High-Spatiotemporal-Resolution PM2.5 Mapping in California: 6-Hourly Estimates with 3-km HRRR-Smoke

Zhiqing Huang
School for Environment and Sustainability/University of Michigan
Ann Arbor MI USA
zhiqingh@umich.edu

Song Gao
Geospatial Data Science Lab, Department of Geography/University of Wisconsin-Madison
Madison WI USA
song.gao@wisc.edu

Qunying Huang
Spatial Computing and Data Mining Lab, Department of Geography/University of Wisconsin-Madison
Madison WI USA
qhuang46@wisc.edu

Abstract

Exposure to fine particulate matter (PM2.5) harms human health, and high–spatiotemporal–resolution estimates are needed for exposure assessment. We present a statewide 6-hourly PM2.5 modeling framework for California (Dec 2, 2020–Jun 30, 2025) that fuses ~3-km HRRR-Smoke with spatially lagged AQS PM2.5 and monitor coordinates. A Random Forest trained with a strictly time-forward 80/20 split attains MAE 0.541 μg/m³, RMSE 1.332 μg/m³, R^2 0.936 on the held-out window. The fitted model produces seamless 6-hourly maps and seasonal/annual composites that consistently highlight the Central Valley and South Coast Air Basin; spring is cleanest, summer–autumn elevations align with wildfire smoke, and winter hotspots reflect Central Valley inversions. The statewide mean time series shows episodic late-summer/early-autumn maxima and a weak downward trend (–0.825 μg/m³ yr^{-1}, R^2 0.055). Case-day analyses reveal statewide smoke plumes on the Dixie Fire peak day (>200 μg/m³) and evening–overnight bumps on Independence Day (<25 μg/m³) contrasted with a Good level benchmark (≈2.5–6.3 μg/m³). These results show that high-resolution inputs like HRRR-Smoke can support operational, 6-hourly PM2.5 mapping with a simple, reproducible pipeline.

CCS CONCEPTS

• **Information systems → Spatial-temporal systems**; • **Applied computing → Environmental sciences**.

KEYWORDS

High-spatiotemporal-resolution PM2.5, PM2.5 estimation, Spatial lag

ACM Reference format:

Zhiqing Huang, Song Gao, Qunying Huang. 2025. High-Spatiotemporal-Resolution PM2.5 Mapping in California: 6-Hourly Estimates with 3-km HRRR-Smoke. In *1st ACM SIGSPATIAL International Workshop on Geospatial Computing for Public Health Proceedings (GeoHealth'25), November 3-November 3, 2025, Minneapolis, MN, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3764918.3770163

GeoHealth '25, November 3–6, 2025, Minneapolis, MN, USA

ACM ISBN 979-8-4007-2181-6/2025/11.
https://doi.org/10.1145/3764918.3770163

1 Introduction

PM2.5 is a major environmental risk factor for public health, and existing studies have established its association with diseases such as cardiovascular [1]. Evidence further indicates that even exposure at low concentrations poses health risks [2]. Since PM2.5 monitoring data are typically available at the station level, achieving high spatiotemporal resolution in prediction, estimation, and analysis is critical for capturing fine-scale dispersion patterns, thereby providing a solid data and technical foundation for public health protection. Research today emphasizes improving the temporal and spatial resolution of PM2.5 mass concentration across different scales [3]. Such efforts typically depend on meteorological inputs at relatively coarse resolutions or daily predictions combined with multi-source spatiotemporal feature engineering, a reliance that has consistently been shown to compromise predictive accuracy and weaken model robustness [4]. Datasets derived from low spatiotemporal-resolution data may result in limited performance for high-precision PM2.5 prediction [5]. And this might be due to uncertainty propagation caused by interpolation or restricted time ranges [6]. Such limitations also hinder further improvements and reduce the model's generalization ability.

In response, this research fuses HRRR (High-Resolution Rapid Refresh) -Smoke with its lightweight spatiotemporal features to produce accurate, maintaining high-spatiotemporal-resolution 6-hourly PM2.5 estimates over California with minimal model complexity. We emphasized producing statewide 6-hourly PM2.5 fields from ~3-km inputs with minimal model complexity, evaluating them in a strictly time-forward manner that mirrors operations, and documenting the workflow end-to-end for reproducibility and efficient deployment.

2 Methodology

2.1 Data

2.2.1 Hourly PM2.5 data of EPA. We retrieved hourly PM2.5 from EPA's Air Quality System (AQS) via API and used it as the dependent variable. Records were filtered to active and primary NAAQS monitors during Dec 2020–Jun 2025 with1-hour sample duration. After filtering, 21 monitors across California remained.

To match HRRR timing, hours were aggregated to non-overlapping 6-hour windows (00–05, 06–11, 12–17, 18–23 UTC) using simple means; each record retains site latitude/longitude and the 6-hour timestamp (window end at 00/06/12/18).

2.2.2 HRRR-Smoke. The HRRR is NOAA's a numerical weather prediction model at ~3 km horizontal resolution with hourly cycles and its radar reflectivity is assimilated roughly every 15 minutes within each cycle, which is real-time and convection-allowing [7]. We used the latest HRRRv4 since December 2, 2020 is the most practical and operational one, for it has inline smoke prediction, improved boundary-layer cloud treatment and lake temperatures, a 3-km ensemble data-assimilation upgrade, and 48-hour forecasts every 6 hours. HRRRv4 has various variables reflecting different meteorological parameters' conditions. Here, we chose the 8m near-surface smoke mass concentration as our proxy for PM2.5. HRRR-Smoke also provides regional plume context that complements the local persistence encoded by spatial lags. The HRRR data from its native projection to the WGS84 coordinate system to match the PM2.5 monitoring data. The Nearest-neighbor resampling method was applied to maximize the retention of the original data values.

2.2 Spatially Lagged PM2.5

PM2.5 exhibits spatial auto correlation across monitoring sites because nearby locations share emissions, transport pathways, and boundary-layer conditions. We represented this dependence with a spatial-lag operator that aggregates recent observations from neighboring sites into a weighted predictor for the target site and time [1]. The spatial-lag predictor is defined by the normalized, inverse-distance-weighted mean of neighboring observations:

$$S_i^{(p)}(t,k) = \frac{\sum_{j \in N_i(R)} w_{ij}^{(p)} y_j(t-k)}{\sum_{j \in N_i(R)} w_{ij}^{(p)}} \quad (1)$$

$$w_{ij}^{(p)} = (max\{d_{ij}, d_{min}\})^{-p} \quad (2)$$

$$p \in \{1, 2\}$$

where $y_j(t-k)$ is the 6-hour mean PM2.5 at monitor j and time window t; $N_i(R) = \{j \neq i : d_{ij} \leq R\}$is the neighbor set for target site i under radius R; d_{ij} is the great-circle distance (kilometer) between sites i and j under WGS84 coordinate system; $d_{ij} > 0$ is a small cutoff to avoid singular weights; $w_{ij}^{(p)}$ is the inverse-distance weight with exponent p; and $k \in \{0,1,2,3,4\}$ indexes temporal lags of 0, 6, 12, 18, and 24 hours.

To prevent information leakage, the contemporaneous feature (k = 0) excludes the target site's own observation, and all features at test times are computed using only observations available up to $t-k$. Missing neighbor observations at a given window are omitted from the sums; when no neighbors are available within the search bounds, the spatial-lag value is imputed from a broader-area weighted mean for that window. Neighbor sets are obtained either by applying a fixed radius R or by adaptively expanding R until a minimum number of neighbors m is reached, subject to an upper cap R_{max}. Here we used p = 1 and enforce a modest minimum neighbor count with R_{max} on the order of 10^2 km, which produced stable features without inflating collinearity across temporal lags. Distances are computed as Haversine distances. Neighbor queries are accelerated with a spherical search tree, and the resulting weights and normalizers are cached so that $S_i^{(p)}(t,k)$ can be evaluated efficiently for all sites and lags.

We constructed five spatial-lag features with p = 1 at k = 0, 1, 2, 3, and 4, referred to elsewhere as sllag_0, sllag_6, sllag_12, sllag_18, and sllag_24, summarizing local persistence and neighborhood context at horizons from 0 to 24 hours.

2.3 Modeling with Random Forest Regressor

We fitted a Random Forest regressor using the open-source Python machine-learning library scikit-learn [8]. This has been leveraged in many studies [9]. The input features of the model comprise (i) HRRR-Smoke's 8m near-surface smoke mass concentration sampled at the monitor's nearest native 3-km grid cell with a 6-hour lead aligned to 00/06/12/18 UTC computed with leave-one-site-out at k = 0 and time-forward availability; (ii) spatially lagged PM2.5 constructed at 0, 6, 12, 18, and 24 hours as inverse-distance-weighted neighborhood means and strict time-forward computation; and (iii) site latitude and longitude. We harmonized AQS and HRRR timestamps, dropped incomplete rows, and built the five spatial-lag features (sllag_0/6/12/18/24) at each time step.

The model's hyperparameters were tuned by a compact grid search with time-aware three-fold cross-validation: in each fold, the model was trained on earlier timesteps and validated on a later, non-overlapping block in a rolling, blocked-split scheme, thereby respecting chronology and preventing information leakage. To approximate true PM2.5 results, we adopted a strict chronological split, reserving the latest 20% of timesteps as the test set, namely the train-test split is 8:2. The label is the 6-hour mean PM2.5 aggregated from hourly AQS observations.

3 Results

3.1 Model Performance

3.1.1 Evaluation Metrices. On the held-out test window, the model attains MAE = 0.541 μg/m^3, RMSE = 1.332 μg/m^3, and R^2 = 0.936; the corresponding train metrics are MAE = 0.349 μg/m^3, RMSE = 1.445 μg/m^3, and R^2 = 0.976. These values indicate strong out-of-sample accuracy with modest error magnitudes relative to day-to-day variability. An R^2 above 0.93 suggests that the chosen features capture most of the spatiotemporal structure relevant for monitor-level PM2.5 in California. The predicted-versus-observed cloud closely follows the 1:1 line from 0 to about 30 μg/m^3 (Figure 1). At higher concentrations the relationship bends below the 1:1 line, indicating mild underestimation above ~45 μg/m^3 and a soft ceiling near ~55 μg/m^3. The same plot reports the test-set coefficient of determination (R^2 = 0.936), which, together with the visual agreement, summarizes overall skill. Residuals r = (prediction vs. observation)

are tightly clustered and centered slightly negative (mean ≈ -0.17; $1\sigma \approx 1.32$).

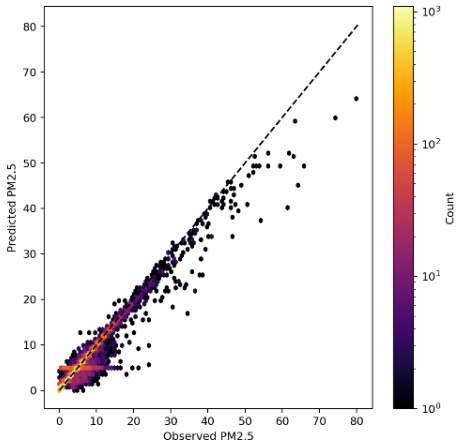

Figure 1: Model Predictions vs. Observed PM2.5 Scatter

3.1.2 Feature importance. We summarize predictor contributions using mean decrease in impurity (MDI) from the trained Random Forest. MDI offers a fast, construction-time indication of relative influence by attributing variance reduction across all trees and normalizing over features. Interpreted qualitatively, the contemporaneous spatial lag provides the dominant signal; short temporal lags follow; HRRR-Smoke contributes a meaningful regional scaffold; and latitude/longitude supply only a light broad-scale correction (Figure 2). Taken together, the model behaves as a hybrid: spatial lags encode local persistence and neighborhood context, HRRR-Smoke lays down regional plume structure, and coordinates stabilize low-frequency spatial trends. Latitude and longitude add only a weak broad scale correction. Taken together, the model functions as a hybrid in which spatial lags provide local persistence and neighborhood context, HRRR Smoke supplies regional plume structure, and coordinates stabilize low frequency spatial trends. Because MDI redistributes credit among correlated lags, we treat the ranking as the robust takeaway rather than the absolute magnitudes.

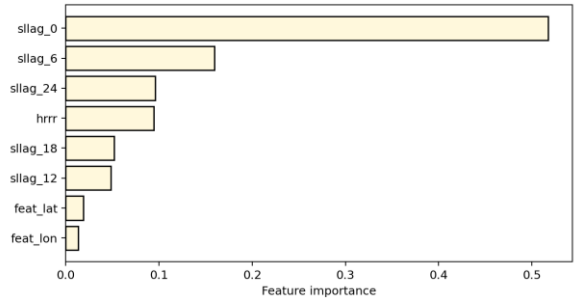

Figure 2: Random-Forest MDI importances

3.2 Spatial Distribution and Temporal Trend

3.2.1 Statewide 6 hourly seasonal–annual distributions. Using the trained Random Forest model, we generated 6-hourly PM2.5 prediction surfaces statewide for Dec 2, 2020–Jun 30, 2025 (Figure 3). Seasonal and annual composites show stable geography: the annual mean concentrates along the Central Valley and the South Coast Air Basin, with lower values on the immediate coast and higher terrain. Spring is cleanest statewide. Summer exhibits marked enhancements over southern California and the Central Valley/Sierra foothills, consistent with smoke transport and warm-season stagnation. Autumn remains elevated over the Central Valley and parts of southern California, aligning with late-season fires and shallow boundary layers. Winter shows localized Central Valley hotspots tied to inversions, while coastal and mountain regions stay comparatively low. The common scale (~2.4–16.4 μg/m³) situates these contrasts within a moderate regime punctuated by episodic spikes evident in the time series.

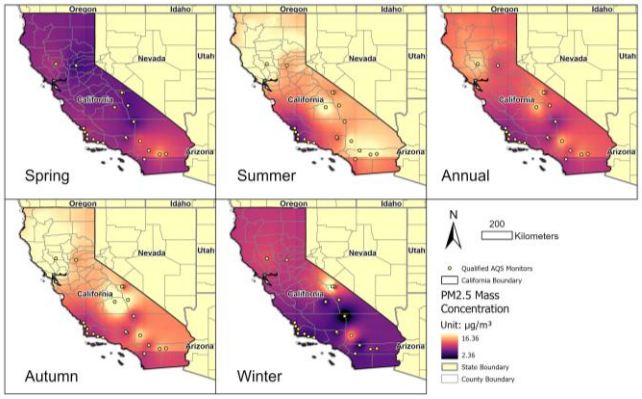

Figure 3: Seasonal and Annual Spatial Distribution of PM2.5 Mass Concentration in CA

3.2.2 Case Study: Wildfire, Holiday Celebrations, and Baseline. Using the estimation results generated by our model, we examine two representative events and a quiet-day baseline in panels keyed to local cycles in California (PDT, UTC–7; UTC windows rotated back one slot so that T06=00:00–06:00, T12=06:00–12:00, T18=12:00–18:00, T00=18:00–24:00) (Figure 4). On the Dixie Fire peak day (2021-08-07), maps show a coherent north-state plume extending across the Sacramento Valley and into downwind basins, broadening from T06 to T12, remaining elevated at T18, and easing by T00 as transport shifts; the event exhibits extreme spatial contrast with a color-scale maximum of 216.9 μg/m³ and a minimum near 0.51 μg/m³. On Independence Day (2024-07-04), enhancements are metro-focused rather than statewide, most noticeable over the South Coast Air Basin in the evening–overnight windows (T18–T00), with maxima up to 24.6 μg/m³ and rapid next-morning dilution. The baseline, defined by windows classified as Good (PM2.5 < 12 μg/m³), remains uniformly low with a persistent coast–inland gradient; the Central Valley sits slightly above coastal and mountainous regions and diurnal modulation is modest, consistent with real-world patterns, with a color scale

spanning 2.45–6.27 μg/m³ [10]. Time-of-day profiles summarize these contrasts: during the Dixie Fire the statewide maximum exceeds 200 μg/m³, the mean is elevated at all times and peaks near T12, and variability (≈±1σ) widens markedly; on Independence Day the mean rises only modestly above baseline and the spread widens mainly at T18–T00, while the baseline remains flat and low with a narrow spread.

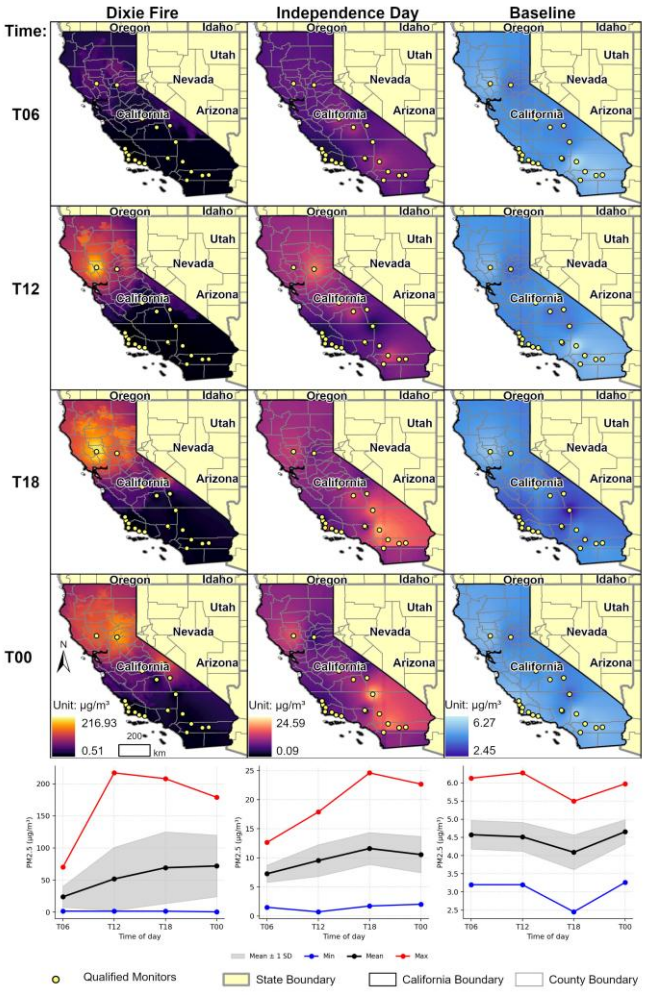

Figure 4: Case-day Spatiotemporal PM2.5 Mass Concentration and Time-of-day Trends in California

4 Conclusion

By fusing ~3 km HRRR Smoke, spatially lagged PM2.5 at 0–24 h, and site coordinates, we produced 6 hourly PM2.5 estimates for California from December 2020 through June 2025 with good sample skill (MAE 0.541 μg/m³; RMSE 1.332 μg/m³; R^2 0.936). Beyond point performance, the derived maps and seasonal/annual composites reveal persistent geography—Central Valley and the South Coast Air Basin as hotspots, coast and mountains lower—with spring cleanest, summer–autumn elevated by smoke transport, and winter inversions creating localized peaks. The statewide mean time series underscores the episodic character of PM2.5 and indicates only a small downward trend (–0.825 μg/m³ yr⁻¹, R^2 0.055). The case study confirms that this 6-hour product resolves both regional smoke intrusions and short-lived urban emissions, underscoring its operational value for situational awareness.

ACKNOWLEDGMENTS

This work was supported by the University of Wisconsin–Madison, Office of the Vice Chancellor for Research and Graduate Education through the Wisconsin Alumni Research Foundation. We also acknowledge NOAA and the U.S. EPA for making the HRRR-Smoke and air quality monitoring datasets publicly available.

REFERENCES

[1] Jing-Xuan Zhou, Zi-Yi Zheng, Zhao-Xing Peng, et al., 2025. Global impact of PM2.5 on cardiovascular disease: Causal evidence and health inequities across region from 1990 to 2021. Journal of Environmental Management 374 (2025), 124168. DOI: https://doi.org/10.1016/j.jenvman.2025.124168

[2] Shaolong Feng, Dan Gao, Fen Liao, et al., 2016. The health effects of ambient PM2.5 and potential mechanisms. Ecotoxicology and Environmental Safety 128 (2016), 67–74. DOI: https://doi.org/10.1016/j.ecoenv.2016.01.030

[3] Qian Di, Hassan Amini, Lanyu Shi, et al., 2019. An ensemble-based model of PM2.5 concentration across the contiguous United States with high spatiotemporal resolution. *Environment International* 130 (2019), 104909. DOI: https://doi.org/10.1016/j.envint.2019.104909

[4] Marissa L. Childs, Jessica Li, Jeffrey Wen, et al., 2022. Daily local-level estimates of ambient wildfire smoke PM2.5 for the contiguous US. *Environmental Science & Technology* 56, 19 (2022), 13607–13621. DOI: https://doi.org/10.1021/acs.est.2c02934

[5] Hui Dai, Yumeng Liu, Jianghao Wang, et al., 2023. Large-scale spatiotemporal deep learning predicting urban residential indoor PM2.5 concentration. *Environment International* 182 (2023), 108343. DOI: https://doi.org/10.1016/j.envint.2023.108343

[6] Zhige Wang, Yue Zhou, Ruiying Zhao, et al., 2021. High-resolution prediction of the spatial distribution of PM2.5 concentrations in China using a long short-term memory model. *Journal of Cleaner Production* 297 (2021), 126493. DOI: https://doi.org/10.1016/j.jclepro.2021.126493

[7] Xinxin Ye, Pouya Arab, Rustem Ahmadov, et al., 2021. Evaluation and intercomparison of wildfire smoke forecasts from multiple modeling systems for the 2019 Williams Flats fire. *Atmospheric Chemistry and Physics Discussions* (2021), 1–69. DOI: https://doi.org/10.5194/acp-21-14427-2021

[8] Leo Breiman, 2001. Random forests. *Machine Learning* 45, 1 (2001), 5-32. DOI: https://doi.org/10.1023/A:1010933404324

[9] Mark R. Segal, 2004. Machine Learning Benchmarks and Random Forest Regression. *UCSF: Center for Bioinformatics and Molecular Biostatistics*, University of California, San Francisco, CA. URL: https://escholarship.org/uc/item/35x3v9t4

[10] Naizhuo Zhao, Ying Liu, Jennifer K. Vanos, et al., 2018. Day-of-week and seasonal patterns of PM2.5 concentrations over the United States: Time-series analyses using the Prophet procedure. *Atmospheric Environment* 192 (2018), 116–127. DOI: https://doi.org/10.1016/j.atmosenv.2018.08.050