AI3011 · MACHINE LEARNING & PATTERN RECOGNITION · PLAKSHA · SPRING 2026

Forecasting where Punjab **burns** next.

01

A weekly grid-level forecasting framework using satellite, climate, and policy data · 7 km resolution, one week ahead, district-ready.

TEAM

Tanush Kalhan · Aditt Singh · Adityapratap Singh Parmar · Arnav Jain

What happens every October **in Punjab**.

WHY FARMERS BURN

Rice harvest ends late October. Wheat sowing must begin within 2–3 weeks. Clearing residue manually costs ₹3,000–5,000 per acre. Burning costs nothing · for small farmers with no margin, there is no real choice.

THE GOVERNMENT MADE IT WORSE

The 2009 Punjab Groundwater Preservation Act delayed rice transplanting to conserve water. It pushed peak burning into November · when boundary layers are shallower and the same smoke causes dramatically worse air quality. A conservation policy accidentally created a pollution catastrophe.

WHERE IT BURNS · EVERY YEAR

Sangrur. Barnala. Mansa. Ludhiana. Patiala.
The same districts. The same fields. The same weeks. Two decades of satellite record confirm it never changes.

7–8M metric tons burned every Oct–Nov

90% of Punjab's harvested area burns

480M people affected across the Indo-Gangetic Plain

500 μg/m³ PM 2.5 at Patiala during peak weeks (WHO limit 25)

42% rise in real fire counts 2003–2016 (Harvard)

3.4× actual emissions vs what global inventories report

THE ENFORCEMENT FAILURE

Punjab Pollution Control Board bans burning. National Green Tribunal issues directives. Fines exist. Satellite monitoring exists.

Still Burning continues at scale · because the entire enforcement system is reactive by design . Satellites confirm fires already burning. By the time a team is dispatched, the field is ash.

THE HEALTH COST

Six Punjab districts alone account for 40% of India's entire agricultural-fire health burden.

A 10% spike in hospital admissions in Punjab's rice-wheat belt within 20–25 days of peak burning.

> *Enforcement is reactive. Forecasting must be **proactive**.*

Source: Punjab Agricultural University; Harvard Atmospheric Chemistry Modeling Group; IIT Kanpur & SAFAR; State of Global Air 2025.

3 Figures that justify operational forecasting, not just retrospective analysis.

AIR QUALITY · DELHI NCR

40%+

of Delhi's peak-season PM 2.5 attributed to Punjab stubble burning

ECONOMIC LOSS · INDIA

$30B

annual cost from stubble-burning health impacts and lost productivity

MORTALITY · INDIA · 2023

2.0M

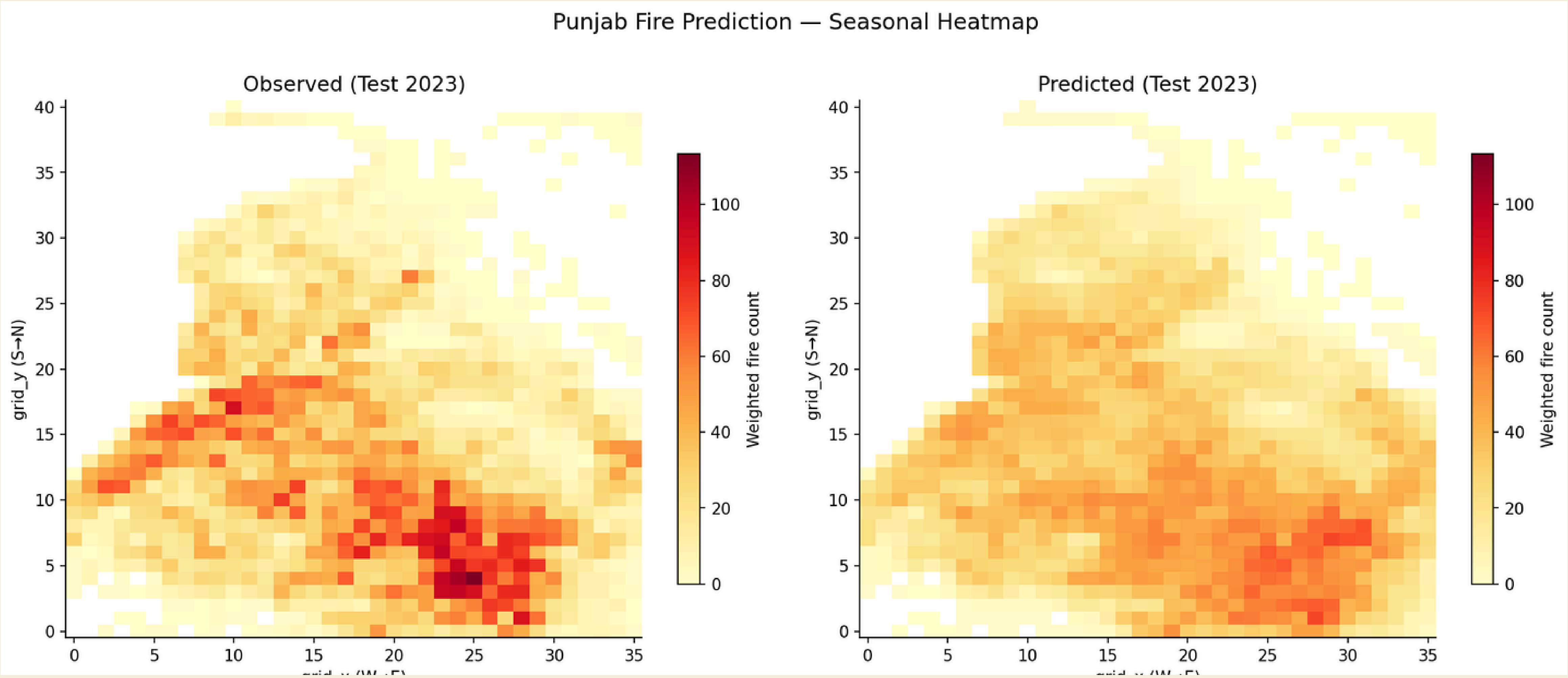
air pollution-related deaths in India during 2023

ADDITIONAL INFO:

- **~35 million kg of particulate matter released per burning season (IIT Kanpur)**
- **AQI spikes to 400–500 in Delhi NCR during peak burning (Oct–Nov)**
- **9,000–11,000 fire events detected per season via satellite (NASA VIIRS/MODIS)**
- **Burning window is narrow: ~21 days between rice harvest and wheat sowing — that's the forecasting hook**

Source: IIT Kanpur & SAFAR (PM 2.5 attribution); IFPRI (economic cost); State of Global Air 2025 (mortality).

1,036 cells. 6 seasons. One **burning belt.**



RESOLUTION
7 km × 7 km
Fine enough for district-level action, coarse enough to align with FIRMS detection footprints.

ACTIVE CELLS
1,036 of 1,600
Cells with at least one fire detection across 2018–2023 · the agricultural Malwa belt.

TEMPORAL COVERAGE
9 weeks × 6 years
October week 1 through November week 4. 56,160 grid-week observations total.

Source: Authors' analysis using NASA FIRMS MODIS Collection 6.1 and VIIRS NOAA-20 Collection 2.

Building on prior work, extending it in seven ways.

DIMENSION	MOR & MOR (2023)	THIS WORK
Region	Punjab + Haryana	Punjab (focused)
Method	ConvLSTM, ConvGRU	XGBoost-Tweedie + ConvLSTM
Features	5 (NDVI, T, wind, pressure, cloud)	**70+** (+ soil moisture, VPD, lag, anomaly, policy)
Forecast horizon	1–3 days	**1 week** (+ 2 wk, 4 wk in ablation)
Validation	Not specified	**Strict temporal split** — train 2018–21, **validate 2022**, test 2023
Best correlation	≈ 0.80 (day 1, Pearson)	**0.80** (week 1, Spearman, ConvLSTM)
Per-family ablation	No	**Yes** · weather is biggest single lift
Leakage audit	Not mentioned	**8 of 8 checks** passed

Same headline correlation as the leading prior work · at a seven-times longer forecast horizon, with richer features and stricter validation.

Source: Mor S., Mor R. S. (2023). Modeling Earth Systems and Environment 9, 4267–4276. DOI: 10.1007/s40808-023-01831-1.

https://link.springer.com/article/10.1007/s40808-023-01831-1

02 Data

Four streams, six seasons, seventy-four features.

Four sources, one **unified** table.

STREAM 01 · SATELLITE FIRE DETECTION

NASA FIRMS

Active-fire detections from MODIS Collection 6.1 (Aqua + Terra) and VIIRS NOAA-20 Collection 2. 1.04 million+ detections across the study period, with brightness, FRP, and confidence scores.

1.04M+

DETECTIONS · DAILY

STREAM 02 · VEGETATION

MODIS MOD13Q1 (NDVI / EVI)

16-day vegetation index composites at 250 m resolution. Provides NDVI, EVI, NDVI velocity (week-on-week change), and NDVI anomaly against the training-period baseline.

5 features

250 M BASE

STREAM 03 · CLIMATE REANALYSIS

ERA5-Land Daily Statistics

Air temperature, soil moisture, soil temperature, wind components, surface pressure, dewpoint, vapor pressure deficit. Plus derived: dry-streak counter, fire weather index. Lagged and anomaly variants included.

48 features

0.1° ≈ 9 KM

STREAM 04 · POLICY VARIABLES

Punjab Policy Indicators

Hand-coded with primary source citations: NGT enforcement level, Super Seeder availability, ex-gratia announcements, election years, CRM central funds (cumulative), MSP for paddy.

8 features

YEARLY VARIATION

FOUR STREAMS UNIFIED AT GRID × WEEK × YEAR RESOLUTION

Source: NASA FIRMS; NASA MODIS MOD13Q1; Copernicus ERA5-Land; PIB India, CEEW, Punjab Government (policy).

From four signals to **seventy-four** features.

FAMILY 01

Base FIRMS lag features (8)

- fire_count_last_week
- same_week_last_year
- 3yr_avg
- neighbor_fires (lag-1)
- avg_frp_last_week
- avg_brightness_last_week
- night_fire_pct_last_week
- week_of_season

FAMILY 02

NDVI / EVI (5)

- NDVI
- EVI
- NDVI_velocity
- NDVI_baseline_train
- NDVI_anomaly *(train-only baseline)*

FAMILY 03

Weather (48)

- 7 raw variables
- mean / max / min / std
- lag-1 versions
- anomalies (train-only)
- dry_streak counter
- fire_weather_index

FAMILY 04

Policy (8)

- super_seeder_available
- ngt_enforcement_level
- ex_gratia_announced
- election_year
- crm_funds_cumulative_cr
- msp_paddy_common

FINAL MASTER TABLE

56,160 rows × 74 columns

1,036 active grids × 9 burning weeks × 6 seasons (2018–2023). Strict temporal split: train on 2018–21, validate 2022, test 2023.

FOUR FEATURE FAMILIES, ENGINEERED WITH LAG, ANOMALY, AND COMPOSITE OPERATORS

Source: Project pipeline · era5_policy_pipeline.ipynb and modeling_master.ipynb.

0.997 PR-AUC was too good. Here is what we caught.

0.997 → 0.894 → 0.869

V1 · TOO GOOD TO BE TRUE

Initial 28K-row table with NDVI merged from fire-positive source only

Boolean "NDVI-is-NaN" perfectly predicted zero-fire rows
PR-AUC 0.997 came from data leakage, not the model

V3 · AFTER LEAKAGE FIX

Rebuilt NDVI from raw MOD13Q1 rasters covering all 1,036 cells

Cartesian grid × week × year expansion to 56,160 rows

8-check leakage audit including shuffled-target control passed

FINAL TUNED + LARGER FEATURE SET

Added 48 ERA5 weather features + 8 Punjab policy variables

Optuna-tuned XGBoost-Tweedie across 100 Bayesian trials

PR-AUC dropped slightly; MAE dropped 27% on count regression

NDVI was NaN *only* on zero-fire rows in the v1 master table · a merge artifact. A single boolean 'NDVI-is-NaN' flag alone gave PR-AUC 0.993. **We rebuilt NDVI from raw MOD13Q1 rasters covering all 1,040 cells.**

PASSED

8 audit checks

01 · Cross-year boundary

02 · Intensity feature lag

03 · Cartesian join sanity

04 · NDVI baseline train-only 05 · Spatial overlap acknowledged

06 · Ablation behaviour sensible

07 Shuffled-target = base rate 08 · No grid_id encoded

All 8 passed

Source: Internal audit log; final_complete.ipynb leakage audit cells (8 of 8 passed).

Strict temporal split. Five models.

TRAIN	TRAIN	TRAIN	TRAIN	VAL	TEST
2018	2019	2020	2021	2022	2023

FIVE MODELS

Trained under identical split

01 · Persistence baseline (uses last week's fire count)

02 · Logistic Regression

03 · Random Forest

04 · XGBoost-Tweedie (tuned via Optuna)

05 · Two-stage Hurdle (XGB binary × XGB Gamma)

STRETCH MODEL

Justified by spatial autocorrelation

ConvLSTM with 4-week input window predicting next-week fire counts at full spatial grid. Motivated by **Moran's I = 0.54** on XGBoost residuals ($p < 0.001$). Implemented in PyTorch, trained on GPU.

Evaluation PR-AUC · MAE · Spearman · ROC-AUC · Precision@50 · Stratified MAE · Lead-time decay · Counterfactual · External validation

Source: modeling_master.ipynb; Optuna 100 trials; reference Mor & Mor (2023).

03 Results

What the model learned, what it cannot.

11 · FIVE-MODEL HEAD-TO-HEAD

Two **different** winners on the same test set.

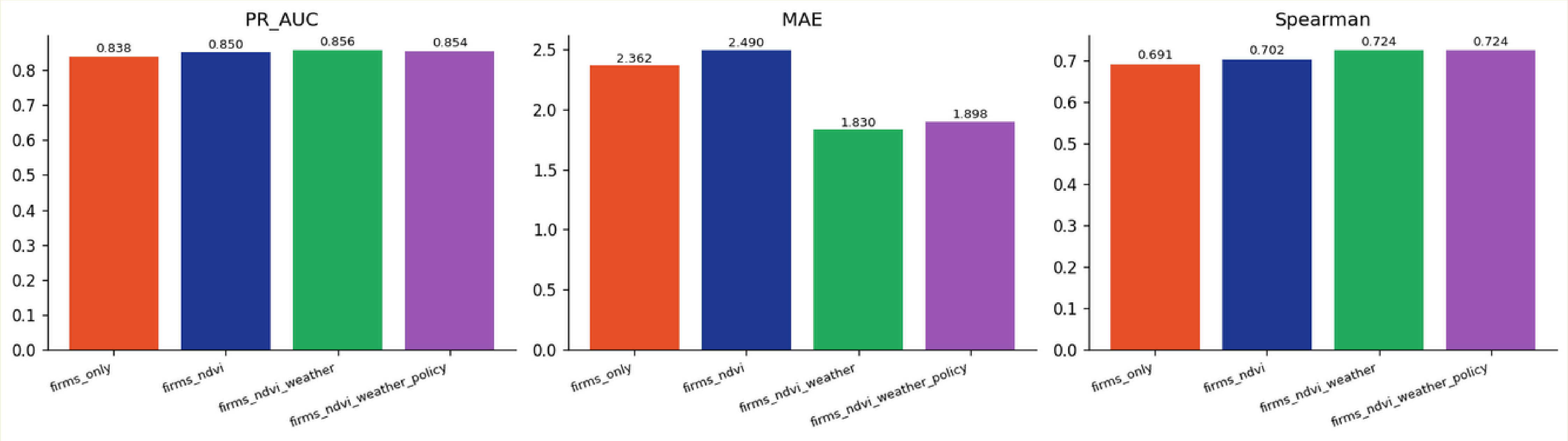
MODEL	PR-AUC	MAE	SPEARMAN	ROC-AUC	VERDICT
Persistence (baseline)	0.750	2.57	0.60	0.75	BASELINE
Logistic Regression	0.900	·	0.73	0.92	Strong rank
Random Forest	0.889	2.87	0.86	0.91	Strong rank
XGBoost-Tweedie (tuned)	**0.869**	**1.88**	**0.74**	**0.88**	**HEADLINE**
ConvLSTM (stretch)	0.824	**0.43**	**0.80**	0.87	**Best count regression**

XGBoost wins on ranking · ConvLSTM wins on count regression. Different tools for different operational uses.

Source: modeling_master.ipynb final results; Optuna best hyperparameters.

12 · ABLATION LADDER

Weather dropped MAE by **27 percent**.



01 · FIRMS-only	PR-AUC 0.838 · MAE 2.36
02 · + NDVI	PR-AUC 0.850 · MAE 2.49
03 · + Weather	PR-AUC 0.856 · MAE **1.83**
04 · + Policy (full)	PR-AUC 0.854 · MAE 1.90

← +27 percent MAE improvement

We pre-registered the hypothesis that weather would help count regression more than ranking. It did. Exactly.

Source: Ablation ladder, modeling_master.ipynb Phase 3.

WEATHER WAS THE BIGGEST SINGLE PREDICTIVE LIFT, MOSTLY ON COUNT REGRESSION

13 · FINAL TUNED MODEL · TEST 2023

TEST-SET PR-AUC

0.869

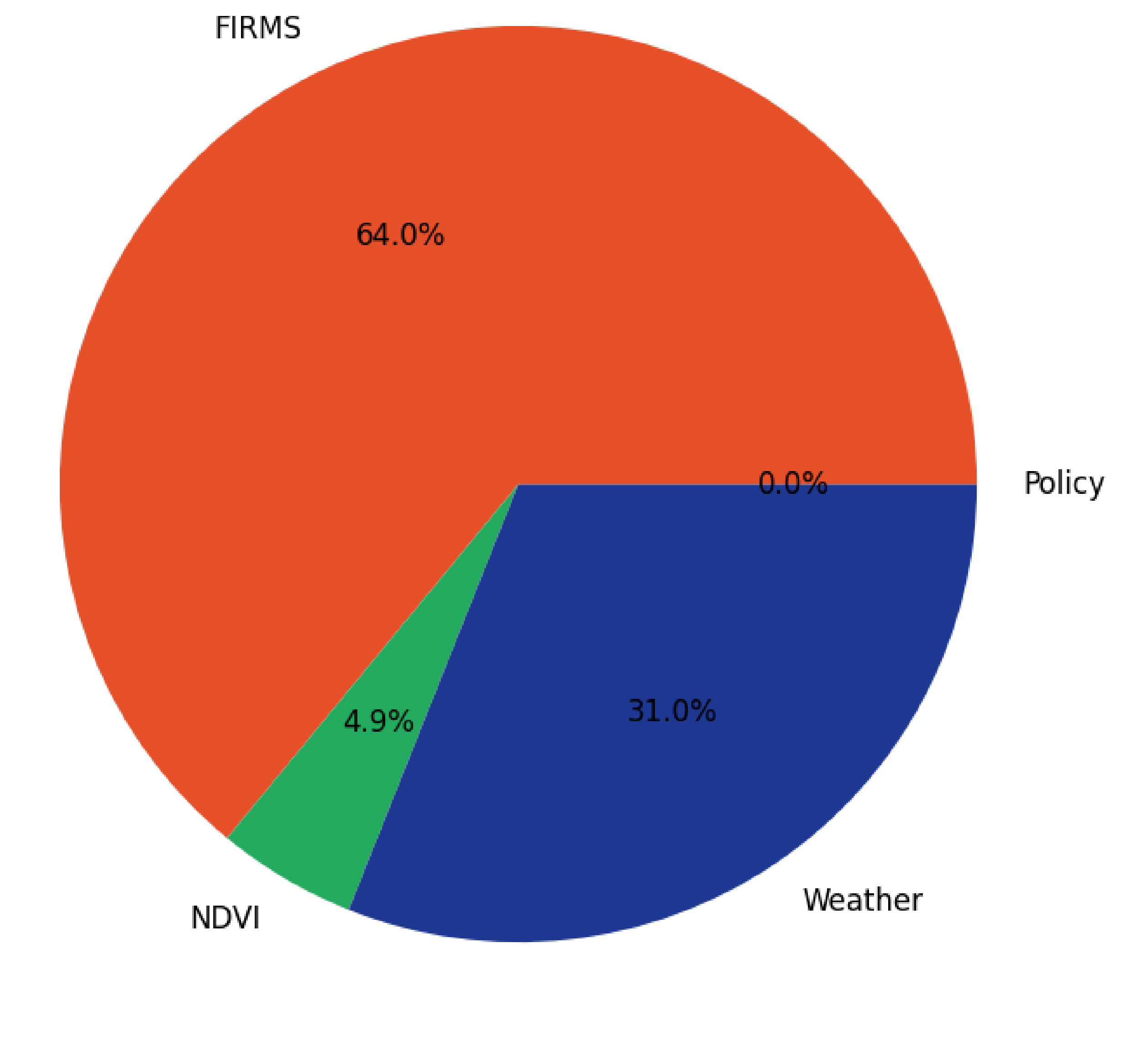
MEAN ABSOLUTE ERROR	SPEARMAN CORRELATION	PRECISION @ TOP 50
1.88	0.74	0.98
21% reduction vs persistence baseline	rank correlation on test	top-50 highest-risk cells confirmed

XGBoost-Tweedie · Optuna-tuned (100 trials) · 8 of 8 leakage audit checks passed.

Source: modeling_master.ipynb final tuned XGBoost-Tweedie on 2023 held-out test set.

14 · WHAT THE MODEL RELIES ON

Fire history dominates. Weather **earns** its place.



64% — **Fire history**
Lag features, neighbor effects, 3-year averages

31% — **Weather**
VPD, soil moisture, temperature, fire weather index

5% — **NDVI / EVI**
Vegetation indices and anomalies

0% — **Policy (direct)**
Direct SHAP zero; counterfactual reveals indirect effects

Policy direct = 0%. Policy via counterfactual = −7.1% (Super Seeder). The model learned policy through correlated fire history.

Source: SHAP TreeExplainer on tuned XGBoost; 2,000-row test sample; polish_additions.ipynb.

15 · SPATIAL DIAGNOSTIC + CONVLSTM

Spatial structure remained. Deep learning found it.

DIAGNOSTIC · MORAN'S I

Spatial autocorrelation in residuals

0.54

p < 0.001 · 999 permutations · signed residual
0.71 on absolute residual

Strong positive autocorrelation means XGBoost errors cluster spatially. The model misses real spatial structure.

STRETCH MODEL · CONVLSTM

Captures the residual

0.43

MAE on test 2023
77% lower than XGBoost (1.88)

ConvLSTM with 4-week input window captures the residual spatial signal that Moran's I predicted was learnable.

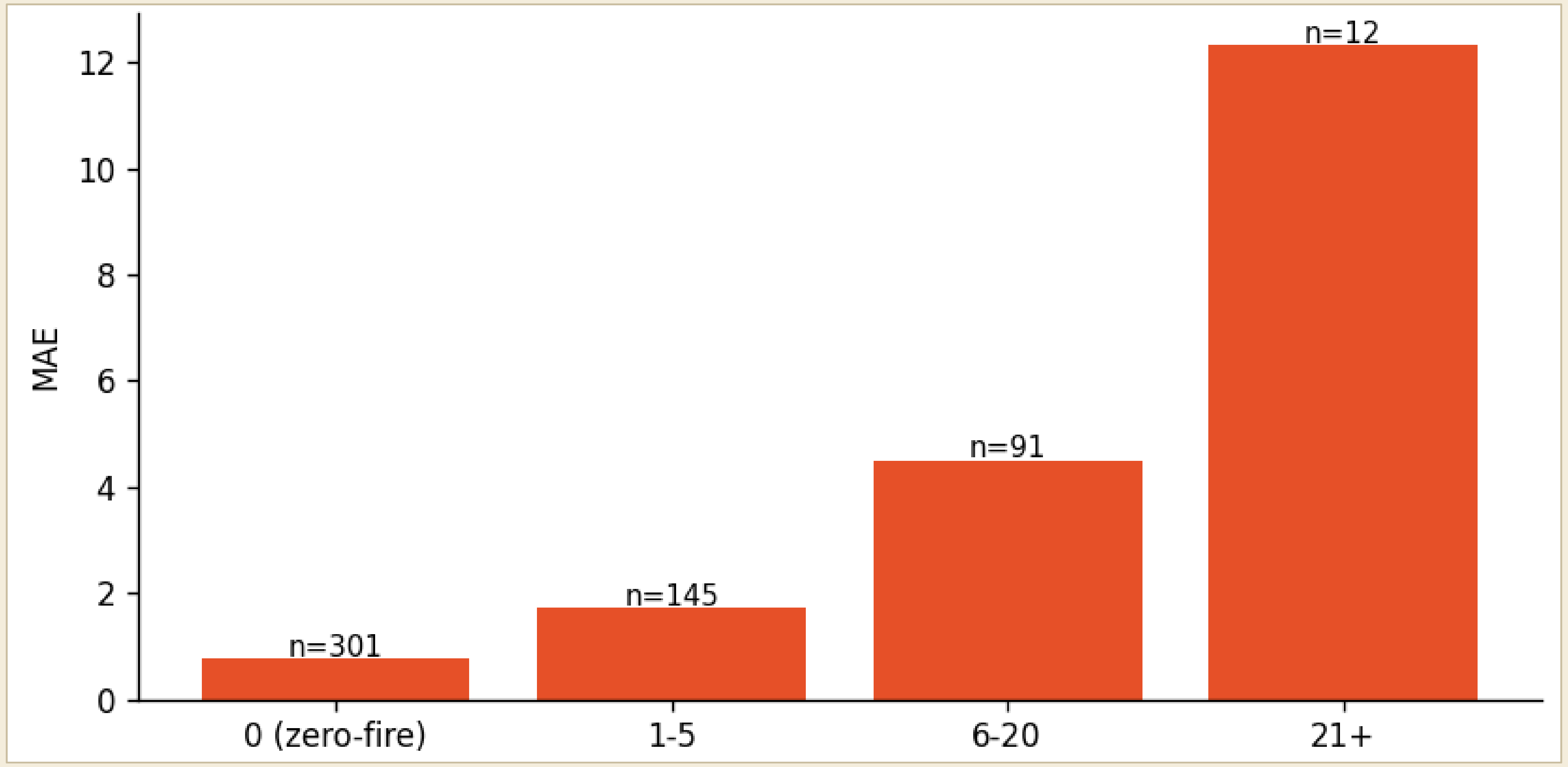
The diagnostic predicted it. Deep learning delivered.

SPATIAL AUTOCORRELATION CONFIRMED, CONVLSTM JUSTIFIED, RESEARCH ARC CLOSED

Source: Moran's I via libpysal / esda (999 permutations); ConvLSTM via PyTorch; polish_additions.ipynb.

16 · HONEST WEAKNESSES

Heavy-burn cells remain the model’s weakest regime.



MAE BY FIRE-COUNT BUCKET (TEST 2023)

BUCKET	N	MAE
y = 0 (zero fires)	301	0.21
0 < y ≤ 1	145	2.78
1 < y ≤ 5	91	3.53
5 < y ≤ 15	·	4.84
y > 15 (heavy-burn)	12	**12.3**

Only 12 cells in the heavy-burn bucket. Statistically thin.

Lead-time decay: t+1 PR-AUC 0.80 · t+2 0.84 · t+4 0.93*

**** t+4 anomaly likely seasonal alignment.***

Aggregate MAE 1.88 is driven by accurate zeros. Heavy-burn cells remain our biggest unsolved problem.

Source: Stratified analysis on test 2023, modeling_master.ipynb Phase 7.

17 · EXTERNAL VALIDATION + COUNTERFACTUAL

Validated against four monitoring stations, sensitive to **policy levers** we tested.

EXTERNAL VALIDATION · CPCB STATIONS

Predictions correlate with observed at station scale

0.83	0.89
PEARSON R	SPEARMAN R

Stations: Amritsar · Ludhiana · Patiala · Bathinda (50 km catchment radii).

Full PM 2.5 integration is deployment work.

COUNTERFACTUAL · POLICY SENSITIVITY

Indirect effects revealed by inference

−7.1%	Remove Super Seeder availability
−3.2%	Add ex-gratia compensation
−0.2%	Remove NGT enforcement

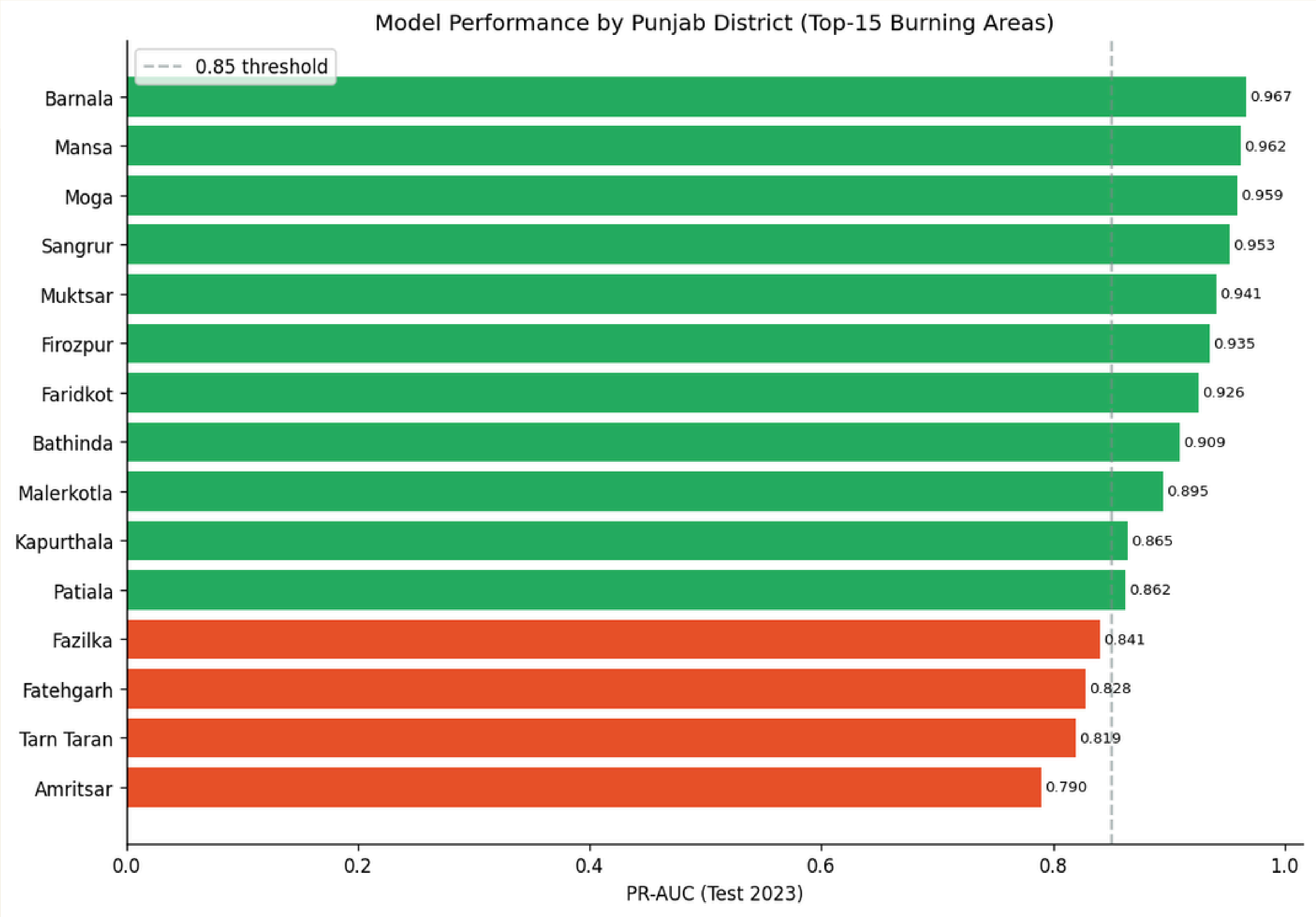
Sensitivity to policy levers, not causal claims.

Direct SHAP from policy was 0%. Counterfactual exposed indirect effects via correlated fire history.

Source: polish_additions.ipynb; CPCB station network; counterfactual via policy-aware model retraining.

18 · NAMED-DISTRICT PERFORMANCE

The model excels where Punjab **burns** hardest.



TOP BURNING DISTRICTS · TEST 2023

All five clear PR-AUC 0.95

DISTRICT	PR-AUC	REGION
Barnala	0.967	Malwa belt
Moga	0.959	Malwa belt
Mansa	0.962	Malwa belt
Sangrur	0.953	Malwa belt
Ferozepur	0.935	Malwa belt
Bathinda	0.909	Malwa belt

Northern Punjab (Amritsar 0.79, Tarn Taran 0.82) weaker · less signal, different cropping patterns.

Source: polish_additions.ipynb per-district analysis; centroid-based mapping to 23 Punjab districts.

19 · LIMITATIONS · FUTURE WORK · DEPLOYMENT

From research notebook to **district risk dashboard**.

LIMITATIONS

Honest weaknesses

01 · Heavy-burn cells (>15 fires): MAE 12.3

02 · Northern Punjab weaker (PR-AUC 0.79–0.82)

03 · Policy correlational, not causal

04 · Conformal intervals deferred 05 · Single test year · no multi-year holdout

FUTURE WORK

Five next steps

01 · Real-time FIRMS pipeline integration

02· PPCB Action Taken Report feedback loop

03 · 2024 + 2025 data for extended validation 05 · GNN extension (Moran's I motivated)

DEPLOYMENT TARGET

Operational by 2027 burning season

Weekly forecast dashboard 23 Punjab districts covered Sub-second inference per district Top-50 risk cells per district per week Pre-positioned enforcement + CRM

The same satellite, climate, and policy data that diagnoses Punjab's burning can guide its prevention.

Source: Project deployment plan; team commitment, May 2026.