

Preregistration: Evaluation of real-time performance of local-level flu forecasts for the 2025-2026 season in a subset of locations in the United States

Authors: Dongah Kim¹, Remy Pasco¹, Spencer J. Fox², Lauren Ancel Meyers^{1,3}, Nicholas G. Reich⁴, Sam Abbott⁵, Becky Wilson¹, Melissa Kerr⁴, Katharine Sherratt⁵, Sebastian Funk⁵, Kaitlyn E. Johnson⁵

1. Department of Integrative Biology, The University of Texas at Austin, Austin, Texas, United States of America
2. School of Informatics, Computing, and Cybersystems, Northern Arizona University, Flagstaff, Arizona, United States of America
3. Santa Fe Institute, Santa Fe, New Mexico, United States of America
4. School of Public Health and Health Sciences, University of Massachusetts, Amherst, United States of America
5. Department of Infectious Disease Epidemiology and Dynamics, London School of Hygiene and Tropical Medicine, London, United Kingdom

Background

Forecasting Hubs have become a useful tool for coordinating, communicating, and evaluating short-term forecasts of infectious disease indicators. The Flu MetroCast Hub [1] was initiated in order to address an interest in both producing and evaluating the robustness of local-level forecasting in the United States, with the 2024-2025 season representing a pilot year for the Hub. Existing forecast Hubs in the United States such as the FluSight Challenge [2] and COVID-19 Forecast Hub's [3] traditionally solicit forecasts at the state and national level, with the flu-metrocast specifically designed to focus efforts to forecast at the local (sub-state) level. Open questions remain regarding the reliability and accuracy of local-level forecasts considering the challenges associated with forecasting small numbers representing smaller population sizes, and which methods are best suited to the task. This project aims to systematically evaluate forecasts generated at the local level in the United States. The goal of these analyses will be to address the following questions applied to the 2025-2026 respiratory virus season in the U.S.:

1. How does local level forecast performance compare to the performance of aggregated forecasts within and across different models, and how do characteristics of the locality and the aggregate population impact the relative value of local forecasting?
2. Which models/methods perform best for both local and aggregate locations, accounting for variables that impact a model's overall forecast performance?

This preregistration is intended to ensure that the methods to address these questions are transparent and clearly stated prior to soliciting submissions. As much as possible, we aim for compatibility with other US forecast hubs (e.g. FluSight and U.S. COVID-19 Forecast Hub) targets and evaluations [4,5].

Forecast targets and submission

Prediction targets and locations

The Hub will solicit city, county, or metro-area (combination of counties)-level predictions plus predictions for the corresponding aggregate location (usually the state), for the target and horizons relevant to each group of jurisdictions. The target will typically be the percent of Emergency Department (ED) visits due to flu within an epi-week from the U.S. Centers for Disease Control and Prevention (CDC)'s National Syndromic Surveillance Program (NSSP) data at the Hospital Service Area (HSA) level (with New York City as an exception, here we solicit estimates of the percent of ED visits due to Influenza-like-illness (ILI)).

The Hub will accept predictions for horizons ranging from -1 to 3 weeks (with week -1 being an optional horizon), however, because horizon -1 will be completely observed, we will only score on horizons 0 to 3. The targets and horizons for evaluation are detailed below. We note that for the HSAs, the local jurisdiction is named by a representative jurisdiction, as HSAs do not necessarily align cleanly with city/county borders

Target name	Local jurisdictions	Aggregate jurisdiction	Target description	Horizon
% of ED visits due to ILI	Bronx, Brooklyn, Queens, Manhattan, and Staten Island	New York City (NYC)	Weekly number of emergency department visits due to influenza-like illness.	0 to 3 (weeks)
% of ED visits due to flu	Austin, Houston, Dallas, El Paso, San Antonio	Texas	Weekly percentage of emergency department visits due to influenza.	0 to 3 (weeks)
% of ED visits due to flu	Boston, Pittsfield, New Bedford, Lynn, Worcester, Springfield	Massachusetts	Weekly percentage of emergency department visits due to influenza.	0 to 3 (weeks)
% of ED visits due to flu	Greenville, Charleston, Rock Hill, Florence, Columbia, Horry	South Carolina	Weekly percentage of emergency department visits due to influenza.	0 to 3 (weeks)

% of ED visits due to flu	Indianapolis	Indiana	Weekly percentage of emergency department visits due to influenza.	0 to 3 (weeks)
% of ED visits due to flu	Roanoke	Virginia	Weekly percentage of emergency department visits due to influenza.	0 to 3 (weeks)
% of ED visits due to flu	Denver, Boulder, Colorado Springs, Mesa, Larimer, Weld	Colorado	Weekly percentage of emergency department visits due to influenza.	0 to 3 (weeks)
% of ED visits due to flu	Portland, Bangor	Maine	Weekly percentage of emergency department visits due to influenza.	0 to 3 (weeks)
% of ED visits due to flu	Baltimore, Hartford, Montgomery, Frederick	Maryland	Weekly percentage of emergency department visits due to influenza.	0 to 3 (weeks)
% of ED visits due to flu	Minneapolis, St. Cloud, Duluth, St. Paul, Rochester	Minnesota	Weekly percentage of emergency department visits due to influenza.	0 to 3 (weeks)
% of ED visits due to flu	Charlotte, Pitt, Wilmington, Asheville, Hickory, Concord, Onslow, Chapel Hill, Gastonia, Fayetteville, Greensboro, Winston-Salem, Durham, Raleigh	North Carolina	Weekly percentage of emergency department visits due to influenza.	0 to 3 (weeks)
% of ED visits due to flu	Ogden, Salt Lake City, Provo	Utah	Weekly percentage of emergency department visits due to influenza.	0 to 3 (weeks)
% of ED visits due to flu	Savannah, Floyd, Hall, Marietta, Macon, Henry, Cherokee, Athens, South Augusta,	Georgia	Weekly percentage of emergency department visits due to influenza.	0 to 3 (weeks)

	Columbus, La Grange,			
--	----------------------	--	--	--

Table 1. Forecast targets grouped by their aggregate jurisdiction. The local jurisdictions were chosen from within the aggregate jurisdiction based on conversations with public health partners at either the local or jurisdiction-level, who identified which locations would be of greatest public health relevance. Additional considerations were made to ensure population size was deemed sufficiently large (> 200,000) for forecasting, though this is not true in all cases.

Forecast submission dates

The challenge period will begin on November 19th, 2025 and end on May 27th, 2026. The NSSP dataset is updated on Wednesdays at 12 pm EST with data available up until the previous Saturday. NYC's data is updated daily. A cleaned target dataset will be made available by Wednesday at 2 pm EST on the Flu MetroCast Hub's GitHub repository [1]. Forecast submissions are due by Wednesday evening at 8 pm EST. Weekly submissions will be specified in terms of the reference date, which is the Saturday following the submission date. This must be included in the file name for any model submission.

Note: The Hub will begin soliciting forecasts on November 19th, 2025. Due to the shutdown of the U.S. government, the NSSP dataset may not be available for all jurisdictions, thus teams will submit for whatever jurisdictions have recent available data. At the time of writing, this may only be the data from New York City. Depending on the situation, the evaluation period may need to be adjusted accordingly. Additionally, there is a chance that the set of locations will have to be adjusted in order to accommodate jurisdictional datasets that are available during the government shutdown. The table above reflects the planned jurisdictions for forecasting and evaluation but is subject to change as the situation evolves. We will ultimately use the set of locations and forecast dates that were solicited in real-time via the MetroCast Hub's GitHub.

Submission format

Predictions will be solicited for weekly values corresponding to the CDC definition of epiweeks [6], which run from Sunday through Saturday. The target end date for a prediction is the Saturday that ends an epiweek of interest. For example, on Wednesday when the forecast is due, a forecast for horizon 0 would correspond to the forecast (or really, a nowcast) for the current week, starting on the previous Sunday and ending on the following Saturday. A "forecast" (or more accurately, a hindcast) for horizon -1 would correspond to the week two Sundays ago through the previous Saturday.

For each location and target, teams will be asked to report 9 quantiles (2.5%, 5%, 10%, 25%, 50%, 75%, 90%, 95%, 97.5%). Teams will submit predictions for each target end date corresponding to the specified forecast horizon in weeks.

Model metadata

All submitted models must contain metadata corresponding to the key characteristics of the team and the model submission. These will include: model name, team name, a brief methods description, and link to more complete methods description, a link to a GitHub repository, whether the model is an ensemble, and whether the local jurisdictions within an aggregate jurisdiction were fit jointly or independently. This last variable will be used as a variable for comparing between local models.

Evaluation

Forecast performance metrics

We will score all forecasts for local and aggregate forecasts using the weighted interval score (WIS) on the log-transformed predictions and observations [7], evaluated against the corresponding final dataset available 90 days after the final submission date. We will decompose WIS into overprediction, underprediction, and dispersion. To assess calibration, the empirical coverage of 50% and 90% prediction intervals will be reported.

In both cases, we will use the R package *scoringutils* [8] to compute WIS and coverage metrics for individual forecasts and to summarise scores across different strata.

Evaluation analysis: local vs. aggregate

For this analysis, we will compare the performance of forecasts produced using each of the local models compared to either a per-capita scaled version of the aggregate level forecast for count forecasts or the exact aggregate forecast for a particular locality for the percent emergency department visit targets. For example, if NYC had a forecast of 8,000 hospital admissions on a specific target date, and one of its boroughs represents 10% of NYC's population, the aggregate-level forecast for that borough on that date would be 800. Alternatively, if the forecast for the state of Texas was 6% of emergency department visits attributable to influenza, we would simply assume a forecast of 6% for Austin, TX as the state-level comparator. Relative WIS will be computed relative to the forecast performance of corresponding aggregate-level forecast evaluated against the locally observed data (e.g. $\text{relative WIS} = \text{score of local Houston forecast evaluated on Houston's data} / \text{score of Texas forecast evaluated on Houston's data}$). Values of 1 indicate equivalent forecast performance, values smaller than 1 indicate that local forecast outperformed the aggregate-level translated forecast, and values greater than 1 indicate aggregate-level forecasts outperformed local ones. This is meant to represent the assumption that, in absence of local forecasts, a reasonable approximation for situational awareness would be to take the aggregate forecast and superimpose it onto the locality (e.g. using Texas's forecast trend for situational awareness on what might occur in Houston).

We examine the relative and absolute WIS across multiple stratifications: overall, by nowcast horizon, and by location. Additionally, we will investigate various relationships between relative WIS averaged across forecast dates and models compared to characteristics of localities such

as population size, proportion of aggregate population, percentage of urbanization, and population density. This analysis will help generate hypotheses and investigate potential characteristics that may indicate a locality is likely to see an improvement in forecast performance due to the use of local scale forecasting.

Model-based evaluation of aggregate vs. local forecast performance

Similar in spirit to Sherratt et al.[9] we will perform a model-based evaluation to account for variables impacting the performance of the local forecasts compared to the aggregate forecasts evaluated against the local data.

To assess the impact on forecast performance of local forecasting, we will include the following variables:

- Score of the corresponding aggregate forecast (as an offset)
- Local location
- Model

Observation model : We'll assume independent and identically distributed errors on the WIS scores on the log transformed predictions and observations.

$$WIS_{h,d,l,m}^{local} \sim Normal(\mu_{h,d,l,m}^{local}, \sigma)$$

Latent model: We will model the expected WIS of a particular forecast horizon h on forecast date d at location l for model m with a generalized linear model.

$$\mu_{h,d,l,m}^{local} = \beta_0 + WIS_{h,d,l,m}^{aggregate} + \beta_1 m + \beta_2 l$$

Where h is the forecast horizon, d is the forecast date, l is the local location of the forecast (the borough or metro area), and m is the model. Location and model are modeled as random effects. The $WIS_{h,d,l,m}^{aggregate}$ offset should roughly account for the average forecast difficulty of a particular forecast location, forecast date, and horizon for the particular model.

We will plot the partial and random effects of each of these components in order to generate hypotheses and better understand the drivers of differences in forecast performance between local and aggregate forecasts.

Evaluation analysis: model comparison across locations

For this analysis, we will compare the performance of forecasts across models on local and aggregate jurisdictions, in order to identify models/methods that perform best in each setting. In this section, relative WIS will be computed relative to the baseline model generated from each location, which projects the last observed week forward for all solicited horizons. For example, the relative WIS for Houston for a particular model would be the WIS of the particular model for

Houston/ WIS of the baseline model for Houston, and the WIS for Texas would be the WIS of the particular model for Texas / WIS of the baseline model for Texas.

Because it's likely that not all teams will submit models for all forecast dates and locations, it is difficult to assess performance in an unbiased manner. However, we will compute the relative skill score which uses the geometric mean of all mean score ratios as this provides a potential work around [8,10].

We examine the relative WIS, absolute WIS, geometric average pair relative comparison, and coverage metrics across multiple stratifications: overall, by nowcast horizon, by location, and by local versus aggregate location. For the local jurisdictions, we will also stratify models by whether or not they performed joint or independent estimation across the local jurisdictions, to assess whether this has an impact on overall forecast performance.

Model-based evaluation of model performance

Similar in vein to Sherratt et al.[9], we will set up a model-based evaluation to account for variables that impact forecast performance. These will include:

- Location
- Granularity of location (local vs aggregate)
- Horizon
- Epidemic phase
- Model

The goal of this analysis will be to estimate the effect of the model while taking into account the many additional variables that impact forecast performance (i.e. certain phases are more difficult to forecast than others).

Observation model: Once again we will assume we have independent and identically distributed error around the WIS scores on the log transformed data.

$$WIS_{h,d,l,m} \sim Normal(\mu_{h,d,l,m}, \sigma)$$

Latent model: We will model the expected WIS of a particular forecast horizon h on forecast date d at location l for model m with a generalized additive model which incorporates predictors of forecast performance in potentially non-linear ways.

$$\mu_{h,d,l,m}^{local} = \beta_0 + \beta_1 m + \beta_2 l + \beta_3 e + \beta_4 g + f(horizon)$$

Where h is the forecast horizon, d is the forecast date, l is the location of the forecast (here could be either local or aggregate), and m is the model. Model, location, epidemic phase, and granularity of location are all modeled as random effects on the overall scores and horizon is modeled as a continuous spline. The epidemic phase of a particular forecast will be categorized using the same algorithm described in [9], based on the 2 week growth rate in the most recent

observations. Once again, we will plot the random effect of the model to compare model performance accounting for variables.

References

1. flu-metrocast: Short-term forecasts for influenza at the city- and county-level. Github; Available: <https://github.com/reichlab/flu-metrocast>
2. FluSight-forecast-hub. Github; Available: <https://github.com/cdcepi/FluSight-forecast-hub>
3. covid19-forecast-hub: A repository run by the US CDC to collect forecasts of incident COVID-19 hospital admissions and emergency department visits. Github; Available: <https://github.com/CDCgov/covid19-forecast-hub>
4. Bracher J, Ray EL, Gneiting T, Reich NG. Evaluating epidemic forecasts in an interval format. *PLoS Comput Biol*. 2021;17: e1008618.
5. Cramer EY, Ray EL, Lopez VK, Bracher J, Brennen A, Castro Rivadeneira AJ, et al. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. *Proc Natl Acad Sci U S A*. 2022;119: e2113561119.
6. Check code: Customizing the data entry process. 16 Sep 2022 [cited 17 Sep 2025]. Available: <https://www.cdc.gov/epiinfo/user-guide/check-code/epiweekfunctions.html>
7. Bosse NI, Abbott S, Cori A, van Leeuwen E, Bracher J, Funk S. Scoring epidemiological forecasts on transformed scales. *PLoS Comput Biol*. 2023;19: e1011393.
8. Bosse NI, Gruson H, Cori A, van Leeuwen E, Funk S, Abbott S. Evaluating forecasts with scoringutils in R. *arXiv [stat.ME]*. 2022. Available: <http://arxiv.org/abs/2205.07090>
9. Sherratt, K., Grah, R., Prasse, B., Becker, F., McLean, J., Abbott, S., Funk, S. The influence of model structure and geographic specificity on predictive accuracy among European COVID-19 forecasts. *mexRxiv*. 2025. doi:10.1101/2025.04.10.25325611
10. Bosse N, Abbott S, Gruson H, Bracher J, Asakura T, Azam JM, et al. scoringutils: Utilities for Scoring and Assessing Predictions. 2025. Available: <https://epiforecasts.io/scoringutils/>