Two decades of flask observations of atmospheric $\delta$(O$_2$/N$_2$), CO$_2$, and APO at stations Lutjewad (the Netherlands) and Mace Head (Ireland), and 3 years from Halley station (Antarctica)

Linh N.T. Nguyen$^1$, Harro A.J. Meijer$^1$, Charlotte van Leeuwen$^1$, Bert A.M. Kers$^1$, H.A. (Bert) Scheeren$^1$, Anna E. Jones$^2$, Neil Brough$^{2,*}$, Thomas Barningham$^2$, Penelope A. Pickers$^3$, Andrew C. Manning$^3$, and Ingrid T. Luijkx$^4$

$^1$Centre for Isotope Research, Energy and Sustainability Research Institute Groningen, University of Groningen, Groningen, the Netherlands
$^2$British Antarctic Survey, Natural Environment Research Council, Cambridge, United Kingdom
$^3$Centre for Ocean and Atmospheric Sciences, School of Environmental Sciences, University of East Anglia, Norwich, United Kingdom
$^4$Meteorology and Air Quality, Wageningen University and Research, Wageningen, the Netherlands

*Now at National Institute of Water and Atmospheric Research, Wellington, New Zealand

Correspondence to: Linh N.T. Nguyen (n.t.l.nguyen@rug.nl)

Abstract. We present 20-year flask sample records of atmospheric CO$_2$, $\delta$(O$_2$/N$_2$) and APO (Atmospheric Potential Oxygen) from the stations Lutjewad (the Netherlands) and Mace Head (Ireland), and a 3-year record from Halley Station (Antarctica). We include details of our calibration procedures and the stability of our calibration scale over time. The results of our inter-comparison involving gas cylinders from various research laboratories worldwide also show that our calibration is of high quality and compatible with the internationally-recognised Scripps O$_2$ scale. The measurement records from Lutjewad and Mace Head show similar long-term trends during the period 2002-2018 of 2.31 $\pm$ 0.07 ppm yr$^{-1}$ for CO$_2$ and -21.2 $\pm$ 0.8 per meg yr$^{-1}$ for $\delta$(O$_2$/N$_2$) at Lutjewad, and 2.22 $\pm$ 0.04 ppm yr$^{-1}$ for CO$_2$ and -21.3 $\pm$ 0.9 per meg yr$^{-1}$ for $\delta$(O$_2$/N$_2$) at Mace Head. They also show a similar $\delta$(O$_2$/N$_2$) seasonal cycle with an amplitude of 54 $\pm$ 4 per meg at Lutjewad and 61 $\pm$ 5 per meg at Mace Head, while the CO$_2$ seasonal amplitude at Lutjewad (16.8 $\pm$ 0.5 ppm) is slightly higher than that at Mace Head (14.8 $\pm$ 0.3 ppm). We show that the observed long-term trends and seasonal cycles are compatible with the measurements from various other stations, especially the measurements from Weybourne Atmospheric Observatory (United Kingdom). However, there are remarkable differences in the progression of annual trends between the Mace Head and Lutjewad records for $\delta$(O$_2$/N$_2$) and APO, which might in part be caused by sampling differences, but also by environmental effects, such as North Atlantic Ocean oxygen ventilation changes to which Mace Head is more sensitive. The Halley record shows clear trends and seasonality in $\delta$(O$_2$/N$_2$) and APO, where especially APO agrees well with continuous measurements at the same location made by the University of East Anglia, while CO$_2$ and $\delta$(O$_2$/N$_2$) present slight disagreements, most likely caused by small leakages during sampling. From our 2002-2018 records, we find good agreement for the global ocean carbon sink: 2.0 $\pm$ 0.8 PgC yr$^{-1}$ and 2.2 $\pm$ 0.9 PgC yr$^{-1}$, based on Lutjewad and Mace Head, respectively. The data presented in this work are available at https://doi.org/10.18160/qq7d-t060 (Nguyen et al., 2021).

1 Introduction

The global carbon cycle is a dynamic system that comprises the exchanges of carbon between various reservoirs and is important for studying human-induced climate change and its impacts (Ciais et al., 2013). Accurate determination of anthropogenic CO$_2$ emissions and their partitioning across different reservoirs plays a vital role in understanding the impact of the remaining atmospheric CO$_2$ mole fraction on climate (Friedlingstein et al., 2020). High-precision atmospheric O$_2$ measurements have been proven to be a strong aide in quantifying CO$_2$ fluxes in the carbon cycle. By combining the decadal trends of atmospheric CO$_2$ and O$_2$, we can quantify the global land and ocean carbon sinks.
(Bender et al., 1996; Keeling and Shertz, 1992; Manning and Keeling, 2006; Tohjima et al., 2019). This is because CO₂ and O₂ cycles are closely coupled – in most processes, there is an anti-correlation in the changes of their mole fraction, except for the oceanic uptake of CO₂ (Manning and Keeling, 2006). To quantify the various components of the global carbon cycle, the changes in atmospheric mole fraction of the two species can be used in combination with their stoichiometric exchange ratio (ER), which is the ratio of CO₂ and O₂ exchanged (consumed/produced) in a process. The ER value varies depending on the process, and is close to 1.1 for photosynthesis/respiration (Severinghaus, 1995) and on average 1.38 for the global mix of fossil fuels (Keeling and Manning, 2014).

There are various techniques to measure atmospheric O₂ to high precision, such as interferometry (Keeling, 1988); mass spectrometry (Bender et al., 1994); paramagnetic analysis (Manning et al., 1999); gas chromatography (Tohjima, 2000); vacuum-UV absorption (Stephens et al., 2003; Stephens et al., 2021); and fuel cell technology (Stephens et al., 2007). Despite many improvements to these techniques over the years, it is still very challenging to obtain O₂ measurements with high accuracy and precision. This is mainly because the atmospheric background mole fraction of O₂ is very high – around 209,392 ± 3 ppm (Tohjima, 2005) – while the observed variations are at the level of a few ppm. These challenges are magnified further for long-term measurements because of possible small biases, drifts or other changes in the analysers or in the calibration scales. Thus the sampling procedures and analysing (laboratory) conditions must be monitored and corrected for by a carefully designed use of calibration and reference gas cylinders over the years (Aoki et al., 2021). As a result, there are only a handful of programmes around the globe which are proficient in coupled CO₂ and O₂ measurements, for example, the network of atmospheric stations maintained by the Scripps Institution of Oceanography (Manning and Keeling, 2006); National Institute of Advanced Industrial Science and Technology (Aoki et al., 2021); National Institute for Environmental Studies (Tohjima et al., 2008); University of East Anglia (UEA) (Pickers et al., 2017); and the University of Groningen (van der Laan-Luijkx et al., 2010). Our laboratory – the Centre for Isotope Research (CIO) of the University of Groningen (RUG) in the Netherlands – has been carrying out flask measurements of CO₂ and O₂ since the early 2000s from various locations (van der Laan-Luijkx et al., 2010). Flask sampling for CO₂ and O₂ has been conducted at Lutjewad (the Netherlands), Mace Head (Ireland), Jungfraujoch (Switzerland) and Halley (Antarctica).

In this paper, we present the O₂ and CO₂ measurements from flasks collected at Lutjewad (the Netherlands), Mace Head (Ireland), both for the period 2000-2020, and Halley (Antarctica) for 2014-2017. From these measurements, a tracer called Atmospheric Potential Oxygen (APO) (the details of which are given in Sect. 2.5) is calculated. We first describe the measurement sites and the sampling procedure as well as the measurement methods, including the calibration procedure. Then we present the data and discuss the trends and seasonality as well as the quality of the datasets. This paper builds on work previously presented in van der Laan-Luijkx et al. (2010), Sirignano et al. (2010), and van Leeuwen (2015).

2 Methods

2.1 Site description

The stations from which our flasks were collected are: Lutjewad Atmospheric Monitoring Station on the northern coast of the Netherlands (53°24'N, 6°20'E) managed by the CIO (RUG); Mace Head Atmospheric Research Station on the western coast of Ireland (53°20'N, 9°54'W) operated by the National University of Ireland’s School of Physics and Ryan Institute Centre for Climate & Air Pollution studies; and (formerly) Halley VI Research Station situated on the
Brunt Ice Shelf (75°34′S, 25°30′W) operated by the British Antarctic Survey. Figure 1 shows the locations of the three stations.

The Lutjewad station is a “class 2” station in the European Union’s Integrated Carbon Observation System (ICOS) network. It comprises a 60-m tall tower, an additional platform of 10-m height, and a laboratory building containing analysers, flask sampling systems, measurement systems and other equipment. Air is pumped from the top of the 60-m tower via inlets connected to a series of tubing towards the laboratory building. The inlet is equipped with a Nafion drying tube (MD 110-72-S, Perma Pure, Toms River, New Jersey) so that the incoming air is first partly dried. The flow in the outer side of the Nafion tube is the outlet of the same air sampling system, after the air is dried with the second stage cryogenic dryer in the laboratory to a dewpoint below -45 °C (Neubert et al., 2004). This ensures that, except for water, all constituents have a negligible gradient over the Nafion membrane. From the inlet, the sampled air is stored in glass flasks via a flask sampling system for further analyses in the CIO laboratories (Neubert et al., 2004). The dominant wind direction in the Netherlands is southwest, meaning that the measurements acquired at the Lutjewad station often represent continental air masses influenced by anthropogenic and biogenic sources and sinks (van der Laan et al., 2010). Otherwise, when the wind comes from the north, the station samples background air that comes from the North Sea and North Atlantic (van der Laan-Luijkx et al., 2010).

The Mace Head station consists of field laboratories and a 20-m tower for sampling. The dominant wind arriving at the station is westerly from the North Atlantic Ocean, carrying air masses that would not have been considerably affected by regional anthropogenic activities. Air masses from other directions carry contamination from local and continental sources (Derwent et al., 2002; Jennings et al., 1993).

The Halley station is a “Global” station within the World Meteorological Organisation’s Global Atmosphere Watch (WMO/GAW) programme, that observes background atmospheric conditions at various locations around the globe. The main Halley station consists of 8 modules that are atop ski-fitted hydraulic legs, within which are the research facilities and living quarters. Air sampling for this project was carried out at the Clean Air Sector Laboratory, which is located 1.5km from the main station in a location that receives minimal contamination from station activities (Jones et al., 2008). The predominant winds are from the east, bringing background air masses from the South Atlantic sector of the Southern Ocean (60%) or from the continental plateaux (30%). Westerly winds that have passed over the Weddell Sea gyre occur 10% of the time (Barningham, 2018; British Antarctic Survey, 2021).
2.2 Flask sampling procedure

At Lutjewad, we employ an automated flask sampling system, hereafter called the autosampler (Neubert et al., 2004). For storing air samples, we use 2.5-litre glass flasks capped with two high-vacuum valves (Louwers, Hapert, NL) sealed with Viton o-rings. Our autosampler is designed to connect to and fill up to 20 flasks without requiring user intervention, and we can remotely control the opening/closing of the flask valves (via custom-made electric motor actuators) and the filling of samples (via a series including a gas pump, flow controllers, and magnetic solenoid valves). The autosampler schedule is controlled via custom-made software (written in Delphi programming language), and carries out the sampling procedure automatically, but it can also be operated remotely using software such as VNC or TeamViewer when needed. A normal filling procedure starts with the air stream being cryogenically dried and flushed through a flask for at least an hour before filling the flask slowly so that the sample remains at atmospheric pressure (to prevent the sample from fractionation and differential permeation through the o-rings cause by a pressure gradient (Sturm et al., 2004)) and moving to the next flask. Individual flasks can be preserved at any time. Samples at Lutjewad are collected under various conditions and time frequencies, but in this paper we present only the data from flasks collected under local background conditions, defined by van der Laan-Luijkx et al. (2010) as flasks taken while the $^{222}$Radon activity monitored at the station was less than 3 Bq m$^{-3}$ and with a CO mole fraction of less than 200 ppb. This filtering procedure is applied to the dataset after the flasks are analysed.

At Mace Head, flasks are collected once or twice per week via a manually operated system as described by Conway et al. (1994), at 35 m above sea level and mostly during restricted baseline conditions (Bousquet et al., 1996). The filter to retain flasks with a CO mole fraction of less than 200 ppb is also applied for flask samples from Mace Head.

At Halley, flasks are collected once a week depending on the meteorological conditions, via a portable manual sampler. This consists of a KNF diaphragm pump (KNF N86), flowmeter, drying agent (magnesium perchlorate), 7µm filter and 3 sampling flasks connected in concession. The air is sampled about 6 metres above the snow surface on the east side of the building via Synflex tubing connected to an aspirated inlet (the details of the aspirated inlet are as described by Blaine et al. (2006)). The system is flushed for about 45 minutes at a flow rate of 2.5 L min$^{-1}$ at atmospheric pressure before each flask is manually closed. The collected samples are stored in insulated aluminium boxes at room temperature until their annual return to the UK on the Antarctic supply ship.

After sample collection, flasks from the three stations are transferred back to our laboratory in Groningen for analysis. Typically, the mole fractions of CH$_4$, CO, CO$_2$ and O$_2$ (reported as $\delta$(O$_2$/N$_2$), see next section) are measured (van der Laan et al., 2009), and additional analyses such as stable isotopes (for example $^{13}$C and $^{18}$O in CO$_2$) and radiocarbon ($^{14}$C in CO$_2$) are also conducted when required (van der Laan et al., 2010).

2.3 CO$_2$ measurement

All flask samples are analysed on an Agilent HP6890N gas chromatograph equipped with a Flame Ionization Detector (referred to as HPGC) to determine the mole fractions of CO$_2$, CO and CH$_4$. The HPGC system has a set-up similar to the GC-systems described by Worthy et al. (2003) and van der Laan et al. (2009). All working standard mixtures (made from dried ambient air) that were used to calibrate the HPGC have been calibrated on the HPGC system at CIO against a suite of 5 primary standards linked to the World Meteorological Organization (WMO) X2007 scale with CO$_2$ ranging between 354 and 426 µmol mol$^{-1}$ (ppm). These primary standards were provided by the Earth System Research Laboratory (ESRL) of the National Oceanic and Atmospheric Administration (NOAA), USA. Since the summer of 2013, working standard gas cylinders were also calibrated for CO$_2$, CO and CH$_4$ mole fractions on a Cavity Ring-Down
Spectrometer (CRDS) model G2401-m from Picarro Inc. using the same suite of primary standards. We refer to Chen et al. (2010) for more details on the CRDS technique. The measurement precision and accuracy for flask measurements of CO₂ on the HPGC are typically <0.06 ppm and <0.07 ppm, respectively (van Leeuwen, 2015).

All CO₂ measurements presented in this paper were originally calibrated against standards on the WMO X2007 scale, and are updated to the WMO X2019 scale (the new scale is explained in details by Hall et al. (2020)).

2.4 O₂ measurements

Atmospheric O₂ is typically reported as the δ(O₂/N₂) value. The δ(O₂/N₂) value of a sample is calculated as the difference between the O₂/N₂ ratio of the sample and that of a reference gas (Keeling and Shertz, 1992):

$$\delta \left( O_2/N_2 \right)_s = \frac{\left( O_2/N_2 \right)_s - \left( O_2/N_2 \right)_{reference}}{\left( O_2/N_2 \right)_{reference}}$$  \hspace{1cm} (1)

Since for natural variations, δ(O₂/N₂) values are very small, they are usually expressed in “per mil”, which is 1/1000 of a per mil, as typically used in the stable isotope community. Atmospheric O₂ is reported as O₂/N₂ ratio because it is not a trace gas, and its mole fraction is thus affected by changes in other atmospheric constituents such as CO₂.

Atmospheric N₂ is relatively very stable (Keeling et al., 1998), therefore changes in the O₂/N₂ ratio would reflect mostly the changes in atmospheric O₂ (only in a detailed budget analysis minor N₂ variabilities are still considered, as described in Keeling and Manning (2014)). For δ(O₂/N₂) measurements, we use a Micromass Optima Dual Inlet Isotope Ratio Mass Spectrometer (DI-IRMS). The DI-IRMS analytical technique (which was first developed by Bender et al. (1994)) follows the principles as explained by Keeling et al. (2004). Due to the sensitivity of the analyser, it is located inside a climate-controlled room in our CIO laboratory, however the measurements still inevitably drift over time. To correct for the instrumental drifts, we perform frequent calibrations using a suite of reference gas cylinders. These cylinders are calibrated against the international Scripps scale using three primary standard cylinders purchased from the Scripps Institution of Oceanography (SIO), with δ(O₂/N₂) values ranging from -792 to -254 per mil. Details of the extensive calibration procedure are thoroughly described by van der Laan-Luijkx (2010) and van der Laan-Luijkx et al. (2010), and are summarised in Sect. 3.

2.5 Atmospheric Potential Oxygen (APO)

Combining highly precise measurements of atmospheric CO₂ and O₂ can isolate the effects of the oceanic processes, by removing the effects of the land biosphere (Stephens et al., 1998). This is achieved by deriving the tracer Atmospheric Potential Oxygen (APO). The APO value of an air sample is determined by combining its δ(O₂/N₂) and CO₂ measurements (Battle et al., 2006; Gruber et al., 2001; Stephens et al., 1998):

$$\delta APO = \delta \left( O_2/N_2 \right) + \frac{1.1 \times (CO_2-350)}{CO_2}$$  \hspace{1cm} (2)

The value of 1.1 represents the mean O₂:CO₂ ER of terrestrial ecosystems (Severinghaus, 1995); for the S0₂, we take 0.2094, which is the standard atmospheric O₂ mole fraction (Tohjima, 2005); and 350 is the consensus (arbitrary) reference value to be subtracted from the measured CO₂ mole fraction, as defined in the SIO per meg scale conversion for APO (Manning and Keeling, 2006). Therefore, APO is not affected by land biosphere processes and mainly captures the seasonal and long-term air-sea exchange of CO₂ and O₂, with a small influence from fossil fuels combustion, caused by their higher average ER of ≈1.4 (Pickers et al., 2017; Sirignano et al., 2010).
3 Calibration of the DI-IRMS

In this section we present the calibration procedure and the stability achieved at our laboratory from 2006 to 2020. The calibration of the measurements made in the 2000-2011 period and reported by van der Laan-Luijkx et al. (2010) and (van der Laan-Luijkx et al., 2013) are kept intact, and the newly calibrated measurements from 2011 onwards are built on the principles of that work.

3.1 The calibration procedure

The DI-IRMS compares the measurement of a sample gas with that of a reference gas (hereby called “machine reference” or “MREF”) in a sequence of several switches back and forth (“change overs”). The result of this process is the δ(O_2/N_2) value of the sample, as presented in equation 1. Each individual measurement is based on seven successive pairs of sample and reference measurements, which are used to calculate seven delta values (equation 1). The seven delta values then go through a filtering process. First, the mean and standard deviation of the seven delta values are calculated. Then, the delta value that is furthest from the mean is marked as a potential outlier. Next, a new mean and a new standard deviation are calculated for the remaining six delta values. If the excluded delta value is more than 2.7 times (equivalent to p = 0.01) the new standard deviation away from the new mean, it is defined as an outlier and removed. This process is repeated to identify and remove a potential second outlier (at most two outliers are removed by this process, otherwise the reliability of the measurement is sacrificed). After removing possible outliers, the remaining delta values are averaged to produce one δ(O_2/N_2) value per measurement. A flask is typically measured two to three times consecutively, and the final measurement for each flask (as presented in this paper) is the average of the filtered δ(O_2/N_2) values of these repeated measurements (van der Laan-Luijkx et al., 2010).

To improve the stability of our measurements, we also measure local reference gas cylinders (hereafter called “working tank” or “WT”) on the sample side of the DI-IRMS. These WTs are also used to connect between periods of different MREF cylinders, where there may be shifts in the scales of the measurements and thus a scale conversion is required to keep all raw measurements on a comparable scale. The summary of different WTs and MREF cylinders used from 1998 to 2020 is shown in Fig. 2 and Table 1.

![Figure 2: Summary of the different WTs and MREF cylinders in the 1998 - 2020 period. MREFs are shown along the top, with WTs below. In the case of WTs, there is typically overlap between more than one WT. Periods in grey colour are adapted from the work of van der Laan-Luijkx (2010).](image-url)
To connect the different MREF periods, we first convert all raw measurements (which are the ratios of the raw values to their respective MREF) to our internal 2534 CIO scale. Subsequently, they are converted to the SIO scale. Cylinder number 2534 has been chosen as the baseline for our internal reference scale, because it was the first MREF gas in 1998 and later on was measured as a WT against several other MREF cylinders (Fig. 2). When converting the measurements to the internal CIO scale, we need to take into account the “zero-enrichment” factor: measurements of a WT (on the sample side) against an MREF cylinder (on the reference side) do not produce the same value as when they are measured the other way around (van der Laan-Luijkx et al., 2010).

In addition to the conversion to our internal CIO scale, the measurements are also affected by instrumental drifts over time. To correct for these drifts, we first divide our long measurement record into several periods, which are defined based on the timing of when the MREF cylinders are changed, and/or apparent fluctuations in the raw data related to, for example, repairs or modifications of the system. In this work, the calibration procedure is carried out for measurements from 2011 onwards, which were divided into seven periods (periods 9-15, Table 1).

Table 1: Summary of the calibration periods defined in this paper and the corresponding MREF cylinder and WT cylinder numbers most recently used for the calibration of the DI-IRMS.

<table>
<thead>
<tr>
<th>Period</th>
<th>MREF</th>
<th>WTs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Previous</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>17-08-1998 – 18-02-2000</td>
<td>2534</td>
</tr>
<tr>
<td>3</td>
<td>03-02-2004 – 18-02-2004</td>
<td>8780</td>
</tr>
<tr>
<td>4</td>
<td>18-02-2004 – 14-05-2004</td>
<td>8780</td>
</tr>
<tr>
<td>5</td>
<td>04-06-2004 – 19-10-2005</td>
<td>4866</td>
</tr>
<tr>
<td>6</td>
<td>30-01-2006 – 30-12-2006</td>
<td>6170</td>
</tr>
<tr>
<td>7</td>
<td>30-01-2007 – 30-12-2007</td>
<td>6170</td>
</tr>
<tr>
<td>8</td>
<td>30-01-2008 – 15-12-2010</td>
<td>6170</td>
</tr>
<tr>
<td>Current</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>03-01-2011 – 11-03-2014</td>
<td>6185</td>
</tr>
<tr>
<td>10</td>
<td>11-03-2014 – 29-08-2015</td>
<td>5279</td>
</tr>
<tr>
<td>11</td>
<td>30-08-2015 – 10-06-2016</td>
<td>5279</td>
</tr>
<tr>
<td>12</td>
<td>11-06-2016 – 05-05-2018</td>
<td>5279</td>
</tr>
<tr>
<td>13</td>
<td>06-05-2018 – 01-01-2019</td>
<td>5279</td>
</tr>
<tr>
<td>14</td>
<td>02-01-2019 – 11-03-2020</td>
<td>5279</td>
</tr>
<tr>
<td>15</td>
<td>12-03-2020 to present</td>
<td>5279</td>
</tr>
</tbody>
</table>

These 7 periods were divided into 144 sub-periods (selected based on breaks in the records) which were then individually processed to derive the final corrections for all measurements in those sub-periods. The complete step in transforming the raw measurements of a sample (S) against a current MREF (M) into comparable data is to combine the drift correction with the shift to the CIO scale (R), by using an equation described by van der Laan-Luijkx (2010):

\[
\delta_S/R = \left( (\delta_{M/R})_{\text{sub-period}} + \text{drift} \times \frac{\text{days}}{365} + 1 \right) \times (\delta_{S/M} + 1) - 1
\]

(3)

Where:
- \(\delta_S/R\) is the \(\delta(O_2/N_2)\) value of the sample against the CIO 2534 scale;
- \(\delta_{M/R}\)_{sub-period} is the average \(\delta(O_2/N_2)\) value of the MREF cylinder against the CIO scale in a sub-period calculated based on the measurements of all WTs in that sub-period;
- drift is the average drift per day in a sub-period (if any), calculated based on the WT values and days is the number of days at the time of the sample since the start of the sub-period.
- \(\delta_{S/M}\) is the \(\delta(O_2/N_2)\) value of sample against the MREF cylinder (raw value).
The final step is to transform the $\delta_{34}^R$ value of a sample onto the SIO scale via a linear conversion using the values of the Scripps primary cylinders measured against the CIO scale. For an extensive and detailed explanation on how to calculate each component of equation 3 and conversion to the SIO scale, we refer to van der Laan-Luijkx (2010). Figure 3 shows the results for the WTs of the new calibration procedure connected to the previously reported data by van der Laan-Luijkx et al. (2010).

![Figure 3: Measurements of the 3 long-term WTs (5279, 6096, and 6168) for periods 7-15 (Table 1), across the final 3 MREF periods plus a recently added WT (4845). Top panel: raw measurements of the WTs against different MREF cylinders. Bottom panel: measurements of the WTs calibrated and converted to the CIO scale (left y-axis) and against the SIO scale (right y-axis). The values on the plot are the corresponding long-term means and 1-sigma standard errors of the WTs against the SIO scale, and in parentheses are the respective standard deviations. All numbers are in per meg. Visible gaps in the data are due to instrument issues, maintenance or instrument relocation. After these adjustments, the measurements of the 3 long-term WTs (5279, 6096, and 6168) show that they were relatively stable over time, all showing standard deviations of 10.2 to 13.5 per meg (corresponding to standard errors of 0.09 to 0.14 per meg) over the 20-year period. WT 4845 was recently measured for a relatively short period only, and appears to be less stable and noisier compared to WT 5279 measured in the same period. Thus, WT4845 was not used for the calculations in the calibration procedure, and its measurements are only shown here for completeness.

In addition to their long-term stability, the 3 WTs also showed no systematic drifts across different MREF periods (Table 2). For WT 5279 and WT 6096, there are no significant changes (at least to ±0.3 per meg) between the MREF periods 6170 and 6185, although there is a small decrease of 4.0 per meg in the mean measurement of WT 5279 in MREF 6123 period. For WT 6168, the mean value increased by 3.6 per meg from MREF 6170 to MREF 6185 period, then dropped slightly (by 0.5 per meg) in MREF 6123 period. The stability demonstrated in both long-term measurements and per MREF periods consolidates the quality of our calibration procedure.
Table 2: Comparison of the WTs over 3 different MREF cylinder periods. The values (in per meg) are averaged over the corresponding period, accompanied by the standard errors. The N/A values in the MREF6123 period for WT6096 are due to its discontinuation in this period. The Difference column is calculated by subtracting the values of the old MREF periods from the new ones.

<table>
<thead>
<tr>
<th></th>
<th>CIO scale</th>
<th>SIO scale</th>
<th>Difference (CIO scale)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT 5279</td>
<td>MREF 6170</td>
<td>20.9 ± 0.2</td>
<td>-438.2 ± 0.2</td>
</tr>
<tr>
<td></td>
<td>MREF 6185</td>
<td>20.7 ± 0.3</td>
<td>-438.4 ± 0.2</td>
</tr>
<tr>
<td></td>
<td>MREF 6123</td>
<td>16.7 ± 0.2</td>
<td>-442.3 ± 0.2</td>
</tr>
<tr>
<td>WT 6096</td>
<td>MREF 6170</td>
<td>-103.4 ± 0.1</td>
<td>-556.9 ± 0.1</td>
</tr>
<tr>
<td></td>
<td>MREF 6185</td>
<td>-103.1 ± 0.1</td>
<td>-556.7 ± 0.1</td>
</tr>
<tr>
<td></td>
<td>MREF 6123</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>WT 6168</td>
<td>MREF 6170</td>
<td>-137.3 ± 0.2</td>
<td>-589.3 ± 0.2</td>
</tr>
<tr>
<td></td>
<td>MREF 6185</td>
<td>-133.7 ± 0.2</td>
<td>-585.9 ± 0.2</td>
</tr>
<tr>
<td></td>
<td>MREF 6123</td>
<td>-134.2 ± 0.4</td>
<td>-586.4 ± 0.3</td>
</tr>
</tbody>
</table>

3.2 Quality check of the Scripps primary cylinders

The final check on the quality of our scale is the regular measurement of the 3 Scripps primary standard cylinders that we purchased from SIO, numbered 7002, 7003, and 7008. These measurements were conducted at least once a year or when there was an additional need for recalibrating e.g. after instrument failure or upgrade. Each measurement period took a different amount of time – some measurements were spread over a couple of days while others were repeated over (or after) a few weeks. From 2007 to 2018, 16 measurement periods were conducted (Fig. 4). The large gap between 2011 and 2014 was due to a lack of funding.

In Figure 4, each data point is the mean value over each measurement period and the error bars are the standard deviations. The coloured lines are the overall linear fit of the measured values of the corresponding cylinders (and their associated 2-sigma uncertainties) and the black horizontal lines are the assigned values of the cylinders (determined by the SIO, updated in 2020). The assigned and measured values of the primary standard cylinders over the whole period are compared in Table 3. The measured values are the weighted means of each cylinder, since each data point is calculated based on different numbers of separate measurements. It can be seen from Figure 4 and Table 3 that cylinder 7008 exhibits a small upward drift over time of 1.4 ± 0.4 per meg yr⁻¹, whereas the other two remain constant. The ensemble thus suggests that there is no clear systematic error in our scale conversion and calibration procedure. To improve the quality of our conversion into the SIO scale, and especially to check the behaviour of cylinder 7008, we are planning to purchase new primary standard cylinders in the future.

The conversion of the CIO scale to the Scripps scale is done using these measurements, and in such a way that the ensemble difference between the assigned values and weighted averages of our measurements of three Scripps cylinders is minimised.
Figure 4: Scripps primary standard cylinder measurements over time. Each point is the averaged value over a measurement period. Error bars represent 1-sigma standard deviations. Solid horizontal lines are the assigned values (black) and the linear least squares fit to the data (coloured) of each cylinder. The grey shading indicates the 95% confidence interval uncertainties of the values.

Table 3: Comparison of the averaged measured values of the Scripps primary standards against their assigned values in per meg.

<table>
<thead>
<tr>
<th>Cylinder ID</th>
<th>Assigned by SIO</th>
<th>Weighted mean measured</th>
<th>Standard deviation</th>
<th>Standard Error</th>
<th>Deviation from assigned</th>
</tr>
</thead>
<tbody>
<tr>
<td>7008</td>
<td>-254.3</td>
<td>-465.0</td>
<td>8.0</td>
<td>1.9</td>
<td>8.4</td>
</tr>
<tr>
<td>7002</td>
<td>-245.9</td>
<td>-468.9</td>
<td>6.0</td>
<td>2.0</td>
<td>-3.9</td>
</tr>
<tr>
<td>7003</td>
<td>-791.6</td>
<td>-797.6</td>
<td>11.8</td>
<td>2.8</td>
<td>-6.0</td>
</tr>
</tbody>
</table>

3.3 Inter-comparison programmes

In addition to measuring the primary standard cylinders, the CIO also took part in two inter-comparison programmes involving oxygen measurements: “Cucumber” Intercomparison which was initialised in the European Union’s CarboEurope project and coordinated by the UEA (http://cucumbers.uea.ac.uk/); and the Global Oxygen Laboratories Link Ultra-precise Measurements (GOLLUM) programme, also coordinated by UEA (Gollum Comparison: Manning et al., 2015). These inter-comparison programmes provide an additional tool for checking the internal stability of our measurements, while also linking the oxygen measurements between global laboratories.
The Cucumber programme involves inter-comparison of nine atmospheric species (of which $\delta(O_2/N_2)$ is one) between atmospheric research stations in Europe and a number of laboratories in Europe, USA, Canada, Japan, and Australia. Within the programme, there are seven sets of three cylinders sent around in different rotations. The CIO participated in three rotations, with two involving oxygen measurements (called “Inter-1” and “Euro-3”) (University of East Anglia, 2021).

The GOLLUM programme is specifically designed for the inter-comparison of oxygen measurements and involves 10 laboratories worldwide that carry out high-precision atmospheric oxygen measurements. Two sets (named “Bilbo” and “Frodo”) of three cylinders are rotated in opposite directions amongst participating laboratories (Gollum Comparison: Manning et al., 2015).

Figure 5 shows the measurements of the Cucumber cylinders (top two panels), the cylinders in the Bilbo and Frodo rotations of GOLLUM (third and fourth panels, respectively) and the measurements of three internal cylinders at CIO: the working tanks 5279, 6096 and 6168 along with the SIO primary standard cylinder 7008 (bottom panel). The measurements of the cylinders in the Inter-1 and Euro-3 rotations are plotted as the difference between the measured values of the cylinders against their own assigned values as originally measured at the Max Planck Institute for Biogeochemistry in Germany in January 2008. These results show that the cylinders in the Inter-1 and Euro-3 rotations were quite variable over time (varying within a range of less than 30 per meg) but in different directions and size, suggesting that there is not a systematic scaling error but rather individual variations between cylinders and/or measurement periods. Due to the individual variations, the overall drifts for Cucumber cylinders is $11 \pm 18$ per meg yr$^{-1}$, higher than the WMO extended compatibility goal of 10 per meg (World Meteorological Organization, 2018). The lower quality of the measurements (not only in our laboratory) might well be connected to the fact that these cylinders are not part of a dedicated oxygen comparison programme, so the treatment of the cylinders (for example, vertical storage and unsuitable pressure reducers) are not of high enough standard for oxygen.

For GOLLUM cylinders, all measurements are also plotted as the difference between the measured values of the cylinders and their assigned values on the SIO scale. The assigned values for Bilbo, Frodo and SIO cylinders are determined at the SIO, while those for the WTs are their averaged long-term value measured at CIO on the SIO scale. Compared to the Cucumber cylinders, GOLLUM cylinders show much less variations between years (varying within a range of less than 20 per meg), and also significantly smaller overall drift over the duration of the measurements ($4 \pm 6$ per meg yr$^{-1}$, which is within the extended WMO compatibility goal). However, all 6 cylinders appear to drift in similar direction, suggesting a drift (however small) in our scale rather than drifts in these cylinders. The SIO cylinder 7008 also shows similar stability and a general drift in the same direction as GOLLUM cylinders, whereas the two other SIO cylinder do not (Fig. 4).

Since the cylinders show an inconclusive “drift”: INTER-1 and EURO-3 do not show an apparent drift direction; Bilbo and Frodo present a minor drift similarly to that observed by our SIO cylinder 7008 (while the other 2 SIO cylinders did not exhibit this behaviour as shown in Sect. 3.2); and our internal WTs all show no overall drifts, we consider our calibration procedure as sufficient. Recalibration of the SIO cylinders might shed further light on these small discrepancies, mostly to see if cylinder 7008 has indeed drifted or not.
Figure 5: Cylinders from the Cucumber programme (top 2 panels) along with two sets of three cylinders in the GOLLUM programme (middle 2 panels) and 3 internal CIO cylinders (WT 5279, WT 6096 and WT 6168) and a primary standard cylinder at CIO (SIO 7008) (bottom panel). Each colour represents a different cylinder, and the legends show the corresponding cylinder IDs. The points are the measurements of the cylinders over time, plotted as the difference from their assigned values. For the Cucumber, GOLLUM and SIO cylinders, the assigned values are determined at the SIO, and for the WTs, the assigned value is its long-term average measured at CIO on the SIO scale. Y-axis ranges are identical on all panels.

3.4 Treatment of analysed flask samples

After the calibration and conversion to the SIO scale, the individual flask sample measurements are scrutinized for outliers and background conditions. For this purpose, we perform several iterations of fitting a combination of quadratic and 3-harmonic regression (following similar curve fitting methods applied to time series in NOAA (Thoning et al., 1989)) and filtering the outliers from the combined fit. This outlier filtering process uses the robust median absolute deviation (MAD) method (Rousseeuw and Verboven, 2002), in which the MAD value for a dataset is determined by first finding the median of the set, then subtracting the median from each individual value, and finally finding the median of the absolute differences. Measurements that are 3 times the MAD value away from the median of the measurement set are considered outliers and removed. The full principle of the procedure is described by van der Laan-Luijkx (2010) (though with a different filtering process that was described in Sect. 3.1). In total, after both filtering processes, around 30% of the flasks were excluded from further analyses from Lutjewad samples, 16% from Mace Head samples, and only 6% from Halley samples. The larger fraction of discarded measurements in the Lutjewad record
is related to the sampling process, where we do not specifically only sample air at background conditions, which is the case at Mace Head. For Halley, since it is by design a background station, there are hardly any local sources and sinks, and the wind coming from the continental plateaux only accounts for 30% of the total. The 6% outlier fraction for Halley is a good indication of the fraction of actually failed sampling and/or analysis. The APO values of all stations are calculated from $\delta^{18}$O/$N_{2}$ and CO$_2$ measurements (equation 2), when there is information on both species for each flask sample.

In the period prior to 2006, our internal calibration scale was not as well-established as in the later period, due to frequent changes in MREF and WT cylinders, especially in 2004 when there is little information to connect the following period to the first period (as presented in Fig. 2 and Table 1). Next to this, we also only obtained the SIO primary standards in late 2007, so all earlier measurements cannot be directly linked to the SIO scale and have to be converted via the internal CIO scale. The results of this quality check prompt us to exclude the first 2 years from the fits of Lutjewad and Mace Head data so that they are less affected by the problematic period. The last 2 years are also excluded because flask sampling was relatively sparse in those years and this could also introduce biases in the fits. After several tests, we decided to establish our fits for Lutjewad and Mace Head based on the years 2002 to 2018.

4 Flask measurement results

4.1 The CO$_2$, $\delta^{18}$O/$N_{2}$ and APO records

In this section, we present the long-term flask measurement records (from 2000 to 2020) of Lutjewad and Mace Head, along with a 3-year record from Halley. In general, Lutjewad and Mace Head show similar patterns for $\delta^{18}$O/$N_{2}$ and CO$_2$, with some differences in APO variations. Figures 6 to 8 show the CO$_2$, $\delta^{18}$O/$N_{2}$, and APO measurements for Lutjewad, Mace Head, and Halley, respectively. The black points illustrate the final, filtered flask measurement values; the coloured lines are the total fit (combined quadratic trend and 3-harmonic seasonal cycles) and the black lines are the trend parts of the total fit. The fit lines are shown for the whole period, but for the fitting process we left the first and last two years out, to make sure that the fit period comprises an exact multitude of the total fit (combined quadratic trend and 3-harmonic seasonal cycles). Due to the quadratic trend fit, the growth of the fitted increase is linear. The trend (given here in ppm yr$^{-1}$ with their 95% confidence interval (CI) uncertainties) in Lutjewad grows from 1.81 ± 0.10 ppm yr$^{-1}$ in 2002 to 2.27 ± 0.03 ppm yr$^{-1}$ in 2010 and 2.74 ± 0.10 ppm yr$^{-1}$ in 2018. These values agree relatively well with the globally averaged values as measured by the NOAA’s Global Monitoring Laboratory: 1.86 ± 0.20 ppm yr$^{-1}$ in 2002, 1.97 ± 0.14 ppm yr$^{-1}$ in 2010, and 2.57 ± 0.19 ppm yr$^{-1}$ in 2018 (https://gml.noaa.gov/ccgg/trends/global.html). The values from NOAA are calculated based on a 5-year average around the time marks 2002, 2010 and 2018. In all three periods, the values at Mace Head are also in agreement with those of Lutjewad (1.86 ± 0.06 in 2002, 2.24 ± 0.02 in 2010, and 2.63 ± 0.06 ppm yr$^{-1}$ in 2018 for Mace Head). When averaging the trends over the 20-year period, both stations show good agreement with each other and with the global average: 2.31 ± 0.07 ppm yr$^{-1}$ for Lutjewad, 2.22 ± 0.04 ppm yr$^{-1}$ for Mace Head, and 2.1 ± 0.3 ppm yr$^{-1}$ for global.

$\delta^{18}$O/$N_{2}$ measurements at Lutjewad also show a clear trend that becomes increasingly more negative throughout the 20 years. The trends (reported here in per meg yr$^{-1}$ with their 95% CI uncertainties) in 2002, 2010, and 2018 are -18.01 ± 0.06 (2002), -21.50 ± 0.09 (2010), and -25.00 ± 0.07 ppm yr$^{-1}$ (2018).
The APO trend and seasonality can be determined either from fitting the APO values of the individual flasks themselves, or by combining the trend/seasonal parameters of the \( \delta(O_2/N_2) \) and CO\(_2\) fits. Both methods yield almost identical results. We present here the results from the first approach. Since APO is calculated from the combination of \( \delta(O_2/N_2) \) and CO\(_2\) measurements, it shows a combination of the patterns as illustrated in the two species. The APO trend (reported here also in per \( \text{meg yr}^{-1} \)) at Lutjewad does not differ significantly over time, varying from \(-9.4 \pm 0.8\) per \( \text{meg yr}^{-1} \) in 2002 to \(-9.31 \pm 0.20\) per \( \text{meg yr}^{-1} \) in 2010, and \(-9.3 \pm 0.8\) per \( \text{meg yr}^{-1} \) in 2018. In Mace Head, however, the same pattern as \( \delta(O_2/N_2) \) is shown for APO: the trend gets significantly less negative throughout the period (-13.15 \( \pm \) 1.20 per \( \text{meg yr}^{-1} \) in 2002, -9.5 \( \pm \) 0.3 per \( \text{meg yr}^{-1} \) in 2010, and -5.83 \( \pm \) 1.20 per \( \text{meg yr}^{-1} \) in 2018).

Measurements at Halley station show a similar trend as Lutjewad and Mace Head, where CO\(_2\) increases over time while \( \delta(O_2/N_2) \) decreases, with much less variability in \( \delta(O_2/N_2) \) and CO\(_2\) measurements, due to the absence of a terrestrial biosphere influence. The averaged CO\(_2\) trend at Halley from 2014 to 2017 is 2.60 \( \pm \) 0.20 ppm \( \text{yr}^{-1} \), similar to the trends at Lutjewad and Mace Head in the same period (2.62 \( \pm \) 0.08 ppm \( \text{yr}^{-1} \) and 2.53 \( \pm \) 0.05 ppm \( \text{yr}^{-1} \), respectively). On the other hand, \( \delta(O_2/N_2) \) and APO trends at Halley are significantly smaller in size than those at Lutjewad and Mace Head.

The \( \delta(O_2/N_2) \) trend at Halley over the 2014-2017 period is \(-15 \pm 3\) per \( \text{meg yr}^{-1} \) while at Lutjewad and Mace Head, the trends are \(-23.2 \pm 0.9\) per \( \text{meg yr}^{-1} \) and \(-20.3 \pm 1.0\) per \( \text{meg yr}^{-1} \), respectively. For APO, the corresponding values are \(-1.4 \pm 2.4\) per \( \text{meg yr}^{-1} \), \(-9.3 \pm 0.6\) per \( \text{meg yr}^{-1} \), and \(-6.7 \pm 0.9\) per \( \text{meg yr}^{-1} \).
Figure 6: Flask record from Lutjewad station, showing $\text{CO}_2$, $\delta(\text{O}_2/\text{N}_2)$, and APO measurements from 2000 to 2020. The black points are the individual flask measurements, while the black lines are the long-term trend and the coloured lines indicate the trend with seasonal components derived from the combined quadratic and harmonic regression. The uncertainty ranges (2-sigma) in the fits are indicated by lighter shades of the same colours. For comparability, the y-axes ranges are scaled to represent the 5 per meg : 1 ppm ratio.
Figure 7: As for Fig. 6 but for Mace Head station.
Figure 8: As for Fig. 6 and 7, but for Halley station and from 2014 to 2017.

4.2 Seasonal cycles

The seasonal cycles of CO$_2$, δ(O$_2$/N$_2$), and APO for all three stations are presented in Fig. 9. The seasonal components are extracted from the total fits (detrended) and presented as 1-year cycles. In general, the CO$_2$ seasonal cycles at Lutjewad and Mace Head are similar in size and shape, although the average seasonal amplitude is higher at Lutjewad (16.8 ± 0.5 ppm) than Mace Head (14.8 ± 0.3 ppm). The CO$_2$ seasonal cycle at Halley station, on the other hand, has a much smaller amplitude of 3.0 ± 0.3 ppm, as is generally the case for the ocean-dominated Southern Hemisphere due to the absence of a terrestrial biosphere influence. Lutjewad and Mace Head show very similar, and significantly higher δ(O$_2$/N$_2$) seasonal amplitudes (131 ± 6 per meg and 130 ± 6 per meg, respectively) than that at Halley (76 ± 4 per meg), due to the influences of the terrestrial biosphere. In APO this influence is cancelled because APO is invariant to terrestrial biosphere processes, and the Halley amplitude is even somewhat higher than that of Lutjewad and Mace Head (65 ± 3 per meg compared to 54 ± 4 and 61 ± 5 per meg, respectively). All numerical seasonality parameters of the three stations are given in Table 4 below.
Here, we discuss our measurement records in more detail. At first, the difference in the progression of trends in $\delta(O_2/N_2)$ and APO between Lutjewad and Mace Head (Fig. 6 and 7) suggests that there could be an issue with the flask sampling procedure at Mace Head, such as the way the samples are dried. At Lutjewad, the sampling process has been under much closer controlled thanks to the vicinity of our laboratory enabling frequent visits, multiple tests and other measurements taken from the same sample lines. Furthermore, a comparison of the Lutjewad data with data from the nearby Weybourne coastal station in the UK (presented in Sect. 5.2) showed very good agreement. As both Lutjewad and Mace Head samples share the same measurement procedure, measurement and calibration issues cannot explain their differences, so the differences must either be real, or related to the flask sampling procedure. It takes longer to transport the flasks from Mace Head to Groningen than from Lutjewad and thus contaminations of the samples through the valve caps might have occurred. For the samples from the Halley station, the transport time is even longer, but here, additional protective caps are used on the valve caps of the flasks to counteract permeation effects. We tested the preservation of the samples using the protective caps by sending flasks to Halley station that were pre-filled with air of
known composition, without actually using them. Back in Groningen, we could conclude the integrity of the samples by comparing the measurements before and after shipment, and we found no significant change in $\delta(O_2/N_2)$ after 26 to 51 months. Unfortunately, the protective caps were not applied to Mace Head samples. Still, it is hard to imagine how such permeation effects could cause a deviating long-term trend in the data given that the flasks were filled to ambient pressure. If anything, one would expect more scatter in the record. The same holds for sampling problems, such as incomplete drying. To summarise, the trends at Lutjewad are as expected while those at Mace Head are not, so if there are no systematic sampling errors, the differences in $\delta(O_2/N_2)$ and APO in Mace Head compared to Lutjewad might be partially caused by the sparse and irregular sampling frequency at Mace Head. However, it is also worthwhile to consider effects that maybe caused by real environmental differences between the two stations. Two effects come to mind: the first is a difference in fossil fuel use (both in quantity and type), which would influence $\delta(O_2/N_2)$ and to a lesser extent also APO. The average fossil fuel exchange ratio (ER) for the Netherlands, when accounting for all fossil fuel types, is $1.60 \pm 0.02$ for the 2000-2020 period, much higher than that for Mace Head ($1.49$, see van der Laan-Luijkx et al. (2010) and the CO$_2$ release and Oxygen uptake from Fossil Fuel Emission Estimate (COFFEE) database by (Steinbach et al., 2011)), and the global average value for all fossil fuel emissions (of 1.38), as also mentioned by Sirignano et al. (2010) and van der Laan-Luijkx et al. (2010). However, it is unlikely that this is the main explanation of the difference between the two records. Firstly, because at Lutjewad, sampling was selective so as to avoid continental (and thus local fossil fuel) influences as much as possible, and second, because a difference in trends would need a gradual change in the ER. Data from Statistics Netherlands (CBS, 2021) show that the ER of the Netherlands has changed by no more than 0.02 over the period 2000-2020, too small to be of influence on the observed difference in the trends at Lutjewad and Mace Head.

A more likely cause for differences between Mace Head and Lutjewad are changes in North Atlantic oxygen ventilation (Keeling and Manning, 2014) to which the Mace Head observations are more sensitive. Such changes would influence $\delta(O_2/N_2)$ and APO, but not CO$_2$. This is consistent with the fact that the CO$_2$ trends of Mace Head and Lutjewad agree, whereas there are differences in $\delta(O_2/N_2)$ and APO. Changes in the oxygen inventory of the North Atlantic have been reported by Stendardo and Gruber (2012) and Montes et al. (2016) and a relationship with the North Atlantic Oscillation (NAO) has been reported. Data obtained from the NOAA Climate Prediction Center (https://www.cpc.ncep.noaa.gov/data/teledoc/nao.shtml) show that the NAO exhibited gradual changes over the period 2000-2020, from a noisy, more or less balanced positive-negative pattern in the first decade, through to a negative phase in the years 2010-2011 towards gradually mostly positive values for the period 2013-2019.

When comparing the seasonal cycles of the three stations, we can see that while CO$_2$ and $\delta(O_2/N_2)$ seasonal amplitudes at Halley are significantly smaller than those at Lutjewad and Mace Head, the APO seasonal amplitude is slightly higher, agreeing with the model simulation by Tohjima et al. (2012) that the APO seasonal variations in the Southern Hemispheric ocean are larger than those in the Northern Hemisphere due to larger air-sea O$_2$ exchange. As mentioned in Sect. 2, APO values also contain a small influence from fossil fuels, however, by selecting for flasks based on the background conditions, we eliminate as much as possible this influence, especially for the Lutjewad record. As such, our APO values from these three stations represent mostly ocean influences.

To conclude the comparison between Lutjewad and Mace Head, we calculated the partitioning of CO$_2$ uptake by the terrestrial biosphere and the ocean from the observations at the two stations, using the measurements of CO$_2$ and APO concentrations at Lutjewad and Mace Head from 2002 to 2018, via the method described by Keeling and Manning (2014). This partitioning is illustrated in Fig. 10.
The black points are the annual averages of the de-seasonalised measurements of APO and CO₂ concentrations at Lutjewad and Mace Head for the period 2002-2018, and clearly show the unexpected trend in APO for Mace Head, as discussed above. For calculating the partitioning of fossil fuel CO₂, we use, from Keeling and Manning (2014), equations (2) to (10), and the ocean O₂ outgassing component (Z) of 0.44 ± 0.45 · 10¹⁴ mol yr⁻¹ (equivalent to an effect on the carbon sinks of 0.46 ± 0.48 PgC yr⁻¹). Furthermore, we use the total fossil fuel component for the years 2002-2018 of 8.6 ± 0.6 PgC yr⁻¹ as derived from the Global Carbon Budget 2020 by Friedlingstein et al. (2020), and the ER for globally averaged fossil fuel combustion of 1.38 ± 0.03 from Keeling and Manning (2014). From the Lutjewad record, the global land biotic sink (B) is 1.7 ± 1.1 PgC yr⁻¹, the oceanic sink (O) is 2.0 ± 0.8 PgC yr⁻¹, and the CO₂ remaining in the atmosphere (A) amounts to 4.91 ± 0.15 PgC yr⁻¹. The values calculated from Mace Head record are very similar to those from Lutjewad: at Mace Head, B is 1.7 ± 1.1 PgC yr⁻¹, O is 2.2 ± 0.9 PgC yr⁻¹ and A is 4.72 ± 0.09 PgC yr⁻¹. For both stations, the values for the B and O components agree well with those reported by Friedlingstein et al. (2020): 1.7 ± 1.6 GtC yr⁻¹ for B (including emissions from land-use changes) and 2.3 ± 0.8 GtC yr⁻¹ for O. The values for A as measured at both Lutjewad and Mace Head are slightly higher than the reported average value of 4.63 ± 0.03 PgC yr⁻¹ for the 2002-2018 period.

The challenges in making O₂ measurements have presented themselves clearly in this work: the sensitivity of the mass spectrometer that require intensive calibration; the quality maintenance of the internal calibration scale to make sure that our measurements can be reported with sufficient quality on the international scale; and the unexpected patterns (especially in APO for Mace Head) that could not be fully explained, partly due to the lack of consistent sampling frequency before 2004 (for both stations), during 2012 (for Lutjewad) and between 2017 and 2019 (for Mace Head). The trend and seasonality fitting procedure are also of great importance, as these are also highly sensitive to irregular sampling frequency and biases in the timing in which the majority of the samples is collected. Nevertheless, our flask measurement records of Lutjewad, Mace Head, and Halley have proven to be informative and valuable in evaluating APO, and with future technical improvement (especially regarding the sampling frequency and the quality maintenance...
of our internal scale), they will be extended further. In the near future, in addition to more regular sampling frequency at Lutjewad and Mace Head, we aim to improve the frequency at which we perform the measurements on the SIO primary standard cylinders, and also to purchase new primary standard cylinders from them, to produce higher precision conversion to the SIO scale. We also aim to employ more WTMs as the current ones are either running out or experiencing considerable noise (see WT 4845 in Fig. 3). More protective measures to the flasks, such as using additional caps or switching to another type of valve, will also be considered, to reduce the risks of potential leakages, permeations, and contaminations during storage and transportation.

5.2 Comparison with other long-term records

In Table 5, we compare the seasonal amplitudes of our CO$_2$, δ(O$_2$/N$_2$), and APO measurements with those of some other stations worldwide. As can be seen, the measurements for all three species at Lutjewad and Mace Head agree well with the measurements conducted at other Northern Hemisphere stations Weybourne (UK), Sendai (Japan), and Ny Ålesund (Norway). In the Southern Hemisphere, our δ(O$_2$/N$_2$) and APO measurements for Halley station show an excellent agreement with those at the Syowa station. On the other hand, our CO$_2$ measurements exhibit a much larger and noisier seasonal amplitude, which is most likely caused by small leaks during sampling (the details of which are given at the end of this section). Nonetheless, the general concurrence with these stations helps to consolidate the quality of our measurements.

<table>
<thead>
<tr>
<th>Station</th>
<th>Time period</th>
<th>Latitude</th>
<th>CO$_2$ (ppm)</th>
<th>δ(O$_2$/N$_2$) (per meg)</th>
<th>APO (per meg)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ny Ålesund, Spitsbergen</td>
<td>2001-2010</td>
<td>79°N</td>
<td>15.2 ± 0.4</td>
<td>129 ± 4</td>
<td>52 ± 3</td>
<td>Ishidoya et al. (2012b)</td>
</tr>
<tr>
<td>Weybourne, UK</td>
<td>2008-2015</td>
<td>53°N</td>
<td>15.2 ± 1.1</td>
<td>130 ± 8</td>
<td>51 ± 6</td>
<td>(Barningham, 2018)</td>
</tr>
<tr>
<td>Lutjewad, the Netherlands</td>
<td>2002-2018</td>
<td>53°N</td>
<td>16.8 ± 0.5</td>
<td>131 ± 6</td>
<td>54 ± 4</td>
<td>This paper</td>
</tr>
<tr>
<td>Mace Head, Ireland</td>
<td>2002-2018</td>
<td>53°N</td>
<td>14.8 ± 0.3</td>
<td>130 ± 6</td>
<td>61 ± 5</td>
<td>This paper</td>
</tr>
<tr>
<td>Sendai, Japan</td>
<td>1999-2012</td>
<td>38°N</td>
<td>13.9 ± 2.5</td>
<td>128 ± 22</td>
<td>52 ± 10</td>
<td>Ishidoya et al. (2012a)</td>
</tr>
<tr>
<td>Syowa station, Antarctica</td>
<td>2001-2010</td>
<td>69°S</td>
<td>1.1 ± 0.04</td>
<td>70 ± 4</td>
<td>64 ± 4</td>
<td>Ishidoya et al. (2012b)</td>
</tr>
<tr>
<td>Halley station, Antarctica</td>
<td>2014-2017</td>
<td>75°S</td>
<td>3.0 ± 0.3*</td>
<td>76 ± 4</td>
<td>65 ± 3</td>
<td>This paper</td>
</tr>
</tbody>
</table>

*The CO$_2$ seasonal amplitude at Halley is most likely incorrect, details are given at the end of this section.

Additionally, we compare our long-term measurement record with an extended record of Weybourne station (Fig. 11), the first part of which has been published by Pickers (2016) and Barningham (2018). Because of differences between continuous and flask sampling frequency, there are large differences between the two records, obviously in the amount of data points, but also in the height of the peaks and troughs during the diurnal cycle, and also the temporal length of the peaks and troughs. This is mostly because our flask samples at Lutjewad are collected during background conditions and may be subject to other local events. However, when considering background conditions at Weybourne (the lower limits of CO$_2$ and upper limits of δ(O$_2$/N$_2$)), there is good agreement with the Lutjewad record.
Figure 11: Measurements of CO$_2$, $\delta$(O$_2$/N$_2$), and APO at Lutjewad (black diamonds) and Weybourne (orange crosses) from 2010 to 2020. The black line and curve are the trend and the combined fit for Lutjewad, respectively. The grey shadings are the 95% CI associated with the total fit.

For Halley, we compare our CO$_2$, $\delta$(O$_2$/N$_2$), and APO measurements with those conducted by UEA (Fig. 12) (Barningham, 2018). APO measurements between our laboratory and UEA show good agreement, while CO$_2$ measurements show unexpected discrepancies in the first half of 2016. $\delta$(O$_2$/N$_2$) measurements also show a slight disagreement, but it is less visible due to a large seasonal cycle and higher scatter. Because APO agrees well, we conclude that the CO$_2$ and $\delta$(O$_2$/N$_2$) anomalies were most likely caused by a small inwards leak when the flask samples were collected at the station. Laboratory air with higher CO$_2$ mole fractions and lower $\delta$(O$_2$/N$_2$) ratios due to human breathing, probably leaked in. An additional indication pointing to this is that the CH$_4$ and CO mole fractions from the same flasks agree very well with long-term flask measurements made at Halley by NOAA (NOAA, 2021) (not shown here). Such leaks do not influence APO, as the ER from human breathing is close to the value of 1.1 used for the exclusion of the biosphere signal in APO. To better check how much these anomalies would have affected our measurements, we use the long-term flask measurements made at Halley from the NOAA website (https://gml.noaa.gov/dv/iadv/graph.php?code=HBA&program=cegg&type=ts), since the UEA’s measurement period is too short to make a reliable comparison. For CO$_2$, we perform the trend and seasonality fitting procedure, the same as for our own measurements. For APO, since there is no available information on O$_2$, we combine the NOAA’s CO$_2$ measurements with our own $\delta$(O$_2$/N$_2$) measurements to calculate APO, then proceeded with the fitting of the trend and seasonality. Plotted in blue are the results using the NOAA’s CO$_2$ measurements. Leakage would lead to higher CO$_2$ mole fraction and lower $\delta$(O$_2$/N$_2$) ratio than the background sample, but APO would largely cancel them. We now combine the background CO$_2$ measurements from NOAA and the $\delta$(O$_2$/N$_2$) values from our (supposedly contaminated) flask measurements, and indeed a clear bias in the APO can be seen coinciding with the CO$_2$ anomalies. This confirms the presence of some contamination in a number of our flasks, especially during March-June 2016 and in early 2014.
625 Figure 12: Measurements of CO$_2$, $\delta$(O$_2$/N$_2$), and APO at Halley conducted by CIO (black diamonds) from 2014 to 2017, and continuous measurements conducted by UEA (orange crosses) in 2016. The black lines and curves are the trends and the combined fit for measurements by CIO, respectively. The grey shaded area is the 95% CI associated with the total fit. The blue points are the in-situ continuous measurements at Halley, taken by NOAA. The blue lines and curves are the trends and the combined fits for the continuous measurements, with the lighter blue shaded area the 95% CI associated with the total fit. From the combination of all data we conclude that our CO$_2$ measurements must have been contaminated with inside air (human breathing).

630 6 Conclusion

We have presented 20-year flask measurement records for $\delta$(O$_2$/N$_2$), CO$_2$ and APO from Lutjewad and Mace Head, along with 3-year records from Halley. We also presented results of the calibration procedure of our instruments. Due to the sensitive nature of oxygen measurements, we conducted an extensive and intensive calibration procedure, which demonstrated a long-term standard deviation for $\delta$(O$_2$/N$_2$) of 13.5 per meg based on our own internal cylinders.

Measurements of the global primary standard cylinders (from SIO) and inter-comparison cylinders (from the Cucumber and GOLLUM programmes) consolidate the stability, quality, and comparability of our calibration procedure, although there are some indications that our calibration scale might not be entirely stable over the past 20 years. However, the results from those various programmes are not consistent, and therefore inconclusive. The long-term records from Lutjewad and Mace Head provided useful information on the two-decadal trends and seasonality of CO$_2$, $\delta$(O$_2$/N$_2$), and APO, showing good agreements with other stations around the world, especially the Weybourne Atmospheric Observatory in the UK. We found long term trends during the period 2002-2018 of 2.31 ± 0.07 ppm yr$^{-1}$ for CO$_2$ and -21.2 ± 0.8 per meg yr$^{-1}$ for $\delta$(O$_2$/N$_2$) at Lutjewad, and 2.22 ± 0.04 ppm yr$^{-1}$ for CO$_2$ and -21.3 ± 0.9 per meg yr$^{-1}$ for $\delta$(O$_2$/N$_2$) at Mace Head. The notable differences in the year-to-year progression of $\delta$(O$_2$/N$_2$) and APO trends between Lutjewad and Mace Head might in part be caused by the sparse sampling frequency at Mace Head, but also may be indications of real influences from the various types of fossil fuel used, or from changes in ocean O$_2$ ventilation. Using the measurements at Lutjewad for 2002-2018, the partitioning of atmospheric CO$_2$ sinks into the global terrestrial
biosphere and the oceans are 1.7 ± 1.1 PgC yr⁻¹ and 2.0 ± 0.8 PgC yr⁻¹, respectively, and when calculated using the Mace Head record, these are 1.7 ± 1.1 PgC yr⁻¹ and 2.2 ± 0.9 PgC yr⁻¹, respectively. These values agree well with the numbers reported in the most recent Global Carbon Budget. The Halley record shows that the APO seasonal variations in the Southern Ocean are slightly larger than those in the Northern Hemisphere due to larger air-sea O₂ exchange there, and illustrates clearly the influences of oceanic processes on the variations in APO and atmospheric O₂. With better maintenance of our internal scale, more regular sampling frequency, and better quality-control of the sampling process, the reliability of our future flask measurements will be improved.

Data availability

The accompanying database comprises three csv files. The files contain the information on the CO₂, δ(O₂/N₂), and APO measurements (measured values and associated uncertainties) of the three stations, and are named after the corresponding station and the measured parameter (9 files in total).

All files are published by the ICOS Carbon Portal, and are available at https://doi.org/10.18160/qq7d-t060 (Nguyen et al., 2021).

The additional data presented in this paper are available upon request.

Author contributions

LNTN, HAJM, and ITL conducted the data analyses, produced all figures and tables, and wrote the manuscript. ITL and HAJM designed the methodology and framework for the calibration procedure of the DI-IRMS. BAMK conducted the technical work and prepared the flask samples at Lutjewad station, and carried out the CO₂ and δ(O₂/N₂) measurements from flasks collected at all three stations. HAS calibrated the CO₂ data at CIO. AEJ, NB, and TB performed the measurements at Halley station, prepared the flask samples, produced the data for comparison. PAP and ACM provided the data from Weybourne station. All co-authors contributed to the writing of the manuscript.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgements

We would like to thank our colleagues and collaborators at CIO (Janette J. Spriensma for logistics, Henk Jansen for help with the Optima DI-IRMS, Marcel de Vries for the Halley sampler construction, and Ramon R. Richie for help with the measurement database). Furthermore, we thank the collaborators at the stations in collecting and transporting the flask samples, specifically T.G. Spain at Mace Head, and the overwintering teams at Halley station. We thank Eric J. Morgan (SIO) for updating us with new values for the primary standard cylinders. We are also grateful to the UEA staff and students (Michael Pateki, Philip Wilson, Grant Forster and Leigh Fleming) and the Atmospheric Measurement and Observation Facility at the UK National Centre for Atmospheric Science (NCAS-AMOF), for kindly providing data from Weybourne. The work at Lutjewad received partial financial support over the years by the Dutch Research council (NWO), the European Union Integrated Project Carbo-Ocean (511176) and the Dutch national CATO-2 programme. Atmospheric O₂ and CO₂ measurements at Weybourne were supported by UK Natural Environment...
Research Council (NERC) grants NE/F005733/1, NE/I013342/1, QUEST010005 and NE/S004521/1. The Weybourne atmospheric O₂ and CO₂ measurements have been supported by the National Centre for Atmospheric Science (NCAS), funding agreement R8/H12/83/037 since December 2013 onwards. I.T.L. received funding from the Dutch Research Council (NWO) (Veni grant 016.Veni.171.095). A.E.J and N.B. were supported by the BAS core programme “Polar Science for Planet Earth”. T.B. and P.A.P. were supported by UK NERC studentships, NE/L50158X/1 and NE/K500896/1. P.A.P. and A.C.M. have received support from the NERC-funded DARE-UK (Detection and Attribution of Regional greenhouse gas Emissions in the UK) project, grant agreement no. NE/S004211/1.

References


Battle, M., Fletcher, S. M., Bender, M. L., Keeling, R. F., Manning, A. C., Gruber, N., Tans, P. P., Hendricks, M. B., Ho, D. T., Simmonds, C., Mika, R., and Paplawsky, B.: Atmospheric potential oxygen: New observations and their implications for some atmospheric and oceanic models, Global Biogeochemical Cycles, 20, n/a-n/a,


Halley VI Research Station: https://www.bas.ac.uk/polar-operations/sites-and-facilities/facility/halley/, last access: 02/04/2021.


Ishidoya, S., Aoki, S., Goto, D., Nakazawa, T., Taguchi, S., and Patra, P.: Time and space variations of the O$_2$/N$_2$ ratio in the troposphere over Japan and estimation of the global CO$_2$ budget for the period 2000–2010, Tellus B: Chemical and Physical Meteorology, 64, 10.3402/tellsb.v64i0.18964, 2012a.
Ishidoya, S., Morimoto, S., Aoki, S., Taguchi, S., Goto, D., Murayama, S., and Nakazawa, T.: Oceanic and terrestrial biospheric \( \text{CO}_2 \) uptake estimated from atmospheric potential oxygen observed at Ny-Ålesund, Svalbard, and Syowa, Antarctica, Tellus B: Chemical and Physical Meteorology, 64, 10.3402/tellusb.v64i0.18924, 2012b.


Carbon Cycle Gasses Halley Station, Antarctica, United Kingdom:

[Link](https://www.esrl.noaa.gov/gmd/dv/iadv/graph.php?code=HBA&program=ccgg&type=ts), last access: 02/04/2021.


Cucumbers Intercomparison: http://cucumbers.uea.ac.uk/, last access: 24/02/2021.


van Leeuwen, C.: Highly precise atmospheric oxygen measurements as a tool to detect leaks of carbon dioxide from Carbon Capture and Storage sites, University of Groningen, Groningen, The Netherlands, 2015.