Contract No. 582-23-45978
Work Order No. 8
Tracking No. 2025-04
Task 6.2

Prepared for:

Texas Commission on Environmental Quality
12100 Park 35 Circle MC 164
Austin, TX 78753

Prepared by:

Ramboll
7250 Redwood Blvd., Suite 105
Novato, California 94945

July 18, 2025

Streamline Texas Point Source Processing of STARS Extract for Photochemical Modeling: SAS to R Migration

Final Report

PREPARED UNDER A CONTRACT FROM THE
TEXAS COMMISSION ON ENVIRONMENTAL QUALITY

The preparation of this document was financed through a contract from the State of Texas through the Texas Commission on Environmental Quality.
The content, findings, opinions and conclusions are the work of the author(s) and do not necessarily represent findings, opinions or conclusions of the TCEQ.

RAMBOLL

Streamline Texas Point Source Processing of STARS Extract for Photochemical Modeling: SAS to R Migration
Final Report

Ramboll
7250 Redwood Boulevard
Suite 105
Novato, CA 94945
USA

T +1 415 899 0700
https://ramboll.com

Contents

Table of Figures

Table of Tables

List of Acronyms and Abbreviations

AFS	AIRS Facility Subsystem
AIRS	Aerometric Information Retrieval System
CAMx	Comprehensive Air Quality with Extensions
CIN	Control device Identification Number
CONUS	Continental United States
DMS	Degrees, Minutes, and Seconds
EPA	Environmental Protection Agency
EPN	Emission Point Number
EPS3	Emissions Processing System version 3
FIN	Facility Identification Number
HECT	HRVOC Emissions Cap and Trade
HRVOC	Highly Reactive VOC
IQR	Interquartile Range
LCP	Lambert Conformal Projection
NAAQS	National Ambient Air Quality Standards
NOx	Nitrogen Oxides
OSD	Ozone Season Day
RPO	Regional Planning Organizations
SAROAD	Storage and Retrieval of Aerometric Data
SCC	Source Classification Code
SIC	Standard Industrial Classification
SIP	State Implementation Plan
STARS	State of Texas Air Reporting System
TCEQ	Texas Commission on Environmental Quality
U.S.	United States
UTM	Universal Transverse Mercator
VOC	Volatile Organic Compound

Project Summary

This project focused on updating a legacy processing system used by the Texas Commission on Environmental Quality (TCEQ) to process point source emissions data for input to air quality modeling. The original system was built using SAS, which is a specialized programming platform that requires costly licenses and specialized training. Additionally, since the system was developed over two decades ago, much institutional knowledge has been lost. To address these challenges, Ramboll migrated part of the processing system to R, a free and widely used programming language, to make the process more efficient and user-friendly. The new R scripts replicate key functions from the original system while making the process easier for TCEQ staff to use and maintain. This transition helps TCEQ save on licensing expenses and increases flexibility without compromising on functionality. Additional components of the system will be migrated to R in future phases of the project.

Executive Summary

The State of Texas Air Reporting System (STARS) database contains emissions data for point sources in Texas, and the TCEQ has a specialized methodology for processing this data for input to air quality models such as Comprehensive Air Model with extensions (CAMx). Currently, point source emissions processing is carried out using SAS, a proprietary software that requires licensing and specialized training.

This project focused on migrating part of the existing SAS-based processing system to the R programming language, which is an open-source platform that is already familiar to TCEQ's emissions modeling team. Ramboll reviewed the current SAS scripts and methodologies to identify areas for improvement and documented the findings in a design document. Following this review, we developed R scripts to read and process STARS database extracts, calculate emissions for specific time periods, and convert point source coordinates from latitude/longitude to Lambert Conformal Projection, making the data suitable for use in CAMx.

Ramboll also collaborated with TCEQ to compare the intermediate AIRS Facility Subsystem (AFS) files generated by the legacy SAS system and the new R-based scripts. Any discrepancies that arose were investigated, their root causes identified, and the R scripts were revised as needed. In addition, Ramboll developed a QA script to compare outputs generated by the SAS and R processing systems and provided it to TCEQ to support their review and verification efforts.

For this phase of the project, the speciation and Volatile Organic Compound (VOC) allocation components of the current system were not implemented in R. We recommend completing the full migration by implementing the remaining scripts in R and refining the design as needed in coordination with TCEQ.

1.0 Introduction

The point source emissions data submitted to the TCEQ is stored in the STARS database. The major point sources are encouraged to report individual VOC species directly from each source in addition to the total VOC emissions, other ozone precursors such as NOx, and pollutants with National Ambient Air Quality Standards (NAAQS). One key use of speciated VOC data is as input to regional scale photochemical models. The integration of individual VOC emissions improves speciation of VOC emissions in photochemical modeling that supports the State Implementation Plans (SIP).

The STARS database contains speciated and un-speciated VOC emissions data for each applicable point source in Texas. The TCEQ has developed a unique processing methodology to retain all reported speciated data and characterize the remaining portion of un-speciated data according to default EPA speciation profiles (Cantu, 2003; PES, 2001). Currently, point source emissions processing is carried out using SAS, a proprietary software that requires licensing and specialized training. The processing approach is implemented through a series of SAS scripts that generate point-specific VOC speciation profiles. There are also quality assurance and quality control (QA/QC) activities that are included in the SAS scripts such as removing non-VOC species from the modeling inventory before processing and splitting generic chemical mixtures reported by industry (such as "crude oil" or "gasoline") into component hydrocarbons using existing chemical profiles.

This project focused on migrating part of the existing SAS-based processing to the R programming language, which is an open-source platform that is already familiar to TCEQ's emissions modeling team. As part of this effort, Ramboll reviewed the current SAS scripts and methodologies to identify areas for improvement and documented the findings in a design document. We then developed R scripts to read and process STARS database extracts, calculate emissions for specific time periods, and convert point source coordinates from latitude/longitude to Lambert Conformal Projection. These steps prepare the data for use in the Emissions Processing System version 3 (EPS3) system, which generates CAMx-ready emissions input files. EPS3 converts emissions inventories into input files that are speciated and spatially and temporally allocated for CAMx. This phase did not include implementing speciation or VOC allocation components in R.

2.0 Design of the STARS Extract Processing Workflow

TCEQ provided details on the existing STARS extract workflow along with relevant sample data. Ramboll reviewed the associated SAS scripts and processing methodologies to identify areas for improvement. These findings were summarized in a design document. Throughout this process, Ramboll worked closely with TCEQ to ensure the proposed design aligned with the project's goals and operational needs. Following TCEQ's approval of the design, Ramboll initiated development of the new processing system using R, building upon the structure of the existing STARS extract framework.

The R-based system follows the structure of the original STARS extract workflow but incorporates several key improvements, as described in the following section. During the design phase, Ramboll evaluated the current workflow and methodologies to identify opportunities for enhancement. Based on this assessment, we developed a design document that outlined the requirements for the R implementation. This document included an overview of the current system, detailing input and output datasets, calculation steps, and processing methods, and proposed updates to improve overall functionality. It also introduced a more streamlined QA/QC approach designed to enhance both efficiency and reliability.

The following sections provide an overview of the current system and highlight key improvements introduced in the new R-based implementation.

2.1 Current System Overview

2.1.1 System Description

This section describes the current processing performed using SAS and provides an overview of the overall workflow, as illustrated in Figure 2-1. The numbered steps in the flow diagram correspond to the detailed steps described in Section 2.1.2. Below is a high-level summary of the current processing system:

1. Data Ingestion and Preprocessing

In the initial stage, the STARS extract file is divided into several components to enable detailed analysis. These components include account data, facility data (FIN), emission point data (EPN), control equipment data (CIN), and emissions history data. This stage also includes the creations of a cross-reference file that maps Storage and Retrieval of Aerometric Data (SAROAD) codes to TCEQ contamination codes used in the STARS database along with the corresponding species names.

2. Data Processing and Transformation

At this stage, emissions data are loaded, and pollutant classes (e.g., VOC, NOx, PM10) are defined. EPA default VOC speciation profiles are imported and assigned to sources with un-speciated VOC. Emissions are then aggregated by point source for further processing. Multiple datasets are combined to create a comprehensive emissions dataset that includes calculated emissions for different time periods. Additional refinements include:

- Supplementing EPA default speciation profiles with source-specific profiles from CARB and the DFW region. Specifically, the current SAS scripts replace profile 1003 (surface coating operations – solvent-based paint) with D404 (DFW) and profile 0007 (natural gas turbine) with 0719 (CARB). These replacements are made when the scripts detect profiles that consist of only one compound after removing non-VOC and non-reactive components. For example, profile 0007 is included in EPA's default speciation cross-reference file but is replaced by CARB's 0719 profile since it contains 100% formaldehyde (a highly reactive VOC) after methane (a non-VOC) is removed.

- Removing non-VOC compounds. Ethane and acetone are retained, as they are explicitly represented in CB6/CB7 chemical mechanisms.
- Modifying SCC-level speciation profiles to exclude highly reactive VOCs (HRVOCs) for sources subject to the HRVOC Emissions Cap and Trade (HECT) program in Harris County (Thomas et al., 2008).
- Recalculating flare stack parameters (e.g., adjusting stack diameter and exit velocity based on VOC emissions) to better reflect actual operating conditions.

3. Spatial Data Processing

Spatial coordinates are standardized to a consistent format. Latitude/longitude values reported in degrees, minutes, and seconds (DMS), as well as Universal Transverse Mercator (UTM) coordinates, are converted into decimal degrees. QA checks are performed to detect and correct potential location errors.

4. Quality Assurance and Validation

A series of QA steps is performed sequentially to ensure accuracy and reliability. These include validating Aerometric Information Retrieval System (AIRS) codes and analyzing emissions data to identify outliers and inconsistencies. A dedicated QA script also evaluates stack heights data in the STARS dataset by comparing reported values to statistical norms for each Standard Industrial Classification (SIC) and Source Classification Code (SCC) pair. Outliers are flagged for further review and reporting.

5. Emissions Data Structuring and Summary Reporting

Emissions data are organized to support analysis and air quality modeling. The system relies on reported ozone season day (OSD) emissions. One SAS script processes the merged emissions inventory file by splitting it into annual datasets and generating a log to track the processing steps. This improves data organization and enables trend analysis over time. Another SAS script aggregates emissions data across key categories (e.g., pollutant type, account, and owner) to generate a summary report. This report highlights major emission sources and provides a detailed breakdown by geographic area and permit type.

6. Speciation and VOC Allocation

Speciated emissions are generated by creating a point-specific VOC profile for each unique "path" in STARS, where a path is defined as a combination of a process unit and an emission point. Additional steps include:

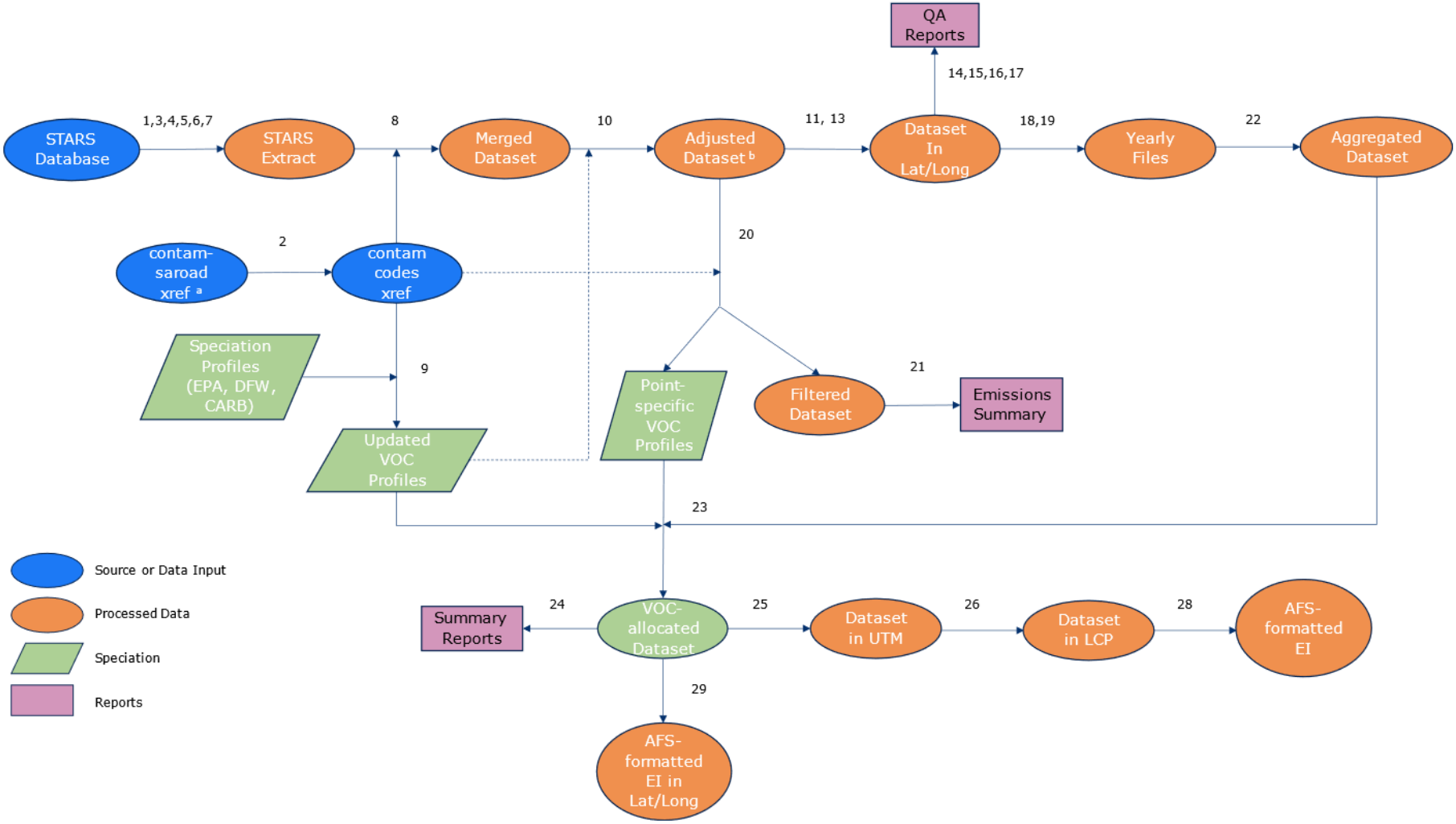
- Generating speciation cross-reference files for downstream processing
- Replacing generic VOC mixtures (e.g., crude oil, gasoline, naphtha, Stoddard solvent, and "refinery") with detailed speciation profiles
- Summarizing emissions by SAROAD code before and after speciation adjustments
- Aggregating emissions by point source and merging them back into the full dataset
- Performing QA on the final aggregated file

VOC emissions are then allocated using both point-specific and EPA SCC-level profiles. A summary report of the allocated emissions is generated by geographic area and permit type.

7. Final Output Generation

In the final step, emissions data are converted to UTM coordinates and then transformed into Lambert Conformal Projection (LCP) coordinates for photochemical grid modeling. Annual emissions data are

calculated and categorized into ozone season and non-ozone season day (non-OSD) periods. The final emissions outputs are prepared in Aerometric File System (AFS) format for EPS3 processing.



[a] The current input file is contam_saroad_xref.map.30Apr2024

[b] Including HECT adjustment, flare corrections, and SCC updates

Figure 2-1. Flow diagram of the current SAS workflow.

2.1.2 Components of the Current System

The table below provides a detailed description of the main components in the current STARS extract processing system.

Table 2-1. Overview of Processing Steps.

Step No.	Script/Program	Input	Purpose	Output
1	breakout.pl	STARS Database	Splits the STARS Extract file into account, fin, epn, cin, and history files.	account_YYYY.txt fin_YYYY.txt epn_YYYY.txt cin_YYYY.txt history_YYYY.txt
2	contam_xref	contam_saroad_xref.map: the current input file is contam_saroad_xref.map.30Apr 2024	This SAS script generates a cross-reference (xref) file for SAROAD (System for Automated Retrieval of Operating Data) and TNRCC (Texas Natural Resource Conservation Commission) contamination codes and names. The script reads raw data, processes and cleans it, assigns species names based on contamination codes, and saves the final dataset.	stars.contam_xref
3	account	account_YYYY.txt	This SAS script processes emissions account data and SIC (Standard Industrial Classification) data for a given year. It reads input data, organizes it into various datasets by categories, merges the datasets into a comprehensive account dataset, and performs quality assurance actions such as removing duplicate records. The final dataset is saved for further use and analysis, and additional SIC-related data is prepared for storage. Loads account and SIC data, identifying pathways by RN. Creates multiple datasets (lastei, plant, optype, sic, etc.). Sorts all datasets by rn before merging. Merges all datasets into a final account dataset. Handles latitude/longitude (in DMS format): sets values to missing if they are zero. Removes records where the account number starts with "9." Assigns counties to regions (e.g., ANDERSON → ETX, DALLAS → DFW). Removes duplicate records and saves the final dataset. Creates an SIC table linking SIC codes to their descriptions.	stars.account_YYYY stars.sic_YYYY
4	fin	fin_YYYY.txt	This SAS script is designed to load and process data related to FIN (Facility Identification Number) and SCC (Source Classification Code) for a given year. The script reads input data, organizes it into various datasets, processes the data by extracting and sorting necessary attributes, merges the datasets into comprehensive FIN and SCC datasets, and performs quality assurance actions such as removing duplicate records.	stars.fin_YYYY stars.scc_YYYY

Step No.	Script/Program	Input	Purpose	Output
			Converts profile types to roof codes (ED, EF, FX, IF) Removes commas from designcap_mmbtu_hr The result is two datasets: FIN and SCC, with documentation of the process and any duplicate records removed.	
5	epn	epn_YYYY.txt	This SAS script processes EPN (Emission Point Number) data for a given year. It reads input data, organizes it into various datasets, processes the data by extracting and sorting necessary attributes, merges the datasets into a comprehensive EPN dataset, and performs quality assurance actions such as removing duplicate records. The final dataset is saved for further use and analysis. Converts measurements to metric: Feet to meters (height, diameter, length, width) Fahrenheit to Kelvin (temperature) Feet/second to meters/second (velocity) Meters to kilometers (UTM coordinates)	stars.epn_YYYY epn.dup_rec_count epn.contents
6	cin	cin_YYYY.txt	This code processes control equipment data from a text file, creating a structured dataset that includes abatement device information, efficiency metrics for various pollutants, and connections to facility (FIN) and emission point (EPN) data.	stars.cin_YYYY cin.dup_rec_count cin.contents
7	history	history_YYYY.txt	This code processes emissions history data, assigns pollutant classes, handles PM10 speciation, and generates multiple output datasets including summaries and CSV files for reporting. Write speciated PM data into a separate file.	stars.history_YYYY history.summary_YYYY.txt stars.PM10speciated_YYYY.csv
8	merge_stars_from_extract	stars.account_YYYY stars.fin_YYYY stars.epn_YYYY stars.cin_YYYY stars.history_pm_YYYY stars.contam_xref	This code combines multiple STARS datasets (ACCOUNT, FIN, EPN, CIN, HISTORY, CONTAM_XREF) into a single merged dataset, calculates "best" emissions (best_ems) for ozone season, introduces a new variable for non-ozone season emissions (other_ems), and prepares data for full-year modeling with a refined threshold. Computes best_ems based on year: Pre-2011: 3-month ozone season (June-August) 2011+: 5-month ozone season (May-September) Calculates other_ems for non-ozone season months.	merged_stars_extract_YYYY_{version} merge_dups_emis_summary_YYYY_{version}.lst
9	profiles	EPA default profiles DFW D404 profile CARB 0719 profile	This SAS script takes EPA default VOC profiles, enhances them with DFW (D404) and CARB (0719) data, excludes non-VOC compounds (except ethane and acetone), normalizes the weight fractions, and links SCCs to profiles. It generates two	Speciation profiles file: prof.emscvt.modified_4VOC_epa.pt

Step No.	Script/Program	Input	Purpose	Output
		Compound database stars.contam_xref	output files: a speciation profile file for EMSCVT and an SCC-to-profile cross-reference file for CHMSPL. Reassigns specific profiles: 1003 → D404 (Dallas), 0007 → 0719 (CARB). Identifies profiles dominated by one compound (weight fraction ≥ 0.95) for review.	Speciation cross-reference file: xref.voc.eps3f.modified_4VOC_epa.pt stars.voc_tmp_profs stars.voc_tmp_xref
10	merge_adjustments_for_model	merged_stars_extract_YYYY_{version} prof.emscvt.modified_4VOC_epa.pt xref.voc.eps3f.modified_4VOC_epa.pt stars.voc_tmp_profs stars.voc_tmp_xref	This SAS script adjusts a STARS extract dataset (merged_stars_extract_YYYY_v#) for air quality modeling by: HECT Adjustments: Modifies SCC profiles to exclude HRVOCs for HECT sources in Harris County, creating unique identifiers (e.g., 0719H). Ethane/Acetone Addition: Generates emissions records for ethane and acetone using EPA default profiles or STARS data, ensuring CB06 compatibility. Flare Corrections: Recalculates stack diameter and velocity for flares based on VOC emissions. SCC Updates: Adds new SCCs, reassigns retired ones, and logs unmatched SCCs with profile 0000. The final output, merged_stars_YYYY_v#, integrates these changes, alongside updated profile (prof.emscvt.modified_epa.pt) and cross-reference (xref.voc.eps3f.modified_epa.pt) files, ready for tools like EMSCVT and CHMSPL. Identify and output SCCs that need default profile; Flag point sources that qualify for the HECT program and adjust their profiles; Creates emissions records for ethane and acetone based on default compositions if they are not already reported; Correct flare stack parameters; Write the final adjusted dataset	merged_stars_YYYY_{version} SCC_list_assignedEPAdefault_YYYY_{version}.csv xref.voc.eps3f.modified_epa.pt prof.emscvt.modified_epa.pt
11	cvt_coord	merged_stars_YYYY_v#	The SAS script converts latitude/longitude coordinates from Degrees, Minutes, and Seconds (DMS) format and Universal Transverse Mercator (UTM) coordinates into decimal degrees (DD) format. The final dataset is designed for modeling and ensures all coordinates are consistently formatted for further use.	stars_cc_no_qa_YYYY_v#
12	write_stars	merged_stars_YYYY_v#	This SAS script writes out emissions data from the "merged_stars" dataset for the development of VOC profiles. The script filters and formats the data appropriately, ensuring	Output contains VOC only

Step No.	Script/Program	Input	Purpose	Output
			only relevant emissions data is written to the output file. The script uses a new method to delete small emissions based on annual, EE_SMSS, and O3_EMS values instead of Best_EMS.	
13	qa_coord	stars_cc_no_qa_YYYY_v#	This SAS script performs QA on latitude and longitude coordinates in the emissions dataset. It identifies potential issues with the coordinates (e.g., missing, large deviations from the median), generates reports on the problematic data points, and corrects such problems by assigning appropriate default values (plant median).	stars_cc_YYYY._v#
14	qa_airs	stars_cc_YYYY_v#	The SAS script performs QA on AIRS (Aerometric Information Retrieval System) codes in the emissions dataset. It identifies and reports any errors related to missing or incorrect AIRS codes, specifically focusing on emissions data.	None
15	qa_ems_by_year	stars_cc_YYYY_v#	This SAS script performs QA on emission data by analyzing data reported over multiple years. It identifies suspicious emission data points that deviate significantly from the mean by more than a specified number of standard deviations (n). The script generates reports to highlight these outliers.	A report of outliers
16	qa_ems_by_sic_scc	stars_cc_YYYY_v#	This SAS script performs QA on emissions data by analyzing SIC/SCC pairs. It identifies suspicious emissions data points that deviate significantly from the mean by more than a specified number of standard deviations (n). The script generates reports to highlight these outliers.	A report of outliers
17	qa_stacks	stars_cc_YYYY_v#	This script performs QA on stack heights in a STARS dataset by comparing them to statistical norms (mean ± n * standard deviation) for each SIC (Standard Industrial Classification) and SCC (Source Classification Code) pair, then reports outliers.	A report of outliers
18	yearly	stars_re_YYYY_v#	The SAS script splits a merged Emissions Inventory (EI) file into yearly files and generates a log of the process. The script processes emissions data by filtering records for a specific year and prepares the dataset for further use and analysis. Additionally, it generates a report containing the dataset's structure and a preview of its contents.	stars_YYYY_v#
19	summary	merged_stars_YYYY_v#	This SAS script generates a summary report of emissions data by area and permit type. The script processes the emissions inventory data, aggregates the emissions by various categories (e.g., pollutant, account, owner), and generates a report highlighting the emissions for each area. The report focuses on significant accounts and summarizes the emissions data.	stars_sum_account_YYYY._v# A summary report of emissions

Step No.	Script/Program	Input	Purpose	Output
20	pt_profs	merged_stars_YYYY_v#	This SAS script processes speciated emissions data to create point-specific profiles. It refines, applies default profiles, and aggregates speciated emissions for input into emissions modeling tools such as EPS3. The script handles various tasks, including filtering emissions data, speciation of general mixtures, and creating final profiles and cross-reference (xref) files.	pt_profiles_YYYY._v#.&update prof.emscvt.stars_YYYY_v#..pt xref.voc.eps3f.stars_YYYY_v#..pt
21	sum_by_species	filtered_ei_YYYY_v#.&update pt_profiles_YYYY_v#.&update	This SAS script sums emissions data by SAROAD codes. It processes emissions data before and after profile processing, generates summary statistics, and writes the summarized data to output files. The script handles tasks such as summing emissions, sorting by specific variables, and generating output reports.	pt_profs_in_filt_emis_by_species_YYYY_v#.&update pt_profs_out_emis_by_species_YYYY_v#.&update pt_profs_in_filt_emis_no_airs_YYYY_v#.&update
22	aggregate	stars_YYYY_v#	This SAS script processes emissions data to aggregate VOC and NOx species (and other pollutants) by point source. It sums emissions for each point source, merges aggregated data back into the dataset, and produces a final aggregated dataset suitable for further analysis and reporting. It also performs QA checks to ensure the integrity of the aggregated data.	stars_agg_YYYY_v# agg_YYYY_v#.contents aggregated_YYYY_v#.log
23	allocate_voc	stars_agg_YYYY._v# pt_profiles_YYYY_v#.&update voc_epa_xref voc_epa_profs	This SAS script allocates VOC emissions from profiles. It processes the aggregated emissions inventory data, applies point-specific and EPA SCC profiles, allocates emissions accordingly, and generates the final allocated dataset. The script handles various tasks such as summing emissions, merging with profiles, and writing results to output datasets.	stars_alloc_YYYY_v#.&update.
24	allocated_summary	stars_alloc_YYYY_v#.&update	This SAS script generates a summary report of emissions data by area and permit type. It processes the allocated emissions inventory data, aggregates emissions by various categories (e.g., pollutant, account, owner), and generates detailed reports highlighting significant emissions for each area and county. The reports are designed to provide insights into emissions distributions and facilitate analysis.	alloc_summary_YYYY_v#.lst
25	write_utm_afs	stars_alloc_YYYY_.&update	This script processes emissions data from a STARS dataset and writes it out in AFS (Air Facility System) format compatible with EPS3 PREPNT program, alongside creating a corresponding SAS dataset.	afs.stars.&pollutan..&emissions..YYYY._v#.&update..&coord afs_&pollutan._&emissions._YYYY_v#.&update

Step No.	Script/Program	Input	Purpose	Output
26	make_RPOlcp_file_part_1 make_RPOlcp_file_part_2	afs.stars.&pollutan..&em_type..YYYY._v#.&update..UTM RPOlcp_for_13.txt, RPOlcp_for_14.txt, RPOlcp_for_15.txt (transformed coordinates).	This script processes an AFS-formatted file containing UTM coordinates, splits it by UTM zone, and writes the data into separate text files for further conversion to Lambert Conformal Projection used by Regional Planning Organizations (RPOs). This script merges RPOlcp coordinates with UTM AFS data from Part 1, adjusts units to kilometers, and writes a new AFS file with RPOlcp coordinates. It's the second step in converting emissions data for regional modeling, relying on an external coordinate transformation.	utm_13.txt, utm_14.txt, utm_15.txt afs.stars.VOC.spec_ems.99_v2a.RPOlcp
27	sum_afs_saroad_pollutan		This Perl script sums emissions for a specified pollutant over a given area in Texas from an AFS file.	
28	write_annual_emissions_afs	afs.stars.All.agg_ems.YYYY._v#.&update..RPOlcp (OSD daily emissions) afs.stars.All.agg_other_ems.YYYY._v#.&update..RPOlcp (non-OSD daily emissions).	This script processes emissions data from STARS AFS files, converts daily emissions into annual tons per year (tpy), and splits them into three AFS files: total annual emissions, ozone season day (OSD) emissions (May–Sep), and non-OSD emissions (remaining months).	afs.stars.All.ems_annual_tpy_v#.&update..RPOlcp (total annual emissions). afs.stars.All.OSD_ems_tpy_v#.&update..RPOlcp (OSD, May–Sep). afs.stars.All.non_OSD_ems_tpy_v#.&update..RPOlcp (non-OSD).
29	write_afs	stars.stars_alloc_YYYY_v#.&update	This script processes emissions data from a STARS dataset and outputs it in AFS format for EPS3 PREPNT program, alongside creating a corresponding SAS dataset.	afs.stars.VOC.spec_ems.2000_v3a.LatLong

2.1.3 Limitations of the Current System

As noted earlier, SAS is a proprietary software that requires licensing and specialized training. The original system was developed more than two decades ago and has evolved incrementally over time. While TCEQ staff maintain a solid understanding of the system, some institutional knowledge about its original development and rationale for certain processes has become less accessible over time. Given the system's age and complexity, a comprehensive review was still valuable to identify opportunities for streamlining, improving documentation, and enhancing long-term maintainability.

2.2 Key Improvements in the New System

2.2.1 Selection of the New Platform

We chose to implement the new system using R because it's an open-source language that's widely used for data analysis and already familiar to TCEQ's emissions modeling team.

2.2.2 Enhancements to the Current Processing System

After a thorough review of the existing workflow and consultations with TCEQ, we identified several updates to enhance the processing system. The key changes include:

- The new system will process one year of emissions data per run, as TCEQ indicated that multi-year processing is unnecessary. Consequently, Steps 15 and 18 in Table 2-1 will be eliminated.
- Hard-coded cross-references in the scripts, such as county-to-region assignments and species name updates, will be replaced with external input files. This will improve adaptability, simplify updates, and make it easier to manage.
- The creation of the speciated PM10 file in Step 7 will be removed, as it is no longer required.
- Location coordinate conversions are currently repeated across multiple scripts. Since TCEQ confirmed that direct conversion from latitude/longitude to LCP coordinates is sufficient, the extra UTM conversion step will be removed.
- Steps 25 and 29 in Table 2-1 serve nearly the same purpose. Step 25 generates AFS-formatted files in UTM coordinates, while Step 29 creates them in latitude/longitude format. Since the lat/long AFS files are unnecessary, Step 29 will be removed.
- The QA process will be streamlined to eliminate redundant steps. More details on these improvements are provided in the next section.

2.2.3 New Streamlined QA Approach

Ramboll proposed streamlining the QA process, which is currently performed using five separate QA scripts:

- 13.qa_coord
- 14.qa_airs
- 15.qa_ems_by_year
- 16.qa_ems_by_sic_scc
- 17.qa_stacks

As mentioned earlier, since the new system will process one year of data per run, Step 15 (qa_ems_by_year) was removed from the system. Instead, it will be conducted separately as an independent check to compare emissions across multiple years.

Additionally, Steps 13, 14, 16, and 17, which currently performed QA checks on location coordinates, AIRS codes, emissions data, and stack heights, were consolidated into a single, streamlined QA script.

This new script also includes checks for additional key stack parameters such as diameter, temperature, and exit velocity.

For emissions data and stack height, the earlier QA scripts computed the mean and standard deviation for each SIC/SCC pair and flag outliers. While this method is simple and widely used, it has limitations, particularly for skewed or non-normally distributed data. The standard deviation method assumes a normal distribution, which may not always be the case for emissions data. To address this, the new QA script adopts a more robust alternative: the Interquartile Range (IQR) method. IQR identifies outliers based on the spread of the middle 50% of the data and is less affected by extreme values, making it more suitable for datasets with heavy tails or skewed distributions. The IQR is calculated as the difference between the 75^{th} percentile (Q3) and the 25^{th} percentile (Q1). Outliers are defined as values falling below Q1 - 1.5 × IQR or above Q3 + 1.5 × IQR.

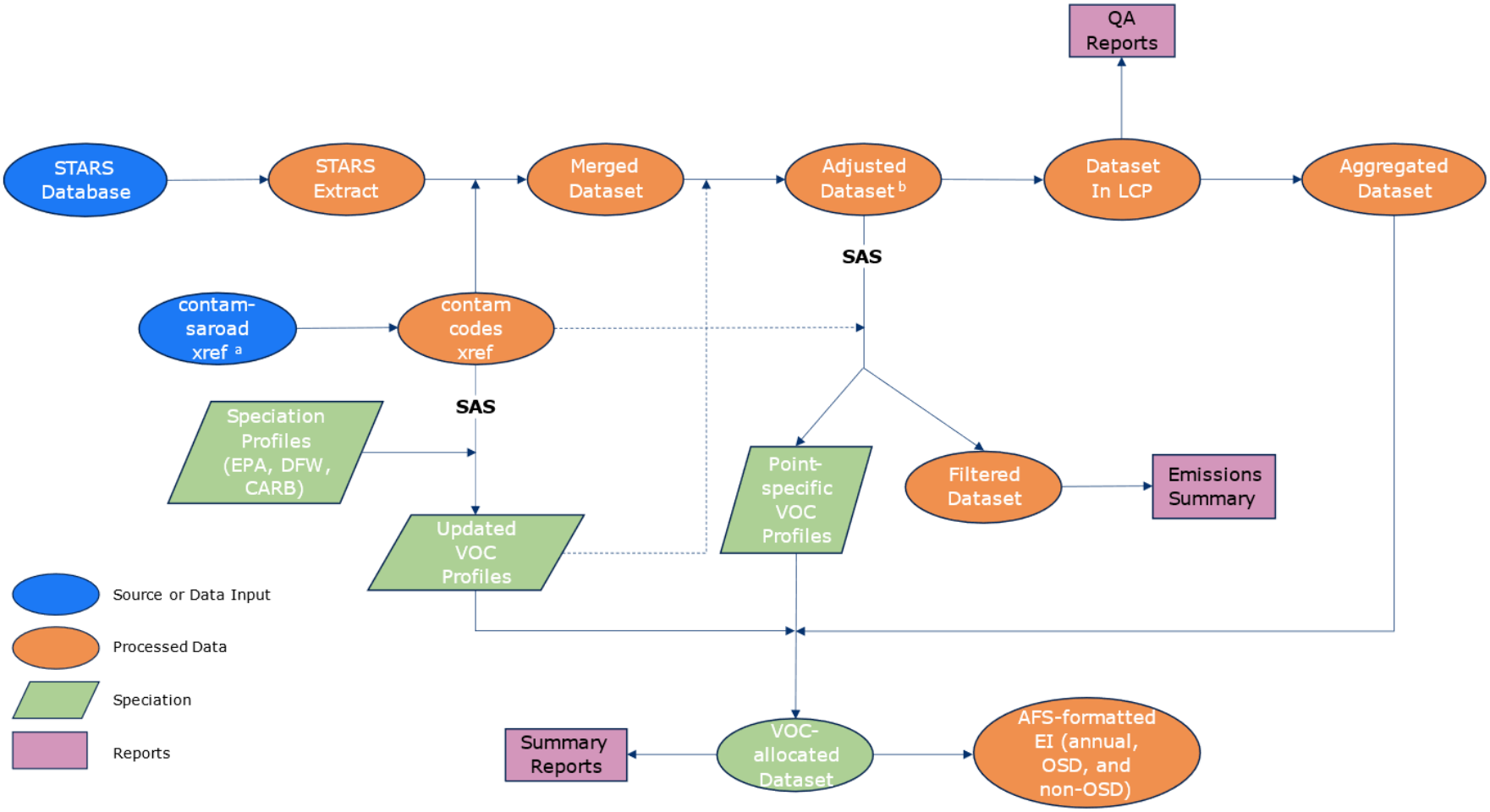
This approach helps ensure that only truly anomalous values are flagged, reducing false positives due to natural variations in emissions levels. To further improve reliability, a minimum data threshold is applied before computing outlier metrics. SIC/SCC pairs with very few data points are flagged for manual review, such as comparison with historical emissions data.

2.2.4 Overview of Key Updates

Table 2-2 provides a consolidated list of the proposed updates and script modifications, highlighting key changes. Figure 2-2 presents a flow diagram of the updated workflow incorporating these changes.

Table 2-2. Summary of Updates and Script Modifications.

Step No.	Script/Program	Proposed Updates
1	breakout.pl	No changes required; it will remain in Perl and does not need to be converted to R.
7	history	Creation of the speciated PM10 file will be removed.
11	cvt_coord	Convert coordinates directly to LCP.
12	write_stars	Removed in the new workflow.
13	qa_coord	Integrate into a new QA script.
14	qa_airs	Integrate into a new QA script.
15	qa_ems_by_year	Removed in the new workflow.
16	qa_ems_by_sic_scc	Integrate into a new QA script.
17	qa_stacks	Integrate into a new QA script.
18	yearly	Removed in the new workflow.
26	make_RPOlcp_file_part_1 make_RPOlcp_file_part_2	Removed; coordinates conversion will be performed in Step 11.
27	sum_afs_saroad_pollutan.pl	No changes required; it will remain in Perl and does not need to be converted to R.
29	write_afs	Removed in the new workflow.



[a] The current input file is contam_saroad_xref.map.30Apr2024

[b] Including HECT adjustment, flare corrections, and SCC updates

Figure 2-2. Flow diagram of the new workflow.

3.0 R Implementation

The first stage of point source processing includes all steps required to process the STARS extract and prepare the data in a format suitable for downstream emissions processing. Following approval of the design document, Ramboll developed R scripts to carry out STARS point source data processing beginning with this stage. The R scripts were designed to closely follow the structure and logic of the existing STARS extract framework. These scripts organize emissions and supporting data, calculate emissions for specified time periods, and convert latitude/longitude coordinates to CAMx Cartesian coordinates (LCP). The resulting output is generated in AFS file format.

Table 3-1 provides a detailed description of the processing steps implemented in R for the STARS extract point source data. For this phase of the project, the speciation and VOC allocation components of the current system were not included in the R implementation.

Ramboll followed good programming practices by incorporating metadata to describe the purpose of each script and by adding comments throughout the code to support future updates and maintenance. The R scripts generate AFS files and other key outputs required for the subsequent steps in point source base case emissions processing. Ramboll also supported the TCEQ Project Manager in running the R scripts within TCEQ's computing environment.

Table 3-1. STARS Extract Point Source Processing Steps Implemented in R.

Script No.	Script/Program	Input	Purpose	Output
1	contam_xref	contam_saroad_xref.map: the current input file is contam_saroad_xref.map.30Apr 2024	This script generates a cross-reference (xref) file for SAROAD (System for Automated Retrieval of Operating Data) and TNRCC (Texas Natural Resource Conservation Commission) contamination codes and names. The script reads raw data, processes and cleans it, assigns species names based on contamination codes, and saves the final dataset.	contam_xref.rds
2	account	account_YYYY.txt	This script processes emissions account data and SIC (Standard Industrial Classification) data for a given year. It reads input data, organizes it into various datasets by categories, merges the datasets into a comprehensive account dataset, and performs quality assurance actions such as removing duplicate records. The final dataset is saved for further use and analysis, and additional SIC-related data is prepared for storage. Load account and SIC data, identifying pathways by RN. Creates multiple datasets (lastei, plant, optype, sic, etc.). Sorts all datasets by rn before merging. Merges all datasets into a final account dataset. Handles latitude/longitude (in DMS format): sets values to missing if they are zero. Removes records where the account number starts with "9." Assigns counties to regions (e.g., ANDERSON → ETX, DALLAS → DFW). Removes duplicate records and saves the final dataset. Creates an SIC table linking SIC codes to their descriptions.	account_YYYY.rds sic_YYYY.rds
3	fin	fin_YYYY.txt	This script is designed to load and process data related to FIN (Facility Identification Number) and SCC (Source Classification Code) for a given year. The script reads input data, organizes it into various datasets, processes the data by extracting and sorting necessary attributes, merges the datasets into comprehensive FIN and SCC datasets, and performs quality assurance actions such as removing duplicate records. Converts profile types to roof codes (ED, EF, FX, IF) Removes commas from designcap_mmbtu_hr	fin_YYYY.rds scc_YYYY.rds

Script No.	Script/Program	Input	Purpose	Output
			The result is two datasets: FIN and SCC, with documentation of the process and any duplicate records removed.	
4	epn	epn_YYYY.txt	This script processes EPN (Emission Point Number) data for a given year. It reads input data, organizes it into various datasets, processes the data by extracting and sorting necessary attributes, merges the datasets into a comprehensive EPN dataset, and performs quality assurance actions such as removing duplicate records. The final dataset is saved for further use and analysis. Converts measurements to metric: Feet to meters (height, diameter, length, width) Fahrenheit to Kelvin (temperature) Feet/second to meters/second (velocity) Meters to kilometers (UTM coordinates)	epn_YYYY.rds
5	cin	cin_YYYY.txt	This code processes control equipment data from a text file, creating a structured dataset that includes abatement device information, efficiency metrics for various pollutants, and connections to facility (FIN) and emission point (EPN) data.	cin_YYYY.rds
6	history	history_YYYY.txt	This code processes emissions history data, assigns pollutant classes, handles PM10 speciation, and generates multiple output datasets including summaries and CSV files for reporting. Write speciated PM data into a separate file.	history_pm_YYYY.rds history.summary_YYYY.txt
7	merge_stars_from_extract	account_YYYY.rds fin_YYYY.rds epn_YYYY.rds cin_YYYY.rds history_pm_YYYY.rds contam_xref.rds	This code combines multiple STARS datasets (ACCOUNT, FIN, EPN, CIN, HISTORY, CONTAM_XREF) into a single merged dataset, calculates "best" emissions (best_ems) for ozone season, introduces a new variable for non-ozone season emissions (other_ems), and prepares data for full-year modeling with a refined threshold. Computes best_ems based on year: Pre-2011: 3-month ozone season (June-August) 2011+: 5-month ozone season (May-September) Calculates other_ems for non-ozone season months.	merged_stars_extract_YYYY_v{version}.rds merge_dups_emis_summary_YYYY_v{version}.csv
8	profiles	EPA default profiles DFW profile CARB 0719 profile Compound database	This script takes EPA default VOC profiles, enhances them with DFW (D404) and CARB (0719) data, excludes non-VOC compounds (except ethane and acetone), normalizes the weight fractions, and links SCCs to profiles. It	Speciation profiles file: prof.emscvt.modified_4VOC_epa.pt

Script No.	Script/Program	Input	Purpose	Output
		contam_xref.rds	generates two output files: a speciation profile file for EMSCVT and an SCC-to-profile cross-reference file for CHMSPL. Reassigns specific profiles: 1003 → D404 (Dallas), 0007 → 0719 (CARB). Identifies profiles dominated by one compound (weight fraction ≥ 0.95) for review.	Speciation cross-reference file: xref.voc.eps3f.modified_4VOC_epa.pt voc_tmp_profs.rds voc_tmp_xref.rds
9	merge_adjustments_for_model	merged_stars_extract_YYYY_v{version}.rds prof.emscvt.modified_4VOC_epa.pt xref.voc.eps3f.modified_4VOC_epa.pt voc_tmp_profs.rds voc_tmp_xref.rds	This script adjusts a STARS extract dataset (merged_stars_extract_YYYY_v#) for air quality modeling by: HECT Adjustments: Modifies SCC profiles to exclude HRVOCs for HECT sources in Harris County, creating unique identifiers (e.g., 0719H). Ethane/Acetone Addition: Generates emissions records for ethane and acetone using EPA default profiles or STARS data, ensuring CB06 compatibility. Flare Corrections: Recalculates stack diameter and velocity for flares based on VOC emissions. SCC Updates: Adds new SCCs, reassigns retired ones, and logs unmatched SCCs with profile 0000. The final output, merged_stars_YYYY_v#, integrates these changes, alongside updated profile (prof.emscvt.modified_epa.pt) and cross-reference (xref.voc.eps3f.modified_epa.pt) files, ready for tools like EMSCVT and CHMSPL. Identify and output SCCs that need default profile; Flag point sources that qualify for the HECT program and adjust their profiles; Creates emissions records for ethane and acetone based on default compositions if they are not already reported; Correct flare stack parameters; Write the final adjusted dataset	merged_stars_YYYY_v{version}.rds SCC_list_assignedEPAdefault_YYYY_v{version}.csv prof.emscvt.modified_epa.pt xref.voc.eps3f.modified_epa.pt voc_epa_profs.rds voc_epa_xref.rds
10	cvt_coord	merged_stars_YYYY_v{version}.rds	The script converts latitude/longitude coordinates from Degrees, Minutes, and Seconds (DMS) format and Universal Transverse Mercator (UTM) coordinates into decimal degrees (DD) format. The final dataset is designed for modeling and ensures all coordinates are consistently formatted for further use.	stars_cc_no_qa_YYYY_v{version}.rds

Script No.	Script/Program	Input	Purpose	Output
11	qa_merged	stars_cc_no_qa_YYYY_v{version}.rds	This script merges the functionalities of four SAS QA scripts: qa_coord.sas, qa_airs.sas, qa_ems_by_sic_scc.sas, qa_stacks.sas. The qa_coord component performs QA on latitude and longitude coordinates in the emissions dataset. It identifies potential issues with the coordinates (e.g., missing, large deviations from the median), generates reports on the problematic data points, and corrects such problems by assigning appropriate default values (plant median). The qa_airs component performs QA on AIRS (Aerometric Information Retrieval System) codes in the emissions dataset. It identifies and reports any errors related to missing or incorrect AIRS codes, specifically focusing on emissions data. The qa_ems_by_sic_scc component performs QA on emissions data by analyzing SIC/SCC pairs. It identifies suspicious emissions data points via one of two methods [chosen by the user, either (i) differs from the mean more than a specified number of standard deviations n or (ii) IQR]. The script generates reports to highlight these outliers. The qa_stacks component performs QA on stack heights in a STARS dataset by comparing them to statistical norms via one of two methods [chosen by the user, either (i) mean ± n * standard deviation or (ii) IQR] for each SIC (Standard Industrial Classification) and SCC (Source Classification Code) pair, then reports outliers.	qa_coord: stars_cc_YYYY_v{version}.rds qa_airs: A report of unique contaminants with missing AIRS codes and emissions above a threshold qa_ems_by_sic_scc: Two reports, one of outliers (where the sample size is greater than or equal to a hard-coded minimum value, currently 5) and one of all pollutant/contam/SIC/SCC combinations with a sample size less than the hard-coded value qa_stacks: Two reports, one of outliers (where the sample size is greater than or equal to a hard-coded minimum value, currently 5) and one of all pollutant/contam/SIC/SCC combinations with a sample size less than the hard-coded value
12	summary	merged_stars_YYYY_v{version}.rds	This script generates a summary report of emissions data by area and permit type. The script processes the emissions inventory data, aggregates the emissions by various categories (e.g., pollutant, account, owner), and generates a report highlighting the emissions for each area. The report focuses on significant accounts and summarizes the emissions data.	stars_sum_account_YYYY_v{version}.rds A summary report of emissions

Script No.	Script/Program	Input	Purpose	Output
13	write_utm_afs	stars_alloc_YYYY_v{version}{update}.rds	This script processes emissions data from a STARS dataset and writes it out in AFS (Air Facility System) format compatible with EPS3 PREPNT program, alongside creating a corresponding R dataset. Additionally, if "RPOlcp" is in the output filename, this script converts decimal degree coordinates to LCP.	afs.stars.{pollutan}.{emissions_type}.{YYYY}_v{version}{update}.{coord_type} afs_{pollutan}_{emissions_type}_YYYY_v{version}{update}.rds
14	write_annual_emissions_afs	afs.stars.All.agg_ems.{YYYY}_v{version}{update}.RPOlcp (OSD daily emissions) afs.stars.All.agg_other_ems.{YYYY}_v{version}{update}.RPOlcp (non-OSD daily emissions)	This script processes emissions data from STARS AFS files, converts daily emissions into annual tons per year (tpy), and splits them into three AFS files: total annual emissions, ozone season day (OSD) emissions (May–Sep), and non-OSD emissions (remaining months).	afs.stars.All.ems_annual_tpy_v{version}{update}.RPOlcp (total annual emissions) afs.stars.All.OSD_ems_tpy_v{version}{update}.RPOlcp (OSD, May–Sep) afs.stars.All.non_OSD_ems_tpy_v{version}{update}.RPOlcp (non-OSD)

4.0 QA/QC and Output Comparison

Ramboll implemented QA/QC checks within the R scripts as outlined in the design document. We also collaborated with TCEQ to compare the intermediate AFS files generated by the legacy SAS system and the new R-based programs. Any discrepancies that arose were investigated, their root causes identified, and the R scripts were updated accordingly. The outputs from the new system are now consistent with those from the legacy system for the processing steps that were included in this phase.

5.0 Conclusions and Recommendations

This project successfully migrated the initial stages of the STARS point source emissions processing workflow from SAS to R, an open-source platform widely used by TCEQ's modeling team. The newly developed R scripts replicate key functionality from the legacy SAS system, including organizing emissions and ancillary data, calculating emissions by time period, and converting source coordinates. These steps prepare the data for use in the EPS3 system, which generates CAMx-ready emissions input files. Ramboll also implemented a more streamlined QA/QC framework within the R workflow, reducing redundancy and improving detection of outliers through the use of robust statistical methods such as the Interquartile Range (IQR).

A comparison between the AFS files generated by the SAS and R-based systems was performed in collaboration with TCEQ. Any discrepancies were reviewed, and the R scripts were updated to resolve identified issues. The outputs from the new system are now consistent with those from the legacy system for the processing steps that were included in this phase.

However, the current phase of work did not include the implementation of speciation and VOC allocation components. We recommend the following next steps:

- Evolve the Design Document into Full System Documentation: Expand the design document into comprehensive system documentation, incorporating recommendations for future improvements.
- Complete R Migration: Implement the remaining speciation and VOC allocation steps in R to fully transition away from SAS and eliminate reliance on proprietary software.
- Refine Workflow Based on User Feedback: Engage TCEQ staff in testing and using the R scripts and incorporate their feedback to improve usability, modularity, and performance of the new system.

6.0 References

Cantu, G; TCEQ Report, "Speciation of Texas Point Source VOC Emissions for Ambient Air Quality Modeling"; July 2003

PES, 2001. Pacific Environmental Services, Inc. Development of Source Speciation Profiles from The TNRCC Point Source Database.

Thomas, R., Smith, J., Jones, M., MacKay, J., & Jarvie, J. (2008, June). Emissions modeling of specific highly reactive volatile organic compounds (HRVOC) in the Houston-Galveston-Brazoria ozone nonattainment area. In 17th Annual International Emission Inventory Conference" Inventory Evolution-Portal to Improved Air Quality", Portland, OR.