www.itcon.org - Journal of Information Technology in Construction - ISSN 1874-4753

SATELLITE-BASED NETWORK-LEVEL PAVEMENT ROUGHNESS ESTIMATION USING DUAL-POLARIZATION SENTINEL-1 SAR AND DEEP LEARNING

SUBMITTED: April 2026
PUBLISHED: November 2026
EDITOR: Frédéric Bosché
DOI: 10.36680/j.itcon.2026.054

Md Tashfique Enam Tutul
Department of Civil and Architectural Engineering and Construction Management, University of Wyoming, Laramie, WY 82071, USA
https://orcid.org/0009-0007-6713-2269
mtutul@uwyo.edu

Ahmed Abdelaty, PhD (corresponding author)
Department of Civil and Architectural Engineering and Construction Management, University of Wyoming, Laramie, WY 82071, USA
https://orcid.org/0000-0002-6744-3944
aahmed3@uwyo.edu

Austin Madson, PhD
Yale Center for Geospatial Solutions (YCGS), Yale University, New Haven, CT 06511, USA
https://orcid.org/0000-0001-8039-4908
amadson@uwyo.edu

Shuvo Dip Datta
Department of Civil and Architectural Engineering and Construction Management, University of Wyoming, Laramie, WY 82071, USA
https://orcid.org/0000-0001-8604-3148
sdatta1@uwyo.edu

SUMMARY: *Timely pavement-condition information supports network-level maintenance planning, but repeated profiler surveys require dedicated field campaigns. This study develops a scripted workflow that converts openly available Sentinel-1 synthetic aperture radar (SAR) observations into segment-level International Roughness Index (IRI) estimates and pavement-management outputs. A seven-configuration experiment separates SAR inputs (S), pre-inspection engineering information (P), and contemporaneous condition indicators (C) according to their availability relative to the target inspection. The study comprises 8,188 flexible-asphalt pavement segments in Alaska, Utah, and West Virginia, including 1,228 corridor-held-out test observations. SAR alone achieved a mean absolute error (MAE) of 14.8 in/mi. Under retrospective, inspection-aligned evaluation, combining SAR with pre-inspection information reduced MAE to 13.4 in/mi, compared with 16.1 in/mi for pre-inspection information alone. Full fusion achieved an MAE of 12.9 in/mi, an RMSE of 17.2 in/mi, and a squared Pearson correlation of 0.69, indicating only a further 0.5 in/mi improvement over the same-campaign condition indicators. Matched SAR-only AnySat and SkySense adaptations achieved three-seed mean MAEs of 14.2 and 14.5 in/mi, respectively. For the full-fusion predictor, MC dropout and split-conformal intervals achieved 89.6% and 90.2% empirical coverage, respectively, at a nominal 90% level. Uncertainty-guided screening flagged 8.4% of test segments while capturing 35.2% of errors exceeding 25 in/mi. The results support using an availability-aware satellite screening layer to prioritize field measurements within the evaluated domain.*

KEYWORDS: *pavement roughness, synthetic aperture radar, deep learning, input availability, uncertainty quantification, pavement management systems infrastructure informatics.*

REFERENCE: *Tutul, M. T. E., Abdelaty, A., Madson, A., & Datta, S. D. (2026). Satellite-based network-level pavement roughness estimation using dual-polarization Sentinel-1 SAR and deep learning. Journal of Information Technology in Construction (ITcon), 31, 1313-1335. https://doi.org/10.36680/j.itcon.2026.054*

COPYRIGHT*: © 2026 The author(s). This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

1. INTRODUCTION

Pavement management depends not only on knowing where deterioration has occurred but also on knowing when condition information has become too old to support a decision. Inertial profilers remain the reference technology for deriving the International Roughness Index (IRI) from measured longitudinal profiles. Their precision is essential for acceptance testing and engineering assessment, yet network-wide deployment requires vehicles, trained personnel, scheduling, and data processing. The resulting observation cycle can leave substantial intervals between field surveys. A complementary information stream that identifies where new measurements are most valuable could improve the timeliness of maintenance planning without replacing profile-based measurement (Sayers et al., 1986; Sayers & Karamihas, 1998).

Satellite synthetic aperture radar (SAR) provides a potential source of that information. Sentinel-1 supplies repeat C-band observations independently of daylight and cloud cover, with open data access and acquisition geometries suitable for systematic processing (Torres et al., 2012). Its backscatter responds to surface texture, dielectric properties, and geometric configuration. These responses can be associated with pavement condition, but they are not a direct measurement of the longitudinal profile used to compute IRI. Moreover, the pavement occupies only part of many radar resolution cells. A useful SAR-to-IRI model must therefore learn a statistical association at the segment reporting scale while controlling, as far as practical, environmental and geometric variability (Meyer et al., 2020; Moreira et al., 2013).

Earlier research has already demonstrated SAR-based pavement-condition classification, interferometric assessment of road deformation, and deep-learning regression of IRI from Sentinel-1 data (Bashar & Torres-Machi, 2022; Fiorentini et al., 2021; Fiorentini et al., 2020; Karimzadeh & Matsuoka, 2020). The present study extends that foundation by examining how satellite-derived estimates can be organized into a reproducible pavement-management workflow, how much useful accuracy is retained when same-campaign distress measurements are unavailable, and how prediction quality changes under alternative model and preprocessing choices.

Input availability is central to this question. Pavement age, layer thickness, traffic loading, and mixture properties may be available before an IRI survey. Rut depth, mean profile depth, and fatigue cracking are often collected during the same inspection campaign as IRI. Combining all these variables can improve prediction, but the resulting accuracy does not describe an autonomous satellite-only service. Reporting one fused result obscures the distinction between information that can support an update between surveys and information that becomes available only after field inspection. This study makes that distinction explicit through separate SAR, pre-inspection, and contemporaneous input groups.

The contribution is an application-oriented integration and evaluation of a satellite-based pavement management system (PMS). The workflow connects automated SAR preparation, availability-specific regression, condition classification, uncertainty reporting, and a tiered field-confirmation strategy. Its evaluation combines seven input configurations, two contemporary SAR-capable model adaptations, corridor-grouped sensitivity analysis, and an alternative uncertainty estimator. Throughout this paper, the evaluated domain denotes flexible-asphalt segments on corridor-held-out roads in Alaska, Utah, and West Virginia observed with Sentinel-1 C-band under the reported acquisition, preprocessing, temporal-matching, and quality-screening conditions. Subsequent performance interpretations use this definition.

The central distinction concerns the timing of information rather than immediate deployment readiness. S+P combines SAR with variables available before the target inspection and excludes same-campaign C. Because SAR scenes were matched within a symmetric ±25-day window around the survey date, its reported MAE of 13.4 in/mi is a retrospective current-condition result. Full fusion provides an upper-information benchmark for data-rich PMS settings, while a prospective service would require one-sided acquisition-by-issue-date validation. The paper first situates this distinction within the pavement and remote-sensing literature, then describes the workflow and evaluation design, and finally examines agency implementation and further validation.

CC BY

2. BACKGROUND AND RELATED WORK

2.1 Pavement roughness and SAR-based observation

IRI summarizes the response of a standardized vehicle model to a measured longitudinal road profile. It is therefore a segment-level ride-quality measure, not a direct descriptor of individual cracks, surface color, or radar-scale roughness (E-08, 2015; Sayers et al., 1986). This distinction matters when comparing sensing technologies. Inertial profilers measure the profile required by the IRI definition; image- and backscatter-based approaches instead estimate IRI from related surface and contextual information. Smartphone sensing and mobile laser scanning provide additional measurement pathways, with different requirements for calibration, spatial coverage, and deployment (Chang et al., 2006; Kumar et al., 2015; Sattar et al., 2018).

SAR backscatter combines contributions from material properties, surface configuration, moisture, and surrounding objects. Figure 1 illustrates the conceptual relationship between surface condition and scattering. The relationship is not one-to-one: two segments with similar IRI can have different radar signatures, and similar signatures can arise from different pavement and environmental conditions. Vehicles and metallic roadside objects contribute further, as illustrated in Figure 2. Preprocessing improves observation comparability while retaining the segment-scale, mixed-scattering interpretation of moderate-resolution imagery (Karimzadeh & Matsuoka, 2020; Meyer, 2019; Meyer et al., 2020).

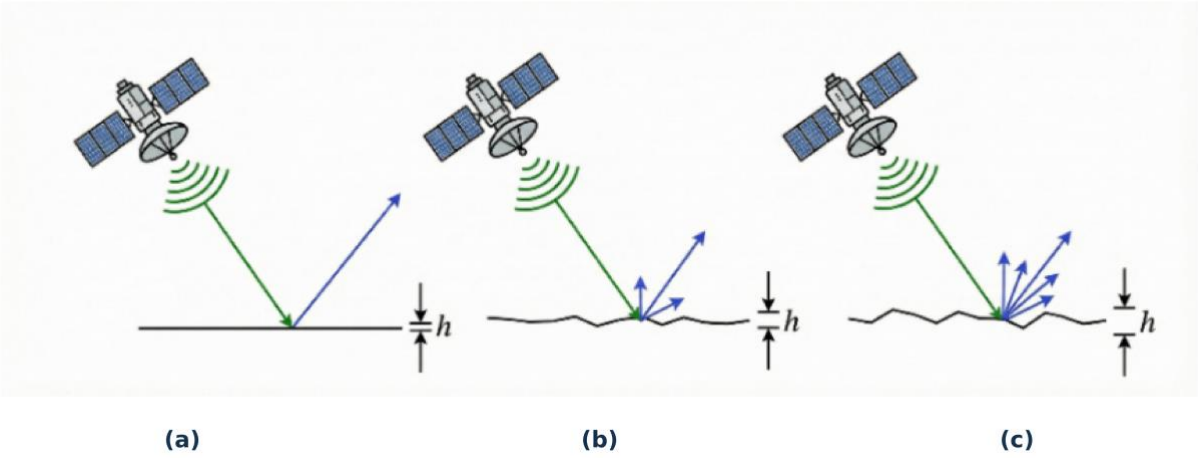


Figure 1: Conceptual SAR scattering for (a) smooth, (b) moderately rough, and (c) deteriorated pavement surfaces; adapted from (Meyer, 2019).

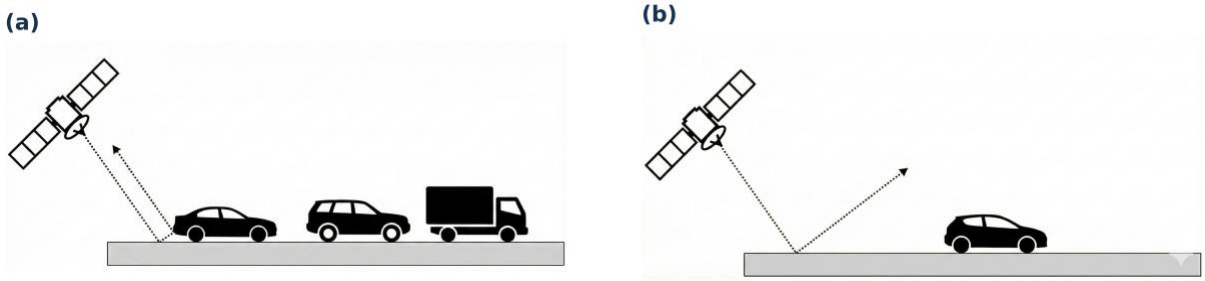


Figure 2: Roadway SAR backscatter with (a) multiple transient vehicles and (b) one transient vehicle, illustrating traffic-related signal variation; adapted from (Karimzadeh & Matsuoka, 2020).

2.2 SAR-based pavement monitoring and representation learning

Backscatter regression and interferometric monitoring address related but distinct engineering questions. Backscatter-based studies investigate associations between scattering signatures and surface condition. Interferometric approaches estimate displacement from phase observations and can identify deformation processes that affect road performance. A relationship between displacement and IRI is useful for management, but it does

CC BY

not make the two measured quantities interchangeable. Table 1 summarizes the methodological position of representative studies and the comparison used here.

Deep learning has enabled learned representations of SAR imagery to replace or complement hand-crafted texture descriptors. (Bashar & Torres-Machi, 2022) demonstrated pavement roughness estimation from SAR data, establishing an important precedent for this application. More recently, general-purpose Earth-observation models have expanded the range of available representations. AnySat supports heterogeneous resolutions and modalities, while SkySense learns multimodal remote-sensing representations (Astruc et al., 2025; Guo et al., 2024). Their availability motivates task-specific benchmarking, but using an adapted SAR branch is not equivalent to evaluating every capability of the complete pretrained system. Input information, channel adaptation, fine-tuning, and test partitions must remain explicit.

Accordingly, the present comparison has two levels. Published pavement studies provide cross-study context without implying a controlled performance ranking (Bashar & Torres-Machi, 2023; Bashar & Torres-Machi, 2022). Independently fitted SAR-only implementations of EfficientNet-B0, AnySat, and SkySense provide a within-study comparison on the same held-out cohort. This separation avoids attributing differences in geography, reference data, or preprocessing to the learning architecture alone.

Table 1: Positioning of the present study relative to representative SAR and Earth-observation research.

Study	Observation or method	Relevance to this study
(Suanpaga & Yoshikazu, 2010)	ALOS/PALSAR L-band SAR with IRI-based riding-quality modeling	Provides an early demonstration of relating spaceborne SAR backscatter to measured pavement riding quality.
(Meyer et al., 2020)	X-band SAR and pavement-condition assessment	Establishes the value and limitations of radar signatures for network-scale pavement assessment.
(Karimzadeh & Matsuoka, 2020)	X-band SAR, deformation, and traffic-related processing	Motivates separation of pavement information from transient scattering and deformation.
(Fiorentini et al., 2021)	PS-InSAR with machine learning and profilometric evaluation	Links road deformation information to assessment while retaining a distinction from profile measurement.
(Bashar & Torres-Machi, 2022)	Deep-learning IRI estimation from SAR	Provides the closest application precedent; published errors are cross-study context.
(Astruc et al., 2025); (Guo et al., 2024)	AnySat and SkySense Earth-observation representations	Provide contemporary SAR-capable model families for task-specific matched evaluation.
Present study	Availability-specific SAR-to-PMS workflow	Evaluates S, P, and C separately, with matched backbones, sensitivity analysis, and uncertainty-based screening.

2.3 Operational information and uncertainty

An accurate regression model becomes useful to an agency only when its inputs and outputs match the intended decision process. A between-survey estimate must be based on information available at the estimate date. A model that incorporates measurements from the next inspection campaign instead answers a different question: how well can multiple contemporaneous observations jointly explain the condition? Both questions are meaningful, but their performance should not be presented as interchangeable. Multimodal learning provides a flexible mechanism for combining the inputs; the availability experiment determines when that combination is operationally legitimate (Baltrušaitis et al., 2018; Ngiam et al., 2011a).

Uncertainty adds a second dimension to this interface. A condition estimate can support screening more effectively when it is accompanied by an interval or a ranking statistic that identifies difficult observations. MC dropout provides stochastic model predictions, while split conformal prediction calibrates residual-based intervals using held-out observations (Angelopoulos & Bates, 2022; Gal & Ghahramani, 2016). Their empirical behavior must be assessed rather than inferred from their names. This study therefore reports interval coverage and width separately from the ability to concentrate large errors in a small inspection subset.

CC BY

3. METHODOLOGY

3.1 Study workflow and validation framework

The workflow consists of acquisition, SAR preprocessing, segment-level feature preparation, model development, and PMS-oriented reporting (Figure 3). Sentinel-1 products are obtained through programmatic queries to the Alaska Satellite Facility, linked to pavement segments and survey dates, and processed in a scripted SNAP and Python environment. Regression outputs retain the segment reporting unit rather than being interpreted as independent pixel-level IRI measurements. Pavement condition and inventory records were obtained from the participating state agencies and supplemented with Long-Term Pavement Performance reference information.

The study contains 8,188 flexible-asphalt pavement segments in Alaska, Utah, and West Virginia (Table 2). These regions represent sub-arctic, arid continental, and humid temperate settings within the study. The overall IRI distribution has a mean of 118 in/mi, a standard deviation of 39 in/mi, and a range of 42 to 285 in/mi. Good, Fair, and Poor IRI categories account for 34.2%, 52.1%, and 13.7% of the full cohort, respectively. The reporting unit is approximately 0.1 mile (161 m), with segment counts representing observations in the analysis cohort rather than unique centerline mileage.

A corridor-stratified partition reserves 1,228 segments for final testing. No corridor is shared between the development and test subsets. The remaining 6,960 observations support five-fold GroupKFold development, with all observations from one corridor assigned to the same fold (Kapoor & Narayanan, 2023). Training subsets contain approximately 5,568 observations, and validation subsets contain approximately 1,392. Feature normalization and model fitting use only the training portion of each fold. The three states are represented in both development and testing, so the design evaluates new corridors within the represented domain rather than using a leave-one-state-out transfer.

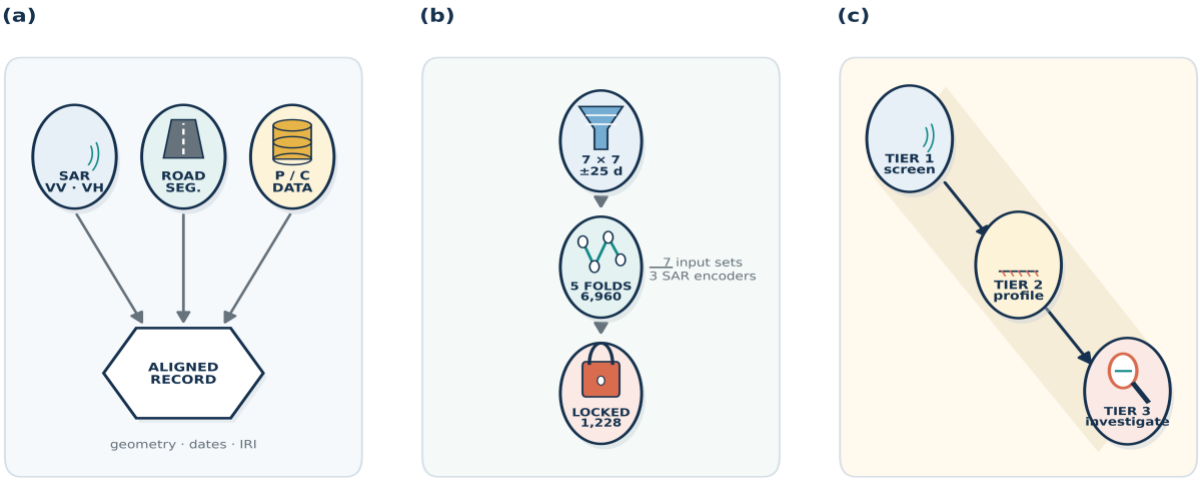


Figure 3: Satellite-to-PMS workflow: (a) acquisition and evidence assembly; (b) corridor-grouped development and held-out evaluation; (c) PMS screening and escalation.

Table 2: Study cohort and corridor-held-out evaluation partition.

State	Full cohort	Development	Test	Represented setting
Alaska	2,763	2,349	414	Sub-arctic
Utah	2,706	2,300	406	Arid continental
West Virginia	2,719	2,311	408	Humid temperate
Total	8,188	6,960	1,228	Flexible-asphalt pavements

Corridors, rather than individual adjacent segments, define the development/test grouping. Counts refer to segment observations in the analysis cohort.

CC BY

3.2 SAR acquisition, normalization, and spatial support

3.2.1 Acquisition and preprocessing

The input archive comprises 1,823 Sentinel-1A/B Level-1 Interferometric Wide-swath high-resolution Ground Range Detected products. The study uses ascending acquisitions with VV and VH polarization. Sentinel-1 operates at approximately 5.405 GHz, corresponding to a wavelength of 5.6 cm, with a nominal IW swath width of 250 km (Torres et al., 2012). The processed grid has 10 m pixel spacing. Pixel spacing specifies sampling on the output grid and must not be equated with independent spatial resolution or an ability to resolve individual lane defects.

Processing uses SNAP v9.0 and its Python interface. Precise orbit information and noise corrections precede radiometric normalization, terrain flattening, speckle filtering (Lee et al., 1994), and Range-Doppler terrain correction. The terrain-normalized backscatter is represented as gamma-naught (γ^0), with VV and VH processed consistently. Terrain correction resamples the observations to state-appropriate UTM grids at 10 m spacing. The elevation workflow uses the SRTM 1-arc-second v3.0 product (Farr et al., 2007) and regional fallback elevation tiles for Alaska. DEM-based masks identify terrain-related exclusions. The processing sequence follows the principles described by (Filipponi, 2019) and (Small, 2011).

A Lee Sigma filter with a 7 × 7 window reduces speckle while retaining spatial variation relevant to the segment representation (Lee et al., 2009). The development experiment varies the filter window to assess the balance between noise suppression and prediction error. A separate fixed-model test perturbation examines how changes in preprocessing affect a model trained with the reference filter. These two experiments are distinguished because refitting a model for each filter and applying a new filter to a fixed model answer different questions.

The geometric selection excludes elevated structures, segments within 50 m of intersections, and road gradients exceeding 8%. Layover and shadow masks further restrict usable observations. These operations reduce problematic mixtures but do not isolate a pure pavement return. In particular, a centerline buffer cannot remove all scattering from shoulders, adjacent land cover, or guardrails. Figure 4 illustrates the processing stages; the native acquisition display and map-projected output have different display geometries and are not intended to be interpreted as a registered before-and-after spatial comparison.

3.2.2 Temporal and environmental screening

SAR acquisitions are linked to IRI survey dates using a symmetric ±25-day window. Temporal minimum compositing suppresses transient bright contributions across available acquisitions, followed by a residual anomaly screen that replaces observations beyond ±3 temporal standard deviations with a temporal median. The objective is to stabilize the radar signature in the presence of traffic and environmental variation, not to identify every vehicle independently. Optical reference imagery provides roadway context rather than synchronous ground truth for moving traffic.

Acquisitions associated with precipitation exceeding 5 mm during the preceding 24 hours are excluded using NOAA Stage IV for Utah and West Virginia and the North American Regional Reanalysis for Alaska (Lin & Mitchell, 2005; Mesinger et al., 2006). Observations affected by snow, ice, or standing water are outside the intended dry-surface interpretation. Major maintenance or rehabilitation events within the SAR-to-survey window are excluded using agency treatment records, since these events can change condition abruptly rather than through gradual deterioration.

The symmetric ±25-day matching window centers SAR observations on the inspection date and defines a retrospective assessment of current conditions. A prospective issue-date estimate would use only acquisitions available by that date and would require independent validation of the resulting temporal support. Window-length sensitivity is estimated on a common development cohort, preserving the final test set exclusively for evaluation.

3.2.3 SAR patch generation and feature extraction

Each approximately 161 m segment is represented by a centerline-aligned image subset containing about 16 × 3 samples on the 10 m grid. Bicubic interpolation maps this elongated subset to a 64 × 64 tensor (Keys, 1981). Consequently, the number of interpolated samples increases more strongly across the road than along it. This anisotropic resampling standardizes the neural-network input shape; it does not extrapolate beyond the patch or recover sub-meter pavement detail. The geospatial segment remains the prediction support before and after interpolation.

CC BY

The three image channels are VV γ^0, VH γ^0, and normalized local incidence angle. Geometric context helps distinguish acquisition-related variation from condition-related associations. SAR-derived descriptors also include polarimetric quantities and 30 gray-level co-occurrence matrix (GLCM) texture measures. The texture inventory consists of five measures, Contrast, Dissimilarity, Homogeneity, Energy, and Correlation, computed at three window sizes and for two polarizations, with directional offsets averaged (Haralick et al., 1973). These are satellite-derived quantities, not contemporaneous pavement-inspection measurements. Input groups are distinguished by their physical information source rather than by whether they enter a learned image representation or a derived-feature representation.

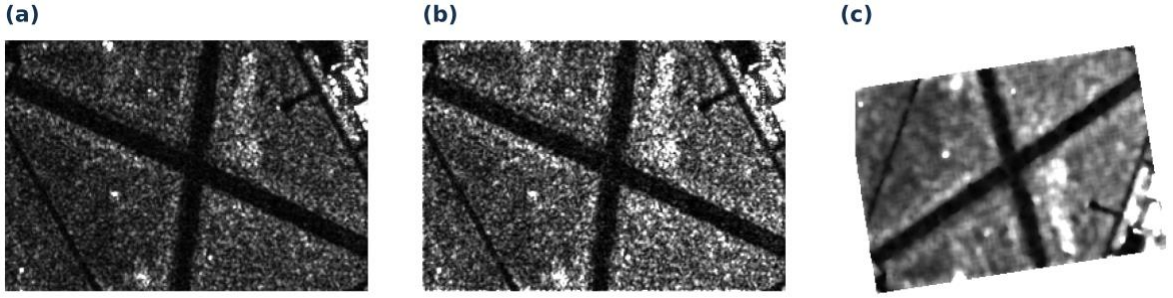


Figure 4: SAR preparation: (a) native GRD display; (b) corrected, terrain-normalized backscatter; (c) Lee Sigma-filtered, map-projected output. The views are not pixel-registered.

3.3 Input configuration and model architecture

3.3.1 Input groups and information availability

Table 3 defines the three information groups used in the ablation experiment. S contains the satellite observations and SAR-derived descriptors. P contains nine inventory, loading, mixture, and exposure variables available before the target inspection. C contains rut depth, mean profile depth (MPD), and fatigue cracking measured during the same inspection campaign as IRI. Availability is defined relative to the prediction date, not simply by whether a variable is stored in an agency database.

Seven independently fitted configurations are evaluated: P, C, P+C, S, S+P, S+C, and S+P+C. The primary retrospective comparison is P versus S versus S+P because all three exclude the same-campaign C. Configurations containing C quantify the upper-information benefit available when contemporaneous condition indicators have already been collected. Each configuration is fitted to its specified inputs, so the full-fusion performance applies specifically to S+P+C.

Auxiliary records are linked using route identifiers, spatial references, survey dates, and treatment information. Numerical features are standardized using statistics from the training partition, and the fitted transformation is then applied to the corresponding validation or test data. In prospective use, an inventory value must genuinely predate the estimate. The existence of a historical field in a current database is not, by itself, evidence that the value was available at an earlier decision date.

Table 3: Input groups and their availability relative to the target IRI survey.

Group	Variables	Availability and interpretation
S: satellite	VV and VH backscatter, local incidence angle, and SAR-derived polarimetric/texture information	Remote observations processed for the target segment and time window; no same-campaign pavement distress measurement.
P: pre-inspection (9)	Pavement age; surface and base thickness; AADTT; cumulative ESAL; binder content; air voids; VFA; freeze-thaw cycles	Inventory, loading, mixture, and exposure information are available before the target inspection.
C: contemporaneous (3)	Rut depth; mean profile depth; fatigue cracking	Condition indicators from the same inspection campaign as IRI. Their use defines an inspection-dependent configuration.

CC BY

AADTT: annual average daily truck traffic. ESAL: equivalent single-axle loads. VFA: voids filled with asphalt. The P and C groups contain 12 engineering variables in total.

3.3.2 EfficientNet-B0 fusion architecture

The reference SAR encoder uses ImageNet-pretrained EfficientNet-B0 (Tan & Le, 2019). Early convolutional blocks 1 through 6 are frozen during the initial training phase, while higher-level features and the regression layers are fine-tuned. Global average pooling produces a 1,280-dimensional SAR embedding. In the full-fusion configuration, the engineering branch produces a 64-dimensional representation; concatenation yields 1,344 features, followed by fully connected layers with 256 and 128 neurons and a linear IRI output (Figure 5). The reference backbone contains approximately 5.3 million parameters, with approximately 1.3 million trainable parameters in the benchmark implementation.

EfficientNet-B0 supplies a compact reference implementation for the task. Late fusion permits satellite and engineering representations to be learned separately before joint regression (Ngiam et al., 2011b). The ablation experiment determines the contribution of each information group, while the architecture experiment tests whether more recent pretrained representations change SAR-only accuracy. These questions are complementary and cannot be answered by comparing a full-information model with a satellite-only competitor.

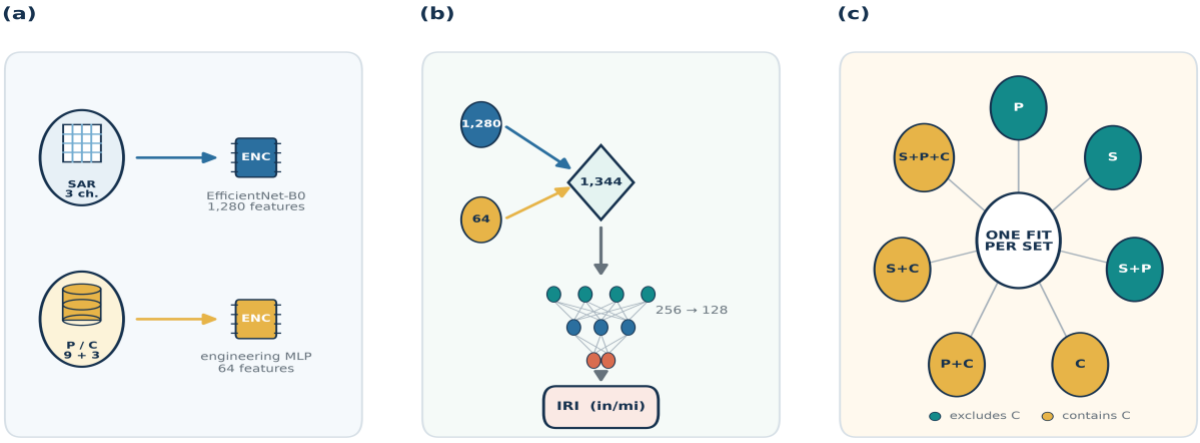


Figure 5: Model design: (a) information-specific encoders, (b) late fusion and regression, and (c) availability-specific fits.

3.3.3 Model optimization and sensitivity analysis

Training minimizes Huber loss with $\delta = 1.0$ in/mi. The Adam optimizer uses $\beta_1 = 0.9$, $\beta_2 = 0.999$, an initial learning rate of 10^{-4}, and a ReduceLROnPlateau schedule (Kingma & Ba, 2014). The batch size is 32, the dropout probability is 0.3 (Srivastava et al., 2014), and the L2 regularization coefficient is 10^{-4}. Within-fold validation loss guides early stopping with a patience of 15 epochs. Table 4 summarizes the reference settings. Development and test data remain separated throughout normalization, tuning, and model selection.

Sensitivity analysis varies one factor at a time while holding the other reference settings constant. The factors are the temporal window length, Lee Sigma window size, initial learning rate, Huber threshold, dropout probability, and batch size. Each comparison uses five corridor-grouped development folds. The temporal-window analysis uses a common cohort of approximately 5,639 observations with at least one valid acquisition in the most restrictive ±10-day window, so changing the window length does not simultaneously change the evaluation population. Mean validation MAE and its between-fold standard deviation summarize each setting. This local sensitivity design characterizes the neighborhood of the chosen configuration; it is not an exhaustive search over all parameter interactions.

CC BY

Table 4: Reference training and model-development settings.

Component	Setting
Image encoder	ImageNet-pretrained EfficientNet-B0; early blocks 1–6 initially frozen
Fusion dimensions	1,280-dimensional SAR embedding + 64-dimensional engineering embedding; regression head 256 → 128 → 1
Loss	Huber, $\delta = 1.0$ in/mi
Optimizer	Adam; $\beta_1 = 0.9$, $\beta_2 = 0.999$; initial learning rate 10^{-4}
Regularization	Dropout 0.3; L2 coefficient 10^{-4}
Training control	Batch size 32; ReduceLROnPlateau; early-stopping patience 15 epochs
Development design	Five-fold corridor-grouped cross-validation on 6,960 observations
Final evaluation	1,228 corridor-held-out test observations
Architecture-comparison seeds	42, 123, and 456

The appendix reports fold results, the complete sensitivity tables, and the per-seed architecture results. Parameter selection uses development data, not the final test cohort.

3.4 Comparative experiments and evaluation metrics

The architecture comparison evaluates task-specific adaptations of AnySat and SkySense, alongside EfficientNet-B0, on the same locked cohort of 1,228 segments. The cohort, prediction target, and SAR-only task are matched, while each architecture retains its documented channel interface: EfficientNet-B0 receives VV γ°, VH γ°, and local incidence angle (LIA); the AnySat ascending Sentinel-1 branch receives VV γ° and VH γ° with LIA as the third channel; and the SkySense SAR encoder receives two-channel VV γ° and VH γ° input. Optical inputs and contemporaneous engineering condition indicators are excluded. These results characterize the study-specific task adaptations rather than the complete multimodal foundation models (Astruc et al., 2025; Guo et al., 2024). Seeds 42, 123, and 456 quantify fitting variation for every architecture. Checkpoint sources, channel adaptations, normalization, frozen and trainable components, parameter counts, optimization settings, and deviations from the original implementations are reported in the appendix, Table A4.

Model error is summarized by mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE), median absolute error, and the 90th percentile of absolute error. For observation i, y_i is the measured IRI and $\hat{y}_i$ is the predicted IRI. The correlation statistic is the squared Pearson correlation, r^2. It is not interpreted as the residual-based coefficient of determination, $1 - \Sigma(y_i - \hat{y}_i)^2/\Sigma(y_i - \bar{y})^2$, because those quantities need not be equal for held-out predictions.

$$MAE = \left(\frac{1}{n}\right)\sum_{i=1}^{n}|y_i - \hat{y}_i|\,; \quad RMSE = \sqrt{\left(\frac{1}{n}\right)\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{1}$$

$$r^2 = corr(y, \hat{y})^2; \quad \Delta MAE = MAE_{comparator} - MAE_{EfficientNet-B0} \tag{2}$$

The architecture table reports the mean and standard deviation across seeds, as well as the bootstrap confidence interval for ΔMAE. Negative ΔMAE favors the comparator. The input-availability table reports the independently fitted configuration results without assigning significance to differences for which paired confidence intervals are not reported. Published results from other studies remain contextual because their target populations and observations differ.

IRI estimates are also mapped to three condition categories: Good for IRI < 95 in/mi, Fair for 95–170 in/mi (inclusive), and Poor for IRI > 170 in/mi. These are IRI-based categories, not a complete federal pavement-condition rating incorporating every distress criterion. Classification reporting includes the count matrix, overall accuracy, class-specific precision and recall, and macro-F1. Spatial residual structure is summarized by Moran's I under the study's 5 km neighborhood definition (Moran, 1950). This diagnostic complements the corridor-based validation design by assessing residual spatial structure within the represented study domain.

CC BY

3.5 Prediction intervals and inspection screening

Uncertainty analysis was performed on a separately fitted full-fusion S+P+C point predictor using the EfficientNet-B0 fusion architecture described in Section 3.3.2 and the optimization procedure described in Section 3.3.3. The overall development pool contained 6,960 observations. Before fitting this uncertainty predictor, 1,188 observations were reserved by corridor-level stratified hold-out exclusively for split-conformal calibration, leaving 5,772 observations for point-model fitting. MC-dropout and split-conformal intervals were evaluated on the same locked 1,228-segment corridor-held-out test cohort. MC-dropout inference used 50 stochastic forward passes with dropout probability 0.3 active at the same layer locations used during training (Gal & Ghahramani, 2016). The calibration and test cohorts were non-overlapping, and neither contributed observations to fitting the uncertainty predictor. For nominal coverage $1 - \alpha$, let L_i and U_i denote the lower and upper prediction bounds. The prediction interval coverage probability (PICP) and the mean prediction interval width (MPIW) are defined as follows.

$$PICP = \left(\frac{1}{n}\right)\sum_{i=1}^{n} 1\{L_i \le y_i \le U_i\}; \quad MPIW = \left(\frac{1}{n}\right)\sum_{i=1}^{n}(U_i - L_i) \tag{3}$$

A residual-based split-conformal comparator calibrates absolute residuals on the 1,188 reserved calibration observations. At nominal 90% coverage, the finite-sample order statistic is $k = \lceil (1{,}188 + 1) \times 0.90 \rceil = 1{,}071$; the corresponding residual quantile defines a fixed-width interval around the fitted point prediction. The theoretical marginal-coverage interpretation requires the relevant exchangeability conditions, so dependence among road segments and deployment shifts remains important beyond the evaluated cohort (Angelopoulos & Bates, 2022).

$$C_\alpha(x) = [f(x) - \hat{q}, f(x) + \hat{q}]; \quad \hat{q} = r_k; \quad r_i = |y_i - f(x_i)| \tag{4}$$

Signed coverage error is reported in percentage points as empirical minus nominal coverage. The interval score combines width with a penalty for observations outside the interval; lower values are preferable at a fixed nominal level. Coverage and width are kept separate from uncertainty-based ranking because an interval can have adequate marginal coverage while providing limited discrimination between easy and difficult segments.

$$IS_\alpha = (U - L) + \left(\frac{2}{\alpha}\right)(L - y)1\{y < L\} + \left(\frac{2}{\alpha}\right)(y - U)1\{y > U\} \tag{5}$$

For the full-fusion S+P+C uncertainty predictor, a segment is flagged when its 90% MC-dropout prediction interval exceeds 40 in/mi, and a large error is defined as an absolute prediction error greater than 25 in/mi. The operating point reports both the proportion of segments flagged and the proportion of large errors captured. Random selection at the same flagged fraction provides the reference expectation. These quantities measure error concentration under the stated screening rule and inspection fraction.

4. RESULTS

4.1 Preprocessing and development sensitivity

The processing workflow generated the segment representations used in development and testing from the 1,823-product SAR archive. The selected Lee Sigma configuration achieved an equivalent number of looks of 12.4 and an edge-preservation index of 0.89 in the image-quality assessment. These measures describe the processed imagery; downstream prediction error provides the task-specific criterion for assessing whether stronger smoothing remains useful.

Temporal-window sensitivity shows a diminishing improvement as the common-cohort window widens (Figure 6). Mean development MAE decreased from 14.2 ± 0.5 in/mi at ±10 days to 13.7 ± 0.4 in/mi at ±15 days, 13.4 ± 0.4 in/mi at ±25 days, and 13.3 ± 0.4 in/mi at ±35 days. Thus, widening the window beyond the selected ±25-day setting changed the mean error by only 0.1 in/mi while extending temporal support. The selected window reflects this practical compromise, rather than a claim that it attains the smallest numerical error in every comparison.

The Lee Sigma development comparison also exhibits a plateau (Figure 7). Unfiltered inputs yielded MAE 15.8 ± 0.6 in/mi; 3 × 3 and 5 × 5 windows yielded 15.4 ± 0.5 and 15.0 ± 0.4, respectively. The 7 × 7 setting yielded 13.2 ± 0.4, compared with 13.3 ± 0.5 for 9 × 9. The largest filter improved the equivalent number of looks but did not improve the mean IRI error. This result supports evaluating speckle suppression through both image quality and the engineering target rather than assuming that a smoother image is a better predictor.

CC BY

The learning rate, Huber threshold, dropout, and batch size sweeps are reported in the appendix. The reference learning rate of 10^{-4}, Huber threshold of 1.0 in/mi, dropout of 0.3, and batch size of 32 each attained a mean development MAE of 13.2 in/mi in their respective one-factor comparisons. Nearby settings often produced similar errors, whereas the extreme learning rates and the absence or excess of dropout increased mean error. These observations characterize local robustness around the reference configuration; joint parameter interactions remain a separate optimization question.

The fixed-model filter perturbation yields a different, operationally relevant result. With the SAR-only model held fixed, test MAE was 14.8 in/mi under the reference 7 × 7 preprocessing, 14.9 for 9 × 9, 15.0 for 5 × 5, 15.4 for 3 × 3, and 16.3 without filtering. The model therefore remained close to its reference error under adjacent window sizes but degraded when the input processing changed more substantially. Because the model was not refitted for each filter, these results describe preprocessing compatibility rather than the best achievable performance of each filter family.

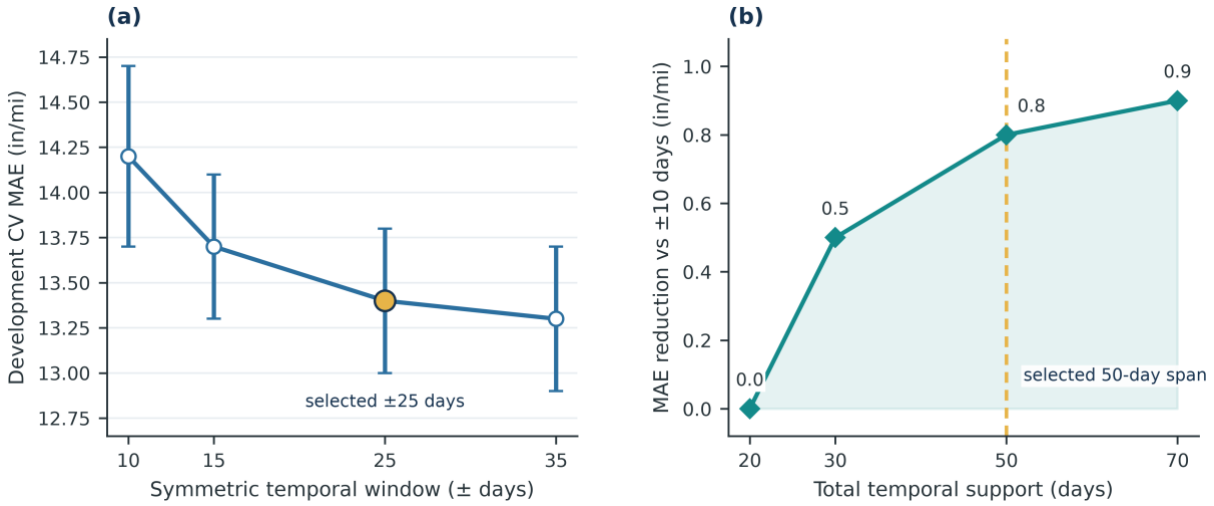


Figure 6: Temporal-window sensitivity on the common development cohort (n ≈ 5,639): (a) mean validation MAE with between-fold SD; (b) improvement relative to ±10 days. The selected window is ±25 days.

4.2 What SAR contributes when current distress measurements are unavailable

Table 5 and Figure 8 separate performance by information set. Pre-inspection variables alone achieved MAE 16.1 in/mi, while SAR alone achieved 14.8 in/mi. Combining them reduced MAE to 13.4 in/mi and RMSE to 18.1 in/mi and increased squared Pearson correlation to 0.62. Relative to P alone, S+P reduced mean absolute error by 2.7 in/mi (16.8%); relative to S alone, adding pre-inspection context reduced it by 1.4 in/mi (9.5%). Because P, S, and S+P exclude C, these results isolate the value of SAR and existing engineering context without same-campaign distress measurements.

Full fusion achieved an MAE of 12.9 in/mi, an RMSE of 17.2 in/mi, and an r^2 of 0.69. Adding C to S+P reduced MAE by a further 0.5 in/mi, or 3.7%, establishing the incremental value of contemporaneous indicators in a data-rich setting. The P+C model achieved MAE 15.7 in/mi; adding SAR reduced that error by 2.8 in/mi. SAR therefore contributed predictive information both with and without C, while full fusion served as the upper-information benchmark.

The C-only model yielded an MAE of 17.8 in/mi when predicting IRI from rut depth, MPD, and fatigue cracking. These indicators and longitudinal ride quality quantify related but distinct aspects of pavement condition. The comparison, therefore, concerns the predictive information contained in alternative input sets; the field profiler remains the source of the reference IRI.

Across the five grouped development folds, full fusion achieved an MAE of 13.2 ± 0.4 in/mi, an RMSE of 17.8 ± 0.5 in/mi, and an r^2 of 0.68 ± 0.02. Its held-out performance, MAE 12.9 in/mi, RMSE 17.2 in/mi, and $r^2 = 0.69$, was consistent with those development summaries, indicating stable error across the evaluated partitions. Fold-level results and additional distributional error measures are reported in the appendix.

CC BY

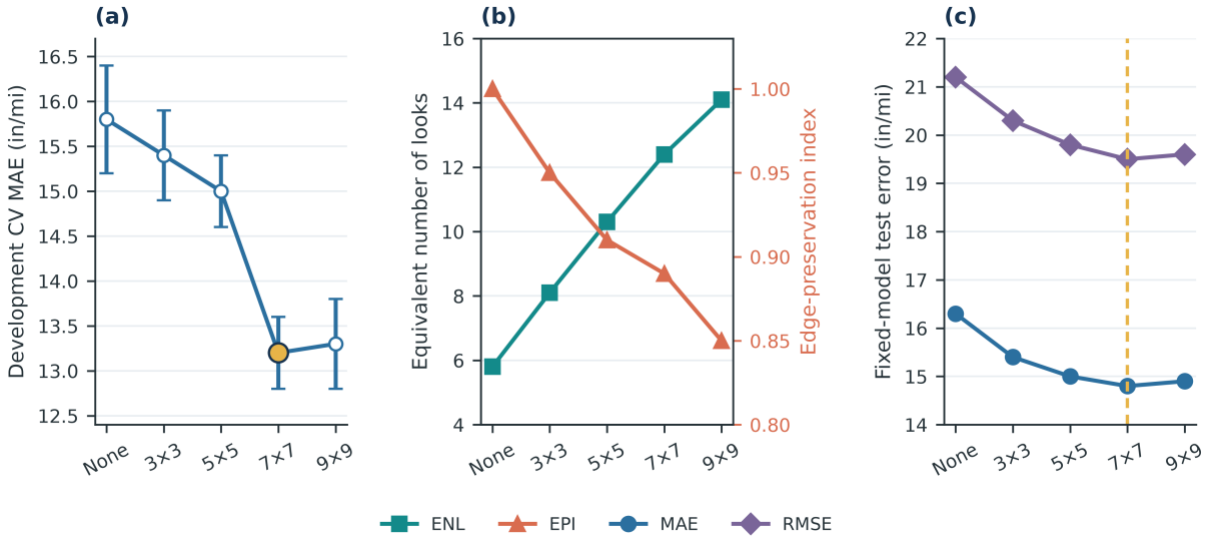


Figure 7: Lee Sigma filtering: (a) development MAE; (b) ENL and EPI; (c) fixed-model test perturbation. The 7 × 7 setting is the training/reference filter.

Table 5: Input-availability experiment on the same 1,228 corridor-held-out test segments.

Configuration	MAE (in/mi)	RMSE (in/mi)	r^2	MAPE (%)	Median AE (in/mi)	90th-percentile AE (in/mi)
P	16.1	21.0	0.43	19.8	13.8	31.6
C	17.8	22.8	0.35	21.5	15.1	34.6
P+C	15.7	20.3	0.50	18.9	13.2	30.4
S	14.8	19.5	0.55	17.1	12.1	28.7
S+P	13.4	18.1	0.62	15.4	10.8	25.8
S+C	13.8	18.5	0.59	16.2	11.4	27.3
S+P+C	12.9	17.2	0.69	14.2	10.4	24.3

S: SAR; P: pre-inspection information; C: contemporaneous condition indicators. AE: absolute error. r^2 is the squared Pearson correlation. P, S, and S+P do not use the same campaign distress measurements; all configurations are independently fitted.

4.3 Comparison with contemporary SAR models

AnySat achieved the lowest mean SAR-only MAE among the three evaluated implementations (Table 6). Its three-seed mean MAE was 14.2 ± 0.2 in/mi, compared with 14.8 ± 0.2 for EfficientNet-B0. The ΔMAE of −0.6 in/mi had a bootstrap 95% confidence interval of −1.1 to −0.1 in/mi. SkySense achieved 14.5 ± 0.2 in/mi, with ΔMAE −0.3 and a confidence interval of −0.9 to +0.3 in/mi (Figure 9). The SkySense contrast, therefore, does not isolate a clear direction of advantage at that interval level.

These results qualify the role of the reference backbone. EfficientNet-B0 is a practical implementation within the workflow, not the most accurate SAR-only architecture in this comparison. AnySat's modest improvement indicates that stronger pretrained representations can enhance the satellite-only branch. Separating input availability and quantifying the benefit of pre-inspection information remain necessary across encoders, so improved representations can be incorporated without changing the operational distinction between S, P, and C. Runtime or cost superiority is not inferred from backbone parameter counts, because matched computational measurements for all three adaptations are not reported.

CC BY

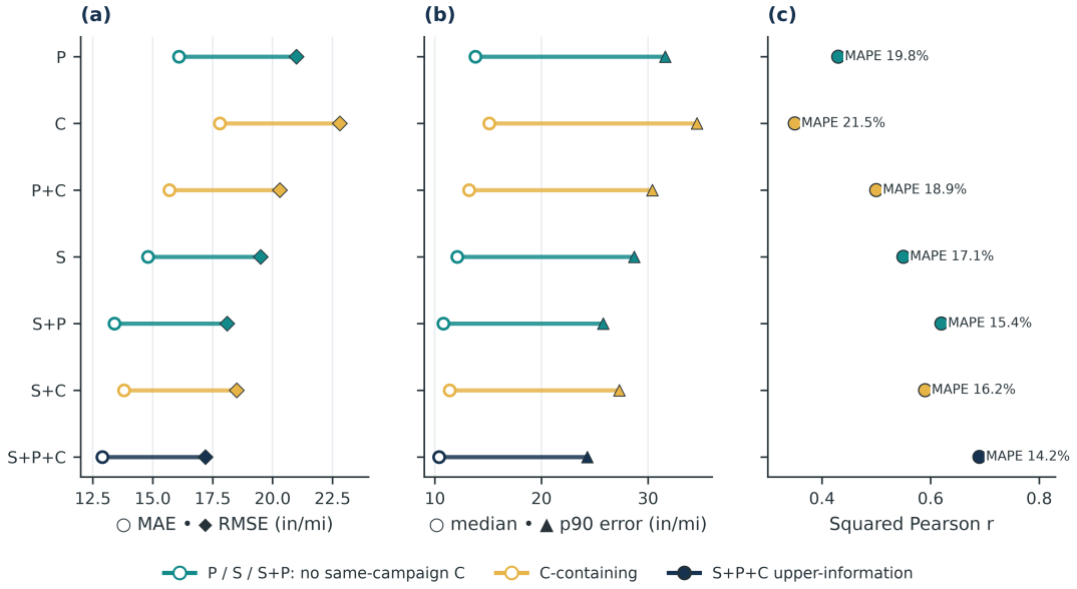


Figure 8: Held-out input-availability results: (a) MAE and RMSE; (b) median and 90th-percentile absolute error; (c) squared Pearson correlation and MAPE.

The published SAR-based pavement result of (Bashar & Torres-Machi, 2022), MAE approximately 14.6 in/mi, provides useful application context but is not a paired baseline. Its data, setting, and implementation differ. Accordingly, percentage improvements relative to that published number are not interpreted as the effect of the proposed methodology.

Table 6: Contemporary SAR-only architecture comparison on 1,228 test segments.

Task implementation	MAE (in/mi)	RMSE (in/mi)	r^2	ΔMAE (in/mi)	Bootstrap 95% CI for ΔMAE
EfficientNet-B0	14.8 ± 0.2	19.5 ± 0.2	0.55 ± 0.01	Reference	NA
AnySat adaptation	14.2 ± 0.2	18.8 ± 0.2	0.57 ± 0.01	−0.6	[−1.1, −0.1]
SkySense adaptation	14.5 ± 0.2	19.2 ± 0.2	0.56 ± 0.01	−0.3	[−0.9, +0.3]

Mean ± SD is across seeds 42, 123, and 456. ΔMAE is the comparator minus EfficientNet-B0. The benchmark concerns the study-specific SAR adaptations, not the complete multimodal foundation models. Per-seed results appear in the appendix.

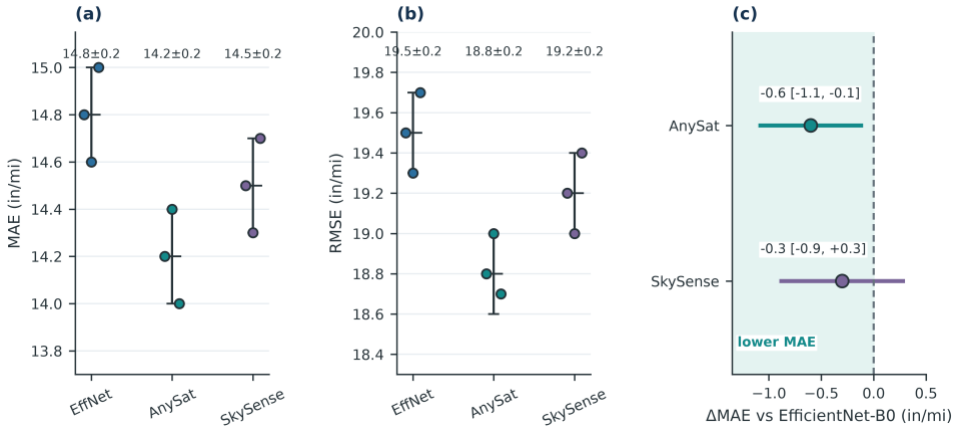


Figure 9: Matched SAR-only comparison: (a) seed-level MAE; (b) seed-level RMSE; (c) comparator-minus-EfficientNet-B0 ΔMAE with bootstrap 95% confidence intervals.

CC BY

4.4 Geographic behavior within the represented domain

The full-fusion model achieved state-specific MAEs of 13.4 in/mi in Alaska, 12.2 in/mi in Utah, and 13.1 in/mi in West Virginia, with corresponding r^2 values of 0.66, 0.72, and 0.68 (Figure 10). Weighting the state errors by their test counts reproduces the pooled full-fusion MAE to the reported precision. The cross-state range is 1.2 in/mi, indicating that the pooled result is not solely attributable to one represented state.

Residual spatial autocorrelation was small under the specified neighborhood definition, with Moran's I = 0.04 and p = 0.31. This is compatible with limited detectable residual clustering at that scale. It is not a validation of a national deployment model, nor does a nonsignificant result establish independence at every spatial scale. The corridor partition is the design element that separates development from test locations; the residual diagnostic is a complementary description of the resulting errors. Figure 11 provides spatial examples of the segment-level prediction products.

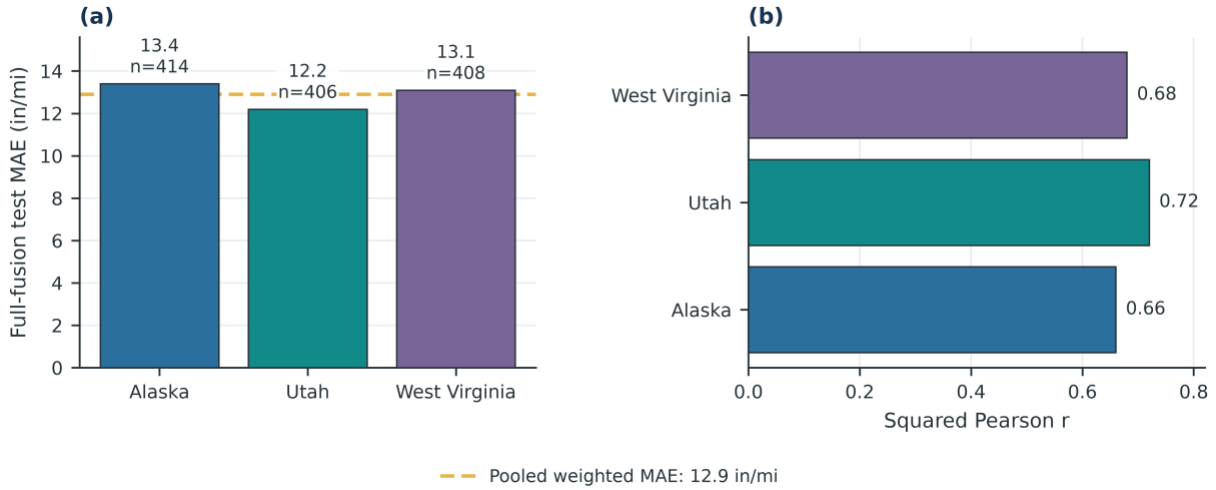


Figure 10: Full-fusion performance by represented state: (a) held-out MAE and test counts; (b) squared Pearson correlation.

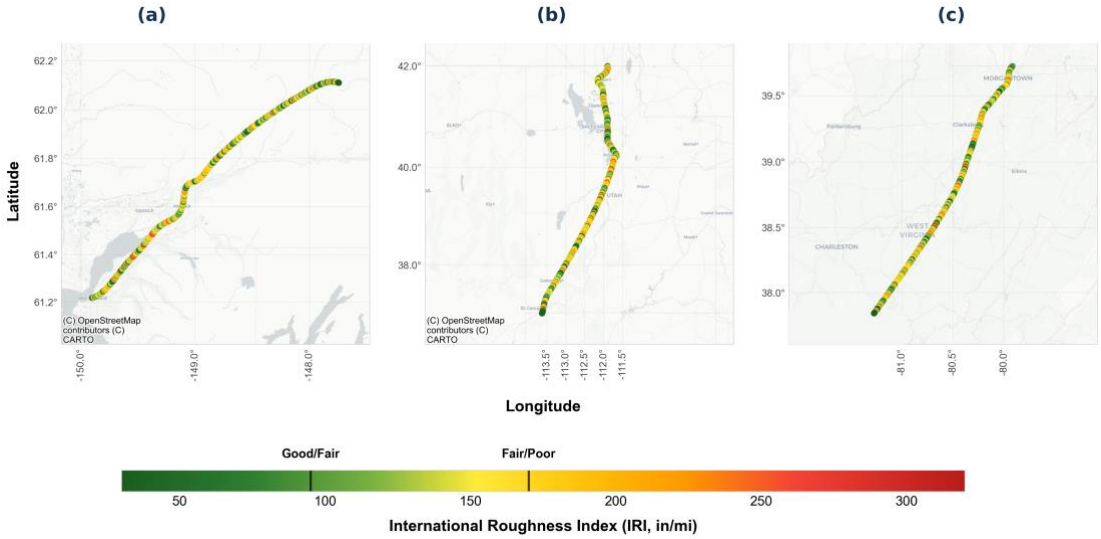


Figure 11: Predicted segment-level IRI in (a) Alaska, (b) Utah, and (c) West Virginia. Basemap: © OpenStreetMap contributors; CARTO.

4.4.1 Geometric error diagnostics

Potential sensitivity to acquisition geometry was evaluated on the locked 1,228-segment test cohort. Road orientation was defined relative to the Sentinel-1 ground-look azimuth, and the acute road-axis-to-look angle, θ,

CC BY

was folded to 0°–90°. Orientation sensitivity was evaluated with the SAR-only S configuration because the hypothesized mechanism concerns the radar observation itself. MAE was 15.2 in/mi for θ = 0°–30° (n = 312), 14.5 in/mi for 30°–60° (n = 498), and 14.9 in/mi for 60°–90° (n = 418). Bootstrap 95% confidence intervals for MAE were [14.0, 16.4], [13.5, 15.4], and [13.9, 15.9] in/mi, respectively. The intervals overlapped, and the continuous orientation–absolute-error association was small (Spearman $\rho = -0.04$, $p = 0.18$). These diagnostics provide no evidence of pronounced systematic orientation dependence within the represented acquisition geometry.

Road-width sensitivity was evaluated separately for the full-fusion S+P+C model. MAE was 14.6 in/mi for paved carriageway widths below 8 m (n = 287), 12.6 in/mi for widths of 8–12 m (n = 510), and 12.2 in/mi for widths above 12 m (n = 431), compared with 12.9 in/mi overall. Corresponding RMSE values were 19.1, 16.8, and 16.4 in/mi. Absolute error showed a weak negative association with width (Spearman $\rho = -0.11$, $p = 0.06$). The direction is compatible with a larger pavement contribution on wider carriageways, but remains a weak observational association.

A joint SAR-only geometric-error regression used W = paved carriageway width/20 m as a dimensionless scaling of width; 20 m is a normalization denominator, not a physical estimate of cross-road support or pure-pavement fraction. The θ coefficient was −0.021 in/mi per degree (SE = 0.020, p = 0.31), the W coefficient was −3.2 in/mi per unit W (SE = 1.8, p = 0.09), and the θ × W coefficient was 0.009 in/mi per degree per unit W (SE = 0.012, p = 0.44); with state fixed effects, the model R^2 was 0.04. A separate full-fusion width model yielded −5.7 in/mi per unit W (SE = 3.2, p = 0.07; model $R^2 = 0.09$). Thus, no strong independent geometric association was detected after adjustment for state. Complete stratified metrics and regression coefficients are reported in the appendix, Tables A17–A19, and Figure A6.

4.5 IRI-based condition classification

Full-fusion predictions assigned the correct three-class IRI category to 1,084 of the 1,228 test segments, giving 88.3% accuracy. The confusion matrix is shown in Figure 12. Errors were confined to neighboring categories in this cohort, with no observed Good-to-Poor or Poor-to-Good assignments. However, adjacent errors remain operationally meaningful, particularly where a Poor segment is predicted as Fair. Classification should therefore be evaluated alongside continuous error and the intended escalation rule.

Of the 170 truly Poor segments, 146 were identified as Poor, corresponding to a recall of 85.9%. Of the 191 segments predicted Poor, 146 were truly Poor, corresponding to 76.4% precision. These quantities describe different inspection consequences: recall reflects missed Poor segments, while precision reflects how many Poor flags are confirmed by the reference category. Macro-F1 was 86.8% (Table 7). The classification result applies to the full-fusion configuration and is not assigned to the S-only or S+P model without a separate count matrix.

Table 7: Full-fusion three-class IRI classification metrics.

Observed class	Support	Precision (%)	Recall (%)	F1 (%)
Good	419	91.5	90.5	91.0
Fair	639	89.7	87.5	88.6
Poor	170	76.4	85.9	80.9
Overall / macro	1,228	Overall accuracy: 88.3	Balanced accuracy: 87.9	Macro-F1: 86.8

Calculated from the count matrix in Figure 12. Metrics are rounded to one decimal place. The categories use IRI thresholds only.

4.6 Interval calibration and uncertainty-guided screening

For the full-fusion S+P+C predictor, both uncertainty methods attained empirical coverage close to the nominal 90% level (Table 8). MC dropout achieved PICP 89.6%, mean interval width 32.8 in/mi, median width 31.4 in/mi, and interval score 34.1. The fixed-width split-conformal comparator achieved PICP 90.2%, mean and median widths of 56.8 in/mi, and an interval score of 58.0. At approximately comparable marginal coverage, MC dropout produced narrower intervals and a lower interval score in this evaluation.

MC-dropout coverage tracked nominal coverage across the five reported settings (Figure 13). Empirical coverage was 49.8%, 68.9%, 78.6%, 89.6%, and 93.8% at nominal levels of 50%, 70%, 80%, 90%, and 95%, respectively. The largest absolute signed coverage difference was 1.4 percentage points. These summaries establish aggregate interval calibration across the reported levels; state- and condition-specific calibration remains a separate diagnostic.

Predictive uncertainty was positively associated with absolute error (Pearson r = 0.67). The interval-width rule flagged 103 of 1,228 test segments (8.4%) and captured 35.2% of errors exceeding 25 in/mi (Figure 14). Random selection would capture 8.4% of large errors in expectation at the same flagged fraction. The observed rule, therefore, concentrated large errors at the stated 8.4% inspection fraction; a threshold sweep would extend the analysis to other inspection budgets.

The fixed-width conformal interval provides a benchmark for marginal coverage, whereas MC dropout supplies segment-specific widths for inspection ranking. Interval calibration and screening discrimination are therefore complementary performance dimensions. Future work can evaluate locally adaptive conformal intervals as a route to combining residual calibration with segment-specific prioritization.

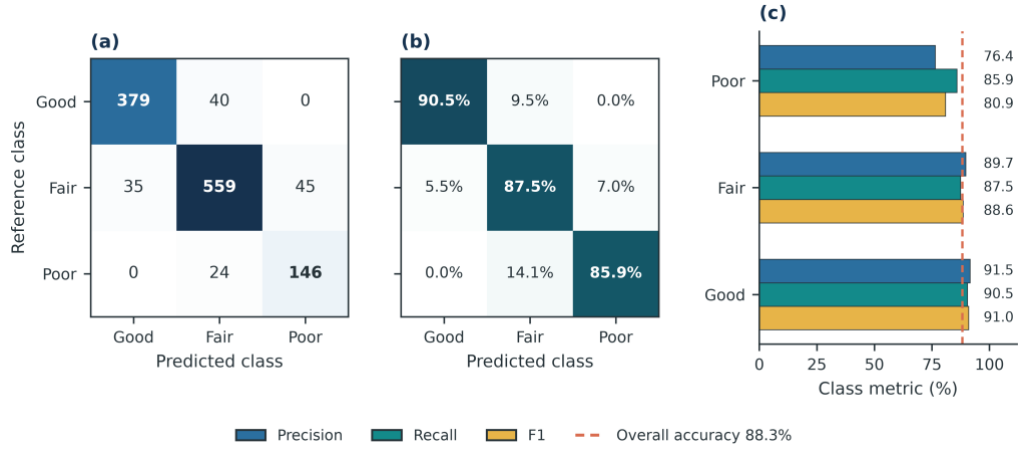


Figure 12: Full-fusion classification on 1,228 held-out segments: (a) count matrix; (b) row-normalized matrix; (c) precision, recall, and F1 with overall accuracy.

Table 8: Prediction-interval comparison at nominal 90% coverage for the full-fusion S+P+C predictor on the locked corridor-held-out test cohort.

Method	PICP (%)	Mean width (in/mi)	Median width (in/mi)	Interval score (in/mi)	Signed coverage error (pp)
MC dropout, K = 50	89.6	32.8	31.4	34.1	−0.4
Split conformal, fixed width	90.2	56.8	56.8	58.0	+0.2

PICP: prediction interval coverage probability. pp: percentage points. Interval scores are compared at the same nominal coverage. A fixed-width interval does not provide a segment-specific width ranking.

4.7 Computational throughput and cost scope

The study processed the 1,823-scene archive in approximately 72 hours using the University of Wyoming's research computing environment. On an NVIDIA A100 GPU, the prepared-segment inference times were 0.22 s for S, 0.08 s for P+C, and 0.30 s for S+P+C. The latter corresponds to approximately 12,000 segments per hour. These measurements quantify prepared-segment inference throughput; end-to-end processing additionally includes data acquisition, preprocessing, and quality control.

The marginal computational estimate was approximately USD 0.21 per centerline mile. It excludes one-time model development, agency integration, data maintenance, and field confirmation. The estimate therefore characterizes the processing component of a screening service, not its total life-cycle cost or a like-for-like replacement for a

CC BY

lane-mile profiler survey. Its practical value is the low marginal cost of generating another screening layer once the required infrastructure and quality controls are established. The achievable update interval remains constrained by the availability of usable acquisitions, environmental screening, temporal support, and agency review.

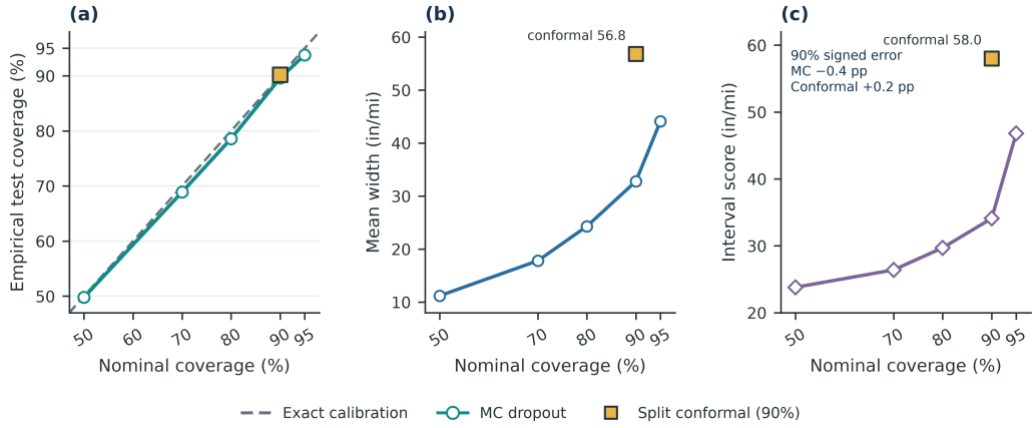


Figure 13: Uncertainty analysis for the full-fusion S+P+C predictor: (a) empirical versus nominal MC-dropout coverage; (b) mean prediction-interval width; and (c) interval score.

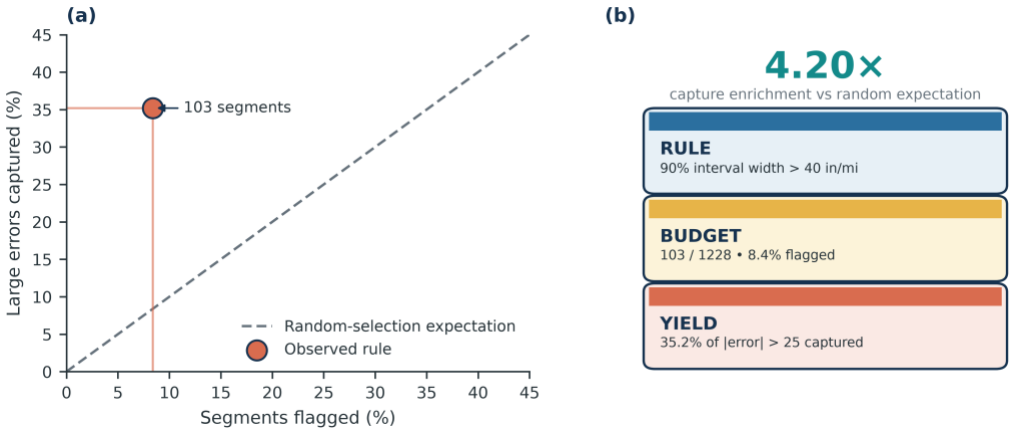


Figure 14: Uncertainty-guided screening for the full-fusion S+P+C predictor: (a) observed large-error capture versus random selection at the same inspection fraction; and (b) operating rule and outcome.

5. DISCUSSION

5.1 Input availability and model performance

The seven-configuration experiment makes the timing of information the organizing principle for interpretation. Among configurations that exclude the same campaign, C, S+P achieved the lowest MAE at 13.4 in/mi. Full fusion reached 12.9 in/mi after adding contemporaneous rut depth, MPD, and fatigue cracking, an incremental improvement of 0.5 in/mi. Because S+P uses a symmetric ±25-day SAR window around the inspection date, it provides the strongest retrospective basis for a future between-survey workflow rather than a measured live-deployment accuracy.

Agency use follows the information available at the estimate date. S represents the satellite-only option; S+P adds complete, temporally valid pre-inspection records; and S+P+C represents the upper-information setting once contemporaneous indicators are available. Prospective operation requires one-sided SAR acquisition and independent issue-date validation. Where only part of P is available, the agency should fit and validate the corresponding reduced-input model rather than assign it the reported S+P performance.

The results from the contemporary model reinforce this interpretation. AnySat improves SAR-only accuracy modestly relative to the reference encoder, showing that the framework can benefit from advances in Earth-

observation representation learning. Such improvements should be incorporated through matched experiments rather than obscured by comparisons across unequal input sets. The workflow is not conceptually tied to EfficientNet-B0, but substituting an encoder still requires calibrating and validating its output behavior.

5.2 Three-tier framework for agency implementation

Satellite-derived IRI is best positioned as a Tier 1 screening product. A segment can be escalated because the predicted condition is poor, uncertainty is high, an acquisition fails quality checks, or treatment history makes the satellite-to-survey alignment unreliable. Tier 2 uses targeted profiling to obtain the measured longitudinal profile required for engineering confirmation. Tier 3 applies more detailed investigation, such as UAV imaging or lidar, where a localized failure mechanism or project-level decision warrants additional spatial detail. Table 9 connects these roles to their information and decision requirements.

The operational output should retain the segment and corridor identity, the target and acquisition dates, the input configuration, the model version, predicted IRI and units, interval bounds, the IRI class, and the reason for any escalation. These fields distinguish a valid low-condition flag from a prediction generated under inadequate observational support. They also make later profiler confirmation traceable, enabling agencies to evaluate whether the screening system is improving inspection allocation rather than merely producing more maps (Visintine et al., 2018). The appendix defines a recommended exchange schema for this purpose.

Agency commissioning should begin with a geographically bounded pilot. The local evaluation set should contain profiler measurements excluded from adaptation and should report continuous error, Poor-class recall and precision, interval coverage, and large-error capture at the available inspection budget. Segment geometry, route direction, treatment dates, and input timing require particular attention because errors in those joins can mimic changes in pavement condition. Human authorization remains appropriate for inspection scheduling and maintenance decisions, even when the upstream processing is scripted.

The observed screening point demonstrates that model uncertainty can concentrate large errors within a relatively small inspection subset: 8.4% of segments contained 35.2% of errors exceeding 25 in/mi. This study, therefore, quantifies prioritization efficiency at a single stated threshold. A prospective, budget-matched evaluation can translate that efficiency into inspections avoided, confirmation yield, and maintenance-planning value.

Table 9: Proposed three-tier integration of satellite screening and field confirmation.

Tier	Primary information	Role in the agency workflow	Decision boundary
1: network screening	Availability-specific SAR-derived IRI, interval, and quality flags	Identify candidate segments for review or measurement; update a PMS screening layer	Not construction acceptance or a measured road profile
2: targeted confirmation	Inertial-profiler measurements and agency inspection records	Confirm condition and resolve uncertain or high-priority flags	Engineering decisions use measured reference information
3: detailed investigation	Project-specific UAV, lidar, or other forensic evidence	Characterize localized defects or failure mechanisms when needed	Selected according to the engineering question, not triggered by an accuracy claim alone

5.3 Model scope and operational boundaries

Within the evaluated domain, the framework demonstrated consistent performance on usable acquisitions that met the reported environmental and geometric quality criteria. The present validation encompasses flexible-asphalt pavements across sub-arctic, arid continental, and humid temperate conditions, as represented by Alaska, Utah, and West Virginia. Extension to other pavement materials, climatic regimes, design practices, radar bands, or satellite platforms will require context-specific evaluation and calibration.

Quantitative diagnostics showed no pronounced orientation sensitivity within the evaluated Sentinel-1 acquisition geometry. SAR-only MAE varied from 14.5 to 15.2 in/mi across the three road-axis-to-ground-look groups; the bootstrap confidence intervals overlapped, and the continuous orientation–absolute-error association was small (Spearman $\rho = -0.04$, $p = 0.18$). The joint SAR-only regression likewise detected no strong orientation or orientation-by-width association after adjustment for state.

CC BY

Road width exhibited a modest directional pattern. For the full-fusion model, MAE decreased from 14.6 in/mi below 8 m to 12.2 in/mi above 12 m. The continuous width–absolute-error association was weak ($\rho = -0.11$, $p = 0.06$), and the state-adjusted width estimate remained marginal ($p = 0.07$). This pattern is physically consistent with a greater pavement contribution and reduced mixed scattering on wider carriageways, although the observed association was weak.

Metallic guardrails are especially relevant because they can persist across every acquisition in a temporal composite. Their scattering behavior is therefore not addressed in the same way as transient vehicles. Guardrail presence, material, offset, continuity, and look-direction geometry are useful candidate predictors when a reliable inventory exists. Guardrails were not encoded as a separate engineering variable in the present experiment. Their potential influence lies in the mixed-scattering interpretation of the SAR inputs, rather than in a criticism of the profiler reference measurements.

DEM choice affects local incidence angle, terrain normalization, geolocation, and the usable-pixel mask. A meaningful comparison requires common scenes, spatial support, and an explicit decision about whether the learning model is held fixed or refitted. Accordingly, the present results are interpreted within the reported elevation workflow rather than as a comparison among DEM products. Similarly, the Lee Sigma results characterize window selection within the studied workflow and fixed-model sensitivity, rather than universal superiority over all speckle-filter families. Table 10 summarizes the operational consequences of these boundaries.

Table 10: Operational boundaries and their implications.

Boundary	Implication	Practical response
Mixed spatial support	Pixels can include pavement, shoulders, vegetation, and roadside objects; interpolation adds no independent detail	Retain segment-scale interpretation and confirm localized defects with suitable field measurements
Symmetric temporal compositing	The evaluated estimate uses observations before and after the target date	Report the observation window; separately validate trailing-window or real-time use
Material, climate, and sensor domain	Unseen pavement types, climates, radar bands, and constellations are not evaluated	Use a local external evaluation before assigning the reported accuracy to a new setting
Geometry and elevation	Look direction, paved width, terrain, and DEM choice can alter usable signatures	Retain geometry and quality flags; evaluate matched support for future product comparisons
Engineering-data availability	Full-fusion accuracy depends on all supplied input groups	Select the independently evaluated configuration consistent with the data available at prediction time
Calibration and inspection policy	Marginal coverage and one screening point do not establish all subgroup or decision outcomes	Monitor calibration and evaluate inspection performance under the agency's actual budget

5.4 Future sensing and monitoring strategies

Commercial higher-resolution SAR could reduce pavement-shoulder mixing and improve separation of some roadside objects. Improved IRI accuracy is nevertheless a hypothesis rather than a direct consequence of smaller pixels. Wavelength, polarization, incidence geometry, radiometric response, temporal availability, and licensing also change. Commercial acquisition and licensing costs were outside the scope of this study; therefore, cost-effectiveness remains an empirical question for the controlled comparison. A controlled follow-on experiment would evaluate the same road segments and reference dates, compare prediction error and calibration, and include acquisition, processing, and confirmation costs on a common basis.

A post-repaving reference signature offers a complementary route to localized monitoring. The initial SAR observation should be linked to a measured post-treatment IRI, then compared with later acquisitions under compatible orbit, look geometry, season, and moisture screening. The engineering target would be a change in measured IRI rather than a change in backscatter alone. Curing, surface aging, crack sealing, and nearby construction can change the radar signature without producing an equivalent change in ride quality. A longitudinal evaluation should therefore retain treatment events and later time periods outside model development.

CC BY

Further methodological work can build on the strongest results of the present evaluation. Contemporary encoders can be tested within S+P as well as S; adaptive conformal methods can be compared with MC dropout, and uncertainty-guided inspection can be evaluated prospectively against random allocation and agency rules. The relevant progression is from accurate retrospective estimation to validated updating and decision support, with each step retaining explicit information, timing, and independently evaluated outcomes.

6. CONCLUSION

This study developed and evaluated a scripted Sentinel-1-to-PMS workflow for segment-level pavement roughness screening. Its central contribution is the integration of remote observations, input-availability-specific prediction, comparative model evaluation, and field-confirmation pathways. The workflow was evaluated on 1,228 corridor-held-out test segments from an 8,188-segment flexible-asphalt cohort in Alaska, Utah, and West Virginia.

The seven-configuration experiment establishes S+P as the primary retrospective result without same-campaign C. SAR alone achieved MAE 14.8 in/mi, S+P achieved 13.4 in/mi, and P alone achieved 16.1 in/mi. Full fusion provided the upper-information benchmark, with MAE 12.9 in/mi, RMSE 17.2 in/mi, and $r^2 = 0.69$; adding same-campaign condition indicators to S+P further reduced MAE by 0.5 in/mi. Contemporary SAR-only AnySat and SkySense adaptations achieved mean MAEs of 14.2 and 14.5 in/mi, respectively, with AnySat providing the strongest matched SAR-only result.

The full-fusion classifier correctly assigned 88.3% of test segments and achieved 85.9% recall for the Poor category. For the full-fusion S+P+C predictor, MC-dropout and fixed-width split-conformal intervals achieved empirical coverages of 89.6% and 90.2%, respectively, at a nominal 90% level. MC dropout produced segment-specific intervals with a mean width of 32.8 in/mi, compared with 56.8 in/mi for the fixed-width conformal benchmark; a rule flagging 8.4% of segments captured 35.2% of absolute errors above 25 in/mi. Together with the observed processing throughput and marginal computational cost, these results support satellite-based prioritization of field measurement.

All reported performance belongs to the evaluated domain. Within that domain, separating information availability from attainable accuracy provides a practical foundation for incorporating satellite-derived condition estimates into construction asset management. The next deployment step is an independently evaluated agency pilot that uses one-sided, issue-date acquisition rules and measures both prediction quality and inspection-decision value.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the Alaska Department of Transportation & Public Facilities, the Utah Department of Transportation, and the West Virginia Department of Transportation for pavement condition and inventory data. The authors also acknowledge the European Space Agency and Alaska Satellite Facility for Sentinel-1 observations and distribution services, and the University of Wyoming research-computing infrastructure. Opinions, findings, and conclusions are those of the authors and do not necessarily reflect the views of the data-providing agencies.

DATA AVAILABILITY

Sentinel-1 imagery and LTPP data are publicly available through Copernicus/ASF and InfoPave, respectively. State-agency PMS data remain subject to applicable reuse restrictions. The appendix and accompanying files provide aggregate results and figure-reproduction materials. Study data and code may be requested from the corresponding author, subject to agency approval.

AUTHOR CONTRIBUTIONS

Md Tashfique Enam Tutul: Conceptualization, Methodology, Data curation, Validation, Formal analysis, Visualization, Writing: original draft. **Ahmed Abdelaty:** Conceptualization, Methodology, Resources, Supervision, Validation, Writing: review and editing. **Austin Madson:** Methodology, Validation, Writing: review and editing. **Shuvo Dip Datta:** Validation, Formal analysis, Writing: review and editing.

CC BY

DECLARATION OF GENERATIVE AI IN SCIENTIFIC WRITING

During manuscript preparation, the authors used AI-assisted tools for language editing, structural refinement, document preparation, and consistency checking. The authors reviewed and edited the outputs and take full responsibility for the content of this publication.

CONFLICT OF INTEREST

The authors declare no competing financial or non-financial interests.

REFERENCES

Angelopoulos, A. N., & Bates, S. (2022). *A gentle introduction to conformal prediction and distribution-free uncertainty quantification*. https://doi.org/10.48550/arXiv.2107.07511

Astruc, G., Gonthier, N., Mallet, C., & Landrieu, L. (2025). *AnySat: One Earth observation model for many resolutions, scales, and modalities* Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, *41*(2), 423-443.

Bashar, M., & Torres-Machi, C. (2023). Quantifying the value of satellite-based pavement monitoring in partially observable stochastic environments. *Journal of Computing in Civil Engineering*, *37*(3), 04023004.

Bashar, M. Z., & Torres-Machi, C. (2022). Deep learning for estimating pavement roughness using synthetic aperture radar data. *Automation in Construction*, *142*, 104504.

Chang, J.-R., Chang, K.-T., & Chen, D.-H. (2006). Application of 3D laser scanning on measuring pavement roughness. *Journal of Testing and Evaluation*, *34*(2), 83-91.

E-08, A. (2015). Standard practice for computing international roughness index of roads from longitudinal profile measurements. *West Conshohocken, PA: ASTM International*.

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., & Roth, L. (2007). The shuttle radar topography mission. *Reviews of geophysics*, *45*(2).

Filipponi, F. (2019). Sentinel-1 GRD preprocessing workflow. Proceedings,

Fiorentini, N., Maboudi, M., Leandri, P., & Losa, M. (2021). Can machine learning and PS-InSAR reliably stand in for road profilometric surveys? *Sensors*, *21*(10), 3377.

Fiorentini, N., Maboudi, M., Leandri, P., Losa, M., & Gerke, M. (2020). Surface motion prediction and mapping for road infrastructures management by PS-InSAR measurements and machine learning algorithms. *Remote Sensing*, *12*(23), 3976.

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. Proceedings of the 33rd International Conference on Machine Learning,

Guo, X., Lao, J., Dang, B., Zhang, Y., Yu, L., Ru, L., Zhong, L., Huang, Z., Wu, K., Hu, D., He, H., Wang, J., Chen, J., Yang, M., Zhang, Y., & Li, Y. (2024). *SkySense: A multi-modal remote sensing foundation model towards universal interpretation for Earth observation imagery* Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Haralick, R. M., Shanmugam, K., & Dinstein, I. h. (1973). Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-3*(6), 610-621.

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, *4*(9).

Karimzadeh, S., & Matsuoka, M. (2020). Remote sensing X-band SAR data for land subsidence and pavement monitoring. *Sensors*, *20*(17), 4751. https://doi.org/10.3390/s20174751

Keys, R. G. (1981). Cubic convolution interpolation for digital image processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *29*(6), 1153-1160. https://doi.org/10.1109/TASSP.1981.1163711

Kingma, D. P., & Ba, J. (2014). *Adam: A method for stochastic optimization*. https://doi.org/10.48550/arXiv.1412.6980

Kumar, P., Lewis, P., McElhinney, C. P., & Rahman, A. A. (2015). An algorithm for automated estimation of road roughness from mobile laser scanning data. *The Photogrammetric Record*, *30*(149), 30-45.

Lee, J.-S., Jurkevich, L., Dewaele, P., Wambacq, P., & Oosterlinck, A. (1994). Speckle filtering of synthetic aperture radar images: A review. *Remote sensing reviews*, *8*(4), 313-340.

Lee, J.-S., Wen, J.-H., Ainsworth, T. L., Chen, K.-S., & Chen, A. J. (2009). Improved sigma filter for speckle filtering of SAR imagery. *IEEE Transactions on Geoscience and Remote Sensing*, *47*(1), 202-213. https://doi.org/10.1109/TGRS.2008.2002881

Lin, Y., & Mitchell, K. E. (2005). The NCEP Stage II/IV hourly precipitation analyses: Development and applications. Proceedings of the 19th Conference on Hydrology,

Mesinger, F., DiMego, G., Kalnay, E., Mitchell, K., Shafran, P. C., Ebisuzaki, W., Jović, D., Woollen, J., Rogers, E., Berbery, E. H., Ek, M. B., Fan, Y., Grumbine, R., Higgins, W., Li, H., Lin, Y., Manikin, G., Parrish, D., & Shi, W. (2006). North American regional reanalysis. *Bulletin of the American Meteorological Society*, *87*(3), 343-360.

Meyer, F. (2019). Spaceborne synthetic aperture radar: Principles, data access, and basic processing techniques. In *The SAR handbook*. NASA SERVIR and SilvaCarbon.

Meyer, F. J., Ajadi, O. A., & Hoppe, E. J. (2020). Studying the applicability of X-band SAR data to the network-scale mapping of pavement roughness on US roads. *Remote Sensing*, *12*(9), 1507.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, *37*(1/2), 17-23.

Moreira, A., Prats-Iraola, P., Younis, M., Krieger, G., Hajnsek, I., & Papathanassiou, K. P. (2013). A tutorial on synthetic aperture radar. *IEEE Geoscience and remote sensing magazine*, *1*(1), 6-43.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011a). Multimodal deep learning. ICML,

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011b). Multimodal deep learning. Proceedings of the 28th International Conference on Machine Learning,

Sattar, S., Li, S., & Chapman, M. (2018). Road surface monitoring using smartphone sensors: A review. *Sensors*, *18*(11), 3845.

Sayers, M. W., Gillespie, T. D., & Queiroz, C. A. (1986). *The international road roughness experiment: Establishing correlation and a calibration standard for measurements*.

Sayers, M. W., & Karamihas, S. M. (1998). *The little book of profiling: Basic information about measuring and interpreting road profiles*.

Small, D. (2011). Flattening gamma: Radiometric terrain correction for SAR imagery. *IEEE Transactions on geoscience and remote sensing*, *49*(8), 3081-3093.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929-1958.

Suanpaga, W., & Yoshikazu, K. (2010). Riding quality model for asphalt pavement monitoring using phase array type L-band synthetic aperture radar (PALSAR). *Remote Sensing*, *2*(11), 2531-2546.

Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. Proceedings of the 36th International Conference on Machine Learning,

Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., Potin, P., Rommen, B., Floury, N., Brown, M., Navas-Traver, I., Deghaye, P., Duesmann, B., Rosich, B., Miranda, N., Bruno, C., L'Abbate, M., Croci, R., Pietropaolo, A., . . . Rostan, F. (2012). GMES Sentinel-1 mission. *Remote Sensing of Environment*, *120*, 9-24. https://doi.org/10.1016/j.rse.2011.05.028

Visintine, B. A., Rada, G. R., & Simpson, A. L. (2018). *Guidelines for Informing Decisionmaking to Affect Pavement Performance Measures*.

APPENDIX A

This appendix reports the processing settings, evaluation partitions, per-seed model results, development sensitivity, quantitative geometric diagnostics, uncertainty summaries, and classification counts that accompany the main manuscript. S denotes SAR information, P denotes pre-inspection engineering information, and C denotes contemporaneous condition information. Full-fusion labels apply only where stated: orientation diagnostics use S, whereas width, interval, and screening results use S+P+C. All errors are expressed in inches per mile (in/mi) unless another unit is stated; r^2 denotes squared Pearson correlation, while uppercase R^2 in Table A19 denotes regression-model explained variation.

The final test cohort comprises 1,228 observations and is separate from the 6,960-observation development pool. For the uncertainty experiment only, 1,188 development observations were reserved before point-model fitting for split-conformal calibration, leaving 5,772 observations to fit the S+P+C uncertainty predictor. Architecture mean ± SD summarizes three fitting seeds; development sensitivity mean ± SD summarizes five corridor-grouped folds; bootstrap confidence intervals are reported only where retained with the corresponding experiment. All additional numerical values are taken from retained aggregate experiment outputs. Their arithmetic and cross-document consistency were checked during resubmission preparation.

A.1 Acquisition and preprocessing

Table A1: Acquisition and preprocessing configuration.

Component	Study setting	Interpretation
SAR archive	1,823 Sentinel-1A/B IW high-resolution GRD products; ascending; VV and VH	C-band observations processed consistently across the study regions
Output grid	10 m pixel spacing in state-appropriate UTM coordinates	Output sampling, not independent 10 m spatial resolution
Software	SNAP v9.0, controlled through its Python interface	Scripted scene preparation and feature extraction
Radiometric/geometric steps	Orbit correction; noise treatment; calibration; terrain flattening; Lee Sigma filtering; Range-Doppler correction	Terrain-normalized γ^0 with geometric and environmental quality checks
Elevation	SRTM 1-arc-second v3.0 where available; regional Alaska fallback elevation tiles where required	SRTM 1-arc-second v3.0 was used where available, with regional Alaska fallback elevation tiles where required. The reported results correspond to this elevation workflow; DEM-product sensitivity was outside the matched experimental design
Speckle reference	Lee Sigma, 7 × 7 window	Development-selected reference; adjacent-window and fixed-model sensitivity are reported below
Temporal support	Symmetric ±25 days around the IRI survey	Retrospective current-condition support, not a forecast horizon
Temporal compositing	Temporal minimum followed by residual ±3-standard-deviation screening and temporal-median replacement	Suppresses some transient bright returns; does not remove every persistent roadside scatterer
Weather screening	Precipitation >5 mm in the preceding 24 hours excluded; Stage IV for UT/WV and NARR for AK	Dry-surface interpretation; snow, ice, and standing water limit usable observations
Geometric/treatment exclusions	Elevated structures; within 50 m of intersections; gradients >8%; layover/shadow; major rehabilitation inside the matching window	Reduces confounding without claiming pure pavement pixels
Patch representation	Approximately 16 × 3 grid samples over a 0.1-mile segment, interpolated to 64 × 64	Anisotropic bicubic interpolation standardizes the tensor; it does not add independent spatial information
Image channels	VV γ^0, VH γ^0, normalized local incidence angle	Three-channel image representation
GLCM inventory	Contrast, Dissimilarity, Homogeneity, Energy, and Correlation; 3 × 3, 5 × 5, and 7 × 7 windows; two polarizations	30 texture descriptors with averaged directional offsets
Feature scaling	Training-partition statistics applied to the associated validation/test observations	Scaling does not use the held-out target observations

CC BY

The observation unit remains the pavement reporting segment after interpolation. In particular, an increase from approximately three across-track samples to 64 tensor entries is not interpreted as a corresponding increase in effective across-road resolution. Road masks emphasize the corridor but can retain contributions from shoulders and roadside objects. A persistent guardrail is therefore not equivalent to a transient vehicle artifact.

Road orientation is defined as the acute angle between the road-axis bearing and the Sentinel-1 ground-look azimuth. Paved-carriageway width is distinct from centerline-buffer width, pixel spacing, and effective radar resolution. Quantitative orientation and width diagnostics for the locked test cohort are reported in Section A.10; the normalized width W used there is a numerical rescaling of recorded carriageway width, not an estimate of physical pixel support or pure-pavement fraction.

The symmetric temporal window defines a retrospective estimate around the survey date because observations after the target date may be included. Prospective operation using only acquisitions available by the issue date requires separate validation with one-sided temporal support.

A.2 Development partitions and reference training

Table A2: Full-fusion five-fold development results and fold sizes.

Fold	Training n	Validation n	MAE (in/mi)	RMSE (in/mi)	r^2
1	5,570	1,390	13.7	18.4	0.66
2	5,565	1,395	12.7	17.0	0.70
3	5,568	1,392	13.0	17.6	0.68
4	5,570	1,390	13.5	18.1	0.66
5	5,567	1,393	13.1	17.9	0.70
Mean ± SD	Approximately 5,568	Approximately 1,392	13.2 ± 0.4	17.8 ± 0.5	0.68 ± 0.02

Each row partitions all 6,960 development observations. Corridors remain within one side of each training/validation split. The 1,228 final test observations are not included in these folds.

The reference encoder is ImageNet-pretrained EfficientNet-B0 with early blocks 1–6 frozen in the initial training phase. Global average pooling gives a 1,280-dimensional representation. The full-fusion engineering embedding is 64-dimensional, with a concatenated 1,344-dimensional representation and fully connected regression layers of 256 and 128 neurons before the scalar output.

The optimizer is Adam, not AdamW, with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and an initial learning rate of 10^{-4}. Training uses Huber loss with $\delta = 1.0$ in/mi, batch size 32, dropout 0.3, L2 coefficient 10^{-4}, a ReduceLROnPlateau schedule, and early-stopping patience of 15 epochs. Validation loss within each development fold guides training control. The local sensitivity study changes one of these factors at a time rather than claiming a fully factorial or globally optimal search.

A.3 Contemporary architecture comparison

Table A3: Per-seed SAR-only test results.

Task implementation	Seed	MAE (in/mi)	RMSE (in/mi)	r^2
EfficientNet-B0	42	14.8	19.5	0.55
EfficientNet-B0	123	14.6	19.3	0.55
EfficientNet-B0	456	15.0	19.7	0.54
AnySat adaptation	42	14.2	18.8	0.57
AnySat adaptation	123	14.4	19.0	0.57
AnySat adaptation	456	14.0	18.7	0.58
SkySense adaptation	42	14.5	19.2	0.56
SkySense adaptation	123	14.3	19.0	0.56
SkySense adaptation	456	14.7	19.4	0.55

All task implementations are evaluated on the same 1,228 corridor-held-out observations and fitted with seeds 42, 123, and 456. The target and SAR-only tasks are matched, while input tensors and checkpoint normalization remain architecture-specific as documented in Table A4. The main manuscript reports the supplied three-seed summaries, rounded independently of the displayed per-seed values.

Table A4: Reproducibility matrix for the task-specific backbone adaptations.

Item	EfficientNet-B0	AnySat adaptation	SkySense adaptation
Repository / implementation	pytorch-image-models (timm)	github.com/gastruc/AnySat	github.com/Jack-bo1220/SkySense
Reference	(Tan & Le, 2019)	(Astruc et al., 2025)	(Guo et al., 2024)
Pretrained checkpoint	ImageNet-1k default weights distributed through timm	CVPR 2025 released weights	CVPR 2024 released weights
Original modalities	RGB	Multi-resolution Earth observation: SAR and optical	Multimodal SAR and optical
Sentinel-1 inputs used	VV γ°, VH γ°, LIA	VV γ°, VH γ°; LIA supplied as third channel	VV γ°, VH γ°
Input tensor	64 × 64 × 3	64 × 64 × 3	64 × 64 × 2
Channel normalization	ImageNet mean and SD	Sentinel-1 ascending-branch statistics	SAR-encoder statistics from the released checkpoint
Frozen / trainable components	Blocks 1–6 frozen; later encoder blocks and regression head trainable	Pretrained backbone encoder frozen; regression head trainable	Pretrained SAR encoder frozen; regression head trainable
Total parameters	≈5.3 M	≈86 M	≈123 M
Trainable parameters	≈1.3 M	≈2.1 M	≈2.3 M
Regression head	FC 256 → FC 128 → IRI	FC 256 → FC 128 → IRI	FC 256 → FC 128 → IRI
Loss	Huber, δ = 1.0 in/mi	Huber, δ = 1.0 in/mi	Huber, δ = 1.0 in/mi
Optimizer	Adam; lr = 1×10^{-4}; L2 = 1×10^{-4}	Same	Same
Batch size / early stopping	32; patience 15 epochs	32; patience 15 epochs	32; patience 15 epochs
Augmentation	None	None	None
Random seeds	42, 123, 456	42, 123, 456	42, 123, 456
Development evaluation	n = 6,960; five corridor-grouped folds	Same	Same

CC BY

Item	EfficientNet-B0	AnySat adaptation	SkySense adaptation
Final test	n = 1,228; locked corridor-held-out cohort	Same	Same
Selection criterion	Development-fold MAE	Development-fold MAE	Development-fold MAE
Deviation from original	Classification head replaced by regression head	Optical branch disabled; ascending Sentinel-1 branch retained; input adapted to study tensor	Optical branch disabled; SAR encoder retained; input reduced to two SAR channels

The parameter counts are approximate counts for the study-specific task implementations. They are not measured runtime or cost comparisons and, for AnySat and SkySense, do not represent deployment of the complete original multimodal systems. Optical inputs were disabled, and Table A4 distinguishes the channels, frozen components, trainable regression heads, and other deviations used for this matched SAR-only task. LIA denotes local incidence angle, and FC denotes fully connected. 'Same' refers to the EfficientNet-B0 column.

The comparison uses the released checkpoint families and implementation sources recorded in Table A4. EfficientNet-B0 is cited to (Tan & Le, 2019), AnySat to (Astruc et al., 2025), and SkySense to (Guo et al., 2024).

Mean ΔMAE is computed as comparator minus EfficientNet-B0. The AnySat contrast is −0.6 in/mi with a bootstrap 95% confidence interval [−1.1, −0.1]; the SkySense contrast is −0.3 in/mi with an interval [−0.9, +0.3]. The main manuscript uses those contrasts without extending them to differences in field deployment cost. Results from (Bashar & Torres-Machi, 2022) are treated separately as published cross-study context.

A.4 Complete development sensitivity

Each sensitivity experiment uses corridor-grouped development evaluation and leaves the final test set outside parameter selection. All factors other than the factor being varied remain at their reference settings. The tables report mean validation MAE and between-fold SD; Figures 6a–b and 7a–c summarize the temporal and filter experiments. Small differences are descriptive and carry no unreported significance claim.

The temporal comparison uses common support: approximately 5,639 development observations with at least one valid acquisition under the most restrictive ±10-day condition. Keeping these observations fixed separates changes in temporal support from changes in the sampled cohort. The chosen ±25-day setting lies near the observed error plateau, while the ±35-day setting has a slightly lower mean MAE on that common cohort.

Table A5: Temporal-window sensitivity on common development support.

Window	Common cohort n	MAE (in/mi)	Status
±10 days	≈ 5,639	14.2 ± 0.5	Most restrictive window
±15 days	≈ 5,639	13.7 ± 0.4	Sensitivity setting
±25 days	≈ 5,639	13.4 ± 0.4	Selected reference
±35 days	≈ 5,639	13.3 ± 0.4	Sensitivity setting

The same common cohort is used in every row. These counts are not the all-eligible population at each window.

Table A6: Lee Sigma window sensitivity under development cross-validation.

Filter/window	ENL	EPI	MAE (in/mi)
Unfiltered	≈ 5.8	1.00	15.8 ± 0.6
Lee Sigma 3 × 3	≈ 8.1	0.95	15.4 ± 0.5
Lee Sigma 5 × 5	≈ 10.3	0.91	15.0 ± 0.4
Lee Sigma 7 × 7	12.4	0.89	13.2 ± 0.4
Lee Sigma 9 × 9	≈ 14.1	0.85	13.3 ± 0.5

ENL: equivalent number of looks. EPI: edge-preservation index. The selected reference is 7 × 7. Increased ENL alone does not establish lower downstream prediction error.

CC BY

Table A7: Initial learning-rate sensitivity.

Initial learning rate	MAE (in/mi)	Status
1×10^{-5}	14.6 ± 0.5	Sensitivity setting
5×10^{-5}	13.6 ± 0.4	Sensitivity setting
1×10^{-4}	13.2 ± 0.4	Selected reference
5×10^{-4}	13.6 ± 0.5	Sensitivity setting
1×10^{-3}	14.4 ± 0.7	Sensitivity setting

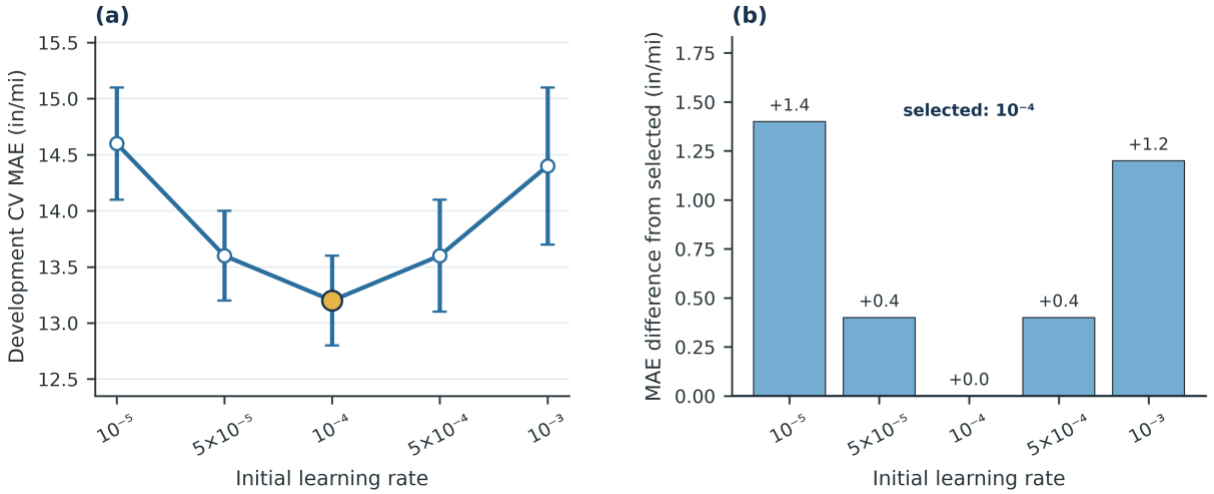


Figure A1: Learning-rate sensitivity: (a) five-fold development MAE ± SD and (b) difference from 10^{-4}.

Table A8: Huber-threshold sensitivity.

Huber δ (in/mi)	MAE (in/mi)	Status
0.5	13.5 ± 0.4	Sensitivity setting
1.0	13.2 ± 0.4	Selected reference
1.5	13.4 ± 0.4	Sensitivity setting
2.0	13.7 ± 0.5	Sensitivity setting

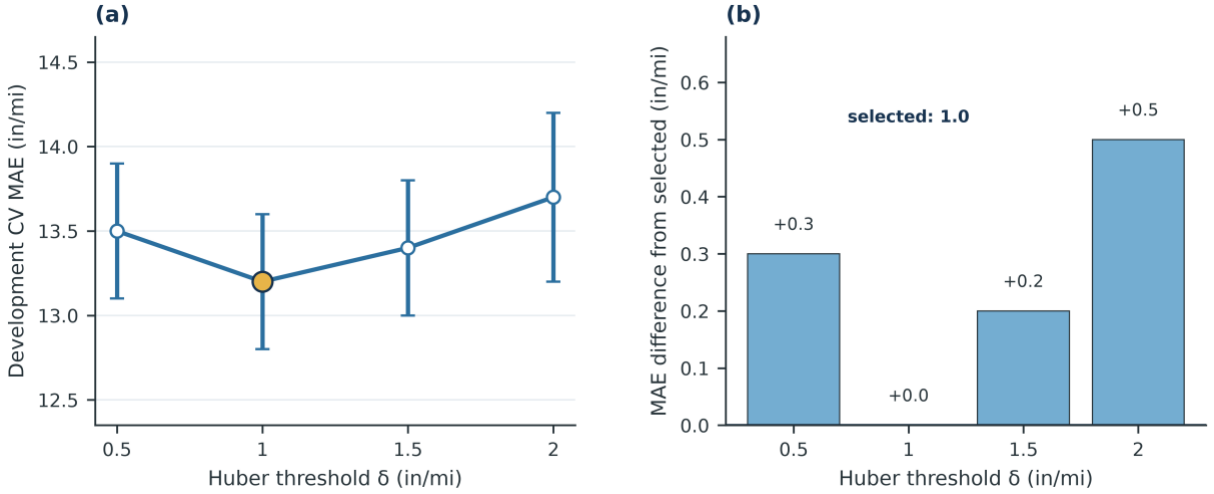


Figure A2: Huber-threshold sensitivity: (a) five-fold development MAE ± SD and (b) difference from δ = 1.0 in/mi.

Table A9: Training-dropout sensitivity.

Dropout probability	MAE (in/mi)	Status
0.0	13.9 ± 0.5	Sensitivity setting
0.1	13.4 ± 0.4	Sensitivity setting
0.2	13.3 ± 0.4	Sensitivity setting
0.3	13.2 ± 0.4	Selected reference
0.5	13.8 ± 0.5	Sensitivity setting

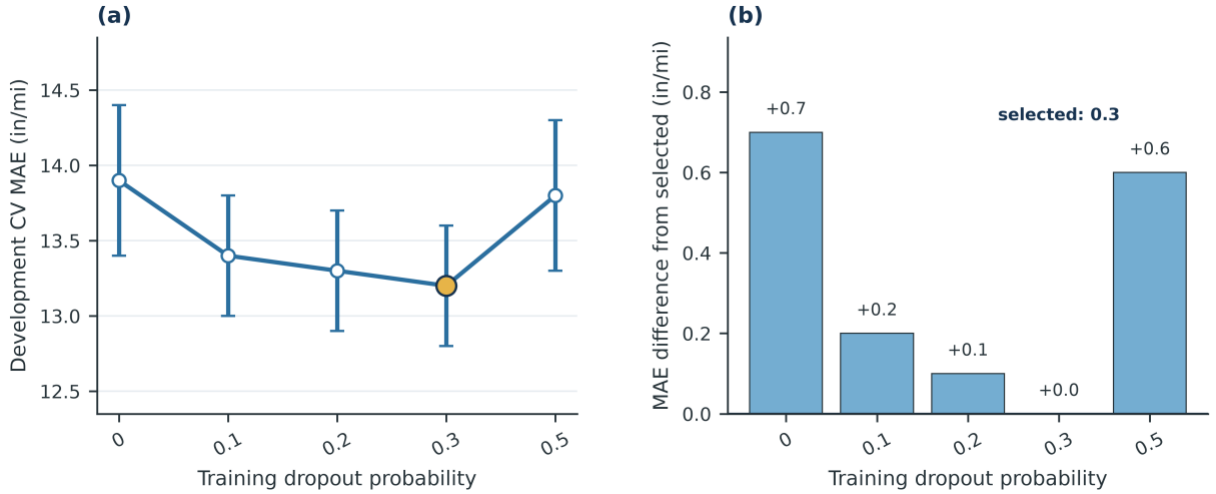


Figure A3: Training-dropout sensitivity: (a) five-fold development MAE ± SD and (b) difference from p = 0.3.

Table A10: Batch-size sensitivity.

Batch size	MAE (in/mi)	Status
16	13.3 ± 0.4	Sensitivity setting
32	13.2 ± 0.4	Selected reference
64	13.5 ± 0.5	Sensitivity setting

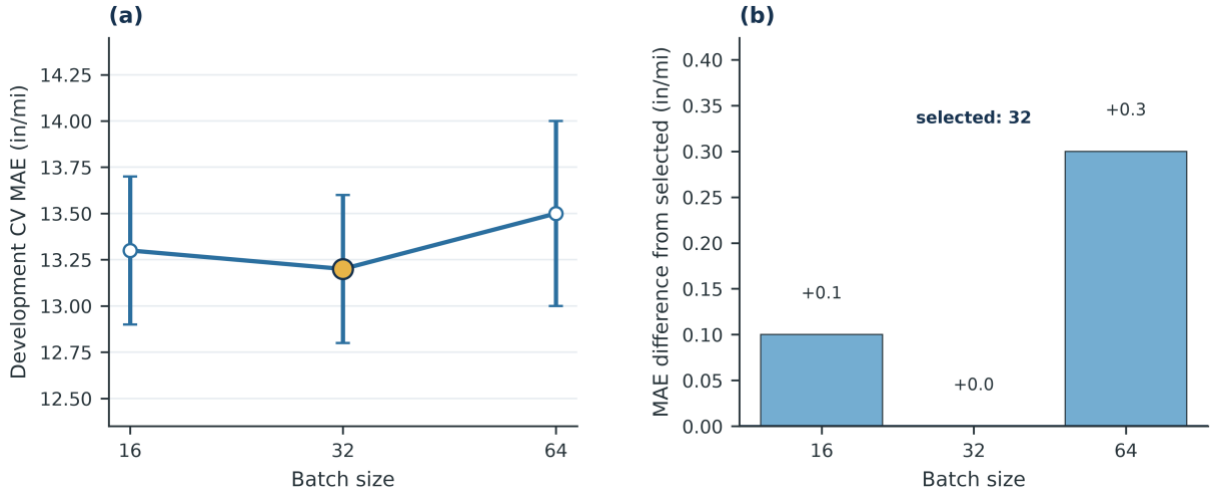


Figure A4: Batch-size sensitivity: (a) five-fold development MAE ± SD and (b) difference from batch size 32.

A.5 Fixed-model filter perturbation

This experiment changes the preprocessing supplied to the fitted S-only model without refitting. It measures compatibility with the reference input transformation; the best achievable comparison would require matched refitting and development selection. Figure 7c and Figures A5a–b show the retained fixed-model results.

Table A11: Fixed-model test sensitivity to Lee Sigma filtering.

Filter/window	ENL	EPI	MAE (in/mi)	RMSE (in/mi)
Unfiltered	≈ 5.8	1.00	16.3	21.2
Lee Sigma 3 × 3	≈ 8.1	0.95	15.4	20.3
Lee Sigma 5 × 5	≈ 10.3	0.91	15.0	19.8
Lee Sigma 7 × 7	12.4	0.89	14.8	19.5
Lee Sigma 9 × 9	≈ 14.1	0.85	14.9	19.6

The S-only model is fixed, the test cohort is the same 1,228 observations, and 7 × 7 is the training/reference filter. These results do not compare independently optimized filters from different families.

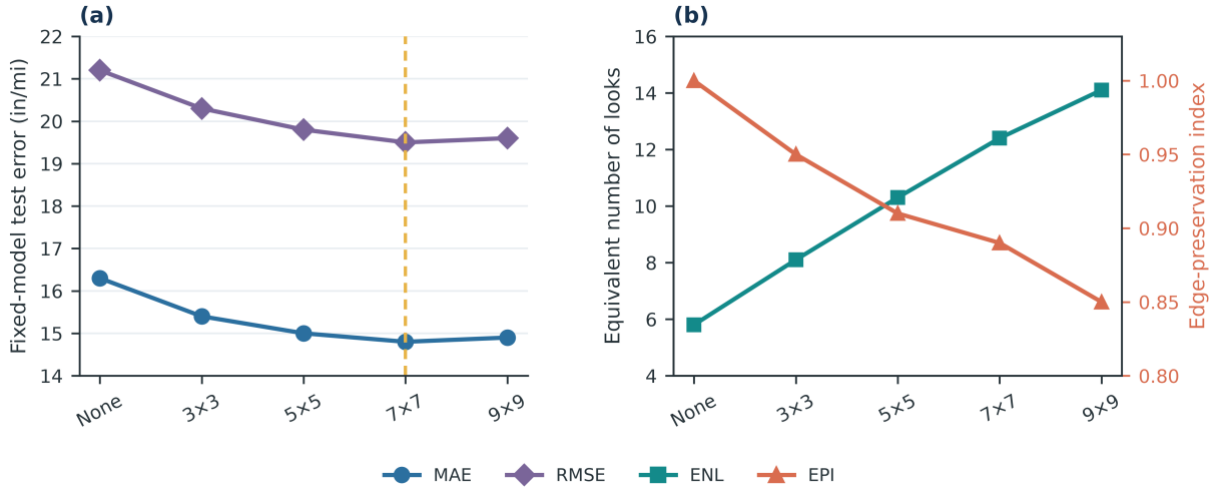


Figure A5: Fixed-model Lee Sigma perturbation: (a) test MAE and RMSE by preprocessing condition; (b) ENL and EPI. The model is not refitted, and 7 × 7 is the training/reference transformation.

A.6 Input-availability effect sizes

The ablation is organized by whether the same-campaign condition indicators C are used. P and S+P exclude C, although the S component retains the retrospective symmetric ±25-day temporal support described in Section A.1. C-containing models quantify the additional information available in an inspection-rich setting. The differences below are descriptive contrasts between independently fitted configurations on the same test cohort; paired confidence intervals and p-values are reported only where supplied.

CC BY

Table A12: Descriptive MAE contrasts for the input-availability experiment.

Comparison	Reference → comparator MAE	Reduction (in/mi)	Reduction relative to reference	Interpretation
P → S	16.1 → 14.8	1.3	8.1%	SAR alone is more accurate than the pre-inspection-only model in this cohort
P → S+P	16.1 → 13.4	2.7	16.8%	Value of adding SAR to pre-inspection information
S → S+P	14.8 → 13.4	1.4	9.5%	Value of pre-inspection context for SAR-based estimation
S+P → S+P+C	13.4 → 12.9	0.5	3.7%	Additional contribution of same-campaign condition indicators
P+C → S+P+C	15.7 → 12.9	2.8	17.8%	Value of SAR in a full engineering-information setting
S → S+P+C	14.8 → 12.9	1.9	12.8%	Full-information comparison, not a SAR-only deployment claim

A.7 Interval coverage and screening metrics

These interval and screening results use the full-fusion S+P+C uncertainty predictor. From the 6,960-observation development pool, 1,188 observations were reserved by corridor-level stratified hold-out exclusively for split-conformal calibration, and the remaining 5,772 were used to fit the point predictor. MC-dropout and split-conformal intervals were evaluated on the locked 1,228-segment test cohort. MC dropout uses 50 stochastic forward passes with probability 0.3. Split conformal uses the absolute calibration residual $|y - f(x)|$; at nominal 90%, the residual order statistic is the 1,071st smallest residual. The fixed interval width of 56.8 in/mi corresponds to a half-width of 28.4 in/mi, a residual-calibration quantile rather than a predictive standard deviation.

For interval endpoints L and U, empirical coverage is the fraction of target values that fall within the interval, and interval width is U − L. Signed coverage error, in percentage points, equals 100 × [empirical coverage − nominal coverage]. The interval score at miscoverage α is defined as follows.

$$IS_\alpha = (U - L) + \left(\frac{2}{\alpha}\right)(L - y)1\{y < L\} + \left(\frac{2}{\alpha}\right)(y - U)1\{y > U\} \quad \text{(A1)}$$

Scores are compared at the same nominal level. These definitions keep regression coverage distinct from classification-confidence calibration.

Table A13: MC-dropout coverage across nominal levels for the full-fusion S+P+C predictor.

Nominal coverage (%)	PICP (%)	Mean width (in/mi)	Signed error (pp)	Interval score (in/mi)
50	49.8	11.2	−0.2	23.8
70	68.9	17.8	−1.1	26.4
80	78.6	24.3	−1.4	29.7
90	89.6	32.8	−0.4	34.1
95	93.8	44.1	−1.2	46.8

For the full-fusion S+P+C predictor, 103 of 1,228 segments are flagged by the 90% MC-dropout interval-width rule >40 in/mi. The flagged fraction is 8.3876%, rounded to 8.4%, and the captured fraction of errors above 25 in/mi is 35.2%. Random selection at the same budget has an expected capture fraction of 8.3876%, yielding an observed-to-expected capture ratio of approximately 4.20 at this operating point.

A fixed-width conformal interval has no interval-width variation across observations, so its width alone cannot supply the same ranking exercise. The main comparison, therefore, assesses its coverage and width without reporting an artificial uncertainty-error correlation for a constant-width predictor. The screening graph in the manuscript uses the observed operating point and a random-selection reference, not a synthetic scatter cloud or a reconstructed continuum of thresholds.

CC BY

A.8 Classification counts and arithmetic

Table A14: Full-fusion three-class counts, with observed categories as rows.

Observed category	Predicted Good	Predicted Fair	Predicted Poor	Observed total
Good	379	40	0	419
Fair	35	559	45	639
Poor	0	24	146	170
Predicted total	414	623	191	1,228

Table A15: Metrics calculated from the three-class count matrix.

Quantity	Calculation	Value
Overall accuracy	(379 + 559 + 146) / 1,228	88.3%
Poor-class recall	146 / 170	85.9%
Poor-class precision	146 / 191	76.4%
Poor-class F1	2 × 146 / (191 + 170)	80.9%
Macro-F1	Unweighted mean of the three class F1 scores	86.8%
Balanced accuracy	Unweighted mean of the three class recalls	87.9%

Class thresholds are Good <95 in/mi, Fair 95–170 in/mi inclusive, and Poor >170 in/mi. Figures 12a–c visualize the retained full-fusion counts and metrics. No Good/Poor cross-confusion occurs, while 24 Poor segments are predicted as Fair.

A.9 Recommended pms exchange schema

Table A16 specifies a recommended agency exchange schema for the three-tier workflow. The schema is an implementation recommendation, not a claim that a particular production PMS integration has been deployed. A versioned record should preserve enough context to determine which model, observations, and input-availability mode produced the estimate. A missing or invalid field should not silently be interpreted as a valid low-uncertainty prediction.

Table A16: Recommended fields for a segment-level screening record.

Field group	Recommended fields	Purpose
Stable identity	Segment ID; corridor ID; state; route; direction; start/end reference; geometry version	Preserve the reporting unit and allow later confirmation to be joined correctly
Temporal provenance	Target date; acquisition IDs/dates; temporal-window start/end; data issue date	Distinguish retrospective support from information available at the time
Model provenance	Model version; preprocessing version; input configuration S/P/C; feature-schema version	Prevent a full-information accuracy claim from being assigned to an incomplete record
Continuous estimate	Predicted IRI; units; lower and upper interval bounds; nominal coverage; uncertainty statistic definition	Keep point estimation, interval coverage, and ranking statistics interpretable
Condition output	IRI class; threshold convention; class-rule version	Separate IRI-only categories from a broader regulatory pavement rating
Quality control	Scene validity; environmental/terrain mask; treatment-window check; usable acquisition count	Identify estimates with inadequate observational support
Escalation	Screening threshold; flag reason; confirmation status; confirmed IRI/date when available	Connect screening to field measurement and decision evaluation

The recommended commissioning evaluation retains local profiler observations that were excluded from model adaptation. The agency can then compare random selection, uncertainty-based screening, and its existing risk rules under the same inspection budget. Useful outcomes include confirmed large-error capture, Poor-category recall, calibrated interval coverage, and the consequences of missed or false inspection priorities. These are prospective evaluation criteria and are not presented as outcomes already demonstrated by the current study.

CC BY

A.10 Geometric error diagnostics

All diagnostics use the locked 1,228-segment test cohort. Orientation is evaluated with SAR-only S errors because the hypothesized mechanism concerns the radar observation; width is evaluated separately with full-fusion S+P+C errors. For each segment, Δφ is the absolute difference between road-axis and ground-look azimuths, and θ is the corresponding acute angle constrained to 0°–90°. For regression, W is recorded as the paved carriageway width divided by 20 m. W is dimensionless; 20 m is only a normalization denominator and is not interpreted as cross-road extraction support or pure-pavement fraction.

$$\theta_i = \min(\Delta\varphi_i, 180° - \Delta\varphi_i); \quad \Delta\varphi_i = \left|\varphi_{road,i} - \varphi_{look,i}\right| \tag{A2}$$

$$\left|e_{i(S)}\right| = \beta_0 + \beta_1\theta_i + \beta_2 W_i + \beta_3\theta_i W_i + \gamma_{AK} I_{AK,i} + \gamma_{WV} I_{WV,i} + \varepsilon_i \tag{A3}$$

$$\left|e_{i(S+P+C)}\right| = \beta_0 + \beta_1 W_i + \gamma_{AK} I_{AK,i} + \gamma_{WV} I_{WV,i} + \varepsilon_i \tag{A4}$$

Table A17: Orientation-stratified SAR-only performance on the locked test cohort.

θ bin	n	MAE	Bootstrap 95% CI for MAE	RMSE	Median AE	P90 AE
0°–30°	312	15.2	[14.0, 16.4]	20.1	12.5	30.1
30°–60°	498	14.5	[13.5, 15.4]	19.1	11.7	28.0
60°–90°	418	14.9	[13.9, 15.9]	19.6	12.2	28.6
Total	1,228	14.8	—	19.5	12.1	28.7

AE: absolute error. All error metrics are in in/mi. The three bin counts sum to 1,228; count-weighted MAE and RMSE² reproduce the pooled values to rounding. Spearman ρ(θ, |error|) = −0.04 (p = 0.18). The supplied aggregate record does not identify the bootstrap resampling unit.

Table A18: Paved-width-stratified full-fusion S+P+C performance on the locked test cohort.

Width class	Paved width	Normalized W	n	MAE	RMSE	Median AE	P90 AE
Narrow	<8 m	<0.40	287	14.6	19.1	11.9	30.2
Medium	8–12 m	0.40–0.60	510	12.6	16.8	10.1	23.1
Wide	>12 m	>0.60	431	12.2	16.4	9.8	21.8
Total	All represented	All represented	1,228	12.9	17.2	10.4	24.3

AE: absolute error. All error metrics are in in/mi. The three class counts sum to 1,228; count-weighted MAE and RMSE² reproduce the pooled values to rounding. Spearman ρ(width, |error|) = −0.11 (p = 0.06).

Table A19: State-adjusted geometric-error regression diagnostics.

Error configuration	Term	Estimate	SE	p
SAR-only joint	Intercept	15.2	1.1	<0.001
SAR-only joint	θ, per degree	−0.021	0.020	0.31
SAR-only joint	W	−3.2	1.8	0.09
SAR-only joint	θ × W	0.009	0.012	0.44
SAR-only joint	AK vs UT	+0.2	0.5	0.71
SAR-only joint	WV vs UT	+0.1	0.5	0.83
SAR-only joint	Model R²	0.04	—	—
Full-fusion width	Intercept	16.1	0.9	<0.001
Full-fusion width	W	−5.7	3.2	0.07
Full-fusion width	AK vs UT	+0.5	0.4	0.23
Full-fusion width	WV vs UT	+0.3	0.4	0.49
Full-fusion width	Model R²	0.09	—	—

CC BY

The outcome is absolute error in in/mi. θ is in degrees, and W = paved carriageway width/20 m is dimensionless. W coefficients are in/mi per unit W; the interaction coefficient is in/mi per degree per unit W. Utah is the reference state. Uppercase R^2 is the regression coefficient of determination. These are association diagnostics, not causal estimates.

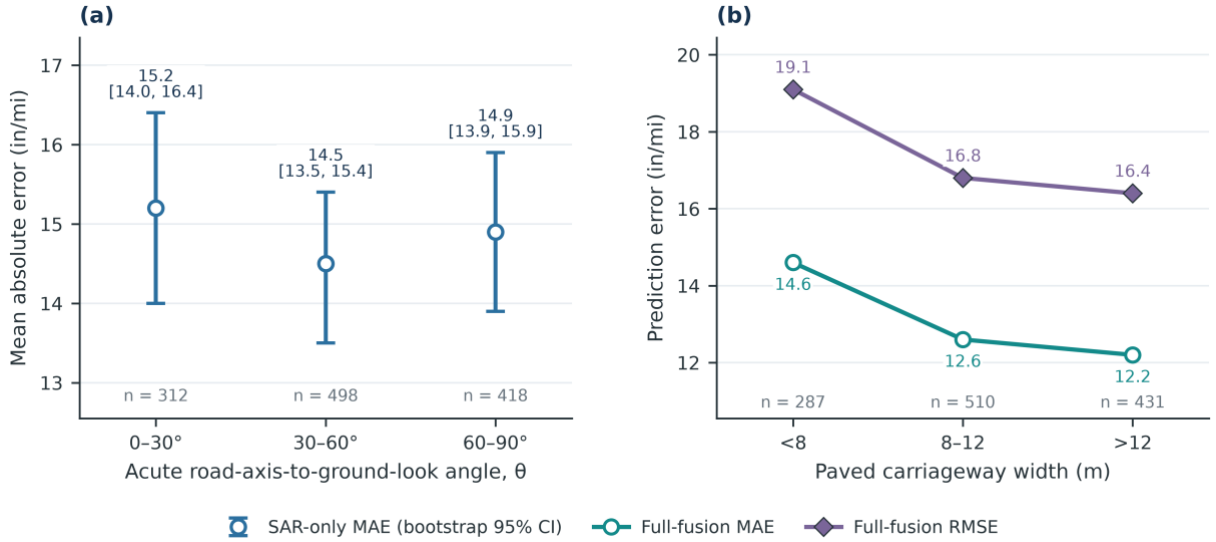


Figure A6: Geometric error diagnostics on the locked test cohort: (a) SAR-only MAE by acute road-axis-to-ground-look-angle bin, with bootstrap 95% confidence intervals; and (b) full-fusion S+P+C MAE and RMSE by paved-carriageway-width class.