Annual Statistics and Quality Assurance of the 2021 Ambient Monitoring Archive

October 2023

Table of Contents

Purpose

This document outlines the major steps implemented in calculating annual statistics for each pollutant, site, sampling duration, and year for the 2021 version of the Ambient Monitoring Archive (AMA) for the Hazardous Air Pollutants (HAPs) (i.e., “the 2021 Archive”).

Steps for Calculating Annual Statistics

Calculate derived local conditions

Air quality samples are collected using either local conditions or standard conditions, each of which have different applications (e.g., risk, model evaluation, defining detection limits, etc.). In instances where local conditions are not available, but standard conditions are, the local conditions 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.

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 standard conditions and the average local conditions 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 local conditions/standard conditions ratios to derive local conditions

1. Identify all the samples collected using standard conditions but not local conditions.
2. If the ratio exists for that pollutant/site/POC/day/sampling duration, derive local conditions by multiplying the sample collected using standard conditions and the ratio.
3. If the above ratio does not exist but the ratio exists for that site/day, derive local conditions by multiplying the sample collected using standard conditions and the ratio.
4. If the above ratio does not exist but the ratio exists for that site/quarter, derive local conditions by multiplying the sample collected using standard conditions and the ratio.
5. If the above ratio does not exist but the ratio exists for that site/year, derive local conditions by multiplying the sample collected using standard conditions and the ratio.
6. If the above ratio does not exist but the ratio exists for that site, derive local conditions by multiplying the sample collected using standard conditions and the ratio.
7. If the ratio by site is missing for the sample collected using standard conditions, assume $ratio = 1$ (i.e., local conditions equal standard conditions).

Samples in either local conditions or derived local conditions 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 samples 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 *seven* 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 so 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). Due to the stability of the ROS, these statistics are only calculated if the 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 standard conditions while the original samples may be collected using local conditions. 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 and derived local	zero	arithmetic mean, variance, maximum, 10th percentile, 25th percentile, median, 75th percentile, and 90th percentile
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 local conditions (LC) to standard conditions (STD) 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.
- 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 derived local conditions and standard conditions and treatment of non-detect data.
- AMA_file.R: Creates the final annual excel file.

Quality Assurance of the 2021 Archive

Quality Assurance (QA) of the 2021 Archive was primarily performed by comparing the ROS annual averages from the 2021 Archive to the 2020 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 Categories and QA Suspect Categories sections below, respectively). Currently, there is no corrective action for annual averages identified by the suspect categories.

The *difference categories* summarize the reasons why annual averages are different between the 2021 Archive and the 2020 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 – 2020, with the newest year, 2021, 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 2020 Archive and the 2021 Archive (Table 2).

The *suspect categories* identify the set of annual averages that vary largely from the different annual averages (i.e., in the overlap years between 1990 – 2020) and the newest annual averages (i.e., 2021). In the newest annual averages, the suspect categories first establishes if (1) there is a corresponding annual average in 2020 for that pollutant/site/sampling duration, (2) there is no corresponding annual average in 2020 but there is a time series for that pollutant/site/sampling duration, or (3) there is no annual average for 2020 and no time series for that pollutant/site/sampling duration. The two sets of checks (i.e., for the different annual averages and the newest annual averages) are further described below.

Table 2. Counts of annual averages in each QA category in the 2021 Archive.

QA Category Name	Count
All annual averages	303,800
annual averages (ROS*)	216,247
Overlap years^ (ROS)	208,345
Overlap years – Same (ROS)	202,549
Overlap years – Different (ROS)	2,288
Overlap years – Added (ROS)	3,508
Overlap years – Removed% (ROS)	623
Newest year& (ROS)	7,902

*ROS = Regression on Order Statistics

^Overlap years = 1990 – 2020

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

&Newest year = 2021

QA Difference Categories

The QA difference categories categorize the reasons in which an annual average in the 2021 Archive is different than an annual average in the 2020 Archive for a given pollutant/site/year/sampling duration. Difference categories explain how individual records may change between different versions of the Archive. Because of the comparison to the previous Archive, difference categories only apply to annual averages in the overlap years (i.e., years 1990 – 2020). 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 seven difference categories in the different annual averages, seven difference categories in the added annual averages, and four 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 are new. In this example, the annual average would fall into two difference categories: Rounded Values and Reported Added. The difference category counts for the different annual averages presented in Figure 1 and Table 3 count the number of annual averages in which there is at least one instance of a difference category occurring. Thus, the summation of the annual averages across the difference categories for the different annual averages exceeds 2,288. Different annual averages are primarily driven by differences in LC/STD ratios, values being rounded, and MDLs being rounded.

In practice there is typically one difference category that is driving an annual average to be added or removed (i.e., the dominant difference category). These are presented in Figure 1 and Table 3. There are 3,508 and 623 added and removed annual averages, respectively. Added annual averages are primarily driven by the inclusion a new data source or records being added from an existing data source. Removed annual averages are primarily driven by reported values being removed or reported values assigned to zero.

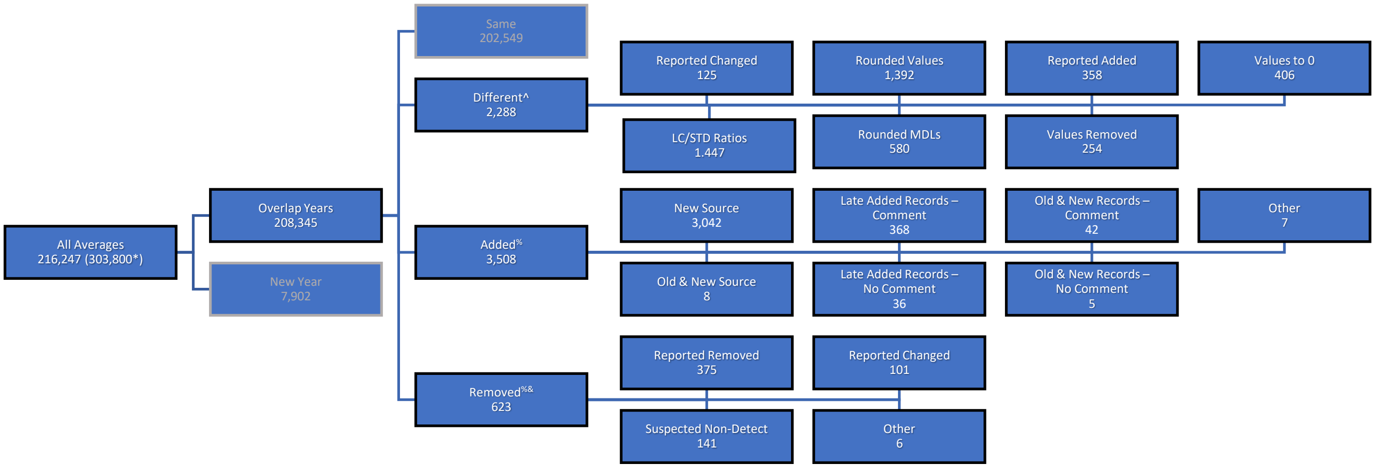

Figure 1. Schematic of the QA process for the difference categories and counts for the 2021 Archive.

*There are 303,800 total annual means. The 216,247 ROS annual means are used in the difference categories.

^One annual average may contain multiple difference categories among different annual averages. An annual average is counted if it contains at least one instance of a difference category.

%The added and removed difference categories are quantified by the dominant difference category in each annual average.

&The removed annual averages are not included in the 216,247 ROS annual means in the overlap years.

Table 3. Annual average difference category descriptions and counts.

Major Difference Category	**Difference Category – Name**	**Difference Category – Description**	**Count**
Different – Overall		**annual averages that are different by pollutant/site/year/ sampling duration in the overlap years between the 2021 Archive and the 2020 Archive**	**2,288**
Different	LC/STD Ratios	where (1) non-null ratios are different exceed a small threshold	1,447
Different	Rounded MDLs	where (1) old and new values are non-null and at least one value is zero and (2) old and new MDLs are different	580
Different	Rounded Values	where (1) old and new reported are equal and (2) old and new value are different and non-zero	1,392
Different	Values to 0	where (1) old value is not zero and non-null, (2) new value is zero, and (3) new reported is positive	406
Different	Reported Added	where (1) old reported is null and (2) new reported is non-null	358
Different	Values Removed	where (1) old value is non-null and (2) new value is null	254
Different	Reported Changed	where (1) old and new reported are different and non-null and (2) old and new values are non-null	125
Added – Overall		**annual averages added to the 2021 Archive in the overlap years**	**3,508**
Added	New Source	All records of an annual average originate from a data source that is new to the 2021 Archive	3,042
Added	Old & New Source	Records of an annual average originate from both a previously existing data source and a new data source to the 2021 Archive	8
Added	Late Added Records – Comment	All records of an annual average originate from data that have been added since the previous Archive, as identified in the comment field	368

Major Difference Category	Difference Category – Name	Difference Category – Description	Count
Added	Late Added Records – No Comment	All records of an annual average originate from data that have been added since the previous Archive without a comment	36
Added	Old & New Records – Comment	Records of an annual average originate from both previously existing records and records added since the previous Archive, as identified in the comment field	42
Added	Old & New Records – No Comment	Records of an annual average originate from both previously existing records and records added since the previous Archive without a comment	5
Added	Other		7
Removed – Overall		**annual averages removed in the 2021 Archive in the overlap years**	**623**
Removed	Suspected Non-Detect	A large portion of records of an annual average are suspected of being a surrogate for non-detects and are therefore changed to zero	141
Removed	Reported Changed	A large portion of records of an annual average have changed since the previous Archive, as identified in the comment field	101
Removed	Reported Removed	A large portion of records of an annual average have been removed since the previous Archive	375
Removed	Other		6

QA Suspect Categories

The QA suspect categories count the annual averages in the 2021 Archive that have either (1) a large difference compared to the 2020 Archive (in the overlap years) or (2) a large different within the 2021 Archive (in the newest years). In contrast to the difference categories, the suspect categories create threshold of high values and looks at the newest annual averages (i.e., 2021). In the suspect categories, the comparison sets vary. For example, the comparison set to the different annual averages is the 2020 Archive and the comparison set to the newest annual averages is the 2021 Archive, as described below. The count of annual averages falling into each suspect category is presented in Figure 2 and Table 4.

For the newest annual averages, the suspect threshold depends on whether the annual average (1) can be compared to 2020, or (2) is contained within a time series, or (3) neither. Most suspect new annual averages (i.e., 47) are not contained within a time series and do not have a 2020 annual average.

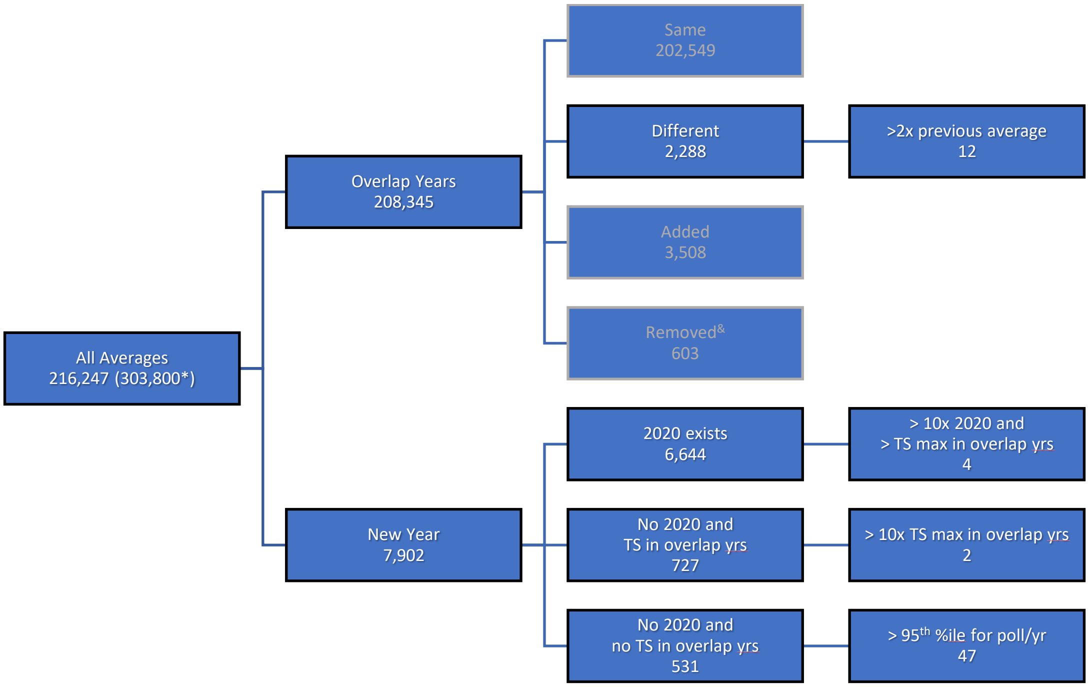

Figure 2. Schematic of the QA process for the suspect categories and corresponding annual average counts for the 2021 Archive.

*There are 303,800 total annual mean. The 216,247 ROS annual means are used in the suspect/difference categories.

[&]The removed annual averages are not contained within the 216,247 ROS annual means in the overlap years.

Table 4. Counts of annual averages by suspect category.

Major Suspect Category	Suspect Condition Description	Count	Suspect Threshold	Count
Different – Overall		**2,288**	**>2x previous annual average**	**12**
Newest Year – Overall				**7,902**
Newest Year	2020 exists	6,644	>10x 2020 and >TS* max in overlap years	4
Newest Year	no 2020 and TS in overlap years	727	>10x TS max in overlap years	2
Newest Year	no 2020 and no TS in overlap years	531	>95th %ile for pollutant/year	47

*TS = Time Series

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 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 an EPA regulatory network also collect relevant data. See technical report for additional information.

Why are local conditions derived?

Local conditions are derived from standard conditions using an LC/STD ratio to utilize data for toxics applications where local conditions are required. These 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 AMA2021_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 May 2023)
- NATTS: https://www.epa.gov/system/files/documents/2023-03/NATTS-site-list-2023.pdf (accessed May 2023)
- NEI: PollutantCode.xlsx from https://www.epa.gov/system/files/other-files/2022-05/EISCodeList_1.zip (accessed May 2023)

- AirToxScreen: 2018AirToxScreenPollutants.xlsx from https://www.epa.gov/system/files/other-files/2022-10/2018%20AirToxScreen%20Supplemental%20Data%20files.zip (accessed May 2023)

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

A minimum of *seven* 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 daily averages outside of the three “valid” quarters.

What data in the 2021 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.