**Use of Secondary Data to Generate Publicly Owned Treatment Work Sewersheds for the United States**

[*QA Project Plan Coversheet w/ Signature Page*]



***[This page left intentionally blank]***

**A.2 Table of Contents**

**List of Tables**

**List of Figures**

# SECTION A – PROJECT MANAGEMENT

**(A.1 and A.2 are previously provided as Approvals page and Table of Contents)**

**A.3 Distribution List**

Quality Assurance (QA) Project Plans and Standard Operating Procedures (SOPs) shall be controlled (through documented approvals) as required by Section 5.3 of the Office of Research and Development (ORD) Quality Management Plan. The project lead will be responsible for distribution of the current signed approved version of the QA Project Plan to project participants shown in Section A.4. Signed approved versions of SOPs will be available to project staff through the [ORD@Work SOP intranet site](https://webx.ord.epa.gov/quality-assurance/standard-operating-procedures-sops?combine=&field_sop_previous_number_value=&title=&field_lab_value=ceser&field_sop_contact_value=&field_discipline_value=&items_per_page=10). Signature approved electronic copies of this QA Project Plan, SOPs, and any associated QA assessment reports, will also be maintained in ORD QA Track.

The project lead will also be responsible for timely communications with all involved participants and will retain copies of all management reports, memoranda, and correspondence between research task personnel.

**Table 1. Distribution list for this QAPP. In addition, this QAPP will be provided to any other project support EPA staff members that perform work. In addition, a secured copy of this QAPP will be maintained in the project server files.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name** | **Organization** | **Role** | **Responsibility** | **Distribution List** |
| Andrew Murray | EPA ORD | Project Lead | Scientific leadership for the research effort | Yes |
| Jill Hoelle | EPA ORD | QA Manager | Oversight of QA program implementation | Yes |
| John McKernan | EPA ORD | Branch Chief | Supervisor to Project Lead in HSMMD WAIDB | Yes |
| Alexander Hall | EPA ORD | Project Co-Lead | CESER LRTD Model Support | Yes |
| Elisabeth Schlaudt | EPA OW | Project Co-Lead | OWM WID Subject Matter Expert | Yes |
| Smiti Nepal | EPA OW | Project Co-Lead | OWM WID Subject Matter Expert | Yes |

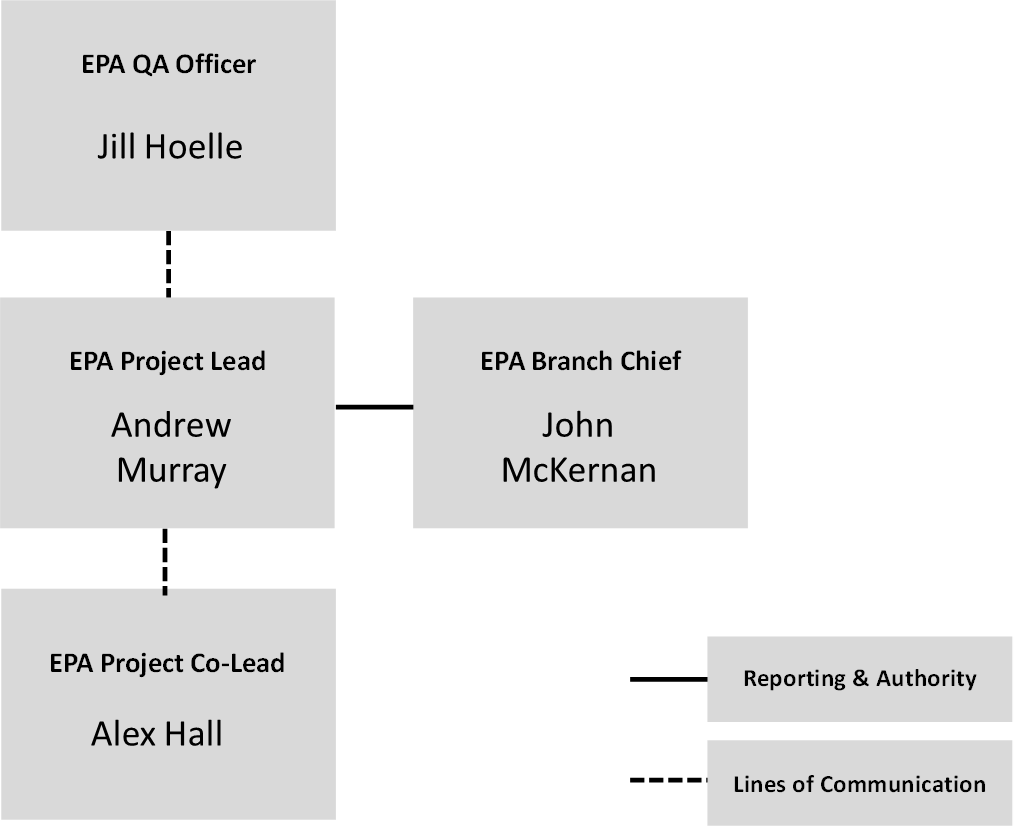
**A.4 Project/Task Organization**

The project organization for this work is depicted in **Figure 1**. Andrew Murray will serve as the Project Lead (PL) and will be responsible for the technical and administrative aspects of this project, ensuring that the quality of work meet the requirements of the EPA. The PL will ensure that the project staff will adhere to the requirements of this QAPP and will keep the EPA Branch Chief (BC) and QA Manager apprised of any data quality problems that arise. The PL will prepare all QA-related plans and transmitting them to the BC and the QA Manager for review and approval. The PL is responsible for overseeing the work performed and providing assurance that all required QA/Quality Control (QC) procedures are being implemented.

The QA Manager (QAM), Jill Hoelle, is responsible for monitoring and facilitating QA activities on this Project, and generally helping the PL and co-Lead understand and comply with EPA QA requirements. As QA Manager, Ms. Hoelle has a direct line of reporting to her Division Director who is organizationally independent from Andrew Murray. This gives the QA Manager sufficient independence to function effectively and efficiently.

The Project Co-Lead (PCL) is responsible for performing electronic data analysis and synthesis activities. Thus, the PCL is primarily responsible for the quality of all work under this Project. The PCL is also responsible for producing all deliverables and for ensuring that quality procedures are implemented. He will discuss any concerns about quality with the PL and QAM to identify and implement quality improvements.

Figure 1. Project QA Organization.

******

**A.5 Problem Definition/Background**

There are approximately 12,000 publicly owned treatment works in the United States. Currently, the US EPA doesn’t know the geographic extent of these systems—known as a sewershed. A sewershed is the community area served by a wastewater collection system. More specifically, it is [all](https://www.lawinsider.com/dictionary/sewershed) the land area that is drained by a network of municipal storm sewer system conveyances to a single point of discharge to a water of the United States. Having accurate delineations of sewersheds within the US aid in advancing our understanding of 5 key areas:

1. Clean Water State Revolving Fund (CWSRF)-- The CWSRF program is a federal-state partnership that provides low-cost financing to communities for a wide range of water quality infrastructure projects, including municipal wastewater facilities, nonpoint source pollution control, decentralized wastewater treatment systems, stormwater runoff mitigation, green infrastructure, estuary protection, and water reuse. Sewershed delineations can provide a means to more effectively analyze the impacts of SRF projects for wastewater infrastructure.
2. Wastewater Surveillance— During the pandemic, it became evident the role of wastewater surveillance in understanding community level health trends.  At this point, many communities continue to wastewater surveillance beyond SARS-CoV-2 into other disease tracking.   EPA has taken the effort to define and develop networks of sampling sites which can be rapidly accessed at the onset of an epidemic and used for pre-epidemic monitoring.  The effort consist of two approaches: national level and small scale to identify hot spots.  As part of this, it is important to generate data to improve wastewater-based models and there is a clear need in understanding sewersheds around the nation to develop future wastewater emerging contaminants surveillance.
3. Pretreatment Program—Understanding influent loads of discharge impacting treatment plants.
4. Decentralized Wastewater Systems Programs—Identifying gaps where centralized systems are not available and providing support to decentralized communities.
5. EJ communities—Identifying gaps in service and support EJ communities access resources and funding.
6. Supplement the Clean Watersheds Needs Survey—As directed by Congress, the CWNS provides an assessment of the capital investments necessary for states, the District of Columbia, and U.S. Territories to meet the water quality goals of the CWA over the next 20 years. To complete the survey, each state collected and submitted data on needed investments for wastewater infrastructure. Knowing the sewershed for a given POTW is critical for analyzing the potential impact of reported needed investments. The CWNS team will aggregate socio-demographic data from the Census block groups to the sewershed level to analyze trends in sewershed characteristics. This analysis will provide important insight into the representativeness of the 2022 CWNS dataset of states’ investment needs, which is the only existing source of nationwide wastewater infrastructure needs and technical information. By identifying gaps in state-reported data and the populations affected, this analysis can contribute to planning of the next CWNS and development of state SRF programs.

**A.6 Project Description and Objectives**

The objective of this project is to develop and execute an approach to model POTW sewersheds for systems that do not publicly share their sewersheds.

***Task 1. Identify and gather municipal sewershed boundaries***

Models require good training data. A model uses explanatory information to replicate empirical observations so that these observations can be replicated in areas where no observational data is present. In order to develop a model that delineates sewersheds, the model needs to learn what input variables explain the presence or absence of POTW sewersheds. The “presence and absence” information serving as the training data will be POTW sewersheds we find from state or municipal agencies. Some states and municipalities have publicly published their POTW sewersheds. This task will entail gathering this data from online research and compiling and curating the data for model training.

***Task 2. Differentiate census blocks that are either primarily served by POTWs or decentiralized systems***

Wastewater disposal can be bifurcated into two primary categories: centralized and decentrailized. Our second objective is to develop and execute a modelling approach that assigns each US census block with their associated wastewater category based on training data from task 1. The long form census asked this question had this question but has not been asked since 1990. This 1990 data will be used as a baseline. The modelling effort will update these 1990 estimates for 2024. Using census data, parcel fabrics, ZTRACs housing data, remote sensing imagery, Clean Water Needs Survey, digital elevation models, and other secondary data a model will be trained on states where POTW sewershed data is available. The training data will provide the model parameters that will be applied to the rest of the country where POTW sewersheds are not available.

***Task 3. Associate POTW supplied census blocks with the utility provider***

Task 2 only provides the type of wastewater system. Task 3 will associate the correct utility name and identifier to block groups served by POTWs. Using a random forest modeling technique, utility identifiers from the Clean Water Needs Survey will be used to associate blocks served by POTWs with the POTW name.

***Task 4. Create a public geospatial online application of POTW sewershed boundaries***

Here we take the model output of sewersheds with relevant utility information and muncipal-supplied boundaries and develop an online application to view the data product geospatially.

**Table 3. Proposed Timeline for the Project.**

|  |  |
| --- | --- |
| **Task** | **Dates** |
| Develop QAPP | May-24 |
| Task 1 | Sep-24 |
| Task 2 | Dec-24 |
| Task 3 | Mar-25 |
| Task 4 | Jun-25 |

**A.7 Special Training/Certification**

N/A

**A.8 Documents and Records**

Research activities must be documented according to the requirements of ORD QA Policies titled *Scientific Recordkeeping: Paper*, *Scientific Recordkeeping: Electronic*, and *Quality Assurance/ Quality Control Practices for ORD Laboratory and Field-Based Research*, as well as requirements defined in this QA Project Plan. The ORD QA Policies require the use of research notebooks and the management of research records, both paper and electronic, such that project research data generation may continue even if a researcher or an analyst participating in the project leaves the project staff.

Electronic project records will be maintained by the project lead on the ORD network drive:

L:\Public\CESER-PUB\amurray

Electronic Records shall be maintained in a manner that maximizes the confidentiality, accessibility, and integrity of the data. ORD PPM Section 13.6 provides guidance on the maintenance of electronic records for ORD.

Records retention:

Records that are generated under this research effort will be retained in accordance with EPA Records Schedule 1035, and as required by Section 5.1 of the ORD Quality Management Plan for QA Category B Projects.

# SECTION B – DATA GENERATION & AQCUISITION

***Refer to subsection content specified in the applicable*** [***QA Project Plan Checklist (by Research Type)***](https://intranet.ord.epa.gov/ceser/quality-assurance-qa-project-plan-development)***.***

This project will utilize publicly available data previously QA/QC’d by federal or state agencies. This project follows ORD’s QAPP requirements for Secondary Data Research Projects under Appendix B “QAPP Requirements for Projects Using Secondary Data”. This project will use secondary data. No primary data will be generated in this project. Secondary data will be used to generate new electronic datasets as specified below:

**B.1 Sources of Secondary Data**

The source(s) of the secondary data is provided in table 4. The rationale for selecting these datasets is that they are relevant datasets for the model development. All data sources used in this project will be identified in the project deliverable.

Table 4 Data sources

|  |  |
| --- | --- |
| **Dataset** | **Source** |
| Clean Watersheds Needs Survey | https://www.federalregister.gov/documents/2024/03/11/2024-05049/proposed-information-collection-request-comment-request-clean-watersheds-needs-survey-cwns-renewal |
| NLCD & Imperviousness | https://www.mrlc.gov/ |
| Digital Elevation Model | <https://github.com/jhollist/elevatr> |
| 1990 Census | https://data2.nhgis.org/main |
| 2020 Census | https://data2.nhgis.org/main |
| Building Footprints | https://github.com/microsoft/USBuildingFootprints |
| Parcel Data | https://regrid.com/ |
| Combined Sewer Overflows | <https://echo.epa.gov/> |
| NPDES | https://echo.epa.gov/ |
| National Hydrography Dataset Plus | https://www.epa.gov/waterdata/nhdplus-national-hydrography-dataset-plus |

**B.2 Quality of Secondary Data (Following Appendix B Section 3.0)**

This project will utilize secondary data that has undergone QA/QC review previously. Data portals for the datasets list the QA requirements and the quality of the datasets are described in Table 5. As such, the quality of the secondary data has previously been demonstrated, as the QAM will determine. A disclaimer will be added to the project deliverable that indicates that the quality of the secondary data was evaluated outside of the EPA.

Table 5

|  |  |
| --- | --- |
| **Dataset** | **Source** |
| Clean Watershed Needs Survey | The EPA conducted quality assurance and quality control reviews of the data presented in this Report to ensure their precision and accuracy. Throughout these reviews, the EPA followed a quality assurance project plan, which it developed in keeping with its Information Quality Guidelines and the *EPA Requirements for Quality Assurance Project Plans* (EPA QA/R-5, EPA/240/B-01/003). The DEP allows states to directly enter needs and technical data and upload the required supporting documentation. The EPA developed the DEP to include automated checks of expected entries and value ranges to minimize incorrect or incompatible data entry. In addition, the EPA followed specific, documented protocols for reviewing technical and needs data submitted by states. These quality control checks were focused on technical data for wastewater submissions in addition to needs data for all infrastructure types. Data were reviewed throughout the data entry period as they were submitted and underwent additional checks at the close of data entry. |
| NLCD & Imperviousness | Peer reviewed federal data product: https://www.mdpi.com/2072-4292/11/24/2971 |
| Digital Elevation Model | Peer reviewed federal data product: https://ca.water.usgs.gov/projects/sandiego/data/gis/dem/ned13/NED\_DataDictionary.pdf |
| 1990 Census | Census data products undergo rigorous QA/QC |
| 2020 Census | Census data products undergo rigorous QA/QC |
| Building Footprints | "Our metrics show that in the vast majority of cases the quality is at least as good as data hand digitized buildings in OpenStreetMap. Pixel recall/precision = 95.5%/94.0%. We estimate <1% false positive ratio in 1000 randomly sampled buildings from the entire output corpus." |
| Parcel Data | This data is taken from county assessors' offices. Property data is QA/Qced at the county level. |
| Combined Sewer Overflows | This is federally QA/Qced data collected for under the CWA. |
| NPDES | This is federally QA/Qced data collected for under the CWA. |
| National Hydrography Dataset Plus | "NHDPlusV2 contains statistical QA measures of the flow estimates. These QA statistics are included with the NHDPlusV2 data and provide a comprehensive assessment of the quality of the flow estimates. The NHDPlusV2 processing tools perform extensive automated quality assurance/quality control (QA/QC) on the VAAs to ensure the accuracy and consistency of attributes." |

**B.2 Secondary Data Acquisition and Management**

Microsoft Excel spreadsheets, GIS files, and code will be used to store secondary data and the results of computer analysis. The design of the spreadsheets and code will be adequate and appropriate for use. All parameters will be defined and documented. Tables and fields will be clearly and unambiguously named. Output tables, graphs, spatial displays, and text will be cross-checked to ensure accuracy. Electronic copies of cited data sources will be maintained in project files.

**B.2 Project Design**

***Task 1. Identify and gather municipal sewershed boundaries***

Online datasets from municipalities and/or states of sewershed boundaries will be collected. The metadata on these datasets will be examined to see how the boundaries were delineated and the quality of the datasets. Boundaries that meet our quality threshold will be chosen to serve as our training dataset. Boundaries will be selected for their quality, vintage, and granularity. Any querying of datasets will be cataloged in addition to any augmentation of the features or attribute information.

***Task 2. Differentiate census blocks that are either primarily served by POTWs or decentiralized systems***

This task requires the development of a decision tree model. The model will be developed within R Studio and all code will be stored on GitHub. The model will use datasets from task 1 and explanatory datasets laid out in Table 4. Model parameters and model efficacy results will be maintained in the source code on GitHub. Model parameters will be determined by model performance.

***Task 3. Associate POTW supplied census blocks with the utility provider***

This task requires the development of a random forest model. The model will be developed within R Studio and all code will be stored on GitHub. The model will use outputs from task 2 and POTW IDs cataloged by the Clean Water Needs Survey.Model parameters and model efficacy results will be maintained in the source code on GitHub. Model parameters will be determined by model performance.

***Task 4. Create a public geospatial online application of POTW sewershed boundaries***

This task will use the geospatial files developed form task 4 to create an online application where the public can access the boundary files. This application and associated metadata will be developed within EPA’s GeoPlatform and esri online and will go through GeoPlatform metadata standards and publication requirements.

# SECTION C – ASSESSMENT AND OVERSIGHT

**C.1 Assessments and Response Actions**

This QAPP does not include any laboratory or field analyses. Thus, audits are not required for utilizing the secondary data for this project.

Draft publications resulting from this project will undergo ORD clearance in RAPID prior to dissemination as required by ORD Policy titled *ORD Clearance Policy and Procedures* and CESER SOP titled *Standard Operating Procedure for Product Clearance*.

**C.2 Reports to Management**

N/A

Required approvals for draft publications undergoing ORD clearance is documented in STICS.

# SECTION D – DATA VALIDATION AND USABILITY

**D.1 Data Review and Verification**

All work conducted under this project will be subject to technical review. This review will be conducted by technical reviewers with specific expertise prior to the STICS clearance process.

The term “verification” refers to whether conclusions can be correctly drawn from the data. The quality of data will generally be determined by its source. Data published by government sources, national laboratories, and peer-reviewed journals will generally be assumed to be of sufficient quality unless other factors indicate otherwise. We will characterize the uncertainty using sensitivity analysis.

**D.2 Verification Methods**

Verification of input data is performed by expert review of data sources as well as prior peer-review. All data is being sourced from either federal government sources (have undergone peer-review), open-source data which has been peer-reviewed, or from sources the federal government has contracted with for access and has deemed acceptable. Verification of validation data (state supplied sewershed boundaries) will be performed using spatial analysis to identify erroneous polygons by evaluating areas against census data and by visual inspection.

Model validation will be performed for tasks 2 & 3. Both tasks utilize tree-based machine learning methods and can be validated using accepted methods. Models will be trained on a randomized subset of data, selected from boundaries collected in task 1. These methods will include the determination of an Area Under Curve (AUC) value, as well as specificity and sensitivity metrics, which are obtained from a confusion matrix derived from model training.

**D.3 Reconciliation with User Requirements**

The product output objective is for spatial areas estimating sewersheds. Data will be summarized using spatial polygons in a shapefile format. Model performance statistics, as described in D.2 will be provided within model training documentation, to accompany final sewershed dataset. Tables will include confusion matrices and AUC values as tables for tasks 2 & 3. Additionally, a table will be provided that presents the probability returned by the decision tree model (task 2) and the probability returned for the random forest model (task 3) for each 2020 census block.

REVISION HISTORY

|  |  |  |
| --- | --- | --- |
| Revision # | Description | Effective Date |
| 0 | Initial Version |  |
|  |  |  |