Discrimination, Not Recalibration: What a Weather Market Adds to the Official Forecast*

ORHAN TORUL

August 31, 2026

Abstract

A trader pricing a weather contract can consult the official forecast first, so a market-versus-agency horse race need not compare forecasts formed from independent information sets. The better question is what each adds given the other. On 15,782 settled Kalshi daily-maximum temperature contracts at seven American stations, market prices beat the National Weather Service's operational percentile ladder by 0.0295 in Brier score. Conditional on the ladder, the price adds 0.0327 of held-out discrimination; conditional on the price, the ladder adds 0.0000. The gap is not a calibration deficit a user could repair: perfect recalibration of the official ladder would recover at most 8% of it, and no post-processing specification fitted here recovers more than 6.9% out of sample. The market's advantage is discrimination. It separates days that the official forecast prices alike.

Keywords: probability forecasting; forecast evaluation; forecast encompassing; prediction markets; proper scoring rules; calibration and discrimination.

JEL Codes: C53; G13; G14; Q54.

*Address: Boğaziçi University, Department of Economics, 34342 Bebek, Istanbul, Türkiye. E-Mail: orhan.torul@bogazici.edu.tr. Data availability and the replication archive are described at the end of the paper. I hold no position in any contract studied here and have never traded on this exchange. All errors are mine.

1 Introduction

When market transaction prices score better against realised outcomes than an official forecast, the comparison by itself settles very little. Two probability forecasts can differ in accuracy for reasons whose implications for a user run in opposite directions, and the accuracy gap alone cannot say which is at work.

In the first case the price separates outcomes more sharply than the published forecast does, and no map of the stated probability can close the gap. In the second the two are equally discriminating and the official forecast is simply miscalibrated: its stated probabilities do not deliver the frequencies they promise, the market's do, and any user could capture the whole gap with a regression run on last year's forecasts. Only the second is a defect a user can repair, and the repair is cheap. The distinction is the standard one between calibration and discrimination (Murphy, 1973; Gneiting et al., 2007), and drawing it requires scoring a probability against a probability and then decomposing the difference.

That comparison has not been run before in this literature, because it could not be. Prediction markets have beaten published meteorological forecasts before: Ritter (2013) prices Chicago Mercantile Exchange weather futures against forecasts of the same monthly temperature index and finds the market generally more accurate, building on Ritter et al. (2011). But a futures price and a point forecast are both single numbers, and a single number cannot be split into calibration and discrimination. Daily temperature contracts change that. Kalshi lists binary contracts on the daily maximum at named weather stations, so its prices form a distribution over the settlement variable; the National Weather Service's National Blend of Models publishes, for the same station and the same day, a ladder of percentiles that is also a distribution. Both can be scored against the same settlement record, though the benchmark forecasts an eighteen-hour maximum rather than the contract's local calendar day, a mismatch Section 3 sets out and Appendix C bounds.

I assemble 15,782 settled contracts on seven stations between 2025-05-06 and 2026-06-20 and score both forecasts against the temperature the exchange settled on. The gap is there, as expected: the market's Brier score is 0.1013 against 0.1308 for the official ladder, a paired difference of 0.0295 (standard error 0.0013). It survives clustering on date, a Diebold and Mariano (1995) test with the small-sample correction of Harvey et al. (1997), and it is negative at every station taken separately. Scoring the two implied distributions by the continuous ranked probability score gives the same ordering, 0.917°F against 1.439°F.

The decomposition is where the paper starts, not where it ends. Writing each score as miscalibration minus discrimination plus a common uncertainty term, the two forecasts are close on miscalibration, 0.0011 for the market against 0.0024 for the official ladder, and far apart on discrimination, 0.0456 against 0.0173. The number that carries the argument is the level 0.0024: it is all the official ladder could gain from being recalibrated perfectly, and against a gap of 0.0295 it is small. The share of the gap attributable to discrimination is 96%, but that share has an arithmetic floor of 92.0% that no market could fall below, so the level and not the share is the informative quantity, and Section 4.2 gives the floor before the share. Both point the same way. The market's advantage is not that its probabilities are better tuned; both forecasts deliver roughly the frequencies they promise. It is that the market's probabilities separate days the published ladder prices alike.

Two features of the calculation were not available to earlier work. It uses the CORP decomposition of Dimitriadis et al. (2021), in which the calibrated version of a forecast comes from the pool-adjacent-

violators algorithm instead of bins, so the identity is exact for any forecast and there is no bin count for the author to choose. And the same construction applies at every threshold of the predictive distribution and integrates, which delivers the identical split for the continuous score with no reference to any contract the exchange happened to list. Done that way, discrimination carries 84% of the continuous gap against a floor of 34.5%.

The obvious objection to any market-versus-forecast comparison is that the two do not condition on the same information, and it cannot be answered by advancing the benchmark's vintage, because the difficulty is not that the model run is stale. A trader pricing a contract at midday can consult the published ladder itself, along with the morning's observations and whatever else, so the two forecasts need not be formed from independent information sets. So I stop running the contest. What spoils the horse race is exactly what makes a better question well posed: what does each forecast add given the other? Both go into the conditioning set, $E[Y \mid \cdot]$ is estimated non-parametrically on a training sample, and the discrimination of each conditional mean is evaluated on a disjoint holdout. The price adds 0.0327 of discrimination to a user who already holds the official forecast; the official forecast adds -0.0006 to a user who holds the price. Averaging over 20 folds of repeated grouped cross-fitting gives 0.0277 for the price and 0.0000 for the ladder. The design is calibrated against its own null: permuting the price within station and within decile of the official probability, which leaves it everything it shares with the ladder and removes only what it adds, returns an increment of 0.00004.

That is the paper's central result. It does not put the two forecasts at a common information origin. It measures what the price adds, out of sample, to a forecast that already uses the published ladder, at the hour the contracts traded. The ladder enters the conditioning set by construction, so the exercise never has to establish whether any particular trader consulted it. What it cannot say is where the increment comes from: later observations, better models and local judgment all remain live, and the price cannot distinguish them.

A user holding the published ladder wants to know whether the gap is repairable, and the answer is no. For maps of the stated probability alone the bound is tight: any such map is measurable with respect to a coarser conditioning set, so it can at best drive miscalibration to zero (DeGroot and Fienberg, 1983), which recovers at most 8% of the gap. The bound covers maps of p and only those, so Section 4.6 also fits maps on progressively richer conditioning sets, all of them information a user of the public forecast holds, by monotone gradient boosting. Out of sample none beats the univariate map, and the best of any conditioning set recovers 6.9% of the gap. Combination gives the same answer from the other direction: the Brier-optimal weight on the market sits at or above one on two of three splits, where the unconstrained optimum places no weight on the ladder at all.

The paper connects to two literatures. Work on prediction markets has established that prices aggregate dispersed information well (Wolfers and Zitzewitz, 2004; Leigh and Wolfers, 2006), examined their calibration (Page and Clemen, 2013; Berg et al., 2008) and documented their pricing anomalies (Snowberg and Wolfers, 2010), while Manski (2006) cautions that a price need not equal a mean belief, which is why the test here is against outcomes, not stated beliefs. The meteorological literature has scored probability forecasts since Brier (1950) and Murphy (1973), formalised the scoring rules (Gneiting and Raftery, 2007), put the decomposition on firmer ground (Bröcker, 2009), made it estimable without binning (Dimitriadis et al., 2021), and developed the post-processing (Gneiting et al., 2005) and distributional diagnostics (Hamill, 2001) used below. Two recent papers sit closest. Le (2026) measures how calibration on these

exchanges varies with domain, horizon and trade size across 353 million trades, and finds short-horizon weather to be the one domain where prices are too extreme rather than compressed; Appendix A reconciles this sample with that finding. Luo (2026) asks whether these particular contracts can transfer weather risk rather than predict it, and two of its findings corroborate what follows: the realised high escapes the exchange's bucket ladder on 17% of days in its sample, against 14.6% here, and isotonic and Platt recalibration of the market's prices does not improve them out of sample. Section 4.6 runs the mirror image of that exercise on the official forecast and reaches the same verdict. Recalibrating either forecast on its own stated probability leaves the gap standing, and the decomposition below says why: the residual is discrimination, which no map of a stated probability can reach.

Section 2 describes the contracts and the settlement record. Section 3 builds the benchmark and is candid about the two respects in which it does not match the contract. Section 4 reports the gap, decomposes it, establishes what each forecast adds given the other, asks what a user could recover, and closes with what the market appears to be correcting. Section 5 concludes. Four appendices carry the supporting evidence: the scoring-rule and calibration details, the behaviour of the holdout estimator under a null, the timing, measurement and heterogeneity checks, and a ledger collecting every specification the paper runs.

2 Contracts, prices and settlement

Kalshi lists a ladder of binary contracts on the daily maximum temperature at each of several American stations. A ladder for a given city-day contains narrow interval contracts, typically two degrees wide, together with contracts paying above the highest interval and below the lowest. Each contract pays one dollar if the realised maximum falls in its region and nothing otherwise. I read each price as a probability forecast for the corresponding binary event, which at a one-day horizon on a claim of this size is a mild assumption but not a free one; Manski (2006) is the standard caution, and the object scored throughout is the market-implied forecast, not any trader's belief. I take the seven stations with continuous coverage: Central Park in New York, Chicago Midway, Los Angeles International, Miami International, Denver, Austin Bergstrom and Philadelphia International. The sample runs from 2025-05-06 to 2026-06-20 and contains 16,212 settled contracts on 2,877 city-days, of which 15,782 on 2,800 city-days also carry a benchmark and form the estimation sample. The contracts that drop out are those on days when the archived model run is missing or the trailing bias correction is not yet estimable, which is unrelated to how they traded.

Prices come from the trade tape, not from closing quotes. For each contract I take the volume-weighted average of all trades between 12:00 and 15:00 UTC on the contract's own date. A fixed clock window matters here: the last print of a market is dated at whatever moment trading happened to stop, and that moment differs systematically between contracts that stayed live into the afternoon and contracts that collapsed to zero in the morning. Appendix C repeats the exercise in six windows through the day.

Contracts settle on the National Weather Service climatological report for the named station, and I reconstruct each payoff from the corresponding daily maximum in NCEI's archived daily record (NOAA National Centers for Environmental Information, 2026), then check the reconstruction against the exchange's own resolution, contract by contract. The two agree on every contract but one, a boundary

case at Miami, for an agreement rate of 0.9999. The check is worth performing because the obvious shortcut fails it: daily maxima derived from hourly observations, which is what most free interfaces supply, run cold by 0.75°F on average, enough on two-degree buckets to resolve 12.8% of contracts to the wrong side. Anyone building a panel of these contracts should verify the settlement source against the exchange first.

These are actively traded contracts. The sample contains 11.2 million trades and $99 million of notional volume; the median contract prints 349 trades over its life, 93.5% of listed contracts trade inside the pricing window, and the median single trade there is $1.14. The prices being scored are neither stale nor thin.

One further descriptive fact bears on how the prices are used below. The interval buckets do not partition the outcome space: they span about seven degrees, and the settlement falls outside them on 14.6% of city-days, so the analysis needs the two tail contracts wherever it needs a distribution. The object that does partition is the full ladder, buckets and tails together; exactly one contract pays on 2,872 of 2,877 city-days, and the exceptions are city-days on which a listed tail never traded inside the pricing window. A city-day is called *complete* below when it carries at least three interval buckets and both tails, which 2,151 of them do.

Priced, that ladder comes to rather more than a dollar: on complete city-days the prices sum to a median of 1.059, exceeding one on 91.3% of them. An overround of 6.1% is what a volume-weighted average of executed trades should show on an exchange that charges the aggressor, since a taker fee and a crossed spread both push the traded price of every contract up. It is not an executable arbitrage, because the trades that generate it did not occur at the same instant, and it sits close to uniform across the ladder, not concentrated in the rungs a bookmaker would load. Where the analysis needs a coherent distribution, each city-day's full ladder is rescaled to sum to one; where it scores contracts one at a time the raw transaction price is used and the rescaled version reported alongside. Appendix A returns to what the rescaling does to the calibration slope, the one place where it changes a reading.

3 The official benchmark

The benchmark is the National Blend of Models, the operational statistical guidance the National Weather Service issues for public forecasting (Craven et al., 2020). It blends deterministic and ensemble output from several numerical models, post-processes the result, and is the product from which the public-facing forecast is built. It is guidance: the forecast the public actually reads carries adjustment by the local forecast office on top of this blend, and that adjustment is not in the benchmark. A comparison against the adjusted product is not available, because NOAA's open archive of it mirrors current files and keeps no dated history. Two things follow. The comparison bounds what a user of the published probabilistic guidance forgoes, not what a user of the forecast-office product forgoes, and the sign of the difference between those two is unknown. And the site-corrected station bulletin scored in Appendix C, the closest available object, does not close the gap in discrimination.

For each run NBM publishes the 1st through 99th percentiles of the 18-hour maximum temperature on a 2.5 kilometre grid. I read versions 4.3 and 5.0 from NOAA's public archive (NOAA National Weather Service, 2026), take the 00Z run, whose maximum window spans the daylight hours of the target day, read the percentile ladder at the grid cell containing each station, and interpolate the implied cumulative

distribution to the contract's strike. Interpolation is the only modelling step, but it is a step: what is scored is a researcher-implemented reading of the agency's operational gridded guidance, not the local forecast-office product.

Three caveats about what this holds fixed. First, a grid cell is not a station, and the ladder carries a warm bias at these locations on the order of one degree. I report the benchmark raw and shifted by the trailing twenty-station-day median error, one parameter estimated strictly backward looking. The shift moves the benchmark's Brier score from 0.1337 to 0.1308, a tenth of the 0.0295 gap, and does not touch the ordering; Appendix C pursues the point with the agency's own site-corrected bulletin, which needs no interpolation at all. Second, the benchmark's predictand is not quite the contract's. NBM forecasts the maximum over an 18-hour window running from 12Z on the target day to 06Z the next, while the contract settles on the calendar-day maximum in local standard time, so the window misses the small hours of the target day. The natural repair, a calendar-day predictand built from the archive, is not available: the archive carries exactly one valid time per cycle at a fixed 30-hour offset, so the uncovered pre-dawn hours appear in no archived field. Appendix C instead reports the restriction to days whose maximum falls inside the window, wherever the headline appears. Third, the market and the benchmark do not condition on identical information at any hour. A trader pricing a contract at midday has seen the morning temperatures and later model runs; the 00Z ladder has not, and neither would a fresher cycle repair the deeper problem, which is that the trader has also seen the ladder. Section 4.4 takes that seriously enough to change the estimand.

I also carry a second benchmark for contrast: the raw Global Ensemble Forecast System, 31 members at half-degree resolution, post-processed by the author with nonhomogeneous Gaussian regression in the sliding-window form of Gneiting et al. (2005). This is what a researcher can build without the agency's own post-processing, and the distance between it and the operational product is itself instructive.

4 Results

4.1 The accuracy gap

Table 1 establishes the gap the rest of the section decomposes. Over the 15,782 contracts on 2,800 city-days for which both forecasts exist, the market scores 0.1013 and the official ladder 0.1308; the raw and bias-adjusted versions of the benchmark score 0.1337 and 0.1308, so nothing turns on the location correction. The paired difference is 0.0295 (standard error 0.0013 clustered on city-day, 0.0014 clustered on date), and the market is the more accurate forecast on 86% of days. The Diebold and Mariano statistic is 14.8 on 400 daily loss differentials. It is computed on daily aggregates instead of all 15,782 contracts because the contracts within a city-day are rungs of one ladder and their loss differentials are close to perfectly dependent; treating them as 15,782 observations would overstate the test's information by roughly the number of rungs. Weighting each city-day equally instead of each contract leaves the gap at 0.0295 (0.0017).

Panel B scores the two implied distributions instead of the contracts one at a time. A Brier score over a partition of an ordered variable is indifferent to how far wrong a concentrated forecast is; the continuous ranked probability score is not. The market scores 0.917°F against 1.439°F on 2,784 city-days, a daily paired difference of 0.5218°F with a Newey-West standard error of 0.0523 and a Diebold and Mariano statistic of 10.0 on 400 daily differentials. Inference is taken on the daily differential, not the stacked

Table 1: Forecast accuracy: market prices against operational guidance

Forecast	Score	Skill vs. climatology	Observations
Panel A. Binary contracts, Brier score			
Kalshi market price	0.1013	0.486	15,782
NBM operational, bias adjusted	0.1308	0.336	15,782
NBM operational, raw	0.1337	0.321	15,782
Constant base rate	0.1457	0.260	15,782
Day-of-year climatology	0.1969	0.000	15,782
Panel B. Full predictive distributions, continuous ranked probability score (°F)			
Kalshi market ladder	0.917	0.799	2,784
NBM percentile ladder	1.439	0.685	2,784
Day-of-year climatology	4.572	0.000	2,784

Lower is better for both scores. Panel A covers 15,782 settled daily-high temperature contracts on 7 United States stations, 2025-05-06 to 2026-06-20; Panel B covers the 2,784 city-days on which both sides quote a complete distribution, and its unit is degrees Fahrenheit. The market price is the volume-weighted average of trades between 12:00 and 15:00 UTC on the contract's own date. The operational benchmark is the National Blend of Models 18-hour maximum-temperature percentile ladder from the 00Z run, interpolated to each contract's strike; the adjusted row shifts the ladder by the trailing 20-station-day median error, meaning the median of the signed errors of the station's own previous 20 forecasts, estimated strictly backward looking. The climatology is a Gaussian in the smoothed day-of-year mean and standard deviation of 30 years of NCEI daily maxima at the same stations, ending before the sample begins. It is a worse Brier forecast than the constant base rate because the exchange lists strikes around the current forecast, so an unconditional distribution is confidently wrong about which rung pays; the constant base rate absorbs that by construction. The paired differences in favour of the market are 0.0295 on the Brier score and 0.523 on the continuous score, with Diebold–Mariano statistics of 14.8 and 10.0 respectively, each computed on 400 daily loss differentials with Newey-West errors at ten lags. Standard errors are in the text.

city-day panel, for the same reason as in Panel A, and because a lag window applied to observations ordered by station and then by date runs across station boundaries while capturing none of the same-day dependence a continental weather pattern induces.

The two win rates differ in an informative way. The market is more accurate on 86% of days by Brier and on only 69% of city-days by the continuous score, while its relative improvement runs the other way: it cuts the ladder's continuous score by 36% and its Brier score by 23%. The market wins less often under the continuous score and by more when it wins. Its distributional advantage is concentrated in avoiding large misses, not in shaving small ones, which is what a discrimination advantage should look like.

Building the market's distribution requires the tail contracts as well as the interval buckets, since a cumulative distribution built from the buckets alone is truncated exactly where it matters. Beyond the outermost quoted point each distribution is completed linearly to zero and one over six degrees, the same completion applied to both forecasts. Six degrees is a convention, and nothing turns on it: completing over three degrees gives a paired difference of 0.510 and over twelve degrees 0.475, against 0.523 at six. Two further checks are in Appendix A: the ordering holds at essentially every cost-loss ratio a user might have, not only under the quadratic score, and both forecasts show large skill against a climatological reference.

Table 2: Where the accuracy gap comes from

	Miscalibration		Discrimination		Share of gap
	Market	NBM	Market	NBM	from discrimination
Panel A. The 15,782 contracts, Brier score					
CORP decomposition	0.0011	0.0024	0.0456	0.0173	96%
Bootstrap 95 percent interval					[94%, 99%]
Panel B. The 2,784 city-days, continuous score, degrees Fahrenheit					
CORP integrated over thresholds	0.2498	0.3137	7.358	6.904	88%
Restricted score reconstructed	0.914	1.432			
Panel C. Robustness of Panel A					
Ladder renormalised to sum to one	0.0015	0.0024	0.0457	0.0173	97%
Murphy, twenty bins, as estimated	0.0006	0.0021	0.0449	0.0169	95%
Murphy, twenty bins, bias corrected	0.0004	0.0019	0.0447	0.0167	95%

The CORP decomposition of Dimitriadis et al. (2021) writes a proper score as $S = \text{MCB} - \text{DSC} + \text{UNC}$, where the calibrated version of the forecast is obtained by the pool-adjacent-violators algorithm. Miscalibration is what recalibration would remove and lower is better; discrimination rewards separating outcomes from the base rate and higher is better; uncertainty depends only on the events and is common to both forecasts, 0.1457 in Panel A and 8.022 in Panel B. The identity is exact for any forecast, so unlike the binned decomposition it leaves no residual: the two components of the difference reproduce the difference in score exactly in Panel A, and to within 1e-14 in Panel B, which is numerical rather than statistical. The share column is the discrimination difference over the score difference; a share above one hundred percent means the market is the worse calibrated of the two, so that its calibration subtracts from an advantage its discrimination more than supplies. Panel B applies the same construction at every threshold t to the event $\{T \le t\}$ and integrates, with the base rate taken within station so that neither forecast collects discrimination for knowing that Miami is warmer than Denver; its units are degrees Fahrenheit. Its components reconstruct the *restricted* continuous score printed in the second row of the panel, which integrates only over thresholds at which the outcome varies within a station. Thresholds where every day in a station falls on the same side carry score but no information the decomposition can speak to, and they account for the difference from the full scores in Table 1. Subtracting the components from the full score will appear to break the identity; against the restricted score it holds exactly. Panel C reports the same decomposition after rescaling each city-day's full ladder to sum to one (15,777 contracts), and the binned decomposition of Murphy (1973) that earlier drafts led with, before and after the finite-sample correction of Ferro and Fricker (2012). The correction is not common to the two forecasts, because it depends on how each occupies the unit interval: it is 0.00018 for the market against 0.00027 for the ladder. Under six binnings of the Murphy decomposition, ten, twenty and fifty bins of equal width and of equal mass, the share runs from 94% to 98%. The bootstrap resamples city-days with replacement, 400 replications.

4.2 Decomposing the gap

Table 2 is the paper's central decomposition. A proper score splits into a miscalibration term, which penalises stated probabilities that do not deliver the frequencies they promise, minus a discrimination term, which rewards separating outcomes from the base rate, plus an uncertainty term that depends only on the events and is common to both forecasts.

How the split is estimated matters, and the choice available now is better than the one available to Murphy (1973). Murphy's identity is exact only for a forecast taking finitely many values, so applying it to a continuous forecast requires bins, and both terms then move with the bin count. The CORP decomposition of Dimitriadis et al. (2021) removes the choice: the calibrated version of the forecast comes from the pool-adjacent-violators algorithm rather than from a partition of the unit interval, and the identity is then exact for any forecast. Panel A's two components of the difference sum to the difference in score with no residual at all.

The two forecasts are close on miscalibration, 0.0011 for the market against 0.0024 for the official ladder. Both are small; both make probability statements that roughly deliver what they promise. They are far apart on discrimination, 0.0456 against 0.0173, a factor of 2.6. Of the total gap, 96% is discrimination,

with a block bootstrap interval of [94%, 99%].

That share has an arithmetic floor, and the floor belongs before the number. Write the difference in score as

$$\Delta S = \Delta\mathrm{MCB} - \Delta\mathrm{DSC}, \qquad \Delta X = X^{\mathrm{mkt}} - X^{\mathrm{NBM}}, \tag{1}$$

and impose only $\mathrm{MCB}^{\mathrm{mkt}} \geq 0$. The discrimination share then obeys

$$\frac{\Delta\mathrm{DSC}}{\Delta S} \geq 1 - \frac{\mathrm{MCB}^{\mathrm{NBM}}}{\Delta S}, \tag{2}$$

which here is 92.0% whatever the market does. The realised 96% clears it by 3.9%. So the share is informative about the benchmark, not about the market: it says that the official ladder's miscalibration *level*, 0.0024, is small beside a gap of 0.0295. That level is the substantive quantity, and Section 4.6 shows it is also the binding constraint on what a user can repair. Panel C shows that a binned estimator with the finite-sample correction of Ferro and Fricker (2012) lands within two percentage points of the same share; the details are in Appendix A.

4.3 The decomposition without a contract partition

Everything so far conditions on the ladder the exchange happened to list. That is the natural object when the question is which forecast a contract buyer should consult, but it invites the objection that the decomposition is a property of the strikes rather than of the forecasts. The objection can be removed entirely.

Let F be a predictive distribution for the daily maximum T. For any threshold t the event $\{T \leq t\}$ is binary and $F(t)$ is a probability forecast of it, so the same decomposition applies at t. The continuous ranked probability score is the integral of that Brier score over t, so the decomposition integrates term by term, delivering a miscalibration, a discrimination and an uncertainty for the whole predictive distribution, in the units of the variable, with no reference to any strike. Panel B of Table 2 reports it, with the base rate taken within station so that neither forecast collects discrimination for knowing that Miami is warmer than Denver. The construction is close kin to Hersbach (2000); integrating the CORP identity divides Hersbach's potential score further, into the uncertainty of the events and the discrimination the forecast achieves against it, which is the distinction the paper needs. The identity is exact at every threshold; only the integration over thresholds is a Riemann sum, on a half-degree grid fixed by the resolution of the temperature record.

Building the market's distribution needs both tails, and on a quarter of the eligible city-days a tail never traded inside the pricing window. Setting an absent tail to zero truncates the distribution exactly where truncation matters, so the figures quoted in the abstract and conclusion require both tails. On the 2,097 city-days that carry a complete ladder, the market scores 0.948°F against 1.546°F, a paired difference of 0.597°F, wider than on the zero-filled panel, not narrower. Discrimination carries 84% of the gap against a floor of 34.5%. The continuous share is lower than the 96% on the contract partition because it integrates over regions no exchange would list, where the official ladder has more miscalibration to lose; it is also the more informative of the two, because its floor sits far below it.

Table 3: What each forecast adds given the other

	Split at the chronological midpoint	Alternating calendar weeks
Panel A. Held-out discrimination of the best use of each information set		
Conditioning on the official ladder alone	0.0130	0.0187
Conditioning on the price alone	0.0462	0.0456
Conditioning on both	0.0456	0.0457
Discrimination the price adds to the ladder	**0.0327**	**0.0270**
95 percent interval	[0.0293, 0.0365]	[0.0237, 0.0308]
Discrimination the ladder adds to the price	**-0.0006**	**0.0001**
95 percent interval	[-0.0016, 0.0000]	[-0.0005, 0.0006]
Paired difference in score, same holdout	-0.0353	-0.0271
Held-out contracts	8,270	7,887
Panel B. Linear forecast encompassing, in sample, all 15,782 contracts		
Coefficient on the price, alone / with the ladder	1.079 / 1.050	
Coefficient on the ladder, alone / with the price	0.768 / 0.060	
F: the ladder adds nothing / the price adds nothing	7.7 / 3612	

Panel A grants the timing objection rather than contesting it. A trader pricing a contract has seen the published ladder, so the ladder's information set is contained in the trader's and a comparison of the two forecasts is not a contest at a common information set. Putting both in the conditioning set removes the difficulty by construction. Each row reports the held-out discrimination, UNC $-$ S, of the best estimate of $E[Y \mid \cdot]$ on the stated information set: isotonic regression for a single forecast and a monotone gradient-boosted regression for the pair, fitted on the training half and evaluated on the disjoint holdout. Intervals come from 300 bootstrap replications that resample *dates* with replacement in the training and evaluation halves separately and refit the learner inside each replication. Dates rather than contracts or city-days are the resampling unit because contracts on one city-day are rungs of a single ladder and a continental weather pattern moves all seven stations on the same date. The price's increment is above zero in every replication; the ladder's interval contains zero and bounds its increment below a twentieth of the price's. Nothing turns on the second of those choices: across seven estimators of $E[Y \mid p, q]$, including an unconstrained gradient-boosted regression, a random forest, a logistic model quadratic in both forecasts with their interaction, and fully non-parametric averages over an 8×8 and a 12×12 grid of cells, the price adds between 0.0299 and 0.0328 and the ladder between -0.0033 and -0.0005. Uncertainty on the holdout is 0.1405 and 0.1458 in the two columns. The price adds 92% and 100% of the raw paired difference to a user who already holds the official forecast; the ladder adds nothing measurable to a user who holds the price. Because the second column splits on alternating calendar weeks, both its samples span the year. Panel B is the linear counterpart, in the sense of Chong and Hendry (1986): the contract payoff regressed on each probability alone and on both together, with standard errors clustered on city-day. The ladder retains a coefficient of 0.060 (0.022), which is small but not zero, and the restriction that it adds nothing is rejected. Panel A says what that increment is worth out of sample once the functional form is not forced to be linear.

4.4 What the price adds given the ladder

The comparison so far is a horse race, and the race is not quite the object the data support. A trader pricing a contract in the middle of the day can consult the published ladder, the morning's observations and whatever else, so the two forecasts need not be formed from independent information sets. That is a statement about what was available, not a measurement: this design does not observe any trader's information, nor when the ladder was consulted.

Two things follow, and they should be kept apart. The first is a statement about conditional means: a forecast measurable with respect to a coarser conditioning set has weakly lower discrimination, so if the ladder is available to the trader then the conditional expectation given the trader's information weakly dominates the one given the ladder's. The second is a statement about the objects actually scored, and it does not follow from the first. The price is a volume-weighted average of executed trades, not the trader's conditional expectation, and a price can compress or distort what the trader knows; nothing

here establishes that the ladder is recoverable from the price. So the availability of the ladder fixes the direction of a population benchmark, not the ordering of the two forecasts in Table 1, which is therefore reported as descriptive. What it does establish is that the horse race is the wrong question: it compares two forecasts that need not draw on independent information, and it does so without saying what either one contributes.

What replaces it is a narrower question that can be evaluated directly out of sample: not which forecast is better, but what each adds given the other. Put both in the conditioning set, estimate $E[Y \mid \cdot]$ non-parametrically on a training sample, and evaluate the discrimination of each conditional mean on a disjoint holdout. The ladder is then inside the econometric conditioning set by construction, so the exercise asks whether the observed price improves an out-of-sample forecast that already uses the published ladder, with no premise about whether a particular trader consulted it.

Panel A of Table 3 reports the result on 8,270 held-out contracts. Conditioning on the official ladder alone delivers 0.0130 of discrimination; on the price alone, 0.0462; on both, 0.0456. The price adds 0.0327 to a user who already holds the official forecast, 92% of the raw paired Brier difference on the same holdout, with a bootstrap interval of [0.0293, 0.0365] over 300 replications that resample dates and refit the learner in each. Because a chronological split in a sample spanning one annual cycle trains on one set of seasons and evaluates on another, the second column splits on alternating calendar weeks instead; there the price adds 0.0270. Averaging over 20 folds of repeated grouped cross-fitting, with folds grouped on date, the price adds 0.0277 (range 0.0210 to 0.0328, positive in 20 of 20 folds).

The increment in the other direction is a precision check, not a finding. It is bounded below by zero for a reason that has nothing to do with what traders saw: $\{p, q\}$ is a richer conditioning set than $\{p\}$, so in population the conditional mean given both weakly dominates the one given the price alone. The point estimate, -0.0006, is on the wrong side of that bound, which is the estimator's own noise; Appendix B shows the mechanism and repairs it, and the repaired increment is -0.0001. What the exercise delivers is precision: across splits and folds the ladder's increment lies in [-0.0009, 0.0006]. Whatever the official forecast adds to a user holding the price, it is smaller than a twentieth of what the price adds to a user holding the official forecast.

The design is calibrated against its own null. Permuting the price within station and within decile of the official probability leaves the permuted price its marginal distribution, its station, and everything it shares with the ladder, and removes only what it adds. Over 20 permutations the increment is 0.00004 with a maximum of 0.00054, against 0.0327 on the true price: 61 times the largest value the permutation ever produced. Whatever the design is measuring, it is not the dimension of its own conditioning set. Nor is the central number an artefact of the learner: across seven estimators, from gradient boosting through a random forest to fully non-parametric cell averages, the price's increment stays within [0.0299, 0.0328] (Appendix B).

Panel B is the linear counterpart and is reported because it disagrees, informatively. In a linear probability model of the contract payoff on both forecasts the ladder retains a coefficient of 0.060 (0.022) beside the price, and the restriction that it adds nothing is rejected ($F = 7.7$); the converse restriction, that the price adds nothing to the ladder, is rejected on a different scale entirely ($F = 3612$). The official forecast is therefore not statistically redundant, and the distinction that matters is between statistical and operational significance. The coefficient is distinguishable from zero. What it buys a user is not: Panel A is what happens when that increment has to survive out of sample without the functional form being

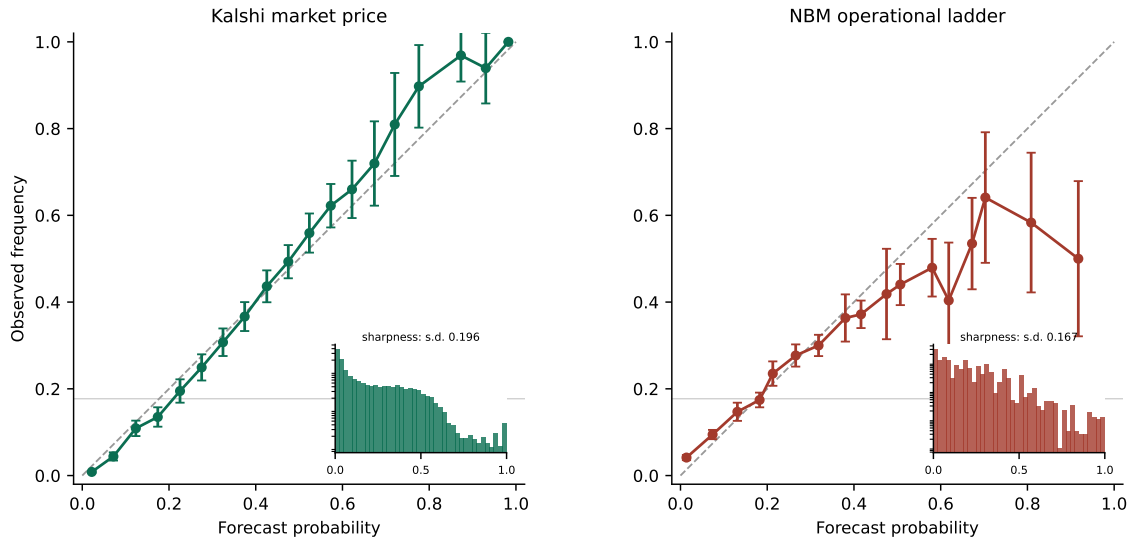


Figure 1: Reliability diagrams with sharpness insets

Observed frequency against forecast probability in bins of width 0.05 holding at least thirty contracts, with 95 percent intervals; the diagonal is perfect calibration and the faint horizontal line is the base rate. The bins are for display only: the decomposition in Table 2 uses none. Insets give the distribution of stated probabilities on a log count scale, so that the sparsely populated interior remains visible alongside the mass near zero and one. Both forecasts track the diagonal closely, which is the calibration result; they differ in how far their probabilities travel from the base rate, which is the discrimination result.

forced to be linear, and Section 4.6 shows that a weight estimated to exploit it makes the forecast worse.

Is the price only a thermometer? The sharpest version of the timing objection is that a trader pricing at midday has watched the morning's temperatures arrive while the overnight run has not, so what the price adds is nowcasting. That is testable: the hourly station record (Iowa Environmental Mesonet, 2026) contains the observations a trader could actually have seen, so they can go into the conditioning set beside the ladder. They add nothing. On the 8,264 held-out contracts, the observations available before the pricing window opens deliver 0.0003 of discrimination on their own; added to the ladder they are worth -0.0028; and the price still adds 0.0347 on top of the ladder and the observations together. Extending the cut-off to everything observable before the window closes changes nothing. The reason is meteorological: the pricing window opens at 07:00 in New York and 04:00 in Los Angeles, so the observations in hand cover the overnight minimum, not the afternoon. Whatever the market knows about the day's maximum at that hour, it has not read it off a thermometer.

Two limits on what this establishes. It does not identify the source of the increment: later observations, better models and local judgment all remain live, and nothing here shows the agency could not obtain the same information by other means. And the conditioning set is the published ladder rather than everything the agency holds, so the exercise bounds what a user of that product forgoes, not what the agency could achieve.

4.5 Calibration diagnostics

The decomposition says the two forecasts are close on calibration and far apart on discrimination. Figures 1 and 2 show that without the arithmetic, and each concedes something.

In Figure 1 both curves sit near the diagonal, and the visible difference is in the insets: the market's

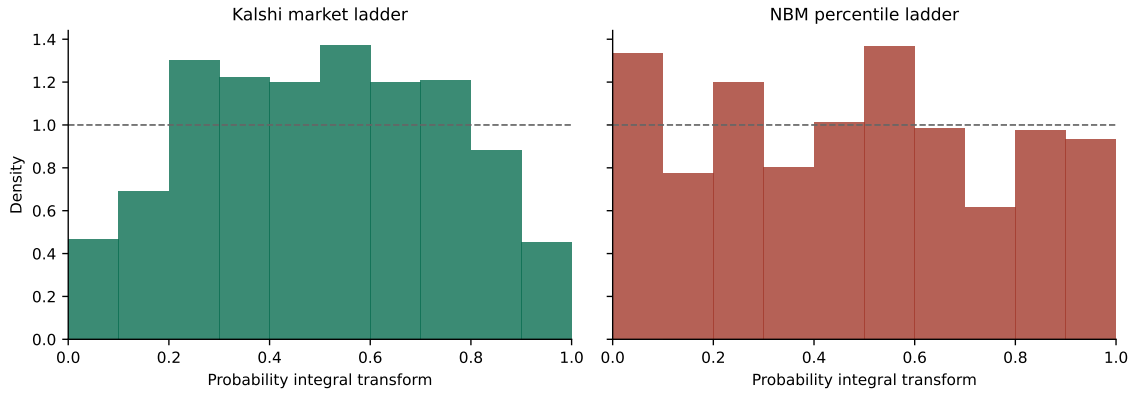


Figure 2: Probability integral transform of the realised maximum

Each forecast's implied cumulative distribution evaluated at the realised daily maximum, over 2,784 city-days. A calibrated predictive distribution returns a uniform transform, marked by the dashed line. The market's transform is concentrated toward the middle relative to uniform, the signature of a distribution slightly too wide; the official ladder's mean of 0.474 indicates a residual warm tilt. Both are close to uniform, which is the distributional counterpart of the calibration result in Figure 1.

stated probabilities are more dispersed, with a standard deviation of 0.196 against 0.167 for the ladder. Sharpness and discrimination are distinct (Gneiting et al., 2007), and the distinction matters, because a proper scoring rule rewards more extreme probabilities only if they are also right. The market is sharper by a factor of 1.17; its edge in discrimination, a factor of 2.6, is far larger. It is not simply being bolder.

The figure also shows where the official ladder's calibration does break down. Above a stated probability of one half its curve falls away from the diagonal: at around 0.9 the observed frequency is closer to 0.5. That region is thinly populated, holding 5.9% of its contracts, and it accounts for 75% of the ladder's binned reliability penalty. The deterioration is real, it is where a user of the published ladder should be most careful, and it is still not what separates the two forecasts: correcting all of the ladder's miscalibration recovers 8% of the gap.

Figure 2 runs the same check one level up, on the distribution, as Hamill (2001) recommends. Neither transform is far from uniform. The market's has a standard deviation of 0.243 against the uniform's 0.289, so its predictive distribution is if anything a little too wide: at the level of the whole distribution the market is the *less* sharp of the two, the price it pays for its near-zero miscalibration. The official ladder's standard deviation is 0.292, with a mean of 0.474, a residual warm tilt. A Kolmogorov–Smirnov test rejects uniformity for both, which at 2,784 city-days it would do for small departures; the informative quantity is the statistic, 0.085 for the market against 0.061 for the ladder. On this diagnostic the official ladder is the marginally better calibrated of the two. That strengthens the argument: the market can be substantially the more accurate forecast while being slightly the worse calibrated one, which is what it means for the advantage to be discrimination.

Appendix A completes the picture: the linear and logistic calibration slopes, the reconciliation with Le (2026), and the tail behaviour of both forecasts, where the market carries a classical favourite-longshot bias and the official ladder a larger defect running the other way.

Table 4: What a user of the official forecast can recover

Share of the held-out gap recovered	Split at the chronological midpoint	Alternating calendar weeks
Panel A. Post-processing the published probability		
Isotonic in p	6.5%	1.8%
Logistic in p with station effects	-0.2%	-3.1%
Isotonic in p, fitted within station	4.2%	5.3%
Boosted in p and station	5.9%	6.9%
and moneyness and contract type	3.9%	4.8%
and the station's trailing error	0.9%	1.2%
and everything together	2.6%	-1.6%
Memorandum: held-out gap	0.0353	0.0271
Held-out contracts	8,270	7,887
Panel B. Combining the official forecast with the price		
Weight on the market, $\hat{\lambda}$, unconstrained	0.959	1.010
Constrained to $[0,1]$	0.959	1.000
Held out, market alone	0.0945	0.1006
Held out, ladder alone	0.1299	0.1277
Held out, combination	0.0947	0.1006
Combination minus market	0.00018	0.00000
	(0.00006)	(0.00001)

Panel A asks how much of the market's advantage a user holding the published ladder could capture by post-processing it. Each row fits a map of the official probability on the stated conditioning set in the training sample and evaluates it on the disjoint holdout; the entry is the reduction in the ladder's held-out Brier score as a share of the held-out gap between the ladder and the price. Moneyness is the distance of the official probability from one half, contract type distinguishes interval buckets from the two tails, and the station's trailing error is the mean absolute error of the official implied mean over the previous ten days at that station, all of it information a user of the public forecast holds. The boosted rows are gradient-boosted regressions constrained to be monotone in p. The exercise matters because the DeGroot and Fienberg (1983) bound in Section 4.6 covers maps of p alone, and conditional miscalibration that varies with a covariate is invisible to it: the richer conditioning sets are where such a miscalibration would show up, and none of them recovers more than the univariate map does. Panel B combines rather than recalibrates, $\hat{\lambda}F^{\mathrm{mkt}} + (1-\hat{\lambda})F^{\mathrm{NBM}}$, with the weight chosen to minimise the Brier score in the training sample. It is estimated *unconstrained*, so a value at or above one means the unconstrained optimum places no positive weight on the ladder. The first column splits at the chronological midpoint, which in a sample spanning one annual cycle trains on one set of seasons and evaluates on another; the second trains on even calendar weeks and evaluates on odd ones, so both samples span the year. Standard errors are clustered on city-day.

4.6 What can a user recover?

Two quantities here are easy to confuse, and the confusion decides how this section is read. The miscalibration *difference* between the forecasts accounts for the small remainder of the gap. The benchmark's miscalibration *level* is 0.0024, and the level is what a recalibration removes, because recalibrating the official ladder eliminates its own deficit without touching the market's.

The level is also a bound, and the bound needs no data. Let p be the official probability and g any function of it. The forecast $g(p)$ is measurable with respect to a coarser conditioning set than p, so by the tower property its discrimination cannot exceed p's, while its miscalibration cannot fall below zero. The improvement is therefore at most the miscalibration of p, attained only with the calibration function

itself in hand:

$$\sup_{g} \frac{S(p) - S(g(p))}{\Delta S} = \frac{\mathrm{MCB}^{\mathrm{NBM}}}{\Delta S} = 8\%. \tag{3}$$

This is the calibration-refinement structure of DeGroot and Fienberg (1983). The right-hand side is the complement of the floor in (2), and deliberately so: the statement that discrimination carries 96% of the gap and the statement that recalibration can recover 8% of it are one fact about the official ladder's miscalibration level, seen from two sides. They are not two findings, and the paper does not count them as two.

The bound covers maps of p and only those, and a miscalibration that varies with a covariate is invisible to it. Two such candidates turn up elsewhere in this paper: the ladder's bias reverses sign with whether a contract is still trading late in the day (Appendix C), and the gap is widest in the days after the ladder has been running wrong (Section 4.8). So Panel A of Table 4 runs the uncovered case. I fit maps of the official probability on progressively richer conditioning sets, every one built from information a user of the public forecast holds, and evaluate each on a disjoint holdout. Isotonic recalibration in p recovers 6.5% of the held-out gap. A logistic map with station effects recovers -0.2%. Monotone gradient boosting on the probability and the station recovers 5.9%; adding moneyness and contract type, 3.9%; adding the station's own trailing forecast error, the covariate Section 4.8 identifies as the one that predicts where the ladder underperforms, 0.9%; everything together, 2.6%. The best figure any conditioning set achieves on either split is 6.9%.

The pattern is not that the richer maps help a little. None improves on the univariate map out of sample, and the richest overfits. That is a stronger result than the bound, because it is the bound's uncovered case, and it is bounded in turn by what was tried: it says that these covariates, on these splits, do not recover the gap. It does not rule out other covariates, other training windows, or a target-matched benchmark.

Combination gives the same answer from the other direction. Panel B reports the Brier-optimal weight on the market, estimated unconstrained in a training sample and applied to a disjoint holdout: 0.959 on the chronological split, 1.010 on alternating weeks, 1.038 when the objective is the continuous score. Two of the three sit at or above one, where the unconstrained optimum places no positive weight on the ladder at all and the constraint to the unit interval returns the price alone. The chronological column is the exception: a weight of 0.959 leaves the ladder a small positive share, and what that share buys, reported two rows down, is negative. Refitting the weight every day on an expanding window gives a mean of 0.968 and a combination scoring 0.0985 against 0.0984 for the market alone. A linear pool of two calibrated forecasts is itself underconfident (Ranjan and Gneiting, 2010), so the pool used here is a benchmark, not the best available combination; since the estimated weight already sits at or above one on the market, sharpening the pool (Gneiting and Ranjan, 2013) moves it towards the price, not away.

The section therefore establishes something specific. A user holding the official probabilities, their own history of them, and every covariate that history supports recovers a few percent of the market's advantage and no more. The deflationary reading of the accuracy comparison, that a public forecast is merely poorly tuned and a market has tuned it better, does not survive.

4.7 Timing, measurement and heterogeneity in brief

Appendix C works through the remaining explanations a reader will reach for, and none overturns the ordering or its composition. Four results are worth stating here.

Vintage: advancing the benchmark to the 12Z cycle, which initialises at the moment the pricing window opens, improves it by 0.0040 over the overnight run, so the design detects an assimilation gain when there is one to detect; the market still beats it by 0.0251 on 15,543 contracts. Intraday: computed window by window through the trading day, the gap is already -0.0211 in the earliest window, before the target day's heating begins, and widens thereafter. Measurement: the agency's own site-corrected station bulletin, which needs no interpolation, improves the benchmark's reliability but not its discrimination, and the market's discrimination exceeds the bulletin's by 0.0310; restricting to the 95% of station-days whose maximum falls inside the benchmark's forecast window leaves a gap of 0.0239, so the predictand mismatch carries 19% of the headline difference. Heterogeneity: the gap is negative and precisely estimated at all seven stations and in all four seasons, survives dropping any station, is present in the thinnest volume tercile, and is largest where the outcome is genuinely in doubt.

4.8 What is the market correcting?

The decomposition cannot say what the market is discriminating on, but the price can be made to say more. Both forecasts imply a distribution over the day's maximum and therefore a mean, and the difference between those means is the correction the market applies to the official forecast. It averages 1.40°F. Table 5 asks what that correction contains and what it is worth.

The correction predicts most of the official forecast's realised error. Regressed on the correction alone, that error has an R^2 of 0.624, with an intercept of -0.016 (0.031) and a slope of 1.053 (0.020). Under forecast encompassing the intercept would be zero and the slope one; the intercept is indistinguishable from zero and the slope sits above one. Seven stations are too few for cluster asymptotics, so the slope restriction is also tested by a null-imposed wild cluster bootstrap on station, enumerated over all 2^7 sign vectors: the bootstrap p-value is 0.039. The market slightly over-corrects; it is not merely undoing what the model got wrong. Out of sample, a coefficient estimated on the first half of the sample cuts the official forecast's mean absolute error from 2.20°F to 1.27°F on held-out city-days, closing 99% of the distance to the market's own implied mean. That sits beside the decomposition without contradicting it: a forecast whose mean moves on the right days is discriminating, and the decomposition shows the discrimination extends beyond the mean.

Could the station's own public error history reproduce the correction? The benchmark for that has to be a real post-processor, so I use an autoregression of order three in the station's realised errors together with its trailing ten-day and thirty-day means, all strictly backward looking. It reproduces 15% of the correction's variance, and residualising the correction on that history and re-entering it leaves a coefficient of 1.054 (0.021): the part of the market's adjustment that the public record cannot reproduce predicts the official forecast's error as well as the whole of it does. Roughly 85% of the correction is orthogonal to the station's recent performance. That is a statement about an autoregressive post-processor, not about everything a user could have done; satellite imagery, hourly observations, neighbouring stations and other models are all outside this comparison, and nothing here identifies the source of the correction.

One pattern is the closest thing to a mechanism the paper has, and it concerns Los Angeles. Four facts accumulate in four different places: it has the largest paired gap of any station (-0.0413); the lowest

Table 5: What the market's correction to the official forecast contains

Dependent variable: official forecast error	(1)	(2)	(3)	(4)
Market's correction δ	1.053		1.054	
	(0.020)		(0.021)	
δ, orthogonal to error history				1.054
				(0.021)
Constant	-0.016	-0.229	-0.026	-0.229
	(0.031)	(0.057)	(0.033)	(0.034)
Error history: AR(3) and two trailing means	no	yes	yes	yes
R^2	0.624	0.092	0.625	0.625
F for $\alpha = 0,\ \beta = 1$	3.58			
Error of the implied mean in degrees Fahrenheit				
		Absolute, in sample	Absolute, held out	Squared, held out
Official forecast, raw		1.94	2.20	3.15
Official forecast, own error history		1.84	2.10	3.00
Official forecast plus the correction		1.24	1.27	1.73
Market's implied mean		1.24	1.26	1.73

2,679 city-days. Both forecasts imply a distribution over the day's maximum and therefore a mean; δ is the difference between them, the correction the market applies to the official forecast. The public-information benchmark is an autoregression of order three in the station's own realised forecast errors together with its trailing ten-day and thirty-day mean errors, all strictly backward looking. Column (4) residualises δ on that benchmark before entering it, so its coefficient uses only the part of the correction the public error record cannot reproduce. By the Frisch–Waugh theorem that coefficient equals column (3)'s and the two columns share an R^2; what differs is the error-history coefficients, which return to their standalone values because the residualised correction is orthogonal to them by construction. Under forecast encompassing the constant is zero and the slope on δ is one; the constant is indistinguishable from zero and the slope sits 2.7 standard errors above it, so the joint restriction is rejected at the five percent level ($p = 0.029$) and the market slightly over-corrects. The lower block estimates the coefficient on the chronologically first half of the sample and applies it to the second, 1,348 held-out city-days; the third row of that block is the official forecast plus $\hat{\beta}\delta$ and the second is the same exercise with the station's own error history in place of δ. Standard errors clustered on date. Seven stations are too few for cluster asymptotics in the cross-sectional dimension, so the restriction $\beta = 1$ is also tested by a null-imposed wild cluster bootstrap on station with Rademacher weights, enumerated over all 2^7 sign vectors: the station-clustered t is 2.43 and the bootstrap p-value is 0.039. The same procedure puts the intercept's p-value against zero at 0.88.

discrimination share (80%), meaning more of its gap is calibration than anywhere else; it is the one station where recalibrating the ladder within station recovers a meaningful part of the gap (16.1% against a median of 2.1%); and the share of the market's correction that public error history reproduces is highest there (58%). The four are consistent with one reading: at the station where a gridded product represents the site worst, a meaningful part of what the market supplies is bias repair the public record could have supplied too, and it shows up as calibration; at the other six the correction is something the record does not contain. This rests on a single station and is offered as description, not as an identified mechanism.

Where is the advantage concentrated? Not where the models disagree: regressing the paired Brier difference on the GEFS ensemble spread, with station effects, gives -0.00005 (0.00305), a precisely estimated null. It is concentrated after the benchmark has been wrong: the coefficient on the station's mean absolute forecast error over the prior ten days is -0.01089 (0.00135), and survives station effects (-0.00902, se 0.00146). The market's edge does not track how uncertain the models say the weather is; it tracks how wrong the official forecast has lately been, at the same station. Two readings are available and the data do not separate them: the market may be correcting a persistent error the ladder has not yet shaken off, or

the ladder may simply be in a bad patch that persists from one day to the next. Either way, Table 4 has already shown that knowing this does not help a user, since conditioning a recalibration on exactly that trailing error recovers 0.9% of the gap.

5 Conclusion

Market prices score better than the published forecast in this sample. That much was already known, and on its own it settles almost nothing, because the two need not be built on independent information: a trader can consult the published forecast, and more besides. This paper turns the fact that spoils the comparison into the comparison itself. Once the ladder is inside the conditioning set, the question is no longer which forecast wins but what each adds given the other, and the answer is asymmetric: given the official ladder, the price adds 0.0327 of discrimination out of sample; given the price, the official ladder adds -0.0006. Permuting the price away, leaving it everything it shares with the ladder, returns 0.00004. Averaged over 20 folds of repeated cross-fitting the increment is 0.0277, positive in every fold.

The label on that result matters, and the quantity that fixes it is a level. The official ladder's total miscalibration is 0.0024 against a gap of 0.0295, so there is very little for any recalibration to remove. The agency's probability statements are reasonably calibrated, at the contract level and at the level of the whole distribution; on the distributional diagnostic the official ladder is the marginally better calibrated of the two, and the market carries a favourite-longshot bias the ladder does not. Integrated over the whole predictive distribution on complete ladders, discrimination carries 84% of the gap against an arithmetic floor of 34.5%. The market does not win by being better tuned. It wins by separating one day from another more sharply, as published and at the moment a user would consult it.

For a user of the published forecast the reading is unusually specific. Recalibration is not the repair: with perfect hindsight a map of the stated probability recovers at most 8% of the gap, and fitted out of sample no conditioning set I have tried does better than 6.9%, including one built on the very covariate that predicts where the ladder underperforms. Combination fails from the other direction, with the unconstrained optimal weight sitting at or above one on the price. For a forecast producer the reading is narrower, and worth stating carefully. What this design establishes is that the published probabilities cannot be improved much by any post-processing of themselves, so a producer reading a poor Brier score against a market price should not conclude that the product is badly stated. What it does not establish is where the missing discrimination would come from. Later observations, better models, richer assimilation and local judgment all remain live, the price cannot distinguish them, and nothing here shows the agency could not obtain the same information by other means.

Three limits bound all of it. The cross-section is seven large stations with liquid contracts over a single annual cycle, so the magnitudes average over a handful of climates and one realisation of each season, and nothing here speaks to thinner contracts, smaller stations, longer horizons, or another country's operational system. The benchmark is one operational product read at one run time, so the comparison bounds what a user of that product forgoes, not what the agency could achieve. And the dominance of discrimination is an empirical finding about these two forecasts, not a structural property of superior forecasting: two forecasts can differ in accuracy through either channel, which is precisely why the decomposition is worth running instead of assuming. Nor does the paper price the gap: a Brier difference of 0.0295 is a statement about forecast quality, not about what it is worth to a decision-maker, and Luo

(2026) takes up the related question of whether these contracts can transfer weather risk, which is where the decision-theoretic work belongs.

The distinction generalises beyond weather. Wherever an agency publishes a probability and an exchange prices the same event, the same two exercises can be run, and together they separate claims that are routinely conflated: that a market discriminates outcomes the official forecast does not, and that the official forecast is merely badly stated. Only the second is something a user can regress away. In this application it is the first that the evidence supports.

Data availability

The replication archive contains the analysis code, the derived contract-level analysis file, the archived object identifier of every model run used, and a one-command build that regenerates every number, table and figure in this paper from that file. It is deposited in the Harvard Dataverse at https://doi.org/10.7910/DVN/UTGJVR.

Three of the four sources are open and can be retrieved without credentials. The benchmark is NOAA's National Blend of Models, read from the NOAA Open Data Dissemination archive on Amazon S3 (NOAA National Weather Service, 2026); settlement observations are NOAA's Global Historical Climatology Network Daily record (NOAA National Centers for Environmental Information, 2026), accessed through the NCEI Access Data Service. Both are in the public domain as works of the United States government, and their use here is gratefully acknowledged. Hourly station observations, used for the settlement comparison in Section 2, the nowcasting test in Section 4.4 and the predictand-window restriction in Appendix C, come from the Iowa Environmental Mesonet's ASOS archive at Iowa State University (Iowa Environmental Mesonet, 2026), whose maintainers are gratefully acknowledged. The contract tape is not open: reaching back beyond a rolling window of recently resolved markets requires Kalshi's historical data tier, which is a paid subscription, so the raw trade tape cannot be redistributed. Anyone holding a historical-tier key can rebuild the derived analysis file from raw with the code in the archive.

The author holds no position in any contract studied here, has never traded on the exchange, and has no financial relationship with Kalshi or with NOAA.

Declaration of competing interest

The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Specifically, the author holds no position in any contract studied here, has never traded on the exchange whose prices are analysed, and has no financial or consulting relationship with Kalshi or with the National Oceanic and Atmospheric Administration.

Declaration of generative AI in scientific writing

During the preparation of this work the author used generative AI assistants for language editing, for critical review of exposition and framing, and to help audit the reproducibility of the analysis code. No text, result, number or citation was accepted without verification against the manuscript's own code

and data, or against the cited source. Every quantity reported in the paper is generated by the released replication code rather than transcribed into the text. After using these tools the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Submission declaration

This paper is single-authored, is not under consideration for publication elsewhere, has not been published previously in any form, and its submission is approved by the author. If accepted, it will not be published elsewhere in the same form without the written consent of the copyright holder.

References

Joyce E. Berg, Forrest D. Nelson, and Thomas A. Rietz. Prediction market accuracy in the long run. *International Journal of Forecasting*, 24(2):285–300, 2008.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1): 1–3, 1950.

Jochen Bröcker. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 135(643):1512–1519, 2009.

Yock Y. Chong and David F. Hendry. Econometric evaluation of linear macro-economic models. *The Review of Economic Studies*, 53(4):671–690, 1986.

Jeffrey P. Craven, David E. Rudack, and Phillip E. Shafer. National blend of models: A statistically post-processed multi-model ensemble. *Journal of Operational Meteorology*, 8(1):1–14, 2020. doi: 10.15191/nwajom.2020.0801.

Morris H. DeGroot and Stephen E. Fienberg. The comparison and evaluation of forecasters. *The Statistician*, 32(1–2):12–22, 1983.

Francis X. Diebold and Roberto S. Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263, 1995.

Timo Dimitriadis, Tilmann Gneiting, and Alexander I. Jordan. Stable reliability diagrams for probabilistic classifiers. *Proceedings of the National Academy of Sciences*, 118(8):e2016191118, 2021. doi: 10.1073/pnas.2016191118.

Werner Ehm, Tilmann Gneiting, Alexander Jordan, and Fabian Krüger. Of quantiles and expectiles: Consistent scoring functions, Choquet representations and forecast rankings. *Journal of the Royal Statistical Society, Series B*, 78(3):505–562, 2016. doi: 10.1111/rssb.12154.

Christopher A. T. Ferro and Thomas E. Fricker. A bias-corrected decomposition of the Brier score. *Quarterly Journal of the Royal Meteorological Society*, 138(668):1954–1960, 2012.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Tilmann Gneiting and Roopesh Ranjan. Combining predictive distributions. *Electronic Journal of Statistics*, 7:1747–1782, 2013.

Tilmann Gneiting, Adrian E. Raftery, Anton H. Westveld, and Tom Goldman. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133(5):1098–1118, 2005.

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society, Series B*, 69(2):243–268, 2007.

Thomas M. Hamill. Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129(3):550–560, 2001.

David Harvey, Stephen Leybourne, and Paul Newbold. Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, 13(2):281–291, 1997.

Hans Hersbach. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559–570, 2000.

Iowa Environmental Mesonet. Automated surface observing system (asos) hourly observation archive, 2026. URL https://mesonet.agron.iastate.edu/ASOS/. Accessed 2026.

Nam Anh Le. Decomposing crowd wisdom: Domain-specific calibration dynamics in prediction markets. Technical report, arXiv:2602.19520v2 [stat.AP], August 2026. URL https://arxiv.org/abs/2602.19520.

Andrew Leigh and Justin Wolfers. Competing approaches to forecasting elections: Economic models, opinion polling and prediction markets. *Economic Record*, 82(258):325–340, 2006.

Ziye Luo. Prediction markets as weather hedges: Geographic basis risk, settlement constraints, and contract replication. Technical report, SSRN Working Paper No. 7138562, July 2026. URL https://ssrn.com/abstract=7138562. Version of 17 July 2026.

Charles F. Manski. Interpreting the predictions of prediction markets. *Economics Letters*, 91(3):425–429, 2006.

Allan H. Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4):595–600, 1973.

NOAA National Centers for Environmental Information. Global historical climatology network daily, version 3. NCEI Access Data Service, `daily-summaries`, 2026. TMAX at stations USW00094728, USW00014819, USW00023174, USW00012839, USW00003017, USW00013904 and USW00013739. Accessed 25 August 2026.

NOAA National Weather Service. National blend of models, conus domain, versions 4.3 and 5.0. NOAA Open Data Dissemination, Amazon S3 bucket `noaa-nbm-pds`, 2026. 18-hour maximum temperature percentile fields `maxt18p01`–`maxt18p99`, 00Z cycle. Accessed 25–26 August 2026.

Lionel Page and Robert T. Clemen. Do prediction markets produce well-calibrated probability forecasts? *The Economic Journal*, 123(568):491–513, 2013.

Roopesh Ranjan and Tilmann Gneiting. Combining probability forecasts. *Journal of the Royal Statistical Society, Series B*, 72(1):71–91, 2010.

Matthias Ritter. Weather forecasting with market prices of weather futures. *Journal of Energy Markets*, 6 (4):25–41, 2013. Winter 2013/2014 issue.

Matthias Ritter, Oliver Mußhoff, and Martin Odening. Meteorological forecasts and the pricing of temperature futures. *The Journal of Derivatives*, 19(2):45–60, 2011. doi: 10.3905/jod.2011.19.2.045.

Erik Snowberg and Justin Wolfers. Explaining the favorite-long shot bias: Is it risk-love or misperceptions? *Journal of Political Economy*, 118(4):723–746, 2010.

Justin Wolfers and Eric Zitzewitz. Prediction markets. *Journal of Economic Perspectives*, 18(2):107–126, 2004.

Appendix

A Scoring-rule and calibration details

The ordering across scoring rules. The Brier score is one member of a family. Every consistent scoring function for a probability forecast is a mixture of elementary scores indexed by a threshold θ, and the Brier score is the mixture that weights them equally (Ehm et al., 2016). For a binary outcome θ is a cost-loss ratio, so each elementary score is the loss of a user who protects when the forecast exceeds θ. Figure A1 plots them. The market scores better at 98 of the 99 thresholds on a hundredth grid, and the bootstrap band excludes zero at 98 of them; the single exception sits at $\theta = 0.99$, where the difference rounds to zero. The advantage is largest at $\theta = 0.26$, where it reaches 0.0340. The ordering is not an artefact of the quadratic weighting, and it holds for essentially every cost-loss ratio a user might have. What the figure does not do is price the gap for any particular user, which needs a loss function this design does not supply.

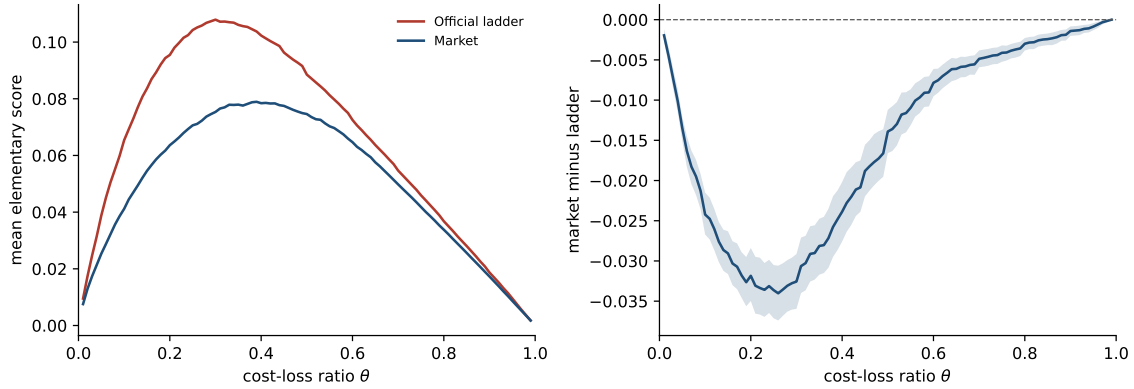


Figure A1: Elementary scores across cost-loss ratios

Mean elementary score $S_\theta(x, y) = (\mathbf{1}\{x > \theta\} - \mathbf{1}\{y > \theta\})(\theta - y)$ at each threshold θ, following Ehm et al. (2016). For a binary outcome θ is the cost-loss ratio of a user who protects when the forecast exceeds it. Left panel: each forecast. Right panel: the paired difference, with a 95 percent band from 400 bootstrap replications resampling dates. The Brier score is the equally weighted mixture of these scores, so the diagram shows whether the ranking survives every other weighting.

The climatological reference. The reference is a day-of-year climatological distribution built from 30 years of NCEI daily maxima at the same stations, ending before the sample begins. It scores worse in Brier than the constant in-money rate, which looks wrong until one notices what the exchange lists. Strikes are set around the current forecast, and the interval ladder spans about 1.04 climatological standard deviations at the median city-day, so a distribution that ignores the current forecast places substantial mass on rungs far from the one that pays and pays the full quadratic penalty for doing so. Against the climatology the market's Brier skill is 0.486 and the ladder's 0.336; in continuous score, 0.799 and 0.685.

The binned decomposition and its correction. The binned decomposition on twenty equal-width bins gives a discrimination share of 94.8%, and the finite-sample correction of Ferro and Fricker (2012) moves it to 95.1%, within two percentage points of the CORP share. The correction is worth computing: it depends on how each forecast occupies the unit interval, and the market's probabilities are the more

dispersed, so the bias is 0.00018 for the market against 0.00027 for the ladder. It happens not to move the share because it enters the miscalibration and discrimination differences with opposite signs. Where the binned decomposition is reported, bin edges are defined on each forecast's own distribution; one set of equal-mass edges from the pooled distribution gives 96% instead. Under the binned estimator the continuous share is much the less stable of the two, moving across bin counts in a way the contract-level share does not; that instability belongs to the estimator, and it does not survive the change to CORP.

Calibration slopes. Both forecasts fail a joint test of the linear calibration restriction, and neither failure carries the gap. With an intercept included the market gives −0.0254 (0.0023) with a slope of 1.079 and the official ladder 0.0419 (0.0036) with 0.768; the ladder's slope below one means its probabilities are too extreme, not too diffuse. The market's slope moves with what the ladder is rescaled by, and one of the two available rescalings is wrong. Rescaling each city-day's full ladder to sum to one, the rescaling Section 2 justifies, moves the slope from 1.079 to 1.139 (standard error 0.013): it moves away from one, because the full ladder sums to more than a dollar and dividing by a number above one shrinks every probability. Dividing instead by the sum over the interval buckets alone moves the slope to just under one, but the interval buckets do not partition the outcome space, so their sum is not a normalising constant for anything and that version is discarded. The rescaling makes no difference to the decomposition: repeating CORP on the 15,777 rescaled contracts gives a share of 97% against 96% as traded.

Reconciliation with Le (2026). A linear slope and a logistic slope are not comparable in magnitude, so the comparison must be run on Le's estimator. Fitted that way, the logistic recalibration slope on this tape is 1.261 (0.029) as traded and 1.294 (0.029) with the ladder rescaled; both sit well above one, which on that estimator means compressed rather than too extreme. Le reports 0.87 and 0.91 for weather contracts at six-to-twelve and twelve-to-twenty-four hours to resolution, the bins that bracket the horizon here, and both are below one. The disagreement is about the sample rather than the method. Le's weather universe pools temperature-record, precipitation and natural-event contracts across horizons from under an hour to beyond a month, with a base rate of 24%, against a same-day daily-maximum ladder with a base rate of 17.7% here; and the object rescaled there is not this ladder. The official forecast, fitted the same way, gives 0.622 (0.024), further below one than anything Le reports for weather, which is the upper-tail breakdown of Figure 1 seen through a logistic lens.

Tail behaviour. The market's one calibration defect is the one the prediction-market literature predicts (Snowberg and Wolfers, 2010), and it survives the rescaling. Among the 9,116 rescaled contracts priced below 0.15 the mean price is 0.044 against a realised frequency of 0.030, an overpricing of longshots by 0.015; in the interior the market underprices by 0.019, and above 0.85 it underprices by 0.056. That is a favourite-longshot bias of the classical shape. The official ladder's tail defect runs the other way and is much the larger. Below 0.15 it *under*prices, 0.050 against a realised 0.074; above 0.85, on 58 contracts, it states 0.935 and delivers 0.448, an overstatement of 0.487. Neither forecast is calibrated in its tails, they fail in opposite directions, and the ladder fails by an order of magnitude more. This is why the ladder carries the larger total miscalibration despite both figures being small, and it means the market wins the comparison while carrying a calibration defect of its own.

Implementation. The Diebold and Mariano statistic is computed on the equal-weighted mean contract-level squared-error differential within each calendar date, giving one observation per day and 400 in total, with Newey-West errors at ten lags; the small-sample correction of Harvey et al. (1997) changes it in the third significant figure. The bootstrap resamples city-days with replacement, four hundred replications. Leaving the market's ladder unnormalised gives 0.907 in continuous score against the normalised 0.917, so renormalisation does not flatter it; completing the tails over three, six and twelve degrees moves the discrimination share of the continuous gap only from 82% to 89%.

B The holdout estimator under a null

Two properties of the estimator behind Table 3 deserve to be established rather than assumed, because the increment is the paper's central number and a flexible learner on a richer conditioning set is exactly the setting where an estimator can manufacture one.

The first is what the held-out quantity is. Reading discrimination off as uncertainty minus the holdout score treats the fitted conditional mean as calibrated, and out of sample it is not. By identity (1) that difference is discrimination minus miscalibration, and the miscalibration is larger for the richer conditioning set, which is what put the ladder's increment on the wrong side of zero in Panel A. Passing the holdout fitted values through the pool-adjacent-violators algorithm separates the two. Conditioning on the ladder alone then gives miscalibration 0.0015 and discrimination 0.0144; on the price alone, 0.0006 and 0.0468; on both, 0.0012 and 0.0468. The two-dimensional fit carries the largest holdout miscalibration, which is the mechanism. Measured this way the price adds 0.0324 of discrimination to a user holding the ladder and the ladder adds -0.0001 to a user holding the price, against 0.0008 on alternating weeks. The impossibility disappears once the estimand is the one the text names, and the headline increment is unmoved.

The second is the null. Adding any second regressor changes an estimator's variance and its holdout behaviour, so the increment needs a placebo that holds everything constant except the information. Permuting the price within station and within decile of the official probability does that: the permuted price keeps its marginal distribution, its station, and everything it shares with the ladder, and loses only what it adds. Over 20 permutations the increment is 0.00004 with a standard deviation of 0.00033 and a maximum of 0.00054, against 0.0327 on the true price.

The central number is also not an artefact of how $E[Y \mid p, q]$ is estimated. Across seven estimators, from an unconstrained gradient-boosted regression through a random forest and a logistic model quadratic in both forecasts to fully non-parametric averages over a grid of cells, the price adds between 0.0299 and 0.0328 and the ladder between -0.0033 and -0.0005 on the chronological split. In the alternating-week column the ladder's increment is 0.0001, on the other side of zero and inside the same interval. In the nowcasting test of Section 4.4, extending the observation cut-off from the window's open to its close leaves the observations worth 0.0011 alone and -0.0026 beside the ladder, with the price still adding 0.0345.

Table A1: The benchmark on its best terms

Forecast	Brier	Reliability	Resolution
Panel A. Advancing the benchmark's vintage			
Kalshi market price, 12:00–15:00 UTC	0.1014	0.0006	0.0448
NBM 00Z run, the headline benchmark	0.1309	0.0022	0.0169
NBM 06Z run, archived during the window	0.1311	0.0027	0.0172
Panel B. A benchmark initialised as the pricing window opens			
Kalshi market price, 12:00–15:00 UTC	0.1015	0.0006	0.0447
NBM 00Z run	0.1306	0.0020	0.0170
NBM 12Z run	0.1266	0.0022	0.0212
Panel C. Removing the interpolation: station guidance			
Kalshi market price	0.0928	0.0014	0.0506
NBM gridded ladder, interpolated	0.1220	0.0024	0.0226
NBM station bulletin, site-corrected	0.1243	0.0018	0.0196

Each panel gives the official forecast an advantage it does not have in the main comparison. The valid timestamp is part of the archive's object key and the same one is requested from every cycle, so all three runs forecast the same 18-hour period: on all 405 run dates archived for both the 00Z and 06Z cycles, the valid times agree. Panel A, on 15,664 contracts, replaces the overnight run with the 06Z cycle, which this archive receives about 7.1 hours after its cycle time and therefore during the pricing window. The fresher cycle is not inert, gaining 0.0004 of resolution and losing 0.0005 of reliability, which nets to a paired Brier difference of 0.0003 (0.0007); the market beats it by 0.0298 (0.0013). Panel B goes further, to the 12Z cycle, whose initialisation coincides with the opening of the pricing window on the 15,543 contracts and 2,758 city-days where it is archived, spanning 14 calendar months. That run is not one a user could have held while these prices printed: every cycle reaches this archive about 7.1 hours after its cycle time, so the 12Z field lands after the window closes, and this archive cannot speak to when the forecast first became available on the agency's own systems. What it delivers instead is a benchmark that has assimilated everything observable up to the moment the window opens, against a price formed over the three hours after it. Panel C, on 2,080 contracts over six weeks, replaces the interpolated grid with NBM's site-corrected station bulletin, whose predictive distribution is formed as Gaussian in the bulletin's forecast and standard deviation; that construction is the author's. Correcting to the site does not measurably change resolution: the bulletin minus the grid is -0.0029, with a city-day bootstrap interval of [-0.0080, 0.0008] that contains zero, so the six-week sample cannot resolve a small difference in either direction. What it can resolve is the market's margin over the bulletin, 0.0310 [0.0255, 0.0368]. Only one of the four candidate valid-time offsets is populated in the archive, so the offset is forced rather than selected. Standard errors are clustered on city-day.

C Timing, measurement and heterogeneity

Section 4.4 changed the estimand from a raw comparison to what each forecast adds given the other. The evidence below bears on the descriptive comparison, not on the central result; it is what motivated the change, together with the other explanations a reader will reach for.

Vintage. Panels A and B of Table A1 advance the benchmark's vintage, first by six hours and then by twelve. The benchmark's predictand does not move when the cycle does: the valid timestamp is part of the archive's object key, and on all 405 run dates archived for both the 00Z and 06Z runs the two agree. On the 15,664 contracts where both cycles exist, the 06Z run scores 0.1311 against 0.1309 for the 00Z run, a paired difference of 0.0003 (0.0007). The fresher cycle is not inert: its discrimination rises by 0.00036 and its reliability deteriorates by 0.00050, which nets to 0.00015 in Brier, the remaining 0.00011 of the observed 0.00026 being the binning residual of the Murphy estimator used for this panel. Six hours of assimilation buy the official ladder a small amount of discrimination and cost it a little calibration.

Table A2: Market accuracy through the trading day

Price window (UTC)	Contracts	Brier score Market	NBM	Paired difference		Share from discrimination
04:00–07:00	14,845	0.1159	0.1370	-0.0211	(0.0013)	94%
07:00–10:00	11,483	0.1131	0.1404	-0.0274	(0.0015)	93%
10:00–13:00	15,059	0.1084	0.1360	-0.0277	(0.0013)	95%
13:00–16:00	15,452	0.1005	0.1335	-0.0330	(0.0013)	96%
16:00–19:00	14,829	0.0768	0.1387	-0.0619	(0.0015)	101%
19:00–22:00	13,977	0.0303	0.1456	-0.1153	(0.0015)	104%

Each row is computed on its own window's contract set rather than on the panel of contracts priced in all six windows. That matters: requiring a price in every window keeps the contracts still trading in the evening, which are the ones that had not resolved, and discards the far-from-the-money buckets that collapse in the morning, so a balanced panel raises the base rate from 0.121 among the contracts it drops to 0.206 among those it keeps and inflates every score in the table. Computed this way there is no such selection, and the levels are comparable with the headline. The benchmark is the same 00Z ladder in every row, so its score moves only with the contract set. The price is the volume-weighted average of trades in each three-hour window on the contract's own date. Standard errors in parentheses are clustered on city-day. The share column is the discrimination difference over the score difference under the CORP decomposition of Table 2; a share above one hundred percent means the market is by then the worse calibrated of the two. The 00Z run reaches the public archive in the early hours of the target day, so the first row is the least temporally separated comparison the archive allows. It is still not a common information set, which is what Table 3 addresses; the later rows are informative about how quickly the market incorporates what arrives, not about which forecast is better at a common origin.

The 12Z cycle goes further. It initialises at the moment the pricing window opens, so it has assimilated everything observable up to 12:00 UTC, while the price it is scored against is a volume-weighted average of trades over the three hours that follow. Two things should be said plainly about that panel. It is not a forecast a user could have acted on while these prices printed: every cycle reaches this archive about 7.1 hours after its cycle time, so the 12Z field lands in the evening. And the residual asymmetry in observation still runs in the market's favour, three hours of it rather than the twelve of the headline comparison. The 12Z run is genuinely better than the overnight one, by 0.0040 (0.0008), which shows the design detects an assimilation gain when there is one to detect. The market still beats it by 0.0251 (0.0013) on 15,543 contracts. None of this constitutes a common information set, which is why Section 4.4 exists.

The trading day. Table A2 reports the profile across six three-hour windows, each computed on its own window's contract set. The alternative is a balanced panel of the contracts priced in every window, and it cannot be read for the level of any row: a contract still printing at 19:00 UTC is one that had not resolved, so requiring a price in all six windows keeps the hard contracts and discards the easy ones, raising the base rate from 0.121 among the dropped contracts to 0.206 among those kept, and inflating every score in the table with it. Computed window by window there is no such selection. In the earliest window, before the target day's heating begins, the gap is -0.0211 (0.0013) on 14,845 contracts with discrimination carrying 94% of it. The gap then widens in every subsequent window, reaching -0.1153 by evening, when much of the day's maximum has already been observed.

One consequence of the balanced panel is a fact about the benchmark, not about the table, and it is the reason Section 4.6 runs a multivariate horse race. On contracts that trade all day the ladder under-forecasts, with a mean probability -0.0116 below the realised frequency; on contracts that stop trading early it over-forecasts by 0.0187; pooled, the two nearly offset and the aggregate bias is -0.0012.

The ladder's pooled miscalibration is therefore flattered by averaging, and a conditional miscalibration an order of magnitude larger than the pooled figure is invisible to any function of the stated probability. Table 4 is what happens when a user tries to exploit it.

Measurement. A grid cell is not a station, and interpolation error is the one channel that could manufacture the result mechanically. It runs in the direction a reader might not expect. Because the decomposition compares conditional outcome frequencies instead of fitting a line, it does not inherit the attenuation that measurement error induces in a regression slope. It is not thereby immune: noise in the stated probability shrinks the conditional frequencies toward the base rate, which lowers measured discrimination. The benchmark's discrimination is therefore a lower bound on what the agency's forecast achieves at the station itself, and the question is how much of one.

Panel C of Table A1 answers it with a product that needs no interpolation. NBM issues station guidance alongside its grids, corrected to the site and carrying a maximum-temperature forecast with its standard deviation. Correcting to the site improves the benchmark's reliability, from 0.0024 to 0.0018, and its calibration slope, from 0.872 to 0.945. Its discrimination does not rise: the bulletin scores 0.0196 against 0.0226 for the interpolated ladder, a difference of -0.0029 whose city-day bootstrap interval, [-0.0080, 0.0008], contains zero. The archive covers six weeks, so this sample cannot resolve a small difference in either direction; it can resolve the comparison that matters, since the market's discrimination exceeds the site-corrected bulletin's by 0.0310, with an interval of [0.0255, 0.0368]. One caveat runs the other way: forming the bulletin's predictive distribution as a Gaussian in its forecast and standard deviation is my construction, not the agency's, and applying the same construction to the gridded ladder costs 0.0021 of discrimination.

The second measurement channel is the predictand mismatch set out in Section 3. Dating each day's maximum from the hourly observations, 95% of station-days have their maximum inside the benchmark's window; restricting to those leaves a gap of 0.0239 (0.0012) with discrimination carrying 97% of it. The mismatched 5% of station-days carry 0.0056 of the headline 0.0295, which is 19% of it. The full sample nonetheless remains the principal descriptive comparison, because the restriction conditions on when the maximum actually occurred, which is a realisation and not something a forecaster or a trader knows when either forecast is issued. Building a calendar-day predictand would be the better repair and the archive does not permit it. Both figures are reported wherever the headline appears, and neither the ordering nor its composition depends on which is used.

One further comparison establishes that the benchmark is not a straw man. On the 7,078 contracts where the raw ensemble also exists, the agency's post-processing earns its keep: discrimination rises from 0.0161 for the hand post-processed ensemble to 0.0202 for the operational ladder, a gain of 0.0041. The market sits at 0.0452, more than twice either.

Heterogeneity and construction. Table A3 reports the gap station by station and season by season. It is negative and precisely estimated at all seven stations and in all four seasons, from 0.0187 in summer to 0.0390 in winter, with discrimination carrying between 89% and 98% of it. No station's Diebold and Mariano statistic falls below 4.6 in absolute value. The sample spans one annual cycle, so Panel B describes a single realisation of each season and not a stable seasonal pattern; it is reported to show the average is not the product of one regime.

Table A3: The gap across stations and seasons

	Contracts	Paired difference	Diebold–Mariano	Share from resolution
Panel A. By settlement station				
Austin	2,275	-0.0201 (0.0030)	-6.6	98%
Chicago	2,265	-0.0386 (0.0037)	-8.4	89%
Denver	2,277	-0.0230 (0.0028)	-6.8	97%
Los Angeles	2,263	-0.0413 (0.0038)	-7.1	80%
Miami	2,199	-0.0124 (0.0027)	-4.6	98%
New York	2,278	-0.0354 (0.0038)	-9.1	91%
Philadelphia	2,225	-0.0353 (0.0037)	-7.3	91%
Panel B. By season				
Winter (DJF)	3,645	-0.0390	-15.6	89%
Spring (MAM)	4,452	-0.0326	-9.7	92%
Summer (JJA)	4,272	-0.0187	-8.9	98%
Autumn (SON)	3,413	-0.0287	-7.3	96%

Paired differences in Brier score, market minus benchmark, with standard errors clustered on city-day in Panel A, and the share of each difference attributable to resolution under twenty equal-width bins. The Diebold–Mariano statistic is computed on the mean contract-level loss differential within each calendar date, with Newey-West errors at ten lags; each station contributes 400 daily observations. The difference is negative at every station and in every season, and no station's statistic falls below 4.6 in absolute value. The sample spans one annual cycle, so Panel B describes a single realisation of each season rather than a stable seasonal pattern.

Dropping one station at a time addresses the concern that a single site carries the result. Across the 7 jackknife samples the gap runs from 0.0275 to 0.0323 and the discrimination share from 93.4% to 97.8%. Los Angeles is the station Section 4.8 singles out, and removing it raises the share rather than lowering it, to 97.8%: the interpolation channel there is a calibration phenomenon, so taking it out leaves a comparison in which discrimination matters more, not less.

Four further splits address the concern that the result belongs to a favourable subsample. Sorting city-days by traded notional, the gap is 0.0221 in the thinnest tercile and 0.0403 in the thickest, and discrimination carries 98% of it in the thinnest against 87% in the thickest: the advantage is larger where the market is deeper, and if anything more purely a discrimination phenomenon where the market is thin. Sorting contracts by the distance of the strike from the official ladder's median, the gap is 0.0420 at the money and 0.0167 far from it, so the market's discrimination shows up where the outcome is genuinely in doubt. The benchmark changed version inside the sample, from 4.3 to 5.0 in May 2026: the gap is 0.0295 before the cutover and 0.0292 on the 2,080 contracts after it; those are the same contracts as the station-bulletin window in Panel C, because the bulletin archive begins at the version cutover, so the two rows are one subsample seen twice, not two independent checks. And because the ladder is priced from non-synchronous trades, restricting to the 7,755 contracts whose city-day ladder needs no rescaling gives a gap of 0.0326 with discrimination carrying 86%; tightening the tolerance leaves 1,648 contracts and 78%. The share falls as the tolerance tightens for a reason already in the table: a ladder that sums close to one is one that traded heavily, and the thickest volume tercile is where the share is lowest. The gap itself grows.

Table A4: Where the result survives

Specification	Obs.	Difference	Discrimination share	Reported in
Headline				
Market at 12:00–15:00 UTC against the 00Z ladder	15,782	-0.0295	96%	Tabs. 1, 2
Information set				
What the price adds given the ladder, held out	8,270	-0.0327		Tab. 3A
splitting on alternating weeks instead	7,887	-0.0270		Tab. 3A
Timing				
Overnight price, 04:00–07:00 UTC	14,845	-0.0211	94%	Tab. A2
Benchmark advanced to the 06Z cycle	15,664	-0.0298	93%	Tab. A1A
Benchmark initialised as the window opens (12Z)	15,543	-0.0251	94%	Tab. A1B
Measurement				
Days whose maximum falls inside the benchmark's window	14,992	-0.0239	97%	§C
Agency's site-corrected station bulletin	2,080	-0.0314		Tab. A1C
Construction				
Full ladder rescaled to sum to one	15,777	-0.0293	97%	Tab. 2C
Ladders already summing to one within 0.05	7,755	-0.0326	86%	§C
Each city-day weighted equally	15,782	-0.0295	95%	§4.1
Murphy decomposition, twenty bins, bias corrected	15,782	-0.0295	95%	Tab. 2C
Continuous score, no contract partition	2,784	-0.5226	88%	Tab. 2B
restricted to complete ladders	2,097	-0.5972	84%	§4.3
Leave-one-station-out, worst case		0.0275	93%	§C
Sample				
Thinnest tercile of traded notional	5,356	-0.0221	98%	§C
NBM version 5.0 only	2,080	-0.0292	97%	§C

Every specification the paper runs, with the paired difference in score and the share of that difference attributable to discrimination. Differences are in Brier units on contracts except the continuous row, which is in degrees Fahrenheit on city-days. Four rows use the exact CORP decomposition of Table 2: the headline, the continuous score, the rescaled ladder and the overnight window. The remainder use the binned decomposition of Murphy (1973), because the panels they are drawn from do; Panel C of Table 2 shows the two agree to within two percentage points on the headline sample. The two information-set rows report the discrimination the price adds to a user who already holds the official ladder, signed negative so that, as in every other row, a negative entry favours the market; the ladder adds -0.0006 and 0.0001 in the other direction, and no decomposition applies to an increment, so the share column is blank. The station-bulletin row compares the market with the agency's own site-corrected product over the six weeks the bulletin archive covers, which are the same six weeks as the version-5.0 row rather than an independent sample; its resolution comparison, which is the quantity of interest there, is 0.0310 with a bootstrap interval of [0.0255, 0.0368].

D The ledger

Table A4 collects every specification the paper runs in one place, with the paired difference and the share of it carried by discrimination. The difference is negative in every row, and discrimination carries the large majority of it in every row where the decomposition is defined.