|  |  |
| --- | --- |
|  |  |
| An annual review of provincial long-term air quality data Review procedure for data submitted into Air Data Warehouse |

#

# 1.0 Background

Air quality and deposition monitoring in Alberta is delivered through a distributed system, which includes industry, Airshed organizations and the provincial and federal governments. Data from the continuous ambient air quality monitoring stations operated by Alberta Airsheds and Alberta Environment and Parks (AEP) are available for download from the Alberta Air Data Warehouse. The Air Data Warehouse is a central repository for air quality and deposition data, and associated metadata, collected in Alberta. These data are used by the government, stakeholders, and the public to protect human health and the environment. In order to ensure high quality data, the data provided into the Air Data Warehouse are validated and verified, as required by the Air Monitoring Directive, prior to submission. In addition, the Air Data Warehouse performs automatic data validations upon data submission. In some cases, poor quality data can pass these verification and validation steps, and are only identified later on through data analysis or station audits. An annual review of provincial long-term air quality data has been developed to identify the few poor quality data that have been missed by the existing validation steps. This annual review is being done to 1) ensure the best possible data are being used and 2) ensure data provided through the Air Data Warehouse are of known quality. These assessments are done annually to minimize the distribution of erroneous data and to catch issues in a timely manner when resolution is possible.

The data review procedures developed for this annual review were piloted using 2020 data in Spring 2021. Data validation procedures used by other jurisdictions were consulted in developing the data review, including BC’s Ministry of Environment and Climate Change Strategy, the US Environmental Protection Agency, and the National Air Pollution Surveillance program. Development of the annual review process also considered automatic data validations implemented by the Air Data Warehouse, including validations that only result in a warning to the data submitter (as opposed to a failed submission), as well as validations anticipated for development in future phases of the Air Data Warehouse. Through internal discussions, the most suitable checks were chosen for the development of the annual data review process. Revisions to these procedures will occur in Fall 2021, based on the learnings of this pilot year. A revised version of the data review procedure will be used to perform the annual review of 2021 data in Spring 2022. It is expected that over time a reduced number of data quality issues will be found as growing pains associated with the new Air Data Warehouse are worked out, the root cause of data quality issues are resolved, and/or data providers incorporate similar data checks into their validation procedures.

# 2.0 Checks Performed in Data Review

Data for all Core Long Term monitoring stations for the considered parameters were downloaded for 2017-2020 (2017-2019 data included to facilitate checks that required historical data) through internal access to the Air Data Warehouse using R from February 24 – March 2, 2021. Only data identified as valid by the data provider are included in the review. Parameters considered include:

* CO
* NO, NO2, and NOx
* O3
* PM2.5
* SO2
* H2S and TRS
* NMHC, CH4, and THC
* Outdoor Air Temperature
* Relative Humidity
* Wind Speed (at 10 m, or lowest height) and Wind Direction (at 10 m, or lowest height)

The purpose of the annual review is to identify potential data quality issues that need to be investigated/followed up on with the data provider. There are various data checks included in the review process (Sections 2.1-2.6). Each check results in an output file. The output files contain instances of which data pass/fail each check. R scripts and various R packages are used to perform most of the checks; the use of scripts has made the handling of large datasets and iterative calculations easier, but these data checks could be done using other methods.

Data that failed the checks are tagged for additional investigation. Investigation activities may include closer visual examination of the time series, comparison between parameters and nearby stations, and reviewing monthly monitoring and calibration reports. Where warranted, data providers may be contacted to further discuss the validity of data and to apply corrections as needed.

## 2.1 Range in Values Checks

The range in values checks look for instances where the values do not fall within the expected range for a given parameter. There are three checks in this category: (1) Wind Direction Range; (2) Outliers; (3) Rate of Change/Spike.

### 2.1.1 Wind Direction Range

* To check that the wind direction from a station spans the full 0-360º, the number of hours with wind direction values in each quadrant (Figure 1) during the month were counted.
* Any station with a month of no data in any one of the quadrants was tagged for supplementary examination.
* Supplementary examination of wind direction time series is conducted to evaluate the extent of the reporting issue.



Figure : Diagram showing the four wind quadrants, where Quadrant 1 = 0 < wind direction (WD) >= 90°, Quadrant 2 = 90 < WD >=180°, Quadrant 3 = 180< WD >=270°, and Quadrant 4 = 270 < WD >=360°.

### 2.1.2 Outliers

* The outlier check is applied to all parameters except wind direction and temperature.
* For each month of the year, the mean and standard deviation of the hourly values from the previous three years of data at the station were calculated. These are hereafter referred to as the “reference mean” and “reference sd”, respectively.
	+ As the previous three years of data at the station are required for this check, if a station does not have these data available, the check is skipped.
* For each month, hours which exceed the threshold are tagged. The threshold is defined as follows
$$reference mean \pm \left(10×reference sd\right)= threshold$$
* Notes from the pilot year:
	+ In some cases, there may be a large number of hours which exceed the threshold marginally (<5 ppb) and likely do not require review.
		- For example, there were many outlier hours identified for SO2. In cases such as this, it was helpful to filter the hours that exceeded the threshold by at least 10 ppb to focus review of data on the most notable outliers.
	+ The outlier check requires data for the previous three years. Stations that did not have historical data were skipped. Instead, the Rate of Change/Spike check (Section 2.1.3) and/or Visual Time Series Review (Section 2.5) were used to identify possible data anomalies.
		- In future years, alternatives to this approach may be used, including using less than three years of data for the reference mean and reference sd.

### 2.1.3 Rate of Change/Spike

* Unlike the outlier check (Section 2.1.2) which is applied to all parameters except wind direction and temperature, the rate of change/spike check is applied to only a select set of parameters. In addition, this check does not compare the data being reviewed to data reported for previous years.
* This check is applied to all parameters for which a maximum rate of change has been defined (Table 1).
* For each parameter at each station, the rate of change between each consecutive hour was calculated.
* The sample period is tagged for further examination when the absolute value of the rate of change exceeds the limit defined in Table 1.
* Notes from the pilot year:
	+ There were many instances in which the Rate of Change/Spike check and the Outlier check flagged the same hours. In these cases, one of the two checks was used to conduct any further investigation as needed, with the outlier check taking precedence.

Table 1: Rate of change maximums for select parameters

All values are in ppm/hr except for PM2.5 (μg/m3hr).

|  |  |
| --- | --- |
| **Parameter** | **Max Rate of Change1** |
| CO | 10 |
| PM2.5 | 50 |
| NO2 | 0.030 |
| O3 | 0.050 |
| SO2 | 0.040 |

1 These values were adopted from B.C. Ministry of Environment and Climate Change Strategy’s data validation process for air monitoring data (ftp://ftp.env.gov.bc.ca/pub/outgoing/AIR/Validation/) and may change as additional information becomes available.

## 2.2 Sticking Values Checks

The sticking values checks look for instances where the same value is repeating multiple hours in a row (in the case of non-zero values) or occurs more often in a month than expected (in the case of zeros). There are two checks in this category: (1) Sticking Parameter; (2) Percent Zero.

### 2.2.1 Sticking Parameter

* This check is applied to all parameters and examines the reported concentrations for extended periods with the same non-zero value.
* Sample periods with the same non-zero value for 12 or more consecutive hours are tagged for review.
* Notes from the pilot year:
	+ This check identified numerous sample periods for review. In most cases, the identified periods had the same non-zero value due to rounding (for example NO2 concentrations reported at 1 ppb resolution).
	+ The next time the annual data review is applied, this check will likely be modified.

### 2.2.2 Percent Zero

* This check examines the reported concentrations for extended periods of below detection (value of zero) and is applied to all parameters except H2S, TRS, SO2, and NMHC, because these parameters are known to be below the analyzer detection for extended period of time. The check is also not applied to temperature, relative humidity, and wind direction.
* This check tags for review the sample months where the percentage of zero values is greater than 50%.

## 2.3 Related Parameter Checks

The related parameter checks are applied for both NOx and THC and check whether the sum of the related parameters (NO and NO2 for NOx; NMHC and CH4 for THC) is equal to the submitted NOx and THC values.

### 2.3.1 NOx

* For each hour, a NOx value calculated from the reported NO and NO2 values are compared to the reported NOx value. Differences that are greater than 1 ppb are tagged for review.

$$NO+NO\_{2}=calculatedNO\_{x}$$

$$NO\_{x}-calculatedNO\_{x}>\pm 1 ppb$$

* Notes from the pilot year:
	+ This check also identified hours for which one or two of the related parameters have no reported concentrations (due to calibration) but the other related parameter(s) do have a value. Such incidents were tagged for further investigation as all related parameters are expected to be similarly affected by calibration or instrument maintenance.

### 2.3.2 THC

* For each hour, a THC value calculated from the reported NMHC and CH4 values are compared to the reported THC value. Differences that are greater than 0.1 ppm are tagged for review.

$$NMHC+CH\_{4}=calculatedTHC$$

$$THC-calculatedTHC>\pm 0.1 ppm$$

* Notes from the pilot year of the annual review:
	+ As with the NOx check (Section 2.3.1), this check also identified hours for which one or two of the related parameters have no reported concentration, but the other parameter(s) do have a value, and were tagged for further investigation.

## 2.4 Shifts in Data Checks

The shifts in data checks are used to identify if there are any stepwise changes in reported concentrations in either the baseline concentrations or the monthly mean concentrations.

### 2.4.1 Baseline Shift

* This check is applied to all parameter except wind speed, wind direction, temperature, and relative humidity.
* The previous three years of data are used to determine the historical 10th percentile. The subset of historical and target year concentrations below the historical 10th percentile are used to calculate monthly base mean concentration for both the historical period and target year. A sample month is tagged for further investigation if the base historical mean is greater than twice the base target mean or less than half of the base target mean.
	+ As the previous three years of data at the station are required for this check, if a station does not have this data available, the check is skipped.
* Supplementary visual time series review (Section 2.5) was also used to identify shifts that occur between subsequent months or mid-month that aren’t tagged through comparison with historical data.

### 2.4.2 Shift in Mean

* This check is applied to all parameters except wind direction.
* This check compares concentrations for the target year with concentrations for the previous three years.
* The monthly mean values were calculated for the target year ($μ\_{Jan}$, $μ\_{Feb}$, etc.) and for the previous three years ($μ\_{HJan}$, $μ\_{HFeb}$, etc.). Sample months were tagged for further investigation when the percent difference between the target year monthly mean value and the monthly mean of the previous three years was greater than a set threshold value.

$$Pdiff=\frac{\left|μ\_{Jan}-μ\_{HJan}\right|}{μ\_{HJan}}×100\%$$

$$Pdiff>Pdiff\_{thresh}$$

* In the pilot year, $Pdiff\_{thresh}=100\%$.
* Supplementary visual time series review (Section 2.5) was also used to identify shifts that occur between subsequent months or mid-month that aren’t tagged through comparison with historical data.

## 2.5 Visual Time Series Check

* A visual check of time series of each parameter for each station is performed to determine if the reported concentrations are generally as expected or if there are obvious data anomalies. Many of the potential anomalies identified through this check were also pickup by the more detailed checks presented in Sections 2.1-2.4.
* The few cases that were not identified by the detailed checks were tagged for further discussion with the data provider.
	+ For example, a station with a wind direction range that did not span the full 0-360° but instead was limited to 25-325° was not tagged as a potential issue in the wind direction range check (Section 2.1.1) as there was data in each quadrant. However, a visual time series check was able to tag this data for further discussion with the data provider.

## 2.6 Expected Parameters Check

* For each station, the parameters reported during the year are compared to the parameters expected to be reported by that station. This allows the identification of measured yet not reported parameters.