Cold-Start Hyperlocal PM2.5: A Leakage-Controlled Comparison of Interpolation, Multimodal Fusion, and Contrastive Representation Learning

Siam Shibly Antar
School of Computer Science
McGill University
Montreal, Canada
siam.antar@mail.mcgill.ca

Syem Shibly Ador
School of Computer Science
Macquarie University
Sydney, Australia
syemshibly.ador@students.mq.edu.au

***Abstract*—Estimating PM2.5 at unmonitored locations during a wildfire smoke episode is usually framed as a fusion problem. The goal is to produce a single estimate. This estimate combines three data sources. The first source is sparse regulatory monitors. The second source is dense, low-cost sensors. The third source is satellite aerosol retrievals. We test that framing under a leakage-controlled cold-start protocol over the 2020 San Francisco Bay Area smoke season, holding out each of the 15 reference monitors in turn and scoring the transfer to the held-out site. The result is deflationary for the learned approaches. Classical spatial interpolation is the strongest estimator: ordinary kriging reaches an RMSE of 6.10 μg/m^3, against 8.39 for the best supervised tabular fusion and 7.39 for learned graph kriging under the same protocol, and the ordering persists as the network is thinned and on smoke days. Richer inputs do not narrow the gap. Variance-optimal reduction is statistically indistinguishable from the supervised bottleneck at 15 folds, so no representation headroom is detectable, an absent gap rather than proven equivalence. A reference-anchored contrastive pilot aligns the modalities but does not become target-predictive: it fails to encode concentration even in-sample, across eight seeds, which places the failure in the objective and not in generalization. Among the inputs, the low-cost network carries the predictive signal while satellite aerosol optical depth adds little, a non-confounded attribution from train-time ablation and permutation importance rather than a missing-modality floor effect. We present these as calibration findings for the monitored near field, not the far field, and release the dataset assembly, protocol, and experiments for reproduction.**

***Index Terms*—air quality, PM2.5, spatial interpolation, multimodal fusion, contrastive representation learning, cold-start benchmark**

I. INTRODUCTION

Fine particulate matter (PM2.5) is a leading environmental health risk, and its burden falls unevenly across neighborhoods [1]. Wildfire smoke sharpens the problem. Surface concentrations vary by large factors over a few kilometers and a few hours, so a useful exposure estimate must be hyperlocal, resolved at the scale of a city block. Regulatory reference monitors deliver the accuracy this needs, but they are sparse, and almost every location is left unmonitored.

Two data sources have emerged to address this gap. Dense low-cost sensor networks such as PurpleAir report PM2.5 once their systematic bias is corrected [2], [3], and satellite aerosol retrievals, MODIS MAIAC aerosol optical depth (AOD) and the Sentinel-5P absorbing aerosol index, cover wide areas that many models tie to surface PM2.5 [4]. A sparse accurate reference network is one data source. A dense noisy low-cost network is another. A broad satellite signal is a third. These are naturally read as three modalities to fuse. A parallel line of work takes a different approach. It recasts estimation at unmonitored sites as learned spatiotemporal kriging with graph networks [5], [6]. The success of contrastive multimodal representation learning [7], [8] raises the further possibility that a shared latent aligned across the three sources could outperform any single source.

The setting that matters in deployment is the cold-start case: estimate PM2.5 where there is no monitor. This is spatial transfer, and it is exactly where in-sample or random splits mislead. Concentrations are spatially autocorrelated. This creates a problem for evaluation. A poor split lets a model see nearby monitors. Those nearby monitors effectively lend the model their labels. This inflates the model's apparent accuracy [9], [10]. We therefore score every method under one leave-one-site-out (LOSO) protocol that removes an entire monitor. We assemble a multimodal dataset for the 2020 San Francisco Bay Area wildfire season and compare three families under it: classical spatial interpolation, supervised multimodal fusion, and a reference-anchored contrastive representation. Across the methods we test, the results are consistently deflationary for the learned approaches. Most tellingly, the contrastive latent fails to encode PM2.5 even in-sample, which locates the failure in the training objective rather than in generalization or seed variance. We offer these results as a benchmark and a characterization, not as a new method.

Contributions.

1) **A leakage-controlled cold-start benchmark.** We assemble a documented multimodal dataset for the 2020 wildfire

 Accepted for publication in the Proceedings of the 2nd International Conference on Data Science and Geoinformatics (ICDSG) 2026. This is the accepted version; the final published version will be available via IEEE Xplore.

season and define one leave-one-site-out protocol that evaluates transfer to entirely held-out reference monitors, the regime that matters for hyperlocal deployment.
2) **Interpolation over fusion.** Under this protocol, classical spatial interpolation outperforms every supervised tabular fusion model and both learned graph-kriging baselines (KCN, IGNNK) we test, on both non-smoke and smoke days and under network sparsification.
3) **No detectable representation headroom.** Variance-optimal reduction is statistically indistinguishable from the supervised accuracy ceiling, so a learned latent shows no advantage over the leading variance directions at this sample size.
4) **Alignment without prediction.** A single-seed reference-anchored contrastive pilot aligns the modalities but does not become target-predictive, and its missing-modality stability reflects uniform inaccuracy rather than a robustness advantage.
5) **A non-confounded modality attribution.** Using train-time ablation and permutation importance, we show that the low-cost network, not satellite aerosol optical depth, carries the predictive signal, and we release the dataset assembly and all experiments for reproduction.

We scope every claim to the methods evaluated here and to the near and mid field that LOSO validates; the far field is discussed but not tested.

II. RELATED WORK

Hyperlocal PM2.5 estimation draws on three strands of work that have advanced largely apart: the calibration and use of low-cost and satellite air-quality data, classical and learned spatial interpolation, and contrastive representation learning. We take each in turn and contrast its usual evaluation with the held-out-monitor cold-start setting studied here.

Low-cost sensors and satellite AOD. Low-cost optical sensors overestimate PM2.5 and drift with humidity, yet their bias is systematic enough to correct. The US-wide Barkjohn correction calibrates PurpleAir against reference monitors and extends to wildfire smoke [2], [3], and independent work confirms good smoke-season performance while flagging dust as out of distribution [11]. Satellite column AOD, processed with MAIAC [12], is tied to surface PM2.5 by statistical and deep models [4], but it thins on the heaviest smoke days, which motivates gap-tolerant designs [13]. Our coverage analysis reproduces that thinning.

Interpolation and learned kriging. Against these inputs, geostatistical interpolation and land-use regression remain strong air-quality baselines [14], yet they are seldom the explicit comparison point for learned models. A growing literature instead recasts estimation at unmonitored locations as inductive spatiotemporal kriging with graph networks [5], [6]. These models are normally scored by masking nodes inside a monitored region, so the held-out target still keeps same-region neighbors at training time. Our protocol is harder: it removes an entire monitor, and we re-evaluate two such models, KCN and IGNNK, under it rather than under their native node-masking (Section VII). A separate line of work takes another approach. It replaces station-ID embeddings with frozen geospatial foundation-model priors. The goal is to help learned forecasters transfer to unseen monitors [15]. This is a learned-prior direction. It is distinct from the interpolation and fusion families we compare.

Representation learning and evaluation. Contrastive objectives align heterogeneous modalities in a shared space, as in CLIP [7] and ImageBind [8], built on the InfoNCE loss of contrastive predictive coding [16]. We instantiate one such latent, anchored to the reference monitors, for the three input modalities, and probe whether that alignment yields a target-predictive representation. Variance-optimal reduction, a standard tool in atmospheric science [17], is our reference point for representation quality. Evaluation is the other concern. Because cold-start estimation is spatial transfer, it must guard against autocorrelation leakage. Spatially blocked cross-validation does this [9], but a full-range buffer can be over-pessimistic and bias accuracy downward [10], which motivates distance-matched alternatives [18]. We therefore adopt plain leave-one-site-out as primary and report buffered variants as sensitivities (Section VI).

III. PROBLEM SETUP

We study hyperlocal PM2.5 estimation as a cold-start spatial-transfer problem. The task is to estimate daily mean PM2.5 at a target location with no co-located reference monitor, from the modalities observed there, using a model trained only on other monitored locations. Write y for the reference concentration at a monitored site and x_m for modality $m \in \{\text{low-cost}, \text{satellite}, \text{meteorology}\}$ with its spatiotemporal context c. Training sees paired $(\{x_m\}, c, y)$ at monitored sites. Prediction at a cold-start cell sees only $\{x_m\}$ and c, and never y or a co-located reference.

The defining difficulty is leakage. Because concentrations are spatially autocorrelated, a model scored on cells near its training monitors can score well simply by interpolating those training labels, rather than by using the local modalities, and random or in-sample splits reward exactly this shortcut. A faithful benchmark must therefore separate the held-out target from the training labels in space, which is what the leave-one-site-out protocol of Section VI enforces. We pose the question as a comparison. We use one fixed leakage-controlled protocol. We ask which family transfers best to an unmonitored location. We also ask whether added modeling complexity pays off. Specifically, does it yield accuracy that simple interpolation does not already capture?

IV. DATA AND STUDY REGION

The benchmark is built on a single, well-characterized smoke episode, so that every method is stressed on the same high-concentration days. We describe the study region, the four modalities, and the co-location structure that constrains how each modality can be used.

Study region and modalities. We study the San Francisco Bay Area, from 37.2 to 38.3°N and -122.7 to -121.5°E, over the 92 days from 1 August to 31 October 2020, which span the most severe Northern California smoke episodes in the region's modern record. Predictions are made on a common grid of 16,482 cells at roughly 1 km, defined by the native MAIAC AOD pixels. All concentrations are daily means, and a day is a *smoke day* when the reference-network daily mean PM2.5 exceeds 35 μg/m^3, which holds on 11 of the 92 days. The smoke-day subset is where the benchmark is

most demanding, since it holds the concentrations that matter most for exposure.

The dataset assembles four modalities. *Reference* PM2.5 comes from 15 EPA monitors, 1,370 monitor-days after averaging valid records per site-day. *Low-cost* PM2.5 comes from 282 PurpleAir sensors, 23,800 sensor-days and 23,006 after A/B channel quality control and the Barkjohn correction with its wildfire extension [2], [3]. *Satellite* inputs are MODIS MAIAC AOD at 550 nm, which also defines the grid, together with the Sentinel-5P absorbing aerosol index; AOD covers 76.4% of cell-days overall but only 69.1% on smoke days [4], [13]. *Meteorology* comes from ERA5 reanalysis, namely boundary-layer height, relative humidity, wind speed, and temperature, bilinearly interpolated to the grid.

The four modalities co-locate at different radii by design, and this matters for how each can be used. The gridded fields are read at a monitor's own cell, where AOD is present on 85.3% of monitor-days. The low-cost signal, by contrast, is aggregated within 5 km of each monitor, because a PurpleAir sensor sits in the monitor's own cell on only 6.7% of monitor-days, so it is a neighborhood aggregate rather than a co-location. All four modalities co-occur in a single cell-day only 1.2% of the time, so complete multimodal vectors are rare and robustness to missing modalities is of practical interest. The full assembly is documented and reproducible from the released code and read-only sources.

V. METHOD

We compare three families of estimators under the common cold-start protocol: classical spatial interpolation, supervised multimodal fusion, and contrastive representation learning. The first two are standard references. The third is a reference-anchored contrastive pilot, STRATUS, which we introduce as a feasibility probe rather than a tuned system. Figure 1 sketches its pipeline and Algorithm 1 gives its training and cold-start inference.

Estimator families. Three families share one protocol and differ only in what they use to predict. Interpolation uses ordinary kriging with a fitted variogram and inverse-distance weighting, on the reference values and coordinates alone, so it reads no modalities. Supervised fusion covers ridge, random forest, gradient-boosted trees (XGBoost), an MLP, and partial least squares, all tabular regressors trained on an enriched design matrix of 273 predictors that augments the raw modality values with engineered spatiotemporal context: 50 low-cost, 196 satellite, and 24 meteorology features over a shared three-feature block of latitude, longitude, and day of season, with documented imputation. We additionally evaluate two learned graph-kriging models, KCN and IGNNK, adapted to the held-out-monitor protocol.

Representation methods, PCA, PLS, an MLP bottleneck, an end-to-end MLP, UMAP, and the contrastive latent STRATUS, each reduce the inputs to $d \in \{8, 16\}$ dimensions that a frozen linear probe maps to PM2.5, which isolates whether a latent linearly encodes the target. STRATUS is our reference-anchored pilot (Algorithm 1, Fig. 1). It encodes each modality, with its spatiotemporal context, through an identical two-hidden-layer MLP, and aligns the embeddings to an encoded reference anchor with an InfoNCE objective [16] over co-located monitor-days. At prediction the reference is absent, and the latent is the masked mean of the present modality embeddings. STRATUS runs at one fixed configuration (width 64, 200 epochs, temperature 0.07) and is not tuned, so we read it as a feasibility probe rather than a tuned upper bound. Full details are in the released code.

Algorithm 1 STRATUS contrastive training and cold-start inference

Require: monitor-days with modality features $\{x_m\}_{m \in M}$, context c, reference PM2.5 y; temperature τ; dimension d

initialize modality encoders $\{f_m\}_{m \in M}$ and reference encoder g, each an MLP $\cdot \to 64 \to 64 \to d$
for each training epoch **do**
  for minibatch of monitor-days **do**
    for $m \in M$ **do**
      $z_m \leftarrow \mathrm{normalize}(f_m([x_m;\, c]))$ ▷ L2-normalized modality embedding
    end for
    $a \leftarrow \mathrm{normalize}(g([y;\, c]))$ ▷ reference anchor
    $\mathcal{L} \leftarrow \sum_{m \in M} \mathrm{InfoNCE}(z_m,\, a;\, \tau)$ ▷ pull each modality to the anchor
    update $\{f_m\}$, g to minimize $\mathcal{L}$
  end for
end for
fit a frozen ridge probe h on pooled training embeddings to predict y
Cold-start prediction at an unmonitored cell (y and a unavailable):
$z_m \leftarrow \mathrm{normalize}(f_m([x_m;\, c]))$ for each present modality m
$\bar{z} \leftarrow$ mean of the present embeddings $\{z_m\}$ ▷ masked mean pool
return $\hat{y} \leftarrow h(\bar{z})$

VI. EXPERIMENTAL SETUP

Evaluation measures cold-start transfer without leakage, and reports every difference with calibrated uncertainty. We describe the leave-one-site-out protocol and its buffered sensitivities, then the paired significance test.

Cold-start protocol. Our primary protocol is leave-one-site-out (LOSO) over the 15 reference monitors. Each fold holds out one monitor, with all 92 of its monitor-days as test, and trains on the other 14. This scores spatial transfer to an unmonitored location. We report RMSE, MAE, and R^2 pooled over the held-out monitor-days, split into smoke and non-smoke days.

Spatially blocked cross-validation prevents leakage [9], but it can be overly pessimistic [10], which is the case in this network. The empirical PM2.5 variogram range, 41.5 km, exceeds the median inter-site distance of 39.6 km, so a full-range buffer starves 11 of the 15 folds down to as few as three training monitors, which is the over-pessimistic regime rather than a valid control. We therefore keep plain LOSO as primary and retain a 41.5 km buffered LOSO and a 3 km small-buffer LOSO, the latter following distance-matched validation [18], as sensitivities; a random five-fold split is reported only for comparability.

The open grid is a harder setting than the held-out monitors. Its cells lie a median of 15.7 km from the nearest monitor, 8.1% within 5 km and the farthest beyond 60 km, whereas the held-out monitors lie 2.9 to 29.1 km from their nearest training site. The benchmark therefore probes the near and mid field, and does not validate the far field, a point we revisit in Section VIII.

Significance. For each contrast we run a paired Wilcoxon signed-rank test over the 15 per-fold RMSE values, with Holm correction inside each comparison family, and report the median per-fold difference and the Holm-adjusted p-value. We read a non-significant difference as an absent gap at this sample size, not as proven equivalence. Two caveats keep the reading conservative. Each method runs at a single seed, so the per-fold values carry seed variation as well as

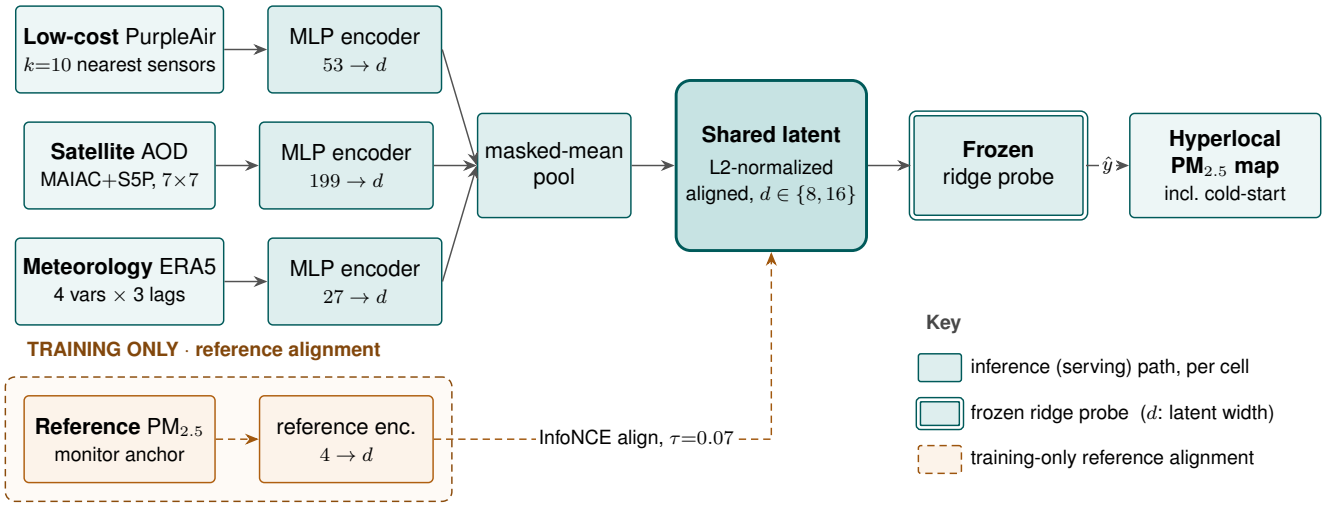

Fig. 1. **The STRATUS reference-anchored contrastive model.** Each modality block (with spatiotemporal context) is mapped by an identical MLP encoder ($\cdot\to64\to64\to d$, L2-normalized, $d \in \{8, 16\}$) that differs only in input width. An InfoNCE objective ($\tau = 0.07$) over co-located reference monitor-days aligns each modality embedding to the encoded reference anchor; at prediction the latent is the masked mean of the present modality embeddings, mapped to PM2.5 by a frozen ridge probe. The reference stream (dashed) is training-only and absent at cold-start cells.

TABLE I
COLD-START LEAVE-ONE-SITE-OUT ACCURACY BY METHOD, POOLED OVER THE 15 HELD-OUT MONITORS. RMSE, MAE, AND THE SMOKE AND NON-SMOKE RMSE ARE IN μG/M^3. INTERPOLATION USES ONLY THE REFERENCE FIELD; SUPERVISED FUSION USES THE ENRICHED MULTIMODAL FEATURES; LEARNED GRAPH KRIGING (KCN, IGNNK) IS EVALUATED UNDER THE IDENTICAL HELD-OUT-MONITOR PROTOCOL. THE HYBRID IS THE ENRICHED MLP WITH THE ORDINARY-KRIGING ESTIMATE AND THE THREE NEAREST TRAINING-MONITOR REFERENCE VALUES ADDED AS FEATURES, UNDER THE SAME HELD-OUT EXCLUSION. BOTH INTERPOLATORS OUTPERFORM EVERY LEARNED MODEL.

Method	RMSE	MAE	R^2	Smoke	Non-smoke
Interpolation					
Ordinary kriging	6.10	3.02	0.943	14.52	3.66
IDW	6.46	3.13	0.936	15.21	3.96
Regression kriging	10.60	5.03	0.827	25.77	6.08
Supervised fusion					
Ridge	12.47	6.78	0.761	30.51	7.02
Random forest	9.43	4.37	0.863	22.87	5.43
XGBoost	8.94	4.62	0.877	20.82	5.62
MLP (enriched)	8.39	4.79	0.892	18.73	5.66
PLS (enriched)	9.31	5.75	0.867	20.89	6.23
Hybrid (MLP)	7.41	4.26	0.916	17.00	4.78
Learned graph kriging					
KCN	7.39	4.53	0.916	15.94	5.22
IGNNK	10.05	5.98	0.845	23.79	6.11

TABLE II
FROZEN-LATENT LINEAR-PROBE ACCURACY UNDER COLD-START LOSO AT d=16 AND d=8. RMSE IN μG/M^3. VARIANCE-OPTIMAL PCA MATCHES THE SUPERVISED BOTTLENECK AND THE REFERENCE-ANCHORED CONTRASTIVE LATENT (STRATUS) IS THE WEAKEST, AT BOTH LATENT WIDTHS.

	d=16		d=8	
Latent (linear probe)	RMSE	R^2	RMSE	R^2
PCA	9.09	0.873	9.23	0.869
PLS	9.31	0.867	9.28	0.868
MLP bottleneck	8.29	0.894	8.45	0.890
MLP end-to-end	8.39	0.892	8.49	0.889
UMAP	17.22	0.544	20.52	0.353
STRATUS (contrastive)	20.87	0.331	21.94	0.261

fold variation. The contrasts are also pairwise within families rather than one factorial model, which is conservative for the interaction structure of the dropout and ablation experiments.

VII. RESULTS

We organize the results around the five claims previewed in Section I: the accuracy ordering across families (C1, C2), the behavior of the contrastive latent (C3), its response to missing modalities (C4), and the source of the predictive signal (C5). Tables I and II give the headline numbers, and the figures isolate where and why the ordering holds.

Interpolation outperforms supervised fusion and graph kriging. Table I and Figure 2 report cold-start accuracy, and the ordering is clear. Ordinary kriging is the most accurate estimator at 6.10 μg/m^3 RMSE (R^2 0.943), followed closely by IDW at 6.46. The strongest supervised model, the enriched MLP, reaches only 8.39, and the rest fall behind it (XGBoost 8.94, random forest 9.43, enriched PLS 9.31, ridge 12.47). Both interpolators therefore outperform every fusion model we test (C1). The paired test confirms it: kriging against the enriched MLP gives a median per-fold RMSE difference of -2.11 μg/m^3 in kriging's favor (Holm $p = 0.0046$).

Graph kriging is the strongest learned competitor under the same protocol. KCN reaches 7.39 μg/m^3, outperforming every tabular fusion model, and IGNNK reaches 10.05, yet both still fall short of the interpolators (kriging vs KCN median -1.52, Holm $p = 0.0084$; vs IGNNK -4.21, Holm $p = 0.0001$). Every method worsens by a factor of three to four on smoke days, but the ordering holds (kriging 14.52 against the enriched MLP 18.73). The gap is not an artifact of network density either. Kriging stays at or below the fusion methods across held-out distances, and thinned to four training monitors it remains more accurate than the tree models (Figure 3; near 8.5 μg/m^3 at k=4 against its full-network 6.10).

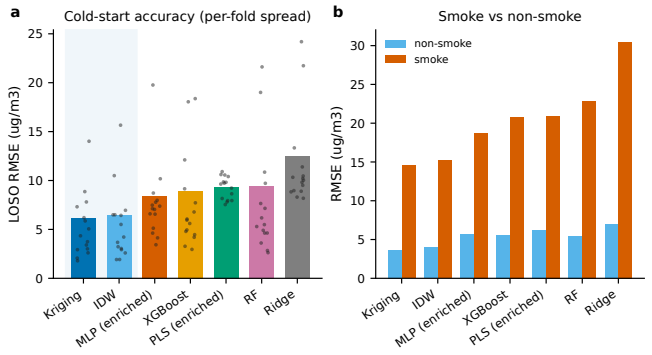

Fig. 2. **Cold-start leave-one-site-out accuracy.** (a) RMSE by method with the per-fold spread shown as points; the two interpolators (kriging, IDW) are the most accurate. (b) RMSE split into non-smoke and smoke days: every method worsens by a factor of three to four on smoke days, but the ordering is preserved.

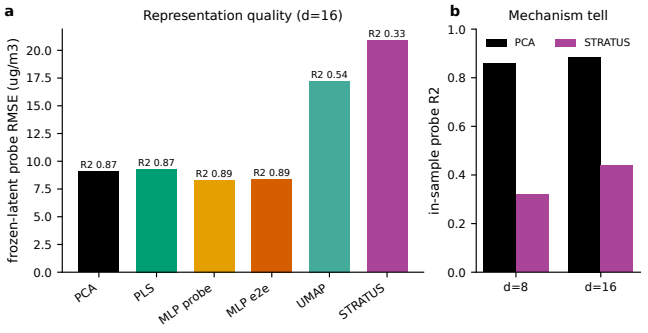

Fig. 4. **Representation probing at d=16.** (a) Frozen-latent linear-probe RMSE by method (with R^2 annotated); variance-optimal PCA matches the supervised bottleneck, while the contrastive latent (STRATUS) is the weakest. (b) In-sample probe R^2 for PCA and STRATUS at $d \in \{8, 16\}$: STRATUS fails to encode PM2.5 even in-sample.

No detectable representation headroom. Table II and Figure 4 probe the learned representations, and two results stand out. The first is that there is no detectable headroom for dimensionality reduction. Variance-optimal PCA reaches 9.09 μg/m^3 (R^2 0.873), and its gap to the supervised MLP bottleneck at 8.29 is not statistically significant (paired PCA against MLP bottleneck, Holm $p = 0.0707$; PCA against PLS, $p = 0.0707$). We read this as an absent gap rather than proven equivalence, especially as the point estimate slightly favors the bottleneck. Either way, a learned latent shows no advantage over the leading variance directions here (C2).

The second result is that the single-seed contrastive pilot is the weakest representation by a large margin. STRATUS probes to 20.87 μg/m^3 (R^2 0.331), below even UMAP (17.22), with a median per-fold gap to PCA of 12.07 μg/m^3 (Holm $p = 0.0004$). This is not a generalization gap. On training data the contrastive latent reaches an in-sample probe R^2 of only 0.439 at d=16 (0.319 at d=8), against PCA's 0.886 (0.859), so it fails to linearly encode PM2.5 even in-sample, at both widths (Figure 4b). Across eight seeds this in-sample R^2 at d=16 is 0.424 $\pm$0.014, a stable property of the objective rather than a seed artifact. Figure 5 makes the mechanism visible: the contrastive latent co-mingles the modalities (panel a, alignment succeeds) but is not organized by concentration (panel b), whereas PCA shows a clean PM2.5 gradient (panel c). In this configuration the latent is aligned but not target-predictive (C3). We do not claim contrastive alignment cannot help in general, only that this pilot does not.

At d=8 the frozen-probe ordering matches d=16 (Table II). The supervised bottleneck keeps the same small margin over variance-optimal PCA (8.45 versus 9.23 μg/m^3), and the contrastive probe remains far worse (21.94). The no-headroom reading (C2) therefore does not depend on the latent width.

Missing-modality robustness is a floor effect. We next withhold modalities at inference time. This is an intentionally confounded stress test, whose distribution shift we untangle from importance in the train-time analysis below. When the low-cost block is removed, the supervised MLP collapses (Figure 6), from 8.39 μg/m^3 overall to 28.06, and from 18.73 to 74.85 on smoke days, while the contrastive latent barely moves (20.87 to 22.00). The contrasts are statistically detectable (MLP degradation median -21.22 μg/m^3, Holm $p = 0.0004$; STRATUS against the MLP under the same dropout, median -6.12, $p = 0.0006$; STRATUS's own change -0.71 μg/m^3, $p = 0.0046$).

We do not read this as a robustness advantage of contrastive alignment. STRATUS is uniformly inaccurate, near 20.87 μg/m^3, so a dropped modality changes little, and it surpasses the MLP only under the dropout that destroys the MLP, never in absolute accuracy. The stability is a floor effect (C4): the aligned latent degrades gracefully, but that would

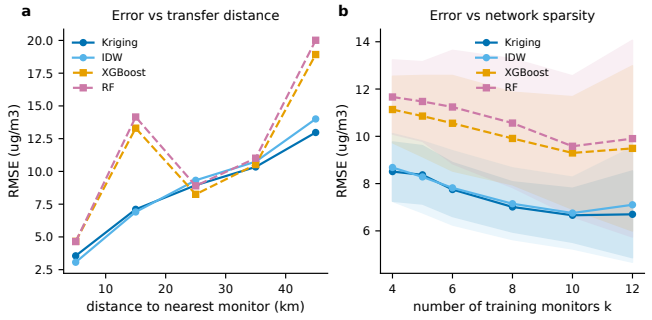

Fig. 3. **The accuracy gap is not an artifact of network density.** (a) RMSE against the distance from the held-out monitor to its nearest training site; (b) RMSE as the training network is thinned to k monitors, with interquartile bands. Kriging and IDW stay at or below the tree models throughout, including at k=4.

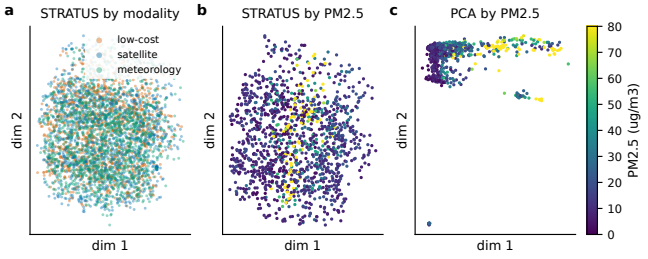

Fig. 5. **Illustrative geometry of the contrastive latent.** (a) STRATUS embeddings colored by modality co-mingle, showing that alignment succeeds; (b) the same embeddings colored by PM2.5 show no clear concentration structure; (c) the PCA embedding colored by PM2.5 shows a clean gradient. The contrastive latent is aligned but not target-predictive.

not help a deployable estimator.

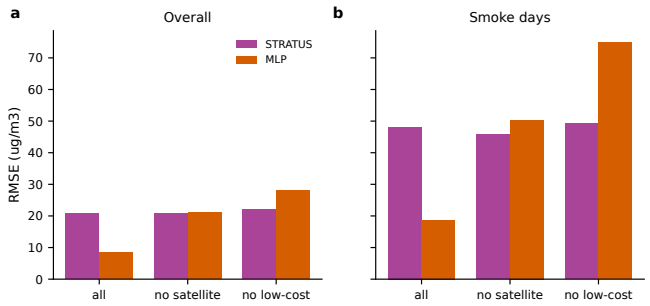

Fig. 6. **Missing-modality behavior under inference-time dropout.** Overall (a) and on smoke days (b), for the contrastive latent (STRATUS) and the supervised MLP across the all, no-satellite, and no-low-cost conditions. The MLP is accurate with all modalities but degrades sharply when the low-cost block is withheld, whereas STRATUS barely changes, a floor effect rather than a robustness advantage.

Low-cost carries the signal. Figure 7 attributes the predictive signal, using train-time ablation and permutation importance. Unlike the inference-time dropout above, these are not confounded by distribution shift, and this analysis, not the stress test, is what licenses the importance claim. Removing satellite from the enriched MLP barely changes accuracy (8.39 to 8.93 μg/m^3), whereas removing low-cost degrades it sharply (to 12.71), and gradient-boosted trees show the same pattern (Fig. 7b). A low-cost-plus-context MLP nearly matches the full model (9.70), while satellite-plus-context is the weakest single modality (19.09). Permutation importance on the full MLP agrees: a low-cost RMSE increase of 24.67 μg/m^3 dwarfs satellite (3.15) and meteorology (2.11), a roughly eightfold margin. Both model families agree on direction. The low-cost network carries the predictive signal, satellite AOD adds little, and meteorology is informative but largely redundant (C5).

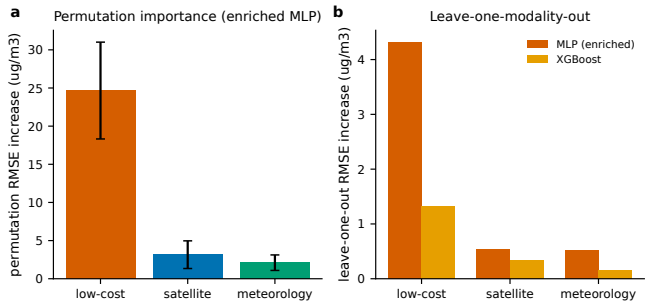

Fig. 7. **Modality contribution for the enriched MLP.** (a) Permutation importance, as RMSE increase, for the three modality blocks; (b) leave-one-modality-out RMSE increase for the MLP and XGBoost. Low-cost dominates; satellite AOD adds little.

VII-A. Computational cost

Cost separates the methods along a train-once versus predict-often axis rather than a single ranking. Ordinary kriging has almost no fitting cost but does its work at query time. Its per-fold prediction (0.564 s) is therefore the slowest in Table III apart from the graph models, roughly three orders of magnitude above the enriched learners' inference (0.0002 s for the MLP). The learned graph models are the most expensive overall (KCN 17.74 s and IGNNK 8.44 s to fit). This does not weaken C1. The models with the cheapest inference are exactly the ones that are least accurate, so the query-time work is what makes kriging accurate. The hybrid pays that same cost through its kriging feature (0.573 s per fold to build) on top of the MLP, and still falls short of plain kriging. These are small-data timings measured on a single held-out site of roughly 1,370 rows on CPU; kriging's prediction cost is the one that grows with query volume, which is the relevant caveat for dense deployment.

TABLE III

PER-FOLD FIT AND PREDICT WALL-CLOCK TIME BY METHOD UNDER THE LEAVE-ONE-SITE-OUT PROTOCOL, AVERAGED OVER THE 15 FOLDS. ALL METHODS RAN ON THE CPU (13TH GEN INTEL CORE I7-13650HX); THE MACHINE'S GPU (RTX 4060) WAS AVAILABLE BUT UNUSED BY THE PIPELINE. RANDOM FOREST AND XGBOOST USE ALL CORES ($n_{\text{jobs}}=-1$); EVERY OTHER METHOD IS SINGLE-THREADED. INTERPOLATION HAS NO SEPARATE TRAINING PASS, SO ITS FIT TIME IS ZERO AND THE REGRESSION-KRIGING FIT IS ITS LINEAR DRIFT ONLY. THE MLP HYBRID ADDITIONALLY BUILDS ITS KRIGING FEATURES AT INFERENCE, ABOUT 0.573 S/FOLD, NOT INCLUDED IN ITS PREDICT TIME.

Method	Fit (s)	Predict (s)
Ordinary kriging	0.0000	0.564
IDW	0.0000	0.031
Regression kriging	0.012	0.431
Ridge	0.020	0.0015
Random forest	0.457	0.046
XGBoost	0.101	0.0023
Enriched MLP	0.908	0.0002
Enriched PLS	0.020	0.0002
Hybrid (MLP)	0.869	0.0002
KCN	17.74	0.013
IGNNK	8.44	0.027

VIII. DISCUSSION

Our findings are deflationary for the learned approaches. The same mechanism explains why interpolation prevails, why there is no representation headroom, and why the contrastive latent fails, and we develop it here before turning to the information-parity objection.

Interpretation. Interpolation prevails at cold-start for two reasons. The held-out monitors lie within the reference network's spatial autocorrelation range, and the most informative learned modality, the low-cost network, is itself a dense neighborhood aggregate (Section IV). The field of nearby reference values therefore already encodes most of what the local features could add, while a flexible model pays for its flexibility in variance. The contrastive latent aligns but does not predict because InfoNCE organizes the latent by source monitor-day rather than by concentration [16]. The no-headroom result shows why this is hard to fix by this route: variance-optimal PCA already sits at the supervised ceiling [17], so the predictive signal lies in the leading variance directions that a target-unaware objective cannot improve on. Three caveats scope the claim: the near and mid field only, a single-seed pilot, and an absence of a detectable gap rather than equivalence (Section IX); this is a result for this regime, not a universal verdict.

Information parity. A natural objection to C1 is that kriging reads the training monitors' same-day reference values while the fusion models do not, so the comparison

might reward information access rather than method. We test two natural remedies under the same leave-one-site-out exclusion. Regression kriging, a global drift over the tabular features plus kriged residuals, is worse than plain kriging at 10.60 μg/m^3 (1.35 μg/m^3/fold in kriging's favor, Holm p =0.0079). The global trend degrades accuracy where the local reference field already dominates, most of all on smoke days. The kriging-as-feature hybrid, which adds the ordinary-kriging estimate and the three nearest training-monitor reference values as inputs, moves the enriched MLP from 8.39 to 7.41 μg/m^3, a significant gain over plain fusion (0.98 μg/m^3/fold, Holm p =0.0125). It remains short of plain kriging (1.54 μg/m^3/fold in kriging's favor, Holm p =0.0079).[1] Even when given the reference-neighbor information that kriging uses, the pre-registered fusion model does not reach it, so the remaining C1 gap is not merely one of information access.

IX. LIMITATIONS

Several boundaries scope the claims rather than weaken them; we state each with the follow-up that would settle it.

Single region and season. The study covers one smoke season in one metropolitan network, so the method ordering could shift where the reference network is sparser or the satellite retrieval is cleaner. *Planned:* replication across multiple regions and seasons as the primary external-validity test.

Single-seed figures. The reported STRATUS figures come from one seed; the eight-seed sweep in Section VII confirms the in-sample failure is a property of the objective, not a seed artifact. *Planned:* a light configuration search.

Absence of a gap, not equivalence. The no-headroom result is a non-significant paired test at 15 folds, which is evidence of no detectable gap, not a formal equivalence. *Planned:* a two-one-sided-tests procedure against a pre-registered margin to turn this into a positive claim.

Near and mid field only. The protocol validates transfer to held-out monitors inside the autocorrelation range, not the open grid's far field. *Planned:* denser reference or campaign data to probe the far-field tail.

X. CONCLUSION

Under a single leakage-controlled cold-start protocol for wildfire-season hyperlocal PM2.5, the learned approaches did not earn their complexity. Classical spatial interpolation outperformed supervised tabular fusion and learned graph kriging. Variance-optimal reduction was statistically indistinguishable from the supervised accuracy ceiling, with no detectable headroom. A reference-anchored contrastive pilot aligned the modalities without yielding a target-predictive representation. Among the inputs, the low-cost network carried the predictive signal while satellite AOD added little. These results map where added modeling complexity helps and where it does not in this regime, and they argue for investing in monitoring density and sound interpolation before fusion or representation learning. Future work should validate the far field, evaluate target-aware representation objectives, and extend the benchmark across regions and seasons. The dataset assembly, evaluation protocol, and all experiments are released to support reproduction and reuse.

[1] A gradient-boosted hybrid on the identical augmented features matches kriging (6.04 versus 6.10 μg/m^3, difference not detectable, $p = 0.5614$). We read this as an echo rather than an independent match: the kriging estimate is a supplied input a tree ensemble can track closely, whereas the regularized MLP blends it.

REFERENCES

[1] A. Jbaily, X. Zhou, J. Liu *et al.*, "Air pollution exposure disparities across US population and income groups," *Nature*, vol. 601, pp. 228–233, 2022.

[2] K. K. Barkjohn, B. Gantt, and A. L. Clements, "Development and application of a United States-wide correction for PM2.5 data collected with the PurpleAir sensor," *Atmospheric Measurement Techniques*, vol. 14, no. 6, pp. 4617–4637, 2021.

[3] K. K. Barkjohn, A. L. Holder, S. G. Frederick, and A. L. Clements, "Correction and accuracy of PurpleAir PM2.5 measurements for extreme wildfire smoke," *Sensors*, vol. 22, no. 24, p. 9669, 2022.

[4] Q. Di, H. Amini, L. Shi *et al.*, "An ensemble-based model of PM2.5 concentration across the contiguous United States with high spatiotemporal resolution," *Environment International*, vol. 130, p. 104909, 2019.

[5] Y. Wu, D. Zhuang, A. Labbe, and L. Sun, "Inductive graph neural networks for spatiotemporal kriging," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 4478–4485.

[6] G. Appleby, L. Liu, and L.-P. Liu, "Kriging convolutional networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 3187–3194.

[7] A. Radford, J. W. Kim, C. Hallacy *et al.*, "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

[8] R. Girdhar, A. El-Nouby, Z. Liu *et al.*, "ImageBind: One embedding space to bind them all," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[9] D. R. Roberts, V. Bahn, S. Ciuti *et al.*, "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure," *Ecography*, vol. 40, no. 8, pp. 913–929, 2017.

[10] A. M. J.-C. Wadoux, G. B. M. Heuvelink, S. de Bruin, and D. J. Brus, "Spatial cross-validation is not the right way to evaluate map accuracy," *Ecological Modelling*, vol. 457, p. 109692, 2021.

[11] D. A. Jaffe, C. Miller, K. Thompson, B. Finley, M. Nelson, J. Ouimette, and E. Andrews, "An evaluation of the U.S. EPA's correction equation for PurpleAir sensor data in smoke, dust, and wintertime urban pollution events," *Atmospheric Measurement Techniques*, vol. 16, pp. 1311–1322, 2023.

[12] A. Lyapustin, Y. Wang, I. Laszlo *et al.*, "Multiangle implementation of atmospheric correction (MAIAC): 2. aerosol algorithm," *Journal of Geophysical Research: Atmospheres*, vol. 116, p. D03211, 2011.

[13] Q. Xiao, Y. Wang, H. H. Chang, X. Meng, G. Geng, A. Lyapustin, and Y. Liu, "Full-coverage high-resolution daily PM2.5 estimation using MAIAC AOD in the Yangtze River Delta of China," *Remote Sensing of Environment*, vol. 199, pp. 437–446, 2017.

[14] G. Hoek, R. Beelen, K. de Hoogh *et al.*, "A review of land-use regression models to assess spatial variation of outdoor air pollution," *Atmospheric Environment*, vol. 42, no. 33, pp. 7561–7578, 2008.

[15] S. S. Antar, S. S. Ador, S. H. H. Ding, and B. C. M. Fung, "SatCLIP-GNN: Cold-start PM2.5 forecasting with satellite-derived location priors," in *Proc. IEEE Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS)*, 2026, pp. 116–120.

[16] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[17] A. Hannachi, I. T. Jolliffe, and D. B. Stephenson, "Empirical orthogonal functions and related techniques in atmospheric science: A review," *International Journal of Climatology*, vol. 27, no. 9, pp. 1119–1152, 2007.

[18] C. Milà, J. Mateu, E. Pebesma, and H. Meyer, "Nearest neighbour distance matching leave-one-out cross-validation for map validation," *Methods in Ecology and Evolution*, vol. 13, no. 6, pp. 1304–1316, 2022.