x

Mid-Semester Deliverable

Our Team


Jessica Liu


Jenny Chung


Nicholas Huang


Aadil Jamari


Alivia Ding


Khushboo Teotia


Cristina Prieto


Avery Hipolito


Michael Murphy


Sona Wyse


Kelly Hu


Lizbeth Velazquez


Jeffrey Gao


Catherine Wang


Dhruv Syngol


Sanay Bordia

SAAS

Agenda

1. Introduction
2. Project Groups and Timeline
3. Data Overview
4. Time Series Approach
5. Spatial Approach

SAAS

Introduction

Project Objective

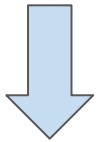
Investigate the temporal and spatial behavior of PFAS chemicals to identify potential relationships between PFAS and other environmental, geological, and geographical factors

SAAS

Project Groups

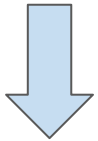
Split up into groups for two project approaches:

Time Series



Training time series models for each of the top 5 PFAS chemicals

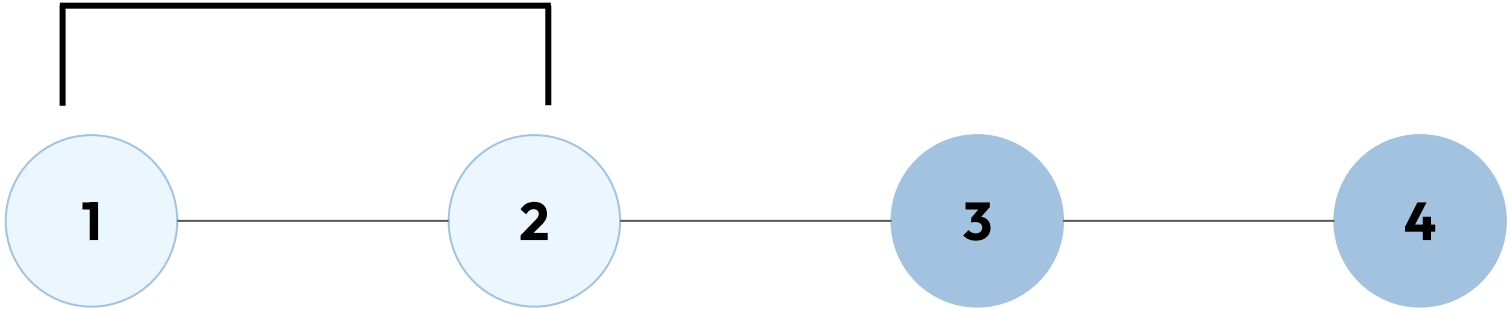
Spatial Analysis



Mapping samples, airport, locations, and industrial site locations on California terrain map

SAAS

Project Timeline



SAAS

Brief Data Overview

SAAS

Data Overview

Data from Geotracker PFAS Map

Filtering

- Top 5 most prevalent PFAS chemicals (PFOA, PFOS, PFHxS, PFNA, HFPO-DA)
- Values from 2019 - onwards
- Qualifier: “=”

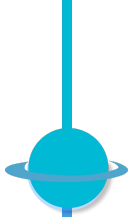
Aggregations

- Grouped by chemicals and by day
- Calculated median and mean

Anomalies

- Extremely large values recorded for certain sites (ex: military camps)
- Negative latitude and longitude values

SAAS



Time Series

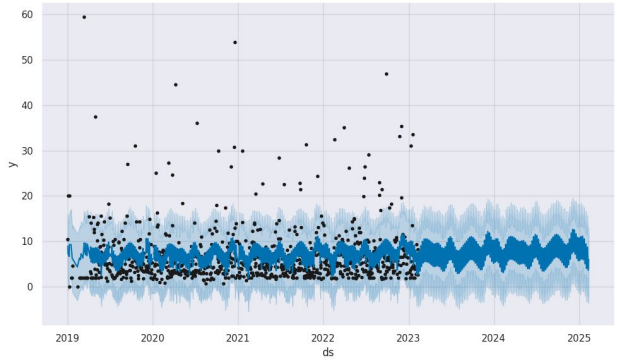
SAAS

Forecasting with Prophet

Utilized Meta's Prophet Time Series Model to predict future trends of mean/median chemical concentrations

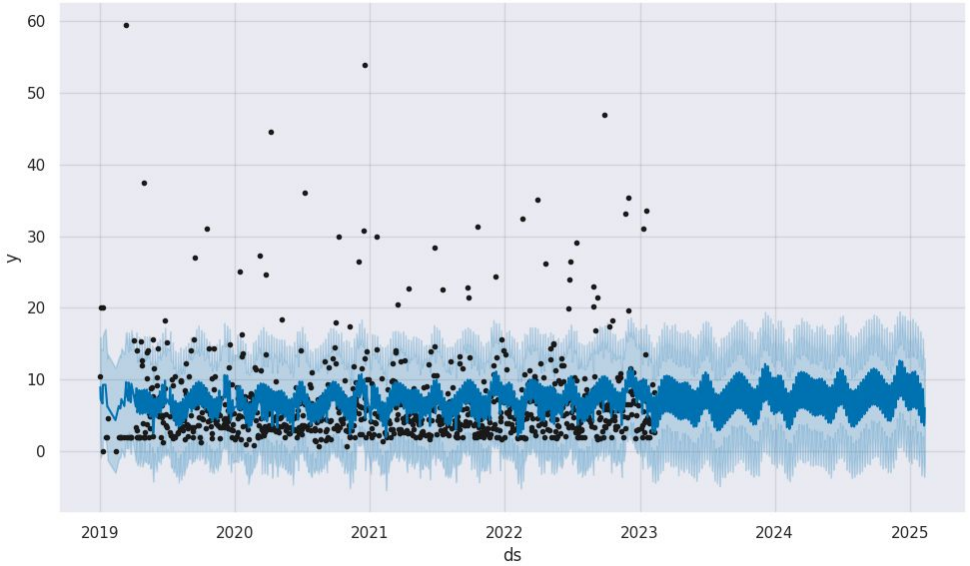
Trends for:

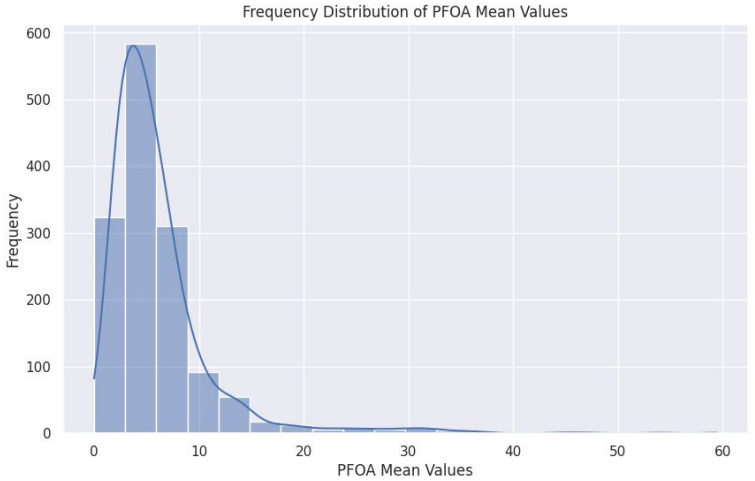
- PFOA
- PFOS
- PFHxS
- PFNA
- HFPO-DA



SAAS

PFOA



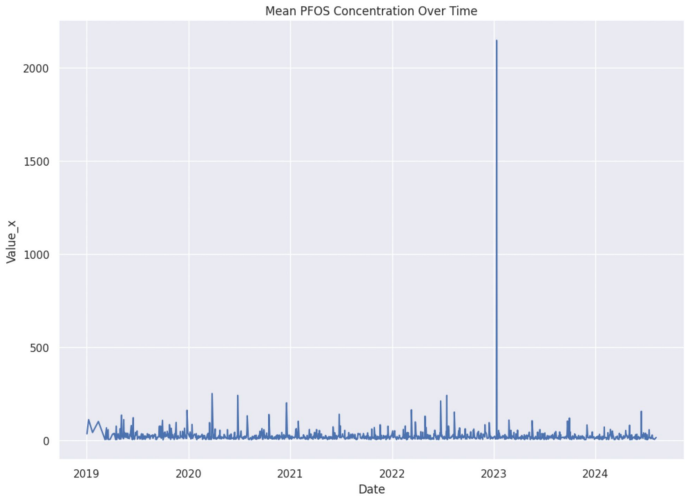


Aggregating by median
MAPE (Mean Absolute Percentage Error): 100.85%
capped at value of 80

SAAS

PFOS

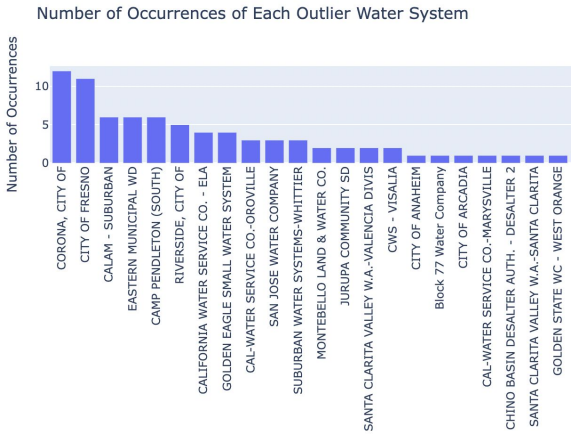
From 2019-2024, we noticed mean values of PFOS had many abnormally large values. This greatly affects how we forecast future trends of PFOS levels.



- Including outliers in our data led to predictions of annual spikes in PFOS levels
- *Not* including outliers in our data led to predictions of more stable PFOS levels

SAAS

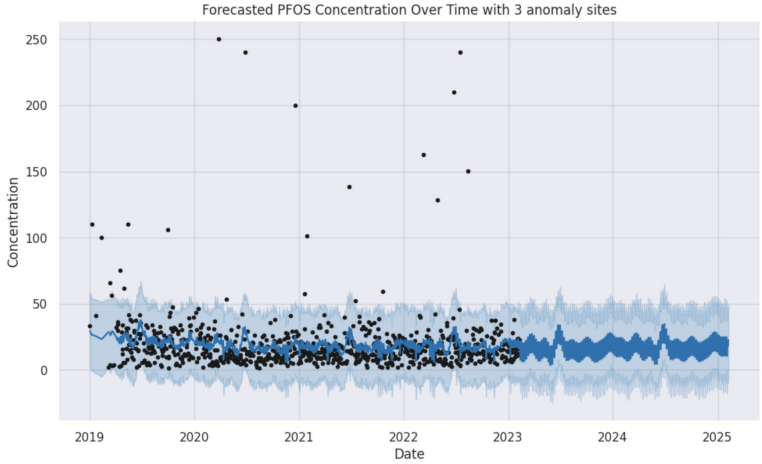
PFOS



Identified water systems that have abnormally large values.

- > 3 occurrences of abnormal values could be indicative of a trend

Forecast Model:

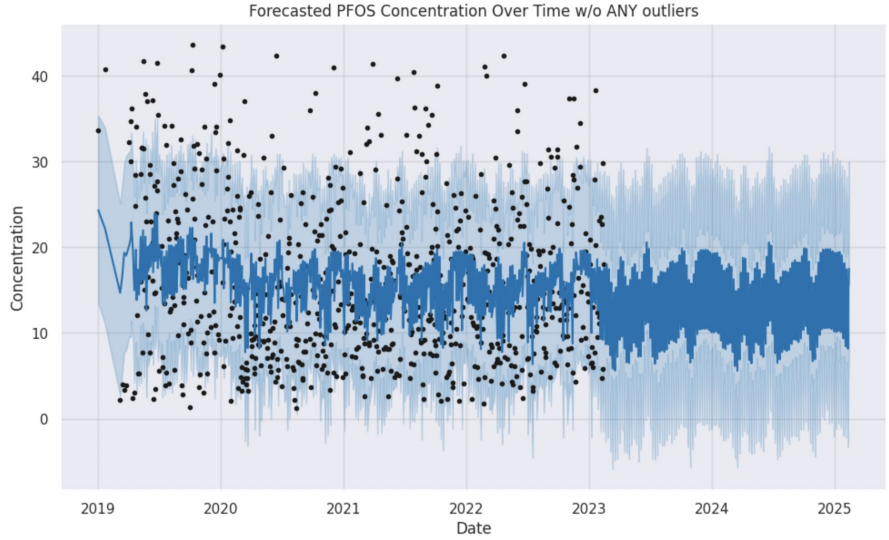


Mean Average Percentage Error: ~112.65%

SAAS

PFOS

Forecasting model using data with all outliers removed:

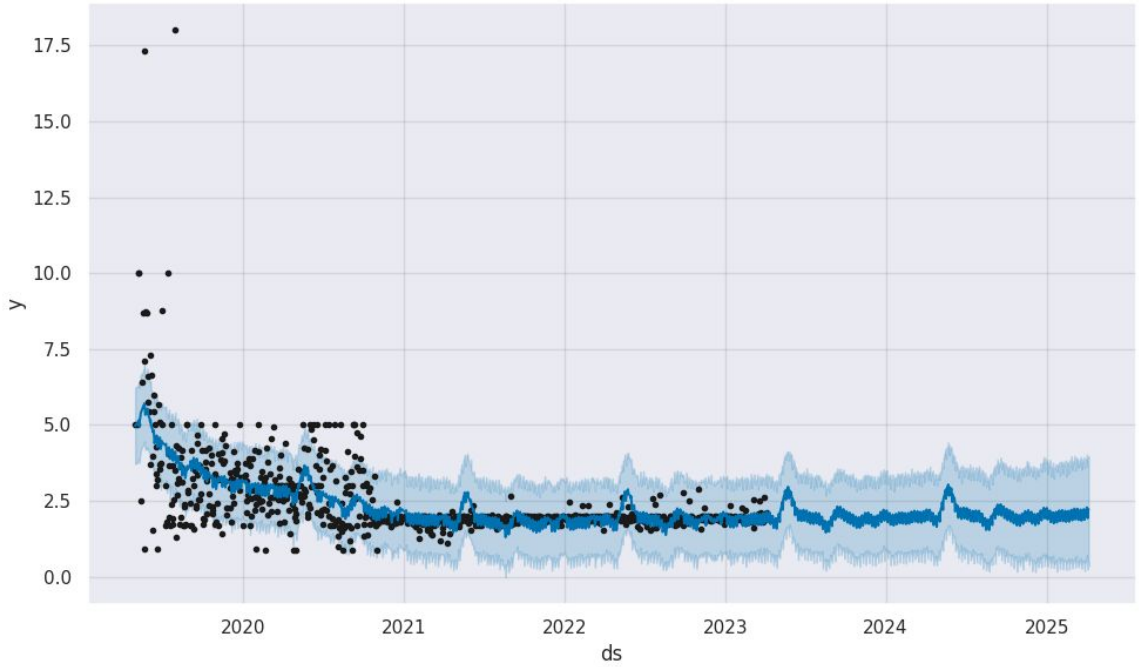


Mean Average Percentage Error: ~68.2%

Better! Cross validation can help us continue to fine tune this going forward.

SAAS

HFPO-DA



Median
MAPE: 10.73%
No cap needed

SAAS

Takeaways From Using Prophet

General Takeaways

- Significant peaks occurring around mid-year for each of the top 5 chemicals
- Challenges in predicting volatility of trends with infrequent yet disruptive abnormal levels of a few of the chemicals

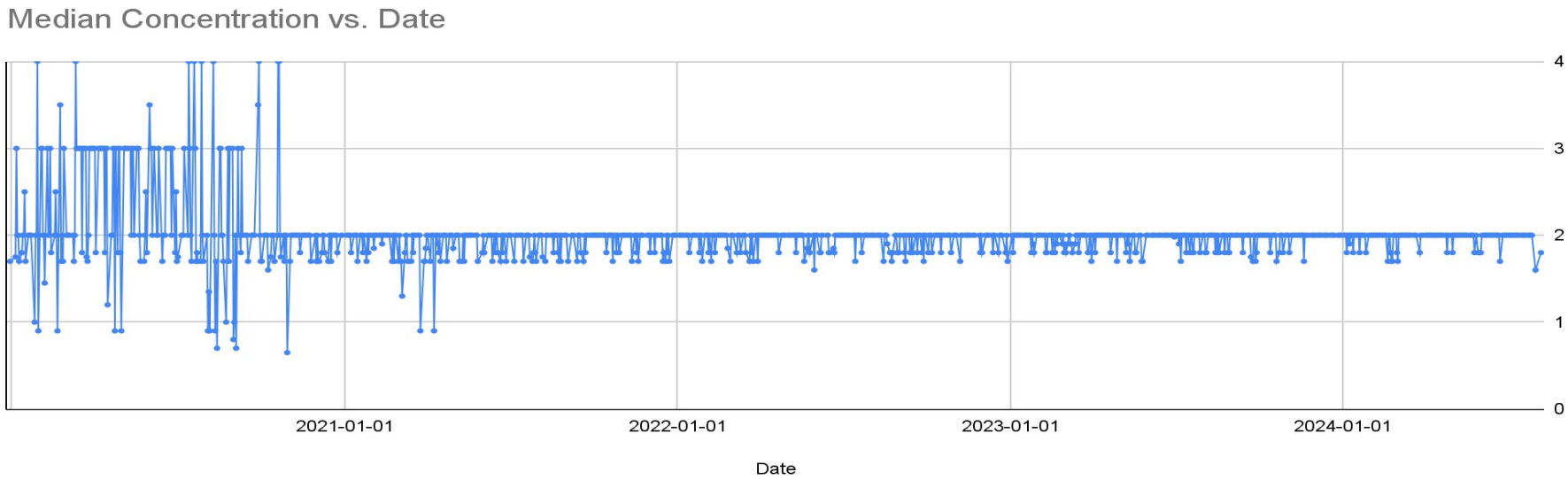
Next Steps/Goals

- Finetune the accuracy of our models using Prophet's features and cross-validation
- Focus on forecasting and predicting chemical levels for certain geographic areas and water system sites only
- Get a better understanding of what appropriate levels are for each chemical

SAAS

Median Daily Concentration

- Taking a step back, we decided to look at the data as a whole, aggregating the concentrations of all five PFAS chemicals of concern.



SAAS

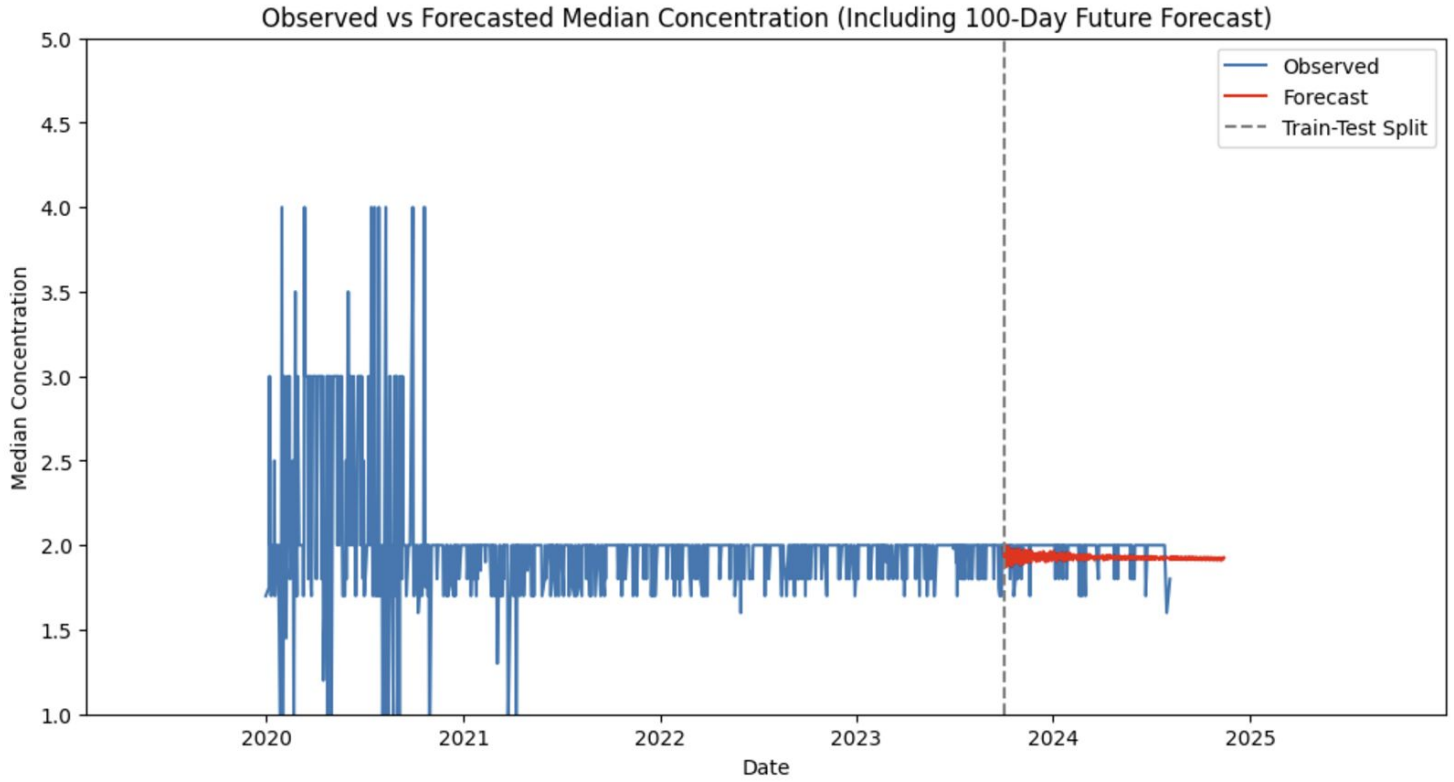
SARIMAX

- Traditional but simple time-series model for modeling and forecasting data
  - Incorporates seasonality and exogenous variables (unexpected events such as wildfires)

- Uses past values + forecasted errors to predict future values

SAAS

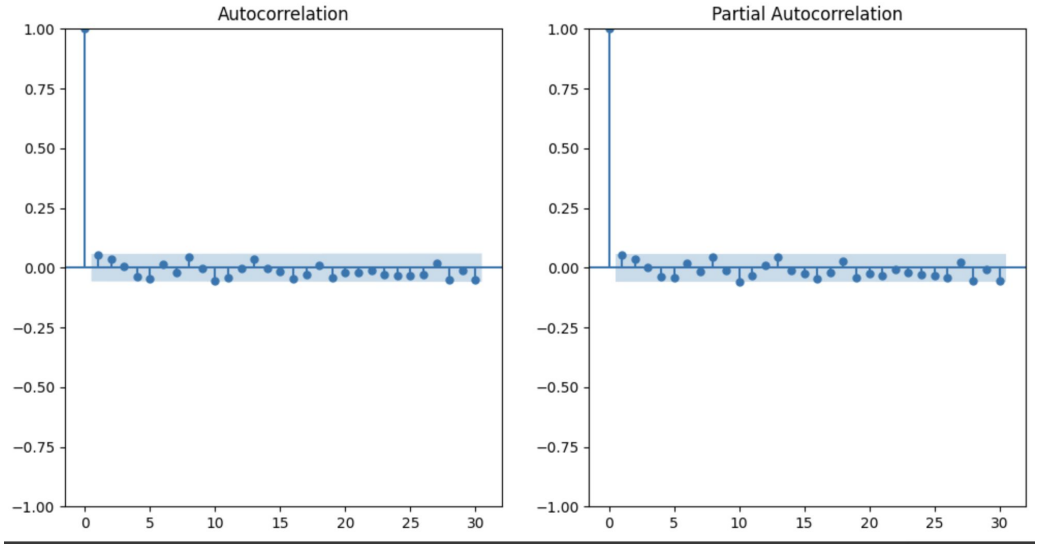
SARIMAX Model – Median



SAAS

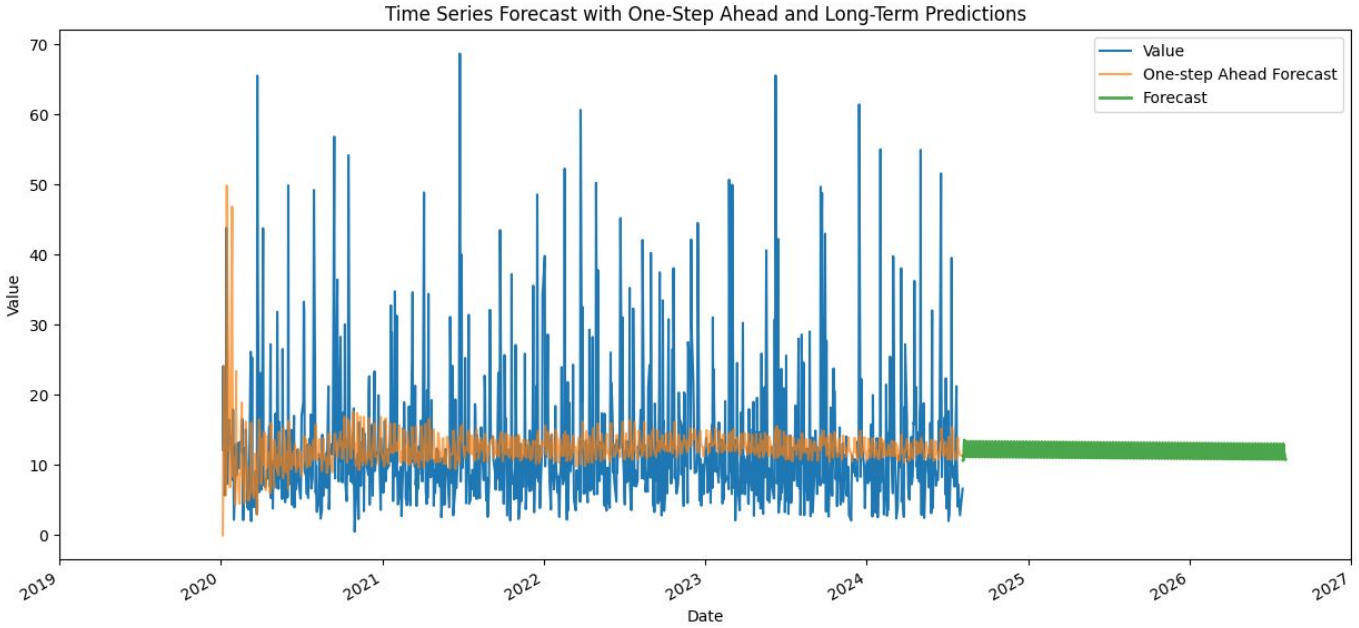
ACF and PACF

- We use the **Autocorrelation Function Plot** to tell the correlation between the y (response) variable with its own lags to determine the ‘p’ value to use when making the SARIMAX model.
- The **Partial Autocorrelation Function Plot** tells us how much of the correlation is unexplained by any shorter lagged values.



SAAS

SARIMAX Model - Mean



SAAS

XGBoost

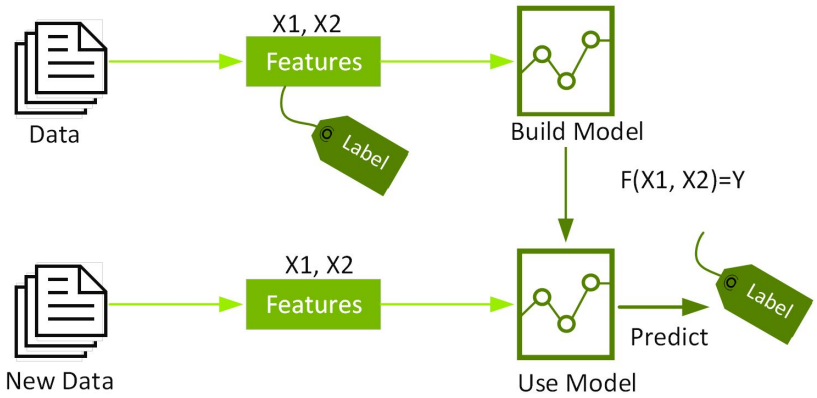
- Tree-based supervised learning method with high performance for time series

Advantages over SARIMA:

- Non-linear relationships
- Feature importance

Disadvantages:

- Overfitting (not generalizing well to unseen data)



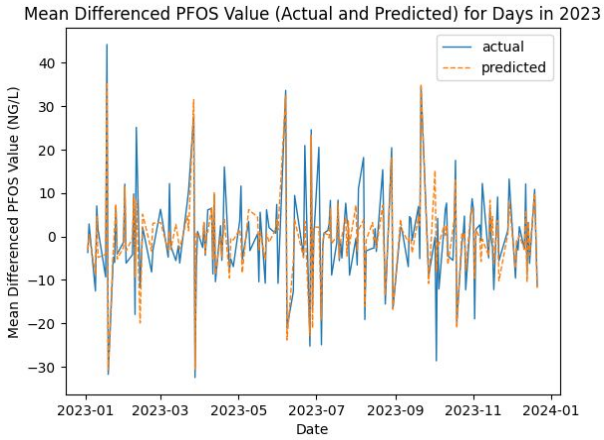
(Diagram source: NVIDIA)

SAAS

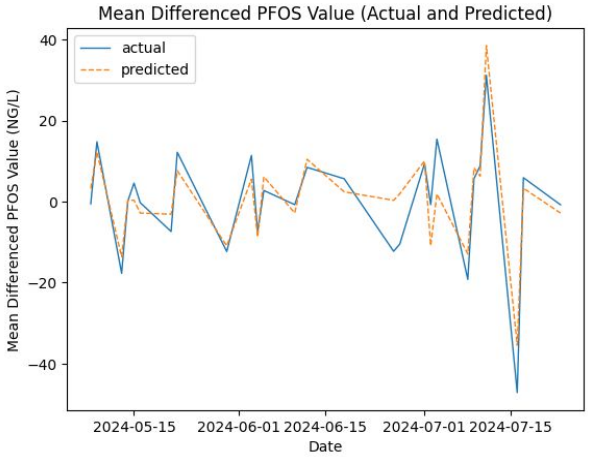
Forecasting Differenced Values

- The differenced value for day x + 1 is the value at day x + 1 minus the value at day x
- Why difference? Make time series stationary → improves prediction accuracy

Training Data (2023 Subset)



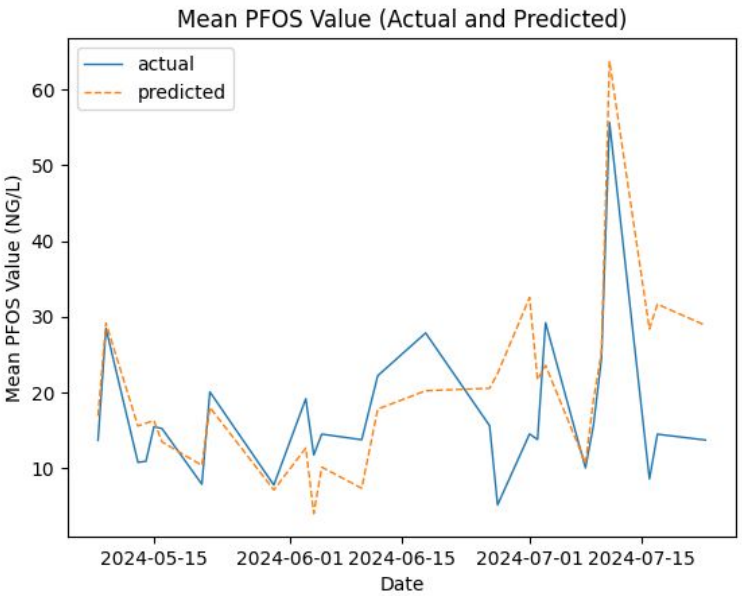
Testing (Unseen) Data



SAAS

Predicting PFOS Values

Testing (Unseen) Data:

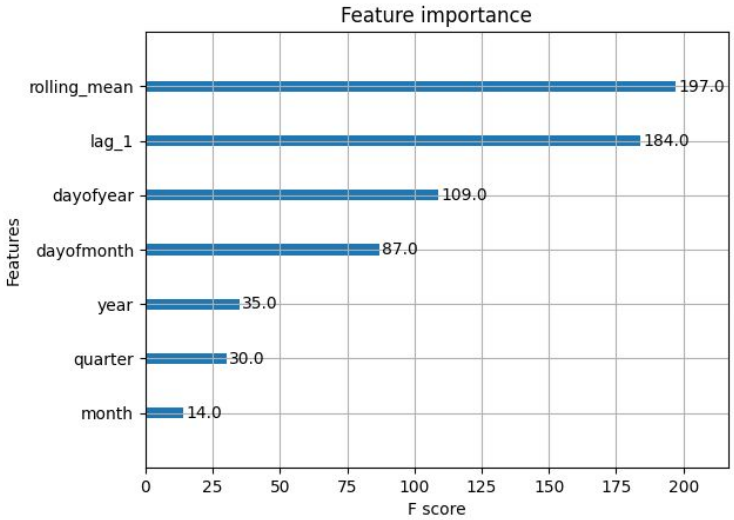


- We trained XGBoost on PFOS values from 2019 to mid 2024, and then evaluated performance on a held-out test set
- XGBoost is able to **predict spikes** with much greater accuracy than SARIMA, even for unseen data
- Rough trend is captured, but precision can still be improved

SAAS

What factors are used by our model to predict PFOS values for a given date?

- Most important predictors relate to very recent behavior
- Intra-monthly and intra-yearly patterns more important than exact year or month



The average PFOS value over the last 5 measurements

The PFOS value of the last measurement

SAAS

Conclusions

- Raw data is **messy**, but with significant **preprocessing** we enabled it to be useful for predictive modeling
- **Time series models** (particularly XGBoost) can produce accurate **predictions** of PFAS concentrations within our existing data, including **spikes & seasonality**
- **Feature importances** show that PFOS concentrations have **patterns** depending on the time of month and year

Next Steps:

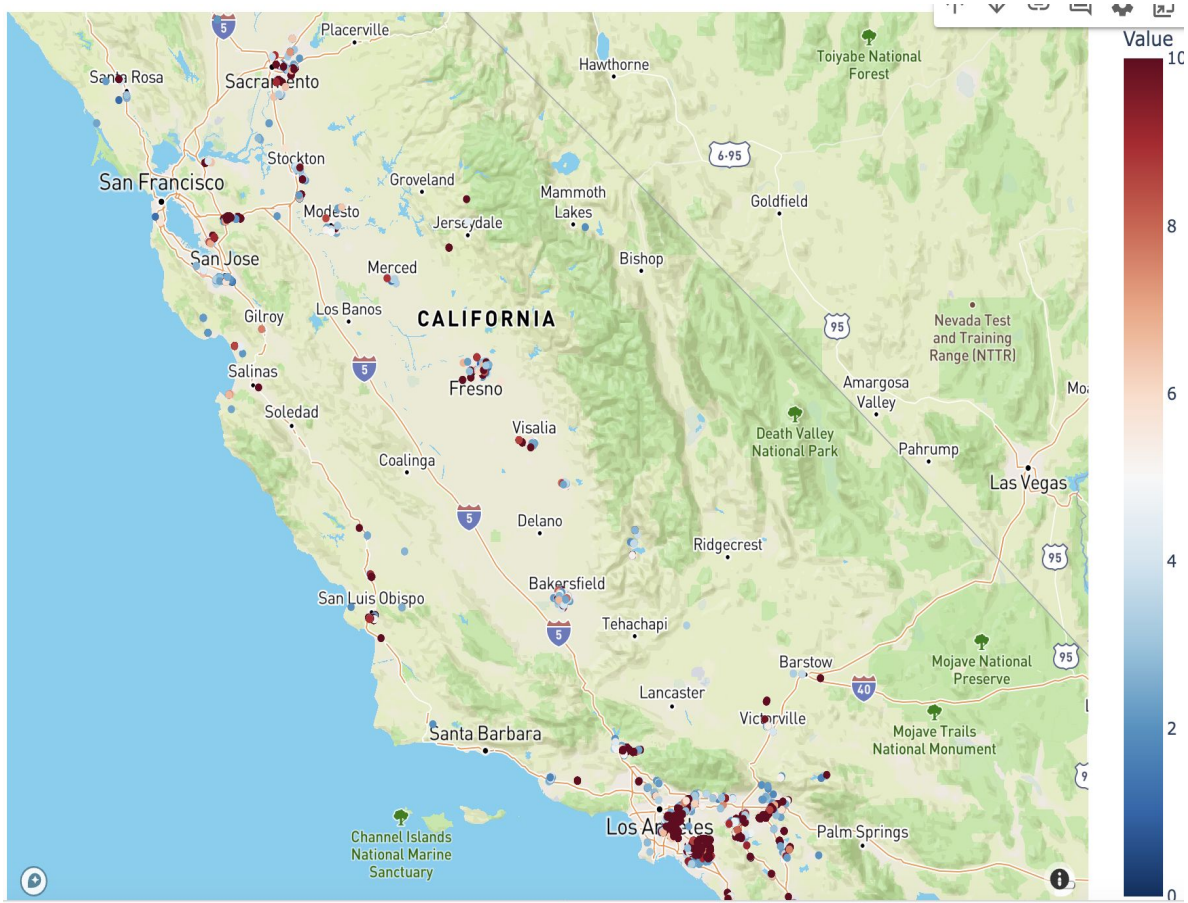
- Further model experimentation to improve forecasting accuracy
- Begin forecasting outside of our existing data: how do we expect PFAS levels to change in 2025?
- Dive deeper into feature importances: which days of the month and year are predicted to have more PFAS, and does this relate to our existing knowledge?

SAAS

Spatial Analysis

SAAS

Terrain Map — PFOS Concentration

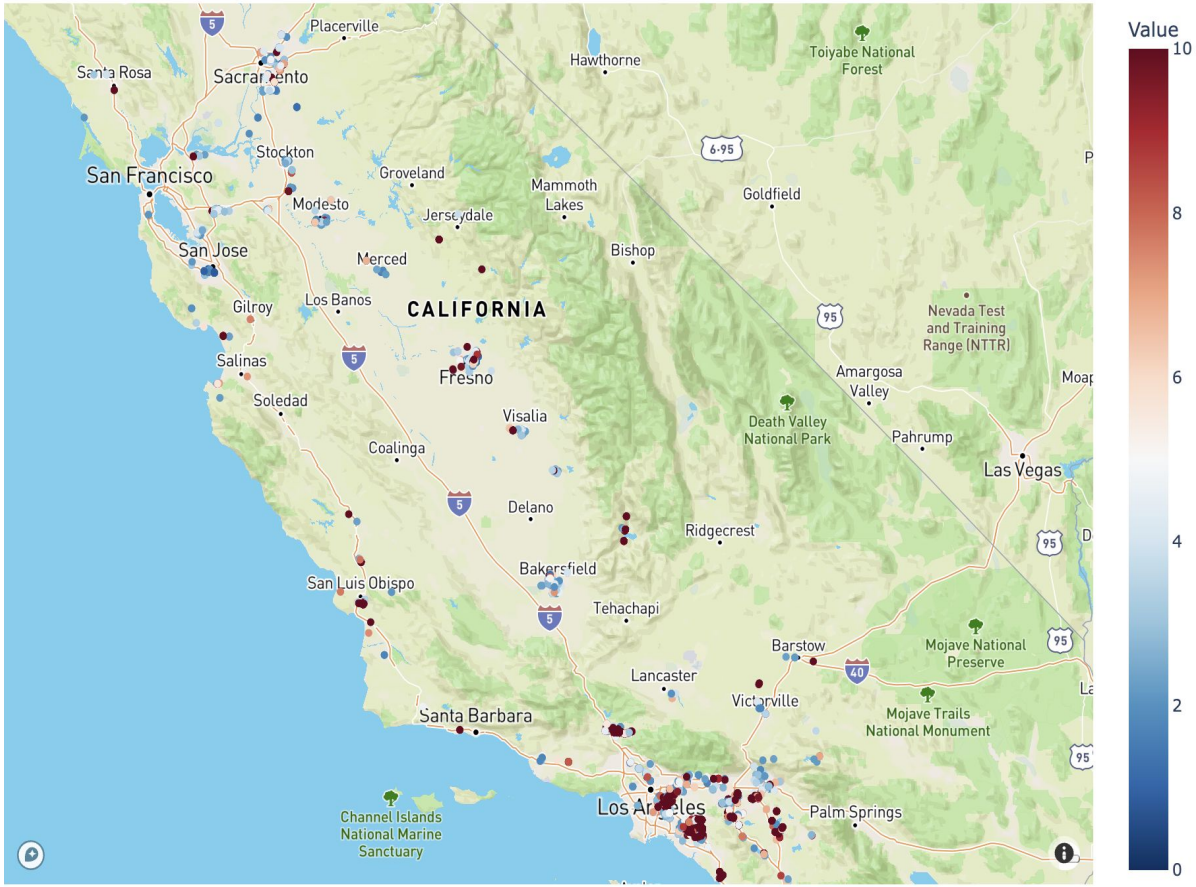


Plotly

- scatter_mapbox
- Outdoor style

High PFOS level in urban and industrial regions (ex. LA) and near bodies of water

Terrain Map — PFOA Concentration

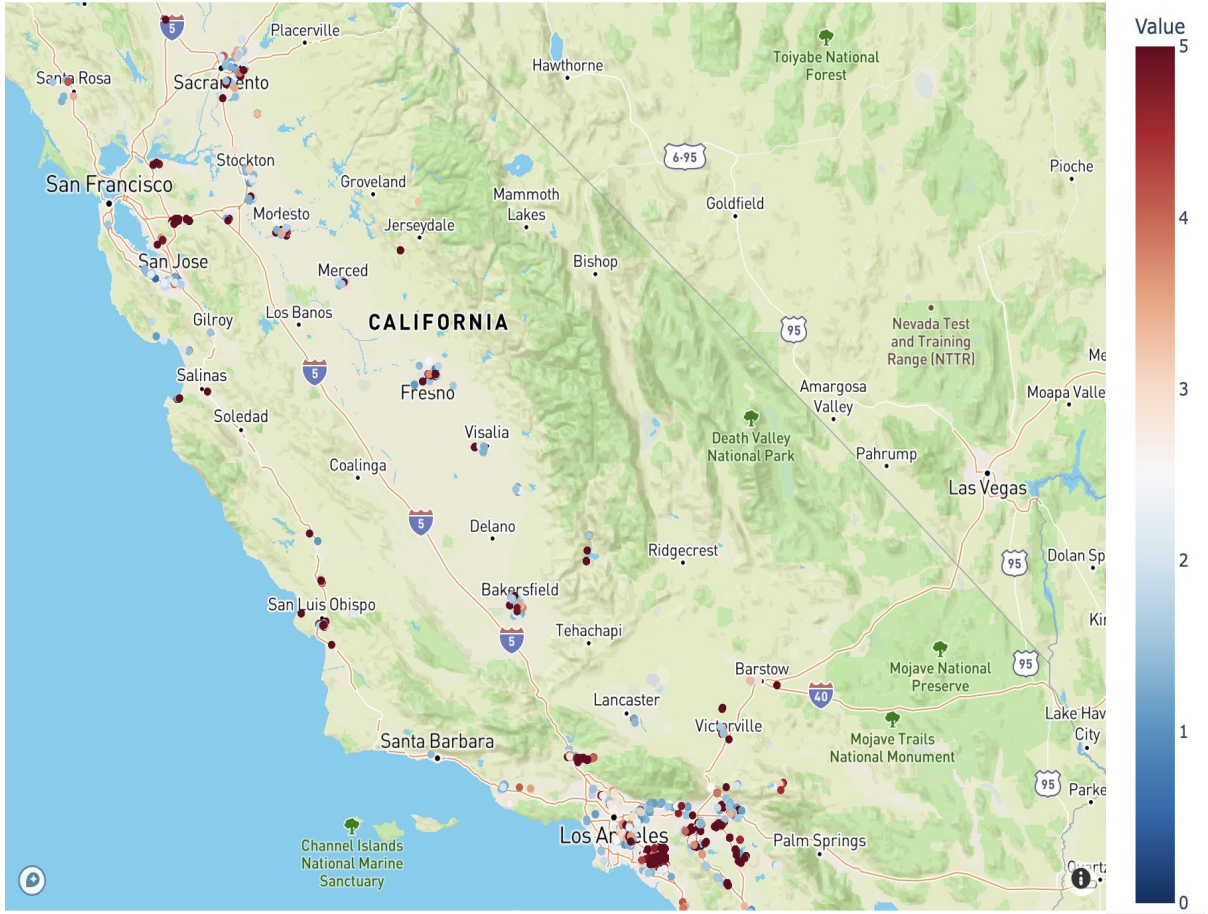


Plotly

- scatter_mapbox
- Outdoor style

High PFOA level in Southern California urban regions (ex. LA) and near bodies of water

Terrain Map — PFHxS Concentration

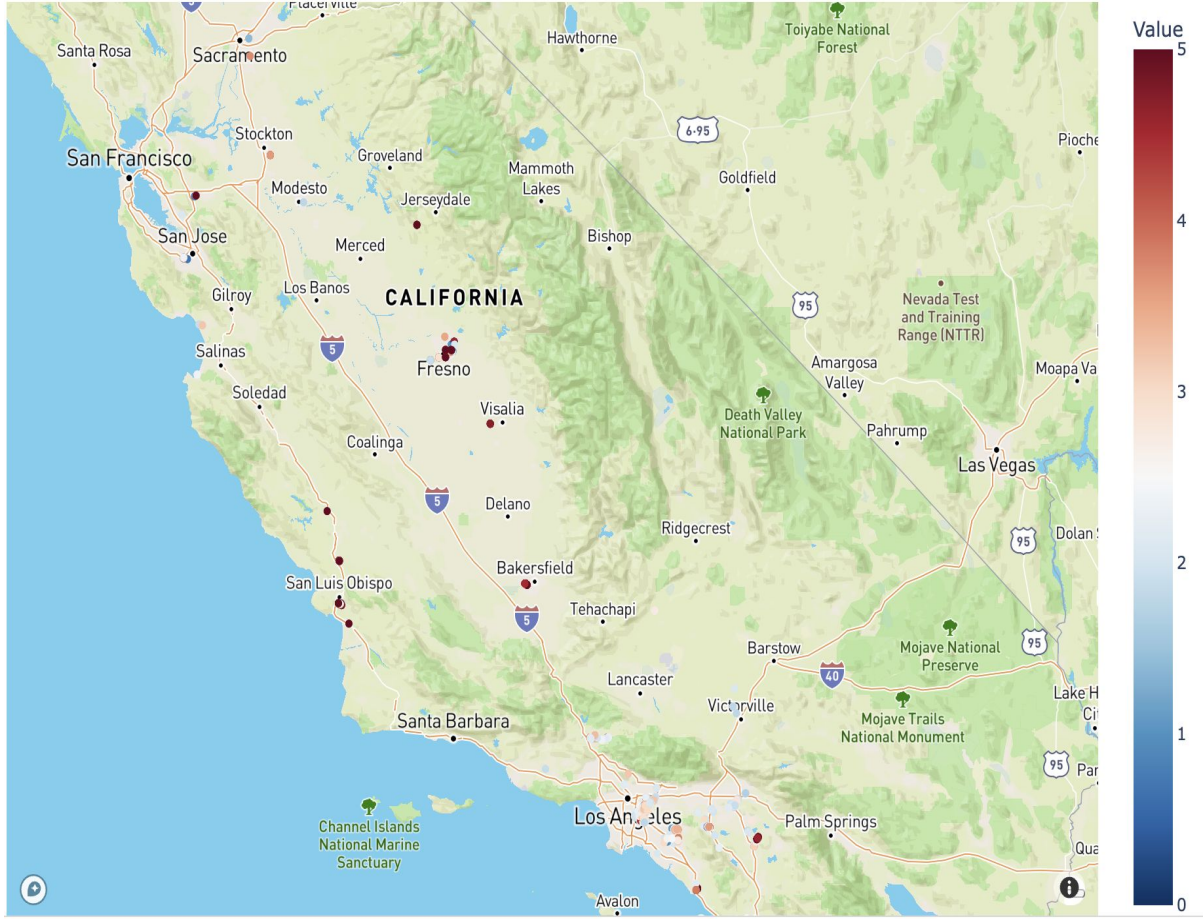


Plotly

- scatter_mapbox
- Outdoor style

High PFHxS level in Southern California urban regions (ex. LA) and Bay Area

Terrain Map — PFNA Concentration

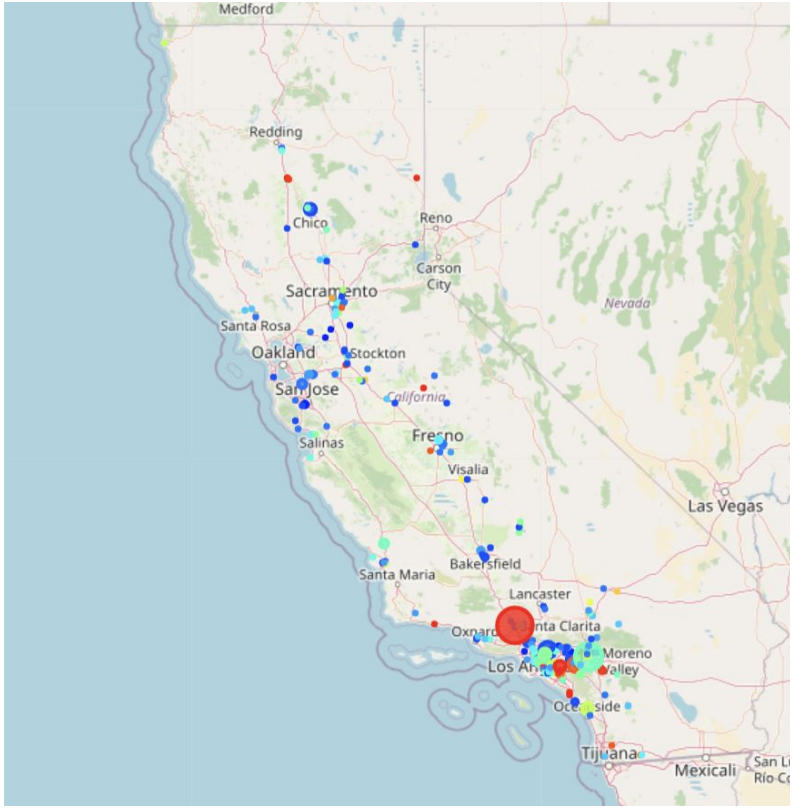


Plotly

- scatter_mapbox
- Outdoor style

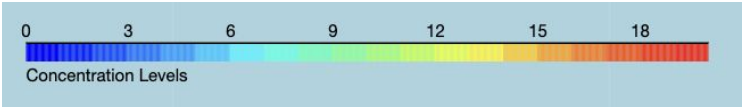
High PFNA level in Fresno and San Luis Obispo

Interactive Map — Procedure



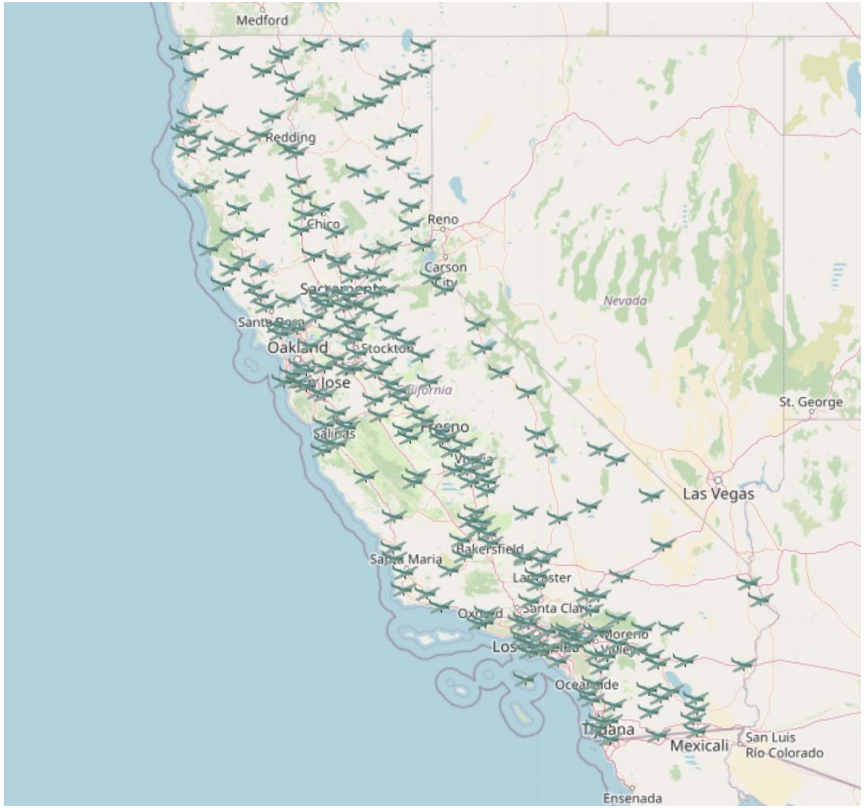
Folium

- Larger markers -> more tests exceeded limit at location
- Divergent color range (>18 ng/L is red)



SAAS

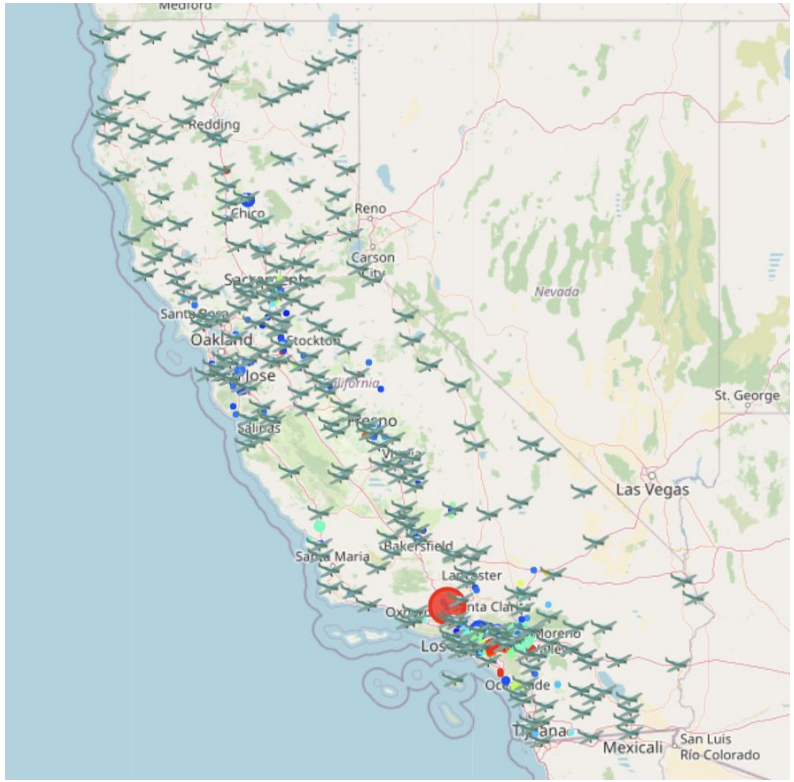
Interactive Map — Procedure

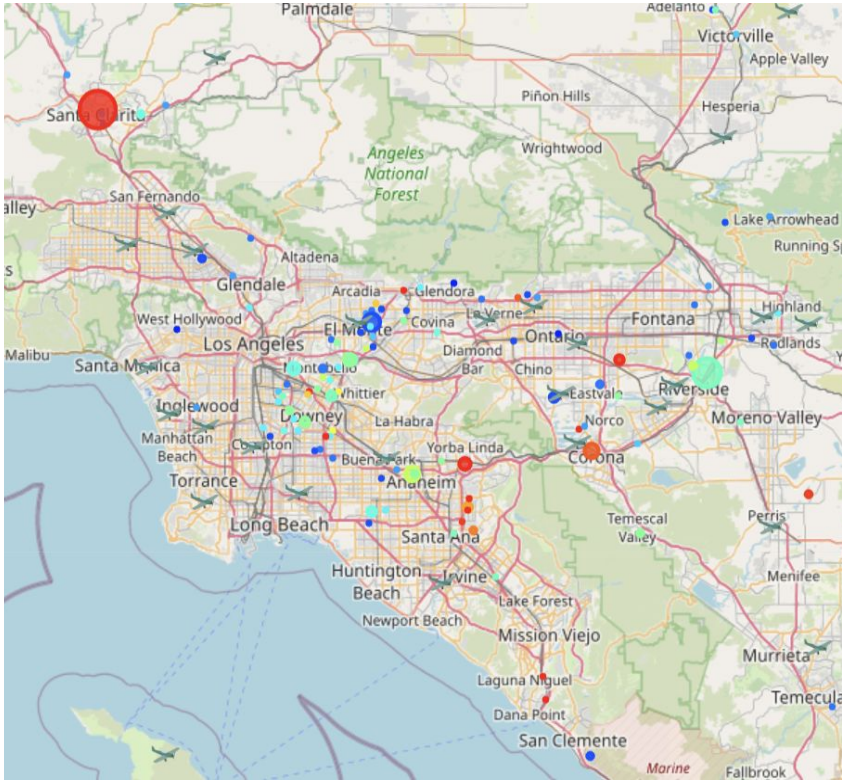


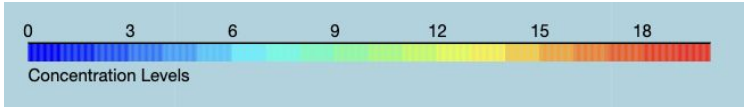
- CalTrans Dataset on airport locations in CA
- Plotted on map utilizing longitude and latitude

SAAS

Interactive Map — Procedure

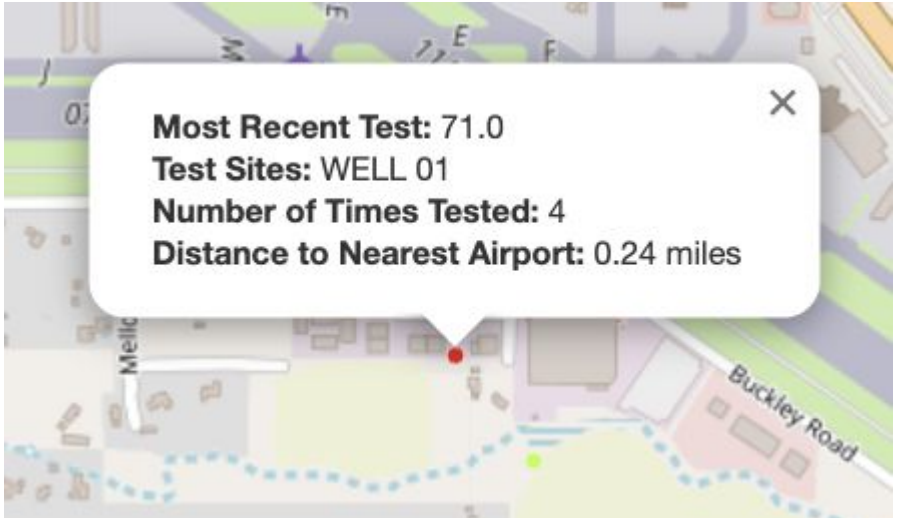


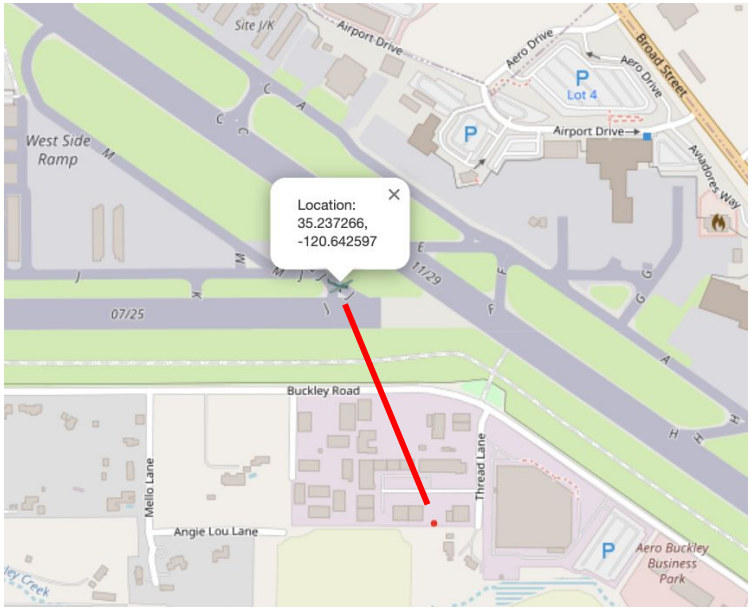


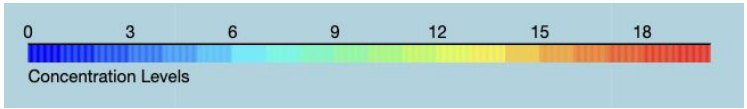


SAAS

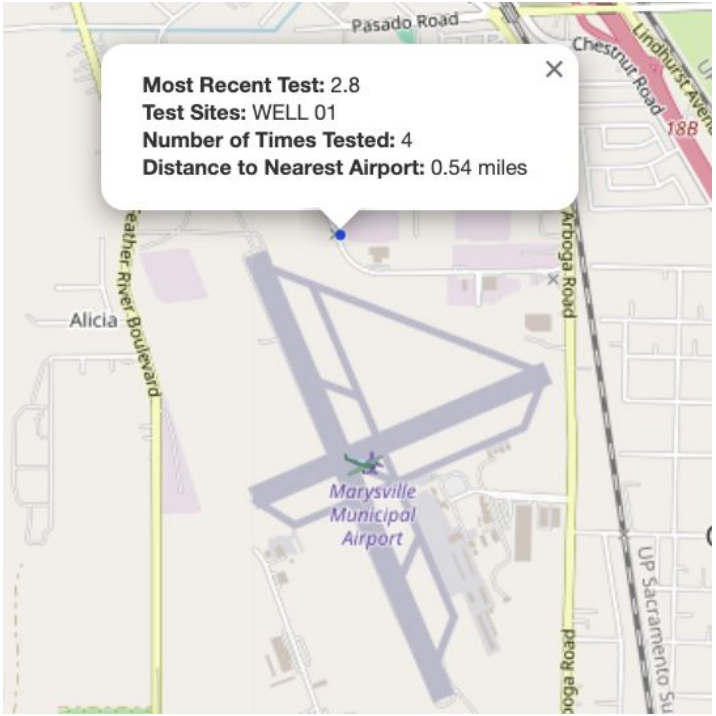
Interactive Map — Wells and Airports

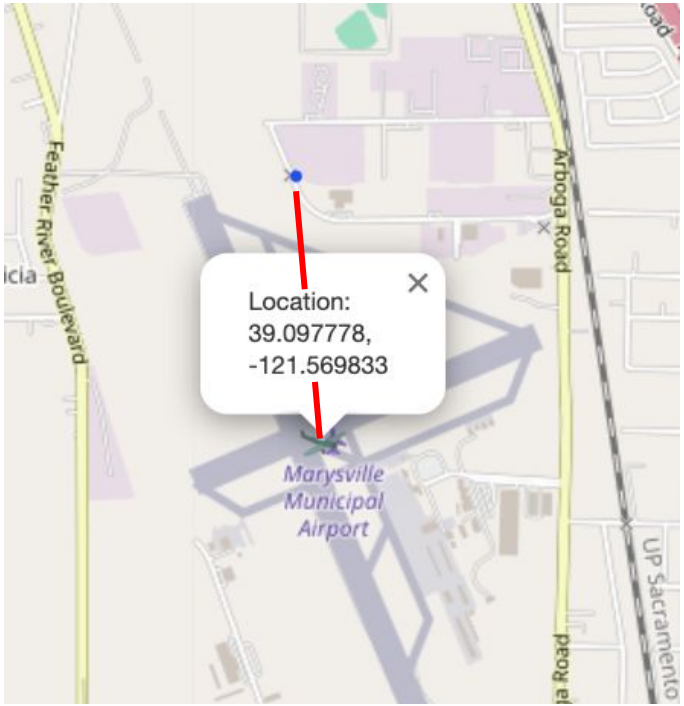


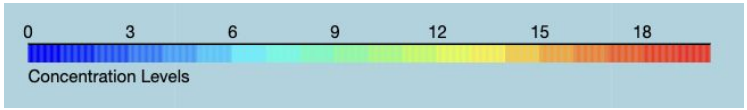




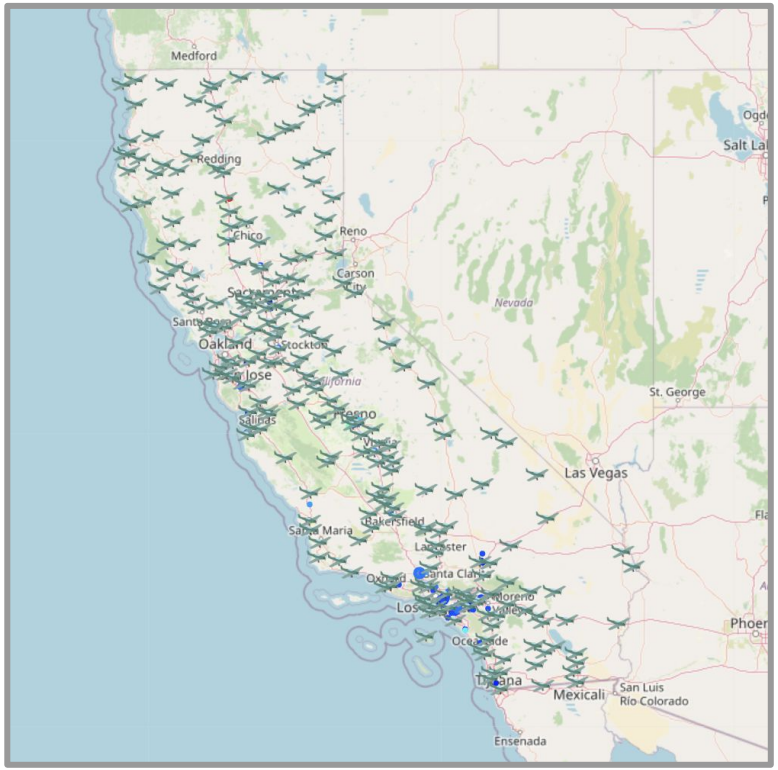
Interactive Map — Wells and Airports

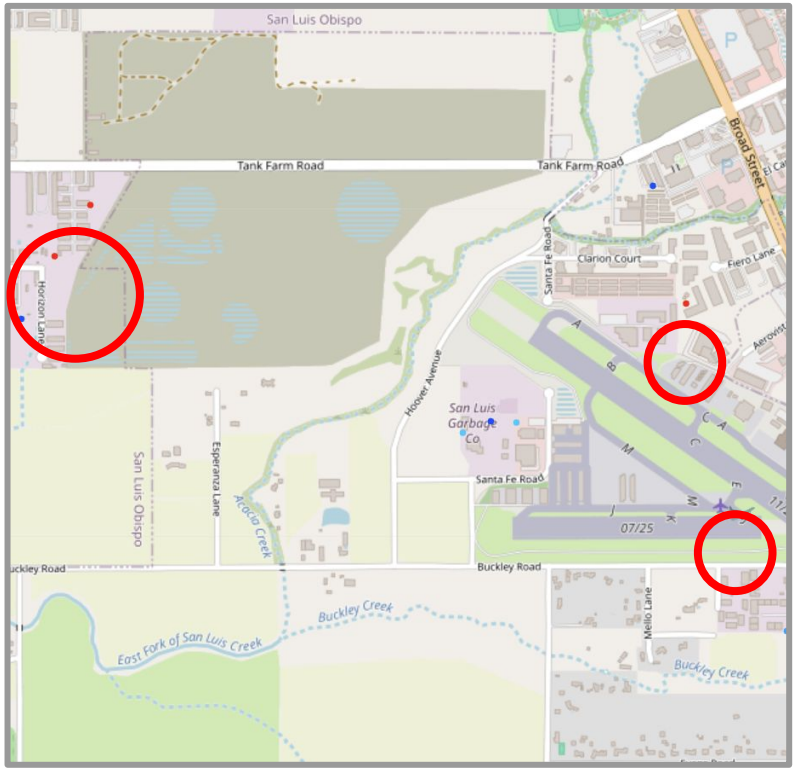


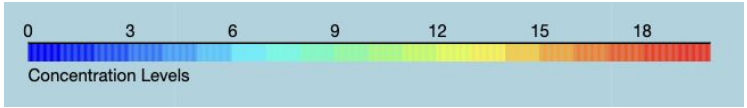




Interactive Map — PFNA

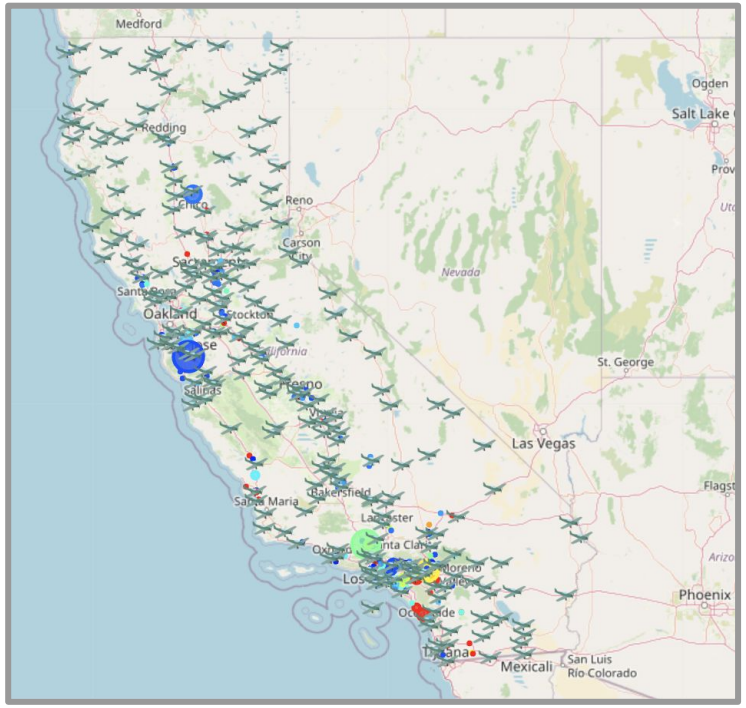


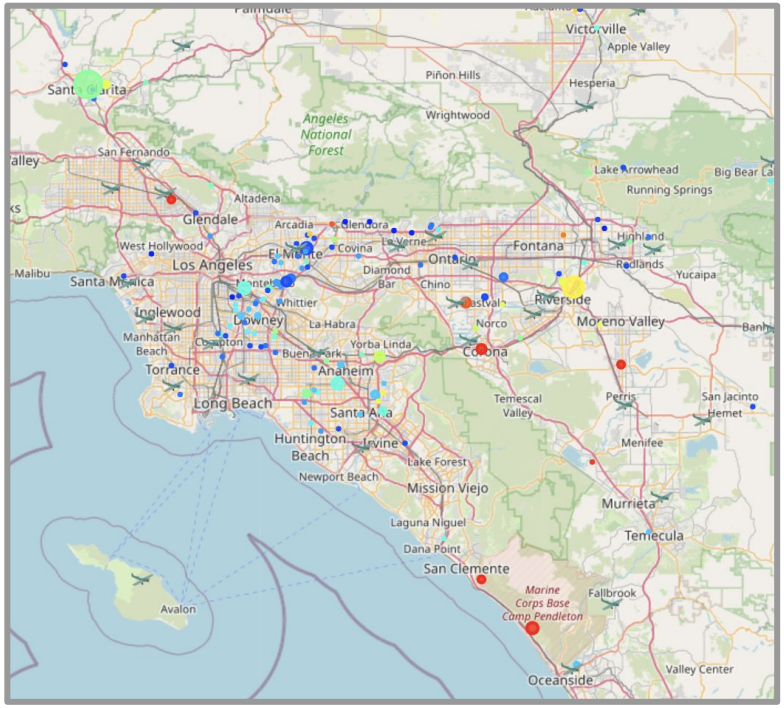


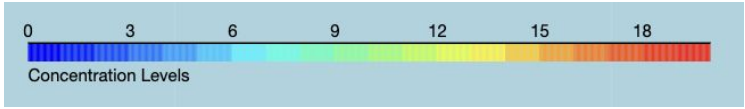


SAAS

Interactive Map — PFHxS

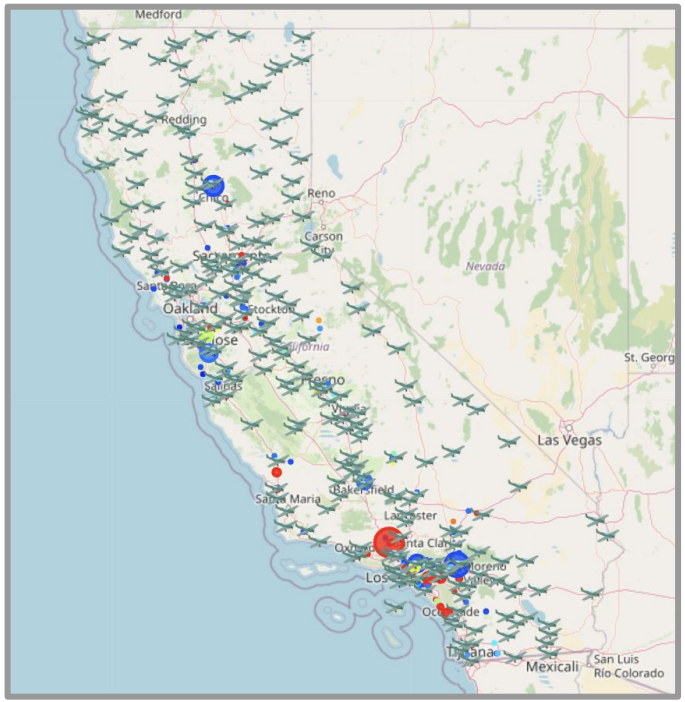


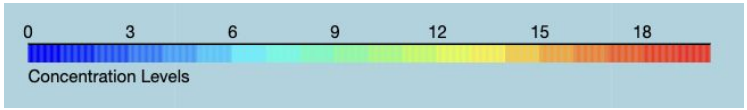




SAAS

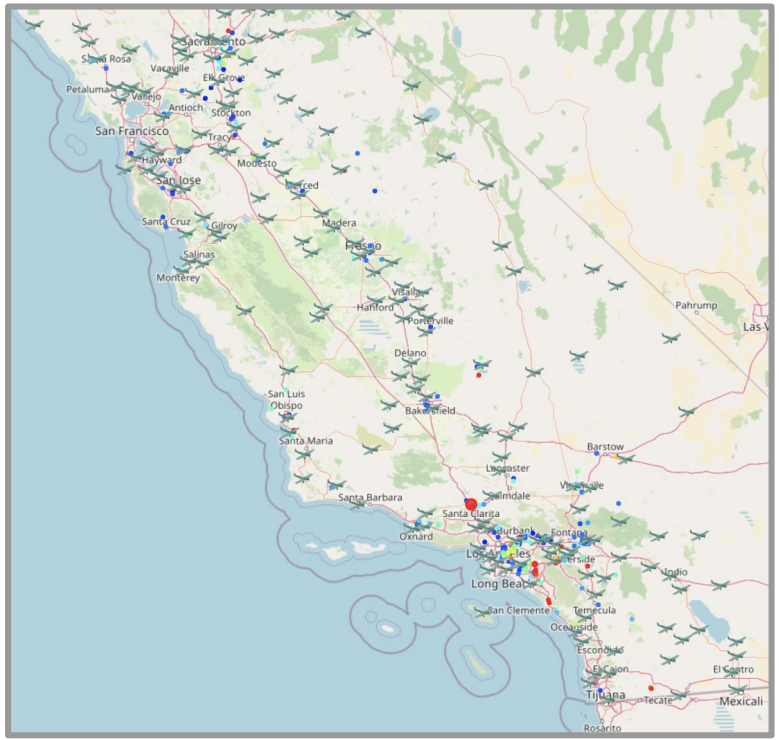
Interactive Map — PFOS

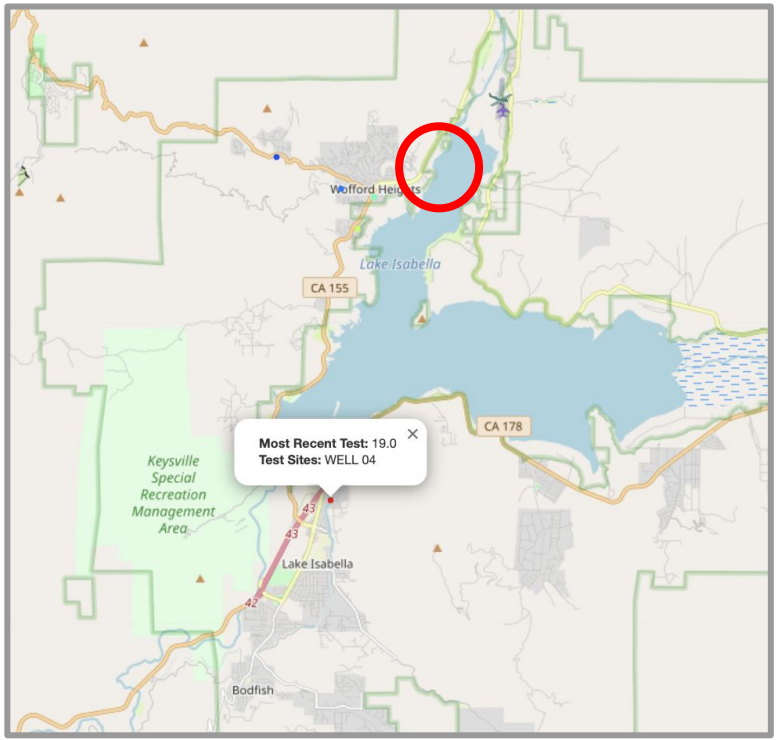


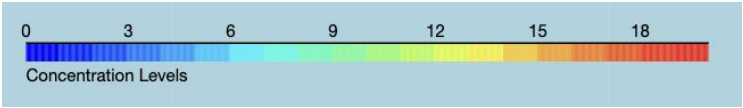


SAAS

Interactive Map — PFOA







SAAS

Interactive Map — Conclusions

- HFPO-DA has few tests
- High contamination in LA County
- Other notable areas include:
  - Chico
  - San Jose
  - San Luis Obispo
- Airports seems to be a reason
- Presence of water bodies and bioaccumulation
- PFHxS and PFOS tested most
- PFNA and PFOS seem to be most prevalent

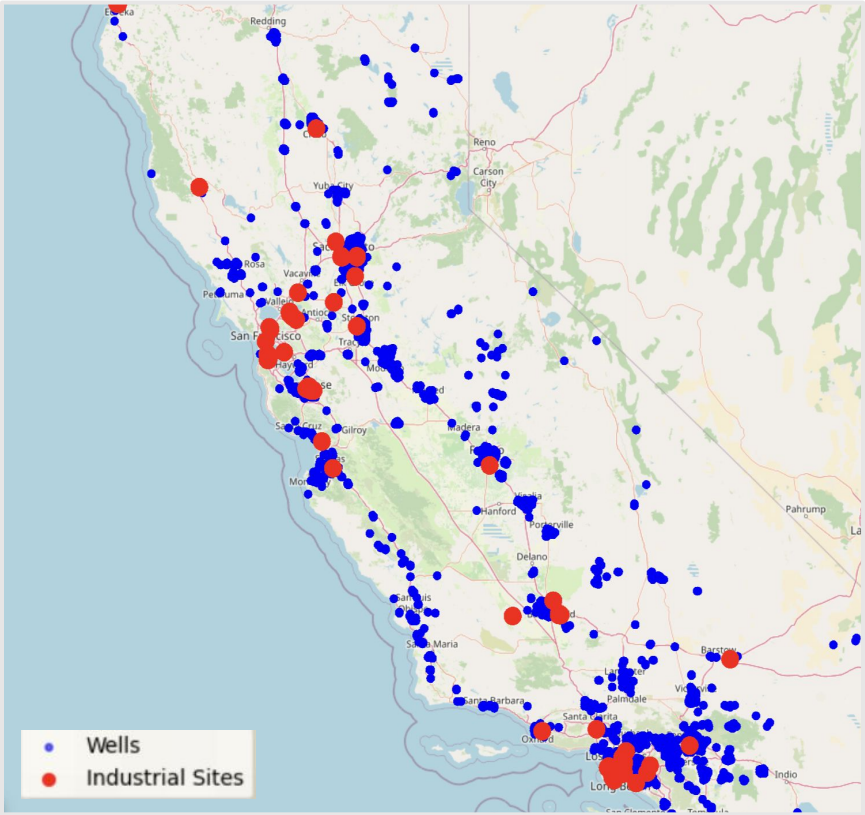
SAAS

From Interactive Maps:

- Santa Clarita Valley
- San Jose
- San Gabriel Valley
- Riverside
- Jurupa Valley
- Bakersfield
- North Garden
- Marysville
- Anaheim
- Clovis
- Camp Pendleton
- Stockton
- Chico
- Alameda

SAAS

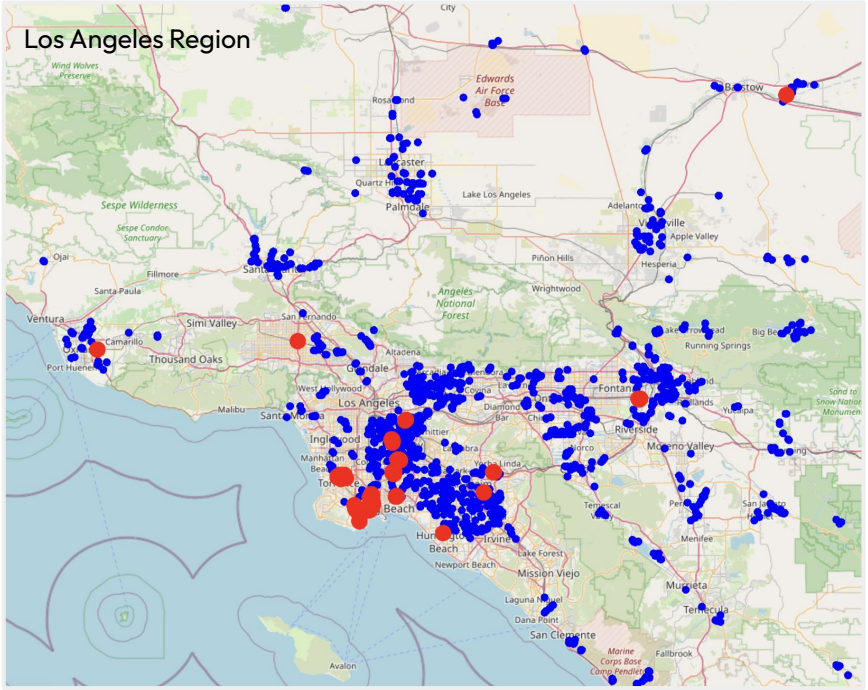
Interactive Map — Industrial Site Distance to PFAS- Detected Wells

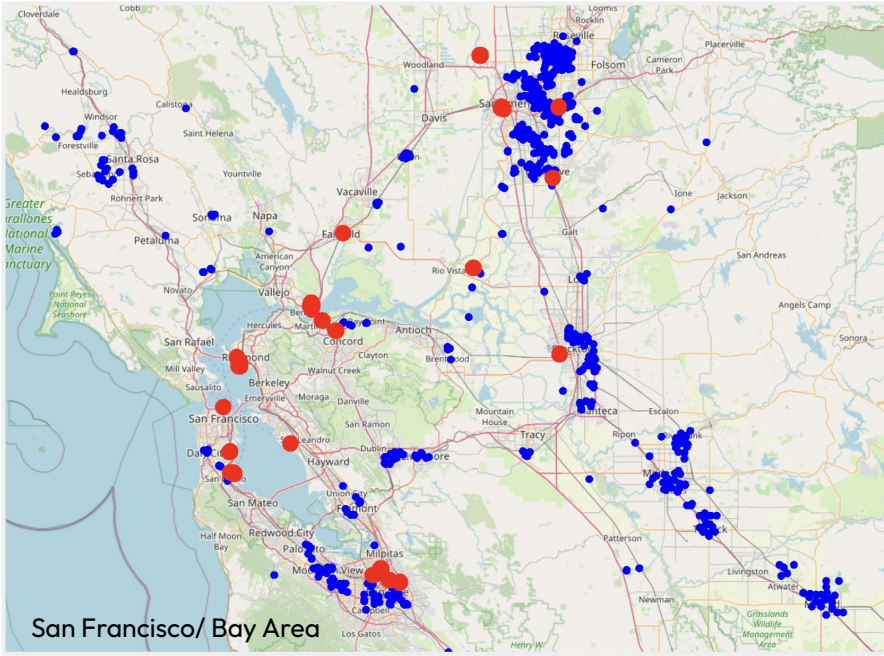


- The red dots, representing industrial sites, are located near urban centers
- Strong correlation between industrial sites and nearby wells
  - May point to areas where pollution risks are greater due to the density of industrial activity.
- Most concentrated along the coast

SAAS

Zooming In On Industrially Impacted Zones

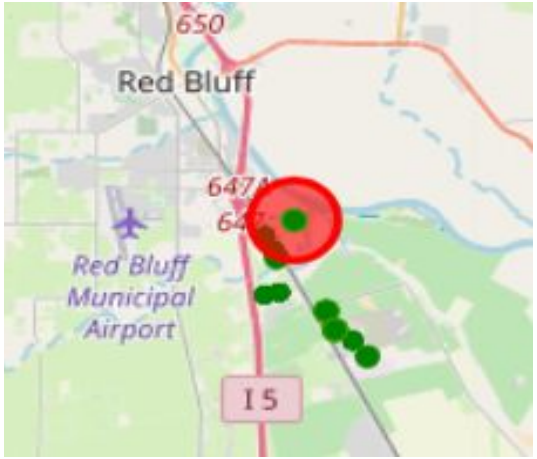




- Bay Area refineries, industrial chrome plating, tech and manufacturing hubs
- Disproportionate exposure in communities near industrial zones

SAAS

High Risk Areas — PFOS

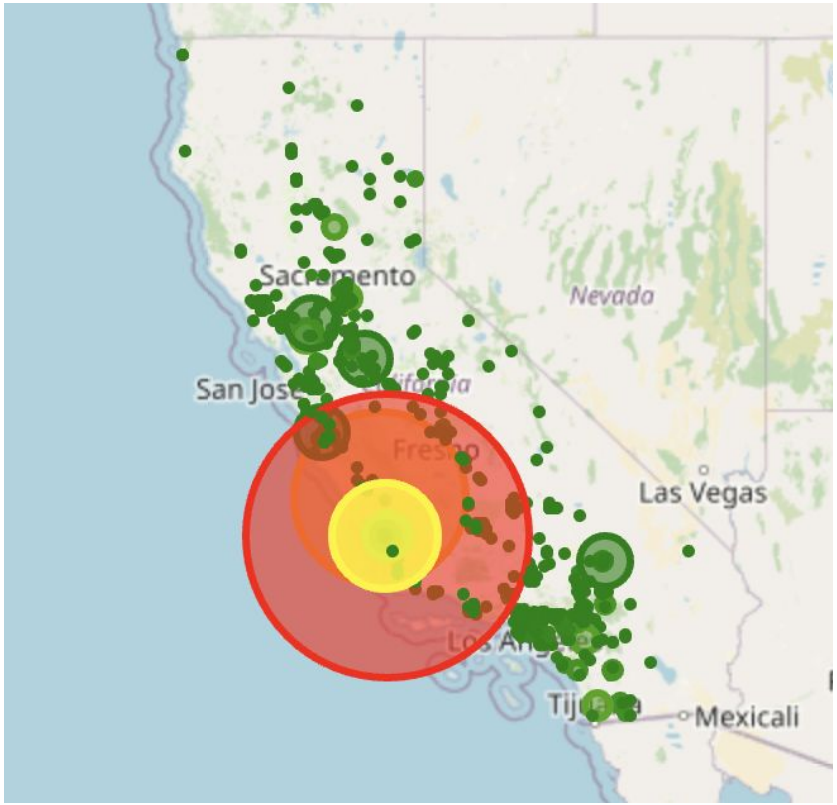


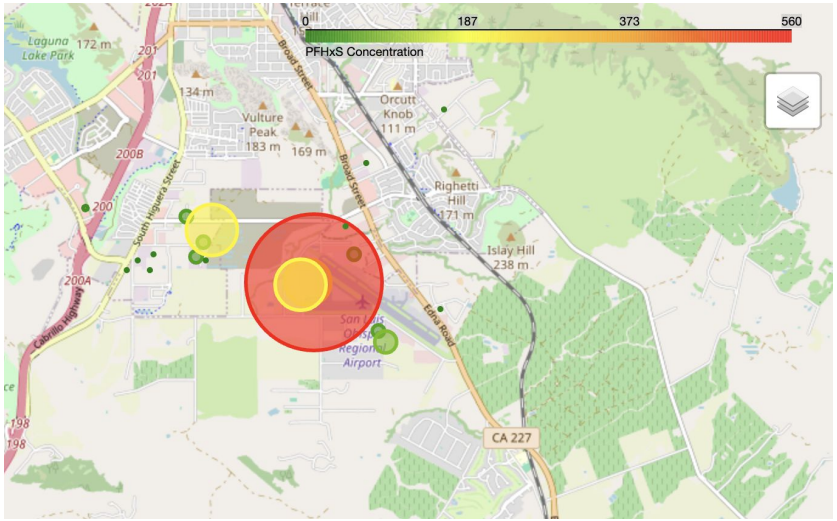
- Close to Reynolds Consumer Products industrial facility, which has a history of PFAS contamination

- Military installation that has high levels of PFOS due to historical use of aqueous film-forming foam (AFFF) in firefighting training and emergency responses

SAAS

Interactive Map — PFHxS Concentration Levels Across CA



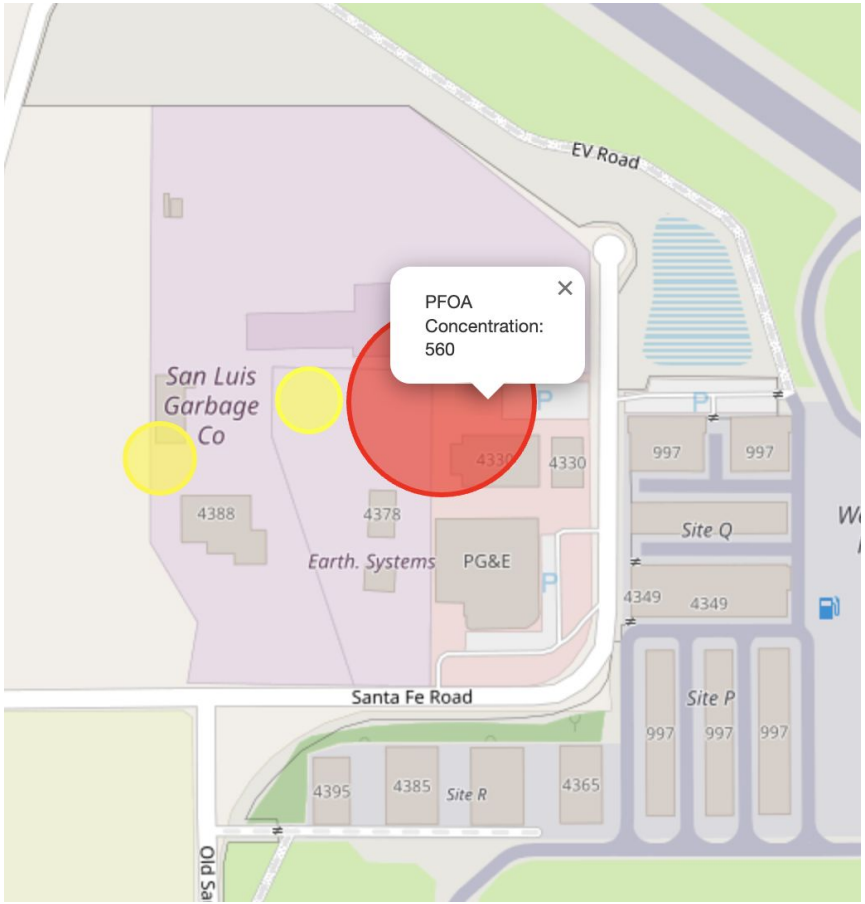


- Overall, concentrations remain consistent throughout California, with the exception of two distinct hotspots.

SAAS

High Risk Areas — PFHxS

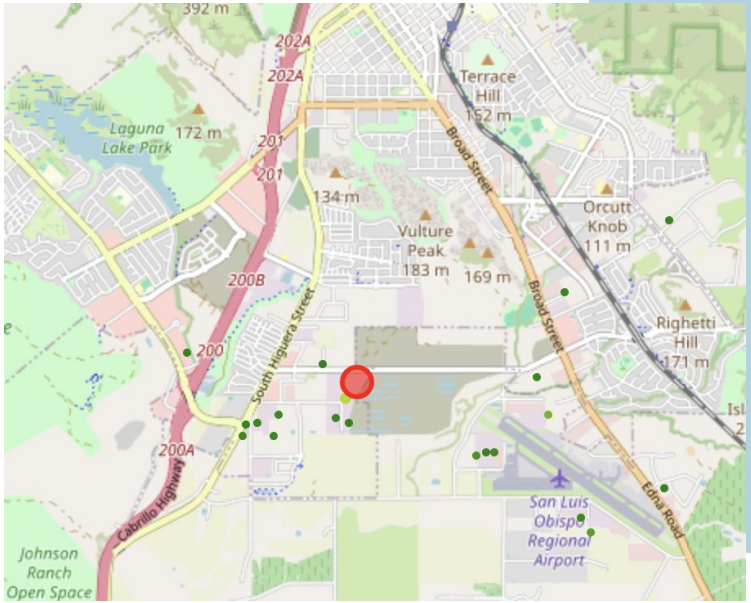
- Highest concentration of chemical in California found at San Luis Garbage Co. (560)
- Nearby landfills lacking proper leachate management could release PFAS from airport materials, firefighting debris, or equipment waste
- PG&E Diablo Canyon Power Plant: engages in industrial-scale electricity generation.

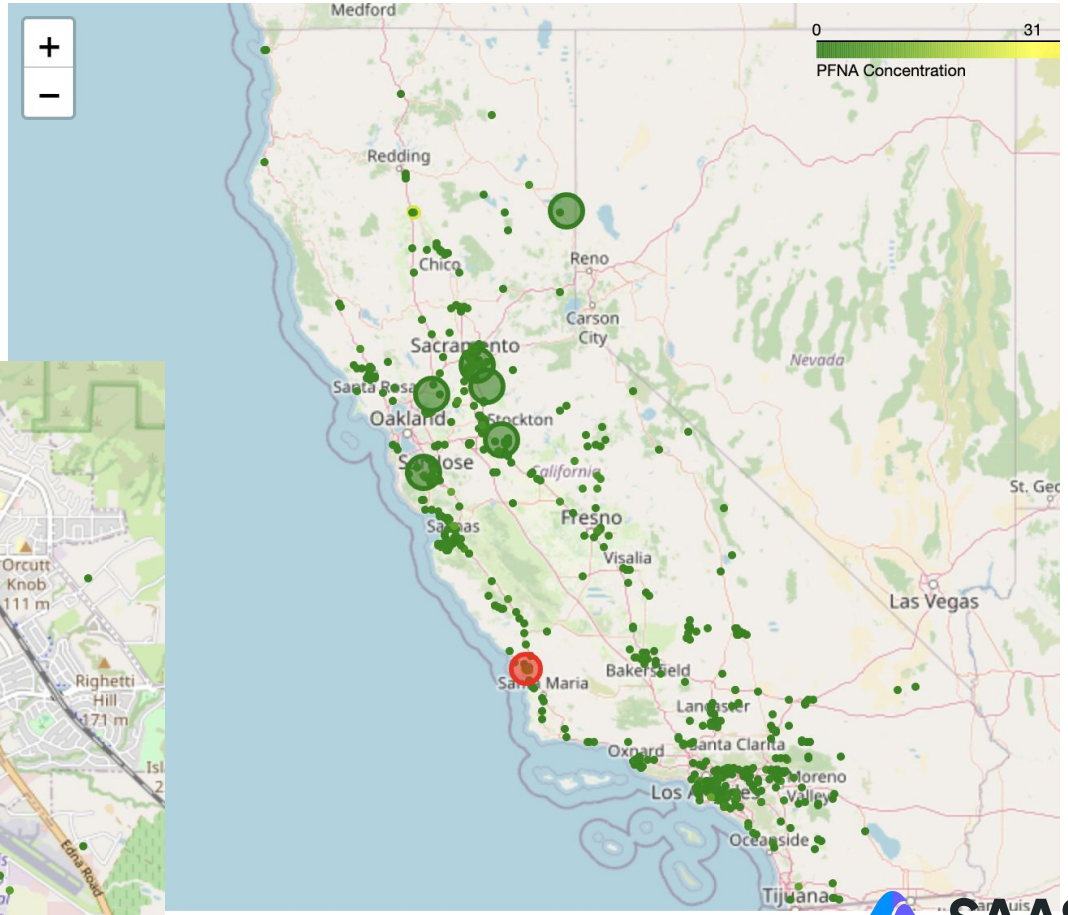


SAAS

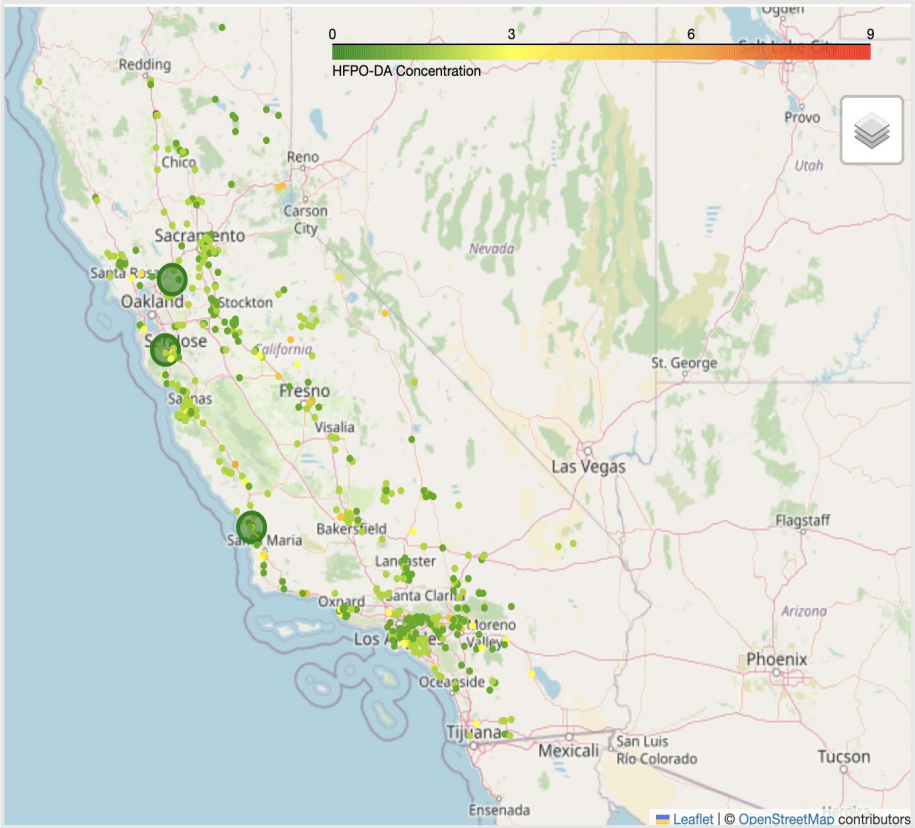
PFNA

- High concentration of PFNA located in San Luis Obispo
- Reason: airport and farms are nearby





Interactive Map — HFPO-DA Concentration Levels Across CA

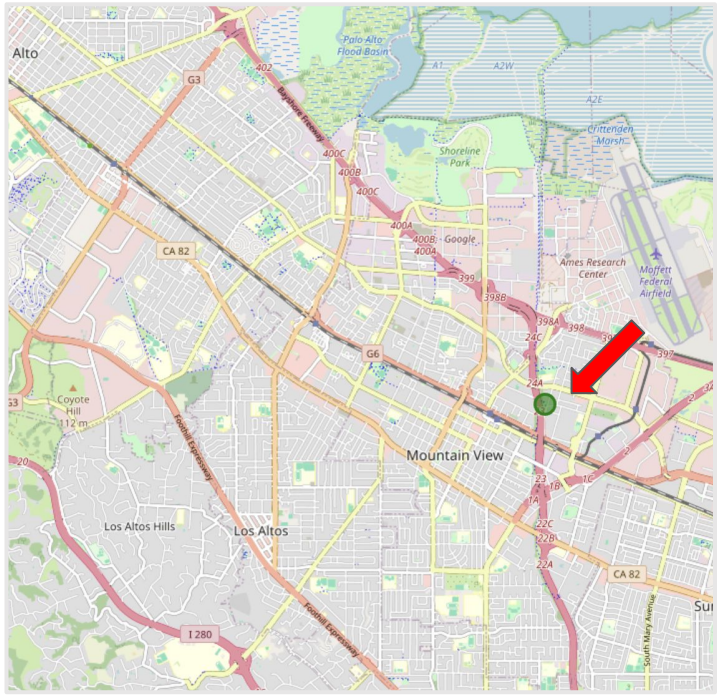


- The majority of markers are smaller
  - Less data
  - Does not appear in high concentrations statewide

- Although California as a whole experiences lower concentrations of HFPO-DA…
  - Small spikes near military bases, and industrial activities are still present.

SAAS

High Risk Areas — HFPO-DA



Tech and semiconductor industrial area- Use chemicals and coatings, including fluoropolymers for making circuit boards and electronic components

Moffett Federal Airfield & Ames Research Center- involve the use of industrial-grade chemicals and materials

SAAS

Interactive Map — Conclusions

- Strong correlation between industrial sites and nearby wells
- Causal zones
  - Bay Area refineries
  - Industrial chrome plating/ tech and manufacturing hubs
  - Military bases
- High PFOS near industrial facilities
- High concentration of PFNA located in San Luis Obispo
- Prominent locations
    - Reynolds Consumer Products industrial facility
    - Silicon Valley
    - PG&E Diablo Canyon Power Plant
    - Camp Roberts

SAAS

Next Steps

SAAS

Remaining Project Timeline

11/18-11/22	Refine models, approaches, and maps based on feedback and additional tuning
11/25-11/29	HOLIDAY - continue developing models and maps. Draw additional analysis from progress and findings. Begin work on dashboard.
12/2-12/6	Construct interactive dashboard with results.
12/9-12/13	Present final models and findings, technical hand-off

SAAS

Thank you!

SAAS

Questions?

SAAS