

Paper #12 · Oral



02-06
NOV²⁰₂₆

MEXICAN INTERNATIONAL
CONFERENCE ON
ARTIFICIAL INTELLIGENCE
ITESM Campus Chihuahua

Forecasting Weekly Depression Incidence in Mexico: A Multi-Model Framework with Auditable Per-Series Model Selection

Javier Augusto Rebull-Saucedo¹ · Juan Carlos Pérez-Nava¹ · Luis Gerardo Sánchez-Salazar¹
Grettel Barceló-Alonso¹ · Ruth Pérez-Hernández²

¹ Tecnológico de Monterrey, Monterrey, Mexico

² Instituto Mexicano del Seguro Social, OOAD Guerrero, Mexico

Independent research; affiliations do not imply institutional endorsement.

*When historical scores meet recorded
2026 forecasts*

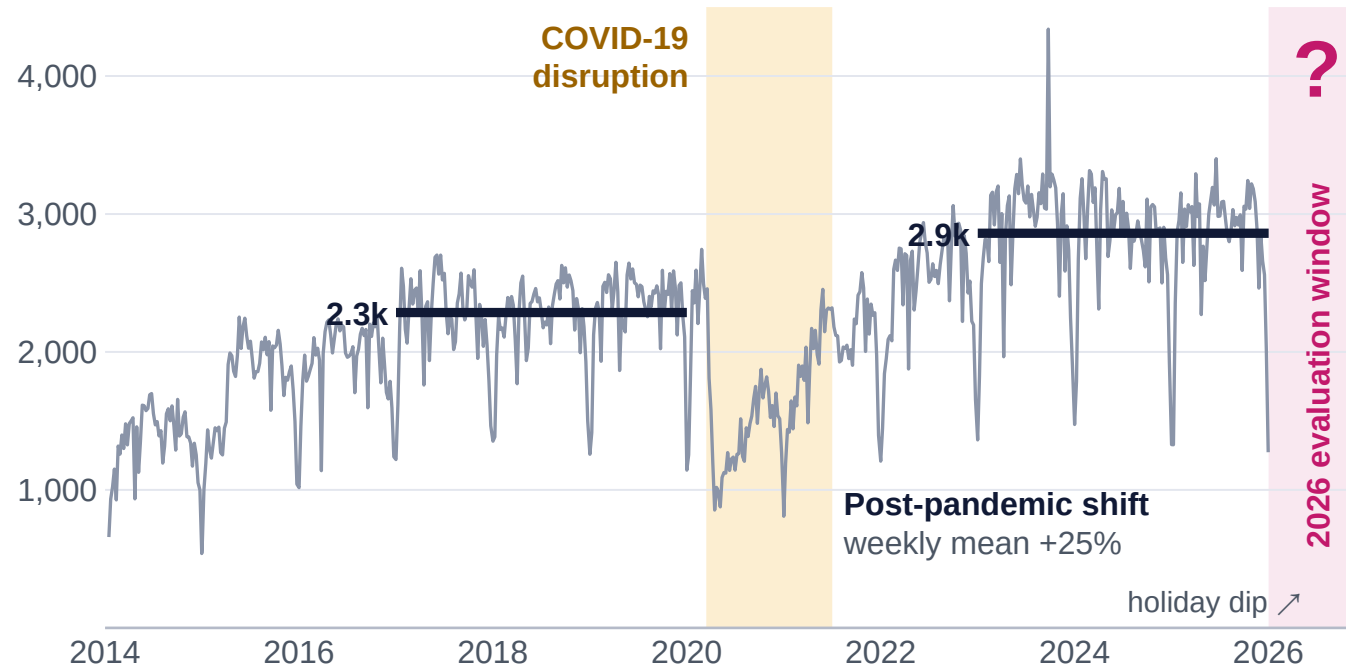


Tecnológico
de Monterrey



Depression is counted weekly, but only in hindsight

Weekly reported cases, national ICD-10 F32, SINAVE 2014–2025



~150k
cases per year
2023–2025

74%
are women
annual totals by sex

~22× total reported cases, 2014–2025



Source: SINAVE bulletins [1]; reported counts, not prevalence. Lines: weekly means 2017–19, 2023–25. Curve omits 2014-W02 and 2016-W20 (flagged partial).

Four candidates per series, one auditable rule



SINAVE bulletins

Weekly ICD-10 F32 counts, 2014–2025 (626 weeks)

111 targets: 32 states, 4 regions, national; general, women, men

99 reported directly



Four candidates

- **DeepAR** [3]
probabilistic recurrent net
- **Prophet** [2]
trend + seasonality
- **Ensemble** [4]
Prophet + XGBoost blend
- **Stacking** [5, 6]
Prophet, ETS, LightGBM



Deterministic selection rule

- 1 **Lowest historical sMAPE**
- 2 **Within 5% relative: lowest MASE**
- 3 **MASE tie: lowest RMSE**



Recorded forecast

52 weeks ahead



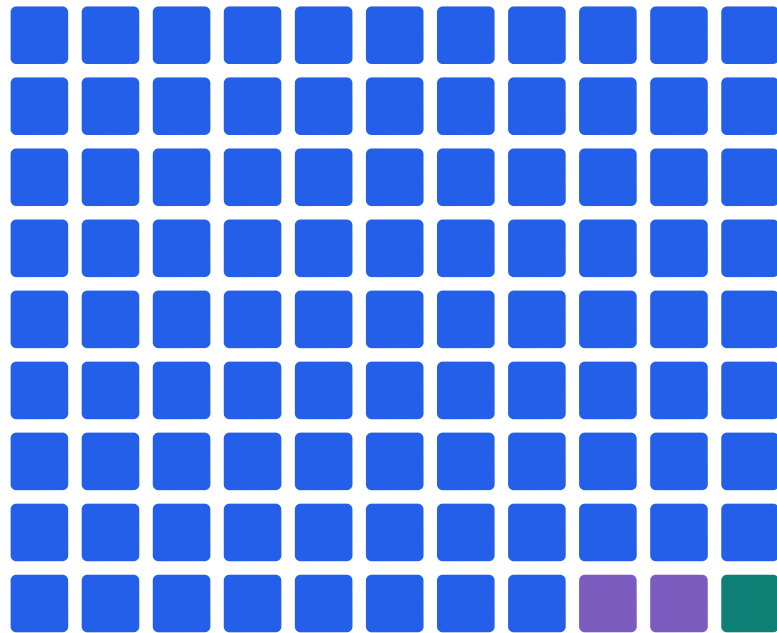
Audit log

which model, on what evidence

Selection evidence: per-model historical scores (paper Table 1); protocols differ by model. MASE as in [7]. The sparse fallback never triggered (0 of 111).

Recorded initial assignments

99 evaluated targets one square each · recorded assignment



 **DeepAR 96**
all 32 states × 3 strata

 **Ensemble 2**
national, general and women

 **Prophet 1**
national, men

State targets assigned to DeepAR

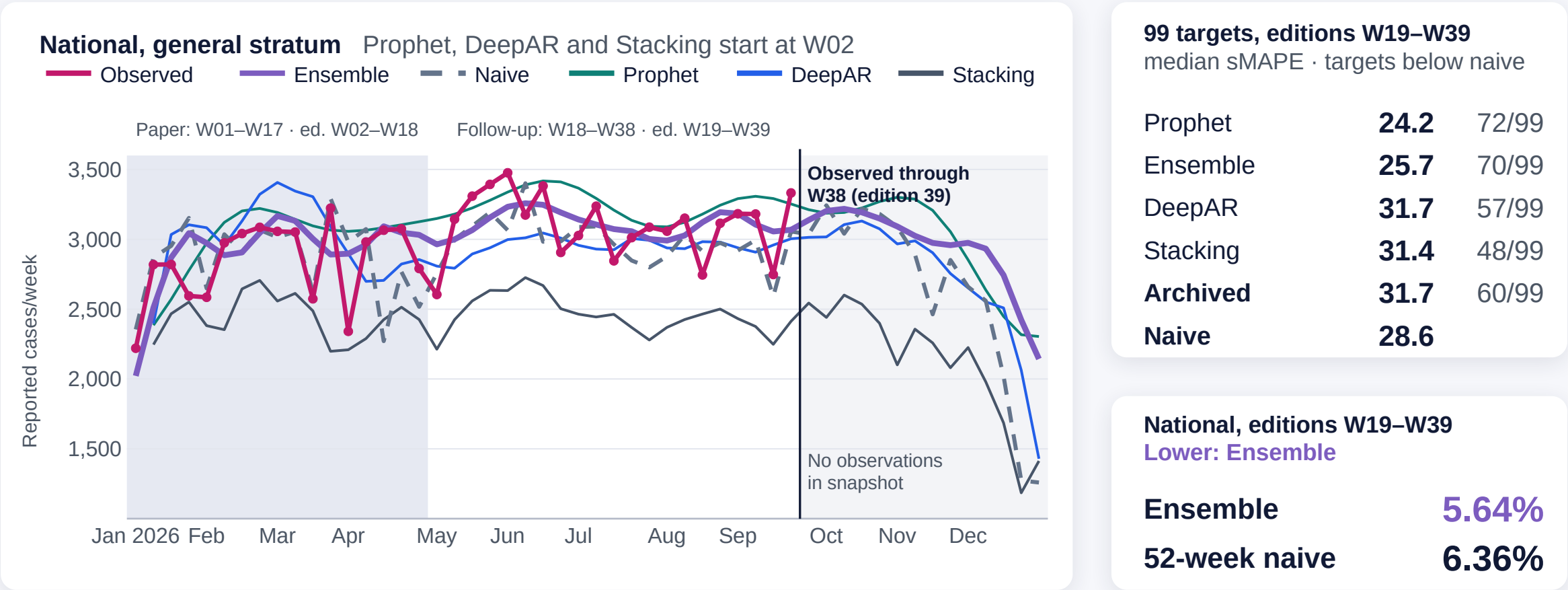
96 / 96

chosen from historical selection scores

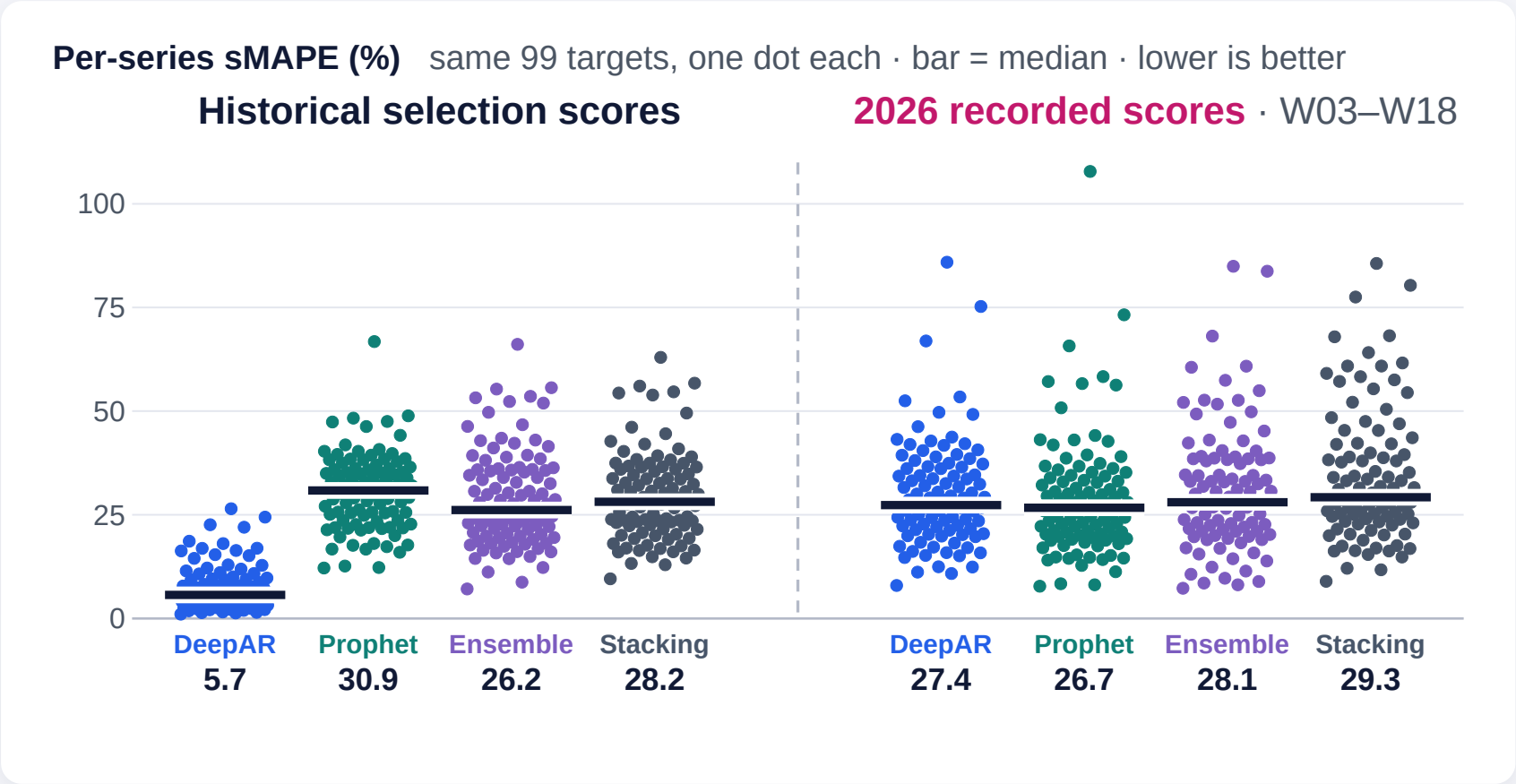
Note: Historical evaluation protocols differ. The state DeepAR code scores training fit.

Selection rule applied to the historical scores of paper Table 1 (107 of 111 targets, regions included, to DeepAR). Historical scores in backup B8.

Recorded 2026 trajectories and follow-up



Historical and 2026 model rankings differ



Historical protocols differ.
Descriptive comparison
across the same 99 targets.

Range of 2026 medians
< 3 pts
Lowest median: Prophet

No model is best everywhere

35	27	19	18
----	----	----	----

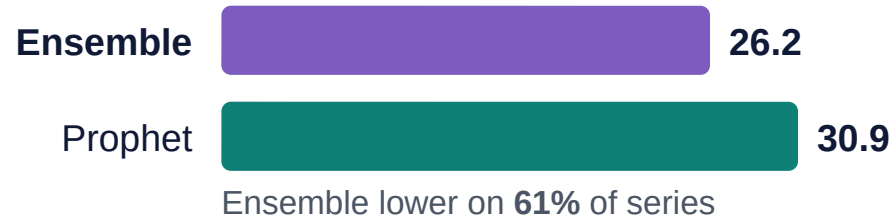
targets won in 2026: Prophet,
DeepAR, Ensemble, Stacking

Right panel and cards: recorded errors, bulletin editions W03–W18 (paper window). Post hoc and exploratory: targets overlap; tests in B2. Full figure in B9.

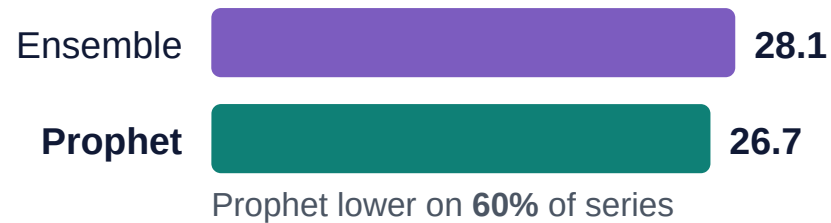
Excluding DeepAR, the two rankings differ

Median sMAPE (%) Ensemble vs. Prophet · same 99 series · lower is better

Historical held-out windows



2026 recorded scores W03–W18



Same lowest-error model, historical vs. 2026

Prophet, Ensemble, Stacking · 99 series

34% **observed**
34 of 99 series

31% **reference**
from each model's win frequencies

Descriptive reference, not a formal chance test.

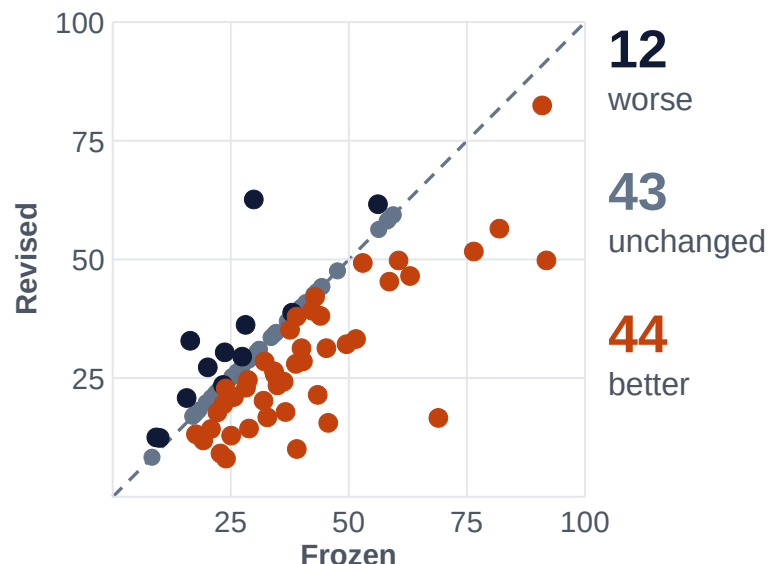
Additional analysis beyond the paper. Point rankings; post hoc and exploratory. Averaging each state's three targets weakens the reversal (backup B2).

Revision lowered the median in a seven-week replay

W03–W11 re-select: same rule, prior weeks only

W12–W18 evaluation: excluded from reselection

Each dot is one series SMAPE, W12–W18



Policies median SMAPE, W12–W18

Frozen assignment 32.7

Global re-selection 32.7

Per-series revision 28.5

Comparators

Prophet only 27.6

Pool without DeepAR 26.4

–4.2 pts difference of policy medians

Reading

In this seven-week replay, revision reduced the median from 32.7% to 28.5%.

Simple alternatives had lower medians.

Revised vs. frozen: 44 better, 12 worse, 43 equal. Overall medians were lower at all nine splits (W06–W14); at W11 they were lower in each stratum. Exploratory.

TAKEAWAYS

Selection is a signal, not a guarantee

An auditable per-series selection framework, evaluated on recorded 2026 forecasts.

1

Historical and 2026 rankings differ

Historical selection favored DeepAR. Prophet had the lowest median error in the recorded panel.

2

Revision lowered the replay median

Seven-week replay: revision 32.7 → 28.5; pool without DeepAR 26.4.

3

The year so far

21-week follow-up: national Ensemble 5.64 vs naive 6.36; archived panel median 31.7 vs 28.6.

THANK YOU

Questions?

Selection is a signal, not a guarantee.

Forecasting Weekly Depression Incidence in Mexico: A Multi-Model Framework with Auditable Per-Series Model Selection

To appear in Advances in Soft Computing (MICAI 2026), Springer LNCS 17149



rebull@exatec.tec.mx



[github.com/EpiForecastMX/
per-series-model-selection2026](https://github.com/EpiForecastMX/per-series-model-selection2026)



epiforecast.mx/micai2026
follow-up, data and slides

References

- [1] Dirección General de Epidemiología, Secretaría de Salud: SINAVE Boletín Epidemiológico (2026).
- [2] Taylor, S.J., Letham, B.: Forecasting at scale. *Am. Stat.* 72(1), 37–45 (2018).
- [3] Salinas, D., Flunkert, V., Gasthaus, J., Januschowski, T.: DeepAR: probabilistic forecasting with autoregressive recurrent networks. *Int. J. Forecast.* 36(3), 1181–1191 (2020).
- [4] Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: *KDD*, pp. 785–794 (2016).
- [5] Ke, G., et al.: LightGBM: a highly efficient gradient boosting decision tree. In: *NeurIPS 30*, pp. 3146–3154 (2017).
- [6] Wolpert, D.H.: Stacked generalization. *Neural Netw.* 5(2), 241–259 (1992).
- [7] Hyndman, R.J., Koehler, A.B.: Another look at measures of forecast accuracy. *Int. J. Forecast.* 22(4), 679–688 (2006).
- [8] Nie, Y., Nguyen, N.H., Sinthong, P., Kalagnanam, J.: A time series is worth 64 words: long-term forecasting with transformers. In: *ICLR* (2023).
- [9] Box, G.E.P., Jenkins, G.M., Reinsel, G.C., Ljung, G.M.: *Time Series Analysis: Forecasting and Control*, 5th edn. Wiley (2015).

This work. Rebull-Saucedo, J.A., Pérez-Nava, J.C., Sánchez-Salazar, L.G., Barceló-Alonso, G., Pérez-Hernández, R.: Forecasting weekly depression incidence in Mexico: a multi-model framework with auditable per-series model selection. In: *Advances in Soft Computing, MICAI 2026, LNCS 17149*, Springer (to appear).

The selection rule, exactly

Per series, with recorded historical sMAPE, MASE, and RMSE per model:

- 1 If the second-lowest sMAPE exceeds 1.05 times the lowest, choose the lowest.
 - 2 Otherwise, among models with $\text{sMAPE} \leq 1.05 \times \text{best}$, pick the lowest MASE.
 - 3 If MASE is tied, pick the lowest RMSE.
- + Fewer than 5 cases in the 52 weeks before the forecast origin: use the regional model (past data only; never triggered here, 0 of 111).

Each choice and its metrics are stored with the forecast: the audit log.

Selection evidence

Per-model historical backtests (paper Table 1), not one shared protocol.

Prophet, Ensemble, Stacking: held-out windows, with different protocols.

Historical state DeepAR code skips CV, trains on the full series, then scores that fit (commit 60d6ea0f; changed in 9fdd6dca). The exact published runs are not independently linked.

The paper reports a data cutoff of 2025-W53; execution dates are not independently verified.

Statistical tests and uncertainty

Comparison	99 targets: raw / Holm	32 states: nominal
Historical, Ensemble vs. Prophet	0.020 / 0.060	0.15
2026, Prophet vs. Ensemble	0.007 / 0.035	0.06
2026, Prophet vs. DeepAR	0.070 / 0.14	0.24
Revision vs. frozen, split W11	< 0.001 / < 0.001	< 0.001

Exploratory, post hoc; targets overlap and Holm does not remove dependence. Not significant $\neq$ equivalent. Two-sided Wilcoxon, zsplit; Holm within six model pairs (nine for revision).

National, W03–W18, recomputed

Mean Poisson deviance: Ensemble 25.7, Prophet 32.3, DeepAR 46.0, Stacking 97.8.

Reported in the paper only

Diebold–Mariano tests; the 80% band covered 10 of 17 weeks. Inputs not released.

Limitations

One condition

Depression (F32) only; transfer is untested.

Training-fit evidence

Historical state DeepAR code uses training fit; run provenance is unverified.

Unequal search budget

Prophet had the broadest search; tuning effect untested.

Partial-year evaluation

Paper windows plus 21 follow-up periods (to W38); full year still pending.

Provenance

Forecast execution dates not independently verified.

Uncertainty

Paper-reported 80% band coverage: 58.8%; revision does not calibrate it.

Reported incidence

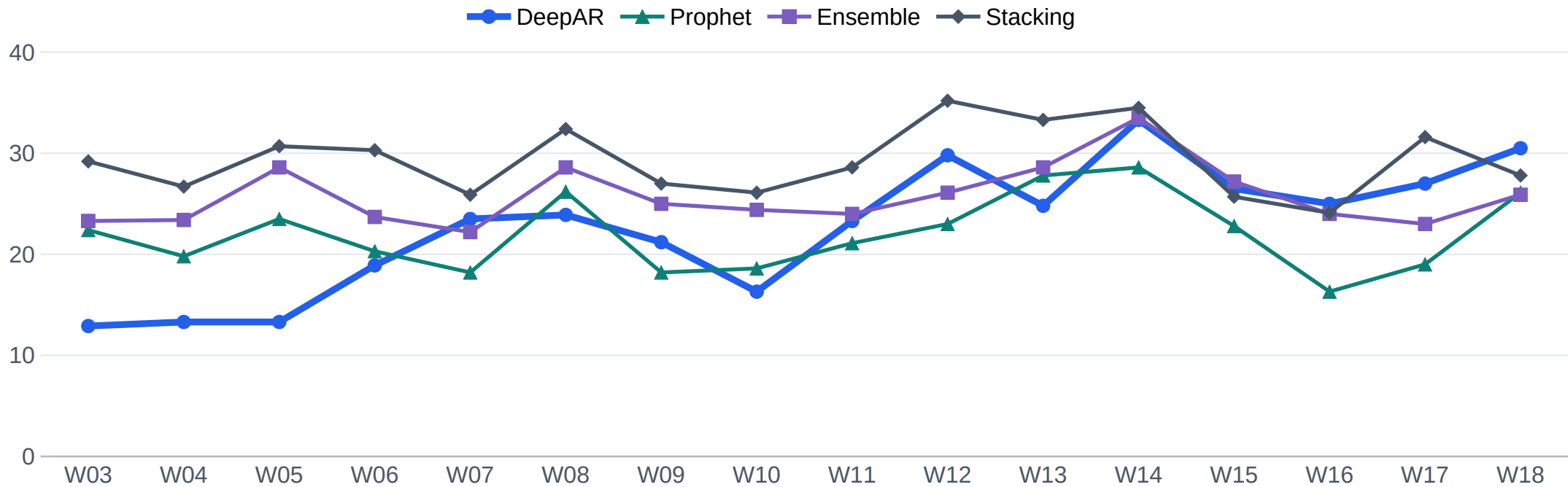
Models learn reporting and disease together.

No external validation

Only SINAVE / Mexico; a research prototype.

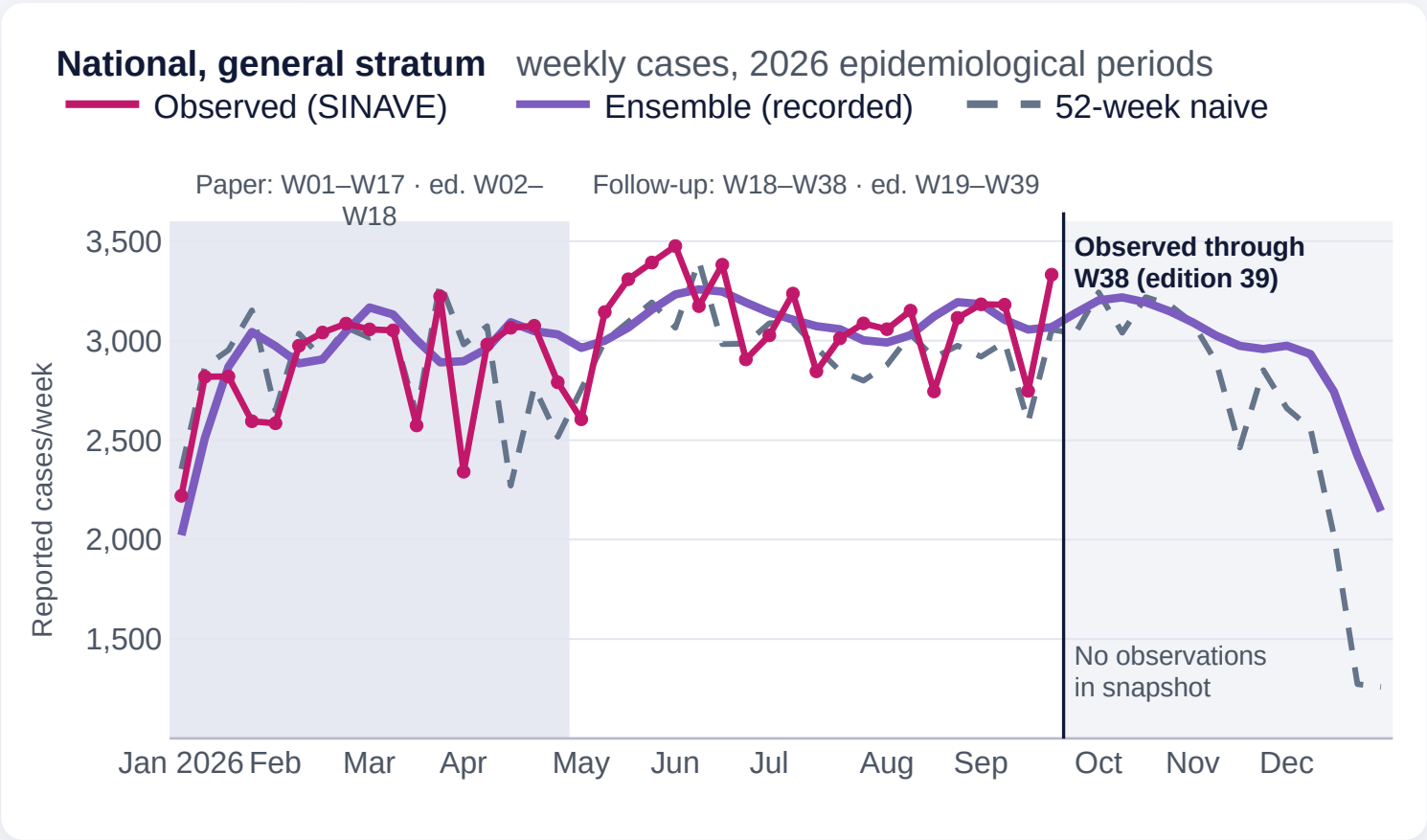
DeepAR leads the first four editions (W03–W06)

Median per-week sMAPE (%) across the 99 series 2026 bulletin editions · lower is better



Revision chooses among frozen 52-week paths; horizon effects may contribute to its gain (not isolated). Recomputed from the public files.

National forecast versus the 52-week naive



Paper	sMAPE (%)
editions W02–W18 · Lower: naive	
Ensemble	7.40%
52-week naive	7.33%

Follow-up	sMAPE (%)
editions W19–W39 · Lower: Ensemble	
Ensemble	5.64%
52-week naive	6.36%

Forecast periods mapped by rule; execution origin unverified.

sMAPE, lower is better. Cumulative deviation, Ensemble / naive: paper +2.45% / +0.6%; follow-up +0.21% / –3.81%. Snapshot: edition 39. Descriptive.

Every model against the naive, by window and sex

Median sMAPE (%) and targets beating the naive, of 33 per stratum · lower is better

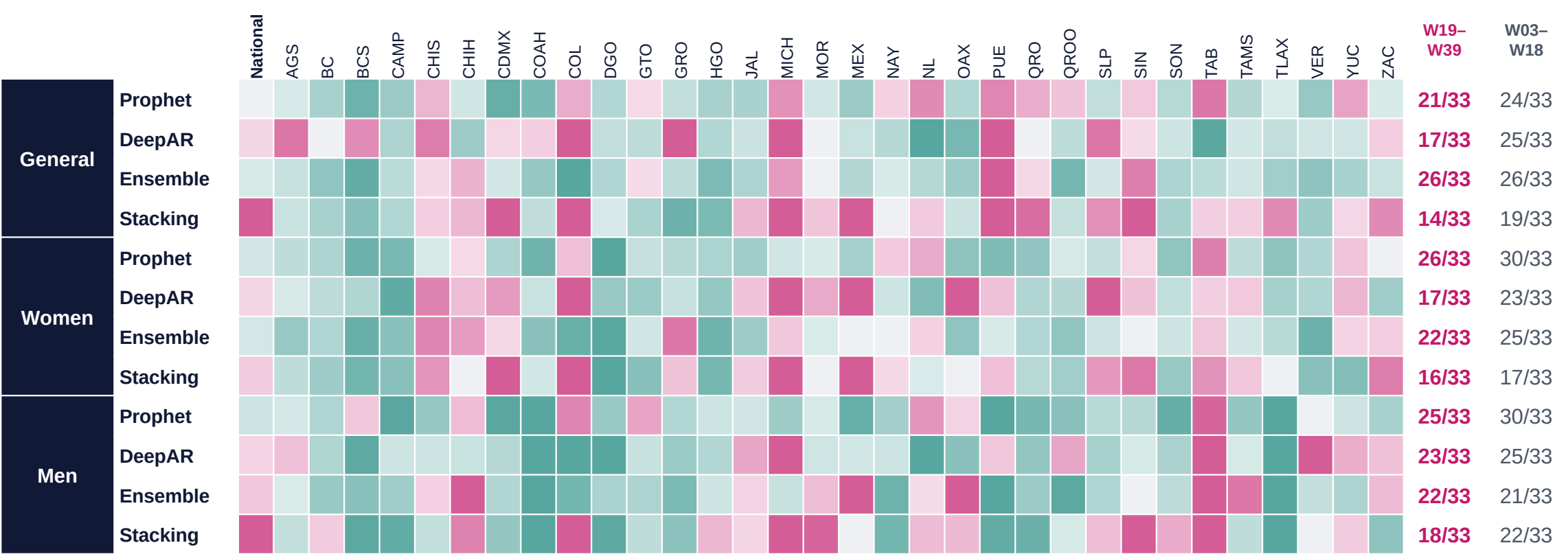
	Paper window · editions W03–W18			Follow-up · editions W19–W39		
Estimator	General	Women	Men	General	Women	Men
Prophet	23.8 · 24	24.4 · 30	30.4 · 30	21.3 · 21	22.7 · 26	30.0 · 25
DeepAR	25.9 · 25	26.0 · 23	32.4 · 25	23.3 · 17	27.1 · 17	35.2 · 23
Ensemble	26.7 · 26	23.8 · 25	38.0 · 21	21.8 · 26	25.7 · 22	36.3 · 22
Stacking	29.0 · 19	27.9 · 17	34.3 · 22	27.8+ · 14	30.6+ · 16	39.7+ · 18
Archived assignment	25.9 · 26	26.0 · 24	32.4 · 25	23.3 · 18	27.1 · 18	35.2 · 24
52-week naive	31.2	30.6	40.1	23.7	27.8	39.0

Cell: median sMAPE across 33 targets · targets with lower sMAPE than the naive. + = median above the naive; bold = lowest median. Archived assignment = DeepAR in every state. Descriptive; overlapping targets.

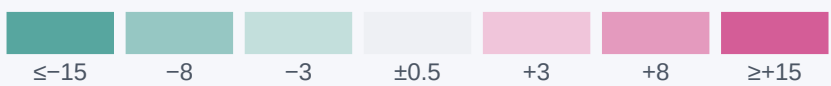
How the year is going: every model against the naive

Follow-up, editions W19–W39 · cell = model sMAPE minus naive sMAPE, one target each

Targets beating the naive



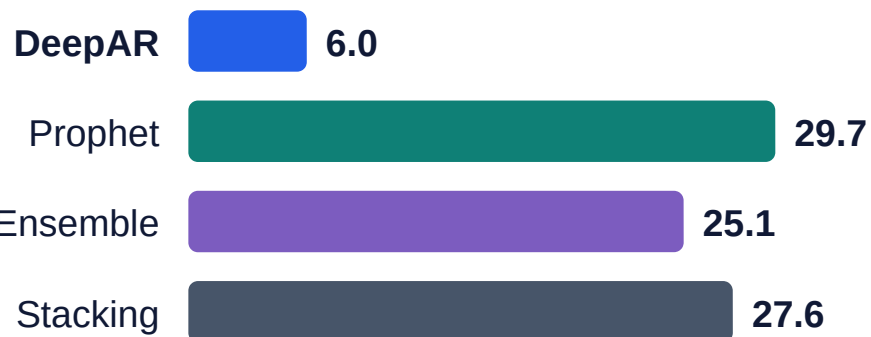
Model minus naive (pts):



Teal: lower error than the naive (better). Magenta: higher. Gray: under 0.5 pts, not necessarily a tie. Archived assignment = DeepAR in every state. Descriptive; overlapping targets. Full matrix for download; state-level reading requires zoom.

Historical selection scores (paper Table 1)

Median historical sMAPE (%) Paper Table 1 · 111 targets · lower is better



Baselines, not in the pool



Series the rule assigns to DeepAR

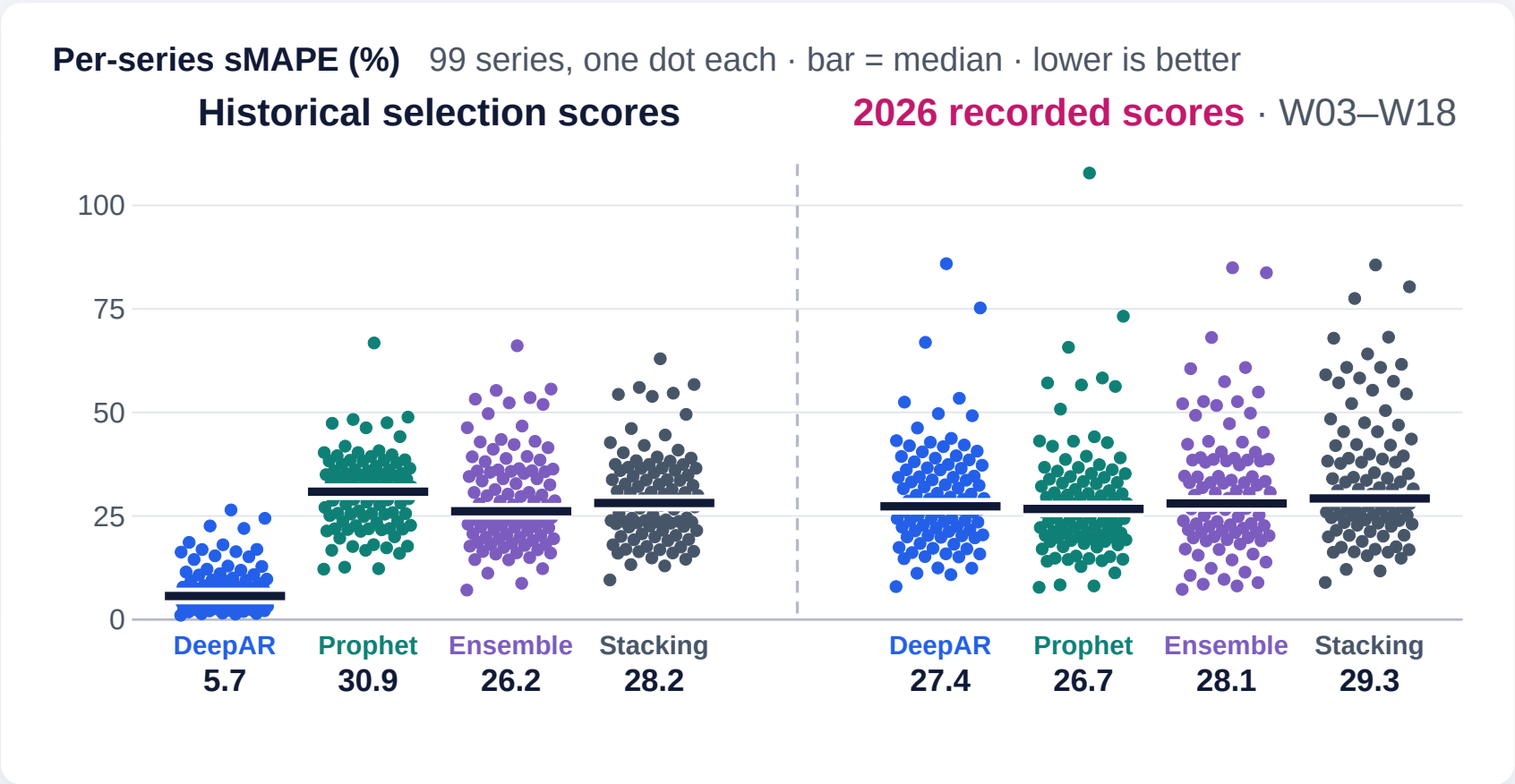
107 / 111

96.4%, including every state series

Reading the 6.0: in the historical code, DeepAR's state scores reflect training fit, not forecast error.

DeepAR's 6.0% historical sMAPE is the median over all 111 targets. Historical selection scores are not a guarantee of 2026 accuracy.

Historical and 2026 model rankings differ



Historical score gap
≈20 pts

Range of 2026 medians
< 3 pts

No model is best everywhere

35	27	19	18
----	----	----	----

series won in 2026: Prophet, DeepAR, Ensemble, Stacking

Historical protocols differ by model. Post hoc and exploratory: targets overlap; paired tests in backup B2.