LavaLabs Atmospheric Lab: Cloud-Quality Ablations and Site-Aware Prediction in OBSCURA

Saishreek Singh

Solo student entry, OBSCURA Satellite AI Research Challenge, 23 November 2026

***Abstract*—We investigate whether cloud-quality handling and relative spatial image summaries improve prediction of five anonymous atmospheric and thermal indicators beyond the supplied table inputs and official Sentinel-2 summaries. Using 3,218 training observations from 81 sites, we fit target-specific regularized gradient-boosted regressors with missing-label masks. A fixed 64-site development partition uses nested grouped selection of conservative prediction mixtures; 17 separate training sites provide an acceptance check. The official-summary family achieves development MCRMSE 0.93378 and confirmation MCRMSE 0.86206, compared with 0.93758 and 0.88725 for the table-only mixture. Our no-fallback/quality and spatial extensions do not outperform the official family in development and are rejected. The raw table baseline scored 1.32129 publicly despite site-CV 0.96638, illustrating a material transfer gap. We provide finite 1,221-row exports, a reproducible private notebook and software checks. The confirmation partition is used for acceptance, and neither it nor grouped training validation establishes future-time or private-test accuracy.**

***Index Terms*—Earth observation, atmospheric regression, site-grouped validation, cloud-quality ablation, reproducibility.**

I. Introduction

Do a controlled change to cloudy-image fallback and sixteen relative spatial descriptors add unseen-site predictive value beyond the official 54 image summaries? The contribution is a measured comparison and a conservative selection workflow. The tested custom features did not earn promotion. We do not claim a new physical methane retrieval, NDVI/terrain novelty, significance result or superior leaderboard rank.

LavaLabs reuses Scientific Noir branding and evidence-review conventions from earlier SATQuery flood and terrain research. The atmospheric regression, feature ablations, validation, exports and current report were developed for OBSCURA. Prior flood cases are not atmospheric validation or extra competition training observations.

A. Decision Support and Scope

The system first predicts the five released standardized targets. Its research interface then prioritizes what must be verified: transfer performance, missing-observation support and whether added complexity survives the control comparison. Anonymous targets cannot support facility rescue routing, named-source enforcement, physical concentration thresholds or a dated future forecast.

II. Data And Targets

Only organizer-supplied OBSCURA inputs are used: 3,218 training rows, 1,221 test rows, 29 numerical table predictors and 4,439 14-layer Sentinel-2 chips. Training contains 81 site groups. Sites and periods differ across released training, public test and private test; true dates, locations, footprint scale and target standardization constants are withheld. No real facility, date or geographic scale is recovered. Identifiers are used only for joins/export and site keys only for partitioning.

TABLE I
OBSERVED TARGET SUPPORT

Target	All training observed	Confirmation observed
CH4	1113	216
NO2	3130	605
CO	3188	614
AAI	3218	621
FIRMS	3218	621

Missing targets are excluded separately for each fit and error calculation, not replaced with zero. All held-out and test predictions must nevertheless be finite. MCRMSE is the arithmetic mean of the five target RMSEs. Each RMSE pools squared errors over its observed rows before taking a square root; fold RMSEs are not averaged as a substitute.

III. Features And Validation

The table already includes NDVI, terrain, weather/season and spectral means. Image bands B1-B12 follow the supplied 14-layer mapping, including B8A, SCL and CLDPRB; reflectance is scaled by 10,000 after conversion to floating point. The official family adds 54 summaries: statistics of six visible/NIR/SWIR bands, NDVI and SWIR-tail/threshold summaries, scene-class fractions and coverage. Its mask excludes SCL 3/8/9/10 and falls back to nonzero red-band pixels below 500 selected pixels. This implementation adapts the Apache-2.0 organizer starter.

The no-fallback variant preserves missing spectral summaries below that same 500-pixel threshold and adds three quality flags. The spatial variant adds sixteen relative centre/outer means and contrasts, B11/B12 tail contrasts, gradients, centre B12 spread and coverage values. Regional statistics require 20 selected pixels. The block uses relative image position, not metres, and is not a plume or gas retrieval. It differs from the earlier research memo's unimplemented residual/2x2 proposal. A strict cloud-probability mask exists in source but was not an evaluated family.

Only 14 of 4,439 chips (0.3154%) fall below the selected-pixel threshold. The no-fallback comparison changes

both fallback behavior and three flags; its result cannot isolate masking causally. Source arithmetic, all-cloud and zero-data behavior were checked with synthetic TIFFs.

A. Site-Grouped Validation and Selection

A fixed GroupShuffleSplit (seed 2026, test_size 0.2 by sites) reserves 17 confirmation sites; the other 64 form development. Three seeded outer GroupKFold partitions evaluate all families on identical sites. Within each outer fit, three grouped folds generate target-specific out-of-fold predictions to select mixtures. No outer or confirmation label is used to fit its prediction or tune its mixture weights. Random-row early stopping is disabled.

For target j, prediction = mu + a[(1-w)(B-mu)+w(E-mu)], where B is table prediction, E is the selected family prediction and mu is the fitting-partition observed mean. Fixed grids: a in {0, .25, .5, .75, 1}; w in {0, .5, 1}. Numerical ties prefer smaller image weight and amplitude. This prevents extrapolation and permits a fold-fitted mean fallback. Zero is a separate fixed standardized reference, not a mean inferred from test labels.

The v2 gate selects one control from table/official and at most one novel challenger from no_fallback/spatial on development only, with tolerance 1e-10 and simpler ties. An eligible novel challenger must beat both controls on confirmation. An official development control that fails table confirmation falls back to table. The other novel family is never searched on confirmation. Here official was selected and no custom challenger qualified. This gate was fixed during the initial image pass before image development or confirmation scores were inspected.

IV. Experimental Results

TABLE II
MATCHED DEVELOPMENT RESULTS

Family (total predictors)	Raw model MCRMSE	Nested mixture MCRMSE
Table (29)	0.968127	0.937579
Official (83)	0.951403	0.933783
No-fallback + quality (86)	0.955164	0.936270
Spatial + quality (102)	0.956259	0.936608

Fold-fitted mean reference: MCRMSE 1.022587 on the same development rows. Official mixture beats the table mixture in each of the three development folds; spatial beats official only in the third. Feature-family selection on these outer scores creates selection optimism for the winner, even though mixture weights are selected inside each outer fit.

A. Rejected Feature Extensions

No-fallback/quality MCRMSE 0.93627 and spatial/quality 0.93661 both exceed official 0.93378. Neither is retained. The evidence supports keeping the existing official image summaries with conservative blending for this protocol; it does not support claiming that extra spatial engineering improved the model.

B. Confirmation and Public Scoring

TABLE III
CONFIRMATION RESULTS BY TARGET

Target	Observed	Table mixture RMSE	Official mixture RMSE
CH4	216	0.935736	0.935736
NO2	605	0.971846	0.933614
CO	614	1.178384	1.162906
AAI	621	0.660183	0.627326
FIRMS	621	0.690108	0.650699
MCRMSE	Per-target masks	0.887252	0.862056

Official has lower confirmation error for NO2, CO, AAI and FIRMS; CH4 is tied. CO remains weak on this partition. The five-target confirmation score improves by 0.025195 in absolute MCRMSE. This is a descriptive matched-partition difference, not a significance test or an unbiased estimate after selection.

TABLE IV
LOCAL AND PUBLIC EVALUATION POPULATIONS

Evaluation	Raw table baseline	Retained image mixture
All-training 5-fold site CV	0.966382	Different nested protocol; see above
Public leaderboard	1.32129	1.15808
Private leaderboard	Unknown	Unknown

The raw baseline's accepted public score is substantially worse than its local score. One public result cannot identify the source of this shift; geographic, temporal and other differences are plausible but not established causes. We do not tune weights against repeated public submissions. Grouped site CV within training does not emulate the withheld latest private period.

V. Reproduction And Runtime

Import the source-only notebook into a private Kaggle notebook with the official input attached, then Save Version / Save & Run All. The pipeline writes a strict aggregate receipt and two identical private CSVs. Export checks enforce 1,221 unique ordered string IDs, exact columns and five finite targets, then reopen the CSV and calculate SHA256.

Corrected interactive replay: 608.654 seconds, including 446.265 seconds of extraction. Fresh saved version 372947210 succeeded with identical scores and prediction hash: 794.109 seconds pipeline, 821.9 seconds including setup. These are measured runs, not hardware-independent latency promises. Fitting uses two numerical threads. Environment: Python 3.13.18, NumPy 2.1.3, pandas 2.3.3, scikit-learn 1.6.1, Rasterio 1.5.2, SciPy 1.16.3.

Source-module SHA256: e878c89e8a00143e35c1f6819f7c926fdf9e7fd90c2cb081d053d5d3407b405c

Private prediction CSV SHA256: 007abb84e2f53b01d7ca71929647d83ea6803e2cb186871fd61f3cda38892c20

VI. Fixed Model Follow-up

After the four-family experiment, we predeclared one bounded Ordered CatBoost comparison using the same 83

table/official-summary inputs and development sites. We did not rescore the earlier confirmation partition. This is iterative development on reused sites, not independent replication or new confirmation evidence.

Fixed CPU recipe: CatBoost 1.2.10, squared-error loss, 400 rounds, learning rate .03, depth 3 symmetric trees, Ordered boosting, L2 regularization 10, Bernoulli subsample .8, per-tree sampling, feature fraction .8, random strength 1 and seed 2026. Native missing features are permitted; missing targets are excluded. Two CPU threads, no evaluation set and no early stopping. The same inner-only mixture grid is used. Promotion requires improved pooled development MCRMSE and at least two of three matched fold wins; no parameter sweep was performed.

TABLE V
FIXED MODEL FOLLOW-UP

Evaluation	HistGB official mixture	Fixed CatBoost mixture
Development pooled MCRMSE	0.933783	0.931200
Outer fold 1	0.843511	0.843672
Outer fold 2	1.026726	1.023215
Outer fold 3	0.896520	0.891904
Accepted public MCRMSE	1.15808	1.16407
Private MCRMSE	Unknown	Unknown

The fixed CatBoost model passed the development gate with two fold wins, but its public score was slightly worse. Both accepted candidates are selected for final private evaluation; HistGB remains the best observed public result. We stopped model search. This reversal reinforces the need to distinguish development gains from transfer gains and does not establish which model will be better privately.

Saved private run 372977582 completed 90 CatBoost fits in 758.705 seconds of pipeline time (442.297 seconds of image extraction; viewer 784.9 seconds including setup). The full-training model exported 1,221 finite prediction rows with the same strict contract. Its settings, actual aggregate receipt and output-free source are frozen separately from v2.

CatBoost source-module SHA256: 2a8e41d2af52ef93dfc54468c48d0e01d3a749ca8176d63d30a236339b173536

CatBoost private prediction CSV SHA256: be5f8926b4d90991f191d9e223f71b772d5577d830e2a652f16ed0afe4b141e6

VII. Verification And Limitations

The frozen v2 recipe passed 26 local tests. The later 30-test atmospheric suite also checks native missing features, missing targets, fixed trees, saved-model replay and outer-label perturbation for CatBoost. Thirty app/legacy checks passed after interface integration. Coverage includes unequal target masks, invalid predictions, site disjointness, exact CSV round-trips, unsigned-band arithmetic, all-cloud/empty imagery, synthetic TIFF extraction, cache invalidation/integrity and model-to-export execution. Synthetic checks do not establish atmospheric accuracy. Code review led to corrections of the confirmation gate and cache key.

The confirmation gate influences model acceptance and is not an untouched test. Initial raw-baseline inspection included all training sites, including the later confirmation sites. Development feature-family selection has selection optimism. Training contains only 81 site groups and sparse CH4 observations. No calibrated uncertainty, real-time health-alert utility, physical emissions, independent field validation or temporal transfer is established. Private-test performance remains unknown. All restricted rows, imagery, identifiers, caches and predictions stay on private Kaggle storage; the public evidence is aggregate only.

Acknowledgment And AI Disclosure

The official feature recipe and fixed regularized estimator settings are adapted from Zelzy's Apache-2.0 OBSCURA starter, not claimed as wholly original. OpenAI Codex assisted implementation, synthetic tests, research synthesis, analysis automation, interface work and writing. Scientific observations retain organizer provenance; no generated target labels or invented performance values are used. No pretrained model or outside environmental data is used by this recipe.

References

[1] OBSCURA data, evaluation and rules. [Online]. Available: kaggle.com/competitions/obscura.

[2] Organizer starter notebook (Apache-2.0), version 359068856. [Online]. Available: kaggle.com/code/zelza0/obscura-starter-baseline?scriptVersionId=359068856.

[3] LavaLabs source-only reproduction (both frozen recipes). [Online]. Available: kaggle.com/code/saishreeksingh/lavalabs-atmospheric-lab-source-and-evidence.

[4] Kaggle profile: Saishreek Singh. [Online]. Available: kaggle.com/saishreeksingh.

[5] Devpost: LavaLabs Atmospheric Lab. [Online]. Available: devpost.com/software/lavalabs-atmospheric-lab.

[6] GroupKFold documentation. [Online]. Available: scikit-learn.org/stable/modules/generated/sklearn.model_selection.GroupKFold.html.

[7] Nested cross-validation documentation. [Online]. Available: scikit-learn.org/stable/auto_examples/model_selection/plot_nested_cross_validation_iris.html.

[8] CatBoost training parameters. [Online]. Available: catboost.ai/docs/en/references/training-parameters/.

[9] Copernicus Sentinel-2 L2A bands and masks. [Online]. Available: documentation.dataspace.copernicus.eu/APIs/SentinelHub/Data/S2L2A.html.

[10] Varon et al. (2021), High-frequency monitoring of anomalous methane point sources. [Online]. Available: amt.copernicus.org/articles/14/2771/2021/.

[11] Cawley and Talbot (2010), On Over-fitting in Model Selection. [Online]. Available: jmlr.org/papers/v11/cawley10a.html.

Code source, public-reference access and platform finalization must be verified separately before submission. A private notebook URL alone is not a public reproduction link. This report records measured experiments; it does not assert final Devpost acceptance.