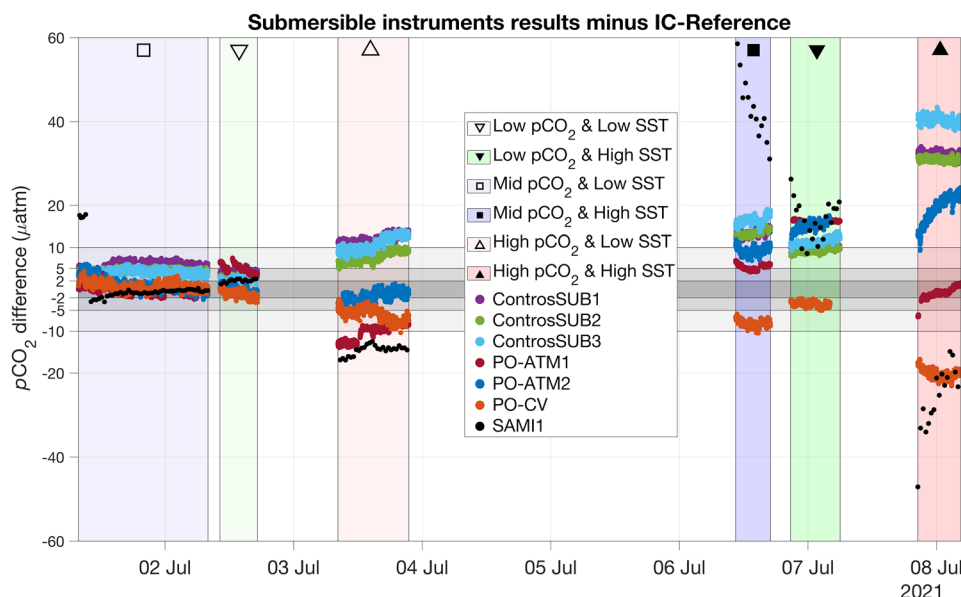


Fig. 8. The difference from the IC-Reference $p\text{CO}_2$ for the submersible instruments' measurements at the six chosen comparison intervals (vertical shading and symbols). Positive (negative) values represent $p\text{CO}_2$ overestimation (underestimation) compared to the IC-Reference. The ± 2 , ± 5 , and ± 10 intervals are also shown (horizontal shading). The Carioca results are not shown. Note the y-scale being -60 to $+60$ μatm . The x-axis shows that date in 2021.

the exercise at the 3rd interval and had similar behavior to SCT1 and SCT2, except for the last interval. Also, for the SCT2, after the 1st interval, and for all other intervals, there is an indication of drift/hysteresis signal. The ControsFT1 underestimated the $p\text{CO}_2$ at the start of the experiment but measured within the ± 10 band for four of the comparison intervals, including all of the high SST ones. ControsFT2 was always within the ± 10 band. The ControsFT3 was within the ± 10 band during all intervals apart from the last one, when it overestimated the $p\text{CO}_2$. All ControsFT instruments displayed an increase in their measured $p\text{CO}_2$ relative to the IC-Reference during the comparison interval on

03 July, which coincides with a rapid increase in sensor temperature while the SST was warming.

Submersible instruments

With the exception of some outliers from the SAMI1 at the start of the 1st comparison interval, all submersible instrument differences from the IC-Reference were within the ± 10 band during the 1st two comparison intervals (Fig. 8). After the 1st two comparison intervals, the Contros SUB sensors overestimated the $p\text{CO}_2$. PO-CV was within the ± 5 band during the 1st two intervals; then within the ± 10 band for the next three, but underestimated $p\text{CO}_2$ by around $20 \mu\text{atm}$ during

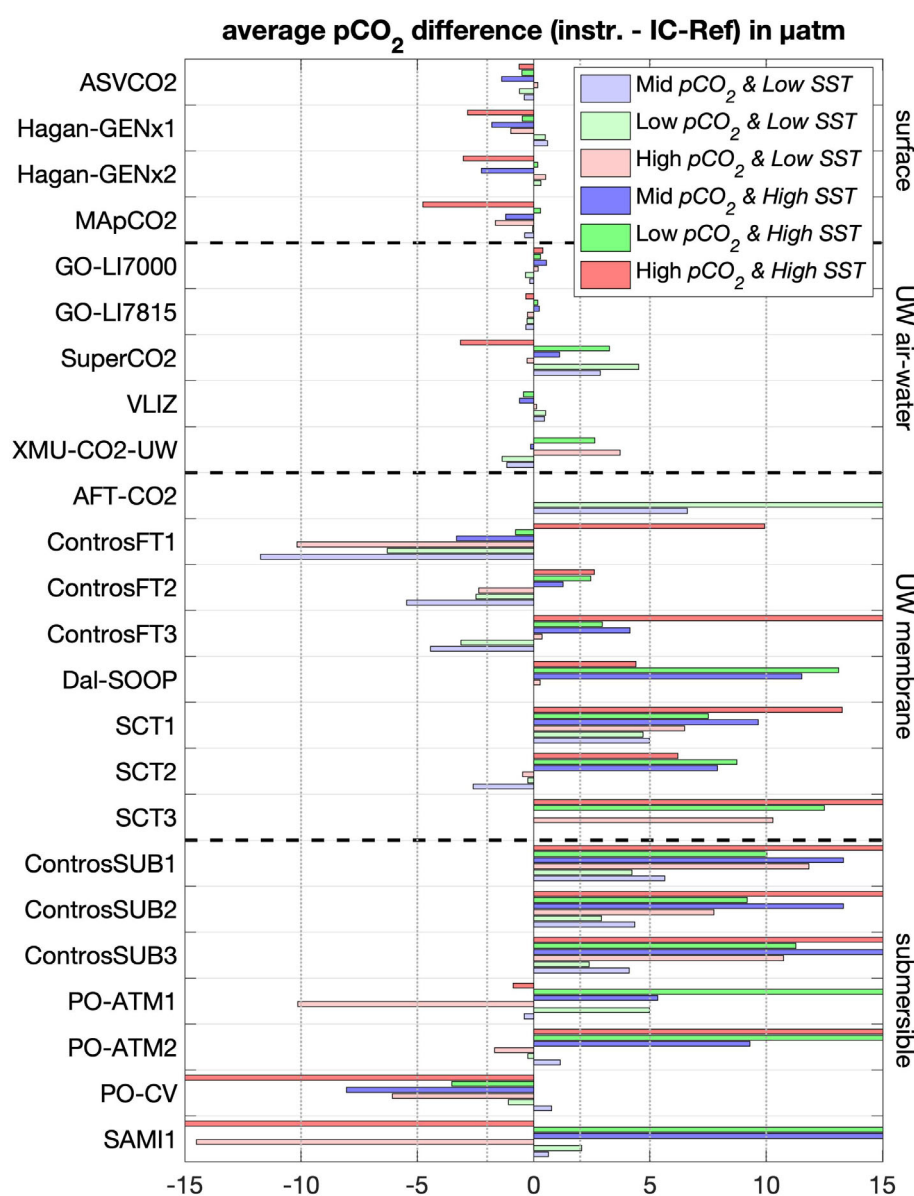


Fig. 9. Statistical summary showing the mean $p\text{CO}_2$ difference from the IC-Reference for each of the six measurement intervals defined in Fig. 4, ordered by instrument per group. The horizontal dashed lines divide the instruments' groups while the vertical gray dashed lines denote the 2, 5, and $10 \mu\text{atm}$ limits.

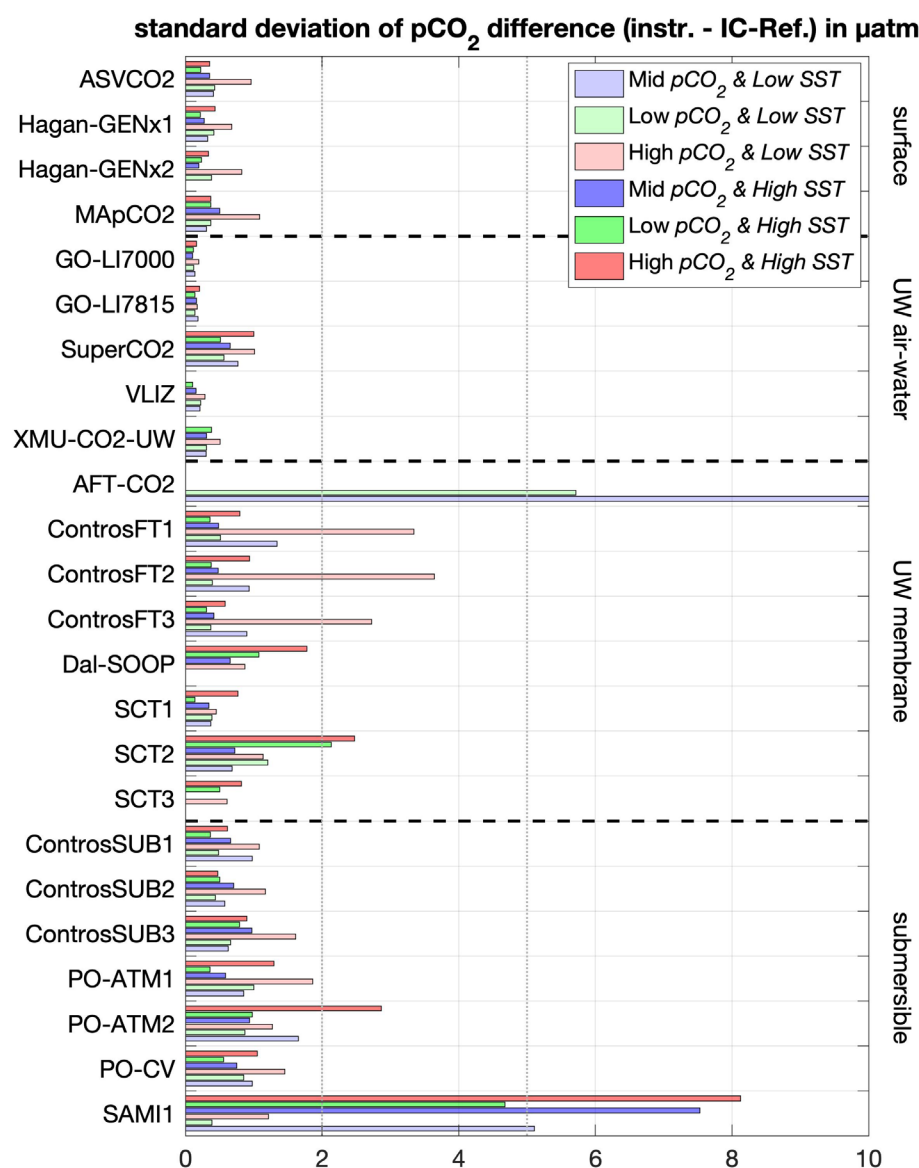


Fig. 10. Standard deviation of the mean differences shown in Fig. 9. The horizontal dashed lines divide the instruments' groups while the vertical gray dashed lines denote the 2, 5, and 10 μatm limits.

the last interval. PO-ATM1 overestimated $p\text{CO}_2 > 10 \mu\text{atm}$ during the low $p\text{CO}_2$ and high SST interval, but was the only submersible instrument to measure within ± 2 of the IC-Reference during the high $p\text{CO}_2$ and high SST interval. PO-ATM2 was mostly within the ± 5 band during the low SST comparison intervals, but overestimated the $p\text{CO}_2$ by > 10 during the high SST intervals. The SAMI1 instrument had a lower measurement frequency; during the 3rd comparison interval, the results were underestimated by 10 to 20 μatm while the high SST comparison intervals were either under- or overestimated by $> 20 \mu\text{atm}$.

Figures 9 and 10 show the average difference between each instrument and the IC-Reference for each chosen interval and

the respective standard deviation over the six time intervals. For each interval, we averaged the $p\text{CO}_2$ difference during the whole interval. The intervals were chosen during stable periods, and therefore the mean differences do not include artifacts due to different instrument response times during abrupt changes in $p\text{CO}_2$, assuming that response times of all instruments are faster than 1 h. The standard deviation represents the repeatability or consistency of the difference from the IC-Reference and each instrument. In practical terms, the way the standard deviation is used here can be defined as the noise or precision of the instrument, and the average difference is the accuracy. Instruments using air–water equilibration (surface, underway air–water) stay within $\pm 5 \mu\text{atm}$ around the

Table 4. Percentage of all evaluated systems that measured within a given uncertainty band at different $p\text{CO}_2$ and temperature levels.

$p\text{CO}_2$ level	Temperature	$\pm 10 \mu\text{atm}$	$\pm 5 \mu\text{atm}$	$\pm 2 \mu\text{atm}$
Low (150–350 ppm)	Low: 10°C	95%	68%	41%
	High: 25°C	61%	57%	35%
Mid (350–600 ppm)	Low: 12°C	86%	57%	38%
	High: 24°C	68%	50%	32%
High (> 600 ppm)	Low: 11°C	70%	52%	30%
	High: 28°C	57%	38%	14%

IC-Reference, and the majority even within $\pm 2 \mu\text{atm}$ (Fig. 9), with their standard deviation is always below $2 \mu\text{atm}$ (Fig. 10). However, three instruments (2 Hagan-GENx and MAPCO₂) of the surface category show an increased difference from the IC-Reference during times of high SST, while the same was not observed in the standard deviation. This might be due to the use of only one span gas or thermal interferences within the system's electronics unit or CO₂ detector. The underway membrane instruments showed both a higher mean difference from the IC-Reference as well as a higher standard deviation. Among the submersible instruments, the Contros systems consistently measured higher than the IC-Reference, the PO-CV mostly measured lower, while the other submersible instruments mainly measured lower at high $p\text{CO}_2$. They all had similar standard deviations (most of the time less than $\pm 2 \mu\text{atm}$, less than ± 3 at high SST high $p\text{CO}_2$), apart from the SAMI1 (up to $8 \mu\text{atm}$).

Table 4 gives a general overview of systems' performances during the exercise. Here, we did not differentiate between different types of instrumentation. For any given $p\text{CO}_2$ level, more systems measured within an uncertainty band at low temperatures vs. high temperature comparisons. More than half of the participating instruments stayed within $\pm 10 \mu\text{atm}$ for all $p\text{CO}_2$ and temperature levels, but only 30% within $\pm 2 \mu\text{atm}$ with the exception of high $p\text{CO}_2$ and high temperature conditions where only 14% were within $\pm 2 \mu\text{atm}$.

For three instrument types (GO, SCT, and ControsFT) there were multiple instruments participating, giving potentially more insight into systematic biases of these instruments. While the GO systems showed differences of less than $2 \mu\text{atm}$ to the IC-Reference and low standard deviations (below $2 \mu\text{atm}$), the membrane-based systems SCT and ControsFT show differences to the IC-Reference ranging from $-14 \mu\text{atm}$ to more than $15 \mu\text{atm}$. The SCT systems overestimate the $p\text{CO}_2$ during most intervals, while the ControsFT systems do not show a clear pattern.

Discussion

Measurement technology has improved over recent years, with new instruments becoming commercially available and

others were further developed from a prototype stage and are now also commercially available (e.g., MAPCO₂, Hagan-GENx, SCT3). There are more instruments with different technologies available for the community to observe surface ocean $p\text{CO}_2$ than a decade before (Clarke et al. 2017, hardware directory for $p\text{CO}_2$ at International Ocean Carbon Coordination Project's website: www.ioccp.org). Instruments (e.g., GO instruments) that were already present during previous intercomparisons (Table 1) and showed good results (uncertainty better than $2 \mu\text{atm}$), performed well also during this intercomparison. This showcases the reliability of these instruments. The aim of this exercise was to better understand how different $p\text{CO}_2$ measurement systems can meet the demands of different applications for which they have been selected. Within this scope, the current uncertainty requirements for data submitted to Surface Ocean CO₂ Atlas (SOCAT) (different flags for $p\text{CO}_2$ uncertainty better than 2, 5 and $10 \mu\text{atm}$; Bakker et al. 2016) and the Global Ocean Observing System Essential Ocean Variable definitions ($p\text{CO}_2$ uncertainty better than 2 or $5 \mu\text{atm}$ depending on the platform) provided a concrete performance evaluation framework. It is worth reiterating that this exercise evaluates the quality of $p\text{CO}_2$ measurement techniques alone under controlled conditions in the laboratory. Additional sources of uncertainties and errors associated with in-field operation and sampling determine how representative the measurements are of the surface ocean $p\text{CO}_2$, but are not explicitly addressed in this work. The combined $p\text{CO}_2$ uncertainty (measurements and sampling) is likely to be higher, and the results described here represent a rather optimistic estimate and are limited to $p\text{CO}_2$ between 260 and $730 \mu\text{atm}$ and water temperature between 10 and 28°C . With this in mind, and looking into the information in Table 4, more than half of the systems in this exercise were able to provide measurements with an uncertainty better than $\pm 10 \mu\text{atm}$. This does not necessarily imply that the other 50% of the instruments are not capable of that. Our results show the status in July 2021 and many groups and manufacturers have already used these results to consider improvements. In addition, we advise the reader to carefully consult the instrument-specific descriptions. Some systems showed a higher difference due to problems that are easy to fix (e.g., some benchtop systems that measure the water temperature at equilibration to correct for warming of the water had uncalibrated temperature sensors and thus they might perform better after the necessary adjustments). An important finding from this exercise was that all systems need a certain degree of knowledge and expertise in order to produce optimal quality data. Especially membrane-based systems are easier to operate in the field but the data processing needs critical understanding of the instrument's characteristics. Attention to detail is required by the user in the entire data life cycle—from system setup over data processing to final data quality control.

It is important to state that one needs understanding of the instrument as a whole. Accurately measuring the temperature and pressure of equilibration is as critical as the $x\text{CO}_2$ measurement. Optimal equilibration depends on equilibrator design, seawater flow, having a leak-tight system, and materials that do not contaminate the sample. One of the main messages of this work is that: Surface ocean $p\text{CO}_2$ measurements are still not trivial!

A major finding of this study comes from Fig. 9 and shows that equilibration through direct air–water contact (e.g., Sabine et al. 2020; Frankignoulle et al. 2001; Johnson 1999) appears to be more consistent and independent of external factors than equilibration through a membrane or photometric detection. There were 9 instruments with air–water equilibrators, shown in the upper 2 sections of Fig. 9, and every one of these performed better than any of the 15 other instruments. For shipboard systems, better results are obtained with equilibrators using direct air–water equilibration compared to membrane systems, but the former tend to be larger and more expensive. For sensors that can be deployed directly on the surface or subsurface platforms, the bubble equilibrators generally give better results than membranes, but the former are limited to surface water while the latter can be used at lower depths. This suggests that the equilibration in air–water equilibrators is well understood and for the detection of CO_2 having at least one nonzero reference gas helps to reduce measurement uncertainty. This clear difference between air–water and membrane equilibration is not observed in the standard deviation (Fig. 10), which can be read as instruments' noise or precision. With the exception of the AFT- CO_2 and SAMI1, all instruments have a standard deviation of less than $5 \mu\text{atm}$ for all intervals and most of the instruments even less than $2 \mu\text{atm}$. The highest standard deviation is observed during intervals with high $p\text{CO}_2$ values and/or high water temperature.

Fifteen systems show the highest difference to the IC-Reference (Fig. 9) during intervals with high water temperature and high $p\text{CO}_2$. The air–water equilibrators all have chilled condensers that reduce the water vapor content of sample streams to a low, constant level. Underway systems shown in Fig. 5 use either Nafion dryers to reduce water vapor levels or in case of the use of the laser-based CO_2 detector LI-7815, the water vapor is measured very precisely, and the CO_2 data are corrected for that. However, the water vapor amount is still relatively high at higher SST. The membrane-based systems (Figs. 7, 8) have no sample air drying (or spectrophotometric detection, to which this argument does not apply) and show the greatest increase in variability at higher SST. The results show that underway systems with condensers are not influenced by SST variations, some SST dependence for those systems with Nafion dryers, and the highest dependence for those with no sample drying. While there is no proof that this relationship is causal, it indicates some instrument types are likely to deliver better data than others in warm waters. One

can speculate that temperature affects the degree of equilibration for membrane-based systems, but it could also be an insufficient correction for water vapor. Temperature and water flow are the two factors that affect the equilibration of all systems (Johnson 1999) but some instruments can cope with that better than others.

A reason for the better performance (smaller difference from the IC-Reference) of air–water equilibration systems could be due to them utilizing one or more reference gases for calibration. However, the three SubCtech underway membrane systems also have multiple reference gases but show higher differences than the air–water equilibration systems. This suggests that there are still uncertainties in the equilibration process across membranes and/or accurately measuring the pressure of equilibration in membrane systems. Furthermore, the quality of the membrane might change over time and can affect the degree of equilibration. For the IR membrane-based systems that do not use nonzero reference gases, the absence of reference gases and consequently the impossibility to optimally characterize the behavior of the IR adds to the observed differences from the IC-Reference $p\text{CO}_2$. It is important to state that all NDIR membrane-based systems use an internally generated zero gas to control the sensor drift. Most membrane systems are small instruments designed for submersible pressure housings and necessarily use lower quality sensors (NDIR, temperature and pressure) than those typically found in shipboard underway systems. The three types of surface systems with air–water equilibration (ASVCO₂, Hagan-GENx, and MApCO₂), however, do manage climate quality with the lower accuracy sensors, but use one nonzero gas for span adjustment and Nafion drying of the gas streams.

Comparing instruments of the same type (e.g., GO, ControsFT, SCT) can be an indicator for systematic overestimation or underestimation of $p\text{CO}_2$ of certain instruments, which impacts its performance. While the GO systems showed low standard deviations (below $2 \mu\text{atm}$), the membrane-based systems SCT and ControsFT show differences to the IC-Reference ranging from $-14 \mu\text{atm}$ to more than $15 \mu\text{atm}$. The SCT systems overestimate the $p\text{CO}_2$ during most intervals, while the ControsFT systems do not show a clear pattern. It was not the aim of this exercise to understand the instrument specific response to the $p\text{CO}_2$ signal rather than documenting it.

Pressure measurements for pressure compensation of IRs can also be challenging, requiring highly accurate measurements at both the detector and the point of equilibration (or the ambient atmosphere in the case of atmospheric measurements) $p\text{CO}_2$ is the product of $x\text{CO}_2$ and pressure, so the pressure measurements must be at least as accurate as the CO_2 measurement for a given accuracy target, or more accurate if one would like to make the contribution of pressure errors negligible in the overall uncertainty budget. For climate quality data, this implies a pressure sensor accuracy requirement of the order of 0.5 hPa for equilibrator measurements

and 0.05 hPa for atmospheric measurements. Membrane systems operate over a much wider pressure range than air–water equilibrators (Clarke et al. 2017) and rely on measuring the equilibration air pressure in a small closed loop while a diaphragm pump is running. The pump can cause noisy pressure readings; depending on the placement of the pressure sensor in the gas loop and how it is sampled, the mean of these readings may not represent the effective mean equilibration pressure.

All instruments that rely on equilibrating seawater with air, whether directly or through a membrane, also require accurate measurement of the temperature of equilibration. In the 1996 intercomparison of underway systems on RV Meteor (Körtzinger et al. 2000), the largest source of differences between systems was found to be poorly calibrated temperature sensors, and this remains a problem nearly 30 yr later. One underway system's equilibrator temperature sensor had errors up to 1°C (these data were re-calculated after the experiment), leading to large errors in in situ $p\text{CO}_2$ that were corrected in post-processing after calibrating the sensor. Another underway system recorded equilibrator temperature with a resolution of only 0.05°C, leading to noisy in situ $p\text{CO}_2$ data: $p\text{CO}_2$ changes by 4.2% per°C of temperature change, so at a $p\text{CO}_2$ of 480 μatm , having a temperature sensor resolution of 0.05 C will introduce step changes of 1.0 μatm into the data, and an error of 1°C in temperature sensor calibration will lead to a $p\text{CO}_2$ error of 20 μatm . Some other systems had no temperature sensor in the equilibrator (or membrane housing) but relied on temperature measurements from nearby sensors. This might be sufficient when the water flow rate is high and the external temperature measurement is very close. However, missing temperature measurements at the point of equilibration always add to increased uncertainty of the final $p\text{CO}_2$ data. This should also be considered when comparing the results for the “surface” systems. They were not equipped with their own temperature measurement and used the closest temperature sensor in the tank. However, the observed differences of the temperature sensors in the tank were less than 0.05°C. Spot checks of temperature sensors from various underway systems with a reference thermometer showed that most were not accurate enough to provide climate-quality $p\text{CO}_2$ data.

In the case of systems based on spectrophotometric detection, these rely on measuring the pH of a dye solution that is equilibrated with seawater through a membrane, and the conversion from pH to $p\text{CO}_2$ brings additional uncertainties (Orr et al. 2018). This makes this challenging for climate-quality measurements.

The above-mentioned characteristics need to be considered when deploying membrane and spectrophotometric systems. These are the only sensors that can currently be used for underwater applications, as they are smaller and generally easier to operate without extensive expert training. Applications

in dynamic coastal or reef environments, where the amplitude of variability is generally higher, do not necessarily require the same quality of measurements as the open-ocean climate research, and the choice of the instrument should be guided by the scientific question they are addressing.

For all systems, it is important to have an adequate set of diagnostic variables (e.g., flow rates). The set of these variables differs from instrument to instrument, but the more information the user has during data processing, the easier it is to evaluate the state of the instrument and the resulting quality of the data. Continuous documentation and transparency of an instrument's maintenance are crucial for uncertainty estimates.

When using bulk water conditions for $p\text{CO}_2$ and temperature in flux calculations, further uncertainty can be introduced. The bulk water conditions are known to deviate under certain conditions (e.g., wind, rainfall, skin temperature) from near-surface values that govern the flux of CO_2 (Parc et al. 2024; Renfrew et al. 2002; Watson et al. 2020).

During the intercomparison we also discussed the water used (natural coastal water from the North Sea) for the experiment. The use of acid and base to adjust $p\text{CO}_2$ might have impacted the AFT- CO_2 and SAMI1. Even if we cannot exclude that biofouling happened during the tank experiment since we were using natural seawater, all systems should have been impacted in the same way. Biofouling can also alter results, especially for near-surface applications. Biofouling can occur on the installation frame and cause a microenvironment that is different from the surrounding water, but it also can happen close to the point of equilibration such as in the membranes. Due to the high flow rate through the chiller, the water in the tank was in constant movement, so we can exclude the build-up of stagnant water films in front of the systems. The resistance of a particular instrument to drift caused by biofouling was not evaluated during the present intercomparison exercise.

We caution use of these intercomparison exercises as the sole criteria by the community to decide which instrument to choose. All instruments were run in a controlled environment for only 2 weeks with close supervision. The observations from this experiment may differ from “real-life” conditions with unattended deployments of up to several months on ships, buoys, or other ocean platforms. We can only speculate that this will add to the observed differences between instruments. A different experiment, for example, a dedicated cruise covering different biogeochemical regions, would be useful to provide an additional intercomparison in more realistic conditions.

For the comparison intervals we explicitly excluded times of rapid changes (addition of acid/base, heating/cooling). However, all instruments recorded data during these times, and the data are available in the data file. But it was not the aim of this exercise to evaluate response times of different instruments.

We encourage the users to carefully consider which instrument is best suited for the specific application. This includes (1) performance of the instrument, (2) the expected environment and scientific question, (3) pre- and post-deployment handling of the instrument and data, and (4) the experience of the deployment personnel and time available to operate the instruments. Also, the measurement frequency achievable by each instrument needs to be taken into account. While some instruments can record data with 1 Hz, others measure one data point every hour. It was not within the scope of this exercise to evaluate measurement intervals. For open ocean applications, 1 Hz data might be too fast and only results in large data files without gaining extra insight, while hourly measurements might be too slow for applications where environmental changes that occur on shorter time scales need to be resolved.

Recommendations

The following recommendations and remarks conclude the findings of the 2021 pCO₂ intercomparison exercise:

- Air–water equilibration systems that closely follow community-agreed SOPs (e.g., Dickson et al. 2007; Pierrot et al. 2009) in general provide data that are capable of measuring seawater pCO₂ with climate quality.
- Instruments being able to measure CO₂ in air (xCO₂) directly were in better agreement with the IC-Reference.
- Instruments are available with different levels of complexity and different data quality. However, when aiming to observe open ocean surface pCO₂ for climate research, air–water equilibration following the SOPs are advisable.
- Systems with laser-based detectors (Picarro and LI-7815) performed consistently well during the intercomparison, operating with an air–water equilibrator and following SOPs.
- One of the results of the intercomparison of Körtzinger et al. (2000) was that incorrect temperature readings (either due to miscalibration or low quality) were one of the main sources of differences between the instruments. During the 2021 intercomparison, we observed the same issue again. It is crucial to use temperature (and pressure) sensors that are regularly calibrated and that measure with the appropriate uncertainty.
- High water temperature and high pCO₂ values caused a larger difference from the IC-Reference for most instruments. This shows the importance of calibrating the instruments for the expected range in temperature and CO₂. But it also points out that a more detailed understanding, especially for membrane-based systems, is needed.
- Additional experiments with individual systems are needed to elucidate the role of temperature, the frequency of span gas calibration, and the quality of pressure measurements, and so forth on instrument performance.

- The 2021 intercomparison is a one-time snapshot under controlled and strongly varying conditions over short time periods often not encountered in many instrument deployments. However, the results taken in context of the protocols of the experiment can serve as a guide to expected performance in the field.
- These pCO₂ instrument intercomparisons should be repeated regularly to test the evolution of existing instruments and new and emerging technology. A next step would be conducting a similar experiment under real-world conditions: (1) a ship-board intercomparison under different environmental conditions for underway instruments and (2) deployments at different environmental conditions of surface and submersible instruments. This should include different salinities.

Author Contributions

Conceptualization: Tobias Steinhoff, Thanos Gkritzalis, Craig Neill, Ute Schuster, and Siv K. Lauvset. Data curation: Tobias Steinhoff, Thanos Gkritzalis, Steve Jones, Vlad A. Macovei, and Craig Neill. Formal analysis: Tobias Steinhoff, Thanos Gkritzalis, Steve Jones, Vlad A. Macovei, Craig Neill, Ute Schuster, John Akl, Dariia Atamanchuk, Mark Barry, Laurence Beaumont, Carolina Antoni, Jana Fahning, Jac Fought, Constantin Frangoulis, Lucía Gutiérrez-Loza, Clinton Hagan, Martti Honkanen, Sami Kielosto, Nadja Kinski, Arne Körtzinger, Peter Landschützer, Noah Lawrence-Slavas, Quanlong Li, Anna Luchetta, Damien Malarde, Melf Paulsen, Markus Ritschel, Anna Rutgersson, Kiminori Shitashima, Reggie Spaulding, Natalia Stamataki, Ken Stenbäck, Adrienne Sutton, Witold Tatkiwicz, Hannelore Theetaert, Bronte Tilbrook, and Rik Wanninkhof. Funding acquisition: Siv K. Lauvset and Richard Sanders. Investigation: Tobias Steinhoff, Thanos Gkritzalis, Craig Neill, and Ute Schuster. Methodology: Tobias Steinhoff, Thanos Gkritzalis, Craig Neill, and Ute Schuster. Project administration: Tobias Steinhoff and Siv K. Lauvset. Resources: Tobias Steinhoff, Thanos Gkritzalis, Craig Neill, Ute Schuster, John Akl, Dariia Atamanchuk, Mark Barry, Laurence Beaumont, Carolina Antoni, Jana Fahning, Jac Fought, Constantin Frangoulis, Lucía Gutiérrez-Loza, Martti Honkanen, Noah Lawrence-Slavas, Quanlong Li, Anna Luchetta, Damien Malarde, Melf Paulsen, Markus Ritschel, Anna Rutgersson, Kiminori Shitashima, Reggie Spaulding, Natalia Stamataki, Adrienne Sutton, Witold Tatkiwicz, Maciej Telszewski, and Bronte Tilbrook. Software: Steve Jones. Supervision: Tobias Steinhoff and Thanos Gkritzalis. Validation: Tobias Steinhoff, Thanos Gkritzalis, Steve Jones, Vlad A. Macovei, Craig Neill, Ute Schuster, John Akl, Ricardo Arruda, Dariia Atamanchuk, Mark Barry, Laurence Beaumont, Carolina Antoni, Andrew Dickson, Jana Fahning, Jac Fought, Constantin Frangoulis, Lucía Gutiérrez-Loza, Clinton Hagan, Martti Honkanen, Sami Kielosto, Nadja Kinski, Arne Körtzinger, Peter Landschützer, Noah Lawrence-Slavas, Quanlong Li, Anna Luchetta, Damien

Malarde, Melf Paulsen, Markus Ritschel, Anna Rutgersson, Kiminori Shitashima, Reggie Spaulding, Natalia Stamatakis, Ken Stenbäck, Adrienne Sutton, Witold Tatkiwicz, Hannelore Theetaert, Bronte Tilbrook, and Rik Wanninkhof. Visualization: Tobias Steinhoff, Thanos Gkritzalis, Steve Jones, Vlad A. Macovei, Craig Neill, and Ute Schuster. Writing – original draft: Tobias Steinhoff, Thanos Gkritzalis, Vlad A. Macovei, Craig Neill, and Ute Schuster. Writing – review and editing: Tobias Steinhoff, Thanos Gkritzalis, Steve Jones, Vlad A. Macovei, Craig Neill, Ute Schuster, John Akl, Ricardo Arruda, Dariia Atamanchuk, Mark Barry, Laurence Beaumont, Carolina Cantoni, Andrew Dickson, Jana Fahning, Jac Fought, Constantin Frangoulis, Lucía Gutiérrez-Loza, Clinton Hagan, Martti Honkanen, Nadja Kinski, Arne Körtzinger, Peter Landschützer, Siv K. Lauvset, Noah Lawrence-Slavas, Quanlong Li, Anna Luchetta, Damien Malarde, Melf Paulsen, Markus Ritschel, Anna Rutgersson, Richard Sanders, Kiminori Shitashima, Reggie Spaulding, Natalia Stamatakis, Adrienne Sutton, Witold Tatkiwicz, Maciej Telszewski, Bronte Tilbrook, and Rik Wanninkhof.

Acknowledgments

This Study was financed by the Norwegian Research Council via ICOS Norway (grant numbers 245927 and 296012). The contribution from VLIZ—Flanders Marine Institute: Research Foundation Flanders—FWO contract I001821N supporting ICOS Flanders. VLIZ the tank facilities, the VLIZ UW $p\text{CO}_2$ system, lab facilities for analysis of samples of DIC, TA, pH. Australia's Integrated Marine Observing System supported the participation of Craig Neill, John Akl, and Bronte Tilbrook and the GO systems used in the intercomparison. Integrated Marine Observing System is enabled by the National Collaborative Research Infrastructure Strategy. We thank NORCE Norwegian Research Centre AS (NORCE) for the administrative support. AGD was supported through the US National Science Foundation support OCE-2050280. The authors acknowledge funding from the US National Science Foundation (grant number OCE-2140395) to the Scientific Committee on Oceanic Research (USA) for the International Ocean Carbon Coordination Project. Participants from the University of Exeter were funded through UK NERC CaPASOS project (NE/P020755/1) and the Horizon Europe GEORGE project (1510413). The contribution from Sweden was funded by the Swedish Research Council via ICOS Sweden. Dalhousie University (Halifax/Canada) provided technical support from the CERC.Ocean group and financial support through Canada Foundation for Innovation, project number 29011. We acknowledge funding from the National Key R&D Program of China (2016YFC1400805). National Oceanographic and Atmospheric Administration - Pacific Marine Environmental Laboratory (NOAA-PMEL) provided the ASVCO₂ sensor and technical support from the Engineering and Carbon groups, funded through ASVCO₂ Gen2 IOOS OTT: Technology

Transfer of a Surface $p\text{CO}_2$ Flux Instrument for Autonomous Platform Applications. This is PMEL contribution 5750. We acknowledge funding from the Helmholtz Association EU Partnering “SEA-ReCap – Research capacity building for healthy, productive and resilient seas” project. We thank Sunburst Sensors for the donation of instruments and technical support. We thank GEOMAR for their support with the handling of shipments/custom clearance and the possibility to use their instrumentation during the test phase. The authors thank Emma Dölger and Silke Verbrugge for their technical support and motivation during the intercomparison. Open Access funding enabled and organized by Projekt DEAL.

Data Availability Statement

Raw and processed data are available at the ICOS carbon portal (Steinhoff et al. 2025): <https://doi.org/10.18160/C94W-S2TV>.

References

- Arruda, R., D. Atamanchuk, M. Cronin, T. Steinhoff, and D. W. R. Wallace. 2020. “At-Sea Intercomparison of Three Underway $p\text{CO}_2$ Systems.” *Limnology and Oceanography: Methods* 18: 63–76. <https://doi.org/10.1002/lom3.10346>.
- Bakker, D. C. E., B. Pfeil, C. S. Landa, et al. 2016. “A Multi-Decade Record of High-Quality $f\text{CO}_2$ Data in Version 3 of the Surface Ocean CO₂ Atlas (SOCAT).” *Earth System Science Data* 8, no. 2: 383–413. <https://doi.org/10.5194/essd-8-383-2016>.
- Bender, M., S. Doney, R. A. Feely, et al. 2002. A large Scale Carbon Observing plan: In Situ Oceans and Atmosphere (LSCOP), 201. National Technical Information Service.
- Clarke, J. S., E. P. Achterberg, D. P. Connelly, U. Schuster, and M. Molem. 2017. “Developments in Marine $p\text{CO}_2$ Measurement Technology; towards Sustained In Situ Observations.” *Trends in Analytical Chemistry* 88: 53–61. <https://doi.org/10.1016/j.trac.2016.12.008>.
- Dai, M., L. Zhongming, Z. Weidong, et al. 2009. “Diurnal Variations of Surface Seawater $p\text{CO}_2$ in Contrasting Coastal Environments.” *Limnology and Oceanography* 54. <https://doi.org/10.4319/lo.2009.54.3.0735>.
- DeGrandpre, M. D., T. R. Hammar, S. P. Smith, and F. L. Sayles. 1995. “In Situ Measurements of Seawater $p\text{CO}_2$.” *Limnology and Oceanography* 40: 969–975. <https://doi.org/10.4319/lo.1995.40.5.0969>.
- Dickson, A. G., C. L. Sabine, and J. R. Christian. 2007. Guide to Best Practices for Ocean CO₂ Measurements. Vol. 3, 191. PICES Special Publication. <https://doi.org/10.1159/000331784>.
- Dong, Y., D. C. E. Bakker, and P. Landschützer. 2024. “Accuracy of Ocean CO₂ Uptake Estimates at a Risk by a Reduction in the Data Collection.” *Geophysical Research Letters* 51, no. 9: e2024GL108502. <https://doi.org/10.1029/2024GL108502>.
- Fay, A. R., D. R. Munro, G. A. McKinley, et al. 2024. “Updated Climatological Mean $\Delta f\text{CO}_2$ and Net Sea–Air CO₂ Flux over the

- Global Open Ocean Regions." *Earth System Science Data* 16, no. 4: 2123–2139. <https://doi.org/10.5194/essd-16-2123-2024>.
- Frankignoulle, M., A. Borges, and R. Biondo. 2001. "A New Design of Equilibrator to Monitor Carbon Dioxide in Highly Dynamic and Turbid Environments." *Water Research* 35, no. 5: 1344–1347. [https://doi.org/10.1016/S0043-1354\(00\)00369-9](https://doi.org/10.1016/S0043-1354(00)00369-9).
- Friedlingstein, P., M. O'Sullivan, M. W. Jones, et al. 2023. "Global Carbon Budget 2023." *Earth System Science Data* 15: 5301–5369. <https://doi.org/10.5194/essd-15-5301-2023>.
- Gloege, L., M. Yan, T. Zheng, and G. A. McKinley. 2022. "Improved Quantification of Ocean Carbon Uptake by Using Machine Learning to Merge Global Models and pCO₂ Data." *Journal of Advances in Modeling Earth Systems* 14: e2021MS002620. <https://doi.org/10.1029/2021MS002620>.
- Gruber, N., D. C. E. Bakker, T. DeVries, et al. 2023. "Trends and Variability in the Ocean Carbon Sink." *Nature Reviews Earth and Environment* 4: 119–134. <https://doi.org/10.1038/s43017-022-00381-x>.
- Hauck, J., C. Nissen, P. Landschützer, et al. 2023. "Sparse Observations Induce Large Biases in Estimates of the Global Ocean CO₂ Sink: An Ocean Model Subsampling Experiment." *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences* 308: 20220063. <https://doi.org/10.1098/rsta.2022.0063>.
- Heimdal, T. H., G. A. McKinley, A. J. Sutton, A. R. Fay, and L. Gloege. 2024. "Assessing Improvements in Global Ocean pCO₂ Machine Learning Reconstructions With Southern Ocean Autonomous Sampling." *Biogeosciences* 21: 2159–2176. <https://doi.org/10.5194/bg-21-2159-2024>.
- IPCC. 2022. "Climate Change 2022: Impacts, Adaptation, and Vulnerability." In Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, edited by H.-O. Pörtner, D. C. Roberts, M. Tignor, E. S. Poloczanska, K. Mintenbeck, A. Alegría, M. Craig, S. Langsdorf, S. Löschke, V. Möller, A. Okem, and B. Rama, 3056. Cambridge University Press. <https://doi.org/10.1017/9781009325844>.
- Jersild, A., and P. Landschützer. 2024. "A Spatially Explicit Uncertainty Analysis of the Air–Sea CO₂ Flux from Observations." *Geophysical Research Letters* 51: e2023GL106636. <https://doi.org/10.1029/2023GL106636>.
- Johnson, J. E. 1999. "Evaluation of a Seawater Equilibrator for Shipboard Analysis of Dissolved Oceanic Trace Gases." *Analytica Chimica Acta* 395, no. 1–2: 119–132. [https://doi.org/10.1016/S0003-2670\(99\)00361-X](https://doi.org/10.1016/S0003-2670(99)00361-X).
- Körtzinger, A., L. Mintrop, D. W. R. Wallace, et al. 2000. "The International At-Sea Intercomparison of fCO₂ Systems during the R/V Meteor Cruise 36/1 in the North Atlantic Ocean." *Marine Chemistry* 72: 171–192. [https://doi.org/10.1016/S0304-4203\(00\)00080-3](https://doi.org/10.1016/S0304-4203(00)00080-3).
- Körtzinger, A., H. Thomas, B. Schneider, N. Gronau, L. Mintrop, and J. C. Duinker. 1996. "At-Sea Intercomparison of Two Newly Designed Underway pCO₂ Systems – Encouraging Results." *Marine Chemistry* 52: 133–145. [https://doi.org/10.1016/0304-4203\(95\)00083-6](https://doi.org/10.1016/0304-4203(95)00083-6).
- Laakso, L., M. Honkanen, S. Kielosto, et al. 2019. JERICO-NEXT TNA: Intercomparison of Instruments for Carbonate System Measurements (INTERCARBO). SEANO. <https://doi.org/10.17882/62526>.
- Macovei, V. A., U. Callies, P. H. R. Calil, and Y. G. Voynova. 2022. "Mesoscale Advective and Biological Processes Alter Carbon Uptake Capacity in a Shelf Sea." *Frontiers in Marine Science* 9: 827075. <https://doi.org/10.3389/fmars.2022.827075>.
- Mayot, N., C. Le Quéré, C. Rödenbeck, et al. 2023. "Climate-Driven Variability of the Southern Ocean CO₂ Sink." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 381: 38120220055. <https://doi.org/10.1098/rsta.2022.0055>.
- Newton, J. A., R. A. Feely, E. B. Jewett, P. Williamson, and J. Mathis. 2014. Global Ocean Acidification Observing Network: requirements and governance Plan. 2nd ed. GOA-ON.
- Orr, J. C., J.-M. Epitalon, A. G. Dickson, and J.-P. Gattuso. 2018. "Routine Uncertainty Propagation for the Marine Carbon Dioxide System." *Marine Chemistry* 207: 84–107. <https://doi.org/10.1016/j.marchem.2018.10.006>.
- Parc, L., H. Bellenger, L. Bopp, X. Perrot, and D. T. Ho. 2024. "Global Ocean Carbon Uptake Enhanced by Rainfall." *Nature Geoscience* 17: 851–857. <https://doi.org/10.1038/s41561-024-01517-y>.
- Pierrot, D., C. Neill, K. Sullivan, et al. 2009. "Recommendations for Autonomous Underway pCO₂ Measuring Systems and Data-Reduction Routines." *Deep Sea Research Part II: Topical Studies in Oceanography* 56, no. 8–10: 512–522. <https://doi.org/10.1016/j.dsr2.2008.12.005>.
- Renfrew, I. A., G. W. K. Moore, P. S. Guest, and K. Bumke. 2002. "A Comparison of Surface Layer and Surface Turbulent Flux Observations Over the Labrador Sea With ECMWF Analyses and NCEP Reanalyses." *Journal of Physical Oceanography* 32: 383–400. [https://doi.org/10.1175/1520-0485\(2002\)032](https://doi.org/10.1175/1520-0485(2002)032).
- Sabine, C., A. Sutton, K. McCabe, et al. 2020. "Evaluation of a New Carbon Dioxide System for Autonomous Surface Vehicles." *Journal of Atmospheric and Oceanic Technology* 37, no. 8: 1305–1317. <https://doi.org/10.1175/JTECH-D-20-0010.1>.
- Steinhoff, T., T. Gkritzalis, S. Jones, et al. 2025. The ICOS OTC pCO₂ Instrument Inter-Comparison Data Archive. ICOS-ERIC – Carbon Portal. <https://doi.org/10.18160/C94W-S2TV>.
- Steinhoff, T., T. Gkritzalis, S. K. Lauvset, et al. 2019. "Constraining the Oceanic Uptake and Fluxes of Greenhouse Gases by Building an Ocean Network of Certified Stations: The Ocean Component of the Integrated Carbon Observation System, ICOS-Oceans." *Frontiers in Marine Science* 6: 544. <https://doi.org/10.3389/fmars.2019.00544>.

- Takahashi, T., S. C. Sutherland, R. Wanninkhof, et al. 2009. "Climatological Mean and Decadal Change in Surface Ocean pCO₂, and Net Sea–Air CO₂ Flux over the Global Oceans." *Deep Sea Research, Part II: Topical Studies in Oceanography* 56: 554–577. <https://doi.org/10.1016/j.dsr2.2008.12.009>.
- Wanninkhof, R., P. A. Pickers, A. M. Omar, et al. 2019. "A Surface Ocean CO₂ Reference Network, SOCONET and Associated Marine Boundary Layer CO₂ Measurements." *Frontiers in Marine Science* 6: 400. <https://doi.org/10.3389/fmars.2019.00400>.
- Wanninkhof, R., and J. Triñanes. 2017. "The Impact of Changing Wind Speeds on Gas Transfer and Its Effect on Global Air–Sea CO₂ Fluxes." *Global Biogeochemical Cycles* 31: 961–974. <https://doi.org/10.1002/2016GB005592>.
- Watson, A. J., U. Schuster, J. D. Shutler, et al. 2020. "Revised Estimates of Ocean–Atmosphere CO₂ Flux Are Consistent With Ocean Carbon Inventory." *Nature Communications* 11: 4422. <https://doi.org/10.1038/s41467-020-18203-3>.
- Weiss, R. F. 1974. "Carbon Dioxide in Water and Seawater: The Solubility of a Non-Ideal Gas." *Marine Chemistry* 2: 203–215. [https://doi.org/10.1016/0304-4203\(74\)90015-2](https://doi.org/10.1016/0304-4203(74)90015-2).
- Weiss, R. F. 1981. "Determinations of CO₂ and Methane by Dual Catalyst Flame Ionization Chromatography and Nitrous Oxide by Electron Capture Chromatography." *Journal of Chromatographic Science* 19: 611–616. <https://doi.org/10.1093/chromsci/19.12.611>.

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Submitted 09 April 2025

Revised 07 August 2025

Accepted 08 August 2025