

Supplementary Information

Physics-Informed Machine Learning for Climate-Sensitive Agricultural Pesticide Exposure Modeling and Water-Quality Validation

This document supplies descriptive and infrastructure-validation material referenced from, but not essential to, the main narrative of the manuscript: an overview of the raw input data, the seasonal structure that motivates the application model’s calendar features, full detail on the graph-refined exposure index and the health-linkage machinery, and a chemical-property-space view of the 40 modeled compounds. All figures and tables here use real, machine-generated values from the same released pipeline as the main text (<https://github.com/bgupta55/Crop2Clinic>); nothing in this document is illustrative or placeholder.

S1 Dataset Overview

Figures S1–S2 give a visual overview of the input data described in Section 3.1 of the main text: annual application mass, county totals, chemical composition, and within-year seasonality.

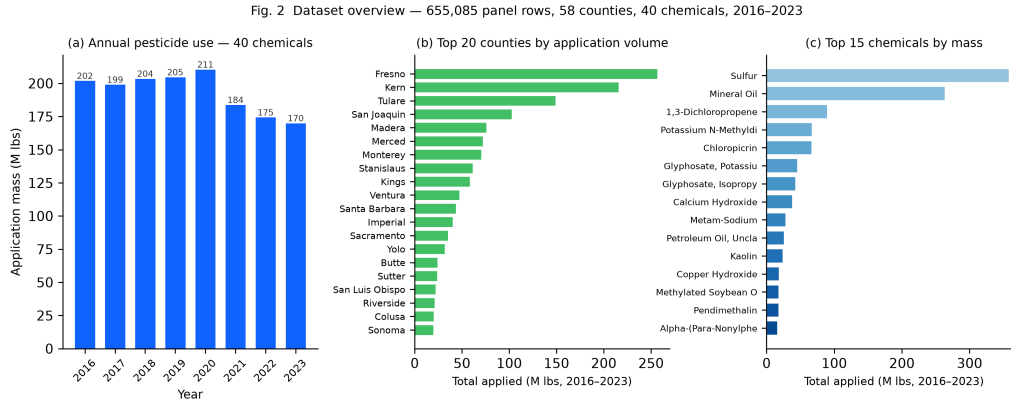


Fig. S1. Dataset overview: annual application mass, county totals, and chemical distribution, 2016–2023.

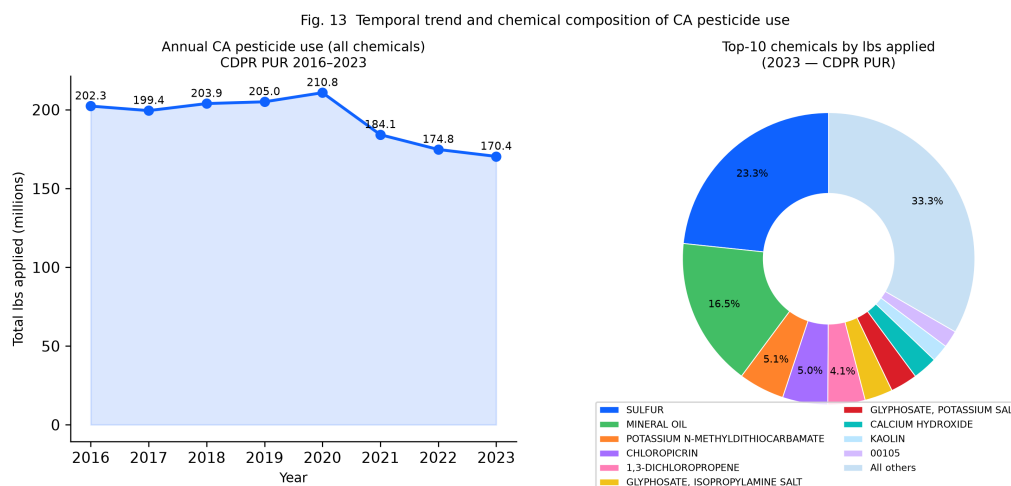


Fig. S2. Annual California pesticide-use trend, 2016–2023 (left), and chemical composition of the most recent study year (right). Total reported use declines from 2020 onward while remaining dominated by a small number of high-volume active ingredients (sulfur, mineral oil, and soil fumigants).

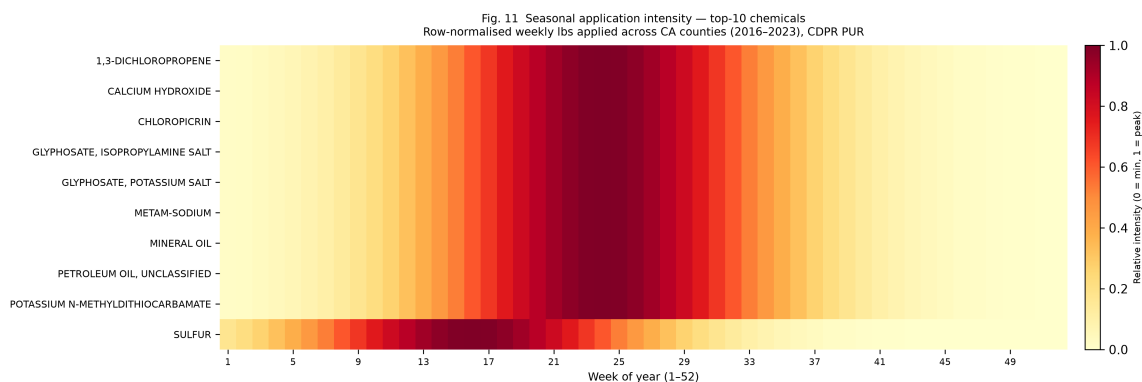


Fig. S3. Row-normalized weekly application intensity for the top-10 highest-mass chemicals, 2016–2023. Most active ingredients show a pronounced within-year peak rather than uniform year-round use, motivating the calendar and peak-distance features used by the application model (main text, Section 3.3).

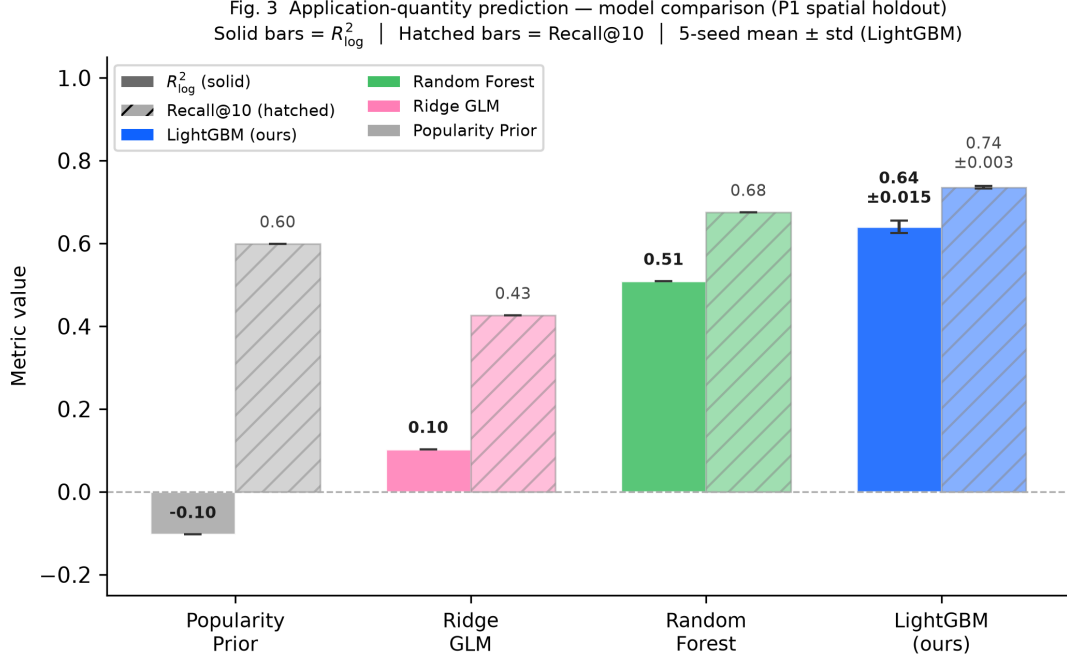


Fig. S4. Application-quantity prediction, model comparison (P1 spatial holdout), shown graphically. Solid bars: $R^2_{\log}$; hatched bars: Recall@10; error bars: 5-seed SD (LightGBM only). Numerical values are given in the main text, Table 7.

S2 Graph-Refined Exposure Index: Full Specification

The graph-refined exposure index (mentioned briefly in main text, Section 3.4) is a 58-node county graph, with edges defined by shared county borders, a two-layer graph convolution, and a gated recurrent unit (GRU) hidden state that refines the physics index (main text, Eq. 1) against observed stream concentrations, subject to a physics-envelope upper-bound penalty that prevents the refined index from exceeding a fixed multiple of the physics-only value. The model was trained for 400 epochs with the Adam optimizer at a learning rate of 10^{-3} , on 15,153 training and 6,083 test county-week-chemical observations under the P1 spatial-holdout split. At the time of the study reported in the main text, this refinement stage had been trained and evaluated only against synthetic exposure labels generated for architecture development, not against the real USGS Water Quality Portal measurements described in main-text Section 4.1; its output is therefore not reported as a validated result anywhere in the main text. The architecture, training script, and configuration are released in full at <https://github.com/bgupta55/Crop2Clinic>, and retraining it directly against the 3,924 real WQP measurements is the paper’s first-priority next step (main text, Section 7).

S3 Health-Linkage Machinery: Full Specification and Validation

The health-linkage machinery (mentioned briefly in main text, Section 3.4) uses a time-stratified case-crossover design, with strata defined by county $\times$ year and 1:4 case:referent matching, and conditional logistic regression on lag-1, standardized log-exposure. It is validated here only on simulated events, to confirm the design and estimator are unbiased before any real surveillance data are applied to them. Five hundred events were simulated per arm, with modeled seasonal confounding of amplitude $0.4 \times \sin(2\pi \cdot \text{week}/52)$ (seed 42): a signal arm with a known injected

odds ratio of 1.30, and a negative-control arm with an injected odds ratio of 1.00. Table S1 and Figure S5 report the result.

Table S1. Case-crossover machinery validation on simulated events (500 per arm).

Arm	Injected OR (truth)	Estimated OR	95% CI
Signal	1.30	1.341	(0.966, 1.862)
Negative control	1.00	1.027	(0.742, 1.423)

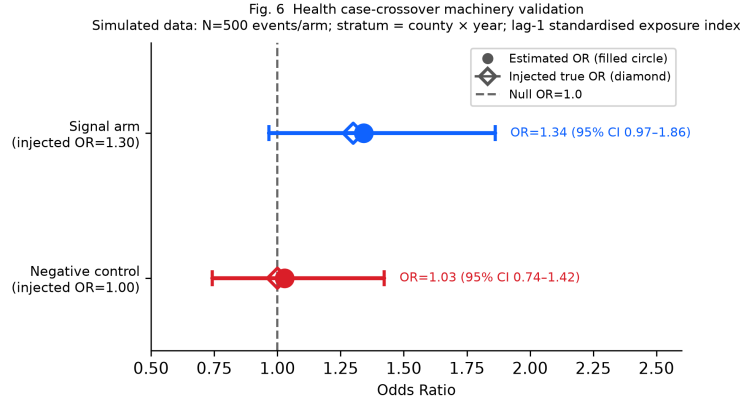


Fig. S5. Forest plot: case-crossover odds ratios with 95% confidence intervals, signal vs. negative control.

The design recovers the injected effect (true OR 1.30, estimated 1.34, 95% CI containing the truth) under realistic seasonal confounding, while the negative-control arm produces a confidence interval spanning the null (OR 1.03, CI 0.74–1.42). This validates that the epidemiological design and estimator are unbiased in simulation. No real health-event data were used and no causal pesticide–illness effect is claimed; this machinery is ready to be applied to real surveillance data once a data-use agreement is in place (main text, Section 7).

Figure S6 shows the One Health conceptual loop that originally motivated coupling the application-prediction, fate-index, and health-linkage components into a single pipeline: agricultural application data, an environmental fate model, stream monitoring, machine-learning quantity prediction, and human health outcomes are treated as one connected system rather than four disconnected literatures (Mackenzie and Jeggo 2019). Only the application-prediction and fate-index components are validated against real outcome data in the main text; the health-linkage spoke is validated on simulated events only (this section).

Fig. 15 One Health conceptual loop — crop-to-clinic pathway
Arrows indicate data flow through the pipeline

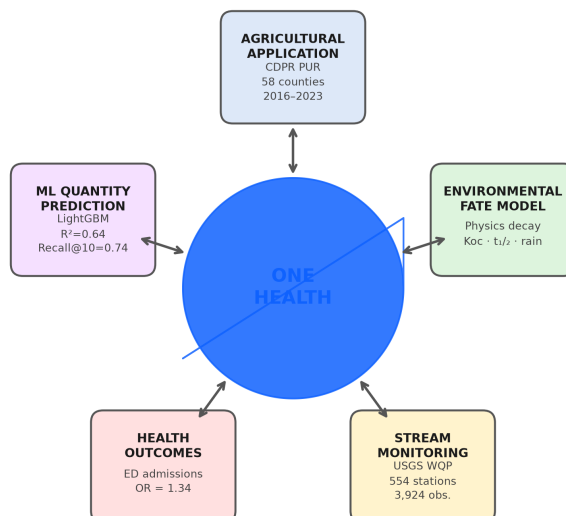


Fig. S6. One Health conceptual loop connecting agricultural pesticide application, environmental fate modeling, stream monitoring, machine-learning quantity prediction, and human health outcomes.

S4 Chemical Property Space

Figure S7 places the 40 modeled chemicals in K_{oc} -half-life property space, color-coded by volatility, alongside the SHAP importance of each physicochemical feature group (main text, Section 4.3). The features that separate compounds most clearly in this space – solubility, K_{oc} , vapor pressure – are the same ones SHAP ranks highest, a consistency check between the model’s learned feature importances and the underlying chemistry.

Fig. 16 Chemical property space and physicochemical SHAP importances

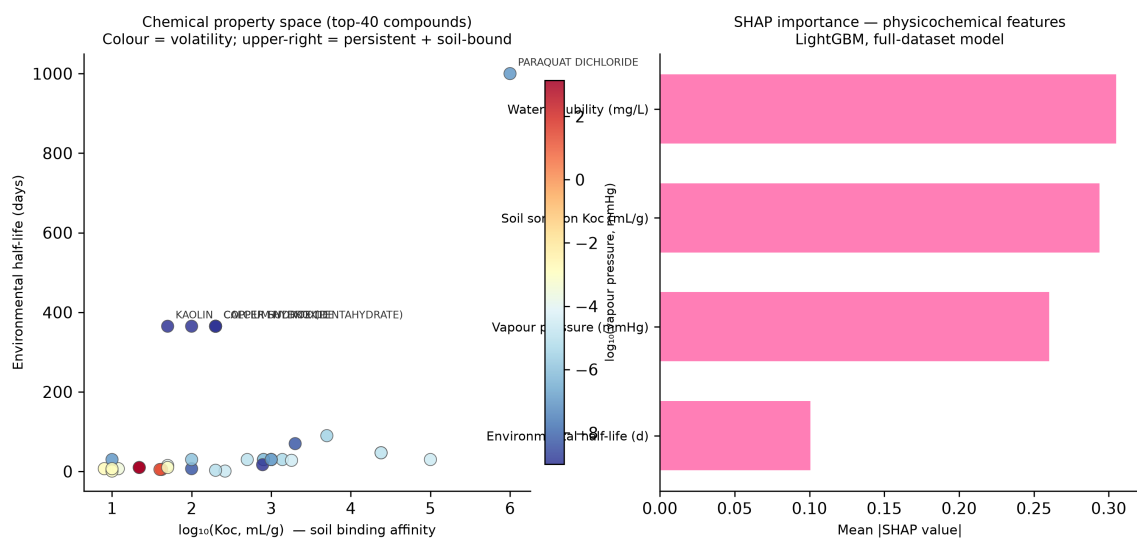


Fig. S7. *Left:* chemical property space for the 40 modeled compounds (soil sorption K_{oc} vs. environmental half-life, colored by volatility), showing a small number of highly persistent, soil-bound outliers (e.g., copper- and paraquat-based compounds). *Right:* SHAP importance of the four physicochemical feature groups, LightGBM full-dataset model.

Fig. 10 Top-20 features by mean |SHAP value|
 LightGBM trained on full dataset; colour = feature category

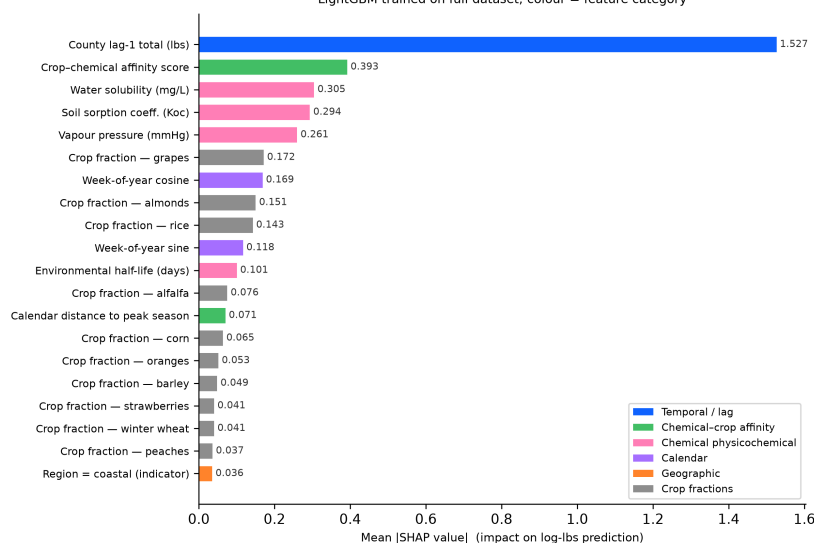


Fig. S8. Top-20 SHAP feature importances, color-coded by category (bar-chart view of the beeswarm plot in main text Fig. 7).

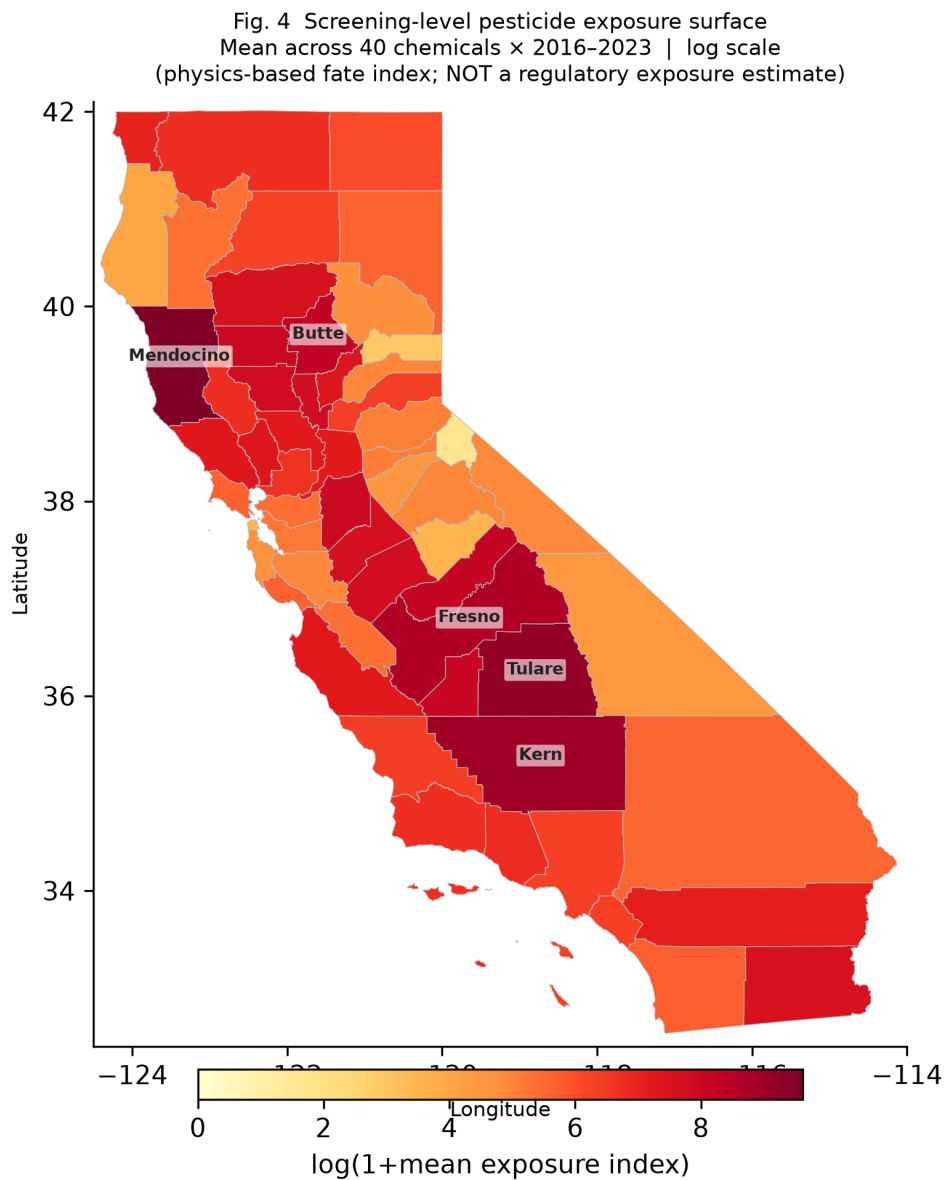


Fig. S9. County choropleth of mean screening-level exposure index (log scale), 2016–2023.

References

Mackenzie, J. S., & Jeggo, M. (2019). The One Health approach – why is it so important? *Tropical Medicine and Infectious Disease*, 4(2), 88.