PROBABILISTIC FORECASTING

From Next Tokens to Future Events: What Does a Forecast Probability Mean?

A matched test of direct event prediction and probabilistic observation tokens on six years of held-out precipitation records

Revanth Matha · Matha Labs | Preprint awaiting review 1.0 | 2026-11-20

Abstract

What changes when an AI is trained to give the probability of a specified future event? We define the event, its deadline and its resolver, derive its probability from a sequence of possible observations, and test direct and token-based estimators on the same historical dates. The event was daily precipitation of at least 1 mm at two US airport stations. Using 4,348 eligible station-days in 2020-2025, we compared monthly climatology, a season-and-history reference, direct logistic and neural forecasts, and a four-symbol softmax whose two wet probabilities were summed. The nonlinear direct model had the lowest Brier score at Los Angeles (0.057873). The four-symbol model had the lowest at Seattle (0.186636). A factorized observation model reproduced the direct logistic event probability to numerical precision by construction. Probability bins revealed local overprediction despite improvements over seasonal baselines. These results show how to make a world-event forecast verifiable and why direct supervision alone gives no guarantee of greater skill. A separate seven-day-lag event network issued forecasts of 11.74% and 51.49%, respectively, for the November 21, 2026 station day before its outcomes were known. Skill on open-ended future events remains untested.

Keywords: event probability, forecast verification, calibration, observation tokens, neural networks, precipitation

Contributions

1. A formal and executable example of how successive observation probabilities imply different event probabilities at different deadlines.

2. A chronological two-station comparison of direct binary models, a four-category observation model, and two seasonal references using the same eligible dates.

3. A factorized four-symbol control, paired uncertainty estimates, and a protocol for prospective forecasts that retain their evidence cutoff and resolution rule.

1. Plain-English summary

Suppose you ask two questions about tomorrow morning. "How would a weather report usually finish this sentence?" has an answer in words. "Will this station record enough precipitation tomorrow to count as wet?" has an answer in observations. A model can sensibly assign different probabilities to the two. A reporter can mention rain as a possibility on a dry day, or describe a wet day using a different word. To forecast the station's result, we have to identify the station, choose a threshold, set a clock, and eventually check the measurement.

A probability is useful partly because it lets a model admit that several futures remain possible. A 20% forecast can be followed by the event and still be a reasonable forecast. A series of comparable forecasts gives us a stronger test. If events occur about one time in five when the model reports 20%, that part of its forecast scale has empirical support. We can also ask whether it distinguishes high-risk occasions from ordinary ones. People make different decisions about moving a picnic or delaying a costly delivery, even if they agree on the chance of rain. [1][6][7]

This paper studies the step from a dated history to the probability of a verifiable event. A simple statistical model gives the idea a test with real observations and inspectable assumptions. The same definition can guide the design of a larger model that reads language, sensor streams or remembered episodes. The evidence needed to establish skill changes with the event. Performance at one weather station leaves the question of skill at anticipating a person's actions open.

The paper ends with a live test. Before the November 21, 2026 station day, we published a probability for each airport and a rule for checking the results. A reader can return to the same dated forecast after the measurement becomes available. [15][16]

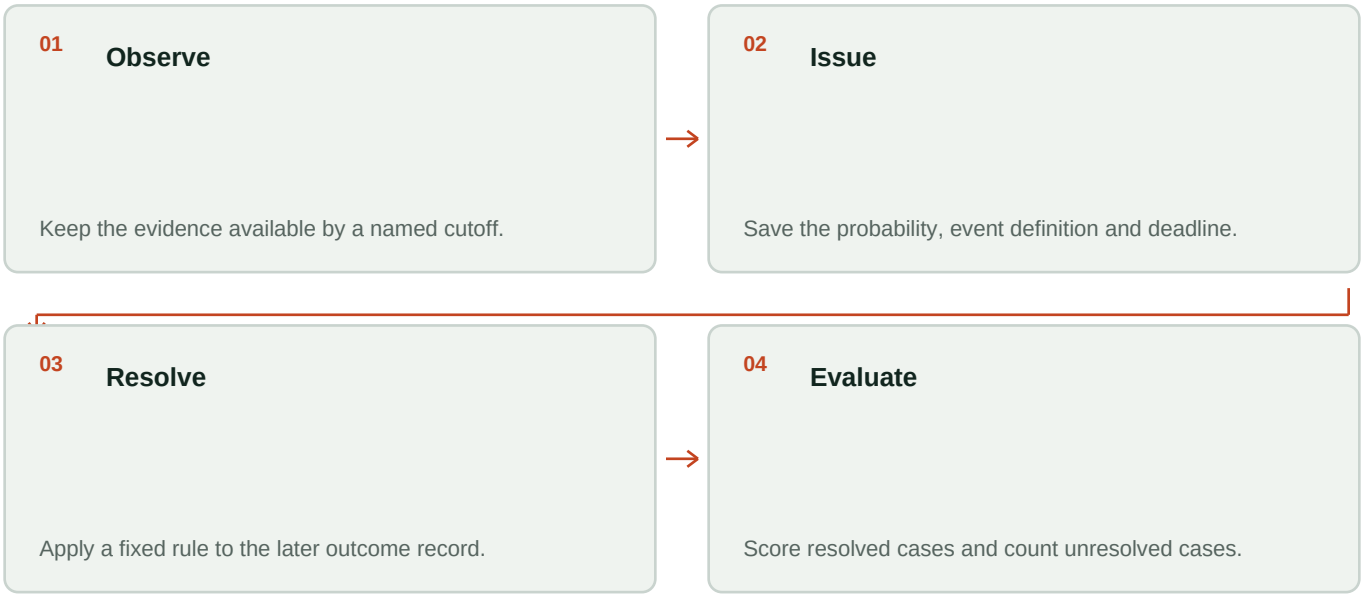
2. Define the event before making the forecast

A forecast record needs an issue time, an evidence cutoff, a target interval, and a rule for resolving the outcome. The issue time says when the model must commit. The evidence cutoff says which observations it could have received by then. Yesterday's measurement can arrive today, and a corrected value can arrive next month. The target interval identifies the piece of the future under study. The resolver specifies exactly which recorded quantity will count as evidence that the event occurred. Storing these items with the probability makes a later audit possible. Dawid's sequential approach supplies the guiding principle of evaluating each forecast against the outcome that follows it. [1]

For a given issue time t, call the available information I(t). Let Y equal one if the resolver finds that the event occurred in the specified interval and zero if complete observations establish its absence. The target is the conditional probability that Y equals one given I(t). One can learn it directly from past issue-time records paired with resolved labels. Such a model can use a logistic function to map a history summary to a number between zero and one. The mathematical definition also accommodates histories represented by images, language embeddings or writable neural weights.

An ambiguous outcome needs its own disposition. If the measuring instrument goes silent, a blank record by itself cannot establish a dry day. Specify in advance how late readings, revisions and missing intervals will be treated. A useful ledger retains unresolved entries and records their reason. It can then show what fraction of issued forecasts actually entered the score.

Figure 1. How a forecast becomes a scored observation



The issue record freezes the information used by the forecaster. The resolution record can arrive later. Independent records allow a reviewer to check timing, corrections and coverage.

3. How a path of future observations becomes one probability

Consider a deliberately small model of two coming days. Today is dry. After any dry day the probability that the following day is wet is 0.20. After any wet day it is 0.60. Assume these transition rules are known and that the most recent day's state contains all the relevant history for this example. Four possible ordered paths then exhaust the outcome space. Dry followed by dry has probability 0.80 x 0.80 = 0.64. Dry followed by wet has probability 0.80 x 0.20 = 0.16. Wet followed by dry has probability 0.20 x 0.40 = 0.08. Wet followed by wet has probability 0.20 x 0.60 = 0.12. They sum to one.

The chance of a wet first day is 0.20. The chance that the second day itself is wet is 0.16 + 0.12 = 0.28. The chance of at least one wet day by the end of day two is 0.16 + 0.08 + 0.12 = 0.36. The extra 0.08 comes from paths that begin wet and end dry. Adding the first-day and second-day probabilities would double count the path with two wet days. Multiplying their marginal dry probabilities would assume independence, which the transition rules violate.

For longer discrete histories the same reasoning applies. Multiply the conditional step probabilities along each possible path, then add the probabilities of all paths satisfying the event definition. Sampling from a fitted joint model can approximate the sum. A direct event model can estimate the answer without generating paths. For an "at least once by" target issued at a fixed time, extending the deadline includes every path that already satisfied the earlier target. Its probability can only stay the same or rise. A question about one particular day has a different constraint.

```
P(event | history) = sum over future paths of [event indicator x path probability]
```

A path probability is the product of its successive conditional probabilities in a discrete sequence model. This is an identity for observations with a joint probability law. Applying it to an actual system requires a model of the observations relevant to the resolver.

4. Three ways to produce a forecast

A direct event head predicts a Bernoulli probability from the evidence available at issue time. Its examples carry resolved event labels, and its objective rewards probability assigned to those outcomes. This gives the model a clear target and can be practical with limited data. It also commits the researcher to a particular question. A head trained on occurrences above one threshold requires fresh validation before its output can stand for a materially different event. Independent heads for several nested deadlines can even issue internally inconsistent probabilities unless the design or fitting process enforces their ordering.

A model of future measurements offers another route. DeepAR trains a recurrent probabilistic model on related series and generates possible future sample paths. Chronos scales and quantizes time-series values into tokens, learns categorical predictions, and samples trajectories that can be mapped back to the measurement scale. In each case a researcher can apply the same event rule to every simulated trajectory and estimate the event probability by its frequency. Discretization deserves scrutiny when a threshold cuts through a bin. Increasing the number of simulated paths reduces sampling noise while leaving errors in the learned distribution intact. [2][3]

A language-based forecaster can read the question, retrieve dated evidence and output a numerical estimate. Halawi and colleagues built a system that searched for evidence and aggregated forecasts, then compared its probabilities with human aggregates on forecasting questions. ForecastBench uses questions whose outcomes are still unknown at submission. These designs evaluate the numerical claim against its later resolution. Simply reading off the probability of the word "yes" also depends on the prompt, formatting and vocabulary. To treat it as a world-event probability, the corresponding readout needs evaluation against resolved events under the same issue-time rules. [4][5]

5. A station, a threshold and an honest clock

We chose a question with a public resolver and a useful complication. A wet day is commonplace in one test city and comparatively rare in the other. A single accuracy figure could reward a model just for naming the more common state. The target is the total daily precipitation observed at Los Angeles International or Seattle-Tacoma International, measured by NOAA's station record. An amount of 1.0 mm or more resolves the event as one. The target includes liquid-equivalent precipitation. Someone seeing a shower elsewhere in the city would supply evidence about a different event. [11][12][13][14]

For each local station date d, we simulate issuing a probability at the start of that target day. The forecast window is the archive's observation day labeled d. We do not infer a universal midnight-to-midnight UTC window from a station date. We reserve the immediately preceding date as a reporting buffer. The newest precipitation measurement entering a forecast therefore belongs to d minus two days. Calendar features for the target date are knowable in advance. Precipitation from the target day and the preceding date stays out of the input vector. Original station reporting and subsequent quality control can change how quickly a usable record arrives. This is a historical simulation under an explicit availability assumption. It does not reconstruct the exact archive any operator saw on each issue date. [11][12]

We read the fixed-width PRCP field in tenths of a millimetre. The reader rejects missing values, negative amounts, nonblank quality flags and observations labeled presumed zero. A recorded zero, including a zero marked as a trace, lies below the threshold. All seven lag dates d-2 through d-8 require usable observations. Eligibility for retrospective scoring also requires a usable target. At Los Angeles all 2,192 test dates were eligible. At Seattle, five lacked a usable target alone, thirty lacked usable history alone, and one lacked both. The matched evaluation therefore contains 2,156 Seattle dates. A missing target can prevent verification after a forecast would otherwise have been possible. An unavailable history can prevent this particular model from issuing a forecast at all. [12][13][14]

Table 1. The forecast contract

Field	Operational definition
Entity	One named NOAA airport station
Outcome	Daily PRCP >= 10 tenths of a millimetre
Issue and deadline	Start of target station date; end of its observation day
Available evidence	Station history through d-2 and the known calendar
Resolver	Target-date PRCP from the pinned archive
Score set	Dates with a usable target and seven usable lag observations

The definition stays fixed across all five forecasts at a station. The repository contains the precise archived versions. [11][12][13][14]

6. Two prediction targets on the same information

We fit each station separately. Eight predictors describe the calendar and precipitation history. Four sine and cosine terms encode annual and semiannual seasonality. Three more features are the log of one plus the amount on d-2, the amount on d-3, and the mean amount on d-4 through d-8, respectively, with amounts in millimetres. The eighth is the fraction of the seven lag dates meeting the wet threshold. Standardization uses only the fitting period. Using the same features makes the input information available to all learned heads comparable.

The direct logistic model maps these eight numbers to a wet-day probability. The four-symbol model maps the same numbers to four softmax probabilities. Its labels are recorded zero, positive but below 1 mm, 1 to below 5 mm, and at least 5 mm. Its loss rewards the probability assigned to the observed category. At evaluation we add the probabilities of the last two categories. The direct neural model uses eight tanh hidden units and a sigmoid wet-event output. It learns a nonlinear relationship between the eight features and the event. None of these three estimators uses written weather reports, radar or numerical weather model outputs.

Two frequency forecasts set the reference level. One estimates the wet rate for the target month. The other estimates it for that month among dates sharing the wet or dry state on d-2. Each table cell receives one added wet and one added dry count before division, so small cells can still produce a finite probability. Both references also use only the fitting window. This stronger second baseline matters because a learned model can appear to add information by merely rediscovering the seasons or persistence.

For model selection, training used dates from January 1981 through December 30, 2014, and validation used January 2015 through December 30, 2019. We chose among L2 coefficients 0.001, 0.01 and 0.1 by minimum validation Brier score on the binary event for every learned family. We then refit the selected models through December 30, 2019. The newest training label at the first validation issue and the newest selection or refit label at the first test issue both satisfy the two-date lag. The station clock advances through 2020-2025 and available weather history changes daily. Parameters remain fixed during that holdout. The optimizer is L-BFGS with analytic gradients. The neural initialization uses one fixed seed, an explicit limit on any comparison of training stability.

7. Can an observation model preserve a direct event forecast exactly?

One can add detail to an event probability while leaving the probability itself unchanged. Let q be a fitted model's probability of a wet day, a be the conditional probability of a positive measurement given a below-threshold day, and b be the conditional probability of at least 5 mm given a wet day. The four category probabilities, in order of amount, become (1-q)(1-a), (1-q)a, q(1-b), and qb. They are nonnegative and exhaust the possible observations. The last two sum to q for any values of a and b.

We implemented this factorization with the fitted direct logistic model as the gate. Using its fixed standardized inputs and L2 strength, we fitted two conditional logistic models on the corresponding branches of the pre-test records. We compared the marginal from this four-symbol model with the direct gate on all 4,348 test rows. The

largest absolute difference was below 0.000000000001. This control verifies the implementation of marginalization. Equality is built into its design. Independently training the unconstrained four-way softmax supplies the empirical test in the next sections.

Why can the unconstrained softmax give a different answer when it uses exactly the same observations? Its log odds for wet versus dry are the difference between two log-sums of exponentials of linear predictors. That difference need not be linear in the eight inputs. It also receives information about severity within each side of the threshold. The binary logistic model uses one linear log odds and one binary target. A performance gap can reflect parameterization, supervision, or their interaction. The factorized control shows precisely when encoding a richer future remains compatible with the same probability for the event of interest.

8. Calibration, skill and the cost of a confident error

For a binary outcome y and an issued probability p, the Brier loss is (p - y) squared. A forecast of 0.8 costs 0.04 when the event happens and 0.64 when it does not. Logarithmic loss applies a heavier penalty to an extremely confident error. Both are strictly proper for binary forecasts. In expectation, the loss is minimized by assigning the event its true probability under the forecaster's information. The two losses can rank imperfect models differently. Reporting their averages on the same resolved occasions gives readers two views of the errors. [6]

A reliability diagram groups issued probabilities and compares the mean probability in each group with its observed event frequency. The number of cases and the range of forecasts in each group matter. A few rare high-probability days can make a point appear far from the diagonal by chance. Empty groups contain no evidence about calibration. A constant forecast equal to the population event rate can be perfectly calibrated and still offer no information about how this day differs from another. Assess calibration alongside score differences against the historical base rate and a baseline that uses recent observations. [7]

A time series creates a separate problem for uncertainty. Adjacent forecast errors can share the same weather system or the same observed episode. Compare the models on matched issue times and preserve runs of consecutive observations when estimating uncertainty. A single historical test period may describe one particular regime. A prospectively stored forecast offers a stronger check that the model truly had its stated inputs at the moment it issued the probability. [1][5]

9. How the predictions were scored

Each eligible date produces one probability per model and one later binary outcome. Binary Brier loss is the squared distance between the probability and the zero-or-one outcome. Binary log loss is the negative natural logarithm of the probability assigned to the realized outcome. We report means within a station. Both scores reward honest probabilities in expectation. Their units differ and their sensitivities differ. A nearly certain forecast of the wrong outcome incurs at most one squared-error unit but can incur arbitrarily large log loss. [6]

The headline comparison uses differences in squared error for the same dates. For descriptive uncertainty we sample circular blocks of thirty calendar days with replacement, keep models paired, and count only eligible days in the resulting sequence. We draw a test-period-length sequence 4,000 times per station. The 2.5th and 97.5th percentiles give the reported interval. The calendar grid retains gaps, so two dates a week apart due to

missing records never become adjacent days in the original series. Wraparound joins the end and beginning of the held-out calendar. We repeat the calculation at seven and ninety days to show dependence on that choice.

The intervals describe variation across resampled stretches of this particular holdout with the fitted models fixed. They omit uncertainty from station selection, historical revisions, model fitting and future changes in climate or observation practice. There are multiple model contrasts, a single fit per station and no geographic replication beyond these two sites. We report the direction and scale of the differences and expose the underlying ledger so a reviewer can choose another scoring rule or time interval. [1][6][7]

10. Forecast skill on six held-out years

There were 150 wet days among 2,192 evaluated dates in Los Angeles (6.84%) and 698 among 2,156 in Seattle (32.37%). Five models saw the same eligible dates at each station. All three learned forecasts had lower mean Brier loss than both frequency references at both airports. The smallest recorded score belonged to the nonlinear direct event model in Los Angeles (0.057873) and to the summed four-symbol model in Seattle (0.186636). These rankings identify what happened in this holdout. They are insufficient by themselves to establish the size or transferability of a model advantage.

The specific comparison motivated by the paper is direct logistic versus the independently fitted four-symbol output. We subtract the token marginal's Brier from the direct model's Brier, so a positive difference favors the token marginal. The difference was +0.000024 [-0.000013, +0.000066] in Los Angeles and +0.000142 [+0.000055, +0.000229] in Seattle. The Seattle difference is small even though its thirty-day interval lies above zero. The fitted models answer the same verifiable question with very similar probabilities.

The nonlinear event model's Brier relative to the month-and-state reference was -0.000956 [-0.001897, -0.000097] in Los Angeles and -0.001548 [-0.003667, +0.000622] in Seattle. Negative values favor the neural forecast. The Seattle thirty-day interval includes zero. In Los Angeles the interval excludes zero at seven and thirty days, but includes it with ninety-day blocks. The Seattle direct-minus-token contrast remains positive at all three block lengths and in each of the six separate calendar-year estimates. A tiny difference reproduced across years at one station still leaves transport to a new location untested.

Table 2. Held-out average forecast losses

Estimator	LA Brier	LA log	Seattle Brier	Seattle log
Monthly frequency	0.060248	0.219208	0.190585	0.558453
Month and recent wet/dry state	0.058830	0.214421	0.188356	0.552807
Direct logistic event model	0.058269	0.210618	0.186778	0.549245
Four precipitation symbols, summed	0.058245	0.210484	0.186636	0.548879
Direct neural event model	0.057873	0.209680	0.186807	0.549931

Lower is better in every score column. Each Los Angeles estimate uses 2,192 dates and each Seattle estimate uses 2,156. Comparisons should be made within a station because base rates and underlying conditions differ.

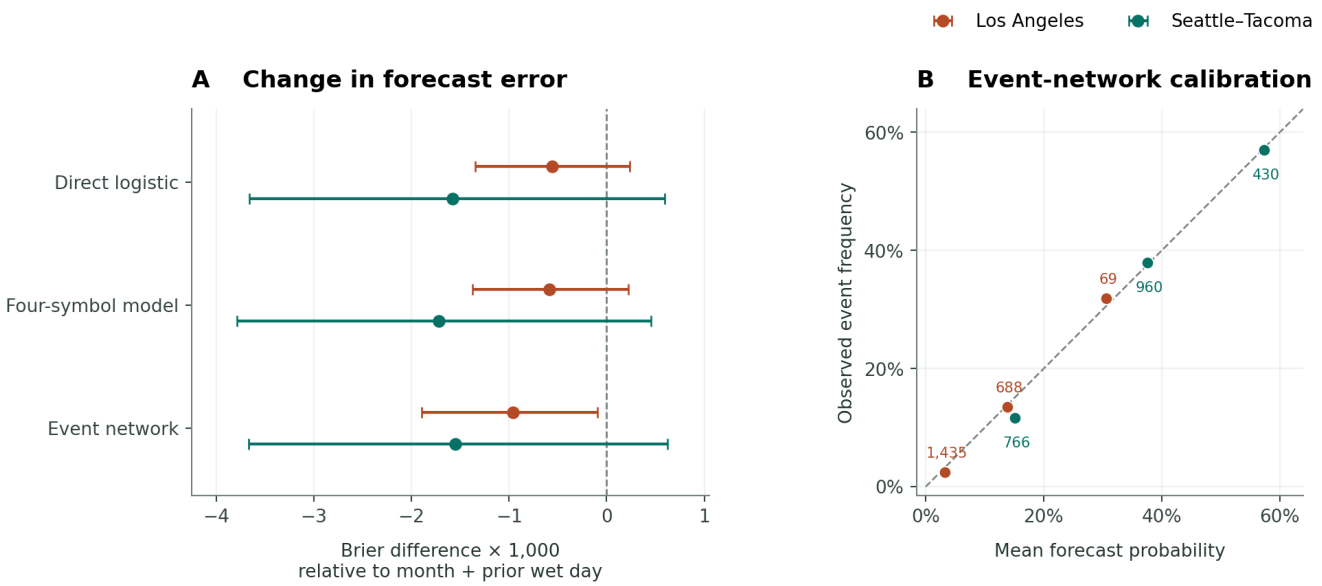
11. What a probability meant on these dates

Consider the forecasts whose Seattle neural probability fell between 10% inclusive and 25% exclusive. There were 766 such dates. The average forecast was 15.10% and 89 of them were wet, or 11.62%. A calibration plot places this group below the diagonal because the observed frequency was lower than the issued probability. A person reading a single 15% forecast could never see this discrepancy from the outcome of that one date. Pooling forecasts in similar ranges makes the scale of the error visible. [7]

In Los Angeles most direct neural probabilities fell below 10%. High-probability bins at both stations can be empty. We display counts rather than drawing a reliability estimate for a group with no observations. Binning combines distinct forecasts, the observations are serially related, and bins can conceal differences by season. The diagram is a diagnostic of what occurred in the six-year evaluation. It does not certify calibration on an arbitrary future day. [7]

A concrete ledger row illustrates how the resolver works. On the first held-out station date, January 1, 2020, the reconstructed model forecasts for the two airports differed. Both stations eventually recorded zero precipitation for that date. Each model's individual squared loss for that observation was its issued probability squared. The formula treats a modest chance on a dry day as a modest error and reserves larger penalties for stronger mistaken confidence. A run of forecasts is necessary to learn how the forecaster performs on the days when precipitation actually occurs.

Figure 2. Paired forecast skill and calibration on held-out station-days



A Bars show paired 95% block-bootstrap intervals. Lower scores indicate smaller errors.

B Labels count eligible test days in each occupied probability bin. The diagonal is perfect calibration.

Left: mean Brier difference from the month-and-state baseline, multiplied by 1,000, with 30-day paired block intervals. Negative values favor the learned forecast. Right: neural event forecast by probability bin. Labels give the number of eligible days. Only occupied bins appear. Station names denote the gauges used in the holdout. [13][14]

12. A forecast issued before the outcome exists

The test above reconstructs past decisions. We also issued two predictions for a future date and preserved their inputs and probabilities in a dated record. When we obtained the station snapshots on November 20, their most recent usable precipitation observations were for November 15. A forecast for November 21 using the two-day design would have required observations through November 19. We therefore specified and trained a separate direct event network with a seven-day input cutoff. All seven observations from November 8 through 14 were available for both stations. They were entered unchanged. The target is next day precipitation; seven days describes the age of the newest admissible reading relative to that target. [11][12][15]

The supplemental model has the same eight predictors and eight tanh units, with measurement lags shifted to seven through thirteen days before the target. It was trained at each station through December 25, 2014, selected by validation Brier through December 25, 2019, and refitted through that second cutoff. In a separate six-year retrospective evaluation, Los Angeles had 150 events in 2,192 eligible dates. Its Brier was 0.060206 for the neural model, 0.060248 for the monthly reference and 0.060203 for the month-and-state reference using the seven-day lag. Seattle had 706 events in 2,157 eligible dates. The three corresponding scores were 0.189944, 0.190824 and 0.191224. The Seattle denominator changes with the required history. A comparison of forecast horizons would require rematching the dates. Here the longer-lag neural model offered little improvement over a monthly reference; its trained probability remains a useful test of the issuance protocol. This supplemental design was chosen after examining archival availability, and its historical scores are exploratory.

At 8:24:43 a.m. Pacific on November 20, 2026, the neural forecast was 11.74% for Los Angeles International and 51.49% for Seattle-Tacoma International for the November 21 station day. The respective month-only references were 8.87% and 53.33%. These probabilities were recorded before either station's target day began. The table is a dated prediction, and the full-precision record provides the exact inputs and model digest. Two future outcomes will provide individual scores when verified. They cannot establish calibration by themselves. [15]

The companion resolution protocol fixes the archive check in advance. For each station, use the first archived snapshot retrieved by a reviewer on or after noon Pacific on December 21, 2026. Preserve its retrieval time and digest. Score a valid target measurement with the original 1 mm rule. A missing or ineligible measurement remains unresolved at that check. Record any later revision separately. The issuance record keeps its original contents. Both outcomes were unresolved when this edition was published. [15][16]

Table 3. Prospective precipitation forecasts for November 21, 2026

Station	Direct neural	Month-only	Status at issue
Los Angeles International	11.74%	8.87%	Unresolved
Seattle-Tacoma International	51.49%	53.33%	Unresolved

Station day, event threshold and available observations are frozen in the issue record. A later valid observation is needed to score either probability. [15][16]

13. A future test for an assistant that remembers

Imagine a wearable assistant that last saw a striped mug on a counter. An actionable forecast might ask whether the wearer will touch that particular mug within ten seconds of a designated cue. A candidate model could use remembered identity and previous encounters together with current hand position. Ego4D provides a related benchmark for anticipating human-object interactions, including the next active object, the action and the time to contact. A probability-of-contact study would need explicit issue times and complete eligible windows, including windows with no contact. Collecting only clips centered on annotated interactions would change the event rate being estimated. [8]

Compare a memory-enabled model with an otherwise matched model using recent frames only. Freeze the object identity rule and keep whole people and scenes separate across splits. Score predictions at several deadlines and report both resolved coverage and uncertainty across held-out people. Distinguish a witnessed absence of contact from loss of camera coverage. A second camera can help resolve an outcome while remaining unavailable to the wearable at issue time. Prior memory experiments motivate the choice of inputs. Establishing a forecasting advantage requires new observations and scored predictions.

Start a prospective run by issuing forecasts silently. If the device later prompts the wearer, record the intervention and analyze that policy separately. A reminder might itself cause the person to pick up the mug. Perdomo and colleagues formalize this feedback between deployed predictions and outcomes. Lakkaraju and colleagues show how prior decisions can also determine which outcomes receive labels at all. Preserve the

entire issue ledger and report the dispositions of missing, delayed and disputed outcomes. Those records help distinguish improved anticipation from a change in behavior or in what was possible to observe. [9][10]

14. What a future-event model would need to earn a broader claim

This study tests a limited prediction system with a resolver that is inexpensive to check. A more general model would have to accept a structured question specifying entity, event, horizon and evidence policy. A shared encoder could process language, current sensor readings and retained episodes. A binary event head could output the probability of the named outcome. A sequence or time-to-event head could support related questions when labels are sparse. The operational test is whether a fixed system improves forecasts across a declared set of future events using only information available at issue time. A forecast log must record both the question and its deadline to make that test possible.

Several inferences remain open. Neither station supplies evidence about all regions, neither event head was trained on textual next-word prediction, and these features contain no atmospheric model or radar forecast. There is no comparison with a professional weather service. The study also gives no estimate of superiority on unconstrained public affairs questions. Archive revisions and unknown reporting lags leave a gap between this retrospective exercise and a prospective trial. Freeze the data-vintage and model policy at issue, expose the resolution rules, preserve skipped and unresolved cases, then collect results as deadlines pass. That prospective protocol would test the same idea against a harder standard. [1][5][11]

The result most relevant to the proposed AI design is that both kinds of output can make genuine forecasts. Direct event supervision tells a model exactly which uncertainty matters to the user. A model of possible future observations can support several event definitions by summing compatible outcomes. On these two stations, the nonlinear direct model scored best at one and the category model at the other. The practical choice depends on which events will be queried, which labels exist, and which probabilities remain accurate when the deadline arrives.

References

[1] Dawid AP. Statistical theory: The prequential approach. Journal of the Royal Statistical Society, Series A, 1984. Sequential probabilities and verification against later observations. Accessed 2026-11-20.

[2] Salinas D, Flunkert V, Gasthaus J, Januschowski T. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. International Journal of Forecasting, 2020. Generative forecasting with autoregressive distributions and sample paths. Accessed 2026-11-20.

[3] Ansari AF and colleagues. Chronos: Learning the language of time series. Transactions on Machine Learning Research, 2024. Scaling, quantization and probabilistic time-series tokens. Accessed 2026-11-20.

[4] Halawi D and colleagues. Approaching human-level forecasting with language models. NeurIPS, 2024. Evidence retrieval, aggregation and evaluation on forecasting questions. Accessed 2026-11-20.

[5] Karger E and colleagues. ForecastBench: A dynamic benchmark of AI forecasting capabilities. ICLR, 2025. Prospective forecasting questions, dated submissions and standardized resolution. Accessed 2026-11-20.

[6] Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 2007. Proper scoring rules and their expected-loss interpretation. Accessed 2026-11-20.

[7] Gneiting T, Balabdaoui F, Raftery AE. Probabilistic forecasts, calibration and sharpness. Journal of the Royal Statistical Society, Series B, 2007. Forecast concentration and empirical calibration. Accessed 2026-11-20.

[8] Grauman K and colleagues. Ego4D: Around the world in 3,000 hours of egocentric video. CVPR, 2022. Section 5.5 describes first-person forecasting tasks including object interaction anticipation. Accessed 2026-11-20.

[9] Perdomo J, Zrnic T, Mendler-Dünner C, Hardt M. Performative prediction. ICML, 2020. Predictions that alter the distribution of subsequent outcomes. Accessed 2026-11-20.

[10] Lakkaraju H, Kleinberg J, Leskovec J, Ludwig J, Mullainathan S. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. KDD, 2017. Previous decisions constrain which labels are observed. Accessed 2026-11-20.

[11] NOAA National Centers for Environmental Information. Global Historical Climatology Network daily. Archive description, daily station records, quality controls and update process. Accessed 2026-11-20.

[12] NOAA National Centers for Environmental Information. GHCN-Daily file format and flags. Fixed-width fields, precipitation units, missing values, trace and quality flags. Accessed 2026-11-20.

[13] NOAA National Centers for Environmental Information. Los Angeles International Airport station USW00023174. Original station measurements. The reviewer snapshot records the uncompressed file digest and retrieval date. Accessed 2026-11-20.

[14] NOAA National Centers for Environmental Information. Seattle-Tacoma International Airport station USW00024233. Original station measurements used for the independent Seattle holdout; supplied as a fixed snapshot to reviewers. Accessed 2026-11-20.

[15] Matha R. Forecasts issued before the November 21, 2026 station day. Immutable issue record for two prospective precipitation forecasts. Includes feature dates, model fingerprint, source fingerprints and resolution status at issue. Accessed 2026-11-20.

[16] Matha R. Resolution protocol for the November 21 prospective forecasts. Dated companion protocol links to the original issuance digest and specifies the eligible archive vintage and recording of revisions. Accessed 2026-11-20.

Reproducibility record

Private review materials: https://github.com/RevanthM/future-event-probability | Research snapshot: 80da1b2ac0d6d3c8b30db793f20dec261cf859a3