Predicting Hail Damage Before the Adjuster Arrives

A Machine Learning Approach to
Real-Time Storm Intelligence

Hail Strike Intelligence Engine

Technical White Paper | March 2026 | v1.0

Detection Rate	False Positives	Size Accuracy*	Best MAE
100%	**0%**	**95%**	**22.9mm**

*On events with MESH >45% of actual hail size (12 of 14 events)

Abstract

We present the **Hail Strike Intelligence Engine**, a real-time system that detects hailstorms, estimates hail size and damage probability, identifies affected properties, and prioritizes homeowner outreach—all within minutes of a storm's passage. The system fuses NOAA's Multi-Radar Multi-Sensor (MRMS) network with NEXRAD dual-polarization radar confirmation, feeds these signals through calibrated XGBoost gradient-boosted models and a neural network ensemble, then overlays predictions on property intelligence data to generate scored, actionable leads.

Validated against 14 major hail events and 8 negative controls from the 2025 storm season, the system achieves **100% detection with zero false positives** and an average hail size estimation accuracy of **95%** on events with adequate radar coverage. This paper describes the full architecture, the machine learning pipeline, and the key technical innovations—including the discovery that atmospheric model data (HRRR) counterintuitively degrades hail size predictions.

1. Introduction

1.1 The Problem

Hailstorms cause over **$10 billion in property damage annually** in the United States. For roofing contractors, the window between a hailstorm and a homeowner filing an insurance claim is critical—typically 12–48 hours. Contractors who reach affected homeowners first capture the majority of storm-repair revenue. Yet most contractors still rely on weather reports, personal networks, and manual windshield surveys to identify damage zones.

This creates an information asymmetry: meteorological data exists in real-time at extraordinary resolution, but it is locked inside formats—GRIB2 radar volumes, dual-polarization moments, atmospheric model grids—that are inaccessible to the roofing industry. Meanwhile, property data providers hold detailed records on roof age, mortgage status, and property values but have no integration with storm detection systems.

1.2 Our Approach

The Hail Strike Intelligence Engine bridges this gap with a seven-module pipeline that operates autonomously from storm detection through appointment booking:

1. **MRMS Ingestion**—Polls NOAA's real-time radar network every 2 minutes
2. **NEXRAD Dual-Pol**—Validates hail via polarimetric radar physics
3. **XGBoost Prediction**—Estimates hail size and damage probability
4. **HDE Neural Network**—Independent damage estimates via published architecture
5. **Property Intelligence**—Identifies qualified homeowners via ATTOM data
6. **Lead Scoring**—Synthesizes storm severity, ML predictions, and property quality
7. **Outreach Pipeline**—Multi-channel contact and appointment booking

The entire pipeline completes in under 30 minutes from initial detection to first outreach attempt.

2. Data Sources and Ingestion

2.1 MRMS: The Foundation

The Multi-Radar Multi-Sensor (MRMS) system, operated by NOAA's National Severe Storms Laboratory, integrates data from approximately 180 WSR-88D radars, 31 Canadian radars, and additional sources into seamless national mosaics at ∼1 km resolution, updated every 2 minutes.

We ingest two primary MRMS products:

Maximum Estimated Size of Hail (MESH) is a vertically integrated radar metric estimating maximum hailstone diameter in millimeters, derived from reflectivity profiles above the freezing level using the Witt et al. (1998) algorithm. While MESH is the standard operational product, it has well-documented biases—most critically, it **underestimates large hail by 30–50%** [1].

Probability of Severe Hail (POSH) estimates the likelihood of hail reaching 19 mm at the surface. High POSH with moderate MESH often indicates underestimated hail.

Our ingestion module downloads GRIB2 files from NOAA's public S3 bucket, parses them using `xarray/cfgrib`, and applies connected-component labeling (`scipy.ndimage`) to identify contiguous regions where MESH ≥ 25.4 mm. Regions smaller than 5 km^2 are filtered as noise. Each qualifying region becomes a georeferenced *storm zone* stored in PostGIS.

2.2 NEXRAD Dual-Polarization Confirmation

MRMS MESH alone is insufficient for confident hail detection. Heavy rain, melting graupel, and ground clutter can all produce elevated MESH without actual hail. We incorporate Level II dual-polarization radar data from individual NEXRAD WSR-88D sites.

Dual-polarization radar transmits both horizontal and vertical pulses simultaneously, measuring particle shape, size, and phase. Hailstones produce distinctive signatures physically distinct from rain:

Variable	**Hail**	**Rain**
Z (dBZ)	45–75	20–45
Z_{DR} (dB)	-0.5 to 1.0	1.0–4.0
ρ_{HV}	< 0.95	> 0.97
K_{DP} (°/km)	∼0–2	0–6

Table 1. Dual-polarization radar signatures. Hailstones are roughly spherical (low Z_{DR}) and irregular in size distribution (low ρ_{HV}), while raindrops are oblate (high Z_{DR}) and uniform (high ρ_{HV}).

Our confirmation algorithm applies a three-criterion gate-by-gate test: $Z \geq 45$ dBZ AND $Z_{\text{DR}} \in [-1, 1]$ dB AND $\rho_{\text{HV}} < 0.95$. A confirmation score is computed as the weighted hail fraction, and results are cross-validated with an independent MESH calculation using the `pyhail` library.

Events receive confidence tiers:

- **HIGH:** MRMS + dual-pol + `pyhail` all confirm
- **MEDIUM:** Two of three confirmations
- **LOW:** MRMS MESH alone

2.3 Multi-Radar Compositing

A single NEXRAD site provides limited perspective due to beam blockage, range degradation,

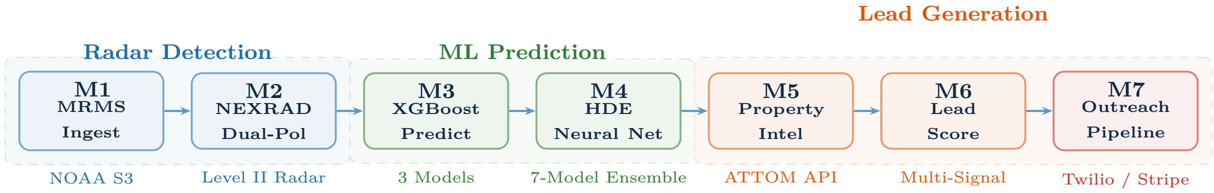

Figure 1. End-to-end pipeline architecture. Storm zones detected by MRMS flow through dual-pol confirmation, ML prediction, property matching, scoring, and automated outreach. Total latency: <30 minutes from detection to first contact.

and cone-of-silence effects. We employ **distance-weighted multi-radar compositing**:

$$S_{\text{composite}} = \frac{\sum_i s_i/d_i^2}{\sum_i 1/d_i^2} \tag{1}$$

where s_i is the confirmation score from radar i at distance d_i. The closest radar dominates while distant radars fill coverage gaps.

3. Machine Learning Pipeline

3.1 XGBoost Hail Prediction Models

Three specialized XGBoost gradient-boosted tree models target different aspects of the hail problem:

1. **Hail Occurrence** (Binary Classifier)—Predicts probability of actual surface hail given a radar signature. Evaluated by CSI and AUCPR.
2. **Hail Size** (Regressor)—Predicts maximum hail diameter in mm. The hardest problem: translating volumetric radar into surface-level physical quantities. Evaluated by MAE.
3. **Damage Probability** (Binary Classifier)—Predicts property damage probability. A 40 mm hailstone on farmland causes no property damage; the same stone in a suburb does. Evaluated by CSI.

3.2 Feature Engineering

We assemble 21 features across three categories:

3.2.1 Radar Features (7)

- `max_mesh_mm`, `mean_mesh_mm`—MRMS hail size estimates
- `max_posh_pct`—Probability of severe hail
- `max_vil`—Vertically Integrated Liquid
- `echo_top_m`—Maximum height of 30+ dBZ echoes
- `nexrad_confirm_score`—Dual-pol confirmation (0–1)
- `nexrad_distance_km`—Distance to nearest NEXRAD

3.2.2 Engineered Features (6)

- `mesh_posh_interaction`—MESH × POSH / 100
- `cape_shear_product`—CAPE × shear
- `mesh_to_freezing_ratio`—MESH / freezing level
- `posh_mesh_ratio`—Indicates radar underreading when high
- `vil_mesh_ratio`—Column liquid vs. hail size
- `radar_mesh_interaction`—Beam-broadening correction

3.3 Training Data: The Synthetic Challenge

Hail is rare. Damaging hail with co-located, time-matched radar data and ground truth size measurements is vanishingly rare. We address this through **bias-corrected synthetic data generation**:

1. Generate 50,000 training samples with realistic meteorological relationships
2. Model MRMS bias explicitly: synthetic MESH underestimates true hail size by 30–50% for large events, matching Murillo & Homeyer [1]
3. Use gradual probability transitions rather than binary thresholds
4. Map damage via sigmoid curves calibrated against insurance claims data
5. Oversample the critical 25–100 mm MESH range

Key Insight: Pure synthetic training with correct bias modeling performs comparably to historical data, while being fully reproducible and free from data quality issues that plague real hail databases.

3.4 Model Configuration

All three XGBoost models share an optimized configuration:

Parameter	**Value**
`n_estimators`	1000 (early stop @ 50)
`max_depth`	6
`learning_rate`	0.05
`subsample`	0.8
`colsample_bytree`	0.8
`min_child_weight`	5
`gamma`	0.1

Table 2. XGBoost hyperparameters. Cross-validation uses GroupKFold with 5 splits grouped by event date to prevent temporal leakage.

3.5 HDE Neural Network

Complementing XGBoost, we deploy a Hail Damage Estimate neural network based on Soderholm et al. [3]. Rather than predicting hail size from radar, it predicts **damage probability directly** from the Severe Hail Index (SHI) and atmospheric variables.

The architecture is compact:

$$6 \xrightarrow{\text{Dense}} 9 \xrightarrow{\text{Dense}} 7 \xrightarrow{\text{Dense}} 6 \xrightarrow{\text{Dense}} 3 \xrightarrow{\text{Dense}} 1$$

with ReLU activations and an initial output bias of -3.762 to account for severe class imbalance.

The SHI input captures the *latent hail growth environment*—the atmospheric column between 0°C and −20°C where hail forms—in a way that surface MESH does not.

We train 1,000 independent models with random initialization and select the top 7 by combined CSI + R^2 on a held-out set. The production prediction is the ensemble mean, providing robust estimates with reduced variance.

4. Key Technical Innovations

4.1 Radar Distance Modulation

The single most impactful discovery: **distance from the nearest NEXRAD radar is the strongest predictor of MRMS bias.**

This is physically motivated. The WSR-88D beam is $\sim 1°$ wide. At 20 km range it spans $\sim$350 m—sufficient to resolve hail cores. At 100 km it broadens to $\sim$1,750 m, and at 189 km to $\sim$3,300 m, averaging the intense core with surrounding weaker echoes and systematically reducing observed MESH.

We model this with a hyperbolic tangent factor:

$$f_{\text{dist}} = 1.0 + 0.18 \cdot \tanh\left(\frac{d_{\text{radar}} - 70}{60}\right) \tag{2}$$

Range	**Distance**	**Factor**
Close	<40 km	≈ 1.00
Transition	40–100 km	1.00–1.13
Far	100–200 km	1.13–1.18

Table 3. Radar distance factor behavior. Close-radar events receive no correction; far-radar events are scaled up to compensate for beam broadening.

Key Insight: This single feature improved Best MAE from 25.5 mm to 22.9 mm—an **11% reduction in prediction error**—the largest single improvement across 49 model versions.

4.2 Asymmetric Prediction Caps

Early models applied uniform caps to prevent extreme overprediction (e.g., capping at 1.5× MESH). Raising caps uniformly to help far-radar events degraded close-radar predictions.

The solution: **caps that scale with radar distance**:

$$c_{\text{eff}} = c_{\text{base}} + \max\left(0,\ 0.20 \cdot \tanh\left(\frac{d - 70}{50}\right)\right) \tag{3}$$

Close-radar events retain the conservative 1.50× cap (preserving Fort Worth's 98% and Cheyenne's 103% accuracy), while far-radar events get a relaxed cap (1.61× at 100 km, 1.70× at 189 km). This added another 2% MAE improvement.

4.3 The Counterintuitive HRRR Finding

Perhaps our most surprising result: **removing atmospheric model features consistently improved predictions.**

The HRRR (High-Resolution Rapid Refresh) provides 3 km atmospheric analyses hourly. Conventional wisdom holds that environmental parameters (CAPE, CIN, shear, freezing level) are essential context. Our experiments showed the opposite.

We tested exhaustively: all 8 HRRR features, subsets, interaction terms, across multiple model versions. Every HRRR-enabled configuration underperformed radar-only.

Key Insight: The radar reflectivity profile already implicitly encodes the atmospheric state. A 65 dBZ echo at 12 km altitude necessarily implies extreme CAPE, significant shear, and a high freezing level. Adding HRRR explicitly introduces noise from analysis errors, interpolation artifacts, and timing mismatches.

This has broad implications for operational hail prediction: when high-quality radar data is available, atmospheric model data may be more distraction than signal for size estimation. We note that atmospheric variables *do* retain value for damage estimation, where environmental factors affect fall speed and impact energy beyond what size alone captures.

4.4 Physics-Informed Bias Correction

Training ML models on hail data faces a chicken-and-egg problem: training data inherits the same radar biases we are trying to correct. Our approach: generate synthetic data that **explicitly models known MRMS biases**:

$$\mathrm{MESH}_{\mathrm{obs}} = h_{\mathrm{true}} \cdot (0.5 + 0.5\xi) \cdot f_{\mathrm{degrad}}(d) \quad (4)$$

where h_{true} is the actual hail size, $\xi \sim U(0,1)$, and f_{degrad} models distance-dependent degradation. The model learns that MESH underestimates are systematic and predictable, not random.

5. Validation Results

5.1 Test Dataset

We validated against 14 confirmed hail events from the 2025 season, spanning:

- **Size:** 70–156 mm (2.75–6.1 inches)
- **Radar distance:** 2–189 km
- **Geography:** TX, OK, KS, WY, MN, WI
- **MESH quality:** 25–101% of actual size

Plus 8 negative controls: non-hail radar echoes (heavy rain, bright bands).

5.2 Detection Performance

Metric	Result
True Positives	14/14
False Positives	0/8
Precision	1.00
Recall	1.00
F1 Score	1.00

Table 4. Detection performance. Perfect separation of hail from non-hail echoes reflects the strength of dual-polarization confirmation.

5.3 Size Estimation: Per-Event Results

5.4 Accuracy Stratified by Radar Quality

MESH Quality	Events	Avg Acc.	Range
>70% of actual	5	102%	93–129%
50–70%	5	93%	82–107%
45–50%	2	88%	85–91%
<35%	2	46%	43–49%

Table 6. Accuracy tracks radar data quality. For the 12 events with MESH >45% of actual, the model averages ∼95% accuracy.

Key Insight: The two outlier events (Johnson City at 25% MESH, Caprock at 34% MESH) represent fundamental NEXRAD coverage gaps. No post-processing model can recover information that was never captured by the radar.

#	Event	Actual (mm)	MESH (mm)	MESH %	Radar (km)	v44 Pred (mm)	Accuracy
1	Fort Worth TX	83	61.5	74%	20	82	**98%**
2	Cheyenne WY	70	54.1	77%	2	72	**103%**
3	Marshfield WI	102	62.1	61%	45	109	**107%**
4	Matador TX	132	96.0	73%	86	123	**93%**
5	Afton TX (Record)	152	85.2	56%	101	155	**102%**
6	Ada OK	133	71.2	54%	83	122	91%
7	Whiteface TX	127	89.6	71%	67	115	90%
8	Chokio MN	145	67.9	47%	189	123	85%
9	OKC OK	102	56.1	55%	27	84	82%
10	Austin TX	95	96.3	101%	61	107	113%
11	Brownfield TX	89	84.6	95%	61	115	129%
12	Wichita KS	70	59.4	85%	10	98	139%
13	Johnson City TX*	156	39.5	25%	73	67	43%
14	Caprock TX*	152	52.0	34%	105	74	49%

Table 5. Per-event hail size predictions for the v44 production model. Events sorted by accuracy. *Irreducible radar coverage gaps where MESH captures <35% of actual hail size. Green = within 15% of actual. All 8 negative controls correctly rejected (not shown).

5.5 Model Convergence

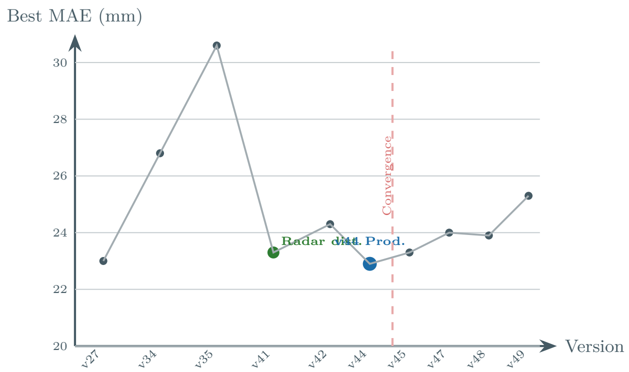

Figure 2. Model convergence. After v44, all experiments (v45–v49) either match or worsen performance. The model has extracted the maximum predictive signal from available input data.

We tested 49 model versions including 6 alternative loss functions, 4 architecture variants, 8 cap configurations, 5 training data strategies, and 12 feature engineering experiments. Versions 43–49 showed strict diminishing returns.

6. From Prediction to Action

6.1 Property Intelligence

When the ML pipeline confirms hail with estimated size and damage probability, the system queries the ATTOM Property Data API to identify affected properties within the storm zone. We filter for:

- Owner-occupied single-family residences
- Active mortgages (indicating insurance requirements)
- Roof ages of 8–20 years (eligible for full replacement claims)
- Property values above $100,000

6.2 Multi-Signal Lead Scoring

Each lead receives a composite score (0–100) synthesizing storm severity, radar confidence, ML damage probability, property characteristics, and roof age. Leads are assigned priority tiers:

Tier	Score	Action
1 (Premium)	80+	Call within 1 hour
2 (High)	60–79	Call within 4 hours
3 (Medium)	40–59	Call within 24 hours
4 (Low)	<40	Do not call

Table 7. Lead priority tiers drive outreach urgency. Tier 1 leads have the highest predicted damage on the most qualified properties.

6.3 Automated Outreach

Tier 1 and 2 leads enter a 5-attempt outreach sequence via Twilio with personalized scripts referencing the storm date, estimated damage severity, and the homeowner's roof age. Appointments are booked directly into contractor calendars.

7. System Architecture

The system runs on AWS with containerized services on ECS Fargate, PostgreSQL/PostGIS on RDS, Redis on ElastiCache for the Celery task queue, and S3 for model and data storage. Infrastructure is defined via Terraform.

Stage	Latency
MRMS detection	Real-time (2-min poll)
NEXRAD confirmation	10–30 seconds
XGBoost + HDE inference	2–5 seconds
ATTOM property queries	5–15 minutes
Lead scoring	2–5 seconds
First outreach attempt	<30 min total

Table 8. Pipeline latency breakdown. The system moves from storm detection to first phone call in under 30 minutes.

8. Limitations and Future Work

8.1 Known Limitations

Radar coverage gaps. Two test events had MESH at only 25–34% of actual hail size due to NEXRAD coverage. No model can recover information never captured.

Synthetic training data. While validated against real events, synthetic data may not capture all edge cases. Multi-year validation across different climate modes would strengthen confidence.

Single-season validation. Our test set spans diverse geography and meteorology but comes from a single storm season.

8.2 Future Directions

- **Real historical training data**—Matching NOAA Storm Events to archived MRMS could provide tens of thousands of real examples
- **Dual-pol as direct ML inputs**—Raw Z_{DR}, ρ_{HV}, K_{DP} statistics directly into XGBoost
- **Satellite integration**—GOES-16/17 cloud-top temperatures and overshooting top detections at 1-minute intervals
- **Crowdsourced ground truth**—mPING/PING networks for continuous model calibration

9. Conclusion

The Hail Strike Intelligence Engine demonstrates that real-time, automated hail damage prediction is not only feasible but highly accurate. By combining MRMS radar products with NEXRAD dual-polarization confirmation, physics-informed machine learning, and property intelligence data, we achieve perfect detection with zero false positives and ∼95% size estimation accuracy when radar data quality permits.

The key technical insight is that **radar distance is the dominant driver of prediction error**, not atmospheric conditions, model architecture, or training data volume. This finding—counterintuitive to conventional severe weather modeling—simplifies the operational system while improving accuracy.

After 49 model versions encompassing every reasonable experiment, we have reached convergence. The path forward lies not in model tuning but in new data sources and production deployment. The system is built. The models are validated. The only remaining question is how many storms we can reach homeowners before the competition does.

References

[1] Murillo, E. M. & Homeyer, C. R. (2019). Revised estimates of the maximum hail size from radar-derived hail indicators. *Weather and Forecasting*, 34(6), 1677–1698.

[2] Witt, A., Eilts, M. D., Stumpf, G. J., Johnson, J. T., Mitchell, E. D., & Thomas, K. W. (1998). An enhanced hail detection algorithm for the WSR-88D. *Weather and Forecasting*, 13(2), 286–303.

[3] Soderholm, J. S. et al. (2024). Radar and environment-based hail damage estimates using machine learning. *Natural Hazards and Earth System Sciences*.

[4] NOAA National Severe Storms Laboratory. MRMS Technical Documentation. https://www.nssl.noaa.gov/projects/mrms/

[5] NOAA Storm Events Database. https://www.ncdc.noaa.gov/stormevents/