Annual Statistics and Quality Assurance of the 2022 Ambient Monitoring Archive

November 2024

Table of Contents

Purpose

This document outlines the major steps in calculating annual statistics and Quality Assurance (QA) procedures for each pollutant, site, sampling duration, and year for the 2022 Ambient Monitoring Archive (AMA) for the Hazardous Air Pollutants (HAPs) (i.e., “the 2022 Archive”).

Steps for Calculating Annual Statistics

Calculate derived local conditions

Air quality samples are collected using either local conditions (LCs) or standard conditions (STDs), each of which have different applications (e.g., risk, model evaluation, defining detection limits, etc.). In instances where LCs are not available, but STDs are, the LCs can be derived as outlined below. In short, when both conditions are available, the ratio between local and standard conditions is calculated and the ratio is applied elsewhere where needed. In the 2022 Archive, this is done for approximately 5% of records.

Calculate local conditions/standard conditions ratios

1. Subset Archive data to those records that have both standard and local conditions available.
2. Calculate the average STDs and the average LCs by pollutant/site/POC/day/sampling duration. (POC is the Parameter Occurrence Code.)
3. Average the ratios calculated above by site/day. Note: this is averaged across pollutants and sampling durations.
4. Average the ratios calculated above by site/quarter.
5. Average the ratios calculated above by site/year.
6. Average the ratios calculated above for the site across all available years.

Pre-process the data

Subset data

1. Removed data that are not relevant or appropriate for the calculation of annual statistics (e.g., a sampling duration of 1 month and integrated 2-weeks samples, pollutants that are not HAPs, etc.).

Apply the LC/STD ratios to derive LCs

1. Identify all the records which are presented in STDs (i.e., SAMPLE_VALUE_STD_FINAL_TYPE=S).
2. If the ratio exists for that pollutant/site/POC/day/sampling duration, derive LCs by multiplying the sample collected using STDs and the ratio.
3. If the above ratio does not exist but the ratio exists for that site/day, derive LCs by multiplying the sample collected using STDs and the ratio.
4. If the above ratio does not exist but the ratio exists for that site/quarter, derive LCs by multiplying the sample collected using STDs and the ratio.
5. If the above ratio does not exist but the ratio exists for that site/year, derive LCs by multiplying the sample collected using STDs and the ratio.
6. If the above ratio does not exist but the ratio exists for that site, derive LCs by multiplying the sample collected using STDs and the ratio.
7. If the ratio by site is missing for the sample collected using STDs, assume $ratio = 1$ (i.e., LCs equal STDs).

Sample are either originally collected in LCs, converted to LCs using collocated or nearby temperature and pressure data, or LCs are derived using the methodology described above. These values are carried forth to calculate daily averages unless otherwise noted.

Calculate daily averages

Separate data

1. Separate remote and non-remote data. Remote data include records from NOAA and MIT data sources. All other data collected from sources other than NOAA and MIT are considered “non-remote.”
2. Separate the non-remote data by minute sampling durations and hourly sampling durations.

Calculate daily averages from the minute sampling durations (non-remote data)

1. Calculate the number of samples per pollutant/site/POC/day/hour/sampling duration. If the minimum number of samples is met (see Table 5), average up to the hour. This constitutes a valid hourly average. If the minimum number of samples is not met, all samples are removed.
2. From all valid hours, calculate the number of samples per pollutant/site/POC/day/sampling duration. If there are at least 18 valid hourly averages in a day (see Table 5), averaged up to the day. This constitutes a valid day from the minute data. If the minimum number of samples is not met, all samples are removed.

Calculate daily averages from the hourly sampling durations (non-remote data)

1. Calculate the number of samples per pollutant/site/POC/day/sampling duration. If the minimum number of samples is met (see Table 5), average up to the day. This constitutes a valid day from the hourly data. If the minimum number of samples is not met, all samples are removed. Note: Sampling durations of 90 MINUTES and 150 MINUTES are considered to have an hourly sampling duration.

Average across POCs (non-remote data)

1. If there are multiple POCs per pollutant/site/day/sampling duration, remove all samples equal to zero if at least 50% of samples are NOT zero.
2. If there are multiple POCs per pollutant/site/day/sampling duration, remove all samples NOT equal to zero if more than 50% of the samples ARE zero.
3. After the appropriate valid collocated daily averages are removed (if necessary), average the remaining valid collocated daily averages by pollutant/site/day/sampling duration across POCs.

Note: Removing some collocated POCs ensures that collocated values that are mostly zero will result in a zero daily average and collocated values that are mostly NOT zero will result in a daily average that does not contain zeros.

Calculate daily averages for remote data

1. Average by pollutant/site/day/sampling duration. Note: The remote data do not require a minimum number of minute or hourly samples. The remote data may have collocated monitors distinguished by POCs for a given pollutant/site/day/sampling duration.

Calculate annual averages from daily averages

An annual average is calculated by averaging valid daily averages by pollutant/site/year/sampling duration meeting the criteria outlined below.

Valid quarters

Annual averages are calculated for a given pollutant/site/year/sampling duration only if there are at least *three* valid quarters for a given pollutant/site/year/sampling duration. A valid quarter is defined as a quarter having at least *six* daily averages per pollutant/site/quarter/sampling duration. If there are less than three valid quarters, all daily averages for a given pollutant/site/year/sampling duration are removed and an annual average is not calculated. Note: All valid daily averages are used in the calculation of an annual average as long as the three valid quarters threshold is met. Therefore, it is possible for an annual average to contain data from one “invalid” quarter of data if the criterion for the remaining three valid quarters is met. Note: a daily non-detect average assigned as zero can be included in the calculation of a valid quarter.

Calculate annual averages

Annual averages and corresponding statistics are calculated by pollutant/site/year/sampling duration (see Table 1 for all the annual statistics calculated). Some annual statistics are also calculated from valid daily averages using regression on order statistics (ROS) that allows for censored values (i.e., daily averages that are non-detect). To ensure a stable statistic, the ROS annual average is only calculated if the percentage of valid daily averages below the method detection limit (MDL) does not exceed 80. The MDLs for the daily censored values are assigned: $min\{MDL, \min\,(day\ in\ year)\}$, where $\min\,(day\ in\ year)$ is the minimum daily non-zero values for a given pollutant/site/year/sampling duration. Note: The MDLs are in STDs while the original samples may be collected using LCs. If there were multiple MDLs used in the construction of the daily average (e.g., through multiple POCs or through a sub-daily sampling duration), the average MDL is taken.

Table 1. Set of calculated annual statistics given the condition of the collected data and the treatment of non-detects.

Condition of Data	Treatment of Non-Detects	Annual Statistics Calculated
Local, converted local, and derived local	zero	arithmetic mean, variance, maximum, 10th percentile, 25th percentile, median, 75th percentile, and 90th percentile
Local, converted local, and derived local	censored	arithmetic mean, 10th percentile, 25th percentile, median, 75th percentile, and 90th percentile
Standard only	zero	arithmetic mean, variance, maximum, 10th percentile, 25th percentile, median, 75th percentile, and 90th percentile
Standard only	censored	arithmetic mean, 10th percentile, 25th percentile, median, 75th percentile, and 90th percentile

Descriptions of R Code Used to Calculate Annual Statistics

- AMA_analysis.R: Sources and calls all functions used to create annual statistics.
- AMA_LC2STDRatio.R: Creates an average ratio between LCs to STDs based on an average across days, quarters, years, and all years for a given site.
- AMA_preprocessing.R: Pre-processes the data including removing some select sites and parameters and applies the LC/STD ratio to data if applicable.
- AMA_duration2daily.R: Calculates daily averages by pollutant/site/day/sampling duration from sub-hourly and sub-daily sampling durations.
- AMA_daily2annual.R: Calculates annual statistics by pollutant/site/year/sampling duration based on local and standard conditions and treatments of non-detect data.
- AMA_file.R: Creates the final annual file.

Quality Assurance of the 2022 Archive

QA of the 2022 Archive was primarily performed by comparing the ROS annual averages from the 2022 Archive to the 2021 Archive by pollutant/site/year/sampling duration using the criteria outlined below. Two sets of checks were performed: (1) summarizing “difference categories” and (2) summarizing “suspect categories” (see the QA Difference and Suspect Categories section below). Currently, there is no corrective action for annual averages and/or records flagged.

The *difference categories* summarize why annual averages vary between the 2022 Archive and the 2021 Archive. Due to the comparison to the previous version of the Archive, the difference categories are only applied to the “overlap years” (i.e., 1990 – 2021, with the newest year, 2022, excluded). In short, annual averages between the old and new version of the Archive may be (1) the same, (2) different, (3) added, or (4) removed. The difference categories identify the reasons why annual averages fall into these four categories. Most annual averages in the overlap years are the *same* between the 2021 Archive and the 2022 Archive (Table 2).

The *suspect categories* are categories used to identify the newest annual averages (i.e., annual averages from 2022) that are very different than the rest of the annual averages (i.e., averages from 1990 – 2021). The suspect categories depend on the comparison sets for the newest annual averages. The suspect categories first establish if (1) there is a corresponding annual average in 2021 for that pollutant/site/sampling duration, (2) there is no corresponding annual average in 2021 but there is a time series for that pollutant/site/sampling duration, or (3) there is no annual average for 2021 and no time series for that pollutant/site/sampling duration.

The difference categories and the suspect categories are further described below.

Table 2. Counts of annual averages in each QA category in the 2022 Archive.

QA Category Name	Count
All annual averages	316,564
Annual averages (ROS*)	225,406
Overlap years^ (ROS)	217,738
Overlap years – Same (ROS)	212,121
Overlap years – Different (ROS)	4,091
Overlap years – Added (ROS)	1,526
Overlap years – Removed[%] (ROS)	35
Newest year[&] (ROS)	7,668

*ROS = Regression on Order Statistics

^Overlap years = 1990 – 2021

[%]"Removed" is not included in the count of total ROS annual averages

[&]Newest year = 2022

QA Difference and Suspect Categories

The QA difference categories describe the reasons for which an annual average and/or individual records in the 2022 Archive is (are) different than an annual average in the 2021 Archive for a given pollutant/site/year/sampling duration. Because of the comparison to the previous Archive, difference categories only apply to annual averages in the overlap years (i.e., years 1990 – 2021). When comparing the ROS annual means in the overlap years, annual averages are either (1) the same, (2) different, (3) added, or (4) removed. The number of annual averages that fall into each of the difference categories is presented in Figure 1 and Table 3. There are six difference categories in the added annual averages and seven difference categories in the removed annual averages.

Multiple differences categories may occur for a single different annual average. For example, an annual average may contain 60 daily values of which 40 values are identical, 15 values have rounding differences, and 5 reported values are new. Different annual averages are primarily driven by unit conversions being updated. Detailed differences in records among the different annual averages are presented in Figure 2 and Table 4.

In practice there is typically one difference category that is driving an annual average to be added or removed. These are presented in Figure 1 and Table 3. There are 1,526 and 35 added and removed annual averages, respectively. Added annual averages primarily occur because new records were added to the 2022 Archive.

The QA suspect categories in the 2022 Archive explore whether or not the newest year of data (i.e., 2022) has (1) a large difference compared to the previous year (i.e., 2021), (2) a large difference within the time series for a given pollutant/site/sampling duration (if the previous year does not exist), or (3) greater than the 95th percentile for a given pollutant/year/sampling duration if a time series does not exist.

Unlike the difference categories, the suspect category thresholds are determined only from the 2022 Archive. The count of annual averages falling into each suspect category is presented in Figure 1 and Table 3.

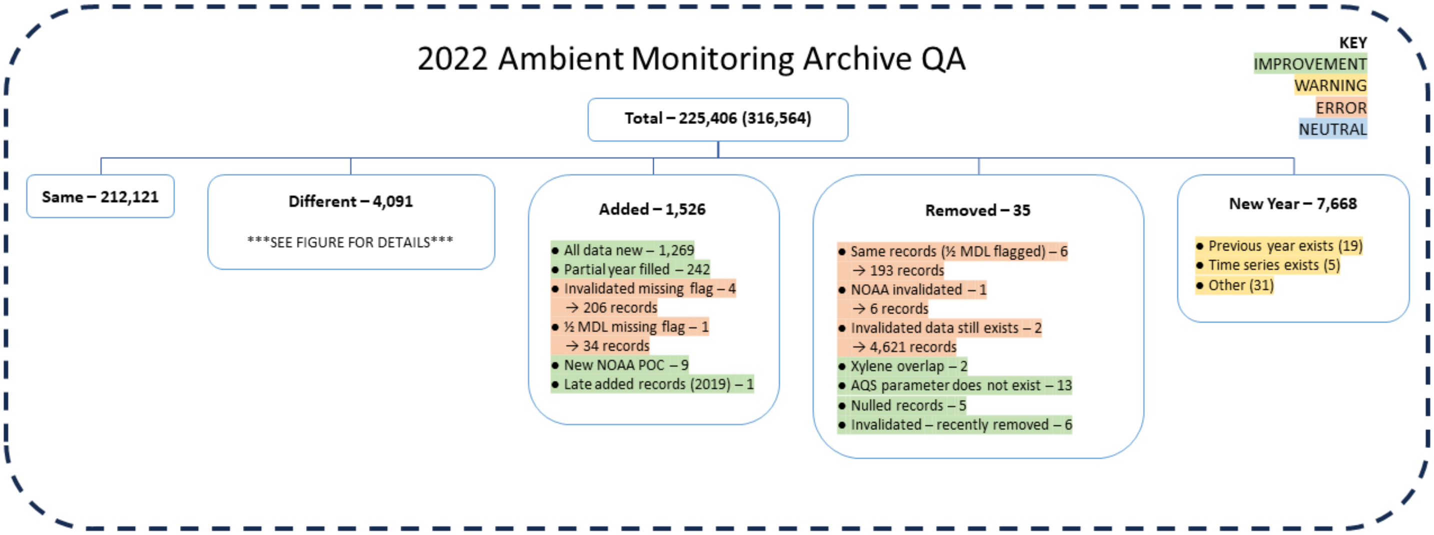
2022 Ambient Monitoring Archive QA
KEY
IMPROVEMENT
WARNING
ERROR
NEUTRAL
Total – 225,406 (316,564)
Same – 212,121
Different – 4,091
SEE FIGURE FOR DETAILS
Added – 1,526
• All data new – 1,269
• Partial year filled – 242
• Invalidated missing flag – 4
→ 206 records
• ½ MDL missing flag – 1
→ 34 records
• New NOAA POC – 9
• Late added records (2019) – 1
Removed – 35
• Same records (½ MDL flagged) – 6
→ 193 records
• NOAA invalidated – 1
→ 6 records
• Invalidated data still exists – 2
→ 4,621 records
• Xylene overlap – 2
• AQS parameter does not exist – 13
• Nulled records – 5
• Invalidated – recently removed – 6
New Year – 7,668
• Previous year exists (19)
• Time series exists (5)
• Other (31)

Figure 1. Schematic of the QA process for the difference and suspect categories and counts for the 2022 Archive.

Notes: (1) There are 316,564 total annual means. The 225,406 ROS annual means are used in the difference categories.

(2) The removed annual averages are not included in the 225,406 ROS annual means in the overlap years.

Table 3. Annual average difference and suspect category descriptions and counts.

Difference Category	Difference Category – Name	Difference Category – Description	Count (Records)
Same	**Overall**	**Annual averages are the same between the 2021 Archive and the 2022 Archive in the overlap years**	**212,121**
Different	**Overall**	**Annual averages are different by pollutant/site/year/ sampling duration between the 2022 Archive and the 2021 Archive in the overlap years**	**4,091**
Added	**Overall**	**Annual averages added to the 2022 Archive in the overlap years**	**1,526**
Added	All data new	All records of an annual average originate from new data added in the 2022 Archive	1,269
Added	Partial year filled	Annual average existed in part in the 2021 Archive and completed from new data added in the 2022 Archive	242
Added	Invalidated missing flag	Records were invalidated, but the invalidation flag is missing	4 (206)
Added	Suspected ½ MDL missing flag	Record is suspected of being ½ MDL, but the flag is missing	1 (34)
Added	New NOAA POC	New POC added to a NOAA data site	9
Added	Late added records	Late added records filled in partial year	1
Removed	**Overall**	**Annual averages removed by pollutant/site/year/sampling duration in the 2022 Archive in overlap years**	**35**
Removed	Suspected ½ MDL flag	The reported values are the same, but the new values are suspected of being ½ MDL	6 (193)
Removed	NOAA invalidation	NOAA invalidated records	1 (6)
Removed	Invalidated data still exist	Invalidated data still exist	2 (4,621)
Removed	Xylene overlap	Xylene overlap data were removed	2

Difference Category	Difference Category – Name	Difference Category – Description	Count (Records)
Removed	AQS parameter code does not exist	AQS parameter code was relabeled with the correct parameter code	13
Removed	Nulled records	Records were nulled in AQS	5
Removed	Invalidated – recently removed	Data were invalidated in AQS and therefore removed from the 2022 Archive	6
New Year	**Overall**	**Annual averages added in the 2022 Archive from newest year**	**7,668**
New Year	Previous year exists	2021 exists and 2022 annual average >10x 2021 annual average	19
New Year	Time series exists	2021 does not exist but time series exists in overlap years for a given pollutant/site/year/sampling duration and 2022 annual average >10x time series max in overlap years	5
New Year	Other	2021 and time series in overlap years do not exist and 2022 annual average >95th %ile for pollutant/year	31

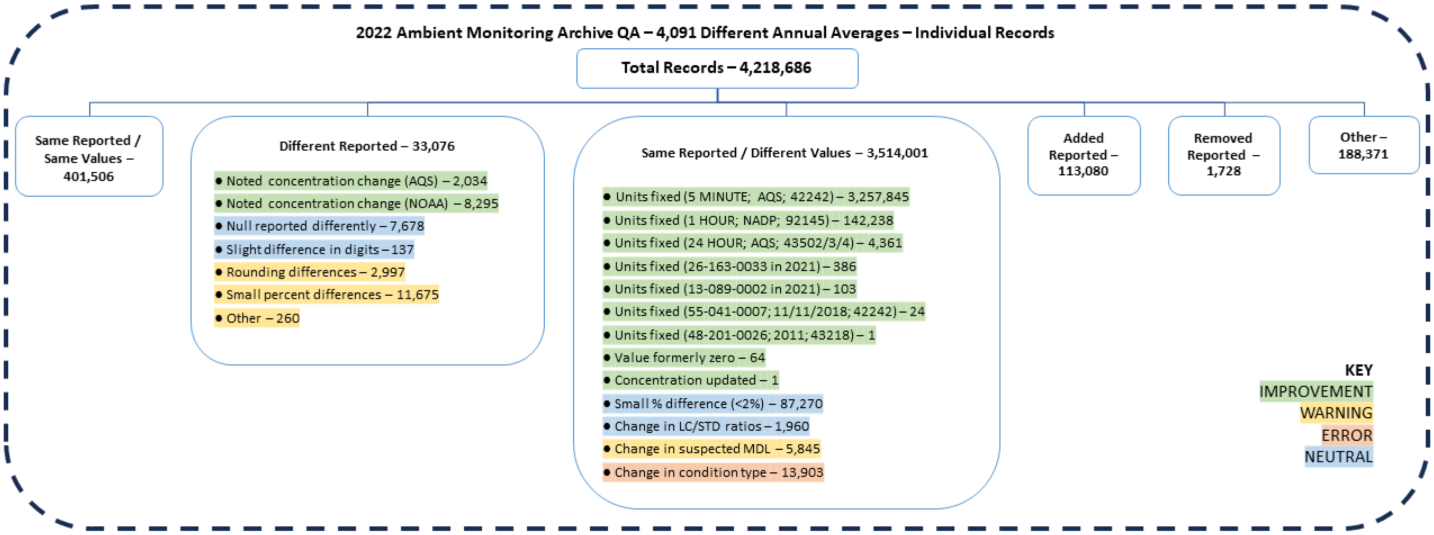
2022 Ambient Monitoring Archive QA – 4,091 Different Annual Averages – Individual Records
Total Records – 4,218,686
Same Reported / Same Values – 401,506
Different Reported – 33,076
● Noted concentration change (AQS) – 2,034
● Noted concentration change (NOAA) – 8,295
● Null reported differently – 7,678
● Slight difference in digits – 137
● Rounding differences – 2,997
● Small percent differences – 11,675
● Other – 260
Same Reported / Different Values – 3,514,001
● Units fixed (5 MINUTE; AQS; 42242) – 3,257,845
● Units fixed (1 HOUR; NADP; 92145) – 142,238
● Units fixed (24 HOUR; AQS; 43502/3/4) – 4,361
● Units fixed (26-163-0033 in 2021) – 386
● Units fixed (13-089-0002 in 2021) – 103
● Units fixed (55-041-0007; 11/11/2018; 42242) – 24
● Units fixed (48-201-0026; 2011; 43218) – 1
● Value formerly zero – 64
● Concentration updated – 1
● Small % difference (<2%) – 87,270
● Change in LC/STD ratios – 1,960
● Change in suspected MDL – 5,845
● Change in condition type – 13,903
Added Reported – 113,080
Removed Reported – 1,728
Other – 188,371
KEY
IMPROVEMENT
WARNING
ERROR
NEUTRAL

Figure 2. Schematic of the QA process for the individual records of the different annual averages in 2022 Archive.

Table 4. Category descriptions and records counts among different annual averages (4,218,686 records total).

Category	Description	Record Count
Same Reported / Same Value	**Overall**	**401,506**
Different Reported	**Overall**	**33,076**
Different Reported	Noted concentration change (AQS)	2,034
Different Reported	Noted concentration change (NOAA)	8,295
Different Reported	Null reported differently	7,678
Different Reported	Slight difference in digits	137
Different Reported	Rounding differences	2,997
Different Reported	Small percent differences	11,675
Different Reported	Other	260
Same Reported / Different Values	**Overall**	**3,514,001**
Same Reported / Different Values	Units fixed (5 MINUTE; AQS; 42242)	3,257,845

Category	Description	Record Count
Same Reported / Different Values	Units fixed (1 HOUR; NADP; 92145)	142,238
Same Reported / Different Values	Units fixed (24 HOUR; AQS; 43502/3/4)	4,361
Same Reported / Different Values	Units fixed (26-163-0033 in 2021)	386
Same Reported / Different Values	Units fixed (13-089-0002 in 2021)	103
Same Reported / Different Values	Units fixed (55-041-0007; 11/11/2018; 42242)	24
Same Reported / Different Values	Units fixed (48-201-0026; 2011; 43218)	1
Same Reported / Different Values	Value formerly zero	64
Same Reported / Different Values	Concentration updated	1
Same Reported / Different Values	Small % difference (<2%)	87,270
Same Reported / Different Values	Change in LC/STD ratios	1,960
Same Reported / Different Values	Change in suspected MDL	5,845
Same Reported / Different Values	Change in condition type	13,903
Added Reported	**Overall**	**113,080**
Removed Reported	**Overall**	**1,728**
Other	**Overall**	**188,371**

Frequent Questions

What is the purpose of the Archive? Why does the Archive include data not in AQS?

The Archive brings together multiple data sets of ambient monitoring HAPs data across the US in a unified format that allows for user-ready data analysis. The Archive pulls in high quality data not housed in AQS in addition to AQS data. Due to data reporting requirements of HAPs, some state air quality agencies collect data that are appropriate for AQS but are not uploaded to AQS. Other federal agencies and established national programs outside of a given EPA regulatory network also collect relevant data. See Technical Report for additional information.

Why are local conditions derived?

Sample values are either collected in local or standard conditions. Data collected in standard conditions can be converted to local conditions from collocated temperature and pressure data or they can be derived from calculated LC/STD ratios. Converting to local conditions and deriving to local conditions allows the data to be utilized for air toxics applications where local conditions are required. The LC/STD ratios derive LC where the final standardized value is STD due to a lack of temperature and pressure data.

Why is there different criterion to calculate daily averages for remote and non-remote data?

Non-remote sub-daily data, depending on the sampling duration, require a minimum number of samples to calculate a daily average. This ensures that a daily average is representative. However, remote data do not have this requirement. For example, a daily average for remote data may be represented by one 5-minute sample. This criterion is loosened because the pollutants and monitoring locations of remote data typically represent background concentrations that do not meaningfully fluctuate at sub-daily time scales.

Why are annual statistics separated by sampling duration?

Annual statistics are separated by pollutant/site/year/sampling duration to account for broad method differences among different sampling durations.

How is the minimum number of samples of the different sampling durations determined when calculating a daily average?

A minimum number of samples is needed to calculate a daily average when the sampling duration is sub-daily. The minimum number of samples varies depending on the scale of the sampling duration (i.e., hourly or sub-hourly). These are found under the “Durations” tab in AMA2022_lookups.xlsx. If the duration description is under 60 minutes, the minimum count is defined as $\left\lceil 0.75 * \frac{60}{DUR\_DESC} \right\rceil$, where DUR_DESC is the sub-hourly duration description in minutes. If the duration description is 60 minutes or above, the minimum count is defined as $\left\lceil 0.75 * \frac{24}{DUR\_DESC} \right\rceil$, where the DUR_DESC is the hourly duration description in hours. In other words, the minimum count is rounding up the minimum number of samples needed to have 75% completeness for an hour (or day).

Table 5. Minimum sampling count by sampling duration needed to construct an hourly or daily average.

Duration Description	Average Up To	Minimum Count
5 MINUTES	HOURLY	9
10 MINUTES	HOURLY	5
15 MINUTES	HOURLY	3
30 MINUTES	HOURLY	2
150 MINUTES	DAILY	8
90 MINUTES	DAILY	12
1 HOUR	DAILY	18
2 HOUR	DAILY	9
3 HOURS	DAILY	6
4 HOUR	DAILY	5
5 HOUR	DAILY	4
6 HOUR	DAILY	3
8 HOUR	DAILY	3
12 HOUR	DAILY	2
24 HOURS	DAILY	1

Where is the information in the lookup table tabs obtained?

- AQS_CAS: https://aqs.epa.gov/aqsweb/documents/codetables/parameters.csv (accessed November 2024)
- NATTS: https://www.epa.gov/system/files/documents/2023-03/NATTS-site-list-2023.pdf (accessed November 2024)
- NEI: PollutantCode.xlsx from https://www.epa.gov/system/files/other-files/2022-05/EISCodeList_1.zip (accessed November 2024)
- AirToxScreen: 2020AirToxScreenPollutants.xlsx from https://www.epa.gov/system/files/other-files/2024-06/2020-airtoxscreen-supplemental-data-files.zip (accessed November 2024)

What is the minimum number of daily averages required to create a valid quarter?

A minimum of *six* daily averages is required per pollutant/site/sampling duration to create a valid quarter.

How many valid quarters of data are needed to calculate an annual average?

A minimum of *three* valid quarters in a year is needed per pollutant/site/sampling duration to create an annual average. This is to ensure that the annual average is representative of the year. All daily averages are used in the calculation of an annual average, assuming the threshold of valid quarters is met. Therefore, an annual average may include data from three “valid” quarters and one “invalid” quarter.

What data in the 2022 Archive are excluded from annual averages?

Only ambient HAPs monitoring with a sampling duration of less than or equal to 24 HOURS are considered for annual averages. Therefore, records with a 2-week sampling duration are excluded from the annual average calculation. This sampling duration may be included in future versions of the Archive after conducting a more thorough analysis. Fenceline monitoring is currently excluded from annual averages.