# Final Report for Contract 15-C0055

## 1. Input data

- **Dependent variable: Pesticide Concentration**
  The measured pesticide concentration in surface water at a particular place and time within California. Data is acquired internally from DPR. The concentration data were grouped by sample collection method (grab or composite samples), sample analytical type (whole water, dissolved, or particulate portion), and censorship (uncensored, or censored by detection and reporting limits of analytical methods).

- Currently developing model for specific pesticides based on active ingredient *Fipronil*, but plan to extend this later on. For fipronil, composite samples were not considered at this time due to lack of detail on how the composite samples were collected. Particulate samples were also excluded because all 24 samples were from one monitoring site and were all below detection limit.

- **Explanatory variables**
  Considered factors that have the potential to affect the fate and transport of pesticides in surface water. The factors can either be static (i.e., stable over a relatively long period of time) or time series (i.e., changes over time and whose history affect the pesticide's signal in surface water).

  - **Static data: StreamCat**
    "EPA's Office of Research and Development (ORD) has developed the *Stream-Catchment (StreamCat)* dataset, an extensive collection of landscape metrics for 2.6 million streams and associated catchments within the conterminous U.S. StreamCat includes both natural and human-related landscape features. The data are summarized both for individual stream catchments and for cumulative upstream watersheds, based on the National Hydrography Dataset Plus Version 2 geospatial framework." The database has > 500 attributes.

  - **Time-series data: Weather data (PRISM)**
    A database of climate related measurements from the *PRISM Climate Group***,** which have been localized to California.

  - **Time-series data: Pesticide Use Report (PUR) database**
    Self-reported usage of individual applications of pesticides within the state of California.

- *Initial data analysis:* Random Forest Model was used to find the most influential attributes using static data. Previous modeling results suggest starting point for a model and potential variables of interest. The next model will include both static data and time-series data.

## 2. Error correction for PUR data

- Fipronil is registered for urban uses only in California. The professional use records are reported by each company for each pesticide product at the monthly and county scale. Self-reporting errors have been identified and a phone call survey was carried out in order to correct the self-reporting errors in the fipronil PUR records in 2010 – 2015 according to the methodology described in a CDPR report (Ensminger et al., 2019). The survey corrected PUR data were used to train a machine learning model that was then used to correct fipronil PUR data in years other than 2010 – 2015. The model considered is the Random Forest model that can perform well for non-linear relationships (Breiman, 2001). R package "ranger" was used for the correction of PUR (Wright and Ziegler, 2015). The detailed methodology will be described in a separate report.

## 3. Development of model

- *Spatial aggregation*: All variables in the model have a spatial dimension that coincides with the response variable of interest (pesticide concentration); namely, the physical coordinate of where said sample was taken. This involves localizing all variables to this same spatial scale (down-scaling for weather data etc.). Another important feature of the model is 'stream aggregation'. It is well known that rivers and streams form a network and hence what happens upstream will soon propagate downstream. Because of this, features present in the StreamCAT database at a particular spatial location can be aggregated with respect to the stream network. This was done with regards to both watershed and catchment, separately, for most attributes. Some attributes, including PUR, are summarized at watershed level only. Fipronil PUR data is predominantly from structural uses (> 99% statewide in 1997 – 2017). Thus, total fipronil PUR was downscaled as if all are structural use. Assuming that fipronil structural use is statistically consistent among all the buildings in a given county and month, the structural PUR was downscaled from county level to watershed level using the proportion of urban parcel areas of a county that are in that specific watershed.

- *Temporal aggregation for time-series data* (PRISM and PUR) was done in three ways. Firstly, the short-term (daily) and mid-term (monthly) anomaly as used in SEAWAVE-QEX model; secondly, the total, max, min, and mean over the past 1 week, 1, 3, 6 months, 1, 2, 3 years; lastly, the total, max, min, and mean over the current and the previous 12 dry/wet periods. The total PUR over the past 13 dry/wet periods were further summed to total PUR over the past 0 to 1, 0 to 2, … 0 to 12 periods.

- *Novelty of dataset with regards to literature:* There are few pesticide concentration models in the literature but due to the sparsity of data, wide diversity of distinct pesticides in use within California, and other factors, the existing models cannot be applied and only serve as inspiration for our current modeling approach.

- Currently considering Random Survival Forest Model (motivated by prior analysis). It is well suited for non-linear relationships and can handle censored response data.

## 4. Model performance and next steps

- Many variables of interest for pesticide monitoring/regulation are considered in the model: their contribution to measured pesticide concentrations is potentially useful for DPR. The large number of input variables were used as input with default model parameterization and repeat the modeling 100 times to evaluate the influence of the variables to the fipronil concentration. The top 25, 50, and 100 variables were used in fine-tuning of the model parameterization. The available data were divided 80%/20% as training and validation data. For uncensored data, the model for grab water analyzed as filtered water performed satisfactory in that the predicted concentrations are well aligned with the observed concentrations. The model for water analyzed as whole water performed less ideal. There are some outliers observed. The next step is to investigate if those outliers shifted the model. For censored data, the model was optimized to improve the correlation between the observed and predicted censored data. However, optimizing the model on rank-based correlation does not guarantee that the quantitatively predicted concentrations are close to the observed values. The next step is to investigate the optimization method for censored data. An alternative approach is a two-step approach: (1) first do a classification to identify samples whose concentration will be lower than water quality objective, and (2) only quantitatively predict the concentration for samples whose concentration is above the water quality objective.

- More work is needed to understand how stable the model performed in identifying the most important variables and predicting the concentration, considering that such large number of covariates and relatively small number of samples.

- PUR is expected to have a measurable effect on pesticide concentrations. For fipronil, PUR covariates helped with model performance. However, with the inclusion of many urban covariates, such as impervious surface, urban land use subtypes (high, medium, low intensity and open space; commercial, industrial and residential), vegetation, population density, housing density, road density, etc., the model can perform comparably well without PUR covariates. This could partially due to the fact that fipronil is predominantly used on structural pest control in California.

- Model may be used to identify 'dangerous conditions' that lead to extreme pesticide concentrations. Such 'dangerous conditions' may: reinforce current understanding, provide new insight for further study, and improve mitigation/prevention efforts.

- Model identified most important variables can be considered in further evaluation for causal-relationship and used to prioritize mitigation.