

Wellington ¹⁴CO₂ record methods and updates

An overview of the measurement methods and summary table is given in the main paper (table 1). This supplementary material provides full details of the sampling methods used through time, compiling methodological information provided in previous reports on the Wellington record (Rafter and Fergusson 1959, Manning et al., 1990, Currie et al., 2011) along with methods newly applied in this new extension and refinement of the dataset. This supplementary material intentionally contains some repetition from the main paper so that all methodological details are available in a single document for users of the dataset.

1. Sampling sites

Samples from 1954 – 5 June 1987 were collected at Makara, on the west coast of the North Island (MAK, 41.25°S, 174.69°E, 300 m asl). The sampling site was moved to Baring Head on the South Coast of the North Island and 23 km southeast of Makara (BHD, 41.41°S, 174.87°E, 80 m asl). All samples since 8 July 1988 have been collected at Baring Head (figure 1). A set of tree ring samples was collected from two kauri trees from Nikau St, Eastbourne (NIK, 41.31°S, 174.89°E, 5 m asl). Sampling site for each sample is included in the dataset. The overall dataset uses site code BHD to be consistent with other greenhouse gas and isotope measurements from Baring Head.

2. Date of collection

Samples are collected over a period of several weeks, so the date of collection is determined as the middle of the sampling period. In some cases, the dates of start and end were not recorded or there is ambiguity as to whether the date recorded is the middle date, start date or end date.

For gas counting records, usually the date of collection was taken from Currie et al. (2011). These are usually the same as the dates reported in Manning and Melhuish (1990), but some samples reported at the end of the Manning and Melhuish (1990) dataset used the start date as date of collection (rather than the mean date), and this was corrected in Currie et al. (2011). For the present update, additional sample date information for a number of early measurements (1950s-1970s) was recorded in a comment field in the GNS Science radiocarbon database (RLIMS), and the date of collection was re-determined from this information.

For the AMS measurements starting in 1995, the date of collection was not included in RLIMS in most cases. We used Currie et al. (2011) and a separate sampling spreadsheet log kept at NIWA to determine the date of collection.

For samples where both start and end dates were recorded, the mean date is used as the date of collection. Where only the start date was recorded, we use two options. For samples measured by gas counting, we use the start date (as the number of days exposed was variable). For samples measured by AMS (dates after 1995), there are only a few for which the end date is missing:

- For NZA9240 (start date 1/4/96), NZA11523 (start date 15/6/96) and NZA11524 (start date 1/12/99). We assume a 14-day collection period since most other samples during this period were collected over 14 days.
- NZA14572 (start date 11/9/00) and NZA14573 (start date 22/9/00). We assume the end date is the same date as the start date for the following sample, which give days exposed of 11 and 18 days respectively.

3. Collection methods

3.1. Static NaOH absorption

The primary collection method is static absorption of CO₂ into sodium hydroxide (NaOH) solution, which is left exposed to air at the sampling site for ~2 weeks. The NaOH solution is first prepared to remove any CO₂ that was initially present when making up the solution. The standard preparation method involves bubbling CO₂-free air or nitrogen through deionized water for ~5 hours. We continue using this step, although it may be unnecessary since any initially-present CO₂ is removed in the following steps. NaOH pellets are then added to make 0.5 or 1 M. The NaOH pellets always absorb some ambient CO₂ when exposed to air, inevitably introducing some CO₂ contamination during the preparation, and this is removed from the solution by the addition of ~ 1 g barium chloride (0.004 mol) in ~50 mL of bubbled, deionized water. Any carbonate is precipitated as barium carbonate and the resultant nominally carbonate-free solution is decanted into smaller bottles under a stream of CO₂-free air or nitrogen. Since 2015, a further step adds 0.2 mL of 85% phosphoric acid to the solution after the barium chloride precipitation step. Any remaining barium is then precipitated as barium phosphate, ensuring that all the CO₂ absorbed during sample collection remains as carbonate rather than precipitating out as barium carbonate. The resulting solution is stored in sealed bottles until ready for use.

From 1954-1995, ~ 2 L of the 0.5 M NaOH solution was poured into a large (~450 cm² surface area) Pyrex® tray and left exposed to air for 1 - 2 weeks. Some early (1954-1970) samples may have been collected using different vessels. In a few early samples, air may have been pumped through the solution rather than passive absorption (Manning et al., 1990). Two samples were collected using Ba(OH)₂ rather than NaOH, in 1957 and 1958. As these are not outliers, they are included in the dataset, with a comment and flag.

Beginning in 1995, when the measurement technique changed from gas counting to AMS measurement (see section 3), the large trays were replaced with high density polyethylene (HDPE) bottles containing ~200 mL 0.5 M NaOH solution, with the depth of liquid kept the same as it was in the previously used trays. The lid is removed and the bottle is left open inside a Stevenson meteorological screen for ~2 weeks. Date of opening and closing bottle is recorded. The two methods (large tray and small bottle) have been compared, with no significant difference observed between the two methods (Currie et al., 2011).

Static NaOH absorption necessarily fractionates relative to CO₂ in the atmosphere; typical $\delta^{13}\text{C}$ values are -15 to -25 ‰ for these samples. This is corrected for in the data analysis.

3.2. Whole air flask samples

Flasks of whole air are collected by flushing ambient air through the flask for several minutes. The flask is then closed and allowed to fill to slightly over ambient pressure. All flask samples presented here were collected at the Baring Head site. Most of those collected during 1984 – 1993 were collected during southerly, clean air, conditions (Stephens et al., 2013). Flasks collected since 2013 were all collected during southerly conditions.

3.3. Tree rings

When trees photosynthesize, they faithfully record the $\Delta^{14}\text{C}$ of ambient CO_2 in their cellulose, the structural component of wood. Annual tree rings therefore provide a summertime (approximately September – April in the Southern Hemisphere) daytime average $\Delta^{14}\text{CO}_2$. Photosynthetic uptake varies during the daylight hours depending on factors including growth period, sunlight, and temperature (Bozhinova et al., 2013). For the Wellington location with typically high wind speeds and little short-term variability in CO_2 and other trace gases, this variability in photosynthesis is likely not a big impact on $\Delta^{14}\text{CO}_2$. The ring age we assign differs from standard nomenclature for Southern Hemisphere dendrochronology. We assign the mean age of the ring as January 1 of the year in which growth finished (i.e. the mean age of a ring growing from September – April), whereas dendrochronologists assign the “ring year” is as the year in which ring growth started (i.e. the previous year).

We collected cores from three trees close to the Baring Head site. A pine (*Pinus radiata*) located 10 m from the Baring Head sampling station (figure 1) yielded rings back to 1986 (Norris, 2015). New Zealand kauri (*Agathis australis*) is a long-lived hardwood species that has been widely used in dendrochronology and radiocarbon calibration studies (Hogg et al., 2013). The rings of this species are well defined and seldom missing, making this an ideal species for our comparison. We collected cores from two specimens planted in 1919 and 1920, located 20 m from one another in Eastbourne, 12 km from Baring Head (figure 1).

We evaluated the ring counts by the timing of the ^{14}C bomb spike. Shifting the Eastbourne record by one year in either direction moves the maximum in the tree ring record out of phase with the final Wellington $\Delta^{14}\text{CO}_2$ record (figure S1), confirming that the ring counts are correct. For the Baring Head pine, rings go back to only 1986, so we compare with the Eastbourne record, and find that the two records agree very well, with an insignificant mean difference of $-0.4 \pm 0.8 \text{ ‰}$ (figure S1).

Sampling of the tree rings for ^{14}C analysis is also a potential source of bias; ideally only the full ring for a given year would be sampled. In practice, it is difficult to ensure that the full ring is isolated without losing any material from that ring and no wood from the surrounding rings is included. To evaluate this, we measured replicate samples different cores from the same tree (Baring Head pine) or two different trees (Eastbourne kauri). For samples collected since 1985, all these replicates are consistent within their uncertainties (figure S2). We also tested shifting the Baring Head pine ring counts by

one year in either direction, and find that when we do so, the mean difference between the Baring Head pine and Eastbourne kauri is $+5.6 \pm 0.7$ ‰ or -8.4 ± 0.8 ‰ when moving the chronology one year back or forward respectively. These differences are too large to be plausible differences between these two sites, and give us further confidence that the ring counts are correct.

We do find large differences of 9.2, 44.5 and 4.9 ‰ for three replicates from Eastbourne in 1963, 1965 and 1971, which we attribute to small differences in sampling of the rings which were magnified by the rapid change in $\Delta^{14}\text{C}$ of up to 200 ‰ yr^{-1} during this period. Thus, the ring values during this period should be treated with caution.

4. ^{14}C Analysis methods

4.1. Extraction

From 1954 - 1995 CO_2 was extracted from the NaOH solution by acidification followed by cryogenic distillation (Rafter and Fergusson, 1959). AMS sample extraction from 1995 onward follows the general same method, but with smaller volumes (Currie et al., 2011), and this extraction is performed at NIWA. The $\delta^{13}\text{C}$ is measured at NIWA on an aliquot of CO_2 from the same extraction.

CO_2 is extracted from whole air flask samples by cryogenic extraction. Samples collected from 1984- 1993 were extracted at GNS Science by Graeme Lyon and archived as ampoules of pure CO_2 . In 2012, these tubes were cracked under vacuum to liberate the CO_2 . Any leakage during storage is readily by air present in the tube when it is cracked for transfer and tubes with leakage were discarded. An aliquot of each was measured for $\delta^{13}\text{C}$ to confirm that no fractionation had occurred during storage. Whole air samples collected since 2013 are analyzed for $\delta^{13}\text{C}$ and other trace gases and isotopes at NIWA and for the $^{14}\text{CO}_2$ measurement, CO_2 is extracted from whole air at GNS Science (Turnbull et al. 2015).

Cellulose was isolated from whole tree rings by first removing labile organics with a series of solvent washes (hexane, isopropanol and acetone) in a Soxhlet system (Norris, 2015). This was followed by oxidation to isolate the cellulose from other materials (Hua et al., 2000). The cellulose was combusted and the CO_2 purified following standard methods in the GNS Science laboratory (Baisden et al., 2013).

4.2. Graphitization (AMS only)

The LG1 graphitization system was used from 1995 to 2011 ($\text{NZA} < 50,000$) (Lowe et al., 1987), and replaced with the RG20 graphite system in 2011 ($\text{NZA} > 50,000$) (Turnbull et al., 2015). In both systems, CO_2 is reduced to graphite over iron catalyst in the presence of hydrogen gas. RG20 has online pressure monitoring to ensure the reaction goes to completion. On LG1, iron powder was used as catalyst; on RG20 iron oxide is reduced to iron online immediately prior to graphitization. All samples measured on the EN Tandem were prepared using LG1, and all except one sample measured on XCAMS were prepared using RG20.

4.3. Decay counting

Static NaOH samples were measured by conventional decay counting on the CO₂ gas from 1954-1995, described in detail in earlier publications (Manning and Melhuish, 1990; Currie et al., 2011). Samples measured by gas counting are identified by unique numbers with the prefix “NZ”.

4.4. AMS

All measurements made since 1995, including recent measurements of flask samples collected in the 1980s and 1990s, were measured by accelerator mass spectrometry (AMS). The unique identifiers for these samples have the prefix “NZA”, but the NZ and NZA numbers never overlap.

From 1995 to 2010, an EN Tandem AMS was used for measurement (NZA < 35,000, Zondervan and Sparks, 1996). Four graphite targets were prepared from each sample by splitting a single large CO₂ aliquot under equilibrium conditions, then reducing to graphite using the LG1 graphite system. Until 2005 (NZA < 30,000), only ¹³C and ¹⁴C were measured on the EN Tandem system, so the normalization correction for isotopic fractionation (Stuiver and Polach, 1977) was performed using the IRMS $\delta^{13}\text{C}$ value of the sample measured at NIWA. The reduction in scatter from 2005 onwards (NZA > 30,000) reflects the addition of online ¹²C measurement in the EN Tandem system. This allows direct online correction for isotopic fractionation that may occur during sample preparation and in the accelerator itself (Zondervan et al., 2015), and results in improved long-term repeatability.

Measurements of multiple targets are averaged to provide a weighted mean F¹⁴C and $\Delta^{14}\text{C}$. This averaging was done outside of the AMS database, and records of how these original calculations were done exist for only some of the results. In many cases, one or two targets appeared to be outliers and were excluded from the averaging calculation. In some cases, simple means of the multiple targets were used instead of weighted means.

We have recalculated the results for all of these multiple target samples, using consistent criteria for all. To determine and flag outliers, the reduced chi square statistic (χ^2_v) was calculated for each individual sample date (typically four targets) and for the full multi-target dataset. Note that in all except three cases, the multiple targets were all measured within the same AMS wheel, so wheel-to-wheel variability is not accounted for here.

First, including all multi-target measurements, we obtain $\chi^2_v=1.2$ from 598 targets. We identified outliers as those where χ^2_v for that sample date was greater than three and excluded the targets within that set for which the residual was greater than two. Recalculating χ^2_v after removing these outliers obtains $\chi^2_v=0.99$ from 585 targets, a satisfactory result. From EN Tandem measurements only, $\chi^2_v=1.02$ from 529 targets. For XCAMS, multiple targets were collected from both within and between wheels. The Baring Head NaOH samples measured on XCAMS give, $\chi^2_v=0.95$ from 56 targets. The weighted means have been recalculated from the F¹⁴C values, and $\Delta^{14}\text{C}$ calculated from F¹⁴C.

For samples where the outliers changed, differences of up to 5‰ from the values reported by Currie et al. (2011) occur. For other samples, differences of 0.1 – 0.5 ‰ occur, apparently because when results were initially calculated, straight rather than weighted means (weighted by reported uncertainties) were used in some cases and rounding errors in other cases.

A few samples were measured on the EN Tandem in two different wheels, and in some cases were assigned different R (laboratory) numbers. In those cases, Currie et al. (2011) did not always average all four replicates. This has now been revised.

In 2010, the EN Tandem was replaced with a National Electrostatics Corporation AMS, dubbed XCAMS (NZA > 34,000). XCAMS measures all three carbon isotopes, such that the normalization correction is performed using the AMS measured ^{13}C values (Zondervan et al., 2015). XCAMS measurements are made on single graphite targets, measured to high precision, typically better than 2‰ overall uncertainty. Replicate analyses within and between wheels have been used to determine long-term repeatability and overall uncertainty levels (Turnbull et al., 2015), including replicate measurements of samples from this dataset. Where more than one measurement was made for a given date, we report the weighted mean of all measurements.

Three different preparation methods have been used for the primary standards (oxalic acid I, OxI) using in determining $\Delta^{14}\text{CO}_2$. Sealed tube combustion was used for all the EN Tandem measurements, whereby sufficient solid OxI material was combusted to provide five or six aliquots. The resultant CO_2 was split and then graphitization was performed separately for each aliquot. Since 2010, OxI has also been prepared by combustion of single aliquots by elemental analyzer and subsequent cryogenic CO_2 collection (Baisden et al., 2013). Both these methods show larger scatter than air standard materials, and there is an offset of -1.5‰ in samples measured with OxI made by sealed tube combustion, and those samples have been corrected for this offset (Turnbull et al., 2015). To reduce the scatter in OxI due to variability in combustion (which air samples do not undergo), we prepared two large flasks of CO_2 from OxI, and in 2015, began taking aliquots of this CO_2 as primary standard material. There is no offset relative to the elemental analyzer combustion method, and repeatability is slightly improved.

During the transition between the EN Tandem and XCAMS instruments, three Wellington $^{14}\text{CO}_2$ samples were measured on both instruments, and XCAMS results were 4.2‰ higher for these three samples. However, this jump during the transition does not appear consistent across wheels, and the larger quality control dataset which shows no significant difference in several solid sample quality control standards (Baisden et al., 2013), and is not addressed further here.

5. Results format

5.1. Date of collection

NaOH samples are collected over a period of typically two weeks, and sometimes much longer. We determine the date of collection as the average of the start and end collection

dates. In cases where the end date was not recorded, we use the start date. For a few samples, the sampling dates were not recorded or are ambiguous, and those results have been flagged and excluded from the final dataset.

5.2 $\Delta^{14}\text{C}$ and $F^{14}\text{C}$

Results are reported here in $F^{14}\text{C}$ (Reimer et al., 2004) and $\Delta^{14}\text{C}$ (Stuiver and Polach, 1977). $F^{14}\text{C}$ is corrected for isotopic fractionation and blank corrected. $F^{14}\text{C}$ is calculated from the original measurement data recorded in our databases, and has been updated for only a handful of records where errors were found in the $\delta^{13}\text{C}$ value originally used in the normalization correction. No changes in measured values were made except in cases where multiple targets were averaged.

$\Delta^{14}\text{C}$ is derived from $F^{14}\text{C}$, with a correction for radioactive decay since the time of collection (Stuiver and Polach, 1977). The date of collection is required for this calculation. In previously published versions of the dataset, the gas counting and EN Tandem results used only the year of collection in the $\Delta^{14}\text{C}$ calculation, and in many cases, $\Delta^{14}\text{C}$ was initially reported using the date of measurement. Currie et al. (2011) recalculated $\Delta^{14}\text{C}$ in a separate database, using the date of collection, but assumed that the date of measurement, rather than the year of measurement, had been used. Here we went back to the original measured $F^{14}\text{C}$, and to the sample collection database to find the date of collection. All $\Delta^{14}\text{C}$ results are recalculated using the decimal date of collection. This results in changes of a few tenths of permil in most $\Delta^{14}\text{C}$ values relative to those reported by Currie et al (2011) and Manning and Melhuish (1990). Samples for which changes have been made relative to the previously published results are indicated by the third column quality flag provided in the supplementary dataset.

5.3. Uncertainties

The primary source of uncertainty is the ^{14}C counting statistical uncertainty determined from the number of beta decays (gas counting) or ^{14}C counts (AMS). For AMS measurements, we add an additional error term, determined from the long-term repeatability of secondary standard materials and added in quadrature to the AMS uncertainty. In early AMS measurements, this was 4%, and has decreased through time to 0.12%. For the EN Tandem results, this was determined from the performance of OxI primary standard targets through time, which may somewhat overestimate the uncertainty for air samples (Turnbull et al., 2015). For XCAMS measurements, long-term repeatability of air samples has been assessed based on repeated splits of CO_2 from a subset of these NaOH samples, as well as repeated extractions of aliquots of CO_2 from pressurized tanks of whole air. These repeated aliquots are measured both within the same AMS wheel and across multiple wheels. Within individual wheels, the counting statistical uncertainty is sufficient to explain the full variability, but an additional uncertainty of 0.12% is required to explain the spread across multiple wheels (Turnbull et al., 2015).

6. Smooth curve fits

In addition to the raw measured $\Delta^{14}\text{CO}_2$ values, we calculate a smooth curve fit (“fit” in

the supplementary datasets) and deseasonalized trend from the Wellington $\Delta^{14}\text{C}$ and F^{14}C datasets. The deseasonalized trend (“trend” in the supplementary datasets) may be more useful than the raw data for aging of recent materials (e.g. Reimer et al., 2004; Hua et al., 2013). Acknowledging that the 1995-2005 period is variable and possibly biased in the Wellington record, we also provide in the supplementary material an alternative mid-latitude Southern Hemisphere smooth curve fit and deseasonalized trend in which the Wellington data for 1995-2005 has been removed and replaced with the Cape Grim data for that period (Levin et al., 2010).

Curvefitting is particularly challenging for the $\Delta^{14}\text{CO}_2$ record, since (a) there are data gaps and inconsistent sampling frequency, (b) the growth rate and trend vary dramatically and (c) the seasonal cycle changes both in magnitude and phase (section 4.2). We chose to use the CCGCRV fitting procedure (Thoning et al., 1989; www.esrl.noaa.gov/gmd/ccgg/mb1/crvfit/), which uses fast Fourier transform and low-pass filtering techniques to obtain a smoothed seasonal cycle and long term trend from atmospheric data. This technique can readily handle the data gaps, inconsistent sampling frequency, and rapid changes in the seasonal cycle and trend.

CCGCRV assigns a single set of harmonic terms across the full time period, which is inappropriate in this case of large variation in the seasonal cycle. Thus, we separate the record into five time periods: 1954-1965, 1966-1979, 1980-1989, 1990-2004, 2005-2014. These divisions were chosen based on major changes in the raw observational growth rate, seasonal cycle and data quality. The peak of the bomb spike in the Southern Hemisphere (1965) results in a very large change in seasonality that makes an obvious cutoff point. There is an obvious change in seasonality in the raw observations in 1979 – 1980. The 1990 to 2004 period was grouped to include the time when flask measurements have supplemented original NaOH measurements, and the 1995 – 2004 period with noisy data.

The other widely used fitting procedure, seasonal trend decomposition using locally weighted scatter plot smoothing (STL, Cleveland et al. 1990; Pickers et al., 2015) assumes that the seasonal cycle and trend change only gradually over a specified time period. This assumption is problematic for the $\Delta^{14}\text{CO}_2$ time-series, due to the rapid changes in the trend and seasonal amplitude during and following the bomb spike. Using this method would necessitate both gap-filling the record and dividing the record into time periods (as we have done for CCGCRV), giving no advantage over CCGCRV.

For each time period, we use CCGCRV with one linear and two harmonic terms and fit residuals are added back using a low-pass filter with an 80-day cutoff in the frequency domain. At each transition, we overlapped a two-year period and linearly interpolated the two fits across that two-year period to smooth the transitions caused by end effects. We tested different overlap periods, and found that two years was optimal to minimize end effects and retain the benefit of separating the time periods. The deseasonalized trend was determined from the full dataset rather than the five time periods, as it does not include the seasonality and produces the same result in either case.

We tested other time period divisions, and our chosen time divisions have the lowest

mean residual difference from the measured $\Delta^{14}\text{CO}_2$, indicating the best fit to the data (we tested only periods of >10 years since it is difficult to draw conclusions about seasonal cycles from shorter periods when the seasonal cycle amplitude is small relative to the measurement uncertainty). The mean difference between the fitted curve and the measured $\Delta^{14}\text{CO}_2$ values is 3.8 ‰, consistent with the typical measurement uncertainty for the full dataset. Further, the residuals are highest for the early period (1954-1970) at 6 ‰, consistent with the larger measurement errors at that time of ~6 ‰. The residuals improve as the measurement errors reduce, such that since 2005, the mean residual is 2 ‰, consistent with the reported 2 ‰ uncertainties. The exception is the 1995-2005 period where the mean residual difference of 5 ‰ is substantially higher than the mean reported uncertainty of 2.5 ‰, reflecting the apparent larger scatter during this period as discussed in section 3.5.3.

The one-sigma uncertainty on the smoothed curve and deseasonalized trend were determined using a Monte Carlo technique (n=100). Each data point was perturbed by a random normal error based on the reported uncertainty of that data point, such that the standard deviation of all perturbations would equal the reported uncertainty to derive the one-sigma uncertainty for the smooth curve. This is provided for further users of the dataset, and may be particularly helpful when the dataset is used for aging of recent materials.

6. Data validation

6.1. Tree ring comparison

Over the more than 60 years of the Wellington $\Delta^{14}\text{CO}_2$ record, there have necessarily been many changes in methodology, and the tree rings provide a way to validate the full record, albeit with lower resolution. Due to the possible sampling biases in the tree rings, use them to validate the existing measurements, but do not include them in the main Wellington $\Delta^{14}\text{CO}_2$ dataset. We include the tree ring data in a separate table for researchers who may want to use these data in addition to the main Wellington $\Delta^{14}\text{CO}_2$ dataset.

During the rapid $\Delta^{14}\text{CO}_2$ change in the early 1960s, there are some differences between the kauri tree ring and Wellington $\Delta^{14}\text{CO}_2$ records. The 1963 and 1964 tree ring samples are slightly lower than the concurrent $\Delta^{14}\text{CO}_2$ samples. The peak $\Delta^{14}\text{CO}_2$ measurement in the tree rings is 30‰ lower than the smoothed $\Delta^{14}\text{CO}_2$ record, and 100‰ lower than the two highest $\Delta^{14}\text{CO}_2$ measurements in 1965. These differences are likely due to small errors in sampling of the rings, which will be most apparent during periods of rapid change.

Prior to 1960 and from the peak of the bomb spike in 1965 until 1990, there is remarkable agreement between the tree rings and Wellington $\Delta^{14}\text{CO}_2$ record, with the wiggles in the record replicated in both records. And since 2005, there is excellent agreement across all the different records. Some differences are observed in 1990-1993 and 1995-2005, which we discuss in the following sections.

6.2. 1990-1993 excursion

An excursion in the gas counting measurements between 1990 and 1993 has previously been noted as a deviation from the Cape Grim $\Delta^{14}\text{CO}_2$ record (Levin et al., 2010) during the same period. Cape Grim is at a similar latitude and also draws air from Southern Ocean. The Wellington and Cape Grim records overlap during almost all other periods.

We use archived CO_2 from flask samples to evaluate this period of deviation. First, the recent flask samples agree very well with the NaOH static samples from the same period (figure 2). This indicates that despite the difference in sampling period for the two methods, flask samples reflect the $\Delta^{14}\text{CO}_2$ observed in the longer-term NaOH static samples. We selected a subset of the archived 1984 - 1992 CO_2 samples for measurement, mostly from Southerly wind conditions, but including a few from other wind conditions. Each sample was liberated and the $\delta^{13}\text{C}$ measured before $\Delta^{14}\text{CO}_2$ measurement was performed. A few of the samples had been compromised, and this was readily apparent as they contained air rather than CO_2 . $\delta^{13}\text{C}$ from all the remaining samples was in agreement with $\delta^{13}\text{C}$ measured from separate flasks collected at Baring Head and measured for $\delta^{13}\text{C}$ by Scripps at close to the time of collection. These flask $\Delta^{14}\text{CO}_2$ measurements do not exhibit the excursion seen in the NaOH static samples, implying that the deviation observed in the original NaOH static samples may be a sampling error. Annual tree rings from both the kauri and pine follow the flask measurements for this period (figure 2), confirming that the NaOH static samples are anomalous.

The 1990-1993 period was characterized by major changes in New Zealand science, both in the organizational structure and personnel. Although we are unable to exactly reconstruct events that time, we hypothesize that the NaOH solution preparation was conducted slightly differently, perhaps omitting the BaCl_2 precipitation step for these samples. This would result in contaminating CO_2 absorbed on the NaOH before the solution was prepared, which would result in higher $\Delta^{14}\text{CO}_2$ observed in these samples than in the ambient air. In any case, these values are anomalous and we remove the original NaOH static sample measurements between 1990 and 1993 and replace them with the new flask measurements for the same period.

6.3. 1995-2005 variability

In 1995, the measurement method was changed from gas counting to AMS. During the first ten years of AMS measurements, the record is much noisier than during any other period (figure 2). In 2005, online ^{12}C measurement was added to the AMS system, substantially improving the measurement accuracy, and the noise in the $\Delta^{14}\text{CO}_2$ record immediately reduced.

The remaining NaOH solution for all samples collected since 1995 has been archived, and typically only every second sample collected was measured, with the remainder archived without sampling. In 2011-2016, we revisited the 1995-2005 period, remeasuring some samples that had previously been measured and some that had never been measured for a total of 52 new analyses.

The new measurements on this period do show reduced scatter over the original analyses, particularly for the period from 1998-2001 where the original analyses appear anomalously low and in 2002-2003 when the original analyses appear anomalously high. Yet there remain a number of both low and high outliers in the new measurements. These are present in both the samples that were remeasured and in those for which this was the first sample from the bottle. This suggests that a subset of the archived sample bottles were either contaminated at the time of collection, or that some bottles were insufficiently sealed, causing contamination with more recent CO₂ during storage. Comparison with the tree ring measurements and with the Cape Grim record (Levin et al., 2010) suggest that the measurements during this period may, on average, be biased high as well as having additional scatter (figure 3). Nonetheless, in the absence of better data, we retain both the original and remeasured NaOH sample results in the full record.

8. Existing results added to the Wellington record

A number of samples were identified in RLIMS (The GNS Science radiocarbon database) that were measured but not included in the dataset provided by Currie et al. (2011) and have been added to the record:

- NZA11524 was a “replacement” for NZA11278 which was discarded due to analytical problems (but these were collected on different dates).
- NZA11523 was a remeasurement of another extraction from bottle 12 (NZA9382) which was an obvious outlier in the record. Both samples had the same sample date, sampling bottle and $\delta^{13}\text{C}$.
- NZA8462-8468 are all extractions from the same sample bottle. They were originally given two different R numbers, and each was given a different NZA number. These have all been averaged together in this updated record.
- NZA26053 (bottle 96) and NZA39544 (bottle 42) are missing from the NIWA sample collection spreadsheet but are recorded in RLIMS. Collection dates can be guessed at, but not confirmed so have been flagged as having an analytical problem and been excluded from the dataset.
- NZA 11191, 11192, 11278, 39189 had analytical problems identified and have been flagged and removed.
- NZA11522 had a sampling problem and has also been flagged.
- Six gas counting samples were identified in RLIMS as contaminated and flagged (these were also excluded from the Currie et al. (2011) dataset).
- NZ 2284, 2290, 5700, 6535 were discovered in the RLIMS database that were not included in the Currie et al., (2011) dataset. These are not outliers and are added here.

9. Quality flags used in the datasets

A three-character flagging system is used. When no flag is indicated, each character is “.”, resulting in “...” for a result with no flags.

The first character is a “hard” flag. Any value in this column indicates a sampling or measurement problem and this result should be ignored in any analysis. These are retained in our internal database, but have been removed from the distribution dataset and may be obtained on request.

501 A.. Analytical measurement problem
502 N.. Problem at point of collection
503 C.. Contamination in sample
504 O.. This target was excluded as an outlier from the set of multiple targets measured
505 for this sample
506

507 The second character is a “soft” flag, used to indicate that this result should be excluded
508 from the background record analysis. Here it is used to indicate that the result is an
509 outlier relative to the rest of the record, but there is no positively identified sampling or
510 measurement problem and we therefore cannot remove the datapoint a priori. These
511 values are included as a separate table in external distribution dataset but should not
512 usually be included in any analysis.

513 .F. this measurement is a >3 sigma outlier from the smooth curve fit to the dataset

514 .A. period from 1990-1993 where the static NaOH samples deviate from recently
515 measured flasks collected during the same time period and also deviate from the Cape
516 Grim ¹⁴CO₂ measurements (Levin et al., 2010).

517 .T. period from 1995-2005 when the static NaOH samples have larger scatter than
518 expected, and may be biased high, *but are retained in the record* in the absence of other
519 data.
520

521 The third character is informational only and *does not indicate a problem with the result*.
522 Here we use this flag to indicate changes from the results reported by Manning and
523 Melhuish (1990) and/or Currie et al. (2011) and/or results initially reported from the
524 Rafter AMS database to the NIWA database.

525 ..D date of collection and/or decay correction has changed

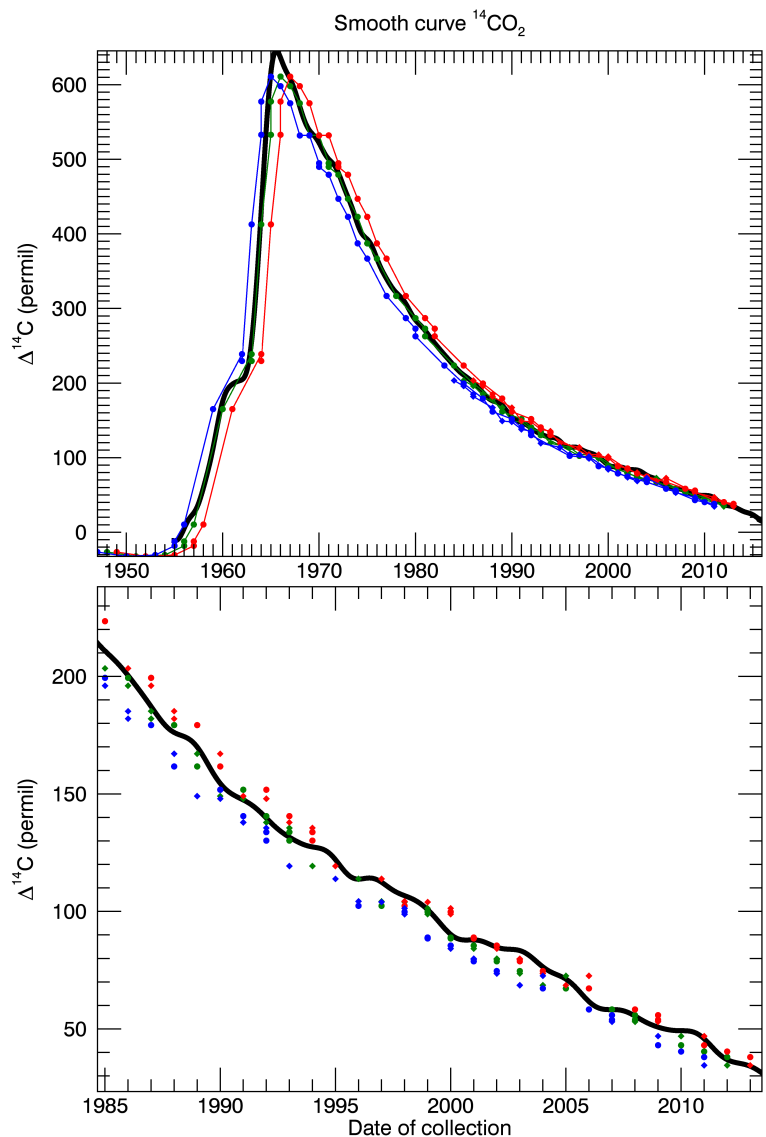
526 ..N normalization correction has changed

527 ..A change in how the average of multiple targets from the same sample was
528 calculated.

529 ..O sample was excluded from multi-target average in Currie et al. (2011), but is now
530 included in the analysis (applies only to the individual measurements, not listed for the
531 mean values)
532

533 10. Data Availability

534 The datasets presented here are included in the supplementary material as a separate data
535 file. Regularly updated versions of these datasets can be accessed through the World
536 Data Centre for Greenhouse Gasses (<http://ds.data.jma.go.jp/gmd/wdcgg/>) or directly
537 through GNS Science ([https://gns.cri.nz/Home/Products/Databases/Wellington-](https://gns.cri.nz/Home/Products/Databases/Wellington-atmospheric-14CO2-record)
538 [atmospheric-14CO2-record](https://gns.cri.nz/Home/Products/Databases/Wellington-atmospheric-14CO2-record)) or NIWA (<ftp://ftp.niwa.co.nz/tropac/>).



540
 541 Figure S1. Evaluation of tree ring counts for the full record (top) and since 1985
 542 (bottom). Black line is the smooth curve fit to the final Wellington $\Delta^{14}\text{CO}_2$ dataset.
 543 Green points and line are the combined tree ring data from both Baring Head and
 544 Eastbourne, using the assigned ring counts. Blue and red lines/points are the tree ring
 545 data shifted by one year in either direction. Adjustment in either direction clearly
 546 misaligns the upswing of the bomb pulse and the timing of the peak $\Delta^{14}\text{CO}_2$ values. In
 547 the later part of the record, since 1985, the trend through time is smaller, but the blue
 548 points (shifted one year older) consistently underestimate the Wellington $\Delta^{14}\text{CO}_2$ values
 549 throughout the record. The red points (shifted one year younger) overestimate until about
 550 1995, when it is difficult to discern from this comparison whether the red or green
 551 (assigned ring count year) is correct, but there is excellent agreement between the records

from four cores and three different trees, giving us confidence that the ring counts are correct in all cases.

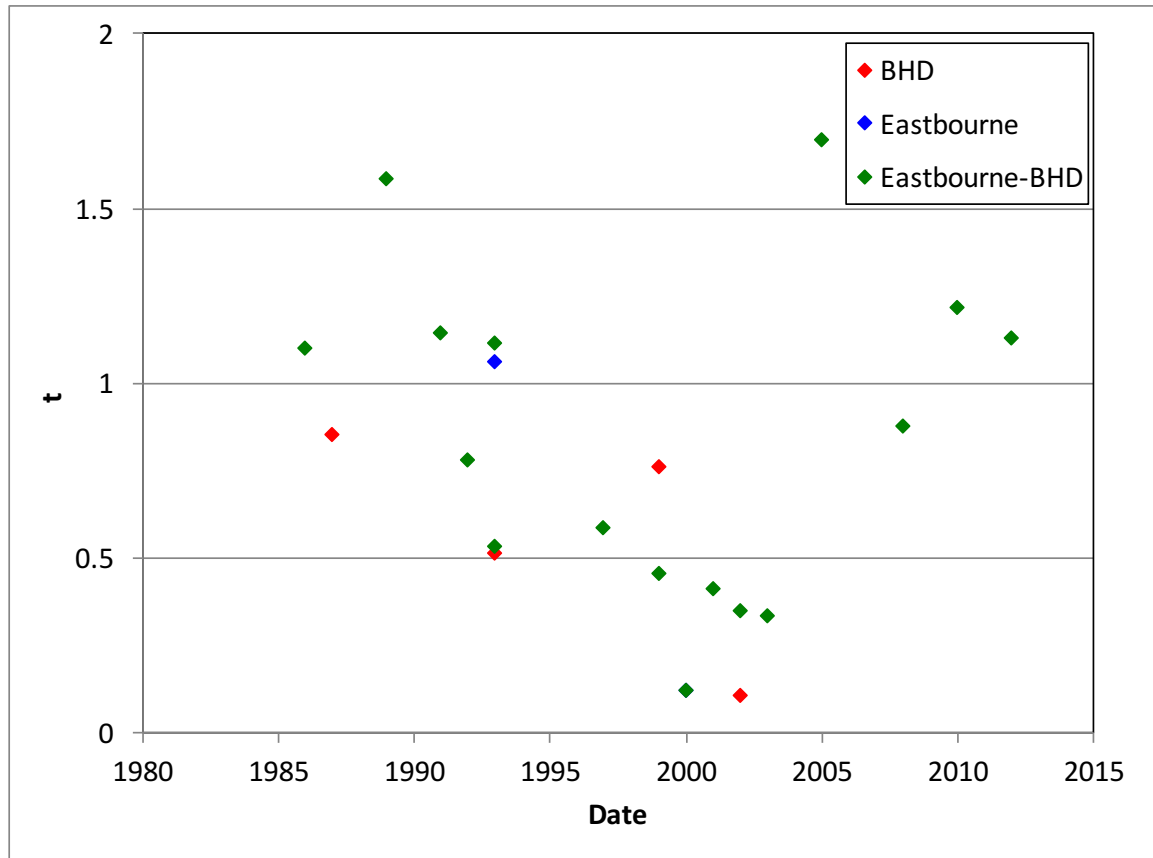


Figure S2. Evaluation of replicate tree ring measurements, shown as Student's t statistic for each replicate pair (pair difference divided by the sum of squares of the individual uncertainties). Red: replicates are the same ring taken from two different cores of the same pine tree at Baring Head. Blue: replicates are the same ring taken from cores of two different kauri trees from Eastbourne (located 20 m apart). Green: replicates are the same ring taken from the Baring Head pine and one Eastbourne tree. A t-value of 1 indicates that the two replicates agree within one standard deviation. Of the 22 replicate pairs, 14 (64%) agree within one standard deviation and 100% agree within two standard deviations. Using a paired (dependent) sample t-test to combine all the pairs of results, the overall t statistic is 0.97 ($p=0.34$), that is, given the assigned uncertainties, there is no significant difference between the pairs of measurements. There is no significant bias between the Baring Head and Eastbourne trees (mean difference is -0.4 ± 0.8 for the 16 replicate pairs that span both sites).