

**CA Wastewater Needs Assessment
Review of Wastewater Infrastructure Modeling Effort
Summary for April 2026 Advisory Group Meeting**

University of Massachusetts Amherst: Nelson da Luz, Emily Kumpel, Jay Taneja

1. Model Purpose and Intended Use

Throughout the Wastewater Needs Assessment (WWNA), data on the different types of wastewater infrastructure and systems serving Californians have been collected and processed. This data collection effort has been extensive to build the Facilities List¹, but there are still gaps in the data to be filled. To fill in one of the major gaps in data, the WWNA project team will use a model developed and refined by the University of Massachusetts Amherst (UMass) since 2021 to identify unsewered areas, i.e., likely locations of onsite wastewater treatment and disposal systems. These systems include but are not limited to septic tanks, cesspools, leach fields, and seepage pits.

The model serves as a screening and planning tool. Predictions by models are good approximations to the best of our knowledge. It is not meant to be used as a regulatory determination tool or to assign “concern” labels at the parcel level. For maintaining privacy, the model will not publicly display parcel-level information – in the WWNA or elsewhere. The model fills a data gap in the absence of complete ground truth data for unsewered areas in the state, and more broadly across the U.S.

The Unsewered Model will be used as an input to guide the Onsite Sewage Treatment Systems (OSTS) Connection Opportunities portion of the broader WWNA project, which will identify opportunities for connecting OSTs to existing sewer collection systems or connecting OSTs to a community cluster system.

2. Methodology

A description of the methods used to develop the model can be found in the Executive Summary for the July 2025 WWNA Advisory Group meeting². In short, this model leverages the patterns found among homes and businesses with known wastewater system types to predict the wastewater system type where it is unknown. The model is built to make 1 of 3 possible classifications: 1) Sewer - served by a sewer connection (collection system), 2) Onsite - served by an onsite solution (unsewered - e.g., septic tank, cesspool, or straight pipe), or 3) Not Applicable - not served by any wastewater infrastructure. UMass employed a two-stage machine learning approach to model

¹ See the Phase 1 Executive Summary: [Phase-1-Executive-Summary.pdf](#). Also see California Wastewater Needs Assessment Phase 1 Report, Chapter C: [wastewater-needs-assessment-phase-1-report](#)

² See the July 2025 Unsewered Model Summary [UMass-Unsewered-Model-Executive-Summary-ENGLISH.pdf](#)

wastewater infrastructure coverage for this purpose ³. Since July 2025 WWNA Advisory Group meeting, based on feedback from many stakeholders, UMass implemented a series of changes to the approach to boost model performance within California. This included adding higher-quality sewer data for model training, accounting for differences between urban and rural areas, and fine-tuning the model to have greater focus on California training examples and the extreme ends of the wastewater service spectrum – namely, urban unsewered parcels and rural sewerred parcels.

3. Description of Model Performance and Model Uncertainty

UMass worked with the WWNA project team, State Water Board, and the California Regional Water Quality Control Boards (collectively Water Boards) to collect available wastewater infrastructure data (lists of parcels that include sewer systems or onsite wastewater treatment or disposal systems from parts of 10+ states including California). From California specifically, UMass included ground truth data (i.e., information on the presence of collection systems or onsite systems) samples from 53 counties (more than 900 thousand land parcels represented). The final version of the model created by UMass can be assessed by the following performance metrics in Table 1.

Table 1. Unsewered Model Performance Metrics.

Performance Metric	Value	Description
Accuracy	95%	The number of correct predictions divided by the total number of predictions made
Type 1 Error	4%	The proportion of sewer-served parcels incorrectly predicted as onsite-served
Type 2 Error	11.3%	The proportion of onsite-served parcels incorrectly predicted as sewer-served
Matthews Correlation Coefficient (MCC)	0.81	A balanced metric which tells you how well a model predicts both onsite and sewerred outcomes. Can range from -1 to +1

Overall, a model with an MCC value above 0.8 is considered excellent, a standard which our unsewered modeling effort achieves.

The parcel-level predictions presented by this model are probabilistic. Uncertainty in any model prediction is unavoidable. The uncertainty associated with predictions may be affected by quantity and quality of available ground truth data in various parts of California. California is a geographically diverse state in many ways and while the model captures some of these dynamics, errors may cluster geographically. As a result

³ Many more details on the overall approach used to develop the model can be found here: <https://doi.org/10.1021/acsestengg.5c00637>

of the model uncertainty, again, this model should not be used as the sole basis for regulatory compliance decisions. It is not to be used as a regulatory determination tool or to assign “concern” labels at the parcel level.

4. Summary of Unsewered Model Results

The Unsewered Model predicts that 85.1% of parcels with wastewater needs in California are served by sewer (collection systems) with the remaining 14.9% being unsewered. These percentages are specific to land parcels, not population. The percentage of the population served by unsewered areas is likely lower than 14.9%. Particularly, urban sewered parcels that serve several households would need to be considered when calculating population estimates. As a point for comparison, the 2023 American Housing Survey (AHS) estimates that 11% of California households are reported as not connected to sewer systems⁴. Parcels that have a classification of Sewer or Onsite based on model outputs were assigned a confidence value. The confidence values describe on a scale of 0 to 100 of how confident the model is that the assigned classification is correct. The average confidence of predictions across California was 81.1%. 44 California counties had an average confidence equal to or greater than 80%.

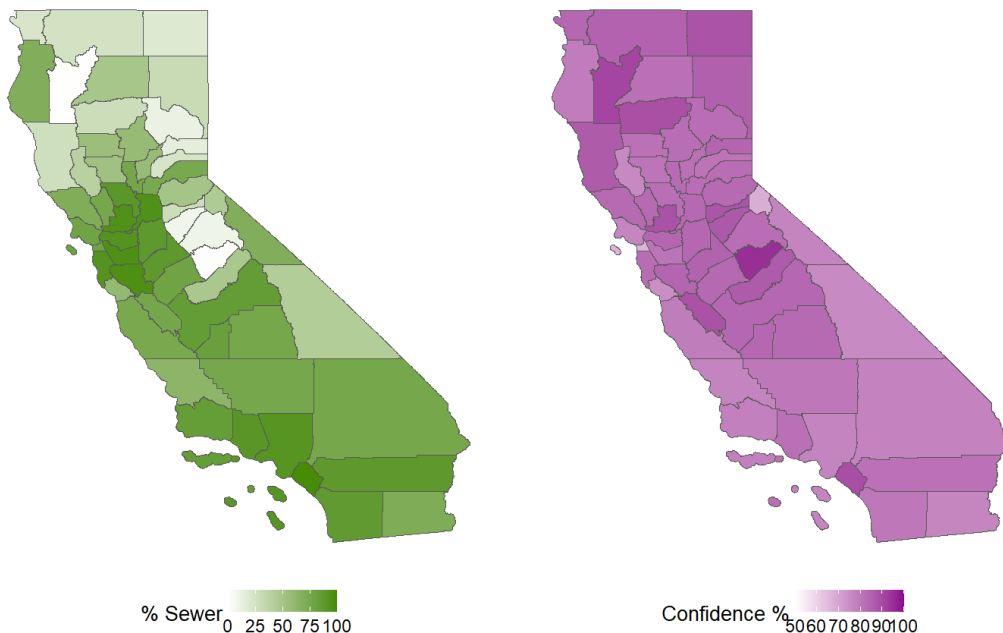


Figure 1: (Left) Map of California indicating the percentage of sewer parcels served for each county. (Right) Map of California indicating the average confidence of predictions in each county.

⁴ [American Housing Survey \(AHS\) – AHS Table Creator](#)

These results were shared with the Water Boards and the WWNA project team in PDF map format and MS Excel files with county-by-county summary statistics. Information on geodatabase format was also provided to the State Water Board. The model will be used in the Onsite Sewage Treatment Systems (OSTS) Connection Opportunities portion of the broader WWNA project, as well as by the Water Boards for future broader purposes such as supporting water quality and nutrient loading analyses or informing funding prioritization and infrastructure planning.