Supplement: Collaborative estimation and evaluationof SARS-CoV-2 variant nowcasting in the U.S.

2025-11-11

# Additional models not evaluated in this work

The following models began submissions following the initial assessment period (October 9th, 2024 to June 4th, 2025).

| Model name | Description | Citation | Data Sources | Locations | Output Type | Ensemble? |
| --- | --- | --- | --- | --- | --- | --- |
| open\_hier\_mlr | A Bayesian hierarchical multinomial logistic regression (MLR) model for nowcasting COVID variants using variant counts based on GISAID sequences. Regression coefficients are modeled hierarchically across locations. | Abousamra E, Figgins M, Bedford T (2024) Fitness models provide accurate short-term forecasts of SARS-CoV-2 variant frequency. PLOS Computational Biology 20(9): e1012443. https://doi.org/10.1371/journal.pcbi.1012443 | GISAID | All | Point and probabilistic | No |
| gisaid\_hier\_mlr | A Bayesian hierarchical multinomial logistic regression (MLR) model for nowcasting COVID variants using variant counts based on INSDC sequences. Regression coefficients are modeled hierarchically across locations. | Abousamra E, Figgins M, Bedford T (2024) Fitness models provide accurate short-term forecasts of SARS-CoV-2 variant frequency. PLOS Computational Biology 20(9): e1012443. https://doi.org/10.1371/journal.pcbi.1012443 | INSDC | All | Point and probabilistic | No |
| ensemble | An ensemble of the hub forecasts, created by taking an equally weighted sample of all forecasts that submit samples for a given week, using the function linear\_pool from the hubEnsembles package. | https://github.com/hubverse-org/hubEnsembles/tree/main | Other model submission files | All | Point and probabilistic | No |
| PyHMLR | A Bayesian hierarchical multinomial logistic regression model with Dirichlet-Multinomial observation process. The model uses hierarchical hyperpriors to enable partial pooling across locations and clades. Location-specific concentration parameters allow the model to adaptively learn appropriate uncertainty levels for each location. Linear trends are modeled in logit space with standardized time variables. | https://github.com/trobacker/pymc\_modeling | COVID Variant Nowcast Hub S3 target data | All | Point and probabilistic | No |

# Additional figures

## Data landscape



Fig. S1 Distribution of sequence counts compared to population size in the U.S. A. Total number of sequences collected and submitted by the final evaluation date for each state during the period from September 2024 to June 2025. Color indicates whether the location’s sequences are among the top 90th percentile of all sequences submitted nationally B. Population size for each jurisdiction, ordered by total number of sequences submitted. Color indicates whether the population size is among the top 90th percentile of the total population.



Fig. S2 Example final sequence data from CA

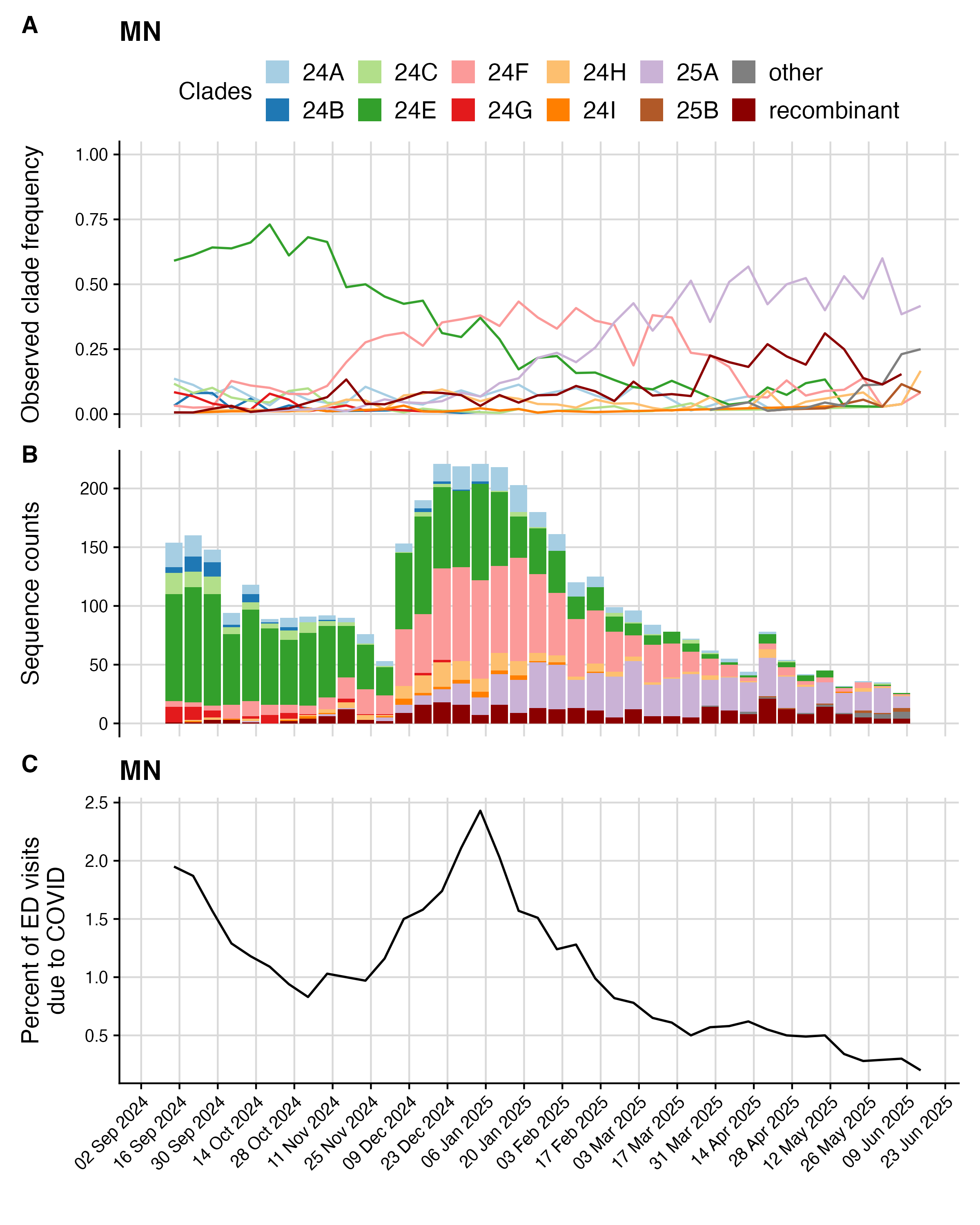


Fig. S3 Example final sequence data from IL



Fig. S4 Example final sequence data from MN

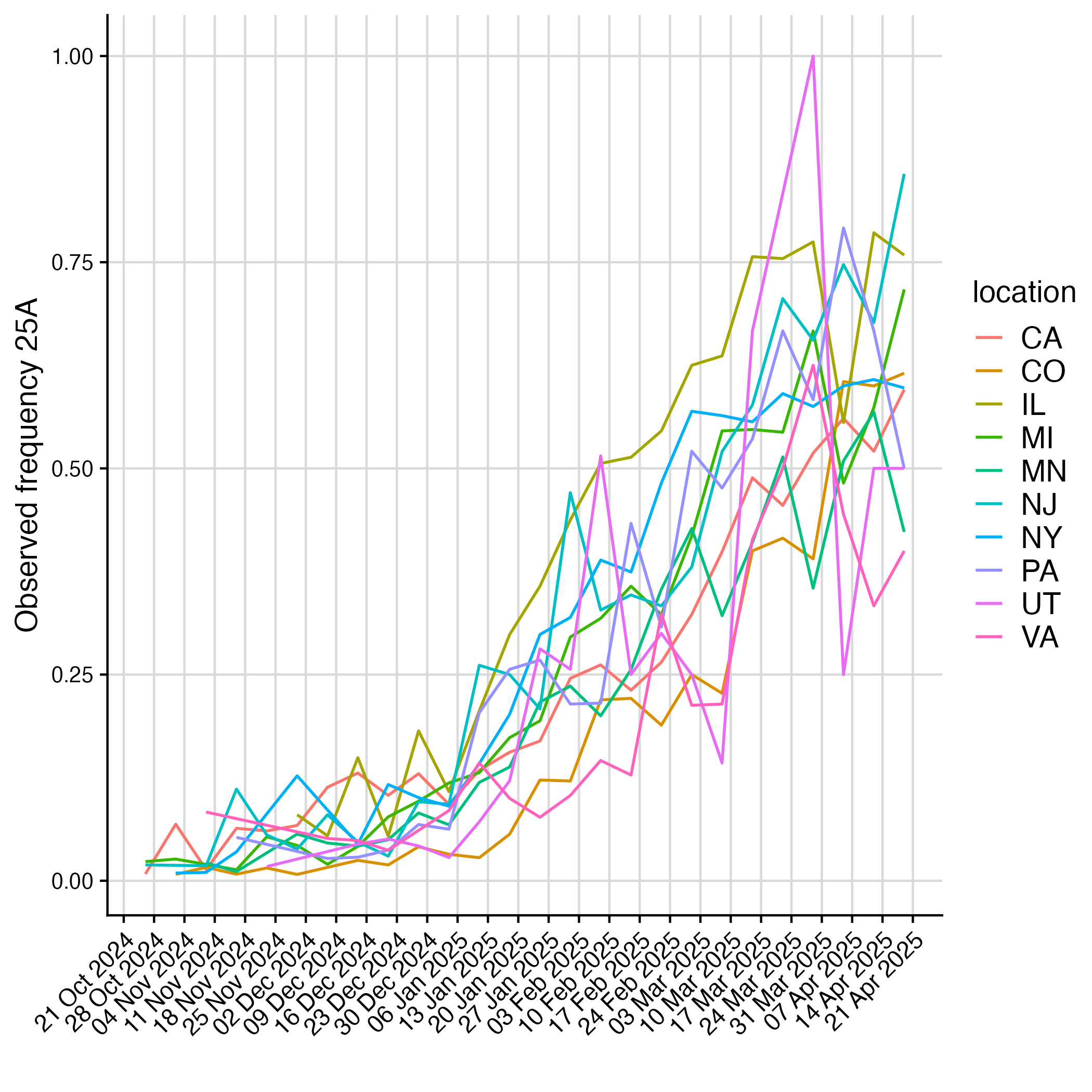


Fig. S5 Different dynamics of 25A emergence

## Model nowcasts

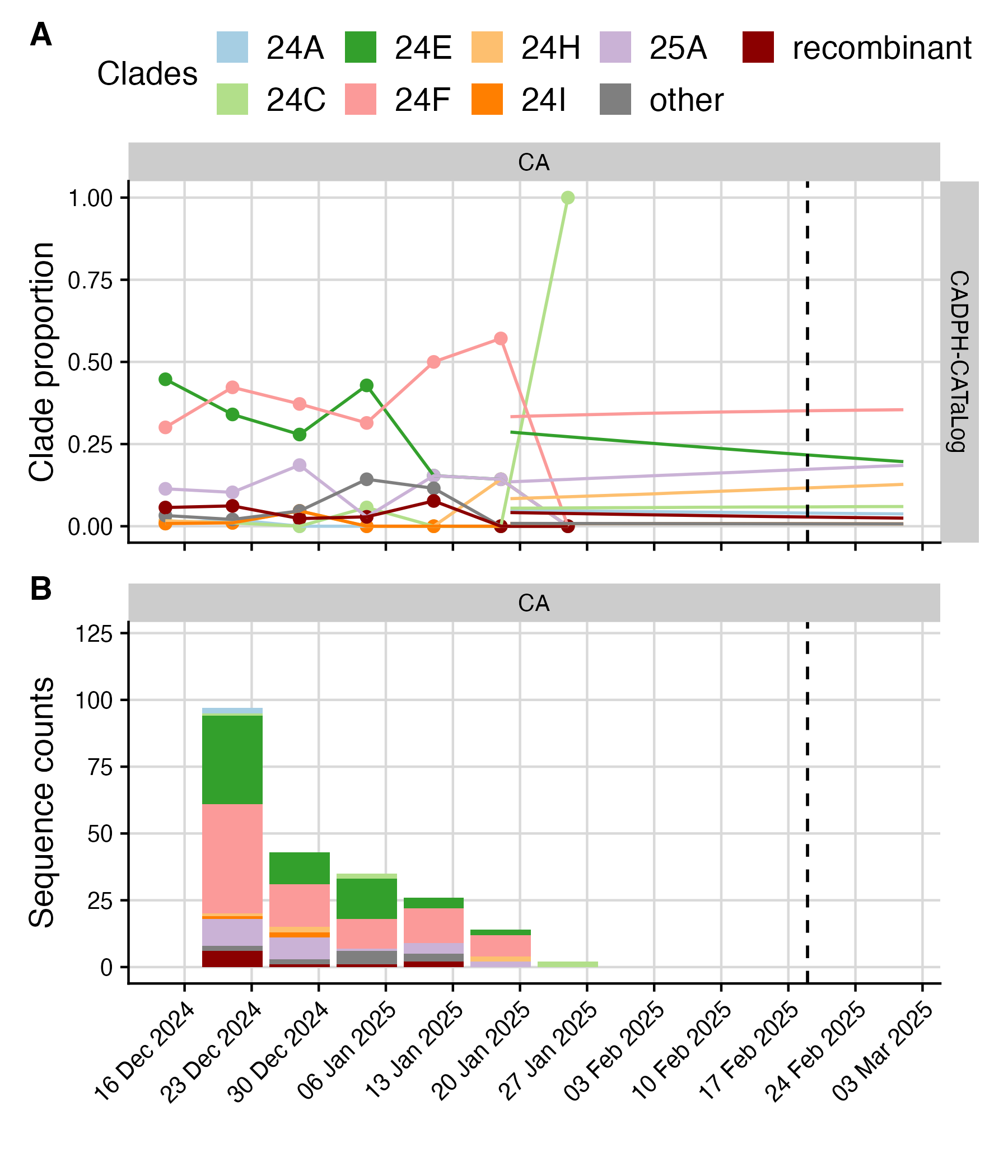


Fig. S6 Model nowcasts for the CADPH-CATaLog model from February 19th, 2025.

## Submission metadata

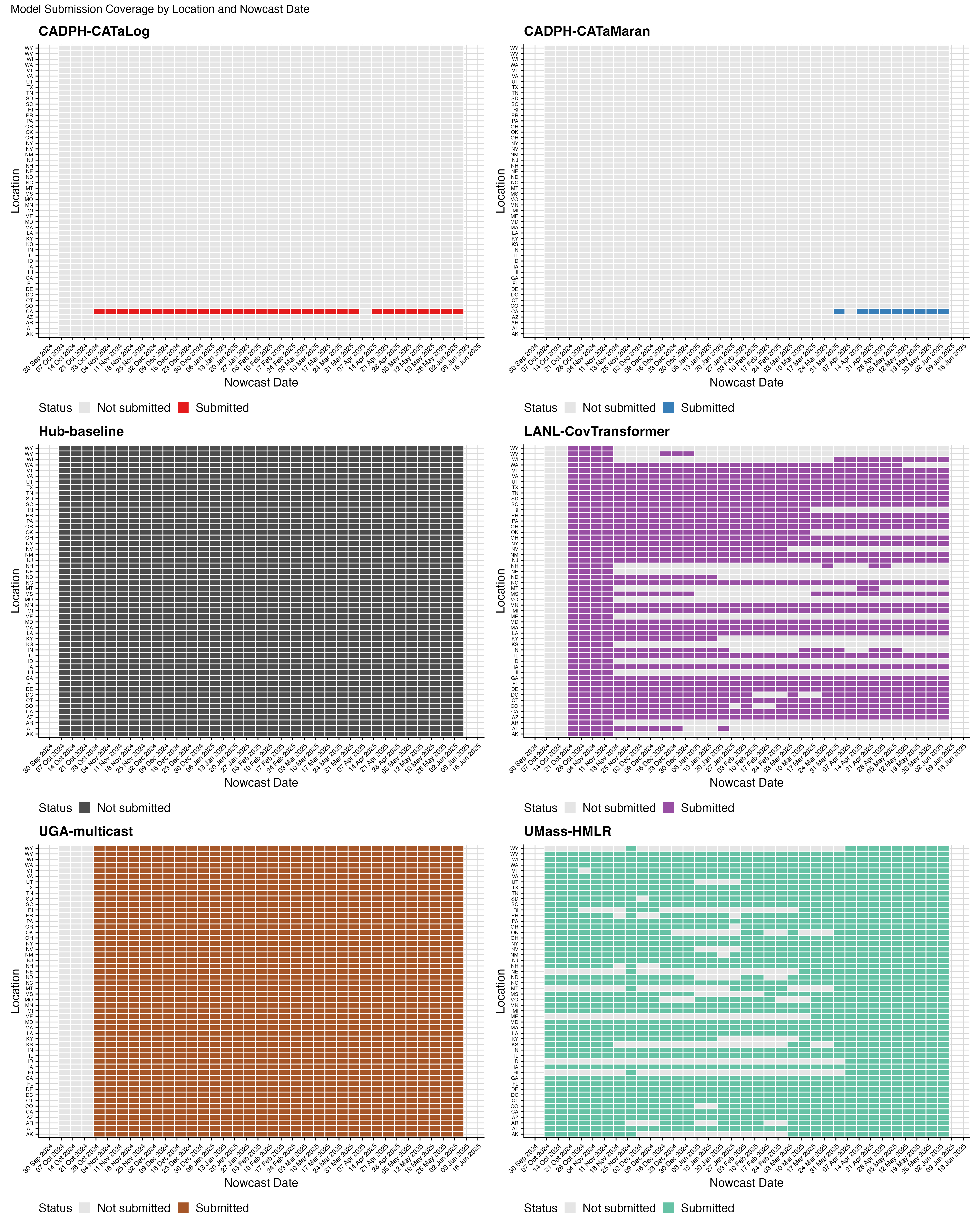


Fig. S7 Model submission coverage by location and nowcast date for each participating model. Tiles colored by model-specific colors indicate dates and locations where the model submitted a forecast. Gray tiles indicate missing submissions.

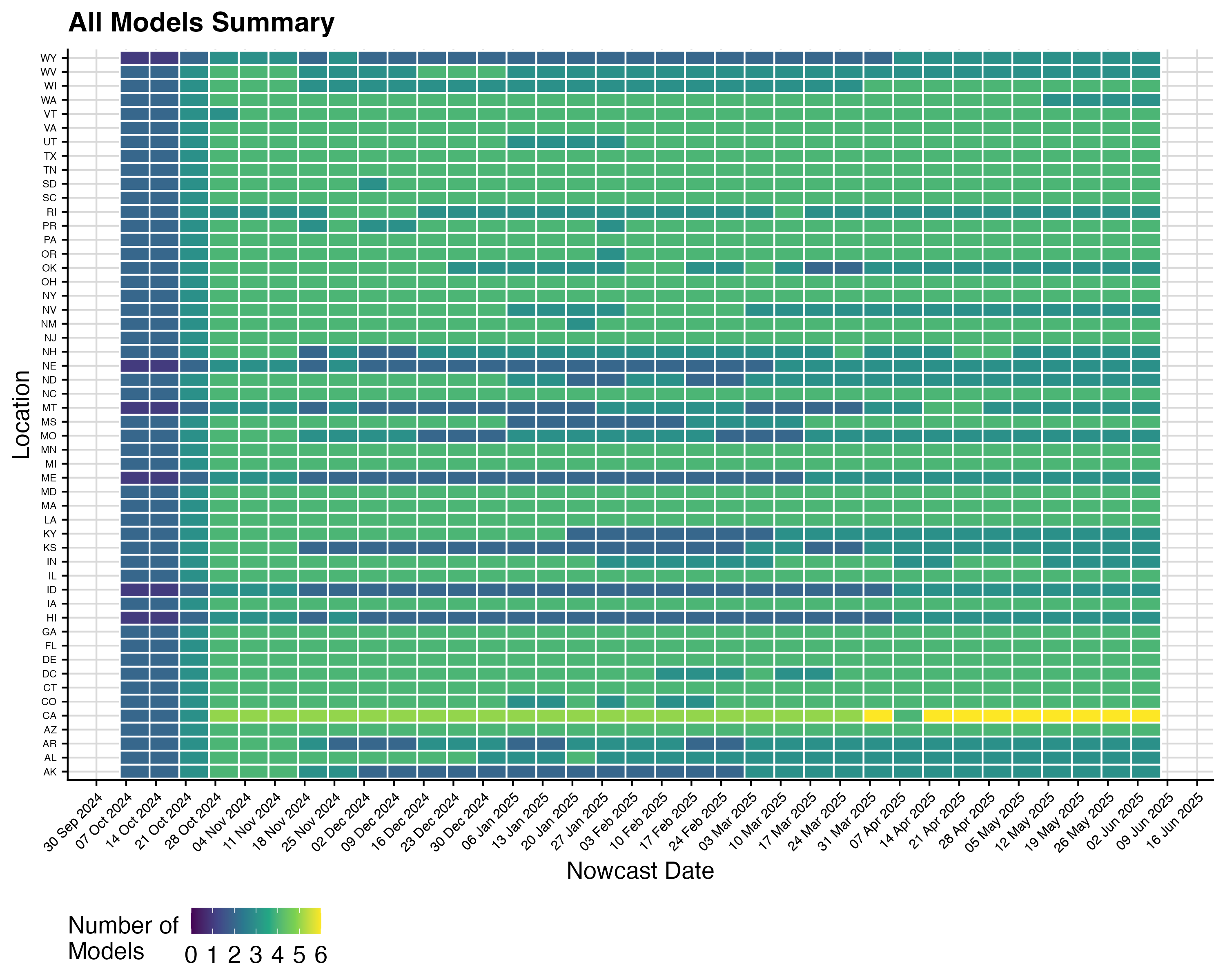


Fig. S8 Summary of model submission coverage across all models. Color indicates the number of models that submitted forecasts for each location and nowcast date combination, ranging from 0 (no models) to 6 (all models).

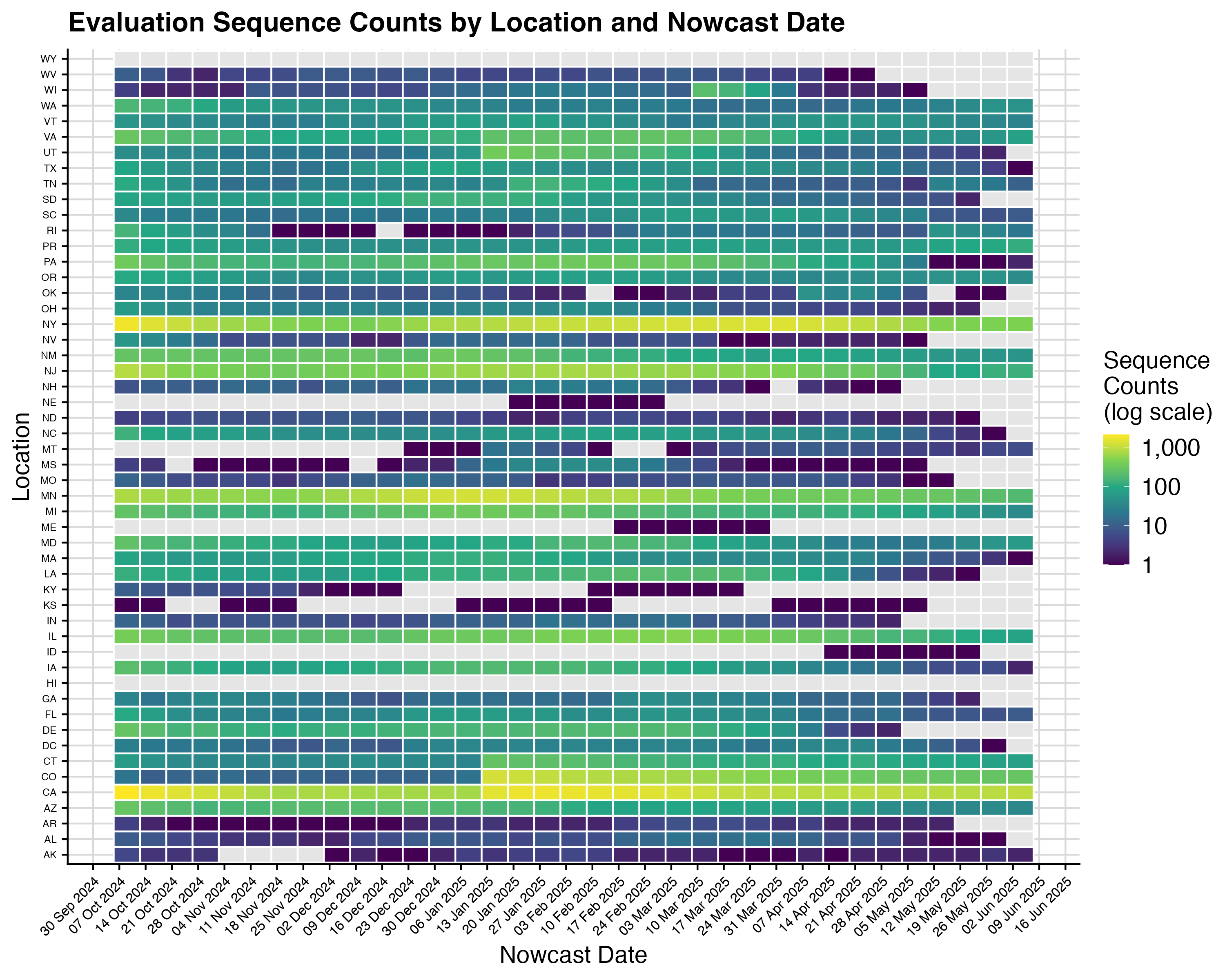


Fig. S9 Evaluation sequence counts by location and nowcast date. Color indicates the total number of sequences available in the evaluation data (log scale) for each location and nowcast date combination. Gray tiles indicate nowcast date-location pairs with no sequences across all horizons.

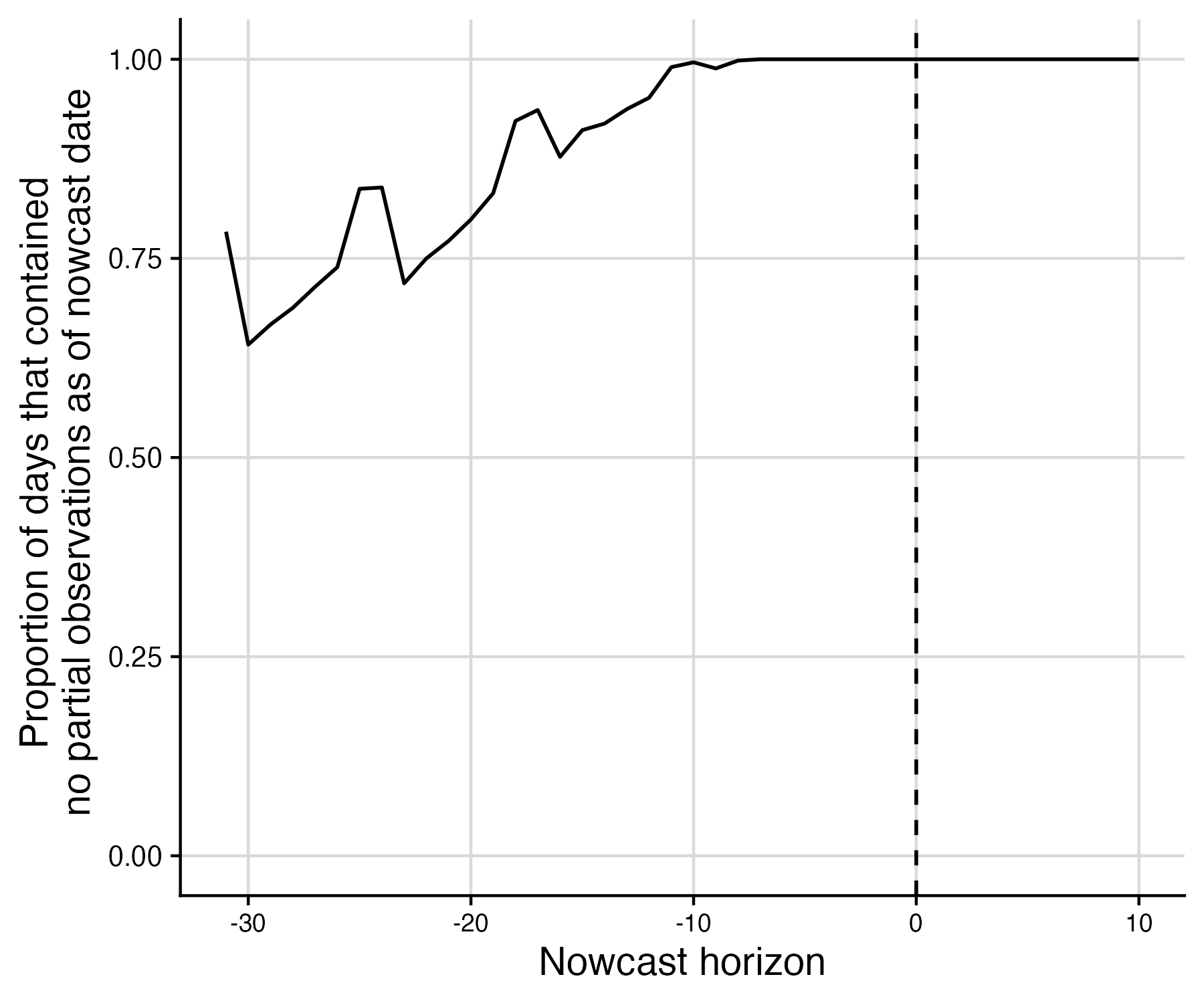


Fig. S10 Proportion of evaluated days that contained no partial observations as of the nowcast date by nowcast horizon, summarised across all locations and nowcast dates. Days which had partial observations as of the nowcast date are excluded from the main text evaluation analyses, consistent with our stated evaluation plan and in order to prevent incentivising modelers from predicting the already observed clade frequency.

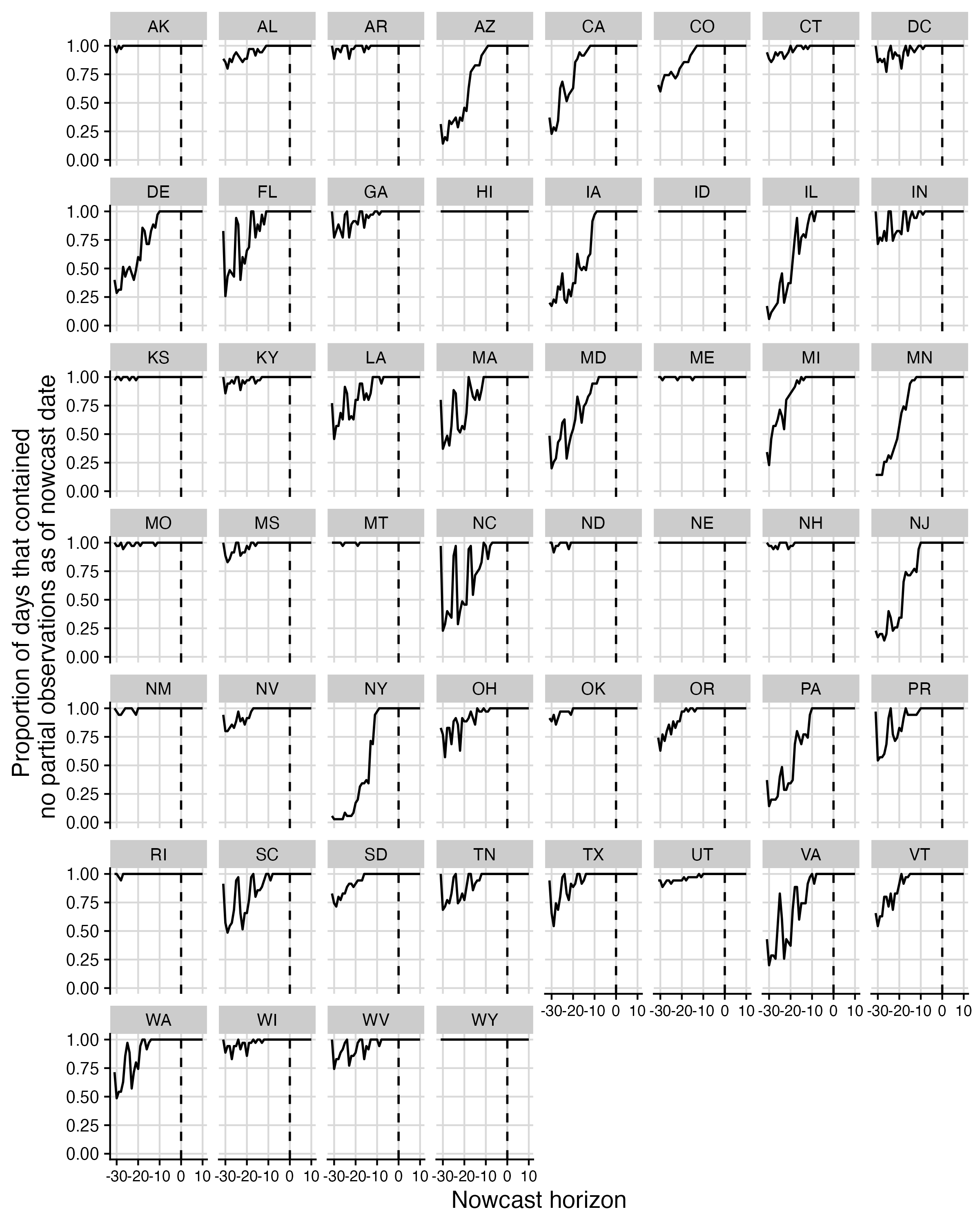


Fig. S11 Proportion of evaluated days that contain partial observations as of the nowcast date by nowcast horizon for each location across all nowcast dates. Days which had partial observations as of the nowcast date are excluded from the main text evaluation analyses.

## Additional results

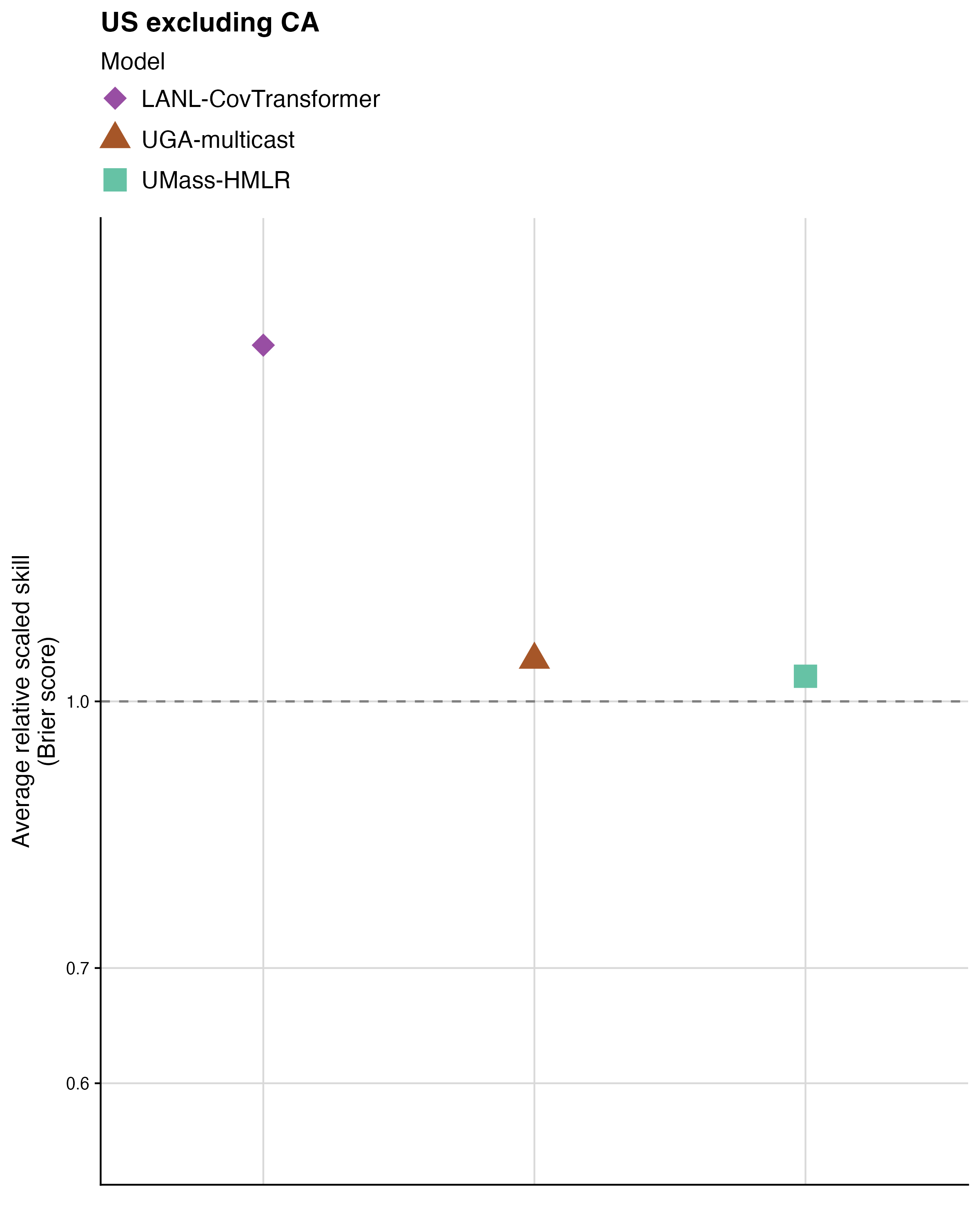


Fig. S12 Average of relative scaled skill scores on the Brier score in the US excluding California. In this figure, we compute the relative scaled skill for each horizon, location, and nowcast date and then average across the relative scaled skill scores, which has the effect of weighting each horizon, location, and nowcast date equally rather than weighting by the number of observed sequences

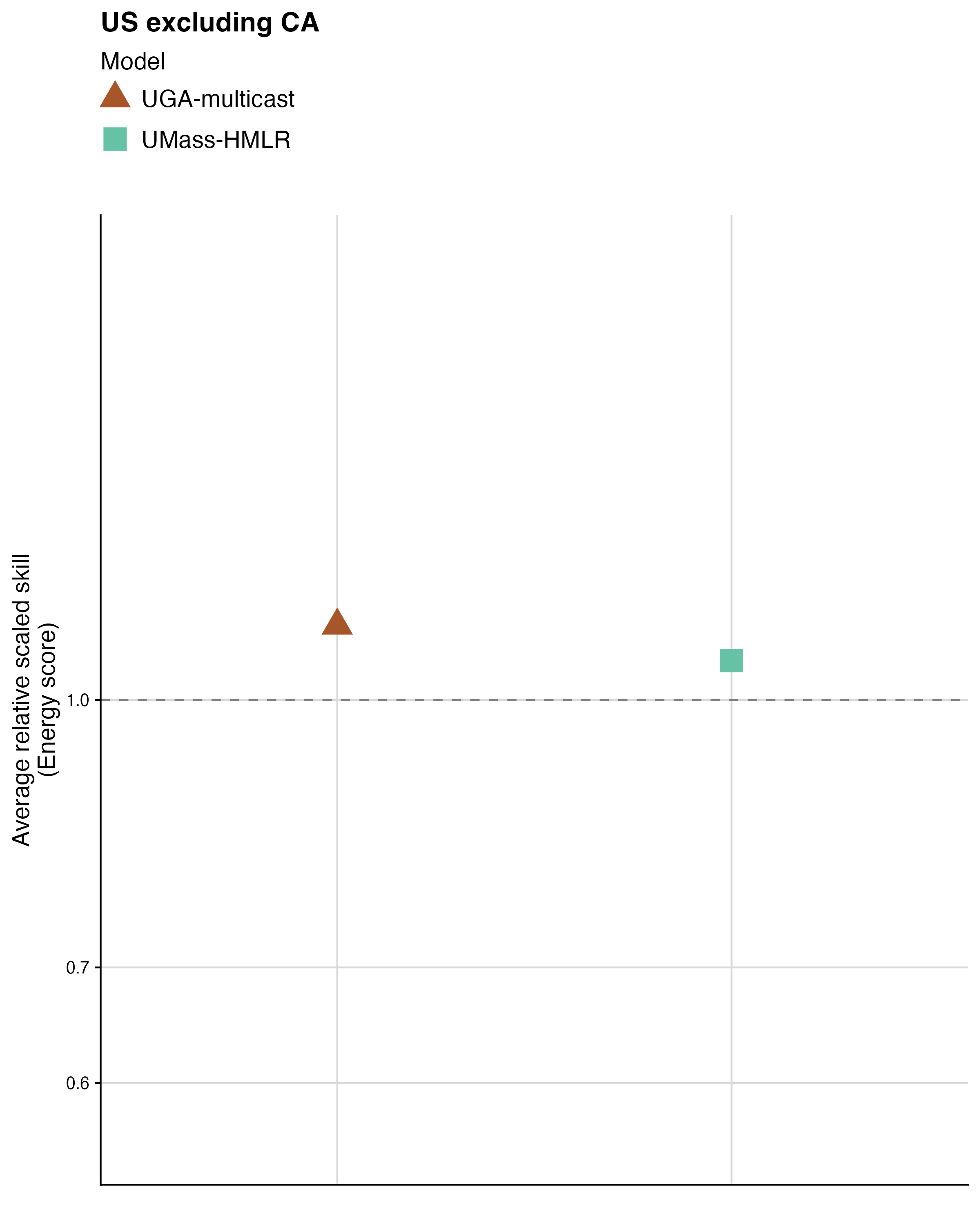


Fig. S13 Average of relative scaled skill scores on the energy score in the US excluding California. In this figure, we compute the relative scaled skill for each horizon, location, and nowcast date and then average across the relative scaled skill scores, which has the effect of weighting each horizon, location, and nowcast date equally rather than weighting by the number of observed sequences

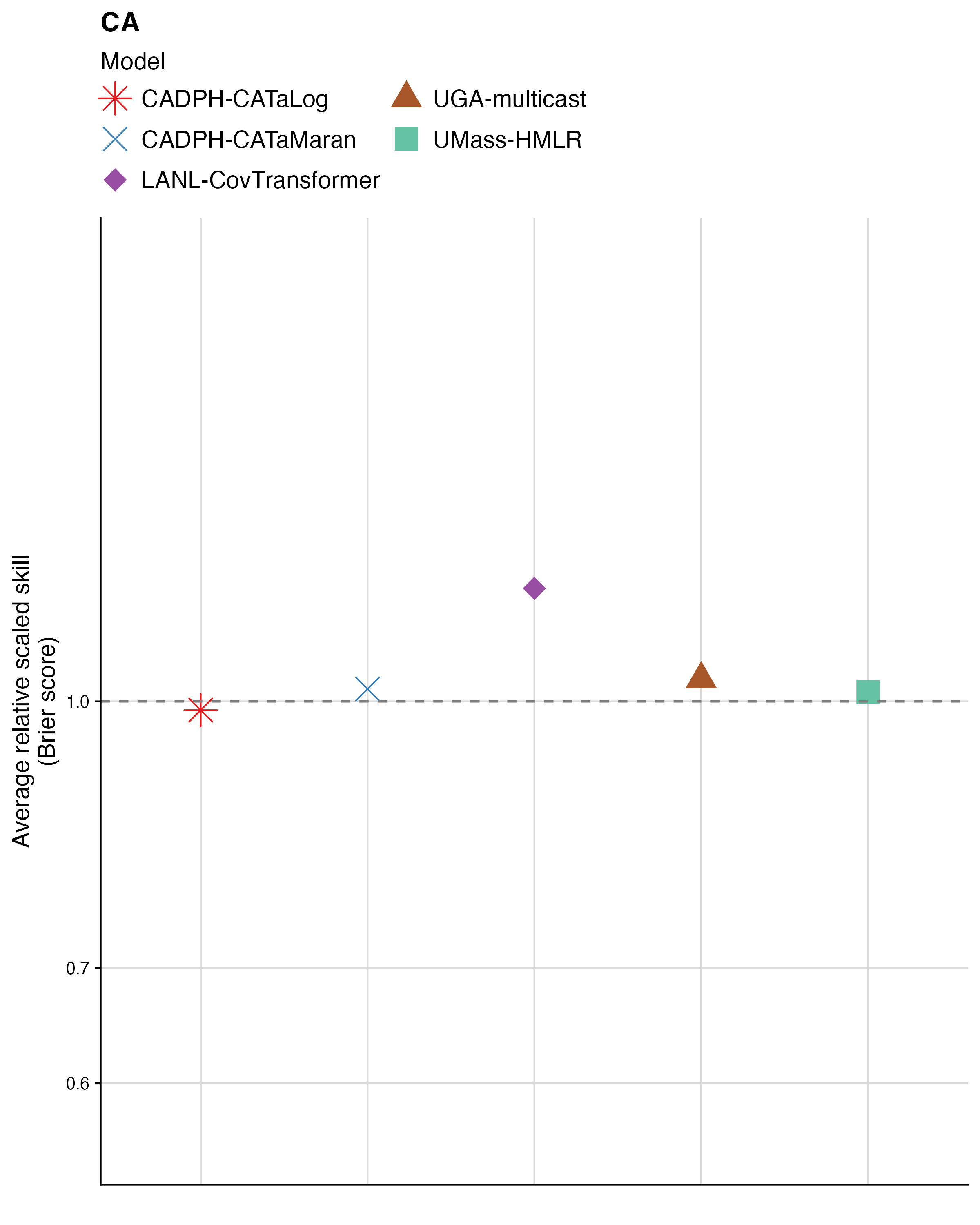


Fig. S14 Average of relative scaled skill scores on the Brier score in California. In this figure, we compute the relative scaled skill for each horizon, location, and nowcast date and then average across the relative scaled skill scores, which has the effect of weighting each horizon, location, and nowcast date equally rather than weighting by the number of observed sequences

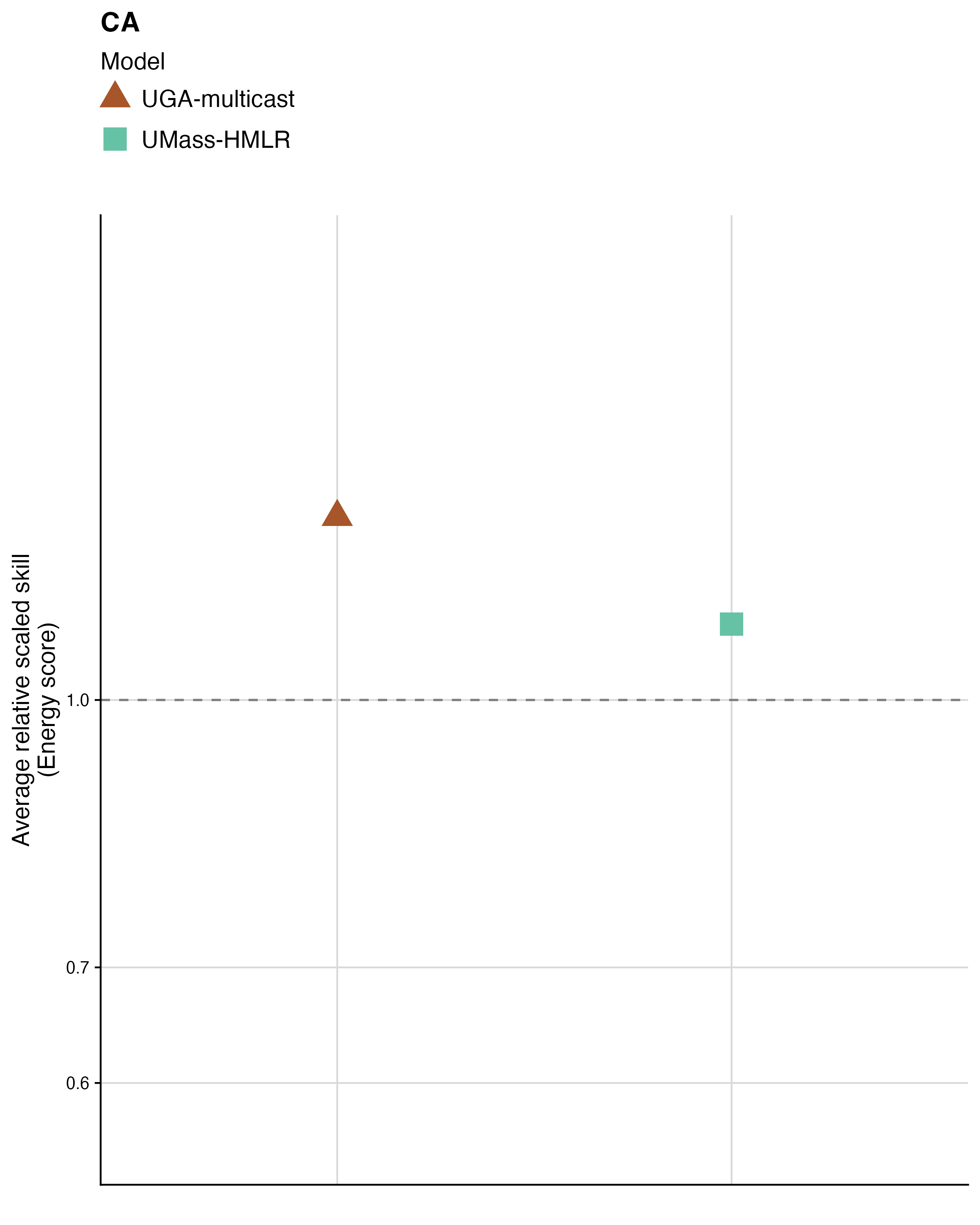


Fig. S15 Average of relative scaled skill scores on the energy score in California. In this figure, we compute the relative scaled skill for each horizon, location, and nowcast date and then average across the relative scaled skill scores, which has the effect of weighting each horizon, location, and nowcast date equally rather than weighting by the number of observed sequences

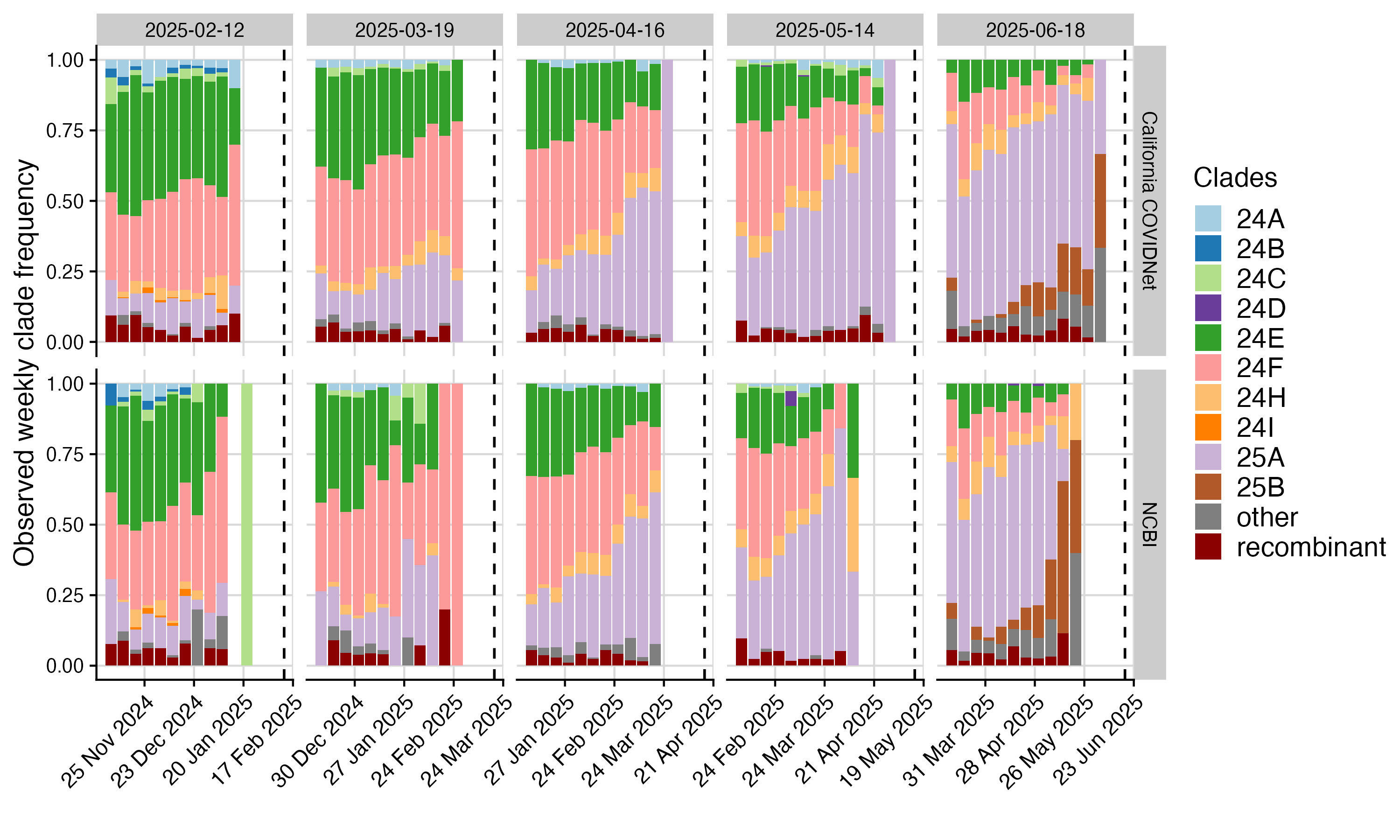


Fig. S16 Comparison of the observed clade proportions by sequence collection week across nowcast dates (columns) in the California specific data source for sequencing called California COVIDNet (top row) and the NCBI GenBank data provided by the Hub and used by most other models (bottom row). Colors indicate clade, dashed line indicates the nowcast date. In most weeks, the California specifc dataset, California COVIDNet, has observed clade proportions for a more recent week than the NCBI GenBank dataset does for California.

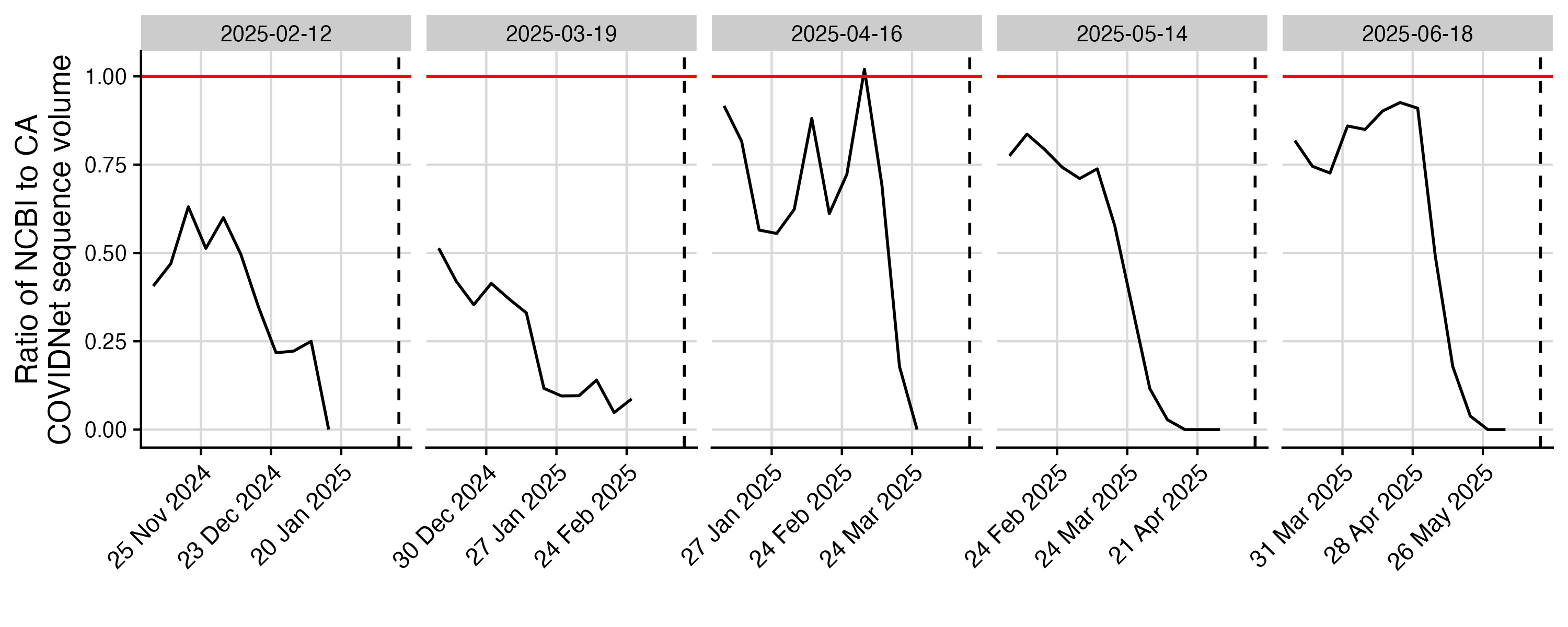


Fig. S17 Ratio of the number of sequences available in the NCBI GenBank dataset for California compared to the California-specific dataset across collection dates for a subset of nowcast dates (colums). Patterns indicate that even further back in time, the GenBank dataset at most makes up between 50% and 90% of the California specific data volume, with a particular drop off closer to the nowcast date.

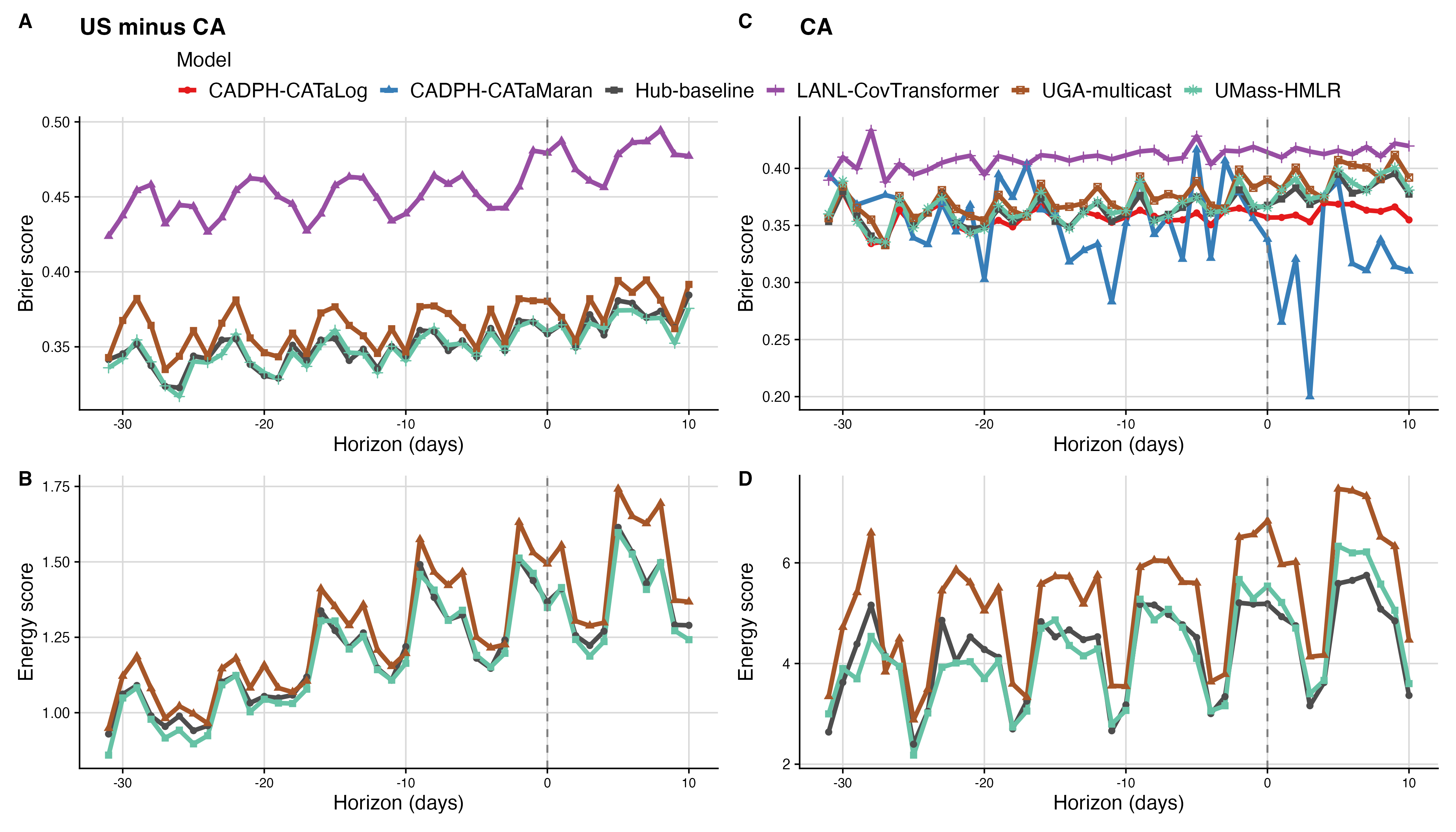


Fig. S18 Absolute Brier (top) and energy (bottom) in the U.S. excluding California (left) and California (right) by horizon.

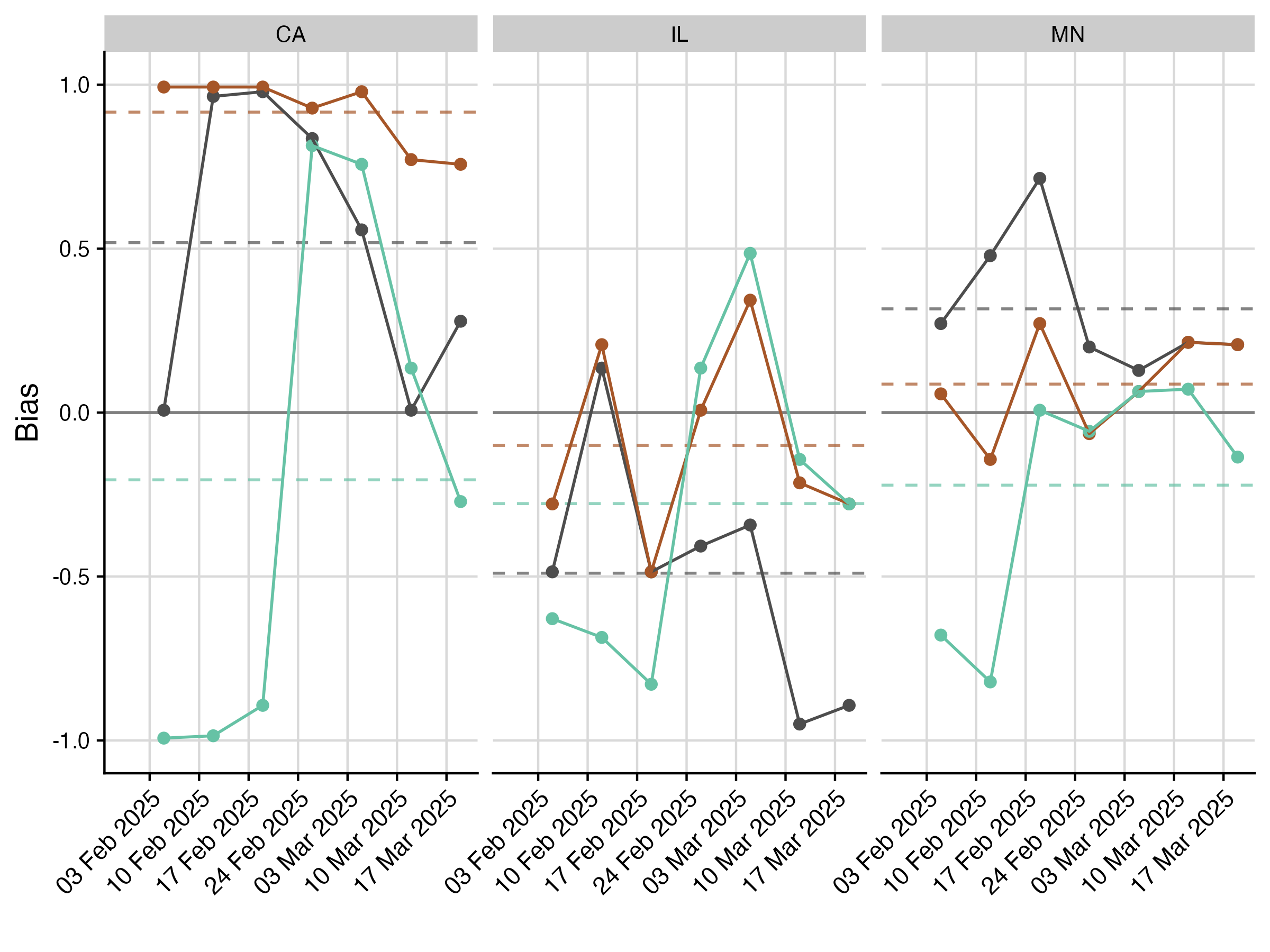


Fig. S19 Bias over time for three example states in the US during the 25A emergence.