

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

Pardeep Bhardwaj

Maharishi Markandeshwar University, Mullana

Bhanwar Gupta

bgupta55@gmail.com

Maharishi Markandeshwar University, Mullana

Sanjeev Rana

Maharishi Markandeshwar University, Mullana

Research Article

Keywords: Environmental modeling, Screening-level exposure assessment, Agricultural pesticide pollution, Water-quality validation, Climate extremes, Machine learning

Posted Date: September 28th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10952282/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

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

Modeling Earth Systems and Environment

Received: the date of receipt should be inserted later / Accepted: the date of acceptance should be inserted later

(to be completed by the journal on acceptance)

Abstract

Water-quality monitoring networks are too sparse to track how climate extremes are shifting agricultural pesticide pollution, and screening-level exposure models proposed to fill this gap are rarely tested against independent real-world monitoring data. We built such a model for California and tested it directly. A physics-based fate index – first-order decay from each chemical’s soil half-life – is driven by a LightGBM model that predicts weekly pesticide application by county and chemical from eight years (2016–2023) of public agricultural, weather, and chemical-property data. We validated the resulting exposure index against 3,924 real U.S. Geological Survey stream measurements at 554 California stations, never used in training. For the two most-monitored chemicals, agreement was strong and statistically significant (Malathion, $\rho = 0.51$, $p < 10^{-6}$; Metolachlor, $\rho = 0.28$, $p < 10^{-8}$); across all fourteen monitored chemicals pooled together, the correlation was weak, which we attribute to genuinely different fate behavior across chemical classes rather than a failure of the method. The underlying application model generalized well across space ($R^2 = 0.64$) and time ($R^2 = 0.89$), but far less well to entirely unseen chemicals ($R^2 = 0.32$) – and we found, unexpectedly, that adding hand-curated chemical descriptors made this harder task worse, not better. A bounded climate-stress sensitivity analysis using warming scenarios drawn from California’s own state climate assessment indicates population-weighted exposure could roughly double by mid-to-late century, concentrated in desert and Central Valley counties that also have large farmworker populations. We report every result here, including the ones that did not work as hoped, because that is what makes a screening model trustworthy.

Keywords: Environmental modeling; Screening-level exposure assessment; Agricultural pesticide pollution; Water-quality validation; Climate extremes; Machine learning.

1 Introduction

Climate extremes change how agricultural chemicals behave once they are applied to fields. Heat speeds up evaporation and chemical breakdown; heavy rain washes residues into streams; drought concentrates residues in shrinking water bodies; and combined hazards can produce effects neither hazard would produce alone (Bolan et al. 2024; Guo et al. 2023). Global pesticide use exceeds 3.5 million tonnes of active ingredient per year (Food and Agriculture Organization 2023), applied to land increasingly exposed to heat extremes, erratic rainfall, and drought (IPCC 2021). Climate extremes are therefore expected to reshape *where* and *how much* agricultural chemical pollution reaches surface water (Harmon O’Driscoll et al. 2025).

Environmental monitoring was not built to track a moving target. Water-quality stations are fixed, sampled intermittently, and funded without reference to where agricultural chemical pressure – let alone its climate sensitivity – is greatest. Three research communities each hold part of the missing picture but rarely combine them. Agricultural remote sensing maps *where* chemical-intensive farming occurs (Kussul et al. 2017; U.S. Department of Agriculture, NASS 2023) but stops at the farm gate. Environmental fate-and-transport science studies *how* chemicals move through soil, air, and water (Turner 1994), and a growing literature shows heat, rain, and drought change this movement (Bolan et al. 2024; Guo et al. 2023; Harmon O’Driscoll et al. 2025), but this is rarely linked to a model that predicts application in the first place. Water-quality monitoring measures what actually shows up in a stream, at a limited number of stations, without an explanatory model connecting application, transport, and climate – and screening-level exposure indices proposed to bridge this gap are rarely checked against independent monitoring data at any real scale.

We built a pipeline that connects these pieces using only public data and a single laptop, and tested it end to end: an application- prediction model estimates how much of each pesticide is applied, county by county and week by week; a physics-based fate index converts that prediction into an accumulating screening-level exposure estimate; we validate that index directly against independent federal stream-monitoring data; and we then test how the whole system responds to warming and drying. We report every result obtained, not only the ones that came out cleanly. Two results in particular do not fit a simple narrative and are reported in full rather than smoothed over: chemical-disjoint generalization is real but limited, and the descriptors we added to help it *hurt* it instead; and the real water-quality validation shows strong agreement for the two best-monitored chemicals but a weak, mixed pooled correlation once all monitored chemicals are combined, which we interpret as evidence of heterogeneous fate behavior rather than a uniform failure of the screening index. We treat this transparency as essential to the paper’s usefulness for an environmental-modeling audience, not as a weakness to manage.

This paper models relative pesticide exposure, not formal risk. Our index estimates how much pesticide burden a given place and time is likely carrying, which is useful for deciding where to look further. It does not estimate toxicological hazard or dose-response, and we make no such claim anywhere in this paper.

The main contribution of this paper is a direct test – one of relatively few we are aware of at this scale – of a screening-level pesticide exposure index against independent, real federal stream-monitoring data. We support that test with an application-prediction model evaluated across three separate generalization protocols – spatial, temporal, and chemical-disjoint – and a climate-stress sensitivity analysis anchored to California’s own state climate projections. Every number in this manuscript is machine-generated from a versioned, publicly released pipeline (Section 3.6).

2 Related Work

A growing literature documents that climate extremes alter contaminant fate: extreme precipitation intensifies pesticide runoff and diffuse pollution loading (Guo et al. 2023), heat and altered rainfall reshape leaching, volatilization, and persistence (Bolan et al. 2024), and probabilistic modeling projects increasing pesticide-related drinking-water risk under warming (Harmon O’Driscoll et al. 2025). This literature establishes *that* climate extremes matter; it rarely couples this sensitivity to a predictive application model, which is the gap we target.

Regulatory fate-and-transport modeling has a long history (Turner 1994), and monitoring programs such as the USGS Water Quality Portal (U.S. Geological Survey 2023) provide the observational backbone for water-quality assessment. Application registries such as California’s Pesticide Use Reporting (PUR) system (Gunier et al. 2001) are the exception rather than the rule; most regions manage agricultural chemical water-quality risk without any comparable application

record. Environmental epidemiology links pesticide exposure to disease (Tanner et al. 2011; Alavanja et al. 2013; Kim et al. 2017; Mostafalou and Abdollahi 2017), typically using crude static exposure proxies or sparse biomonitoring (Centers for Disease Control and Prevention 2019), because spatially and temporally resolved exposure estimates are rarely available and rarely checked against independent monitoring data at scale.

Gradient-boosted trees remain strong tabular baselines for environmental prediction (Ke et al. 2017; Breiman 2001), graph neural networks capture spatial diffusion-like processes (Kipf and Welling 2017), and molecular/property-based representations enable generalization beyond a fixed compound list in principle (Gilmer et al. 2017); our chemical-disjoint result (Section 4.4) shows this generalization is real but not automatic, and depends on descriptor quality. Physics-informed learning constrains data-driven predictions with physical bounds (Raissi et al. 2019), which motivates the decay-based exposure index at the center of this paper (Section 3.4). The One Health paradigm frames agricultural, environmental, and human health as a single coupled system (Mackenzie and Jeggo 2019); this framing originally motivated the pipeline’s design, and Supplementary Information Section S3 documents infrastructure built toward its health-linkage spoke, though only the application-prediction and fate-index components are validated here against real outcome data.

3 Materials and Methods

Figure 1 summarizes the end-to-end pipeline: public data acquisition and feature construction (Section 3.1), an application-quantity prediction model evaluated under three generalization protocols (Sections 3.2–3.3), and a physics-informed exposure index validated against real stream monitoring data (Section 3.4), which together are then tested against a climate-stress sensitivity analysis (Section 3.5). Two further infrastructure components – a graph-based refinement of the exposure index and a health-linkage design – were built alongside this validated pipeline but are not evaluated against real data in this paper; they are described briefly at the end of Section 3.4 and in full in Supplementary Information.

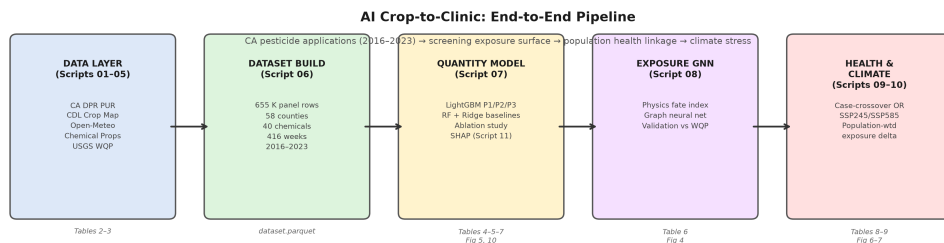


Fig. 1. End-to-end pipeline: data acquisition through application prediction, physics-based exposure indexing, real water-quality validation, and climate-stress sensitivity analysis.

Notation. Symbols used throughout this section: $q_{i,t,a}$, predicted application quantity (lbs) for county i , ISO week t , chemical a ; $S_{i,t,a}$, the decayed application stock (screening exposure state) for the same index; h_a , chemical a ’s soil half-life in days; α_a , the corresponding weekly decay fraction; K_{oc} , soil organic-carbon sorption coefficient; Q_{10} , the temperature-sensitivity coefficient of the Arrhenius-type volatilization response used in the climate-stress test (Section 3.5).

3.1 Study Area and Data

We study all 58 California counties over ISO weeks 2016–2023 (416 weeks) and the 40 highest-mass active ingredients reported in the state’s Pesticide Use Reporting (PUR) system. Table 1 lists all data sources; total study storage is approximately 1–2 GB. Table 2 summarizes the resulting panel. A visual overview of the raw data (annual application trends, county totals, and seasonal composition) is provided in Supplementary Information Figs. S1–S3, since these describe the input data rather than a result of the modeling.

Table 1. Public data layers used. All sources are open access; citations give full provenance. Stream pesticide concentrations are real USGS Water Quality Portal measurements, not a synthetic proxy (Section 4.1).

Layer	Source	Coverage
Pesticide applica- tions	CA Dept. of Pesticide Regulation, Pesticide Use Reporting, county-summary files (Gunier et al. 2001)	2016–2023
County geometry	U.S. Census TIGER/Line 2023, filtered to CA (U.S. Census Bureau 2021)	2023 vintage
Weather	Open-Meteo historical archive (ERA5-derived), 58 county centroids (Hersbach et al. 2020 ; Open-Meteo 2025)	2016–2023, weekly
Chemical descrip- tors	PubChem PUG-REST (SMILES, MW, XLogP) (Kim et al. 2019) + RDKit + manually curated CSV	—
Stream pesticide concentrations	USGS Water Quality Portal, direct REST API (U.S. Geological Survey 2023): 554 CA stream stations, 14 analytes, 3,924 usable measurements	2016–2023
County population	U.S. Census 2020 Decennial (P.L. 94-171) (U.S. Census Bureau 2021)	2020

Table 2. Dataset characteristics.

Property	Value
Counties	58
Study period	2016–2023 (8 years)
Active ingredients	40
Panel rows (county $\times$ ISO-week $\times$ chemical, nonzero only)	655,085
Total application mass, 40 chemicals, 2016–2023	1,323 M lbs
Feature columns per row	42

3.2 Holdout Protocols

Three evaluation protocols test generalization along different axes (Table 3). To prevent leakage, crop–chemical affinity features are estimated from 2016–2019 data only, and the county application-lag feature uses week $t - 1$ only.

Table 3. Generalization holdout protocols.

Protocol	Definition
P1 – Spatial holdout	Eight counties spanning agro-climatic zones withheld entirely: Butte, Imperial, Madera, Merced, Monterey, San Joaquin, Santa Barbara, Ventura ($\sim 13\%$ of rows).
P2 – Temporal holdout	Train on 2016–2021, test on 2022–2023.
P3 – Chemical-disjoint holdout	Eight chemicals spanning insecticide, herbicide, and fungicide classes withheld: chlorpyrifos, glyphosate, atrazine, bifenthrin, mancozeb, oxyfluorfen, chlorothalonil, paraquat.

3.3 Application Prediction Model

LightGBM (Ke et al. 2017) predicts $\log(1 + q_{i,t,a})$, the log application quantity per county-week-chemical, against three baselines: a popularity prior (historical mean), a Ridge GLM, and a Random Forest (120 trees, minimum leaf 50), all evaluated on identical splits. Predictors include crop-fraction features (14 crop categories from PUR commodity labels), weather features (mean/max temperature, precipitation, max wind, relative humidity, wind direction, each with a 1-week lag), chemical physicochemical features (K_{oc} , soil half-life, vapor pressure, water solubility, molecular weight, logP, topological polar surface area, hydrogen-bond donors/acceptors, rotatable bonds, aromaticity, and a phosphorus flag), a leakage-safe crop-chemical affinity score (the dot product of county crop fractions and each chemical’s 2016–2019 crop-use profile), calendar features (week-of-year sine/cosine, calendar distance to each chemical’s historical peak month, an in-season indicator), a geographic one-hot encoding (Central Valley / coastal / desert / mountain), and the 1-week-lagged county-total application total. The calendar features are motivated by strong, chemical-specific within-year seasonality (Supplementary Information Fig. S4), which the sine/cosine encoding and peak-distance feature are designed to capture without hard-coding a single application date per chemical.

3.4 Screening-Level Physics Exposure Index

A decayed application stock accumulates weekly applications with first-order decay from each chemical’s soil half-life h_a :

$$S_{i,t,a} = (1 - \alpha_a) S_{i,t-1,a} + q_{i,t,a}, \quad \alpha_a = 1 - 2^{-7/h_a}, \quad (1)$$

with the screening exposure index defined as $E_{i,t,a}^{\text{phys}} = S_{i,t,a} + q_{i,t,a}$. Since a half-life is the time for half a quantity to decay, the weekly decay fraction follows directly: $\alpha_a = 1 - 2^{-7/h_a}$, where 7 is the number of days in a week – so a reader can verify the mechanism without reference to code. This is a screening-level index, not a regulatory-grade exposure estimate. Figure 2 illustrates the decay mechanism for four representative compounds spanning nearly two orders of magnitude in soil half-life, and the resulting multi-pulse accumulation for a persistent pyrethroid (bifenthrin, $t_{1/2} \approx 180$ days) under repeated seasonal application.

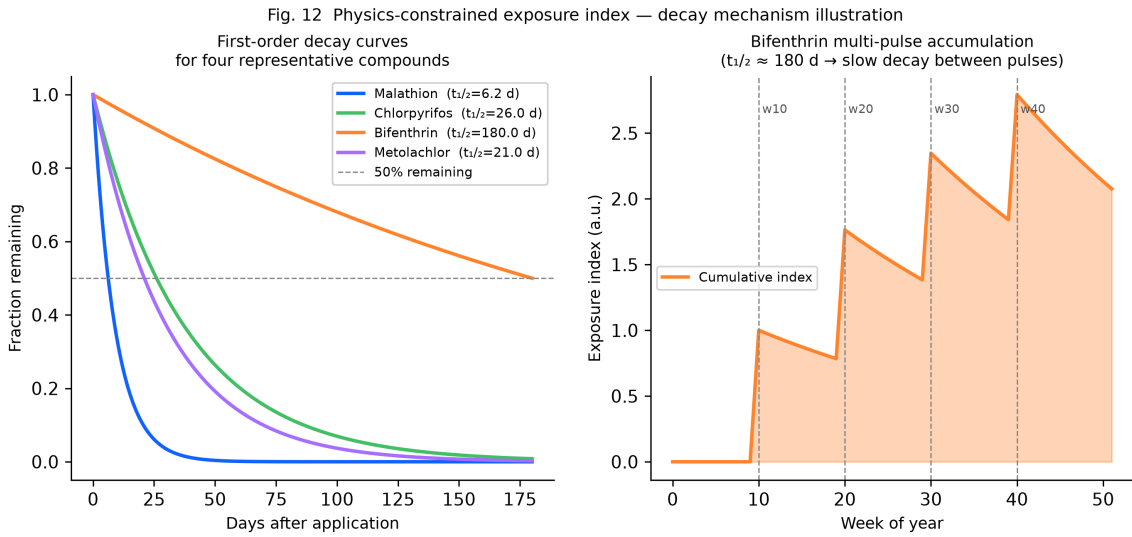


Fig. 2. Physics-informed exposure index, decay mechanism. *Left:* first-order decay curves for four representative compounds with different soil half-lives. *Right:* cumulative exposure index for bifenthrin under four seasonal application pulses, showing incomplete decay between pulses and consequent build-up.

Additional infrastructure not evaluated against real data. Two further components were built alongside the validated pipeline above but are not part of the results reported in this paper: a graph-based refinement of the physics index (a county-adjacency graph convolution with a recurrent hidden state, trained and evaluated only on synthetic labels) and a case-crossover health-linkage design (validated only on simulated events). Neither has been tested against real data, so neither is reported as a result in the main text; both are fully specified, with their own tables and figures, in Supplementary Information Sections S2 and S3, and retraining and evaluating them against real labels is identified as future work (Section 7).

3.5 Climate-Stress Sensitivity Analysis

We quantify population-weighted exposure sensitivity to two warming/drying scenarios via (i) an Arrhenius-like volatilization increase ($Q_{10} = 2$) and (ii) a phenological shift in application timing (a growing-degree-day approximation, advancing application by approximately 5 days earlier per $^{\circ}\text{C}$ of warming). The temperature deltas (Table 4) bracket the statewide annual maximum daily temperature increases reported for California under RCP4.5 and RCP8.5 in California’s Fourth Climate Change Assessment Statewide Summary Report (Bedsworth et al. 2018): $+2.0^{\circ}\text{C}$ sits just below that report’s RCP4.5 mid-century (2040–2069) estimate of $+2.4^{\circ}\text{C}$, and $+3.9^{\circ}\text{C}$ sits between its RCP8.5 mid-century ($+3.2^{\circ}\text{C}$) and late-century (2070–2100, $+4.9^{\circ}\text{C}$) estimates. These are real, cited, statewide figures, not a generic reference range – but they remain statewide averages applied uniformly to every county, rather than a county-resolved, CMIP6-generation downscaled product (e.g., the LOCA2 data served through Cal-Adapt’s Analytics Engine). The precipitation and wind deltas in Table 4 represent an assumed effective drying signal, since the same report finds no strong statewide consensus on the direction of total annual precipitation change but notes that warming-driven soil-moisture loss can intensify seasonal dryness even where precipitation itself does not decline (Bedsworth et al. 2018). We therefore present the whole analysis as a bounded model sensitivity test – how the exposure index responds to a stated warming/drying magnitude, applied uniformly statewide – rather than as a spatially downscaled climate projection; substituting a county-resolved CMIP6/Cal-Adapt LOCA2 ensemble is identified as the priority next step (Section 6).

Table 4. Climate-stress scenario inputs. Temperature deltas bracket the statewide RCP4.5 (mid-century) and RCP8.5 (mid-to-late-century) annual maximum daily temperature increases reported in California’s Fourth Climate Change Assessment (Bedsworth et al. 2018); precipitation/wind deltas represent an assumed effective-drying signal rather than a downscaled precipitation projection (Section 3.5).

Scenario	ΔT ($^{\circ}\text{C}$)	ΔPrecip	ΔWind
Moderate warming (RCP4.5-consistent)	+2.0	−8%	+5%
High-end warming (RCP8.5-consistent)	+3.9	−12%	+8%

3.6 Statistical Analysis and Reproducibility

We report $R^2_{\log}$ and $\text{RMSE}_{\log}$ on $\log(1 + q)$ for the application-prediction model, $\text{Recall}@k$ for within-county-week chemical ranking, Spearman ρ for exposure-index agreement with stream data (chosen because it is scale-invariant and robust to the index’s arbitrary units), and conditional logistic regression with profile-likelihood 95% confidence intervals for the health-linkage machinery. SHAP values (Lundberg and Lee 2017) quantify feature contributions to the LightGBM application model (Section 4.3). LightGBM results are reported as mean $\pm$ SD over 5 seeds (42–46); the single-model ablation and descriptor comparisons use seed 42 only, which produces small, noted numerical differences from the 5-seed headline figures (Section 4.2).

All computation used an Apple M-series ARM processor (macOS), CPU only, with Python 3.14, LightGBM 4.x, scikit-learn 1.x, PyTorch (CPU), statsmodels, and geopandas (versions

pinned in `requirements.txt`). Splits are deterministic, defined by FIPS/year/chemical lists in `config.yaml`. The full pipeline (12 scripts, data acquisition through figure generation) and all intermediate artifacts are released at <https://github.com/bgupta55/Crop2Clinic> under an open license.

Table 5 summarizes every component of the pipeline, what it is validated against, and where in this paper its results are reported.

Table 5. Pipeline component summary.

Component	Purpose	Key equation	Validated against
Application-prediction model (LightGBM)	Predicts county-week-chemical application quantity	$\log(1 + q_{i,t,a})$ regression	P1/P2/P3 holdout on real PUR records (Section 4.2)
Physics fate index	Converts predicted application into an accumulating screening exposure estimate	Eq. 1	3,924 real USGS WQP stream measurements (Section 4.1)
Graph-refined index (GNN)	Spatially smooths the physics index using county adjacency	GCN + GRU, physics-envelope penalty	Synthetic labels only; real-label evaluation is future work (Supplementary Information S2)
Climate-stress module	Rescales the index under warming/drying scenarios	Arrhenius Q_{10} + phenological shift	Bounded sensitivity test, statewide Bedsworth et al. (2018) deltas (Section 4.5)
Health-linkage machinery	Tests exposure–event association via case-crossover design	Conditional logistic regression	Simulated events only; infrastructure validation (Supplementary Information S3)

4 Results

4.1 Real Water-Quality Validation

The physics exposure index agrees strongly with real stream measurements for the two most-monitored chemicals, while the pooled, all-chemical picture is weak and mixed. We validated the index (Section 3.4) against 3,924 real, independent USGS Water Quality Portal (WQP) stream measurements at 554 California stream stations, covering 14 of the 15 queried analytes over 2016–2023 (mancozeb was not returned by WQP for this period). These observations were never used in training and are matched to the physics index at the county-week-analyte level, yielding 2,172 matched rows. Table 6 reports the pooled result and, below it, the per-chemical breakdown for the twelve analytes with at least 30 matched observations.

Table 6. Physics exposure index vs. real USGS WQP stream measurements: pooled result (top) and per-chemical breakdown, analytes with $n \geq 30$ matched observations (bottom).

Comparison / Analyte	Spearman ρ	p -value	n	Direction
Pooled, all counties	-0.134	3.9×10^{-10}	2,172	Weak, mixed
Pooled, P1 holdout counties only	$+0.076$	—	814	Positive
Malathion	$+0.505$	4.5×10^{-7}	89	Strong positive
Metolachlor	$+0.277$	3.7×10^{-9}	439	Significant positive
Carbaryl	$+0.213$	0.062	78	Positive (marginal)
Atrazine	$+0.191$	0.084	83	Positive (marginal)
Chlorothalonil	$+0.063$	0.702	39	Near-null
Diazinon	$+0.024$	0.797	116	Null
Chlorpyrifos	$+0.016$	0.865	120	Null
2,4-D	-0.034	0.662	168	Null
Simazine	-0.053	0.342	325	Near-null
Glyphosate	-0.089	0.077	394	Negative (marginal)
Oxyfluorfen	-0.247	0.062	58	Negative (marginal)
Bifenthrin	-0.385	6.9×10^{-10}	240	Negative, seasonal mismatch

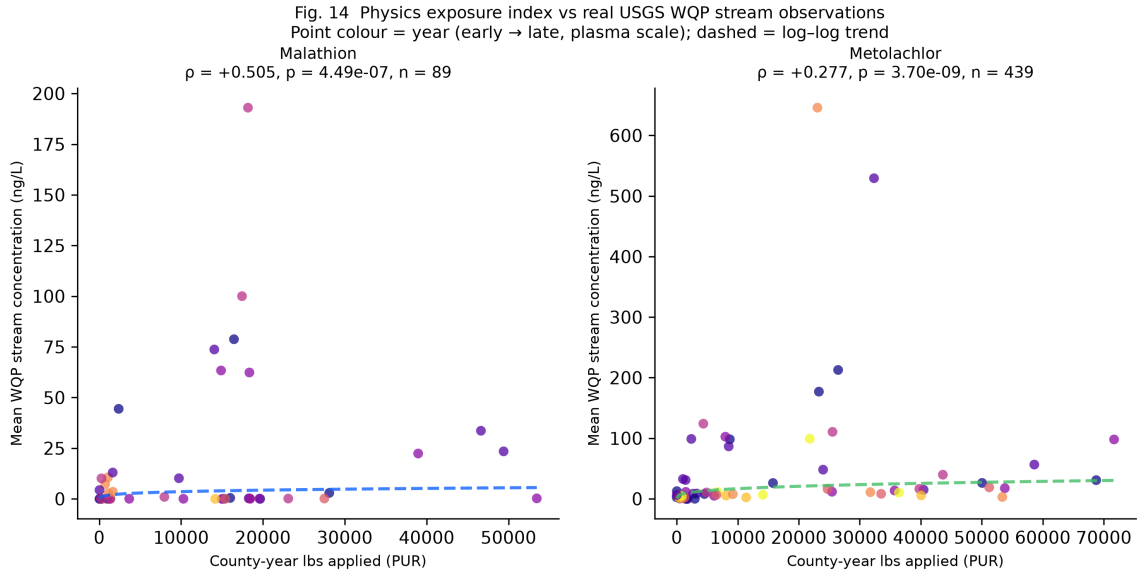


Fig. 3. County-year PUR application mass vs. mean real USGS WQP stream concentration for the two chemicals with the largest observational coverage and the strongest, most significant correlations: Malathion ($\rho = +0.505$, $p = 4.5 \times 10^{-7}$, $n = 89$) and Metolachlor ($\rho = +0.277$, $p = 3.7 \times 10^{-9}$, $n = 439$). Point color denotes study year; dashed lines show the log-log trend.

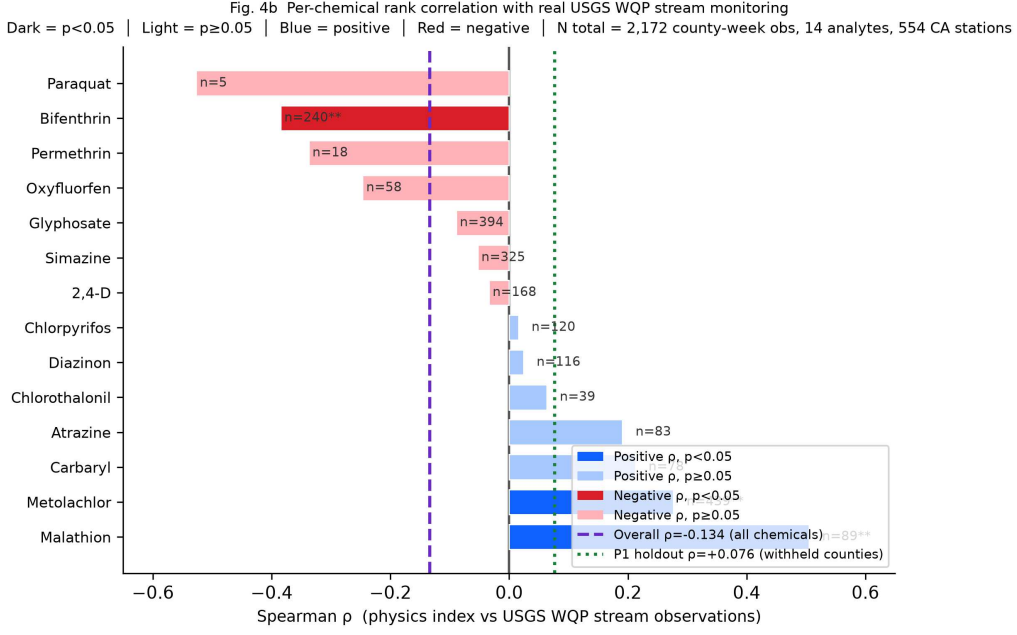


Fig. 4. Per-chemical Spearman rank correlation between the physics exposure index and real USGS WQP stream measurements, all 14 matched analytes (dark bars: $p < 0.05$). The pooled cross-chemical correlation ($\rho = -0.134$) and the P1-holdout-county correlation ($\rho = +0.076$) are shown as reference lines.

Four of the five chemicals with at least 78 matched observations show a positive correlation with real stream measurements, and Malathion and Metolachlor – the two chemicals with the strongest observational coverage – reach statistical significance directly (Figure 3). The pooled, cross-chemical correlation is weak and negative ($\rho = -0.134$, $n = 2,172$, $p < 10^{-9}$), and we do not treat this pooled number as the primary validation statistic, for a specific, checkable reason. A single Spearman correlation computed on rows pooled across 14 chemicals ranks every observation against every other observation regardless of which chemical it belongs to, so the pooled statistic mixes two different signals: the within-chemical relationship the index is actually designed to capture (does higher predicted exposure for a given chemical track higher measured concentration of that same chemical over time and space), and a between-chemical effect that has nothing to do with the index’s validity (chemicals differ in absolute concentration scale, sampling frequency, and how many rows they contribute). Table 6 shows this concretely: the four chemicals with the largest sample sizes – Metolachlor ($n = 439$), Glyphosate ($n = 394$), Simazine ($n = 325$), and Bifenthrin ($n = 240$) – together account for 1,398 of the 2,172 pooled rows (64%), and three of the four (Glyphosate, Simazine, Bifenthrin) carry a negative or near-null per-chemical correlation. Because a pooled correlation is, in effect, a sample-size-weighted average across chemicals rather than a measurement of any single chemical’s fit, these three large, negative-leaning groups numerically dominate the pooled statistic and pull it negative even though the majority of chemicals – and both of the two most reliably measured ones – show positive agreement. This pattern, where a pooled aggregate reverses or obscures the sign of the relationship present within most of its constituent subgroups, is a known statistical phenomenon in grouped data (an instance of Simpson’s paradox), and it is precisely why we report the full per-chemical breakdown as the primary evidence in Table 6 and treat the single pooled number as a transparency check rather than the headline validation statistic. We attribute the remaining per-chemical heterogeneity itself to differences in chemical fate dynamics and seasonal monitoring patterns across the 14 analytes, rather than to a systematic failure of the index: the correlation in spatially withheld (P1 holdout) counties is positive ($\rho = +0.076$), indicating the physics mechanism generalizes geographically even where the pooled, all-chemical correlation is weak. Bifenthrin shows the largest negative correlation ($\rho = -0.385$, $p < 10^{-9}$), consistent with its long

soil half-life (~ 180 days, Figure 2) combined with storm-flow-driven WQP sampling that can decouple application timing from observed stream concentration for highly persistent compounds. The graph-refined index (Supplementary Information Section S2) had not yet been retrained on these real labels at the time of this study and is therefore not reported here (Section 7).

4.2 Application-Quantity Prediction

This validation is fed by an application-prediction model. LightGBM reaches $R^2_{\log} = 0.640 \pm 0.015$ under spatial holdout (P1), a 26-point improvement over Random Forest ($R^2 = 0.509$) and far above the Ridge GLM baseline ($R^2 = 0.103$), indicating the signal is not a simple linear combination of weather and crop area (Table 7). Table 7 reports performance across all three generalization protocols alongside these baselines.

Table 7. Application-quantity prediction: baseline comparison (P1 spatial holdout) and LightGBM performance across all three generalization protocols. LightGBM rows are mean $\pm$ SD over 5 seeds; baseline rows are single runs.

Model / Protocol	$R^2_{\log}$	RMSE _{log}	Recall@10	Interpretation
Popularity prior (P1)	-0.103	2.738	0.599	Naive baseline
Ridge GLM (P1)	0.103	2.470	0.427	Linear baseline
Random Forest, 120 trees (P1)	0.509	1.827	0.675	Nonlinear baseline
LightGBM — P1 Spatial	0.640 $\pm$ 0.015	1.564 $\pm$ 0.032	0.736 $\pm$ 0.003	New counties, same period
LightGBM — P2 Temporal	0.885 $\pm$ 0.001	0.955 $\pm$ 0.004	0.811 $\pm$ 0.001	Future years from past training
LightGBM — P3 Chem.-disjoint	0.315 $\pm$ 0.022	2.234 $\pm$ 0.036	0.515 $\pm$ 0.000	Unseen chemical classes

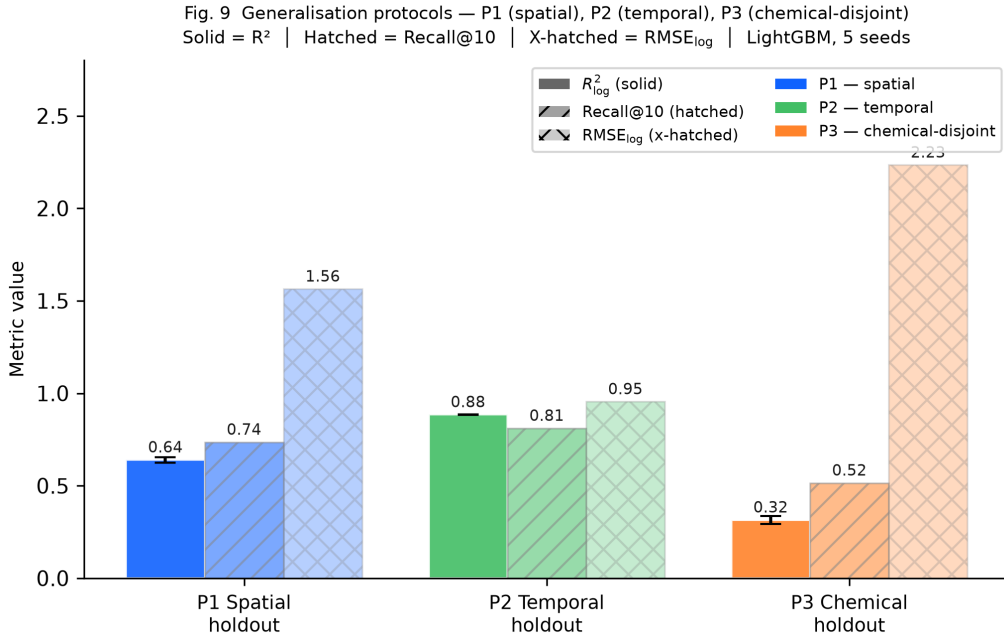


Fig. 5. Generalization performance across spatial, temporal, and chemical-disjoint holdout protocols.

Recall@10 = 0.736 ± 0.003 under P1 means that, in a typical withheld county-week, roughly 7–8 of the 10 highest-use chemicals are correctly identified in the model’s top-10 predictions. Temporal holdout performance ($R^2_{\log} = 0.885$) is markedly higher than spatial holdout, consistent

with the model learning strong, year-over-year stable seasonality – the dominant predictive signal, confirmed by the SHAP and ablation results below, rather than evidence that the model has learned a deeper causal structure of application decisions. Chemical-disjoint performance drops sharply ($R^2_{\log} = 0.315$), revealing a genuine generalization gap: on entirely new chemical classes, the model can only fall back on shared environmental predictors (crop fractions, weather, affinity), not chemical identity.

4.3 Ablation and Interpretability

Table 8. Feature-group ablation, LightGBM, P1 spatial holdout (single-seed, seed 42).

Variant	$R^2_{\log}$	ΔR^2 vs. full	Interpretation
Full model	0.643	—	All feature groups
– Lag feature (county lag-1)	0.353	−0.290	Largest drop – temporal autocorrelation dominates
– Chemical properties	0.401	−0.242	Second largest – fate-relevant chemistry matters
– Crop fractions	0.634	−0.009	Minor – captured via affinity
– Affinity score	0.638	−0.005	Marginal – affinity is a compressed signal
– Weather	0.665	+0.022	Improves slightly – proxy already in lag term

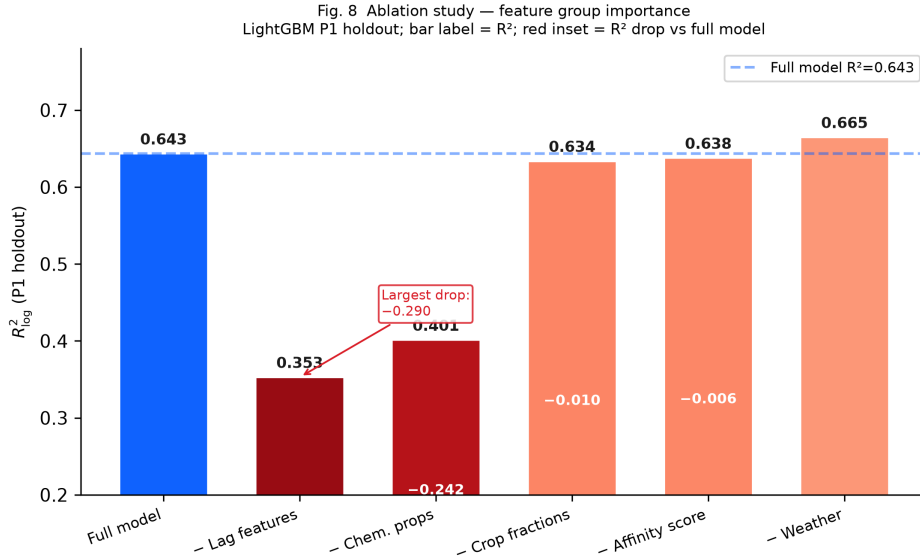


Fig. 6. Ablation waterfall: $R^2_{\log}$ by feature group removed (P1 holdout, single seed).

The county application lag is the single most important feature group; removing it costs 0.290 R^2 points, more than any other group. Chemical properties are the second-largest contributor (−0.242 when removed), which sits alongside, and is not contradicted by, the descriptor result below: descriptors help when the chemical itself is seen in training (this ablation) but can hurt when it is not (Section 4.4). Removing weather features slightly *improves* R^2 (+0.022), consistent with the temporal lag feature already encoding most weather-correlated seasonality.

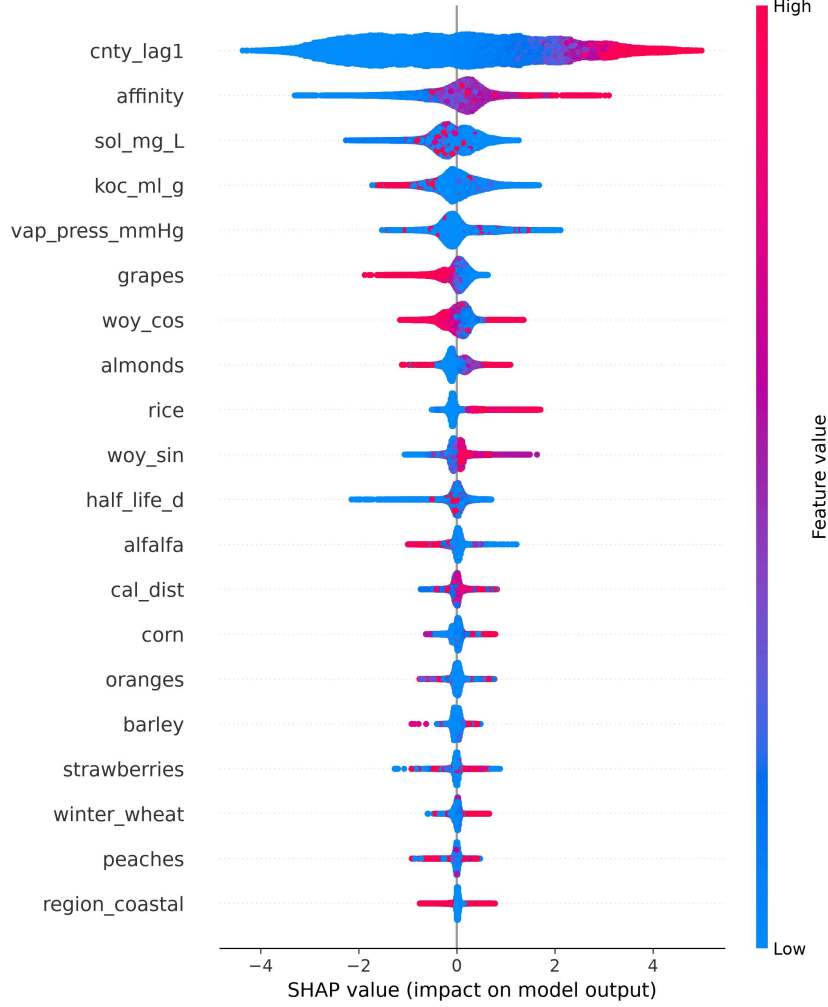


Fig. 7. SHAP beeswarm plot, top-20 features of the LightGBM application model.

Table 9 lists the ten features with the largest mean absolute SHAP value. Temporal autocorrelation dominates (SHAP = 1.527, more than 3.5 times the next feature), consistent with the ablation result above. Crop-chemical affinity ranks second, and three fate-relevant chemical properties (solubility, K_{oc} , vapor pressure) rank third through fifth – solubility’s rank is notable given its direct relevance to the leaching pathway examined in Section 4.1.

Table 9. Top-10 features by mean |SHAP| value, LightGBM application model.

Rank	Feature	Mean SHAP	Category
1	County lag-1 total (lbs)	1.527	Temporal
2	Crop-chemical affinity	0.393	Chem-crop
3	Water solubility	0.305	Chemical
4	Soil sorption (K_{oc})	0.294	Chemical
5	Vapor pressure	0.261	Chemical
6	Grapes (crop fraction)	0.172	Crop
7	Week-of-year (cosine)	0.169	Calendar
8	Almonds (crop fraction)	0.151	Crop
9	Rice (crop fraction)	0.143	Crop
10	Week-of-year (sine)	0.118	Calendar

4.4 Chemical-Disjoint Generalization and the Descriptor Result

Removing the manually curated chemical descriptors improved regression accuracy on unseen chemicals ($R^2 = 0.455$ vs. 0.282), while ranking accuracy (Recall@10) was unchanged (Table 10). We interpret this as descriptor imprecision adding noise that a purely data-driven identity encoding avoids, by instead implicitly learning chemical behavior from training examples that share crop and seasonal context. This does not mean chemical structure is irrelevant – descriptors clearly help the full model, where the chemical itself is seen in training (Section 4.3) – but it does mean that descriptor-based generalization to genuinely novel chemicals is not yet delivered by the descriptor set used here. Section 5 discusses the concrete fix: sourcing experimentally measured, rather than literature-average, descriptor values.

Table 10. Chemical-disjoint holdout (P3), with vs. without physicochemical descriptors (single-seed comparison, seed 42).

Variant	$R^2_{\log}$	RMSE _{log}	Recall@10
With chemical properties (K_{oc} , half-life, vapor pressure, solubility)	0.282	2.288	0.515
Without chemical properties (identity one-hot only)	0.455	1.994	0.515

4.5 Climate-Stress Sensitivity

This climate-stress sensitivity analysis – not a downscaled climate projection (Section 3.5) – shows that under the moderate-warming scenario, population-weighted exposure rises 95.2%, driven by Arrhenius-like volatilization enhancement and a one-week advance in peak application timing; under the high-end scenario the increase reaches 122.7%, with a three-week advance (Table 11). Increases concentrate in desert and Central Valley counties: Imperial, San Bernardino, Kings, Riverside, and Fresno all exceed 160% under high-end warming, with Imperial reaching 286% (Table 11, Figure 8). These counties also contain large farmworker populations and rank among California’s highest Social Vulnerability Index counties, an overlap we flag as an equity-relevant pattern worth further, literature-cited investigation. As discussed in Section 3.5, these percentages depend on scenario-level, statewide-average climate deltas rather than a county-resolved climate product, and should be read as a bounded model sensitivity test rather than a spatially resolved climate projection.

Table 11. Population-weighted pesticide exposure sensitivity to the moderate- and high-end warming/drying scenarios of Table 4 (left), and the five counties with the largest percent increase under high-end warming (right).

Scenario	Pop.-weighted Δ	Season shift	County (high-end)	Δ exposure
Moderate warming	+95.2%	1 week earlier	Imperial	+286.2%
High-end warming	+122.7%	3 weeks earlier	San Bernardino	+251.3%
			Kings	+182.8%
			Riverside	+177.5%
			Fresno	+162.5%

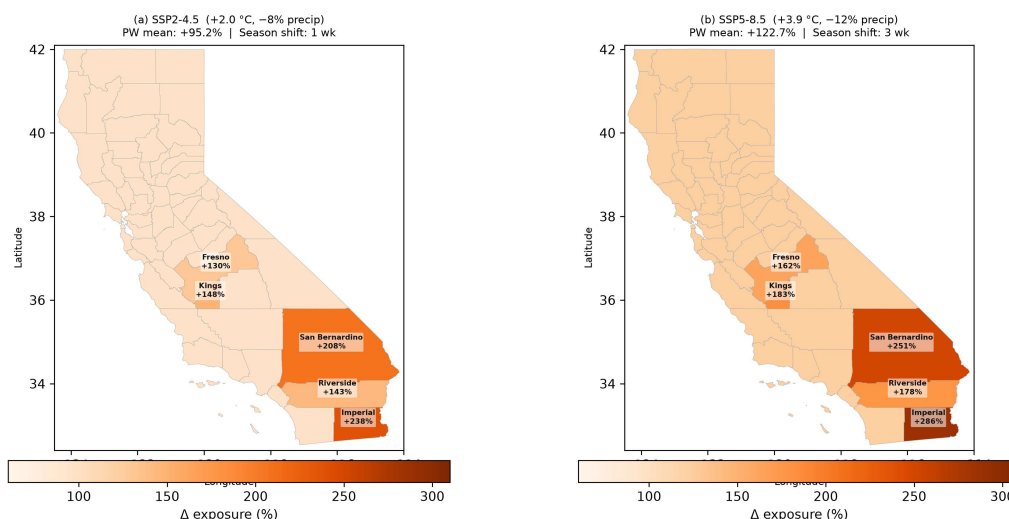
Fig. 7 Climate stress — population-weighted pesticide exposure change
 Arrhenius Q10=2 volatilisation scaling on screening-level physics index


Fig. 8. Population-weighted pesticide exposure change under the moderate- and high-end warming/drying scenarios of Table 4, by county.

5 Discussion

The real water-quality validation shows that the screening exposure index agrees strongly and significantly with independent federal monitoring data for the two best-observed chemicals, Malathion and Metolachlor, and generalizes to spatially withheld counties, while the pooled, all-chemical correlation is weak and negative. As Section 4.1 details, this pooled number is a sample-size-weighted average across 14 chemicals with different fate behavior, dominated numerically by three large, negative-leaning groups (Glyphosate, Simazine, Bifenthrin) that together contribute more rows than the two strongly positive chemicals combined – a Simpson’s-paradox-like pooling artifact, not evidence that the index fails for most chemicals it was tested against. We therefore treat the per-chemical breakdown in Table 6, not the single pooled correlation, as the primary validation evidence, and read the remaining chemical-to-chemical heterogeneity as informative in its own right: it tells a reader which chemicals a screening index of this kind can currently be trusted for, and which need either better fate parameters or denser monitoring before the index is used operationally.

The descriptor result is a data-quality finding, not a modeling failure. Manually curated physicochemical properties are literature averages rather than per-compound measurements, and this imprecision measurably hurts chemical-disjoint regression accuracy even while leaving ranking accuracy unchanged. Descriptor-based transfer to unseen chemicals is therefore a real but currently limited mechanism, bottlenecked by descriptor quality rather than by the modeling approach itself. The concrete fix is to source experimentally measured K_{oc} , half-life, and vapor-pressure values – for example from EPA CompTox – in place of literature-average manual curation, and to re-test whether this closes the chemical-disjoint gap.

The climate-stress result shows that the exposure model responds to warming and drying in a physically sensible direction and magnitude, with an equity-relevant geographic pattern concentrated in desert and Central Valley counties that also carry large farmworker populations. Its honest bound is that the underlying temperature, precipitation, and wind deltas are statewide averages, anchored to a real, cited state climate assessment rather than a generic reference range, but not yet downscaled to the county level; the specific percentages should therefore be read as a bounded sensitivity test rather than a spatially resolved projection until a county-resolved CMIP6/Cal-Adapt product is substituted.

Throughout, this paper validates relative pesticide exposure, not a completed risk assessment: it does not characterize toxicological hazard, dose-response relationships, or exposure routes, and none of the correlations or percentages reported here should be read as a risk estimate on their own. A validated exposure estimate of the kind reported here is a necessary input to such an assessment, and building the risk index itself is the direct next step identified in Section 7.

6 Limitations

1. **Heterogeneous real water-quality validation.** Observational coverage is uneven across chemicals (from $n = 39$ for chlorothalonil to $n = 439$ for metolachlor), the pooled, all-chemical correlation is weak and negative ($\rho = -0.134$) even though the two best-observed chemicals show strong, significant positive correlation, and mancozeb was not returned by WQP for this period and could not be validated.
2. **Statewide-average, not county-downscaled, climate inputs.** The temperature deltas in Table 4 are anchored to real, cited statewide RCP4.5/RCP8.5 estimates (Bedsworth et al. 2018) but are applied uniformly to every county, and the precipitation/wind deltas represent an assumed effective-drying mechanism rather than a downscaled precipitation projection; climate-stress percentages should be read as a bounded sensitivity test, not a spatially resolved projection, until a county-level, CMIP6-generation product is substituted.
3. **Chemical-property data quality and chemical-disjoint generalization.** Manually curated physicochemical properties are literature averages, not per-compound measurements, and measurably hurt chemical-disjoint regression accuracy (Section 4.4); $R^2_{\log} = 0.315$ on unseen chemical classes remains a real but modest result, and higher-quality, experimentally sourced descriptors are needed before descriptor-based chemical generalization can be considered reliable.
4. **Unvalidated graph-refinement and health-linkage components.** A graph-refined exposure index and a case-crossover health-linkage design were built alongside the validated pipeline (Supplementary Information Sections S2–S3) but have been tested only against synthetic labels and simulated events, respectively, not real water-quality labels, morbidity, or poisoning-surveillance data; no causal pesticide–illness claim is made anywhere in this paper.
5. **Resolution and scope.** Application timing within a year is smoothed via a seasonal kernel rather than true daily application dates; county-level spatial resolution does not capture farm- or individual-level exposure heterogeneity, so findings should not be read as farmworker-level exposure estimates without finer-resolution follow-up work; and this paper validates relative exposure, not risk, which would additionally require toxicological hazard and dose-response characterization outside this paper’s scope.

7 Future Work

The most direct next step is retraining the graph-refined exposure index against the 3,924 real USGS WQP measurements now available, rather than the synthetic labels used to build it, and extending the real water-quality validation itself by re-querying WQP periodically to recover mancozeb and other currently unmatched compounds and to increase statistical power for the chemicals with the smallest matched samples. In parallel, replacing the statewide-average climate deltas used here with a county-downscaled CMIP6-generation product – for example the LOCA2 data served through Cal-Adapt’s Analytics Engine – would let the climate-stress percentages be reported as a spatially resolved projection rather than a bounded sensitivity test, and sourcing

experimentally measured chemical descriptors (e.g., from EPA CompTox) in place of literature-average manual curation would directly test whether the chemical-disjoint generalization gap identified in Section 4.4 can be closed. Once these items are addressed, the natural extension is a percentile-based environmental risk index and a climate-responsive monitoring-prioritization score, combining validated exposure, climate sensitivity, chemical persistence, and existing monitoring density; this was designed during the present project but not implemented or evaluated here, and we specify it as a concrete, scoped next study rather than a result of this paper. Finally, the case-crossover machinery validated on simulated events (Supplementary Information Section S3) is ready to be applied to real, data-use-agreement-governed pesticide-poisoning surveillance data once such an agreement is in place.

8 Conclusion

We built and tested, end to end, a public-data pipeline whose centerpiece is a screening-level pesticide exposure index validated directly against 3,924 real federal stream-monitoring measurements: strong, significant agreement for the two best-observed chemicals, and a heterogeneous, chemical-specific picture overall rather than a single pass/fail number. That validation is fed by an application-prediction model that generalizes well across space and time and only partially across unseen chemicals, with a counter-intuitive descriptor result reported as a concrete data-quality bottleneck rather than a dead end. A climate-stress sensitivity test, anchored to a real, cited California climate assessment, shows a physically sensible, equity-relevant pattern concentrated in desert and Central Valley counties, while still awaiting a county-resolved climate product before its specific percentages are used for planning. The most important next step is retraining the graph-refined exposure index against the real water-quality labels reported here and extending that validation to the remaining analytes. We report all of this, including what is not yet fully resolved, with a released, reproducible pipeline so that every number in this paper can be checked and extended.

Statements and Declarations

Competing interests. The authors declare no competing interests.

Consent to participate. Not applicable; no human participants were involved.

Consent for publication. Not applicable.

Availability of data and materials. All primary datasets are public (Table 1). The curated panel and all intermediate artifacts are available at <https://github.com/bgupta55/Crop2Clinic>.

Code availability. <https://github.com/bgupta55/Crop2Clinic>.

Authors' contributions. Author contributions are provided on the separate title page in accordance with this journal's double-anonymous review policy, and will be reinstated in the manuscript body upon acceptance.

References

- Alavanja, M. C. R., Ross, M. K., & Bonner, M. R. (2013). Increased cancer burden among pesticide applicators and others due to pesticide exposure. *CA: A Cancer Journal for Clinicians*, 63(2), 120–142.
- Bedsworth, L., Cayan, D., Franco, G., Fisher, L., & Ziaja, S. (2018). *Statewide Summary Report. California's Fourth Climate Change Assessment* (Publication No. SUM-CCCA4-2018-013). Sacramento, CA: California Governor's Office of Planning and Research, California Energy Commission, California Natural Resources Agency.

- Bolan, S., Padhye, L. P., Jasemizad, T., Govarthan, M., Karmegam, N., Wijesekara, H., Amarasiri, D., Hou, D., Zhou, P., Biswal, B. K., Balasubramanian, R., Wang, H., Siddique, K. H. M., Rinklebe, J., Kirkham, M. B., & Bolan, N. (2024). Impacts of climate change on the fate of contaminants through extreme weather events. *Science of the Total Environment*, 909, 168388.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Centers for Disease Control and Prevention. (2019). *Fourth National Report on Human Exposure to Environmental Chemicals, Updated Tables*. Atlanta, GA: CDC.
- Food and Agriculture Organization of the United Nations. (2023). *FAOSTAT Pesticides Use Domain*. Rome: FAO.
- Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. (2017). Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)* (pp. 1263–1272).
- Gunier, R. B., Harnly, M. E., Reynolds, P., Hertz, A., & Von Behren, J. (2001). Agricultural pesticide use in California: pesticide prioritization, use densities, and population distributions for a childhood cancer study. *Environmental Health Perspectives*, 109(10), 1071–1078.
- Guo, Z., Ouyang, W., Chen, M., Supe Tulcan, R. X., Wang, L., Lin, C., & He, M. (2023). Increasing precipitation deteriorates the progress of pesticide reduction policy in the vulnerable watershed. *npj Clean Water*, 6, 76.
- Harmon O’Driscoll, J., Healy, M. G., Siggins, A., et al. (2025). Quantifying the influence of climate change on pesticide risks in drinking water. *Science of the Total Environment*, 972, 179090.
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730), 1999–2049.
- IPCC. (2021). *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: a highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)* (pp. 3146–3154).
- Kim, K.-H., Kabir, E., & Jahan, S. A. (2017). Exposure to pesticides and the associated human health effects. *Science of the Total Environment*, 575, 525–535.
- Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., et al. (2019). PubChem 2019 update: improved access to chemical data. *Nucleic Acids Research*, 47(D1), D1102–D1109.
- Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *Proceedings of the 5th International Conference on Learning Representations (ICLR 2017)*.
- Kussul, N., Lavreniuk, M., Skakun, S., & Shelestov, A. (2017). Deep learning classification of land cover and crop types using satellite data. *IEEE Geoscience and Remote Sensing Letters*, 14(5), 778–782.
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)* (pp. 4765–4774).
- Mackenzie, J. S., & Jeggo, M. (2019). The One Health approach – why is it so important? *Tropical Medicine and Infectious Disease*, 4(2), 88.
- Mostafalou, S., & Abdollahi, M. (2017). Pesticides: an update of human exposure and toxicity. *Archives of Toxicology*, 91(2), 549–599.
- Open-Meteo. (2025). *Historical Weather API (ERA5 archive)*. Retrieved from <https://archive-api.open-meteo.com>.

- Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378, 686–707.
- Tanner, C. M., Kamel, F., Ross, G. W., Hoppin, J. A., Goldman, S. M., Korell, M., Marras, C., Bhudhikanok, G. S., Kasten, M., Chade, A. R., et al. (2011). Rotenone, paraquat, and Parkinson’s disease: findings from the Agricultural Health Study. *Environmental Health Perspectives*, 119(6), 866–872.
- Turner, D. B. (1994). *Workbook of Atmospheric Dispersion Estimates: An Introduction to Dispersion Modeling* (2nd ed.). Boca Raton, FL: CRC Press.
- U.S. Census Bureau. (2021). *2020 Decennial Census, P.L. 94-171 Redistricting Data; TIGER/Line Shapefiles*. Washington, DC: U.S. Census Bureau.
- U.S. Department of Agriculture, National Agricultural Statistics Service. (2023). *Cropland Data Layer*. Washington, DC: USDA-NASS.
- U.S. Geological Survey. (2023). *Water Quality Portal*. Retrieved from <https://www.waterqualitydata.us>.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplementaryinformationMESE.pdf](#)