Weather Exposure and Commercial Real Estate Returns

An exploratory modeling study and validation roadmap

AgilityAI Lab | Working report AAI-LAB-2026-01 | Version 0.2 | October 28, 2026

Abstract

This working report reviews an experimental archive that combines historical weather observations with commercial real estate financial records. The research question is whether temperature extremes, precipitation and selected economic variables can inform modeling of real estate returns. The archive includes linear and regularized regression, tree ensembles, and LSTM/GRU experiments. An inspected expanded panel contains 12,148 records across 136 U.S. Core-Based Statistical Areas, covering stored quarters from 1981 Q1 through 2023 Q4. Earlier five-market tabular experiments retain R-squared values from -0.0705 to 0.2246 across eight variants. These are cached outputs, not newly reproduced benchmark results. Same-period return components, pooled-market sequences, preprocessing leakage and incomplete weather coverage limit interpretation in several branches. We document these distinctions and propose a market-aware, forward-holdout evaluation with controlled climate-feature ablations. No causal climate effect, prospective forecasting advantage or operational deployment is established by this review.

Research question

Can historical weather exposure contribute information beyond financial and macroeconomic baselines when estimating future commercial real estate returns? Answering this requires a defined forecast origin, features available at that origin and a comparable out-of-time evaluation. A relationship in historical records is not, by itself, a causal effect of climate change.

Scope and report status

This is a new technical review prepared in October 2026. Version 0.2 also incorporates the supplied AgilityAI Real Estate Analytics Technical Exhibit [4], using its technical sections as supporting source material. It is not a peer-reviewed publication and has no assigned DOI. No model was retrained or executed for the report. The source filenames contain earlier experiment labels; this report does not authenticate those labels as creation or publication dates. Numerical results are presented with their source-task qualifications rather than as operational performance claims.

1. Data and geographic coverage

The archive identifies PRISM as a weather source and NCREIF financial exports as a real estate source. PRISM provides historical climate datasets [1]. The NCREIF Property Index is a quarterly, unleveraged composite total-return measure for investment properties [2]. These provider descriptions explain the named source families; they do not independently authenticate the repository's particular extracts, transformations or usage rights.

The expanded filtered table contains one retained row per observed CBSA-quarter: 12,148 rows, 136 distinct metropolitan identifiers and stored quarters from 1981 Q1 through 2023 Q4. The market count increases over time. Consequently, this is an unbalanced historical panel, not 136 complete market histories and not evidence of validated accuracy in each market.

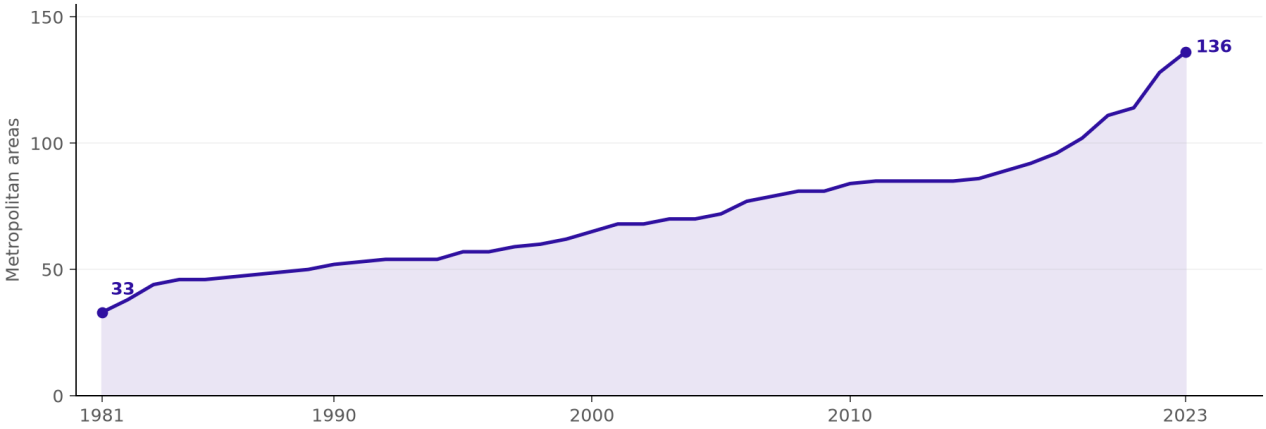


Figure 1. Distinct represented CBSAs by year in the inspected filtered panel. This newly drawn figure describes archive coverage. It is not a forecasting result. A CBSA is a Core-Based Statistical Area.

Separate experimental populations

The tabular comparison in Section 3 belongs to an earlier five-market research branch, with quarterly modeling tables of approximately 817 rows. It must not be described as evaluation across the expanded 136-market panel. A separate augmented table is explicitly designated synthetic in its filename; its scores are excluded from this comparison.

Coverage and unit controls

The expanded annual weather extracts inspected in the archive end at December 1, leaving late-December coverage incomplete. Quarterly counts and totals therefore need completeness controls. The broader financial-weather join also contains multiple records per CBSA-quarter before filtering. The observational unit and filtering lineage must be reconciled before calendar-window features can be interpreted consistently.

2. Implemented analytical workflow

The supplied technical exhibit [4, sections 2-5] describes a reusable CBSA-quarter workflow: concatenate annual weather extracts, order dates, associate CBSA names with representative coordinates, construct exposure fields, aggregate quarterly observations, normalize financial identifiers and join weather to financial records. Tabular models then consume observation-level feature vectors; recurrent models consume sample-by-row-by-feature tensors. These are related experimental branches, not one verified production pipeline.

Weather exposure and calendar semantics

The code constructs cold indicators at minimum temperature at or below 0 and heat indicators at maximum temperature at or above 30. Interpreting quarterly sums as counts of extreme-temperature days requires confirmed Celsius units and complete daily records. Precipitation is accumulated over available observations; temperature and vapor-pressure sums are cumulative quantities, not quarterly averages. Fields named l4q and l24q use four-row and 24-row rolling sums without a lagging shift. Repeated financial rows mean these windows need not span that many distinct quarters [4, section 4].

Explicit statistical and target contracts

MSE is the mean of squared differences between predicted and observed targets. RMSE is its square root and has the target's units. R-squared compares squared error with holdout variation around its mean. Compiled neural-network training loss can include regularization and is not necessarily prediction-only MSE. All require an explicit target, transformation, sample construction and information cutoff [4, sections 5 and 8].

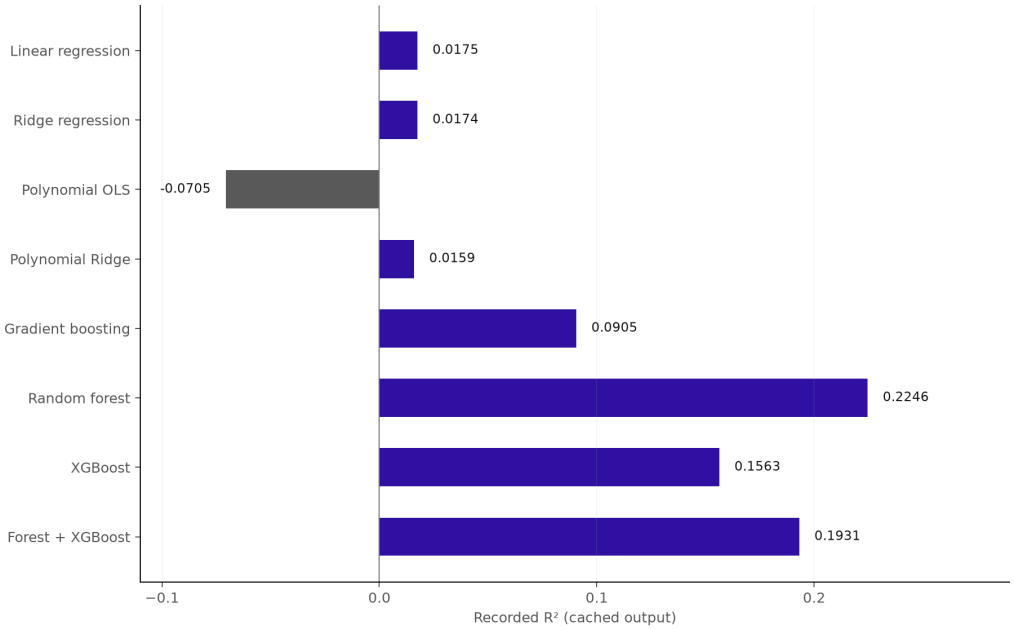
The September tabular baseline uses an 80/20 random split with random_state 42, with standardization fitted on the training partition in that branch. Selected variants use five-fold cross-validation and hyperparameter searches. This is exploratory fitting evidence; a random holdout does not establish forward-time generalization [4, sections 5-6].

Source populations remain distinct

The exhibit reports 64,979 rows and 147 CBSAs in the source financial export, 24,596 rows and 143 CBSAs in the expanded quarterly weather table, and 61,865 rows and 136 CBSAs in the broader joined panel. These are separate processing stages. The filtered panel in Section 1 is the 12,148-row analytical selection. A stored prediction export has 356 rows across 89 CBSAs and four labeled periods in 2025; its presence is not verified forecast performance or adoption [4, section 3].

3. Models and recorded tabular results

The earlier branch explores ordinary least squares, Ridge regression, polynomial feature expansions, gradient boosting, random forests, XGBoost and an average of forest/XGBoost predictions. Several variants use random splits and cross-validation. Financial and weather feature selections vary; recorded differences cannot be attributed solely to climate inputs.



Model variant	MSE	R-squared
Linear regression	0.0016450983	0.017450
Ridge regression	0.0016451395	0.017426
Polynomial OLS	0.0017923775	-0.070514
Polynomial Ridge	0.0016477323	0.015877
Gradient boosting	0.0015227619	0.090517
Random forest	0.0012982849	0.224588
XGBoost	0.0014125745	0.156327
Forest + XGBoost	0.0013509678	0.193122

Figure 2 and Table 1. Retained five-market tabular outputs. MSE is in squared target units on the fractional-return scale; R-squared is a fit statistic, not an accuracy percentage. Results were not rerun for this report. Values across different targets, transformations or source populations are not comparable by default.

The random forest has the highest retained R-squared in this set (0.224588). The forest/XGBoost average (0.193122) does not improve on it. Polynomial OLS has negative R-squared (-0.070514), indicating performance below the split's mean-prediction reference under that statistic. The linear and Ridge results are weak. This mixture of outcomes is useful exploratory evidence, but does not establish prospective accuracy or a climate-related forecasting advantage.

4. Recurrent models and target interpretation

The exhibit describes a saved weather-feature GRU with input shape 12 rows by 11 features, recurrent layers of 70 and 35 units, dropout 0.3 and one Dense output. Its inputs comprise the quarter code, six rolling weather fields, inflation, interest rate, a COVID impact score and unemployment; the branch uses MinMax scaling [4, section 7]. These dimensions describe the inspected artifact, not proof that its rows are consecutive quarters from a single market.

Experiment	Input rows x fields	Cached MSE
Weather-feature GRU	12 x 11	0.00029784
Five-feature GRU	12 x 5	0.00025447

Table 2. Two selected recurrent results transcribed from the exhibit [4, section 8], rounded as reported. The weather-feature branch records MSE 0.00029784 on its unscaled tret-mean target; the separate five-feature branch excludes weather predictors and records 0.00025447. These are not a controlled ablation, comparable to the tabular fractional-return scores, or evidence of a benefit caused by weather inputs.

Row averages, return factors and currency

The recurrent label is generally the arithmetic mean of the H target rows following L input rows. It is not a compounded holding-period return. In the filtered panel, the exhibit reports treturn approximately equal to 100 multiplied by (tret minus 1), with a maximum discrepancy of about 0.00000005. This supports distinguishing a gross return factor from a percentage-return representation [4, sections 7 and 8.3].

The dollar-named branch uses a scaled treturn target and has four input rows followed by a mean of 12 target rows. Its reported MSE is measured before inverse transformation. Currency, valuation base and a valid conversion are absent from that error, so it cannot be presented as a dollar-loss estimate [4, section 8.3].

Architecture and output lineage

Several experiments overwrite a shared archive filename. A current two-layer architecture cannot be assigned to every cached result that references that path. An immutable run record must connect the source, data selection, preprocessing state, model configuration and output before performance is reproducible. Declining cached training traces do not repair this lineage or establish independent final-test evaluation [4, sections 8-8.2 and 12].

5. Evaluation limits that change the interpretation

Return reconstruction is not a forecast

Other regression branches include same-period income and capital return while predicting total return. In stored tables, total return closely matches the sum of those components. Near-perfect scores can therefore reflect reconstruction of an accounting relationship. They do not establish that weather predicts an unknown future return. Such branches are not used as headline performance evidence here.

Rows are not necessarily quarters within one market

The recurrent experiments include both LSTM and GRU architectures. Several construct windows from a pooled, date-sorted panel without grouping by CBSA. Adjacent rows can belong to different markets in the same quarter. An input length of 12 rows cannot automatically be called a 12-quarter history for one market, and an average of subsequent rows cannot automatically be called a market-specific annual forecast.

Preprocessing and holdout separation matter

Several recurrent or regression branches fit scaling before the split. Overlapping sequence windows and use of test data for validation further weaken holdout independence. A proper forward evaluation must fit preprocessing on training observations only, separate model selection from final testing and purge overlapping target periods at split boundaries.

Date decoding must be explicit

The exhibit records spreadsheet serial dates that correspond to quarter-end dates when decoded with the appropriate day origin. General notebooks instead apply a numeric timestamp conversion without that origin, producing epoch-adjacent labels. Preserved numeric ordering is not validation of calendar labels; the revised benchmark must reconcile decoded dates with quarter codes [4, section 10].

Climate contribution has not been isolated

The archive contains weather-feature and non-weather recurrent variants with different selected inputs. Their retained losses do not form a controlled climate-feature ablation. Error scales and target definitions also differ across branches. We do not combine these losses into a model leaderboard or claim an improvement caused by weather features.

Storm and target-definition branches need repair

The exploratory storm merge uses geographic distance logic whose coordinate units do not match the stated radii, and some neural branches have target-selection or input/horizon inconsistencies. A branch named "actual-dollars" reports scaled-target error, not independently verified monetary error. None of these results establishes property-level damage estimation, catastrophe forecasting or a monetary valuation service.

6. Proposed validation roadmap

The following protocol is future work. It has not been completed by preparing this report.

1. Define one target, its unit, a forecast origin and an operational horizon. Exclude contemporaneous return components and any information unavailable at that origin. Document publication lags for economic variables.

2. Construct a complete calendar-quarter weather series per market before rolling aggregation. Track missing dates, stable geographic boundaries and all joins. Apply shifts when the prediction task requires information strictly before the target period.

3. Build each sequence within a single market and financial category where applicable. Require consecutive quarters. Preserve the market identifier, origin and target-period identifiers for every sample.

4. Split by time, fit transformations on training observations and use a separate validation period for selection. Keep a final untouched forward holdout; purge or embargo overlapping target windows at boundaries.

5. Compare historical-mean, persistence, macroeconomic-only, climate-only and combined models under identical samples and splits. Report MAE, RMSE, suitable R-squared, sample counts, uncertainty and per-market/extreme-weather errors. Test climate's incremental value with a controlled ablation.

Reproducibility and attribution

The published result JSON contains selected aggregate counts, cached tabular metrics, selected recurrent summaries, source notebook hashes and the supporting exhibit fingerprint. It does not distribute raw financial records, model weights, credential-bearing notebooks or private exhibits. Hashes identify the inspected source bytes; they do not prove historical existence or original authorship. The underlying model card identifies the earlier project as Climate Index AI. This AgilityAI Lab report is a new technical review and does not reattribute the original experiments, assert a transfer of ownership or grant a license to underlying assets. Distribution and use of those assets remain subject to their rights holders and source terms.

References

[1] PRISM Climate Group. Time series data. https://prism.oregonstate.edu/data/ (accessed October 28, 2026).

[2] NCREIF. NCREIF Property Index. https://user.ncreif.org/data-products/property/ (accessed October 28, 2026).

[3] Cvijanovic, D., and van de Minne, A. Temperature Extremes and Commercial Real Estate. February 28, 2024. https://doi.org/10.2139/ssrn.3956714. Related external literature; not an AgilityAI-authored paper and not reproduced here.

[4] AgilityAI Real Estate Analytics Technical Exhibit. Supplied supporting technical review, consulted October 28, 2026. Technical sections 2-12 and their source references inform this revision. The original document is not distributed with this public report.

Suggested citation: AgilityAI Lab (2026). Weather Exposure and Commercial Real Estate Returns: An exploratory modeling study and validation roadmap. Working report AAI-LAB-2026-01, v0.2.