Community Air Monitoring Plan
Appendix F

California Statewide Mobile Monitoring Initiative (SMMI)
Aclima's Data Management Plan (v3.0)

aclima

July 1, 2025

CALIFORNIA AIR RESOURCES BOARD

CALIFORNIA CLIMATE INVESTMENTS
Cap and Trade Dollars at Work

The Statewide Mobile Monitoring Initiative is part of California Climate Investments, a statewide initiative that puts billions of Cap-and-Trade dollars to work reducing greenhouse gas emissions, strengthening the economy, and improving public health and the environment — particularly in disadvantaged communities.

Contents

Glossary of abbreviations

ADR - AMN Data Review ticket
AMN - Aclima Mobile Node
AMP - Aclima Mobile Platform
API - Application Programming Interface
CAMP - Community Air Monitoring Plan
CERP - Community Emissions Reduction Plan
CNC - Consistently Nominated Community
DEM - Digital Elevation Model
DEP - Deployment ticket
DQO - Data Quality Operations
EtO - Ethylene Oxide
GCP - Google Cloud Platform
GCS - Google Cloud Storage
IAM - Identity and Access Management
IoT - Internet of Things
L1 - Level 1 mobile air quality monitoring data
L2 - Level 2 mobile air quality monitoring data
L3 - Level 3 mobile air quality monitoring data
L4 - Level 4 mobile air quality monitoring data
PEG - Project Expert Group
PML - Partner Mobile Laboratory
SAPP - Sampling, Analysis, and Presentation Plan
SMMI - State Mobile Monitoring Initiative
SOW - Scope of Work
TVOC - Total Volatile Organic Compounds
VER - Deployment Verification ticket

About Aclima

Aclima is a Public Benefit Corporation and climate tech company catalyzing bold action to reduce emissions, protect public health, and deliver clean air for all. Powered by its network of roving sensors, Aclima measures air pollution and greenhouse gases across cities, regions, and states with block-level resolution. The company's online tools translate billions of scientific measurements into interactive maps and analytics that are designed to inform and empower clean air decision-making, with a focus on intuitive tools that support communities — particularly those hardest hit by air pollution. Aclima's customers include regulators, utilities, companies, and communities working to reduce emissions and protect public health and the environment

About the Statewide Mobile Monitoring Initiative

The Statewide Mobile Monitoring Initiative (SMMI) is designed to attain a comprehensive dataset of criteria pollutants, toxic air contaminants, and greenhouse gases, create a data portal for the public to access and visualize SMMI data, and conduct inclusive community engagement to better understand and address community concerns. This project provides an opportunity to complement AB 617 statewide air monitoring activities by engaging communities beyond those currently selected under the Community Air Protection Program, providing data to fill air monitoring gaps and support additional actions to reduce emissions and exposure.

1. Requirements and Deliverables

1.1. Data management plan outcomes

The purpose of this documentation is to provide a high-level overview of Aclima's proprietary data management system and capabilities, and explain how this is applied to provide high-quality finalized raw data and derived data products.

Aclima's data management system will collect and push raw data through multiple levels of processing, storage, and tagging throughout the SMMI collection campaign, including data from Partner Mobile Laboratories (PMLs). The system will then transfer finalized raw data to CARB via cloud-based data bucket and provide data directly to support visualizations and reporting via StoryMaps.

1.2. Relevant SOW elements

This documentation is in direct or partial fulfillment of the following SMMI Scope of Work (SOW) elements and Aclima sub-tasks:

- CARB SOW element 6.1: Establish Data Management Procedure
    - Aclima sub-task 6.1.1: Draft data management plan developed and documented
    - Aclima sub-task 6.1.2: Data Management Plan iteration via the CAMP process, with focus on analysis of data and communication of results to support action
    - Aclima sub-task 6.1.3: Final Sign off on Data Management Plan
- CARB SOW element 6.2: Establish Data Transfer Mechanisms
    - Aclima sub-task 6.2.1: Data transfer to CARB: interface requirements gathering
    - Aclima sub-task 6.2.2: Data transfer to CARB: interface requirements document finalized
    - Aclima sub-task 6.2.3: Data transfer to CARB: data schema requirements gathering

2. Key concepts

2.1. Aclima data levels

The categories of data Aclima will make available to CARB and communities will span 1 Hz measurements for analysis and development, aggregations of data taken throughout the observation period, and low latency alerts for detection of high concentration signals.

Aclima further organizes these data into levels reflecting the degree of processing applied, from the lowest level (Level 0, or L0) at sensor readout to high level (Level 4, or L4) modeled analyses which synthesize individual data points into actionable insights and data summaries for dissemination through visualization and reporting. See Table 2.1 below for description of all data levels. This classification was informed by published research, and has been adapted for the purposes of describing Aclima's mobile data processing levels. L0-L2 data levels are also continuous time series data. L3 and L4 are distinct from L0-L2 data levels in that they are more aggregated data sets (either in space or in time) or are abstracted from the original times series data, for example a geospatial cluster of enhancement detections. The purpose of defining these data levels is to communicate the general data processing steps associated with delivered data and data products for CARB and other end users of SMMI data and visualizations and to help inform CARB and end users of the type and nature of data that is provided.

Table 2.1: Aclima's Data Processing Levels. Asterisks (*) indicate data levels provided to CARB or in support of non-scientific communication and community visualization.

Data Level	Name	Definition	Example
0	Raw Signal	Original signal produced by the sensor.	Voltage, digital number, raw mass spectral data.
1*	Intermediate geophysical quantities	Derived from Level 0 data using basic physical principles or calibration equations.	Concentration in ppb or $\mu g/m^3$.
2a*	Standard geophysical quantities	Estimate using sensor plus associated on-board physical measurements. May also include simple signal processing steps such as smoothing or frequency filtering to	NO_2 derived from O_3 and O_x (O_3+NO_2). Temperature and humidity corrections to sensor estimates using empirically derived relationships rather than basic physical principles.

		correct for sensor artifacts.	Measurements corrected for sensor drift using a high band pass filter method of signal processing.
2b	Standard geophysical quantities, extended	Similar to Level 2a but using external data sources for artifact correction Not anticipated to be used for SMMI.	Quantities normalized using collocated meteorological or pollutant data.
3*	Aggregated geospatial quantities	Geospatial data products derived typically from L1, L2a, or L2b data from single pollutant measurements or a combination of multiple pollutants and using standard statistical aggregation approaches of measurements (i.e. mean, median, max, etc)	Map of locations of individual enhancement events of TVOCs, speciated toxics or other pollutants. Clusters of methane enhancements, in combination with ethane, indicating locations of natural gas leaks. Continuous maps of ambient concentrations using basic statistical aggregation of measurements only, with no advanced modeling methods applied.
4*	Aggregated geospatial quantities combined with modeled spatio-temporal phenomenology	Aggregated geospatial data products derived typically from L1, L2a, or L2b data using advanced statistical modeling methods across space and time. External data sources or contextual information such as meteorological measurements, topographic data, road	Map of pollutant averages derived from measurement data in combination with a statistical reconstruction method or land use regression method Continuous map of probabilistic estimates of the likelihood of a methane leak in an area Map showing statistically

		type, etc may be incorporated.	significant hot spots of average pollutant concentrations via the Getis-Ord (Gi*) method

2.2. Data management pipeline

Key steps in Aclima data management are summarized in Table 2.2 below, and visually depicted in Figure 2.1 below along with the technical modules enabling the pipeline. These are discussed in the following sections.

Table 2.2: An overview of the major components of the data management from device through transfer of finalized raw data to visualizations fed by APIs.

Data management pipeline		
1	**Publish**	Publish 1 Hz sensor data along with metadata, including device serial numbers, and auxiliary measurements such as sensor-specific temperature, relative humidity, pressure, latitude and longitude (GPS), and flow rate from remote devices to the cloud.
2	**Ingest**	Ingestion of data into the cloud. Level 0 data in the cloud archive are never removed or altered.
		Custom code running on cloud compute services to pull third party data for contextual data layers as needed.
3	**Transform**	Process data into geolocated measurements with physical meaning along the pathway from L0→L1→L2.
		Flag invalid data from instrumentation faults, warmup periods, out of range readings, etc through QA/QC activities
4	**Model**	Data transformations to create higher level data products from L1/L2 data, including spatial and temporal aggregations and multi-pollutant data analyses.
5	**Store**	Label and store all levels of processed data in a data lake. Make data accessible for further processing, analysis, and direct delivery to customers.
		Preserve snapshots of data at critical processing stages such as at finalization of calibration.

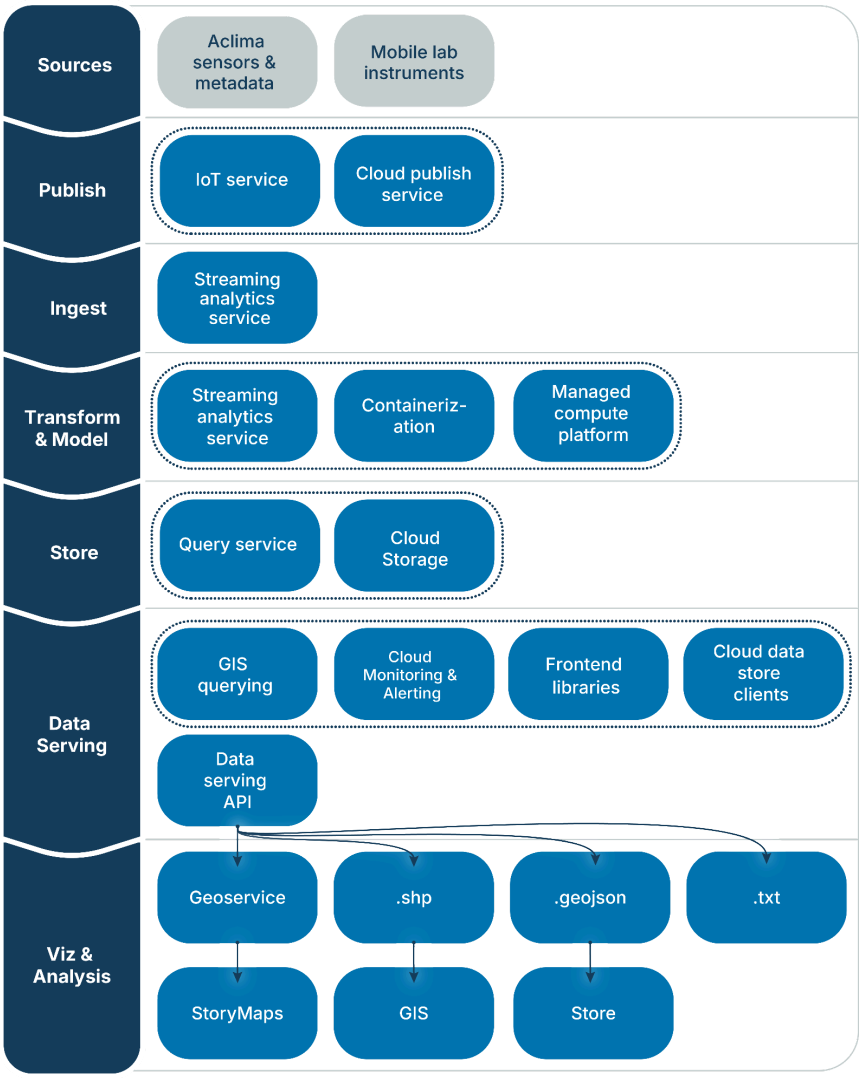

Figure 2.1: The primary flow of information from the publication of 1 Hz sensor and instrumentation data at the point of measurement through processing, storage, alerting, and serving of data and analysis products to summarizing findings to the community.

3. Data ingestion

Data ingestion describes the process by which the unaltered lowest level data from the Aclima Mobile Node (AMN) sensors is delivered to the Aclima data storage and processing backend.

3.1. Definition and application of Data Level 0

Data Level 0 (L0) describes the unaltered raw signal produced by a sensor. Depending on the sensor, this may be a voltage, a digital number, raw mass spectral data, or another format. An example of L0 data is raw counts from an electrochemical sensor. L0 data is generally not useful for the end user and will not be included as part of the delivered data set for SMMI, but it will be retained indefinitely in its original state.

Note that in some cases when the sensor has its own internal on-board data processing, the L0 data may not be accessible, and the lowest level data available for certain measurement streams may be L1 or L2. For example, for an optical particle counter the raw signal is considered to be counts per unit time (e.g. counts/s), but the on board data processing for this sensor transforms the data into counts per unit volume (e.g. counts/L), which is considered L1 in the framework used here. In all cases, the lowest level data available is what gets published by the AMN, ingested in the cloud, and retained indefinitely in its original state. The following describes these steps in more detail.

3.2. Data ingestion workflows

3.2.1. Publish

The AMN polls the physical sensors and lab-grade reference equipment for data every second. Raw mobile sensor measurements are recorded along with the associated date, time, and location of each measurement as well as other critical diagnostic parameters such as sensor-specific temperature, relative humidity, pressure, latitude and longitude (GPS), and flow rate. All raw sensor readings, location, and diagnostic data are packaged into structured messages with an associated network timestamp and transmitted to Aclima's cloud-based backend through an internal LTE module, using a widely adopted Internet-of-Things (IoT) lightweight transport protocol. If a network connection is not available at the time of data collection, it is stored locally and published at a later time once a connection is available. There is enough local storage available for up to 60 days of data collection.

3.2.2. Ingest (Aclima Mobile Node data)

Packets of data from sensors in the mobile node are sent to the cloud through a gateway and then forwarded to a messaging service that separates message publishing from consumption. Messages are processed by data pipelines running as cloud-based jobs. Data are stored in multiple cloud archives, to ensure long term preservation of lowest-level data.

Additional third-party datasets are ingested to support operational and product data processing as well as QA/QC operations. These might include modeled meteorological data,

digital elevation models (DEM), and stationary air quality monitoring data from various sources.

3.2.3. Ingest (PML data)

Partner Mobile Labs (PMLs) are handling data ingestion themselves, per their internal standard operating procedures and given their unique hardware and software capabilities.

Aclima will not ingest PML data directly into the standard pipeline. PML data will be received post-processing, and handled through command line or drag-and-drop interfaces provided by Aclima's cloud provider. Each PML team will be provided with a dedicated cloud-based data bucket, and appropriate read/write accounts to transfer data to these buckets.

PML data will be formatted in the same post-processing schema as AMN data, with QA flags and error codes tracked in the same format. Aclima will relay this data to CARB without modification via the same cloud-based data bucket as AMN data (see section 8.2). The "vehicle id" field of the data can be used to identify PML data, as well as the "organization" field contained in the data file names. The PML teams will follow the same best practices as Aclima of retaining L0 data (or the lowest level data available) indefinitely and unaltered.

4. Data transformation

Data transformation describes how Data Level 0, now ingested and stored in the Aclima backend, is transformed into Data Levels 1, 2a, and 2b.

4.1. Definition and application of Data Levels 1, 2a, and 2b

Data Level 1

Data Level 1 (L1) is derived from Level 0 (raw sensor signal) data using basic physical principles and/or calibration equations. Whereas L0 data might be a voltage or raw mass spectral data, L1 data are intermediate geophysical quantities, such as a concentration in ppb or $\mu g/m^3$. An example of L1 data is concentration in units of ppb that has been derived from the raw signal as current (in amps) from an electrochemical sensor and converted to concentration based on a calibrated sensitivity (in amps per ppb) and an offset current (in amps), but prior to any empirical correction for humidity or temperature.

Data Level 2a

Data Level 2a (L2a; also referred to as "raw" and "finalized" in other project documentation) is generally the final form of data for the purposes of SMMI data delivery, though in some cases delivery of L1 data may be included (see examples below). Aclima defines L2a data as sensor signals transformed to geophysical quantities of measurement, estimated using the sensor signal plus associated physical measurements directly related to the measurement principle such as temperature and relative humidity measurements at the sensor aperture for correction of interferences, with flagged artifacts included, and with final calibration parameters applied. Additional signal processing may also be included in the transformations to produce L2a data, such as smoothing or frequency filtering.

Examples of L2a data include:

- $PM_{2.5}$ derived from size-resolved particles counts using an assumed particle density
- NO_2 derived as a difference between O_3 and O_x (O_3+NO_2) measured from two sensors in the same AMN.
- Temperature and humidity corrections to sensor estimates
- Smoothing of black carbon data using optimized noise reduction (ONA) or other method

Depending on the measurement stream, either L1 or L2a data (or both) could be delivered to CARB. For example, size-resolved particle counts per unit volume is considered L1 and

is planned to be delivered in addition to $PM_{2.5}$ (categorized as L2a) because both measurements can be useful for emissions source categorization. Data from certain research grade instrumentation operated by the PMLs may be considered L1 if they use only the basic physical principles and a calibration factor to translate the measured quantity (e.g. the amount of light at a specific wavelength absorbed via Beer's Law) into physical concentration units.

Data Level 2b

Data Level 2b (L2b) is L2a data but using external data sources for normalization or artifact correction, such as stationary monitoring data or meteorological data. This data level is included here for completeness, but is not currently proposed to be used in any outputs derived for SMMI.

4.2. Data transformation workflows

4.2.1. Processing into Level 1 data

Raw signals are transformed into calibrated physical quantities (L1 data) representing pollutant concentrations, using basic physical principles and derived sensor-specific calibration parameters. As part of this transformation, Aclima's metadata database is typically referenced in order to pair calibration parameters with the unique reference id of the reporting sensor. The method of transformation from raw sensor signal to calibrated physical quantities depends on the specific sensor type and detection principle. For certain measurement streams, this transformation is done on-board the device by manufacturer specified methods and is not controlled by Aclima's data transformation pipeline.

4.2.2. Processing into Level 2a data

Level 1 data is transformed into Level 2a through a variety of methods, depending on the detection method and measurement stream. These methods can range from simple differences between two simultaneous measurements (e.g. NO_2 from O_3 and O_x), combination of multiple pollutants into a single quantity (e.g. $PM_{2.5}$), temperature and humidity corrections, and/or signal processing methods.

4.2.3. Time Shifting

One additional transformation step that occurs on both L1 and L2a data prior to being used to produce higher level derived data products and prior to delivery to CARB is the application of a time shift to account for the finite amount of time that passes between air being sampled at the intake of the sample lines outside the vehicle and when it is detected inside the AMN. Aclima uses a fixed time shift across the entire fleet for each

sensor based on the volumetric flow rate and the length and inner diameter of the sample lines associated with that sensor. The PMLs use a similar approach customized to each instrument and the layout in the mobile labs. This transformation is important for aligning data sets across multiple pollutants and to the true sampling location, rather than the detection location.

4.2.4 Data Flagging

As part of the data transformation process, an automated system flags individual data points as invalid according to predetermined thresholds of pollutant values and auxiliary diagnostic measurements (e.g. temperature, flow rate, pressure, and humidity). This automated flagging captures some common cases where invalid data is produced, but does not capture all of the possible instances. For this reason, there is a manual system for identifying and flagging any possible data point that is integrated into the transformation process. A detailed description of this process, known as data review, is included in Section 7. Both the automated and manually determined flags are applied to L1 and L2 data and published along with the delivered data files for SMMI. Details of how these flags are integrated into the data are described in Section 8.

Table 2.3: An overview of each modality and the L data level (coming from AMN and PMLs)

Parameter	Method	Data Level
pm_1.0	amn-pm	L2a
pm_2.5	amn-pm	L2a
pm_ch_1_count	amn-pm	L1
pm_ch_1_mass	amn-pm	L2a
pm_ch_2_count	amn-pm	L1
pm_ch_2_mass	amn-pm	L2a
pm_ch_3_count	amn-pm	L1
pm_ch_3_mass	amn-pm	L2a
pm_ch_4_count	amn-pm	L1
pm_ch_4_mass	amn-pm	L2a
pm_ch_5_count	amn-pm	L1
pm_ch_5_mass	amn-pm	L2a
pm_ch_6_count	amn-pm	L1
pm_ch_6_mass	amn-pm	L2a
c2h6	amn-c2h6	L2a
ch4	amn-ch4	L2a
no	amn-no	L2a
no2	amn-no2	L2a

o3	amn-o3	L2a
voc	amn-voc	L2a
blackcarbon	amn-blackcarbon	L2a
co2	amn-co2	L2a
co2	amn-co	L2a

5. Data modeling

Data modeling describes the further processing of lower level data via standard statistical aggregation and/or advanced analyses into spatially distributed geophysical quantities. The primary distinction between lower level data (L2 and below) and these higher data levels is the transformation of data that is indexed as a continuous time series, into data that is non-continuous (e.g. specific events in time) or has been aggregated along the time dimension, thus representing geospatial snapshots in time. The purpose of these higher level data sets is to transform mobile monitoring data into easy to interpret visualizations and analyses to derive insights that more directly address the monitoring objectives.

5.1. Definition and application of Data Levels 3-4

Data Level 3

Data Level 3 (L3) constitutes aggregated geospatial data products derived from lower level time series data using standard statistical aggregation approaches. These data products can be from single pollutant measurements or a combination of multiple-pollutants to infer pollutant source types.

Examples of L3 data include:

- Locations of individual enhancement events of TVOCs, speciated toxics, or other pollutants
- Spatially grouped clusters of methane and ethane enhancement events, indicating locations of persistent natural gas leaks
- Spatially continuous maps of ambient concentrations using basic statistical aggregation of measurements only, with no advanced probabilistic models applied

Data Level 4

Data Level 4 (L4) is defined similarly to L3, but uses advanced probabilistic models. L4 data may also incorporate external data sources, such as (but not limited to) meteorological data, stationary site data, land-use type or road type data layers, or topographical data.

Examples of L4 data include:

- A continuous map of pollutant averages with associated confidence intervals derived using a statistical reconstruction method
- A map of probabilistic estimates of likelihood of a methane leak in an area
- A map showing statistically significant hot spots of average pollutant

concentrations via the Getis-Ord (Gi*) method

5.2. Data modeling workflows

Modeling workflows operate to transform data from L1/L2 data into L3/L4 data to create data analyses focused on identification of locations of pollutant sources and areas overburdened by known pollutants.

Both L3 and L4 include two general categories of analyses.

1. *Enhancement-based analyses* identify events in time and space where concentrations of a pollutant are measurably distinguishable from the ambient background concentrations.
2. *Ambient concentration-based analyses* identify spatial gradients in average or typical ambient concentrations observed for specific air pollutants or combinations of pollutants. The specific L3 and L4 data that will be produced for SMMI will be detailed in the CAMPs after better understanding community concerns and priorities. The possible options for these analyses and additional details about the types of analyses are discussed in Section 9.

The underlying source data for all L3 and L4 data sets will be the finalized data (as L1 and L2a) included in delivery to CARB. Prior to inclusion in the modeling work flow, data that has been flagged as invalid is first removed.

Basic modeling workflow for enhancement-based analyses

1. Identification of enhancement events in the pollutant time series through “peak finding” (L1/L2a→L3)
2. Multi-pollutant analyses for categorization of enhancement events (L3→L3)
3. Spatial grouping or clustering of enhancement events and generation of basic statistics per cluster (L3→L3)

Basic modeling workflow for ambient concentration-based analyses

The basic modeling workflow for ambient concentration based analyses can vary depending on the type of analysis and the data level to be produced. Below are some examples of workflows:

Basic statistical aggregation of ambient concentrations (L3)

1. Aggregate time series data to spatial units (i.e. road segments, hexbins, etc) to define a single “pass” or “visit”, by calculating, for example, the mean (L1/L2a→L3)
2. Aggregate the single pass means over the entire monitoring period using, for example, the median (L3→L3)

Statistical reconstruction of ambient concentrations[1] (L4)

1. Group the time series data by spatial unit of choice over entire monitoring period (road segments, hexbins, etc) (L1/L2a→L1/L2a)
2. Spatially grouped time series data is combined with meteorological data, topographical data, and road type information. The output is a probabilistic estimate along with confidence intervals at the spatial unit defined in step 1 (L1/L2a→L4).

While these are potential workflows and may not fully reflect what is included in the CAMPs and final visualizations for SMMI, it sets up a framework for communicating the modeling workflows associated with different derived data products. This framework will be used in the CAMPs and in descriptive material accompanying the final public visualizations.

[1] The specific details of this approach are outside the scope of this document, but will be communicated in the CAMPs if this method is chosen for analysis.

6. Data storage

An essential step in the data management pipeline is *data storage*, from ingestion of Level 0 data through the compute operations that support processing data into Levels 3 and 4.

Aclima's data storage strategy includes the following components:

1. Use of industry standard, scalable cloud data stores to provide persistent long term storage of multiple levels of data.
2. Use of cloud compute tools to support both streaming and batch processes for efficient large volume data transformations between levels, with ability to write to all data stores.
3. A storage architecture that supports backfilling operations to re-transform data between levels, when updated calibration parameters and data flagging indicate the need.
4. Use of an off-the-shelf, central query interface for data at various levels.
5. Use of a cloud-based metadata database to be able to connect data to physical sensor hardware and vehicles where the sensors were installed, and to track calibration parameters.
6. The ability to write to flat files for final delivery to customers, and the availability of cloud-based UI, command line tools, and APIs to access this data.

For the SMMI project, Aclima will provide CARB access to the stored data for 3 months after the end of the contract period. Aclima will then remove CARB's access and de-provision the cloud data store.

7. Data review and quality assurance

The data management system incorporates support for data review checks, defined as the manual or automated flagging of automated signals from sensor time series. This section emphasizes the engineered components of the data review and quality assurance process. Scientific details of data review can be found in the appropriate QA/QC documents (to be included in the CAMPs).

7.1. Overview

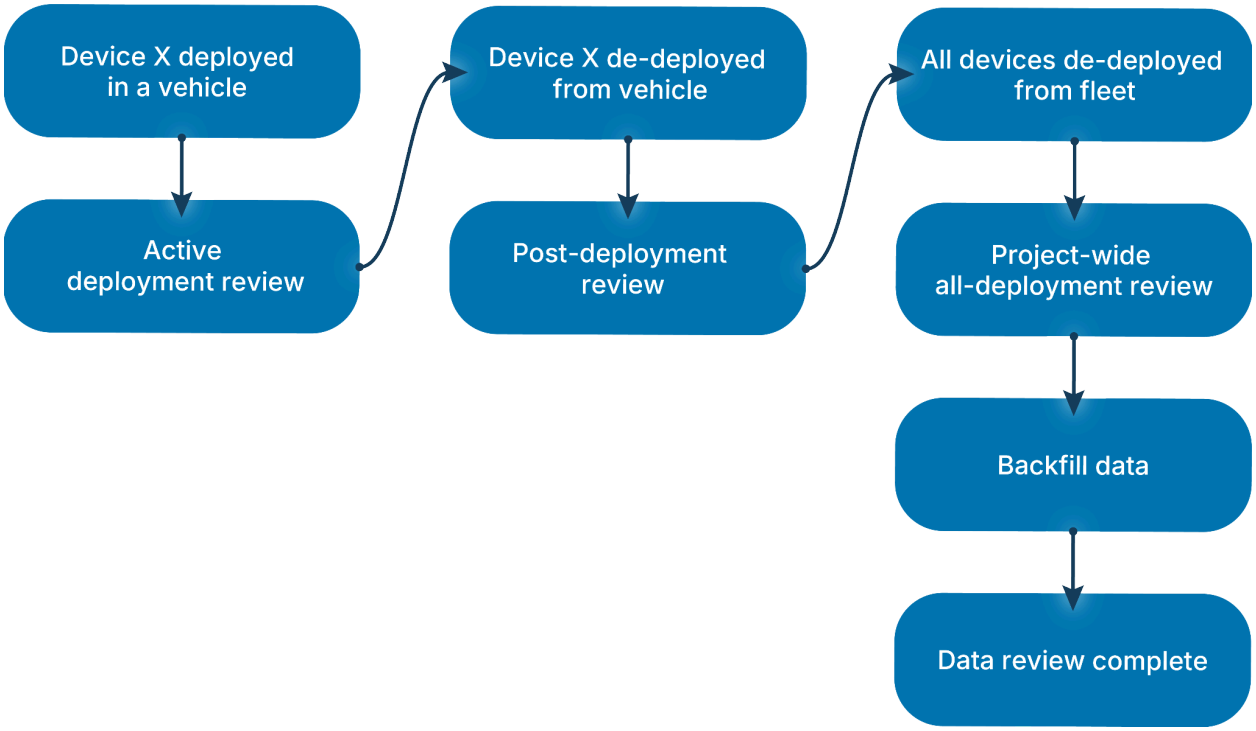

Figure 7.1: overview of the data review and quality assurance process. Each review step includes both manual and automated activities.

7.2. Active deployment review

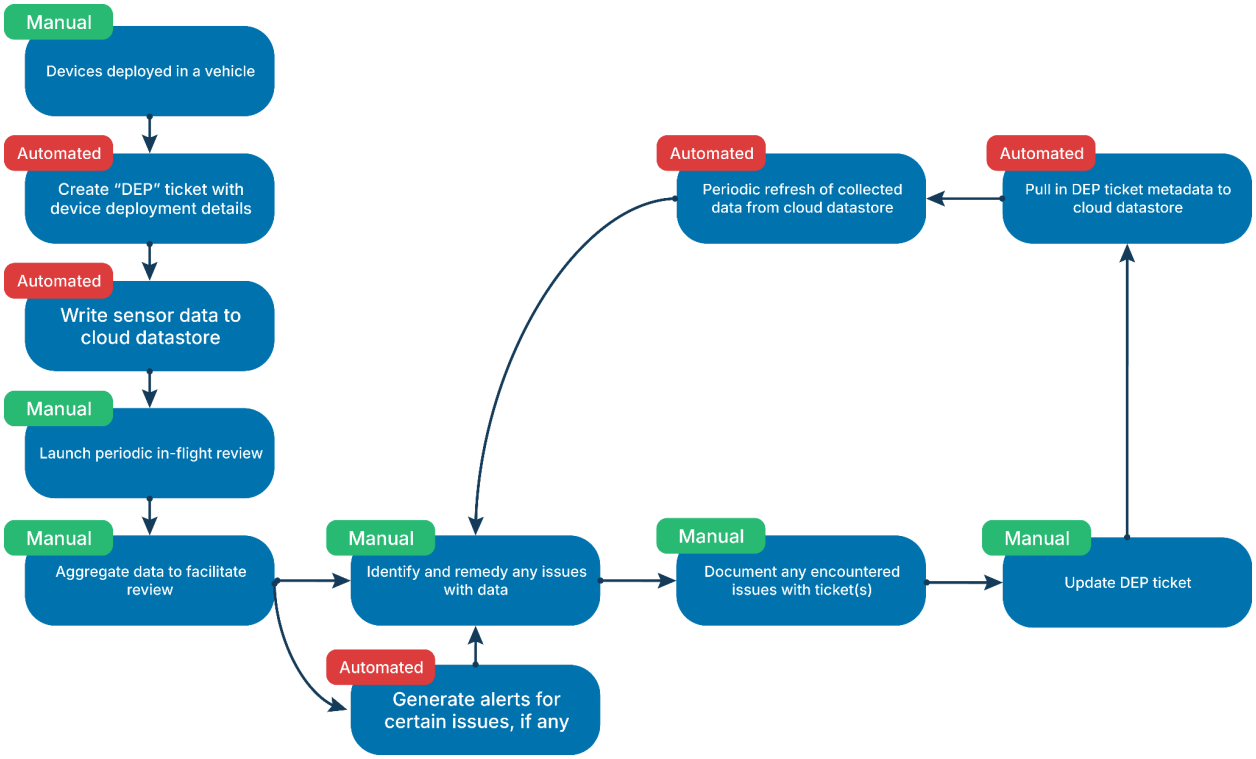

Figure 7.2: data review process for a single device while its deployment is in-progress, or active. "Manual" flags indicate work undertaken by the Data Quality Operations (DQO) team.

Each device deployment begins when that device is installed or deployed into a vehicle. From the time of installation until it is uninstalled, the device deployment is considered to be "active". Data from devices' sensor(s) are periodically reviewed as active deployments progress. The main stages of this flow are described below.

7.2.1. Active deployment review workflow

Step 1: Create "DEP" ticket with device deployment details

The installation of a device into a vehicle signals the start of that device's deployment in that vehicle. Each installation triggers an automated creation of a unique deployment (DEP) ticket in Aclima's off-the-shelf ticketing system. Each DEP ticket is then used as a central repository for data reviewers to record information on all data quality and hardware issues detected, maintenance performed, standardized system checks performed (e.g. flow checks, connectivity checks, etc), and any other relevant information during a deployment.

Step 2: Write sensor data to cloud data store

Once a device has been installed and associated with a vehicle, it can be powered on to begin data collection. During data collection, sensor data streams from the AMN and is written to the cloud data store.

Step 3: Launch periodic in-flight review

Once devices are collecting data, the Data Quality Operations (DQO) Team begins periodic checks of the data, to flag and remedy any sensor or data quality issues as they arise.

Step 4: Aggregate data to facilitate review

To view collected data, sensor data must be imported into a database and temporally aggregated (e.g. to 1 minute or 1 hour) to make data review more efficient. It is then available for manual manipulation, and is checked against automated data issue flagging. The underlying unaggregated data (at 1 Hz) is still accessible for data reviewers when the need arises for focusing on short (minutes to hours) time intervals of interest.

Step 5: Identify and remedy any issues with data

The DQO Team manually visualizes collection data as it comes in to quickly flag, rule-out, and resolve issues that arise in the field. Data is automatically and frequently refreshed in these tools.

Step 6: Generate alerts for certain issues, if any

Certain data quality or sensor issues tend to display the same behavior each time they occur, and can be translated to an automated rule to alert users when this behavior is identified.

Step 7: Document any encountered issues with ticket(s)

When issues are detected via automated or customized alerts, or during manual data review, they are documented via an AMN Data Review (ADR) ticket. Manually-generated flags are also referred to as “omissions” and are attached to stretches of data that the DQO Team wishes to omit.

Reasons for flagging data as non-valid (omission) can include:

- Longer than usual sensor warmup period: sometimes a device’s sensors take longer than usual to warm up. The measured data will reflect this, showing smoothed, steadily increasing or decreasing time series behavior within the first hour of a vehicle’s shift.

- Leak: sample tubing gets disconnected from the device, and the device begins sampling from inside the vehicle, resulting in unrealistic measurements.
- General sensor failure: a sensor fails completely while deployed and begins reporting non-physical data.
- Flow blockage or failing flow: the flow path becomes obstructed due to a clogged filter or object entering the flowpath, or the flow pump begins to fail, resulting in flow rates outside of acceptable limits.

Step 8: Update DEP ticket

Omission tickets generated during active data review are linked to the ADR Issue tickets which prompted those omissions, to provide traceability in Aclima's database back to the reason for omission. In addition, all issue and omission tickets get linked to their relevant DEP ticket.

Step 9: Pulls in DEP metadata to cloud data store

Omission tickets created during active data review are periodically ingested into the cloud data store, where their details are stored until they are incorporated into generating validity and QA flags for the final data product during the post-deployment backfill.

Step 10: Periodic refresh of collected data from cloud data store

Data tables storing sensor data and ticket metadata are refreshed periodically as new data is collected, and the active data review process is repeated for this next section of data.

7.2.2. Customized alerting

In addition to the manual review and the basic automated alerts, the data review team can define threshold-based alerts, to alert for specific conditions that may indicate atypical behavior for a given sensor that needs to be resolved. The team uses visualizations to get more information on the issue, initiate an issue ticket, and start the resolution process.

7.3. Post-deployment review

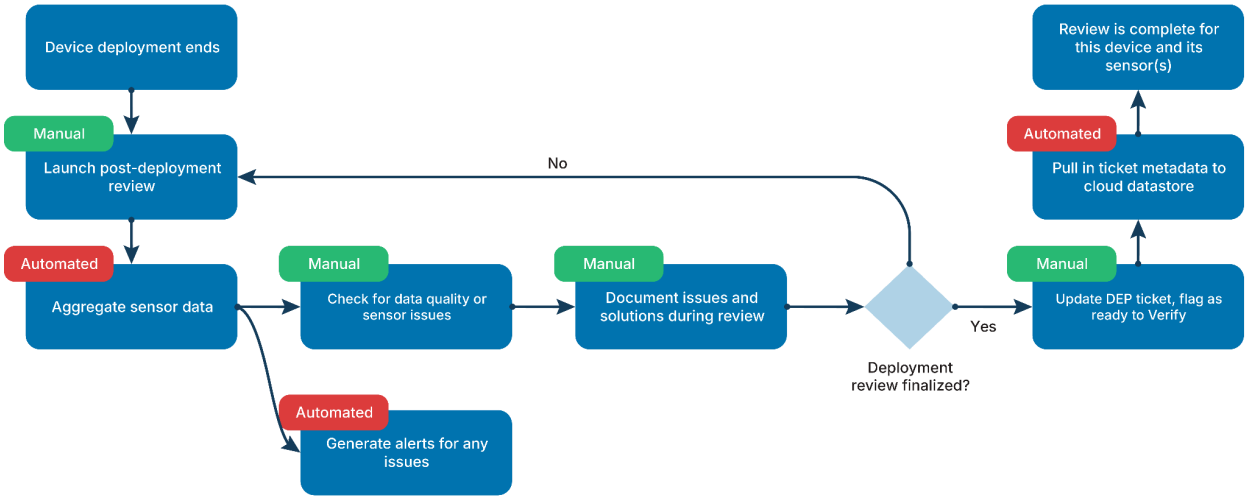

Figure 7.3: data review process for a single device once a deployment concludes. See text below for detailed discussion of each step.

The post-deployment review phase is a final data review of all sensor data collected during a given deployment, to ensure that any issues not identified during active data reviews are detected and addressed.

7.3.1. Post-deployment review workflow

Step 1: Launch post-deployment review

Once a device deployment ends, the DQO Team initiates a post-deployment review of all sensor data collected during that device's deployment, to ensure any issues that may have slipped through the cracks during the deployment period are detected before data is finally verified.

Step 2: Aggregate sensor data

Ticket metadata detailing the device's deployment is used to query all sensor data associated with that deployment over the deployment's duration. This data is then compiled into a database.

Step 3: Generate alerts for any issues

Any occurrences of well-characterized data quality issues are flagged by alerts in a dashboard accessible to the DQO Team. These issues are quantified in a summary view for the entire deployment, to help the DQO Team efficiently target remedying these issues during post-deployment review.

Step 4: Check for data quality or sensor issues

The review team uses a variety of visualization tools to look for data quality or sensor issues. Any issues detected are tracked with an AMN Data Review (ADR) issue ticket. This ADR ticket is linked to the deployment's DEP ticket, so that all data quality or hardware issues encountered by that deployment can be tracked in a centralized place.

Step 5: Document issues and solutions during review

Similar to active deployment review, issues encountered during post-deployment review are documented using ADR issue and/or omission tickets.

Step 6: Deployment review finalized check

Once a DQO Team member has concluded a post-deployment review, they update the deployment's DEP ticket to reflect that the deployment device's sensor(s) have received their initial data review.

Step 7: Update DEP ticket, flag as ready to Verify

If the review is finalized, the deployment is ready for a backfill, and the DEP ticket receives a flag of "Ready to Verify."

Step 8: Pull in ticket metadata to cloud data store

Metadata contained within Aclima's ADR, DEP, and VER tickets are pulled into the cloud data store. This information is referenced by later backfill workflows which reprocess sensor data and flag it as "Verified" based on the metadata contained within attached tickets.

7.4. Post-all deployments review

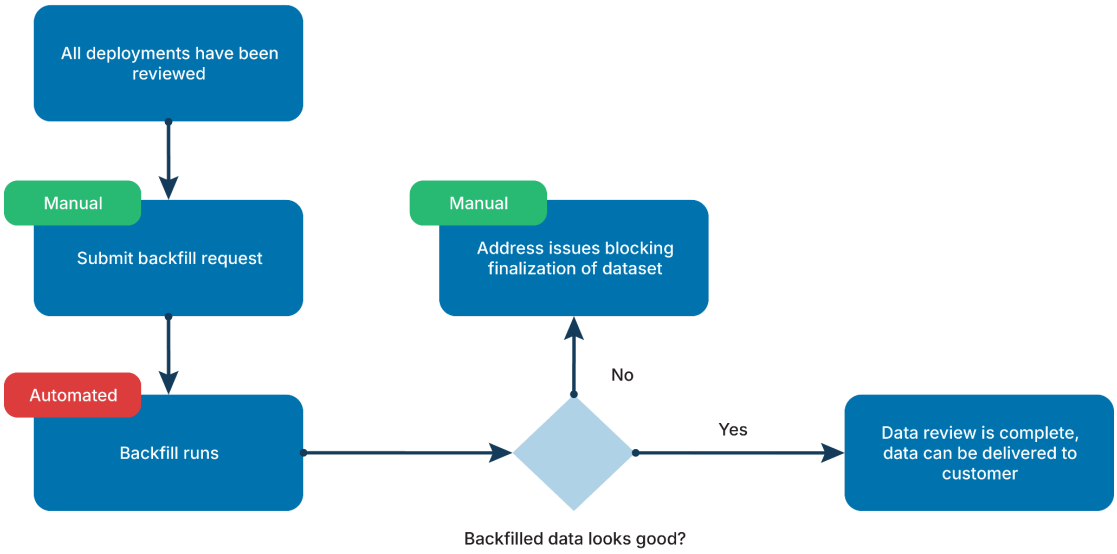

Figure 7.4: data review process once all AMN deployments are complete.

Once all deployments relevant to the SMMI contract have been reviewed over the relevant time period, a *backfill* of collected data is run. The backfill re-processes all collected data, along with deployment-related metadata, to output Verified data that is ready to be provided to the Client.

Backfilling definition

Data review and calibration processes during and post-deployment result in a set of manual determined data flags and updated calibration parameters to apply. In order for these corrections to be applied, a *backfill* must be triggered to reprocess Level 1 data transformations that make use of sensor calibration parameters and any downstream aggregations (i.e. modeling workflows to produce L3 or L4 data) for all data collected in the field.

A backfill is a process by which the entirety (or some subset) of the daily workflow is rerun to incorporate the updated parameters and is critical for Aclima's data platform. Tables in the data warehouse include data versions associated with the code used to process the data to track the lineage of data. Note that a backfill will never alter the values stored immediately upon ingestion (typically L0 data and in some cases L1 data).

The backfill process additionally generates metadata that categorize data flags and detail the period for which they were made. Verification status from data review tickets is also incorporated. These annotations will be made available to CARB in the delivered data files.

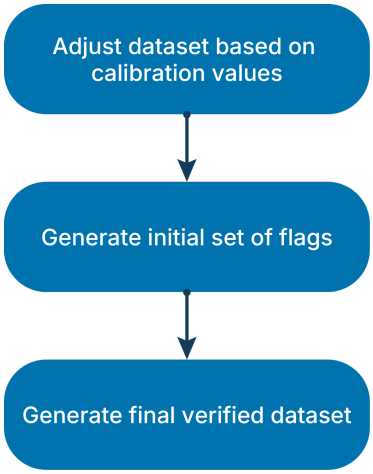

Figure 7.5: overview of the three main stages in backfilling. In step 1, calibration value adjustments are considered and applied if necessary. In step 2, previously generated metadata that indicates potentially invalid data are converted into formal flags. In step 3, a final verification step checks over all flags and prepared the final Level 2a data for delivery.

7.4.1. Post-all deployments review workflow

Step 1: Submit backfill request

The DQO Team submits a request to backfill all reviewed data, including details used to format the required backfill workflow. Once these details are in place, the backfill is scheduled and initiated.

Step 2: Backfill runs

A sequence of workflows is run to re-process sensor data along with metadata from data review tickets generated during the active- and post-deployment review processes.

During the backfill, omission ticket metadata is referenced by the backfill workflow to flag any applicable data as non-valid, to ensure that omitted data can be easily excluded or filtered from final data products if desired.

In addition to omission ticket metadata, metadata from deployment metadata (DEP) and verifications metadata (VER) tickets that have been flagged as Ready to Verify also gets incorporated, to flag all applicable data points as Verified.

Step 3: Backfilled data looks good check

After the backfill concludes, the reprocessed data is re-queried by the review team to refresh their data visualization tools as one last check to make sure the now Verified data

is ready to be delivered to CARB. If the backfilled data passes this check, the DQO Team can conclude its deployment review process for that contract period and its deployments.

Step 4: Address issues blocking finalization of dataset

If any errant data is identified, the DQO Team essentially repeats Steps 1 through 3 to correct the issue through another focused round of review, request a targeted follow-up backfill to resolve the outstanding issue(s), then check the backfilled data once again.

7.5. Data revisions

The Partner Mobile Lab (PML) partners will deliver finalized data on the same cadence, and to the same schema, to the CARB-accessible cloud data store bucket as described in Section 8.2. However, for some period after the end of the data collection period (as long as 1-2 years, or more), additional research may continue. This may lead to necessary revisions, such as to calibration parameters, that CARB may wish to propagate internally or externally.

Before the contract ends, PMLs will submit updates to Aclima. Aclima will upload the new file to the cloud data store, increment the revision number in the filename, and update the data store readme (see Section 8.2.3).

After the contract period is over, Aclima will have no further involvement in these PML-sourced revisions. However, Aclima recommends that PMLs submit updates to CARB directly and for consistency follow the same revisions naming protocol (see Section 8.2.3).

8. Data transfer

This section describes how Aclima will format and deliver *finalized* data to CARB. Finalized data refers to verified data corresponding to the highest available non-modeled data level for each measurement stream, typically Level 2a, but may be Level 1 in some cases. It focuses on a description of the data schema and its typical application; and on how cloud storage tools will facilitate the transfer of data, along with usability and security considerations.

8.1. Level 2a Data Schema

8.1.1. Overview of schema

A critical subtask of the SMMI is the delivery of finalized data (i.e. geolocated 1 Hz concentration time series) for all pollutants measured by Aclima or Partner Mobile Labs (PMLs). The delivered data should be formatted in a schema that contains the necessary information to describe the concentration of a location at a particular time and understand any important contextual information including data quality flags and vehicle metadata. This information will allow air quality regulators to make effective use of this data.

The schema shown in Table 8.1 contains all of this information in a minimal format that attempts to strike a balance between limiting redundant information and remaining accessible for analysts that should not need to apply multiple join operations to preprocess the data.

Table 8.1: data schema for Level 2a data

Field Name	Field Type	Field Description
timestamp	STRING	The ISO 8601 encoded UTC timestamp at which the measurement was taken with timezone information. The format string is *YYYY-MM-DDTHH:mm:SSZ* (e.g. ***2024-09-03T14:46:39Z***)
parameter	STRING	The measurand described by the row, typically pollutant (e.g. no, no2, co) or pollutant subtype (pm2.5_ch1). For each parameter value, a separate lookup table links this field to AQS parameter codes (1:1 relationship) and a longer text description (see Table 8.2; see section 8.1.5 for a general note on AQS codes).

method	STRING	A method identifier for the sensor used to collect the measurement. See discussion in section 8.1.4 below. For each method value, a separate lookup table links this field to the relevant AQS method code (1:1 relationship) and a longer text description (see Table 8.2; see section 8.1.5 for a general note on AQS codes).
duration	INTEGER	Sample duration for the associated measurand. This is relevant in particular for certain methods that require sampling times longer than 1 second or may have variable sampling times over the course of the data collection period. In these cases, the **timestamp** field represents the start of the sampling period.
value	FLOAT	The measurement value for the measurand **parameter** at the time **timestamp** in the location **location**.
unit	STRING	The unit of measurement that **value** represents.
latitude	FLOAT	The latitude at which the measurement was taken. This is used in conjunction with longitude to completely describe the corrected position of the measurement.
longitude	FLOAT	The longitude at which the measurement was taken. This is used in conjunction with latitude to completely describe the corrected position of the measurement
status_indicator	INTEGER	An integer value that represents a boolean (either 0 → false or 1 → true). A value of 1 indicates that this measurement has been flagged, meaning there is a relevant annotation (in **qualifier_codes**) that contextualizes data quality or conditions.
qualifier_codes	STRING	A blob of comma-separated AQS Codes that documents the reason that data was excluded or flags any notable conditions annotated via AQS codes. Note that data is only reported when instruments are active within a defined data

		collection area. Note that any data collected while instruments are active and in the data collection area, but that is invalidated for another reason, *will* be included.
vehicle_id	STRING	A string that uniquely identifies the vehicle that collected the measurement described by the row.

8.1.2. Schema fields and utility

There are two main categories of fields in this simple schema: primary and metadata. The primary fields are `timestamp`, `parameter`, `longitude`, `latitude`, `units`, and `value`. These fields describe the time at which the measurement was collected, the pollutant being measured, the location of the measurement, the value of the measurement, and the units of measurement for that value.

The metadata fields are `status_indicator`, `qualifier_codes`, `method` and `vehicle_id`. These fields provide important context that help in the interpretation of the measurement described in the primary fields. `status_indicator`, `qualifier_codes`, and `vehicle_id` describe the quality of the measurement via `status_indicator` and `qualifier_codes` and the platform that collected the measurement via `vehicle_id`. The field `method` is used to identify the sensor type or method of sensing used for collection of the measurement.

8.1.3. AQS code lookup table

The AQS code lookup table is used to bridge between the Aclima schema and AQS parameter and method codes, as described in Table 8.2.

Table 8.2: mapping between Aclima schema keys and lookup table fields to enable bridging to AQS parameter and method codes

Aclima schema key	**Lookup table fields**	
parameter, e.g. “no2”	AQS *parameter code*, e.g. “42602”	*Description*, e.g. “Nitrogen dioxide (NO2)”
method, e.g. “amn-no2”	AQS *method code*, e.g. “010”	*Description*, e.g. “Integrated passive monitor”

8.1.4. Populating the method field

The method of sensing used to collect the measurement described by a row is important metadata that cannot be excluded, especially given the diversity of measurements between Aclima and the 3 PMLs. We propose including this information directly in the data schema, as follows:

1. Using and extending where applicable the Method codes described in the AQS coding dictionary.
2. Designing unique string identifiers for different sensors and including those as identification fields.

8.1.5. General note on AQS codes

The exact choice of parameter, method, qualifier codes will be developed in coordination with CARB and the PML teams. There will be some overlap with existing regulatory AQS parameters, methods, and qualifier codes, but likely also some new codes for unique measurements made as part of this project.

8.2. Level 2a data transfer

8.2.1. Scope and objectives of finalized data transfer

Aclima has designed its data transfer approach primarily to ensure CARB has access to finalized data as quickly and seamlessly as possible. Aclima will make data available via cloud storage on a regular cadence. Aclima expects CARB to pull data from cloud storage at their convenience.

8.2.2. Making finalized data available via Google Cloud Storage

Aclima will use Google Cloud Storage (GCS) to store snapshots of the finalized data. GCS is a managed service from Google, for storing unstructured data of any size, for retrieval at any time. GCS provides a range of UI, client library, and command line tools for accessing data and/or integrating with other cloud providers. Figure 8.1 below showcases optional flows for data transfer, described in more detail in the following sections.

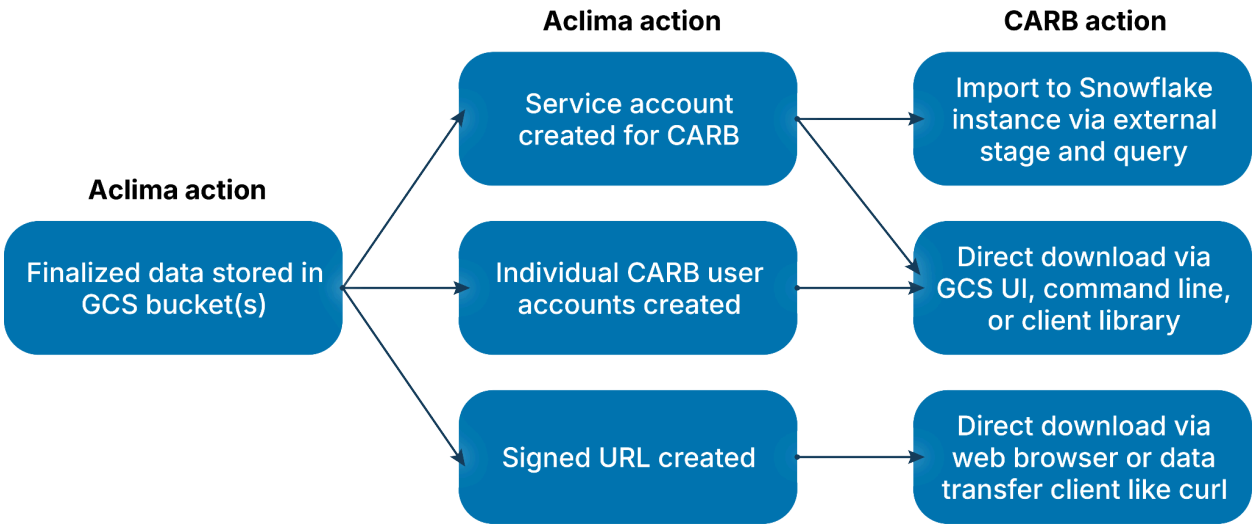

Figure 8.1: options for accessing Level 2a data from GCS bucket(s), using a variety of means per CARB preference

The GCS bucket will initially be configured as a multi-region bucket for lowest latency access. Its default storage class will be standard, which is optimized for frequent access at low cost with higher at-rest storage fees. If access patterns indicate that a file is accessed infrequently (less than once per month), the storage class may be transitioned to nearline or coldline, potentially using the GCP autoclass mechanism.

8.2.3. Data file formatting

Raw 1 Hz data will be provided in gzipped comma-separated value (csv) files, one per `organization_method` name per month.

For example, all $PM_{2.5}$ data from the AMN PM instrument for month 1 of data collection would be in a single csv file; all $PM_{2.5}$ from the AMN PM instrument for month 2, would be in a separate, single csv file.

The following naming convention will be followed, and is described in detail in Table 8.3:

```
gs://[bucket]/[organization]_[method]_[measurement
date]_[revision_number]_[revision_date]
```

E.g.

```
gs://bucket123/aclima_AMN-bc_202605_r1_20261111
```

In addition, a readme file will be included at the top level of the cloud data store to document the nature of any revisions made, starting with the first version of any uploaded file ("r1").

Table 8.3: definition of naming convention elements

`gs://`	The GCS domain
`bucket`	The GCS bucket containing all data; this will be the same for all deliveries
`organization`	Denoting the organization which conducted the monitoring, e.g. "aclima"
`method`	Collection instrument name, e.g. "amn-bc" for the black carbon sensor in the Aclima Mobile Node. Organization+method will be guaranteed to be unique.
`measurement date`	YYYYMM date of collection e.g. "202605"
`revision number`	First uploaded version of the file will be "r1". Subsequent revised versions, if any, will be r2, r3, etc. Reasons for and nature of the revision will be included in the data bucket readme file.
`revision date`	YYYYMMDD date of revision, if any, e.g. "20261111". For the first uploaded version of the file, the revision date will be the upload date.

8.2.4. Delivery cadence and estimated file sizing

Finalized data will be transferred to CARB on a monthly basis beginning four months after monitoring has commenced. Data received from sensors flows through all data processing levels and to stage finalized data snapshots in a GCS bucket where it will be available to CARB. The last data delivery will occur at the end of the contract. Figure 8.2 below describes a hypothetical series of data deliveries, assuming a hypothetical monitoring period beginning on June 1, 2025 and ending on October 30, 2025. Note: this is for the purpose of illustration only; the actual proposed monitoring period is 9 months.

In this scenario, the first data delivery occurs 4 months after start (October 1, 2025, at latest), and includes all data gathered during June 2025. Data deliveries continue monthly in the same way until the final data delivery 3 months after end of monitoring (February 1, 2026, at latest), which covers monitoring conducted during October 2025. New months of data will be added to the bucket as new files (see filenaming convention, Section 8.2.3),

while past monthly deliveries still remain available. By the time of last month's delivery, all months of data are available as individual files.

All data for each delivery will be delivered at the same time, into the GCS bucket. Once the data is available, CARB will be notified automatically, by e-mail to the CARB project manager. CARB can expect sizing of monthly data deliveries to be in the tens to hundreds of GB.

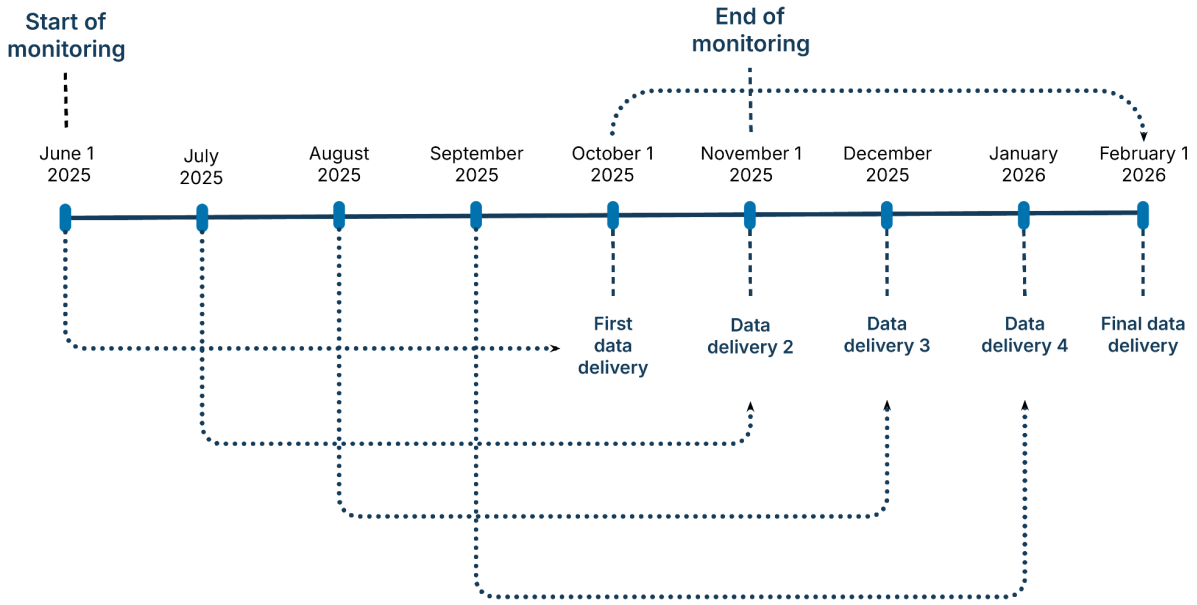

Figure 8.2: example cadence of data deliveries, assuming a hypothetical monitoring period only from June 1, 2025 to November 1, 2025. *Note that this is for the purpose of easy illustration only; the actual monitoring period is proposed for 9 months*. In this hypothetical example, the first data delivery is on October 1, 2025, covers all data collected June 1-30, 2025. The last data delivery on February 1, 2026, covers all data collected October 1–31, 2025. Data deliveries are always cumulative within the bucket, in that they add to the existing set of files.

8.2.5. Maintenance obligation

Aclima will provide CARB access to the stored data for 3 months after the end of the contract period. Aclima will then remove CARB's access and deprovision the storage bucket.

8.3. Proposed transfer pathways

Aclima understands that CARB prefers to transfer directly from GCS to Snowflake, using programmatic means. This is described below, along with other possible pathways.

8.3.1. Preferred pathway: Snowflake integration

Aclima will work with the CARB IT department to establish an integration with Aclima's GCS and CARB's Snowflake instance. CARB IT in Snowflake would create a service account that can then be granted permissions by Aclima to access the Aclima GCS bucket(s) that store the finalized data. Aclima will then grant the Service Account created by Snowflake permissions to access bucket objects using Google Cloud IAM. This approach has the advantage of not requiring any individual users on the CARB side to have GCP accounts: only the service account is authenticated.

Key technical steps in this integration are described below and in Figure 8.3, and in detail in the official GCS<>Snowflake integration tutorial from Snowflake.

1. **(CARB action) Create a Storage Integration in Snowflake**: Use CREATE STORAGE INTEGRATION to create an integration object that Snowflake uses to authenticate with GCS, setting allowed (and optionally, blocked) bucket locations for access control.
2. **(CARB action) Retrieve the GCS Service Account ID**: Run DESC STORAGE INTEGRATION <integration_name> to get the Snowflake-managed GCS service account ID.
3. **(Aclima action) Grant Permissions in Google Cloud**: In GCS, assign the Storage Object Viewer role to this service account for the specified bucket(s).
4. **(CARB action) Create an External Stage**: Define an external stage in Snowflake that references the storage integration and specifies the GCS bucket URL.
5. **(CARB action) Load or Unload Data**: Use COPY INTO or other commands to interact with data in GCS through the Snowflake stage.

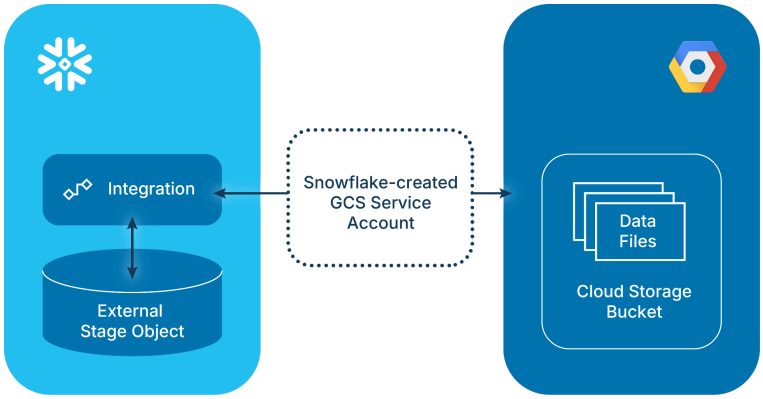

Figure 8.3: integration flow for a Snowflake Cloud Storage stage.

8.3.2. Alternative pathway: Google Cloud console UI and tools capabilities

An alternative means of accessing the data is via tools built into Google Cloud Platform (GCP), as described below. These access pathways require anyone interacting with the data via GCS to have a GCP account.

- Google Cloud console user interface: this interface allows CARB, when appropriately credentialed, to visually navigate to the specific storage bucket that contains the Level 2a data, and select specific folders or data objects within that bucket for download.

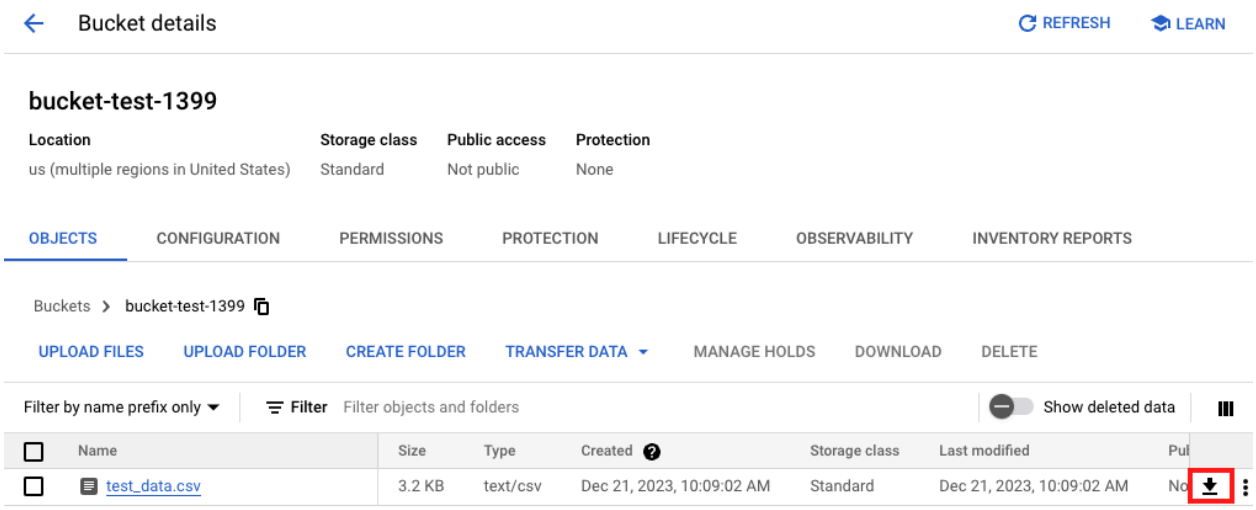

Figure 8.4: screenshot of the GCS console user interface, highlighting the download feature available for a specific dataset within a bucket.

- Client libraries: GCS supports a range of languages for programmatic download. For example, to use the Python client library, CARB would:
    a. Install the client library locally, if not already installed.
    b. Set up authentication, which requires a service account key with the appropriate permissions, and setting the appropriate local environment variable.
    c. Download the data using Python functions from the client library, such as `blob.download_to_filename(destination_filename)`
- Command line utility: the Google Cloud Command Line Interface (CLI) is available via a local command line shell such as Bash, Zsh, or Fish. CARB might use the CLI to accomplish a download as follows:
    a. Install the Google Cloud CLI, if not already installed.

b. Authenticate with Google Cloud with a user or service account that has permissions to access the bucket.
c. Given the file path to the bucket, use the `gsutil` function to download one or more files from the bucket path. Various `gsutil` subcommands can be used to download files, such as `cp` and `rsync`. Downloads are resumable. When downloading a large number of files, the download can be parallelized using the `-m` option.
d. An alternative command line utility that offers similar functionality to `gsutil` with improved performance is `gcloud storage`, which is also part of the Google Cloud CLI.

More information and full instructions on using client libraries and the command line utility is available via the Google Cloud website.

8.3.3. Transfer pathway security

Aclima uses Google Cloud Identity and Access Management (IAM) tools and minimum access best practices for securing all the mobile monitoring data it collects. These practices govern access to data stored in Cloud Storage, BigQuery, open source databases running in GCP, and all other infrastructure used to process data.

For the CARB-preferred integration pathway via Snowflake, Aclima will assign IAM access to a service account. This allows CARB to upload a JSON access key to Snowflake and use Snowflake external stages and SQL query tools to bring data directly into Snowflake instances.

Depending on CARB need, Aclima may consider using a signed URL for temporary access: this would grant a CARB user time-limited access to individual objects inside the storage bucket, without requiring long-term IAM roles.

Alternatively, Aclima can assign a predefined IAM role to designated CARB users, likely at the Storage Object Viewer level (i.e. read-only access to objects in the bucket). This then will allow Aclima to grant user-specific access to the storage bucket's permissions.

Aclima will enable audit logging, monitoring and notifications within Google Cloud for tracking access and actions.

9. StoryMap datasets and visualizations

This section describes how Aclima datasets (including both finalized data and modeled phenomena, i.e. Levels 2a and above) will be used to support visualizations using CARB's preferred tool, ESRI StoryMaps, and how these visualizations will be developed and delivered to CARB.

Mobile monitoring data gathered as part of SMMI are intended to facilitate focused actions by communities and CARB, including any future work to identify and prioritize locations for more comprehensive community-scale air monitoring, or develop Community Emissions Reduction Programs (CERPs).

To support this potential future work, Aclima will provide a defined suite of datasets and associated StoryMaps visualizations, for which mobile monitoring data is well-suited, and that can help communities better understand and interpret the data in the context of their goals. This will be available for consideration by the Project Expert Group (PEG), and for use during the CAMP process.

These datasets and visualizations are designed to support a) identifying sources and b) identifying locations of disproportionate impact, as described below and listed in more detail in Table 9.1.

9.1. Datasets to help identify sources

9.1.1. Enhancement-based datasets

These use a combination of measurements that are *indicative* of a particular source to suggest a source location (and, as a result, are often referred to as "indicators"). Aclima defines an enhancement as a localized elevation in concentration of a pollutant that is measurably distinguishable from the ambient background. These typically represent the detection of emissions in the form of a coherent plume in the vicinity of the source. Enhancements are typically identified in the time-resolved data, and repeated observations of an enhancement in a given location suggest a degree of persistence in the signal and therefore more confidence in a hotspot and underlying source. Temporal characterization can also be an important aspect of enhancement-based analyses.

Examples of source types that can be identified by this type of analysis include diesel particulate matter, natural gas leaks, and coarse particulate matter (PM) from construction sites or other industrial facilities.

Enhancement-based datasets can be visualized spatially in a variety of ways, from heatmaps to individual points.

9.1.2. Ambient concentration-based datasets

The presence of spatial gradients in the ambient concentrations observed for specific air pollutants (or combinations of different measured pollutants) may indicate the presence of a source and the spatial extent of detectable emissions from that source. These spatial gradients may be observed in the temporally averaged concentrations aggregated over multiple measurements or within individual time series of a single drive. Whereas enhancement-based datasets may more precisely indicate the location of a source and the extent to which pollution impacts the areas immediately downwind, ambient concentration-based datasets can provide a wider spatial view of certain sources and the areas they persistently impact.

An example of this type of signal is the observation of mobile source tracers, such as NO_2 or black carbon in a neighborhood bordering a freeway and the decay of concentrations with distance from the freeway. In some cases, a single pollutant can be associated with a specific source of concern, like arsenic from cement plants. For many source types, a combination of measurements are used to identify the source, such as the combination of methane and sulfur-containing compounds identifying animal feedlots.

Ambient concentration-based datasets can be visualized as gradients in a variety of spatial aggregations, such as hexbins or road segments at different levels of resolution.

9.2. Datasets to help identify locations of disproportionate impact

These datasets can be best visualized as spatial gradients (via different spatial aggregations, such as hexbins or road segments) and statistically-derived hotspot markers (geographic points). Identifying locations of disproportionate impact involves visually assessing proximity of spatial gradients and hotspot markets to a geography of concern. These datasets can be useful when combined with other geospatial datasets to generate quantitative measures of impact equity.

Table 9.1: description and classification of datasets, and appropriate (possible) visualizations in StoryMaps

Identify sources	
Enhancement-based datasets	Appropriate visualizations
Indicators[2]	

[2] I.e. indicative of a particular source type and can be used to identify a source location

Black carbon (BC)	Analysis of block level data of single pollutants for signatures.	Point, heatmap, table
$PM_{2.5}$		Point, heatmap, table
TVOC (single pollutant, no type)		Point, heatmap, table
Natural gas leak	Clear indication of a source and characterized as natural gas, plus localization. A probabilistic timeline of when a leak started and when it was resolved.	Point, heatmap, table
Multiple modalities from reference-grade detection	Reference-grade instruments operated by PMLs (and, to a limited extent, the Aclima fleet) will support a range of additional modalities as indicator data products, such as speciated TVOCs and toxic metals	Point, heatmap, table
Hotspot identification from indicators	Localized, time-resolved hotspots for indicators tagged with wind direction at occurrence.	Point, heatmap, table
Ambient concentration-based datasets		Appropriate visualizations
$PM_{2.5}$, NO_2, O_3, and BC as spatial gradients	Geospatially and temporally aggregated averages, supporting visual identification of clear gradients from a particular source to impacted areas	Hexbin (e.g. H3 hexagon levels 10 and 11[3]); road segments. See Figures 9.1 and 9.2 for examples.
Hotspot identification from ambient concentrations, for $PM_{2.5}$, NO_2, O_3, and BC	Neighborhood-scale, persistent hotspots	Hexbin (e.g. H3 hexagon levels 10 and 11) or point

[3] H3 level 10 hexagons are on average ~0.01 km^2 in area and with average edge lengths of ~0.075 km; H3 level 11 hexagons are on average ~0.002 km^2 in area and with average edge lengths of ~0.029 km. See this reference for more computed statistics for hexagons under the H3 system.

Identify locations of disproportionate impact		
$PM_{2.5}$, NO_2, O_3, and BC as spatial gradients	These ambient-concentration based datasets, as described above, are also useful for visually identifying locations of disproportionate impact, or for further downstream correlative work with other geospatial datasets.	Hexbin (e.g. H3 hexagon levels 10 and 11[4]); road segments

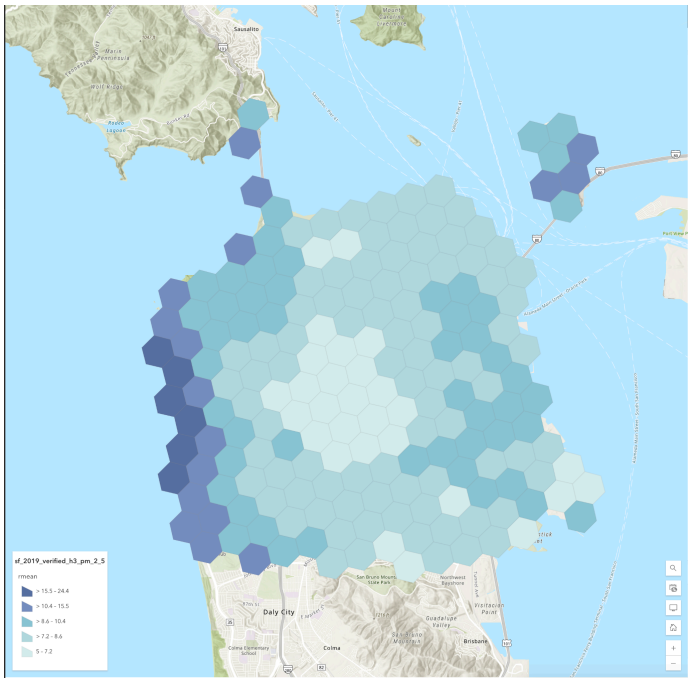

Figure 9.1: example of an ambient concentration data set ($PM_{2.5}$) plotted using hexbins, as a StoryMaps element. This visualization can be done with Data Levels L3 or L4.

[4] H3 level 10 hexagons are on average ~0.01 km^2 in area and with average edge lengths of ~0.075 km; H3 level 11 hexagons are on average ~0.002 km^2 in area and with average edge lengths of ~0.029 km. See this reference for more computed statistics for hexagons under the H3 system.

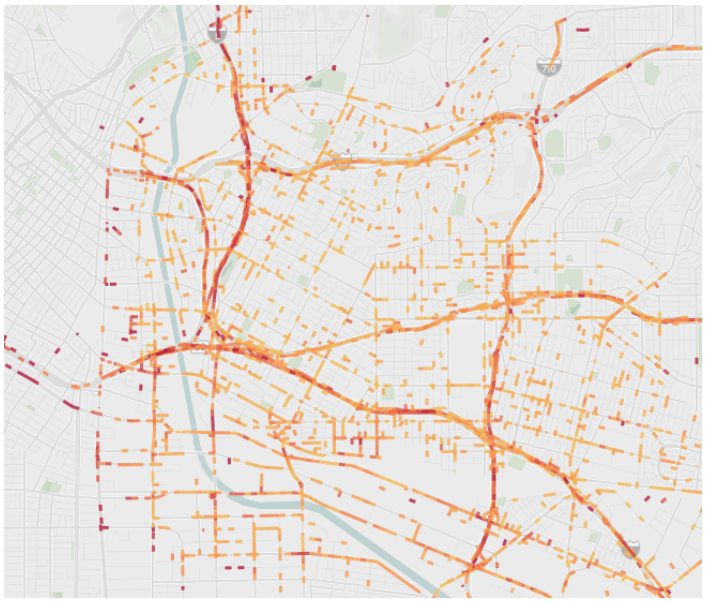

Figure 9.2: example of an enhancement-based dataset (Multi-pollutant: Black Carbon and Nitric Oxide) plotted using road segments. This visualization is done with Data Level L3.

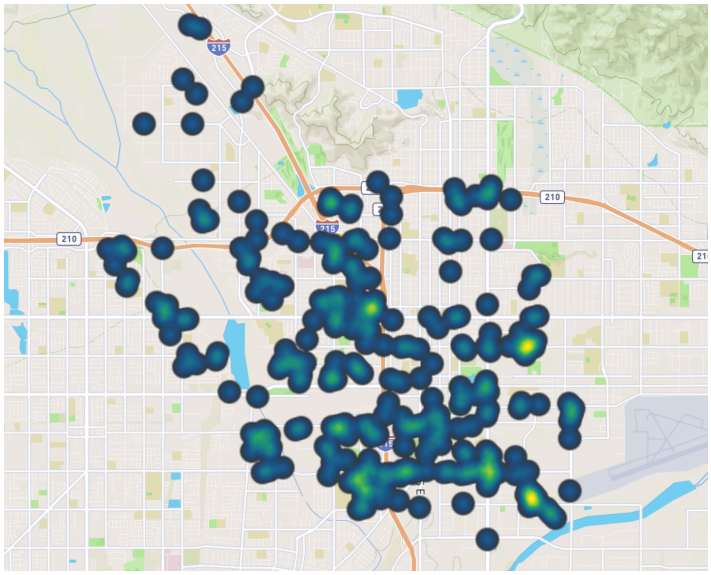

Figure 9.3: example of plotting an enhancement-based dataset (TVOC) as a heatmap. This and other visualizations can also be shown in tabulated form. This visualization is done with Data Level L3.

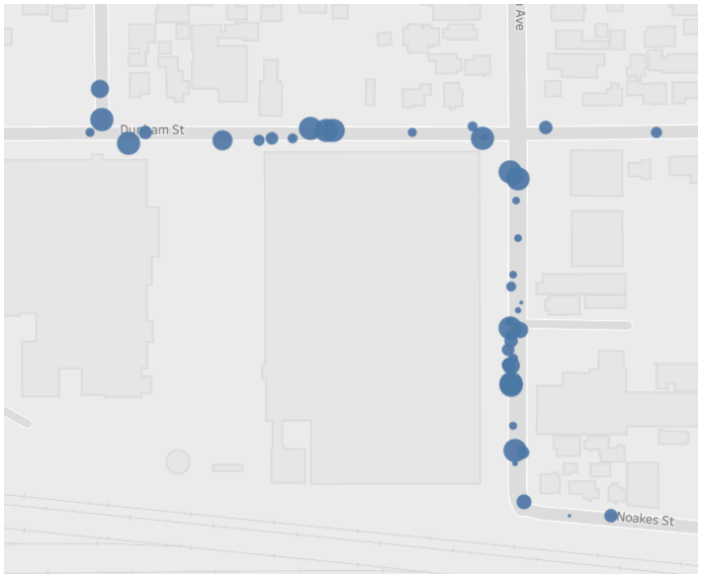

Figure 9.4: example of plotting a dataset using a point visualization (TVOC enhancement events). This visualization is done with Data Level L3.

9.3. Dataset visualization approach

To make these datasets useful to communities, they will need to be visualized in an accessible way. CARB has selected ESRI StoryMaps as the visualization platform. There may be as many as 64 individual StoryMaps depending on PEG and CAMP decisions. To make this feasible given time and resource constraints, Aclima proposes to adopt a streamlined, template-based approach for constructing these visualizations. This will ensure datasets are only presented in appropriate formats and a common experience is available to all communities, while retaining some ability to customize to community, PEG, and CARB needs. Each visualization will be implemented using StoryMaps tools, ideally on CARB-hosted accounts so that ownership can be seamlessly transferred.

Aclima proposes two main components available for inclusion in any given community's StoryMap:

1. A common StoryMap template available to all communities:
    a. Explains the project background and scientific context through narrative elements
    b. Visualizes all ambient concentration-based datasets using different StoryMap spatial elements
    c. Provides a standard structure for the StoryMap flow
2. An optional data layer for visualizing enhancement-based datasets:
    a. Visualizes data sourced from either Aclima sensors or PML instruments
    b. Available for integration with any existing spatial elements, as a separate spatial element, or in tabulated form

These components are described in more detail in the following sections.

9.3.1. Common StoryMap template

Aclima will provide a template, common to all StoryMap instances for all communities. This template will comprise:

1. **Narrative modules** which introduce the project, provide background information, share scientific context (e.g. information on monitored pollutants, mobile monitoring sampling design, Q&A, etc), and can be interspersed with specific spatial or tabulated visualizations to provide more detailed information
2. A **spatial gradient overlay** (spatially aggregated across H3 hexagons and temporally aggregated across the entire monitoring period), above a standard basemap or basemaps. This will primarily provide access to ambient concentration-based datasets. It may be available as a standalone map element or as a base layer in more complex maps.
3. A **point overlay** of L2a data for all pollutants detected by Aclima sensors, i.e. each individual measurement as a single geographically-located point. This will typically be overlain on a spatial gradient layer, but is also available as a standalone map element.

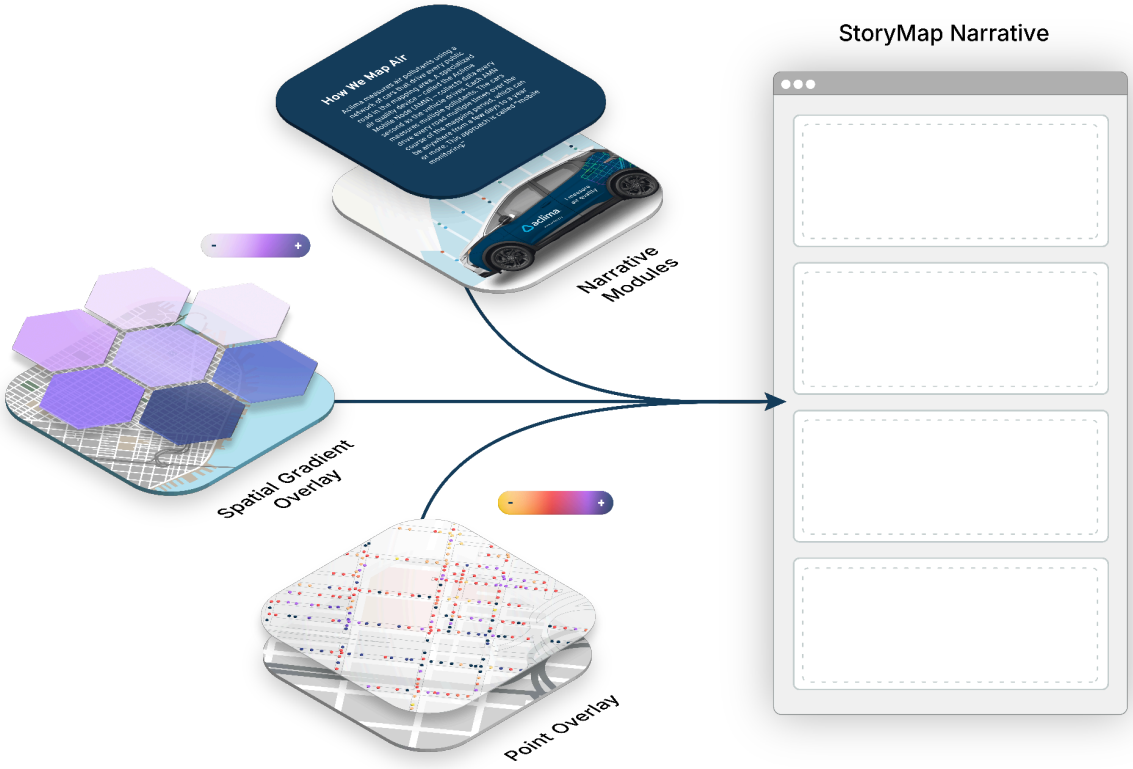

Figure 9.5: overview of StoryMap template design. Narrative modules, spatial gradient overlay, and point overlays are interwoven and combined to construct a linear StoryMap narrative.

9.3.2. Enhancement-based data layer, derived from Aclima sensors

Aclima will plot enhancement-based datasets ("indicator" data products, such as black carbon or $PM_{2.5}$; see Table 9.1 above) for pollutants detected using Aclima Mobile Platforms (AMP) via a broad area driving sample design, as an overlay on the base map. The data will only be visible at higher zoom levels, and may be presented as point data, road segment concentrations, or a 3D peak visualization, depending on the pollutant and PEG/CAMP input. Narrative content will accompany this overlay to provide context and guidance.

This overlay will be available to all visualizations for all communities, if enhancement-based indicators are detected.

9.3.3. Enhancement-based data layer, derived from PML instruments

Aclima will plot enhancement-based datasets for pollutants detected using PML reference-grade instruments via a range of targeted area sampling designs, as an overlay on the base map. These data will only be visible at higher zoom levels, and may be presented as point data, road segment concentrations, or a 3D peak visualization, depending on the pollutant and PEG/CAMP input. Narrative content will accompany this overlay to provide context and guidance, tailored to each specific community.

Because PML data collection uses targeted area driving sampling designs that draw on a variety of target sources (e.g. data from the Aclima fleet; known sources; community concerns, etc) and with only limited capacity for campaigns run with these designs, overlays of this type will be unique to specific communities, will not be available to all communities, and may not be evenly distributed among communities. Aclima anticipates 15-20 targeted sampling campaigns that may be distributed anywhere in the project area, and therefore 15-20 different overlays of this type.

Templates for PML-derived dataset overlays

For data gathered by PMLs *only*, Aclima and the PMLs have developed a series of proposed Sampling, Analysis, and Presentation Plan (SAPP) templates which will both guide PML data collection and influence how PML-sourced StoryMap content is visualized. A single SAPP:

- Captures a typical community concern
- Frames this as a scientific question or monitoring objective
- Identifies implicated pollutants and source types

- Suggests appropriate sampling design
- Identifies analyses to be run on collected data
- Suggests appropriate statistical and/or visual outputs
- Indicates possible regulatory actions for reference only

An initial set of community concerns being considered in draft form is summarized in Table 9.2, and draft content for a specific concern is described in Table 9.3. SAPPs will be finalized through further PEG, CAMP, and CARB input.

Table 9.2: summary of community concerns in an early draft list of Sampling, Analysis, and Presentation Plan (SAPP) templates for supporting PEG and CAMP discussions, and ultimately guiding presentation of PML-sourced (and specific pollutants sourced from targeted area driving by AMPs) in StoryMaps

SAPP	Community concern
1	Are the nearby refineries impacting my neighborhood?
2	There is a fuel storage farm in my community. Should I be concerned about benzene or other pollutants in my neighborhood?
3	Are the facilities emitting pollutants other than what their permit allows?
4	When are sources impacting my neighborhood?
5	My neighborhood was redlined. Is it still disproportionately impacted?
6	Are freeways impacting my neighborhood, especially on and off ramps?
7	There is a warehouse in my community where sterilized medical equipment is stored. Should I be worried about Ethylene Oxide (EtO)?
8	I can see dust being blown away from a cement plant in my neighborhood. Is there anything toxic in the dust I should be worried about?
9	There are so many different potential pollution sources in my neighborhood. How can I tell which ones are the most hazardous and should be prioritized?
10	How is my community impacted by truck traffic?

Table 9.3: draft content for SAPP #7, relating to Ethylene Oxide (EtO) concerns

Community concern	There is a warehouse in my community where sterilized medical equipment is stored. Should I be worried about Ethylene Oxide (EtO)?
Possible scientific questions	• Is there a persistence enhancement of EtO in vicinity of the warehouse? • How frequently is it found, and how far does it spread?

	• What is the background of EtO in absence of source? • What events correlate with detecting a plume of concern? • Where else in the community has similar concerns?
Implicated pollutant	EtO
Implicated source types	• Petrochemical facilities • Compressed gas suppliers • Sterilization facilities
Sampling design	• Deploy PML instruments (TILDAS, AROMA-ETO) • Drive in and out of plume • Stay stationary in plume • Drive transects upwind and downwind
Analyses	• Average and max EtO concentration in plume • Derive frequency of EtO detections • Compute average background EtO concentration in absence of plume • Ambient concentration estimates of EtO in downwind areas
Outputs	• Map of plume extent using 1 Hz data • Heatmaps around a facility • Location of source • Visualized frequency of detection • Wind direction and strength overlays • Histograms with acute and long-term limit overlays • Bar and whiskers plot for stationary or neighborhood averages
Regulatory actions	• Instigate a study or inspection on facility of concern

9.4. Data provisioning and StoryMaps ownership transfer

Aclima will develop all StoryMaps, and all supporting datasets and data products, per a final list developed through PEG, CAMP, and CARB consultation, within constraints of reasonableness and feasibility.

Aclima will make supporting datasets and data products available in a GCS bucket, for CARB access and transfer to Snowflake or CARB's Esri servers. StoryMap visualizations will be designed and developed by Aclima, either using dedicated Esri accounts provided by CARB, or on an Aclima-provided Esri account and then transferred to CARB ownership

and hosting. Aclima recommends that Aclima team members be given access to CARB's Esri account, to make this work more efficient and reduce post-contract risks.

Aclima will not set up or manage StoryMap hosting, during or after the contract period. All products will be live and maintainable beyond the end of the contract period (Aclima will have no further involvement after that time).

Aclima understands that CARB will not provide Aclima with any direct pathway to the AQview product, and is not requiring any specific AQview work by Aclima.