Mind the Baseline: The Hidden Impact of Reference Model Selection on Forecast Assessment

Manuel Stapper[1*] and Sebastian Funk[1]

[1]Department of Infectious Disease Epidemiology and Dynamics, London School of Hygiene & Tropical Medicine, London, UK.

*Corresponding author

Abstract

Baseline models are essential reference points for evaluating forecasting methods, yet their selection often receives insufficient attention. We present a systematic framework for baseline model selection in epidemiological forecasting, establishing criteria for suitable baselines and demonstrating the consequences of different choices. Analysing data from COVID-19 and influenza forecast hubs, we evaluated ten baseline model frameworks. Our results reveal that baseline selection profoundly impacts forecast evaluation: for influenza, the proportion of models outperforming the baseline ranged from 10% to 93% depending on the baseline chosen. No single baseline satisfied all evaluation criteria. The choice of baseline also affected model rankings, with some baselines producing substantially different orderings of forecast model performance. We found that well-calibrated baselines do not necessarily align with good forecast performance, highlighting a fundamental tension in baseline selection. These findings highlight the need for careful baseline selection in forecast evaluation, particularly in collaborative efforts where fair comparison across multiple models is essential. We provide practical recommendations for baseline selection and suggest strategies for improving evaluation fairness when ideal baselines cannot be identified.

Keywords: Baseline models, COVID-19, epidemic forecasting, forecast evaluation, influenza, model comparison

1 Introduction

Evaluating the performance of forecasting models is a cornerstone of predictive modeling in epidemiology and beyond. When a new forecasting method or model is proposed, researchers typically compare its performance with existing approaches to demonstrate

its value. Central to this evaluation process is the choice of a baseline model: a simple reference model against which more sophisticated methods are compared. Despite its fundamental role in forecast evaluation, the selection of baseline models (sometimes referred to as null models) often receives insufficient attention, potentially leading to misleading conclusions about model performance.

A baseline model serves as a reference point for evaluating a single or multiple forecasting methods. It is typically a simple model that provides a standard of comparison, allowing researchers to quantify the added value of more complex approaches. For instance, when evaluating regression models in-sample, the overall mean serves as an implicit baseline through the R^2, quantifying how much variance a model explains beyond simply predicting the average. In time series forecasting, common baselines include the "no-change" forecast (using the most recent observation) or seasonal averages from historical data. The choice of baseline can profoundly influence conclusions: a sophisticated model may appear impressive when compared to a bad baseline, yet show marginal improvement against a well-chosen reference model.

The implications of baseline selection extend beyond individual model evaluation. In collaborative forecasting efforts, such as the COVID-19 and influenza forecast hubs, multiple research teams submit predictions that are evaluated against common baselines. The choice of baseline affects not only which models are considered successful but also how different models are ranked relative to each other. However, the baseline choice can systematically alter model rankings only when forecast models do not fully overlap in their spatial or temporal coverage, a common situation in forecast hubs where different teams may focus on different regions, periods. In such cases, an improperly chosen baseline can introduce systematic biases, making it difficult to identify where and when certain forecasting approaches truly perform well.

Despite its importance, there is no consensus on what constitutes a "good" baseline model. This lack of standardisation is problematic for several reasons. First, it reduces comparability across studies, different baseline choices can lead to conflicting conclusions about the same forecasting method. Second, without clear guidelines, researchers may inadvertently select baselines that favour or disadvantage their proposed methods, either by choosing overly simplistic references or by selecting baselines that perform in a particular way on the specific dataset. Third, the absence of structured baseline selection procedures makes it difficult to determine whether observed improvements stem from genuine methodological advances or from comparison against inadequate benchmarks.

The distinction between a good forecasting model and a good baseline model is crucial yet often overlooked. A good forecasting model aims to capture the underlying data generating process accurately, optimising for predictive performance. In contrast, a baseline need not have exceptional predictive performance, nor should it necessarily perform poorly. Rather, it should be reliable and fair across the various conditions under which models are evaluated, and thus reflect a minimum standard that any "good" forecast model should be able to outperform.

This paper addresses these challenges by proposing a structured framework for baseline model selection in epidemiological forecasting. We establish criteria for what constitutes a suitable baseline, provide recommendations for baseline selection and

evaluation, and demonstrate the consequences of different baseline choices using data from recent collaborative forecasting hubs. Our analysis reveals that baseline selection impacts model evaluation outcomes and that careful consideration of baseline properties, particularly calibration across different subsets of data, is essential for fair and meaningful forecast assessment.

1.1 Previous approaches in infectious disease epidemiology

The choice of baseline model often reflects the underlying characteristics of the epidemic time series, including seasonality, trend patterns, and data availability. We here provide a quick overview of some baseline models used across different infectious diseases and other fields. These diverse approaches to baseline model selection highlight that there is an overlap in baseline models for disease but there is no consensus for a suitable baseline across diseases or fields of applications. Further, there is no consensus on how simplicity in the baseline is balanced with the need to capture essential features of the time series.

COVID-19

For COVID-19 forecasting, researchers have predominantly employed simple methods that reflect the rapidly evolving nature of the pandemic. The most common approach uses the most recent observation as point forecast, with various methods for uncertainty quantification. Cramer et al (2022) compared a median forecast using the most recent value against a random walk model with dispersion based on the previous five observations for COVID-19 mortality in the US. Similarly, Meakin et al (2022) used a no-change model from the last observation for hospital admissions in England, while Funk et al (2020) employed a constant model with uncertainty derived from a truncated normal distribution for UK forecasts.

More sophisticated baseline approaches incorporated trend components. Bracher et al (2021b) evaluated three baselines for German COVID-19 data: a constant model, multiplicative extrapolation, and multiplicative smoothing without seasonality. Petropoulos and Makridakis (2020) applied exponential smoothing with trend but no seasonal component for global COVID-19 forecasts, capturing rapid growth phases.

Influenza

Influenza forecasting baselines often incorporate seasonal patterns, reflecting the periodicity often observed in influenza epidemics. Reich et al (2019) and Ray and Reich (2018) both employed seasonal kernel density estimation (KDE) based on historical seasons for US influenza forecasting, with Reich et al (2019) also considered the Delphi Uniform baseline that assigns equal probability to all outcomes. Colón-González et al (2021) used seasonal averages for dengue forecasting in Vietnam, citing similar approaches in influenza studies.

Several studies suggested more complex baselines. Kandula and Shaman (2019) argued for non-naïve baselines and employed ARMA-like models for influenza-like illness (ILI) in the US. Morris et al (2023) used both a constant model and a simple elastic net as comparisons for neural network models. Recent work by Ray et al (2025)

compared a flat baseline with one incorporating local trends, while Wang et al (2024) evaluated both the most recent observation and historical averages across 32 countries.

Other diseases

For Ebola forecasting, Viboud et al (2018) used an AR(3) model on synthetic data informed by the 2014-15 outbreak in Liberia. Funk et al (2019) compared three baselines for Sierra Leone data: a deterministic SEIR model, an unfocused incidence stochastic volatility model, and a Bayesian AR(1) model.

Baseline models for vector-borne diseases often need to account for complex seasonal patterns and environmental drivers. Dengue forecasting has employed various seasonal baselines. Johansson et al (2019) used both a null model with equal probability and a SARIMA baseline for South American dengue forecasts. Lauer et al (2018) utilized a 10-year median for Thailand, while Colón-González et al (2021) applied seasonal averages in Vietnam. For Zika forecasting in Latin America, both McGough et al (2017) and Lewnard et al (2014) employed autoregressive models, with McGough et al (2017) using an AR(3) as a simple baseline and Lewnard et al (2014) considering (S)ARMA models.

Other Fields

The selection of baseline models in epidemiological forecasting has parallels in other forecasting domains. Gneiting et al (2006) established the no-change forecast as a standard reference in meteorology. In economic forecasting, Lerch et al (2017) used AR(p) models with constant variance for inflation and GDP data, while Krüger (2017) employed the last observation with uncertainty estimated from Gaussian variance fitted to the previous 20 observations for Euro area growth and inflation forecasts.

2 Methods

In the context of distributional forecasting, a good forecast model and a good baseline serve different purposes and should be evaluated by different standards. A good forecast model is typically more complex, aiming to capture the underlying data generating process (DGP) accurately. Such models are evaluated on the basis of how closely their predicted distributions align with the true DGP, usually rewarding performance in accordance with the forecasting paradigm of maximising sharpness subject to calibration (Gneiting et al, 2007). In contrast, a good baseline is generally naïve and simple, designed to provide a stable and interpretable reference point. It should be robust in the sense of reliable performance across different conditions, making it hard to beat globally to avoid forecast models from overfitting.

2.1 Evaluation

The *Weighted Interval Score* (WIS) is a proper scoring rule for evaluating predictive distributions, particularly useful when forecasts are given in terms of quantiles or prediction intervals (Bracher et al, 2021a). When evaluating multiple forecast distributions jointly, individual WIS values are aggregated by taking the average, which

maintains propriety. If forecasts are evaluated on the log-transformed scale, we denote the WIS as *lWIS* in the following.

When evaluating multiple models with forecasts on different spatial or temporal subsets, pairwise relative scores, say $\theta_{ij} = \frac{\text{score}_i}{\text{score}_j}$ for models i and j, need to be harmonised. Each pairwise comparison is based on the set of overlapping forecasts of models i and j and are set to one in case of no overlap. The relative skill score of model i is the geometric mean $\theta_i = \sqrt[N]{\prod_{j=1}^{N} \theta_{ij}}$, where N is the total number of models to be compared including the baseline model. Dividing all relative skill scores by the baseline's relative skill score gives the scaled relative skill score (SRSS). This approach ensures comparability even when models differed in forecast coverage. A lower SRSS value indicates better forecast performance.

Forecast calibration refers to the agreement between predicted probabilities and observed frequencies. A calibrated forecast provides prediction intervals that contain the true outcome at the expected rates over time. In contrast to performance metrics such as the WIS, which evaluate the accuracy forecasts encompassing aspects of both calibration and sharpness, calibration on its own reflects the ability of a forecaster to quantify their own uncertainty, irrespective of how precise the forecasts are. For baseline models, the primary goal is to achieve good calibration, regardless of its precision.

A useful visual tool to assess the calibration of forecasts for integer-valued target variables is the *non-randomised probability integral transform (PIT) histogram*, see Czado et al (2009). Suppose that forecasts are given in terms of K central prediction intervals and a median forecast m. This yields a set of $M = 2K + 1$ quantile forecasts with levels $q_1, \ldots, q_M$. For completeness, we define $q_0 = 0$ and $q_{M+1} = 1$, assigning forecast values of $-\infty$ and ∞, respectively. Given a time series of observations y_t, let l_t be the largest forecast value smaller than y_t, and u_t the smallest forecast value greater than l_t, with corresponding quantile levels $q_t^{(l)}$ and $q_t^{(u)}$. Then, the PIT function for a single observation y_t is defined as

$$F(v|y_t) = \begin{cases} 0 & & v < q_t^{(l)} \\ \frac{v - q_t^{(l)}}{q_t^{(u)} - q_t^{(l)}} & \text{if} & v \in \left[q_t^{(l)}, q_t^{(u)}\right] \\ 1 & & v > q_t^{(u)} \end{cases} .$$

The PIT function for the full time series is obtained by averaging over time:

$$F(v|y_1, ..., y_T) = \frac{1}{T} \sum_{t=1}^{T} F(v|y_t).$$

If the forecasts are well calibrated, the PIT function approximates the cumulative distribution function (CDF) of a uniform distribution on $[0, 1]$.

We can quantify deviations from perfect calibration using the *Cramér-von-Mises (CvM) divergence* between the PIT function and the uniform CDF:

$$\mathrm{CvM}(y_1, ..., y_T) = T \int_0^1 \left(F(v|y_1, ..., y_T) - v\right)^2 \mathrm{d}v.$$

This quantity is easy to compute since the PIT function is piecewise linear between quantile levels. Henceforth, we refer to the CvM divergence as (global) calibration. When assessing calibration on different subsets of forecast data, such as separately for states, we define *uniform calibration* (with respect to a partitioning) as the sum of CvM divergences of subsets.

2.2 Recommendation / Guidelines

When constructing a baseline model, it is essential to assess its suitability for evaluating the performance of more complex models. The following steps outline key criteria to consider when selecting a baseline model.

1. **Key Characteristics.** Typical properties of time series, particularly in epidemiological contexts, may include positivity, integer-valued observations, periods of exponential growth, and seasonal effects. A baseline should account for key characteristics if they are undeniable.
2. **Format Compatibility.** Forecast performance metrics require comparability of outputs. If the baseline does not generate forecasts in the same format (e.g. quantiles, predictive distributions), fair evaluation is not feasible.
3. **Training Data Availability.** The baseline must not have access to more or better training information than the evaluated model. This usually includes using only the type of data that is being forecast for training and excluding any other potentially informative data sets.
4. **Simplicity.** A good baseline should be easy to fit, interpret, and communicate. Overly complex baselines defeat the purpose of providing a reference.
5. **Calibration.** A good baseline should be calibrated globally, i.e. across all spatial units, time points and forecast horizons. Ideally, a baseline model is also well-calibrated in different strata.

If a baseline does not account for key characteristics of the data, it may overly favour the forecast model. Conversely, if the baseline makes implausible assumptions about these characteristics, it may lack calibration, leading to poor forecast performance. A baseline that is not well-calibrated overall can distort the evaluation. More importantly, if the baseline is not well-calibrated in different strata, the performance of a single forecast model is not comparable between these strata in a fair way.

These recommendations can be systematically evaluated using diagnostic procedures. To assess the appropriateness of log-transformation, one can apply the Box-Cox transformation and examine whether the estimated power parameter is near zero. For multiple time series, histograms of power parameters can be used to assess whether consistent transformation across the dataset is reasonable. The approach extends

to other transformations, such as a square root transformation. Time series characteristics can be tested visually using the autocorrelation function at lag one and suspected seasonal periods. Through ordinary least squares regression of the (appropriately transformed) series on lagged observations, sine-cosine seasonal terms, and linear time trends, one can test these characteristics with standard t- and F-tests. Global calibration can be evaluated visually through PIT histograms and quantified using the Cramér-von-Mises divergence. Deviations from perfect (global) calibration can be tested with bootstrap tests sampling quantile bins according to their nominal probability levels to assess deviations from perfect calibration. For uniform calibration, PIT functions stratified by location, time, or forecast horizon can reveal systematic biases across subsets, and bootstrap procedures can test both for deviations from perfect uniform calibration and for differences between strata, ensuring fair model comparisons across all evaluation contexts. However, formal testing is not always necessary. Domain knowledge often provides clear guidance about expected time series characteristics, such as known seasonal patterns in influenza or obvious autocorrelation in epidemic data, allowing researchers to proceed directly with theoretically motivated baseline specifications.

2.3 Baseline Models

We restrict ourselves to models that are suitable as a baseline due to their simplicity in terms of computational complexity and interpretability. In total, we include ten model frameworks which can be grouped into naïve models, simple regression models, autoregressive models and time series decomposition models. Despite the spatial correlation we observe in infectious disease data, we do not account for it in the baseline models. Instead, the models are fit to data for spatial units separately. To introduce the models formally, let y_t be the number of infections observed for a selected spatial unit at time $t \in \{1, ..., T\}$. Further, we denote $\hat{y}^q_{t+h|t}$ as the forecast of the $q\%$ quantile for week $t+h$, based on the data available up to week t. As point forecast, we denote the median prediction as $\hat{y}_{t+h|t} = \hat{y}^{50}_{t+h|t}$ Given a set of probabilities $Q = (q_1, ..., q_k)'$, we denote the discretised forecast distribution as $\hat{\mathbf{y}}_{t+h|t} = \left(\hat{y}^{q_1}_{t+h|t}, ..., \hat{y}^{q_k}_{t+h|t}\right)'$.

<u>Constant Model:</u> As a very simple naïve baseline model, we use the constant model, also referred to as "random walk",

$$\hat{y}_{t+h|t} = y_t \ .$$

Forecast quantiles are derived by computing past h-step-ahead forecast errors

$$\hat{\epsilon}_{\tau+h|\tau} = \hat{y}_{\tau+h|\tau} - y_{\tau+h}$$

for $\tau = 1, ..., t$. We collect the errors as $\mathcal{E}_t = \{\pm\hat{\epsilon}_{\tau+h|\tau} | \tau < t\}$ and obtain the forecast quantiles as $\hat{y}^q_{t+h|t} = \hat{y}_{t+h|h} + \mathcal{E}^q_t$, where $\mathcal{E}^q_t$ is the $q\%$-quantile of $\mathcal{E}_t$.

Marginal Model: The marginal model is a generalisation of the constant model, where the point forecast is the average of the p most recent observations,

$$\hat{y}_{t+h|t} = \frac{1}{p^*} \sum_{i=0}^{p^*-1} y_{t-i} \qquad \text{where} \qquad p^* = \min\{p, t\} \ .$$

It reduces to the constant model for $p = 1$. Forecast quantiles are computed as above by computing past h-step ahead forecast errors.

Kernel Density: A similar alternative to the marginal model is the kernel density model (KDE) that approximates the marginal distribution of a time series. Using a kernel G with bandwidth w, and training data $y_1, ..., y_t$ the marginal density is approximated as

$$\hat{f}(y) = \frac{1}{wt} \sum_{s=1}^{t} G\left(\frac{y - y_s}{w}\right) \ .$$

From there, the CDF is derived as $\hat{F}(y) = \int_{z=-\infty}^{y} \hat{f}(z)\mathrm{d}z$. Forecast quantiles are found by inversion and they are independent of the horizon h.

Last Similar Date: To account for seasonality in time series data, we include a model (LSD) that is similar to the marginal model but accounts for seasonality. Given an integer periodicity S, we may predict that the number of infections is similar to the number of infections at the same week of the year in prior seasons. While being useful in applications with a long history of available data, infectious disease data is often not available for a longer time. Thus, we generalise the idea and include not only the same week of the year, but also similar dates. To this end, let w be a positive integer. Then, point forecasts are computed as

$$\hat{y}_{t+h|t} = \frac{1}{|\mathcal{T}(t)|} \sum_{\tau \in \mathcal{T}(t)} y_\tau$$

where $\mathcal{T}(t) = \{t - kS - w, ..., t - kS + w | k = 0, ..., \lceil t/S \rceil\} \cap \{1, ..., t\}$ collects all similar dates. The model contains the "last equivalent date" model as special case for $w = 1$. Forecast quantiles are found as in the previous models.

Ordinary Least Squares: Allowing for local trends, we include a simple regression baseline. Given the most recent p observations and a polynomial order d, we fit an OLS model $y_t = \beta' x_t + \epsilon_t$ with $x_t = (1, t, ..., t^d)'$ locally to $y_t - p + 1, ..., y_t$. The point forecast then is

$$\hat{y}_{t+h|t} = \hat{\beta}' x_{t+h}$$

where $\hat{\beta}$ is the OLS estimate. We do not assume normality of errors ϵ_t to infer forecast quantiles, but use historic forecast errors instead.

Increase-Decrease-Stable: The Increase-Decrease-Stable model (or IDS for short) fits an OLS model locally and includes a linear trend only if the most recent p observations have a clear direction, i.e. $d = 1$ if $y_t < y_{t-1} < ... < y_{t-p+1}$ or $y_t > y_{t-1} > ... > y_{t-p+1}$ and $d = 0$ otherwise.

ARMA: A popular and flexible model framework in forecasting is the ARMA model. We allow the mean of the process to be described by a function $\mu_t(\theta)$, parametrised by θ. This mean function may be constant, contain a sine-cosine seasonality term or a linear trend. We then fit

$$y_t - \mu_t(\theta) = \epsilon_t + \sum_{i=1}^{p} \alpha_i (y_{t-i} - \mu_{t-i}(\theta)) + \sum_{i=1}^{q} \beta_i \epsilon_{t-i}$$
$$\epsilon_t \sim \mathcal{N}(0, \sigma^2)$$

to all observations up to t. We estimate the ARMA parameters α and β together with θ by Maximum Likelihood. The forecast distribution is derived from the normality assumption, where σ^2 is estimated from residuals.

INARCH: We include a model designed for integer-valued time series, the INARCH model. We denote the autoregressive order as p. In case seasonality is taken into accounted, we include sine-cosine waves where S denotes the periodicity and K gives the number of harmonic waves. Let then $\log(\mu_t) = \beta' z_t$ be the seasonality term where $z_t \in \mathbb{R}^{2K}$ collects $\sin(2\pi kt/S)$ and $\cos(2\pi kt/S)$ for $k = 1, ..., K$. We fit

$$y_t \sim \text{NegBin}(\mu_t \lambda_t, \phi)$$
$$\log(\lambda_t) = \beta_0 + \sum_{i=1}^{p} \alpha_i y_{t-i}/\mu_{t-i}$$

by Maximum Likelihood. Forecast quantiles are found by generating a large number of trajectories (here 1000) following the fitted model and computing quantiles across those trajectories.

Exponential Smoothing: We include exponential smoothing of the target time series with the ETS model, see for example Chapter 7 in Hyndman and Athanasopoulos (2018) and the *forecast* package in R (Hyndman and Khandakar, 2008). The general idea is to decompose the observable time series y_t into error, seasonality and trend. Each component, if included, can be additive or multiplicative, the trend component can further be damped or not. Therefore, we have two choices for the error, three for the seasonality and five for the trend and thus 30 models in total. The periodicity of the seasonal component needs to be selected. The general framework can be expressed as

$$y_t = w(z_{t-1}) + r(y_{t-1})\epsilon_t$$
$$z_t = f(z_{t-1}) + g(Y_{t-1})\epsilon_t$$

where z_t is a latent vector containing the trend and seasonality component and ϵ_t is the error. The functions w, r, f and g are determined by the model selection. Assuming Gaussianity of errors allows to compute forecast intervals.

STL: The season and trend decomposition using LOESS (STL) also decomposes the time series into seasonality, trend and remainder

$$y_t = S_t + T_t + R_t \ .$$

The model is fitted iteratively using a locally weighted regression (LOESS). If a seasonal component is included, only its periodicity must be provided. Tuning coefficients that determine the number of iterations in the estimation and the extent of smoothing are set to reasonable default values. Forecast intervals are found by historic forecast errors. For details, see Cleveland et al (1990).

3 Data

This study utilises publicly available forecast and ground truth data for both seasonal influenza and COVID-19, focusing on weekly hospitalisation counts in the United States. All forecasts from forecast models and baseline models consist of a forecast distribution stored as a set of quantiles for probabilities $5\%, 10\%, ..., 95\%$ as well as $1\%, 2.5\%, 97.5\%$ and 99%. We combine the forecast distributions made at the same reference date, for one location across different horizons as one forecast.

3.1 Forecast Data

3.1.1 Influenza

Influenza forecasts were sourced from the CDC FluSight initiative, including both archived and the current repository, see CDC Influenza Forecasting Team (2021) and CDC Influenza Forecasting Team (2025). Forecasts span three complete influenza seasons (2021/22 through 2023/24) and one incomplete season (2024/25) and include submissions from 86 unique models across 110 forecast dates, covering all 50 U.S. states, Washington D.C., and national-level data. Forecast data from both repositories were harmonised through the following steps:

Only quantile-based forecasts were retained (23 quantiles per forecast). Forecasts were filtered to include only the target variable: weekly incident hospitalisations. Forecasts covering U.S. territories and county-level data were excluded. Only future-facing horizons were kept for analysis. For the newer repository, horizons -1 and 0 (representing retrospective or current-week forecasts) were excluded. The cleaned data were merged into a unified format with metadata including model, forecast date, location, target, horizon, and quantile. A summary of the resulting dataset is provided in Table A1 in the appendix. After cleaning, we were left with 81 models. Out of all models, only six included one location, one model only covers six locations, seven models submitted forecasts for all states and not the national level. With 67 models, the majority submitted forecasts for all locations considered in this study. For season 2021/22 and 2022/23 forecasts are available from around 30 models each. In seasons 2023/24 and 2024/25, we are left with forecasts of around 40 models each.

3.1.2 COVID-19

COVID-19 hospitalisation forecasts were obtained from the COVID-19 Forecast Hub Reich Lab (2025a). The initial dataset comprised 129 models with submitted forecasts for a variety of targets (e.g., cumulative deaths, incident hospitalisations). For this study, only quantile-based forecasts of weekly hospitalisations at the state and national levels were retained. The filtering process involved exclusion of forecasts for other outcomes (e.g., deaths), daily predictions, and county-level data, as well as of models that did not provide the required forecast structures (quantiles over relevant horizons). The final set included 57 models with hospitalisation forecasts across 152 weeks and 52 locations. As with influenza, only quantile forecasts for weekly incident hospitalisations with valid target dates and locations were included. Table A2 in the appendix summarises the availability. For eight models, there were only forecasts for one location available. For the remaining models, forecasts had been submitted for most locations, with nine models where the number of locations is slightly below 52 and 40 models that submitted forecasts for all locations. On average, forecasts were available for 67 weeks. The CovidHub baseline model was the only model with full coverage of weeks and locations.

3.2 Truth Data

Ground truth data for hospitalisations served both as evaluation basis and training inputs. For influenza, these data were included within the forecast repositories. For COVID-19, truth data were obtained using the *covidHubUtils* R package, which provides access to hospitaliation data from publicly available sources (Reich Lab, 2025b).

4 Application

4.1 Application Outline

We followed a similar process for both influenza and COVID-19 data. First, we selected an appropriate model architecture for each of the ten baseline model frameworks. Each baseline model was then applied to both the original and log-transformed time series. We generated forecasts for each individual time series, formatted consistently with the forecast hub data. To evaluate the baseline models, we discussed their suitability qualitatively as well as quantitatively using calibration metrics. Finally, we examined how the choice of a baseline model influences the evaluation of forecasts submitted to the forecast hubs.

4.2 Baseline Settings

A defining feature of influenza time series is their strong seasonal pattern, which, when relevant, should be captured by a baseline model. We also assumed temporal autocorrelation in the data. For the constant model, no architectural choices were required. The marginal model generated forecasts using the median of the ten most recent observations. The kernel density model was applied with default bandwidth

settings and did not require additional configuration. The least-similar-date model assumed a 52 week periodicity and uses a window of eleven observations, i.e. weeks were considered similar if they differed by no more than five weeks. This approach balances the trade-off between data availability and temporal similarity.

OLS models were fitted using the five most recent observations and included a linear time trend to capture short-term upward or downward movements in infection counts. The IDS model relied on the three most recent observations. For the ARMA model, we specified orders $p = q = 1$ and assumed a 52 week periodicity and no time trend in the marginal mean function. The INARCH model was fitted with order $p = 1$ incorporating a 52 week seasonal pattern represented by $K = 8$ harmonic waves. The conditional distribution was modeled as Negative Binomial. Unlike the other baselines, the INARCH model was not applied to log-transformed data as these were no longer integer valued.

The ETS model was applied with additive errors, no trend component, and used pre-filtered seasonality: the seasonal component was estimated and removed prior to fitting, and then added back to the forecasts. Similarly, the STL model was applied with default settings, and the seasonal component was handled through the same pre-filtering and post-correction process.

Unlike influenza, COVID-19 time series do not exhibit a regular seasonal pattern but often feature phases of exponential growth. As with the influenza case, the constant model required no architectural choices. For the marginal model, point forecasts were computed as the average of all past observations. The kernel density model was again applied using default bandwidth settings. The LSD model reduced to a marginal model due to the absence of seasonality. We included it as a marginal model based on the five most recent observations.

The OLS model was fitted using the five most recent observations and included a linear time trend. The IDS model was based on the three most recent observations. For the ARMA model, we specified orders $p = q = 1$, no seasonality, but include a linear time trend in the marginal mean function. The INARCH model was used with order $p = 1$, no seasonality, and a Negative Binomial conditional distribution.

For the ETS model, we used multiplicative errors, a damped additive trend, and no seasonality when applied to the original data. When applied to log-transformed data, we assumed additive errors instead. The STL model was applied with default settings and excluded a seasonal component.

Some of the baseline models described above do not enforce non-negativity when applied to time series on the original scale. While point forecasts mostly remained non-negative, lower quantile forecasts occasionally fell below zero. In such cases, negative quantiles were truncated at zero to ensure plausibility.

4.3 Baseline Evaluation

Point forecasts were always non-negative across all models. When models were applied to log-transformed time series, forecast intervals respected non-negativity upon back-transformation. For models applied to the original scale, forecast intervals remained non-negative only for the KDE and INARCH models.

For influenza time series, the constant, marginal, KDE, OLS, and IDS models did not explicitly account for seasonal patterns. In contrast, the ARMA, INARCH, ETS, and STL models incorporated seasonality in their structure. Among all baseline models, only the INARCH model was specifically designed to handle integer-valued time series.

For COVID-19 time series, which often exhibit phases of exponential growth or decline, the OLS, IDS, and STL models could capture local linear trends. The ARMA and ETS models instead reflected broader, overall trends and might not adapt well to sharp changes. Short-term temporal correlation was modeled by the constant, LSD, OLS, IDS, ARMA, INARCH, ETS, and STL models, each incorporating some form of dependency on recent observations.

All models produced forecasts in a format compatible with the forecast hub data. They relied only on past observations from the individual time series for training. The simpler models (constant, marginal, KDE, LSD, OLS, and IDS) were easy to fit and interpret. The ARMA and INARCH models were estimated via maximum likelihood, while the decomposition based models (ETS and STL) used iterative procedures. Although more computationally intensive, these models were simple to interpret.

We assessed calibration through the PIT function and the resulting CvM divergence. For a single baseline model, all forecasts across points in time, regions and forecast horizons were jointly evaluated. Figure 1 shows the PIT functions of all baseline models for influenza and COVID-19. Highlighted are the baseline models that were calibrated best and worst. See also Figure B1 in the Appendix for a detailed overview. In case of influenza infections, we see that the Kernel density model (applied to log-transformed time series) exhibited poorest calibration. The ARMA model (applied to log-transformed time series) had the best calibration among all baselines. We further see that the KDE, LSD and marginal models were poorly calibrated. A possible explanation is that the time series had low infection counts in the beginning. Forecasts in the later phases rely to greater extend on those times with low counts and thus infection counts were systematically underestimated. The ARMA model as the best calibrated baseline was able to adapt to such a change in overall level quicker, preventing systematic over- or underestimation.

The best calibrated baseline for COVID-19 time series was the constant model applied to log-transformed time series while the worst calibrated baseline was the STL decomposition. This reflects that individual time series had a short-term dependence that needed to be captured for calibration but otherwise were not exhibiting much temporal structure.

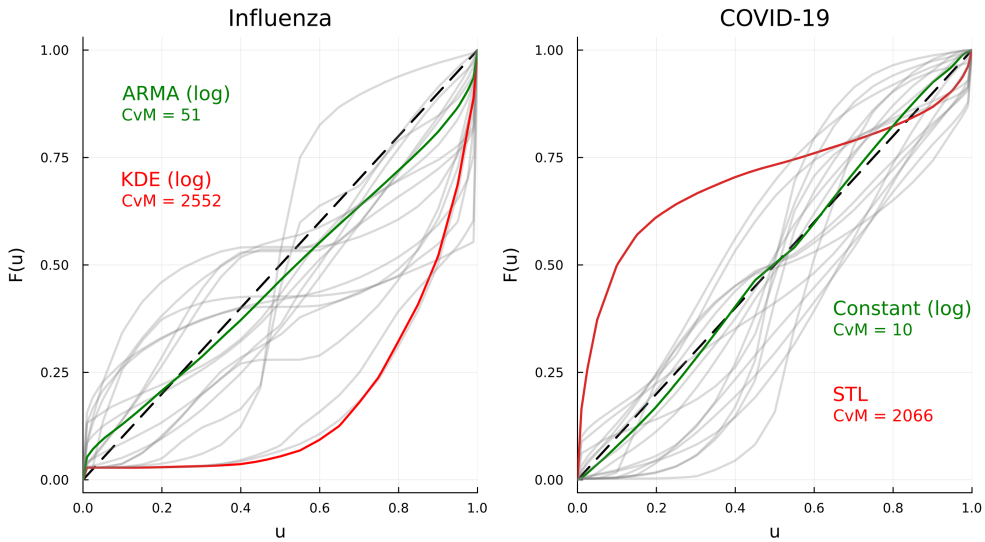

Fig. 1 PIT functions for different baselines

When comparing the performance of a forecast model between subsets of the data, such as for a different horizons, regions or phases, it is important to assess the calibration of the baseline model for the subsets individually. Otherwise, we would risk an unfair assessment caused by differences in the calibration. Figure B2 in the appendix shows PIT functions for different subsets, split by horizon and region, for influenza and COVID-19 and the corresponding baseline model that was best calibrated on the complete set of forecasts. The ARMA model showed differences in calibration for different forecast horizons, with forecasts at shorter horizons being better calibrated. For the constant model and COVID-19, these differences were less pronounced.

When looking at PIT functions for different regions, we see that there were larger deviations from the uniform CDF for influenza, whereas for COVID-19 the PIT functions appeared more closely scattered around it.

4.4 Baseline Correction

When calibration is not ideal across regions, time periods, or forecast horizons, several strategies can be applied to improve it. One option is retrospective re-calibration of baseline forecasts. By appropriately scaling and shifting the quantile forecasts of a given baseline model, calibration can be improved within specific subsets of the data, potentially improving overall calibration. Another straightforward approach is to allow for different baseline model selections across subsets, such as regions.

Figure 2 illustrates this idea, showing that the best calibrated baseline model differed across regions. For COVID-19, the constant model (applied to log-transformed data) achieved the best calibration in most regions. In contrast, for influenza, the optimal baseline varied considerably. Across the 52 regions, one of eight baseline models

was the best calibrated in at least one case. Notably, although the ARMA model (in logs) was the best calibrated overall, it was the best calibrated model in only 20 of the 52 regions.

Allowing for region-specific baseline model selection and then aggregating the resulting forecasts can lead to improved overall calibration. For influenza, the CvM divergence for the best single baseline model was around 51 and was reduced to around one by selecting best individual baselines. For COVID-19 the effect was not as pronounced with a reduction from ten to one. The moderate effect is not surprising, since the single best baseline was the best calibrated baseline in most states.

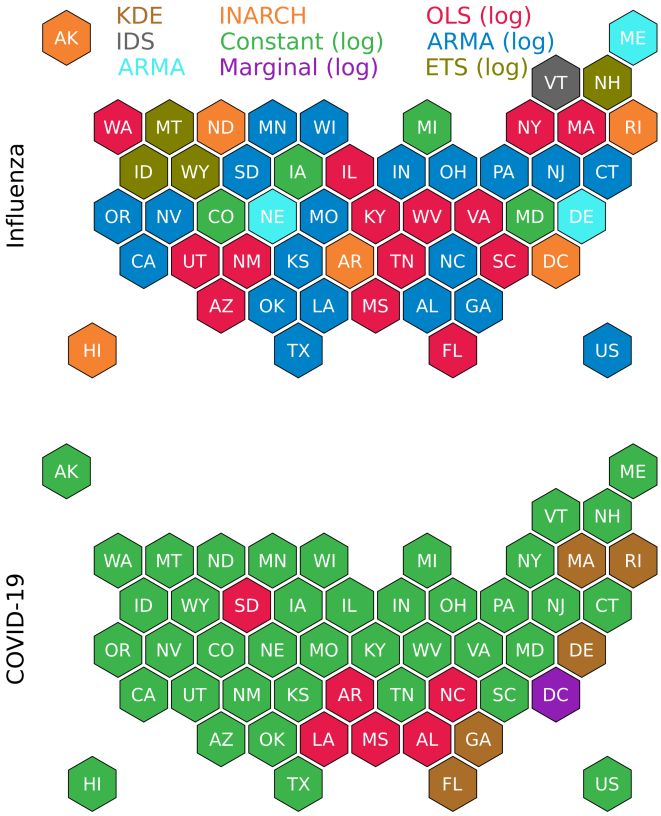

Fig. 2 Best baseline model for each US state for influenza (top) and COVID-19 (bottom)

Another strategy for achieving well-calibrated forecasts is based on the following idea: if all observations in a time series for a given region were drawn from the same marginal distribution, and if this distribution was known, we could generate perfectly calibrated forecasts by simply using its quantiles. The marginal and KDE models are motivated by this principle. However, in practice they struggle with calibration because

the marginal distribution must be estimated from past data, and the magnitude of epidemiological time series levels can vary over time.

To ensure fair comparison with submitted forecasts, Property 3 of the evaluation recommendations restricted us from using the full data history when generating baseline forecasts. Nonetheless, for illustrative purposes, we included an additional marginal distribution baseline in the analysis. The baseline computes quantile forecasts as empirical quantiles of the time series, excluding the target observation and every observation for which there is no forecast available from forecast hubs.

Both additional baseline models were included to achieve better uniform calibration, i.e. improved calibration across subsets of the forecast data. Our focus was on calibration for all states. A baseline that demonstrated good calibration across all states would enable fair comparison of a single model's forecast performance between states. We assessed this by computing PIT functions separately for each state. Figure 3 presents PIT functions for the two additional baseline models alongside the baseline exhibiting the best global calibration. The PIT functions were summarised as median values with pointwise 80% intervals and ranges across all states. While a baseline with good global calibration would exhibit a median PIT close to the uniform CDF, a baseline with good uniform calibration would additionally display narrow intervals around it. For both influenza and COVID-19, the best single baselines (ARMA-l and Constant-l, respectively) exhibited wider intervals compared to the two additional baselines. Selecting different baselines for individual states did not substantially improve uniform calibration in either case. However, the marginal model that utilises future data significantly improved uniform calibration. For influenza, despite showing near-perfect global calibration, there were deviations from perfect uniform calibration, particularly for small quantiles. These deviations are likely attributable to the large number of zero observations. For COVID-19, we observed near-perfect global and uniform calibration.

4.5 Forecast Evaluation

When comparing baseline forecasts with forecasts from hubs, our objective was to answer three questions:

1) How does the choice of a baseline and scoring rule affect the overall assessment of forecast models?
2) How does the choice of a scoring rule and baseline affect the ranking of forecast models?
3) When considering a single forecast, how does the choice of a baseline affects the evaluation of the model on different subsets, such as regions?

In total, we evaluated forecasts from 81 models for influenza and 57 models for COVID-19 using the scaled relative skill score. For each baseline model, we added it to the pool of forecasting models and computed the scaled relative skill score once based on the WIS on the original scale and once applied to log-transformed forecasts. This gives a total of 38 ranked list of forecasting models. If all baseline models were perfectly calibrated but varied in sharpness, we would expect the overall level in relative skills to vary across the 19 lists, but the ranking to be similar.

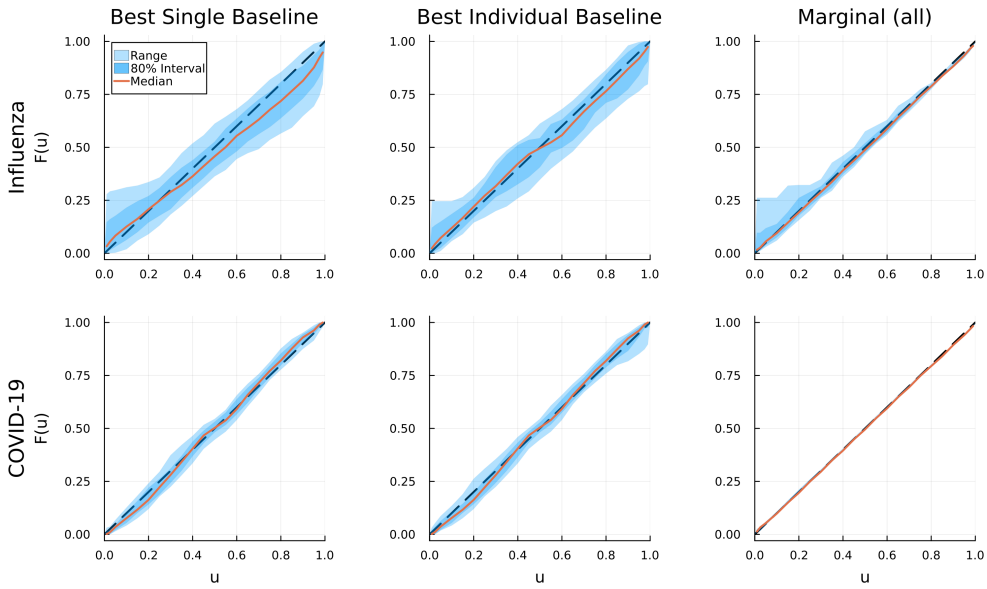

Fig. 3 Uniform calibration plot: Pointwise 80% interval (light blue), range (dark blue) and median (orange) of PIT functions across all states

Figure 4 shows histograms of SRSS values across forecasting models, split by disease and scoring rule. The comparison focused on models evaluated against either the best or worst calibrated baseline.

For influenza, the results show that baseline selection strongly influenced how many models outperformed the baseline. Using the WIS, 8 out of 81 models outperformed the best calibrated baseline, while 75 models outperformed the worst calibrated one. When using the lWIS 24 models outperformed the best-calibrated baseline, whereas 80 beat the worst calibrated baseline. For COVID-19 under the WIS, all 57 models outperformed the best calibrated baseline and 52 models outperformed the worst calibrated one. This underlines that calibration and forecast performance do not necessarily align. When using the lWIS, 14 models outperformed the best calibrated baseline and 34 outperformed the worst calibrated one.

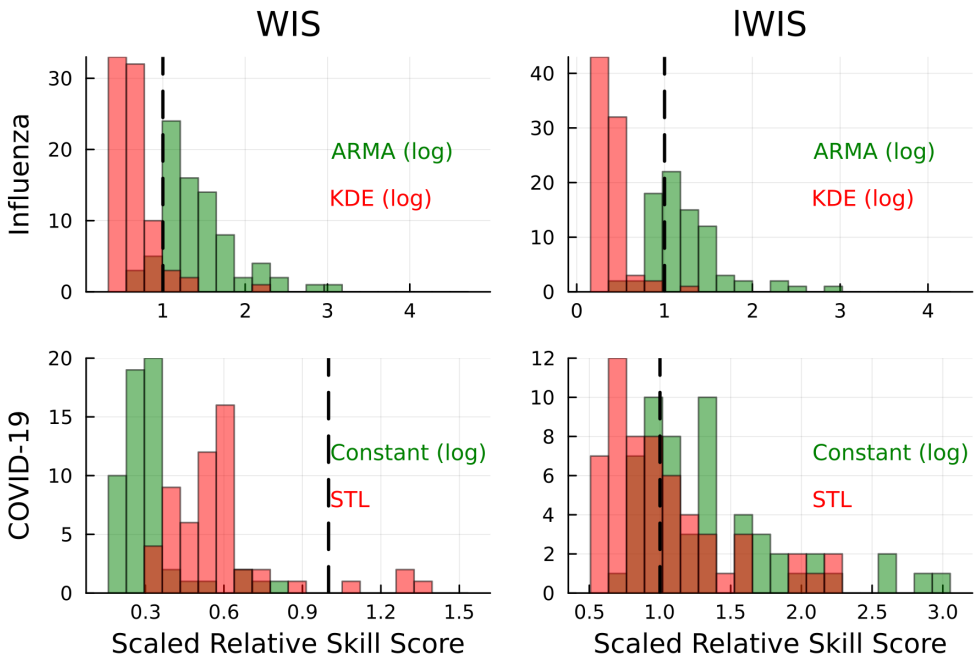

Fig. 4 Scaled Relative Skill Scores for best (green) and worst (red) calibrated baselines - Split by disease and evaluation metric

The choice of baseline model can influence not only the overall assessment of forecast models but also their relative performance rankings. Figure B3 in the appendix displays how forecast models rank under different baseline choices, ordered by the baselines' calibration quality. If all baselines were perfectly calibrated, we would expect only minor differences between rankings. For influenza, the best and worst performing forecast models maintained their positions regardless of the baseline choice for both scoring rules, with one exception: the IDS-l baseline under the WIS produced substantially different rankings. While there were differences across all calibration levels, most baselines agreed on the top performers. The seven highest-ranked models under the WIS (or eight under the lWIS) changed by no more than one position across all baselines. For COVID-19 under the WIS, two baseline models—OLS-l and IDS-l—produced markedly different rankings. These baselines ranked second and 16th respectively in terms of global calibration. Hence, good calibration does not ensure alignment of rankings. However, under the lWIS, forecast model rankings remained consistent across all baselines. Figure 5 summarises these ranking differences by Kendall's tau for each pair of rank lists. Correlations were capped at 0.9 for clearer visualisation, affecting only 10% of pairs of the bottom left panel. For influenza under the WIS, the IDS-l baseline clearly stood out, confirming the pattern seen in Figure B3. The OLS-l baseline also differed from others, whilst the six least calibrated baselines (the marginal, LSD and KDE models and their log-transformed versions) produced similar rankings. For COVID-19 under the WIS, there was less agreement in rankings compared to

influenza. Again, the IDS-l baseline showed the least agreement with others. Notably, the two best calibrated baselines, the constant and OLS models in logs, produced different rankings. The largest ranking difference involved the MUNI-VAR model, which ranked 45th under the ARMA-l baseline but 31st under the OLS-l baseline. Given that forecasts were submitted during only eight weeks at the end of 2021 when COVID-19 infections increased exponentially, the OLS-l baseline may have been more sensitive to sudden, sharp increases, making the MUNI-VAR model appear more favourable. When using the lWIS as the scoring rule, rankings showed greater consistency across baseline models for both influenza and COVID-19. Therefore, while baseline choice can affect forecast model rankings, models with few submitted forecasts have a substantial influence on rankings. Using alternative scoring rules such as the lWIS can make rankings more robust.

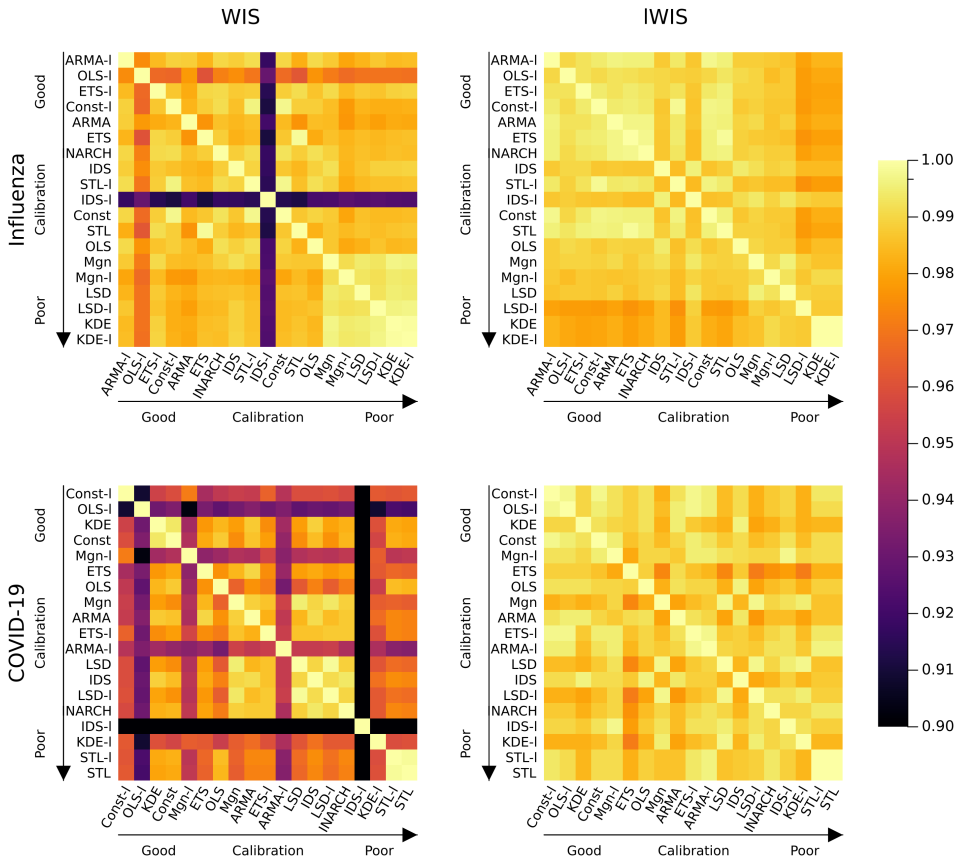

Fig. 5 Kendall's tau correlation heatmap for influenza and COVID-19 based on WIS and lWIS - Correlations are truncated at 0.95 (< 15% affected)

We compared the forecast performance of a single forecast model across individual states. For this analysis, we selected the FluSight-Ensemble and the COVIDhub-Ensemble for influenza and COVID-19 respectively, as these models have good coverage across states and time. We aimed to answer whether the choice of baseline affects our evaluation of where these forecast models perform well. Beyond good global calibration, we assumed that good uniform calibration with respect to state partitioning is necessary to address this question. Figure B4 in the appendix summarises the uniform calibration of all baseline models for influenza and COVID-19, alongside their global calibration. Overall, baseline models with good global calibration were also calibrated well uniformly. We compared the forecasts of the two forecasting models with each baseline's forecast using the relative WIS and lWIS. Computing the scaled relative skill score was not necessary due to equal coverage across states and time. The relative WIS and lWIS were then used to rank states by the forecast model's performance. These ranks were again compared using Kendall's tau correlation matrices. For reference, we included not only the 19 baselines as before, but also the baseline based on individually best calibrated models and the marginal baseline using future data. Figure 6 summarises all correlation pairs, with baselines sorted by their uniform calibration. Negative correlations were truncated at zero. For influenza, there was consensus amongst the six best (uniformly) calibrated baselines, with strong agreement between the two additional baselines for both scoring rules. For baselines with poorer calibration, there was less agreement amongst themselves and with well-calibrated baselines. For COVID-19, the two additional baselines were not the best calibrated. The constant model (in logs) had better uniform calibration than the baseline with state-wise best calibrated models. This likely occurred because the constant baseline was best-calibrated in most states, as shown in Figure 2. For both scoring rules, there was strong agreement between the marginal model using all observations and the constant baseline. There was less agreement between other baselines and either of the two best calibrated models. In contrast to the comparison of forecast models, the effect of baseline selection remains strong when using the lWIS as the scoring rule. For both scoring rules and both diseases, the best uniformly calibrated valid baseline produced a ranking similar to the (invalid) marginal baseline.

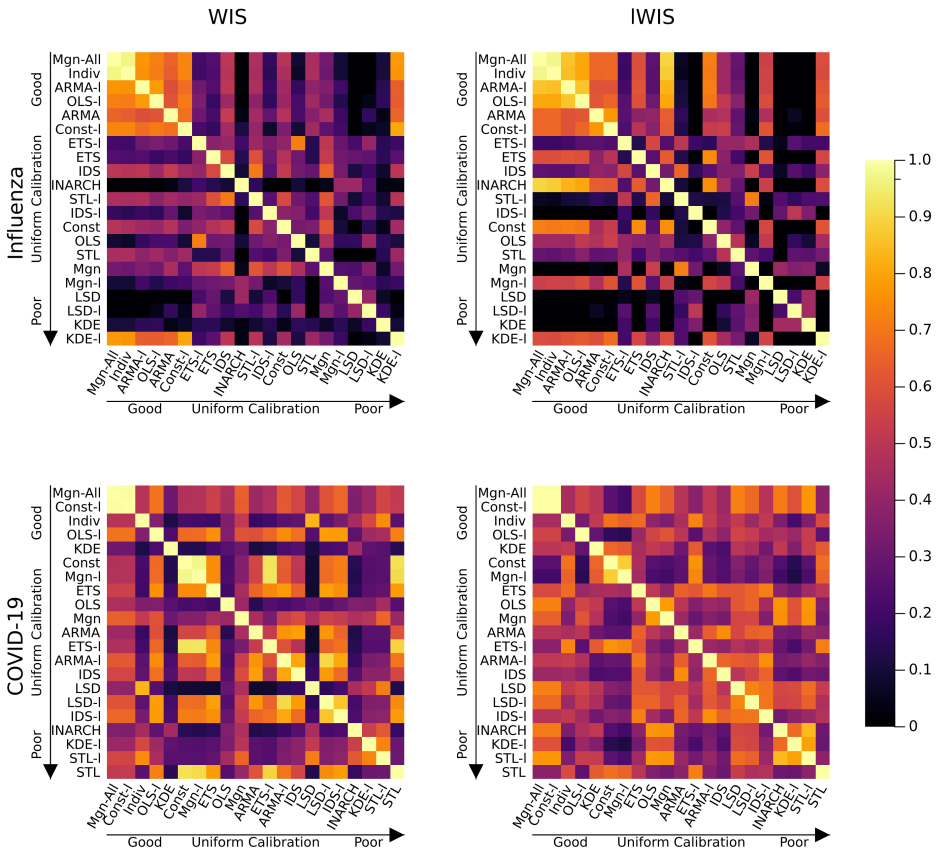

Fig. 6 Kendall's tau correlation heatmap for location ranking - evaluated for FluSight-Ensemble (influenza) and CH-Ensemble (COVID-19). Correlations were truncated at zero.

5 Discussion

This study provides recommendations for baseline model selection in epidemiological forecasting, addressing a critical but often overlooked aspect of forecast evaluation. Our analysis revealed that baseline choice profoundly influences both the assessment of individual forecast models and their relative rankings.

Our findings demonstrate that selecting an appropriate baseline model is far from trivial. Among the ten baseline frameworks evaluated, no single model satisfied all recommended criteria across both influenza and COVID-19 applications. The best globally calibrated baseline for influenza (ARMA model on log-transformed data) differed from that for COVID-19 (constant model on log-transformed data), highlighting that baseline selection is application-specific. We found that calibration and forecast performance do not necessarily align, a well-calibrated baseline may perform poorly

in terms of forecast accuracy, while a poorly calibrated baseline may inadvertently favour certain forecasting models.

The impact of baseline selection extends beyond simple performance metrics. When evaluating multiple forecasting models, the choice of a baseline model can alter conclusions about which models are successful. For influenza, the number of models outperforming the baseline ranged from 8 to 75 depending on the baseline selected. Variations stem partly from differences in model coverage across spatial and temporal domains, but also reflect fundamental differences in how baselines capture the underlying data characteristics.

Our analysis of uniform calibration revealed another layer of complexity. A baseline that appears well-calibrated globally may exhibit poor calibration across specific subsets such as individual states or forecast horizons. In our application, however, global and uniform calibration was similar. Heterogeneity in calibration can lead to biased assessments when comparing model performance across different regions or time periods. Such biases have direct practical consequences: if a baseline systematically underperforms in certain regions, forecast models in those areas may appear artificially superior. This potentially leads to resource allocation for prevention and preparedness that is based on a false impression of certainty in forecasts. The finding that regions had different best calibrated baseline models, especially for influenza, suggests that a one-size-fits-all approach may be inadequate for fair evaluation.

From a practical standpoint, we recommend that researchers follow a structured process when selecting baseline models. First, identify the key characteristics of the time series being forecast, including seasonality, growth patterns, and data availability. Second, ensure format compatibility between the baseline and forecast models. Third, verify that the baseline uses only appropriate training data. Fourth, prioritise simplicity to maintain interpretability. Finally, and most importantly, assess both global and uniform calibration across relevant data partitions.

When a single well-calibrated baseline cannot be identified, several strategies can improve evaluation fairness. Regional or temporal stratification of baselines may provide better uniform calibration, though at the cost of increased complexity. Alternatively, researchers might consider using multiple baselines to assess the robustness of their conclusions, or employ ensemble baselines that combine simple models. The choice of scoring rule also matters; our results suggest that log-scale evaluation (lWIS) often produces more stable rankings across different baselines than evaluation on the original scale.

This work has several limitations that suggest directions for future research. First, our analysis focused on epidemiological time series, and the generalisability to other forecasting domains remains unclear. Different fields may have distinct requirements for baseline models based on their specific data characteristics and evaluation needs. Second, we restricted our analysis to relatively simple baseline models. While this aligns with the principle that baselines should be interpretable and easy to implement, more complex baselines might achieve better calibration in some settings.

Future research could explore automated procedures for baseline selection, potentially using cross-validation or other model selection techniques to identify baselines

that achieve good calibration. The development of diagnostic tools to assess baseline suitability before conducting forecast evaluations would be valuable. Furthermore, the relationship between baseline selection and the predictability of time series is an interesting direction. The forecast performance of a suitable and simple baseline gives insight into how much of the variation in a time series can be explained by a simple model.

The challenge of baseline selection reflects broader issues in forecast evaluation. Just as human judgment serves as an informal baseline in many decision-making contexts, statistical baselines provide a reference point for assessing the value of computational forecasting methods. Our results suggest that this reference point must be chosen carefully, with attention to both statistical properties and the specific evaluation context. At the same time, systematic collection of human judgement forecasts could enable their use as an additional baseline (Bosse et al, 2022, 2024). As collaborative forecasting efforts continue to grow, establishing standards for baseline selection will become increasingly important for ensuring fair, meaningful, and reproducible model comparisons.

In conclusion, baseline model selection deserves the same careful consideration given to forecast model development. By providing tools to assess baseline calibration and demonstrating the consequences of different choices, this work aims to improve the rigour and fairness of forecast evaluation in epidemiology and beyond. The lack of a universally optimal baseline underscores the need for thoughtful, context-specific selection guided by clear principles and thorough evaluation.

Acknowledgments

The authors would like to thank the *epiforecast* team members for their valuable contributions to the literature review and project conceptualisation in the early stages that formed the foundation of this research.
Funding: This work was supported by funds from Wellcome (210758/Z/18/Z) and the National Institute for Health and Care Research (NIHR) Health Protection Research Unit (HPRU) in Health Analytics and Modelling.
Use of AI: The authors acknowledge the use of Claude Sonnet 4 (Anthropic) for formatting and proof reading. All AI-generated content was thoroughly reviewed, fact-checked, and revised by the authors. The final analysis, conclusions, and academic interpretations are entirely the authors' own work.

Declarations

Conflict of interest: The authors declare no conflict of interest.
Availability: All data is publicly available and summarised at https://doi.org/10.5281/zenodo.16407890. Application code is available at https://github.com/ManuelStapper/Baselines_Application.
Contributions: S.F. conceived the study, S.F. and M.S. developed the methodology and study design, M.S. implemented the software, conducted the analysis and wrote the initial draft. S.F. and M.S. revised the manuscript. S.F. provided supervision. Both authors reviewed and approved the final version.

Appendix A Data availability

Model	States (52)	Availability 21/22 (34)	22/23 (32)	23/24 (31)	24/25 (6)	Model	States (52)	Availability 21/22 (34)	22/23 (32)	23/24 (31)	24/25 (6)
CADPH	1	0	31	29	6	META	1	0	0	0	5
CEID	52	21	10	0	0	MI-J	52	0	0	0	5
CEPH	52	0	28	30	6	MI-E	52	0	30	30	6
CFA-F	52	0	0	28	0	MOBS	52	20	31	30	5
CFA-R	52	0	0	30	0	NEU-A	52	0	0	0	5
CMU	52	24	31	28	6	NEU-F	51	0	0	0	4
CU	52	24	26	30	6	NIH	51	0	28	26	6
FH	52	0	0	30	4	NU	52	0	0	23	0
FS-BL	52	24	32	30	6	OHT	52	0	0	0	6
FS-ENS	52	24	31	30	6	PSI-D	52	24	31	0	0
FS-LOP	52	0	0	30	6	PSI-P	52	0	0	30	6
GT-PT	51	0	0	0	6	PSI-PB	52	0	0	20	4
GT-PB	51	0	0	0	6	SG-R	52	20	30	29	0
GH-FS	52	21	0	0	0	SG-S	52	24	0	0	0
GH-MOD	52	0	0	9	0	SIG-C	51	24	31	30	0
GT-FNP	51	24	29	27	0	SIG-T	51	24	31	30	6
IEM	52	24	0	0	0	STV-GBR	52	0	0	29	0
ISUN-E	52	0	0	18	0	STV-ILI	52	0	0	0	6
ISUN-F	52	0	20	0	0	UGA-CEID	52	0	0	0	4
ISUN-G	52	0	0	0	5	UGA-CC	52	0	0	26	3
ISUN-N	52	0	0	25	0	UGA-I	52	0	0	27	6
ISUN-S	52	0	0	18	0	UGA-O	52	0	30	3	0
JHU-D	52	0	0	0	6	UGA-S	52	0	0	0	4
JHU-G	52	6	0	0	0	UG-CC	52	0	0	28	4
JHU-C	52	0	0	17	4	UG-FP	1	0	13	0	0
JHU-I	52	0	26	0	0	UGENS	1	0	0	27	1
LA	52	24	23	28	3	UI	52	0	0	0	4
LU-C	52	0	0	30	6	UM	52	0	0	30	6
LU-ER	52	0	9	0	0	UM-AR2	52	0	0	0	6
LU-EH	6	0	6	0	0	UM-ARIMA	1	9	0	0	0
LU-M	52	0	14	0	0	UM-F	52	0	0	30	6
LU-HJ	52	15	17	0	0	UM-G	52	0	9	0	0
LU-HW	52	0	12	0	0	UM-T	52	22	31	31	0
LU-KF	52	0	11	0	0	UNC	52	0	25	24	4
LU-ST	52	0	8	0	0	UT	52	24	0	0	0
LU-TEVA	52	22	0	0	0	UVA-E	52	24	30	29	6
LU-V2	52	23	0	0	0	VTS-ENS	52	0	0	26	0
LU-V2K	52	22	0	0	0	VTS-EXOG	52	18	25	0	0
LU-V2K-C	52	22	0	0	0	VTS-P	52	0	0	0	6
LU-V2-C	52	23	0	0	0	VTS-T	52	0	5	0	0
MDPRED	1	0	0	0	6						

Table A1 Availability of forecasts for influenza

Model	Availability States (52)	Availability Dates (152)	Model	Availability States (52)	Availability Dates (152)	Model	Availability States (52)	Availability Dates (152)
AMM	51	20	IEM-MED	52	68	MUNI-ARIMA	52	89
BPag	52	123	IHME-Curve	52	22	MUNI-VAR	52	8
CEID	52	93	IQVIA	50	2	DeepSTIA	52	98
CH-4wk	52	135	IUPUI	51	10	OQN	52	14
CH-base	52	152	ISLW-STEM	50	42	PR-UMD	51	21
CH-ens	52	135	JBUD-HMXK	1	10	QJHong	1	86
CH-tens	52	96	JCB-PRM	52	9	RW-ESG	52	117
CH-CDCens	52	75	JHU-Bucky	52	105	SDSC-ISG	52	52
CU-noch	52	132	JHU-ICATTML	52	3	SigSci	52	14
CU-scenH	52	51	JHU-CSSE	52	72	UCF-AEM	1	97
CU-scenL	52	130	JHU-IDD	52	114	UCLA	52	67
CU-scenM	52	128	KITens	52	83	UC-C	52	49
CU-sel	52	132	Karlen	52	106	UMich	51	70
Col-SC	1	55	LANL	52	65	USACE	52	18
C19Sim	52	56	LNQ-ens1	52	48	USC	52	117
CA-DELPHI	52	103	LU	52	31	USC-RF	52	2
DDS-NBDS	52	35	MIT-Cass	49	82	UVA-Ens	52	96
FD-Sweight	1	37	MIT-ISO	51	107	SU-GRU	1	20
FRBSF	1	51	MOBS	51	42	prolix	1	18

Table A2 Availability of forecasts for COVID-19

Appendix B Results

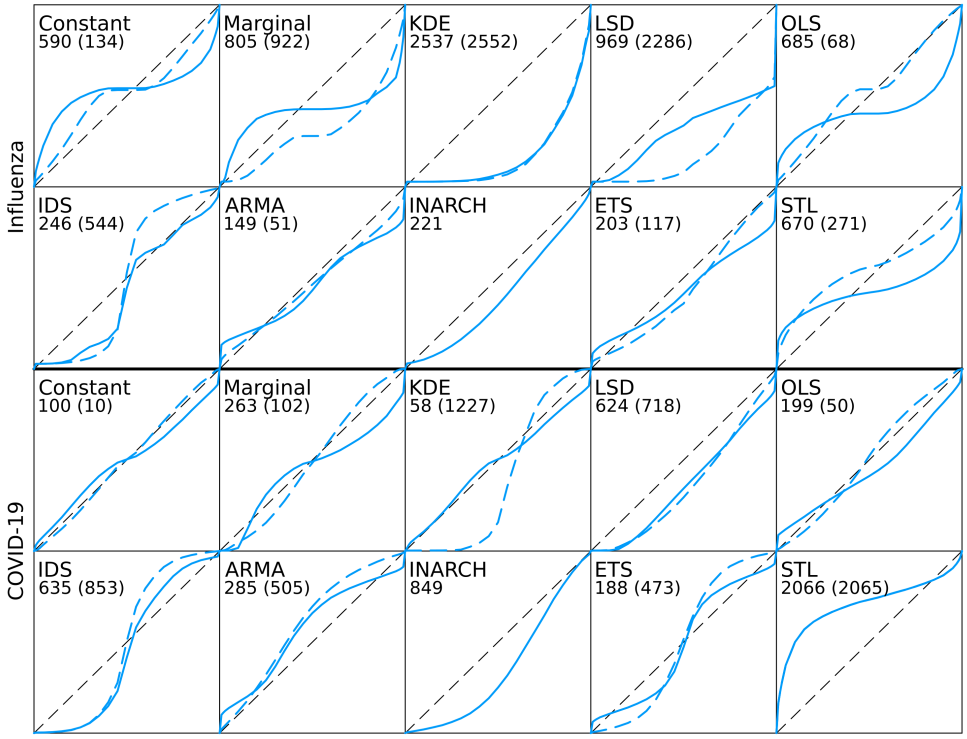

Fig. B1 PIT functions for all baselines on original time series (solid) and log-transformed time series (dashed) - CvM divergence for baselines fitted to original time series and log-transformed time series in parentheses

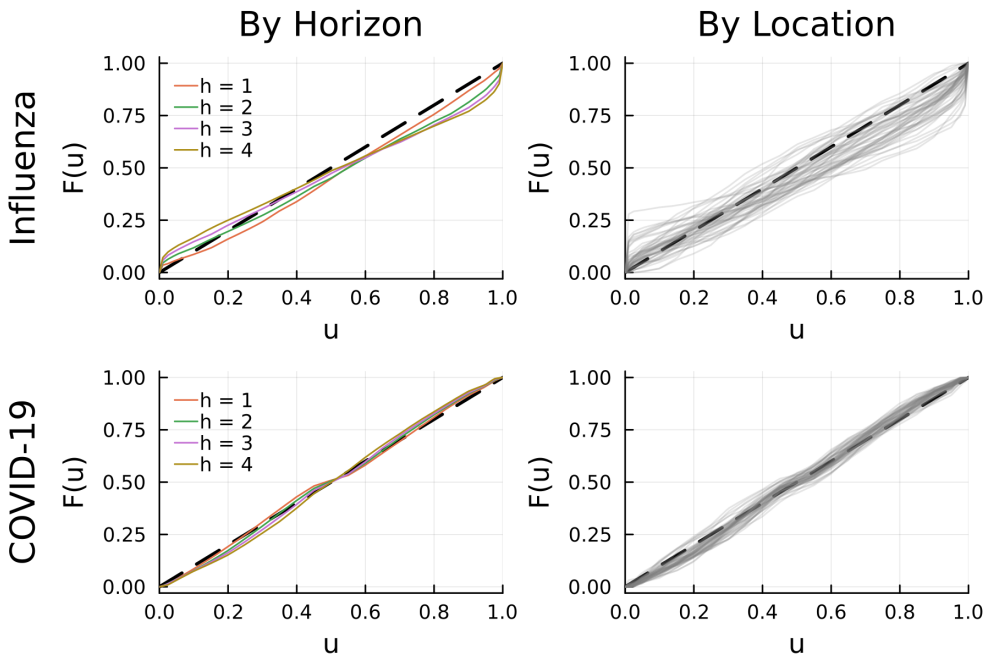

Fig. B2 PIT functions for best calibrated baseline - Split by location and horizon.

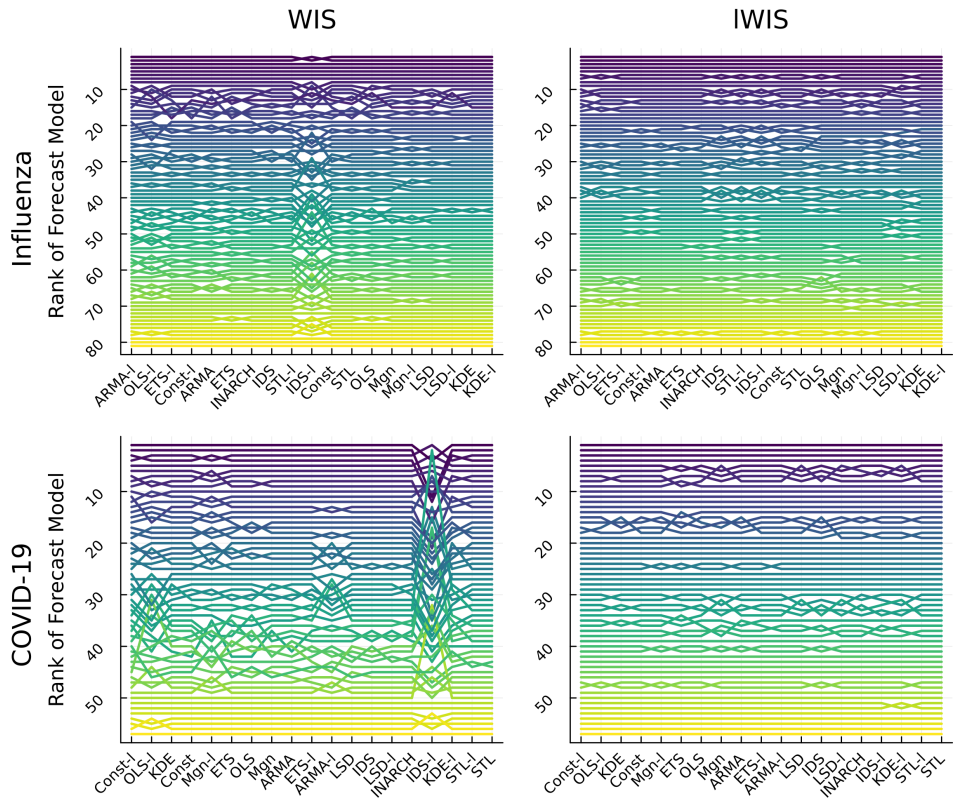

Fig. B3 Ranks of forecast models for different baselines - each line represents one forecast model

References

Bosse NI, Abbott S, Bracher J, et al (2022) Comparing human and model-based forecasts of covid-19 in germany and poland. PLOS Computational Biology 18(9):e1010405. https://doi.org/10.1371/journal.pcbi.1010405, URL http://dx.doi.org/10.1371/journal.pcbi.1010405

Bosse NI, Abbott S, Bracher J, et al (2024) Human judgement forecasting of covid-19 in the uk. Wellcome Open Research 8:416. https://doi.org/10.12688/wellcomeopenres.19380.2, URL http://dx.doi.org/10.12688/wellcomeopenres.19380.2

Bracher J, Ray EL, Gneiting T, et al (2021a) Evaluating epidemic forecasts in an interval format. PLOS Computational Biology 17(2):1–15. https://doi.org/10.1371/journal.pcbi.1008618, URL https://doi.org/10.1371/journal.pcbi.1008618

Bracher J, Wolffram D, Deuschel J, et al (2021b) A pre-registered short-term forecasting study of COVID-19 in Germany and Poland during the second wave.

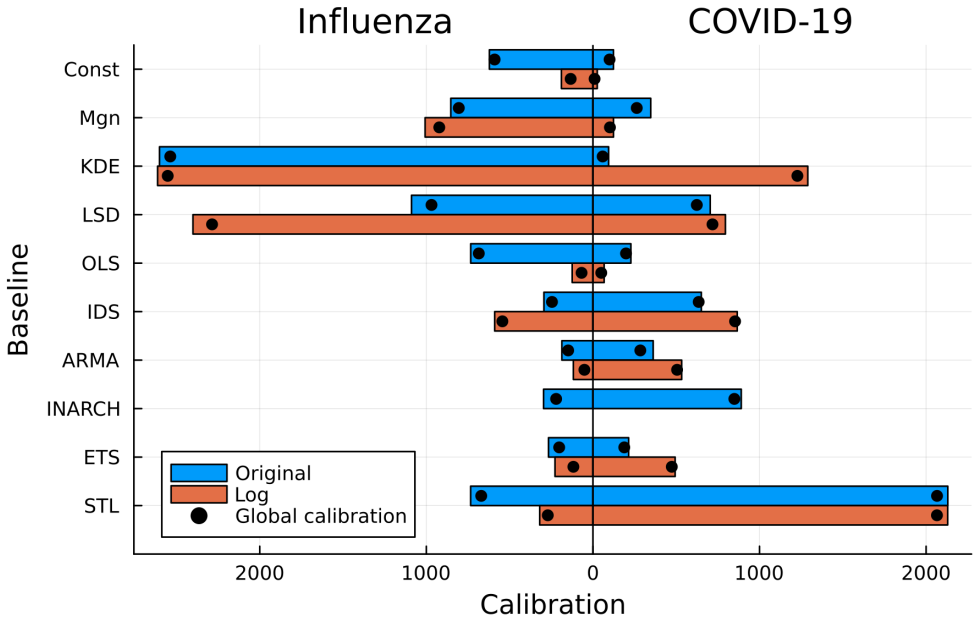

Fig. B4 Global and uniform calibration of baseline models for forecasts on the original scale (blue) or log-transformed cases (orange)

Nature Communications 12(1):5173. https://doi.org/10.1038/s41467-021-25207-0, URL https://www.nature.com/articles/s41467-021-25207-0

CDC Influenza Forecasting Team (2021) Flusight forecast data archive. URL https://github.com/cdcepi/Flusight-forecast-data, accessed: 2025-01-15

CDC Influenza Forecasting Team (2025) Flusight forecast hub. URL https://github.com/cdcepi/FluSight-forecast-hub, accessed: 2025-01-15

Cleveland RB, Cleveland WS, McRae JE, et al (1990) Stl: A seasonal-trend decomposition. J off Stat 6(1):3–73

Colón-González FJ, Soares Bastos L, Hofmann B, et al (2021) Probabilistic seasonal dengue forecasting in Vietnam: A modelling study using superensembles. PLOS Medicine 18(3):e1003542. https://doi.org/10.1371/journal.pmed.1003542, URL https://dx.plos.org/10.1371/journal.pmed.1003542

Cramer EY, Ray EL, Lopez VK, et al (2022) Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. Proceedings of the National Academy of Sciences 119(15):e2113561119. https://doi.org/10.1073/pnas.2113561119, URL https://pnas.org/doi/10.1073/pnas.2113561119

Czado C, Gneiting T, Held L (2009) Predictive model assessment for count data. Biometrics 65(4):1254–1261. https://doi.org/10.1111/j.1541-0420.2009.01191.x, URL https://doi.org/10.1111/j.1541-0420.2009.01191.x

Funk S, Camacho A, Kucharski AJ, et al (2019) Assessing the performance of real-time epidemic forecasts: A case study of Ebola in the Western Area region of Sierra Leone, 2014-15. PLOS Computational Biology 15(2):e1006785. https://doi.org/10.1371/journal.pcbi.1006785, URL https://dx.plos.org/10.1371/journal.pcbi.1006785

Funk S, Abbott S, Atkins B, et al (2020) Short-term forecasts to inform the response to the Covid-19 epidemic in the UK. https://doi.org/10.1101/2020.11.11.20220962, URL http://medrxiv.org/lookup/doi/10.1101/2020.11.11.20220962

Gneiting T, Larson K, Westrick K, et al (2006) Calibrated Probabilistic Forecasting at the Stateline Wind Energy Center: The Regime-Switching Space–Time Method. Journal of the American Statistical Association 101(475):968–979. https://doi.org/10.1198/016214506000000456, URL http://www.tandfonline.com/doi/abs/10.1198/016214506000000456

Gneiting T, Balabdaoui F, Raftery AE (2007) Probabilistic Forecasts, Calibration and Sharpness. Journal of the Royal Statistical Society Series B: Statistical Methodology 69(2):243–268. https://doi.org/10.1111/j.1467-9868.2007.00587.x, URL https://academic.oup.com/jrsssb/article/69/2/243/7109375

Hyndman R, Athanasopoulos G (2018) Forecasting: Principles and Practice, 2nd edn. OTexts, Australia

Hyndman RJ, Khandakar Y (2008) Automatic time series forecasting: Theforecast-Package forr. J Stat Softw 27(3)

Johansson MA, Apfeldorf KM, Dobson S, et al (2019) An open challenge to advance probabilistic forecasting for dengue epidemics. Proceedings of the National Academy of Sciences 116(48):24268–24274. https://doi.org/10.1073/pnas.1909865116, URL https://pnas.org/doi/full/10.1073/pnas.1909865116

Kandula S, Shaman J (2019) Near-term forecasts of influenza-like illness. Epidemics 27:41–51. https://doi.org/10.1016/j.epidem.2019.01.002, URL https://linkinghub.elsevier.com/retrieve/pii/S1755436518301336

Krüger F (2017) Survey-based forecast distributions for Euro Area growth and inflation: ensembles versus histograms. Empirical Economics 53(1):235–246. https://doi.org/10.1007/s00181-017-1228-3, URL http://link.springer.com/10.1007/s00181-017-1228-3

Lauer SA, Sakrejda K, Ray EL, et al (2018) Prospective forecasts of annual dengue hemorrhagic fever incidence in Thailand, 2010–2014. Proceedings of the National Academy of Sciences 115(10). https://doi.org/10.1073/pnas.1714457115,

URL https://pnas.org/doi/full/10.1073/pnas.1714457115

Lerch S, Thorarinsdottir TL, Ravazzolo F, et al (2017) Forecaster's Dilemma: Extreme Events and Forecast Evaluation. Statistical Science 32(1). https://doi.org/10.1214/16-STS588, URL https://projecteuclid.org/journals/statistical-science/volume-32/issue-1/Forecasters-Dilemma-Extreme-Events-and-Forecast-Evaluation/10.1214/16-STS588.full

Lewnard JA, Jirmanus L, Júnior NN, et al (2014) Forecasting Temporal Dynamics of Cutaneous Leishmaniasis in Northeast Brazil. PLoS Neglected Tropical Diseases 8(10):e3283. https://doi.org/10.1371/journal.pntd.0003283, URL https://dx.plos.org/10.1371/journal.pntd.0003283

McGough SF, Brownstein JS, Hawkins JB, et al (2017) Forecasting Zika Incidence in the 2016 Latin America Outbreak Combining Traditional Disease Surveillance with Search, Social Media, and News Report Data. PLOS Neglected Tropical Diseases 11(1):e0005295. https://doi.org/10.1371/journal.pntd.0005295, URL https://dx.plos.org/10.1371/journal.pntd.0005295

Meakin S, Abbott S, Bosse N, et al (2022) Comparative assessment of methods for short-term forecasts of COVID-19 hospital admissions in England at the local level. BMC Medicine 20(1):86. https://doi.org/10.1186/s12916-022-02271-x, URL https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-022-02271-x

Morris M, Hayes P, Cox IJ, et al (2023) Neural network models for influenza forecasting with associated uncertainty using Web search activity trends. PLOS Computational Biology 19(8):1–23. https://doi.org/10.1371/journal.pcbi.1011392, URL https://doi.org/10.1371/journal.pcbi.1011392, publisher: Public Library of Science

Petropoulos F, Makridakis S (2020) Forecasting the novel coronavirus COVID-19. PLOS ONE 15(3):e0231236. https://doi.org/10.1371/journal.pone.0231236, URL https://dx.plos.org/10.1371/journal.pone.0231236

Ray EL, Reich NG (2018) Prediction of infectious disease epidemics via weighted density ensembles. PLOS Computational Biology 14(2):e1005910. https://doi.org/10.1371/journal.pcbi.1005910, URL https://dx.plos.org/10.1371/journal.pcbi.1005910

Ray EL, Wang Y, Wolfinger RD, et al (2025) Flusion: Integrating multiple data sources for accurate influenza predictions. Epidemics 50:100810. https://doi.org/https://doi.org/10.1016/j.epidem.2024.100810, URL https://www.sciencedirect.com/science/article/pii/S1755436524000719

Reich NG, Brooks LC, Fox SJ, et al (2019) A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. Proceedings of the National Academy of Sciences 116(8):3146–3154. https://doi.org/10.1073/pnas.1812594116, URL https://pnas.org/doi/full/10.1073/pnas.1812594116

Reich Lab (2025a) Covid-19 forecast hub. URL https://github.com/reichlab/covid19-forecast-hub, accessed: 2025-01-15

Reich Lab (2025b) covidhubutils: Utility functions for the covid-19 forecast hub. URL https://github.com/reichlab/covidHubUtils, accessed: 2025-01-15

Viboud C, Sun K, Gaffey R, et al (2018) The RAPIDD ebola forecasting challenge: Synthesis and lessons learnt. Epidemics 22:13–21. https://doi.org/10.1016/j.epidem.2017.08.002, URL https://linkinghub.elsevier.com/retrieve/pii/S1755436517301275

Wang H, Kwok KO, Riley S (2024) Forecasting influenza incidence as an ordinal variable using machine learning. Wellcome Open Research 9:11. https://doi.org/10.12688/wellcomeopenres.19599.1, URL https://wellcomeopenresearch.org/articles/9-11/v1