Probabilistic Machine Learning for Parcel Forecasting at National Scale

Homecastr Technical Note

Daniel Hardesty Lewis
Founder & ML Engineer, Homecastr
March 2026

Abstract

Homecastr produces probabilistic price forecasts for U.S. residential properties at one- to five-year horizons. The system combines a frozen tabular transformer encoder for deterministic trend estimation with a simulation-free Schrödinger Bridge generative model that draws samples from the distribution of price-path deviations from the trend, coordinated through shared latent tokens that capture neighborhood-level spatial dynamics.

In data-dense jurisdictions, the model achieves well-calibrated 80% prediction intervals: NYC RPAD origins 2020–2024 show actual coverage of 87–97% across all five horizons, with median point error of 7–16%. In data-sparse and macro-sensitive settings, coverage collapses to 40% at a three-year horizon (ACS Nationwide, origin 2021) and to 1.5% at a one-year horizon (TxGIO Texas, origin 2021), while median point error remains competitive with a naive persistence baseline.

The central diagnostic finding is that strong point accuracy and miscalibrated intervals are orthogonal failure modes: a model can beat persistence on median absolute error while its 80% interval covers fewer than half of realized outcomes. This failure is not visible in aggregate accuracy metrics and is only separable by per-horizon calibration decomposition.

This document covers the production system: architecture, training objectives, data sources, calibration diagnostics, and multi-horizon backtest results. The roadmap addresses the structural root cause of the calibration failure.

1. Problem

Most consumer real-estate products present a single forecast value. A single value is easy to display, but it does not communicate downside risk, upside potential, or forecast uncertainty. For a homeowner deciding whether to sell in two years or five, the shape of the distribution matters more than the point estimate.

Homecastr produces forecast distributions with P10, P50, and P90 quantiles at one- to five-year horizons for U.S. residential properties. Building this required solving several problems. Not all of them are unique to this system; some are prerequisites for any national-scale property forecasting effort.

- **Per-horizon calibration.** Most probabilistic forecasts are calibrated in aggregate. Making an 80% prediction interval contain ∼80% of outcomes for each geography and each forecast year, across properties ranging from $50K rural lots to $10M urban condos, is harder. Calibration is more demanding at longer horizons: the forecast target (cumulative price change over multiple years) grows in scale, and the model must correctly widen its uncertainty estimates to match. These failures are difficult to detect without structured per-horizon evaluation infrastructure.
- **Data heterogeneity.** Producing forecasts at national scale requires ingesting property records from heterogeneous county assessor sources with different schemas, identifiers, and

valuation conventions. This challenge is not unique to this system; it faces any national property data effort. The architectural response was to encode cross-source heterogeneity implicitly through learned embeddings rather than conditioning on source identity, enabling one model to generalize across jurisdictions.

- **Distributional shift.** The model learns historical price dynamics. When the macro environment diverges significantly from that history, forecasts may degrade because the learned distribution no longer matches the current data-generating process. This divergence is not detectable in real time; it becomes visible only as realized values accumulate and the monitoring pipeline measures calibration failure.
- **Multi-year temporal coherence.** Forecast paths need to look like realistic extensions of a property's historical curve. Independent per-horizon regressors produce paths that are not internally consistent across time: e.g., predicting 10% growth in year 2 but only 3% cumulative by year 3. A generative model produces paths that respect temporal dependencies by construction.
- **Cross-property spatial coherence.** Adjacent properties in the same neighborhood tend to move together. A neighborhood-level event affects all parcels in an area simultaneously. Representing this requires a shared latent structure across properties that individual property features cannot provide.

2. Why Naive Approaches Fail

Before arriving at the current architecture, simpler alternatives were explored. Each failed in a specific, instructive way.

Point Forecasts Miss the Distribution

A point forecast cannot tell a homeowner whether their downside risk is 5% or 25%. Post-hoc Gaussian prediction intervals around a point estimate are heavy-tailed and geography-dependent in ways a Gaussian cannot represent. Calibrated intervals require learning the distribution directly.

Independent Horizons Produce Temporally Inconsistent Paths

Training separate regressors for each forecast horizon produces paths that are not internally consistent across time: a property might show 10% growth in year 2 but only 3% cumulative by year 3, which is impossible if year 3 follows year 2 in sequence. A generative model produces paths that respect temporal dependencies.

Naïve Diffusion Fails on Short-Horizon Tabular Data

Standard denoising diffusion models require hundreds of reverse steps and are designed for high-dimensional continuous signal spaces. For a five-dimensional target (H=5 annual deltas), this is architecturally mismatched. More fundamentally, ε-MSE training was found to produce reasonable mean estimates but under-dispersed tails. This failure motivated the switch to simulation-free Schrödinger Bridge matching, which uses closed-form flow and score targets derived analytically from the bridge path.

3. System Architecture

Figure 1. End-to-end system architecture. Source data flows through a canonical panel into the model stack. The tabular transformer encoder produces a frozen trend estimate; the Schrödinger Bridge generative model samples distributions of price-path deviations conditioned on the encoder output and shared spatial token paths. Results stream to GCS as a write-ahead buffer before bulk upsert to Supabase.

Data Ingestion

The five primary sources each require distinct schema adaptation before reaching a canonical parcel-year representation.

Source	Coverage	Scale	Notes
TxGIO	Texas statewide	27M parcels	Missing 2020 vintage; requires year carry-forward
Florida DOR	Florida statewide	10M parcels	String-typed appraisal values; requires explicit casting
HCAD	Harris County, TX	1.8M parcels	Richest structural features; parcel-level since 2005
NYC DOF RPAD	New York City	1.1M parcels	BBL identifiers; spatial join with MapPLUTO
ACS	US nationwide	82K tracts	Negative sentinel codes for suppressed data below −600M
FRED	Macroeconomic	7 series	Annual snapshots; joined to panel by anchor year

Property-level features are combined with macroeconomic covariates including mortgage rates, federal funds rate, 10-year Treasury yield, CPI, unemployment, VIX, and oil prices, sourced from FRED and joined by anchor year.

AI-Native Interaction Layer

Homecastr surfaces its probabilistic outputs via an **AI-Native Interaction Layer** designed for both human and agentic consumption:

- **Semantic Omnibar:** A natural language gateway that parses complex queries, maps them to neighborhood geometries, and retrieves the relevant forecast distributions.
- **Model Context Protocol (MCP) API:** Homecastr operates as an MCP server, exposing forecast distributions, appreciation outlooks, and comparable market data as standardized tools.

4. Model Architecture

Why a Generative Architecture Was Needed

Three limitations of simpler approaches motivated the generative design:

1. Independent per-horizon regressors produce paths that are not internally consistent across time. A joint generative model respects temporal dependencies by construction.

2. Parametric intervals around a point forecast are unlikely to capture the heavy-tailed, geography-dependent distributions in realized price changes.
3. Neighborhood-level dynamics require a shared latent structure across properties that individual property features cannot provide.

Frozen Tabular Transformer Encoder

A Feature Tokenizer Transformer [1] (FT-Transformer) is a self-attention encoder over heterogeneous tabular features. Each numeric feature gets its own learned linear projection to d_{model}=128 dimensions, alongside per-category embeddings (hash-bucketed to absorb cross-source schema differences), a history summary token, a geographic region token, and a learnable summary token. The encoder has n=3 layers, 4 attention heads, GELU activations, and pre-LayerNorm. The summary token output feeds a two-layer head producing a deterministic trend prediction $\hat{\mu} \in \mathbb{R}^H$ and a context embedding for the generative model.

During training of the generative model, the encoder is **frozen**. The generative model learns the residual $R = Y_{\text{path}} - \hat{\mu}$, where Y_{path} is the cumulative log-return trajectory. In property terms: the encoder predicts where a property's price is expected to go based on its features; the generative model learns the distribution of how much actual outcomes deviate from that expectation. The P10–P90 band represents the range of those deviations.

Shared Spatial Token Paths

K=8 shared latent token paths capture neighborhood-level shocks. Each token follows an AR(1) process with learned persistence ϕ_k:

$$z_{k,t} = \phi_k \, z_{k,t-1} + \sqrt{1 - \phi_k^2} \, \eta_{k,t}, \quad \eta_{k,t} \sim \mathcal{N}(0,1) \tag{1}$$

where $\phi_k = 0.99 \cdot \sigma(\text{logit}_k)$. A gating network computes per-property sparse mixing weights $\alpha_i \in \Delta^K$. The per-property shared driver is:

$$u_i = \sigma_u \cdot \left(\sum_{k=1}^{K} \alpha_{i,k} \, z_k \right) \cdot \sqrt{\mathbf{h}} \tag{2}$$

where $\sigma_u \in (0, 2)$ is a learned coherence scale and $\sqrt{\mathbf{h}}$ scales neighborhood shock magnitude with forecast horizon. Properties in the same neighborhood share the same token draws in a given inference run, so their forecast paths are correlated. In production training, effective token count stabilizes near 4 of 8 available.

Simulation-Free Schrödinger Bridge

The Bridge Path

The source distribution is standard Gaussian: $x_0 \sim \mathcal{N}(0, I_H)$. The generative model transports x_0 to the training residual $x_1 = R_{\text{path}}$ along a Brownian Bridge:

$$x_s = (1 - s) \, x_0 + s \, x_1 + \sigma(s) \, \varepsilon, \quad \sigma(s) = \sigma_{\max} \sqrt{s(1-s)}, \quad s \in [0, 1] \tag{3}$$

with $\sigma_{\max}$=1.0 and a noise floor of 0.01.

Closed-Form Training Targets

Both training targets are computable analytically from (x_0, x_1, s, x_s), requiring no forward simulation:

$$u^\star = (x_1 - x_0) + \frac{1-2s}{2s(1-s)}\,(x_s - \mu_s) \quad \text{(conditional velocity)} \tag{4}$$

$$\tilde{s}^\star = -(x_s - \mu_s) \quad \text{(scaled score)} \tag{5}$$

Network Architecture

A shared MLP trunk encodes conditioning into a 256-dimensional hidden vector via six parallel encoders summed together. Four residual MLP blocks followed by LayerNorm produce a shared representation. Two lightweight heads decode this independently to velocity $\hat{v} \in \mathbb{R}^H$ and scaled score $\hat{s} \in \mathbb{R}^H$.

Condition-Aware Optimal Transport Coupling

Within each training microbatch of 512, exact Earth Mover's Distance optimal transport pairs each Gaussian source draw with a training residual, minimizing a cost that blends residual-space and conditioning-space distances [3].

Inference and Quantile Computation

At inference, 20 Euler-Maruyama steps transport $x_0 \sim \mathcal{N}(0, I_H)$ forward. S=50 independent draws per property produce 50 complete price-path deviations. P10, P50, P90 are then computed as **marginal quantiles**: at each horizon h independently, the 10th, 50th, and 90th percentile values across the 50 draws. The P10 at h=3 and the P10 at h=4 will generally come from different draws. The P10–P90 band is not the boundary of a single coherent worst-case scenario; it is the envelope of the marginal distributions, horizon by horizon. Full reconstruction: $\hat{Y}_{\text{path}} = \hat{\mu} + \hat{R}_{\text{path}}$.

Training Objective

$$\mathcal{L} = \mathcal{L}_{\text{flow}} + \mathcal{L}_{\text{score}} \tag{6}$$

Both terms use the closed-form targets above with per-horizon mask weighting. There is no distributional scoring term. The absence of a proper scoring rule is the primary factor in coverage collapse at long horizons in sparse settings.

Training Metrics

Jurisdiction	**Origin**	**n_train**	**Epochs**	$\mathcal{L}_{\text{total}}$	$\mathcal{L}_{\text{score}}$	k_{eff}	σ_u
HCAD Houston	2021	4.13M	31	4.78	0.104	3.97	0.28
NYC RPAD	2021	2.97M	14	5.37	0.117	3.97	0.60
ACS Nationwide	2021	0.47M	10	5.43	0.120	3.99	0.94
Florida DOR	2021	4.84M	10	5.06	0.110	3.94	0.51

Table 1: *

Table 1. Training metrics for origin-2021 models (W&B experiment logs). n_train: parcels with a complete five-year forward window, not a raw panel count. OT off-diagonal fraction > 0.998 in all four jurisdictions.

5. Evaluation

Temporal Holdout Protocol

All model candidates are evaluated on genuinely held-out origin years using an expanding-window protocol. Macro features are lagged by one year to prevent look-ahead contamination.

Calibration Diagnostics

Each model candidate is evaluated against a structured diagnostic suite. These targets guide development and flag regressions; they are not automated acceptance gates. Origins spanning macro regime breaks are expected to fall outside ranges.

- **Forecast attenuation:** step deltas drop to near-zero beyond the first year, producing flat trajectories.
- **Miscalibrated dispersion:** intervals are too narrow (overconfident) or too wide (uninformative) relative to realized outcome rates.
- **Mean bias:** directional over- or under-prediction, typically indicating the learned trend does not match the current regime.

Test	What It Checks	Target	Threshold
Anchor Integrity	Forecast start matches last observed price	Median log-ratio is small	< 0.10
Interval Coverage	Outcomes fall inside the 80% band	Coverage in expected range	65–95%
Horizon Scaling	Uncertainty grows at expected rate	Variance ratio near $\sqrt{h}$	$\pm$30%
Point Accuracy	Median absolute error of P50 vs. realized	MdAE within acceptable range	$< 10\%$ at h=1
Baseline Beat	Model vs. persistence forecast	Model wins on more parcels	$> 50\%$
Tail Accuracy	Extremes occur at expected rate	Each tail near nominal	$\approx 5\%$
Variance Ratio	Forecast spread matches historical	Ratio in plausible range	0.3–3.0
Distribution Match	Growth distribution resembles history	KS test does not reject	$p > 0.01$

Backtest Results

Jurisdiction	Origin	h	MdAE	Cov 80%	Wins
NYC RPAD	2025	1	**7.0%**	**87.2%**	**56.7%**
NYC RPAD	2024	2	**8.5%**	**96.7%**	**51.8%**
NYC RPAD	2023	3	**9.8%**	**94.2%**	**63.6%**
NYC RPAD	2022	4	**12.6%**	**94.0%**	**73.6%**
NYC RPAD	2021	5	**16.1%**	**87.9%**	**59.1%**
NYC RPAD	2020	5	**12.6%**	**90.2%**	**72.0%**
NYC RPAD	2019	5	48.6%	**89.1%**	12.4%
ACS Nationwide	2023	1	**7.3%**	**71.2%**	41.7%
ACS Nationwide	2022	2	**12.7%**	**68.6%**	**87.1%**
ACS Nationwide	2021	3	30.2%	40.4%	**88.8%**
ACS Nationwide	2020	4	35.6%	46.8%	**93.2%**
ACS Nationwide	2019	5	42.8%	52.9%	**87.9%**
TxGIO Texas	2024	1	**5.7%**	**93.6%**	22.6%
TxGIO Texas	2023	1	22.6%	57.7%	9.0%
TxGIO Texas	2021	1	22.9%	1.5%	3.4%

Table 2. Furthest-horizon backtest results per origin year. Bold values meet targets: MdAE $< 10\% \times \sqrt{h}$, Coverage 65−95%, Wins $> 50\%$. Coverage reports the fraction of realized values inside the P10–P90 marginal quantile band. NYC calibrates well across recent origins. ACS and TxGIO show coverage collapse from horizon 2 onward. The key diagnostic finding is that strong point accuracy (high Wins) can coexist with severely miscalibrated intervals (40% coverage on an 80% target): these failure modes are independent and require separate diagnostics.

6. Serving

Inference runs on EC2 (g4dn.xlarge) with batch sharding and a GCS write-ahead buffer. Property-level results are rolled into geographic aggregations at ZCTA, tract, and neighborhood levels. Feature retrieval uses Supabase with tiered caching at the edge.

7. Limitations and Failure Modes

Coverage collapse at long horizons in sparse settings. In both data-sparse jurisdictions evaluated (ACS Nationwide and TxGIO Texas), the 80% interval covers far fewer than 80% of realized outcomes at horizons beyond one year. The pattern is consistent across those two cases but has not been examined across a stratified sample of sparse settings; a broader evaluation is needed before generalizing. Root cause: the training objective contains no proper scoring rule. The model learns the most likely deviation path rather than the full distributional width at each horizon.

Conditioning bottleneck. All conditioning information is compressed into a single 256-dimensional sum. Year-level temporal structure is lost; the model cannot distinguish a steeply rising three-year history from a stable one with the same endpoint.

Static macro features. Macroeconomic series are joined at the anchor year only. The model cannot condition on a rate trajectory over the forecast window.

No comparable-sales conditioning. Nearby transaction prices are not used.

Appraisal vs. transaction prices. The model trains on county-appraised values, which can lag market movements and diverge from transaction prices in rapidly changing markets.

Attribution gap. Feature attributions are computed from the deterministic encoder, not the full generative pipeline.

Note on verifiability. Backtest metrics above are cross-checked against experiment tracking exports; a pipeline for generating verified TEX tables directly from those exports does not yet exist. The total panel size across all jurisdictions and years has not been computed from the pipeline itself and is not reported here.

8. What This System Taught

1. **Uncertainty is the product.** The interval width, its calibration across geography and horizon, and the diagnostic that detects when it fails are more informative than the point estimate.
2. **Longer horizons require structural support.** Coverage collapse at h=3−5 in sparse settings could not be fixed by tuning; it required a structural change: a proper scoring rule, richer macro context, or richer history representation.
3. **Data quality is an architectural constraint.** The divergence between NYC and ACS/TxGIO shows up in calibration shape, not just MdAE. A model that generalizes across sources must condition on observation quality explicitly or fail differently in each regime.
4. **The validation loop is the research.** Building per-horizon diagnostics and failure mode classification is not supporting infrastructure. Without it, coverage collapse at h=3 would be invisible in aggregate metrics.
5. **Decomposition is productive.** Separating the deterministic encoder (frozen) from the distributional model allowed both to be diagnosed independently.

9. Transferable Research Problem

The transferable methodological problem is: how do you build and validate a probabilistic model of spatial trajectories over time, when outcomes are only partially observed, history is heterogeneous across sites, and the validation horizon exceeds the available realized data?

Novelty Assessment

The problems encountered in this project are not equally novel. Most are known challenges with established partial solutions; the value of this work is in confronting them together and in a specific validation protocol.

- **Heterogeneous panel data** is well-studied in econometrics (Arellano-Bond, Hausman-Taylor) and in multi-task and transfer learning. The domain combination (spatial, temporal, and schema heterogeneity at national property scale) has fewer worked examples, but the underlying techniques are established.

- **Distribution shift detection** has an extensive literature in machine learning (covariate shift, dataset shift, concept drift). The post-hoc calibration monitoring protocol here is a pragmatic implementation of known ideas, not a methodological advance.
- **Generalization from dense to sparse data** is a domain adaptation and transfer learning problem. The spatial specifics add design constraints but the problem class is known.

The Methodological Contribution

The closest this project comes to a genuine methodological contribution is in the validation protocol for long-horizon probabilistic spatial models.

Rolling-origin backtests evaluate calibration up to the horizon for which realized values exist. For a system trained through 2021 and evaluated in 2026, that means horizons h=1−5 are available only for older origins (2019–2021), and short horizons dominate the evaluable window. Simulation-based validation provides complementary evidence at longer horizons but requires assumptions about the simulator. Neither mode of evidence alone is sufficient.

The specific finding here is that these two failure modes — strong point accuracy and miscalibrated intervals — are **orthogonal in aggregate metrics and only separable by per-horizon diagnostic decomposition**. A model can beat a naive persistence baseline on median absolute error (high "Wins") in the same setting where its 80% interval covers 40% of realized outcomes. Aggregate accuracy metrics (MdAE, RMSE, directional accuracy) do not detect miscalibrated dispersion. This is not a novel observation in the probabilistic forecasting literature generally [6], but its application to multi-year spatial economic models with heterogeneous data quality has not been systematically evaluated.

The per-horizon diagnostic decomposition established here — interval coverage, horizon scaling, tail accuracy, forecast attenuation, and mean bias as separate diagnostics with separate remediations — is a reusable evaluation framework for any probabilistic spatial forecasting system. It makes visible a failure mode that aggregation conceals, and it provides a taxonomy for distinguishing root causes. Dietze et al. [7] make a parallel argument for ecological forecasting; this work is the spatial-economic analogue of that protocol.

Domain Analogues

The table below maps this system's central problems to analogous problems in ecological and environmental spatial modeling. Each row names a specific problem, its instance here, and the analogous domain with representative literature.

Problem	This system	Ecological analogue + sources
Spatial trajectory forecasting	Property price path per parcel	Species range shift, habitat trajectory. Elith & Leathwick [8]; Araújo & Guisan [9]
Probabilistic output calibration	P10–P90 marginal quantile band	Probabilistic ecological forecasting. Dietze et al. [7]; Clark et al. [10]
Long-horizon validation protocol	Rolling-origin backtest + monitoring	Near-term ecological forecasting. EFI NEON Challenge [11]; ISIMIP [12]
Distributional shift	2021–22 rate shock	Climate-induced range contraction. Parmesan & Yohe [13]; Thuiller et al. [14]
Cross-site latent structure	Shared spatial token paths	Hierarchical ecological models. Gelfand et al. [15]; Cressie & Wikle [16]
Calibration collapse at long horizons	40% coverage at h=3	Long-horizon projection uncertainty. IPBES (2019) [17]; Urban et al. [18]
Source heterogeneity	County schema differences	Integrated SDMs, multi-source occupancy. Isaac et al. [19]; MacKenzie et al. [20]

Table 2: *
Each row is a methodological problem shared across domains. The right column gives specific literature entry points in ecology and environmental science.

10. Monitoring

As new realized values accumulate, the calibration diagnostics from Section 5 are re-run. This is reactive monitoring: it detects that coverage has degraded after enough actuals are available, not in real time. The pattern of test failures is diagnostic: coverage dropping while anchor integrity holds indicates miscalibrated dispersion; anchor integrity failing indicates mean bias from regime drift; horizon scaling failing while short-term metrics hold indicates forecast attenuation.

11. Roadmap

The architectural redesign addresses the structural failure modes identified above:

- **Full-sequence inpainting.** Model the complete property trajectory across all historical and forecast years jointly, eliminating the conditioning bottleneck.
- **Proper scoring rule.** CRPS added to the training objective, penalizing overconfidence directly.
- **Comparable-sales conditioning.** Nearby properties from geographic nearest-neighbor search encoded as explicit attention keys.

- **Macro as a time series.** Macroeconomic series encoded as per-year tokens across the full historical and forecast window.
- **Single unified model.** The two-component architecture becomes one transformer, enabling live scenario sweeps and historical counterfactuals.

References

[1] Y. Gorishniy, I. Rubachev, V. Khrulkov, A. Babenko. *Revisiting Deep Learning Models for Tabular Data.* NeurIPS 2021. arXiv:2106.11959

[2] J. Lee, Y. Lee, J. Kim, A. Kosiorek, S. Choi, Y. W. Teh. *Set Transformer.* ICML 2019. arXiv:1810.00825

[3] A. Tong, N. Malkin, K. Fatras, L. Hutin, Y. Hua, G. Liberti, Y. Bengio, G. Doucet. *Improving and Generalizing Flow Matching.* TMLR 2024. arXiv:2302.00482

[4] Y. Shi, V. De Bortoli, A. Campbell, A. Doucet. *Diffusion Schrödinger Bridge Matching.* NeurIPS 2023. arXiv:2303.16852

[5] M. Albergo, E. Vanden-Eijnden. *Building Normalizing Flows with Stochastic Interpolants.* ICLR 2023. arXiv:2209.15571

[6] T. Gneiting, A. E. Raftery. *Strictly Proper Scoring Rules, Prediction, and Estimation.* JASA 102(477), 2007. doi:10.1198/016214506000001437

[7] M. C. Dietze et al. *Iterative near-term ecological forecasting: Needs, opportunities, and challenges.* PNAS 115(7), 2018. doi:10.1073/pnas.1710231115

[8] J. Elith, J. R. Leathwick. *Species distribution models: ecological explanation and prediction across space and time.* Annu. Rev. Ecol. Evol. Syst. 40, 2009. doi:10.1146/annurev.ecolsys.110308.120159

[9] M. B. Araújo, A. Guisan. *Five (or so) challenges for species distribution modelling.* J. Biogeogr. 33(10), 2006. doi:10.1111/j.1365-2699.2006.01584.x

[10] J. S. Clark et al. *Ecological forecasts: An emerging imperative.* Science 293(5530), 2001. doi:10.1126/science.293.5530.657

[11] Ecological Forecasting Initiative. *NEON Ecological Forecasting Challenge.* https://projects.ecoforecast.org/neon4cast-ci/

[12] ISIMIP Consortium. *Inter-Sectoral Impact Model Intercomparison Project.* https://www.isimip.org/

[13] C. Parmesan, G. Yohe. *A globally coherent fingerprint of climate change impacts across natural systems.* Nature 421, 2003. doi:10.1038/nature01286

[14] W. Thuiller et al. *Climate change threats to plant diversity in Europe.* PNAS 102(23), 2005. doi:10.1073/pnas.0409902102

[15] A. E. Gelfand et al. (eds.) *Handbook of Spatial Statistics.* CRC Press, 2010.

[16] N. Cressie, C. K. Wikle. *Statistics for Spatio-Temporal Data.* Wiley, 2011.

[17] IPBES. *Global Assessment Report on Biodiversity and Ecosystem Services.* IPBES Secretariat, 2019. doi:10.5281/zenodo.3831673

[18] M. C. Urban et al. *Improving the forecast for biodiversity under climate change.* Science 353(6304), 2016. doi:10.1126/science.aad8466

[19] N. J. Isaac et al. *Data integration for large-scale models of species distributions.* Trends Ecol. Evol. 35(1), 2020. doi:10.1016/j.tree.2019.08.006

[20] D. I. MacKenzie et al. *Estimating site occupancy rates when detection probabilities are less*

than one. Ecology 83(8), 2002. doi:10.1890/0012-9658

[21] K. E. Case, R. J. Shiller. *The Efficiency of the Market for Single-Family Homes.* AER 79(1), 1989. NBER w2506

[22] Florida Department of Revenue. *Property Tax Data Portal.* https://floridarevenue.com/property/Pages/DataPortal.aspx

[23] NYC Department of Finance. *Real Property Assessment Data (RPAD).* https://data.cityofnewyork.us/

[24] Texas Geographic Information Office. *StratMap Land Parcels.* https://data.geographic.texas.gov/

[25] Harris County Appraisal District. *Public Data.* https://hcad.org/

[26] U.S. Census Bureau. *American Community Survey 5-Year Estimates.* https://www.census.gov/programs-surveys/acs

[27] Federal Reserve Bank of St. Louis. *Federal Reserve Economic Data (FRED).* https://fred.stlouisfed.org/

Daniel Hardesty Lewis · Founder & ML Engineer, Homecastr · homecastr.com

Multi-Horizon Backtest Results

Full backtest results. Bold values meet targets: MdAE $< 10\% \times \sqrt{h}$, Coverage 65−95%, Wins $> 50\%$. Coverage reports the fraction of realized values inside the P10–P90 marginal quantile band at each horizon.

Jurisdiction	Origin	h	MdAE	Cov 80%	Wins
NYC RPAD	2019	1	21.2%	89.5%	34.3%
NYC RPAD	2020	1	**7.4%**	91.5%	**52.0%**
NYC RPAD	2021	1	11.9%	64.9%	29.9%
NYC RPAD	2022	1	**8.0%**	94.8%	**62.9%**
NYC RPAD	2023	1	**7.1%**	91.9%	**56.0%**
NYC RPAD	2024	1	**7.7%**	90.8%	37.5%
NYC RPAD	2025	1	**7.0%**	87.2%	**56.7%**
ACS Nationwide	2019	1	**7.9%**	86.5%	19.7%
ACS Nationwide	2020	1	**7.2%**	**82.2%**	18.0%
ACS Nationwide	2021	1	16.5%	34.8%	15.5%
ACS Nationwide	2022	1	**8.2%**	72.3%	29.3%
ACS Nationwide	2023	1	**7.3%**	71.2%	41.7%
TxGIO Texas	2021	1	22.9%	1.5%	3.4%
TxGIO Texas	2022	1	34.0%	55.7%	4.8%
TxGIO Texas	2023	1	22.6%	57.7%	9.0%
TxGIO Texas	2024	1	**5.7%**	93.6%	22.6%

Table 2. One-year horizon backtest results for the v12sb production model. Bold values meet the diagnostic targets from the calibration table. Full multi-horizon results are in Appendix A.

Table A1. NYC RPAD backtest results across all available origin years and forecast horizons. Bold values meet diagnostic targets.

Table A2. ACS Nationwide backtest results across all available origin years and forecast horizons. Bold values meet diagnostic targets.

Table A3. TxGIO Texas backtest results across all available origin years and forecast horizons. Bold values meet diagnostic targets.

Jurisdiction	Origin	h	MdAE	Cov 80%	Wins
NYC RPAD	2019	1	21.2%	89.5%	34.3%
NYC RPAD	2019	2	34.8%	**80.7%**	6.0%
NYC RPAD	2019	3	35.8%	85.4%	11.6%
NYC RPAD	2019	4	39.7%	93.8%	10.1%
NYC RPAD	2019	5	48.6%	89.1%	12.4%
NYC RPAD	2020	1	**7.4%**	91.5%	**52.0%**
NYC RPAD	2020	2	**8.9%**	89.2%	**58.0%**
NYC RPAD	2020	3	10.4%	90.1%	**69.8%**
NYC RPAD	2020	4	13.0%	88.0%	**81.6%**
NYC RPAD	2020	5	12.6%	90.2%	**72.0%**
NYC RPAD	2021	1	11.9%	64.9%	29.9%
NYC RPAD	2021	2	11.2%	87.4%	50.0%
NYC RPAD	2021	3	14.4%	86.5%	**58.1%**
NYC RPAD	2021	4	15.5%	86.8%	**50.7%**
NYC RPAD	2021	5	16.1%	87.9%	**59.1%**
NYC RPAD	2022	1	**8.0%**	94.8%	**62.9%**
NYC RPAD	2022	2	11.4%	91.8%	**78.2%**
NYC RPAD	2022	3	11.5%	94.5%	**68.5%**
NYC RPAD	2022	4	12.6%	94.0%	**73.6%**
NYC RPAD	2023	1	**7.1%**	91.9%	**56.0%**
NYC RPAD	2023	2	**8.0%**	94.4%	**56.2%**
NYC RPAD	2023	3	**9.8%**	94.2%	**63.6%**
NYC RPAD	2024	1	**7.7%**	90.8%	37.5%
NYC RPAD	2024	2	**8.5%**	96.7%	**51.8%**
NYC RPAD	2025	1	**7.0%**	87.2%	**56.7%**

Jurisdiction	Origin	h	MdAE	Cov 80%	Wins
ACS Nationwide	2019	1	**7.9%**	86.5%	19.7%
ACS Nationwide	2019	2	12.2%	**82.0%**	32.2%
ACS Nationwide	2019	3	28.9%	55.1%	31.8%
ACS Nationwide	2019	4	36.3%	48.6%	**63.2%**
ACS Nationwide	2019	5	42.8%	52.9%	**87.9%**
ACS Nationwide	2020	1	**7.2%**	**82.2%**	18.0%
ACS Nationwide	2020	2	20.5%	49.2%	**70.0%**
ACS Nationwide	2020	3	29.1%	45.7%	**75.9%**
ACS Nationwide	2020	4	35.6%	46.8%	**93.2%**
ACS Nationwide	2021	1	16.5%	34.8%	15.5%
ACS Nationwide	2021	2	21.8%	42.1%	**86.4%**
ACS Nationwide	2021	3	30.2%	40.4%	**88.8%**
ACS Nationwide	2022	1	**8.2%**	72.3%	29.3%
ACS Nationwide	2022	2	12.7%	68.6%	**87.1%**
ACS Nationwide	2023	1	**7.3%**	71.2%	41.7%

Jurisdiction	Origin	h	MdAE	Cov 80%	Wins
TxGIO Texas	2021	1	22.9%	1.5%	3.4%
TxGIO Texas	2021	2	47.4%	11.1%	3.6%
TxGIO Texas	2021	3	48.9%	4.8%	2.5%
TxGIO Texas	2021	4	47.3%	5.0%	2.3%
TxGIO Texas	2022	1	34.0%	55.7%	4.8%
TxGIO Texas	2022	2	44.5%	58.1%	4.0%
TxGIO Texas	2022	3	42.7%	60.6%	4.8%
TxGIO Texas	2023	1	22.6%	57.7%	9.0%
TxGIO Texas	2023	2	35.4%	50.4%	6.9%
TxGIO Texas	2024	1	**5.7%**	93.6%	22.6%