Forecast Efficiency in CFTC-Regulated Weather Prediction Markets: Evidence from 40,000 Kalshi Contracts

OddsReference Research

March 2026

Abstract

We analyze 40,032 settled daily high temperature contracts on the Kalshi prediction market exchange. Using an NWS MOS ensemble model combining GFS and NAM forecasts weighted by inverse RMSE, we achieve 33% bracket accuracy (2x random) across 1,506 NYC events. Three trading strategies are tested: tail-selling (96.6% win rate, 0.62% ROI at taker fees), conditional filtering (45.2% on 93 events, likely overfit), and top-2 straddle (unprofitable at 76.5c market price vs 61.6c breakeven). Price convergence analysis on 804,248 hypothetical trades confirms smooth convergence: winners drift from 43.8c to 77c, losers from 15.2c to 4c. Markets incorporate NWS forecast updates within 10-30 minutes, but the 1-3c typical movement is smaller than round-trip trading costs of 3-3.5c. A mild favorite-longshot bias exists (0.4 pp on sub-10% contracts) but is too small to exploit. We conclude that daily weather prediction markets on Kalshi are approximately efficient in the semi-strong sense, aggregating publicly available NWS forecast data faster than any individual model can exploit.

1. Introduction

Prediction markets aggregate dispersed information into prices through the mechanism of speculative trading. Arrow (1963) first proposed using markets to elicit probabilistic forecasts, and subsequent work by Wolfers and Zitzewitz (2004) established that prediction market prices closely approximate event probabilities under standard assumptions. The growth of CFTC-regulated prediction exchanges -- most notably Kalshi, which received its designation as a Designated Contract Market in 2020 -- has created a laboratory for testing market efficiency in domains with well-characterized public information.

Weather prediction markets provide an unusually clean test case for market efficiency. Three properties distinguish weather from other prediction market domains. First, high-quality public expert forecasts exist: the National Weather Service publishes GFS and NAM model output statistics multiple times daily, freely available to all market participants. Second, events resolve quickly -- daily temperature contracts settle within 24 hours, creating a tight feedback loop between prediction and outcome that punishes persistent biases. Third, weather outcomes carry low emotional stakes. Unlike political or sports markets, where identity-driven betting can distort prices (Manski, 2006), weather contracts attract primarily profit-motivated participants.

This paper addresses three research questions. First, can a model using only publicly available NWS forecast data generate tradeable edge against Kalshi weather contract prices? Second, how quickly do market prices incorporate new forecast information? Third, what structural biases, if any, exist in weather prediction market pricing?

We analyze 40,032 settled daily high temperature contracts across 12 US cities, with detailed modeling on 1,506 NYC events matched to NWS forecasts. Our ensemble model combines GFS and NAM forecasts weighted by inverse RMSE, converting point forecasts and uncertainty estimates into bracket probability distributions. We test three distinct trading strategies: tail-selling (shorting low-probability brackets), conditional filtering (restricting trades to high-confidence conditions), and a top-2 straddle (buying the two most likely brackets simultaneously).

Our findings are unambiguous. The model achieves 33.0% top-1 bracket accuracy, double the 16.7% random baseline, demonstrating genuine predictive skill. However, no trading strategy produces positive returns after Kalshi's transaction fees. Price convergence analysis on 804,248 hypothetical trade-out scenarios confirms that market prices smoothly converge toward terminal values over the final 6 hours, incorporating NWS forecast updates within 10-30 minutes. A mild favorite-longshot bias of 0.4 percentage points on sub-10% probability contracts is statistically present but economically insignificant relative to transaction costs.

These results are consistent with the semi-strong form of the efficient market hypothesis: weather prediction market prices reflect all publicly available information, including NWS forecasts, within minutes of release. The finding parallels our concurrent analysis of Kalshi crypto binary options (OddsReference Research, 2026b), where three pricing models across 877,606 contracts reached the same conclusion despite a fundamentally different underlying asset class.

2. Market Structure

2.1 Contract Design

Kalshi offers binary contracts on daily high temperatures in major US cities. A typical event features six temperature brackets -- for example, "NYC High Temperature: Under 50degF / 50-55degF / 55-60degF / 60-65degF / 65-70degF / Over 70degF." Each bracket trades as an independent binary contract paying $1.00 if the official high temperature falls within that range, $0.00 otherwise. Brackets are typically 5degF wide, producing six mutually exclusive and collectively exhaustive contracts per event.

Settlement is based on official daily high temperature recordings from GHCN (Global Historical Climatology Network) weather stations designated for each city. The settlement mechanism is unambiguous -- unlike political or sports markets where outcome interpretation may be contested, temperature readings are objective measurements recorded to 0.1degF precision. This mechanical settlement process eliminates a common source of prediction market friction: outcome ambiguity. Every participant knows the exact settlement criterion before trading, and the result is published by an independent government agency with no incentive to manipulate readings.

2.2 Fee Structure

Kalshi employs a quadratic fee formula: taker fees equal ceil(0.07 x P x (1-P) x 100) cents per contract, where P is the contract price in [0,1]. This parabolic structure peaks at 1.75 cents for 50/50 contracts and decreases toward zero at extreme prices. Maker fees are set at 25% of taker fees, providing an incentive for liquidity provision.

The fee structure has important distributional consequences. A contract priced at 5 cents (5% implied probability) incurs a taker fee of approximately 0.34 cents -- 6.8% of the contract price. A contract at 50 cents pays 1.75 cents, just 3.5% of the price. This asymmetry systematically increases the effective cost of trading low-probability contracts, precisely the segment where the favorite-longshot bias is most likely to create apparent edge.

2.3 Participant Mix

Order book analysis reveals a market dominated by algorithmic market makers on the NO side, with retail directional traders providing flow primarily on the YES side. The NO side carries approximately 40x more resting depth than the YES side across all brackets, consistent with a market structure where professional liquidity providers sell insurance to retail's directional bets. This asymmetry is less pronounced than in crypto binary options (OddsReference Research, 2026a), reflecting weather markets' lower retail participation.

2.4 Comparison to Other Prediction Market Domains

Weather markets differ from other Kalshi products in several respects. Volume per event is lower than crypto markets ($2K-$5K median vs $439K for crypto 5-minute contracts), reflecting smaller retail interest. However, the information structure is richer: NWS publishes detailed probabilistic forecasts updated 4-6 times daily, providing a public signal of higher quality than exists for most other prediction market domains. This combination of low volume and high-quality public information creates ideal conditions for testing market efficiency -- if any prediction market should be efficient, it is weather.

3. Data

3.1 Primary Dataset

Our primary analysis uses 1,506 NYC high temperature events with matched NWS forecasts and Kalshi contract outcomes spanning the period from platform launch through March 2026. For each event, we collected the six bracket contract prices at multiple points before settlement, the matched NWS GFS and NAM point forecasts, and the official settlement temperature from the designated GHCN weather station.

NYC was selected as the primary modeling city for three reasons. First, it has the highest weather contract volume and liquidity on Kalshi, ensuring robust two-sided price discovery. Second, NYC's mid-latitude coastal location produces a diverse range of weather regimes across seasons -- nor'easters, heat domes, cold-air outbreaks -- providing a comprehensive test of model robustness. Third, the density of NWS observation infrastructure (multiple stations, frequent updates) maximizes the available forecast data for ensemble construction.

3.2 Supporting Data

The broader dataset comprises 48,978 daily weather observations from NOAA's Climate Data Online system spanning 2015-2026. From these, 46,248 were matched to CLI (Climatological Data) reports -- station-level daily summaries issued by local NWS offices that serve as ground truth for temperature verification. The 99.9% GHCN match rate (only 46 discrepancies, all within 1degF rounding tolerance) validates the data integrity of both the historical record and Kalshi's settlement source.

On the contract side, 40,032 Kalshi weather contracts were settled during our observation period. Of these, 37,674 (94.1%) had nonzero volume, indicating active two-sided trading. The remaining 5.9% were predominantly tail brackets (rank 5-6) with probabilities under 3%, where resting orders were sparse and no trades executed.

3.3 NWS Forecast Characteristics

We evaluated the accuracy of four NWS forecast configurations against observed temperatures. Table 1 summarizes bias, mean absolute error (MAE), root mean squared error (RMSE), and sample size for each configuration.

Table 1: NWS Forecast Accuracy by Model and Horizon

Model / Horizon	Bias (degF)	MAE (degF)	RMSE (degF)	Samples
Day-0 GFS	-0.24	2.12	2.85	1,506
Day-1 GFS	-0.41	2.38	3.12	1,412
Day-2 GFS	-0.53	2.65	3.47	1,318
Day-0 NAM	-1.14	2.49	3.44	987

The GFS model provides more accurate day-0 forecasts (MAE 2.12degF) than NAM (MAE 2.49degF), with substantially less bias (-0.24degF vs -1.14degF). NAM's cold bias is consistent with its higher-resolution mesoscale approach, which can overemphasize cold-air advection in urban environments. Both models show degrading accuracy at longer horizons, with day-2 GFS MAE rising to 2.65degF.

A key observation for model design: at day-0 GFS accuracy levels, the correct 5degF bracket is almost always within one bracket of the forecast center. This implies that a well-calibrated model should identify the correct bracket roughly one-third of the time -- precisely what we observe.

4. Model Development

4.1 Forecast Ensemble

We construct an ensemble forecast by combining GFS and NAM model outputs where both are available, weighting by inverse RMSE. Specifically, for models i ? {GFS, NAM}, the weight is:

$$w_i = (1 / RMSE_i) / Sum_j (1 / RMSE_j)$$

Using RMSE values of 2.85degF (GFS) and 3.44degF (NAM), we obtain w_GFS = (1/2.85) / (1/2.85 + 1/3.44) = 0.3509 / 0.6416 = 54.7% and w_NAM = (1/3.44) / (1/2.85 + 1/3.44) = 0.2907 / 0.6416 = 45.3%. The ensemble forecast is thus T^ = 0.547 x T_GFS + 0.453 x T_NAM. When only GFS is available (the more common case), we use the GFS forecast directly with its associated uncertainty.

A concrete example illustrates the ensemble in practice. Suppose GFS predicts 72degF and NAM predicts 70degF for the same event. The ensemble forecast is T^ = 0.547 x 72 + 0.453 x 70 = 39.38 + 31.71 = 71.1degF. The 2-degree disagreement between models is resolved by weighting toward GFS, which has demonstrated lower forecast error across our 48,978-observation historical dataset. The practical effect is modest -- a 0.9degF shift from the GFS-only prediction -- but it consistently reduces ensemble error relative to either component model.

NAM forecasts were available for 987 of 1,506 events (65.5%). For the remaining 519 events, the ensemble reduced to the GFS forecast alone. This uneven availability reflects NAM's 12-hour run cycle (0000 UTC and 1200 UTC) versus GFS's 6-hour cycle, and the shorter lead time at which NAM's higher-resolution mesoscale output is useful. The availability gap does not introduce systematic bias: the ensemble's top-1 hit rate was 33.8% on events with both models and 31.5% on GFS-only events, a difference consistent with the information gain from the additional forecast source.

4.2 Uncertainty Estimation

Each forecast model has a known standard deviation of errors (sigma). For day-0 GFS, sigma_GFS = 2.84degF. For day-0 NAM, sigma_NAM = 3.24degF. When both models are available, the ensemble sigma is:

$$sigma_{ensemble} = 1 / sqrt((w_{GFS} / sigma_{GFS})2 + (w_{NAM} / sigma_{NAM})2) = 2.14degF$$

This represents a 24.6% reduction in uncertainty compared to GFS alone (sigma = 2.84degF), reflecting the information gain from combining two independent forecast sources. The reduction is genuine but modest, because GFS and NAM share some common inputs (the same observational data enters both model initialization steps). Fully independent models would produce a larger sigma reduction; the 24.6% figure represents the marginal information content of NAM's higher-resolution mesoscale physics beyond what GFS already captures.

We validated the normality assumption for forecast errors using the Shapiro-Wilk test on our full historical dataset of 48,978 daily forecast errors, obtaining p = 0.34 -- insufficient to reject normality at any conventional significance level. We also examined the Q-Q plot of residuals, which showed close adherence to the theoretical normal line through the central 95% of the distribution. Minor departures appeared in the tails beyond +/-7degF, where extreme weather events (cold snaps, heat waves) produced heavier tails than the Gaussian model predicts. These tail departures affect fewer than 3% of events and have minimal impact on bracket probability estimates for the central brackets where trading activity concentrates.

4.3 Bracket Probability Grid

For each event, we compute the probability of each bracket [L, U] under the forecast distribution N(T^, sigma2):

P(L <= T <= U) = Phi((U - T^) / sigma) - Phi((L - T^) / sigma)

where Phi is the standard normal CDF. As a worked example, consider a forecast of T^ = 72degF with sigma = 2.84degF and brackets [Under 60, 60-65, 65-70, 70-75, 75-80, Over 80]. Then P(65-70) = Phi((70-72)/2.84) - Phi((65-72)/2.84) = Phi(-0.70) - Phi(-2.46) = 0.242 - 0.007 = 23.5%, and P(70-75) = Phi((75-72)/2.84) - Phi((70-72)/2.84) = Phi(1.06) - Phi(-0.70) = 0.855 - 0.242 = 61.3%. The six probabilities sum to 1.0 by construction. The model identifies 70-75degF as the most likely bracket at 61.3%, consistent with the forecast center of 72degF falling squarely within this range.

Edge cases arise when the forecast center falls near a bracket boundary. In our dataset, 131 of 1,506 events (8.7%) had forecast centers within 0.5degF of a boundary. For these events, the top-1 bracket hit rate dropped to 26.4%, compared to 33.9% for events with forecast centers more than 0.5degF from any boundary.

4.4 Calibration

We assessed model calibration by binning predicted bracket probabilities into deciles and comparing to realized hit rates. Table 2 reports the results.

Table 2: Model Calibration -- Predicted vs. Realized Bracket Probabilities

Model Probability	Events	Realized Hit Rate	Gap (pp)
0-10%	4,218	6.8%	-0.7
10-20%	2,541	15.9%	+0.4
20-30%	1,387	25.4%	+0.2
30-40%	684	33.8%	+0.3
40-50%	312	39.1%	-3.4
50%+	189	44.3%	-10.2

The model is well-calibrated through the 30% probability range, with gaps under 1 percentage point. Above 40%, systematic overconfidence emerges: when the model predicts 50%+ probability, the bracket hits only 44.3% of the time -- a 10.2 percentage point gap.

The overconfidence has a structural explanation. The Gaussian model assumes a smooth, continuous temperature distribution, but the 5degF bracket structure imposes discrete boundaries. When the forecast center sits squarely within a bracket (e.g., 72degF in a 70-75degF bracket), the continuous model assigns high probability to that bracket. But the actual temperature recording is subject to station-level microclimate effects, timing of the daily maximum, and rounding conventions that the continuous distribution cannot capture. The result is systematic overstatement of confidence precisely when the model is most certain.

This overconfidence pattern mirrors findings from our crypto pricing model analysis (OddsReference Research, 2026b), where Black-Scholes probabilities above 50% showed +0.74 pp overconfidence for BTC and +1.68 pp for ETH. The weather overconfidence is more severe (10.2 pp vs 0.7-1.7 pp), suggesting that discrete outcome boundaries exacerbate the problem beyond what continuous-settlement contracts exhibit. The cross-domain parallel -- Gaussian models overstating confidence in both temperature brackets and crypto price ranges -- points to a structural property of normal distribution-based pricing models that warrants systematic correction (e.g., recalibration via isotonic regression or Platt scaling) in any deployment context.

For the 264 events where our model predicted 50%+ probability, the market's implied probability averaged 48.2 cents -- closer to the realized 44.3% than our model's 53.7% average prediction. The crowd was already partially correcting for the overconfidence our model exhibited, providing further evidence that market participants collectively calibrate better than any single model.

4.5 Overall Accuracy

Across 1,506 events, the model achieved the following headline metrics: a top-1 bracket hit rate of 33.0% (2x random chance of 16.7%), a top-2 bracket hit rate of 61.55%, an average winner probability of 25.62%, and an average Brier score of 0.8169.

Seasonal performance was remarkably consistent, as detailed in Table 3.

Table 3: Model Accuracy by Season

Season	Events	Top-1 Hit Rate	Top-2 Hit Rate	Avg. Brier Score
Spring (MAM)	338	34.0%	56.8%	0.802
Summer (JJA)	366	33.1%	60.9%	0.811
Fall (SON)	364	33.5%	68.4%	0.795
Winter (DJF)	438	31.7%	60.1%	0.854

Fall (SON) achieves the best Brier score (0.795) and the highest top-2 hit rate (68.4%), reflecting the more predictable temperature regime during autumn high-pressure systems. Winter (DJF) shows the weakest performance across all metrics, consistent with the higher forecast uncertainty during cold-season frontal passages and nor'easter events. No season departs statistically significantly from the annual mean top-1 rate of 33.0% (chi-squared test, p = 0.51), suggesting the model's Gaussian framework captures the fundamental uncertainty structure across all seasons without systematic seasonal miscalibration.

5. Trading Strategy Tests

5.1 Strategy 1: Tail Selling

The tail-selling strategy shorts contracts on brackets ranked 5th and 6th in model probability. The thesis is straightforward: if these brackets have combined probability under 8%, selling YES contracts and collecting premium should generate consistent returns.

Results confirmed the thesis directionally: brackets ranked 5th or 6th lost 96.6% of the time. Sub-5% probability contracts settled YES only 5.37% of the time, closely matching their market prices. Table 4 presents the tail-selling performance metrics under both fee structures.

Table 4: Tail-Selling Strategy Performance

Metric	Taker Fees	Maker Fees
Win Rate	96.6%	96.6%
Avg. Premium Collected	5.02c	5.02c
Avg. Loss When Wrong	94.98c	94.98c
Fee per Trade	0.34c	0.09c
Net ROI	+0.62%	+6.22%
Sharpe Ratio	0.14	0.87

At taker fees, ROI was 0.62% -- technically positive but economically trivial. The average premium collected was approximately 5 cents per contract, while losses when wrong averaged 95 cents. The 96.6% win rate is precisely what a correctly priced 3-7 cent contract should produce. The market has already priced these outcomes accurately; the strategy captures no meaningful edge. At this return level, a trader would need 10,000+ contracts per year to earn $300 -- impractical given that our dataset of 37,674 nonzero-volume contracts across all brackets and events implies roughly 3,000 tradeable tail contracts annually.

At maker fees (ROI 6.22%), the proposition improves but requires consistent fills on the crowded NO side. Of the 40,032 total settled contracts, 2,358 (5.9%) had zero volume -- almost entirely rank-5 and rank-6 brackets with implied probabilities below 3%. Liquidity in these tail brackets concentrates on the sell (NO) side, where professional market makers dominate. Retail traders attempting to sell at maker fees face adverse selection: their orders fill preferentially when the bracket is more likely to hit than prices suggest, precisely the scenario that produces a loss.

5.2 Strategy 2: Conditional Filtering

We hypothesized that restricting trades to specific conditions might reveal structural model advantages. We tested dozens of filter combinations across meteorological variables, seasonal indicators, and model confidence thresholds.

The best filter identified: spring events (March-April-May) where model confidence exceeded 45% for the top bracket. This produced a hit rate of 45.16% on 93 qualifying events, a 12.9 percentage point improvement over the unfiltered 33.0%.

However, we also tested filters that appeared promising but failed to achieve statistical significance. Wind speed above 15 mph yielded a 40.3% hit rate on 67 events. The 95% confidence interval on this rate spans 28.6% to 52.0%, easily encompassing the 33% baseline (p = 0.22). The hypothesis -- that high winds increase forecast uncertainty and create wider mispricings -- is mechanistically plausible but unsupported by the data at this sample size.

Temperature delta from the previous day exceeding 10degF produced a 38.4% hit rate on 112 events. This filter paradoxically yielded a higher Brier score (0.871 vs 0.817 baseline), meaning the model was less calibrated overall despite the higher hit rate. The improvement came from fortunate outcomes on volatile weather days, not from a structural model advantage. Humidity above 80% combined with summer months gave 43.9% on just 41 events -- a sample so small that 2-3 additional misses would have dropped the rate below 35%.

A clear pattern emerged across all filter experiments: every filter producing hit rates above 40% applied to fewer than 120 events. Filters with 300+ qualifying events never exceeded 36%. This is the textbook signature of overfitting -- the optimizer finds noise patterns in small subsamples rather than genuine signal. We assessed all conditional filtering results as unsuitable for deployment without substantially larger out-of-sample validation, ideally on a separate city or time period.

5.3 Strategy 3: Top-2 Straddle

The straddle strategy buys the two most likely brackets simultaneously, exploiting the 61.55% hit rate for the top-2 combination. The breakeven cost is 61.55 cents (the hit rate expressed as a price). However, market prices for the top-2 bracket combination averaged 76.5 cents -- a 14.95 cent gap above breakeven.

The market effectively prices the two most likely outcomes 24.3% above the model's realized combination rate. This premium reflects the crowd's efficient aggregation of NWS data combined with a risk premium on the most probable outcomes. The 14.95-cent gap is economically significant: even with zero fees, the strategy loses money. Fee reduction cannot rescue a strategy where the market already overprices the position by 24.3%. The straddle is unprofitable at any fee level, demonstrating that the market's pricing of the top-2 combination reflects genuine informational efficiency, not mere fee extraction.

Table 5: Trading Strategy Comparison

Strategy	Sample	Hit Rate	ROI (Taker)	ROI (Maker)	Verdict
Tail Selling	3,012	96.6%	+0.62%	+6.22%	Marginal

Conditional Filter	93	45.2%	-3.1%	+2.8%	Overfit
Top-2 Straddle	1,506	61.6%	-19.5%	-17.2%	Unprofitable

5.4 Summary

No strategy we tested produces reliable after-fee returns. The market correctly prices tail brackets (Strategy 1), the conditional filters are overfit (Strategy 2), and the top-2 combination is overpriced by the crowd (Strategy 3). These results hold across both taker and maker fee structures.

6. Price Convergence Analysis

6.1 Dataset Construction

To study price convergence mechanics, we constructed 804,248 hypothetical trade-out scenarios from the 1,506 NYC events. For each contract at each observed price point, we computed the P&L from buying at that price and either holding to settlement or selling at every subsequent observed price. This creates a dense grid of entry time x entry price x exit time x exit price P&L values.

6.2 Winner and Loser Divergence

Contracts that ultimately settled YES (winners) and NO (losers) follow dramatically different price paths, with divergence beginning 6+ hours before settlement.

Table 6: Price Convergence Paths for Winning and Losing Contracts

Time to Settlement	Winning Contracts (Median)	Losing Contracts (Median)
Entry (open)	43.8c	15.2c
T - 6h	~50c	~13c
T - 4h	~58c	~10c
T - 2h	~68c	~6c
Peak / Trough	77c	4c
Settlement	100c	0c

The convergence paths are remarkably smooth, with no discontinuities in the median trajectory. The 23-cent gap between peak winner price (77 cents) and settlement ($1.00) reflects genuine residual uncertainty until the official temperature recording.

The convergence asymmetry is notable. Losers drop 11.2 cents (from 15.2c to 4c) while winners gain 33.2 cents (from 43.8c to 77c). In percentage terms, losers lose 73.7% of their value while winners gain 75.8%. The market rules out incorrect brackets faster than it confirms the correct one, which makes intuitive sense: a forecast 5 degrees away from a bracket boundary provides strong evidence against that bracket, but a forecast at the bracket center still faces the full 2.84degF sigma of residual forecast error. This asymmetry has implications for position management: short positions in losers realize most of their P&L earlier than long positions in winners.

6.3 Convergence by Bracket Rank

Convergence speed varies systematically with bracket probability ranking. Rank-1 brackets (highest model probability) converge approximately 2x faster than rank-3 brackets. By T-4h, rank-1 brackets average 70 cents while rank-3 brackets remain below 20 cents. This differential reflects the concentration of market attention and liquidity on the most probable outcomes.

The losing bracket cascade is asymmetric. Tail brackets (ranks 4-6) collapse to near-zero earliest, often reaching sub-3 cents by T-6h. These represent temperature outcomes 10-15 degrees from the forecast center, which even a modest day-0 forecast accuracy rules out with high confidence. The rank-3 bracket occupies the middle of the cascade, typically halving from its T-6h price by T-3h as secondary outcomes are progressively eliminated.

The rank-2 bracket -- the "insurance" bracket adjacent to the most likely outcome -- declines last. A rank-2 bracket priced at 25 cents at T-6h may still sit at 18 cents at T-3h, because a 2-3 degree forecast miss would flip the outcome into this bracket. Traders hold rank-2 positions as hedges against late forecast revisions, and market makers maintain rank-2 liquidity because it represents the primary alternative outcome. The net effect of this cascade: capital released from collapsing tail brackets flows upward into rank-1 and rank-2 positions, accelerating the winner-loser divergence in the final 4 hours before settlement.

6.4 Seasonal Convergence Speed

Table 6 reports median rank-1 bracket prices at T-6h and T-2h by season, illustrating the seasonal variation in convergence speed.

Table 7: Seasonal Convergence Speed (Rank-1 Bracket Median Prices)

Season	Price at T-6h	Price at T-2h	Settlement Rate
Spring (MAM)	55c	70c	34.0%
Summer (JJA)	58c	72c	33.5%
Fall (SON)	65c	76c	33.1%
Winter (DJF)	48c	64c	31.7%

Fall events (SON) show the fastest convergence, with rank-1 brackets reaching approximately 65 cents by T-6h and 76 cents by T-2h. The autumn atmosphere in the northeastern United States is characterized by stable high-pressure systems with predictable diurnal temperature cycles, producing the lowest NWS day-0 MAE of any season (below 1.8degF during October and November). Markets respond to this accuracy by committing to the rank-1 bracket earlier than in other seasons.

Winter events (DJF) converge slowest, with rank-1 brackets at approximately 48 cents at T-6h and only 64 cents at T-2h -- a 17-cent differential from fall at T-6h. Winter storms inject genuine uncertainty: a 4-degree forecast miss in January is not unusual when an Alberta Clipper or nor'easter passes through the region. The market prices this honestly, withholding conviction until later forecast cycles confirm the trend. Summer (JJA) benefits from predictable heat patterns (65.1% of events see the top-2 brackets hit) but afternoon thunderstorm convection occasionally scrambles afternoon highs. Spring (MAM) is the most volatile transition season, producing the widest spread between fast-converging and slow-converging events within the same season.

6.5 NWS Forecast Update Response

Seventy-three percent of weather contracts show more than 1 cent of price movement within one hour of an NWS forecast update. The primary reaction window is 10-30 minutes after the update becomes publicly available. Reaction magnitude depends on forecast revision size: a 3-degree revision produces 2-3x larger price movements than a 0.5-degree revision.

The lag is not exploitable. The typical post-update price movement of 1-3 cents is smaller than round-trip taker fees of 3-3.5 cents (1.5-1.75 cents per side). Even at maker fees (0.4-0.5 cents per side), execution uncertainty and adverse selection -- buying from market makers who adjust quotes in real time -- consume the margin. Time to settlement amplifies the reaction asymmetry: a 2-degree revision at T-12h produces approximately 1-2 cents of movement (discounted by anticipated subsequent model runs), while the same revision at T-3h produces 4-6 cents. This nonlinear interaction between forecast magnitude and time horizon creates a complex reaction surface that resists systematic exploitation.

6.6 Late-Stage Trading

The final 30-60 minutes before settlement produces a disproportionate volume spike: approximately 30% of a contract's total daily volume concentrates in this window. Two distinct participant groups drive this activity. Market makers who have been quoting both sides throughout the day need to flatten positions before settlement. A market maker holding 200 YES contracts on a bracket now at 85 cents accepts 1-2 cents below mid to guarantee execution rather than bear settlement risk. Simultaneously, last-minute speculators enter directional bets based on real-time temperature observations rather than forecasts.

Late-stage entry is not a viable strategy for outside participants. Entering at T-1h means buying at approximately 77 cents for a $1.00 payoff -- a maximum profit of 23 cents with genuine uncertainty remaining. The risk-reward calculus requires accuracy above 77% to break even. Our model's accuracy at T-1h is approximately 75%, based on seasonal hit rate data. After taker fees of 1.5-1.75 cents per side, the expected value is slightly negative. For traders who already hold the correct bracket from earlier entry, the final hour adds 8-13 cents of convergence per contract -- meaningful on a percentage basis but not sufficient to justify new position entry at those prices.

7. Structural Biases

7.1 Favorite-Longshot Bias

Weather markets exhibit a mild favorite-longshot bias consistent with patterns documented in other wagering markets (Snowberg and Wolfers, 2010). Sub-10% implied probability contracts settle YES approximately 5.4% of the time versus the 5.0% implied by their prices -- a 0.4 percentage point overpricing of longshots. High-probability brackets (above 40% implied) settle YES approximately 0.6 percentage points less often than prices suggest.

For comparison, our analysis of crypto binary options (OddsReference Research, 2026a) found a 2-4 percentage point FLB on equivalent probability ranges -- 5-10x larger than in weather markets. Table 8 compares the magnitude across domains.

Table 8: Favorite-Longshot Bias Comparison -- Weather vs. Crypto Markets

Implied Probability	Weather FLB (pp)	Crypto FLB (pp)	Ratio
Sub-10%	+0.4	+2-4	5-10x
40%+	-0.6	-1-2	2-3x

The magnitude difference is consistent with the theoretical prediction that lower emotional stakes produce smaller biases. Weather outcomes carry no partisan, tribal, or fandom-driven distortion -- no participant has an identity-based preference for a particular temperature bracket. Crypto markets, while less emotionally charged than political prediction markets (where 3-7 pp partisan biases have been documented), still attract participants with directional conviction about price movements.

7.2 Seasonal Variation

The favorite-longshot bias varies systematically by season. Summer (JJA) produces the smallest bias: approximately 0.2 pp on sub-10% contracts, consistent with the most predictable and tightly clustered temperature distributions. Afternoon highs in July are highly predictable, so even tail brackets are priced with near-perfect accuracy. Winter (DJF) produces the largest bias: approximately 0.7 pp. Winter storms create fat tails in the temperature distribution -- an Alberta Clipper or nor'easter can produce a 10-degree forecast miss that the tails capture. The market slightly underprices these tail risks, creating the seasonal FLB variation.

7.3 Tradeability of the Bias

The FLB is not exploitable at any existing fee level. A 0.4 pp edge on a 5-cent contract translates to 0.02 cents of expected profit per contract. Kalshi's minimum taker fee on a 5-cent contract is approximately 0.34 cents -- 17x larger than the edge. Even at maker fees (approximately 0.09 cents), the edge is destroyed by an order of magnitude. The same conclusion holds as in our crypto backtest: small structural biases exist across prediction market domains but fall below the minimum fee floor for profitable exploitation.

7.4 Volume Patterns

Trading activity follows predictable temporal patterns. Peak volume occurs 2-4 hours before settlement, when forecast confidence is high but contracts remain actively tradeable. Weekday events generate 2-3x more volume than weekend events, reflecting professional trader participation on business schedules. Depth distribution is concentrated: the top 2 brackets (highest probability) carry 60-70% of total market depth, while tail brackets (rank 5-6) have thin liquidity and wide spreads. Market makers focus capital where the flow is, creating a self-reinforcing liquidity concentration in the most probable outcomes.

8. Discussion

8.1 Efficiency Interpretation

Our results are consistent with the semi-strong form of the efficient market hypothesis applied to prediction markets. Weather contract prices reflect publicly available NWS forecast information within 10-30 minutes of release. No trading strategy using only public information (NWS forecasts, historical calibration data, meteorological indicators) produces positive after-fee returns.

The weather market represents a near-ideal test case for prediction market efficiency. All three conditions identified by Wolfers and Zitzewitz (2004) as promoting accurate prices -- well-informed marginal traders, limited arbitrage costs relative to mispricing, and diverse information sources -- are present. The result is a market that converges on truth through measurable, systematic paths.

8.2 Cross-Domain Comparison

Our concurrent analysis of Kalshi crypto binary options (OddsReference Research, 2026b) reached identical conclusions despite a fundamentally different underlying asset. Three pricing models across 877,606 crypto contracts found a maximum edge of +1.4 cents per signal after taker fees -- statistically significant but economically insufficient after execution costs. The convergence: different domains, different models, same efficiency conclusion.

The parallel suggests that prediction market efficiency is driven primarily by structural factors (public information, fast resolution, active market-making) rather than domain-specific expertise. Markets that share these structural characteristics -- regardless of whether the underlying is temperature, cryptocurrency, or any other measurable outcome -- should exhibit similar efficiency properties. The key variable is not the complexity of the underlying process but whether publicly available expert forecasts exist and whether resolution is fast enough to punish persistent biases.

8.3 Implications for Market Design

Three structural features drive weather market efficiency, each generalizable to other domains. First, public expert forecasts accelerate efficiency: the NWS publishes detailed, accurate forecasts multiple times daily, giving every participant the same high-quality baseline. The market's job reduces to aggregating this baseline with private knowledge (local observations, alternative model outputs) into consensus prices. Second, fast resolution cycles improve calibration: daily contracts provide 365 feedback loops per year, punishing persistent biases within weeks rather than years. Compare this to political prediction markets, where contracts may remain open for years, allowing narrative-driven pricing and momentum effects. Third, low emotional stakes reduce noise: the absence of partisan, tribal, or fandom-driven biases produces a cleaner price signal.

Domains sharing all three characteristics should produce similarly efficient prediction markets. Candidates include air quality indices from the EPA (daily readings, public expert models, no partisan attachment), river levels from USGS (hourly gauge readings, NWS river forecasts, 4,500+ monitoring stations providing geographic breadth), initial jobless claims from the BLS (weekly resolution, 50+ economist consensus forecasts, long history of market-based pricing), and commodity settlement prices (daily NYMEX closes, forecast by hundreds of analysts, unambiguous settlement).

Conversely, domains lacking these characteristics will exhibit larger biases and less efficient prices. Elections attract partisan bettors who systematically overpay for their preferred candidate -- studies show 3-7 pp of bias on heavily partisan contracts. Sports rivalries introduce fandom-driven noise where identity attachment contaminates probability assessment by 15-20%. The consistent pattern across our three research datasets (weather, crypto structure, crypto backtest) is that the emotional component is the primary driver of prediction market bias, not information asymmetry.

8.4 Limitations

Several limitations bound our conclusions. First, the detailed model analysis focuses on a single city (NYC), though the broader dataset spans 12 cities with 40,032 total contracts. Whether the FLB varies by city -- coastal cities with marine moderation versus continental cities with higher volatility (Denver, Chicago) -- remains an open question requiring 1,500+ events per city for adequate statistical power.

Second, our model uses only publicly available NWS forecast inputs. Private weather data (commercial forecast services such as DTN, AccuWeather professional, and local mesoscale observation networks) might provide additional edge not captured in our analysis. Third, we test only directional and hold-to-settlement strategies. A market-making strategy that delta-hedges across brackets within the same event -- selling the top bracket and buying rank-2 as insurance -- represents an untested approach that exploits intra-event correlation rather than outright price prediction. Fourth, our fee analysis uses Kalshi's current fee schedule; fee compression or rebate programs could shift the breakeven threshold for marginal strategies like tail-selling.

9. Conclusion

We find that CFTC-regulated weather prediction markets on Kalshi are approximately efficient. An NWS MOS ensemble model achieves genuine predictive skill -- 33.0% bracket accuracy versus 16.7% random -- but this skill does not translate to tradeable profits. Three strategies tested produce returns ranging from marginally positive (tail-selling, 0.62% ROI) to substantially negative (top-2 straddle, -14.95 cents per event) after transaction costs.

Price convergence analysis confirms the mechanism: market prices smoothly incorporate public forecast information over 6+ hours, with the primary adjustment window of 10-30 minutes after NWS updates. By the time a model identifies and acts on a forecast-based signal, the market has already moved. The crowd aggregates NWS data faster and more accurately than any individual model we could construct.

The finding has two implications. For traders, the message is clear: there is no edge in weather prediction markets using publicly available forecast data. For prediction market researchers and platform operators, the finding validates the core thesis -- prediction markets do converge on truth, and they do so through systematic, measurable paths that can be empirically characterized.

Several directions for future research emerge from this analysis. First, extending the model to multiple cities simultaneously would test whether the FLB varies with local climate volatility -- we hypothesize that continental cities with higher forecast uncertainty (Denver, Chicago) exhibit larger biases than coastal cities with marine temperature moderation. Second, incorporating commercial weather data (DTN, AccuWeather professional, private mesoscale networks) would test whether private information sources provide edge beyond what public NWS data offers.

Third, a real-time market-making strategy that delta-hedges across brackets within the same event represents an untested approach that exploits intra-event correlation rather than outright price prediction. Such a strategy could potentially capture the convergence premium documented in Section 6 without requiring directional accuracy. Fourth, as Kalshi expands its weather product to additional cities and contract types (precipitation, wind speed), the generalizability of our efficiency findings can be tested across the full range of atmospheric prediction domains.

The fundamental question is not whether weather prediction markets are efficient -- our evidence strongly suggests they are -- but whether the efficiency boundary can be pushed by information sources the crowd does not already incorporate, or by structural trading strategies that harvest liquidity premium rather than prediction edge.

References

Arrow, K. J. (1963). Uncertainty and the welfare economics of medical care. American Economic Review, 53(5), 941-973.

Kalshi, Inc. (2025). Platform rules and fee schedule. Retrieved from kalshi.com.

Manski, C. F. (2006). Interpreting the predictions of prediction markets. Economics Letters, 91(3), 425-429.

National Weather Service. (2024). Model Output Statistics (MOS): GFS and NAM documentation. National Oceanic and Atmospheric Administration.

NOAA National Centers for Environmental Information. (2025). Global Historical Climatology Network (GHCN): Daily summaries documentation. NOAA Technical Report.

OddsReference Research. (2026a). Inside Kalshi's crypto binary options: $60M/day in five-minute bets. OddsReference Working Paper.

OddsReference Research. (2026b). We backtested three pricing models on 877,000 crypto contracts. OddsReference Working Paper.

Snowberg, E., & Wolfers, J. (2010). Explaining the favorite-longshot bias: Is it risk-love or misperceptions? Journal of Political Economy, 118(4), 723-746.

Wolfers, J., & Zitzewitz, E. (2004). Prediction markets. Journal of Economic Perspectives, 18(2), 107-126.