Supplementary Material

Probabilistic dengue forecasting and early warning across Mexico (1985–2026): a machine-learning and statistical ensemble with conformal calibration

Jaime Briseno-Ramirez, Pedro Martínez-Ayala, Patricia Noemi Vargas-Becerra, Mauricio Alfredo Ambriz-Alarcón, Oscar Francisco Fernández-Díaz, Roberto Miguel Damián-Negrete, Nancy Evelyn Navarro-Ruiz, Ruth Rodríguez-Montaño, Ana María López-Yáñez, Judith Carolina De Arcos-Jiménez

Supplementary Methods

S1. Open data sources, access points and freeze

All inputs were obtained from open national and international repositories and frozen on 31 May 2026; custody was enforced by content hashing, with each run re-verifying the recorded SHA-256 content (or file-stat) hash of every source and aborting on any mismatch. The climate downloaders additionally verified a SHA-256 checksum of each file immediately after download, so a truncated transfer aborted rather than silently corrupting the panel. Table S1 lists each source with its public address.

Climatic covariates were obtained from TerraClimate, a high-resolution (∼4 km) monthly reanalysis, supplemented for recent years by NASA POWER for the 16 endemic states with complete coverage; aggregation to the state–month used population-weighted zonal statistics over the 2020 INEGI *Marco Geoestadístico*. At the TerraClimate-to-NASA-POWER splice (TerraClimate through 2023, NASA POWER from 2024 onward), monthly mean 2-m air temperature and total precipitation were continuous, whereas the daily-range endpoints showed a source-induced discontinuity: the NASA maximum 2-m temperature was ∼0.7 °C cooler and the minimum ∼1.3 °C warmer than TerraClimate (a narrower diurnal range), documented rather than adjusted (see the Discussion).

S2. Outcome harmonisation across case-classification eras

The 2017–2024 count was reconstructed from the Anuarios as the sum of three mutually exclusive ICD-10 categories (A97.0/A97.9 dengue without warning signs, A97.1 dengue with warning signs, A97.2 severe dengue), restoring comparability with the pre-2016 A90/A91 era in which warning-sign cases were not coded separately. The 2025–2026 outcome was estimated from the open line-lists using the Anuario estimated-case formula (laboratory-confirmed cases plus the test-positivity fraction applied to untested probable cases, by state, year and month) and rescaled to the Anuario scale by a factor of 1.25 derived from the 2024 overlap year. National all-forms totals implied by this harmonisation were 157,018 estimated cases in 2019, 156,218 in 2023 and 329,381 in 2024. The resulting panel is rectangular: all 32 states are present

Table S1 Open data sources used to assemble the state–month panel.

Source	Variable(s)	Public access point
DGE *Anuarios de Morbilidad*	Notified dengue cases, 1985–2024 (all-forms)	https://epidemiologia.salud.gob.mx/anuario/html/anuarios.html
SINAVE/DGE open data	Case-level records, 2020–2026; serotype confirmations	https://www.gob.mx/salud/documentos/datos-abiertos-152127
DGE InDRE *Panoramas Epidemiológicos*	Serotype proportions, 2014–2019	https://www.gob.mx/salud/acciones-y-programas/direccion-general-de-epidemiologia-boletin-epidemiologico
PAHO PLISA	Serotype proportions, 1995–2013 (regional)	https://www3.paho.org/data/index.php/en/
TerraClimate	Monthly climate, 1958–2023 (∼4 km)	https://www.climatologylab.org/terraclimate.html
NASA POWER	Monthly climate, recent years, endemic states	https://power.larc.nasa.gov/
NOAA CPC ONI	El Niño–Southern Oscillation index	https://origin.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ONI_v5.php
INEGI *Marco Geoestadístico* 2020	State polygons and adjacency graph	https://www.inegi.org.mx/temas/mg/
CONAPO projections	Population denominators, 1950–2070	https://www.gob.mx/conapo

in every month from January 1985 through May 2026 (15,904 state–month cells), reconstructed from the primary sources with no imputation and no absent cells.

The provenance of each state–month cell (source system, era and reconstruction rule) was retained as a per-row attribute.

S3. PyMC hierarchical negative-binomial model

For each rolling origin, the training cells with outcome y_i, offset $o_i = \log(\text{population}_i)$ and standardised feature vector $\mathbf{x}_i$ for state $s(i)$ were modelled as

$$y_i \sim \text{NegBin}(\mu_i,\ \alpha), \qquad \log \mu_i = \beta_0 + \mathbf{x}_i^{\top}\boldsymbol{\beta} + u_{s(i)} + o_i,$$

with priors

$$\beta_0 \sim \mathcal{N}(0, 2^2), \quad \beta_j \sim \mathcal{N}(0, 1^2), \quad u_s \sim \mathcal{N}(0, \sigma_u^2), \quad \sigma_u \sim \text{HalfNormal}(1), \quad \alpha \sim \text{Gamma}(2, 0.1),$$

where u_s is a state-level random intercept (the hierarchical component, borrowing strength across the 32 states) and α is the negative-binomial dispersion. Standardising the features makes the unit-scale coefficient prior weakly informative. Posterior sampling used the No-U-Turn Hamiltonian Monte Carlo sampler with eight chains run in parallel (1,000 warmup and 1,000 retained draws per chain in the production

configuration; a reduced-draw configuration was used only for smoke tests), a target acceptance probability of 0.9, and a fixed random seed. Convergence was assessed by the potential-scale-reduction statistic ($\hat{R}$) and the effective sample size; predictive intervals were obtained by drawing from the negative-binomial posterior predictive distribution rather than from the posterior mean alone. To guard against numerical overflow during posterior-predictive sampling, the linear predictor on the log scale was clipped to the interval $[-20, 10]$ (an upper bound of $\exp(10) \approx 2.2 \times 10^4$ cases per cell, consistent with the INLA implementation); prediction cells belonging to a state unseen in the training window were returned as missing rather than assigned an arbitrary random effect.

S4. Level-normalised conformity measure

Let $\hat{y}_{q50}$ denote the ensemble median forecast and y the realised count. The conformity measure was

$$s \;=\; \frac{|y - \hat{y}_{q50}|}{\max(\hat{y}_{q50},\, 0) + c}, \qquad c = 3 \text{ cases},$$

a level-normalised (multiplicative) score. The constant c was fixed *a priori* as a minimum-reportable-count floor and was never tuned on coverage. This choice was motivated by the error structure of the ensemble: the median relative error $|y - \hat{y}_{q50}|/\hat{y}_{q50}$ is approximately 0.9 and roughly constant across two orders of magnitude of predicted incidence, whereas a Poisson-style $\sqrt{\mu}$ normalisation rises from 0.6 to 15.9 across the same range and is therefore inappropriate. The recalibrated interval at coverage $1 - \alpha_0$ (with $\alpha_0 = 0.10$) is $[\hat{y}_{q50} - \hat{q}\,(\hat{y}_{q50} + c),\ \hat{y}_{q50} + \hat{q}\,(\hat{y}_{q50} + c)]$, clipped at zero from below, where $\hat{q}$ is the conformal quantile with the finite-sample correction $\lceil (n+1)(1-\alpha_0) \rceil / n$.

S5. Online adaptive update and regime-switching threshold

Adaptive conformal inference. For each horizon the working miscoverage level α_t was updated along the stream of forecasts ordered by origin month:

$$\alpha_{t+1} \;=\; \mathrm{clip}\big(\alpha_t + \gamma\,(\alpha_0 - \mathrm{err}_t),\ 0,\ 0.20\big),$$

where err_t is the realised miscoverage indicator at step t, γ is the step size, and $\alpha_0 = 0.10$ is the long-run target. The conformal quantile at each step was recomputed from the conformity scores observed in the preceding W months only (strictly earlier than the current month, so no future information entered). The clip to $[0, 0.20]$ bounds the working level so the interval can neither collapse below the calibration data nor exceed twice the nominal miscoverage. The step size and window were tuned per horizon on the pandemic and early-resurgence stream up to 2023, choosing the pair bringing long-run coverage closest to 0.90, over the grids

$$\gamma \in \{0.01,\, 0.03,\, 0.05,\, 0.10\}, \qquad W \in \{18,\, 24\} \text{ months};$$

the 2024 epidemic year never entered the tuning. The default $(\gamma, W) = (0.03, 24)$ was used for any horizon with insufficient tuning data.

Regime-switching threshold. Each state–month was labelled high-transmission or baseline according to whether the DENV-3 proportion lagged six months exceeded $\tau = 0.40$ (or a serotype switch had occurred within the previous twelve months), evaluated at the origin month to avoid leakage. A regime-specific conformal quantile was applied: the baseline quantile was calibrated on the pre-pandemic block; the high-transmission quantile was calibrated only on high-incidence 2022–2023 cells (never on 2024) and stabilised by shrinkage toward the baseline quantile,

$$\hat{q}_{\mathrm{HIGH}} \;=\; \hat{q}_{\mathrm{BASE}} + \lambda\,(\hat{q}^{\mathrm{raw}}_{\mathrm{HIGH}} - \hat{q}_{\mathrm{BASE}}), \qquad \lambda = \frac{n_{\mathrm{HIGH}}}{n_{\mathrm{HIGH}} + \kappa}, \quad \kappa = 200,$$

a James–Stein-type rule in which the shrinkage weight λ grows with the number of high-regime calibration cells n_{HIGH}; the raw high-regime quantile was used only when at least ten such cells were available. The deployed scheme applied this regime-specific quantile over the adaptive online update.

Deployment selection. The deployed regime-switching adaptive conformal layer, labelled `cptc_over_aci` in the archived code, was chosen by an *operational* criterion — prediction intervals that are simultaneously calibrated and sharp (median width 62 versus ∼162 cases for the additive baseline) — a choice made independently of the 2024 epidemic year. For transparency an automatic selection rule is also reported for all seven candidates (an additive baseline, the level-normalised static scheme, conformalised quantile regression, two adaptive variants and two regime-switching variants); this rule is itself blind to 2024, being evaluated on the DENV-3 resurgence stream through 2023 only, and it favours wider-band methods of the conformalised-quantile-regression type that maximise coverage at the cost of interval width. Because neither the hyperparameter tuning nor the choice of the deployed scheme used the 2024 epidemic year, the recovered 2024-peak coverage is a genuine out-of-sample result.

S6. Panel integrity and feature availability

The all-forms harmonisation pooled, under the pre-2009 World Health Organization framework, classical dengue fever and dengue haemorrhagic fever including dengue shock syndrome (1997 guidelines [50]; classical dengue notified from 1985 and haemorrhagic dengue from 1991), and, under the 2009 revision [51] adopted for Mexican surveillance through NOM-017-SSA2-2012 [52], dengue without warning signs, dengue with warning signs and severe dengue. The harmonised serotype panel covered 1995–2026 at the national level with five strata of source quality; two years (2011 and 2013) were retained as missing-not-imputed and absorbed by the year-level random-walk prior of the Bayesian models. Of the 15 features flagged as forbidden look-ahead and excluded from every forecast model, 14 were retrospective whole-series cycle features (reserved for descriptive figures) and one was the notification-completeness attribute, whose target-month value is unknown at the forecast origin; the 23 national–annual serotype and emergence proxies carried a conservative 360-day availability lag. Monthly notification completeness was 1.00 in every month of the reconstructed panel but was nonetheless retained as a per-row attribute so this property could be audited at the row level.

Autoregressive predictors used the outcome lagged by 1, 2, 3, 6 and 12 months on the log(1 + cases) scale; climatic predictors used lags of 1–6 months, the ENSO index lags of 1–9 months, and serotype proportions lags of 6–12 months to respect the laboratory typing delay. Population-immunity and emergence proxies (cumulative incidence, serotype diversity, time-since-dominance and emergence-window indicators) were national–annual quantities conservatively lagged by 360 days and informed only the exploratory serotype ablations, not the operational ensemble. Monthly notification completeness was retained as a per-row attribute but excluded from every forecast model because the target-month value is unknown at forecast time.

S7. Cyclical feature construction

From the national monthly series the slow secular trend was extracted by seasonal-trend decomposition (STL) and the multiannual signal isolated with a band-pass filter retaining oscillations with periods of five to eight years; phase and amplitude were obtained by the Hilbert transform and entered as sine and cosine terms. A complementary spectral analysis (Fourier periodogram and wavelet transform [61], evaluated against a red-noise null) placed the multiannual power in a broad band around 3.5 and 5.6 years. The operational band-pass (five to eight years) emphasises the lower-frequency component of this range, and we additionally generated narrow-band phase features at three representative periods spanning it (4.3, 5.0 and 5.79 years); at the ∼0.6-year frequency resolution of the 1985–2026 record these periods are not individually resolved.

Narrow-band phase variants from the complementary spectral analysis were entered as additional sine and cosine terms; these retrospective whole-series variants were reserved for descriptive figures only.

S8. Forecast-model and baseline specifications

The four parsimonious baselines were a seasonal-naïve forecast, a historic monthly mean, an automatically selected seasonal autoregressive integrated moving-average (SARIMA) model fit per state on log(1 + cases), and a cyclic baseline adding a linear adjustment for the contemporaneous cycle phase. The PyMC hierarchical negative-binomial model (S3) used a Gaussian prior on the regression coefficients and on the intercept, a state-level random intercept drawn from a Gaussian whose scale carried a half-normal prior, and a gamma prior on the dispersion; posterior sampling used the No-U-Turn Hamiltonian Monte Carlo sampler with eight chains, and convergence was checked by the potential-scale-reduction statistic and the effective sample size. For the INLA BYM2 model, the BYM2 spatial prior is a reparameterisation that lets a spatially structured (neighbour-correlated) and an unstructured (independent) random effect share a single interpretable variance; the adjacency graph was the real first-order graph of the 32 states (states sharing a border are neighbours; Baja California Sur was linked to Baja California). Sampling the full negative-binomial posterior predictive distribution rather than the posterior of the mean alone raised the empirical coverage of the nominal 90% interval from ≈ 0.22 to 0.89–0.95 at short horizons.

The GLM-NB elastic-net penalty was re-selected by temporal cross-validation at each origin, with state-specific intercepts and a calendar-month effect; the GAM-DLNM followed the distributed-lag non-linear cross-basis convention; and the XGBoost family added six quantile-regression boosters around a Poisson median head under Bayesian hyperparameter search, giving a seven-quantile predictive distribution spanning q05–q95. The tempered ensemble weighting used temperature $\eta = 3$, so each family's weight at a given state and horizon increased with validation skill, equivalently decreasing as its validation continuous ranked probability score increased, and a hard filter excluding any family whose validation score exceeded five times the best family's score at that cell down-weighted families that degraded at a horizon (notably INLA at $h = 12$); the resulting mean weights were dominated by XGBoost ($\sim$0.60) and GAM-DLNM ($\sim$0.15). Because these per-cell weights vary by predictive quantile, the per-quantile weighted median can in principle induce quantile crossing; the aggregated quantiles were therefore restored to monotonicity by row-wise rearrangement ($q_{05} \leq q_{50} \leq q_{95}$) before the online conformal recalibration, so no quantile crossing remains in the released forecasts. The INLA BYM2 family, which had exhibited severe long-horizon error inflation under a fixed-split configuration, was substantially improved by the homogeneous rolling-origin re-estimation together with sampling from the full negative-binomial predictive distribution (block-C median absolute error at $h = 6$ falling from 1,274 to 276 cases), although a residual inflation persisted at $h = 12$ (1,021 cases; Supplementary Table S4); recalibrated in this way it was admitted as a Bayesian ensemble contributor, competitive at short and medium horizons and automatically down-weighted at $h = 12$. The pooled GLM-NB was, like the other count-regression families, far outperformed by the tree-based learner in the DENV-3 resurgence block (block-C median absolute error 238–283 cases across horizons versus 64–92 for XGBoost).

S9. Homogeneous rolling-origin design and additional metrics

The validation design was refined over the course of the work. An initial implementation re-estimated the parsimonious and statistical families sequentially but, for tractability, trained the computationally intensive learners only once on a fixed temporal split; because this asymmetry favoured the heavier learners and distorted their apparent long-horizon performance, the present analysis adopts a single *homogeneous* design applied identically to all nine approaches, with the XGBoost quantile heads produced without any early stopping or tuning choice that could see future data. Each prediction carried its feature month, last training month and validation design; the per-row leakage audit covered all 315,410 forecast records, and the per-family predictions were consolidated into a single canonical table used as the reference for all downstream metrics and figures. Beyond the headline metrics, deterministic accuracy was summarised by the mean and root-mean-square absolute error and by the mean and symmetric mean absolute percentage error, and probabilistic skill additionally by the Winkler interval score; the Diebold–Mariano test used a heteroskedasticity- and autocorrelation-consistent variance correction. The outbreak-alarm classifiers additionally reported the area under the precision–recall curve, sensitivity, specificity and the false-alarm rate, and calibration by reliability diagrams.

The outbreak-alarm classifier was trained with anti-overfitting constraints; the cycle-augmented specification's collapse toward random discrimination at $h \geq 3$ (AUROC 0.564 at $h = 3$ and 0.479 at $h = 6$, against 0.832 and 0.728 for the parsimonious base) reflects the mismatch between cyclical encodings optimised for the multiannual scale and a binary monthly outbreak target, so cycle features are excluded from the operational EWARS specification.

S10. Software environment, custody and reproducibility

The principal Python packages were polars and pandas for data handling, scikit-learn, statsmodels, XGBoost and LightGBM for the regression and machine-learning models, PyMC for the Bayesian hierarchical model and Optuna for hyperparameter optimisation; in R, R-INLA provided the BYM2 spatio-temporal model and mgcv with dlnm the generalised additive distributed-lag model. Architectural invariants were enforced by a unit-test suite (80 tests), including checks against accidental row-multiplying joins, writes outside the designated output directories, the use of any non-causal cyclical feature in a modelling script, and per-row provenance and feature-availability gates. Custody was enforced by content hashing: raw inputs were frozen on 2026-05-31 against a baseline recording the content hash of every source, re-verified each run, and the climate downloaders additionally checksummed each file and wrote a manifest so a truncated transfer halted with an error rather than corrupting the panel. Final artifacts and the principal external sources carried verified live hashes, the remaining base sources being pinned to the baseline; no script wrote outside its output directory (statically enforced by an automated scanner), and per-row provenance and a claim-to-artifact trace tied every result to its producing artifact and hash. Runtime was dominated by the Bayesian families, PyMC sampling being rate-limiting, with per-step memory and timing archived.

S11. Per-family detail of the planned ablations and cross-family comparisons

In the GLM-NB family the band-limited multiannual cycle features yielded no consistent improvement: the median change in absolute error was negative across all blocks and horizons, significant in at most 5 of 32 states ($h = 1$ in the DENV-3 resurgence block) and, in that block at the longer horizons, in no state at $h = 6$ and in only 2 of 32 at $h = 12$ (a rate consistent with chance under multiple testing). The serotype-plus-emergence specification did not deliver positive incremental skill within the GLM-NB in any block, suggesting that the present serotype features (national–annual resolution, lagged 6–12 months) do not enter additively at the state–month scale under the linear specification with the chosen penalty. In the XGBoost family the cycle features delivered a modest improvement confined to the longer horizons (a median reduction in absolute error of about 3–10% at $h = 6$ and $h = 12$), of comparable magnitude in the pandemic and DENV-3 resurgence blocks and significant in a minority of states (2–5 of 32; Supplementary Fig. S1 and Table S3), while degrading the shorter horizons; the features did not improve the EWARS classifier, and the national–annual serotype

and emergence features added no robust incremental skill in any block (Supplementary Fig. S2).

The cross-family Diebold–Mariano comparison across the four horizons (Supplementary Table S4) confirmed that XGBoost significantly outperformed the GLM-NB at every horizon (Holm-adjusted $p < 10^{-7}$ throughout) and the parsimonious baselines at $h = 1$ ($p < 10^{-9}$). The weighted-median ensemble was statistically indistinguishable from XGBoost at $h \geq 6$ (for example $p = 0.16$ at $h = 6$ and $p = 0.022$, not Holm-significant, at $h = 12$), as expected when a single member carries much of the weight. The recalibrated INLA model was now competitive at short and medium horizons (its $h = 1$ disadvantage relative to XGBoost being smaller than that of PyMC) but remained the weakest point-forecaster at $h = 12$ (paired median absolute error 811 vs 78 cases against XGBoost), where the weighting accordingly reduced its ensemble contribution.

A robustness check confirmed the cycle ablation's point skill was insensitive to the exact cyclical encoding: replacing the three reference periods with the data-driven dominant periods ($\sim$3.5 and $\sim$5.6 years) or a single broadband (5–8 year) phase yielded equivalent point skill (within about 0.5% in relative median absolute error), whereas a wider 3–7 year band degraded it (by roughly 4–7%); the chosen periods therefore acted as a flexible multiannual basis rather than physically resolved frequencies. For the spatial-pooling sensitivity, the five highest-burden states ($\approx$44% of national incidence) were tested under five alternative specifications (cohort-specific and state-specific cycle features, serotype–emergence features, independent per-state gradient-boosted models, and a five-state-only pool), all in a single-learner XGBoost configuration so the relative ordering across pooling levels could be isolated (absolute errors therefore not comparable to the five-family conformal ensemble); the full national pool edged every alternative even on overall median error, and the 2024-peak point under-prediction persisted across every specification, confirming it as a structural out-of-distribution limit rather than an artefact of pooling.

S12. Supplementary Discussion: positioning within the dengue-forecasting literature

The present system extends a maturing body of subnational, climate-informed and ensemble dengue-forecasting research while clarifying what is, and is not, new. Relative to a recent Mexican study that clustered all 32 states on eco-epidemiological similarity and issued cluster-wise one-week-ahead probabilistic bands [47], our system forecasts at monthly resolution and 1- to 12-month lead times for every individual state, combines five model families, and adds distribution-free online recalibration with regime-stratified coverage reporting. Where Johansson et al. found that climate features did not improve seasonal autoregressive forecasts for Mexico at coarser spatial resolution [23], our climate-driven GAM-DLNM became a leading ensemble contributor only after the gridded climate panel was extended from 16 to all 32 federal entities, consistent with — though we did not test head-to-head — the hypothesis that the spatial coverage and resolution of the climate product, rather than climate information per se, conditions the gain.

Internationally, the dominance of tree-based learners over count-regression baselines, with the gradient-boosted family receiving the largest mean ensemble weight and outperforming the generalised linear model at every horizon, is consistent with the global evaluation of Wu et al. [25] and the reproducible boosting pipeline of Sebastianelli et al. [24] for Brazil. The robustness of median aggregation across model families echoes the COVID-19 European hub analysis of Sherratt et al., in which a median ensemble outperformed most of its components on interval score [71], and the Vietnamese superensemble of Colón-González et al. [26]. In Latin America, a Colombian comparison found random forests more accurate than artificial neural networks for departmental dengue burden [36], and climate-based early-warning modelling in three endemic Peruvian departments correctly classified all future outbreaks above a defined incidence threshold with a low false-alarm rate [37]; these efforts share our emphasis on subnational, climate-informed prediction but differ in not providing conformal recalibration with explicit empirical coverage diagnostics or an explicit mechanism for regime shift. The 2024 Brazilian Dengue Forecasting Sprint is an instructive caution: across six teams and five states in the atypical 2024 season no single architecture dominated [70], qualifying our gradient-boosting-dominant weighting under future climate regimes. The modest, state-heterogeneous incremental skill of the band-limited cycle phase within the non-linear learner during the DENV-3 block is consistent with the cyclical pattern previously documented descriptively in Mexico [7] and with the Brazilian dynamic-ensemble framework [29], while confirming that this signal is complementary rather than dominant. Finally, the demonstration by Finch et al. that adding serotype information to a climate-only Bayesian model lifts skill in Singapore [32], together with the independent confirmation of a 2024 DENV-3 epidemic in western Mexico by Mendoza-Hernández et al. [10], retrospectively support the emergence indicators that flagged the 2024 outbreak window. The out-of-ensemble comparators benchmarked but not admitted to the operational five-family ensemble are reported in Supplementary Table S5.

S13. Reporting-guideline compliance

Compliance with EPIFORGE 2020 [48] and TRIPOD+AI 2024 [49] is documented item-by-item in Supplementary Tables S6 and S7, respectively.

Supplementary Tables

Table S2 Forecast skill by family × horizon × block (national median MAE; coverage of the nominal 90% interval in parentheses, pre-conformal).

Family	Model	h	A: pre-pandemic	B: pandemic	C: DENV-3
baseline	B0_seasonal_naive	1	—	109 (79%)	379 (70%)
baseline	B1_historic_mean	1	—	65 (85%)	267 (79%)
baseline	B2_sarima	1	—	22 (92%)	130 (90%)
baseline	B3_cyclic_baseline	1	—	64 (92%)	259 (80%)
glm_negbin	M_base	1	101 (69%)	87 (73%)	238 (75%)
glm_negbin	M_cycle	1	101 (69%)	87 (71%)	240 (75%)
glm_negbin	M_serotype	1	114 (70%)	86 (79%)	231 (79%)
xgboost	M_base	1	33 (98%)	24 (96%)	64 (98%)
xgboost	M_cycle	1	46 (95%)	27 (96%)	125 (94%)
xgboost	M_serotype	1	48 (96%)	30 (96%)	118 (95%)
baseline	B2_sarima	3	—	56 (88%)	241 (88%)
glm_negbin	M_base	3	99 (73%)	89 (79%)	283 (74%)
glm_negbin	M_cycle	3	106 (67%)	93 (58%)	294 (58%)
xgboost	M_base	3	33 (99%)	33 (100%)	72 (98%)
xgboost	M_cycle	3	50 (96%)	41 (96%)	122 (96%)
xgboost	M_serotype	3	47 (98%)	40 (96%)	100 (98%)
baseline	B2_sarima	6	—	76 (88%)	277 (83%)
glm_negbin	M_base	6	116 (71%)	89 (75%)	282 (70%)
xgboost	M_base	6	36 (98%)	32 (100%)	92 (100%)
xgboost	M_cycle	6	34 (100%)	32 (100%)	79 (98%)
baseline	B2_sarima	12	—	59 (85%)	288 (79%)
glm_negbin	M_base	12	123 (68%)	91 (77%)	267 (70%)
xgboost	M_base	12	35 (99%)	26 (100%)	70 (98%)
xgboost	M_cycle	12	38 (98%)	26 (100%)	68 (99%)

Median MAE across 32 states; coverage is the median across the 32 states of each state's fraction of forecast cells within the nominal 90% interval. Lower MAE and coverage closer to 0.90 indicate better forecasts. The complete per-family table is provided in the archived analysis outputs on Zenodo (DOI 10.5281/zenodo.20585066; `05_validation/metrics_consolidated_regression.csv`).

Table S3 Selected feature-set ablations: median ΔMAE% across 32 states and number of states with statistically significant improvement (Holm–Bonferroni at $\alpha = 0.05$).

Family	Ablation A → B	h	A: ΔMAE% (sig)	B: ΔMAE% (sig)	C: ΔMAE% (sig)
glm_negbin	M_base → M_cycle	1	−0.5% (1/32)	−0.0% (0/32)	−3.4% (5/32)
glm_negbin	M_base → M_cycle	3	−15.0% (1/32)	−5.4% (0/32)	−11.1% (0/32)
glm_negbin	M_base → M_cycle	6	−9.3% (0/32)	−16.8% (1/32)	−9.8% (0/32)
glm_negbin	M_base → M_cycle	12	−8.2% (1/32)	−16.9% (0/32)	−0.8% (2/32)
xgboost	M_base → M_cycle	1	−63.3% (0/32)	−38.5% (0/32)	−80.4% (0/32)
xgboost	M_base → M_cycle	3	−58.9% (0/32)	−39.9% (0/32)	−56.9% (0/32)
xgboost	M_base → M_cycle	6	+3.4% (5/32)	+9.9% (5/32)	+9.2% (2/32)
xgboost	M_base → M_cycle	12	−3.6% (2/32)	+9.7% (5/32)	+2.9% (4/32)
xgboost	M_cycle → M_serotype	1	—	−8.5% (0/32)	−2.8% (2/32)
xgboost	M_cycle → M_serotype	3	—	+9.1% (5/32)	+10.3% (6/32)
xgboost	M_cycle → M_serotype	12	—	+7.5% (3/32)	+9.5% (2/32)

Positive ΔMAE% means model B outperforms A. Among the cycle ablations, the XGBoost gain at $h = 6$ and $h = 12$ is the principal positive result; it is of similar magnitude in the pandemic and DENV-3 resurgence blocks (and, at $h = 6$, in the pre-pandemic block).

Table S4 Diebold–Mariano matrix — selected pairwise comparisons (Holm–Bonferroni at $\alpha = 0.05$).

h	Family A	Family B	MAE A	MAE B	DM stat	p-value	Sig (Holm)
1	baselines	xgboost	219	130	12.61	$< 10^{-9}$	✓
1	baselines	glm_negbin	219	260	−3.27	0.001	✓
1	ensemble	xgboost	129	120	7.47	$< 10^{-9}$	✓
1	glm_negbin	xgboost	230	120	13.71	$< 10^{-9}$	✓
1	pymc	xgboost	188	120	10.47	$< 10^{-9}$	✓
1	inla	xgboost	179	120	9.91	$< 10^{-9}$	✓
3	ensemble	xgboost	143	116	5.16	2.5×10^{-7}	✓
3	glm_negbin	xgboost	245	118	10.67	$< 10^{-9}$	✓
3	pymc	xgboost	227	116	10.67	$< 10^{-9}$	✓
3	inla	xgboost	327	116	8.70	$< 10^{-9}$	✓
6	ensemble	xgboost	90	93	−1.40	0.161	
6	glm_negbin	xgboost	259	96	8.58	$< 10^{-9}$	✓
6	gam_dlnm	xgboost	215	93	10.56	$< 10^{-9}$	✓
6	inla	xgboost	234	94	9.30	$< 10^{-9}$	✓
12	ensemble	xgboost	90	81	2.29	0.022	
12	glm_negbin	xgboost	253	85	8.93	$< 10^{-9}$	✓
12	gam_dlnm	xgboost	217	81	9.34	$< 10^{-9}$	✓
12	inla	xgboost	811	78	16.69	$< 10^{-9}$	✓

Negative DM statistic favours Family A; positive favours Family B. Paired MAE over the cells common to both families (so values differ slightly from the per-block medians of Table S2). The complete pairwise matrix is provided in the archived analysis outputs on Zenodo (DOI 10.5281/zenodo.20585066; `05_validation/dm_matrix.csv`).

Table S5 Out-of-ensemble comparators in the DENV-3 resurgence block (block C), by horizon: median absolute error, empirical coverage of the nominal 90% interval, and weighted interval score relative to the seasonal-naive reference. These models were benchmarked but not admitted to the operational five-family ensemble.

Comparator	h	MAE	Cov_{90}	rel-WIS
Seasonal-naive (reference)	1	379	0.70	1.00
	3	379	0.70	1.00
	6	379	0.70	1.00
	12	379	0.70	1.00
Auto-ARIMA	1	130	0.90	0.35
	3	241	0.88	0.59
	6	277	0.83	0.70
	12	288	0.79	0.81
N-HiTS	1	243	0.36	0.65
	3	387	0.29	1.06
	6	417	0.24	1.35
	12	431	0.27	1.49
Graph neural network[†]	1	454	—	—

Baselines (seasonal-naive, auto-ARIMA) report the median across the 32 states and also appear in Supplementary Table S2; N-HiTS and the graph neural network report pooled metrics from their own artifacts. rel-WIS is the weighted interval score on the $\log(1 + x)$ scale relative to the seasonal-naive reference (values below 1 indicate skill beyond it). [†]Exploratory point-forecast proof-of-concept, evaluated only at $h = 1$ and producing no predictive interval (skill versus naive persistence ≈ 1.9, i.e. worse than persistence); not developed further. The operational five-family ensemble's relative skill is reported in Table 5 of the main text.

Reporting checklists

Table S6: EPIFORGE 2020 reporting checklist for epidemic forecasting and prediction research [48].

#	Topic	EPIFORGE 2020 recommendation	Done	Where addressed
A. Overall study description and goals				
1	Study classification	Describe the study as forecast or prediction research in at least the title or abstract.	✓	Title and Abstract ("probabilistic dengue forecasting"; structured Background/Methods).
2	Research objectives	Define the purpose of study and forecasting targets.	✓	Introduction, final paragraph (objectives i–iv); target = monthly notified dengue cases per state at $h = 1, 3, 6, 12$ months.
3	Methods documentation	Fully document the methods.	✓	Methods; full technical detail in the Supplementary Methods.
4	Temporal classification	Identify whether the forecast was performed prospectively, in real time, and/or retrospectively.	✓	Methods (Validation): retrospective rolling-origin (forward-chaining); explicitly noted as retrospective in the Discussion.
B. Data description				
5	Data source origin	Explicitly describe the origin of input source data, with references.	✓	Methods (Data sources): DGE Anuarios, SINAVE, PAHO/PLISA, InDRE, NASA POWER, TerraClimate, NOAA ONI, INEGI, CONAPO — all referenced; Supplementary Table S1.
6	Data availability	Provide source data with publication, or document why not.	✓	Declarations (Availability of data and materials): public sources listed; harmonised panel and outputs released on a public repository (Zenodo DOI).

#	Topic	EPIFORGE 2020 recommendation	Done	Where addressed
7	Data processing	Describe input data processing procedures in detail.	✓	Methods (all-forms harmonisation; strictly causal expanding-window feature construction); Supplementary Methods.
C. Model characteristics				
8	Model description	State and describe the model type and document model assumptions, including references.	✓	Methods (Forecast models; Calibration): nine approaches (four baselines and five advanced model families) described with references; Bayesian priors/assumptions; conformal assumptions; Supplementary Methods.
9	Code availability	Make the model code available, or document why not.	✓	Declarations (Availability of data and materials): open-source code on a public repository.
D. Model evaluation				
10	Validation approach	Describe the model validation and justify the approach.	✓	Methods (Validation): leak-free rolling-origin forward chaining across pre-pandemic, pandemic and DENV-3 resurgence blocks; per-row leakage auditing.
11	Accuracy metrics	Describe the forecast accuracy evaluation method, with justification.	✓	Methods/Results: median absolute error, weighted interval score, CRPS, interval coverage and AUROC; proper scoring rules justified (refs [41,42]).
12	Benchmark comparison	Compare results to a benchmark or comparator model, with justification.	✓	Results: relative WIS against a seasonal-naïve reference (Table 5) and parsimonious statistical baselines.
13	Forecast horizon	Describe the forecast horizon, with justification of its length.	✓	Methods/Results/Discussion: horizons $h = 1, 3, 6, 12$ months; operational value concentrated at 1–3 months, justified by the horizon-decay of alarm skill.

#	Topic	EPIFORGE 2020 recommendation	Done	Where addressed
14	Uncertainty quantification	Present and explain uncertainty of forecasting results.	✓	Methods (Calibration) and Results: online adaptive conformal prediction intervals; coverage stratified by regime, year and incidence decile; PIT diagnostic (main-text Figure 5).
E. Translation and generalizability				
15	Non-technical summary	Briefly summarize results in non-technical terms, including a non-technical interpretation of forecast uncertainty.	✓	Abstract (Conclusions): plain-language summary including the meaning of calibrated coverage and its residual extreme-tail limitation.
16	Data versioning	If results are published as a data object, encourage a time-stamped version number.	✓	Methods (custody): data freeze 2026-05-31 with content-hash manifest; versioned, time-stamped repository archive (Zenodo).
17	Limitations discussion	Describe weaknesses of the forecast, including data-quality- and method-specific weaknesses.	✓	Discussion: under-reporting, extreme-tail under-coverage, residual peak point bias, climate-product seam, provisional 2025–2026 data, retrospective design.
18	Public-health implications	Comment on potential implications for public-health action and decision-making.	✓	Discussion and Abstract (Conclusions): complements the operational EWARS; supports timing of vector-control activities at 1–3 months.
19	Generalizability assessment	Comment on how generalizable the research may be across populations.	✓	Discussion and Abstract (Conclusions): in principle transferable to other Latin American surveillance settings with comparable notification systems, pending external validation.

Table S7: TRIPOD+AI 2024 reporting checklist for prediction models that use regression or machine-learning methods [49]. Items not applicable to a population-level, surveillance-count forecasting study (e.g. blinding, individual treatments, class imbalance for a continuous outcome, patient involvement) are marked N/A.

#	Section	TRIPOD+AI 2024 item	Done	Where addressed
Title and abstract				
1	Title	Identify the study as developing/evaluating a multivariable prediction model, the target population and outcome, and name the modelling methods.	✓	Title (probabilistic forecasting ensemble; 32 Mexican states; monthly dengue cases; machine-learning and statistical methods).
2	Abstract	Provide a structured summary.	✓	Structured Abstract (Background/Methods/Results/Conclusions).
Introduction				
3a	Background	Explain the context and rationale, with reference to existing models.	✓	Introduction (operational EWARS; prior forecasting work; evidence gap).
3b	Objectives	State the study objectives, including development and/or validation.	✓	Introduction, final paragraph (objectives i–iv); development with temporal rolling-origin validation.
Methods — data				
4a	Source of data	Describe the study design and data sources, with rationale.	✓	Methods (Data sources; Validation): retrospective state–month panel; rolling-origin temporal validation.
4b	Source of data	Specify key dates and the time of prediction (forecast origin) with horizons.	✓	Methods: January 1985–May 2026; forecast origins with $h = 1, 3, 6, 12$ months.
5a	Participants/units	Setting, locations and the analysis units.	✓	Methods: all 32 Mexican federal entities, monthly national surveillance.
5b	Participants/units	How the analysis units were assembled (inclusion/exclusion).	✓	Methods (Data sources): all-forms harmonisation; complete rectangular state–month panel (15,904 cells).

#	Section	TRIPOD+AI 2024 item	Done	Where addressed
6a	Outcome	Define the predicted outcome, with how and when measured.	✓	Methods: monthly notified all-forms dengue cases per state (DGE/SINAVE).
6b	Outcome	Blinding of outcome assessment.	N/A	Routinely notified surveillance counts; no human outcome adjudication.
7a	Predictors	Define predictors and how/when measured, available at prediction time.	✓	Methods (Feature engineering): climate, serotype, multiannual cycle and autoregressive features; strictly causal (expanding-window).
7b	Predictors	Blinding of predictor assessment.	N/A	Open administrative and environmental data.
8	Sample size	Explain how the study size was arrived at.	✓	Methods: full available census (15,904 state–months; 315,410 forecast records); no a priori sample-size calculation.
9	Missing data	Describe how missing data were handled.	✓	Methods (Data sources; Integrity): cases not imputed; notification completeness recorded per row; climate gaps handled by documented source splice with seam diagnostic.
Methods — analysis				
10a	Analytical methods	How predictors were handled in the analysis.	✓	Methods (Feature engineering; Forecast models): lagging, standardisation and causal construction; per-family feature sets.
10b	Analytical methods	Model type, model-building (selection, regularisation) and internal validation.	✓	Methods (Forecast models; Validation): nine approaches (four baselines plus five advanced model families: GLM-NB elastic-net, XGBoost, GAM-DLNM, INLA BYM2, PyMC); rolling-origin re-estimation.

#	Section	TRIPOD+AI 2024 item	Done	Where addressed
10c	Analytical methods (AI)	Model output and how it relates to the outcome.	✓	Methods: per-family predictive quantiles aggregated on the common q05/q50/q95 grid → weighted-median ensemble; online conformal prediction intervals.
10d	Analytical methods (AI)	Hyperparameter tuning and the procedure.	✓	Methods (Validation): XGBoost re-tuned per origin; GLM-NB penalty cross-validated at each refit; conformal scheme selected from seven candidates; Supplementary Methods.
11	Class imbalance	How class imbalance/rare events were handled (if applicable).	N/A	Continuous count outcome; the auxiliary outbreak-alarm classifier uses endemic-channel thresholds (Methods/Results).
12	Performance measures	Measures used to assess model performance.	✓	Methods/Results: MAE, weighted interval score, CRPS, interval coverage, AUROC and PIT.
13	Model updating	Any updating or recalibration of the model.	✓	Methods (Calibration): online adaptive conformal recalibration as observations arrive.
14	Fairness/heterogeneity	Assessment of performance across relevant subgroups.	✓	Results: performance stratified by temporal block, state, incidence decile and serotype regime.
Open science and oversight				
15	Protocol/registration	Registration and protocol availability.	✓	Declarations: not registered (observational forecasting study); analysis plan and code in the public repository.
16	Funding	Sources of funding and their role.	✓	Declarations (Funding): no specific grant.
17	Conflicts of interest	Competing interests.	✓	Declarations (Competing interests): none declared.

#	Section	TRIPOD+AI 2024 item	Done	Where addressed
18	Data and code sharing	Availability of data and code.	✓	Declarations (Availability of data and materials): public input sources; harmonised panel, feature dictionary and code archived on Zenodo under a versioned DOI (reviewer access via private link; public release upon acceptance).
19	Patient and public involvement	Whether and how patients/public were involved.	N/A	Population-level surveillance study; no individual participants.
Results				
20	Participants/data flow	Flow of data/units through the study.	✓	Methods/Results: panel composition and feature coverage; 315,410 forecast records across three blocks.
21	Model specification	Present the final model(s)/ensemble and how to obtain predictions.	✓	Results (Tables 3 and 4): ensemble weights and conformal configuration; full specification in the Supplementary Methods.
22	Model performance	Report performance with uncertainty, by horizon and block.	✓	Results (Tables 1, 3–5; Figures): error, interval score and coverage by regime, year and incidence decile.
Discussion				
23	Interpretation	Overall interpretation in context, relative to existing models.	✓	Discussion.
24	Limitations	Limitations, including data-quality and method-specific weaknesses.	✓	Discussion.
25	Usability and generalizability	Intended use and generalizability of the model.	✓	Discussion; Abstract (Conclusions).
Other information				
26	Supplementary information	Pointer to supplementary materials, code and data.	✓	Supplementary Methods; Supplementary Tables S1–S7; Supplementary Figures S1–S3.

Table S8 Outcome rescaling sensitivity. The 2025–2026 outcome was estimated from open SINAVE/DGE line-lists rescaled by ×1.25 (the overlap-period factor against the 2024 published annual scale). We repeated the evaluation excluding 2025–2026 and under rescaling factors of 1.0, 1.25 and 1.5 applied only to the test-period y_{true}. The 2024 DENV-3 peak coverage, the block-C $h = 3$ median absolute error and the global coverage of the nominal 90% interval are essentially invariant across these settings, indicating that the headline conclusions are robust to the rescaling assumption.

Test-period rescaling	2024-peak Cov_{90}	Block-C h=3 MAE	Global Cov_{90}
Exclude 2025–2026	0.865	22.1	0.889
Factor ×1.0	0.865	20.9	0.901
Factor ×1.25 (deployed)	0.865	21.3	0.900
Factor ×1.5	0.865	20.5	0.898

The rescaling factor is applied only to the observed test-period outcome (y_{true}); the forecasts and the conformal recalibration stream are unchanged. MAE is in cases per state per month. The 2024 peak lies wholly within the published-annual-scale era and is therefore unaffected by the 2025–2026 factor. Values from `05_validation/outcome_rescaling_sensitivity.csv`.

Supplementary Figures

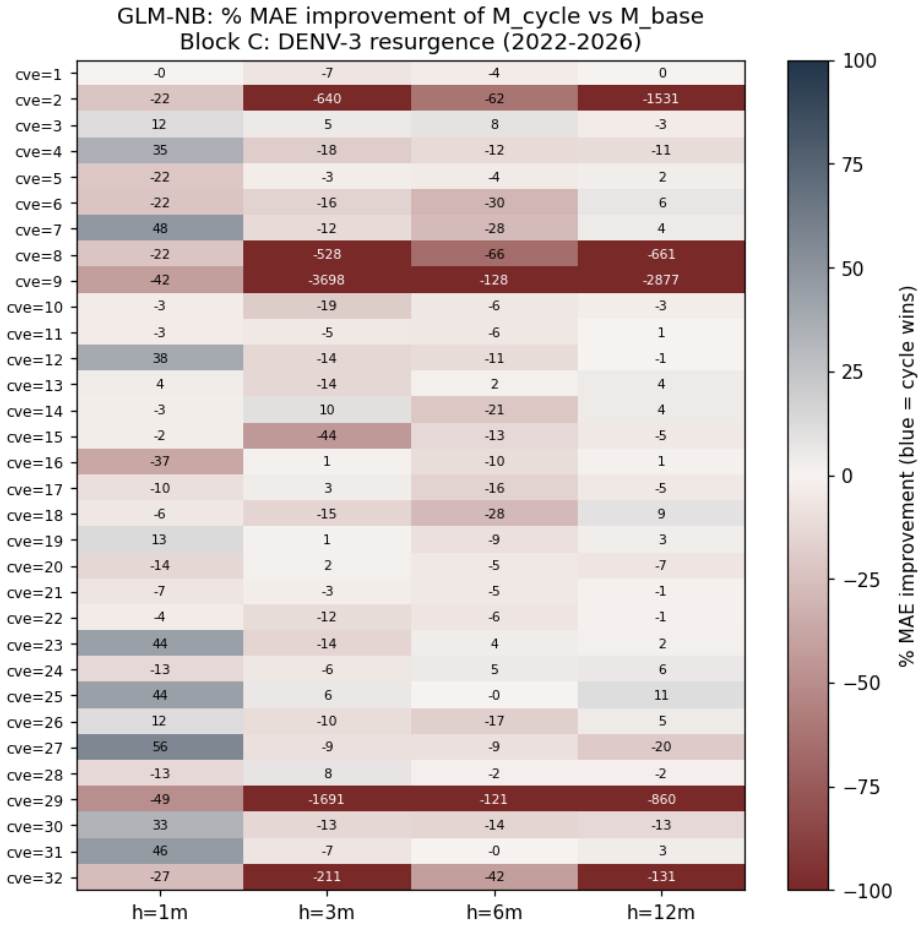

Fig. S1 Heatmap of the M_base → M_cycle ablation (E04) for the GLM-NB family, showing per-state ΔMAE% across horizons and blocks. Positive cells (blue) indicate that adding the band-limited multiannual cycle improves the GLM-NB; negative cells (red) indicate regression.

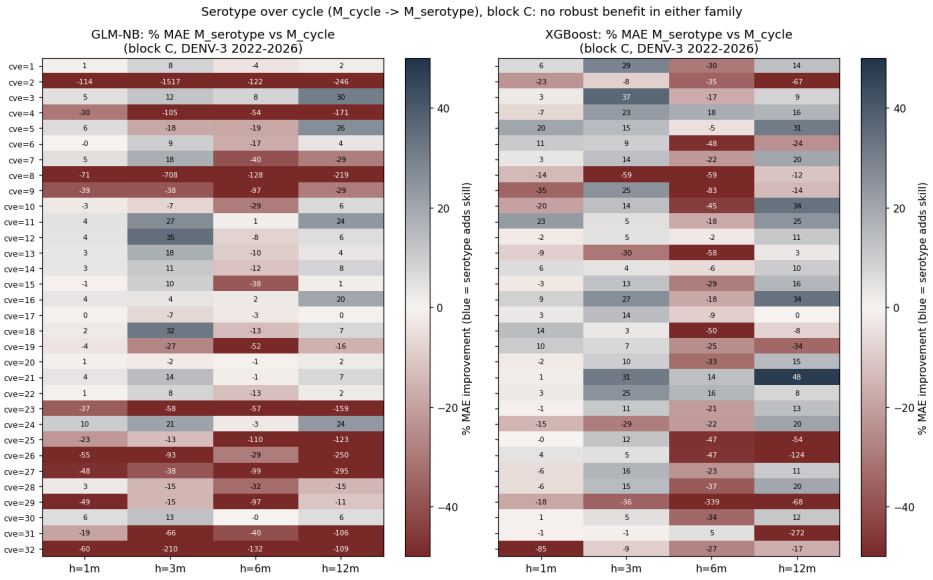

Fig. S2 Heatmap of the M_cycle → M_serotype ablation (E05b) for the DENV-3 resurgence block, per-state ΔMAE% in the GLM-NB (left) and XGBoost (right) families. Adding the national–annual serotype features yields *no robust incremental skill* in block C: the effect is mixed and small in both families (XGBoost median ΔMAE −2.8/+10.3/−14.4/+9.5% at $h = 1/3/6/12$, significant in only 0–6 of 32 states), consistent with the lagged, coarse-resolution serotype proxy described in Methods.

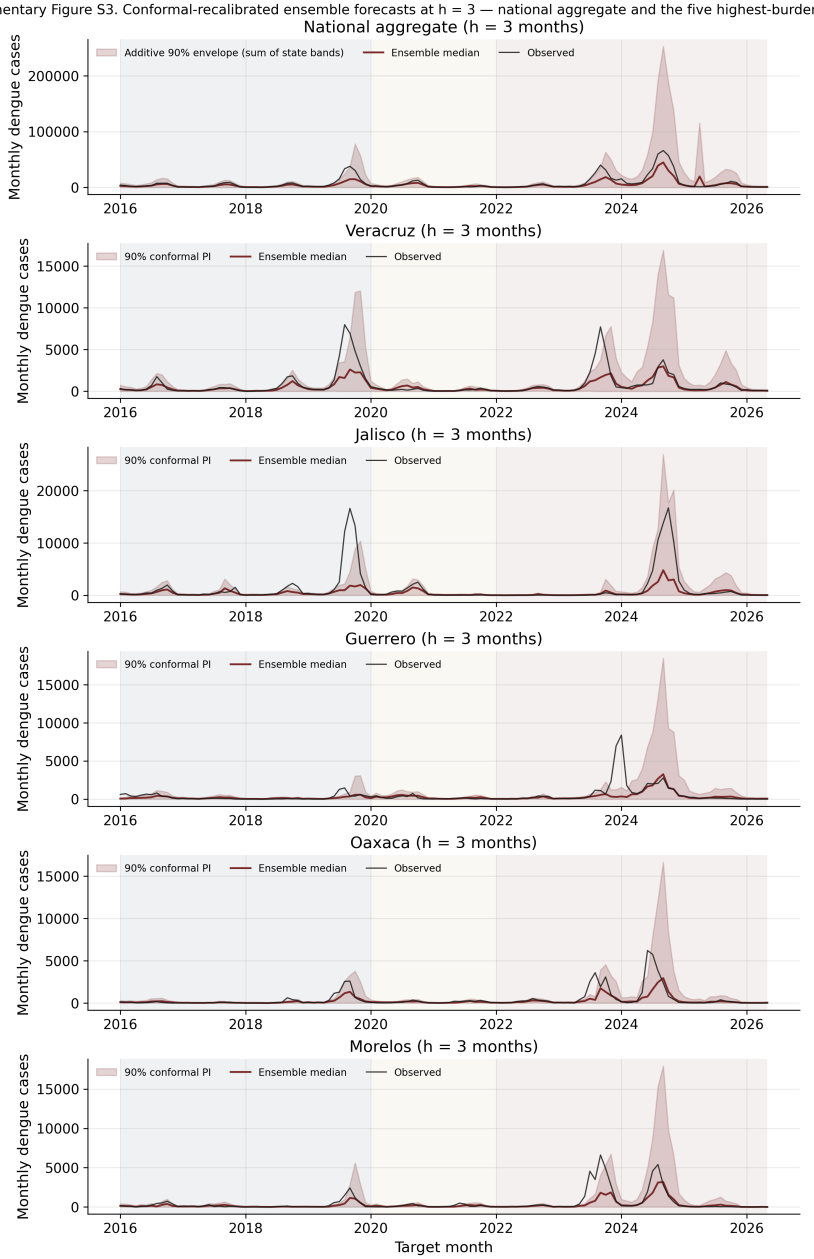

Fig. S3 Conformal-recalibrated ensemble forecasts at $h = 3$ for the national aggregate and the five highest-burden states (Veracruz, Jalisco, Guerrero, Oaxaca and Morelos), showing the median prediction with the recalibrated 90% predictive band. Shaded vertical bands delineate the pre-pandemic (A), pandemic (B) and DENV-3 resurgence (C) blocks. The national-aggregate band is an additive envelope (the sum of the per-state bands), not a separately recalibrated national interval, because the pipeline operates at the state–month scale.