Forecasting Ambient Solar Wind Speed at L1 with Gradient-Boosted Trees and a Persistence-Recurrence Blend

Atish Joottun and Munazir Ramjhun

Team Icarus · November 2026

ABSTRACT

We forecast the hourly bulk speed of the solar wind at the L1 Lagrange point for the 72 hours after an issue time t_0. The served forecast is a fixed mixture: one quarter from a gradient-boosted tree model (LightGBM, six boosters by lead group) trained on in-situ plasma history, and three quarters from a physics baseline that hands over from persistence to a 27-day solar-rotation recurrence. The model was chosen by an internal tournament of about forty entrants under pre-registered selection rules. Every input is gated so that only information published before t_0 can enter, and a five-tier fallback chain replaces failure with degradation. Scored with the competition metric S_{final} (lower is better), the mixture reaches $\mathbf{129.6}$ and $\mathbf{132.7\ km\,s^{-1}}$ on the held-out periods 2020-2024 and 2025-2026, against $\mathbf{140.8}$ and $\mathbf{139.1}$ for the baseline alone. Its timeline RMSE on 2016-2019 is $\mathbf{77.5\ km\,s^{-1}}$, within about $\mathbf{10\ km\,s^{-1}}$ of published imagery-driven deep networks with over a thousand times more parameters, while using no imagery and training in seconds on a laptop CPU.

Keywords: solar wind, space weather, L1, gradient boosting, recurrence, model selection, forecast verification

1 Introduction

The solar wind is a stream of plasma flowing out from the Sun. Its speed near Earth ranges from about $\mathbf{300\ km\,s^{-1}}$ in slow wind to over $\mathbf{700\ km\,s^{-1}}$ in high-speed streams from coronal holes, and fast streams drive geomagnetic storms. Spacecraft at L1, about $\mathbf{1.5 \times 10^6\ km}$ sunward of Earth, measure the wind roughly an hour before it arrives. The task is to forecast that speed every hour for the next three days.

Two physical facts carry most of the skill. Over a few hours the wind changes slowly, so the best guess is the current value (*persistence*). Over days, solar rotation matters: a coronal hole that produced a fast stream faces Earth again one synodic rotation later, about **27.3** days, so the wind one rotation ago is a useful guess (*recurrence*) [3]. Coronal holes are visible as dark regions in extreme-ultraviolet images (Figure 1). Our system combines both with a machine-learned correction, under strict rules on what information may be used and on never failing to deliver a forecast.

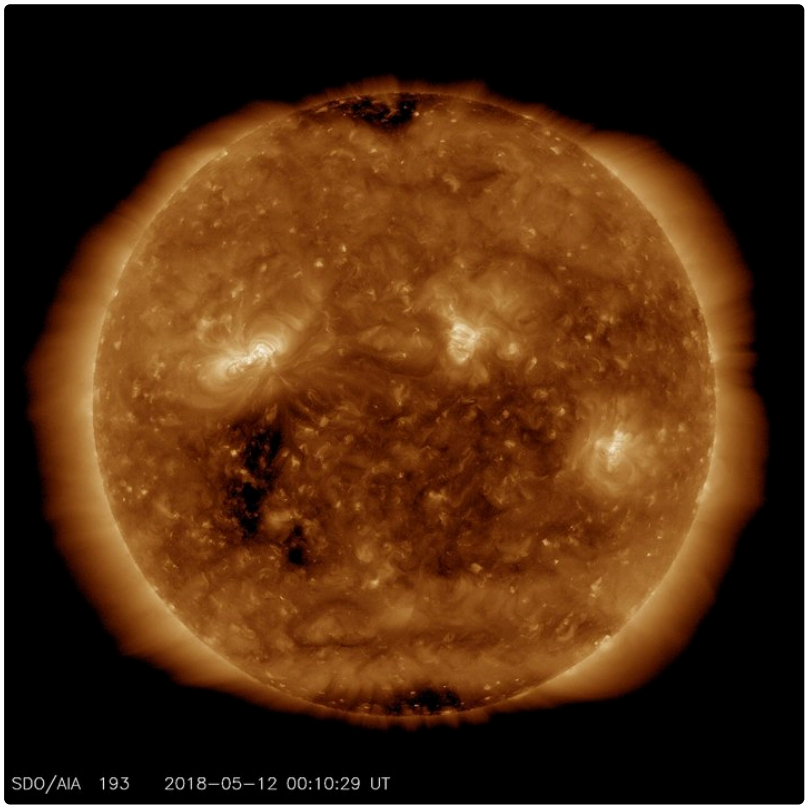

Figure 1. The Sun at 19.3 nm (SDO/AIA 193), 12 May 2018, inside our validation period. The dark regions are coronal holes: open magnetic field from which fast wind escapes. The low-latitude hole left of centre, and the polar holes, are the kind of long-lived structure that makes 27-day recurrence work near solar minimum. Image: NASA/SDO and the AIA, EVE and HMI science teams, public domain, via Wikimedia Commons.

2 Problem and metric

At each issue time t_0, every 6 hours, a fresh container must output 72 values $\hat{v}(t_0+\tau)$, $\tau = 1, \dots, 72$ h, within 15 minutes including the container download. The truth v is the hourly ACE/SWEPAM bulk speed [5]; coronal mass ejections are not removed. For lead τ and N forecasts, the ordinary error is

$$\mathrm{RMSE}_\tau = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{v}_{i\tau} - v_{i\tau}\right)^2}. \tag{1}$$

For the tail error, let $q_0 \le q_1 \le \cdots \le q_{100}$ be the 101 empirical percentiles of the observed speeds at lead τ, and let $M_\tau(q)$ be the mean squared error over the pairs whose *observed* speed is at least q:

$$M_\tau(q) = \frac{\sum_i \mathbf{1}\{v_{i\tau} \ge q\}\left(\hat{v}_{i\tau} - v_{i\tau}\right)^2}{\sum_i \mathbf{1}\{v_{i\tau} \ge q\}}. \tag{2}$$

Integrating the tail error over velocity gives

$$\mathrm{TIMSE}_\tau = \frac{1}{q_{100} - q_0}\sum_{k=0}^{99}\left(q_{k+1} - q_k\right) M_\tau(q_k). \tag{3}$$

With $\mathrm{TRMSE}_\tau = \sqrt{\mathrm{TIMSE}_\tau}$, the score is a geometric mean per lead, averaged over leads:

$$\begin{aligned} S_\tau &= \sqrt{\mathrm{RMSE}_\tau \cdot \mathrm{TRMSE}_\tau}, \\ S_{\text{final}} &= \frac{1}{72}\sum_{\tau=1}^{72} S_\tau. \end{aligned} \tag{4}$$

A forecast biased everywhere is punished through the RMSE half; one that flattens peaks is punished through the tail half. As a check of our implementation, a constant bias b on every value gives $M_\tau(q) = b^2$ for every q, so $\mathrm{TIMSE}_\tau = b^2$ and $S_\tau = |b|$ exactly. Every model choice in this work was ranked on the official metric of the day and on nothing else.

> The metric changed during the competition. Until 22 October 2026 the official score was a single *pooled* TIMSE over all leads, with thresholds at every distinct observed speed. Section 5 explains why that change mattered.

3 Data and the information cutoff

Training data come from ACE at L1, September 2000 to August 2026: the archived real-time SWEPAM product for inputs and the ACE Level 2 speed for labels. The live system reads the real-time product from NOAA SWPC, with DSCOVR and SOLAR-1 data as a secondary source and a history file packaged into the container as a third.

3.1 Availability gating

A forecast may use only information that existed before t_0. Each hourly value averages a window $[s, e)$ and is published some time after e. Every row carries its window start s, end e and availability time a: the stated publication time when known, otherwise $e + \delta$ with a positive latency margin δ (4 h for ACE). A row is usable if and only if

$$s < t_0 \ \wedge \ e < t_0 \ \wedge \ a \le t_0. \tag{5}$$

When a source does not say whether its time tag marks the start or end of the hour, we read it as the start, which places the window later and is the safe choice. Each feature is tested by overwriting every row that (5) bars with NaN and with $9999\ \mathrm{km\,s^{-1}}$, rebuilding, and checking the feature is bit-identical; a feature that passes cannot be reading the future.

3.2 Chronological splits

Because the Sun repeats roughly every 27 days, neighbouring windows are near-duplicates and a random split leaks. We split by time with a 180-day buffer between segments:

Segment	Period	Forecasts	Use
Train	2000-09-01 to 2015-12-31		Fitting
Validation	2016-06-29 to 2019-12-31	5,112	Every choice of model, weight and constant
Test	2020-06-29 to 2024-07-10	5,880	Read once per decision
Recent	2025-01-07 to 2026-08-31	2,396	Read once per decision; closest to the live period

Table 1. Data segments. One forecast every 6 hours.

4 Method

4.1 Physics baseline: persistence to recurrence

Let v_0 be the last usable observed speed. The recurrence forecast weights three lags around the synodic period of 27.28 d, because the rotation period of a coronal hole depends on its latitude:

$$R(\tau) = \sum_{j=1}^{3} w_j \, v\big(t_0 + \tau - P_j\big), \tag{6}$$

with lags $P_j = 26.28,\ 27.28,\ 28.28$ d and weights $w_j = \tfrac{1}{4},\ \tfrac{1}{2},\ \tfrac{1}{4}$.

The baseline moves linearly from persistence to recurrence over the first $L = 24$ hours:

$$\begin{aligned} B(\tau) &= \big(1 - \lambda(\tau)\big)\, v_0 + \lambda(\tau)\, R(\tau), \\ \lambda(\tau) &= \min\!\left(\frac{\tau}{L},\, 1\right). \end{aligned} \tag{7}$$

4.2 Machine-learned model

We use gradient-boosted decision trees, LightGBM [2]. Each training row is a pair (t_0, τ) with the lead as an input, and six boosters cover the lead groups 1-6, 7-12, 13-24, 25-48, 49-60 and 61-72 h. The model fits the squared error, so for a feature vector $\mathbf{x}$ built at t_0 each booster approximates the conditional mean $\mathbb{E}\left[v(t_0+\tau) \mid \mathbf{x}, \tau\right]$ as a sum of trees,

$$F(\tau) = F_0 + \eta \sum_{m=1}^{K_g} f_m(\mathbf{x}, \tau), \tag{8}$$

where F_0 is the mean training speed, each f_m is a regression tree with 15 leaves fitted to the residuals of the previous $m-1$ trees, η is the learning rate and K_g the number of rounds in lead group g. The shipped rounds are $K_g = 242, 158, 72, 27, 3, 14$: 516 trees, about 7.7×10^3 leaf values and 1.3 MB on disk. The inputs, all built from rows passing (5) over a 24 h lookback:

- **In-situ plasma.** Lagged speed, means, spreads and extremes over 6 to 168 h, speed gradients, proton density n_p, temperature T_p, the ratio T_p/T_{exp} with $T_{\mathrm{exp}}/\mathrm{K} = 640\,(v/\mathrm{km\,s^{-1}}) - 1.56 \times 10^5$ [4], and dynamic pressure $p_{\mathrm{dyn}} = m_p n_p v^2$. Compressed, hot plasma ahead of a fast stream makes density and temperature leading indicators.
- **Ballistic back-mapping.** The transit time from the Sun, $t_{\mathrm{tr}} = d/v$ with $d = 1\ \mathrm{AU} - 1.5 \times 10^6\ \mathrm{km}$, which ranges from 2.2 d at $780\ \mathrm{km\,s^{-1}}$ to 5.8 d at $295\ \mathrm{km\,s^{-1}}$, and the solar longitude the source has rotated through, $\Omega_{\mathrm{syn}}\, t_{\mathrm{tr}}$. This was the most used feature, with 21% of the split gain.
- **Time and data quality.** The lead τ, day of year as $\sin$ and $\cos$, the current gap length and the age of the newest observation.

4.3 The served forecast and guardrails

With $g(\cdot)$ the guardrails below, the served forecast is

$$\hat{v}(\tau) = \alpha\, g\big(F(\tau)\big) + (1-\alpha)\, g\big(B(\tau)\big), \tag{9}$$

with $\alpha = 0.25$. The guardrails clip to $[200, 900]\ \mathrm{km\,s^{-1}}$, halve the excess above $730\ \mathrm{km\,s^{-1}}$, and limit every hourly step to under $150\ \mathrm{km\,s^{-1}}$, starting from v_0 so that hour 1 is continuous with the last measurement. A validator then checks 72 finite values in range with no larger step. A forecast that fails is never repaired; the next candidate serves instead.

5 The model tournament

We did not pick a model by hand. Every candidate, from baselines to boosted trees, was a registered *entrant* in one harness: it was trained on the training segment, forecast every validation origin through the same t_0 gate and validator as the live system, and was scored on the official metric. An entrant that produced a single invalid forecast was marked failed. Each result was appended to a leaderboard with its code commit, data specification and an uncertainty range from a bootstrap over solar rotations.

Choosing among many entrants on one validation set is itself a source of optimism, so every decision that mattered was made by a rule written and committed *before* its numbers were read. The rules compare two forecasts X and Y on the same origins through a paired bootstrap of the score difference. Each replicate $b = 1, \dots, B_{\mathrm{boot}}$ resamples whole 27-day solar rotations with replacement, which respects the autocorrelation that a resampling of single hours would ignore:

$$\begin{aligned} \Delta_b &= S\big(X_b\big) - S\big(Y_b\big), \\ X \text{ is preferred} \iff & \; Q_{0.95}\big(\{\Delta_b\}\big) < 0, \end{aligned} \tag{10}$$

where $Q_{0.95}$ is the 95th percentile over replicates, so X wins only if it is better in at least 95% of resampled histories.

5.1 Rounds

BASELINES

Persistence, 27-day recurrence, the three-lag recurrence (6), climatology and the blend (7) at nine handover lengths. These set the bar every learned entrant had to clear.

ROUND 1 · IN-SITU BOOSTED TREES

One model for all leads, six grouped boosters, and 72 per-lead boosters, then a one-axis sweep of learning rate, leaves (15, 31, 127), minimum leaf size, L2 penalty and lookback (24, 72, 168 h). Grouped boosters with 15 leaves won. The per-lead variant was disqualified: it forecast 176.8 km/s, below the physical range.

ROUND 2 · CORONAL-HOLE IMAGERY

An imagery pipeline producing 80 coronal-hole features from AIA images was built and tested on synthetic data, but its ablation needs about 30 GB of images and was not run before the freeze. Imagery is therefore masked.

ROUND 3 · PEAK CALIBRATION

Distribution transforms on the output and quantile objectives, the method that published work credits with a 22-23% gain in peak error [6]. Every transform variant was reliably worse on our metric, with the whole bootstrap interval above zero. None ships.

FINAL SELECTION · PRE-REGISTERED 18 OCT

Candidates A (15 leaves) and B (15 leaves, 24 h lookback). B scored lower; within noise, the rule prefers B because it reads less history. B became the frozen model.

RECURRENCE SWAP · 25 OCT

Candidate C adds the 27-day recurrence features to the model. It was 17% better on validation but 26% worse on test, and was rejected.

METRIC CHANGE · 22 OCT

The organisers replaced the pooled TIMSE with (4). Re-scoring showed the frozen model biased high, and the mixture weight α in (9) was chosen on validation from $\{0.1, 0.25, 0.4, 0.5\}$.

RETRAIN AND BLEND SHAPE · NOVEMBER

A model retrained for the new metric, and a slower capped handover in (7). Neither passed its pre-registered rule, so the configuration above ships unchanged.

5.2 Round 1 results

Entrant	Pooled TIMSE (millions)	Skill vs persistence	Bias (km/s)
Grouped, 15 leaves, 24 h lookback (B, shipped)	**8.01**	**+56.1%**	**+128**
Grouped, 15 leaves (A)	8.06	+55.9%	+132
Grouped, 31 leaves	8.19	+55.1%	+132
Grouped, default settings	8.30	+54.5%	+132
Grouped, 127 leaves	8.32	+54.4%	+130
Single model for all leads	8.60	+52.9%	+136
Per-lead boosters			failed the validator
Constant 600 km/s	9.32	+48.9%	
Recurrence, three lags	12.78	+30.0%	+4
Single model, unweighted	18.12	+0.8%	+3
Persistence	18.26	0	+2
Climatology	24.47	−34.0%	+2

Table 2. Validation 2016-2019 under the metric of the time, pooled TIMSE, lower is better. Skill is 1 − TIMSE / $\mathrm{TIMSE}_{\mathrm{persistence}}$.

Table 2 holds the lesson that shaped everything after it. Under the old metric a flat $600\ \mathrm{km\,s^{-1}}$ beat recurrence, and the model only beat that constant by 11%. An hour counted once in every threshold strip below its observed speed, so over-forecasting slow wind was almost free and under-forecasting fast wind was expensive. Training with weights that reproduced the metric therefore taught the model to sit about $125\ \mathrm{km\,s^{-1}}$ high; the same model without weights scored no better than persistence. The model did what the metric asked.

5.3 Why the recurrence model was rejected

Candidate C read the 27-day recurrence features and scored a pooled TIMSE of $6.62 \times 10^{6}\ \mathrm{km^{2}\,s^{-2}}$ on validation against 8.01×10^{6} for B. On test it scored 29.9×10^{6} against 23.6×10^{6}. The validation years are solar minimum, when long-lived coronal holes make recurrence very reliable; 2020-2024 is the rising cycle, when holes evolve within a rotation and recurrence misleads. This is the 180-day-buffer problem the literature warns about [6] in a new form: a split can be leak-free and still unrepresentative.

5.4 Re-selection after the metric change

Under (4) a bias b costs $|b|$ directly, so the frozen model alone became worse than the baseline. Rather than retrain days before the freeze, the mixture (9) shrinks the bias by $1-\alpha$. With the model refit on the training segment only, so that validation is out of sample, the grid gave

α	0.10	0.25	0.40	0.50
$S_{\mathrm{final}}(\alpha)$ $[\mathrm{km\,s^{-1}}]$	90.08	**89.77**	92.99	96.57

against 92.75 for the baseline and 126.98 for the model alone. On validation the chosen mixture has $\mathrm{RMSE} = 77.5$, $\mathrm{TRMSE} = 104.2$ and bias $+32.9\ \mathrm{km\,s^{-1}}$, down from $+125$ for the model alone.

The retrained model (residual target, no tail weighting, bias under $7\ \mathrm{km\,s^{-1}}$) was served at $\alpha = 0.75$ and scored 129.47 on test against 129.56, and 133.64 on recent against 132.80. The pre-registered bar was a 3% gain on recent with the 90% paired interval below zero; the interval was $[-1.5,\ +3.3]\ \mathrm{km\,s^{-1}}$. The two are tied, so the simpler incumbent ships. A slower handover in (7) improved validation by $1.2\ \mathrm{km\,s^{-1}}$ at best, but a neighbouring grid cell lost 2.0, failing the stability check.

6 Robustness by construction

A container that crashes or runs late scores nothing, and nobody can intervene during the three-month live run. No step on the inference path may raise. Inputs pass through a fallback chain, each tier recorded in the log:

live feed → secondary feed → packaged history → 27-day recurrence → flat 470 km/s

Every network call has a timeout, bounded retries and a byte cap, and a watchdog writes a valid fallback forecast before the time budget runs out. The flat constant minimises S_{final} among constants on validation. We tested this with a 360-run replay of a full live period (all 360 forecasts valid, in every degraded tier), a 44-check container test against a mock upload server, about 6,600 unit and property tests, and mutation testing of the safety modules, which found for example that no test checked the hour-1 continuity with v_0; one now does.

7 Results

All model rows are out of sample: the model was refit on 2000-2015 to score validation and on 2000-2019 to score the later periods. The baselines fit nothing.

Method	Validation 2016-19	Test 2020-24	Recent 2025-26
Served mixture, eq. (9), α = 0.25	**89.8**	**129.6**	**132.7**
Baseline B, eq. (7)	92.8	140.8	139.1
LightGBM alone, eq. (8)	127.0	138.7	149.7
Persistence, our gated data (about 5 h old)	121.5	122.9	162.1
Persistence, latest value at t_0	115.3	117.5	153.4
Flat 470 km/s	132.0	141.9	171.3

Table 3. S_{final} in km/s, lower is better. The two persistence rows differ only in the newest value they hold: ours is the last value our availability gate admits, about 5 h before t_0; the other is the last observation before t_0, as the organisers define their persistence baseline.

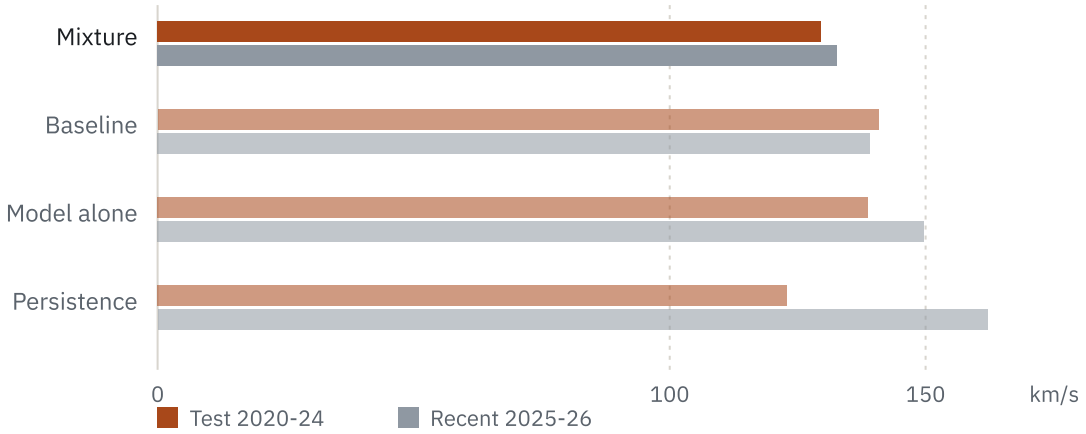

Figure 2. S_{final} by method. The mixture is the only method in the top two in both periods; persistence wins 2020-2024 and loses badly in 2025-2026.

7.1 Skill by lead time

A single S_{final} hides where the skill comes from. Figure 3 shows S_τ from (4) at every lead. Persistence is the best of the four for the first 13 to 16 h in 2016-2019 and 2025-2026, and for about 50 h in 2020-2024. At every later lead the mixture is best. In 2025-2026 it stays flat near $\mathbf{149\ km\,s^{-1}}$ while persistence degrades to $\mathbf{195\ km\,s^{-1}}$.

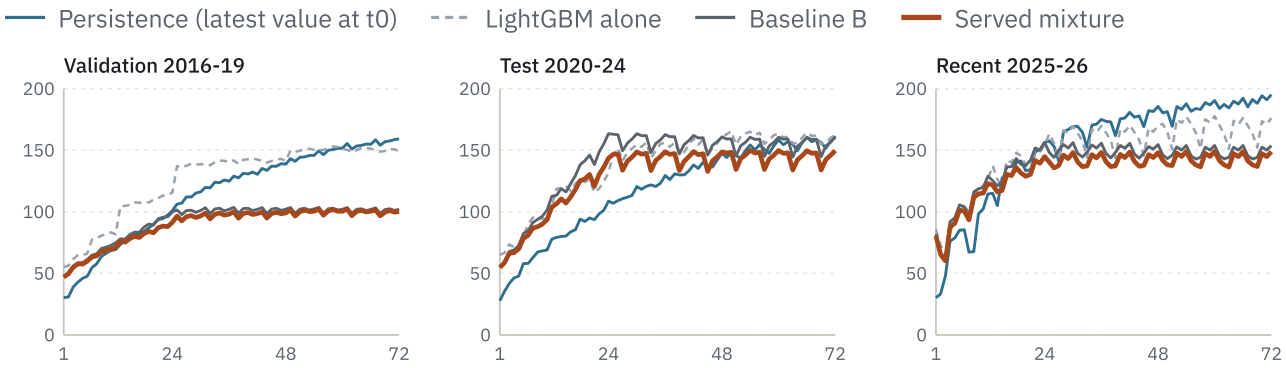

Figure 3. Per-lead score S_τ in km/s against lead time in hours, lower is better. All forecasts out of sample. Persistence here holds the last observation before t_0.

The early-lead gap has a specific cause. Our availability gate (5) treats ACE real-time data as published 4 h after the end of each hour, so the newest value a forecast may use is about 5 h old. At hour 1 our mixture scores $S_1 = 47, 55$ and $80\ \mathrm{km\,s^{-1}}$ in the three periods, against 28 to $30\ \mathrm{km\,s^{-1}}$ for a persistence that holds the value at t_0. The margin is a deliberate fail-closed choice: we use no value we cannot prove was published before t_0. Measuring the real publication latency, and shortening the margin only as far as that evidence allows, is the largest single improvement available to us.

Figure 4 shows why the model is mixed rather than served alone. Its bias grows with lead time to $+110$ to $+170\ \mathrm{km\,s^{-1}}$ at the longest leads, a legacy of training under the old metric (section 5.2). The baseline is nearly unbiased, and the mixture keeps the model's short-lead information while holding the bias under $+45\ \mathrm{km\,s^{-1}}$.

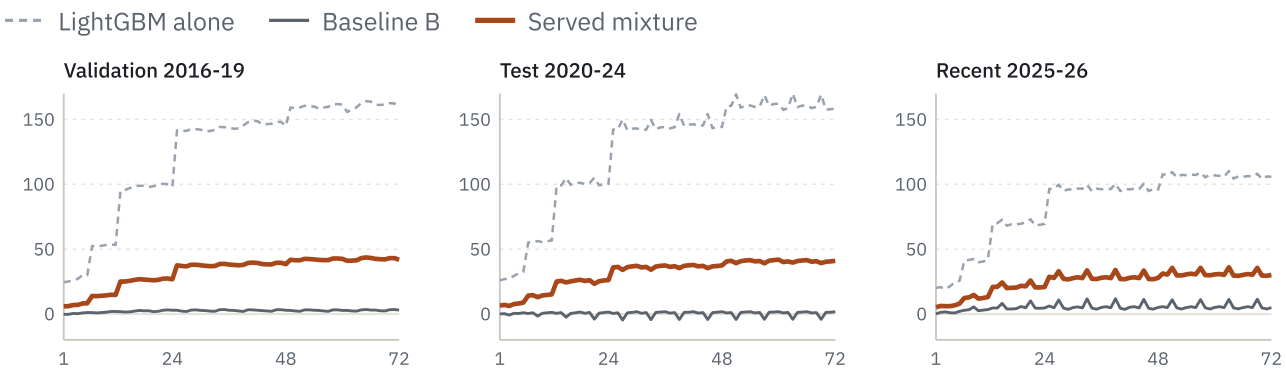

Figure 4. Mean bias, forecast minus observation, in km/s against lead time in hours. Positive means the forecast is too fast.

8 Comparison with published models

Published work reports the timeline RMSE (1), not our score, so Table 4 compares RMSE. The setups differ in period, cadence, lead and whether CMEs are removed, so the comparison is indicative. One anchor makes it fairer: 27-day recurrence, which no study tunes, scores $87.5\ \mathrm{km\,s^{-1}}$ on our validation and 89.1 in the setup of Collin et al. [6], so the two tasks are of similar difficulty. We also report skill against that baseline,

$$\mathrm{SS} = 1 - \frac{\mathrm{RMSE}}{\mathrm{RMSE}_{\mathrm{rec}}}. \tag{11}$$

Model	Inputs	Size and compute	RMSE (km/s)	Skill SS
Dhuri et al. 2025 [7], CNN encoder-decoder	EUV images, wind history	Pretrained GoogLeNet image encoder, GPU training	46.9 (day 1) 53.3 (day 4)	
Collin et al. 2024 [6], Lasso and polynomial	Coronal-hole areas from EUV, sunspots, recurrence	165 coefficients, full imagery segmentation pipeline	68.1	23.6%
Brown et al. 2022 [8], Swin Transformer	EUV images, recurrence	About 29 million parameters, GPU training	68.5	23.1%
Ours, baseline B	**Wind history only**	**No fitted parameters**	**74.2**	**15.2%**
Bailey et al. 2021 [9], gradient boosting	Coronal magnetic model outputs	Boosted trees on WSA-type model runs	75.6	15.2%
Raju and Das 2021 [10], CNN	EUV images	CNN trained from scratch, GPU	76.3	
Ours, served mixture	**Plasma history only**	**516 trees, about 7,700 leaves, 1.3 MB, about 20 s per fit on a laptop CPU, under 1 ms per forecast**	**77.5**	**11.4%**
Lin et al. 2023 [11], CNN	PFSS magnetic maps	CNN, GPU	80.8	
Upendran et al. 2020 [12], CNN	EUV images	Pretrained GoogLeNet, GPU	81.2	
EUHFORIA [6], 3D MHD	Magnetograms	Physics simulation on a cluster	90.8	
WSA-Enlil, operational [13]	Magnetograms	Coronal model and 3D MHD	about 100	
27-day recurrence, our validation	Wind history	None	87.5	0

Table 4. Timeline RMSE as published or as recomputed by Collin et al. Ours: validation 2016-2019, hourly, mean of RMSE_τ over leads 1 to 72 h, CMEs included. Collin: 2010-2019, 4-day lead, CMEs removed. Dhuri: 2019-2024, daily averages. Skill is given only where the study reports its own recurrence RMSE; Bailey uses Collin's recomputation.

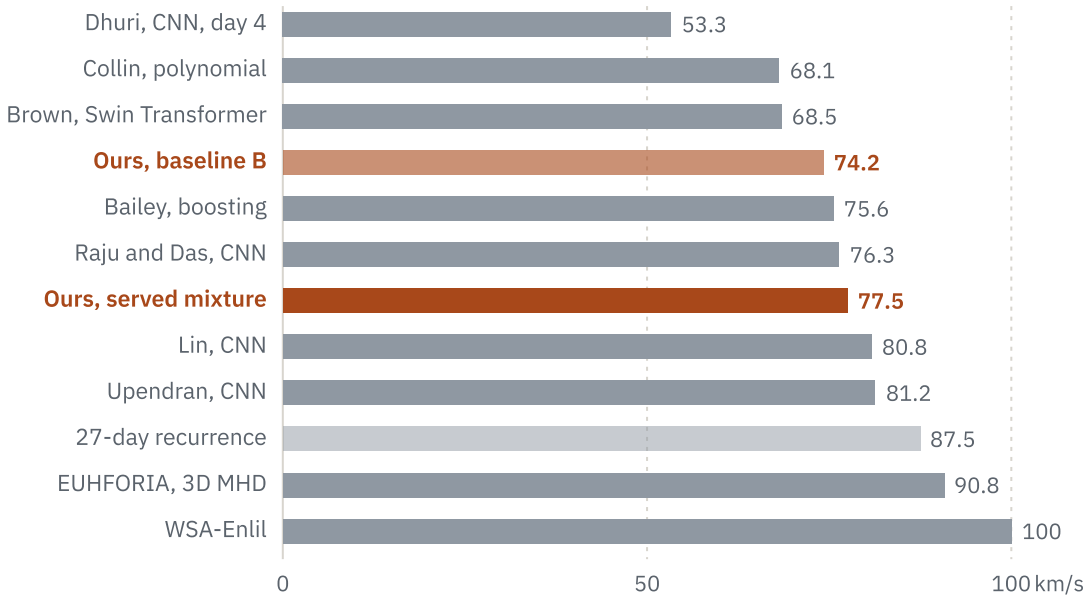

Figure 5. Timeline RMSE, lower is better. Our forecasts in orange. Setups differ (Table 4 caption).

Three readings follow. First, we sit in the middle of the published machine-learning models and ahead of both physics-based operational systems, with no solar imagery and no magnetograms. Second, the best results, about $\mathbf{68\ km\,s^{-1}}$, all use coronal-hole information from EUV images; that is the input we built but could not validate in time (Round 2), and it is where the remaining gap of about $\mathbf{10\ km\,s^{-1}}$ most plausibly lies. Third, capacity is not what separates the models: a 165-coefficient regression matched a 29-million-parameter transformer to within $\mathbf{0.4\ km\,s^{-1}}$ [6], and our 1.3 MB model trains in seconds. Note also that our mixture has a higher RMSE than our own baseline (**77.5** against **74.2**) yet a better score, because the model term lowers the tail error $\mathbf{TRMSE}$; a model tuned for RMSE alone would be tuned for the wrong target here.

9 Discussion

What worked. A weak learned correction mixed into a strong physical baseline beats either alone in every period, by **6** to $\mathbf{11\ km\,s^{-1}}$ against the baseline on held-out data. It is also the only method that is stable from solar minimum (2016-2019) to solar maximum (2025-2026). The tournament's main value was negative: it stopped two attractive ideas, peak calibration and recurrence features, that looked good in the literature or on validation and failed on our metric or on test.

Limits. Persistence beats us for the first 13 to 16 h, and for about 50 h in 2020-2024. Part of this is the 5 h age of our newest usable value (section 7.1); the rest suggests recurrence takes over too early when coronal holes evolve within a rotation. The model inherits a high bias from the old metric that the mixture only partly removes. The model is trained on ACE while operations move to SOLAR-1, and no offset correction is applied.

Future work. A handover that depends on the solar-cycle phase; coronal-hole imagery as a model input; and a SOLAR-1 to ACE correction once enough simultaneous data exist.

10 Conclusion

We built a solar wind forecasting system that is designed first never to fail and second to be accurate. A tournament of about forty entrants under pre-registered rules selected a six-booster LightGBM on in-situ plasma history; after the metric changed, a mixture of one quarter model and three quarters persistence-recurrence baseline gave the best and most stable score. It reaches $S_{\text{final}} = 129.6$ and $132.7\ \mathrm{km\,s^{-1}}$ on the two held-out periods and a timeline RMSE of $77.5\ \mathrm{km\,s^{-1}}$, close to imagery-driven deep networks at a small fraction of their compute.

References

1. Trillium Technologies, Queen Mary University of London, CU Boulder and FDL. *Benchlab Solar Wind Prediction AI Tournament 2026: participant guide* (1 September 2026) and *scoring specification* (22 October 2026).
2. Ke, G. et al. LightGBM: a highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems* 30 (2017).
3. Owens, M. J. et al. A 27 day persistence model of near-Earth solar wind conditions: a long lead-time forecast and a benchmark for dynamical models. *Space Weather* 11 (2013).
4. Elliott, H. A. et al. An improved expected temperature formula for identifying interplanetary coronal mass ejections. *Journal of Geophysical Research* 110 (2005).
5. McComas, D. J. et al. Solar Wind Electron Proton Alpha Monitor (SWEPAM) for the Advanced Composition Explorer. *Space Science Reviews* 86 (1998).
6. Collin, D. et al. Forecasting the solar wind speed from coronal hole areas. arXiv:2410.05068 (2024).
7. Dhuri, D. B. et al. A multi-modal deep learning model for solar wind speed forecasting. arXiv:2507.17298 (2025).
8. Brown, E. J. E. et al. Attention-based machine vision models and techniques for solar wind speed forecasting using solar EUV images. *Space Weather* 20 (2022).
9. Bailey, R. L. et al. Using gradient boosting regression to improve ambient solar wind model predictions. *Space Weather* 19 (2021).
10. Raju, H. and Das, S. CNN-based deep learning model for solar wind forecasting. arXiv:2108.09114 (2021).
11. Lin, R. et al. Solar wind speed prediction from PFSS source-surface maps with a convolutional network. arXiv:2304.01234 (2023).
12. Upendran, V. et al. Solar wind prediction using deep learning. arXiv:2006.05825 (2020).
13. Reiss, M. A. et al. Verification of high-speed solar wind stream forecasts using operational solar wind models. *Space Weather* 14 (2016).