State of Environment Technical Report 2025

Surface Water Quality Dataset

December 2025

J. Patrick Laceby[1], Majid Zaremehrjardy[2], Amelie Litalien[1], Nathan Ballard[3], and John Owrin[1]
[1] Alberta Environment and Parks, Resource Stewardship Division, Oil Sands Monitoring Branch
[2] Alberta Environment and Parks, Lands, Cumulative Effects Management Planning
[3] Alberta Environment and Parks, Resource Stewardship Division, Air and Water Resource Management

The State of Environment Technical Report 2025 Surface Water Quality Dataset and this supporting document are available online at https://osmdatacatalog.alberta.ca/dataset/state-of-environment-technical-report-2025-surface-water-quality-dataset

Recommended citation for the State of Environment Technical Report 2025 Surface Water Quality Dataset:

Laceby J.P., Zaremehrjardy, M., Litalien, A., Ballard, N., and Orwin, J. (2025). State of Environment Technical Report 2025 Surface Water Quality Dataset. Oil Sands Monitoring Program. Calgary, Alberta. Retrieved from: https://osmdatacatalog.alberta.ca/dataset/state-of-environment-technical-report-2025-surface-water-quality-dataset

Contents

Tables

1.0 Introduction

This dataset was developed for the 2025 Oil Sands Monitoring (OSM) Program State of Environment (SOE) Technical Report. It is an integrative compilation of surface water quality data from multiple monitoring projects in the oil sands region (OSR) of northern Alberta, Canada. This data generation summary outlines the process used to develop the unified dataset, including data sources, harmonization, and preparation steps. It also has information on the columns included in the final surface water quality dataset. More information on the approach taken to analyse this dataset can be found in the hard-copy technical State of Environment report that can be found on the OSM Program website (https://www.alberta.ca/oil-sands-monitoring-program) or from Open Alberta (https://open.alberta.ca/publications/osmp-state-of-environment-2025-surface-water).

2.0 Data Sources

As determined during a scoping workshop by the OSM Program's multi-stakeholder Oversight Committee and Science and Indigenous Knowledge Integration Committee, the approach for the State of Environment Technical Report 2025 was to include all existing and available water quality datasets from the OSR. Accordingly, this included datasets from multiple OSM Program surface water quality monitoring projects, one Indigenous Community Based Monitoring (ICBM) project, Alberta Environment and Protected Areas (EPA) Long Term River Network (LTRN) and Tributary Monitoring Network (TMN) ambient monitoring data, historical Regional Aquatics Monitoring Program (RAMP) data, a study completed on the Muskeg River from 1998 to 2015, and surface water data collected by EPA as part of the Kearl Oil Sands monitoring response.

The history of water quality monitoring in the OSR is complex, with geographic ranges and scopes changing significantly over time. Data from the monitoring programs was compiled from 11 separate sources outlined below in Table 1. To create a unified dataset, monitoring locations were assessed to ensure continuity over time despite shifting naming conventions (e.g., where RAMP sampling locations became OSM Program sampling locations). From these programs, 157 independent surface water quality sampling sites were identified. These independent sampling sites include some that may have had multiple samples in one location (e.g., Environment and Climate Change Canada (ECCC's) panel sampling approach) and sites where multiple locations in a similar section of the river over the various iterations of the program have been combined into one independent sampling site. The 157 independent sampling locations, include 30 mainstem sites on the Athabasca River and 105 sites on its tributaries, three (3) sites in the Cold Lake region, five (5) sites in the Peace River region, and 14 sites in the Peace Athabasca Delta (PAD), Slave River and tributaries.

Sampling and laboratory analysis methodologies can differ between the site locations, which may affect the suitability of some comparative analyses. This is reflected in the dataset by different source station information and parameters having different names and/or different Valid Method Variable (VMV) codes. VMV codes are a standardized library of values representing available lab methodologies for sample analyses. These were used to help combine parameters with different names while maintaining the record of varying analytical method during the interpretation of results. When creating the unified dataset, the VMV codes were used to guide the parameter name harmonization based on the analytical methods used, best practices, and expert opinion. The approach to compiling data from these sources is outlined below in Table 1 and Table 2, that together outline how the historical and current data was accessed and combined for the SOE water quality dataset.

2.1 Surface Water Quality Datasets

EPA's OSM Surface water data, generated from from field-gathered samples, was downloaded from the OSM Data Catalogue, including:

- Historical data from a 1998-2015 focus study on the Muskeg River watershed (ABS058),
- OSM tributary data (excluding Muskeg River watershed) from 2011 to 2015 (ABS191),
- OSM tributary data collected by Alberta Environment and Parks/Environment and Protected Areas from 2015 to present (active)(ABS240),
- Enhanced Lower Athabasca River monitoring data from a 2018 to present (active) (ABS264).
- RAMP: Historical data collected as part of the Regional Aquatics Monitoring Program from 2003 to 2015

Data from acid sensitive lakes provided with this dataset was not included owing to the focus of this SOE report on rivers and streams, with lakes being a consideration for future iterations of the OSM Program's SOE reporting initiatives. Data for rivers and streams in the Kearl Oil Sands: Ambient Monitoring Response dataset (link) was also included with care taken not to have duplication of data that is included in both the Kearl Oil Sands and the their EPA OSM Program Surface Water Quality Discrete datasets. In total 52 comma separated value (CSV) files with the same

long format data structure were downloaded from the OSM Data Catalogue and merged together for inclusion into the SOE water quality data environment. Data up to October 31, 2023 was included for all data sources where available.

ECCC's surface water data from field samples was downloaded from the ECCC Data Catalogue (link), including

- Mainstem water quality (6 CSV files)
- Tributary water quality (5 CSV files)
- Expanded geographic extent (9 CSV files)

Code was developed in the R programming language to harmonize the ECCC datasets into a single observation per row (long) format to support merging this data with data from the other sources. To support the long-format transition, some column names and symbols had to be manually modified to separate the English and French names and characters (e.g., changing "è" and "é" into "e" in Microsoft Excel). Other modifications to select column names and formats were performed to ensure these CSV were combinable prior to converting them into a long format to be merged with the other datasets. Additional information was provided by ECCC via personal communication, including data files with VMV codes where they were not listed in the online data portal csv files.

Data were also obtained from EPA's Long Term River Network (LTRN) and Tributary Monitoring Network (TMN), downloaded from Alberta's Water Quality Data Portal (link) in one long-format csv file that was combined with the other datasets. Data from Athabasca Chipewyan First Nation (ACFN) was primarily obtained from Mackenzie DataStream (link) in two separate csv files with an additional file for 2023 provided by personal communication. Data from ACFN was converted into a long-format and combined with the other datasets.

Any data that was provided by personal communication is not included in this dataset as it's not available to the public. All data was downloaded from these sources outlined in Table 1 below between May and July 2024. More information on the sources of data or the data compilation process can be requested from OSM.Science@gov.ab.ca.

Table 1: Data Source Descriptions

Source	Description
EPA - OSM Catalogue ABS058 (link)	Historical surface water quality data from a 1998-2015 focus study on the Muskeg River watershed.
EPA - OSM Catalogue **RAMP (link)**	Historical surface water quality data collected as part of RAMP from 2003 to 2012 and additional data collected as part of JOSM and AMERA from 2012 to 2015.
EPA - OSM Catalogue **ABS191 (link)**	Historical surface water quality data collected in tributaries (excluding Muskeg River watershed and outside of the RAMP network) from 2011 to 2015 as part of JOSM and AMERA.
EPA - OSM Catalogue **ABS240 (active) (link)**	Surface water quality data collected in tributaries by Alberta Environment and Parks/Environment and Protected Areas from 2015 to present.
EPA - OSM Catalogue **ABS264 (link)**	Surface water quality data collected from 2018 to 2022 at specific sites along the Lower Athabasca River as part of the Enhanced Lower Athabasca River monitoring (W-RC-1).
EPA - OSM - Kearl Oil Sands: Ambient Monitoring Response (link)	The Kearl Oil Sands: Ambient Monitoring Response includes a compilation of recent surface water chemistry data from surface waterbodies (lakes, rivers and streams). This dataset includes some samples taken as part of the EPA's core monitoring program (i.e., ABS240 above) which were removed to avoid duplication.

Source	Description
EPA LTRN and TMN (LTRN link, TMN link)	Surface water quality data collected as part of the EPA's LTRN and TMN programs from 1987 to 2024 for the Peace River and Athabasca River available from Alberta's Water Quality data portal.
ECCC Mainstem Water Quality (link)	Surface water quality data collected from the Lower Athabasca River (M2 to M7) between 2011 and 2023 by ECCC as part of the OSM Program.
ECCC Expanded Geographic Extent (link)	Surface water quality data for sites in the Lower Athabasca River (LAR), the Peace and Slave rivers, and their tributaries from as early as 1960 to as recent as 2023.
ECCC Tributary Water Quality (link)	Water quality chemistry data for sites on the tributaries of the Peace and Athabasca Rivers (Ells River, Mackay River, Steepbank River, Firebag River, Muskeg River, and High Hills River) between 2012 and 2015.
ACFN - Mackenzie Data stream **Athabasca Chipewyan First Nation CBM Data (link)**	Water quality data from sites sampled as part of the Athabasca Chipewyan First Nation (ACFN) Indigenous community-based monitoring project. Most of the water quality data is publicly available through the Mackenzie Datastream platform. Data from ACFN for 2023 was provided by personal communication as it was not yet available on Mackenzie Datastream.

3.0 Data Preparation

3.1 Data Cleaning and Wrangling

In total, 76 separate CSVs were compiled into one long-format data file with harmonized parameter and site names, units (e.g. "mg/l", "µg/l" and "ng/l" all converted to "ng/l"), detection limit structure, data formats, and removal of pure duplicates. An outline of the data cleaning and wrangling steps is provided in Table 2 below. Multiple QA/QC cross-checks were also completed throughout this process. In addition, lakes and sites upstream of the OSR with the exception of the Athabasca at the town of the Athabasca (ATR1) were removed from the dataset. Site ATR1 serves as a reference point for the Athabasca River upstream of the OSR. The preliminary dataset for surface water quality, including all parameters, and all sites considered in the initial scoping of this SOE report included over 2.5 million unique values.

Table 2 below summarizes the key data cleaning and harmonization steps applied during the preparation of the SOE surface water quality dataset. Each step is categorized by its function to provide the context for how the data environment was established. More information on these processes can be requested from OSM.Science@gov.ab.ca.

Table 2: Summary of SOE Surface Water Quality Data Cleaning Steps

Step Category	Plain Language Description	Purpose
Initial Setup	Loaded required R packages and sourced all scripts in the '/R' directory.	Prepared the environment for data processing.
Metadata Loading	Loaded, cleaned, and validated site data with information available from workplans and other site documentation.	Enabled spatial joins and station validation.
OSM Data Ingestion	Combined multiple CSVs from OSM Program projects and appended Kearl data.	Created a unified dataset of all available surface water quality data in the OSR.
Column Standardization	Renamed columns to match one schema and included metadata fields and columns.	Ensured consistency across datasets.
Lake and Other Waterbody Filtering	Removed lake, pond and other waterbodies (non-river) stations	Included only river sampling sites.

Step Category	Plain Language Description	Purpose
Exclude QC Samples	Filtered out field and trip blank samples.	Excluded QC samples that were not taken at the sampling sites.
Remove other samples	Excluded sediment and epilithic algae samples.	Only included river water samples taken at the sampling sites.
Sample Metadata Decoding	Decoded sample matrix, type, lab, and collection methods. During this step the metadata and other information in the dataset was used to clearly identify whether parameters such as temperature, pH, and conductivity were obtained in the field or laboratory. If there was no documentation available, no location was assigned.	Standardized categorical metadata.
Unit Harmonization	Standardized unit naming conventions (e.g., µg/L to ug/l, °C to deg C) and detection limit units.	Ensured consistency in unit names.
Unit Conversion	Measurement units were standardized across the dataset using dimensional analysis to convert values to a common unit for each parameter (e.g., converting mg/L to ug/L or ng/L). Detection limits were also updated with any inconsistencies flagged for QA/QC.	Enabled unit values to be standardized, ensuring consistency.
Site Harmonization	Standardized station names and descriptions using site metadata. Merged and harmonized site information from EPA, ECCC, and SOE site to support joins and reporting.	Ensured that only relevant monitoring sites were included, and that spatial metadata was consistent.
Sample Filtering	Removed non-water quality information (i.e., rain, snow, discharge) and samples after Oct 2023.	Included only relevant (river water samples) and maintained temporal consistency in data from multiple sources.
Duplicate Resolution	Identified duplicates by sample number, parameter names, sample date and time, and VMV code. Only pure duplicates (i.e. identical samples and not QA/QC duplicates) were removed.	Prevented double-counting and retained clean records.
Censoring Logic	Standardized flags ('<', '>') and created a logical column for censored values.	Supported accurate statistical treatment of censored data.
Parameter Mapping	Mapped variable names to master names using a parameter key.	Enabled grouping and analysis of related parameters.
Missing Value Checks	Reviewed missing values in key fields and flagged inconsistencies.	Ensured completeness and identified data quality issues.
Final Dataset Assembly	Merged all cleaned datasets.	Produced a harmonized dataset for reporting and analysis.

3.2 Parameter Preparation

The harmonization of the various parameter names used across the different data sources and program iterations over time was fundamental to preparing the dataset for analysis. The initial dataset included parameters that were analyzed in multiple different ways over the analysis period, which is reflected in the dataset by parameters having different naming conventions and, or different Variable Method Value (VMV) codes.

VMV codes are a standardized library that represent a unique combination of variable, method, unit, and detection limit. These were used to combine parameters with different names while maintaining the record of varying analytical method during the interpretation of results. When combining the dataset, the VMV codes were used to guide the parameter name harmonization based on the analytical methods used, best practices, and expert opinion.

Combining different analytical approaches to a parameter presents the potential of introducing bias for trend assessments. However, the assessment in the SOE report is limited to screening level assessments and this potential bias was identified as gap in the report. The initial dataset included over 3500 distinct parameter name and VMV Code combinations that were harmonized into one data environment. Maintaining the VMV Codes, Sample Number, and Source Station number in this dataset allows users to trace back all data to its source.

3.3 Site Preparation

Owing to the multiple iterations of past and even present surface water quality monitoring programs in the OSR there have been numerous site naming conventions used, with numerous organizations sampling the same sites, and at times, simultaneously. To facilitate this site harmonization, all sites sampled within 1 km of each other were combined into one SOE site except for locations where a distinct tributary confluence existed between sampling locations.

In addition, a standardized SOE site naming convention was developed for each site that indicated the waterbody and the site number for the SOE with numbering beginning at the most upstream sites (i.e., Muskeg River Sampling Location 1 (MUR_1)). Waterbody codes are presented in Table 3. Where the waterbody codes would be identical between waterbodies, an additional letter was added for clarity (i.e., Jackfish Creek [JAFC] and Jackpine Creek [JAPC]]). Again, distinct sampling locations were assigned if locations were greater than 1 km from each other or were collected from different positions upstream or downstream of the mouth of a tributary.

Table 3: Waterbody Names

Waterbody	Code	Waterbody	Code	Waterbody	Code
Athabasca River	ATR	Grayling Creek	GRC	Muskeg River	MUR
Beaver Creek	BEC	Gregoire River	GRR	North Steepbank River	NSR
Beaver River	BVR	Hangingstone River	HAR	Peace River	PER
Big Creek	BIGC	High Hills River	HHR	Pierre River	PIR
Birch Creek	BIRC	Horse River	HOR	Poplar Creek	POC
Birch River	BIR	Iyinimin Creek	IYC	Quatre Fourches Channel (PAD)	QFC
Buckton Creek	BUC	Jackfish Creek	JAFC	Redclay Creek	REC
Calumet River	CAR	Jackpine Creek	JAPC	Richardson River	RIR
Christina River	CHR	Jackfish River	JAR	Sawbones Creek	SAC
Clearwater River	CLR	Keane Creek	KEC	Shelley Creek	SHC
Dover River	DOR	Lesser Slave River	LSR	Slave River	SLR
Dunkirk River	DUR	Maybelle Creek	MAC	Smoky River	SMR
Ells River	ELR	Mackay River	MAR	Stanley Creek	STAC
Embarras River	EMR	McClelland Creek	MCCC	Steepbank River	STR
Eymundson Creek	EYC	McIvor River	MCIR	Sunday Creek	SUC
Fisherman's Creek	FIC	McLean Creek	MCLC	Tar River	TAR
Firebag River	FIR	Mills Creek	MIC	Upper Fletcher Creek (PAD)	UFC
Flett Creek	FLC	Moose Creek	MOC	Unnamed Creek	UNC
Fort Creek	FOC	Muskeg River Tributary	MRT	Unnamed Tributary	UNT
Goose Island Channel (PAD)	GIC	Muskeg Creek	MUC	Wapasu Creek	WAC

3.4 Quality Assurance and Quality Control

The data preparation workflow integrated multiple quality assurance and quality control (QA/QC) processes, combining automated checks with manual oversight. Structural consistency was verified repeatedly (over ten times) across all datasets (e.g., OSM, ECCC, ACFN, EPA), both between and within columns, and before and after any merging occurred. Customized filters based on best practices were applied to remove records with missing or invalid values, and logic-based stop conditions were embedded throughout the code to halt execution if critical columns were missing or if unexpected data formats were encountered. Multiple automated flags were built in to detect mismatches in units, method codes, detection limits, and VMV assignments. These checks were complemented by manual quality reviews at various stages of dataset development, often conducted by different individuals to ensure independent validation. The workflow also included iterative cycles of loading, merging, and rechecking the data, reinforcing data integrity at each step.

4.0 State of Environment Report Water Quality Dataset 2025

The State of Environment Report Water Quality Dataset 2025 is a harmonized, long-format dataset that integrates surface water quality data from multiple monitoring programs, including OSM, ECCC, EPA, and an Indigenous Community-Based Monitoring initiative. It includes standardized parameter names, units, detection limits, and site identifiers to support surface water quality initiatives. Table 4 provides a detailed description of the column names included in the dataset. If you have questions or comments regarding this dataset, please contact OSM.Science@gov.ab.ca.

Table 4: Column names in the State of Environment (SOE) Report Water Quality Dataset 2025

Column Name	Description
SAMPLE_NO	A unique sample number that comes directly from or was generated based on the data source.
SOE_CODE	A site identifier that was uniquely generated for this dataset. The SOE_CODE is based on the river identifier in Table 4 above with numbers that increases as you progress from the most upstream site downstream.
SOE_SITE_NAME	A site name that was generated for this dataset that reflects previous site naming conventions.
SOE_LATITUDE	The latitude for the unique sites identified by the SOE_CODE and SOE_SITE NAME.
SOE_LONGITUDE	The longitude for the unique sites identified by the SOE_CODE and SOE_SITE NAME.
SAMPLE_DATETIME	The date and time the sample was taken.
SOE_GROUP_NAME	The group name for SOE Parameters used in the assessment (e.g., dissolved metals, total metals, nutrients, etc.)
PARAMETER	The unique parameter names used in the SOE report for guideline assessment and trend assessment.
VMV_CODE	Variable Method Value (VMV) codes which provide a link to information on the methods used to assess the parameters.
PARAMETER_VALUE	The value of the parameter
UNIT	The measurement unit of the parameter's value.
IS_CENSORED	A logic flag that when TRUE indicates the parameter value was not detected, when FALSE, it indicates he samples parameter was below or above the limits of detection.
CENSORING_FLAG	If this entry is "<", it indicates the measured value is below the detection limit for that parameter. If this entry is ">", it indicates the measured value is above the upper quantification limit or outside the calibrated range of the method. If there is no entry, then the value is not censored (i.e., FALSE in the IS_CENSORED Column)
SOURCE	A column that provides general information on the source of the data (e.g, OSM, ECCC, EPA).
SOURCE _STATION_NO	The original station number from the source data, which can be used to link the data back to the source in combination with the sample number above.
SOURCE_LATITUDE	The source data latitude for the sampling site
SOURCE _LONGITUDE	The source data longitude for the sampling site
SOURCE_PARAMETER_NAME	The original parameter name from the source data